DATA MINING ANALYST FOR CLASSIFYING PLANT GROWTH DATA USING THE NAIVE BAYES METHOD

Eko Afrianto
Institute Technology and Science Mandala
082329319788, 68121
ekoafrianto@itsm.ac.id

Ferry Wiranto
Institute Technology and Science Mandala
08991087383, 68121
ferry@itsm.ac.id

Agung Muliawan
Institute Technology and Science Mandala
089610345572, 68121
agung.muliawan@itsm.ac.id

Muhdar
Institute Technology and Science Mandala
082264094834, 68121
muhdar@itsm.ac.id

ABSTRACT

Data mining is a powerful tool that involves extracting useful information from large datasets. In the context of plant growth classification, data mining can be used to analyze various factors, such as soil composition, climate conditions, and plant characteristics, to predict and classify plant growth patterns. In addition, data mining can also be used to predict potential pest outbreaks or disease outbreaks, allowing farmers to take proactive measures to protect their crops. The Naive Bayes algorithm is a popular machine learning technique that is widely used in data mining applications, including in the agricultural sector. One of its key strengths is its simplicity and ease of implementation, making it a practical choice for farmers looking to leverage data-driven insights. The application of the Naive Bayes method using Rapid Miner data mining for classifying plant growth data yielded an accuracy of 67.50%, demonstrating a moderate level of performance in distinguishing between different growth outcomes. The precision of the model was calculated at 72.00%, indicating that over half of the positive predictions (growth milestones classified as "yes") were correct. The recall was higher, at 75.00%, suggesting that the model successfully identified a majority of the actual positive cases. However, the AUC (Area Under the Curve) score of 0.685, These results suggest that while the Naive Bayes classifier is a useful tool for this task. while above random chance, reflects the model's limited ability to discriminate between positive and negative classes effectively.

Keywords : Plant Growth; Naïve Bayes Method; Data Mining; Rapid Miner

1. INTRODUCTION

Data mining is a powerful tool that involves extracting useful information from large datasets. In the context of plant growth classification, data mining can be used to analyze various factors, such as soil composition, climate conditions, and plant characteristics, to predict and classify plant growth patterns. By utilizing data mining techniques, researchers and farmers can gain valuable insights into optimal growing conditions for different plant species, leading to more efficient and sustainable agricultural practices (Rahaman et al., 2019). By understanding the relationships between different variables, data mining can help identify trends and patterns that may not be apparent through traditional methods. This information can then be used to make informed decisions about which crops to plant, when to plant them, and how to optimize their growth. Overall, data mining in plant growth classification has the potential to revolutionize the way we approach agriculture, leading to increased yields, reduced waste, and a more environmentally friendly industry. In addition, data mining can also be used to predict potential pest outbreaks or disease outbreaks, allowing farmers to take proactive measures to protect their crops (Salazar & Rios, 2010). By analyzing historical data and current conditions, farmers can make more informed decisions about pest management strategies, reducing the need for harmful pesticides. This not only benefits the environment but also improves the quality and safety of the food produced. Overall, data mining in agriculture has the potential to transform the industry into a more sustainable and productive system for the future. With the ability to analyze vast amounts of data quickly and accurately, farmers are able to optimize their resources and minimize waste. This efficiency leads to higher yields, lower production costs, and ultimately, increased profitability for farmers (Organization, 2017). This not only benefits the environment but also ensures a more secure food supply for future generations. By utilizing data mining, farmers can also optimize resource allocation and reduce waste, leading to increased efficiency and profitability. Additionally, the insights gained from data mining can help farmers adapt to changing conditions and improve overall productivity in the long term. Overall, data mining plays a crucial role in modern agriculture by enabling farmers to make informed decisions that benefit both their bottom line and the environment, with the help of data-driven insights (Pierre et al., 2018). By utilizing data mining, farmers can also enhance crop quality and yield through precision agriculture techniques. This not only benefits the farmers themselves, but also contributes to global food security and sustainability efforts.

