THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

DATABASE SYSTEMS JOURNAL

Vol. XIII/2022

LISTED IN

RePEc, EBSCO, DOAJ, Open J-Gate,
Cabell's Directories of Publishing Opportunities,
Index Copernicus, Google Scholar,
Directory of Science, Cite Factor,
Electronic Journals Library

BIG DATA

NoSQL

DATA SCIENCE

MACHINE LEARNING

BUSINESS INTELLIGENCE

CLOUD COMPUTING

DATA MINING

DATA WAREHOUSES

DATABASES

ISSN: 2069 – 3230
dbjournal.ro

Database Systems Journal BOARD

Contact
CaleaDorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: https://dbjournal.ro/
E-mail: editordbjournal@gmail.com

CONTENTS

Analysis of Romanian Air Quality using Machine Learning Techniques

Andreea-Mihaela NICULAE
The Bucharest University of Economic Studies, Romania
niculaeandreea17@stud.ase.ro

Air quality monitoring has become an increasingly important subject and is one of the most important concerns of governments worldwide. Monitoring is especially important in industrial and urban areas. Due to the many forms of pollution generated mainly by fuel consumption, means of transport, coal-fired electricity generation, etc., air quality is negatively affected. As the current trend is an increase in air pollution, it is necessary to install equipment to measure air quality both in areas with a high risk of pollution and in areas where pollution is low. These types of equipment must communicate in real-time their measured values, which then can be accessed to be able to make analyzes and predictions regarding air quality in a certain geographical area, areas with a high industrialization level, or in areas with a growing population. This paper aims to investigate the application of big data and machine learning techniques to make predictions on air quality using, as a source of data, data recorded in the period 2018-2021 from measurement probes throughout Romania for PM_{10}, NO_2, O_3, and SO_2. The results of this paper's analysis show that time-series models outperform traditional models. Moreover, ANN models are successful only in classifying pollutants' AQI levels and not their actual values.

***Keywords**: Big Data, Machine Learning, Romania, Air quality, MLR, SARIMA, C5.0, Random Forest, ANN*

1 Introduction

In recent years, the world's fast development has been both beneficial and harmful to its population. While this development aims to help people live a better life, it also seems to reduce the quality of one's life, by creating problems such as global warming and air pollution. The latest one is a subject that more governments should pay attention to, as it is dangerous to the entire world's health.

According to the World Health Organization, 9 out of 10 people around the globe, live in a place where the air quality exceeds their guideline limits [1]. This means that more than 90% of the world breathes air (along with the pollutants inside it) that harms the body, exposing it to different diseases, which affect both the lungs and other organs.

With such possible side effects of air pollution, it is necessary for most governments to try and lower the damages caused by pollution, or focus on lowering the pollution itself. The latest is of growing importance nowadays, as it has more long-term benefits. In addition, the increase in data sources around the globe makes it easier to analyze air quality and provide accurate models to predict and fight pollution.

Monitors around the world provide information on overall air quality: pollutants and AQI. AQI is an index computed as the maximum of all individual pollutants values of AQI. In Europe, AQI takes values from one to six, where one means that air quality is Good, and six means that air quality is Extremely Poor. To compute AQI, the following formula is used:

$$AQI = \max(AQI_{PM_{2.5}}, AQI_{PM_{10}}, \dots, AQI_{pollutant})$$

The European Environment Agency decided on the concentration intervals of each pollutant to help compute their AQI [2]. The EEA updates these values each year. In *Table 1* are the centralized values for the 2022 pollutants AQI intervals, with the colors according to the EEA regulations.

Forecasting air quality is a complex subject that contains algorithms, techniques, and methods from numerous topics: big data, machine learning, time series, and others.

This paper focuses on some available models used in forecasting air pollution, such as MLR, ARMA, ARIMA, and SARIMA, decision trees (ID3, C4.5, and C5.0), random forest, ANN, while also presenting other popular models used by other papers for predicting air quality. In the end, the practical part of the paper contains an analysis of Romanian air quality data.

Table 1. European AQI and concentration range for each pollutant (AQI sub-index)

AQI Level	AQI	Index level (based on pollutant concentrations in µg/m^3)				
		PM_{10}	$PM_{2,5}$	NO_2	O_3	SO_2
Good	1	0-20	0-10	0-40	0-50	0-100
Fair	2	20-40	10-20	40-90	50-100	100-200
Moderate	3	40-50	20-25	90-120	100-130	200-350
Poor	4	50-100	25-50	120-230	130-240	350-500
Very Poor	5	100-150	50-75	230-340	240-380	500-750
Extremely Poor	6	150-1200	75-800	340-1000	380-800	750-1250

2. Literature Review

There are two main topics of concern regarding the analysis of air pollution: *monitoring air quality* and *forecasting air quality* [3]. Since more and more governments focus on reducing their country's pollution, more funds are available to facilitate better analyses, by opening more monitoring stations around the country and by offering relevant data sources for analysts to model.

A. Monitoring air quality

While this paper has the purpose to model air quality, it is vital to understand where the data comes from. The IoT makes it possible to monitor air pollution both using government stations (of high performance, but costly) or by using portable sensors (of decent performance, but much easier to obtain and at a low cost) [4]. These sensors monitor the most important pollutants, according to each country's needs and regulations. These are, most of the time: $PM_{2.5}$, PM_{10}, O_3, NO_2, and SO2. Using these, one can easily determine AQI and make further analyzes.

B. Air quality forecast

Forecasting air pollutants and thus, air quality itself, is a subject of increasing interest in recent years. Because air quality data presents itself mainly from sensors and monitoring stations, there is a high amount of data available to model, which makes Big Data models more favorable to use [5]. In Figure 1 are presented the most frequently used big data techniques to forecast air quality.

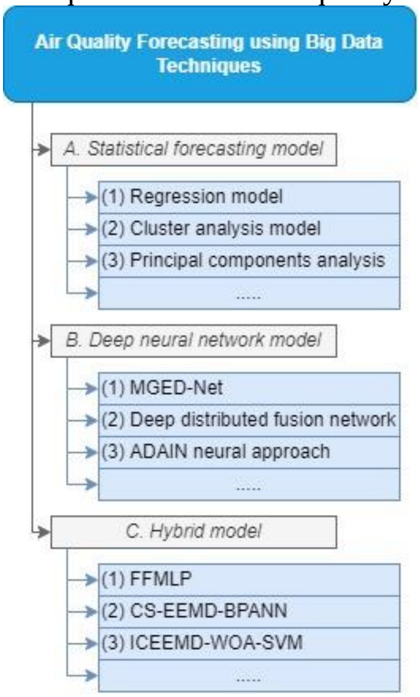

Air Quality Forecasting using Big Data Techniques
A. Statistical forecasting model
(1) Regression model
(2) Cluster analysis model
(3) Principal components analysis
.....
B. Deep neural network model
(1) MGED-Net
(2) Deep distributed fusion network
(3) ADAIN neural approach
.....
C. Hybrid model
(1) FFMLP
(2) CS-EEMD-BPANN
(3) ICEEMD-WOA-SVM
.....

Fig. 1. Air Quality Forecasting using Big Data Techniques [5]

Most models for predicting air quality are linear regression-based [3, 6]. These models assume the linearity of air quality data; however, as some studies suggest [4, 5], pollutants do not behave or evolve linearly, thus these models are not efficient in forecasts. Other widely used linear models are time series-based [7], such as ARMA, ARIMA, and SARIMA. The same linearity problem arises for these models, but even so, most of the time these models perform better than the MLR ones.

Deep neural network models are adaptive and work better with nonlinear data [6], being used more and more in recent papers to estimate and forecast air quality. They are easier to use in model building and their results are significantly better than the MLR model. However, studies show that researchers have a hard time deciding which model is superior between DNN and ARIMA [6, 8], as further analyzes and comparisons between the two are needed.

Another category of models that work well on air quality forecasting is decision trees based on: ID3 [9], C4.5 [10], and C5.0 [11]. One paper discovered that by using the newest and improved technique, C5.0, to model air quality in New Delhi, India, the author obtained a comparatively higher accuracy than other algorithms [11] seen in the respective paper's literature review.

Most of the hybrid, complex models used to forecast air quality start from the ANN model:

- FFMLP (Feed-Forward Multi-Layer Perceptron) [12];
- ANFIS (Adaptive neuro-fuzzy inference system) [13] - which combines the fuzzy logic with the ANN model;
- NARX (Nonlinear autoregressive model with external input) [13] – which combines the logic from the ARIMA series with the ANN model;
- CS-EEMD-BPANN (Cuckoo search - Ensemble Empirical Mode Decomposition - Back-propagation artificial neural networks).

In one of these papers, the author discovered that the RMSE obtained from models based on the ANFIS and NARX methodologies is much lower than the RMSE of traditional, non-complex models [13], especially when using them on big data sets.

Apart from all these supervised models, there are also unsupervised techniques, such as clustering and principal component analysis [14]. Some researchers have used clustering and PCA, for example, to see which meteorological variables were related to the concentration of air pollutants; or to classify monitoring locations - useful for optimizing the monitoring system, by efficiently placing sensors in the network. Some researchers have also used PCA to detect errors in air quality data [14].

3. Methodology

Before presenting the algorithms used in this paper's practical part, one must know about two concepts: Big Data and Machine Learning. These two are the quintessence of proper air pollution analysis, as their concepts cover all the important aspects necessary in an analysis.

A. Big Data analysis

The term *Big Data* refers to enormous data sets that, because of their exponential growth and complexity, are hard to be efficiently processed and used by traditional DBMS. In 2017, data sets were considered Big Data if they meet the “17 Vs” [15] (initially only 3 Vs, then 4, 5, 10, 14, and lately 17): **V**olume, **V**elocity, **V**alue, **V**ariety, **V**eracity, **V**isualization, **V**alidity, **V**olatility, **V**iscosity, **V**irality, **V**enue, **V**ariability, **V**ocabulary, **V**agueness, **V**erbosity, **V**oluntariness, **V**ersatility. Air quality data sets meet many of these characteristics, thus air pollution is a Big Data set. This is a very important part for further analysis, as Big Data technologies provide better, more accurate results, easy to apply (and preferable!) to these data sets [16].

B. Machine Learning

Considered the working horse of Big Data [17], *Machine Learning* is a branch of Artificial Intelligence characterized by the basic idea that working systems can learn from the available data, identify patterns in them, and make decisions, simulating the activity of the human mind.

There are three Machine Learning techniques [17]: *supervised learning*, *unsupervised learning*, and *semi-supervised learning*. Supervised learning distinguishes itself by the fact that the user already knows the result they are trying to obtain, whereas in unsupervised learning techniques the desired result is not so clear from the data. The most widely used unsupervised learning algorithms are clustering, principal component analysis, factor analysis, and even some kind of artificial neural networks [18]. The most used supervised learning algorithms are regressions (MLR, PLS, PCR and GLM), algorithms based on decision trees (ID3, C4.5, C5.0 and Random Forest), support vector machine and artificial neural networks [18, 19].

Machine Learning has a key element: its usage assumes the division of available data into two sets, one for training the model and one for testing the obtained model. This is very useful for obtaining efficient models, which accurately depict the studied phenomena. Moreover, by doing this, machine learning combats an occurring problem seen in modeling air quality data, *overfitting* [5].

C. Multiple Linear Regression (MLR)

The MLR algorithm used to model a linear relationship between a dependent variable, called the response, and multiple independent variables, called predictors, uses the following mathematical equation [20]:

$$y = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + \varepsilon$$

Where b_k is the regression coefficient, x_k is the predictor, y is the response, i takes values from one to k, and ε is the model's error.

As mentioned in the literature review, this model assumes that it uses linear data, which is not always the case with air pollution data sets. MLR is one of the chosen algorithms for this paper to test the hypothesis that linear models do not perform as well as other models in air quality analysis and forecasting.

D. ARMA, ARIMA, SARIMA

The time series models, ARMA (Auto-Regressive Moving Average), ARIMA (Integrated ARMA), and SARIMA (Seasonal ARIMA), have a common basis: they model a variable using both its past values and an error term. These models require the knowledge of only one pollutant evolution in time to make a forecast, making them more desirable to use for air pollution analysis.

ARMA builds a model using a variable's past values, an error term, and the error term's historical values. ARIMA builds a model starting from ARMA, but has, in addition, a backshift operator, whose role is to "send the variable in the past". SARIMA builds a model starting from ARIMA, but also takes into consideration the possible seasonality of a model as well.

The advantage of using these models is that the analysis and forecasting of each pollutant are much faster, depending only on its past values, its evolutionary trend, and the seasonality identified. The reasoning behind choosing these models is to test the hypothesis that time series models sometimes perform better than both linear and nonlinear models found in the literature review.

E. Decision Tree C5.0 algorithm

Decision Trees are instruments used both in classification problems and in predictions. Graphically presented in the form of a tree flow chart, where each node is a test decision, and each branch is a test result, a decision tree is easy to understand and use. The C5.0 algorithm is the latest, improved version of classifying data using a decision tree and it uses the concept of entropy to measure the purity of a variable: the tree's leaves will have values associated between 0 and 1, where 0 means a homogenous class and 1 means the maximum amount of disorder. Figure 2 shows an example of a decision tree constructed using the C5.0 algorithm on dummy air quality data to classify the PM_{10} AQI category using NO_2 and SO_2.

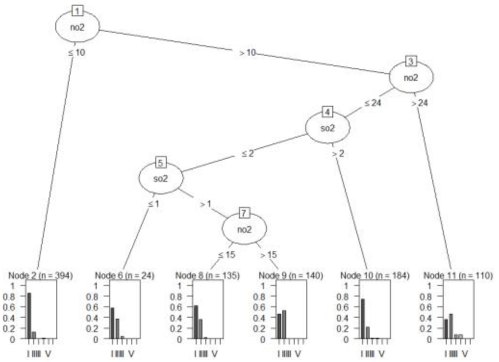


Fig. 2. Example of C5.0 Decision Tree for air quality data

This algorithm is part of the paper's used algorithms to test the hypothesis that, sometimes, this model provides the highest forecasting accuracy among other used models.

F. Random Forest

Random Forest is a robust method that builds multiple decision trees to train, aggregating their results into one, achieving a higher model forecast accuracy. This is a powerful model because it limits the decision tree overfitting issue caused by bias and variance in data, while also increasing the forecast precision. For classification problems, the aggregated result is the majority vote, while for regression problems, the aggregated result is the computed average value.

G. Artificial Neural Network (ANN)

Artificial neural networks are complex models for solving nonlinear problems with a varied number of outputs. One of the most used ANN models is the MLP (Multilayer Perceptron). When building the MLP network, there are multiple layers of artificial neurons: one input layer (with input variables), at least one hidden layer (with computed neurons), and one output layer (used to compute the target variable). To compute the desired result, activation functions are required. Figure 3 shows the architecture of an ANN MLP with three input neurons, one hidden layer, and one output neuron.

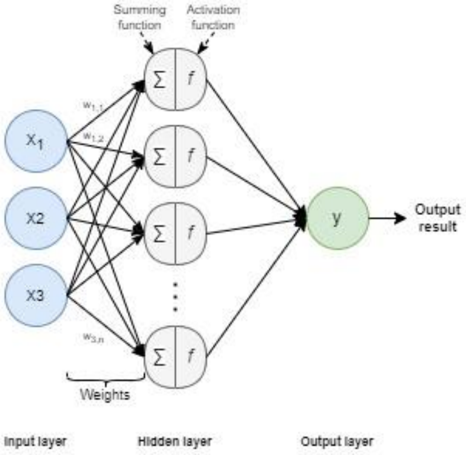


Fig. 3. Artificial Neural Network architecture

ANN is a powerful algorithm, useful in the forecast of air quality, since it does not require a thorough understanding of the dynamics between air pollution concentration levels (or AQI values, if we refer to air quality) and other explanatory variables. ANN is a part of the algorithms used in this paper to test the hypothesis that nonlinear models are superior to linear ones, such as MLR.

4. Romanian AIR Quality Analysis

A. Data source

The data used in the practical part comes from an open-source free website (https://aqicn.org/) which offers historical data of recorded air pollutants around the world. The source is not official, and the given data is subject to change. ANPM (Romanian National Environmental Protection Agency) under the World Air Quality Index Project provided Romanian air quality data.

The data set contains information on several pollutants recorded in Romania: PM_{10}, NO_2, O_3, and SO_2. $PM_{2.5}$ is missing in most of the stations in the country. The models built in this paper use only the four variables mentioned above.

For constructing the data set, the historical daily data from at least one monitoring station per county was extracted from the website. Since some stations did not record any values for O_3, they were removed from the final data set. Time-wise, the final data set contains values from January 2018 to December 2021. All missing data contain values obtained using the interpolation method. Moreover, to analyze Romanian data, the models use the computed average value of each of the four pollutants from the values of all the recorded stations.

B. Results and interpretation

Figure 4 contains the evolution of the extracted and computed Romanian air quality data: it is easy to see that the data is nonlinear. Moreover, O_3 in Romania seems to have a seasonal component in its evolution.

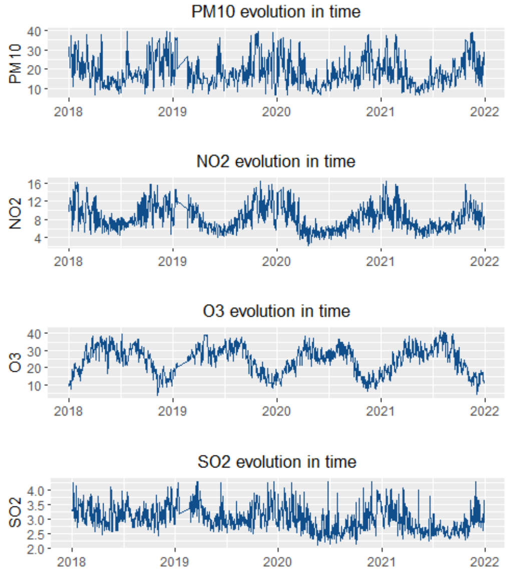


Fig. 4. Air pollutants evolution in 2018-2021

Using Table 1, the data set extends with four factorial variables, which represent the pollutants AQI category. In the available data, all the values for NO_2, O_3, and SO_2 fall within the category "Good", while the values for PM_{10} fall within two categories: "Good" and "Moderate".

All four pollutants have both continuous and factorial values. All the continuous variables were modeled using seven different models, found below: MLR, ARMA, ARIMA, SARIMA, Decision Trees (Regression), Random Forest, and ANN. PM_{10} is the only pollutant that was modeled using classification models for its interval values: Decision Trees (Classification), Random Forest, C5.0, and ANN.

For the pollutant PM_{10}, the best ARMA, ARIMA and SARIMA models were ARMA (4,1), ARIMA (4,0,1), and SARIMA (1,0,3), which can be seen in *Table 2* and Figure 5. RMSE values for all the seven models have a large range, with the lowest RMSE being of 4.405 (corresponding to the time series models) and the highest being of 7.204 (corresponding to the ANN model). ARMA (4,1) can be selected as a good model to predict PM_{10} evolution.

Table 2. RMSE values for PM_{10} models

Pollutant	**Method**	**RMSE**
PM_{10}	MLR	5.968394
	ARMA (4,1)	**4.405115**
	ARIMA (4,0,1)	**4.405115**
	SARIMA (1,0,3)	4.409546
	Classic Decision Tree	6.147519
	Random Forest	6.207245
	ANN	7.204294

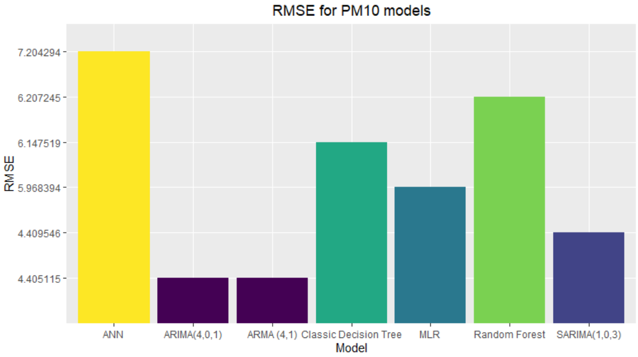


Fig. 5. RMSE for PM10 models using Romanian air quality data

For the pollutant NO_2, the best ARMA, ARIMA and SARIMA models were ARMA (0,3), ARIMA (0,1,3) and SARIMA (0,1,3), which can be seen in *Table 3* and Figure 6. RMSE values for all the seven models have a smaller range, with the lowest RMSE being of 1.523 (corresponding to the time series models) and the highest being of 2.601 (corresponding to the ANN model). ARIMA (0,1,3) can be selected as a good model to predict the pollutant NO_2 evolution.

Table 3. RMSE values for NO_2 models

Pollutant	**Method**	**RMSE**
NO_2	MLR	1.92947
	ARMA (0,3)	1.689982
	ARIMA (0,1,3)	**1.523203**
	SARIMA (0,1,3)	**1.523203**
	Classic Decision Tree	1.998766
	Random Forest	2.014126
	ANN	2.601353

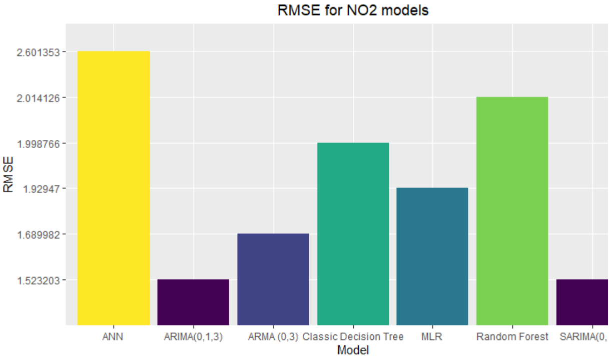


Fig. 6. RMSE for NO2 models using Romanian air quality data

For the pollutant O_3, the best ARMA, ARIMA and SARIMA models were ARMA (1,4), ARIMA (1,0,3), and SARIMA (1,0,2), which can be seen in *Table 4* and Figure 7. RMSE values for all the seven models have a higher range, with the lowest RMSE being of 2.965 (corresponding to the time series model ARMA) and the highest being of 8.444 (corresponding to the ANN model). ARMA (1,4) can be selected as a good model to predict the pollutant O_3 evolution.

Table 4. RMSE values for O_3 models

Pollutant	Method	RMSE
O_3	MLR	6.804479
	ARMA (1,4)	**2.964871**
	ARIMA (1,0,3)	2.971914
	SARIMA (1,0,2)	2.985155
	Classic Decision Tree	6.823751
	Random Forest	6.838346
	ANN	8.444393

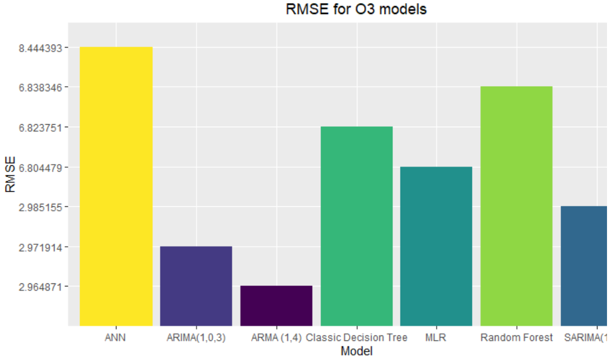


Fig. 7. RMSE for O3 models using Romanian air quality data

For the pollutant SO_2, the best ARMA, ARIMA and SARIMA models were ARMA (1,3), ARIMA (0,1,3), and SARIMA(3,1,1), which can be seen in *Table 5* and Figure 8. RMSE values for all the seven models have a very low range, with the lowest RMSE being of 0.3168 (corresponding to the time series model ARMA) and the highest being of 0.4375 (corresponding to the ANN model). ARMA(1,3) can be selected as a good model to predict the pollutant SO_2 evolution.

Table 5. RMSE values for SO_2 models

Pollutant	Method	RMSE
SO_2	MLR	0.354692
	ARMA (1,3)	**0.3168257**
	ARIMA (0,1,3)	0.3173711
	SARIMA (3,1,1)	0.317244
	Classic Decision Tree	0.3740178
	Random Forest	0.381087
	ANN	0.4375884

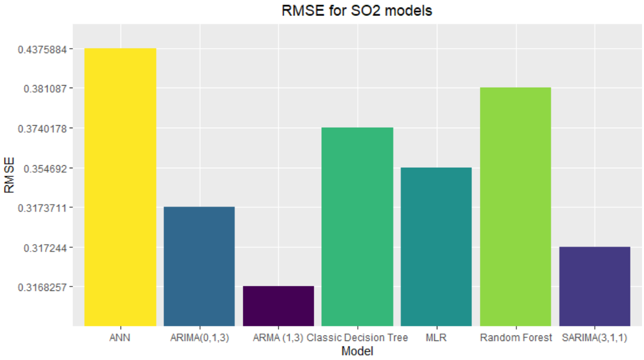


Fig. 8. RMSE for SO2 models using Romanian air quality data

For the available data, the ANN models had the highest RMSE in all four models. This means there is insufficient data to make a proper model for continuous variables using neural networks: to make better models, one would need weather information (temperature, air pressure, wind, humidity, etc.) as well.

PM_{10} is the only pollutant whose factored values can be modeled using classification models. In *Table 6*, one can see the accuracy obtained from the modelling using four different algorithms. The obtained accuracies have appropriate values, between 78.44% and 82.93%. The highest accuracy corresponds to the ANN model, as opposed to the models constructed before, which shows that ANN is a good algorithm to predict PM_{10} interval values.

Table 6. Accuracy for PM10 models

Pollutant	Method	Accuracy
PM_{10}	Classic Decision Tree	79.64072%
	C5.0	81.13772%
	Random Forest	78.44311%
	ANN	**82.93413%**

5. Conclusions

Air pollution is a very important topic that should occupy high priority for governments around the world. Since pollution is in a continuous growth, it is mandatory to have a tool to monitor it, especially because high levels of pollution are harmful to the entire population. After monitoring air quality, it is necessary to analyze the obtained data to help with the proper-decision making to combat pollution.

Such analysis was performed in this paper, using Romania's daily air quality data ranging from 2018 to 2022. Four pollutants were used: PM_{10}, NO_2, O_3, and SO_2; $PM_{2.5}$ was missing significant amount of values, but, in the future, Romanian monitoring stations will add this important pollutant.

After analyzing multiple models, only seven regression models and four classification models qualified for this paper. All regression models had the lowest RMSE for the time series models, and the highest RMSE for the ANN model: given only the information about the four pollutants, time series model perform significantly better than other models, both linear and nonlinear. MLR models also have surprisingly lower RMSE values for the modeled pollutants, as compared to decision trees and ANN. For the classification models, however, the highest accuracy corresponds to the ANN model: factorial data behaves in an easier to model way, which makes it easier for ANN to perform better.

In conclusion, this paper covers the proposed researched topic, by presenting multiple models and computing those using Romanian data. However, this is not enough to have a very good air quality prediction, as some information is missing. To obtain the best air quality forecast, it seems to be important to have weather information in the data set. In the future, the same models should be applied for the integrated data: both air quality and weather data, to test the previously mentioned hypothesis.

Appendix

- IoT – Internet of Things – complex concept regarding a network of intelligent objects;
- $PM_{2,5}$ – fine particle with a diameter less than 2.5 μm;
- PM_{10} – fine particle with a diameter less than 10 μm;
- NO_2 – nitrogen dioxide; fatal in large quantities;
- O_3 – ozone;
- SO_2 – sulfur dioxide; toxic gas;
- AQI – Air Quality Index;
- MLR – Multiple Linear Regression;
- ARMA – Auto-Regressive Moving Average model;
- ARIMA – Auto-Regressive Integrated Moving Average model;
- SARIMA – Seasonal Auto-Regressive Integrated Moving Average model;
- ANN – Artificial Neural Network;
- RMSE – Root Mean Square Error.

References

[1] World Health Organization, "WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide," World Health Organization, Geneva, 2021.

[2] European Environment Agency, "European Air Quality Index," 2022. [Online]. Available: https://airindex.eea.europa.eu/.

[3] T. Kitchilan, M. Abeyratne and P. E. Ediriweera, "Air Quality Monitoring And Prediction Using IOT And

Machine Learning Approaches," *International Journal of Scientific and Research Publications,* vol. 12, no. 3, pp. 34-39, 2022.

[4] V. Barot and V. Kapadia, "Air Quality Monitoring Systems using IoT: A Review," *International Conference on Computational Performance Evaluation (ComPE),* pp. 226-231, 2020.

[5] W. Huang, T. Li, J. Liu, P. Xie, S. Du and F. Teng, "An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability," *Information Fusion,* no. 75, pp. 28-40, 2021.

[6] S. M. Cabaneros, J. K. C. and B. R. Hughes, "A review of artificial neural network models for ambient air pollution prediction," *Environmental Modelling & Software,* no. 119, pp. 285-304, 2019.

[7] A. A. Adebiyi, A. O. Adewumi and a. C. K. Ayo, "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction," *Journal of Applied Mathematics,* 2014.

[8] A. Alimissis, K. Philippopoulos, C. Tzanis and D. Deligiorgi, "Spatial estimation of urban air pollution with the use of artificial neural network models," *Atmospheric Environment,* 2018.

[9] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu and a. G. Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," *International Journal of Environmental Science and Development,* vol. 9, no. 1, pp. 8-16, 2018.

[10] A. Y. Wang and B. T. Kong, "Air Quality Predictive Modelling Based on an Improved Decision Tree in a Weather-Smart Grid," *IEEE Access,* 2017.

[11] A. Singh, R. Kumar and N. Hasteer, "Comparative Analysis of Classification Models for Predicting Quality of Air," in *5th International Conference on Computing Communication and Automation (ICCCA),* 2020.

[12] E. Mosadegh, K. Ashrafi, M. S. Motlagh and I. Babaeian, Modeling the Regional Effects of Climate Change on Future Urban Ozone Air Quality in Tehran, Iran, Cornell University, 2021.

[13] O. Taylan, A. S. Alkabaa, M. Alamoudi and A. Basahel, "Air Quality Modeling for Sustainable Clean Environment Using ANFIS and Machine Learning Approaches," *Atmosphere,* vol. 12, no. 713, 2021.

[14] N. S. Represa, A. Fernández-Sarría, A. Porta and J. Palomar-Vázquez, "Data Mining Paradigm in the Study of Air Quality," *Environmental Processes,* 2019.

[15] A. Panimalar, V. Shree and V. Kathrine, "The 17 V's Of Big Data," *International Research Journal of Engineering and Technology (IRJET)* , pp. 329-333, 2017.

[16] J. T. Ali, "Big data as a tool to improve air quality," 2020. [Online]. Available: https://cepei.org/en/documents/big-data-improve-air-quality/.

[17] I. E. Naqa and M. J. Murphy, "What Is Machine Learning?," in *Machine Learning in Radiation Oncology,* Springer, Cham, 2015, pp. 3-11.

[18] W. Wei, O. Ramalho, L. Malingre, S. Sivanantham, J. C. Little and C. Mandin, "Machine learning and statistical models for predicting indoor air quality," *Indoor Air,* no. 29, pp. 704-726, 2019.

[19] C. Bellinger, M. S. M. Jabbar, O. Zaïane and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC Public Health,* vol. 17, no. 907, 2017.

[20] M. Elbayoumi, N. A. Ramli, N. F. F. M. Yusof, A. S. B. Yahaya, W. A. Madhoun and A. Z. Ul-Sau,

"Multivariate methods for indoor PM10 and PM2.5 modelling in naturally ventilated schools buildings," *Atmospheric Environment*, no. 94, pp. 11-21, 2014.

NICULAE Andreea-Mihaela – student at Bucharest University of Economic Studies, attending Data Bases – Support for Business Master, Bucharest, Romania; obtained a Bachelor's Degree in Economic Cybernetics in 2020; former Erasmus+ student in Athens University of Economics and Business, attending Statistics master courses, Athens, Greece;

Facial Emotion Recognition and Detection Application

Ioana NAGIT, Andreea-Ramona OLTEANU, Ion LUNGU
The Bucharest University Of Economic Studies, Bucharest, Romania
ivana.nagit@gmail.com; ion.lungu@ie.ase.ro

The purpose of the present paper is to study the process of face detection and recognition, followed by the ability to detect the drowsiness state of an individual. This experiment has been part of the bachelor thesis and aims to highlight some of the true power of Artificial Intelligence [1]. *For a full perspective on the topic, an experiment was held in order to emphasize the flexibility and power of facial detection. It aims to analyze the real-time possible drowsiness of a driver by using a key point facial landmark detection and warn the individual when necessary.*

Keywords: *Artificial Intelligence, Face Recognition, Facial Landmarks, Eye Blink Detection*

Introduction

Acknowledging human gestures and instinctual impulses have always captured the attention of analysts because the ability to peruse one's specific gesture helps to amend the human emotional quotient in emotional-computer interaction. To expound the setting of human developments, there are some perspectives that can be taken into consideration, such as voice tone or facial expressions.

The psychological theory of Emotional Intelligence has its origin way back in 1997, when it was developed by the American psychologists Peter Salovey and John D. Mayer [2]. Also known as EI or EQ, the emotional quotient is the individual's ability to recognize one's own emotions but also those of others, to distinguish between feelings and to correctly identify them. All the information received guides the individual's way of thinking and interacting in order to adapt to the environment.

Five main components comprise EI [3], and those are:

- Self-awareness – capacity to recognize and get individual temperaments and feelings, as well as their impact on others;
- Self-regulation - capacity to control or divert troublesome moods and the affinity to think before acting;
- Internal motivation – an affinity to seek after certain objectives with vitality and perseverance;
- Empathy – part of social awareness representing the capacity to understand the temperament of other people;
- Social skills – capability of overseeing connections and building systems [4].

When the English mathematician Alan Turing came up in 1950 with his then mocked idea which supported that machines could be determined to think similarly to a human [5], the concept of Artificial Intelligence first appeared as a possible conceivable reality. His paper - "Computing Machinery and Intelligence" [4] - created some sort of opposition to decide whether this super-new notion is even feasible. The question was on everyone's lips: Could a machine comply with the emotional intelligence of a human being and act accordingly? It took a while for individuals to recognize the genuine control of AI and in 1956, computer and cognitive scientis John McCarthy held the first ever academic conference on this new, controversial subject.

The field of Artificial Intelligence is wide and usually seems to surpass reality. From Turing's Test [6], which assumes that a computer can actually think and give responses similarly to a human, to the inception of the same test and even the contradiction which states that a machine could never invoke an emotional response or originate anything, this marks only the beginning of what now has become a continuous evolution of technology.

A very discussed topic in the field of AI is the Facial Emotion Recognition (FER). Facial Emotion Recognition is a thriving subject of research in which the Human-Robot Interaction is based on evaluating, developing, and planning interactional environments for smart systems to form cognitive and emotional interaction with the help of a few communication channels that link humans to robots. Cognitive scientists keep trying to discover more accurate methods of recognizing facial human expressions and their gestures.

Being able to get the emotions of humans with high accuracy represents a huge advantage in numerous settings, such as monitoring the reactions and feelings of people focused on performing certain tasks or watching an ad. It is important to be able to distinguish between the states of a human because this can be further used in developing applications that may help people struggling with disorders such as dementia and autism spectrum. It might also contribute to building an emotional report of an employee/student or maintaining an accurate profile of a patient that needs to be kept under observation.

