Term Project Report: *Machine Learning for Smarter Crop Planning*

Debasis Pani, Noah Sprenger, Sahithi Challapalli
COMP3009
Professor Ahmed Abdeen Hamed
11/17/2025

Abstract

In this project, the Naïve Bayes classifier is applied to identify the ideal crop for specific given soil and climate features between maize and rice. The soil nutrients include levels of nitrogen, phosphorus, potassium (N, P, K) and pH levels, along with environmental factors including temperature, humidity, and rainfall. We developed a model to predict the suitable crop for the given growing conditions to show how probabilistic machine learning can be used for data-driven farming decisions. Our objective was to use the understanding of the mathematics behind Naïve Bayes classification and implement similar logic in Python with the use of `GaussianNB` classifier from `scikit-learn`. We wanted to evaluate this model using two crops, rice and maize, to analyze the strengths, limitations and real-world application of Naïve Bayes in the agricultural industry.

Our hypothesis is that Naïve Bayes can accurately predict the optimal crop types based on soil and environmental features, providing a simple yet effective tool for agricultural decision-making.

1 Introduction

Agriculture is a fundamental part of food security, economic stability and ensures the well-being of humans. A vital component of farming for centuries has been fertile soil that produced crops in abundance. However, over the years, climate change and changes in soil composition have affected how plants grow and not considering these differences is an inefficient use of resources. As changes continue to occur, poor crop choices can lead to reduced yields, inefficient use of resources and environmental degradation from unsuitable farming practices. As more agricultural datasets become available, data-driven decisions can help farmers in choosing crops that are better aligned with their local growing needs.

In this project, crop suitability is examined through Naïve Bayes classification which is a probabilistic machine learning approach based on the Bayes Theorem. Naïve Bayes estimates prior probabilities from historical data and assumes conditional independence of the features, allowing the dataset to be easily interpretable.

The dataset collected from *Kaggle* contains the soil nutrient measurements and environmental conditions paired with the results for ideal crop of each given value. Throughout this project, the focus was on a condensed dataset of 15 values of 9 rice and 6 maize samples.

N	P	K	Temp	Humidity	pH	Rainfall	Crop
90	42	43	20.88	82.00	6.50	202.94	rice
85	58	41	21.77	80.32	7.04	226.66	rice
60	55	44	23.00	82.32	7.84	263.96	rice
74	35	40	26.49	80.16	6.98	242.86	rice
78	42	42	20.13	81.60	7.63	262.72	rice
69	37	42	23.06	83.37	7.07	251.06	rice
69	55	38	22.71	82.64	5.70	271.32	rice
94	53	40	20.28	82.89	5.72	241.97	rice
89	54	38	24.52	83.54	6.69	230.45	rice
71	54	16	22.61	63.69	5.75	87.76	maize
61	44	17	26.10	71.57	6.93	102.27	maize
80	43	16	23.56	71.59	6.66	66.72	maize
73	58	21	19.97	57.68	6.60	60.65	maize
61	38	20	18.48	62.70	5.97	65.44	maize
68	41	16	21.78	57.81	6.16	102.09	maize

Table 1: Fifteen-sample subset of rice and maize conditions.

As shown in the data, it is clear that rice samples require high humidity and rainfall while maize requires lower humidity and rainfall. Having such strong patterns is ideal for Naïve Bayes making agricultural data to be compatible for such classifications.

2 Related Work

When looking into research done by others, we found several studies that used Naïve Bayes for crop recommendations. One study done by Tedy Setiadi in "*Implementation Of Naïve Bayes Method In Food Crops Planting Recommendation*" used the Bayes Model to recommend crops with the dataset collected from Indonesia. These features included weather, humidity, air pressure, rainfall and historical yields. Though their process and dataset slightly differed, their model produced an accuracy of 85.7 percent. Coming to a similar conclusion as us, they found that the Naïve Bayes model has "proven by high accuracy values, low error rates, and high-performance results".