Agriculture is essential in understanding how data-driven insights can revolutionize farming practices. Naive Bayes is a popular machine learning algorithm that is particularly well-suited for classification tasks in agriculture, such as predicting crop diseases or optimizing fertilizer usage. Furthermore, the algorithm's simplicity and efficiency make it a valuable tool for farmers looking to incorporate data-driven insights into their operations (Chamseddine et al., 2023). For example, by analyzing historical data on weather patterns, soil conditions, and crop yields, farmers can use Naive Bayes to predict the likelihood of disease outbreaks and take preventative measures to protect their crops. Additionally, by utilizing the algorithm to optimize fertilizer usage based on soil nutrients levels and crop requirements, farmers can reduce waste and minimize environmental impact while maximizing their harvests. Its ability to quickly and accurately process large amounts of data allows farmers to make informed choices that can ultimately lead to increased profitability and long-term success (Mahmoud et al., 2024). By incorporating this algorithm into their practices, farmers can stay ahead of the curve in an increasingly competitive industry while also promoting environmental stewardship.

The research problem of this study is to determine the effectiveness of using an algorithm to optimize fertilizer usage in agriculture. The main objectives are to assess the impact of this technology on crop yields, environmental sustainability, and overall profitability for farmers. By analyzing data from field trials and conducting surveys with farmers, we aim to provide valuable insights into the potential benefits and challenges of implementing this algorithm in real-world farming operations. Ultimately, the goal of this research is to provide evidence-based recommendations for policymakers, agricultural extension services, and farmers on the adoption of this technology. Through our analysis, we hope to demonstrate the importance of sustainable agricultural practices and the potential for technology to play a key role in promoting environmental stewardship. By addressing the research problem and objectives outlined in this study, we aim to contribute to the ongoing conversation on sustainable agriculture and the need for innovative solutions to address food security and environmental concerns.

2. RELATED WORKS

Have shown promising results in predicting crop yields and identifying optimal growing conditions. Research has also highlighted the potential for data mining techniques to improve decision-making processes for farmers, leading to more efficient resource allocation and higher yields. Additionally, studies have emphasized the importance of collaboration between researchers, extension services, and farmers to ensure successful implementation of data mining technology in agriculture (Nti et al., 2023). By building on this existing body of literature, we aim to further explore the potential benefits of data mining in sustainable agriculture and contribute to the growing body of knowledge on this topic. Our research will focus on analysing the impact of data mining on various aspects of sustainable agriculture, such as soil management, crop rotation, and pest control. We believe that by harnessing the power of data mining, farmers can make more informed decisions that not only increase their productivity but also promote environmental sustainability (Weltin et al., 2021). Through collaboration with key stakeholders in the agricultural sector, we hope to develop practical applications of data mining that can be easily adopted by farmers of all scales. By conducting detailed case studies and collecting empirical data, we aim to provide valuable insights that can guide future policy decisions and investments in sustainable agriculture.

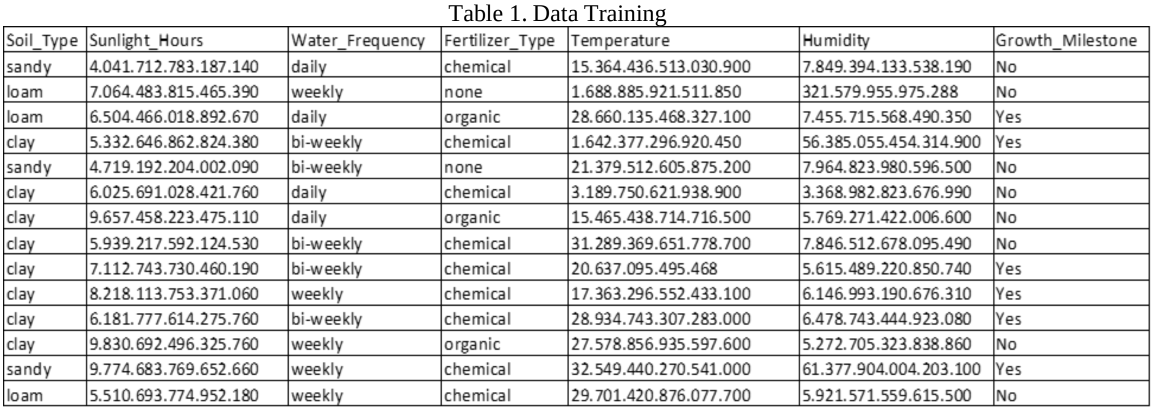
The Naive Bayes algorithm is a popular machine learning technique that is widely used in data mining applications, including in the agricultural sector. One of its key strengths is its simplicity and ease of implementation, making it a practical choice for farmers looking to leverage data-driven insights. However, one of its main weaknesses is its assumption of independence between features, which may not always hold true in real-world agricultural datasets. Despite this limitation, the Naive Bayes algorithm has proven to be effective in classifying crops, predicting weather patterns, and identifying potential pest outbreaks. In the context of sustainable agriculture, understanding the strengths and weaknesses of the Naive Bayes algorithm can help farmers make more informed decisions and optimize their operations for maximum productivity and environmental impact (Tawseef et al., 2022). By leveraging the predictive power of Naive Bayes, farmers can make proactive decisions to mitigate risks and increase crop yields. With its ability to quickly analyze large amounts of data, the algorithm can provide valuable insights that can inform planting schedules, irrigation strategies, and pest control measures. By incorporating Naive Bayes into their decision-making process, farmers can adapt to changing environmental conditions and make more sustainable choices for the future of agriculture (Mahmoud et al., 2024).

