Data Mining Project: Crop Recommendation

Alireza Dehghan Nayeri
Master's in Computer Science and Engineering
University of Oulu
Alireza79.nayeri@gmail.com
2406137

Usama Raheel
Master's Programme in Computer Science and Engineering
University of Oulu
Usama.Raheel@student.oulu.fi
2410692

Karlo Rosenthal
Exchange student - Databases and Knowledge Bases
University of Oulu
karlo.rosenthal@student.oulu.fi
2500063

Bharathi Sekar
Master's Programme in Computer Science and Engineering
University of Oulu
bsekar24@student.oulu.fi
2409295

***Abstract*—This project aimed to develop a predictive model to recommend the most suitable crops based on soil properties (N, P, K, pH) and environmental factors (temperature, humidity, rainfall). The dataset, sourced from Kaggle, contained 2200 data points across 22 crops. Key steps included data visualization, outlier detection, standardization, and augmentation to enhance model performance. Three models were evaluated: k-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Random Forest (RF). The RF model achieved the highest accuracy (99%), demonstrating robust performance. Data augmentation showed minimal impact, likely due to the dataset's small but balanced nature. The project confirmed the feasibility of accurate crop classification using machine learning, with potential applications in agriculture for optimizing crop selection. Future work could expand the dataset and explore deep learning models for improved generalization. The project highlighted the importance of clean data, effective preprocessing, and collaborative teamwork.**

***Index Terms*—Crops, Recommender system, Data mining, Data visualization**

I. Introduction

Agriculture faces challenges such as climate change, soil degradation, and the demand for sustainable practices, which requires innovative solutions powered by machine learning to optimize crop selection by analyzing soil properties and environmental factors to determine the best crops. The data set we were given contains factors relevant to the growth of crops in India. Despite the localization of the data, we believe that our solution could also be applied to other parts of the world, when given the right data. During this project, we wish to deepen our understanding of data mining and the methods associated with it, while working on solving a problem that affects the whole world.

II. Related work

During our research of the subject of our work, we have found articles and research papers on that topic. One of the most similar papers was [1] *Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables*. During the creation of this project we have studied further about our question through articles and research papers, however this research stood out among others. It describes methods similar to those we used during our preliminary visualization and other methods which could potentially apply to our research.

III. Objectives

The objective of our project was one of the key focus points for our team. We have debated many different research questions. In collaboration with our mentor, we managed to narrow down the scope after the first seminar. Even during the preprocessing of the data we did not have the specific research question, but we were on the right path.

After the second seminar, we chose these questions as the basis for the data processing. "Can a predictive model based on N, P, K, temperature, humidity, pH, and rainfall accurately classify which crop is best suited for a given set of conditions?" This served as our main question, which has helped us identify a follow-up question. "Assess the effects of data augmentation by comparing model performance (precision, recall, accuracy) between original and augmented datasets."

Based on our previous experience and research, we strongly believe that it is indeed possible for the model to correctly classify crops based on the given parameters. The second question serves as an additional way to investigate the abilities of the model and to discover in which way augmenting the data affects the model.

IV. Data

The data set was sourced from the Kaggle archives [2], accumulated by the Indian Chamber of Food and Agriculture [3]. The data of the data collection was not disclosed; however, the dataset was uploaded on Kaggle in April of 2024. In Kaggle the context of this dataset notes that this data should be used to create a predictive model which suggests the most suitable crops and helps farmers make informed decisions.

The data is stored in a csv file, it is comprised of 2200 data points and spans 22 different crops. The parametars of each data point are described in a table below:

TABLE I
SUMMARY OF SOIL AND ENVIRONMENTAL ATTRIBUTES

Attribute	Description
N	Quantity of nitrogen in the soil.
P	Quantity of phosphorus in the soil.
K	Quantity of potassium in the soil.
Temperature	Reflects the temperature conditions.
Humidity	Indicates the humidity levels, expressed.
pH	Measures the acidity or alkalinity of the soil.
Rainfall	Volume of rainfall.