Another research article we referred to throughout our project was published through the Institute of Electrical and Electronics Engineers (IEEE) that had an almost identical approach to what we completed through this project. Swetha and Senthilkumar focused on using features of soil nutrients, temperature, rainfall and humidity to find the optimal crop. The researchers explained the need to convert continuous variables into categorical bins prior to applying the Bayes theorem. The limitations of this model were also emphasized as they focused on using other models to compare the results. In the results, they highlighted how the the Naïve Bayes out performs Decision Trees, Random Forest and Logistic Regression.

They stated that this model is the *"most effective due to its simplicity, efficiency, and robust performance with large datasets."*

The combinations of these works showed that the Naïve Bayes was previously used in the agricultural world and our project was already in progress. They showed that our model can use these projects to take into account what has been done so that we take a new approach in this project. Our project builds on this prior work by applying the Bayes Theorem to soil nutrient and environmental data by validating both manual and algorithmic implementations.

3 Methodology

The classification dataset is formulated using the numerical measurements of soil nutrients and environmental conditions. The attributes included nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, pH levels, and rainfall. All features were prepared for classification by converting the continuous variables into categorical values using quantile binning at 1/3 and 2/3 to create three levels of low, medium, and high. This transformation ensured that all attributes were consistent categorically and interpretable to evaluate the data using Naive Bayes.

The following table shows the recorded values for rice and maize that were converted into conditional probabilities. For example, nitrogen was "High" in 4 samples, giving it a likelihood value of 0.444 for rice and "High" in 1 sample with a likelihood value of 0.167 for maize. The conditional probability was calculated by counting the occurrences and dividing by the number of samples. These calculations were done for all other attributes for both crops.

Rice (9 samples)

Feature	Category	Count	$P(x \mid \text{Rice})$
N	High	4	0.444
P	Low	4	0.444
K	Medium	4	0.444
Temperature	Cool (Low)	3	0.333
Humidity	Medium	4	0.444
pH	Acidic (Low)	2	0.222
Rainfall	Medium	4	0.444

Table 2: Example conditional probabilities for rice.

Maize (6 samples)

The A below represents the crops (rice or maize) and B represents the attributes. The theorem can be written as the following formula, which was used to calculate the prior,

Feature	Category	Count	$P(x \mid \text{Maize})$
N	High	1	0.167
P	Low	2	0.333
K	Medium	1	0.167
Temperature	Cool (Low)	2	0.333
Humidity	Medium	1	0.167
pH	Acidic (Low)	3	0.500
Rainfall	Medium	1	0.167

Table 3: Example conditional probabilities for maize.

likelihood and posterior probabilities.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

The prior probability is calculated through the *P(A)* and corresponds to how often each crop appears in the dataset. *P(A—B)* is the likelihood of each attribute and how common each category is for the particular crop. The result of this calculation is *P(A—B)*, the posterior probability, which identifies the crop that is the most probable for the given conditions. Once the data was preprocessed, the probabilities were calculated manually to understand and verify the process.

For Rice:

$$\begin{aligned} P(\text{Rice}|X) &\propto P(\text{Rice}) \times (0.444)^4 \times (0.333) \times (0.444) \times (0.222) \\ &= 0.6 \times 7.707 \times 10^{-4} = 4.624 \times 10^{-4} \end{aligned}$$

For Maize:

$$\begin{aligned} P(\text{Maize}|X) &\propto P(\text{Maize}) \times (0.167)^4 \times (0.333)^2 \times (0.167) \times (0.500) \\ &= 0.4 \times 1.715 \times 10^{-5} = 6.86 \times 10^{-6} \end{aligned}$$

After these calculations, a similar procedure was applied to an algorithm through the use of `GaussianNB` from `scikit-learn` library. With the small dataset of 15 samples, the algorithm computed the posterior probabilities for both and then selected the crop with the higher value as the most suitable for the given climate and soil conditions.

4 Implementation

The Naive Bayes classifier was implemented using Python where all processing and model training was performed through Google Colab. We began with importing the dataset of 15 samples and prepare it to be classified using key libraries such as, `pandas`, `numpy`, and `matplotlib` . These libraries included tools for data manipulation of the numerical, categorical features and data visualization tools.