1. Drowsiness and eye blink detection

Drowsiness state is common among people, especially for drivers that need to stay awake for a long period of time. It is also known as the main cause of car accidents, because driving while being drowsy is very similar to driving under the influence of alcohol, thus it increases the chances to be involved in a car crash by three times. According to the National Sleep Foundation [8], staying awake for 18 hours straight makes the equivalent of a blood-alcohol concentration of 0.05%, while 24 hours of no sleep equals a blood-alcohol concentration of 0.10%.

Among youthful drivers, driving exhausted is quite common due to way of life components. Young people need more rest than grown-ups; weariness may influence youngsters more than grown-ups. Most proficient drivers should adapt to fatigued driving on a visit premise due to work-related components. Approximately half of all professional drivers have less than typical rest time before embarking on a long-distance trip.

Drowsiness is a very discussed topic at the moment, and a robust motivation came from trying to explore the vast world of artificial intelligence and get a glimpse of what it is really like to connect on a whole different level with deep learning. Another incentive was the analogy with the continuous desire of doing multiple things at once instead of taking turns and giving all of the attention to the immediate task that needs to be completed. Moreover, developing two different applications, an Android and a Python one, respectively, showed the flexibility, huge adaptability, and ease of implementing something that acts in both cases as an alert system.

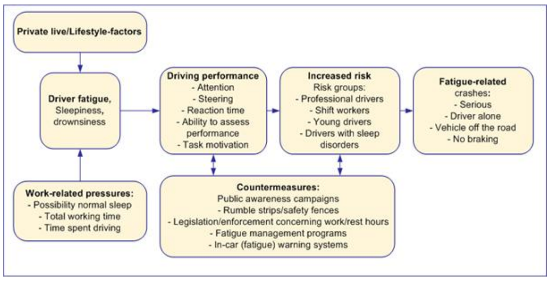

Private live/Lifestyle-factors
Driver fatigue, Sleepiness, drownsiness
Driving performance - Attention - Steering - Reaction time - Ability to assess performance - Task motivation
Increased risk Risk groups: - Professional drivers - Shift workers - Young drivers - Drivers with sleep disorders
Fatigue-related crashes: - Serious - Driver alone - Vehicle off the road - No braking
Work-related pressures: - Possibility normal sleep - Total working time - Time spent driving
Countermeasures: Public awareness campaigns - Rumble strips/safety fences - Legislation/enforcement concerning work/rest hours - Fatigue management programs - In-car (fatigue) warning systems

Fig. 1. Key issues regarding driver

In order to make this first experiment work properly, a method that can determine for how long an individual's eyes have been closed is needed. Therefore, if the eyes of the driver are closed for a given period of time, then an alarm will be played with the sole purpose of bringing the person back to its senses and deciding whether it is time for a nap or if he was just a little distracted. Because, as good as it works for drowsy people, it may come in handy when staring at the phone or in any other part rather than the road while driving.

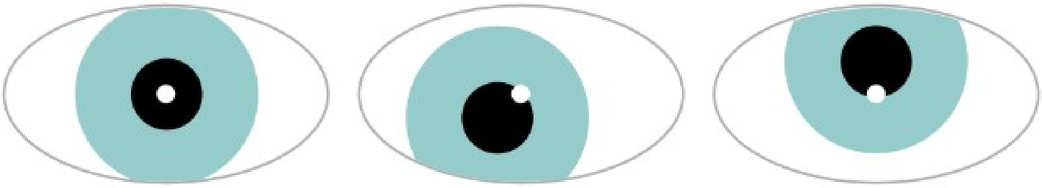


Fig. 2. Infrared light reflection on user's eye correlative to the pupil used to compute the direction of his gaze [6]

The general stream of the drowsiness system is very direct and requires a camera that is able to monitor the face of the driver (e.g. Microsoft LifeCam HD-3000). Once the face of the driver is found, the facial landmark detection is applied, and it extracts the region of the eyes. Further on, the eye aspect ratio is computed in order to decide whether the eyes of the individual are open or closed. An alarm is played if the eye aspect ratio determines that the eyes have been closed for a long period of time.

Eye tracking represents the method of measuring the point where one is looking, the motion of an individual's eye correlative to the head. For more than 20 years, eye tracking has been considered an extreme innovation that helps with the analysis and detection of user interface issues.

The main focus of this method is on the points of fixation (Figure 2). More accurately, they represent specific zones in which a potential user's gaze stops long enough and lingers for a sufficient amount of time so that they can handle what they have just seen. The motion of the user's eyes between these points is also known as „*saccade*" [6].

Facial landmarks have their purpose in localizing and representing specific regions of a potential user's face such as the eyes, nose, eyebrows, and mouth [7]. It has faced a real quick development within the computer vision community since it has numerous applications. For instance, the capability to successfully detect emotion through specific facial gestures, approximating gaze direction, and even the famous face swapping is made with the help of facial landmarks.

Knowing if a person is paying attention or not, especially in systems that require emotion estimation, is done like in the figure below:

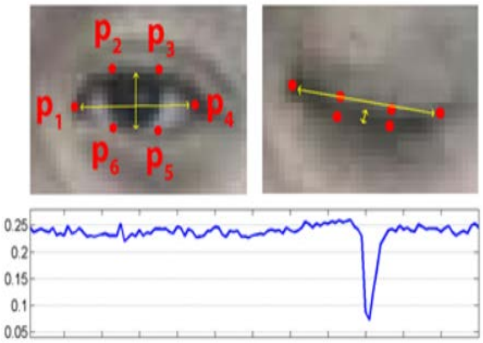


Fig. 3. Eye Aspect Ratio

Figure 3 highlights that every time there is a squint, the proportion between the four points within the eyes becomes smaller. When the eye is open, the EAR remains constant. When the eye is fully closed, the four points turn into a straight line and the EAR drops significantly from 0.25 to 0.05.

Nowadays there are so many applications on both personal computers and smartphones camera programs. For achieving the desired result, the landmark detector has to find the corners of the designated area (mouth, eyes, nose) and connect those points. A lot of algorithms are implemented with the help of the OpenCV free, cross-platform library.

Basically, when talking about facial landmarks, we are actually discussing about the detection of a subset of shape prediction matter. So, the process takes two steps: localizing the face and detecting key facial structures in a specific region of interest. Starting from this point, the developer can choose on which part of the face he wants to focus and create certain tasks.

Of course, before all this can be emphasized, a thorough face detection must be implemented, and it is usually done either by using the built-in Haar cascades of OpenCV library, by applying a pre-trained Histogram of Oriented Gradients [8] and Linear Object Detection, or by using deep learning algorithms. Because the Viola-Jones type detector [6] is a bit old and it really requires a good amount of time to just tune the parameters. Therefore, the second approach was desired and adopted for this experiment. The next subchapter will describe exactly how the process of face detection works.

This paper's focus on eye blinking detection highlights the implementation of DLib library, a modern C++ toolkit that holds machine learning calculations for making complex computer software that can genuinely solve real-world problems. Kazemi and Sullvian's paper [9] was used for implementing the facial landmark detector. This method uses a training data set of manually labeled facial landmarks on a specific image together with the likelihood of distance between sets of input pixels.

The final result may be a facial landmark finder that can be utilized to identify facial regions of interest in real-time with tall quality forecasts. The DLib library uses its pre-trained facial landmark and estimates the facial structure with the help of 68 coordinates (part of the iBUG 300-W dataset) that create the facial structure of an individual. All the coordinates can be seen there as well. They start from the right side of the face, then continue with the right eyebrow, moves to the left eyebrow, which is immediately followed by the node, left and right eye, and finally the mouth.

Using OpenCV to recognize faces means following two simple steps, which is basically detecting the face and extracting the embeddings that quantify each face in a given image. The latter mentioned step involves using a model that is a combination between Python and Torch implementation [10].

Firstly, an image or a video is given as input to the face recognition pipeline. Secondly, the face detection algorithm is applied and the location of the face in the given image is detected. Facial landmarks can optionally be computed so that an accurate alignment of the face can be achieved. Face alignment identifies the geometric structure of faces and attempts to obtain a canonical alignment based on rotation and scale. Even though it is optional, face alignment proved to give a significant increase in accuracy regarding face recognition.

Passing the input image or video through the deep neural network looks something like this:

To prepare a face recognition model by using deep learning means that each input group of information includes an anchor, a positive, and a negative image. The anchor and positive image contain the same face, while the negative image does not share the same identity. With the help of the deep neural network and the Triplet Loss Function [10], the network can assess faces and give vigorous embeddings that are suitable for face recognition.

Choosing the right method or the right solution to implement a specific remains the responsibility of the developer. It is highly recommended to start from a very specific context because deep learning is a very powerful tool, and in order for everything to work out smoothly, there needs to be a simple way of putting things together.

2. Solution Implementation

It is important to be aware of computer vision's field and how it is related to the need to use OpenCV. Computer Vision [11] represents a field which trains computers to transcribe and acknowledge the visual world. It is an indispensable feature for self-driving cars, photo-correction applications, and robotics.

OpenCV is the gigantic open-source library for computer vision and image handling and plays a major, vital part in today's frameworks. By utilizing it, one can handle images and videos to distinguish objects, handwriting of a human, or faces.

When it coordinates with different libraries, such as NumPy, Python is able to handle the OpenCV cluster structure for analysis. To recognize the image pattern and its different highlights, a vector space is used, and then multiple mathematical operations are performed on these highlights.

The Python based experiment begins with importing the OpenCV library together with the other required packages and libraries. The alarm function is defined, and it takes the path to the WAV audio file chosen to alert the driver. It will be called when the computed EAR is lower than the threshold so that the driver can be notified.

Next, the Euclidian distance between the sets of eye landmarks need to be computed, so it is required to create a function which does exactly that. To get a good accuracy, both vertical and horizontal Euclidian distances between the coordinates need to be computed, and then simply return the EAR result. This value stays approximately constant if the driver's eyes are open. A sudden decrease will take place during a blink, and the EAR will get a value close to zero. If the driver's eye is completely closed, then the EAR will stay constant, with a much lower value than the EAR when the eyes are open.

Each eye has six distinct landmark locations, and the counting starts from the outer corner of the eye in a clockwise motion until it is fully covered. With both eyes detected, the function can compute the EAR and further establish in what state the driver's eyes are in.

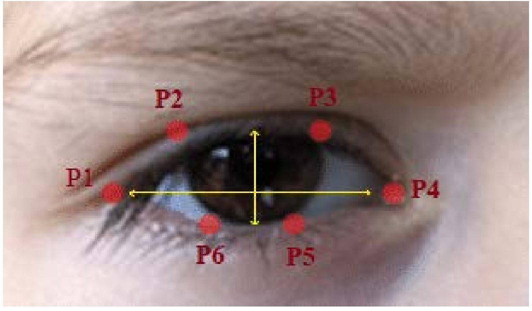


Fig. 4. Landmark locations [22]

In each video frame, EAR is computed by using the formula presented below, where p1-p6 represent landmark locations [15]:

$$EAR = \frac{||p2-p6||+||p3-p5||}{2*||p1-p4||}$$

Constructing the argument parse is also required, so the code will have three command line arguments:

- the shape predictor;
- the alarm;
- the webcam.

It is important to know that the shape predictor is really the path to the pretrained facial detector of DLib. The path to the WAV audio file is optional, but it was used in order to keep the purpose of the application alive. The last command line argument refers to the built-in webcam of the user.

The main part starts from this point on, because the next step aims to define the variables that indicate the blink and the number of consecutive frames, respectively, plus a bool variable for the alarm and a frame counter. If EAR is lower than the given threshold, then the number of frames the driver has kept his eyes closed for will be counted. Consequently, if this counter has a bigger value than the predefined number of consecutive frames, the alarm is sounded.

The input threshold was initialized with a 0.3 value for the very simple reason that it usually gives the best results. The given number of frames is 50, which means that if for 50 consecutive frames the driver's eyes are closed, then the boolean is updated and the alarm plays. Here it is up to the desired level of the application's sensitiveness. For example, if the developer does not want to hear the alarm every time he just looks down, then he needs to increase the number of consecutive frames.

It has been discussed in a previous chapter about DLib's features and the existing relation between the library and the HOG [8] that can train very accurate human detectors. Coming back to the code, the histogram was also instantiated together with the facial landmark predictor. The facial landmarks that are produced by

DLib act similarly to an indexable list, so it was needed to get the indexes of both left and right facial landmarks. Thus, the eye regions were extracted with little to no effort with the help of a slice of an array.

Instantiation of the video stream thread and giving a little time for the camera to properly warm up are next added to the code, followed by a for loop that literally loops over the frames gotten from the video stream. The next frame is read, converted to grayscale, and resized. DLib's face detector is further applied in order to find the face in the image.

Facial landmark detection must be applied next, to get all important regions of the driver's face and convert the obtained result to NumPy array. Using the array slicing described in the paragraph above extracts the exact coordinates of the driver's eyes in order to compute the EARs and get a better estimate by averaging both left and right eye ratios.

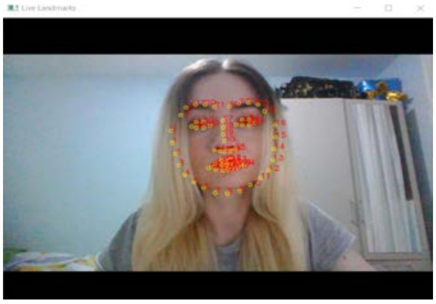


Fig. 5. Facial landmark coordinates

Now both eyes can be visualized after computing the convex hulls, and it is time to see if the state of the driver is drowsy or not. This is done by comparing the input threshold with the EAR. The counter is incremented each time EAR is below the threshold and when it exceeds the number of consecutive frames given as input, it is automatically assumed that the driver is getting drowsy (Figure 6).

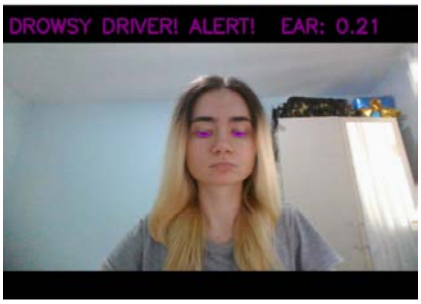

DROWSY DRIVER! ALERT! EAR: 0.21

Fig. 6. Drowsy driver

On top left the message that warns the driver is displayed and on top right the computed EAR can be seen. If the driver is in a normal state, then the driver will only see on the screen the eye aspect ratio displayed.

The alarm sound is checked, and it needs a separate thread where the calling of the alarm can be done. In this way, it is ensured that the main program is not going to stop from execution. If the result of EAR has a bigger value than the input threshold, it means the driver is awake and looking at the road, so the program keeps tracking his eyes. Also, in this happy case, the alarm needs to be turned off.

Displaying the output frame concludes the implementation of the code, and an exit key is added in order to stop the whole process of drowsiness detecting.

Conclusions

The paper "Facial Emotion Recognition and Detection Application" captures the existing trends in the world of machine learning, focusing mostly on face and eye detection. With limited resources, it is hard to compare the application with those created by corporations like Microsoft or Google. However, developing this demo came as a challenge and an unexpected surprise.

My individual contribution is emphasized by developing both applications from scratch, integrating the services offered by Python (OpenCV) and Google Cloud (Google Vision API) and finding a data set for training the neural network with the purpose to accurately identify the face and extract the region of interest, which in this case was represented by the eyes.

Possible future implementations could mean managing to expand the drowsiness detector with a yawning one and with the help of a 3D convolutional neural network to be able to detect the face and eyes, even though the camera is not positioned on the car dash.

Another major improvement could represent the capacity to turn it into a mobile-friendly app and provide certain alternatives to the driver when the drowsiness state is detected, such as suggesting the nearest parking lot or coffee place.

Machine learning and deep learning have already become vital in our everyday life, and as dangerous and exciting as they may seem at a first glance, it takes some courage to dive into the unknown, having as weapons only your IQ, uncontrollable desire to keep adapting, and, of course, a good sense of humor.

Bibliography

[1] I. Nagit, "Facial Emotion Recognition and Detection Application," 2020.

[2] J. D. Mayer, " Annual Review of Psychology," in *Human Abilities: Emotional Intelligence,* 2008, pp. 507-536.

[3] Z. Swijtnik, "Daniel Goleman's five components of emotional intelligence," [Online]. Available: https://web.sonoma.edu/users/s/swijtink/teaching/philosophy_101/paper1/goleman.htm.

[4] [Online]. Available: https://web. sonoma.edu/users/s/swijtink/teaching/philosophy_101/paper1/goleman.htm.

[5] "Alan Turing," [Online]. Available: https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf.

[6] "Eye Tracking: What is it for and when to use it," [Online]. Available: https://usabilitygeek.com/what-is-eye-tracking-when-to-use-it/.

[7] "Facial LAndmark Detection in OpenCV4," [Online]. Available: https://levelup.gitconnected.com/facial-landmark-detection-in-opencv4-616f9c1737a5.

[8] B. T. N. Dalal, "Histograms of Oriented Gradients for Human Detection".

[9] J. S. V. Kazemi, "One Millisecond Face Alignment with an Ensemble of Regression Trees," 2014.

[10] D. K. J. P. F. Schroff, "FaceNet: A Unified Embedding for Face Recognition and Clustering," 2015.

[11] R. Szeliski, "Computer Vision: Algorithms and Applications," September 3, 2010.

[12] K. K. Y.H. Byeon, "Facial Expression Recognition Using 3D Convolutional Neural Network," *International Journal of Advanced Computer Science and Applications(ijacsa),* 2014.

[13] S. Lin, "An Introduction to Face Recognition Technology," vol. 3, no. 1, 2000.

[14] A. Turing, "Mind," *Computing Machinery and Intelligence,* vol. 59, pp. 433-460, 1950.

[15] J. C. T. Soukupova, Real-Time Eye Blink Detection using Facial Landmarks, February 3-5, 2016.

[16] "Fatigue," [Online]. Available: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/ersosynthesis2018-fatigue.pdf.

[17] J. T. M. Stevenson, On the road to prevention: road injury and health promotion, 2014.

[18] "Introduction to OpenCV-Python Tutorials," [Online]. Available: https://docs.opencv.org/master/d0/de3/tutorial_py_intro.html.

[19] D. E. King, Dlib-ml: A Machine Learning Toolkit, 2009.

[20] "Vision AI," [Online]. Available: https://cloud.google.com/vision.

[21] "IoT based Smart Driver Assistance," [Online]. Available: https://www.rfwireless-world.com/Articles/IoT-based-Smart-Driver-Assistance-System-Architecture.html.

Ioana NAGIT is a graduate of the Faculty of Economic Cybernetics, Statistics and Informatics at the Bucharest University of Economic Studies Economic Informatics, with a Bachelor’s degree in Economic Informatics – English module. She is currently pursuing a Master’s degree in Databases - Support for Business and has a keen interest in Big Data, Machine Learning, UX Design and Cloud Computing.

Andreea Ramona OLTEANU is a graduate of the Faculty of Economic Cybernetics, Statistics and Informatics at the Bucharest University of Economic Studies, with a Bachelor’s degree in Economic Informatics in English. She continued down the path of Economic Informatics and is now pursuing a Masters' degree in Databases-Support for Business, being passionate about Business Intelligence, Machine Learning, and Data Warehousing.

Ion LUNGU is a Professor at the Economic Informatics Department within the Faculty of Economic Cybernetics, Statistics and Informatics at the Bucharest University of Economic Studies. He has graduated the Faculty of Economic Cybernetics in 1974, holds a PhD diploma in Economics from 1983 and, starting with 1999 is a PhD coordinator in the field of Economic Informatics. He is the author of 22 books in the domain of economic informatics, 57 published articles (with two ISI-indexed articles) and 39 scientific papers published in conferences proceedings (with five papers ISI-indexed and 15 included in international databases). He participated (as director or as a team member) in more than 20 research projects that have been financed from national research programs. He is a CNCSIS expert evaluator and member of the scientific board for the ISI-indexed journal Economic Computation and Economic Cybernetics Studies and Research. He is also a member of the INFOREC professional association and an honorary member of the Economic Independence academic association. His fields of interest include: Databases, Design of Economic Information Systems, Database Management Systems, Decision Support Systems, Executive Information Systems.

A Correlation Based Way to Predict the Type of Breast Cancer for Diagnosis

Shahidul Islam KHAN
Department of Computer Science and Engineering (CSE)
International Islamic University Chittagong (IIUC)
Chittagong, Bangladesh
nayeemkh@gmail.com

Nowadays, breast cancer is considered one of the most common causes of death among adult women. At the same time, the bright side is that among all the types of cancer, breast cancer is more curable, if diagnosed in the early stages. In this paper, the diagnosis of breast cancer has been proposed using the least possible number of features based on correlation. In the proposed method, we have used correlation to find the strength between the input and the target features. Then we provided a way to create a new subset that consists of only the most relevant features. We have used the Wisconsin breast cancer data set (WBCD) for the experiments. The performance of the model is justified using classification accuracy and the f-score. The result shows that our proposed method obtained the highest classification accuracy (95.26%) with the Random Forest classification using only 4 features from 29 available features, which led to a reduction of 86% in data set size.

Keywords: Health Data; Feature Selection; Correlation; Breast Cancer; Classification

1 Introduction

The use of healthcare information and information technology to organize and analyze health data to improve the quality and safety of patient health care is broadly known as Health Informatics. It deals with the resources from the healthcare sector, machinery and methods used in healthcare to acquire, store, retrieve, and use knowledge in health and medicine. It provides a way to access medical records digitally for health information enthusiasts. In this era of big data, health informatics is a fast-growing field that plays a vital role in the reformation of healthcare [1, 2].

Cancer is the deadliest disease that mankind is facing now. It is a kind of disease in which body cells grow or change out of control. Cancer may form in almost every major part of a human body and is named after the body part it affects. Hence, breast cancer is a term that refers to the abnormal growth in the breast cells. The extra masses caused by abnormal growth are called tumors. The two types of tumors are malignant, which is supposed to be breast cancer, and benign, which is not cancerous and not life-threatening. Today, breast cancer is one of the most common causes of death among middle-aged women (40-55 years). As indicated by the World Health Organization, 2.1 million women are affected each year by breast cancer. In 2018, 627,000 women were estimated to have died from breast cancer, which is roughly 15% of all cancer-caused death among women [3].

Health informatics may play an important role in improving the survival rates of breast cancer patients. Early detection of the type of a tumor by performing the least possible diagnostic test may help to push toward achieving the maximum survival rate.

The feature is a distinguishable attribute or perceptible characteristic of something that is being observed. Feature selection is a way that removes unimportant and redundant features. It generates a subset of features with less dimensionality than the original data set and still provides good prediction results.

The use of machine learning approaches in the health care field is increasing gradually. In the diagnosis of a disease, the most important factors are the patient’s data and the result obtained from the diagnostic tests.

In general, patients have to take a lot of diagnostic tests based on which doctors lead to a decision on whether the patients have benign or malignant type tumors. Current information systems for detecting the type of breast cancer need a lot of features, which is time-consuming. By selecting the most relevant features from the original set, the features needed for the detection of breast cancer type are reduced, which will automatically reduce the number of tests needed as well as the time consumption. Hence, feature selection can be a supportive tool for a doctor in diagnosis and decision-making.

Many authors have used a different type of approach for selecting features, but very few of them have used the correlation method. So, it is a research issue to select features from breast cancer data sets using correlation with the target variable.

In this paper, we have presented a brief overview of a feature selection technique that uses a feature-ranking method based on the correlation of input features with the target. We have provided a way to mitigate the problem of redundant features. We have used two types of correlation methods to find the strength of association between an input feature and the target. Pearson's correlation was used to calculate the correlation between the numeric input and the numeric output. Point Biserial correlation was used to find the correlation between numeric input and binary output and vice versa. We showed that the accuracy of a model can be preserved or improved even after selecting a small subset of features from the original set.

The remainder of the paper is organized as follows. In Section II, we have briefly presented research works related to the diagnosis of breast cancer. Section III describes the feature selection. In Section IV, we have presented the methodology and experiments of our proposed method. The results obtained by applying the proposed method are presented in Section V. Section VI finally concludes the paper.

2 Related works

There has been a great deal of research on breast cancer diagnosis, where the data set was the same as ours in the literature. In [4], the authors have combined two techniques: an evolutionary algorithm and fuzzy systems to automatically report the system of diagnosis. They obtained a classification accuracy of 97.36%.

The authors of [5] have developed a knowledge-based system. The system uses the clustering, noise removal, and classification technique. To cluster the data, they have used Expectation-Maximization, Classification, and Regression Tree to generate the fuzzy rule and PCA to overcome the multiple collinearity issue. They obtained 93.20% accuracy on the WDBC data set.

A combination of an Artificial Immune Recognition System and Synthetic Minority Over-Sampling Technique in a system is presented in [6]. They have compared their result with other classifiers like AIRS, BPNN, C4.5, etc. They obtained 96.53% accuracy using their method.

A novel fuzzy model structure that is an extension of the quadratic Bayes classifier has been presented in [7]. The authors analyze the clusters using Fisher's interclass separability criteria to select the relevant input variable. They obtained 95.57% accuracy by applying the supervised fuzzy clustering technique.

RIAC, a method that stands for Rule Induction through Approximate Classification, was presented in [8]. This method was used for inducing rules from examples, which is based on the theory of rough sets, and obtained 94.99% accuracy.

In [9] they have used 10-fold-cross-validation along with the C4.5 decision tree method. They gained 94.74% classification accuracy. The authors of [10] have presented a convenient approach for learning fuzzy classifiers from data. They obtained 95.06% accuracy using the method called neuron-fuzzy techniques.

In summary, the above-cited works have gained promising results, but none of them have used the feature ranking based on correlation.

3 Feature selection

Feature selection is a process that removes irrelevant and redundant features with little or no predictive information from data sets before an algorithm is applied, generates a subset of features with less dimensionality than the original data set had, and still provides good prediction results.

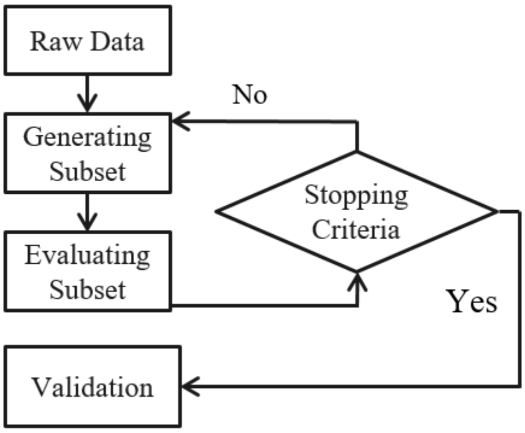


Fig. 1. Basic Procedure of Feature Selection

Fig. 1 demonstrates the progression of the normal technique adopted in feature selection. The dimension space is scaled down to a subset of features that are evaluated based on the criterion. Then, the selected features are validated by the validation process. Stopping criteria are used as an end process indicator; the process may stop if any of the following criterion is fulfilled [11]:

i. Some predefined features have reached;
ii. New features addition/deletion does not bring an improved result;
iii. The selected feature meets the best possible outcome according to the evaluation criterion.

A large number of algorithms have already been proposed to solve the 'curse of dimensionality. Here some of the methods are presented briefly:

3.1 Filter method

The filter method is one of the simplest and computationally less expensive approaches, in this approach feature selection is performed as a preprocessing step. It uses a ranking method to score the features and then the compares the score with a predefined threshold value. If the score is less than the threshold value, then the feature is considered to be irrelevant and gets eliminated. There are three common measures to rank/score a feature; they are distance metrics, correlation, and mutual information.

3.2 Wrapper method

The wrapper method is an approach that evaluates all possible combinations of features to select the subset that leads to the best output [12]. As this method tests all the possible combinations, this can become computationally expensive when the data set is very large. Since the wrapper method evaluates 2n subsets, it becomes an NP-hard problem. To find the subset on the wrapper method, several search algorithms are used. These algorithms can be categorized into Exhaustive Search, Non-Exhaustive Search and Heuristic Search.

3.3 Embedded method

The embedded method is a feature selection method where features are not gets selected or rejected [13]. In this approach, feature selection is integrated into the learning algorithm. Features with less importance are given low weight, which is also called regularization. There are a few types of embedded techniques: Decision tree, LASSO Regression, RIDGE Regression, etc.

Several more techniques deal with dimensionality reduction such as PCA, Clustering, Missing value ratio, Boruta, SelectFromModel, etc.

4 Methodology and experiments

4.1. Breast cancer data set

We have utilized a Breast Cancer Wisconsin (Diagnostic) data set which we took from the Machine Learning Repository of the University of California, Irvine [14]. Researchers who use ML approaches for the diagnosis of breast cancer commonly use this data set. The data set contains 569 data. It consists of 32 features computed from the digitized image of FNA of breast masses. ID was used for identification and diagnosis (M for malignant and B for benign) as the target variable. The rest of the features are 10 real-valued features; these are area, texture, compactness, radius, smoothness, concavity, perimeter, concave points, symmetry, and fractal dimensions. Each of these features was used in three different forms: mean, se (standard error), and worst. For example, radius_mean, radius_se, radius_worst, etc. each of which has a numeric value. In the data set, 357 samples belong to the benign class, and the rest 212 are of the malignant class.

4.2 Correlation

Correlation plays an important role in building a feature selection model. It is advantageous, as it helps to find the input features highly correlated with the target. Measurement of linear dependencies between features correlation coefficient is a widely used method. There are a few types of correlation coefficient measurement methods, such as Pearson's correlation coefficient, Kendall's rank correlation coefficient, etc. we have used Pearson's and Point Biserial Correlation coefficient. Pearson's coefficient can be expressed as:

$$\mathrm{CoR(i)} = \frac{Cov\ (Xi, Y)}{\sqrt{var(Xi) * var(Y)}}$$

Here, CoR(i) is the correlation coefficient, which is obtained by dividing the covariance of the two variables (xi is the ith variable and Y is the output variable) by the product of their variance. Another correlation method is the point Biserial method, which can be expressed as:

$$r_{pb} = \frac{M_1 - M_0}{\sqrt{S_n}} \sqrt{pq}$$

Here, M_1 is the mean for the group that contains a positive binary variable. M_0 is the mean for the group that contains the negative binary variable. Sn is the standard deviation of the complete test. p and q represent the rate of cases in the "0" and "1" groups, respectively.

4.3 Proposed method

In this paper, a feature selection technique based on correlation is presented. We have used the correlation to rank the feature and the sorting method to evaluate all possible ranked features sequentially to select the subset that leads to the best output. We have applied the StandardScalar method to standardize data. To verify the results, we have used different classification techniques such as Random Forest, SVM (Support Vector Machine), Naïve Bayes, and KNN (K-nearest Neighbor). Using these methods, we checked the accuracy of the original data and then the proposed method was applied to make a subset of the reduced feature. Again, we checked the accuracy of the reduced data set using the classification technique mentioned above. Then we compared the result of the original data set and a reduced subset.

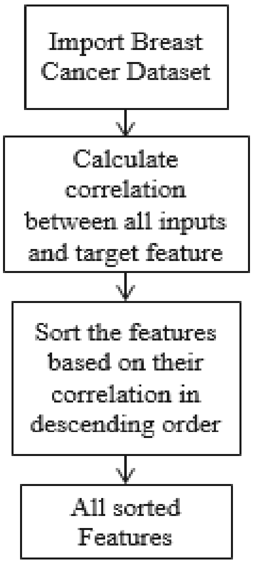

Import Breast Cancer Dataset
Calculate correlation between all inputs and target feature
Sort the features based on their correlation in descending order
All sorted Features

Fig. 2. Flowchart of the proposed method

Fig. 2 demonstrates how the proposed method works. Importations of highdimensional data sets to the selection of a reduced subset of features are shown in this flowchart.

The steps of the method are:

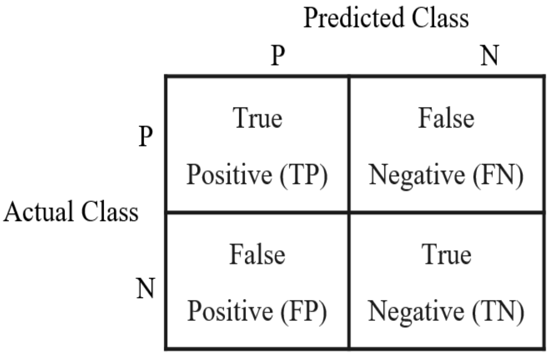
i. Import breast cancer data set;
ii. Compute the correlation between inputs and target variable using Pearson correlation and Point biserial correlation;
iii. Sort the features based on their correlation in descending order.

From the sorted feature, we will add one feature at a time to evaluate the result. The features are sorted based on the high correlativity with the target feature. Hence, a few numbers of features meet the desired output.

4.4 Measures for performance evaluation

We have used different equations to measure the performance of the developed method. These measures are precision-recall, accuracy, and F-score. They can be defined using a confusion matrix.

A confusion matrix is a table that helps to visualize the performance of an algorithm. In this 2*2 matrix, where instances in a predicted class are shown in a row, and the actual class is shown in a column.



		Predicted Class	
		P	N
Actual Class	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

Fig. 3. Confusion Matrix

To demonstrate the result, the following equations are used:

$$Accuracy\ (\%) = \frac{TP + TN}{P + N} \ ...i$$

$$Precision\ (\%) = \frac{TP}{TP + FP} \ ...ii$$

$$Recall\ (\%) = \frac{TP}{TP + FN} \ ...iii$$

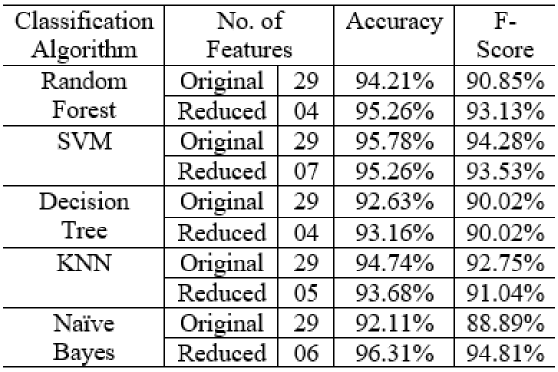
$$F - Score(\%) = \frac{2 * Precision * Recall}{Precision + Recall} \ ...iv$$

5 Results and discussion

We have conducted experiments on the Breast Cancer data set to justify the effectiveness of our approach. The significance of each feature is calculated by its correlation with the target feature. Table 1 shows the results obtained by applying five different classification models. The result shows that the reduction of features does not affect the outcome to a greater extent. Among the five models, naïve Bayes achieved the highest classification accuracy; 96.31%, but in terms of selecting features random forest is more promising which selects only 4 features with an accuracy of 95.26%.

Fig. 4 and Fig. 5 are the performance v/s feature numbers after applying the Random Forest and KNN classification algorithm, where no. of features is on the x-axis and performance is on the y-axis. From the curve, it is seen that the accuracy has been increased initially and been almost the same in the middle and later part of the curve, which indicates that a smaller number of features are enough for obtaining the highest possible outcome.