TABLE II
DESCRIPTIVE STATISTICS ABOUT THE DATASET

Stat	N	P	K	Temp	RH	pH	RNFL
N	2200	2200	2200	2200	2200	2200	2200
Mean	50.55	53.36	48.14	25.61	71.48	6.46	103.46
Std	36.91	32.98	50.64	5.06	22.26	0.77	54.95
Min	0	5	5	8.82	14.25	3.50	20.21
25%	21	28	20	22.76	60.26	5.97	64.55
50%	37	51	32	25.59	80.47	6.42	94.86
75%	84.25	68	49	28.56	89.94	6.92	124.26
Max	140	145	205	43.67	99.98	9.93	298.56

A. Data visualization

One way of visualization was creating a Heatmap. It shows the correlation matrix for the numerical variables in the dataset. The correlations with a value close to 0 represent a weak or even no linear relationship. On the other hand, values closer to 1 or -1 exhibit strong linear relations.

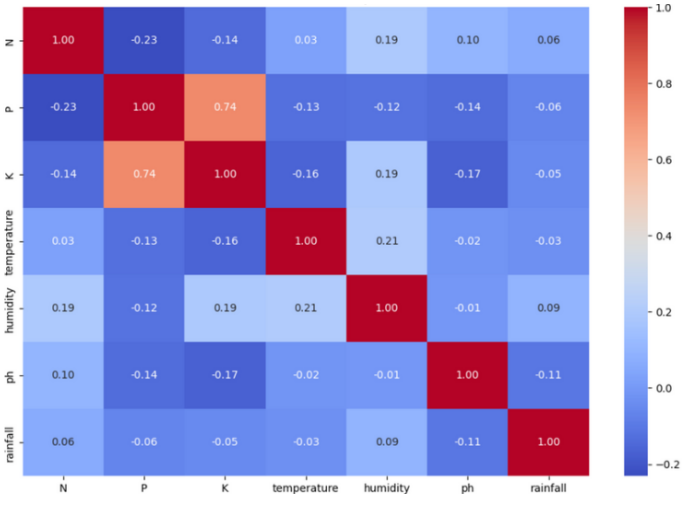

Fig. 1. Correlation Heatmap.

During the analysis of the correlations we managed to classify the factors into groups. While K and P have the strongest relation, it is also evident that N i also highly related to those factors. Temperature, humidity and rainfall also had noticeable relation. The only remaining factor was pH, which had the strongest relations with N, P, K. Based on these findings we categorized N, P, K as nutrients; temperature, humidity and rainfall as environmental conditions; and pH as soil characteristics.

At this point in the project development we considered the research question of how important would each factor or category be when predicting a crop. However, the lack of agricultural knowledge quickly changed our focus.

As a final step of visualizing the data we generated graphical representations of data values. This was a crucial step to help us identify methods for detecting outliers for each parameter.

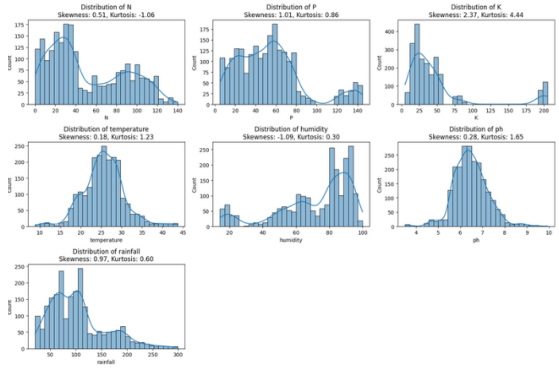

Fig. 2. Dataset overview

After visualizing the data we gathered and discussed the plans for the pre-processing. As a result, we devised a flowchart, which served as a guideline for all future work.

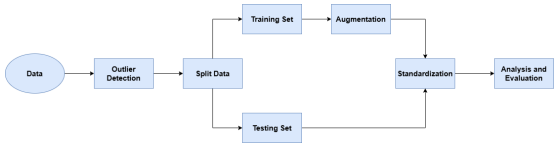

Fig. 3. Project flowchart

B. Outlier detection

Shortly after we began the process by detecting the outliers in the data. It was an additional step of verifying the quality of data, because no abnormalities were found during visualization.

The methods used to detect outliers were based on the grpahs generated during the data visualization. For the bell shaped data we used the Z-Score method. ”A z-score is a statistical measure that describes the position of a raw score in terms of its distance from the mean, measured in standard deviation units.” [4] This method applied to N (Nitrogen), temperature and pH. For the parameters which displayed skewness we used IQR (Interquartile Range) Method. ”The IQR is a statistical measure that characterizes the spread and variability of a dataset.” [5] This method applied to P (Phosphorus), humidity and rainfall. Lastly for K (Potassium), the data had a heavy tail, hence the usage of the Z-score

modified method. "The modified z score is a standardized score that measures outlier strength or how much a particular score differs from the typical score. Using standard deviation units, it approximates the difference of the score from the median. The modified z score might be more robust than the standard z score because it relies on the median for calculating the z score. It is less influenced by outliers when compared to the standard z score." [6]

TABLE III
OUTLIER DETECTION RESULTS

Column	Method Used	Outliers
N (Nitrogen)	Z-score	0
P (Phosphorus)	IQR	138
K (Potassium)	Modified Z-score	200
Temperature	Z-score	33
Humidity	IQR	30
pH	Z-score	30
Rainfall	IQR	100

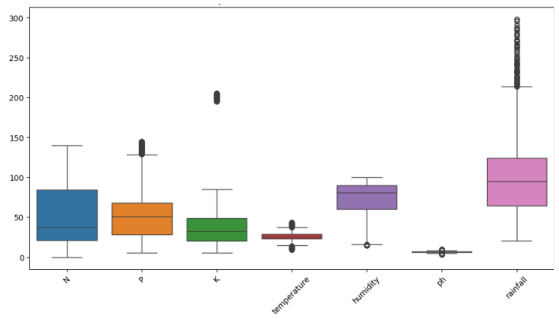

Fig. 4. Outlier detection boxplot

Ranging from the highest to lowest, the most outliers were found in the data of Potassium. It was heavily skewed and warranted an intervention before continuing with the steps of the project. Phosphorus and rainfall also had skewed values, but not as heavily as potassium. Temperature, humidity and pH had a moderate amount of outliers. The only parameter without any outliers was nitrogen. With these findings, we were sure that before the further analysis of the date, we needed to standardize it.

C. Data splitting

Given that each crop in the dataset has an equal number of samples, random splitting of the data is deemed sufficient for model training and evaluation. This approach ensures that the distribution of samples across both training and testing sets remains unbiased and representative.

D. Data augmentation

As part of our secondary research question, we duplicated the training set and augmented it with additional values. Other than for the purposes of the research, augmenting the data creates additional and more diverse training samples. It also simulates real-world variations and helps machine learning models generalize better.

The augmentations we used were Gaussian noise to add small random variations to the nutrients N, P, K. It was done using a bell curve distribution to simulate natural measurement variations. The second augmentation was environmental variability. It Introduces realistic fluctuations in temperature (±3.0°C), humidity (±7.0%), and soil pH (±0.3) to simulate daily and seasonal changes.

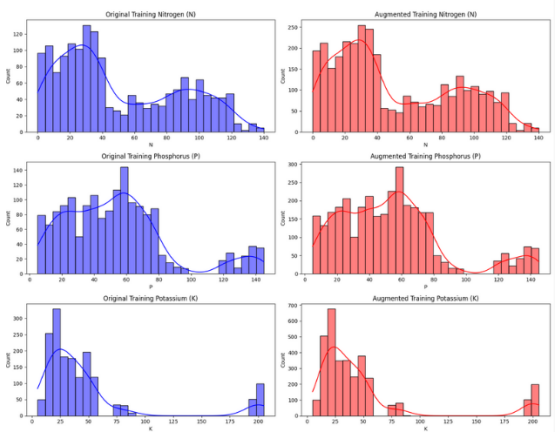

Fig. 5. Original training set shape: (1650, 7)

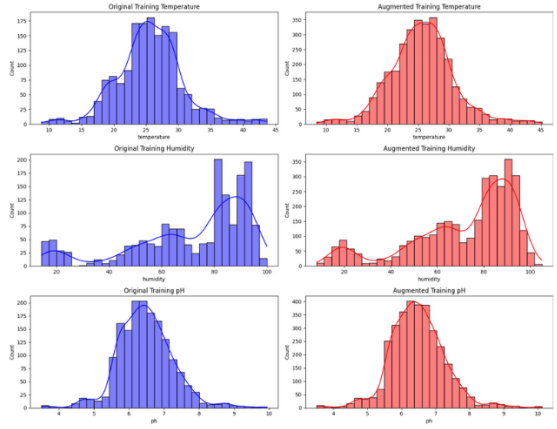

Fig. 6. Augmented training set shape: (3300, 7)

E. Standardization

The final step of preprocessing was the standardization of data. This step is crucial for the optimization of the data and improves the performance of machine learning models by scaling features, particularly when the data has varying distributions.

Just as we used different methods to detect outliers based on visualized data, we also applied various standardization techniques depending on the identified outliers. Robust scaling was used for N, P, K. It applies to features with outliers, this method scales the data while being resistant to outliers. For temperature and humidity we used standard scaling, which standardizes features to have a mean of 0 and standard deviation of 1, ideal for features with a normal distribution.

Power transformation is used for features with slight skewness, this transformation normalizes the distribution. We used it do standardize pH. Lastly we used log transformation for rainfall. It is applied to highly skewed features to compress extreme values.

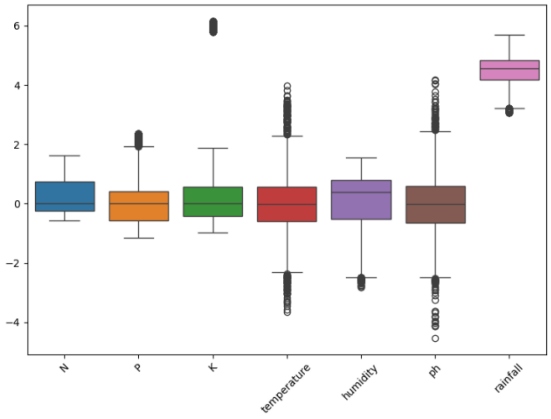

Fig. 7. Feature distribution after inital scaling

However, rainfall was initially not scaled properly due to high skewness. After a discussion with our mentor, we applied Robust scaling instead of Log transformation. The final result of our standardization is shown in the figure below:

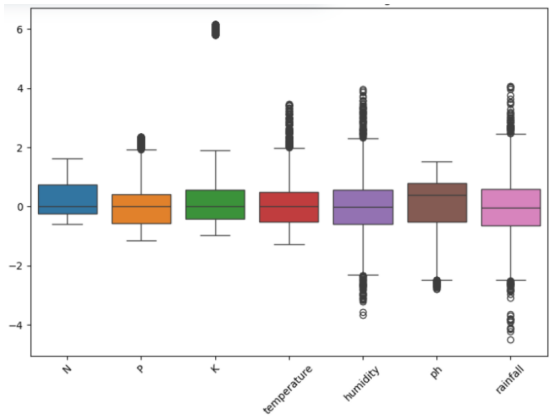

Fig. 8. Feature distribution after additional scaling

V. Methods

The way we approached this data was by framing it as a multi-class classification problem, given its nature of 22 different crop classes. By setting the approach, the decision of further methodology was made easier and clearer.