As mentioned in the methodology, the next step in the process was applying the quantile based thresholds using the `cut()` function from `pandas`. This ensured that the manual

calculations done previously aligned with categorical representations. After continuing the same with the rest of the features, the dataset was separated into the feature variables and targets. The model was then trained with `GaussianNB` classifier from the `scikit-learn` library and computed all probabilities in a consistent manner. The results allowed the model to evaluate the probability of rice vs maize based on these binning attributes. The output of the test is shown below which shows that the model accuracy of the manual calculations and the Python code is a match proving the model translates into code well.

```
SAMPLE OUTPUT
-------------
Model Accuracy: 100.00%

Actual vs Predicted:
Actual  Predicted
0 rice  rice
1 rice  rice
2 rice  rice

Predicted crop for new conditions → rice
```

5 Results

The results of the test cases using the sample of 15 achieved a near perfect accuracy and also matched the predictions of the manual calculations. This confirms that the mathematics of the Naïve Bayes theorem and the `scikit-learn` implementation were consistent. The high accuracy of the model can be expected due to the distinct values of the features for rice and maize. Their values of humidity, rainfall and potassium were so separated that the model was able to produce such accurate results.

To explore this model, we were interested in experimenting with the full dataset that included more crops and samples. We looked to see how the model would behave with additional data and were curious if the high accuracy shown in the results would also occur in the larger dataset. Using a similar process with `GaussianNB`, we trained and tested the model with multiple crops. Despite the varying crops, the results produced were similar in terms of the accuracy of crop prediction.

To further test the validity of these results, we experimented with different training samples and tested to see if the model had issues or the data. To create a better understanding, we also created visuals of the crops for each feature and noticed that there was no overlap in potassium. Since Naïve Bayes assumes the independence of features, the strong separation in potassium and little overlap in humidity and rainfall gives the model a higher accuracy. For readability, the graph below only shows the results of rice, maize and chickpeas. Despite adding multiple crops the results are the same where there are no overlapping values. After realizing the dataset was not realistic, we altered the sample data and challenged the model to test how the performance changed. These tests confirmed that the model was indeed working and with a proper dataset can still predict the suitable crop but might change the accuracy to be more realistic.

