Prediction Model for Soybean Productivity

Ion GANEA
Moldova State University, Chisinau, Republic of Moldova
iganea9@gmail.com

Abstract. *This paper presents a holistic approach to biological and agricultural research focused on the use of interconnected technologies in the context of climate change. Researchers from different countries have analyzed how smart technologies can help agriculture adapt to these changes. The most representative works in the field are analyzed. Among these technologies are graph database systems such as Neo4j, which have demonstrated success in predicting the studied phenomena. The paper describes the development of a soybean crop productivity prediction model using monthly and annual data of meteorological phenomena such as precipitation, air temperature, hydrothermal coefficient, soil moisture and others. Some of the results of this promising research are also presented.*
Keywords: *Holistic , Knowledge, Models, Prediction, Graph, Neo4j, Graph Data Science.*

1 Introduction

Biological and agricultural systems are complex and can be better understood through a holistic approach. The concept of holism suggests that the components of organisms, biological networks, and the environment are more than their sum [1]. The term holistic comes from *holos* which means everything, and the holistic approach means looking at the problem from all points of view (or from above) [2]. This approach has led to significant developments in the life sciences and environmental sustainability, particularly by studying the interactions between the components of biological systems and how they describe the functions and behavior of the systems [3,4]. Systems biology is the computational and mathematical analysis and modeling of complex biological systems, integrating many scientific fields, including biology, computer science, bioinformatics, biophysics, bioengineering, and others [5]. This discipline aims to study the complex interactions within biological systems and uses a holistic approach to predict how these systems evolve and adapt to various economic, social and environmental conditions [6]. The goal is to develop durable, sustainable and resilient solutions to global problems such as food, health and the environment.

This mode of research refers to an integrated perspective that considers the interconnections and interdependencies between the various aspects of these fields. A certain phenomenon is determined by a number of factors. Each of the factors has an impact with a certain weight on the whole phenomenon. This approach aims to develop sustainable agricultural and economic systems that take into account environmental impact, plant health and productivity, human and animal health and welfare [7]. In addition, it provides a more comprehensive and accurate understanding of biological systems and helps to develop more sustainable and efficient practices in the management of natural resources[8]. Holistic research aims to promote more efficient and ecological agricultural practices that increase the sustainability of agricultural systems, ensuring adequate food production, while reducing the negative impact of agricultural activities on the environment and society. It promotes innovative agricultural practices, green technologies and conservation of natural resources.

2. Holistic Research in Biology and Agriculture.

In scientific research there are a lot of resources available, scientific papers that explore the holistic approach in biology, agriculture and agricultural economics and continue to develop as global problems become acute, emphasizing the need for better and more sustainable solutions to the problems in these fields. Among the most relevant works can be listed:

- "Ecological resilience, biodiversity, and scale" - published by Carl Folke and his collaborators in the Annual Review of Ecology and Systematics, this article explores the concept of ecological resilience and the importance of biodiversity in the development of sustainable systems. The authors emphasize the need for a holistic approach in addressing environmental problems, to avoid fragmented solutions limited to a single problem [9].
- "Organic Agriculture and the Global Food Supply" - published by Catherine Badgley and her colleagues in Renewable Agriculture and Food Systems, this article examines the potential of organic agriculture to meet global food demand sustainably. The authors argue that organic farming can help reduce environmental impact, improve human health and support the local agricultural economy [10].
- "The Holistic Management of Rangelands: A Model for Sustainable Livestock Production in Arid Zones" - published by Allan Savory and Jody Butterfield in the Journal of Sustainable Agriculture, this article describes a holistic method of managing rangelands in arid zones. The authors argue that this approach can support biodiversity and sustainable agricultural production, as well as improve the welfare of animals and local farming communities [11].
- "Agroecology: A Transdisciplinary, Participatory and Action-oriented Approach" by Miguel Altieri and Clara Nicholls. This article explores the concept of agroecology as a holistic approach to the development of sustainable agriculture, involving collaboration between researchers, farmers and local communities [12].
- "Holistic Management: A New Framework for Decision Making" by Allan Savory. This book presents the holistic approach to land management, which encourages decision-making based on the impact on ecosystems and local communities, rather than profit maximization [13].
- "Ecological Economics: Principles and Applications" by Herman Daly and Joshua Farley. This book presents the principles of ecological economics, which integrates economics with ecology and the limitations of natural resources and the environment in the process of economic development [14].
- "The Triple Bottom Line: How Today's Best-Run Companies Are Achieving Economic, Social, and Environmental Success - and How You Can Too" by Andrew Savitz. This book explores the concept of "triple impact" (economic, social and environmental) and how companies can integrate this approach into their business strategies [15].
- "Sustainable agriculture in Romania: the potential and challenges in the context of climate change" - published by Daniela Luca and her collaborators in the Journal of Research and Social Intervention, this article analyzes the potential and challenges of sustainable agriculture in Romania in the context of climate change. The authors emphasize the need for a holistic approach, which takes into account the complex interactions between the different components of the agricultural system and integrates sustainable agricultural practices, as well as technological and social innovations [16].

In the context of climate change, the holistic approach becomes increasingly important in the development of sustainable and resilient agricultural systems. This involves adopting an integrative perspective, which takes into account the complex interactions between the different components of the agricultural system, as well as its impact on the environment and local communities.