Table. 1. Summary of the results of Breast Cancer data



Classification Algorithm	No. of Features		Accuracy	F-Score
Random Forest	Original	29	94.21%	90.85%
	Reduced	04	95.26%	93.13%
SVM	Original	29	95.78%	94.28%
	Reduced	07	95.26%	93.53%
Decision Tree	Original	29	92.63%	90.02%
	Reduced	04	93.16%	90.02%
KNN	Original	29	94.74%	92.75%
	Reduced	05	93.68%	91.04%
Naïve Bayes	Original	29	92.11%	88.89%
	Reduced	06	96.31%	94.81%

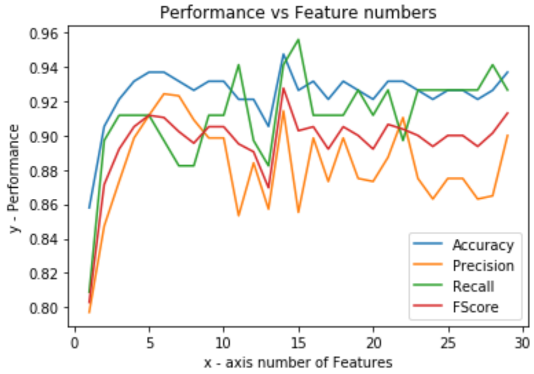


Fig. 4. Performance vs Feature numbers on Breast Cancer data set for Decision Tree

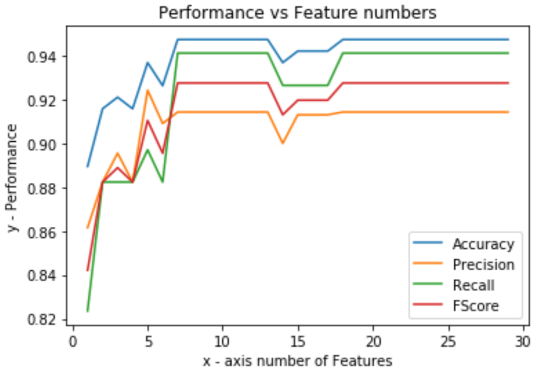


Fig. 5. Performance vs Feature numbers on Breast Cancer data set for KNN

The proposed method helps to reduce data set size. From Table 1 we can see that the difference between the result obtained from the original data set and the reduced data set is negligible. But what impacts the outcome to a greater extent is the size of the data. Table 2 shows a summary of the reduction in data size for different classification algorithms after applying the proposed method. In terms of reduction in size, Random Forest and Decision Tree outperformed other algorithms by reducing the size approximately by 86%. So when we will work with a larger data set in the future, we will get better results with only a few features.

Table. 2 Summary of the data set size-reduction

Classification Algorithm	**No Original Feature**	**No of the Selected Feature**	**Reduction in Data Size**
Random Forest	29	04	86.20%
SVM		07	75.86%
Decision Tree		04	86.20%
KNN		05	82.76%
Naïve Bayes		06	79.31%

From the results above, we presume that in classifying the potential breast cancer patients, promising outcomes have been obtained by the proposed method.

6 Conclusion

A method based on correlation and sorting of features according to their correlation has been applied to the task of predicting the type of breast cancer using the least possible number of features. As we have found the correlation between input and target features and sorted them, hence we were able to detect those features that make the most impact on a particular output. We observed that our proposed method has gained classification accuracies of 95.26%, 95.26%, 93.16%, 93.68%, 96.31% using 04, 07, 04, 05, 06 no. of features for Random Forest, SVM, KNN, decision tree, and naïve Bayes classification, respectively, without using the original (29 features) set. In terms of data set size reduction, the developed model outperformed all other models by reducing approximately 89% of the data set size.

Considering the results, the SVM and Random Forest-based model obtain the best results in classifying breast cancer using the developed method. We have high hope that the method proposed here can be very supportive for the health researchers in their ultimate decisions. They can make a decision within the least possible time using such a tool. Further exploration with a larger data set and finding the reduced subset for data with no target variable can yield more useful results. We will focus on these for our future work.

Bibliography

[1] P. A. Bath, (2008). Health informatics: current issues and challenges. *Journal of information science*, *34*(4), 501-518.

[2] "What Exactly is "Health Informatics?", Healthcare-management-degree.net, 2019. [Online]. Available: https:// www.healthcare-management-degree.net/faq/what-exactly-is-health-informatics/. [Accessed: 14-Oct- 2021].

[3] "Breast cancer", World Health Organization, 2019. [Online]. Available: https://www.who.int/ cancer/prevention/diagnosis-screening/breast-cancer/en/. [Accessed: 14- Oct- 2021].

[4] C. Peña-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis", *Artificial Intelligence in Medicine*, vol. 17, no. 2, pp. 131-155, 1999. Available: 10.1016/s0933-3657(99)00019-6.

[5] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method", Telematics and *Informatics*, vol. 34, no. 4, pp. 133-144, 2017. Available: 10.1016/j.tele.2017.01.007.

[6] K. J. Wang and A. M. Adrian, "Breast cancer classification using hybrid synthetic minority over-sampling technique and artificial immune recognition system algorithm," *Int J Comput Sci Electron Eng (IJCSEE)*, vol. 1, pp. 408-412, 2013.

[7] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers", *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2195-2207, 2003. Available: 10.1016/s0167-8655(03)00047-3.

[8] H. J. Hamilton, N. Cercone, and N. Shan, RIAC: a rule induction algorithm based on approximate classification: Citeseer, 1996.

[9] J. Quinlan, "Improved Use of Continuous Attributes in C4.5", *Journal of Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996. Available: 10.1613/jair.279.

[10] D. Nauck and R. Kruse, "Obtaining interpretable fuzzy classification rules from medical data", *Artificial Intelligence in Medicine*, vol. 16, no. 2, pp. 149-169, 1999. Available: 10.1016/s0933-3657(98)00070-0.

[11] L. Ladha and T. Deepa, "Feature selection methods and algorithms," *International journal on computer science and engineering*, vol. 3, pp. 1787-1797, 2011.

[12] U. Malik, "Applying Wrapper Methods in Python for Feature Selection", Stack Abuse, 2019. [Online]. Available: https://stackabuse.com/applying-wrapper-methods-in-python-for-feature-selection/. [Accessed: 14-Oct- 2021].

[13] S. Rawale, "Feature Selection Methods in Machine Learning.", Medium, 2019. [Online]. Available: https://medium.com/@sagar.rawale3 /feature-selection-methods-in-machine-learning-eaeef12019cc. [Accessed: 14- Oct- 2021].

[14] UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set", Archive.ics.uci.edu, 2019. [Online]. Available: https://archive.ics.uci.edu/ ml/datasets/Breast%20 Cancer%20Wisconsin%20(Diagnostic). [Accessed: 14- Oct- 2021]

Dr. Shahidul Islam KHAN obtained his B.Sc. Engineering Degree in Computer Science and Engineering (CSE) from Ahsanullah University of Science and Technology (AUST) in 2003. He obtained his M.Sc. and Ph.D. from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2011 and 2020. His current fields of research are Data Science, Database Systems, Machine Learning, Information Security, and Health Informatics. Currently, he is serving as the Head of the IIUC Data Science Research Group. He has more than fifty published papers in peer-reviewed journals and at reputed international conferences. He is also an Associate Professor in the Dept. of CSE, International Islamic University Chittagong (IIUC), Bangladesh.

Improving the Treatment Process of Bengali Autistic Children using Specialized Mobile Application

Rashid Al SHAFEE; Rakibul HUDA; Mohammad Imran HOSSAIN; Md. Mahmudul Hasan SHOHAG; Shahidul Islam KHAN
Department of Computer Science and Engineering (CSE)
International Islamic University Chittagong (IIUC)
Chittagong, Bangladesh
rashidodd@gmail.com, rakib10rr3@gmail.com, imranhossain16.ctg@gmail.com, imaginativeshohag@gmail.com, nayeemkh@gmail.com

Many children in Bangladesh have ASD. The rate of Autism Spectrum Disorder (ASD) is increasing in Bangladesh and other countries, day by day. Autistic children find it difficult to talk and express themselves regarding what they want or not. Also, some autistic children are not comfortable dealing with the outside world. For example, they do not feel comfortable in social settings or in any program. There are some schools and organizations where many kind-hearted people are trying to help those autistic children in many ways. We all know that there is no cure for autism. But it can be reduced. After good treatment, an autistic child can recover. To help the treatment process, we have developed an interactive app that will help them to cope with social events and places, as well as help them with verbal tasks. We have developed a model to access the severity of an autistic child and help the child to improve communication. This paper presents our interactive app and also provides a concise comparison of it with existing apps to support children with ASD.

Keywords *- Autism, ASD, Android Apps, Bengali, Machine Learning, ABLLS*

1 Introduction

Autism Spectrum Disorder (ASD) is a brain disorder. It is a lifelong disorder that is understood by weakness in social skills, having difficulties with speech and nonverbal communication, and engaging in repetitive behaviors. Almost one percent of the world population has ASD [1]. The percentage of ASD increased day by day. Children with ASD have problems with their social skills. Statistics show that children with ASD find it difficult to communicate with others and form relationships with others. They find it challenging to make sense of the world around them. Some children with ASD can only speak a few words. Some cannot even speak in their whole life [2].

Because of their lack of social skills and communication skills, it is difficult to teach them. Therefore, they need special care and a special Individual Education Plan. They also need special tools and apps to help them to learn.

We know that smartphones and mobile apps have made our life simpler and better. So why shouldn't the smartphone and mobile apps make the hard life of an ASD child easier?

That is why we are building a mobile application that will help ASD children to learn and get friendly with the social environment, and it will also help non-verbal children to express their emotions and needs to others.

2 Background

Autism is a very common neurodevelopmental disorder. It may be defined as challenges in social interaction and stereotypic behavior. As it is not a single type of disorder, rather it involves various degrees of severity, it is referred to as the Autism Spectrum Disorder (ASD) [3]. According to recent estimates, ASD has a prevalence of about 1 in 54 children and occurs about four times more frequently in boys than in girls [4], [5]. Every year, the total number of children diagnosed with ASD is increasing, which can be seen in Fig. 1.

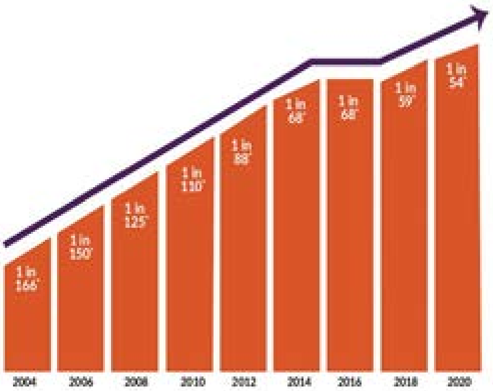


Fig. 1. Statistics of prediction of children with ASD

There is also a risk of recurrence of up to 20% for children when their elder brother or sister is detected with ASD [5]-[8]. ASD is still a non-curable disease. Research shows that if we can diagnose ASD early, it is beneficial for the challenged kids for their treatment [9]. However, many pediatric doctors and health professionals do not share information about a child with ASD with its parents. In some cases, the doctors hope that the children might get cured with age and in some cases, the doctors feel uncomfortable sharing the negative information considering its impacts.

Individuals with autism often remain undiagnosed or incorrectly diagnosed because many clinicians hesitate to discuss this possibility with parents of young children, even when some symptoms are present. These physicians are often concerned about family distress, the negative effects of labeling a child, the possibility of being wrong, or the hope that symptoms will reverse or improve over time. However, healthcare researchers believe that the pros of early and correct diagnosis are much higher than the cons and help the child’s family in the long run.

There are many protocols for assessing and educating children with ASD, one of the popular is - The assessment of Basic Language and Learning Skills (ABLLS) by Dr. Partington [10]. In this protocol, the children face an initial assessment at the beginning. Then an Individual Education Plan or IEP is created for each child. IEP is created based on their disability. Every child has a separate learning timeline from others. The teachers used various materials to train the autistic child. They also use smartphone apps, such as puzzle apps, drawing apps, mix, and matching apps, etc. But most of the apps are in English and many of them have a lot of bugs. So, it is getting difficult to teach ASD children in Bangladesh and find helpful apps in the Bengali language. To address these issues, we have performed research, interviewed the stakeholders, and developed a mobile-based application.

3 Objectives and methodology

Children with autism have a disability in communication and lack social skills. And it is difficult to develop communication skills. Therefore, our objective is to create a media between autistic children and other people to communicate with each other and help them to learn social skills. As Autism is not curable, our motive is to reduce the effect of the disorder on their attitude and improve the communication skill of the challenged kids. Another goal is to develop the app in the Bengali environment as in Bengali, no such app is available. Therefore, Bengali parents face difficulties to get used to another language app to teach their challenged kids.

4 Data collection and analysis

We have already collected data sets from a renowned school for autistic children. This school follows the ABLLS protocol to train students diagnosed with ASD. We have visited the school a few times and met with their teachers and children. We also received suggestions for our work from the school teachers. We have collected their manually evaluated data. Currently, there are 26 students enrolled in the school and we have 26 data sets of 9 children diagnosed with ASD at the school.

In ABLLS methodology, there are 12 steps for initial assessment:

1. Cooperation with Adults
2. Requests (Minds)
3. Motor Imitation
4. Vocal Play
5. Vocal Imitation
6. Matching to Sample
7. Receptive
8. Labeling (Tacts)
9. Receptive by Function, Feature, and Class
10. Conservational Skills (Intra Verbal)
11. Letters and Numbers
12. Social Interaction

We got all data sets in handwritten format, so we had to digitize them. The initial assessment has 12 events. Each event has 5 scores. When a new child comes to the school, a teacher runs the assessment and scores against every event. Then after running the IEP for some months (maybe 3 or 6 months, depending on the performance of the Initial Assessment), the teacher performs the assessment again and sees if there is any progress in the score.

5 Proposed model

The following diagram of Fig. 2 shows our proposed model to automatically predict the level of autism of a child based on the scores of 12 assessment tests proposed in the ABLLS protocol.

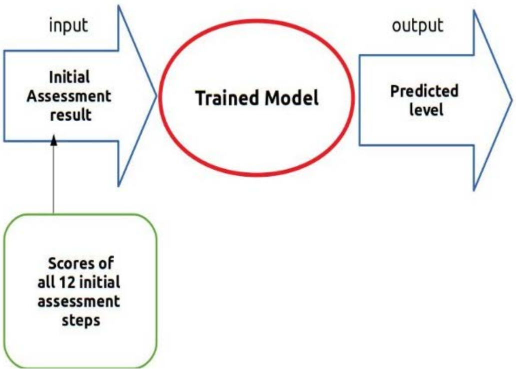

input
Initial Assessment result
Trained Model
output
Predicted level
Scores of all 12 initial assessment steps

Fig. 2: Block diagram of Autism level predicting model

From the values of each category in the Initial Assessment, we see that 1 indicates the most negative behavior. 5 indicates the most positive behavior. 2 to 4 indicate the behaviors gradually improving from negative to positive. As ASD has 3 levels: Mild, Moderate, and Severe. We can correlate the values with those levels. We calculated the severity value for Model 1 of the twelve categories to predict the level by the following equation 1.

$$Severity = \frac{\sum_{i=1}^{n} Value\ (i)}{n} \quad ---(1)$$

Table 1: Autism Leveling in Proposed Model

Level	**Range**
Level - 3: Severe	Severity<2
Level - 2: Moderate	2<=Severity<3.5
Level - 1: Mild	Severity>=3.5

For Level 3 - Severe we have considered 2 as the threshold average value. If we consider 1 which is the most negative value for each category as the threshold average value for Level 3 - Severe, an individual who scores 1 in every category will not be in the severe level. If he scores 1 in the maximum categories and more than 1 in the remaining categories, he will not also be in the severe level. To overcome the level, we considered the threshold value as 2. Similarly, we considered 3.5 as the threshold average value for Level 1 - Mild, because if an individual scores 4 in maximum scores less than 4 in the remaining, he will not also fall in the mild category. To overcome the problem, we considered 3.5 as the threshold value. The values between 2 and 3.5 indicate the Level 2 - Moderate group.

6 Features of the developed app

For developing our app, we have followed the Prototype model. The prototype model is an iterative trial-and -error-based procedure that works between the software developers and the end-users. This model works fine in situations where all the project requirements are not known a priori. The model is presented in Fig. 3. In this model, a basic prototype is developed with the initial requirements of the stakeholders. Then, using an iterative process, customers' feedback is collected and the basic prototype is updated. This process ends when the stakeholders are fully satisfied or all the functionalities meet. We have developed the basic version of *PicTalk* using our initial requirement analysis. Then from the feedback of the teachers of the autistic children and with some feedback from the specially challenged students, we have improved the app in an iterative way.

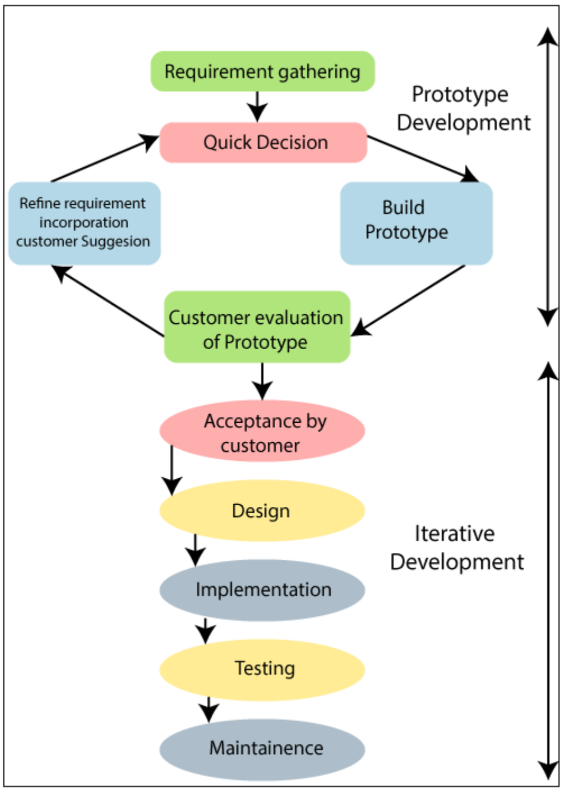


Fig. 3 Prototype model

Our application *PicTalk* has two parts, a social interaction learning part and a communication part.

1. Children with autism learn social interaction using various pictures. Generally, teachers show some images and try to make you understand what to do and what not to do in a place. For example, do not speak loudly in the mosque, etc. So, we make a compilation of places in our app with interactive information. We believe that children will become more interested in learning with our app.

2. Our second feature is for the non-vocal or less talking children. Children can use our app to ask for something, for example, "I want to eat rice" or yes/no answer, etc. Our app has a list of tasks and important elements' names, children can just click and the app will say the task. The app is fully in the Bangla language. So it will also speak in Bangla. Fig. 4 – Fig. 7 provide a pictorial view of our developed app to improve the social skills of children with ASD.

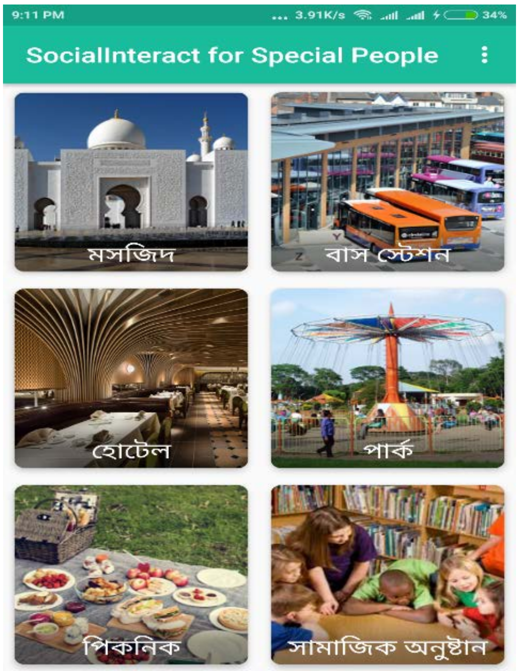


Fig. 4. All areas of the social interaction part

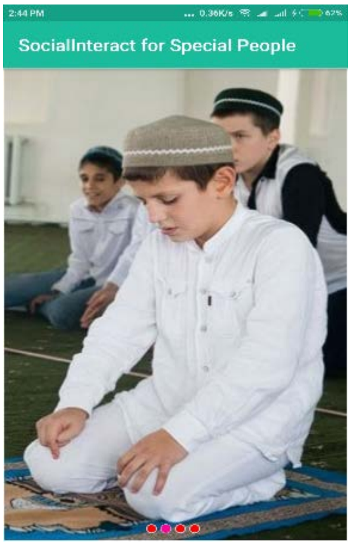


Fig. 5. A part of how to react in a Mosque

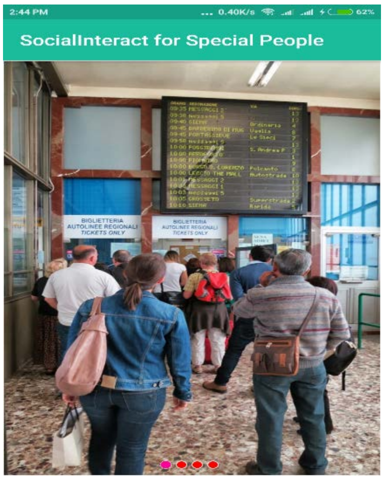


Fig. 6. How to react at the bus station

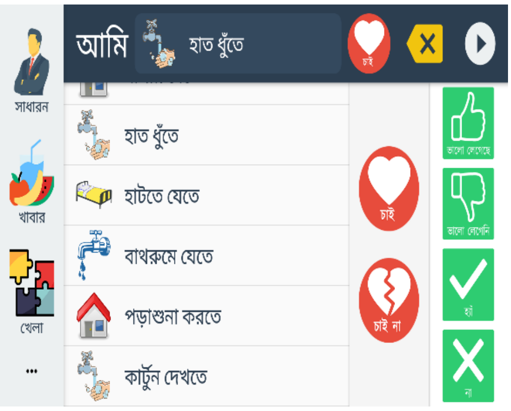


Fig. 7. Home page view of Pic talk part

7 Comparison with existing tools

We searched for other apps for autism training. Comparisons between our app and other existing apps in the English language are presented in Table 2.

Table 2: Comparison between our app and other apps

App Name	Language	Text To Speech	Learning with image	Helpful for non-verbal students	Text-based learning	Online or Offline	Ref
PicTalk	Bangla	Yes	Yes	Yes	No	Offline	[11]
Teach Autistic Children	English	No	No	NO	Yes	Online	[12]
Otsimo	English	No	Yes	Yes	No	Online	[13]
Talking pictures	English	Yes	Yes	Yes	No	Offline	[14]
LetMeTalk	English	Yes	Yes	Yes	No	Offline	[15]

8 Usefulness survey

We interviewed some of the teachers and doctors about the prospective impacts of our developed app. They were all positive about our idea and research findings. DR. Fahmida Islam Chy, vice president of the Foundation of Autism Research and Education (FAREBD), said that it will be a great addition to ASD treatment procedures, as well as for teachers and also for family members of an ASD-diagnosed child. The General Secretary of the Foundation of Autism Research and Education (FAREBD) said that ASD-diagnosed children stay in the training school for a maximum of 5 to 6 hours. The rest of the time, they stay with their parents and family members. If any app can be used to train them about social interaction and verbal talk, it will be quite beneficial for the children. One of the guardians of an ASD-diagnosed child told us that her son always screams if he goes to a train station, bus station or any crowded place. This app will be very helpful for her to teach her son how to react in a crowded place.

9 Limitations

Though the app is fully interactive, it needs a human helper to perfectly use it, and make it help autistic children. As our developed app is the first one of its kind in Bengali, there are still many scopes for improvement by considering the short-term and long-term effects of the app among children with ASD.

10 Conclusions

ASD is a neurological condition, meaning that people with ASD are born with it, and have it for life. But they can be trained and they can become skilled. The main skills they need are social skills, behavioral skills, and communication skills. We are trying to help the children with ASD by supporting them in improving their social skills with the help of our smartphone-based application, which is fun and interactive. Also to help the non-vocal or less talking children to communicate with others using the application. We believe it will help them learn and communicate with others.

As autism is not remediable, we believe that our work will at least help in the training process for ASD-diagnosed children. Teachers, as well as parents, can use this app for children to teach them about society and how they will react in social places such as bus or train stations, restaurants, social programs, etc. Also, it will help non-verbal ASD diagnosed children to express their feelings and what they want to do or not. The Government of Bangladesh should take proper steps to facilitate the development and improvement of apps for the training of special children in their own language.

Acknowledgment

We are grateful to the authorities of the Foundation for Autism Research & Education (FARE) and the Spectra School of Autism for allowing us to work in their place and for providing us with valuable data.

References

[1] Autism Society (CDC, 2014) - www.autism-society.org/what-is/facts-and-statistics

[2] Centers for Disease Control and Prevention. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2010. MMWR Surveill Summ 2014; 63: 1–21.

[3] L. Catherine, et al. "Autism spectrum disorders." Neuron 28.2 (2000): 355-363.

[4] D. L. Christensen, J. Baio, K. V. N. Braun, D. Bilder, J. Charles, J. N. Constantino, J. Daniels, M. S. Durkin, R. T. Fitzgerald, M. Kurzius-Spencer, L.-C. Lee, S. Pettygrove, C. Robinson, E. Schulz, C. Wells, M. S. Wingate, W. Zahorodny, and M. Yeargin-Allsopp, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years – Autism and developmental disabilities monitoring network," *MMWR. Surveillance Summaries*, vol. 65, no. 3, pp. 1–23, 2016.

[5] C. Tan, V. Frewer, G. Cox, K. Williams, and A. Ure, (2021). Prevalence and age of onset of regression in children with autism spectrum disorder: A systematic review and meta-analytical update. *Autism Research*, *14*(3), 582-598.

[6] N. Cardinal, N. Donald, A. J. Griffiths, D. Zachary, and J. Fraumeni-McBride. "An investigation of increased rates of autism in US public schools." *Psychology in the Schools* 58, no. 1 (2021): 124-140.

[7] S. Ozonoff, G. S. Young, A. Carter, D. Messinger, N. Yirmiya, L. Zwaigenbaum, S. Bryson, L. J. Carver, J. N. Constantino, K. Karen Dobkins, T. Hutman, J. M. Iverson, R. Landa, S. J. Rogers, M. Sigman, and W. L. Stone, "Recurrence risk for autism spectrum disorders: A Baby Siblings Research Consortium study," *Pediatrics*, vol. 128, no. 3, pp. e488–e495, 2011.

[8] S. Sumi, H. Taniai, T. Miyachi, and M. Tanemura,“Sibling risk of pervasive developmental disorder estimated by means of an epidemiologic survey in Nagoya, Japan,” *Journal of Human Genetics*, vol. 51, no. 6, pp. 518–522, 2006.

[9] S. B¨olte, “Is autism curable?” *Developmental Medicine & Child Neurology*, vol. 56, no. 10, pp. 927-931, 2014.

[10] W. Partington, James, The Assessment of Basic Language and Learning Skills-Revised (the ABLLS®-R): An Assessment, Curriculum Guide, and Skills Tracking System for Children with Autism Or Other Developmental Disabilities. *Behavior Analysts*, 2010.

[11] PicTalk – to be available to google play soon

[12] Teach Autistic Children - https://play.google.com/store/apps/details?id=com.autisme.myApp

[13] Otsimo - <https://play.google.com/store/apps/details?id=com.otsimo.app>

[14] Talking Pictures - <https://play.google.com/store/apps/details?id=ru.igorsh.kidcommunicator>

[15] LetMeTalk - https://play.google.com/store/apps/details?id=de.appnotize.letmetalk

Rashid Al Shafee obtained his B.Sc Engineering Degree in Computer Science and Engineering (CSE) from International Islamic University, Chittagong (IIUC) in 2018. He is a member of the IIUC Data Science Research Group. He is currently working as a Software Engineer in Kinetik. Previously, after graduation, he completed his Internship at Samsung R&D Institute Bangladesh and worked as a Software Engineer for three years there.

Rakibul Huda completed his B.Sc Engineering in Computer Science and Engineering (CSE) from International Islamic University, Chittagong (IIUC) in 2018. He is a member of the IIUC Data Science Research Group. He is currently working as a Software Engineer in Exabyting.

Mohammad Imran Hossain obtained his B.Sc Engineering Degree In Computer Science and Engineering (CSE) from International Islamic University, Chittagong (IIUC) in 2018. He is a member of the IIUC Data Science Research Group and has a published paper in a peer-reviewed journal. He is currently working as a software engineer in a renowned software company, in Bangladesh.

Md. Mahmudul Hasan obtained his B.Sc Engineering Degree in Computer Science and Engineering (CSE) from International Islamic University, Chittagong (IIUC) in 2018. He is a member of the IIUC Data Science Research Group and has a published paper in a peer-reviewed journal. He is currently working as a software engineer in Softzino Technologies.

Dr. Shahidul Islam Khan obtained his B.Sc. Engineering Degree in Computer Science and Engineering (CSE) from Ahsanullah University of Science and Technology (AUST) in 2003. He obtained his M.Sc. and Ph.D. from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2011 and 2020. His current fields of research are Data Science, Database Systems, Machine Learning, Information Security, and Health Informatics. Currently, he is serving as the Head of the IIUC Data Science Research Group. He has more than fifty published papers in peer-reviewed journals and at reputed international conferences. He is also an Associate Professor in the Dept. of CSE, International Islamic University Chittagong (IIUC), Bangladesh.

Solutions for Relaunching Art Consumption After COVID-19 - From the Perspective of Consumers With Higher Education

Iuliana COMAN
The Bucharest University of Economic Studies, Romania
iuliana.coman.ards@gmail.com

The Covid-19 pandemic brought major changes to most areas of activity. Art was no exception and faced significant changes in both consumer behavior and the behavior of art producers who had to adapt to the difficulties of this period. The paper aims to present the image of art consumer behavior, including the socioeconomic context generated by the coronavirus pandemic in Romania, and to analyze the possible relaunching measures that can be taken for the restoration of the art market after the coronavirus pandemic. Another goal of the paper is to open this subject for future analysis, underlining the influences that art manifests in society. The analysis uses macroeconomic indicators provided by the National Institute of Statistics, Eurostat, estimates of companies playing in the Romanian market, and a survey conducted in the first week of May 2020, during the COVID-19 crisis, on a sample of 200 persons from the south of Romania. The survey goal was to capture the image of art consumer behavior, the influence that art has on the lives of respondents as well as the respondents' attitude towards the possible relaunching measures that can be taken for the restoration of art consumption after the coronavirus pandemic. Art continued to influence the lives of individuals and society during the COVID-19 period, with a wide range of roles played in the evolution of society. The online promotion of all art forms was the relaunch measure that was best received by most of the respondents. In the assessment of the possible relaunch measures, an important role is played by the presence of art in the respondents' lives and their convictions regarding the influences of art upon society.

Keywords*:* Art Consumption, COVID-19, Solutions for relaunching art consumption

1 Introduction

Art was an important component of human history, participating in the construction of society and to the development of people as individuals.

An analysis of art consumers' behavior during the COVID-19 period in Romania is included in this paper. The research also includes an analysis of a series of recovery measures proposed for supporting the art during this crisis period. The study considered the macroeconomic indicators calculated by the National Institute of Statistics and Eurostat, and the results of the survey conducted on a sample of 200 people with high education in southern Romania. The objectives of the survey were the analysis of the changes in the behavior of art consumers in the crisis period and an analysis of the relaunching solutions for art consumption after this pandemic period.

2. Literature Review

The evolution of the art sector during the pandemic period was reflected in a series of studies published during the last years.

[12] analyze the difficulties faced by the arts and creative industries during the COVID-19 pandemic. These were severely affected by the COVID-19 pandemic, a solution for the rebirth of these sectors being cultural tourism. [4] examine the short-term impact of the pandemic on self-employed people in Canada

and finds that the occupations with the largest decreasing number of hours worked during the pandemic period are art, culture, and recreation.

Radermecker in [16] considers that the arts and culture sector has faced a paradoxical situation caused by the pandemic: the demand for cultural and creative content has increased, but traditional consumption patterns have been severely affected. Radermecker believes that consumers will be the main vector that can participate in the return of this sector and identifies four directions of research that will create the necessary framework for development: the collection of data on consumer cultural practices, consumers and the digital cultural experience, consumer involvement and loyalty in art and culture, and consumer welfare.

[11] analyzes participatory online exhibitions in China and claims that these exhibitions published by WeChat in China during the pandemic era helped to create a space for expression that responded to the need for information, and at the same time created alternative ways of understanding and expressing the crisis.

[5] evaluate the government measures for cultural and creative organizations in 5 European countries (the Czech Republic, the Netherlands, Portugal, Slovenia, and Switzerland) during the pandemic and notes that the intensity of state involvement, the economic situation before the pandemic, and the tendency to engage in own account in society are the key factors for understanding the adoption of specific measures in each of the countries included in the research.

Studying the evolution of Italian state museums during the pandemic, the period in which museums were forced to close the physical exhibitions, [1] notes a sharp increase in online cultural material and initiatives, which take place through social networks. During the analyzed period museums doubled the online activity. The study included 100 Italian state museums.

Analyzing state interventions to support the arts and culture sector, [3] found that in the UK there is considered that professional arts and culture need to develop a new strategy that can enhance the potential and value of culture. This new strategy is intended to help the sector become involved in the many social, economic, and environmental challenges that follow beyond COVID-19.

[6] underline the importance of face-to-face interaction in the art sector, with a focus on art market transactions. In the absence of the possibility of physical presence, during the pandemic, a transition to remote online communication is created, but this solution is considered that will disappear depending on the evolution of the virus, continuing to remain present only in the case of some segments.

Considering the pandemic art market, [10] found that art is a sector that has found solutions and managed to respond well to the crisis. The art sector was capable to adapt and overcome the challenges imposed by the new pandemic situation. Thus, contemporary art galleries expanded their digital activities, participated in art fairs, continued their international development strategy, and managed to obtain government support for the crisis period.

[13] consider that in the context of the difficulties generated by the pandemic to the arts sector, the decentralized approach of the United States to funding culture has undermined the ability of cultural organizations to respond to issues of public relevance and demonstrate their civic value, threatening their legitimacy.

3. Objective and Methodology of Research

This paper aims to evaluate the evolution of the art consumers' behavior during the COVID-19 crisis, presenting also the socio-economic context generated by Coronavirus Crisis in the Romanian society. The objectives of the paper also include the presentation of some identified solutions for art relaunching after this crisis period and the evaluation of the influences manifested by demographic factors or convictions in the evaluation of these solutions. Promoting the importance of art in the life of individuals and society is another objective of this paper, which aims to open the topic for future research.

The socio-economic context will be presented taking into consideration different macroeconomic indicators.

For this analysis, a survey was organized, during the first week of May 2020 on a sample of 200 people in the south of Romania. Data collection was done using TalkOnlinePanel, a company specialized in online surveys.

The sampling rates used for defining the sample were the next ones: education - persons with university studies: gender - 50% men and 50% women; age, 35% of respondents with ages between 31 and 40 years, 47% with ages between 41 and 50 years, and 18% others.

The sample comprised people with higher education for guaranteeing, therefore, a consistent contact of the respondents to art and consequently a deeper assessment of the influence of art on the life of individuals and society. The influence of education on the behaviour of art consumers was also confirmed by the Cultural Barometer 2018 that mentions that persons with higher education have the highest contribution to cultural consumption [7], [8]. For the development of the questionnaire, studies and analyses carried out in the field of art were considered, allowing the documentation of the main influences that are present in this field. Among the factors studied are the following ones: consumption of art - with the frequency of consumption, the presence of art in the lives of the respondents, the convictions regarding the participation of the art in the life of individuals and the consolidation the society, the opinions regarding the different solutions for art recovery after COVID-19 crisis.

The method used to identify the relationships between factors was the Chi-Square method and Cramer’s V test. The test was introduced by Karl Pearson and allows the verification of the hypothesis of an association between the variables generated by responses obtained from two different questions.