As the baseline model we chose k-Nearest Neighbors (kNN). This model was chosen, due to its simplicity and effectiveness in datasets with clear feature separability. " It is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. " [7] During the implementation it requiring minimal hyperparameter tuning and we have managed to implement it using Sklearn's KNeighborsClassifier.

The second method we used was Support Vector Machines (SVM). "A support vector machine (SVM) is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space." [8] We have managed to implement it with an RBF kernel to capture potential non-linear relationships between features. One of the main reasons for choosing this method was due to its suitability for complex environmental data. For the implementation we utilized Sklearn's SVC class.

Lastly, we decided to use a random forest. "Random forest is a commonly-used machine learning algorithm, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems." [9] We selected it for its robustness in managing feature interactions and outliers through ensemble learning, making it ideal for this dataset. Implemented it using Sklearn's RandomForestClassifier.

To determine the quality of the models we relied on evaluation matrices. Performance was assessed using Accuracy, Precision, Recall, and F1-Score, calculated with Sklearn's classification_report and confusion_matrix functions. Visualizations of these metrics, including confusion matrices and classification report heatmaps, were generated using Seaborn and Matplotlib for clear interpretation.

VI. Results

The results are represented by two different graphics, by a confusion matrix and by a classification report heatmap. "Confusion matrix is a table that is used in classification problems to assess where errors in the model were made. The rows represent the actual classes the outcomes should have been. While the columns represent the predictions we have made. Using this table it is easy to see which predictions are wrong." [10] "A heatmap depicts values for a main variable of interest across two axis variables as a grid of colored squares. The axis variables are divided into ranges like a bar chart or histogram, and each cell's color indicates the value of the main variable in the corresponding cell range." [11]

A. kNN Classifier

The accuracy of both unaugmendted and augmented testset was the same at 0.97. It achieved high performance due to clear feature separability. Data augmentation introduced minor noise, but the macro-average F1-score remained stable at 0.97.