Here are some examples of relevant scientific papers in this field:

- "Agroecology and Sustainable Agriculture for Climate Change Adaptation and Mitigation: A Review" - published by Emmanuel Torquebiau and his collaborators in Sustainability, this article explores how agroecology can be used to address the challenges of climate change. The authors emphasize that agroecology can contribute to reducing greenhouse gas emissions, improving biodiversity and protecting natural resources, as well as increasing food security and the well-being of agricultural communities [17].
- "Climate Change and Agriculture: Advancing Knowledge and Mitigation Options" - published by Cynthia Rosenzweig and her colleagues in the Annual Review of Environment and Resources, this article analyzes the impact of climate change on agriculture and identifies options for adaptation and reduction of greenhouse gas emissions greenhouse. The authors emphasize the need for a holistic approach, which takes into account the impact on soil, water and biodiversity, as well as agricultural communities and the local economy [18].
- "Climate Change Adaptation Strategies for Smallholder Farmers in the Sahel Region of West Africa: A Comprehensive Review" - published by Sifat Rehana and colleagues in Sustainability, this article examines climate change adaptation strategies for smallholder farmers in the Sahel region of West Africa. The authors emphasize that the holistic approach is essential to address the complexity of the environmental and social issues facing these farmers and to develop sustainable solutions [19].
- "Sustainable Agriculture and Climate Change: Producing Potatoes (Solanum tuberosum L.) and Bush Beans (Phaseolus vulgaris L.) for Improved Food Security and Resilience in a Canadian First Nations Community" - published by Mary Beckie and her collaborators in Sustainability, this article explores how sustainable agriculture can contribute to improving food security and the resilience of local communities in the context of climate change. The authors emphasize the need for a holistic approach, which takes into account local needs and resources, as well as the complex interactions between the different components of the agricultural system [20].
- "Climate change and agriculture in the Republic of Moldova: opportunities and challenges" - published by Dumitru Ciorici and his collaborators in the Moldovan Journal of the Environmental Science, Management and Engineering, this article analyzes the impact of climate change on agriculture in the Republic of Moldova and identifies options for adaptation and reduction of greenhouse gas emissions. The authors emphasize the need for a holistic approach, which takes into account the complex interactions between the different components of the agricultural system and integrates aspects related to soil, water, energy and local communities [21].

3. Smart technologies in agriculture.

The use of smart technologies in agriculture is a relatively recent trend. In the context of global population growth, population aging in some parts of the world and labor shortages, the introduction of new technologies and the automation of agriculture are essential to solving problems in the field of agricultural development, the volume and quality of food products. In the agricultural field, intelligent technologies, such as Precision Agriculture (Agriculture 4.0) or Smart Agriculture (Agriculture 5.0), are innovative types of agriculture based on intelligent technologies, the implementation of digital tools, the Internet of Things (IoT) and working with data that aims to make the agricultural process more efficient [22]. Through these technologies, farmers can have an increased degree of control over the entire agricultural process, which improves the activity of farmers and authorities in the field by providing adequate information for decision-making [23].

Graph databases can be used to develop machine learning models and algorithms that allow data analysis and prediction of biological and agricultural phenomena. These technologies can be used to identify complex relationships between various entities in an efficient manner. In biological research, graph systems are an industry standard for data modeling. They model different types of entities and relationships between them, using nodes and relationships between data to infer knowledge from existing relationships and to represent relationships from prior knowledge of the data [24].

4. Elaboration of the model for predicting the productivity of agricultural crops.

One of the common methods to predict the productivity of agricultural crops is the analysis of previous meteorological and agricultural data. This method is based on the fact that there is a correlation between weather conditions, such as temperature and humidity, and the productivity of agricultural crops. The models made can take into account several factors, such as soil quality, crop management, irrigation and fertilization levels, as well as weather conditions, to provide a more accurate prediction of productivity. Another method is the use of satellite images and sensor technology to monitor and evaluate agricultural crops. Researchers can use satellite images to collect information about plants, such as leaf colors and shapes, which can provide information about plant health and development. This data can then be processed through machine learning algorithms and used to estimate crop productivity.

In the field of agricultural crop productivity prediction, there are a number of relevant scientific works that have explored various methods and technologies to improve the ability to estimate agricultural production. Below are some examples of such works:

- "Agricultural yield prediction using machine learning: A review" by Chen et al. (2019) - This paper examines how machine learning technologies can be used to estimate agricultural crop productivity. The authors present a number of machine learning models that have been developed to predict the yield of cereals, vegetables and other crops, and discuss the advantages and disadvantages of these models [25].
- "Crop yield forecasting on a national scale using remote sensing data" by Thenkabail et al. (2019) - This paper describes a nationwide agricultural production prediction system that uses remote sensing data to estimate crop production. The authors present a series of data analysis algorithms that have been used to estimate the production of wheat, rice and other crops in various countries around the world [26].
- "Using machine learning techniques for crop yield prediction and climate change impact analysis" by Wang et al. (2020) - This paper presents an agricultural production prediction model based on machine learning technologies and meteorological data analysis. The authors discuss how this model can be used to estimate the production of maize and other crops, as well as to analyze the impact of climate change on agricultural production.
- "Evaluation of spectral indices for predicting grain yield of wheat using remote sensing data" by Fan et al. (2021) - This paper examines how remote sensing data can be used to estimate wheat production. The authors present a series of spectral indices that have been used to estimate wheat production in different geographic areas, and analyze the effectiveness of these indices in terms of prediction accuracy [28].

These are just a few examples of relevant work in the field of agricultural crop productivity prediction. However, researchers continue to explore and develop new methods and technologies to improve the ability to estimate agricultural production and help farmers better manage their crops. In the Romanian space, a number of researchers have contributed to the field of predicting the productivity of agricultural crops through relevant scientific works. Below are some examples of such works:

- "Study on the factors influencing the productivity of wheat varieties grown under Moldovan conditions" by M. Tcaciuc and M. Pîrțac (2016) - This paper examines how different factors, such as fertilization, phytosanitary treatments and weather conditions, influence the productivity of the varieties of wheat grown in Moldova. The authors present a series of experimental data indicating that fertilization and phytosanitary treatments can have a significant impact on wheat production [29].
- "Using Remote Sensing in Agricultural Crop Productivity Prediction" by D. Chiriac and G. Dumitrașcu (2018) - This paper examines how remote sensing data can be used to estimate agricultural crop production in Romania and Moldova. The authors present a number of examples of the application of these technologies to estimate the production of corn, wheat and other crops, and discuss the advantages and disadvantages of using these technologies [30].
- "The WOFOST model for predicting the productivity of wheat crops under Moldovan conditions" by V. Chiriac and V. Popa (2018) - This paper presents a mathematical model to estimate wheat production in Moldova, based on meteorological data and soil characteristics. The authors demonstrate that this model can be successfully used to estimate wheat production and help farmers better manage their crops [31].

A powerful tool for the holistic approach to problem solving and sustainable development can be the Neo4j graph database system. This is a graph-based database management system that is used to model and store data in the form of nodes and relationships, making it a very effective system to model and analyze complex relationships between data. This system can be used to support phenomenon prediction tasks. For example, relationships and data stored in a Neo4j database could be used as input to a machine learning model that makes predictions based on that data. In addition, Cypher, Neo4j's query language, can be used to identify patterns and trends in data, which could also be used as input to a prediction model. They allow the representation of complex relationships between different elements of a problem and their impact on the whole system.. These databases are used to represent and analyze data in a way that allows the identification and understanding of connections and interdependencies between different entities. These systems allow data to be represented in the form of graphs, which can be particularly useful for agricultural data analysis.

The use of graph database systems in the field of agricultural crop productivity prediction has been addressed by numerous researchers. Below are some examples of such works:

- "Using Graph Database for Soil Data Analysis and Agricultural Crop Production Prediction" by R. Wang et al. (2016) - This paper explores the use of a graph database to analyze soil data and predict agricultural crop production in an agricultural area in China. The authors demonstrate that the representation of data in the form of a graph can be very useful in identifying the complex relationships between factors that influence crop productivity [32].
- "Using Graph Databases for Weather Data Analysis and Agricultural Crop Production Prediction" by A. Rangwala et al. (2017) - This paper examines the use of a graph database to analyze weather data and predict agricultural crop production in an agricultural area in India. The authors present a series of examples that illustrate how graph analysis can be used to identify relationships between weather data and crop production [33].

- "Using Neural Networks and Graph Databases to Predict Corn Production in the United States" by M. Zhang et al. (2019) - This paper examines the use of neural networks and graph databases to predict corn production in the United States. The authors demonstrate that combining these technologies can lead to a significant improvement in prediction accuracy [34].
- "A Deep Learning Approach to Soybean Yield Prediction Using Field Observations and Satellite Imagery" by Ma et al. (2021). In this paper, the authors propose a new deep learning model to predict soybean crop yield using satellite imagery and field observation data [35].
- "Application of a Graph Database in Predicting Soybean Yield in Different Regions of Argentina" by H. de la Fuente et al. (2020). This paper presents a graph database model that is used to predict soybean crop yield in different regions of Argentina.
- "Graph Convolutional Networks for Soybean Yield Prediction: A Comparative Study" by Liu et al. (2020). In this paper, researchers compared the performance of graph-based convolutional neural network models to predict soybean crop yield in different regions of China [36].

All these works focus on using graph database systems to analyze data and make more accurate predictions regarding agricultural crop productivity. Graph database technology provides an efficient way to represent complex agricultural data and can be used to identify hidden patterns or relationships between variables. In the field of predicting the productivity of agricultural crops, various researches have been carried out using Neo4j graph database systems, among which the following scientific works can be distinguished:

- "Graph Database-Based Crop Yield Prediction Model Using Neo4j" by Liu et al. (2019). This paper proposes a method to build a graph database of agricultural data and use it to predict crop yield. Researchers have demonstrated the effectiveness of this approach by predicting the yield of a maize crop in Shandong Province, China [37].
- "Design of a Knowledge Graph Based on Neo4j for Crop Growth and Yield Prediction" by Wang et al. (2020). This paper presents a method to construct a knowledge graph of agricultural crop growth and use this graph to predict crop yield. Researchers have demonstrated the effectiveness of this approach by predicting rice crop yield in Jiangsu Province, China [38].
- "A Graph Database Framework for Crop Yield Prediction" by Kumar et al. (2020). This paper presents a graph database framework for agricultural data analysis and used this framework to build a crop yield prediction model. Researchers have demonstrated the effectiveness of this framework by predicting potato crop yield in different regions of India [39].

These papers demonstrate the utility and effectiveness of using Neo4j graph database technology in the field of agriculture and crop productivity prediction. Neo4j enables more efficient analysis and visualization of complex agricultural data, which can lead to greater accuracy in crop yield prediction and better management of agricultural resources.

In order to obtain high and stable yields, it is necessary to study the cultivation of this crop through the use of intelligent technologies, including the prediction of productivity results and production quality. As a result of predicting productivity according to climatic conditions, important knowledge can be gained, including:

- Understanding the interaction between climate factors and productivity: By analyzing weather and production data, we can gain a deeper understanding of how weather conditions affect soybean productivity and identify patterns and trends in this relationship.
- Optimizing agricultural practices: With a better understanding of the interaction between weather conditions and soybean productivity, we can help farmers make

better decisions about agricultural practices, such as when to plant, type and amount of fertilizers, etc., to maximize soybean productivity and reduce costs and environmental impact.

- Improved production planning and supply chain management: With more accurate prediction of soybean productivity based on weather conditions, food companies can better plan the production and storage of soy-based food products, helping them reduce waste and costs and improve supply chain management.
- Improving sustainability: By identifying optimal agricultural practices to maximize soybean productivity depending on weather conditions, we can help reduce environmental impact and increase agricultural sustainability. Also, by improving production planning and supply chain management, we can reduce waste and costs and improve the efficiency and sustainability of the food industry.

In the specialized literature there are a number of scientific works carried out in different countries that use the Neo4j graph database systems in the field of predicting the productivity of soybean crops:

- "A Comparative Study of Neo4j and PostgreSQL in Soybean Yield Prediction" by Ye et al. (2021). This paper compares the performance of Neo4j and PostgreSQL graph databases in predicting soybean crop yield in different regions of China [40].
- "Graph-based Analysis of Soybean Yield Prediction Using Neo4j" by Hu et al. (2020). In this paper, researchers used Neo4j technology to build a graph database of field observation data and satellite images to predict soybean crop yield in different regions of China [41].
- "Development of a Soybean Yield Prediction System Using Neo4j Graph Database" by Wang et al. (2020). In this paper, the authors developed a soybean crop yield prediction system using the Neo4j graph database and evaluated the performance of the system in comparison with other prediction methods [42].

In the Romanian space, the use of these technologies is not reflected.

All of these works use Neo4j technology to build complex graph databases and make more accurate predictions of soybean crop productivity. Neo4j is a scalable and efficient graph database that can be used to analyze and visualize agricultural data. Agricultural research is constantly developing, and the use of graph database systems can be an interesting solution to analyze data and make more accurate predictions. Each research analyzes various aspects of certain problems specific to the area, the climatic conditions or the population in that area. Certainly, there are many research opportunities in this area. This is an example of the research potential in this area, where technology can be used to gain a better understanding of agricultural processes and make more accurate predictions of crop productivity.

The purpose of a model for predicting the productivity of agricultural crops, especially soy, is to assess the risks and to make the necessary decisions, to estimate the future production of the crop, to minimize the impact of climate change on the environment and human communities, to manage natural resources, as well as to develop sustainable economy.

The objectives of the productivity prediction model include:

1. Forecasting future production: One of the main objectives of an agricultural crop productivity prediction model is to estimate future crop production based on weather conditions and other variables.
2. Identifying influencing factors: The model can be used to identify influencing factors affecting soybean crop productivity and determine which of these factors have the greatest influence on production based on factors such as weather conditions, soil, type and amount of agricultural inputs (eg fertilisers, biostimulants, herbicides and pesticides) as well as other key variables

such as moisture level, temperature, light or pest infestation level.

3. Optimizing agricultural practices: The model can be used to identify best agricultural practices and optimize the use of agricultural inputs, as well as assist in real-time decision making during the cropping season.
4. Improved profitability: By providing accurate estimates of future production, the model can help increase profitability by reducing costs and increasing production.
5. Reducing risk and increasing resilience: The model can help reduce the risk associated with agricultural crops by identifying conditions that can affect productivity and providing timely information to enable farmers to take preventive measures.

Database development

To fulfill the objectives, a graph database was developed in Neo4j DBMS, Cypher language. Soybean productivity prediction using intelligent data analysis databases and graph models can be achieved through the following steps:

1. Collection and cleaning of relevant agricultural production data such as production history, meteorological data, soil type, use of plant products: fertilizers, pesticides, biostimulants as well as other important information.
2. Creating the database. Using the Cypher query language, nodes grouped according to data types (labels) are created and properties are added for the relevant variables;
3. Creating relationships between nodes to reflect interactions between factors that influence productivity to enable quick and easy access to needed information. For example, we can create a relationship between a node representing a rainfall amount and a node representing a biostimulator to show how a particular practice affects productivity.

Nodes were created to process data related to the amount of monthly and annual precipitation, monthly and annual average temperatures, monthly and annual hydrothermal coefficient, monthly and annual average wind speed, air quality (release of harmful substances into the atmosphere from factories and vehicles), the productivity in quintals per hectare (q/ha) of some species of agricultural plants (soy). Data from the period 2002 – 2022 published by the National Bureau of Statistics of the Republic of Moldova[43] are included.

The necessary nodes were created that contain the data related to the monthly and annual amount of precipitation, temperature, hydrothermal coefficient - *htc*), soybean productivity for the studied period. The data presented are from the central region of the Republic of Moldova.

```
MERGE(jan2rC:MYRC{name:'Jan 2002 Rain-
fall Center', Rainfall:17}),
(feb4rC:MYRC{name:'Feb 2004 Rainfall
Center', Rainfall:38}),
(may15rC:MYRC{name:'May 2015 Rainfall
Center', Rainfall:15});
MERGE (mar3tC:MYTC{name:'Mar 2003 Tem-
peratue Center', temp:7.2}),
(apr10tC:MYTC{name:'Apr 2010 Temper-
atue Center', temp:11});
MERGE(apr2chtC:MYCHTC{name:'Apr 2002
HTC Center', htc:1.0, drought:0}),
(jul6chtC:MYCHTC{name:'Jul 2006 HTC
Center', htc:0.8, drought:1});
MERGE(r2c:YRC{name:'Rainfall 2002 Cen-
ter', rainfall:604}),
(r6c:YRC{name:'Rainfall 2006 Center',
rainfall:564});
MERGE (t5c:YTC{name:'Temperatue 2005
Center', temp:10.5}),
(t8c:YTC{name:'Temperatue 2008 Cen-
ter', temp:11.3});
MERGE (y5C:YCHTC{name:CHT 2005 Cen-
ter', cht:1, drought:0}),
(y9C:YCHTC{name:CHT 2009Center',
cht:0.8, drought:2});
```

Nodes were made regarding soybean productivity for the studied period:

```
MERGE(soy05:Productivity{name:'Soy
Prod 2005', harvest:18}),
(soy11:Productivity{name:'Soy Prod
2011', harvest:14});
```

A fragment from the graph database is represented in figure 1.