For calculating the Chi-Square indicator, computed using EXCEL, contingency tables were organized by intersecting the answers to two questions: X - with the alternatives Xi, placed as rows of the table, and Y - with the alternatives Yj, placed in columns [14].

The next steps were as follows:

- The formulation of the null hypothesis H0, which states that between the two variables-segmentation questions there is no causal link or association;
- Choosing the significance level or threshold α and calculating the number of degrees of freedom of the table according to the formula (r-1) (c-1); based on these data, one assumes from the table of distribution $\chi 2$ its value, theoretically (index t);
- Calculating the expected theoretical frequencies (expected, in case of a homogeneity test), according to the following formula:
  $\theta ij=(\text{Total Line I x Total Line J}) / \text{Total} = Ti.\ x\ T.j$
- Calculation of $\chi 2$ index using the formula:
  $\chi 2=\sum ri=1\sum cj=1\ [(xij- \theta ij)2/ \theta ij]$
- $\chi 2$ is compared with the one obtained from the distribution table $\chi 2$ as follows:
  - if $\chi 2$calculated > $\chi 2$teroretical, the null hypothesis is rejected and, therefore, there is an association or potential relationship between the studied segmentation variables;
  - if $\chi 2$calculated < $\chi 2$teroretical, the null hypothesis is accepted, and therefore there is no association or potential relationship between the studied segmentation variables.
- After identifying the existence of the association between the segmentation variables, we used the Cramer's V test to verify how strong the connection between the two variables is.

$V=\sqrt{(\chi2/[(N)Min\ (r-1,\ c-1))}$

The scale of values that Cramer's V can have is the following:

≤0.10 there is no association
>0.10 and ≤0.30 weak association
>0.3 and ≤0.50 moderate association
>0.5 and ≤0.70 strong association
>0.70 very strong association

4. Results and Discussion

4.1 Art Consumption during COVID-19 Crisis

During the COVID-19 crisis, according to GfK reports (2020), 83% of consumers worldwide have changed their behavior. The main change in consumer behavior was the move of purchases in the online environment, so in the first half of 2020, there is an increase of 64% for online sales and a decrease of 5.4% in offline sales compared to the same period last year.

[15] considers that the COVID-19 crisis has generated several positive aspects, including the development of the digital sector, the development of the medical field, and the increase in sales in the luxury art market. At the same time, [9] point out that the sectors most affected by the pandemic's movement restrictions are the arts, entertainment, and recreation sectors.

In the Romanian market, the image of the evolution of the Romanian art consumption is presented by the macroeconomic indicators available at the National Institute of Statistics (Romania), for the domain Culture. They show a market with growth trends manifested in the last 10 years for spectators and auditors at artistic performances, spectators at cinemas, and visitors to museums and public collections. During the same period, we are witnessing a fluctuating evolution for the number of active readers in libraries, the number of printed books and brochures entered in the Legal Deposit of the National Library, and for the production of newspapers, magazines, and other periodicals.

In 2019 there was an increase of approximately 2% in the number of spectators and auditors at artistic performances, an increase of 3% in the number of museum visitors and public collections. The active readers in libraries and the number of spectators in cinemas registered a decreasing evolution.

Table 1. Evolution of the art consumption – INS Romania Indicators

	2019	**Growth indicators 2019 vs 2018**
Population residing on January 1 in Romania	19,414,458	-0.61%
Spectators and listeners at artistic performances	8,074,487	1.94%
Active readers at libraries	3,101,970	-0.96%
Spectators at the cinemas	13,130	-1.63%
Visitors to museums and public collections	18,197,586	3.34%
Printed books and brochures, entered in the Legal Deposit of the National Library	19,604	174.91%
Production of newspapers, magazines and other periodicals	1,446	-54.49%

A detailed image of art consumption in Romania from the perspective of artwork auctions is available in the report “Perspectives of the Romanian Art Market 2020-2021 in the context of the pandemic” developed by ARTMARK. The report presents the evolution of the Romanian art market in recent years and an estimate of the evolution for 2020. In Romania, the art market has experienced annual growth of

22% in the last five years, until the end of 2019. The year 2020 begins under the auspices of a positive end to 2019, signaling trends of moderate organic growth in all directions.

During the crisis, of the six auction houses existing in 2019, only three continued their activity in the first period of the pandemic: Artmark, Alis (Bucharest), and Quadro (Cluj), each using its own online facilities, functionally developed before the pandemic context.

The ARTMARK report [2] considers that the main reason for the continued growth of the art market in Romania in 2020 arrives both from the lessons of the economic crisis 2008-2010 and from the economic instincts to preserve the value of money in safe deposits, in the face of the possibility of future inflation. We are witnessing the continued growth of the market, but an atypical growth and only partially organic, less based on the concern for the esthetics needs. Consequently, it is forecast that in the next two years, two different factors will influence the Romanian art market, sometimes cumulatively, sometimes opposingly. These factors are generated by the previous period of increase and by the changes generated by the crisis. The factors could slow down or exacerbate evolution. Is expected a preference of hoarding over investment, therefore, the acquisition of works by well-known authors, not necessarily of heritage, but certainly consecrated works that do not exclude contemporary art.

For a detailed picture of art consumption in Romania in terms of frequency of consumption during the COVID-19 crisis, data provided by the survey organized for this research were used, a survey regarding the behavior of art consumers in southern Romania conducted in the first week of May 2020.

Table 2. Frequencies of Art Consumption

		Man	Woman	Grand Total	Share
Movies on Internet platforms (Netflix, HBO Go, others)	Daily	54	38	92	46%
	Weekly	25	33	58	29%
	Monthly	2	1	3	2%
	Occasional	8	8	16	8%
	Not at all	11	20	31	16%
Theater on Internet Platform (Online TV Stations, Online Plays, Others)	Daily	13	4	17	9%
	Weekly	26	25	51	26%
	Monthly	8	11	19	10%
	Occasional	30	24	54	27%
	Not at all	23	36	59	30%
Music on Internet platforms (Online TV stations, Online plays, others)	Daily	59	54	113	57%
	Weekly	13	20	33	17%
	Monthly	5	6	11	6%
	Occasional	20	7	27	14%
	Not at all	3	13	16	8%
Literature in classic format (books) or electronic	Daily	31	36	67	34%
	Weekly	25	26	51	26%
	Monthly	13	12	25	13%
	Occasional	23	19	42	21%
	Not at all	8	7	15	8%
Fine Arts - through online platforms of major museums	Daily	6	3	9	5%
	Weekly	8	9	17	9%

		Man	Woman	Grand Total	Share
	Monthly	14	13	27	14%
	Occasional	32	34	66	33%
	Not at all	40	41	81	41%
	Daily	3	4	7	4%
	Weekly	3	3	6	3%
	Monthly	1	2	3	2%
	Occasional	8	7	15	8%
	Not at all	41	43	84	42%
Others, which one?	--	44	41	85	43%

Source: author's own research

For the analysis of the frequencies of consumption during the COVID-19 pandemic, the Internet platforms developed for offering access to different forms of art were taken into consideration. The survey shows that 75% of the respondents consumed movies online, using internet platforms, with at least weekly frequency. In the case of men, 79% of the respondents consumed movies with a frequency of at least weekly using the Internet platforms. The theater also continued to be consumed during the pandemic crisis. 44% of the respondents consumed monthly theater using the Internet platforms and 47% of men consumed at least monthly theater using the various existing platforms. Music was the form of art used with daily frequency, 57% of the respondents consumed music every day via different platforms of online broadcasting, and 59% of men consumed music daily. The literature was also included in the lives of the respondents. 59% of the respondents read at least weekly literature, traditional books, or in an electronic form like e-books, and 62% of women read literature at least weekly. The fine arts were also available online during this pandemic period, and numerous museums revealed their galleries in virtual tours. 59% of the respondents accessed different galleries and other Internet platforms for works of art. 60% of men accessed different online platforms used for exposure of the works of art. Other forms of art, except consumed during the COVID-19 crisis, except those presented above, are dance, culinary art, and photography. 16% of the respondents consumed these other forms during this COVID-19 pandemic.

4.2 Solutions for relaunching art consumption after COVID-19

The above analyzes reflect a continuation of art consumption in the various forms existing in the Romanian market and globally. At the same time, there was a continuing concern to establish ways in which consumer behavior could approach the classic form known before the crisis.

The survey carried out in the south of Romania for this research analyses a set of recovery measures proposed for the art domain in Romania. These measures include the reopening of shows, theaters, cinemas, concerts, without physical distancing measures or the opening considering physical distancing measures; the reopening of museums and exhibitions; the online promotion of various art forms, grants for theater, cinema, concrete performances, etc. to be watched online for free or at affordable prices and the financial support for artists and art producers, to get over the crisis more easily.

To what degree these relaunching measures after the crisis period were considered efficient by the respondents, is illustrated in Table no. 3, a table that shows the distribution of the respondents' answers according to their agreement regarding each of the proposed measures.

Table 3. Relaunching measures and degree of agreement

Degrees of the agreements for each proposed measure	**Number of answers**
Reopening of shows (theatres, cinemas, concerts, etc.) without any restrictions;	**200**
Total disagreement	20
Partial disagreement	21
No agreement, no disagreement	40
Partial agreement	55
Total agreement	64
Reopening of shows (theatres, cinemas, concerts, etc.) with restrictions on social distance (more space between spectators, etc.);	**200**
Total disagreement	8
Partial disagreement	8
No agreement, no disagreement	36
Partial agreement	72
Total agreement	76
Reopening of museums/exhibitions;	**200**
Total disagreement	9
Partial disagreement	11
No agreement, no disagreement	32
Partial agreement	76
Total agreement	72
Online promotion of all art forms;	**200**
Total disagreement	5
Partial disagreement	7
No agreement, no disagreement	30
Partial agreement	55
Total agreement	103
Subsidies for theatre, cinema, concrete performances, etc. to be able to be watched online for free or at affordable prices;	**200**
Total disagreement	10
Partial disagreement	7
No agreement, no disagreement	41
Partial agreement	55
Total agreement	87
Financial support for all artists, to get over the crisis more easily;	**200**
Total disagreement	10
Partial disagreement	6
No agreement, no disagreement	44
Partial agreement	60
Total agreement	80
Other	**200**
Total disagreement	15
No agreement, no disagreement	43
Partial agreement	9
Total agreement	34
-	99

Degrees of the agreements for each proposed measure	Number of answers
Grand Total	**200**

Source: author's own research

The measures that received the agreement of the largest number of respondents were the online promotion of various art forms (79% of respondents expressed their agreement with this measure); followed by the reopening of shows, theaters, cinemas, concerts taking into account the measures of physical distance (74% of respondents agreed on this measure) and the reopening of museums and exhibitions (74% of respondents agreed with this measure).

The proposal to reopen theaters, cinemas, or concerts without restrictions on physical distance was received with skepticism. 20% of the respondents disagreed with this proposal and 20% were undecided. The number of people who expressed their partial or total agreement was the lowest compared to the other proposed measures.

The reopening of museums and exhibitions is another measure considered successful by respondents. 74% of the respondents agreed and 10% disagreed with this measure.

The online promotion of all art forms was the measure best received by most respondents. 52% of the respondents strongly agreed on this measure, and 28% of respondents agreed in part. Only 6% of the respondents disagreed with this measure.

Subsidy proposal for theater performances, cinema, concerts, etc. to be able to be watched online for free or at affordable prices was considered a good measure by 71% of respondents, to a greater or lesser extent. 8% of the respondents disagreed with this measure.

Financial support for all artists, to get over the crisis period more easily, was the measure for which 70% of the respondents expressed their total or partial agreement. 8% of the respondents disagreed with this measure.

To have a deeper image of the assessment of the relaunching methods proposed, the influence of demographic factors on the evaluation of each of the proposed relaunching solutions was also analyzed. Table 4 shows the influences of gender, age, and education on measures to return to art after the crisis.

Table 4. The influence of age, gender, and education in the assessment of relaunching measures

Proposed Measures	Indicators	Age	Gender	Education
Reopening of shows (theaters, cinemas, concerts, etc.) without any restrictions	χ2	1.85	0.75	2.57
	Cramér's V	**0.17**	0.06	**0.11**
Reopening of shows (theaters, cinemas, concerts, etc.) with restrictions on social distance (more space between spectators, etc.);	χ2	2.55	0.00	0.23
	Cramér's V	**0.16**	0.00	0.03
Reopening of museums/exhibitions;	χ2	3.12	1.30	1.50
	Cramér's V	**0.12**	0.08	0.09
Online promotion of all art forms;	χ2	5.55	0.79	5.75
	Cramér's V	**0.17**	0.06	**0.17**
Subsidies for theater, cinema, concrete performances, etc. to be able to be watched online for free or at affordable prices;	χ2	0.63	0.11	2.87
	Cramér's V	0.06	0.02	**0.12**
Financial support for all artists, to get over the crisis more easily;	χ2	0.85	0.54	6.87
	Cramér's V	0.07	0.05	**0.19**

Proposed Measures	Indicators	Age	Gender	Education
Other	χ2	10.35	1.24	5.58
	Cramér's V	**0.23**	0.08	0.**17**

Source: author's own research

The influences manifested by gender, age, or education on the assessment of relaunching measures for art after the crisis caused by COVID-19 are very small. Gender has no influence on recovery measures after the crisis, and none of the variables generated by the responses regarding the relaunching measures are in association with the variable generated by the gender of the respondents. Education influences to a small extent some of the beliefs about the recovery measures after the pandemic period: the reopening of shows without any restrictions, the online promotion of all art forms, subsidies for spectacle performances, for being available online, and the financial support for all artists to overcome the crisis.

Convictions regarding the efficiency of reopening theaters, cinemas, or concert performances without any restrictions are influenced to a small extent by the education and age of the respondents. Convictions regarding reopening theatrical performances, movies, or concerts with social distancing restrictions are influenced by the age of the respondents. Convictions about reopening museums or exhibitions are also influenced by the age of the respondents. The assessments regarding the online promotion of all art forms or other measures that can be taken for the return of the art field after the crisis period have been influenced to a small extent by both the education and the age of the respondents. Opinions about the suitability of grants so that theater, cinema or concrete performances can be watched online for free or at affordable prices, and financial support for all artists to overcome the crisis more easily were influenced by the education of respondents.

If in the case of demographic variables, the influences were very weak in the case of beliefs related to the presence of art, the usefulness of art, the influence of art on the life of the respondents and society, the influences are substantially more significant.

Table 5. The Influences of convictions regarding art upon the relacunhing measures

Relaunching measures\ Convictions regarding art		**Do you consider that art is useful**	**Do you consider that art influences your life**	**Do you consider that art influences life in the society in which you live**	**Do you consider that art has been present in your life**
Reopening of shows (theaters, cinemas, concerts, etc.) without restrictions on social distance	χ2	*10.99*	*8.28*	*10.42*	*4.85*
	Cramer's V	0.23	0.20	0.23	0.16
Reopening of shows (theaters, cinemas, concerts, etc.) with restrictions on social distance	χ2	*17.41*	*9.91*	*8.86*	*14.50*
	Cramer's V	0.30	0.22	0.21	0.27
Reopening of museums/exhibitions	χ2	*17.42*	*11.79*	*13.53*	*3.26*
	Cramer's V	0.30	0.24	0.26	0.13
	χ2	*25.37*	*32.48*	*35.07*	*7.74*

Relaunching measures\ Convictions regarding art		**Do you consider that art is useful**	**Do you consider that art influences your life**	**Do you consider that art influences life in the society in which you live**	**Do you consider that art has been present in your life**
Online promotion of all art forms;	Cramer's V	0.36	0.40	0.59	0.20
Subsidies for theater, cinema, concrete performances, etc. to be able to be watched online for free or at affordable prices;	χ2	*30.24*	*31.52*	*22.07*	*13.89*
	Cramer's V	0.39	0.40	0.33	0.26
Financial support for all artists, to get over the crisis more easily;	χ2	*18.82*	*24.26*	*19.27*	*8.04*
	Cramer's V	0.31	0.35	0.31	0.20
Others, which one?	χ2	*4.79*	*0.95*	*1.41*	*9.45*
	Cramer's V	0.15	0.07	0.08	0.22

Source: author's own research

The influence of the convictions regarding the usefulness of art, the presence of art in the life of the respondents or regarding the capacity of art to influence the life of society or the individual, have notable influences on the next recovery measures: online promotion for all art forms, subsidies for theater performances, cinema, concerts to be available online, and on the proposal of financial support for all artists to pass more easily over the crisis period. These three measures were most strongly influenced by beliefs about the usefulness, presence, or capability of the art of influencing the lives of the respondents.

Taken individually, each of the relaunching measures after the crisis period was influenced by the respondents' convictions about art. Only in the case of options for other measures of recovery was there a lack of association between beliefs about the ability of art to influence the life of the individual and society and this measure of relaunching.

The measure of reopening performances (theaters, cinemas, concerts, etc.) without social distance was the one most strongly influenced by beliefs about the ability of art to influence society and beliefs about the usefulness of art. The measure of reopening performances (theaters, cinemas, concerts, etc.) with restrictions on social distance was most strongly influenced by beliefs about the usefulness of art and the perception of the presence of art in the lives of respondents.

The reopening of museums and exhibitions has been most strongly influenced by beliefs about the ability of art to influence society and beliefs about the usefulness of art.

The online promotion of all art forms was influenced by the beliefs about the usefulness of art, the ability of art to influence society, but also the lives of respondents. These convictions also had the strongest influences on the appreciations regarding the subsidy’s proposal for the theater, cinema, concerts to be available to be seen online for free or at affordable prices and the financial support for all artists, to overcome the crisis period more easily.

The presence of art, or convictions regarding the usefulness of art, and the ability of art to influence the lives of individuals and society lead to changes in the acceptance and evaluation of proposals to recovery the art consumption. The age of the respondents or the education of the respondents have very little influence or do not influence the way of evaluating the relaunching measures after the crisis period.

Limitations of the research

The study on relaunching measures for art consumption after the crisis period aims to be an invitation for discussions on this topic, an invitation that emphasizes the role that art plays in the life of society. This communication is a researcher speech, an example of a simple approach starting from the statistical composition of the percentages of contingency tables of the variables found in interactions or indifference. The survey regarding the image of the Romanian art consumption during the COVID-19 crisis has no statistical guarantees, the research offered just an image regarding the behavior of the 200 respondents included in the sample.

5. Conclusions

During the pandemic, art has continued to be present in our lives, of each of us, of society, and of the community we belong to, and has played a multitude of roles from responding to the need for beauty to participating in the evolution of humanity or to the involvement in contemporary life. Art was a part of the development of the creative business or educational models and participated in the consolidation of society by the support it provides in times of crisis. Art participates in accurately reproducing the image of different periods of crisis and can participate in preventing these periods, but also in finding solutions to the difficulties raised by these crises.

This research opens the topic of the importance of art in society for future analysis. It offers an image of art consumption in Romania and the context generated by the COVID-19 crisis, and an analysis of a set of measures to return art consumption after the crisis period.

The study conducted on a sample of 200 people with higher education in southern Romania reveals an art consumption that continues in times of crisis, using online platforms developed to provide access to art in various forms: cinema, theater, virtual gallery tours, and broadcasting of musical pieces. Most of the measures to support the field of art to revitalize art consumption after the crisis period were considered useful by respondents.

Thus, the reopening of performance halls with or without physical distance, the reopening of museums or exhibition halls, the online promotion of various art forms, the subsidy of theaters, cinemas, or concert halls to create solutions to be accessible online, the financial support offered to the artists to pass this period more easily, all these measures were considered to participate in the relaunch of art consumption. Over 70% of the respondents in the survey considered that these measures will lead to the resumption of art consumption, except for the proposal to reopen theaters or cinemas without physical distance, a measure for which only 60% of the respondents considered that they can participate in the relaunch of art consumption. The main factors that influenced the evaluation of the proposed relaunch measures were the respondents' beliefs about the usefulness of art, the ability of art to influence the lives of individuals and society, and the presence of art in the lives of respondents. Demographic factors, such as gender, education, or age, had very little or no influence on the evaluation of the proposed relaunch measures.

References

[1] Agostino D., Arnaboldi M., Lampis A. (2020). Museums during the global Covid-19 pandemic. Italian state museums during the COVID-19 crisis: from onsite closure to online openness. Museum Management and Curatorship, Volume 35, 2020 - Issue 4, Pages 362-372 | Received 29 May 2020, Accepted 28 Jun 2020, Published online: 13 Jul 2020

[2] ARTMARK Report (2020). ARTMARK Report [Online]. Available at:

https://www.artmark.ro/en/about-us/art-market-reports/ (Accessed: 6 Oct 2020)

[3] Banks M. and O'Connor J. (2020) "A plague upon your howling": art and culture in the viral emergency. Cultural Trends Volume 30, 2021 - Issue 1: Art and culture in the viral emergency. Pages 3-18

[4] Beland L.P., Fakorede O. and Mikola D. (2020) Short-Term Effect of COVID-19 on Self-Employed Workers in Canada, UTP Journals, Volume 46 Issue S1, pp. S66-S81 https://doi.org/10.3138/cpp.2020-076

[5] Betzler D., Loots E., Prokůpek M., Marques L. and Grafenauer P. (2021) COVID-19 and the arts and cultural sectors: investigating countries' contextual factors and early policy measures, International Journal of Cultural Policy, Volume 27, 2021 - Issue 6, Pages 796-814

[6] Buchholz L., Fine G.A. and Wohl H. (2020) Art markets in crisis: how personal bonds and market subcultures mediate the effects of COVID-19. American Journal of Cultural Sociology volume 8, pages462–476

[7] Croitoru, C., Becuț Marinescu A. (2017). 'Barometrul de Consum Cultural 2016: O radiografie a practicilor de consum cultural', Universul Academic, Bucharest

[8] Croitoru, C., Becuț Marinescu A. (2019). 'Barometrul de Consum Cultural 2018: Dinamica sectorului cultural în anul Marii Uniri', Universul Academic, Bucharest

[9] Danieli, A., Olmstead-Rumsey, J. (2020). 'Sector-Specific Shocks and the Expenditure Elasticity Channel During the COVID-19 Crisis,' [Online], [Retrieved October 2, 2020], Available at SSRN: https://ssrn.com/abstract=3593514 or http://dx.doi.org/10.2139/ssrn.3593514 (Accessed: 6 Oct 2020)

[10] Duarte A., Fialho A.L. and Pérez-Ibáñez M. (2021) External Shocks in the Art Markets: How Did the Portuguese, the Spanish and the Brazilian Art Markets React to COVID-19 Global Pandemic? Data Analysis and Strategies to Overcome the Crisis. Arts 2021, Special Issue Global Art Market in the Aftermath of COVID-19. 10(3).

[11] Feng X. (2020) Curating and Exhibiting for the Pandemic: Participatory Virtual Art Practices During the COVID-19 Outbreak in China. Soc Media Soc,6(3).

[12] Flew T., Kirkwood K. (2020). The impact of COVID-19 on cultural tourism: art, culture and communication in four regional sites of Queensland, Australia, Research Article, https://doi.org/10.1177/1329878X20952529

[13] Katz S.N. and Reisman L. (2020). Impact of the 2020 crises on the arts and culture in the United States: The effect of COVID-19 and the Black Lives Matter movement in historical context. International Journal of Cultural Property. Volume 27 Issue 4. pp. 449 - 465

[14] Mihaita, N., Stanciu-Capota, R. (2005). 'Relations statistiques fortes, cachees, fausses, et illusories Applications de la statistique informationelle'. ASE, Bucharest

[15] Puaschunder, J. M., (2020). 'Value at COVID-19: Digitalized Healthcare, Luxury Consumption and Global Education'. Proceedings of the ConScienS Conference on Science & Society: Pandemics and their Impact on Society, September 28-29, 2020., pp. 43-51.

[16] Radermecker A.S. (2021) Art and culture in the COVID-19 era: for a consumer-oriented approach. SN Business & Economics volume 1

Oracle Machine Learning for Python in APEX - Analyzing and Predicting CO2 Emission by private vehicles

Miruna Teleașă[1], Alexandra Teodora Bardici[2]
[1,2]Bucharest Academy of Economic Studies
teleasamiruna19@stud.ase.ro, bardicialexandra19@stud.ase.ro

Nowadays, the global warming threat is a highly discussed matter. One of the factors that accelerates this process is the air pollution that can be caused by cars' emissions. This paper concerns how the size of the engine, the type of fuel, the fuel consumption and the transmission type influence the emission of CO2. In order to understand and predict that variable, we used several machine learning algorithms, such as Regression for Generalized Linear Model or K-Means for Hierarchical Cluster Model. The technology that empowered this analysis was Oracle's Machine Learning for Python (OML4Py) that allowed us to integrate both database and data management concepts and data analysis algorithms. By doing that, we managed to discover a pattern for the emission of CO2 based on the factors previously mentioned and, after that, predict future levels of CO2 emissions for various car models.

Keywords: *Machine Learning Algorithms, Python, Oracle Autonomous Database, Environment, Regression, K-Means*

1 Introduction

The analysis we are going to perform has the goal of estimating the influence of different vehicle characteristics on the carbon dioxide emissions produced by it, with the purpose of providing a guideline for the most environment friendly vehicles. This is done with the hope that, in the future, car manufacturing companies, on the one hand, and vehicle buyers, on the other hand, would make conscious choices when it comes to vehicle characteristics and attributes, providing the best alternatives to eco-friendly vehicles. We chose these data considering that, in Europe, for example, passenger cars are the largest air polluters, having accounted for 60.7% of total road transport emissions in 2016 [1]. Considering also that the entire transport sector accounts for 21% of the total carbon dioxide emission [2], we can safely say that the personal vehicles make up 12.74% of the total emissions worldwide. This is an impactful percent and understanding the way in which these vehicles can be altered to create more eco-friendly ones could lead to a better, safer world [1]. In order to do this, it is important to understand the importance that various technical attributes of the vehicles have when it comes to the carbon dioxide they emit.

To achieve the previous result, we used a data set provided by the Canadian Government, which contains the model-specific estimated carbon dioxide emissions for new light-duty vehicles for retail sale in Canada in the year 2022 [3]. This dataset contains columns for the Make and Model and Class of the vehicle, all three being qualitative variables, considering for the Model variable six different types: Four-wheel drive (4WD), All wheel drive (AWD), Flexible-fuel vehicle (FFV), Short Wheelbase (SWB), Long Wheelbase (LWB) and Extended Wheelbase (EWB). Moreover, a series of quantitative variables were also considered, such as the Engine size measured in litres, the number of Cylinders of the engine, the combined Fuel Consumption measured in litres/100km and, finally, the CO2 Emissions, which is our considered dependent variable, measured in g/km. Last but not least, the Transmission and Fuel Type were also considered, both qualitative variables, first being either Automatic (A), Automated Manual (AM), Automatic with Select shift (AS), Continuously Variable (AV), Manual (M), or measured in the number of gears, a number from 3 to 10. The Fuel Type, however, is considered either Regular (X) or Premium (Z) Gasoline, Diesel (D), Ethanol (E) or Natural Gas (N).

In order for the machine learning algorithms to be applied on our set of data, a software for data analysis was needed that would support large sets of data and would enable us to use a programming language, for example Python, to conduct the analysis. Therefore, we used Oracle's Machine Learning for Python API (OML4Py), which enables the employment of Python commands for data transformation and analysis, while also providing a statistical approach to data examination. This API interprets Python commands and translates them into operations for SQL in-database execution, thus connecting the two programming languages into a complex environment. The main advantage of this software is that it allows users to operate directly onto the database without using SQL commands, as it transparently translates Python functions into SQL. Also, it provides access to preexistent machine learning algorithms, while minimizing the data movement across platforms, keeping it secure in the autonomous database.

OML4Py environment provides direct access to all Python libraries needed for data analysis and machine learning processes. Out of those libraries, we firstly used the *pandas* library, which is one of the open source Python standard libraries for data importing and manipulation through an indexed data structure called DataFrame [5]. Furthermore, we used the *numpy* standard Python library, a library containing various mathematical functions and data structures that support "vectorized" operations for big data [5]. Another Python specific functionality that was used during the machine learning process was the data plotting and data visualization, for better understanding the way in which our dataset is structured, and the impact of the algorithms applied on it [5]. For all this, the matplotlib library was used. Lastly, the module that made possible the integration between the database and the Python commands was the oml module, which allows the manipulation of Oracle Database objects, such as tables and views, by embedded Python specific commands [6].

The Oracle Autonomous Database is a cloud-base database management technology developed and provided by Oracle that is automated to perform routine tasks like backups and security [4]. The purpose of using this database is to store data and information used by the application and to be able to access it from the front-end. There are two main components on which the autonomous database is built upon: the data warehouse and the transaction processing. While first manages the entirety of not only the data, but also the operations performed on it, like provisions, scaling or configuring, the second provides the analytical and security-driven processes that guarantee the smooth functioning of the app that will be developed based on it.

The application that will be developed contains only a fragment of the entirety of capabilities that this analysis offers, with the purpose of demonstrating the hypothesis. However, very complex and useful applications can be followingly developed on the machine learning algorithms' backbone and making use of the discoveries that will be outlined in this article. The final goal might be developing an application that would be available to every user interested in buying an eco-friendly vehicle and that would allow him to use basic information regarding the said vehicle, like the fuel consumption or engine size in order to find out what volume of emissions the vehicle would produce.

Considering this, to develop the application we used Oracles Application Express. Oracle Application Express is one of those tools, hereinafter referred to as APEX, that has emerged as an integrated development environment, providing not only database management, but also web-development functionalities. The enthralling benefit of using APEX in comparison to other web-development tools is that Oracle has embedded in their technologies a Hyper Text Markup Language (HTML) generator, which allows users to create an application without vast knowledge of HTML, JavaScript, PHP, or CSS. Not only this, but the connection between the database and the web application also empowers users to build a front-end and a back-end application without having to use third parties APIs.

2 Algorithms, Results, and Interpretation

The machine learning process is based on the study of existent datasets from various perspectives by using various algorithms and methods. Some authors might say that training the computer into predicting different data is actually synonym with "the systematic study of algorithms and systems that improve their knowledge or performance with experience" [7]. The dissimilarity between the algorithms is derived from the desired predictions and the type of data analyzed, considering that there are algorithms adapted for qualitative data, quantitative data or for both. By way of illustration, an algorithm adapted for qualitative data is the Natural Language Processing, which, based on sentiment analysis, can predict if a news article is fake or can power various Smart Assistants, like Siri or Alexa. However, the results from the data set approached by this project will be quantitative data types, considering that we are interested in predicting the CO2 emissions of different cars, measured in g/km. In the light of that, we used a series of algorithms adapted to this specific type of data. We decided, for that manner, to use three types of data analysis algorithms, all three having slightly different approaches, but similar purposes.

Firstly, we will apply a simple Regression algorithm, on a model that we will generate, a model that will be considered as linear. Secondly, we will apply a K-Means algorithm, suitable for clustered data, which will imply grouping our data and will result in a classification rather than a prediction. Lastly, we will consider the regression process on a Support Vector Machine model, which is marginally more complex than the linear model, but also generates more accurate and detailed results.

Before applying these machine-learning specific algorithms, it is important to point out that we divided the initial data frame into two different data frames, 80% for training the data and 20% to test the results, and we accustomed both data frames around the dependent variable considered, CO2 emissions.

```
emissions_df = oml.sync(table =
'EMISSIONS_2022')
training_df, testing_df =
emissions_df.split(ratio = (0.8,0.2))

training_x =
training_df.drop(['CO2_EMISSIONS'])
training_y =
training_df['CO2_EMISSIONS']

testing_x = testing_df
testing_y =
testing_df['CO2_EMISSIONS']
```

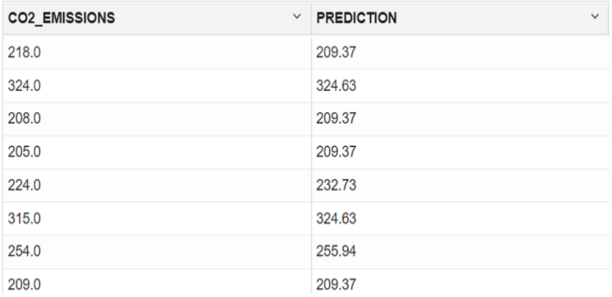
2.1. Regression – Generalized Linear Model (GLM)

A Generalized Linear Model is a certain type of modeling structure adapted to regression analysis, as in this type of analysis, the dependent variable is modeled as a linear function of the independent ones. The main difference between a Linear Model and a Generalized Linear Model is that the latter is more adapted to variables without a specific, normal distribution, thus being more reliable when it comes to big data analysis [8]. In code, the first step of the analysis was to generate the said model, and Oracle’s Machine Learning API provided the oml.glm() method, which takes as parameters the type of analysis carried out, the regression in our case, and the settings needed for this analysis. The result of calling this method is a Generalized Linear Model, upon which the fit() method can be applied, so that it considers the training and testing data we desire to analyze. After fitting the model to the desired datasets, we can further investigate its details to see the basic statistical indicators.

The first indicator observed is the adjusted R2, which stands for the coefficient of determination and marks the percentage of variance in the dependent variable that is influenced by the independent variables, adjusted for the number of observations. The resulting coefficient in the conducted analysis is 98.92%, meaning that almost the entire variation of the CO2 emissions in the analyzed data is explainable by the variables used and that the investigation is statistically significant.

Moreover, the regression coefficients of the generated model, corresponding to an Analysis of Variance (ANOVA) table, coefficients determining the exact dependency function between the CO2 emission and the independent variables used, can be accessed. Consequently, we can better understand the influence each variable has on the final predicted model, and what is the direction in which this influence is oriented. As an example, for the coefficient obtained of the fuel consumption of +24.84, we can say that if the fuel consumption increases by 1 L/100km, considering all other independent variables as constant, the CO2 emissions will increase by 24.84 g/km.

To better understand the model significance and the importance of the predicted results, the residual values of the observation were calculated, that is, the difference between the values that our generated formula predicts and the actual values. For this, the data testing fragment was used, on which the predict() method was applied and the following results were obtained, stating that the error between the actual and the predicted values is relatively small, so the model is fairly correct:



CO2_EMISSIONS	PREDICTION
218.0	209.37
324.0	324.63
208.0	209.37
205.0	209.37
224.0	232.73
315.0	324.63
254.0	255.94
209.0	209.37

Fig. 1. Example of the residuals obtained in the Generalized Linear Model's Regression

Another method used to visualize better the error of the predicted model is a scatter plot, constructed with the seaborn Python library, that plots on one axis the predicted values and on the other the actual values of the dependent variable. Taking this into account, the closer the points are to the regression line, the more fitted and significant the constructed model. Moreover, another fact to take in consideration after examining the scatter plot is that the line that marks the relation between the two variables is positively skewed. That shows that there is a direct, linear, positive relation between the two, further embellishing the idea that the actual emissions and the predicted ones are similar, and that the model we constructed is reliable.