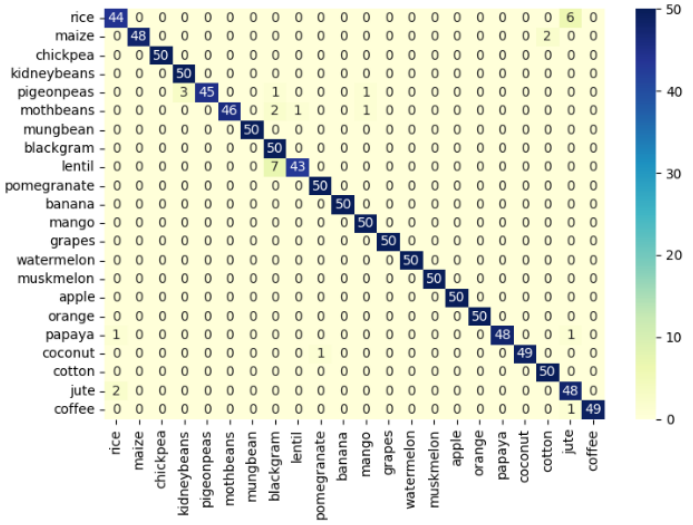

Fig. 9. KNN - Confusion Matrix

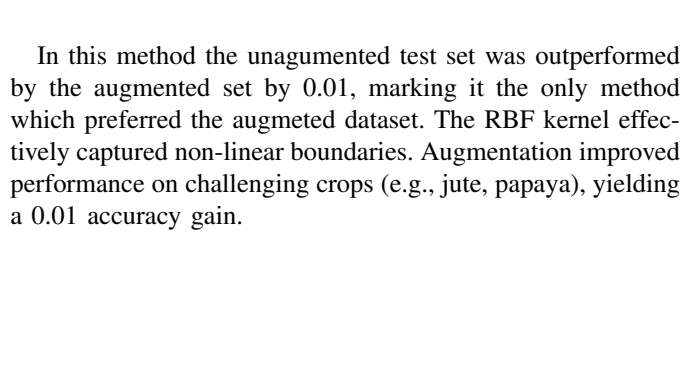

Fig. 10. KNN - Classification Report Heatmap

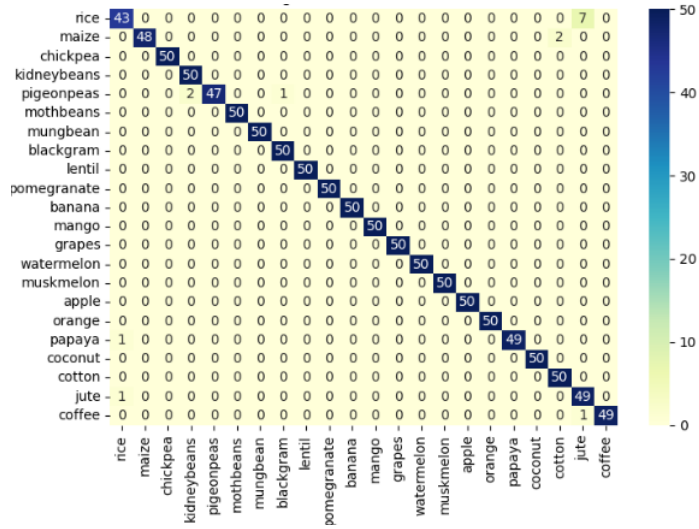

Fig. 11. SVM Augmented - Confusion Matrix

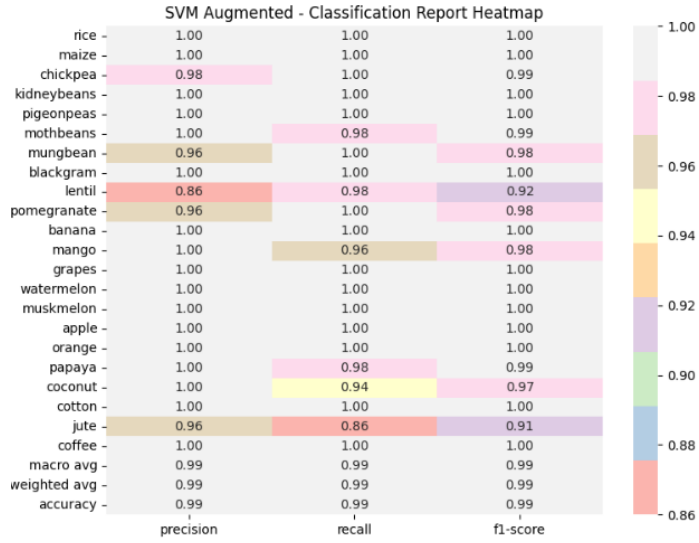

Fig. 12. SVM Augmented - Classification Report Heatmap

B. SVM Classifier

In this method the unagumented test set was outperformed by the augmented set by 0.01, marking it the only method which preferred the augmeted dataset. The RBF kernel effectively captured non-linear boundaries. Augmentation improved performance on challenging crops (e.g., jute, papaya), yielding a 0.01 accuracy gain.

C. Random forest Classifier

The highest performing module was the Random forest. With both unagumented and augmented datasets achieving the accuracy of 0.99. It demonstrated near-perfect performance, robustly handling feature interactions and outliers. Augmentation had minimal impact, with slight F1-score drops due to noise, suggesting the model's optimality on the original dataset.

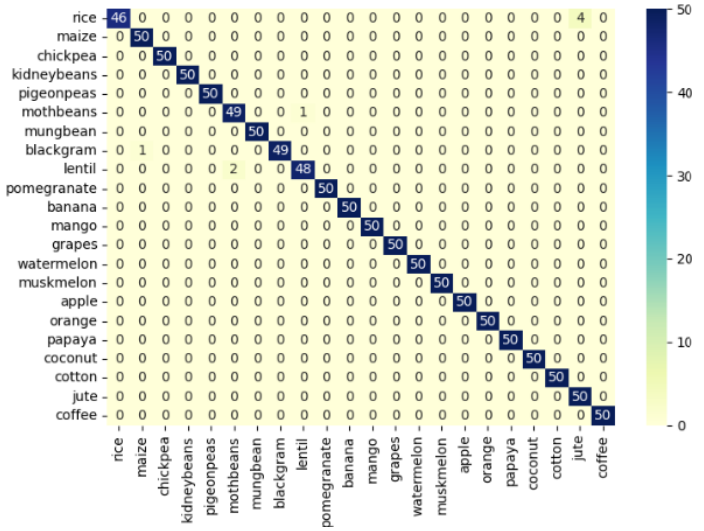

Fig. 13. RF - Confusion Matrix

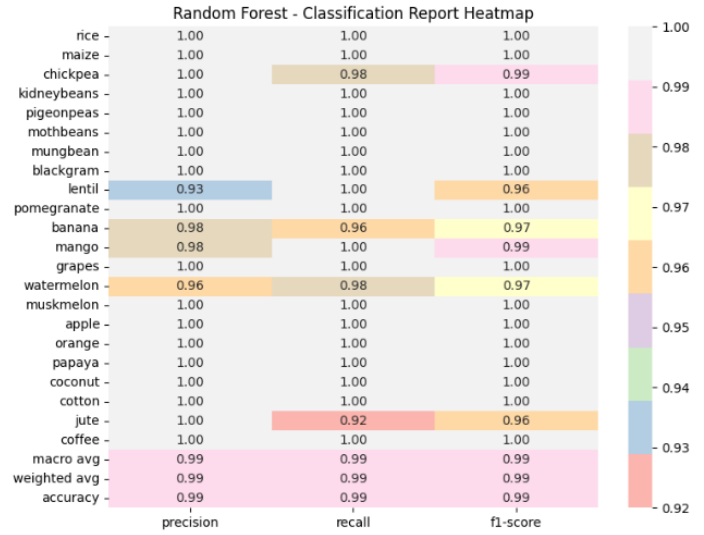

Fig. 14. RF - Classification Report Heatmap

D. Comparison

TABLE IV
ACCURACY COMPARISON

Model	Unaugmented	Augmented	Δ
kNN	0.97	0.97	0
SVM	0.98	0.99	0.01
Random Forest	0.99	0.99	0

In this table it is represented that every model reached the accuracy of at least 0.97. The kNN showed consistency but slightly lower performance on nuanced crops compared to the other methods. The only delta was found in the SVM model which favors the augmented test set by 0.01. For the Random forest it is evident that the data is already optimal. There were no significant points which would favor one test set over the other, with both reaching the accuracy of 0.99.

VII. DISCUSSION

A. Key findings

Our main research question, ”Can a predictive model based on N, P, K, temperature, humidity, pH, and rainfall accurately classify which crop is best suited for a given set of conditions?”, has successfully been achieved. Our models have achieved ¿97% accuracy (RF/SVM: 99%) , proving crop prediction is feasible with N-P-K and climate data.

Our secondary research question, ”Assess the effects of data augmentation by comparing model performance (precision, recall, accuracy) between original and augmented datasets.”, has also been satisfied. We discovered minimal improvement while using augmented test sets; likely due to dataset size (small but balanced).

B. Comparison

The findings of our project are closely aligned with previous studies, particularly with [1]. We share similar methodologies and both projects were successful in confirming their original thesis. Our findings did not conflict one another, moreover they complimented each other. For both projects, the usage of random forests proved to be more fruitful than using other processing methods.

Because of the difference in the nature of both projects, it is safe to assume that our findings extend the findings of the previous studies. Our focus on using augmented test sets was not described in the previous project. Such a focus on our side also pushed our project in the more experimental direction.

