Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

DOI: 10.37943/17LKYF9288

Aigul Mimenbayeva
Master of Sciences, Senior Lecturer, Department of Computational and Data Science
aigulka79_79@mail.ru, orcid.org/0000-0003-4652-470X
Astana IT University, Kazakhstan

Gulnur Issakova
PhD, Senior Lecturer of the Department of Information Systems
is_gul_oral@mail.ru, orcid.org/0000-0001-7272-4786
S.Seifullin Kazakh Agro Technical Research University, Kazakhstan

Balausa Tanykpayeva
Master of Natural Sciences, Senior Lecturer of the Department of Information Systems
balausa1.80@mail.ru, orcid.org/0009-0001-1259-0832
S.Seifullin Kazakh Agro Technical Research University, Kazakhstan

Ainur Tursumbayeva
Master of Technical Sciences, Teacher of the Department of Information Systems
Turcumbaewa_ainur84@mail.ru, orcid.org/0009-0000-3710-3925
S.Seifullin Kazakh Agro Technical Research University, Kazakhstan

Raya Suleimenova
Candidate of Technical Sciences, Acting Professor of School of Engineering and Information Technology
Suleimenova_raya@mail.ru, orcid.org/0009-0004-2780-5391
Eurasian Technological University, Kazakhstan

Almat Tulkibaev
Master of Sciences, Teacher of the Department of Information Systems
Almat_tulkibaev@mail.ru, orcid.org/0000-0002-3783-5429
S.Seifullin Kazakh Agro Technical Research University, Kazakhstan

APPLYING MACHINE LEARNING FOR ANALYSIS AND FORECASTING OF AGRICULTURAL CROP YIELDS

Abstract: Analysis and improvement of crop productivity is one of the most important areas in precision agriculture in the world, including Kazakhstan. In the context of Kazakhstan, agriculture plays a pivotal role in the economy and sustenance of its population. Accurate forecasting of agricultural yields, therefore, becomes paramount in ensuring food security, optimizing resource utilization, and planning for adverse climatic conditions. In-depth analysis and high-quality forecasts can be achieved using machine learning tools.

This paper embarks on a critical journey to unravel the intricate relationship between weather conditions and agricultural outputs. Utilizing extensive datasets covering a period from 1990 to 2023, the project aims to deploy advanced data analytics and machine learning techniques to enhance the accuracy and predictability of agricultural yield forecasts. At the heart of this endeavor lies the challenge of integrating and analyzing two distinct types of datasets: historical agricultural yield data and detailed daily weather records of North Kazakhstan for 1990-2023. The intricate task involves not only understanding the patterns within each dataset but also deciphering the complex interactions between them. Our primary objective is to develop models that can accurately predict crop yields based on various weather parameters, a crucial aspect for effective agricultural planning and resource allocation. Using

DOI: 10.37943/17LKYF9288

the capabilities of statistical and mathematical analysis in machine learning, a Time series analysis of the main weather factors supposedly affecting crop yields was carried out and a correlation matrix between the factors and crops was demonstrated and analyzed.

The study evaluated regression metrics such as Root Mean Squared Error (RMSE) and R^2 for Random Forest, Decision Tree, Support Vector Machine (SVM) algorithms. The results indicated that Random Forest generally outperformed the Decision Tree and SVM in terms of predictive accuracy for potato yield forecasting in North Kazakhstan Region. Random Forest Regressor showed the best performance with an R^2=0.97865. The RMSE values ranged from 0.25 to 0.46, indicating relatively low error rates, and the R^2 values were generally positive, indicating a good fit of the model to the data.

This paper seeks to address these needs by providing insights and predictive models that can guide farmers, policymakers, and stakeholders in making informed decisions.

Keywords: machine learning, crop yield, correlation matrix, linear regression, time series analysis

Introduction

In the face of a growing global population and the escalating demand for food, traditional agricultural practices are facing unprecedented challenges. To meet this demand while ensuring sustainable resource utilization, the adoption of innovative technologies is imperative. Machine learning (ML), a subset of artificial intelligence, has emerged as a powerful tool capable of revolutionizing agriculture [1]. ML algorithms can analyze vast amounts of data to identify patterns and make predictions, offering valuable insights for optimizing crop production and resource management. One of the most critical aspects of agricultural decision-making is crop yield prediction, which influences resource allocation, market strategies, and food security. Accurate yield predictions enable farmers to make informed decisions regarding planting schedules, irrigation requirements, and fertilizer application, leading to increased productivity and reduced environmental impact [2-4]. ML-powered crop prediction models can harness the power of data to process vast amounts of agricultural information, including historical yield records, weather patterns, soil characteristics, and satellite imagery, to identify complex relationships between these factors and crop yield. By analyzing this data, ML models can accurately predict crop yields, providing farmers with valuable decision-making tools. While ML holds immense promise for agriculture, its adoption faces challenges primarily related to data quality and availability. Acquiring and maintaining high-quality, consistent agricultural data is crucial for training and validating ML models [5].

ML algorithms are adept at processing and analyzing vast amounts of agricultural data, including historical yield records, weather patterns, soil characteristics, and satellite imagery. By extracting valuable insights from this data, ML models can identify complex relationships and patterns that influence crop yield. ML-powered crop prediction models form the foundation of precision agriculture initiatives. By providing precise predictions at the field or even sub-field level, these models enable farmers to optimize resource allocation, such as water, fertilizers, and pesticides, based on specific crop requirements, soil conditions, and environmental factors [6-7].

Literature review

The field of agricultural analysis, particularly in relation to weather conditions, has been extensively studied, with numerous researchers exploring the intersection of climatology, agriculture, and data science. The following literature review highlights key scholarly contributions that provide a foundation and context for this research.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

In [8] author's comprehensive review delves into the effects of climate change on crop yields. Their study synthesizes findings from various global research efforts, emphasizing the complex relationship between changing weather patterns and agricultural productivity. The review highlights that temperature fluctuations and altered precipitation regimes significantly affect crop growth cycles and yields.

The [9] reference explores the application of machine learning techniques in predicting agricultural yields. Authors demonstrated the efficacy of models like Random Forest and Support Vector Machines in forecasting crop productivity based on weather data. Their findings suggest a high correlation between weather variables and yield outcomes, underscoring the potential of machine learning in agricultural planning.

Reference [10] focuses on the role of data analytics in developing climate-resilient farming practices. They discuss how big data and predictive analytics can empower farmers to make informed decisions, particularly in the context of adapting to climate variability. The paper also addresses the challenges in integrating diverse data sources for effective analysis.

Addressing this aspect, reference [11] underscores the versatility of ML applications in smart farming, extending beyond crop yield prediction to encompass various facets of agriculture. Notably, the authors discuss the integration of ML in livestock management, water conservation strategies, soil health assessment, and crop management. This holistic approach emphasizes the potential of ML to address multiple challenges in agriculture, leading to more efficient and sustainable practices. The significance of accurate crop yield prediction is highlighted as a cornerstone for informed decision-making in the agricultural sector. By leveraging ML algorithms, farmers gain insights into optimal planting times, irrigation needs, and fertilizer usage, ultimately maximizing productivity while minimizing environmental impact. The authors argue that the adoption of ML crop prediction models represents a paradigm shift in agriculture, offering a data-driven approach to precision farming.

The reference [12] present case studies on technology adoption in agriculture under the challenges posed by climate change. They highlight how advancements in remote sensing and information technologies are revolutionizing farming practices. The paper argues for a more integrated approach that combines traditional knowledge with modern technology to enhance agricultural sustainability.

These studies collectively provide a comprehensive overview of the current state of research at the intersection of climatology, agriculture, and data science. They underscore the importance and potential of using advanced data analytics and machine learning techniques in understanding and predicting agricultural outcomes, which is central to our project.

The intersection of agro-technology and machine learning has gained significance due to advancements in data methodologies and high-performance computing. Machine learning classifiers have emerged as vital tools in crop prediction, playing a crucial role in agriculture. The authors of reference [13] proposes the advanced stacking ensemble learning approach for crop prediction, addressing the need for accurate predictions within a short processing time. The primary objective of their approach is to meet the demand for accurate predictions while also ensuring that the processing time remains low. Ensemble learning involves combining multiple machine learning models to improve prediction accuracy. Stacking, in particular, is a method where the predictions of several base models are used as input for a meta-model, which then produces the final prediction. By utilizing this advanced stacking ensemble learning technique, Sethy et al. aim to achieve higher accuracy in crop prediction compared to individual models while also maintaining efficiency in terms of processing time. In reference [14] discussed the rise of agro-technology and machine learning, facilitated by significant advancements in data methodologies and computing capabilities. The authors delve into the intersection of agro-technology and machine learning, highlighting the emergence of this

interdisciplinary field. They emphasize that this fusion has been made possible by notable advancements in data methodologies and computing capabilities. In their discussion, Zhai and colleagues likely explore how advancements in data methodologies, such as data collection, preprocessing, and analysis techniques, have paved the way for leveraging machine learning in agriculture. Moreover, they may address the role of improved computing capabilities, including faster processors, parallel computing architectures, and cloud computing infrastructure, in handling large agricultural datasets and executing complex machine learning algorithms efficiently. The authors of studies [15], [16], [17] underscores the increasing importance of machine learning classifiers in crop prediction, highlighting their role as a crucial component in modern agriculture. By leveraging machine learning techniques within the context of agriculture, the proposed approach aims to enhance crop prediction accuracy and efficiency, contributing to advancements in the field of agro-technology.

These studies collectively provide a comprehensive overview of the current state of research at the intersection of climatology, agriculture, and data science. They underscore the importance and potential of using advanced data analytics and machine learning techniques in understanding and predicting agricultural outcomes, which is central to our research. The references [18] and [19] utilizes decision tree analysis to identify the key factors influencing changes in crop yield and to predict crop yield outcomes. Their approaches contribute valuable insights to the field of agricultural research and helps stakeholders in optimizing agricultural practices and maximizing crop productivity. By combining random forest with wheat yield data, meteorological variables, and satellite images, [20] aimed to develop a robust and accurate predictive model for wheat yield in southeastern Australia. This approach allows for comprehensive analysis and prediction of crop yields, enabling better agricultural planning and decision-making in the region. In their studies, the authors emphasized the effectiveness of the random forest method in comparison with SVM and linear regression methods for China's main rapeseed-producing area. The authors of reference [21] also used the decision tree algorithm to evaluate crop yield by combining soil and climate data. By leveraging the ensemble nature of random forest, the authors likely achieve accurate and robust predictions of crop yield by integrating soil and climate data, thereby contributing valuable insights to agricultural research and decision support systems.

Analyzing the research of the above-mentioned authors, it can be confirmed that the forecast of agricultural yields largely depends on territorial conditions. In this regard, digitalization of precision agriculture in the North Kazakhstan region is a relevant area. While wheat is predominant, North Kazakhstan also cultivates other crops such as barley, oats, and potatoes. In North Kazakhstan, where crop productivity may be influenced by complex interactions between weather patterns, soil properties, and agricultural practices, the ability of these methods to handle nonlinear relationships is crucial for developing accurate prediction models. This interpretability is particularly valuable for agricultural stakeholders and policymakers in North Kazakhstan, as it allows them to gain insights into the factors driving yield variations and make informed decisions based on the model's outputs. Many studies focus on specific crops grown in Central Asia, such as wheat, cotton, and barley. These crop-specific analyses allow for tailored approaches to yield prediction based on the unique characteristics and requirements of each crop.

To summarize, this study selected the most effective methods of Decision tree, random forest and linear regression to analyze potato productivity in North Kazakhstan region.

Purpose and Objectives of Research

The purpose of this study is to analyze and forecast the yield of agricultural crops in the North Kazakhstan region for 1990-2023.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

The object of the study is time series of agroclimatic data in the North Kazakhstan region over the past 33 years.

To achieve the goal, the following tasks were set:

- prepare an appropriate array of data on agroclimatic data of the North Kazakhstan region for 1990-2023;
- make a linear trends of the main factors influencing crop yields;
- using machine learning algorithms to make a yield forecast for the last years.
- to perform comparative analysis for different test and training data;
- to make a conclusion based on the applied methods.

Materials and Research Methods

Dataset Description: We have multiple CSV data files, primarily using the Pandas library for easy processing. The largest file, our training data, exceeds 100 MB, but our computing power should handle it without requiring special techniques. The training file has 12346 rows, 45 columns: datetime, feelslikemin, feelslikemax, tempmax, tempmin, humidity, and more, with various data types such as numerical, categorical, and datetime values. It includes foreign keys pointing to other CSV files, like one containing metadata about the stores.

Data Preprocessing: Rigorous data cleaning procedures were executed to handle missing values, ensuring the dataset's integrity for meaningful analysis. Feature scaling techniques were applied to normalize the data, a crucial step for the effectiveness of machine learning models in time series forecasting.

Modeling and evaluation: Multivariate time series models were chosen to exploit the internal dependencies between different factors. The correlation matrix was used to identify the strengths of certain factors, thereby increasing the accuracy of prediction. In addition, we used a linear regression method on the pooled data set for univariate time series forecasting. Evaluation metrics for a Linear Regression model are chooses ccoefficient of Determination (R2).

Time Series Analysis: Time Series Analysis were used to analyze long-term source data. Time series analysis employs various statistical and mathematical models to make predictions or forecasts based on historical data. Common techniques include moving averages, autoregressive integrated moving average (ARIMA) models, exponential smoothing methods, and more advanced methods like machine learning algorithms, including recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks [22].

Decision Tree Regressor: A Decision Tree Regressor is a type of machine learning algorithm used for regression tasks. It belongs to the family of decision tree algorithms, which are commonly used for both classification and regression problems. While decision trees for classification partition the data into discrete classes, decision trees for regression predict continuous values. The algorithm starts with the entire dataset and selects the best feature to split the data based on some criterion (commonly mean squared error or variance reduction). It repeats this process recursively for each resulting subset until a stopping criterion is met, such as reaching a maximum tree depth, having too few samples in a node, or other conditions. Once the tree is constructed, to make a prediction for a new instance, it traverses down the tree from the root node to a leaf node. At each node, it follows the decision rule based on the feature value of the instance until it reaches a leaf node, which contains the predicted value for regression tasks. In decision tree regression:

- Features of the input data are used to make decisions at each node of the tree.
- The tree is recursively partitioned based on these features until some stopping criteria are met.
- At the leaf nodes of the tree, the model predicts the continuous output based on the features of the input data that lead to that leaf node.

DOI: 10.37943/17LKYF9288


```
import pandas as pd
file_path = '/content/Crop_Yield_NK +.xlsx'
df = pd.read_excel(file_path)
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
rf = DecisionTreeRegressor(n_estimators = 100, max_features = 'sqrt', max_depth = 5, random_state = 18).fit(x_train, y_train)
```

Figure 1. Implementation of the Decision Tree Regressor algorithm code in Python

In Figure 1 illustrated implementation of the Decision Tree Regressor algorithm code in Python. In the syntax of Random Forest and Decision Tree algorithms commonly used next hyperparameters [23]:

n_estimators: This parameter determines the number of decision trees that will be used in the random forest model. Increasing the number of trees generally improves the performance of the model, but it also increases computational complexity.

max_depth: This parameter sets the maximum depth of each decision tree in the random forest. A deeper tree can capture more complex patterns in the data, but it can also lead to overfitting. Setting an appropriate max_depth is crucial for balancing model complexity and generalization.

max_features: This parameter determines the maximum number of features considered for splitting a node in each decision tree. Random forests typically consider a random subset of features at each split, which helps to reduce overfitting and increase model diversity.

The random_state parameter in scikit-learn's RandomForestRegressor or RandomForestClassifier (and in many other scikit-learn models) is used to ensure reproducibility of the results. In machine learning models, randomness may be introduced during certain operations such as bootstrapping, feature sampling, or initialization of weights.

Random Forest Regression: Random Forests train a collection of decision trees, where each tree is trained on a random subset of the dataset. Random Forest is considered a meta-estimator because it aggregates the predictions of multiple base estimators (individual decision trees) to make a final prediction. It doesn't directly learn relationships from the data like traditional estimators, but rather combines the outputs of simpler models. In Random Forest, multiple decision trees are trained, each on a random subset of the dataset. These trees are typically trained independently of each other. Each decision tree in the Random Forest is trained on a different subset of the original dataset. This process, known as bootstrapping, involves sampling the dataset with replacement to create multiple subsets. Instead of relying on the prediction of a single decision tree, Random Forest combines the predictions of all the trees in the forest. For classification tasks, it uses majority voting, while for regression tasks, it typically takes the average of the predictions. By training each decision tree on a random subset of the data and considering only a random subset of features at each split, Random Forest introduces randomness into the learning process, which helps prevent overfitting. Random Forests are widely used and highly effective for a variety of machine learning tasks, including classification and regression, due to their ability to handle high-dimensional data, mitigate overfitting, and provide robust predictions [24].

```
import pandas as pd
file_path = '/content/Crop_Yield_NK +.xlsx'
df = pd.read_excel(file_path)
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor(n_estimators = 100, max_features = 'sqrt', max_depth = 5, random_state = 18).fit(x_train, y_train)
```

Figure 2. Implementation of the Random Forest Regressor algorithm code in Python

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

In Figure 2 illustrated implementation of the Random Forest Regressor algorithm code in Python.

Support Vector Regression (SVR): SVR is indeed a type of Support Vector Machine (SVM) algorithm used for regression analysis. While traditional SVMs are primarily used for classification problems, SVR extends the SVM framework to handle regression tasks. Instead of finding a hyperplane that best separates classes, SVR aims to find a hyperplane that best fits the data points within a specified margin of error.

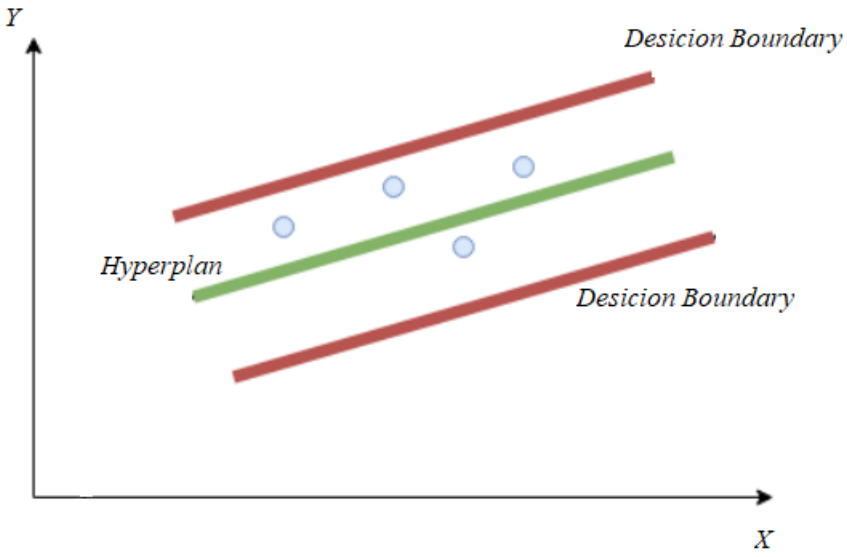

Figure 3. Graphical representation of Support Vector Regression

The hyperplane in SVMs represents the decision boundary that separates different classes in the feature space (Figure 3). It is defined by a set of parameters (weights and bias) learned during the training process. The margin is the distance between the hyperplane and the closest data points from each class. SVM aims to find the hyperplane that maximizes this margin, as it generalizes better to unseen data and improves the model's robustness [25].

Figure 4 illustrates the main steps in solving the objectives of this study.

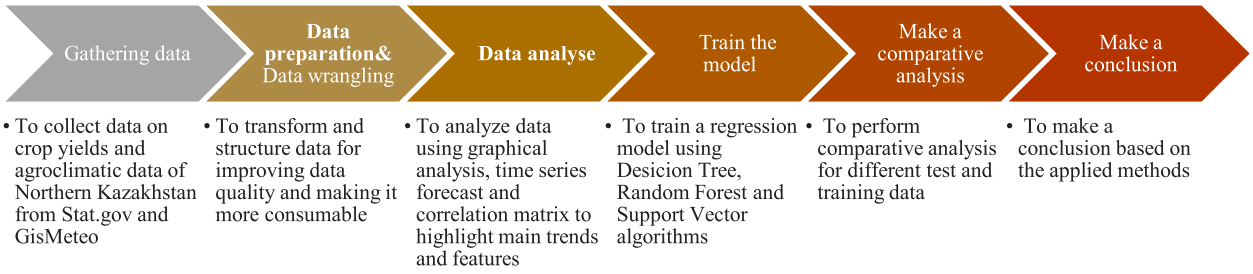

Figure 4. The main steps to solve the research

Results and Discussion

For the research, a database of agroclimatic data and data on the yield of the North Kazakhstan region for the period 1990-2023 was prepared (Fig.1). This dataset consists of 12346 rows and 45 columns that describe agro climatic conditions over the last 33 years in the North Kazakhstan region. Time Series Analysis of the main factors influencing crop yields was carried out.

DOI: 10.37943/17LKYF9288

Year	Grains	Oilseeds	Seeds sunflower	Potato	Open_veg	Melons	Sugar beet	temp_mean	humidity_mean	precip_mean
1990	12,2	3,9	9,2	113,0	154,0	84,0	239,0	3.706575	68.501370	0.683562
1991	5,3	5,7	4,9	99,0	121,0	79,0	148,0	4.324586	63.30828	0.380110
1992	13,2	6,3	3,3	104,0	114,0	72,0	136,0	2.805191	70.669126	2.001366
1993	9,7	7,1	3,2	94,0	106,0	69,0	123,0	1.781319	69.978846	1.114835
1994	7,9	6,2	3,4	94,0	104,0	59,0	77,0	3.193699	68.243836	0.909863
1995	5,0	7,0	2,9	84,0	101,0	59,0	91,0	4.450552	70.887158	0.330663
1996	6,5	6,6	1,9	88,0	96,0	58,0	105,0	1.518579	62.765205	0.932240
1997	8,7	5,5	2,8	84,0	101,0	67,0	116,0	4.652329	65.409091	0.260000
1998	5,6	6,5	4,2	77,0	114,0	78,0	143,0	0.414773	71.531579	0.098295
1999	13,0	4,9	4,9	108,0	134,0	97,0	172,0	4.788623	69.146027	0.662534
2000	9,4	3,9	4,0	106,0	153,0	119,0	154,0	3.966082	69.283288	1.176901
2001	12,2	5,7	6,0	133,0	166,0	127,0	173,0	4.081370	69.143836	1.429863
2002	11,5	6,3	5,9	139,0	172,0	135,0	207,0	4.790411	69.940984	1.267671
2003	10,8	7,1	6,8	139,0	177,0	144,5	210,4	3.713151	70.309041	1.546575
2004	8,8	6,2	5,9	134,0	186,0	153,2	197,4	3.701644	67.959178	0.930055
2005	10,0	7,0	6,3	150,0	196,0	159,3	209,2	3.909863	70.357534	0.810959
2006	11,7	6,6	5,9	153,6	201,0	167,1	240,8	3.401093	70.660109	0.930137
2007	13,3	7,2	5,9	155,8	211,0	171,7	248,9	2.531233	71.796164	0.826849
2008	10,1	5,5	4,1	143,7	204,0	158,9	204,3	2.418082	66.826849	0.000000
2009	12,6	6,5	5,7	160,0	218,7	161,1	182,9	2.012055	71.087397	0.000000
2010	8,0	5,0	4,4	143,0	214,4	177,0	174,3	4.163288	65.545355	0.000822
2011	16,9	6,7	4,6	167,2	222,9	186,1	188,2	2.491233	72.284110	0.009041
2012	8,6	6,1	5,9	165,9	234,0	206,8	168,2	4.163288	71.087397	0.818306
2013	11,6	8,0	7,0	181,5	238,7	212,4	267,7	2.543562	72.284110	1.348219
2014	11,7	7,8	6,7	184,3	243,0	217,1	240,6	3.951233	67.333425	0.955616
2015	12,7	8,1	7,6	185,5	245,8	221,0	232,5	3.951233	69.397260	1.098904
2016	13,5	9,6	9,3	190,4	250,0	221,4	285,5	3.712022	69.771858	1.141530
2017	13,4	9,7	10,2	194,2	253,7	224,2	274,4	4.240274	67.97941918	0.704932
2018	13,5	9,7	10,0	197,9	257,3	224,2	305,3	1.589041	73.291507	1.174351
2019	12,3	9,3	10,3	203,4	260,5	234,6	324,5	3.877260	69.643014	0.920219
2020	12,8	9,5	11,3	206,7	265,9	238,8	323,2	4.679508	68.630874	1.248445
2021	10,4	8,3	11,0	207,4	268,0	252,7	275,5	3.569589	65.909041	0.933945
2022	13,8	9,1	12,0	205,4	271,3	255,6	341,4	3.978630	64.570685	0.714082
2023	14,8	9,7	12,0	206.7	282,2	252,5	345,4	4.178630	63.570685	0.794082

Figure 1. Agroclimatic and crop yield data for the North Kazakhstan region for 1990-2023

Moreover, it is generated line plots for mean, minimum, and maximum humidity over the years (Fig. 2).

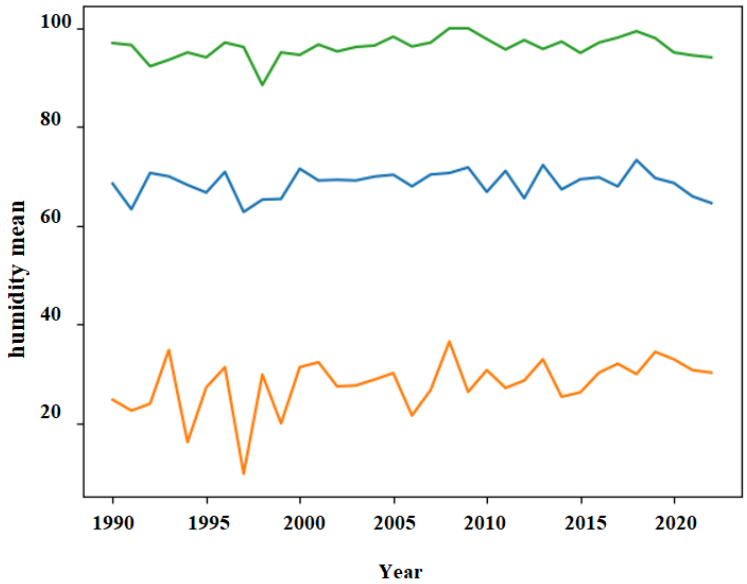

Figure 2. Time Series of humidity of North Kazakhstan region for 1990-2023

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

The top line (green) represents the maximum humidity for each year, showing less fluctuation and maintaining high values throughout the period. The middle line (blue) indicates the mean or average annual humidity, which shows a slight downward trend with some year-to-year variation. The bottom line (orange) shows the minimum annual humidity, which displays more pronounced fluctuations compared to the mean and max lines. The x-axis represents the years, while the y-axis represents the humidity level, which, although not explicitly labeled, is likely in percentage given typical humidity measurements. From the plot, we can observe that while the maximum humidity remains relatively stable, the mean and minimum humidity levels exhibit more variability. Notably, the minimum humidity shows several dips, which could correspond to particularly dry periods or years. This visualization helps in understanding the overall humidity trends and can be particularly useful when analyzing the impact of varying humidity levels on agricultural productivity, climate patterns, or other environmental factors.

Furthermore, we created line plots for mean, minimum, and maximum temperature across years.

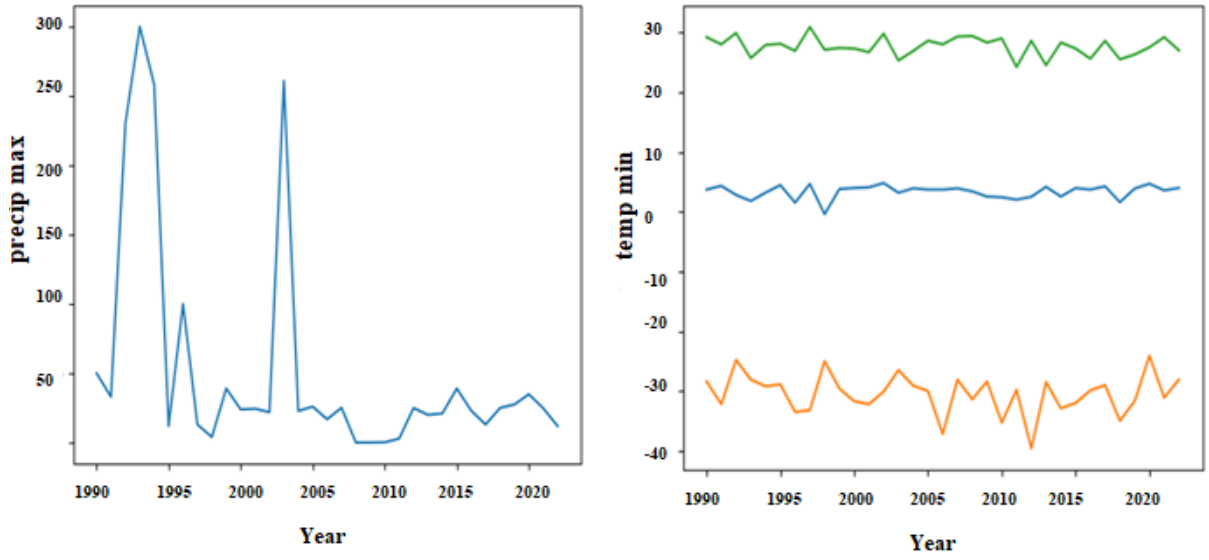

Figure 3. Precipitation and temperature trends North Kazakhstan region for 1990-2023

The plot shows significant spikes in maximum annual precipitation, suggesting years with extreme precipitation events. There is considerable variability from year to year, with some years experiencing very high maximum precipitation, while others remain closer to what may be the region's average. The most noticeable peaks occur around the early 1990s and the mid-2000s, which may be indicative of particularly heavy rainfall or snowfall events during those times.

The second plot displays three lines representing different statistical measures of temperature over the same period:

The top line (green) represents the maximum mean temperature for each year, showing a slight fluctuation but generally maintaining higher values, which suggests warmer average conditions in the region.

The middle line (blue) depicts the mean of the mean annual temperatures, which remains relatively stable over the period, indicating consistent average temperatures from year to year.

The bottom line (orange) shows the minimum mean annual temperature, with noticeable variability and a slight overall increase in the latter years, which could suggest a warming trend.

Next, it is visualized the yield of potatoes over the years using a line plot.

DOI: 10.37943/17LKYF9288

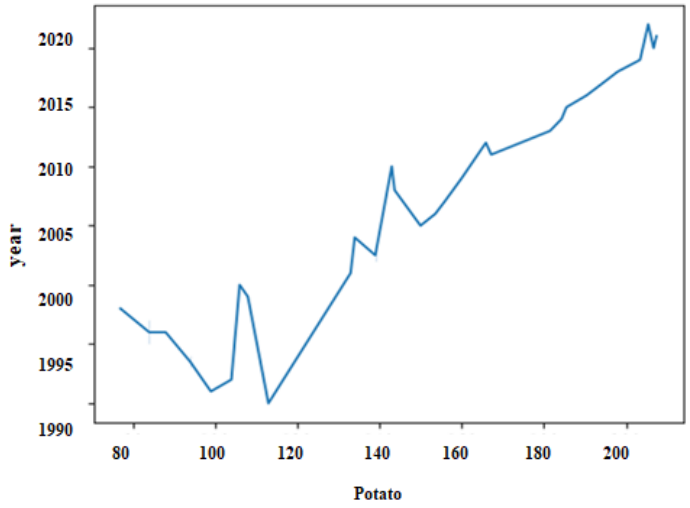

Figure 4. Potato yields of North Kazakhstan region

Overall, there is an upward trend in potato yields over the years, which may indicate improvements in agricultural practices, technological advancements, or favorable climatic conditions for potato farming in Kazakhstan. There are noticeable fluctuations within the trend, with some years showing sharp declines in yield. These could correspond to adverse weather events, pest outbreaks, or other agricultural challenges faced in those years. The most significant dips occur around the mid-1990s and early 2000s. These years could be of particular interest for further investigation to determine the causes of these declines. After the early 2000s, the trend is predominantly upward, suggesting a period of growth and possibly increased efficiency or resilience in potato production.

Furthermore, we generated a heat map to visualize correlations between all variables in the merged dataset (Fig. 5). Values close to 1 or -1 indicate a strong positive or negative correlation, respectively, while values near 0 suggest no correlation. The crop yields (grains, potato, open_veg, melons, sugar) show strong positive correlations with each other, indicated by the deep red color.