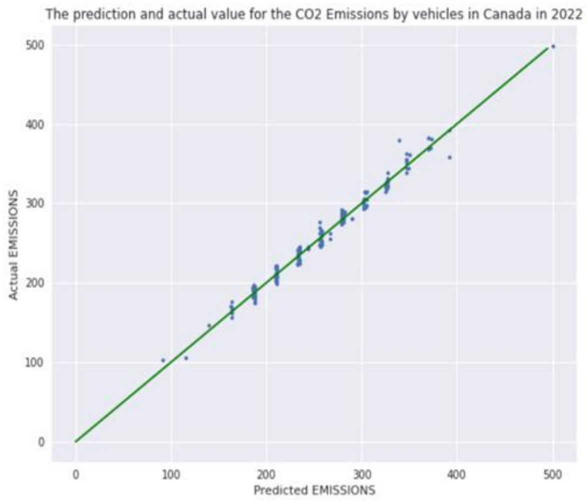


Fig. 2. Scatter plot of the predicted vs the actual values of the CO2 Emissions obtained by regressing the Generalized Linear Model

Finally, the homoscedasticity assumption about the model was tested by a residual plot that graphed the points of the standard residuals obtained and the predicted values. The homoscedasticity is, in broad terms, the constant variance in the errors of the obtained model [8], which is what is to be desired for our analysis. Graphically, in the generated plot, the more randomly distributed along the regression line the points are, the more probable the model is to be homoscedastic and the data to be linear, an assumption that is applicable for our generated residual plot.

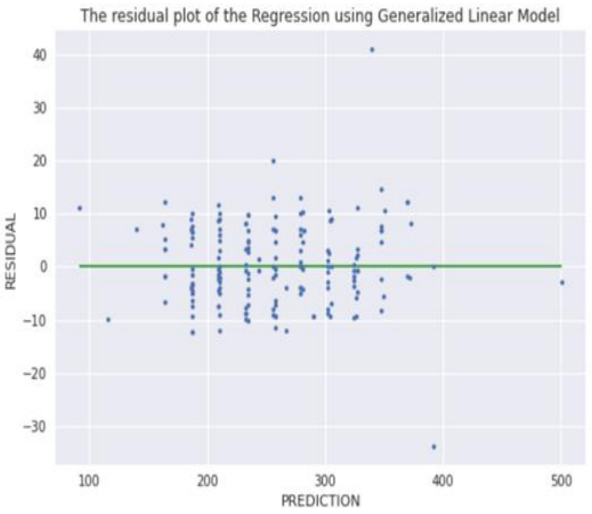


Fig. 3. Residual scatter plot of the CO2 Emissions regression model

2.2. K-means – Hierarchical Cluster Model

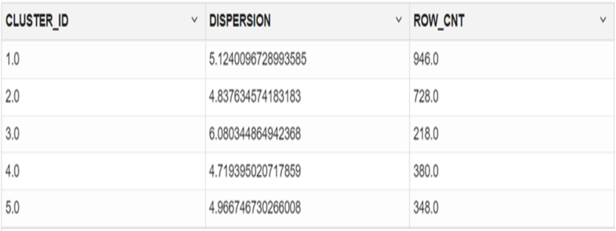
A cluster model is one that divides an unlabeled dataset into different groups based on certain characteristics and maps each observation to a certain group, or cluster, according to specific information subtracted from it. The K-means algorithm is a type of predictive machine learning algorithm that learns a "clustering model from training data that can be subsequently used to assign new (testing) data to clusters" [7]. The said algorithm is a hierarchical, distance-based one, that uses existing observations to predict groups in which future observations will be placed, according to certain variables. For this analysis we considered the groups generated by the Fuel Consumption and the Engine Size variables, considering that both have a positive coefficient in the linear model, thus both positively influencing the CO2 Emissions. This means that the further right a cluster is, the more likely it is to generate higher emissions for vehicles falling within that cluster. For this, we broke down the dataset into a smaller data frame for the predicting set of observations, with the Fuel Consumption and the Engine Size data.

The OML4Py API provides the oml.km class that takes into account a variety of settings with the purpose of generating the clusters. In the context of this analysis, we are using 20 iterations with 3 clusters to be generated, considering that there is a fairly small set of data, with only 946 observations.

```
emissions_df = oml.sync(table =
'EMISSIONS_2022')
training_df, testing_df =
emissions_df.split(ratio = (0.8,0.2))
training_y =
training_df[['FUEL_CONSUMPTION',
'ENGINE_SIZE']]
testing_y =
testing_df[['FUEL_CONSUMPTION',
'ENGINE_SIZE']]
```

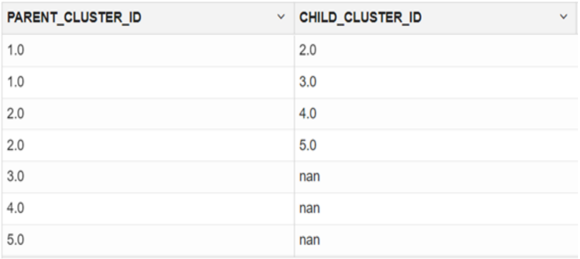
```
setting = {'KMNS_ITERATIONS': 20}
km_mod = oml.km(n_clusters = 3,
**setting).fit(training_df,
model_name="EMISSIONS_KMEANS_CLUSTER_M
ODEL")
```

The generated model can be further analyzed by invoking the clusters and its taxonomy, displayed as a data frame. While the taxonomy table shows the hierarchy of the child clusters in relation to their parents, the clusters table presents various attributes regarding the model, like the dispersion of clusters and the number of observations in each of them.



CLUSTER_ID	DISPERSION	ROW_CNT
1.0	5.1240096728993585	946.0
2.0	4.837634574183183	728.0
3.0	6.080344864942368	218.0
4.0	4.719395020717859	380.0
5.0	4.966746730266008	348.0

Fig. 4. K-Means generated Clusters' Table



PARENT_CLUSTER_ID	CHILD_CLUSTER_ID
1.0	2.0
1.0	3.0
2.0	4.0
2.0	5.0
3.0	nan
4.0	nan
5.0	nan

Fig. 5. K-Means generated Clusters' Taxonomy Table

Here, it can be remarked that the dispersion value, for example, is a variable that measures how spread the observations are inside the cluster, which means that, in our analysis, Cluster 3 is more dispersed than Cluster 4, which is the most compact of the three clusters generated. These certain clusters are observed in the taxonomy table, displayed as having no children, because they are the leaves of the hierarchical cluster tree.

The same conclusion is drawn by looking at the ROW_CNT column of the cluster table, which displays the smallest row count for the 3rd cluster, and the highest for the 4th. However, it is important to be considered that all the three clusters contain similar numbers of observations, which means that the clustering process was a reliable and precise one.

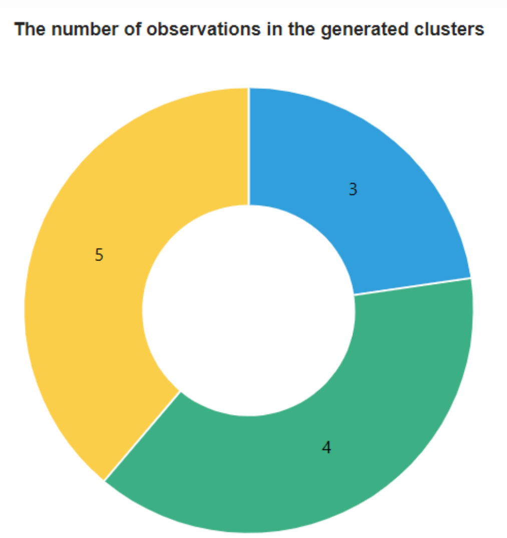


Fig. 6. Donut Chart displaying the size of each cluster

Taking this into consideration, we can say that there is a higher probability for a vehicle to be placed in the 4th cluster than in the 3rd. This is important when observing the predicted variables of the data to be tested, which shows in which cluster is more probable for an observation to fail, according to its fuel consumption and its engine size value, a prediction which can be visualized as a scatter plot.

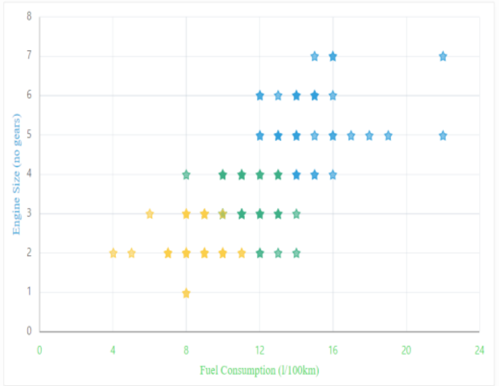


Fig. 7. Scatter Plot of the predicted clusters generated by the k-means algorithm

```
setting = {'svms_kernel_function'
:'dbms_data_mining.svms_linear',
'odms_partition_columns':'FUEL_TYPE'}
svm_mod = oml.svm("regression",
**setting)

svm_mod.fit(training_x, training_y ,
model_name =
'EMISSIONS_SVM_PARTITIONED_REGRESSION_
MODEL')
prediction =
svm_mod.predict(testing_data,
testing_data)
```

2.3. Regression – Support Vector Machine Model

In the previous regression analysis, the model was considered in its entirety, as it built a linear equation that showed the dependence of Emissions on the vehicle's attributes. Oracle's Machine Learning API offers, however, another class to solve a regression model, the Support Vector Machine. This model can be used not only for regression, but also for classification, based on decision planes, and for anomaly detection, which can be used as a security tool. However, SVM regression can be used to better predict an outcome or the value of a dependent variable, taking into account different categories of objects, thus providing a more accurate prediction. For example, our model can be partitioned according to the number of cylinders, the make of the vehicle or even the vehicle class, but for this analysis we used the fuel type.

We considered the observations to be X – regular gasoline, Z – premium gasoline, D – diesel, E – ethanol and N – natural gas., therefore generating five different partitions. According to this, we again divided the model into 80% for the training and 20% for the testing and we created a separate series with the values of the CO2_EMISSIONS column, which we dropped from the training data frame. The settings that the oml.svm class takes into consideration are the kernel function to be applied, which by default is linear, and the column on which the partition should be done, which in our case is the FUEL_TYPE column. As in the previous algorithms, we built the model, based on the training data, using the fit() function, and we generated a prediction for the testing data with the predict() function.

Besides showing a comparison between the actual value and the predicted value of emissions for each observation, the model provides the ability to generate global statistics for each partition, such as whether the data are converged or how many observations fit in each partition. We can remark consequently that there is no vehicle using natural gas and that most of the vehicles observe use either regular or premium gasoline, with ethanol being the least used fuel.

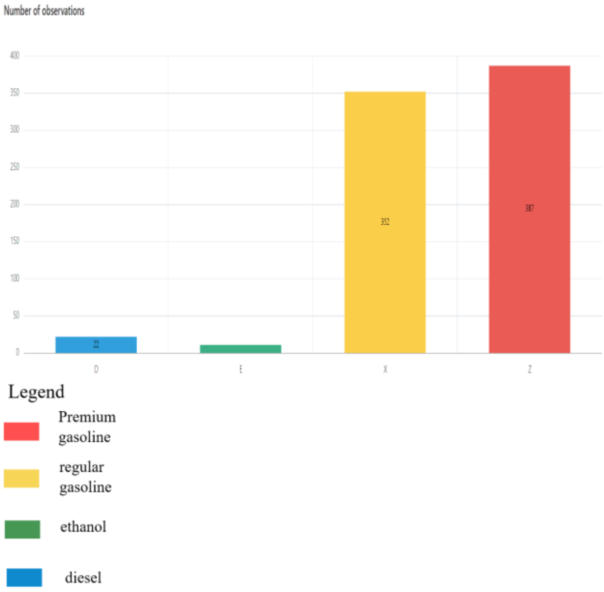


Fig. 8. Number of observations for each fuel type generated by the SVM Model

Furthermore, using the topN_attrs parameter, we can generate prediction details for each vehicle, and understand which variable is the most important in the final result and which has the least involvement. This is the classification capability of the Support Vector Machine model, and we can therefore observe that the fuel consumption is the most important attribute of a vehicle, regarding the volume of emissions, while the transmission is the least important, having the weight 5 on most of the observations. According to this classification, we can predict that, when buying a vehicle, it is more important to consider its fuel consumption than its transmission or its engine size.

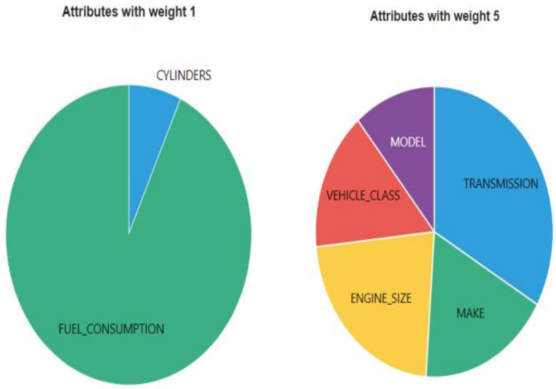


Fig. 9. Number of observations for each fuel type generated by the SVM Model

The pie charts that were built in APEX are a graphical proof that the most important aspect of a vehicle, when it comes to CO2 Emissions is its fuel consumption, which might be intuitive to some degree.

3 Conclusions and comparison between the used algorithms

After applying all these algorithms to our data, we can analyze the main similarities and dissimilarities between them and decide which one is the best to be used in the situation presented by our data.

Firstly, the most obvious comparison that can be remarked is the one between the two regression methods, the one applied on the Generalized Linear Model, and the one applied on the Support Vector Machine Model. While the former is more effective on a set of simple data, as it does not provide information about the importance of the coefficients, the SVM gives more appropriate results, classified by attribute importance, consequently being more powerful as an analysis tool. Not only that, but SVM allows the modelling of non-linear relationships, as they can discover linear separation between data. However, GLM is less memory and time consuming and may be more suited for big masses of data, while SVM heavily uses the computer’s assets and it is usually considered more of a classification algorithm, rather than a regression one.

Second, the k-means algorithm stands out from the other two, as it creates a cluster analysis, rather than a regression analysis. As we saw in the algorithm walkthrough, using the k-means algorithms requires building a model against two dependent variables, rather than only one, as in the case of regression. Moreover, considering that the Cluster Model did not predict exact values for the CO2 emissions of each vehicle, but rather placed said vehicles in certain clusters and groups, it can be considered the least reliable for our analysis. Furthermore, the clustering algorithm considers that an analysis of attribute importance was already conducted and that the clusters are formed against the two most important attributes. Consequently, this algorithm requires on more step than the other two, so it also becomes again more memory and time consuming. Not only this, but it has been observed that the k-means algorithm generates significantly different results, depending on how many groups have been used, which further proves that this algorithm is the least reliable. Even if for our data, the clustering algorithm was not the most suited, when it comes to grouping data, or discovering more general patterns of behavior in data, this algorithm comes in handier.

That being so, in the particular case of our data, we can observe that the best machine learning technique would be the Regression and Classification by Support Vector Model, as it is not applied to a big volume of data, and it provides more information about the fitness of the model and the final prediction. Besides that, we can remark that the predicted values and the actual values are similar for both regression techniques, so we can safely assume that both regression algorithms worked as expected. As a final decision, however, we can safely say that the first algorithm employed, the regression on the linear model, is the best for our data and our desired results. This is because our initial purpose was not classification of data, but rather prediction of actual results, on one hand, and because we already knew that the relationship between actual data and predicted data was linear, so we did not need a multidimensional analysis, as support-vector model is.

To conclude, after conducting this machine learning analysis on the carbon dioxide emissions of different vehicles, using the OML4Py software, we can confidently say that there is a strong relationship between certain characteristics of light-duty cars and their environmental impact. By separating datasets into training and testing data and using machine learning algorithms like regression and k-means on various models, from Generalized Linear Model to Support Vector Machine or Clustered Model, we were successful in predicting the future values for CO2 Emission, based on vehicle attributes as fuel type, consumption, or engine size. Furthermore, we were able to predict in which cluster, or group, a vehicle could be placed, based on how eco-friendly that certain vehicle is. This makes us hopeful that there is a plethora of opportunities for further work in the field, and that machine learning offers the possibility of designing, manufacturing, and buying the most environmentally friendly vehicle and, consequently, decelerating the climate change process and making a better world for the future inhabitants of Earth.

References

[1] European Parliament , "CO2 emissions from cars: facts and figures," 2016.

[2] IEA, "Transport sector CO2 emissions by mode in the Sustainable Development Scenario, 2000-2030," Jan 2022. [Online]. Available: https://www.iea.org/data-and-statistics/charts/transport-sector-co2-emissions-by-mode-in-the-sustainable-development-scenario-2000-2030. [Accessed April 2022].

[3] Government of Canada, "Natural Resources Canada," 2022. [Online]. Available: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64. [Accessed April 2022].

[4] Oracle, "What is an Autonomous Database," [Online]. Available: https://www.oracle.com/database/what-is-autonomous-database/ . [Accessed April 2022].

[5] D. F. G. V. Igor Milovanović, "Python Data Visualization Cookbook," Birmingham, Packt Publishing Birmingham, 2015, p. 34.

[6] M. Hazarika, "Oracle Cloud. Using Oracle Machine Learning on Autonomous Database," Oracle, 2017.

[7] P. Flach, Machine Learning. The Art and Science of Algorithms that Make Sense of Data, Cambridge University Press, 2012.

[8] J. Fox, *Applied Regression Analysis and Generalized Linear Models*, California: SAGE Publications, 2016.

Miruna Teleașă is a recent graduate of Bucharest Academy of Economic Studies, having a Bachelor's degree in Economic Informatics. She has also finished a Bachelor of International Relations at the University of Bucharest and she is currently studying for a Master's in International Security at Sciences Po University in Paris. She is passionate about ethical technologies and how they can improve and embelish human life and, furthermore, about how Artificial Intelligence can evolve and provide a more safe and secure environment for humans. Consequently, she has researched and developed an APEX application for her Bachelor’s thesis, using mostly web and database technologies, with the purpose of facilitating communication between humanitarian organizations and people in need.

Bardici Alexandra Teodora has studied Communication and Emerging Media at SNSPA and at Informatics Economics at Bucharest Academy of Economic Studies. She finished one bachelor thesis this year, at SNSPA, which was called "The Use of Internet Marketing in Small Businesses". This fall she will attend the master program Digital Communication and Innovation from SNSPA. When it comes to IT, she is drawn to the web design aspects, which is why her bachelor thesis will also be headed in this direction. For her, the web technologies blend perfectly with the communication strategies that she has studied, which is why her interests gravitate towards it.

Deep Learning-based Solution for Mental Health Issues

Andreea RIZEA
The Bucharest University of Economic Studies, Romania
rizeaandreea19@stud.ase.ro

The current paper proposes a solution for the nowadays mental health problems using artificial intelligence algorithms. Making use of natural language processing (NLP) techniques, the main idea is to construct a conversational agent which can act as a psychologist. This can be possible by implementing sentiment analysis on the patient's input text. In order to cope with the user's feelings, the chatbot is developed to perform cognitive behavioral therapy (CBT) exercises with him or her. These exercises are effective even in an online environment. The analysis performed by the sentiment model will detect a dominant emotion in the user's behavior and in this way the bot will adapt the conversation for obtaining better results.

Keywords: *Deep Learning, Natural Language Processing, Chatbot, Artificial Intelligence, Cognitive Behavioral Therapy*

1 Introduction

During the last years, the artificial intelligence (on short, AI) field has become more and more popular, being used in a vast number of newly developed software applications. Besides other examples, chatbots are one of the most frequent tools used in today's solutions.

There are multiple reasons behind the increased number of conversational agents used in software products now. Depending on the application's domain, the reasons can vary from increased efficiency among support teams up to better understanding of the customer's needs and behaviour while using the software. By simulating human-like conversations, the users find these conversational agents very practical when it comes to requesting help, needing online guidance or just some advice regarding a specific topic.

Recently, AI chatbots started to be present in many applications used in the healthcare field. From offering real time diagnosis to conducting small online appointments, they ease the access to medical help in many cases.

In a study conducted in 2018, it was concluded that despite the improvements and evolutions of the technology, there is a significant number of internet users who do not trust online healthcare tools, mainly chatbots, for several reasons such as the lack of trust in the prediction of the diagnosis or in technology in general [1]. However, when it comes to mental health advise, people seem to be open to use these online tools. After having a conversation with an AI chatbot, many individuals declared that their level of stress decreased [2]. More than that, persons with severe symptoms of depression registered better results in the improvement process while using a chatbot designed for mental health issues compared to those obtained after a clinical session [3].

In this paper, I aim to create a rule-based chatbot which will be trained according to Cognitive Behavioural Therapy (CBT) techniques for helping its users to cope with different predominant emotions encountered in their daily life. The chatbot will be integrated in a web application so that it can be accessible whenever and wherever the user feels the need, as long as they have access to a device with internet connection.

I strongly believe that mental health status affects the daily activity of an individual regardless his age or profession. The most recent global event, the COVID-19 pandemic, had a negative impact on our mental health. Alongside the panic caused by the new virus, the lockdowns and the imposed restrictions prevented the contamination, but caused other problems such as unemployment, change of daily routines, adaptation to new (in the case of virtual meetings related to work and school) which are well-known triggers for anxiety and depression for an individual.

According to an article published in the National Library of Medicine [4], the social distancing measure caused social anxiety, emotional disturbance, insomnia or even depression in some extreme cases. The same article states that there is enough evidence to sustain that the infection with the virus can cause some severe mental health issues. In the acute stages, infected individuals presented delirium. After the recovery, there is still some data which shows that depression or anxiety are present up to 1 year after infection.

In term of numbers, National Alliance of Mental Illness conducted a study [5] in United States which shown that 1 in 6 teenagers (aged 12-17) experienced massive depression episodes in 2020 and over 3 million of them had serious thought of suicide. Moreover, there was an increase of 31% in the number of emergency mental health department visits for patients at this age. In addition, 23% of young adults (18-25 years old) admitted that the pandemic had a negative significant impact over their mental health status.

These alarming numbers strengthen the idea that mental health should be carefully analyse at all ages. Giving the fact that sometimes children and adults do not have access to supervised help in the domain of psychiatry, AI conversational agents could improve their mood and act as a treatment for their issue.

2 Cognitive behavioral therapy

Cognitive behavioural therapy (CBT) is a psychological therapy technique which based on talk, has as aim the management of the patient's problems by changing his problematic thoughts and behaviour. It was developed in the United States during the second part of the 20th century.

According to this technique, the problematic thoughts are broken into small parts and for each of them it should be found a positive coping aspect. The hole process is based on the current issues encountered by the patient so that the actual negative feelings would change quickly and stop affecting his life.

The advantage of the CBT therapy is that its specific techniques that are implemented can vary from one therapy session to another, guided by a psychologist or self-help exercises done individually. Within the recovery process, the patient will be encouraged to use diverse tools such as journaling, relaxation techniques or even role-playing. In addition to the wide range of coping mechanism, the strategies based on cognitive behavioural therapy rules are effective both during in-person training session and also in online solutions such as monitoring applications or virtual meetings with a mental coach. [6]

However, the CBT strategy does not have as final objective to label the patient's issues and make him aware of his disease. This therapy is built over the idea that it is more efficient to find strong and healthy arguments which can improve his behaviour. The key words of CBT are identifying negative thoughts, practicing new skills and goal setting.

Cognitive behavioural therapy can help with a large set of frequent mental health issues which are affecting people nowadays. These are:

- Anxiety
- Depression
- Eating disorder (bulimia)
- Phobias
- Bipolar disorders

However, cognitive behavioural based exercises can also solve a range of common problems which cause discomfort among individuals such as:

- Grief and loss
- Low self-esteem

- Stress management
- Relationship problems
- Addiction issues (drug, alcohol)

The patient who is involved in any kind of cognitive behavioural therapy should have a significant level of motivation for changing his mental-damaging thoughts and actions. The entire process may consume a lot of energy because the problematic thoughts should be carefully analysed. The alternative thoughts that are aimed to combat the problem need to be constructed around a neutral idea which will not generate any other harm to the mental health.

The current AI driven chatbot presented in this paper will focus on treating users' issues applying a CBT technique called cognitive restructuring.

Cognitive restructuring [7] is focused on the identification of harmful behaviours and thoughts followed by building a correction mechanism against them. This mechanism consists of a set of alternative thoughts that will trigger fewer negative emotions to an individual's mind.

The cognitive restructuring should not be associated with a positive thinking manner [8]. In this case, the focus is not on a drastically emotional change. For example, the objective is not to make an anxious patient instantly happy. He needs to understand the reasons for his anxiety, find some alternative actions which can help in lowering the emotion intensity and apply them every time he encounters himself in a similar situation.

By applying constantly cognitive restructuring exercises over a predominant problem, the patient will notice a more relaxed, happy mood. Also, he will experience a newly developed way of thinking and behaving. The situations which usually caused him negative emotions now will be overpass without any significant damage on the mental state.

A very common way used for solving negative thinking using cognitive restructuring is called thoughts recording. In the application described in my thesis, the chatbot will apply this exercise with every user which presents signs of depression, anxiety or anger issues.

The steps of completing this exercise are the following ones [8]:

a) Identifying the problematic thought or action

This step is represented by the clear identification of the thought or idea that generated the negative emotion felt by the patient. For reference, this idea will be called automatic thought, as it is the first impulse that the patient's mind reveals when facing a certain event.

b) Constructing different points of view over the identified damaging pattern using a set of predefined questions

During this phase, the patient should answer to different questions which will make him have a better understanding over the essence of his problem.

Questions like "what is the effect of believing this thought?", "what would happen if you stopped believing this idea?", "what is the rational evidence of keep trusting this thought?" will make the patient aware of his present thoughts. In this way, he will observe them from an objective perspective. This is the starting point of changing his focus from being overwhelmed by the amount of negative emotions experienced to the process of solving the current issue.

To dive deeper into the problem, asking the patient to find an alternative explanation for his automatic thought could force him to take into account other variables that were not considered before. As an example, he may find out that overthinking some actions may trigger a high level of anxiety in his mind. The entire process generated in the patient's mind while he is trying to respond to this particular question will open multiple doors for new possibilities which can solve the main problem.

The final point of this step is to identify the worst-case scenarios that can be caused by the automatic thought. Going through this phase, the patient starts to have a clearer picture over the situation built in his mind. In case of anxiety or depression, this could be the key moment when the level of negative emotions starts to decrease as the patient realises which parts of his imagined scenario are probably to happen and which are just augmentation of a fake reality.

c) Imagining a scenario in which you would give advice to a friend which is confronting with a problem similar to the previously identified automatic thought

The depersonalisation process involved in this step has many benefits on the patient way of thinking. Here, it is implemented using a well-known psychological strategy called role-play.

After analysing the patient's thought, we can observe that he is stuck in a vicious circle formed by his own harmful thoughts. This may isolate him from seeing the bigger picture and thus everything will look and feel worse than it actually is.

By transposing the same problem to an external person (in this case, a close friend), the patient will escape from the infinite loop created in his mind and will try to find the best arguments to make his friend feel better. As soon as the encouraging words are found, the patient will observe that his similar problem has a solution, so the intensity of the negative emotions will drop.

The long-term benefit of this type of mental exercise is that it will increase the empathy and the self-esteem of the patient, so emotions like anxiety, depression or fear will affect less his overall well-being.

d) Constructing the final alternative response which will combat the automatic thought

The last step of the thought recording exercise is to build the final alternative response. The alternative response is the concept that will help the patient to cope with the automatic idea that caused his entire set of overwhelming emotions.

Analysing the answers formulated at the previous steps, the patient may try to find an alternative response, suitable for diminishing the source of the problem. It can be used to calm his emotions every time he will find himself in a similar situation.

As a conclusion regarding the effectiveness of the cognitive restructuring, a study published in August 2014 in "Journal of Anxiety Disorders" showed that this technique significantly decreased the possibility of developing post-event processing (PEP) thoughts. The concept of PEP thoughts refers to the reflective ideas that an individual can experience after a social event. Usually, those thoughts are automatic and cause anxiety as they focus on possible bad or embarrassing actions done while interacting with others. [9] [10]

3 Application architecture and data flow

Regarding the application architecture, it is presented in the figure below:

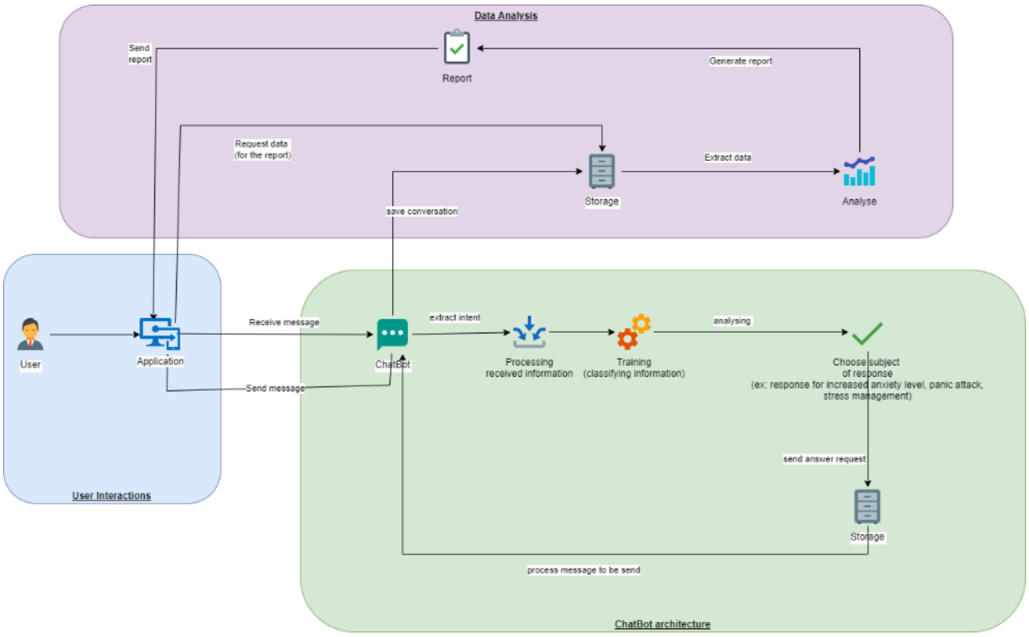


Fig. 1. Application architecture

As it can be noticed, the application consists of three levels:

- User interface
- Chatbot architecture
- Data analysis

a) User interface

The user interface part represents the initial interaction between the user and the application. The user must be connected to the web application in which the chatbot is integrated.

The backend of the web application is built using Flask, a web framework which gives the possibility to create such application based on one or many Python files. However, for adding multiple functionalities to the application, the Flask project can also contain HTML, CSS or JavaScript files. [11]

For styling the application, I used HTML (Hyper Text Markup Language) and CSS (Cascade Style Sheets). The web pages and also the chat box are constructed using these tools for making the experience more appealing to the user. The design is based on the idea of a sky full of clouds. This image is usually relaxing, so the user's feelings may be influenced by this concept.

Moreover, there is a JavaScript attached to the Flask project for handling the events of the chat box. In general, this programming language is used for managing special effects of a web page.

An important aspect to mention here is that the developed web application is a responsive one, so it can be used on multiple type of devices (computer, phones or tablets) as long as they have a stable internet connection.

b) Chatbot architecture

Once the user launches the chat box from the web application, the conversation with the chatbot starts. The chatbot architecture is one of the most complex parts of the application. It has multiple components:

- Processing user messages
- Analysing user messages and identifying the predominant sentiment
- Processing bot replies depending on the identified scenario

Initially, the bot will present itself explaining what its purpose is and what therapeutical techniques is it using. After that, it will involve the user into the conversation by asking what his name is. An important but simple question is addressed afterwards. The bot will ask the user to describe what are his current life issues. This step has as objective the identification of the sentiments felt by the user. In this step, it will be involved the sentiment analysis model detailed in the Sentiment analysis paragraph.

From here, the bot can adapt to two main scenarios depending on the type of emotion. If the emotion is a positive one, the bot will guide the user throughout a journaling exercise and it will encourage him to talk about his day or detail his current feelings. In addition, the bot can make podcasts suggestion based on the user's preferences.

In the other scenario, when negative emotions are present in the user's description, the bot will start using the cognitive restructuring techniques with the help of the exercise called thought recording. During the exercise, the bot will try to extract the source of the user's emotional problems. Firstly, the problematic thought will be identified with the help of the user. After that, the bot will put a set of predefined questions which will offer a clearer picture of the problem. Once all the questions are answered, the user will need to choose an alternative thought that will help him to overcome the situation induced by the initial automatic thought.

The alternative thought developed by the user will be also analysed by the sentiment model. On one hand, if the predicted emotion of the alternative response is a negative one or if there is no significant improvement between the final and the initial thoughts, then the bot will suggest the user to seek for professional help. For this, the user can contact one of the psychologists who are listed in special section of the web application.

On the other hand, if there is a major improvement in the way in which the user thinks, then the bot will congratulate him and finally the conversation will end.

The entire conversational process is done using multiple Python functions. The bot replies are extracted from files as dataframes using the *pandas* package. The methods which are used for building the sentiment model are detailed in the next section.

Each conversation is saved in a JSON file which will be used for generating the report detailed in the next step. Initially, the user responses are stored progressively into a Python dictionary and in the end, it is converted into a JSON file using the *dump* method from the *json* package.

c) Data analysis

The last component of the application architectural scheme is the data analysis part where the final report will be generated.

The previously saved JSON file containing the messages between the user and the bot will be analysed using different data analysis technique. In this way, a report will be generated as a line chart graph. It will highlight the emotional process in which the user was involved during the conversation with the chatbot. This is an important tool if the user would like to take evidence of his mental health status during a certain period of time.

For completing this step, I used different Python packages such as *pandas, pyplot, matplotlib* and *json*.

4 Sentiment analysis

Text mining is the process of developing meaningful insights from unstructured pieces of text. Originally, it was a part of the Natural Language Processing (NLP) technique which is a subcategory of the Artificial Intelligence (AI) field.

Natural Language Processing (NLP) contributes to the communication between machines and humans. With the help of some specific algorithms, it makes a computer understand and process the human language. Examples of fields in which NLP is used are spell check systems, online translation solutions and so on.

Sentiment analysis is one of the many Natural Language Processing (NLP) applications. It has as objective the extraction of the emotional component behind a portion of text. This analysis requires the use of Machine Learning (ML), Artificial Intelligence (AI) and text mining techniques in order to implement it properly.

Nowadays, sentiment analysis algorithms are most frequently used in organizations that interact directly with clients. These algorithms help the enterprises to have an overall image of the customer's feedback regarding their services or products. Depending on the performed analysis, the decisional team may develop new strategies for improving the clients' experience.

a) Implementation

There are several types of sentiment analysis, but in this paper, I will focus on the one in which there will be extracted four categories of emotions out of a text: sadness, fear, anger and joy.

The analysis will be implemented in Python. In general, Python is used for applications developed using machine learning, artificial intelligence, or natural language processing methods because it provides numerous libraries that help with the mathematical computations encountered in the implementation of these fields.

The training dataset consists of a series of tweets placed into a text file. The structure of the file is the following: the number of the tweet in the text file, the corpus of the tweet (including tags, emojis and hashtags), the emotion expressed by the tweet and a threshold which shows the intensity of the expressed sentiment.

Using the *pandas* package, the data is stored into dataframes depending on the expressed sentiment. The distribution of the emotions was draw using the *pyplot* module from the Matplotlib package. It represents how the tweets are distributes over the entire dataset.

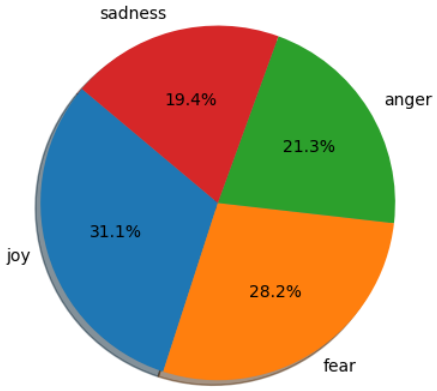


Fig. 2. Distribution of sentiments in the training dataset

As the pie plot suggests, the sentiments seem to be almost equally distributed in the training dataset. However, the number of tweets which express joy is the largest while the tweets that express sadness are the fewest.