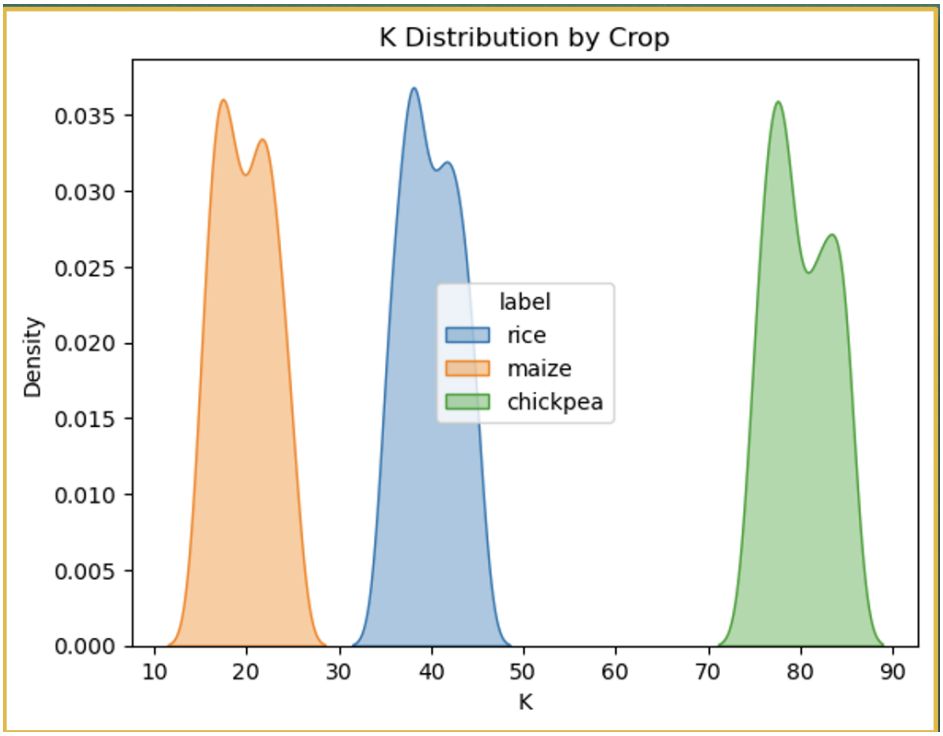

Figure 1: K Distribution by Crop

6 Discussion

At first, our goal was to find a dataset that can be implemented to create a solution to a real-world problem. On our journey, we found the agricultural dataset and thought it would be a great match for the Naïve Bayes theorem while producing results that can make a true difference. When analyzing the dataset, we found that there were more issues than we had realized. Though the results produced a hundred percent accuracy, we questioned the validity and further examined the reason why. As we explored, we found that certain features were not overlapping and were so distinct causing the model to spot these differences rather than learn.

As we faced this issue, we thought to do more research on whether the model would actually work with a better dataset or if our hypothesis was incorrect. Our research led us to find others that have done this exact work and saw that the Naïve Bayes theorem can recommend the crop given certain attributes. Though the accuracy wasn't as high as ours, we found that they used a real agricultural dataset and were still able to produce similar results. We realized that our dataset was synthetically clean and was not a true representation of real agricultural values, but were able to confirm our model was still ideal.

Although Naïve Bayes has many advantages due to its simplicity, interpretability and efficiency, we found that it also has its limitations. The independent assumption of the model can cause the model to be restrictive and since the features are correlated, it can oversimplify the crops when determining the prediction. Another limitation is when a particular feature or combination is not in the training data, the model will give it a probability of 0 which can lead to incorrect predictions. Keeping these limitations in mind, this model can be improved

by using real data that shows overlap to improve the realism of agricultural datasets. The addition of more features can also increase the accuracy and can take different regions into consideration when giving crop recommendations. As mentioned in the related research work that we found, using other models when testing the dataset can show how the results compare to different methods while proving the importance of Naïve Bayes and its simplicity.

7 Conclusion

This project demonstrated how the use of Naïve Bayes classification can be applied to the recommendation of agricultural crops. The mathematical foundations of Bayes theorem we reviewed and the implementation of the logic in Python can produce accurate and interpretable results regardless of the dataset size. A major takeaway from this project is that the extreme accuracy can be a reflection of the highly separable data rather than the capabilities of the model. It is important to recognize and critically evaluate the features when interpreting the results. Regardless of the issues in the dataset, Naïve Bayes remains an effective starting point and provides strong baseline for crop recommendation, especially with a true structured and clean dataset.

There are several ways this work can be expanded in the future to make the system more practical and realistic. One main improvement we considered was to incorporate real agricultural data with natural noise, geographical variation, seasonal patterns that would produce more accurate results. Expanding the features to include soil textures or drainage capacity can also make the model customization to farmers. From the modeling perspective, comparing the Naïve Bayes model with alternative methods, such as, Decision Trees or Logistic Regression, can reveal additional patterns and compatibilities that the simple probabilistic model cannot identify. Combining these changes, the project can produce results beyond what was demonstrated in this project and can truly make a difference in the farming world.

In conclusion, we were able to identify that even a simple probabilistic model can provide meaningful insight into crop suitability. This study also emphasizes the importance of recognizing the algorithm's capabilities, the features that it was trained on, and dataset's limitations.

References

1. Setiadi, D., Wibowo, M., & Sulistyo, S. (2020). *Implementation of Naïve Bayes Method in Food Crops Planting Recommendation*. Department of Agriculture, Yogyakarta. Available at: https://www.researchgate.net/publication/343417429_Implementation_Of_Naive_Bayes_Method_In_Food_Crops_Planting_Recommendation

2. Patil, P. S., & Sherekar, S. S. (2019). *Crop Prediction Using Machine Learning Techniques*. International Journal of Engineering and Advanced Technology (IJEAT), 9(1). Available at: https://www.ijeat.org/wp-content/uploads/papers/v9i1/A1171109119.pdf

3. IEEE Research Article (2024). *Evaluation of Naïve Bayes for Agricultural Prediction Tasks*. IEEE Access. Available at: https://ieeexplore.ieee.org/document/10933086

4. Ingle, A. (2020). *Crop Recommendation Dataset*. Kaggle. Available at: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset