Implementation of machine learning algorithms in precision agriculture using soil data

1st Crisan Dragos
Computer Science dept.
University of Babes Bolyai
Cluj-Napoca, Romania
dragoscrisan.ubb@gmail.com

2nd Adriana Mihaela Coroiu
Computer Science dept.
University of Babes Bolyai
Cluj-Napoca, Romania
Adriana.coroiu@ubbcluj.ro

***Abstract*—This paper aims to showcase the impact machine learning algorithms can have on the agricultural sector, by providing valuable information to the farmers and aiding their decision-making process. Workers in this field face challenges in considering all factors when making decisions, often relying on intuition. Introducing a tool that would make it easier for them to make decisions while taking such data into account would greatly benefit them. In order to achieve that, we present multiple machine learning algorithms and their results on predicting crop yield based on soil nutrient levels and irrigation data, and comparing their results. Based on these findings, we determined that the Extreme Gradient Boosting method yielded the highest accuracy among the three tested methods, achieving over 99% accuracy. The Convolutional Neural Network (CNN) method achieved a respectable 95% accuracy but introduced a higher error rate. The paper also seeks to encourage the creation of more detailed and extensive datasets by providing examples of future possible developments that can be achieved and their potential to help farmers. One such example would be the pivot of the algorithms towards the recommendation of fertilizer quantity necessary to reach desired nutrient levels in the soil for certain crops. This model was trained on a public dataset collected in India, but can easily be adapted to other countries with their specific climates, provided there is access to relevant training data.**

***Index Terms*—Machine Learning (ML), Extreme Gradient Boosting (XGB), Convolutional Neural Network (CNN), Fully Convolutional Neural Network (FCNN), Nitrogen (N), Phosphorus (P), Potassium (K), agriculture, crop yield, application**

I. Introduction

The introduction of Machine Learning (ML) algorithms into the agricultural sector offers vital insights to farmers, aiming to enhance agricultural productivity in response to rising population pressures, increasing desertification of arable land, and significant climate shifts. Providing farmers with actionable information and recommendations is becoming increasingly crucial to ensure sustainable farming practices and food security in the face of these challenges. Efficient management of soil fertility, particularly the levels of essential nutrients such as Nitrogen (N), Phosphorus (P), and Potassium (K), is critical for optimizing crop yields. Despite the importance of these chemical factors, there are few comprehensive public datasets that cover them and other nutrient levels in the soil, posing significant challenges for research and application in precision agriculture.

In our work, we addressed this issue by integrating all three chemical factors: N, P, K into our predictive models. This approach provided a deeper understanding of soil fertility and its impact on crop yields, revealing the varying preferences of different crops. Additionally, it highlighted the differences in crop performance across various zones within the country, underscoring the importance of region-specific agricultural practices. To achieve this, we experimented with three different machine learning (ML) algorithms, in order to compare their accuracy and efficiency and chose the best suited ML for our dataset. We chose to retain three models: Extreme Gradient Boosting (XGBoost) and a Convolutional Neural Network (CNN), each chosen for their unique strengths in predicting crop yield, and a Fully Convolutional Neural Network (FCNN) for its potential in recommending optimal nutrient levels.

The paper is structured as follows: State of the Art reviews existing literature and projects that integrate ML algorithms into agriculture, highlighting their approaches to precision agriculture, Computational Experiments details the process of data extraction and transformation, the application of machine learning algorithms, and the implementation of the application, including its Frontend, Server, and Database components. Finally, Conclusion and Future Developments summarises our findings and discusses potential future enhancements for the application.

II. State of the art

The most important factors in the growth of plants are the three macronutrients: Nitrogen (N), Phosphorus (P) and Potassium (K). The efficiency with which they can be absorbed from the ground depends heavily on the climate and the amount and incidence of rain. Too much rain could impact the field negatively, as the macronutrients travel downwards in the ground thanks to the rain and can get to levels at which the roots of the plants cannot reach. This is the reason why there are countless studies and machine learning algorithms that focus on the meteorological data and its influence on crops.

As an example for calculating the impact of such data, we can look at a project that tries to find solutions for the food security problem in Saudi Arabia [1]. This paper focuses on the impact temperature, insecticides and rainfall have on

the crops, with the dataset containing information about all these and the crop yields for potatoes, rice, sorghum and wheat from 1994 to 2016. To interpret this data, they built a machine learning algorithm, namely an Artificial Neural Network (ANN), which is a highly effective multilayer perceptron, to predict crop yield. The results were then processed through various statistical evaluation metrics: the mean square error, the root-mean-square error, normalized root mean square error, Pearson's correlation coefficient, and the determination coefficient (R^2), with the best model being characterized by a result of R^2=0.9633. This study was motivated by rising concerns about food security and aimed to help farmers by giving them accurate foresight a year in advance to improve their management decisions.

In an attempt to integrate more sources of information in the prediction of crop production, we look at the following paper [2]. With Australia being one of the leading countries when it comes to exporting wheat, this study focuses on the timely and reliable wheat yield prediction in order to improve global food security. When collecting the information, they used both climate and satellite data in combination to create a new and unique set of data with better accuracy. They used multiple machine learning algorithms to study their performance, specifically random forest, neural network and support vector machine (SVM) in order to predict wheat yield across Australia from 2000 to 2014. The conclusion of the results is that combining climate and satellite data can enhance the performance of yield prediction, with R^2 reaching approximately 0.75.

Both papers showcase the influence clime has on the growth of crops. However, they both rely on the farmer to constantly use the desired amount of fertilizer to maintain the preferred levels of chemicals in the soil. In the next example we will look at pieces of work that take soil data into consideration as well.

In a study that aims to predict accurate estimation of crop yield and effective nitrogen management we find a similar approach to ours on studying a key chemical component, which is Nitrogen [3]. This approach uses remote sensing systems to improve yield production and nitrogen management while reducing operating costs and environmental impact. The key factor of realising a program like this is processing enormous amounts of remotely sensed data. This is handled by machine learning techniques, such as Gaussian Processes, Dirichlet Processes and Indian Buffet Process. The swift advancement in sensing technologies and machine learning algorithms offers affordable and thorough solutions for enhanced crop and environmental monitoring, aiding informed decision-making in agriculture. This highlights the potential of monitoring different factors through sensors and processing them with increasingly advanced ML methods to improve the agricultural sector now and in the near future.

III. Computational Experiments

A. Methods

1) Data extraction and transformation: The dataset we found that aligns with our vision and provides important data about the soil and irrigation that was collected from India between 2018 and 2019 [4]. This dataset is divided into four sections, which include data on crop price, crop production, soil analysis and water usage. For our study, we excluded the crop price dataset because it does not contribute to our research objectives. The remaining datasets were merged into one comprehensive dataset that contains information about the balance of macronutrients in the soil, namely pH level, organic matter, nitrogen content, phosphorus content, potassium content along with specifics regarding the district, area (hectares), season, irrigation method, and water availability. Following the merge operation, certain transformations of the data were required to fit the model. Depending on the algorithm, various encoding techniques were applied to handle data containing names and characters. After the data was transformed and ready, we created some statistics to analyse the data and draw some conclusions.

To understand the distribution of different values in the dataset, we first analyzed the frequency of various features. This initial analysis provided insights into the prevalence and range of key variables. In the following histogram(1), we can observe the frequency distribution of several critical parameters such as nitrogen content, phosphorus content, and pH level.

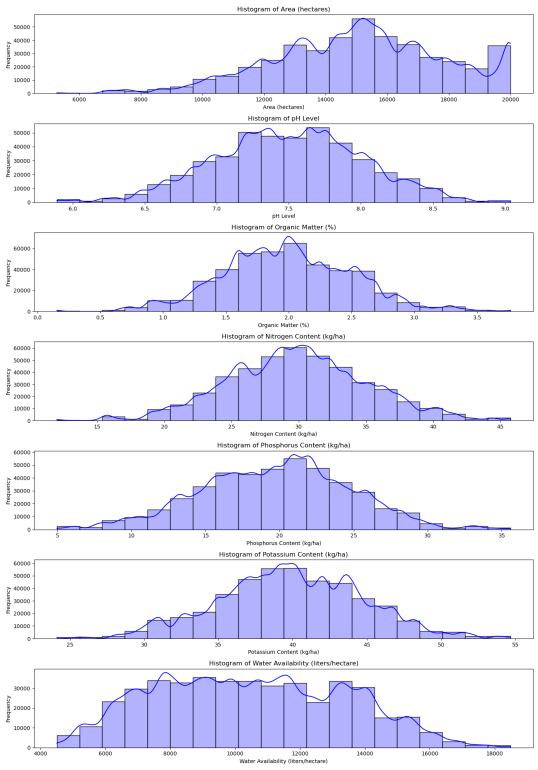

Fig. 1. Depiction of the crop Dataset statistics

Each and every one of these factors has a high impact on the productivity of the crops. To further analyze and visualize their influence on crop yield, we tried to compare different sets of data against crop yield production. Hexbin plots help to represent the density of data points, where the intensity of the color indicates the density, providing a clear visual of where data points are concentrated in relation to crop yield, as we can see in the following figure (2).

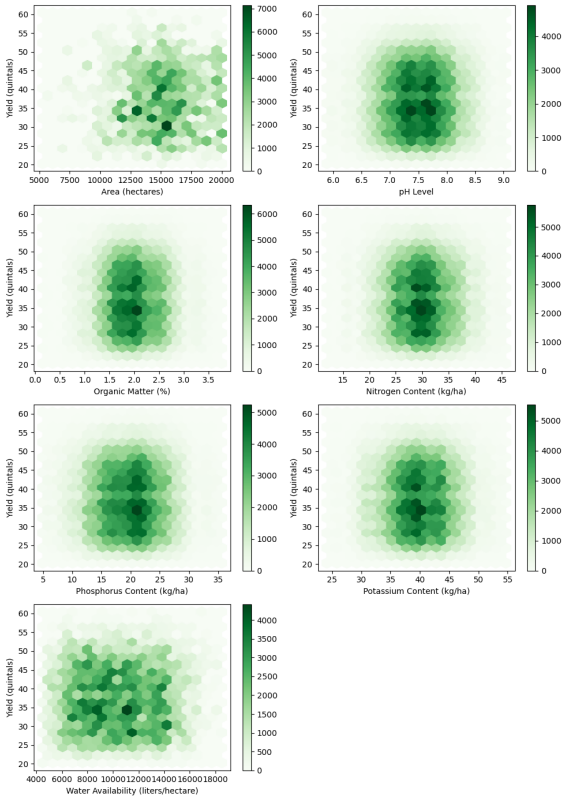

Fig. 2. Diagram of density of data around crop yield

As we can see around the values of Phosphorus, organic matter, Potassium and Nitrogen that the density converges to some central points, indicating potential correlations or optimal ranges that plants may favor for growth and yield. On the other hand, on water availability we can see more scattered central points of density on the water availability, as different crops have a lot of different preferences regarding water.

2) Extreme Gradient Boosting Algorithm: XGB which stands for Extreme Gradient Boosting, is a state-of-the-art machine learning algorithm renowned for its exceptional predictive performance. It creates a series of weak learners, usually decision trees and combines their results to improve the model's performance. It uses a boosting technique to create an extremely accurate ensemble model by having each weak learner after it correct the mistakes of its predecessors [5]. This renders it highly efficient on high amounts of tabular data that needs to be processed, making it a perfect candidate for the format of our dataset.

XGB was the optimal choice for our project, given the high amounts of tabular data with a high proportion of them having numerical value. Since the information is not that complex and does not have intricate correlations between particulars, it proves to be a better fit than the Convolutional Neural Network (CNN) module for the time being. On top of that, this model excels in situations that we might find if we extend to more general scenarios from agriculture, like datasets with missing data.

In our data preprocessing steps, we first eliminated the unnecessary information, handled missing values, encoded categorical features using OneHotEncoder, and standardized numerical features to ensure they were on a comparable scale. This preparation was crucial for assuring an accurate functioning model.

Following that, we divided the data. For the most accurate result, it was split into three: training, testing and validation. To avoid overfitting, we allocated 80% of the data for training and 10% each for testing and validation to prevent overfitting.

Regarding the specifics of the XGBoost model used, we chose the tree method 'hist', as it has the ability to use categorical values, which are very much present in five of our columns. For even more improvements on the accuracy we used hyperparameter tuning involving optimizing parameters such as 'n estimators', 'max depth', and 'learning rate'.

The model's performance was evaluated using metrics such as MSE to calculate the average of the squares of the differences between the actual and predicted values and Root Mean Squared Error to get the square root of the MSE, for an examination on the loss of the algorithm.

3) Convolutional Neural Network Algorithm: Despite the impressive performance of the XGBoost, we wanted to explore other options and chose CNN, which are generaly used in image processing but have shown promise in handling tabular data with spatial hierarchies and temporal patterns.

CNNs are known for their ability to automatically learn and extract features from the input data through convolutional layers. These layers apply different filters on the input data to find patterns and hierarchical structures. In our context, complex interactions between features like soil composition and irrigation can be found.

For preparing the data we had to once again encode the nominal features. To prepare it for the CNN algorithm, we had to transform them into an array. Then we followed a similar approach, only this time the proportions we arrived at in order to achieve the best results, were 60% training, 20% validation and 20% testing.

This experiment had two purposes: first, to compare the efficiency of the XGB model to other models, and second, to pave the way for future development, where multiple data from different periods of time would be collected and an algorithm able to handle more complex connections between the data would be necessary, like the influence certain

fertilizers and irrigation methods has over the chemical components of the soil over time.

4) Fully Connected Neural Network Algorithm: In addition to XGBoost and CNN, we implemented a Fully Connected Neural Network (FCNN) to explore a different direction of the project using the same dataset. We changed the desired result from crop yield to the three big chemical substances in the soil, namely Nitrogen (N), Phosphorus (P) and Potassium (K) and moved crop yield to the input. The purpose of the model is to create recommendations of balances between these chemicals in the soil in order to create prosperous conditions for the desired plant. We chose this model because the task of predicting NPK levels benefits from the FCNN's ability to learn direct mappings from input features to output recommendations.

5) Application Building: For this application we designed a comprehensive architecture comprising three main components: the Database, Server, and Frontend. This enables easier future development on the project and allows for an easier maintenance. Since the Frontend connects directly to the server, the processing of the input data is done directly into the requests of the server. This can later be improved by separating it and moving it to the Backend.

The database was built using DataGrid IDE with SQlite. These were the optimal choices considering the fact that we have a relativelly simple and small dataset, with information only about the users and the options for the dropdowns. This choice simplified deployment and maintanence for us. The database schema, illustrated in figure (3), provides a clear representation of the tables.

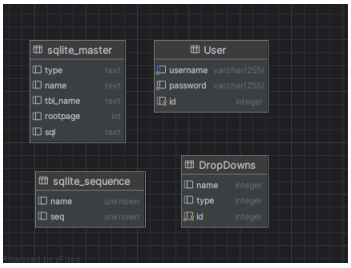

Fig. 3. Database schema

The server was implemented using Python for its well known simplicity and support for rapid development of server-side applications. We utilized the FastAPI framework due to its high performance, intuitive design and ease of use, which further streamlined our development process. Development was done in PyCharm, chosen for the debugging tools which rank highest in speed on the market.

The server's primary responsibilities include importing the machine learning models, loading the scalers and encoders for data processing on the input from the Frontend before making requests for the machine learning algorithms, handling http requests for authentification, fetching the data for the dropdowns and processing requests for executing each of the three algorithms, with correct formatting of the data. The choice of moving the options from the dropdown data to the database, although it adds some steps in order to gather this data for the Frontend, provides future scalability and easier access to modify the information. A concise visualization of the server structure can be seen in the following image (4).

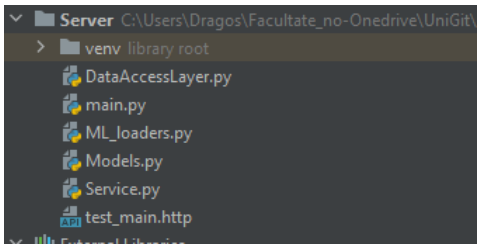

Fig. 4. Server structure

For the Frontend, we used Visual Studio with React for the multitude of its features. We used its component-based architecture to create our reusable components and make them easy to manage, maintain and test, the declarative UI for the changing of states in the HomePage for the change between register and login, the react router dom for the navigation and reacts state management for handling our forms.

We organized the features across three pages: the HomePage, designed with a pleasant aesthetic that complements the subject, as depicted in the illustration (5), and the CropYieldPage and ChemicalComponentPage. These pages share a similar design featuring forms for data entry, as illustrated in the image (6). In the HomePage, the user can login or register with his information. Once this form is completed succesfuly, it fades away and the buttons for navigation between the pages appear.

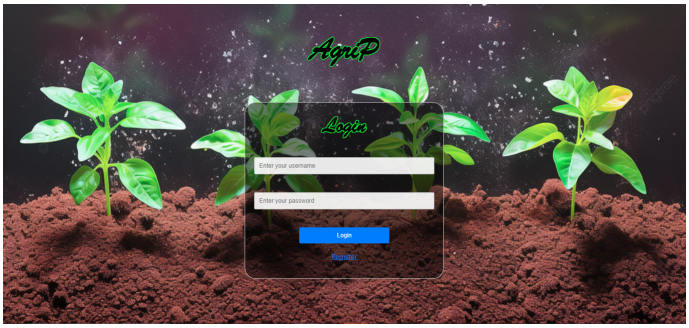

Fig. 5. Visual representation of login page

The other two pages are very similar with each other, each having a form to be completed with the data necessary for the machine learning algorithms. One of them provides predictions about the crop yield and has the option to alternate between the CNN and XGB model annd the other provides recommendations regarding the N, P, and K levels in the soil for the selected plant and conditions.

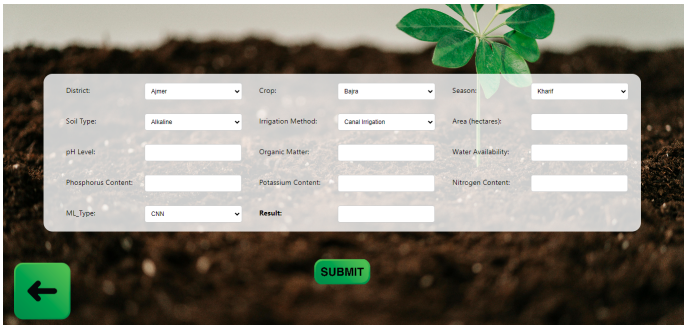

Fig. 6. Visual representation of crop yield prediction page

B. Results

In this section we will talk about the results of the models, starting with the XGB model predicting crop yield. The evaluation of the model is done using R^2 score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) across training, validation, and testing datasets. This assessment was crucial in ensuring that the model did not overfit the training data and maintained its predictive power across unseen data. On top of that, some further improvements were achieved through some hyperparameter tuning.

The performance metrics were calculated for each dataset. For the training dataset, the model produced results having a Mean Square Error (MSE) of 0.2729, a Root Mean Square Error (RMSE) of 0.5224, and an R^2 score of 0.9960. This high value results indicates that the model explains 99.6% of the variance in the training data, demonstrating excellent fit.

Although the value of the R squared is expected to lower on the validation dataset, we find some favorable values surpassing the training one, with an MSE of 0.2553, an RMSE of 0.5053, and an R^2 score of 0.9963. This proves the fact that the model adapts well to new data and that it is not overfited.

For the results of the testing data, we can see a slight decrease in accuracy, with values of MSE of 0.2851, RMSE of 0.5340, and an R^2 score of 0.9958. The small difference between the results of the three datasets indicates that the model maintains a consistency in its predictive accuracy. For a better visualisation of the three results compared, we can look at graphs showcased in the following picture (7).

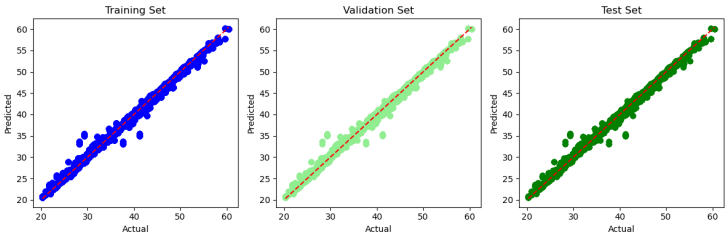

Fig. 7. Visualisation of the comparison between the results

To further improve the accuracy of our model hyperparameter tuning was performed using the following parameters: 300 estimators, a maximum depth of 8, a learning rate of 0.1, a subsample ratio of 1, and a column sample by tree ratio of 1. The model was trained with early stopping set to 10, evaluated on both the training and validation sets. This showed a slight improvement, with the R^2 being raised to 0.9978, as we can see in the following representation (8).

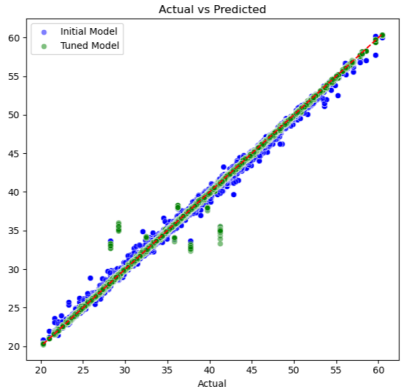

Fig. 8. Graph showcasing initial versus tuned model accuracy

Although not as effective, the CNN model has also managed to provide some beautiful results. Following a similar approach with segmenting the data, although we reached some different proportions of the datasets in order to provide the best results, the model has produced R^2 scores of 0.955. However, an unfortunate drawback was a significant raise in the Mean Square Error compared to the XGB model, with values reaching 3.52. To enhance the clarity of the error change visualization, we can refer to the graph depicted in the following depiction (9).

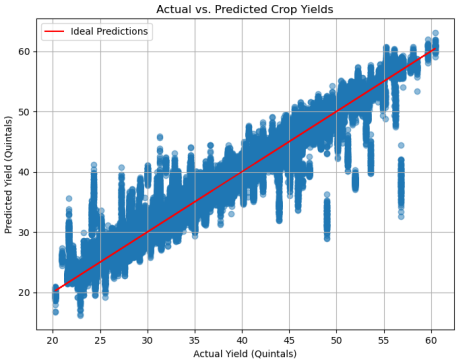

Fig. 9. Graph showcasing actual vs. predicted crop yields using CNN

The last model that uses FCNN, has shown even less accuracy, with the Mean Square Error reaching 4.12 and the score of R^2 being situated at 0.83. However, this is to be expected due to the changes made in order to achieve this model, namely, moving three data inputs as output, thus lowering the amount of data the model can train on by a significant amount.

The model also has to predict three different values as a result, which adds to the difficulty of reaching accurate results.

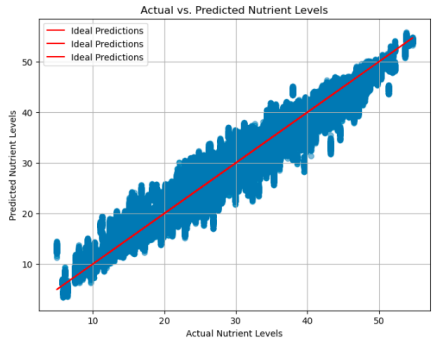

Fig. 10. Graph showcasing actual vs. predicted nutrient levels using FCNN

In the following illustration (10) we can see how this change has impacted the accuracy.

C. Discussions

Comparing the results of the XGB model produced over the training, validating and testing datasets provides a throughout check against overfiting. The slight variations in its performance shows that the model has not memorised the training data but learned patterns to make accurate predictions, even to new data. The hyperparameter succeded in further improving the model that was already at a really high accuracy. This demonstrates the suitability of the XGB model for our current program.

Through our process of experimenting with different models to reach the best result, we have also created a really meaningful new approach with the CNN model. It managed to provide some exceptional results, even though they were outranked by the XGB model and served an important role in our study by providing a comparative benchmark for the results obtained with the XGB model. The model also serves another purpose, to carve a path for future datasets that might be more complex. The model would excel if the temporal structure of the data would be added, with information about the soil being collected at different periods of times. In such a scenario, the model could become more accurate than the rest.

The FCNN model proved that although the results are less accurate, it is feasible to turn the project around and give it a new direction. If the situation would deem this result more valuable, through more research and gathering of more data this model could be improved and brought to a level of accuracy that would compete with the others.

D. Comparison with existing results

In the field of precision agriculture, the focus has not been on soil macronutrients as there are few accessible datasets that provide such information and it is hard to collect. Thanks to a systematic literature review over publications regarding the use of machine learning algorithms in the agriculture district [6], that does statistics on papers starting from 2008 to 2019, we can see the frequency with which certain features were used in similar projects illustrated by the following table (11).

Feature	# of times used
Temperature	24
Soil type	17
Rainfall	17
Crop information	13
Soil maps	12
Humidity	11
pH-value	11
Solar radiation	10
Precipitation	9
Images	8
Area of production	8
Fertilization	7
NDVI	6
Cation exchange capacity	6
Nitrogen	6
Irrigation	5
Potassium	5
Wind speed	5
Zinc	3
Magnesium	3
Shortwave radiation	2
Sulphur	2
Boron	2

Fig. 11. Table showing frequency of use of certain features

While the majority of research papers predominantly focus on climate data, a significant body of work also incorporates soil data into their analyses. One example would be this paper that focuses on mustard crop yield [8], where the authors attempted a similar approach of taking the macronutrients of the soil as input data and building different machine learning algorithms in order to asses the best performing one. In their case, the highest accuracies being achieved by the k-nearest neighbor (KNN) model and random forest model, with values of 88.67% and 94.13% respectively. Graphical representation of the results was provided in the cited paper and can be viewed in the following figure (12)

ML Algo	Accuracy	Precision	Recall	Specificity	F-Score
KNN	88.67%	78.14%	90.92%	80.72%	0.8405
NB	72.33%	70.91%	72.69%	75%	0.7179
MLR	80.24%	24.17%	96.66%	0.0000%	0.3866
ANN	76.86%	99.94%	99.61%	99.78%	0.9976
RF	94.13%	66.55%	99.66%	100%	0.7981

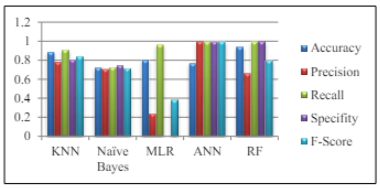

Fig. 2. Comparison of ML techniques under study

Fig. 12. Statistical report on the results of ML's

Other papers focus on the salinity of the soil or on one of the three big chemical components that have the highest impact on the crops development, like in the example provided in the state of the art section that takes nitrogen as its main focus [3].

We also found a project that worked on the same dataset as us, with an implementation of a similar XGB method with different proportions [7]. The model reached an astounding accuracy of 0.9942 before hyperparameter tuning, compared to our model that had 0.9960. This proves the importance of the information contained in the dataset and the impact it has on the result.