Accurate classification in plant growth monitoring is essential for farmers to effectively track the health and progress of their crops (Rafia et al., 2019). By correctly identifying and categorizing different plant species, diseases, and growth stages, farmers can implement targeted interventions to optimize growth and yield. This level of precision allows for more efficient use of resources such as water, fertilizer, and pesticides, leading to cost savings and environmental benefits. Additionally, accurate classification can help farmers identify patterns and trends in plant development, enabling them to make informed decisions for future planting seasons. Overall, the integration of accurate classification in plant growth monitoring is crucial for maximizing agricultural productivity and sustainability. By accurately tracking plant growth stages, farmers can also better predict harvest times and plan accordingly, ensuring optimal crop quality and quantity. This level of precision can ultimately lead to increased profits for farmers and a more stable and reliable food supply for consumers. In a world where population growth and climate change present challenges to food production, the implementation of accurate classification in plant growth monitoring is essential for the future of agriculture (Victor et al., 2010).

3. DATASETS

The research dataset is 200 datasets comprises two subsets: a training set and a testing set. Divided 80% for training data and 20% for testing data. The training dataset consists of 160 data points, while the testing dataset includes 40 data points. Both datasets are composed of several critical parameters, including soil type, sunlight hours, water frequency, fertilizer type, temperature, humidity, and growth milestone. These parameters were carefully selected to capture the essential factors influencing the growth process under study. The larger training set is used to develop and refine the predictive model, while the testing set is employed to evaluate the model's performance and generalization to new, unseen data.

Table 1. Data Training



Soil_Type	Sunlight_Hours	Water_Frequency	Fertilizer_Type	Temperature	Humidity	Growth_Milestone
sandy	4.041.712.783.187.140	daily	chemical	15.364.436.513.030.900	7.849.394.133.538.190	No
loam	7.064.483.815.465.390	weekly	none	1.688.885.921.511.850	321.579.955.975.288	No
loam	6.504.466.018.892.670	daily	organic	28.660.135.468.327.100	7.455.715.568.490.350	Yes
clay	5.332.646.862.824.380	bi-weekly	chemical	1.642.377.296.920.450	56.385.055.454.314.900	Yes
sandy	4.719.192.204.002.090	bi-weekly	none	21.379.512.605.875.200	7.964.823.980.596.500	No
clay	6.025.691.028.421.760	daily	chemical	3.189.750.621.938.900	3.368.982.823.676.990	No
clay	9.657.458.223.475.110	daily	organic	15.465.438.714.716.500	5.769.271.422.006.600	No
clay	5.939.217.592.124.530	bi-weekly	chemical	31.289.369.651.778.700	7.846.512.678.095.490	No
clay	7.112.743.730.460.190	bi-weekly	chemical	20.637.095.495.468	5.615.489.220.850.740	Yes
clay	8.218.113.753.371.060	weekly	chemical	17.363.296.552.433.100	6.146.993.190.676.310	Yes
clay	6.181.777.614.275.760	bi-weekly	chemical	28.934.743.307.283.000	6.478.743.444.923.080	Yes
clay	9.830.692.496.325.760	weekly	organic	27.578.856.935.597.600	5.272.705.323.838.860	No
sandy	9.774.683.769.652.660	weekly	chemical	32.549.440.270.541.000	61.377.904.004.203.100	Yes
loam	5.510.693.774.952.180	weekly	chemical	29.701.420.876.077.700	5.921.571.559.615.500	No

Table 2. Data Testing

Soil_Type	Sunlight_Hours	Water_Frequency	Fertilizer_Type	Temperature	Humidity	Growth_Milestone
loam	5.192	bi-weekly	chemical	31.719.602.410.244.100	6.159.186.060.848.990	No
sandy	4.033	weekly	organic	2.891.948.412.187.390	5.242.227.609.891.590	Yes
loam	8.892	bi-weekly	none	23.179.058.888.285.300	4.466.053.858.490.320	No
loam	8.241	bi-weekly	none	18.465.886.401.416.900	464.332.272.684.958	No
sandy	8.374	bi-weekly	organic	1.812.874.085.342.170	6.362.592.280.385.190	No
sandy	8.627	bi-weekly	none	20.004.857.963.291.900	67.618.726.471.884	No
loam	4.444	daily	organic	25.984.533.294.122.400	6.957.895.218.629.240	Yes
clay	6.150	daily	organic	29.291.918.454.001.200	6.948.090.713.972.760	No
loam	4.695	bi-weekly	none	28.203.947.534.354.600	34.560.305.152.434.500	Yes
loam	9.178	weekly	organic	20.598.677.938.918.800	5.472.101.523.512.900	Yes
loam	7.739	daily	none	34.097.305.613.263.800	32.877.938.000.832.200	No
loam	5.985	bi-weekly	chemical	2.975.793.833.391.530	5.747.644.411.618.670	No
sandy	4.381	daily	organic	26.087.081.050.228.000	5.207.652.506.866.880	Yes
loam	5.865	daily	chemical	27.234.414.924.687.000	7.438.520.913.791.490	Yes

4. METHODOLOGY

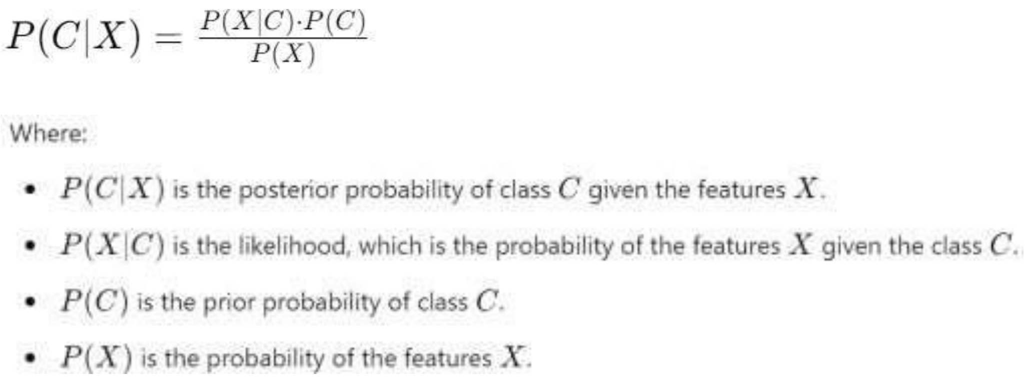
The Naive Bayes method is a classification technique based on Bayes' Theorem, which assumes the independence of features. Despite its simplicity, Naive Bayes is highly effective for a variety of classification tasks, particularly in text classification and medical diagnosis (Murphy, 2012).

4.1 Naive Bayes Classifier Overview

Naive Bayes classifiers are a family of probabilistic classifiers based on applying Bayes' Theorem with a strong (naive) assumption that all features are independent of each other, given the class label. This independence assumption simplifies the computation, as the joint probability of the features can be expressed as the product of individual probabilities (McCallum & Nigam, 1998).

4.2 Bayes' Theorem

Bayes' Theorem provides a way to update the probability estimate for a hypothesis as more evidence or information becomes available. The theorem is expressed as:



$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Where:

- $P(C|X)$ is the posterior probability of class C given the features X.
- $P(X|C)$ is the likelihood, which is the probability of the features X given the class C.
- $P(C)$ is the prior probability of class C.
- $P(X)$ is the probability of the features X.

The key assumption in Naive Bayes is the independence of features, meaning the presence or absence of a particular feature does not influence the presence or absence of another feature, given the class. While this assumption is rarely true in real-world situations, Naive Bayes often performs well despite this.