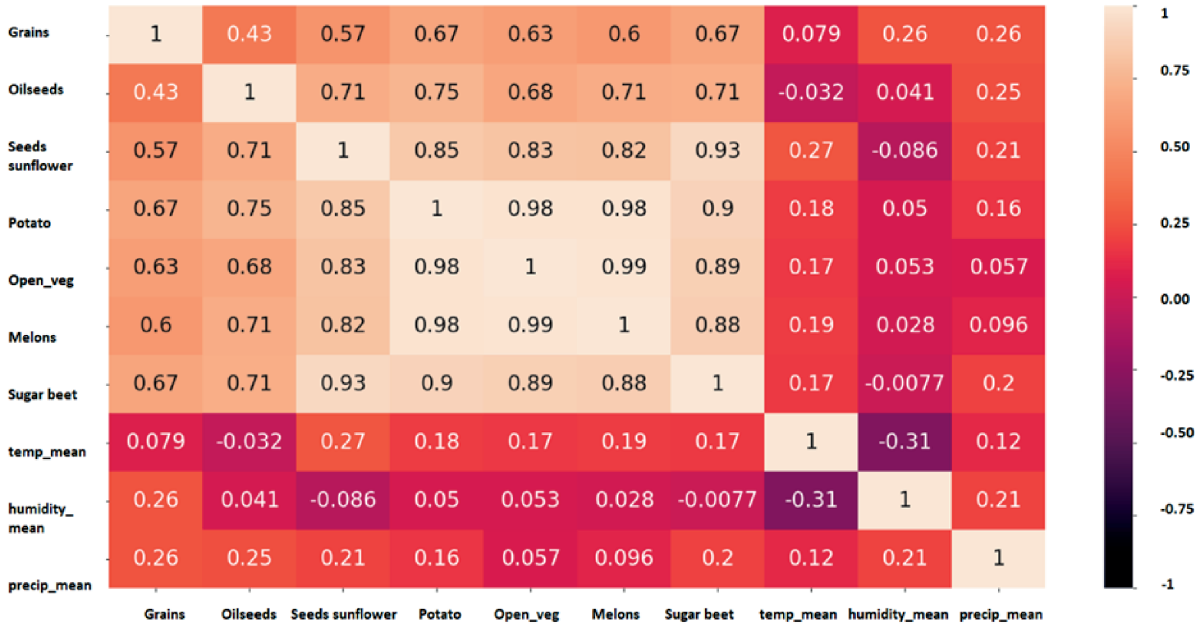

Figure 5. Correlation matrix between agroclimatic parameters and data on agricultural crop yields in the North Kazakhstan region

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

This suggests that when the yield of one crop type is high, others are likely to be high as well, which could be due to favorable weather conditions or effective agricultural practices across the board.

Crop yields generally have low to moderate positive correlations with temperature (temp_mean, temp_min, temp_max) and humidity (humidity_mean, humidity_min, humidity_max) factors. This can imply that certain levels of temperature and humidity are beneficial for the crops but do not necessarily increase yield linearly. Interestingly, there is a noticeable negative correlation between maximum temperature (temp_max) and the crop yields, suggesting that extremely high temperatures might negatively impact the yields. Maximum precipitation (precip_max) shows a moderate negative correlation with crop yields, hinting that excessive rainfall might be detrimental to the crops or could indicate flooding events.

Train model

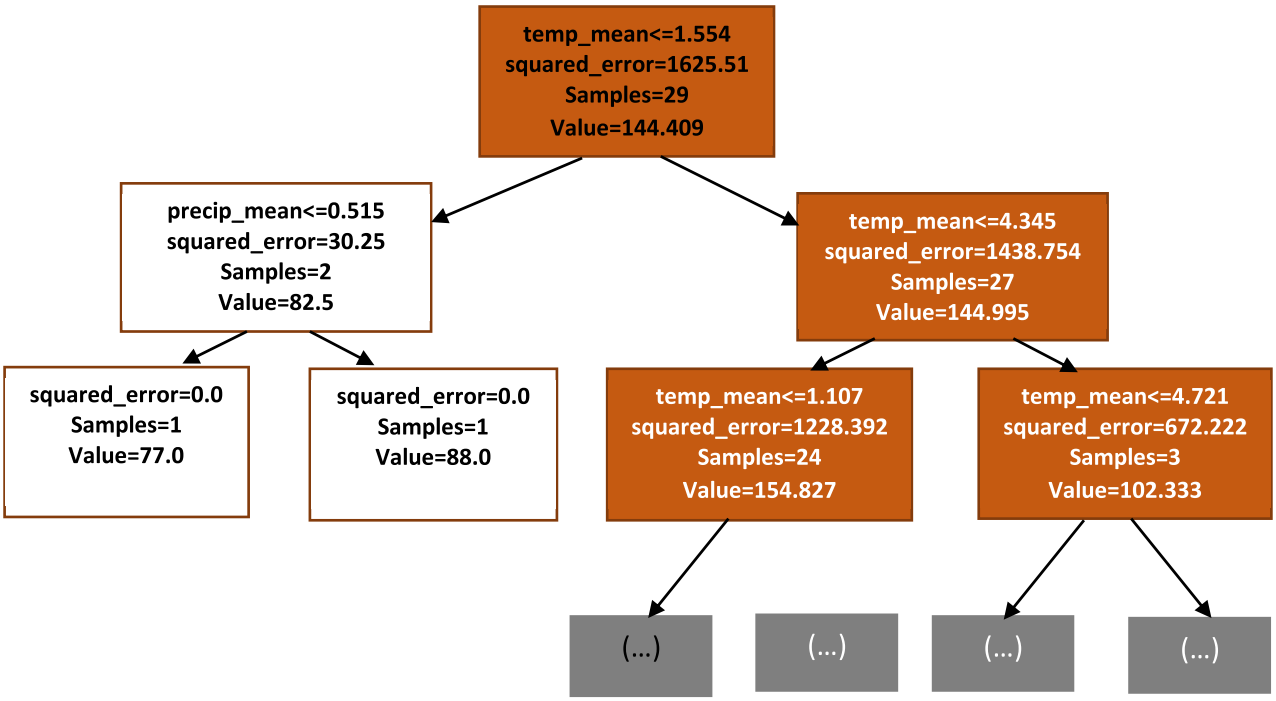

Figure 6. Decision Tree visualization for potato yield prediction task

Figure 6 illustrates Decision Tree visualization for potato yield prediction of North Kazakstan region.

The Decision Tree Regressor, Random Forest Regressor and Support Vector Machine Regressor machine learning algorithms were applied during the research (Figure 7).

DOI: 10.37943/17LKYF9288


```
new_data_rf = pd.DataFrame({ 'Year':[2023],'temp mean': [3.5], 'humidity mean': [69], 'precip mean': [1.5]})
prediction_rf = model_rf.predict(new_data_rf)
print(f'Predicted Potato Yield (Random Forest) for New Data: {prediction_rf[0]}')
from sklearn.model_selection import train_test_split
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error, r2_score
import pandas as pd
features = ['Year','temp mean', 'humidity mean', 'precip mean']
target = 'Potato'
X_train, X_test, y_train, y_test = train_test_split(df[features], df[target], test_size=0.25, random_state=42)
model_svm = SVR()
model_svm.fit(X_train, y_train)
predictions_svm = model_svm.predict(X_test)
mse_svm = mean_squared_error(y_test, predictions_svm)
r2_svm = r2_score(y_test, predictions_svm)
print(f'R-squared (SVM): {r2_svm}')
#print(f'mse_svm (SVM): {mse_svm}')
new_data_svm = pd.DataFrame({'Year':[2024],'temp mean': [3.5], 'humidity mean': [69], 'precip mean': [1.5]})
prediction_svm = model_svm.predict(new_data_svm)
print(f'Predicted Potato Yield (SVM): {prediction_svm[0]}')
results_df = pd.DataFrame({
    'Method': ['Decision Tree', 'Random Forest', 'Support Vector Machine (SVM)'],
    'Potato yeild': [predictions, predictions_rf, predictions_svm],
    'R-squared': [r_squared, r2_rf, r2_svm]
})
print(results_df)
max_r2_method = results_df.loc[results_df['R-squared'].idxmax(), 'Method']
print(f"\nMethod with best R-squared: {max_r2_method}")
```

```
Predicted Potato Yield (Random Forest) for New Data: 201.66879500000027
R-squared (SVM): -0.06502733979166542
mse_svm (SVM): 1487.2325780141427
Predicted Potato Yield (SVM): 141.00259558089448
                        Method  \
0                Decision Tree
1                Random Forest
2  Support Vector Machine (SVM)

                                        Potato yeild  R-squared
0  [205.4, 139.0, 185.5, 153.6, 84.0, 113.0, 143....   0.916788
1  [202.79404300000016, 138.26700000000002, 188.2...   0.943589
2  [141.00212993293704, 140.99979577907945, 141.0...  -0.065027

Method with best R-squared: Random Forest
```

Figure 7. Code fragment for potato yield forecasting

Comparative analysis

The results of several iterations of the above algorithms for forecasting potato yields for 2024 in the North Kazakstan region are presented below (table 1):

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

Table 1. Results of Random Forest, Decision Tree, Support Vector Machine algorithms with different test and training data

Algorithm	Number of iterations	Number of features	Number of targets	Train Size, %	Test size, %	RMSE	R2
Decision Tree	1	3	1	85	15	0.2854	0.97865
	2.	3	1	80	20	0.3245	0.97961
	3.	3	1	75	25	0.4122	0.91678
Random Forest	1.	3	1	85	15	0.3354	0.97365
	2.	3	1	80	20	0.5245	0.96137
	3.	3	1	75	25	0.4878	0.94358
Support Vector Machine	1.	3	1	90	10	77.258	-0.04454
	2.	3	1	80	20	75.242	-0.11308
	3.	3	1	75	25	144.25	-0.06502

The Random Forest, Support Vector Machine, Decision Tree algorithms were applied to predict potato yield in the study area for a given growing season. Random Forest Regressor algorithm showed the best performance with the best R^2 =0.97865.

The performance of the Decision Tree algorithm exhibits sensitivity to changes in the train/test size, particularly affecting classification metrics. Random Forest exhibits less sensitivity to changes in the test size compared to the Decision Tree. Regression metrics for Random Forest, including RMSE and R^2, demonstrate an improvement over the Decision Tree. The RMSE ranges from 0.25 to 0.46, and R^2 values are generally positive, indicating better predictive performance.

Overall Observations:

1. Best Overall Performance: Random Forest generally outperforms both Decision Tree and SVM in regression task.

2. Train/Test Size Trade-off: The sensitivity to test size suggests a trade-off that impacts model performance, and the optimal split may vary between algorithms.

3. Regression Challenges: Challenges in predicting the target variable are evident, as indicated by consistently negative R2 values. Further investigation into model complexity may be needed.

4. Fine-tuning Opportunities: Hyperparameter tuning and experimentation with different features could contribute to enhanced model performance. This analysis provides valuable insights for refining the models based on the specific characteristics of the dataset.

Conclusion

Throughout our paper, we conducted an in-depth analysis using sophisticated data processing and machine learning techniques to unravel the complex interactions between weather conditions and agricultural outputs in Kazakhstan.

We distilled large datasets into actionable insights, revealing strong correlations between various weather factors and crop yields. Our exploratory data analysis, visualized through line plots and heat maps, provided a clear depiction of trends and helped identify factors that could significantly influence agricultural productivity.

One of the critical challenges we faced was the integration of diverse datasets, which required careful preprocessing to ensure data integrity. We also had to navigate the inherent complexities of agricultural data, which presented both a methodological challenge and an opportunity to refine our analytical approaches.

The research's conclusion highlights several key takeaways:

- The importance of rigorous data cleaning and preprocessing to enable accurate modeling.
- The potential of using weather data to predict agricultural yields, offering valuable insights for farmers and policymakers.
- The realization that agricultural data analysis is complex and multifaceted, necessitating a nuanced approach to model building and evaluation.

In future iterations of this work, we could explore the integration of additional data sources, such as satellite imagery or soil quality data, to further refine our predictions. Moreover, delving into more advanced machine learning models and expanding our hyper parameter tuning could potentially yield even more accurate forecasts.

In conclusion, this research stands as a testament to the power of data science in agriculture. By blending traditional statistical methods with modern machine learning techniques, we have made strides in predicting agricultural yields, contributing valuable knowledge to the field and setting the stage for further research and innovation.

References

[1] M. Ziliani, M. Altaf, B. Aragon, R. Houborg, T. Franz, Y. Lu, J. Sheffield, I. Hoteit, M. McCabe. "Early season prediction of within-field crop yield variability by assimilating CubeSat data into a crop model". Agric. For. Meteorol., vol. 313, 2022, pp. 108736. https://doi.org/10.1016/j.agrformet.2021.108736

[2] X. Hao, Zh. Xiaohu, Y. Zi, J. Li, Q. Xiaolei, T. Yongchao, Y. Tian, Zh. Yan, C. Weixing. "Machine learning approaches can reduce environmental data requirements for regional yield potential simulation". Eur. J. Agron, vol. 129., 2021, pp.126335. https://doi.org/10.1016 /j.eja.2021.126335

[3] T. Kusainov, Zh. Zhakupova. "Statisticheskie svoystva i prognozirovanie urozhaynosti zernovyih v severnom zernoseyuschem regione Kazahstana". Vestnik nauki Kazahskogo agrotehnicheskogo universiteta im.S.Seyfullina (mezhdistsiplinarnyiy), vol.1 (96), 2018, pp.33-40.

[4] Uteulin V., Zhientaev S. "Drivers of Cereal Production Efficiency Improvement in Kazakhstan (The Case of the Kostanay Region)". J. Ecol. Eng., vol.23(10), 2022, pp.1-10. https://doi.org/10.12911/22998993/150624

[5] Z.H. Khalila, S.M. Abdullaeva. "Neural network for grain yield predicting based multispectral satellite imagery: comparative study". Procedia Comput. Sci., vol.186, 2021, pp. 269–278. https://doi.org/10.1016/ j.procs.2021.04.146

[6] T. Xiaopei, L.Haijun, F. Dongxue, Zh. Wenjie, Ch. Jie, L. Lun, Y. Li. "Prediction of field winter wheat yield using fewer parameters at middle growth stage by linear regression and the BP neural network method". Eur. J. Agron., vol.141, 2022, pp. 126621. https://doi.org/10.1016/j.eja.2022.126621

[7] Leisner, C. P. (2020). Review: Climate change impacts on food security- focus on perennial cropping systems and nutritional value. Plant Science, 293, 110412. https://doi.org/10.1016/j.plantsci.2020.110412

[8] Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access*, *9*, 4843–4873. https://doi.org/10.1109/access.2020.3048415

[9] Albuquerque, P.C., Cajueiro, D.O., & Rossi, M.D.C. (2022). Machine learning models for forecasting power electricity consumption using a high dimensional dataset. Expert Systems With Applications, 187, 115917. https://doi.org/10.1016/j.eswa.2021.115917

[10] Elbasi E., Zaki C., Topcu A.E., Abdelbaki W., Zreikat A.I., Cina E., Shdefat A., Saker L. Crop Prediction Model Using Machine Learning Algorithms. Applied Sciences. 2023; 13(16):9288. https://doi.org/10.3390/app13169288

[11] Adnan, N., Nordin, S.M., Bahruddin, M.a.B., & Tareq, A.H. (2019). A state-of-the-art review on facilitating sustainable agriculture through green fertilizer technology adoption: Assessing farm-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 17, MARCH 2024

ers behavior. Trends in Food Science and Technology, 86, 439–452. https://doi.org/10.1016/j.tifs.2019.02.040

[12] Hossain, A., Sabagh, A.E., Barutçular, C., Bhatt, R., Çığ, F., Seydoşoğlu, S., Turan, N., Konuşkan, Ö., Iqbal, M. A., Abdelhamid, M. T., Soler, C. M. T., Laing, A. M., & Saneoka, H. (2020). Sustainable crop production to ensuring food security under climate change: A Mediterranean perspective. Australian Journal of Crop Science, 14(03):2020, 439–446. https://doi.org/10.21475/ajcs.20.14.03.p1976

[13] Raju, C.M.A., Ashoka, D.V., & Prakash, A. (2023). CropCast: Harvesting the future with interfused machine learning and advanced stacking ensemble for precise crop prediction. Kuwait Journal of Science, 100160. https://doi.org/10.1016/j.kjs.2023.11.009

[14] Zhai, Z., Martínez, J.F., Beltrán, V., & Martínez, N. L. (2020). Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, *170*, 105256. https://doi.org/10.1016/j.compag.2020.105256

[15] Shi, F., Hao, Z., Zhang, X., & Hao, F. (2021). Changes in climate-crop yield relationships affect risks of crop yield reduction. *Agricultural and Forest Meteorology*, *304–305*, 108401. https://doi.org/10.1016/j.agrformet.2021.108401

[16] Morales, A.G., & Villalobos, F.J. (2023). Using machine learning for crop yield prediction in the past or the future. Frontiers in Plant Science, 14. https://doi.org/10.3389/fpls.2023.1128388

[17] Palanivel K., Surianarayanan Ch. An approach for prediction of crop yield using machine learning and big data techniques. Int. J. Comput. Eng. Technol., 10 (2019), pp. 110-18. http://iaeme.com/Home/issue/IJCET?Volume=10%26Issue=3

[18] Spanaki, K., Sivarajah, U., Fakhimi, M., Despoudi, S., & Irani, Z. (2021). Disruptive technologies in agricultural operations: a systematic review of AI-driven AgriTech research. *Annals of Operations Research*, *308*(1–2), 491–524. https://doi.org/10.1007/s10479-020-03922-z

[19] Afzal, S., Shokri, A., Ziapour, B.M., Shakibi, H., & Sobhani, B. (2024). Building energy consumption prediction and optimization using different neural network-assisted models; comparison of different networks and optimization algorithms. *Engineering Applications of Artificial Intelligence*, *127*, 107356. https://doi.org/10.1016/j.engappai.2023.107356

[20] S. Kujawa, G. Niedbała. "Artificial Neural Networks in Agriculture", Agric., vol. 11(6), 2021, pp. 496-497. https://doi.org/10.3390 /agriculture11060497

[21] M. Karatayev, M. Clarke, V. Salnikov, R. Bekseitova, M. Nizamova. "Monitoring climate change, drought conditions and wheat production in Eurasia: The case study of Kazakhstan". Heliyon, vol.8 (1), 2021, pp.e08660. https://doi.org/10.1016/j.heliyon.2021.e08660.

[22] Tanaka, A., Diagne, M., Saito, K. Causes of yield stagnation in irrigated lowland rice systems in the Senegal River Valley: Application of dichotomous decision tree analysis. *Field Crop Res.* 2015, *176*, 99–107. https://doi.org/10.1016/j.fcr.2015.02.020

[23] Banerjee, H.; Goswami, R.; Chakraborty, S.; Dutta, S.; Majumdar, K.; Satyanarayana, T.; Jat, M.L.; Zingore, S. Understanding biophysical and socio-economic determinants of maize (*Zea mays* L.) yield variability in eastern India. *Njas-Wagen. J. Life Sc.* 2022, *70–71*, 79–93. https://doi.org/10.1016/j.njas.2014.08.001

[24] Pang, A.; Chang, M.W.L.; Chen, Y. Evaluation of Random Forests (RF) for Regional and Local-Scale Wheat Yield Prediction in Southeast Australia. *Sensors* 2022, *22*, 717. https://doi.org/10.3390/s22030717