IV. Conclusions and future developments

The aim of this paper was first and foremost to show the accuracy that can be reached in predicting crop yield if relevant soil data and irrigation data is provided. By exploring different approaches, we were able to find the most accurate one, the Extreme Gradient Boosting model with an astounding accuracy of 99.78%. This highlights the influence soil and irrigation data has on the productivity of fields. We also managed to create an application to show the features provided by the models in a friendly and easy-to-use interface.

The project also excels in providing relevant scenarios for the future development of not only this project but the idea of implementing machine learning into agriculture. On top of improvements such as adding a backend layer to the application, deploying it, and making mobile versions of it, future developments can also be made on the dataset provided.

In the case where a more complex dataset is provided, that contains data about the fertilizer, its specifications and quantity used, the FCNN model could turn from recommending the levels of Potassium (K), Nitrogen (N) and Phosphorus (P) to recommending amounts of fertilizer of a certain type needed to be used to reach desired levels of N, P and K in the soil. With this model having the lowest accuracy, the dependability of the result would be a concern. In order to improve the accuracy of this result and also bring new possibilities of development, more extensive data would be necessary. This could be achieved by having a periodic monitoring of the levels of nutrients in the soil. In this case, the CNN model would be the most fitted. On top of being able to handle more intricate input data with a temporal structure, this could also be the case for the output data, as the model could start recommending the different quantities of fertilizer needed to be used over different periods of time. In the case that the model would still benefit from having its accuracy improved, additional meteorological data can be added and processed together with the rest of the dataset.

Acknowledgment

We are deeply grateful to Nicu Sima for his invaluable guidance and mentorship in the field of agriculture.. Their expertise and insights have been instrumental in deepening our understanding of the agricultural domain and its unique challenges. Their support has played a pivotal role in shaping my approach and finding a meaningful solution. Thank you for your unwavering support and encouragement.

We would like to also thank Dan Angheloiu, for his graphical design support for the frontend part of the application, as he contributed with all of the visual designs of the buttons used.

References

[1] Al-Adhaileh MH, Aldhyani THH. (2022). "Artificial intelligence framework for modeling and predicting crop yield to enhance food security in Saudi Arabia" *PeerJ Computer Science*, 8 https://peerj.com/articles/cs-1104/.

[2] Cai Y, Guan K, Lobell D, Potgieter AB, Wang S, Peng J, Xu T, Asseng S, Zhang Y, You L, Peng B. (2019). "Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches, Agricultural and Forest Meteorology" *Agricultural and Forest Meteorology*, 274, 144-159. ISSN 0168-1923. https://doi.org/10.1016/j.agrformet.2019.03.010.

[3] Chlingaryan A, Sukkarieh S, Whelan B. (2018). "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, Computers and Electronics in Agriculture" *Computers and Electronics in Agriculture*, 151, 61-69. ISSN 0168-1699. https://doi.org/10.1016/j.compag.2018.05.012.

[4] Suraj (suraj520). (2018-2019). "Agricultural Data for Rajasthan, India (2018-2019)" Retrieved from https://www.kaggle.com/datasets/suraj520/agricultural-data-for-rajasthan-india-2018-2019

[5] Zeravan Arif Ali, Ziyad H. Abduljabbar, Hanan A. Taher, Amira Bibo Sallow, Saman M. Almufti. (2023) "Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review", https://journals.nawroz.edu.krd/index.php/ajnu/article/view/1612

[6] van Klompenburg T, Kassahun A, Catal C. (2020). "Crop yield prediction using machine learning: A systematic literature review" *Computers and Electronics in Agriculture*, 177, 105709. ISSN 0168-1699. https://doi.org/10.1016/j.compag.2020.105709.

[7] Aagrawal V. (n.d.). "Predicting Crop Yield in Rajasthan." Retrieved from https://www.kaggle.com/code/atom1991/predicting-crop-yield-in-rajasthan-r2-99-75.

[8] Pandith V, et al. (2020). "Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis." *Journal of Scientific Research*, 64(2), 394-398. Retrieved from https://www.researchgate.net/publication/343219111_Performance_Evaluation_of_Machine_Learning_Techniques_for_Mustard_Crop_Yield_Prediction_from_Soil_Analysis.