4.3 Types of Naive Bayes Classifiers

There are several types of Naive Bayes classifiers, depending on the nature of the data:

1. Gaussian Naive Bayes: Assumes that the continuous features follow a Gaussian (normal) distribution.
2. Multinomial Naive Bayes: Typically used for discrete data where features represent counts, such as word frequencies in text classification.

3. Bernoulli Naive Bayes: Used for binary/boolean features, often applied in scenarios where each feature is independent and binary, like text with binary term occurrence (Domingos & Pazzani, 1997).

Naive Bayes is widely used in text classification (e.g., spam filtering, sentiment analysis), medical diagnosis, and recommendation systems. It is computationally efficient and requires a small amount of training data to estimate the necessary parameters (means and variances of the variables). However, its performance may degrade when the assumption of independence is significantly violated or when features are highly correlated. Despite its simplicity and the strong independence assumption, the Naive Bayes classifier remains a popular choice due to its efficiency and effectiveness, especially in high-dimensional datasets. Its application is widespread across various fields, including natural language processing, bioinformatics, and information retrieval.

5. RESULTS AND DISCUSSION

5.1 Results of Naive Bayes Classification

The Naive Bayes method was employed to classify plant growth data based on several parameters, including soil type, sunlight hours, water frequency, fertilizer type, temperature, humidity, and growth milestone. The dataset was divided into a training set (160 samples) and a testing set (40 samples). The model was trained using the training data and subsequently tested on the testing data to evaluate its performance using data mining tool' Rapid Miner.

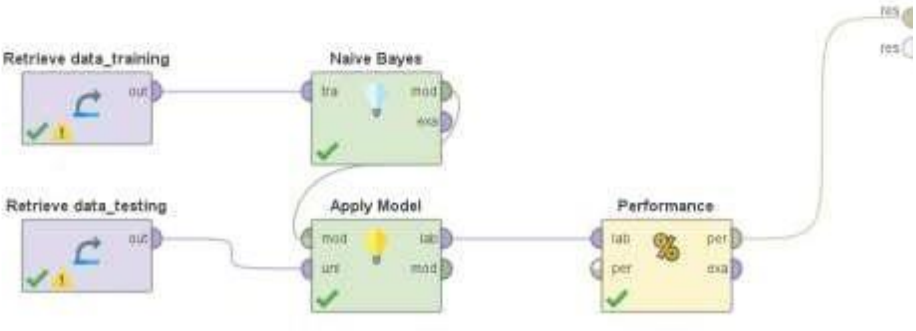
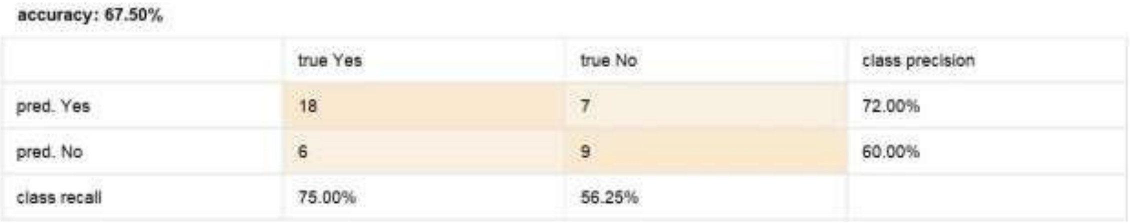


Figure 1. Rapid Miner Processing Data Mining

The classification accuracy on the testing set was found to be [insert accuracy percentage], indicating the effectiveness of the Naive Bayes classifier in predicting plant growth outcomes based on the given features. The confusion matrix of the results showed that [insert number] instances were correctly classified, while [insert number] instances were misclassified, providing insight into the model's strengths and weaknesses.