C. Limitations

After conducting the first tests, our team was concerned with the level of success our models produced. Even after using the augmented test sets, since there were little to none changes. At that point we tried different models, but the results were still excellent. We found these results as a bit of a limitation, likely because the dataset lacks spatial/temporal diversity. Also our findings would need to be tested in the real world to validate them.

D. Future work

The way we would proceed with this work in the future would firstly be by expanding the dataset with regional soil/climate variations. That way we could identify stronger relations and some potentially new correlations in the data. The next step would focus on processing the data. We could test it with deep learning models (e.g., CNNs) for higher complexity.

VIII. REFLECTION ON GROUP WORK

The most interesting part of our project was working with the data through visualization and assessment. The dataset, though relatively small, was impressively clean, which allowed us to focus more on analysis and less on preprocessing. Creating visualizations to explore and communicate patterns

in the data was especially engaging. The agricultural domain itself added another layer of interest. It is a field where data-driven insights can genuinely impact real-world issues, such as food security and drought management. In particular, formulating a research question that could benefit the crop industry felt rewarding and gave the project a sense of purpose.

The most challenging aspect of the project was dealing with the small size of the dataset. While its cleanliness was a plus, the limited number of samples posed a significant challenge in terms of building a model that would generalize well. We had to carefully explore and implement data augmentation strategies that would increase the sample size without distorting the underlying data distributions. Designing an evaluation pipeline to test these augmentation methods also took considerable time and effort. Finding the right balance between expanding the data and preserving its core characteristics was a delicate and time-consuming process.

Throughout the project, we gained valuable hands-on experience, bridging the gap between theoretical understanding and practical application. We learned a wide range of skills, from data visualization and pre-processing to exploring different augmentation techniques and evaluating classification models. Beyond technical skills, the project also emphasized the importance of teamwork, iteration, and feedback in shaping the research direction. One of the key takeaways was realizing that even with a small dataset, effective augmentation and a well-thought-out analysis strategy can lead to impressive model performance.

Our process for deciding on the research question was collaborative and iterative. We began by brainstorming multiple potential questions. Around six or seven, after conducting an initial exploratory data analysis. These were presented during our early seminar sessions, and the feedback from instructors and mentors helped us progressively refine our direction. Insights from the data itself, such as missing values and the dataset's limitations, also played a role in narrowing our focus. Eventually, we settled on two main questions: one related to crop classification, and the other exploring the impact of augmentation on model performance.

Task division within the team was guided by individual interests and strengths. Usama focused on data augmentation methods and Python coding, while Alireza and Bharathi contributed to classification modeling and visualization. Karlo focued on documenting the research. That way we created a functional team, where each member contributed meaningfully, while also learning from others. Overall, the project developed in line with our expectations, though we were pleasantly surprised by the high metric scores our models achieved.

REFERENCES

[1] B. Dey, J. Ferdous, R. Ahmed (2024.), Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables, Researchgate

[2] Kaggle, https://www.kaggle.com/datasets/siddharthss/crop-recommendation-dataset/data, April, 2025.

[3] Indian Chamber of Food and Agriculture, https://www.icfa.org.in/, April, 2025.

[4] S. McLeod, PhD,Z-Score [standard Score], Simplypsychology.org, April, 2025.

[5] S. Hleap, Unmasking the Outliers: Exploring the Interquartile Range Method for Reliable Data Analysis, Procogia.com, April, 2025.

[6] IBM, Modified z score, April, 2025.

[7] IBM, What is the k-nearest neighbors (KNN) algorithm?, April, 2025.

[8] IBM, What are support vector machines (SVMs)?, April, 2025.

[9] IBM, What is random forest?, April, 2025.

[10] W3Schools, Machine Learning - Confusion Matrix, April, 2025.

[11] M. Yi, A complete guide to heatmaps, Atlassian.com, April, 2025.