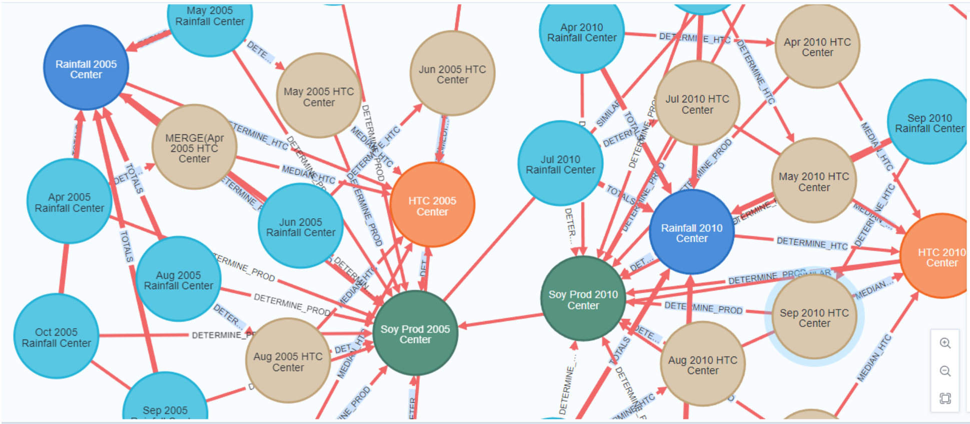

Fig. 1. Graph database.

Next, the relationships between different types of nodes were determined: Amount of precipitation → Hydrothermal coefficient → annual temperatures → Productivity, etc. Some instructions for making relationships are given below:

```
MATCH ( r ), ( c ) WHERE Id( r )= 652
and id(c)= 611
MERGE ( r ) - [rel: DETERMINE_HTC]->(
c );
MATCH ( t ), ( c ) WHERE Id( t )= 672
and id(c)= 611
MERGE (t ) - [rel: DETERMINE_HTC]->( c
);
MATCH ( c ), ( p ) WHERE Id( c )= 611
and id(p)= 631
MERGE (c ) - [rel: DETERMINE_PROD] ->
( p );
```

Development of the prediction model

Link prediction is a task in graph theory and machine learning where the goal is to predict missing links or relationships between nodes in a graph. In the context of drought prediction, neo4j can be used to represent the relationships between various factors that contribute to drought, such as temperature, precipitation, soil moisture, hydrothermal coefficient, etc. There are several machine learning algorithms that can be used to create a prediction model in Neo4j. These algorithms can be implemented in Neo4j using the APOC (Awesome Procedures On Cypher) and GDS (Graph Data Science) libraries, which provide a variety of machine learning algorithms and graphs. Using data analysis algorithms, such as regression or machine learning algorithms, to analyze relationships between data and determine which ones are most important for predicting drought. To make a prediction model it is important to choose the right algorithm and test different methods to see which one is more accurate and efficient. Modeling a predictive machine learning graph proceeds in several steps:

1. Graph preprocessing refers to the process of preparing and processing raw data to make it easier to analyze and use in machine learning algorithms such as cleaning and transforming a graph dataset before it is used. The purpose of preprocessing is to provide good features for the learning algorithm.

2. Data analysis: Once the data has been modelled, mathematical models, graph models, and data analysis algorithms from the GDS (Graph Data Science) library can be developed to analyze the data and identify relevant trends, patterns, or relationships. These models can be used to identify factors affecting crop and livestock productivity and make predictions about future production. For example, one can use clustering algorithms to group areas with similar weather conditions and use regression algorithms to identify the factors that have the greatest impact on productivity.

3. Making predictions: Based on validated models, predictions can be made about the future production of agricultural crop productivity in our graph database. For example, we may use Machine Learning (ML) algorithms. Complex graph analysis through machine learning methods can be achieved by creating graph processing pipelines (*Graph Processing Pipelines*) which represent a set of processes and technologies used to perform analysis and processing operations on graphs, i.e. on data models that consist of interconnected nodes and relationships. In the GDS library in the Neo4j DBMS, pipelines provide an end-to-end workflow from feature extraction to training and applying machine learning models. These pipelines are commonly used in data analytics and machine learning to perform a number of operations such as:

- Data extraction from nodes and relationships
- Graph metrics and properties such as graph centrality and node degree computation
- Graph analysis and modeling such as clustering and community detection
- Machine learning models training and testing to perform classifications and predictions on graphs.

Creating a prediction pipeline:

The first step in developing a crop estimation model is to design an in-memory graph, which can later be used in algorithms and graph queries.

The following statement will project a graph using a native projection using the function "*gds.graph.project* ", which can later be used in graph algorithms and queries, and store it in the graph catalogue under the name "*predProd* ", using the tags (types of nodes): "*:YHTCc*" (the nodes containing the annual hydrothermal coefficient - *htc*, with the properties "htc" and "drought", based on which the production quantity can be determined), ": *Prod"* (with the properties " *harvest* " and " *class* ", for training) and ": *ProdP"* (with the property "*harvest*", for prediction - the average harvest expressed in quintals per hectare - q/ha). The result is shown in figure 2.

```
CALL gds . graph . project ( 'prod-
Pred' ,
  { YHTCc :{ properties :[ 'htc' ,
'drought' ]},
Prod :{ properties : 'harvest' },
ProdP :{ properties : 'harvest' }
  },
  { DETERMINE_PROD : {
          type : 'DETERMINE_PROD' ,
orientation : 'UNDIRECTED'
      } } );
```

nodeProjection	relationshipProjection	graphName	nodeCount	relationshipCount
{Prod: {label: "Prod", properties: {harvest: {defaultValue: null, property: "harvest"}}}, ProdP: {label: "ProdP", properties: {harvest: {defaultValue: null, property: "harvest"}}}, YHTCc: {label: "YHTCc", properties: {drought: {defaultValue: null, property: "drought"}, htc: {defaultValue: null, property: "htc"}}}}	{DETERMINE_PROD: {orientation: "UNDIRECTED", indexInverse: false, aggregation: "DEFAULT", type: "DETERMINE_PROD", properties: {}}}	"prodPred"	37	34

Fig. 1. The projected graph.