An important part of the sentiment analysis could be to highlight the most frequently used words in the training dataset to express each emotion. For this, I used the WordCloud package which helps with this overview of the most recurrent words. This package is generally used for data visualization. It gives as output an image with the most frequent words of a text. The higher the font of the word, the higher the number of occurrences in the analysed text.

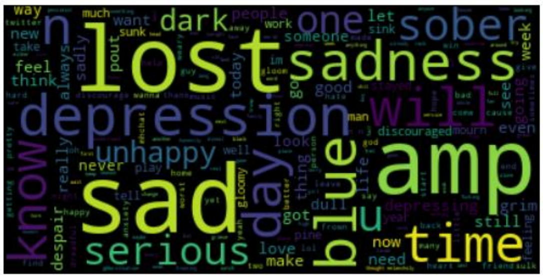


Fig. 3. Most common words used to describe sadness

In the training dataset it seems that words like "sadness", "lost", "depression", "sober", "unhappy", "sad", "discouraged" or "dark" are often used to express sad thoughts. There was a high probability that these words will be pressed in the data set containing sad tweet messages. However, there are some words which are not normally expected from a sad lexical field, namely “time”, “day”, “know” or “will”. An explanation for this may be the fact that these words are among the most frequently used words on Twitter [12].

On the other hand, in the data set collecting the fear thoughts are repeatedly found words like “fear”, “nightmare”, “afraid”, “bully”, “terrorism”, “shocking”, “afraid” or “anxiety”. Generally, those are the words which form the lexical field of fear.

When it comes to expressing their anger, most people used in their tweets words such as “angry”, “anger”, “bitter”, “people”, “offended”, “revenge”, “furious” or “insult”.

The most common used word to describe joy in the data set which represents the tweets expressing thoughts of joy are “happy”, “smile”, “love”, “amazing”, “delight” and “laughter”.

Usually, data visualization procedures contribute to a better understanding of the analysis that follows to be done.

The sentiment analysis algorithm that I have developed is composed of three parts:

- Normalization of the text
- Vectorization of the text
- Building the sentiment model

b) Normalization

The process of normalization means extracting the most relevant information out of a text after it is cleaned, tokenized and stemmed. In the text cleaning part, all the emojis, user tags, hashtags, URL links and abbreviations should be removed from the tweets. In this way, only the main parts of each sentence will be kept.

The tokenization method consists of splitting the pieces of text into smaller groups of words. There are several ways in which this algorithm can implemented. In this paper, I chose to define a custom function in which I provide the options to eliminate the numbers, the punctuation signs and the stop words from the text. The notion of “stop words” refers to the pronouns, articles, prepositions or conjunctions normally presented in a text. They would not contribute to the sentiment analysis, so this is the reason why they should be eliminated before performing the analysis

The main package that I will used to perform this task is the NLTK package. This package gives numerous facilities for tokenizing a text. The most important one is the module which uses the word tokenization principle, meaning that it extracts each word of the provided text. Moreover, the same package makes the extraction of the stop words easier.

```
def custom_tokenize(text,
keep_punctuation=False,
keep_alphanumerical=False,
keep_stopwords=False):
    token_list = word_tokenize(text)

    if not keep_punctuation:
        token_list = [token for token in
token_list if token not in
string.punctuation]
    if not keep_alphanumerical:
        token_list = [token for token in
token_list if token.isalpha()]

    if not keep_stopwords:
        stop_words =
set(stopwords.words('english'))
        stop_words.discard("not")
        token_list = [token for token in
token_list if not token in stop_words]

    return token_list
```

The custom function will give the possibility to choose if the punctuation, the numbers or the stop words should be kept or not. An important aspect of this function is that it drops the word “not” from the list of stop words. By default, there will not be dropped. Considering the objective of this project, the word “not” can have a major impact on the sentiment analysis.

The stemming method refers to the process of reducing a word to its root. Being a normalization technique, it ensures that two or many derived words coming from the same root word will be processed as one (the root one) by the machine. As in the previous text, the normalization step, one of the most useful packages is NLTK. There are several modules in this package that provide different algorithms for stemming, but the one chose for this project is Snowball stemmer which is known also as Porter2. The main reason for which I chose to use this module is that the algorithm behind it returns a better root of the word then Lancaster or Porter stemmer does.

c) Vectorization

The text vectorization method is the process in which the text is represented in a numerical format. This approach makes the text understandable by the machine. Term Frequency – Inverse Document Frequency (TF-IDF) algorithm is usually the most versatile solution for solving this issue. It offers a way of calculating how relevant is a word in a text. The relevance of the word increases proportionally to its number of occurrences in the given text.

In the term frequency (TF) part, it is computed the weight of a term based on its occurrences in the document. The term of “document” refers to a part of the text that needs to be analysed. For example, if the text on which the vectorization will be performed contains three sentences, then each sentence will be called “document”. Term frequency method highlights the relevance of a specific term in the given text as the value of important terms will be higher than the others. It has the following formula:

$$tf_{w,d} = \frac{n_{w,d}}{\sum_k n_{w,d}} \quad (1)$$

where:

- w = current word to be analysed
- d = document on which the vectorization is performed
- n = number of occurrences of the word in the document

Document frequency (df) computation is similar to term frequency one, but the only difference is that in this case the focus will be on the occurrences of a term in the data set.

The next part in the algorithm is the computation of the Inverse Document Frequency (IDF). It tests the relevance of a word according to the entire text. It is computed using the formula:

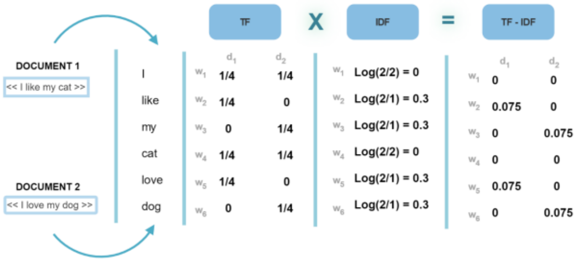
$$idf_w = log_2(\frac{N}{df_w}) \quad (2)$$

where:

- w = current word
- N = number of documents
- df_w = document frequency of the word

The final step in the TF-IDF computation is the multiplication between the TF and IDF. The obtained result will suggest if the word is significant or not to the analysed text. The higher the score of the word, the higher its significance.

The next figure summarizes the computations performed during the vectorization method giving a practical example:



		TF d_1	TF d_2		IDF		TF - IDF d_1	TF - IDF d_2
I	w_1	1/4	1/4	w_1	Log(2/2) = 0	w_1	0	0
like	w_2	1/4	0	w_2	Log(2/1) = 0.3	w_2	0.075	0
my	w_3	0	1/4	w_3	Log(2/1) = 0.3	w_3	0	0.075
cat	w_4	1/4	1/4	w_4	Log(2/2) = 0	w_4	0	0
love	w_5	1/4	0	w_5	Log(2/1) = 0.3	w_5	0.075	0
dog	w_6	0	1/4	w_6	Log(2/1) = 0.3	w_6	0	0.075

Fig. 4. Term Frequency – Inverse Document Frequency (TF – IDF) Vectorization Method

In Python, the vectorization algorithm is implemented using the TfidfVectorizer method from the *sklearn* module.

d) Building the sentiment model

A main part of building the sentiment analysis model is to split the data in two sets: training data set and testing data set. The model is represented by a multiple linear regression which means that the independent and dependent variables should be identified. In the current built model, the independent variables will be represented by the tokenized form of each tweet which needs to be analysed while the dependent variable will be its associated sentiment. Based on this observation, 80% of the data will be allocated for training the model while the rest of it will be used for testing it.

Once the split is done, the vectorization method will be applied on both training and testing sets of the independent variables. This procedure needs to be done before building the sentiment model. It ensures that the text is transformed into numerical values and the necessary computation can be realized without any impediments.

Considering the fact that there are four sentiments to be identified in the analysed text (namely sadness, fear, anger and joy), each emotion needs to have a score based on which the prediction will be performed. The score is allocated as it follows:

- If the tweet expresses sadness, its score will be 0
- If the tweet expresses fear, its score will be 0.33
- If the tweet expresses anger, its score will be 0.67
- If the tweet expressed joy, its score will be 1

The multiple linear regression equation has the following structure:

$$estimated\_sentiment = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \dots + \beta_n * x_n \qquad (3)$$

Where:

- estimated_sentiment = the predicted numerical value which will be associated to a sentiment
- β_i = the coefficient of the regression
- x_i = independent variable

Based on the scores previously allocated, the prediction will be done in the next manner:

- If the estimated_sentiment variable is lower than 0.33, then the expressed emotion is sadness
- If the estimated_sentiment variable is between 0.33 and 0.67, then the expressed emotion is fear
- If the estimated_sentiment variable is between 0.67 and 1, then the expressed emotion is anger
- If the estimated_sentiment variable is higher or equal to 1, then the expressed emotion is joy

Note that both the score allocation and the score prediction may be adjusted depending on the data set on which the model is built.

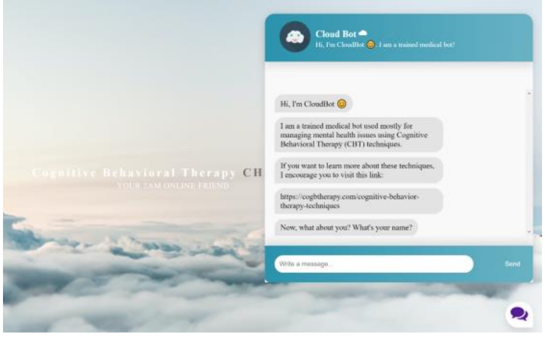
In Python, the *sklearn* package provides the Linear Regression model which helps with the construction of the regression equation.

Having all these steps put together, the function that gives an interpretation to the sentiment analysis is defined as:

```
def predict_text(text):
    processed_text =
function.process_text(text)
    transformed_text =
tf_idf.transform([processed_text])
    prediction =
final_model.predict(transformed_text)

    if prediction >= 1:
        message = "Prediction is: joy"
    else:
        if (prediction >= 0.67) and
(prediction < 1):
            message = "Prediction is:
anger"
        else:
            if (prediction >= 0.33) and
(prediction < 0.67):
                message = "Prediction
is: fear"
            else:
                if prediction < 0.33:
                    message =
"Prediction is: sadness"

    return prediction, message
```

5 Application functionalities

In the last section of this paper, I will present the available functionalities of the described application. As mentioned before, my solution has a web application which has an integrated chatbot which makes use of cognitive behavioural therapy techniques.



Hi, I'm CloudBot 😊. I am a trained medical bot!

Hi, I'm CloudBot 😊

I am a trained medical bot used mostly for managing mental health issues using Cognitive Behavioral Therapy (CBT) techniques.

If you want to learn more about these techniques, I encourage you to visit this link:

https://cogbtherapy.com/cognitive-behavior-therapy-techniques

Now, what about you? What's your name?

Fig.5. Initial page

In **Fig.5.** is illustrated the initial web page which is loaded at the launching of the application. There are brief descriptions of both the cognitive behavioural therapy and chatbot characteristics and purposes. Besides that, there can be found the chat box where the conversation between the user and the bot takes place. By default, the first messages of the bot are directly generated in the HTML file of the chat box. Once the user will send the message containing his name, the conversation will continue.

The bot can adapt to three main scenarios:

1. Anxiety and depression (or negative) scenario

- If the analysis performed on the user's main problem predicted that the dominant sentiment is either sadness or fear

2. Anger (or neutral) scenario

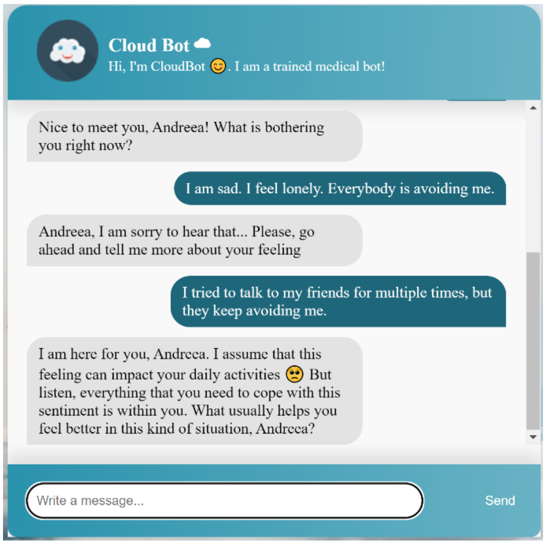
- If the model predicted anger emotions regarding the user's concern

3. Happy (or positive) scenario

- If the user may emotion predicted by the sentiment model is joy

Both negative and neutral scenarios involve the use of the thought recording exercise. In each case, it will be adapted to the specific characteristics of the dominant feeling.

In the following part of this section, I will show the steps that are implemented by the bot if the user has issues which cause him a significant level of anxiety and depression.



Hi, I'm CloudBot 😊. I am a trained medical bot!

Nice to meet you, Andreea! What is bothering you right now?

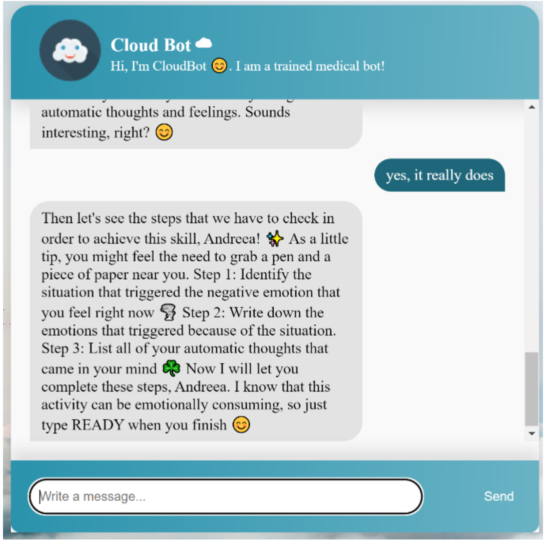
I am sad. I feel lonely. Everybody is avoiding me.

Andreea, I am sorry to hear that... Please, go ahead and tell me more about your feeling

I tried to talk to my friends for multiple times, but they keep avoiding me.

I am here for you, Andreea. I assume that this feeling can impact your daily activities 😟 But listen, everything that you need to cope with this sentiment is within you. What usually helps you feel better in this kind of situation, Andreea?

Fig.6. Conversation based on a negative emotion



Hi, I'm CloudBot 😊. I am a trained medical bot!

automatic thoughts and feelings. Sounds interesting, right? 😊

yes, it really does

Then let's see the steps that we have to check in order to achieve this skill, Andreea! ✨ As a little tip, you might feel the need to grab a pen and a piece of paper near you. Step 1: Identify the situation that triggered the negative emotion that you feel right now 💭 Step 2: Write down the emotions that triggered because of the situation. Step 3: List all of your automatic thoughts that came in your mind 🍀 Now I will let you complete these steps, Andreea. I know that this activity can be emotionally consuming, so just type READY when you finish 😊

Fig.7. Conversation based on a negative emotion

Fig.6. illustrates the starting point of a scenario. Once the user answers to the question regarding his actual issue, the bot will perform the sentiment analysis on his response and thus it will decide what type of scenario is suitable for the current user.

In the given example, the user reports sad feelings by explaining that he is sad and lonely. As a result, the bot initiates the thought recording exercise. Firstly, it will present the characteristics of the exercise and the techniques that will be involved in it.

As it can be seen in **Fig.7.**, the bot exposes the necessary steps to be done for correctly performing this exercise and asks the user to express what is the automatic thought that is causing him emotional discomfort.

Moving on, the conversation will continue with a set of predefined questions put by the bot. The scope is to create an objective perspective over the problem. In this way, the user will be detached from his auto-generated feelings and would be more capable of finding a solution.

However, when the user finishes answering all the questions, the bot will ask him to build an alternative response to his initial thought which will make him feel better. In this manner, the bot will conclude if the exercise was effective or not.

The sentiment analysis will be performed on the alternative thought identified by the user. If there is a significant improvement between the initial mental state and the final one, then the user will be congratulated and the conversation will end, otherwise the bot expresses his apologies and will guide the user to consult a professional therapist.

As mentioned before, in the case of predicted angry issues in the user's behaviour, the bot will follow similar steps with the ones presented above in the explanation of the negative scenario.

However, if the bot identifies that the predominant emotion felt by the user is a positive one, it will act like a journal by encouraging him to detail his daily activities. In addition to that, the bot will make suggestions regarding online sources that would keep the user in a happy mood such as online guided meditations or therapeutical podcasts.

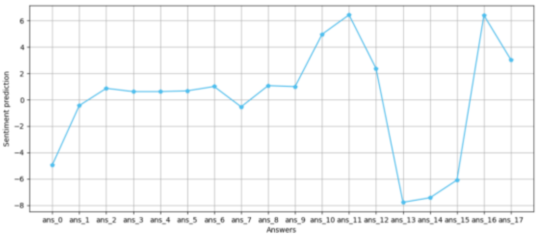


Fig.8. Graph based on the conversation when user identified an efficient alternative thought

The last functionality of the application is the final report that will allow the user (and his therapist if needed) to have a better understanding of the emotions that were involved while he was talking about his thoughts. This report is represented by a line chart which contains the sentiment prediction coefficient corresponding to each of the user's answers.

In general, the higher the coefficient, the happier the mood of the user. The graph generated based on the conversation presented above (**Fig.8.**) clearly shows that during the first messages when the user was exposing his issues, he was overwhelmed by negative emotions. As he started the thought recording exercise, his mood significantly improved. The conversation ended with a high score because the user was able to find a positive thought that could change his initial mental state.

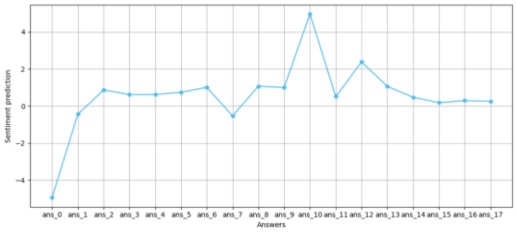


Fig.9. Graph based on the conversation when user didn't identify an efficient alternative thought

On the other hand, **Fig.9.** is the report generated when the user is not able to find an efficient alternative thought. The last answer's coefficient is almost 0, meaning that the assumed sentiment for it is a negative one.

6 Conclusions

This paper aimed to highlight the constant evolution in both technology and healthcare industry. Nowadays, both of them play an important role towards our society and fortunately they cointegrated for offering us complex solution for a better quality of life.

The dynamic of everyone's daily life has changed drastically in the past years and as a consequence, more and more people suffer from mental health issues. They provoke obvious alterations on both personal and work-related plans, so a solution for this represents a primary necessity.

We usually refer to artificial intelligence as a set of algorithms that can imitate the human thinking process. By adapting the field of psychology to one of AI's branches, the present solution demonstrated that these algorithms can be used for developing a complex application which can improve people's life. This merging process was possible because cognitive behavioural therapy techniques were proved to be efficient during in-person session as well as integrated in online exercises.

As a final remark, the solution presented in the current paper has been found effective in improving different types of mental health issues. Giving the fact that the application can also detect if a user has serious psychological problems, it can be admitted that there is a low probability that it can cause additional harm to the patient's behaviour.

In conclusion, the developed chatbot accomplished the initial objectives and it can be considered a safe solution for managing and treating mental health issues.

References

[1] A. Vaidyam, H. Wisniewski, J. Halamka, M. Kashavan and J. Torous, "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape," The Canadian Journal of Psychiatry, Vol 64, Issue 7, 2019, pp. 456-464. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/0706743719828977.

[2] K. H. Ly, A.-M. Ly and G. Andersson, "A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods," 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221478291730091X.

[3] T. W. Bickmore, S. E. Mitchell, B. W. Jack, M. K. Paasche-Orlow, L. M. Pfeifer and J. O'Donnell, "Response to a relational agent by hospital patients with depressive symptoms," 2010. [Online]. Available: https://doi.org/10.1016/j.intcom.2009.12.001.

[4] N. Kathirvel, "Post COVID-19 pandemic mental health challenges," 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7507979/.

[5] "National Alliance of Mental Illness," [Online]. Available: https://www.nami.org/mhstats.

[6] K. Cherry, "What Is Cognitive Behavioral Therapy (CBT)?," 16 May 2022. [Online]. Available: https://www.verywellmind.com/what-is-cognitive-behavior-therapy-2795747.

[7] "Cognitive Behavior Therapy Techniques," [Online]. Available: https://cogbtherapy.com/cognitive-behavior-therapy-techniques.

[8] "Part 6: Cognitive Restructuring to Change Your Thinking," [Online]. Available: https://cogbtherapy.com/cognitive-restructuring-in-cbt.

[9] B. Shikatani, M. M. Antony, J. R. Kuo and S. E. Cassin, "The impact of cognitive restructuring and mindfulness strategies on postevent processing and affect in social anxiety disorder," August 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0887618514000887.

[10] A. Cuncic, “Understanding Cognitive Restructuring,” 1 July 2020. [Online]. Available: https://www.verywellmind.com/what-is-cognitive-restructuring-3024490.

[11] “How To Make a Web Application Using Flask in Python 3,” 17 April 2020. [Online]. Available: https://www.digitalocean.com/community/tutorials/how-to-make-a-web-application-using-flask-in-python-3.

[12] “The 500 Most Frequently Used Words on Twitter,” 8 June 2009. [Online]. Available: https://techland.time.com/2009/06/08/the-500-most-frequently-used-words-on-twitter/.

Andreea RIZEA graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies, with the thesis *AI based Solution for Mental Health Issues,* in July 2022. She previously completed internship programs in cybersecurity and data engineering fields. Her domains of interest are data science, artificial intelligence, machine learning and mathematics.

Human Resources Allocation Solution

BÂLTAC Mihai-Cristian
The Bucharest University of Economic Studies, Romania
baltacmihai19@stud.ase.ro

As technology advances, so do its applications and standards. We are at a crossroads in a civilization that has grown based on the automation of operations and the development of technology to better human lives. As additional programs that do the same thing arrive, both large and small businesses utilize them, promoting their development. The approach in this paper is to address the major issue, which is the most frequently utilized capabilities in a company, whether it is IT or event production. My work involves minimizing these applications and developing a standard that may subsequently be updated on needs and demand.

Keywords: *HRIS, Tasks, Meetings, Resource Allocation, Management*

1 Introduction

The employees, in addition to the programs they use to do the work, may also utilize time management or managerial applications. According to an analysis by Okta Inc. [1], the number of software applications used by major organizations in all industries increased by 68% between 2014 and 2018, with an average of 129 apps per company at the end of 2018. We define a large firm as one with more than 2000 people and a small business as one with fewer than this amount. And small organizations tend to use more applications, with the average number increasing from 53 in 2015 to 73 in 2017. In 2018, more than 10% of businesses had more than 200 enterprise applications in their portfolio.

According to Bill Swanton, vice president and analyst at Gartner Inc., many firms utilize numerous programs that accomplish the same thing, since each person has a favorite application. Many of them are attempting to adopt "application streamlining" in order to standardize applications and minimize total business application costs.

Ticketing, meeting or planner, task tracking, or the application where each employee fills his activity (either the tasks he worked on or the meetings he attended) are all apps that are not missing from the portfolio of enterprise applications, whether we are talking about small or large businesses. Not all companies have a free day management application; they use the classic way: an application for the human resources department.

According to Luke Marson's post *The 7 most effective employee experience applications* [2], the HRIS application (or The human resource information system) is first and foremost, being the most effective application in his opinion. Collaboration apps, which are related to systems that record meetings, tasks, projects, and important information about them, are second. Ticketing applications are ranked fourth.

HRIS is a kind of business program that allows companies to store employee data, handle typical HR processes, and perform important HR tasks such as payroll and benefits administration.

An employee self-service portal, payroll, workforce management, recruitment and hiring, benefits administration, and talent management are all features of HRIS solutions. The individual modules that make up a comprehensive suite of HR solutions are frequently used to provide these features.

By first establishing a document, called a "ticket," a ticketing system documents the interactions on a support or service case.

Both the representative and the customer have access to the ticket, which keeps account of their interactions in one spot.

Either party can return to the thread at any time.

After creating the ticket, advocates can work on the issue on their own. When they have updates or solutions, they can notify the client via the ticket. Meanwhile, if the customer has any questions, they can contact the customer support person by ticket.

The agent is then notified by the ticketing system that a response has been recorded on the ticket, allowing them to resolve it immediately.

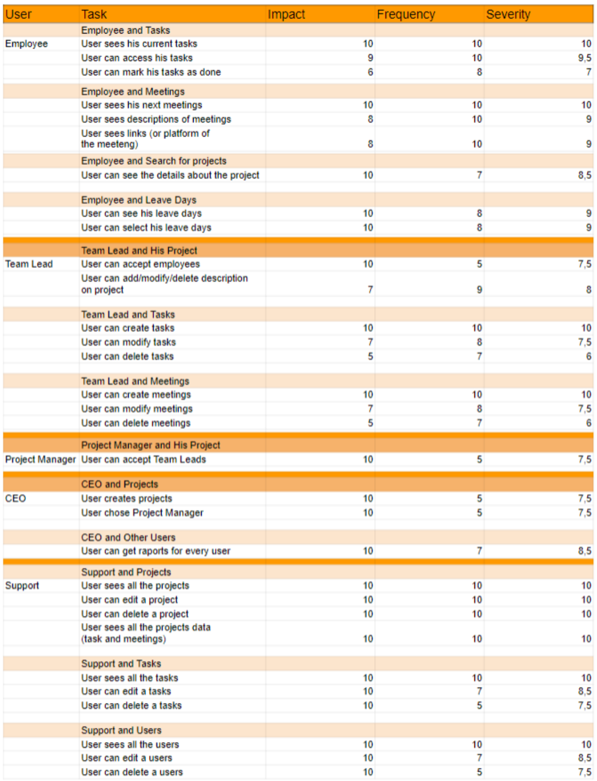
Jira [3] is one of the most widely used ticketing applications. Jira is a project management and issue tracking software program. The Atlassian tool, which was created in Australia, is now widely used by agile development teams to monitor bugs, stories, epics, and other activities.

There is no universal template that can be used to all projects; we can see that a variety of Agile time management approaches (Scrum, eXtreme Programming) have emerged, therefore, different programs are utilized for different types of project (Jira vs SpiraTeam) The approaches stated above are only a few examples; they can be applied to technical projects, IT, and programming in general. But there are methodologies and strategies for other areas as well: Marketing, Sales.

The application I am presenting is one that combines the most commonly utilized features of the previous programs. I have listed the most often used features by different types of users. Users, action, impact, frequency, and severity are described in the next paragraphs.

The following people will be considered users:

- *an employee* with access to tasks, meetings, projects, and vacation time;
- *The Team Lead* is in charge of coordinating a department;
- each project can have many departments, which are created by the *Project Manager* in charge of the project to which it is allocated.
- *The CEO* is in control of adding and modifying projects, as well as generating monthly reports for employees.
- *Support* is a person who has complete access to all databases and the ability to add, alter, and delete anything from the database.



User	Task	Impact	Frequency	Severity
	Employee and Tasks			
Employee	User sees his current tasks	10	10	10
	User can access his tasks	9	10	9,5
	User can mark his tasks as done	6	8	7
	Employee and Meetings			
	User sees his next meetings	10	10	10
	User sees descriptions of meetings	8	10	9
	User sees links (or platform of the meeteng)	8	10	9
	Employee and Search for projects			
	User can see the details about the project	10	7	8,5
	Employee and Leave Days			
	User can see his leave days	10	8	9
	User can select his leave days	10	8	9
	Team Lead and His Project			
Team Lead	User can accept employees	10	5	7,5
	User can add/modify/delete description on project	7	9	8
	Team Lead and Tasks			
	User can create tasks	10	10	10
	User can modify tasks	7	8	7,5
	User can delete tasks	5	7	6
	Team Lead and Meetings			
	User can create meetings	10	10	10
	User can modify meetings	7	8	7,5
	User can delete meetings	5	7	6
	Project Manager and His Project			
Project Manager	User can accept Team Leads	10	5	7,5
	CEO and Projects			
CEO	User creates projects	10	5	7,5
	User chose Project Manager	10	5	7,5
	CEO and Other Users			
	User can get raports for every user	10	7	8,5
	Support and Projects			
Support	User sees all the projects	10	10	10
	User can edit a project	10	10	10
	User can delete a project	10	10	10
	User sees all the projects data (task and meetings)	10	10	10
	Support and Tasks			
	User sees all the tasks	10	10	10
	User can edit a tasks	10	7	8,5
	User can delete a tasks	10	5	7,5
	Support and Users			
	User sees all the users	10	10	10
	User can edit a users	10	7	8,5
	User can delete a users	10	5	7,5

Fig.1. Analyzes of Application Activities Table

There are three types of profiles in the app: employee, CEO, and support. The profile for the roles of Team Manager and Project Manager is similar to that of an employee, with the exception that they have additional functionalities on the project where they hold this position.

For each role, the Action column gives a list of available actions. As a result, depending on the type of user, it can perform a variety of tasks.

The Impact column indicates the user's level of need for that activity. The action is graded on a scale of 1 to 10, with 1 indicating that it is not very important and

10 indicating that it is extremely important.

The Frequency column shows how often the user will perform that everyday action. The user is ranked on a scale of 1 to 10, with 1 indicating that the user does not perform the activity frequently and 10 indicating that the user performs the action multiple times per day.

The arithmetic mean of the impact column and the frequency column is the Severity column. This is estimated to determine which activities are necessary and how the design should be conceived to maximize the user experience.

2 Technical Specification

Frontend technologies

For frontend React.js and Sass technologies were used.

Facebook created the React Js user interface library in JavaScript. It offers profound insights on how to work with the DOM (Document Object Model), organize your app's data flow, and consider user interface elements as separate components. [4]

React (also known as React.js or ReactJS) is a free and open-source front-end JavaScript toolkit for creating UI components-based user interfaces.

React can be used to create single-page or mobile applications as a foundation. React, on the other hand, is solely concerned with state management and rendering that information to the DOM, so constructing React apps frequently necessitates the usage of extra frameworks for routing and client-side functionality.

One issue that React solves, to the detriment of traditional web applications, is efficient DOM processing. Each website contains a DOM that shows the page components; they are organized as nodes and objects and may be updated using javascript. It is shaped like a tree. When javascript makes significant modifications to the conventional DOM, it is rendered, which can become inefficient. React js introduces the concept of Virtual DOM, which duplicates the standard DOM and renders only the updated nodes or objects.

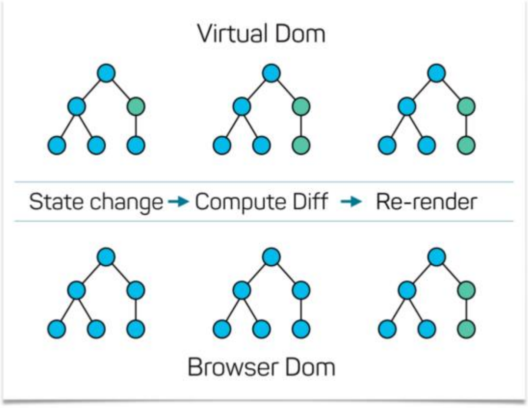


Fig.2. Virtual DOM Example [5]

Fig.2. shows how the virtual DOM modifies a node (State Change) and changes the color from blue to green. Then we modify all the nodes that have the updated node (Compute Difference) as the parent, and we re-render the Virtual DOM in the Browser style (the DOM that the browser sees).

Class React and Functional React are the two types of React. Functional React is becoming increasingly popular because it makes use of the fact that Javascript is a functional programming language. A component of the React class is constructed by extending the React class and rendering HTML code, which is returned by the render() function. The functional React component is given by a function that returns HTML code. It is significantly easier to design a component when using functional react; however, this is a drawback when utilizing a component as a class and want to use the OOP (Object-Oriented Programming) paradigm. Because the two types are compatible, we may have a component in the form of a class that calls another component in the form of a function.

CSS (Cascading Style Sheets) is a fundamental technology of a web page. CSS is the visual component of a website, covering from layout to text color. We may use it to create multiple styles for different devices or screen sizes.

A CSS preprocessor is a scripting language that allows programmers to write code in one language and then compile it into CSS. Less and Stylus are two well-known examples of preprocessors. Sass is probably the most popular right now.

Sass (short for "Syntactically Awesome Style Sheets”) is a CSS extension that allows us to use variables, nested rules, inline imports, and other features. It also helps organize and allows one to produce style sheets more quickly.

When Chris Eppstein released Compass in 2009, a project specifically designed to handle Sass packages and encourage open-source Sass code sharing, Sass attracted widespread notice.

Eppstein identified an opportunity for Sass to adjust pre-build class names' copy-pasting libraries. [6]

When we compile the sass code, we get the CSS code. As a result, the capacity of Sass to affect the DOM should not be confused; it just helps to create CSS code more effectively. Some CSS constraints are also present in Sass, such as the inability to access an element's parent and unconnected text flows.

There are various ways to compile Sass. Installing an extension that achieves this is the simplest, but also the most recommended option. Another option is to utilize bash scripts to do this. The global Sass mode installation is necessary for the device to recognize these instructions. Uncompiled Sass is often permitted in frameworks. As a result, when the app builds the page code, it also produces the Sass.

The syntax of Sass code is similar to that of the Python language since it employs indentation, no brackets, and no semicolons to conclude a line.

Scss is a Sass variant that has the same capability as Sass but has a code syntax that is similar to conventional CSS. Scss code will be utilized in the provided application.

Backend technologies

For backend Node.js, Express and Sequelize technologies were used.

Node JS can be thought of as a JavaScript runtime environment built on top of Google's V8 engine. As a result, it gives us a context in which we can write JavaScript code on any platform that has Node.js installed.

Ryan Dahl gave a presentation at JSXonf in 2009 that permanently impacted JavaScript. He introduced Node.js to the JavaScript community during his lecture. [7]

Node.js allows developers to utilize JavaScript to create command-line tools and server-side scripting, which involves running scripts on the server before sending the page to the user's browser. As a result, Node.js symbolizes a "JavaScript everywhere" paradigm, bringing online application development together around a single programming language rather than separate languages for server- and client-side scripts.

There are benefits and drawbacks to Node.js. Some of its benefits include fast performance for real-time applications, simple scalability for recent software, a quick learning curve for developers already familiar with Javascript, and improved performance of applications. The disadvantages of using a relatively new technology include: limited support from existing libraries and a lack of experienced developers. Javascript also has other drawbacks that affect node.js, such as an unstable API that encourages frequent code changes, Javascript is also a language with the Asynchronous Programming Model paradigm that makes code maintenance challenging.

The core concept of Node is the usage of non-blocks, which are extremely effective and lightweight in comparison to other technologies because Javascript is an event-driven programming language (anything starts with an event). Node.js seeks to address a market issue rather than try to displace competing technologies. It is useful for constructing fast and scalable applications because it is efficient and writes relatively quickly. It is not suggested to utilize it for CPU operations or any type of intensive processing, since these will neutralize all of its benefits.

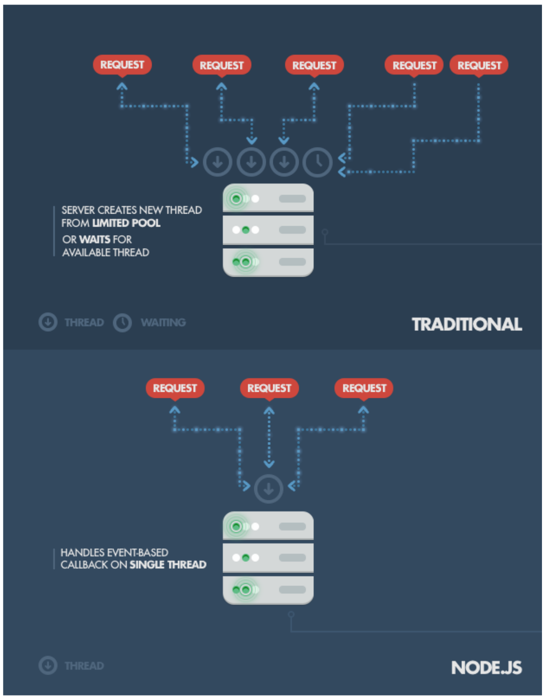

REQUEST
REQUEST
REQUEST
REQUEST
REQUEST
SERVER CREATES NEW THREAD FROM LIMITED POOL OR WAITS FOR AVAILABLE THREAD
THREAD
WAITING
TRADITIONAL
REQUEST
REQUEST
REQUEST
HANDLES EVENT-BASED CALLBACK ON SINGLE THREAD
THREAD
NODE.JS

Fig.3. Traditional Web Server vs Node.js [8]

The architecture of a conventional Web server and Node.js are contrasted in **Fig.3.** Accordingly, in the conventional form, every new request generates a new thread, filling the system memory; when there is no longer any further memory, it waits for a thread to be released. Instead, Node.js uses non-blocking I/O calls and runs on a single thread, allowing hundreds of active connections to be called simultaneously.

Express.js, or simply Express, is a back-end web application framework for Node.js that was distributed under the MIT license as free and open-source software. It is intended for the development of web applications and APIs. It has been dubbed Node.js' de facto standard server framework.

Express. js is a web framework built around the Node.js http module and the Connect components. Middlewares are the term for these components. Developers are the epicenter of the framework's concept, namely, configuration over condensation. In other words, developers are free to choose which libraries they require for a given project, giving them a great deal of freedom and customization. [9]

Express is a simple and adaptable framework that helps in building reliable online applications. It is Node.js' most widely used framework. It offers a range of techniques for quickly and simply building an API. Both SQL and non-SQL databases can use it. Other well-known frameworks like Feathers, KeystoneJs, NestJs, and many others have been developed on top of this framework because of how commonly used it is in industry.

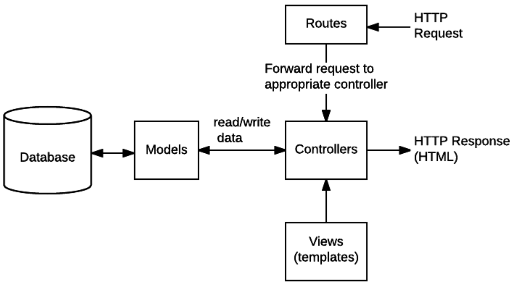


Fig.4. Express Architecture [10]

The architecture of this framework is illustrated in **Fig.10.** Data models, called models, typically define tables. This means that each name of the table is represented by a structure with columns as its elements. Controllers are functions that handle data, act like functions, and receive parameters using a variety of techniques (in the body of a request or as parameters of the function). Routes link the corresponding controller to the request code. As a result, we may specify which requests call which controls on the routes. Controllers render the date using views or templates.

Data conversion between systems that use OOP languages is known as object-relational mapping or ORM. In that programming language, it generates a "virtual object database" that can be used. In other words, we can use objects to manipulate database data. As a result, we can now use object methods instead of writing SQL code. Therefore, if we change the database, all we need to do is check to see if the ORM we use supports it; if it does, then we don't need to do any code changes.

Prism and Sequelize, which have accumulated over 20,000 github starts each, are the top ORMs for Node js. On the other hand, users use Sequelize more frequently. Both ORMs support six different types of databases. [11]

Sequelize is a promise-based Node.js ORM tool that supports Postgres, MySQL, MariaDB, SQLite, DB2 and Microsoft SQL Server. It includes transaction support, relations, eager and lazy loading, read replication, and other features.

Sequelize uses Semantic Versioning and is compatible with Node v10 and higher.

Sequelize uses objects that are extended from the Model class, as is conventional for ORMs. Thus, object methods are used to carry out actions like Select, Insert, Update, and Delete. The "belongsTo ()" and "hasMany ()" methods specify the connections between tables (in our case, models). Later, these techniques assist us in joining the tables (in Sequelize the "include" method is used).

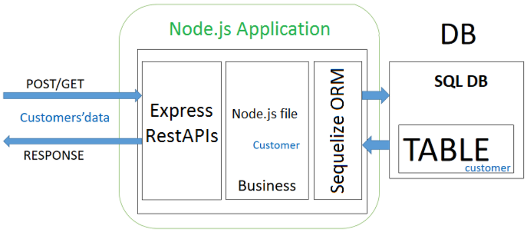


Fig.5. Sequelize Role in Node.js Application [12]

Database Technology

One of the first open-source RDBMSs to be designed and built was MySQL. Although there are many different versions of MySQL available right now, their fundamental syntax is the same. Due to its unique architecture compared to other database servers, MySQL can be used to address a variety of issues. You must understand its design to operate with it because it has a unique architecture.

In contexts with high demand, like web applications, MySQL is flexible enough to operate very effectively. It is adaptable in many aspects, such as you may set it to run effectively on a wide variety of hardware and supports a variety of data kinds. [13]

Because MySQL is based on a client-server architecture, its central component is a MySQL server, which manages all database commands. MySQL was originally built to manage huge databases fast. Various transaction types, including stored procedures, functions, viewers, views, and triggers, are possible. Compared to other databases, MySQL is relatively efficient for read-only commands, but for big testing or sophisticated queries, PostgreSQL is a superior alternative.

3 Implementation of the solution

My objective is to create a web platform that incorporates all of these technologies, allowing managers and human resources to better manage workers on a wide range of projects.

The software has five types of users: employees, team leaders, project managers, CEOs, and support.

The employee can accept assignments, provide comments or be in charge of tasks, look for other individuals, study project strategies, and apply to the project as an employee, team leader, or even project manager.

The team leader may carry out all employee responsibilities, admit people into his project, create a calendar, submit assignments and meetings, and load the strategy.

The project manager not only performs all the duties of a team leader, but also chooses team leaders for the project.

The CEO develops, oversees, and produces monthly reports for each employee.

Support has complete access to all objects and can monitor all projects and events in the application, as well as add, modify, and remove each piece from the database.

A well-defined architecture is required for a better solution to ensure a smooth flow.

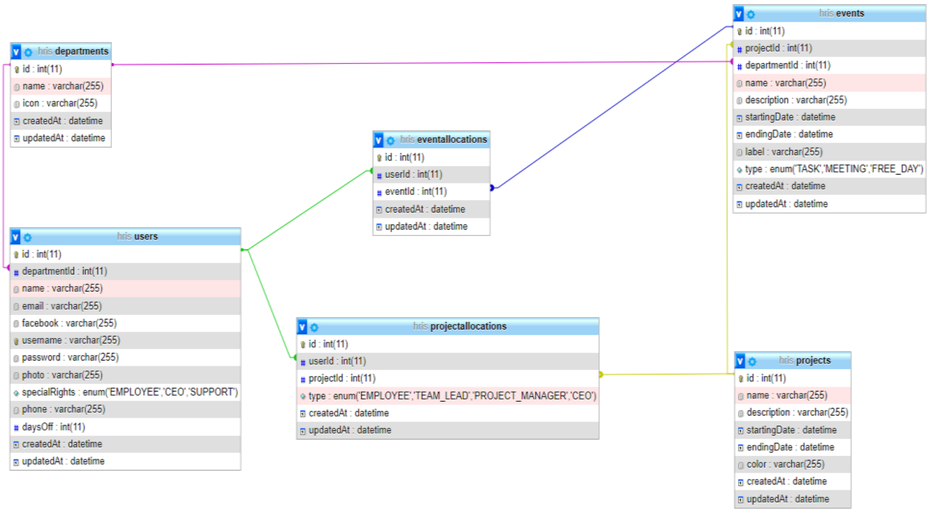


Fig.6. Database Architecture

The User table contains user information such as name, email, Facebook, username and password, photo, phone number, vacation days, and the type of rights, which can be of the following types: employee, CEO, and support. The department is a foreign key for the department table, so we can figure out what interface the user will see based on this information.

Each department's name and icon are listed in the department table.

The Events table has two foreign keys that point to the department and project tables, respectively. Other columns include name, description, label, start and end dates, and type, indicating if the event is a task, meeting, or free day. I chose this format since I didn't want to make tables with the same structure. Aside from the type of event, the only distinction between a task and a meeting is the information in the label column; in the case of tasks, the label column has four possible answers: New, Doing, Done, and Closed. This table, the label table, contains the link to the platform where the meeting is hosted in the case of meetings.

The user and the event are linked through the event allocation table. This table was chosen to prevent the many-to-many relationship between the two tables.

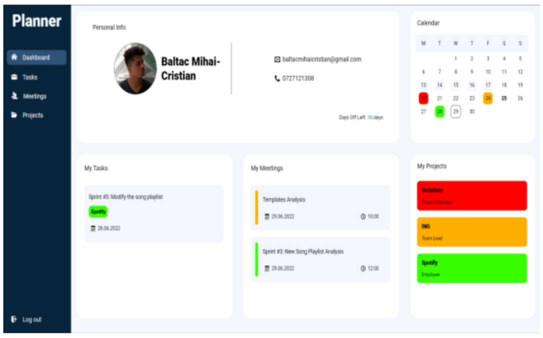
The project table provides the following details: name, description, colour (which is used to visually distinguish the projects), start date, and end date.

In addition to the foreign key relationships between the user and the project, the assigned project table contains a type, which is an enumeration of various types: employee, team lead, project manager, and CEO. Additional activities are established in this project based on these types.

When the application is first launched, a login window displays, requesting the user for login information before sending a request to the backend. If the input is accurate, the frontend receives a json containing the user id and user type. As a result, the program may render the user interfaces and activities. As a result, we have three options for the user: employee, support, or CEO.

The information obtained from json is then saved as cookies, so that if the user refreshes the page or exits the program and returns later, no additional login is necessary; he may erase this cookie by hitting the Logout button.

The employee interface shows when the user has logged in with his employee credentials. It has access to four different pages: the Dashboard, Tasks, Meetings, and Projects. The Dashboard page is the first and most significant page in terms of User Experience. On this page, the User can view: account information, a calendar for the current month (where he can see which days, he has tasks or meetings), three sections: one for meetings, one for tasks, and one for projects.



Baltac Mihai-Cristian

baltacmihaicristian@gmail.com

0727121308

Days Off Left: 16 days

M	T	W	T	F	S	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			

Sprint #3: Modify the song playlist
Spotify
28.06.2022

Templates Analysis
29.06.2022 10:00

Sprint #3: New Song Playlist Analysis
29.06.2022 12:00

Vodafone
Project Manager

ING
Team Lead

Spotify
Employee

Fig.7. Dashboard

As shown in **Fig.7.**, On the Dashboard page, the user can access a number of sections. A picture, a name, some contact information (phone and email), and the number of free days, they are all found in the first section, which is titled Personal Information. We can see three different types of highlights in the calendar section on the right. If a day is highlighted in colour, it indicates that a task or meeting will take place on that day. This time, the project's colour serves as the background for the event. Days 13 through 17 are highlighted in a very light gray, which indicates that the user is on vacation on those days. On the 25th, which is the current date, there is another kind of highlight. If a date has a border but no background, there will be multiple events that day. We can hover over a day to view more information about the events on that day.

We can add events (tasks, meetings, or free days) directly from the dashboard. When we click on a date in the calendar, a context menu with three options will appear, allowing us to add additional events. When you select the "Add Meeting" or "Add Task" option, a pop-up window will appear with all the details for that event. When we click "Take Vacation," the screen will refresh and we can see that the date we chose is highlighted and that the number of days off has been reduced by 1, indicating that the action was successful.

It takes numerous steps to publish a day of vacation. As a result, we have a project with the id 0 of "FREE DAY" and a department with the id 0 of "FREE DAY". I decided to work with ids of 0 since MySQL permits the usage of primary keys with values of 0 and because the first primary key would automatically start at 1 when the database is reset. Therefore, the value 0 will remain empty.

We have the following logic for posting a day off: We build an event that will receive the start and finish dates, the department id, the description data, and the project id, which will have the number 0 in it. If this event is successfully created, we go to the next step; otherwise, we provide a 500 error code and the words "Server Error." We add our user to the event after it has been created. The user's remaining free days are then determined by accessing him. By deducting one day from these days after this, we can modify the user.

If every step was successfully completed, we return a code of 200 with all of our adjustments. If not, we send the client a code 500 with an error message.

```
postFreeDay: async (req, res) => {
    EventDB.create({
      name: "Off day",
      projectId: 0,
      description: "Free day",
      departmentId:
req.body.departmentId,
      startingDate:
req.body.startingDate,
      endingDate: req.body.endingDate,
      type: "FREE_DAY",
      label: "FREE_DAY",
    })
      .then((event) => {
        EventAllocationDB.create({
          userId: req.body.userId,
          eventId: event.id,
        })
          .then((event) => {
            UserDB.findOne({
              where: {
                id: req.body.userId,
              },
            })
              .then((event) => {
                UserDB.update(
                  {
                    daysOff:
event.daysOff - 1,
                  },
                  {
                    where: {
                      id:
req.body.userId,
                    },
                  }
                )
                  .then((event) => {

res.status(200).send(event);
                  })
                  .catch((error) => {

console.log(error);

res.status(500).send({ message:
"Server error" });
                  });
              })
              .catch((error) => {
                console.log(error);

res.status(500).send({ message:
"Server error" });
              });
          })
          .catch((error) => {
            console.log(error);
            res.status(500).send({
message: "Server error" });
          });
      })
      .catch((error) => {
        console.log(error);
        res.status(500).send({
message: "Server error" });
      });
  }
```

We have three sections at the bottom of the dashboard: one for tasks, one for meetings, and one for projects. Depending on the kind, each part has a variety of cards with useful information. We may click on these cards to get to the card page. If I click on a task, it will take me to the task page; if I click on a meeting, it will take me to the meeting page; and if I click on a project, it will take me to the project page.

The next 2 pages in the user menu contain the tasks list and the meetings list. These are similar, the only difference is the data in the tables.

To add a new event, we simply click the + sign on the page. Depending on the type of event, inputs change. We will create a task, which will include the necessary information, if we are on the task page. Only the projects that the user is enrolled in are visible since the selected project dynamically receives the projects we have.

```
  let { projectId, departmentId } =
useParams();
if (projectId && departmentId) {
    tableDetails = {
      type: "Project",
      userId: userId,
      projectId: projectId,
      departmentId: departmentId,
    };
  } else {
    tableDetails = {
      type: "User",
      userId: userId,
      projects: projects,
    };
  }
```

These two pages, which provide lists of tasks or meetings, can be accessed by the current user or by a department within a project. The same component will be used, with a few user-specific differences. I utilized the URL's parameters for this. If it receives the parameters, it means that a user is accessing the page from a project; if it doesn't have parameters, it means that the current user is accessing the page from the navigation bar.

A task or meeting's page is opened up when we click on it. The dashboard also provides access to an event.

The pages of a meeting and a task are similar, but the information displayed and the actions that can be taken are different. For example, while we can access the meeting link in a meeting page, this action is unnecessary in a task page, because we don’t have this information, instead, we have a status and the ability to modify that status. New, Doing, Done, and Closed are the available states.

Change event, see participants, and remove event are the common actions. A popup identical to adding an event's opens when the event is changed, except this time the inputs have a default value, which is the current value. We can view, add, or remove participants to that event by choosing the option labelled "See Members." When we click the button to delete an event, a popup window asks us once more whether we are sure we want to do so.

The projects they are assigned to are listed on the projects page along with basic information about them.

We have more options on some other pages depending on the job role we have. We are unable to edit or add departments to a project's details when we have an employee or a team lead position on the project.

When working as the project manager, we can modify a project's details by clicking the pencil icon, which causes a popup to appear with all the project's editable details. If we want to add a new team lead on the project or a department, we click on the plus. A modal will open, in which we are asked for the username of the person we want to add. If we add a team lead, the department to which it belongs is automatically added.

As a Team Lead, you have two options: eliminate the department (by clicking on the trash icon in the first section) or alter the users (add or delete) (by clicking on the pencil to the right of the Members). These features are not available if we have the position of Employee on the project. As a result, they will not appear.

The task and meeting options will open the same component as on the user's task and meeting pages, but this time the department's tasks will be displayed.

```
findDepartmentProject: async (req,
res) => {
    const { Op } =
require("@sequelize/core");
    const { projectId, departmentId } =
req.params;

    await ProjectAllocationDB.findAll({
      where: {
        projectId: projectId,
        type: {
          [Op.and]: [{ [Op.ne]:
"PROJECT_MANAGER" }, { [Op.ne]: "CEO"
}],
        },
      },
      include: [
        {
          model: UserDB,

          attributes: ["name", "photo",
"id"],
          where: {
            departmentId: departmentId,
          },
          include: [
            {
              model: DepartmentDB,

              attributes: ["name"],
            },
          ],
        },
      ],
      attributes: ["type"],
      order: [["type", "DESC"]],
    })
      .then((event) => {
        res.status(200).send(event);
      })

      .catch((error) => {
        console.log(error);
        res.status(500).send({ message:
"Server error" });
      });
  }
```

There are multiple processes in the above controller. We first look for project allocations and exclude the Project Manager and CEO, because they are automatically in the project. We search all the users of that project for those from the department which is given as a parameter of the route. We provide details like name, image, and ID for these users. The department's name is also included. After completing all of these steps, we have a list of users who belong to our department and do not have the role of Project Manager or CEO on the project. Because the letter "e" from "Employee" occurs in the alphabet before the letter "t" from "Team Lead," it is required to sort the types of project allocation in descending order.

The interface between the CEO and the employee hasn't changed much. The new "Reports" page and new features are what distinguish them from one another: It features all of the above functionalities in addition to the ability to add and delete projects. Its role is an extension of the Employee's (**Fig.8.**).

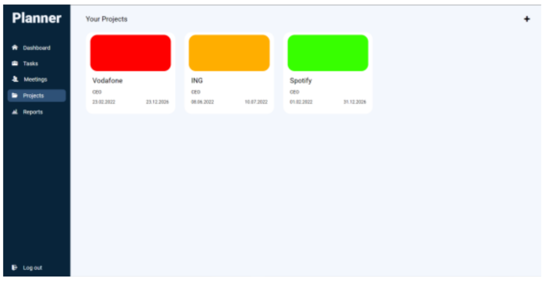


Fig.8. CEO's Projects Interface

The report option, which directs us to a different page, is present in the navbar. The "+" symbol can be seen on the projects page to the right. When the button is touched, a pop-up window appears asking for all the necessary data to establish a project.

```
 postProject: async (req, res) => {
    ProjectDB.create({
      name: req.body.name,
      description:
req.body.description,
      color: req.body.color,
      startingDate:
req.body.startingDate,
      endingDate:
req.body.endingDate,
    })
      .then((project) => {
        ProjectAllocationDB.create({
          projectId: project.id,
          userId: req.body.userId,
          type: "CEO",
        })
          .then((pjAllocation) => {

res.status(200).send(pjAllocation);
          })
          .catch((error) => {
            console.log(error);
            res.status(500).send({
message: "Server error" });
          });
      })
      .catch((error) => {
        console.log(error);
        res.status(500).send({ message:
"Server error" });
      });
  }
```

The logic for starting a project is seen in the above controller. When we initially get the data from the body, we use it to make a project with name, description, color, start and end date. If the project was created successfully then, we assign the user who made the request as CEO.

Whenever we start a new project, the presence of a Project Manager is requested. By selecting the "Add project manager" button in the first section, we can accomplish this. In addition to the project's edit button, we also have a trash button that, when touched, displays a pop-up asking us to confirm that we want to destroy the project. We do this to prevent accidents from occurring if the trash button is unintentionally pressed.

The report page is quite simple, it has two options: departments or users.

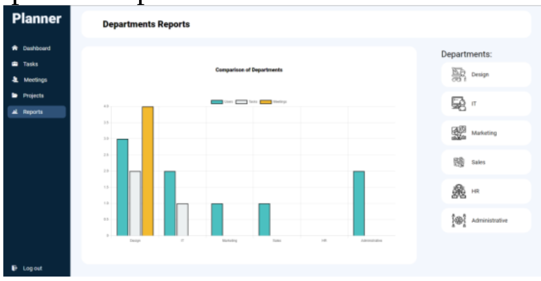


Fig.9. Departments Reports

There are three sections on the department reports page (**Fig.9.**). The page name appears in the first part. A barchart in the second section compares various statistics between departments. The list of departments is represented in the third section. I utilized a third-party library named *chart.js* in the second portion. We can compare tasks and meetings, users and tasks, users and meetings, or even follow the differences between these metrics individually by clicking on one of the rectangles in the legend, which is dynamically erased from the chart. The report page for a department will open if we click on a department in section 3 of the screen.

There are two bar charts on each department's page of individual reports. One describing the distinction between meetings and tasks and another describing projects and members. The functionality of these bar charts is identical to that of clicking on the legend. To make it simpler to calculate additional types of statistics, there is a summary below this. The "See Members" option also directs us to a list of the department's employees. The only distinction between this page and the one in which is reached by selecting the alternate option (the Users on the Reports page), is that on this page, only the users in the department from which we select "See Members," are shown.

On this page, which we access by clicking on "Users" on the "Reports" page, there is a list of all the users on our page. It can be seen that there are users with the color yellow text (which means that person has the role of CEO) and green (indicating that the person is Support). We also have the choice of adding a new user; however, we are not asked for the username or password because those are produced automatically in the backend.

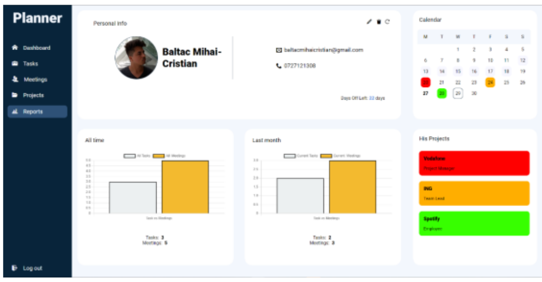
When we click on a user on the page with all users it will lead us to that user's report page. The first two dashboard sections are found on this page; the only difference is that, on the calendar, there is no longer a context menu. Therefore, nothing happens if we click on the date.

The following two ("All time" and "Last month") show the reports from meetings and assignments throughout various time periods. The final panel matches the dashboard's panel for Projects.

```
getStatsLastMonth: async (req, res)
=> {
    const { Op } =
require("sequelize");
    const lastMonth = new Date(new
Date().setDate(new Date().getDate() -
31));

    const { userId } = req.params;
    UserDB.findOne({
      attributes: [],

      include: [
        {
          model: EventAllocationDB,
          attributes: ["eventId"],
          include: [
            {
              model: EventDB,
              attributes: ["type",
"endingDate"],
            },
          ],
        },
      ],
      where: {
        id: {
          [Op.eq]: userId,
        },
      },
    })
      .then((event) => {
        event.dataValues =
event.dataValues.EventAllocations?.map((
e) => {
          if
(e.dataValues.Event.dataValues.endingDat
e > lastMonth)
            return
e.dataValues.Event.dataValues.type;
        });

        let noOfTasks = 0;
        let noOfMeetings = 0;

        event.dataValues.forEach((e) =>
{
          if (e == "TASK") noOfTasks++;
          if (e == "MEETING")
noOfMeetings++;
        });

        event.dataValues.forEach((e) =>
{});
        event.dataValues.Task =
noOfTasks;
        event.dataValues.Meeting =
noOfMeetings;

        res.status(200).send(event);
      })
      .catch((error) => {
        console.log(error);
        res.status(500).send({ message:
"Server error" });
      });
  },
```

We have a bit more difficult process in the controller shown above. The month's past date is the first thing we obtain (we assumed that each month has 31 days). The following stage incorporates gaining access to the database and discovering all the Events that our user attended. If the activity proceeded as intended after getting this information, we compare the deadline for each event. Those that occurred less than a month ago will alter by store only the type of the event, and the other data is discarded as unnecessary. Following the completion of this phase, we go through our list once more and count the instances of events of type "Task" and "Meeting" to add to the variables noOfTasks and noOfMeetings, respectively. The data must then be removed and the calculated numbers for tasks and meetings added.

As an extension of the CEO user, the Support user gets access to all of the CEO's application features. It can also add, edit, and remove departments. It can also alter and delete users. It can also reset passwords. He can view all tasks, meetings, and projects on the dedicated pages, which is another distinction.



Baltac Mihai-Cristian

baltacmihaicristian@gmail.com

0727121308

Days Off Left: 22 days

Tasks: 3
Meetings: 5

Tasks: 2
Meetings: 3

Vodafone
Project Manager

ING
Team Lead

Spotify
Employee

Fig.10. User Reports Support Interface

4 Conclusions

The topic of human resources management is extremely complex, and tactics are always evolving. We aim to allocate resources as effectively as possible so that each project can be structured utilizing a methodology. Numerous applications are continually changing, and new standards are always being released. Both new jobs and modified new processes are established. The application that I've described in my paper is a way to enhance teamwork and effective time management within an organization, assisting both staff members and business owners. The application aims to combine the functionalities currently utilized by the majority of businesses while giving users complete control to develop and alter plans in order to reach their objectives.

By allowing users to customize the sequence of events, project management tactics are made easier to utilize, from choosing which departments we include on each project to structuring tasks and meetings. To make it simpler to build an effective report, the application tries to take into account every event that a user has. Business leaders can track and organize their human resources with the use of the numerous reports produced on various matrices.

Although it has a friendly interface and is a complex tool, anyone can use it. By employing applications that can perform more and can manage these components more simply, I hope to reduce the number of programs needed by businesses.

These standards are always evolving; therefore, the program needs to adapt and provide new functionality as needed. Considering the possibility to extend the application and to make it as efficient as possible, I can state some improvements that the application must have. These are:

- Developing a mobile app (currently the application can only be used on the desktop).
- The user's ability to personalize the application, from the colour scheme to the grouping of dashboard parts.
- Connecting to other calendars, such as Google Calendar and Outlook, so that when an event is added to one, it should also appear in the other.
- Linking up with other user-used applications.
- The ability to send event notifications through email and phone number.
- Making it possible for users to write comments on events would make user communication more effective.
- The ability to connect several events.

- The ability to upload files, including movies and text documents.

References

[1] Loten, A., "*Employees Are Accessing More and More Business Apps, Study Finds"*, The Wall Street Jurnal, 7 February 2019, https://www.wsj.com/articles/employees-are-accessing-more-and-more-business-apps-study-finds-11549580017

[2] Marson, L., "*The 7 most impactful employee experience apps"*. Tech Target 31 August 2020, https://www.techtarget.com/searchhrsoftware/tip/The-7-applications-that-impact-employee-experience-the-most

[3] Atlassian. 2022. "Jira | Issue & Project Tracking Software | Atlassian." <https://www.atlassian.com/software/jira?bundle=jira-software&edition=free&tab=release>.

[4] A. and Bush, A., "React.js essentials", Packt Publishing, 2015, p.5.

[5] DEV Community. 2022. "React Virtual DOM Explained in Plain English https://dev.to/adityasharan01/react-virtual-dom-explained-in-simple-english-10j6

[6] Giraudel, H. and Suzanne M., „Jump start Sass"., SitePoint, 2016, p.139.

[7] Satheesh, M., D'Mello, B. and Krol, J. "Web development with MongoDB and NodeJS" , Packt Publishing., 2015, p.2.

[8] Endeev.com. 2022. "Sails.js – Beginning with Node.js – Endeev". <http://www.endeev.com/blog/sails-js-beginning-with-node-js/>

[9] Azat M., "Express.js Guide". Lean Publishing, 2014, p. 2.

[10] Developer.mozilla.org. 2022. "Express Tutorial Part 4: Routes and controllers - Learn web development | MDN". <https://developer.mozilla.org/en-US/docs/Learn/Server-side/Express_Nodejs/routes>].

[11] Openbase. 2022. "10 Best Node.js MySQL ORM Libraries in 2022 | Openbase". <https://openbase.com/categories/js/best-nodejs-mysql-orm-libraries> .

[12] Thakkar, R., 2022. "Mechanisms of Sequelize.js". DEV IT Journal. Available at: <https://www.blog.devitpl.com/sequelize/> .

[13] Schwartz, B., Zaitsev, P., Tkachenko, V., Zawodny, J., Lentz, A. and Balling, D., 2008. "High performance MySQL". 2nd ed. United States of America: O'Reilly, pp.1,2.

BÂLTAC Mihai-Cristian is a graduate of the Faculty of Economics Cybernetics, Statistics and Information at the Bucharest University of Economic Studies, bachelor's degree in Computer Science, mostly interested in web development and automation.

Making use of digital innovations in Business Process Improvements

Radu SAMOILĂ
Bucharest University of Economic Studies
radusamoila2@gmail.com

Business Process Management (BPM) concept is more and more influenced by the emerging technologies changing the conventional way of improving or optimizing the business processes. Digital innovations and technology have been used to improve and manage people, products, programmes and projects across the globe. Connected devices, big data analytics, cloud computing, robotics process automation, 3D printing or other emerging technologies are commonly used to generate more efficient and effective business processes.

Therefore, nowadays, the businesses are continuously undergoing changes which can be rapid and significant. There are many methodologies/ approaches available to support the businesses improve their processes through change. A strong connection exists between business process improvements and digital innovation as, through a proper combination, has a great potential of generating significant long-term benefits for organizations. Hence, focusing the organization's strategies on digital technology can be a successful direction.

The purpose of this paper is to present potential ways of integrating process improvements methodologies with digital innovation and the main market trends. It focuses on market trends concerning business process improvements and digital innovations. The work encompasses a 'status quo' review in this field together with the main trends in terms of new technologies and their adoption by organizations. Companies started to utilize a wide range of communication channels, integrated technologies or social media platforms to connect with their peers, employees, and clients but also to boost collaborative partnerships. Technology is used to create more participatory businesses by improving collaboration. Furthermore, newest technologies can support effective monitoring of business processes across diverse products and services and counterparties (e.g., suppliers, clients).

This work's conclusions confirm the significant role of digital innovations in business process improvements and provide further insights on how to embed a wide range of new technologies within the organizations' efforts to improve their business processes and operations.

Keywords: *process improvement, improvement methodology, digital innovation, emerging technologies, process mining, business process management*

1 Introduction

The number of technological solutions is evolving in a rapid pace and, nowadays, these solutions are bringing a new industrial revolution as well as the extension and changes to the current ways of doing business. With the rapid creation and adoption of new technologies (e.g., blockchain, Internet of Things (IoT) or artificial intelligence), organizations are struggling to take maximum advantage of new IT [1].

In response, business operations, business structures and processes need to learn how to adapt and implement a new version of business process management (BPM). This is now being called 'ambidexterity'. The BPM is called ambidextrous if it focuses on the main two aspects:

- exploiting the benefits of existing technologies (i.e., exploitative BPM) and
- exploring the benefits of new IT technologies (i.e., explorative BPM) [2].

New technologies enable disruptive digital innovations (i.e., DI or innovations with new technologies) which are elementary prerequisites of sustainable business processes (i.e., company's long-term way of working). While product / service innovations are a potential feature for organizations to lead in the market, digital process innovations support with regards to reducing delays (time) and resources [3]. In addition, digital innovations are transforming both the client needs and the infrastructural requirements. New technologies, such as blockchain, IoT, process mining, robotic process automation, artificial intelligence, virtual reality and 4D printing, have the potential to disruptively change business processes.

With regards to the BPM, it is generally accepted that the business processes follow a cycle from the process identification stage, moving to implementation and further to the monitoring and control stages [4]. There are various studies available that highlight the BPM maturity model [5], the BPM core elements [6] and BPM context factors [7]. In the last period, students started focusing on new topics like green BPM, the human aspects of BPM, social BPM and ambidextrous BPM [8]. It was also noted aspects covering how two streams of BPM and digital innovations can be combined and highlighted benefits of common methodologies. Other studies presented seven paradoxes concerning BPM and related synergies with IT by suggesting smart devices and digital transformation. The changing dynamics of high-speed internet and digital technologies are thus also entering the BPM discipline.

The open innovation is one of the key factors needed to ensure sustainable development through change in business processes and operations [10]. In response, to let the BPM discipline better prepare for a digital knowledge economy, there were several studies conducted through an expert panel with practitioners' opinions on future BPM trends covering to emerging technologies and digital innovations. This study revealed seven BPM-DI trends based on experimental data only, however this paper's work is to supplement these trends with practical examples and further critical thinking to substantiate the extent to which the current level of knowledge lectures each trend, and to get better information regarding the existing differences between what is practical relevant and the availability of the current knowledge. The current research purpose is to explore any uncovered aspects of BPM available methodologies and researches in relation to the newest digital technologies from the past few years and current trends. This research's principal benefits would be to note the main advantages of digital innovations that were less or not explored so far in relation to BPM activities. Therefore, to get to a well-informed research and information with regards to BPM and digital innovation synergies, this paper's main question at this point would be with regards to any unexplored areas in 'status quo' concerning the application of digital innovations in business process management.

2 Background

This paper provides next some relevant details with regards to BPM and DI, then it provides further information on the main trends noted in the market and some useful examples on the available technologies that goes hand in hand with BPM concept.

2.1. Business Process Management

BPM is often defined as a complex set of techniques to discover and understand a business process, to re-shape designs for that processes, monitor it by defining proper metrics and data, as well as by optimizing and automating the processes by considering technological, financial and human resources. In addition, some other researchers have noted this set in a BPM lifecycle with various phases to address a business process, namely iterations that begin with process identification and process discovery, then process analysis and redesign, leading to the actual implementation and then monitoring and control [4].

Above, the paper shows the main phases of a BPM lifecycle; each of these phases require innovation in order to increase the process operating pace or speed [11]. These innovations should also closely adhere to organizational goals, both explorative and exploitative goals so to achieve the needs of ambidexterity in a digital knowledge economy. With enhanced flexibility features, ambidextrous BPM is more dynamic and extends traditional BPM with a more balanced view between incremental and innovative process changes.

The way of approaching the BPMs and the continuous transformation of knowledge support the transformation approach leading to a more dynamic BPM. For example, sharing of know-how coming from experience persons to new joiners among process teams is crucial for the BPM success [12]. Binci et al. [13] presented four project-based factors including (1) task specialization, (2) knowledge transfer, (3) conversion of knowledge and (3) ambiguity and change management, that would support in the adoption of ambidexterity.

Fast pace innovations in business processes increase the productivity and support the overall company's financials improvements. Hence, if BPM would dynamically change in a constant way this would facilitate the organizational performance improvements in multiple perspectives. Given that the business process modelling is seen as prominent BPM sub-areas, which are now reshaping abruptly, prior researches have contributed to these domains while other BPM sub-areas such as ambidextrous BPM have not been under the focus from an innovation point of view [2].

Digital Innovations

The newest technologies are applied by digital innovations in order to solve most part of the existing business issues and to improve current practices so to achieve new transformation or business models, processes, products or services. Some of the emerging technologies proved to have a positive impact on the execution of the processes' activities or task which allow a better coordination among work teams and impact the entire BPM lifecycle, albeit more influential at the re-design phase.

There are various examples of digital innovations that involve easy and rapid integrations between IT environment and operations, secure payments technologies or automatic price updates. Or smart devices can be used to upsurge process improvement for an organization to go faster and within budget. Interoperability between the BPM lifecycle phases and IT innovations is important to achieve best benefits from information and data. There-fore, the strategical and operations levels of organizations are both impacted by digital innovations.

2.2. BPM and Digital Innovations trends

As above mentioned, the BPM concept is under trans-formation in the digital economy in order to create new opportunities for improving and automating business processes. The newest IT technologies are able to auto-mate a high degree of manual interventions within a business process with the support of internet solutions and / or intelligent tools. For example, the use of social media and related tools can connect more easily the business products with their consumers leading to in-creased sales and facile access to consumers' feedback. The technology can support with real time data analysis for tracking and monitoring in a fast and efficient way.

There is a general understanding that digital transformation can reshape the BPM concept, however, there is a need of more research to fully grasp the opportunities and related benefits. This is the reason for consolidating the opinions of the main BPM practitioners regarding how they understand the BPM future linked to the technological developments. Those are being categorized into the following main trends that could govern the markets:

1) Customer experience changing continuously;
2) Strong synergies between BPM and digital innovation;
3) Faster innovations, drive process changes, challenge current way of working;

4) Increased alignment between IT and business operations;
5) BPM is gaining traction in the organizations (e.g., In Process Modelling and Monitoring);
6) Reduced resistance in relation to BPM and DI.

2.2.1. Customer experience changing continuously

The first trend related to BPM and DI refers to the fact that the digital tools continues to change in a constant way the experience of business customers and this possibly with an increased speed. By using enriched data management and big data analytics, companies can make more use of data for incorporating customer-centric offerings [14].

Customer experience sits at the core of the business process improvement concept. Market responsiveness and developing proper customer value propositions are the basics of developing a great customer experience. The organizations can own a significant amount of data and applying big data analytics can help identify and differentiate between customer profiles based on a faster retrieval of information than before. Ultimately, providing a customization facility can improve customer relationships, stimulate customer engagement and determine/predict consumer behaviors [15].

There are various tools available to improve customer experience, such as data mining, machine learning or artificial intelligence. Data mining relates mostly to discovering patterns in large datasets using real-time customer data, machine learning is about the scientific study of algorithms and models that information systems apply to perform tasks without or minimum human instructions but with machines behaving human beings. Artificial intelligence also uses big data to derive decisions and for making predictions. For example, nowadays, many organizations are already using is a Customer Relationship Management system for storing and sharing real-time information of customers. There-fore, with the application of all these tools or solutions, the customer experience can change massively due to the interventions of emerging technologies, and this will only increase in the future.

2.2.2. Strong synergies between BPM and digital innovation

The newest technological developments are responsible for this shift towards ambidextrous BPM. Most part of the organizations already use the traditional BPM methods and techniques, however, the explorative way of doing BPM will support to boost a culture of timely communication and collaboration (though the use of communication technologies or social media) and entrepreneurship to identify new techniques and new ways of doing business (e.g., Delivery Hero). In this case, the success of BPM is boosted by the strategic adoption of newest IT technologies. It is important to note that the business process goals should be in line with the organizational goals but the alignment between IT and business is of a significant important for a successful BPM.

Currently, the BPM concept must deliver value out of employees and customers, this being called value-driven BPM. Furthermore, the key noted from the market practices is the need to get a proper balance between the exploitative and explorative business processes to achieve organizational performance. For example, the trend discussed in this section includes big data management that show how big data can be used and linked to digital innovation and BPM. In an ambidextrous environment, the role of big data for creating a balance between exploitation and exploration is discussed less within the existing literature. Usually, the newest technologies in the market are rapidly adopted by organizations in their aim to gain a relative competitive advantage. IT enables organizations to get the maximum benefits from the available data. Therefore, changing an organization’s strategies towards digital technology can be a successful way towards achieving best benefits.

2.2.3. Faster innovations, drive process changes, challenge current way of working

Business processes can be more efficient and faster through applying the agile principles. By looking at the traditional (exploitative) BPM approaches, Six Sigma, Lean Six Sigma or lean manufacturing have been used since many years and which are in line with the continuous process improvements concept. In the same sense, Total Quality Management (TQM) is used to in-crease the quality of the business processes while the relevant ISO standards (e.g., ISO 9000 standards) are used in relation to various products / services and organizations.

Regarding the exploration reasons, BPM requires a combination of standardization in today's high-speed internet with an increased awareness of the DI potentials. Hence, understanding more about the new technologies is paramount to improve business processes. BPM managers and practitioners must be trained in time management, so they can promote teamwork in their teams and projects. Similarly, project management skills are highly important to manage each BPM lifecycle phase.

In addition, BPM maturity models have an important role in the adoption of digital technologies. For instance, [16] contributed to a well-defined maturity model involving strategic alignment, culture, people, governance, method and IT elements, which revealed how these core elements can contribute to BPM success (albeit with a stronger focus on exploitation).

There can be concluded that digital technologies raise new opportunities for innovation by sharing information externally (i.e., outside of the organization). Innovation in business processes is positively associated with an information exchange towards an organization's environment. The ease of use and perceived usefulness of the emerging technologies also contribute to a positive integration with business processes. Agile business process improvements is possible in different ways. One way is to divide the innovation project into sub-tasks and to integrate them with the help of digital technologies. Another way is using BPM knowledge with user-friendly BPM systems or suites (BPMS). Knowledge transformation in BPM enables faster communication, a deeper understanding and a rapid execution of tasks. Therefore, unspoken know-how should be converted into spoken knowledge in BPM scenarios. BPM is reshaping in such a way that it becomes more agile and faster in critical cases.

2.2.4. Increased alignment between IT and business operations

The IT capabilities should be used in order to gain competitive success and continuous strategic alignment. The alignment between business and IT refers to the required integration between the business strategy and the company's IT strategy, but also between the business and its IT structures. This alignment type remains a major concern to be assessed by IT departments. There are various research studies performed before that examined the alignment between business and IT, such as its measures and outcomes. In addition, alternative studies discussed the ongoing nature or sustainability of business - IT alignment [17].

Nonetheless, business - IT alignment is one of the key areas required to be explored for a successful BPM in the twenty-first century, for which the IT architecture constitutes an important pillar. The company's process architecture must be aligned to the entire enterprise architecture in order to ensure a smooth execution of the activities and related tasks. Business - IT alignment is strengthened through the collaboration in each BPM lifecycle phase and helps achieve a more rapid processing time, improved customer experience, beneficial technological transformations, achieving IT agility and increased collaboration. Furthermore, the overall financial performance of the companies could improve as well. The alignment between IT and business processes provides further support for customer involvement and allows the companies to get closer to the digitized solutions, such as RPAs or Artificial Intelligence tools.

2.2.5. BPM is gaining traction in the organizations (e.g., In Process Modeling and Monitoring)

Previous research studies show that the traditional (exploitative) BPM approach received several criticisms by claiming that it can be way too bureaucratic. Now, the newest technologies offer the possibility to the BPM to become more attracting for the organizations with regards to practicing new ways of process modelling and monitoring. More appealing things are happening on the BPM exploration domain, such as journey mapping through a comic book style [18], which strongly contrasts with the traditional process languages (e.g., process diagrams in BPMN and UML). Real-time application monitoring tools are useful for monitoring an IT infrastructure.

Additionally, network monitoring tools are used more and more by the organizations given the benefits brought. Furthermore, explorative tools have been designed for more demand-driven, case-driven and value-driven BPM. Knowledge management tools are introduced to derive knowledge-intensive processes that perform in unexpected conditions.

Similarly, knowledge-intensive BPM works in unstructured environments by using knowledge to promote employee involvement in process improvements projects. Other examples are intelligent neonatal monitoring systems using multi-sensors for intelligent monitoring. The above-mentioned explorative BPM examples also turn out to be successful. For instance, studies showed that a business intelligence implementation in BPM escalates the performance of corporate performance management [19]. Knowledge management in BPM appeared to ensure the quality of data and information. On the other hand, reducing carbon footprints across the BPM lifecycle stages are vital steps towards achieving green BPM. Nonetheless, while digital process innovations help advance process analytics and trigger a new generation of process modelling and of organizational capabilities by emerging technologies, these types of technologies would decrease in future human interventions in BPM.

2.2.6. Reduced Resistance in relation to BPM and DI

The latter trend in relation to BPM & DI shows a reduced degree of resistance against process change through promoting an adaptation culture in digital technologies and a learning organization. Until now, change management models like Lewin’s change management model and the McKinsey 7-S model have been applied in BPM. Demonstrated techniques for managing process changes are culture mapping, metrics and flow chart, force field analysis. New curricula in IT and BPM confirm that change management remains beneficial in removing the hindering factors in BPM and learning. For example, a future BPM trend could include teaching BPM practices, teaching BPM as a problem-solving domain and teaching about the technology-driven benefits of BPM. A paradigm shift from exploitative BPM to explorative BPM is seen as a must to be considered in future BPM curricula.

Educating the people and organizations about BPM depends on the effective utilization of available data, namely how organizations use the data regarding employees and customers. Intangible metrics or elements, such as job satisfaction, job and performance engagement, can be determined by data with the help of technologies in a BPM environment. Evaluation criteria and measuring standards can be made available to unexperienced employees for reasons of learning. Employee participation in strategic process decisions is inevitable for organizations to avoid an integration cost on a later stage. Experienced-based learning considers experience as the main method of learning for BPM tools and techniques. A learning cycle can be used to transform tacit knowledge into work patterns.

In addition to the impact of digital innovations on BPM, other factors such as social culture and work culture also have a promising role in reinventing BPM. For example, an educated society with an open culture is less resistant to change, and therefore more open to disruptive process changes. Similarly, digital innovations also have an impact into the social culture. In other words, the BPM concept is not only reshaped by technological factors but also cultural changes which reinforce the former.

Over-all, all expected trends can already be observed in the literature, at least to some extent and with different dimensions. Based on these trends, the intention is to incorporate them into an updated business process improvement methodology that would cover the features brought by the newest BPM conceptual ideas.

3. Emerging Technology - Deep dive into Process Mining

The research performed within several high-scale organizations showed some of the mainstream digital innovations and concepts used at the moment. **Process Mining** is a relatively new area of study grounded in a long tradition of businesses striving to optimize business outcomes by improving the efficiency, effectiveness, and productivity of their critical workflows. Process Mining happens in four distinct stages:

a. Collection of time-stamped event log data from key transactional systems
b. Discovery within that data of real processes as they happen
c. Enhancement of those processes to optimize business outcomes
d. Monitoring those changes for further improvement opportunities.

Early core business processes were simple (and often manual); but as businesses have digitized every aspect of working life into IT systems, core processes have become complex operational machinery in and of themselves - too fast, frequent, interconnected and distributed to manage manually.

Process improvement is not a new idea. But the scale and complexity of the modern process environment has quickly accelerated beyond the capabilities of traditional tools. Process mapping software, business intelligence initiative - or worse, whiteboards and post-it notes - just can’t deliver the real-time insight and control that continuous process excellence demands. Process Mining isn’t just about resolving the mountain of complexity and friction that most organizations have unconsciously accepted as the cost of doing business. Its true value is in connecting granular operational performance to high-level business outcomes on a continuous basis.

One of the most used process improvement methodologies is Lean Six Sigma which follows the well-known DMAIC methodological steps, as per *Figure 1* below:

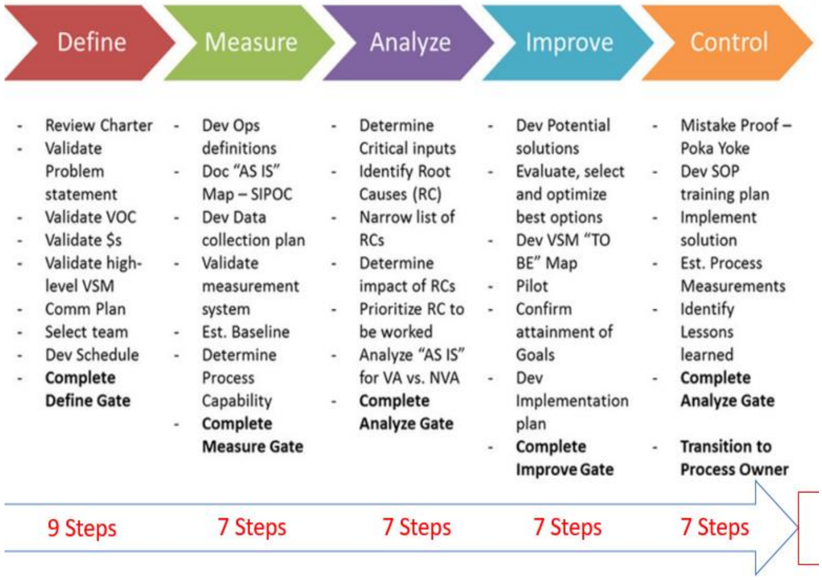

Define
Measure
Analyze
Improve
Control
- Review Charter
- Validate Problem statement
- Validate VOC
- Validate $s
- Validate high-level VSM
- Comm Plan
- Select team
- Dev Schedule
- Complete Define Gate
- Dev Ops definitions
- Doc "AS IS" Map – SIPOC
- Dev Data collection plan
- Validate measurement system
- Est. Baseline
- Determine Process Capability
- Complete Measure Gate
- Determine Critical inputs
- Identify Root Causes (RC)
- Narrow list of RCs
- Determine impact of RCs
- Prioritize RC to be worked
- Analyze "AS IS" for VA vs. NVA
- Complete Analyze Gate
- Dev Potential solutions
- Evaluate, select and optimize best options
- Dev VSM "TO BE" Map
- Pilot
- Confirm attainment of Goals
- Dev Implementation plan
- Complete Improve Gate
- Mistake Proof – Poka Yoke
- Dev SOP training plan
- Implement solution
- Est. Process Measurements
- Identify Lessons learned
- Complete Analyze Gate
- Transition to Process Owner
9 Steps
7 Steps
7 Steps
7 Steps
7 Steps

Fig.1. DMAIC

Hence, the traditional way of improving business processes involves a high number of activities that involves high costs and time to implement. The newest technologies could streamline greatly the improvement process through use of emerging technologies, implying reduction of required steps or activities which would reduce company's cost and time. By making use of the available technologies, the process improvement work could be simplified to fewer activities while the remaining activities are to be automated, letting the technology to play its role.

For example, in one of the largest global FMCG (Fast-Moving Consumer Goods) companies the combination between Lead Six Sigma and process mining technologies was successfully implemented by following three (3) main steps, as per *Figure 2* below:

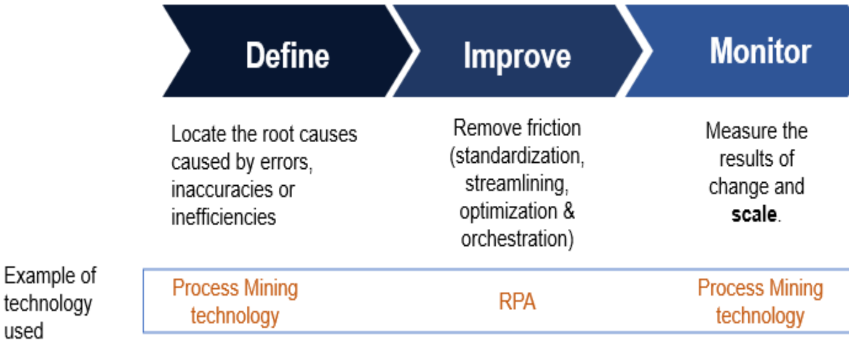

Define
Improve
Monitor
Locate the root causes caused by errors, inaccuracies or inefficiencies
Remove friction (standardization, streamlining, optimization & orchestration)
Measure the results of change and scale.
Example of technology used
Process Mining technology
RPA
Process Mining technology

Fig.2. Simplified process improvement flow

We can see from the two figures above that the technology has played an important role in reducing the number of activities required to identify pain points and to improve the process. Some main steps of Lean Six Sigma methodology have been followed though.

Looking closer to how process mining technology has been used, it followed four (4) steps framework, as follows:

1. Define main goals
2. Define actionable initiatives
3. Identify frictions

4. Define standard metrics and future scalable actions.

To showcase one practical example, the planning process with regards to the process mining process within this company has followed the following main framework:

1. Which are company's main goals?
   a. Improve delivery times to drive enhanced customer satisfaction
   b. Obtain a competitive advantage in relation to the main clients
   c. Scale the process and improve targets and results
2. Which are the actionable initiatives to achieve those goals?
   a. Get full understanding over the order management and fulfilment
   b. Automate so to reduce the cycle times
   c. Get robust information on product management
3. What are the frictions the imped goals realisation?
   a. Most part (>50%) of credit checks were done for low value accounts
   b. Lots of blocked deliveries which required time to process
   c. Repetitive penalties paid to customers due to delays
4. What metrics can be applied to solve the frictions?
   a. Credit checks reduction
   b. Reduced number of blocked orders and improved resolution of those blocked
   c. Increased on-time delivery
5. What further improvement opportunities could be pursued?
   a. Master data clean-up
   b. Increase the number of orders that wouldn't require manual checks
   c. Obtain around GBP 4.5 mln savings.

This type of framework supported the organisation to organise the foundation for aligning the objectives at company level for all pertinent processes.

Automation was a key component of this case which allowed using RPA or bots (UiPath Technology) to perform required actions much faster and error free.

Moving further with the **Order to Cash** example the company followed the adoption of an explorative approach in improving the business processes. The combination of the traditional process improvement methodologies with the use of emerging technologies has led to a successful outcome. In *Figure 3* below there is further information on how the actual Order to Cash process has changed and which are the main benefits.

The new solution combined the best of both sides: reducing manual interventions and cycle times overall, while preserving the high-value human touches where needed.

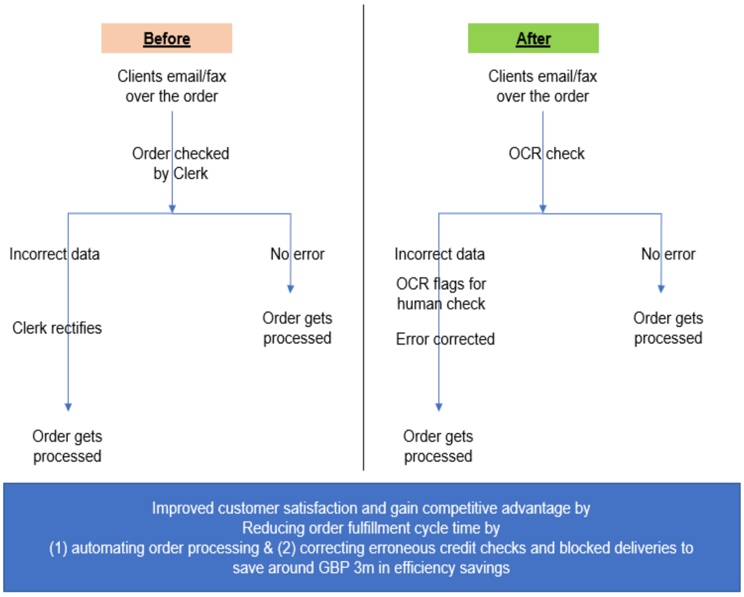
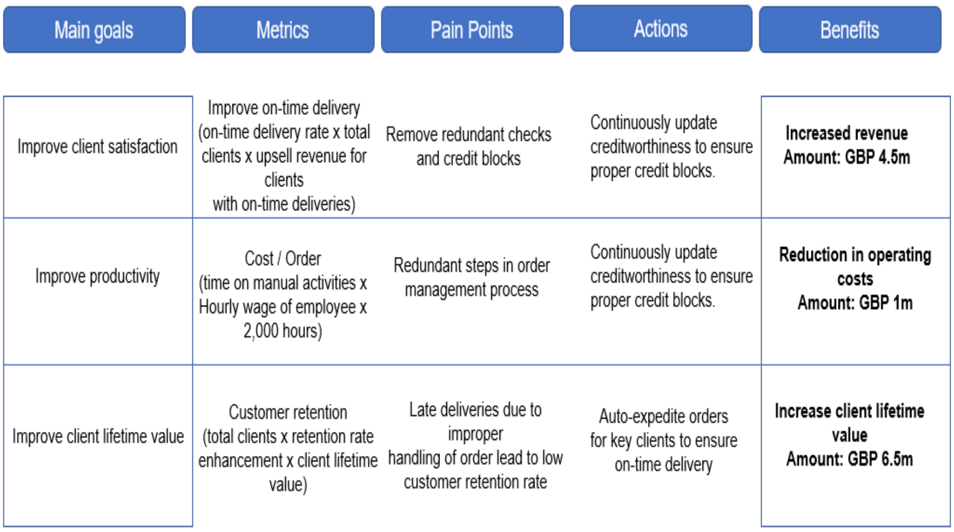

Before
Clients email/fax over the order
Order checked by Clerk
Incorrect data
No error
Clerk rectifies
Order gets processed
Order gets processed
After
Clients email/fax over the order
OCR check
Incorrect data
OCR flags for human check
Error corrected
No error
Order gets processed
Order gets processed
Improved customer satisfaction and gain competitive advantage by
Reducing order fulfillment cycle time by
(1) automating order processing & (2) correcting erroneous credit checks and blocked deliveries to save around GBP 3m in efficiency savings

Fig.3. Order to Cash case

On long term basis, the following benefits are expected to be delivered given the analysis performed over 3-month period (*Figure 4*):

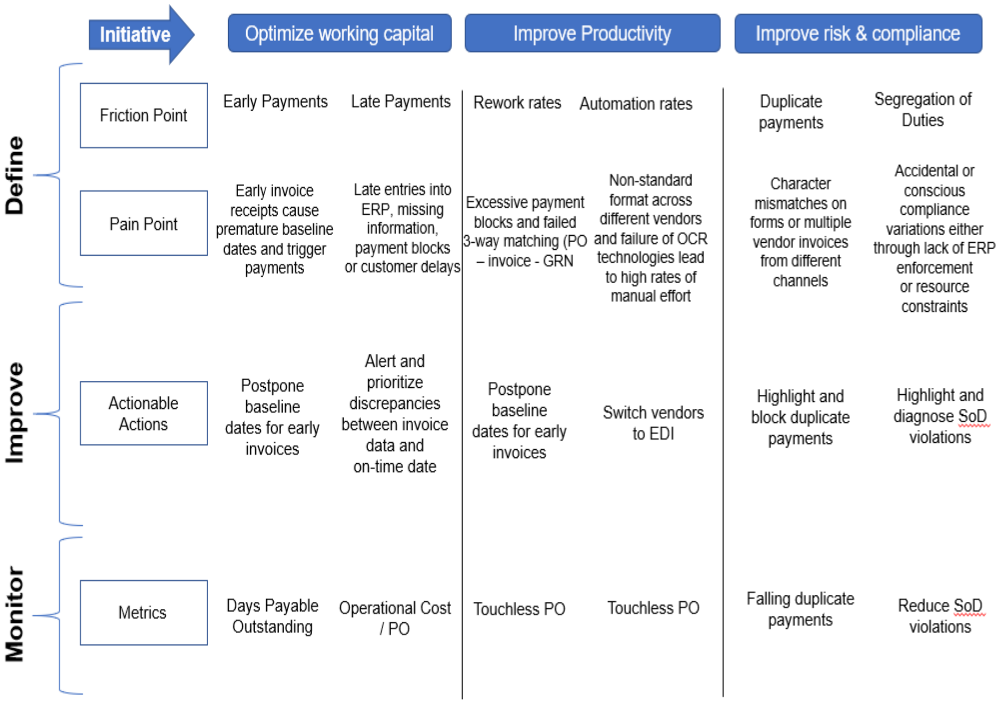


Main goals	Metrics	Pain Points	Actions	Benefits
Improve client satisfaction	Improve on-time delivery (on-time delivery rate x total clients x upsell revenue for clients with on-time deliveries)	Remove redundant checks and credit blocks	Continuously update creditworthiness to ensure proper credit blocks.	**Increased revenue Amount: GBP 4.5m**
Improve productivity	Cost / Order (time on manual activities x Hourly wage of employee x 2,000 hours)	Redundant steps in order management process	Continuously update creditworthiness to ensure proper credit blocks.	**Reduction in operating costs Amount: GBP 1m**
Improve client lifetime value	Customer retention (total clients x retention rate enhancement x client lifetime value)	Late deliveries due to improper handling of order lead to low customer retention rate	Auto-expedite orders for key clients to ensure on-time delivery	**Increase client lifetime value Amount: GBP 6.5m**

Fig.4. Longer term benefits in Order to Cash

These types of process improvement initiatives are often scalable and applied in different cases or business areas. This is what happened next and led to further improvements, as per *Figure 5* below.

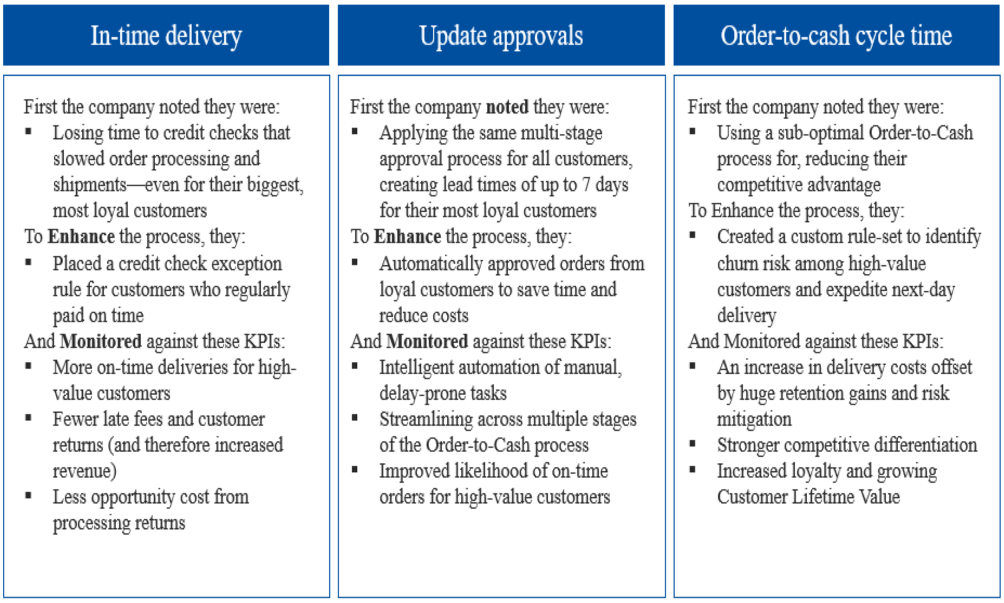
It was very important for the Company that it gradually built-up the success through the hierarchy of business process improvement, as the benefits of the initiative compounded over time.



	Initiative	Optimize working capital		Improve Productivity		Improve risk & compliance	
Define	Friction Point	Early Payments	Late Payments	Rework rates	Automation rates	Duplicate payments	Segregation of Duties
	Pain Point	Early invoice receipts cause premature baseline dates and trigger payments	Late entries into ERP, missing information, payment blocks or customer delays	Excessive payment blocks and failed 3-way matching (PO – invoice - GRN	Non-standard format across different vendors and failure of OCR technologies lead to high rates of manual effort	Character mismatches on forms or multiple vendor invoices from different channels	Accidental or conscious compliance variations either through lack of ERP enforcement or resource constraints
Improve	Actionable Actions	Postpone baseline dates for early invoices	Alert and prioritize discrepancies between invoice data and on-time date	Postpone baseline dates for early invoices	Switch vendors to EDI	Highlight and block duplicate payments	Highlight and diagnose SoD violations
Monitor	Metrics	Days Payable Outstanding	Operational Cost / PO	Touchless PO	Touchless PO	Falling duplicate payments	Reduce SoD violations

Fig.5. Examples of improvement cases

Initially the process improvement begun with a credit check correction exercise that quickly gathered in pace and scope as early quick wins were further reported. The optimization plan went through various stages of analysis over the digital prints of the in-scope process. *Figure 6* presents the process noted by the team, that led to desired optimizations:



In-time delivery

First the company noted they were:
- Losing time to credit checks that slowed order processing and shipments—even for their biggest, most loyal customers

To **Enhance** the process, they:
- Placed a credit check exception rule for customers who regularly paid on time

And **Monitored** against these KPIs:
- More on-time deliveries for high-value customers
- Fewer late fees and customer returns (and therefore increased revenue)
- Less opportunity cost from processing returns

Update approvals

First the company **noted** they were:
- Applying the same multi-stage approval process for all customers, creating lead times of up to 7 days for their most loyal customers

To **Enhance** the process, they:
- Automatically approved orders from loyal customers to save time and reduce costs

And **Monitored** against these KPIs:
- Intelligent automation of manual, delay-prone tasks
- Streamlining across multiple stages of the Order-to-Cash process
- Improved likelihood of on-time orders for high-value customers

Order-to-cash cycle time

First the company noted they were:
- Using a sub-optimal Order-to-Cash process for, reducing their competitive advantage

To Enhance the process, they:
- Created a custom rule-set to identify churn risk among high-value customers and expedite next-day delivery

And Monitored against these KPIs:
- An increase in delivery costs offset by huge retention gains and risk mitigation
- Stronger competitive differentiation
- Increased loyalty and growing Customer Lifetime Value

Fig.6. Optimization process undergone by the company

The optimization of expedited delivery appeared initially to be a 'no-go' as the company was having issues with on-time delivery on standards sales orders. However, the technology used (Celonis Process Mining) made available this important business case.

4. Conclusions

Emerging technologies and digital innovations provide the organizations with further opportunities to reshape and streamline the BPM. Digital process innovations help accomplish tasks in faster and smarter ways. For instance, smart cities take advantage from IoT devices for doing technology-enabled monitoring. The BPM combined with the newest technologies can change the value propositions of customers, which opens new avenues to develop a strategic alignment between the organizational policies or rules on the one hand and the BPM features on the other hand.

The research studies performed so far could support to differentiate between the yet covered BPM - DI themes in the literature and the still uncovered avenues that would support further development in the BPM area, including process improvement and optimizations. Process mining is a technology which gains more and more attraction for organizations given the benefits noted so far in relation to a new way of approaching the explorative BPM. Process mining is a clear example of the explorative BPM approach. The combination of traditional busines process improvements methodologies and newest IT technologies on process mining (such as UiPath Platform or Celonis Process Mining), for example, allows additional flexibility as an action driven system for identifying and solving complex issues and scalability towards obtaining ambitious business outcomes.

The technology revolutions and adoptions during last years gave a paradigm shift for managing business processes digitally, as BPM goals are not only set to target organizational goals BPM strategies aligned with IT, employees, customers, etc., brings value-driven process because a well-defined BPM leads to an innovative and adaptive way of working. The context of digital transformation requires a rethinking of the dominant assumptions that have characterised how we think of BPM and triggers further development in the methodological space to respond to the following questions:

1. Are the traditional methodologies still efficient in the context of emerging technologies that proved to be successful?
2. Is there a need to enhance the current process improvement methodologies that would consider more the emerging technologies' features?
3. Which are the criteria the organisations should use to decide what technology to adopt given its aim of improving processes?

The overall results show the need for further efforts and research with regards to the market practices depending on the industry the organizations operate on. But we consider that adapting the traditional process improvements framework to the available technologies in the market (e.g., process mining) will provide the foundation for a cohesive strategy on a case-by-case basis. An adapted framework like this gets the whole organization aligned in one place, from process practitioners to senior stakeholders. We intend to continue with integrating the ideas of adapting process improvements methodologies or concepts to the emerging technologies which would have the potential to obtain new and better framework that will support a smoother, more efficient and less costly way to improve business processes.

References

1. Singh, S., Rathore, S., Park, J.H., BlockIoTIntelligence: A Blockchain-enabled Intelligent IoT Architecture with Artificial Intelligence. Future Gener. Comput. Syst. 2019.
2. Ferraris, A., Monge, F., Mueller, J.: Ambidextrous IT capabilities and business process performance: An empirical analysis. Business Process Man-agement 2018.
3. Rosemann, M., Proposals for future BPM research directions. In Asia-Pacific Business Process Management; Springer: Cham, Switzerland, 2014.

4. Dumas, M.; La Rosa, M., Mendling, J., Reijers, H.A. Introduction to Business Process Management. In Fundamentals of Business Process Management; Springer: Berlin, Germany, 2013.

5. Tarhan, A., Turetken, O., Reijers, H.A. Business process maturity models: A systematic literature review. Inf. Software Technology 2016.

6. Rosemann, M., Brocke, J., The Six Core Elements of Business Process Management; Springer: Berlin/Heidelberg, Germany, 2015.

7. Vom Brocke, J., Zelt, S., Schmiedel, T. On the role of context in business process management. Int. J. Inf. Manag. 2015.

8. Ahmad, T., Looy, A. Van Reviewing the historical link between Business Process Management and IT: Making the case towards digital innovation. In Proceedings of the IEEE Thirteen International Conference on Research Challenges in Information Science, Brussels, Belgium, 2019.

9. Mendling, J., Pentland, B., Recker, J.: Building a Complementary Agenda for Business Process Management and Digital Innovation. 2020

10. Yun, J.H.J., Jung, W.Y., Yang, J.H. Knowledge strategy and business model conditions for sustainable growth of SMEs. J. Sci. Technol. Policy Manag. 2015.

11. Marrella, A. What Automated Planning Can Do for Business Process Management. In Business Process Management Workshops; Springer: Barcelona, Spain, 2017.

12. Koopman, A., Seymour, L.F., Factors impacting successful BPMS adoption and use: A South African financial services case study. In Enterprise, Business-Process and Information Systems Modeling; Springer: Cham, Switzerland, 2020.

13. Binci, D.; Belisari, S., Appolloni, A. - BPM and change management: An ambidextrous perspective. Bus. Process Manag. J. 2019.

14. Spiess, J., T'Joens, Y. Dragnea, R., Spencer, P., Philippart, L. Using big data to improve customer experience and business performance. Bell Labs Tech. J. 2014.

15. Bleier, A., de Keyser, A., Verleye, K. Customer engagement through personalization and customization. In Customer Engagement Marketing; Macmillan, P., Ed.; Springer International Publishing: Cham, Switzerland, 2018.

16. Rosemann, M.; de Bruin, T. Towards a business process management maturity model. In Proceedings of the 13th European Conference on Information Systems (ECIS 2005), Regensburg, Germany, 26–28 May 2005; The London School of Economics: London, UK, 2005.

17. Wong, C., Skipworth, H., Godsell, J., Achimugu, N. Towards a theory of supply chain alignment enablers: A systematic literature review. Supply Chain Manag. 2012.

18. Veale, T., Feyaerts, K., Forceville, C. Creativity and the Agile Mind: A Multi-Disciplinary Study of a Multi-Faceted Phenomenon; Walter de Gruyter: Berlin, Germany, 2013.

19. Richards, G., Yeoh, W., Chong, A.Y.L., Popovic, A. Business Intelligence Effectiveness and Corporate Performance Management: An Empirical Analysis. J. Comput. Inf. Syst. 2019.

Radu SAMOILA has graduated the Master of Economy and Information Technology in 2011 at the Bucharest University of Economy. Currently he is a PhD Student at this university, since 2019. Main fields of interest are business process optimization, automation of business processes and the continuous improvements programmes. His background is mainly finance and the main expertise is on internal auditing, external auditing and advisory practices. At present he is leading the Europe Internal Audit Team at a Global FMCG Company.