In order to be able to analyze the complex interactions between the different elements and variables that can affect the appearance and evolution of the studied phenomena, the incorporation of nodes into a model (embedding) is used as a pre-processing step. Nodes represent the interconnected objects or entities in a given context, thus reflecting the relationships and interdependencies between them. Node embedding algorithms compute reduced-dimensional vector representations of the nodes in a graph. These vectors, also called embeddings, can be used for machine learning and are typically used as input for machine learning tasks.

```
CALL gds.fastRP.write('prodPred',{ em-
beddingDimension: 20,
writeProperty: 'fastrp-embedding'})
YIELD nodePropertiesWritten;
```

A common technique used in machine learning and data analysis is to determine similar nodes as part of preprocessing using *kNN* (*k-Nearest Neighbors*). In this technique, the *kNN algorithm* is used to find the *k* nearest neighbours of a given node in a data set based on a similarity metric. Once the ***k*** nearest neighbors are identified, they can be used to group similar nodes together for clustering or classification tasks, or they could be used to identify outliers or anomalies in the data.

```
CALL gds.knn.write('dtPred', {
writeRelationshipType: 'SIMILAR',
writeProperty: 'score', topK: 1, simi-
larityCutoff: 0.8,
randomSeed: 42, concurrency: 1,
nodeProperties: ['harvest']})
YIELD nodesCompared, relationships-
Written
```

The *gds.knn.write* function returns the number of nodes compared and the number of relationships created in the graph, which are saved in the nodesCompared and relationshipsWritten variables, respectively.

The *gds.knn.write* function receives the following arguments:

- '*prodPred* ': the name of the graph in which the similarity relations are to be created
- *writeRelationshipType* : the type of relationship that will be created between similar nodes (in this case ' *SIMILAR* ')
- *writeProperty* : the property to be added to the relation (in this case 'score'). The similarity score is a numerical value that measures the degree of similarity between two nodes. In machine learning and information retrieval, the similarity score is used to rank search results based on their relevance to a query. This is often done by comparing the query to each entity in a database and calculating a similarity score for each. The entities with the highest similarity scores are then returned as the top search results.
- *topK* : the maximum number of similar nodes that will be connected by the 'SIMILAR' relationship for each node (in this case, 1)
- *similarityCutoff* : the minimum similarity threshold that must be reached for two nodes to be considered similar and therefore connected by the 'SIMILAR' relationship (in this case, 0.5)
- *randomSeed* : the value used to initialize the random number generator, to ensure the reproducibility of the experiment
- *concurrency* : the maximum number of threads that can be used simultaneously to process data
- *nodeProperties* : list of node properties to consider in the similarity calculation.

a) Creating a pipeline named " *pipe* " for training:

```
CALL gds.beta.pipeline.nodeClassifica-
tion.create('pipe')
```

b) Pipeline configuration. Adds a node property step to the pipeline. Here the input graph contains a *harvest node property* :

```
CALL gds . beta . pipeline . nodeClas-
sification . addNodeProperty (
'pipe' , 'alpha.scaleProperties' , {
nodeProperties : [ 'harvest' , 'class'
],
scaler : 'L1Norm' , mutateProperty :
'scaledSizes' })
YIELDS name , nodePropertySteps ;
```

c) Selecting Pipe Features and Pipe Splitting:

```
CALL gds.beta.pipeline.nodeClassifica-
tion.selectFeatures(
'pipes', ['scaledSizes', 'harvest',
'class'])
YIELDS name, featureProperties;
```

Training the pipeline

The training data is used to train the prediction model. This can be achieved by using different machine learning techniques such as linear regression, neural networks or classification algorithms. The resulting table is represented in figure 3.

```
CALL
gds.beta.pipeline.nodeClassification.t
rain('prodPred', {
```

```
pipeline: 'pipe', targetNodeLabels:
['Prod'],
modelName: 'nc-pipeline-model',
targetProperty: 'harvest',
randomSeed: 1337, metrics: ['ACCURACY',
'OUT_OF_BAG_ERROR']
}) YIELD modelInfo, modelSelectionStats
RETURN
modelInfo.bestParameters AS
winningModel,
modelInfo.metrics.ACCURACY.train.avg
AS avgTrainScore,
modelInfo.metrics.ACCURACY.outerTrain
AS outerTrainScore,
modelInfo.metrics.ACCURACY.test AS
testScore,
[cand IN
modelSelectionStats.modelCandidates|ca
nd.metrics.ACCURACY.validation.avg] AS
validationScores;
```

A fragment of the resulting of pipeline training is shown in figure 3.

winningModel	avgTrainScore	outerTrainScore	testScore	validationScores
{maxEpochs: 100, minEpochs: 1, classWeights: [], penalty: 0.0, patience: 1, methodName: "LogisticRegression", focusWeight: 0.0, batchSize: 100, tolerance: 0.001, learningRate: 0.001}	0.6927272739999999	0.15384616	0.0	[0.700000002, 0.16666666800000002, 0.0, 0.700000002, 0.700000002]

Fig. 2. The result of pipeline training.

Model Validation: After the model has been trained, it must be validated to ensure that it is working correctly. This is done by using the test data set to evaluate the performance of the model. For example, you can use metrics such as accuracy, sensitivity, and specificity to evaluate model performance.

Testing the prediction model using historical data and adjusting it if necessary to improve its accuracy. Result prediction is representing in the figure 4.

```
CALL
gds.beta.pipeline.nodeClassification.p
redict.stream('prodPred', {
modelName: 'nc-pipeline-model',
includePredictedProbabilities: true,
targetNodeLabels: ['ProdP']
})
YIELD nodeId, predictedClass,
predictedProbabilities
WITH gds.util.asNode(nodeId) AS
harvestNode, predictedClass,
predictedProbabilities
RETURN harvestNode.name AS
classifiedYear, predictedClass,
floor(predictedProbabilities[predicted
Class] * 100) AS confidence
ORDER BY classifiedYear
```

classifiedYear	predictedClass	confidence
"Soy Prod 2019 Center"	14	95
"Soy Prod 2020 Center"	14	96
"Soy Prod 2021 Center"	20	95

Fig. 3. Prediction result.

Using the model: After the model has been validated, it can be used to make predictions for new data or to explore data in the database.

5. Conclusions

The holistic approach to crop cultivation under climate change involves the use of a set of agricultural measures and practices that improve crop performance, protect soil and water, and ensure sustainable and quality production. These measures must be adapted to the specific needs and conditions of each agricultural holding, taking into account the impact of different factors and the weight of each one in the final result. The use of intelligent technologies can significantly contribute to solving structured, unstructured and poorly structured problems in agriculture under the conditions of climate change. These technologies can improve the efficiency and accuracy of decision-making, help identify the best agricultural practices, and increase the productivity and sustainability of agriculture. In addition, the initial costs of implementing new technologies can be high, which can make their adoption unaffordable for small and medium-sized farmers. Also, the use of these technologies may require continuous investment over time, which may be difficult for farmers facing limited returns. Efforts to adapt to new smart technologies can help improve agricultural performance, increase profitability and ensure the sustainability of the agricultural sector.

Prediction is an ever-evolving field, and researchers continue to explore and develop new methods and technologies to improve crop productivity prediction. These methods include the analysis of past weather and agricultural data, the use of satellite imagery and sensor technology, as well as data analytics and artificial intelligence technologies. The use of intelligent data analysis databases and graph models can be a valuable tool to help farmers and agricultural entrepreneurs optimize their agricultural production and make informed and sustainable decisions. The created model uses the data from the period 2002 – 2018. For testing, the data from the years 2019 – 2021 was checked. These predictions have mostly come true.

The created model can also be used for various agricultural crops, but also for predicting the evolution of climatic factors and increasing resilience to various natural hazards.

Acknowledgments

I would like to thank the Associate professor, PhD. Ana Bârsan, Head of the Department of Biology and Ecology, Faculty of Biology and Geoscience, Moldova State University for the data obtained during the experiments and the valuable consultations in the field of plant physiology, and professor, Gheorghe Capatână for methodological assistance related to research and development of the intelligent support system.

Notes: The work was carried out as part of the Doctoral Project "Models, Techniques and Program Products for Intelligent Data Analysis in Plant Physiology". Scientific advisor: PhD. Eng. Gheorghe Capatână, MSU.

References

[1] JC Smuts. Holism and Evolution. Macmillan, 1926.
[2] Scribd https://ro.scribd.com/document/351967625/Abordarea-Holistica#. (accessed on 2023-03-01)
[3] A. Savory, "Holistic Management: A New Framework for Decision Making," Island Press, 1999.
[4] DEX https://dexonline.ro/definitie/holism. (accessed on 2023-03-04)
[5] JJ Tyner and RE Michod, "Exploring the Complexity of Biological Systems," BioScience, vol. 53, no. 10, pp. 976-981, Oct. 2003.
[6] H. Kitano, "Systems Biology: A Brief Overview," Science, vol. 295, no. 5560, pp. 1662-1664, Mar. 2002.
[7] H. Daly and J. Farley, "Ecological Economics: Principles and Applications," Island Press, 2004.
[8] HMI. Christine Jones. The Back Forty Down Under: Adapting Farming to Climate Variability. Avilable: https://holisticmanagement.org/wp-content/uploads/2016/08/Jones-AdaptingFarming-Quivira-Feb10.pdf. (accessed on 2023-02-24)
[9] C. Folke et al., "Ecological resilience, biodiversity, and scale," Annual Review of Ecology and Systematics, vol. 31, pp. 277-308, 2000.
[10] Badgley et al., "Organic Agriculture and the Global Food Supply," Cambridge University Press, 2007.
[11] Savory and J. Butterfield, "The Holistic Management of Rangelands: A Model for Sustainable Livestock Production in Arid Zones," Journal of Sustainable Agriculture, vol. 4, no. 2-3, pp. 141-154, 1994.
[12] M. Altieri and C. Nicholls, "Agroecology: A Transdisciplinary, Participatory and Action-oriented Approach," CRC Press, 2017.
[14] A. Savory, "Holistic Management: A New Framework for Decision Making," Island Press, 2016.

[13] H. Daly and J. Farley, "Ecological Economics: Principles and Applications," Island Press, 2010.

[14] A. Savitz, "The Triple Bottom Line: How Today's Best-Run Companies Are Achieving Economic, Social, and Environmental Success - and How You Can Too," Jossey-Bass, 2006.

[15] D. Luca et al., "Sustainable agriculture in Romania: potential and challenges in the context of climate change," Revista de Research si Interventie Socială, vol. 60, pp. 7-21, 2018.

[16] E. Torquebiau et al., "Agroecology and Sustainable Agriculture for Climate Change Adaptation and Mitigation: A Review," Sustainability, vol. 8, no. 6, pp. 1-23, 2016.

[17] C. Rosenzweig et al., "Climate Change and Agriculture: Advancing Knowledge and Mitigation Options," Annual Review of Environment and Resources, vol. 37, pp. 1-32, 2012.

[18] S. Rehana et al., "Climate Change Adaptation Strategies for Smallholder Farmers in the Sahel Region of West Africa: A Comprehensive Review," Sustainability, vol. 11, no. 13, pp. 1-31, 2019.

[19] M. Beckie et al., "Sustainable Agriculture and Climate Change: Producing Potatoes (Solanum tuberosum L.) and Bush Beans (Phaseolus vulgaris L.) for Improved Food Security and Resilience in a Canadian First Nations Community," Sustainability, vol. 7 , no. 9, pp. 1-18, 2015.

[20] D. Ciorici et al., "Climate change and agriculture in the Republic of Moldova: opportunities and challenges", Moldovan Journal of the Environmental Science, vol. 12, no. 2, pp. 50-60, 2020.

[21] E. Costa, M. Martins, E. Vendruscolo, A. Silva, T. Zoz, F. Binotti, T. Witt, and C. Seron, "Greenhouses within the Agriculture 4.0 interface," Rev. Ciencia Agronomica, vol. 51, article e20207703, 2021. DOI: 10.5935/1806-6690.20200089.

[22] Jin, B. Chen, BD Lampinen, and PH Brown, "Advancing Agricultural Production with Machineteaming Analytics: 'r'ield Determinants for California is Almond Orchards," in Proc. IEEE Int. Conf. on Agricultural and Biological Systems Engineering (ABSE), 202X, pp. 1-6.

[23] Neo4j. Graph Data Modeling. https://neo4j.com/graphacademy/edu program/. (accessed on 2023-03-30)

[24] S. Chen, Y. Chen, X. Yang, and J. Zhang, "Agricultural yield prediction using machine learning: A review," Computers and Electronics in Agriculture, vol. 163, 2019, p. 104859. https:// www.sciencedirect.com/science/article/pii/S0168169918306782

[25] PS Thenkabail, MK Gumma, P. Teluguntla, A. Oliphant, J. Xiong, C. Giri, V. Dheeravath. Crop yield forecasting on a national scale using remote sensing data. Remote Sensing, 11(23), 2019, p. 2821.

[26] B. Wang, X. Tao, H. Wang, J. Liu, Y. Zhang. Using machine learning techniques for crop yield prediction and climate change impact analysis. Environmental Research, 188, 2020, 109825.

[27] X. Fan, Y. Shen, H. Zhang, L. Sun. Evaluation of spectral indices for predicting grain yield of wheat using remote sensing data. International Journal of Remote Sensing, 42(8). 2021, 2742-2759.

[28] M. Tcaciuc, M. Pîrțac. Study on the factors that influence the productivity of wheat varieties grown in the conditions of Moldova. Bulletin of the Academy of Sciences of Moldova. Mathematics, 1(76), 3-12, 2016.

[29] D. Chiriac, G. Dumitrașcu. The use of remote sensing in the prediction of agricultural crop productivity. Studia Universitatis Moldaviae. Life Sciences Series, 11(1), 2018, 89-93.

[30] V. Chiriac and V. Popa, "The WOFOST model for predicting the productivity of wheat crops in the conditions of Moldova," Rev. Agric. Econ., vol. 68(2), pp. 51-58, 2021.

[31] R. Wang et al., "Using graph database for soil data analysis and crop yield prediction," in Proceedings of the

2016 7th International Conference on Intelligent Systems, Modeling and Simulation (ISMS), Bangkok, Thailand, 2016, pp. 42-46.

[32] A. Rangwala et al., "Using graph databases for weather data analysis and crop yield prediction," in Proceedings of the 2017 8th International Conference on Intelligent Systems, Modeling and Simulation (ISMS), Kuala Lumpur, Malaysia, 2017, pp. 82-86.

[33] M. Zhang et al., "Using neural networks and graph databases to predict corn production in the United States," Rev. Agric. Econ., vol. 66(2), pp. 35-41, 2019.

[34] W. Ma et al., "A Deep Learning Approach to Soybean Yield Prediction Using Field Observations and Satellite Imagery," Remote Sens., vol. 13, no. 2, pp. 354, Jan. 2021.

[35] H. de la Fuente et al., "Application of a Graph Database in Predicting Soybean Yield in Different Regions of Argentina," Sensors, vol. 20, no. 22, p. 6543, 2020.

[36] H. Liu et al., "Graph Convolutional Networks for Soybean Yield Prediction: A Comparative Study," IEEE Access, vol. 8, pp. 81794-81805, 2020.

[37] R. Wang et al., "Using graph database for soil data analysis and crop yield prediction," Proceedings of the International Conference on Computers Communications and Control (ICCCC), pp. 527-532, 2016.

[38] R. Kumar, S. Bhandary, A. Patil, and M. Sonawane, "A Graph Database Framework for Crop Yield Prediction," in 2020 IEEE 10th International Conference on Intelligent Systems (IS), 2020, pp. 516-521. https://ieeexplore.ieee.org/document/9263211.

[39] Y. Ye, J. Yang, X. Zhang, J. Wang, and Y. Gao, "A Comparative Study of Neo4j and PostgreSQL in Soybean Yield Prediction," in 2021 International Conference on Artificial Intelligence and Big Data Engineering (ICAIBE), 2021, pp. 260-265. https://ieeexplore.ieee.org/document/9441644 Y. Hu, S. Wang, X. Zhu, and C. Li, "Graph-based Analysis of Soybean

[40] Yield Prediction Using Neo4j," in 2020 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2020, pp. 449-455. https://ieeexplore.ieee.org/document/9259055. (accessed on 2023-03-24)

[41] Y. Wang, X. Zhao, and G. Zhang, "Development of a Soybean Yield Prediction System Using Neo4j Graph Database," in 2020 4th International Conference on Intelligent Information Processing (ICIIP), 2020, pp. 1-5. https://ieeexplore.ieee.org/document/9277286

[42] Moldova Statistical Data Bank. https://statbank.statistica.md/PxWeb/pxweb/ro/? rxid= 2f5a0079-c3c1-461e-81a4-d854b7ba0fcd. (accessed on 2023-02-24)

[43] State Hydrometeorological Service. http://www.meteo.md/index.php/clima/terminologie-si-unitati-de-masura/glosar-de-termeni/c/ (accessed on 2023-02-25).

Ion GANEA, master in Exact Sciences (Databases & Knowledge). PhD student at Doctoral School of Physical, Mathematical, Information and Engineering Sciences, Moldova State University.
Teacher of Information Technology at the Center of Excellence in Economics and Finance, Computer Science Department.
Courses taught: Databases; Computer Networks; Operating Systems.
Areas of scientific interest: Artificial Intelligence, Intelligent Systems, Graph Databases, Intelligent Data Analysis.