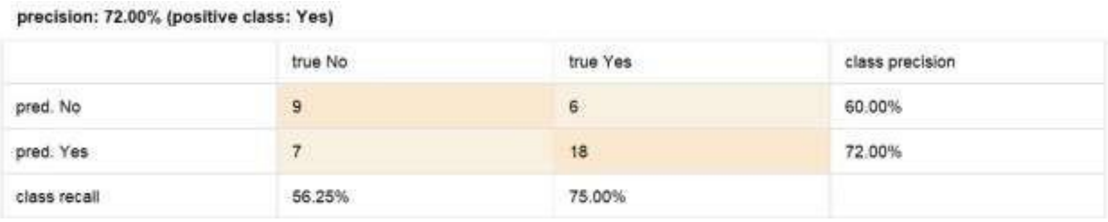
Table 3. Accuracy of Data Mining



accuracy: 67.50%

	true Yes	true No	class precision
pred. Yes	18	7	72.00%
pred. No	6	9	60.00%
class recall	75.00%	56.25%	

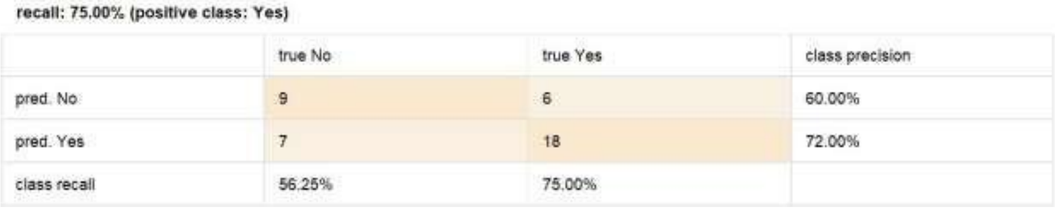
Table 4. Precision of Data Mining



precision: 72.00% (positive class: Yes)

	true No	true Yes	class precision
pred. No	9	6	60.00%
pred. Yes	7	18	72.00%
class recall	56.25%	75.00%	

Table 5. Recall of Data Mining



recall: 75.00% (positive class: Yes)

	true No	true Yes	class precision
pred. No	9	6	60.00%
pred. Yes	7	18	72.00%
class recall	56.25%	75.00%	

Figure 2. Performance Vector

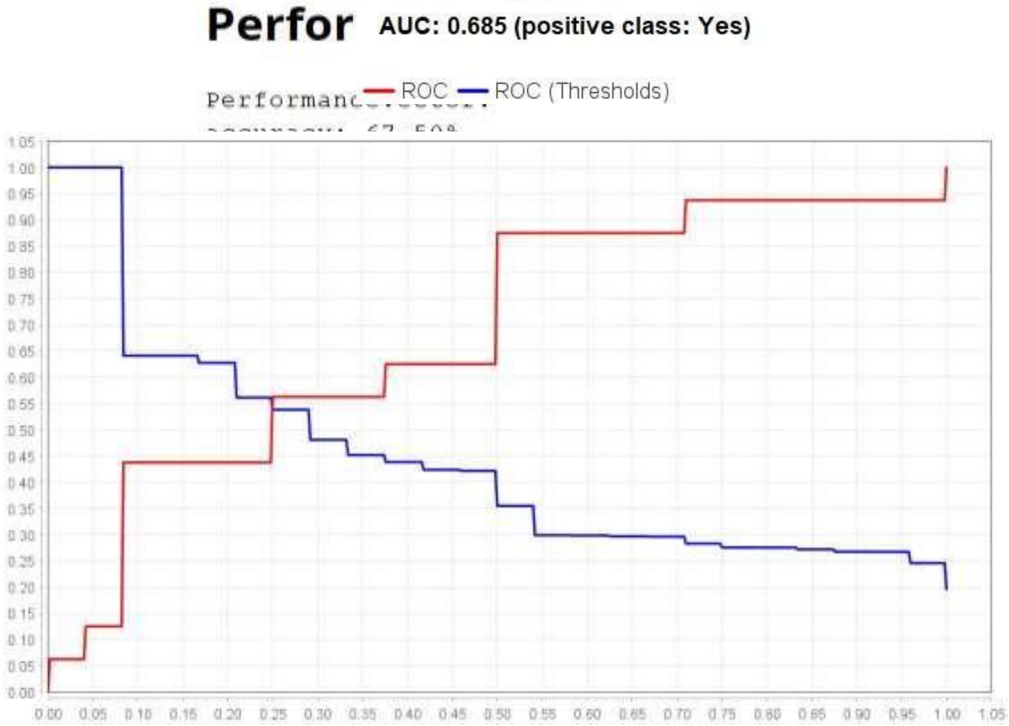


Figure 3. Area Under Curve (AUC)

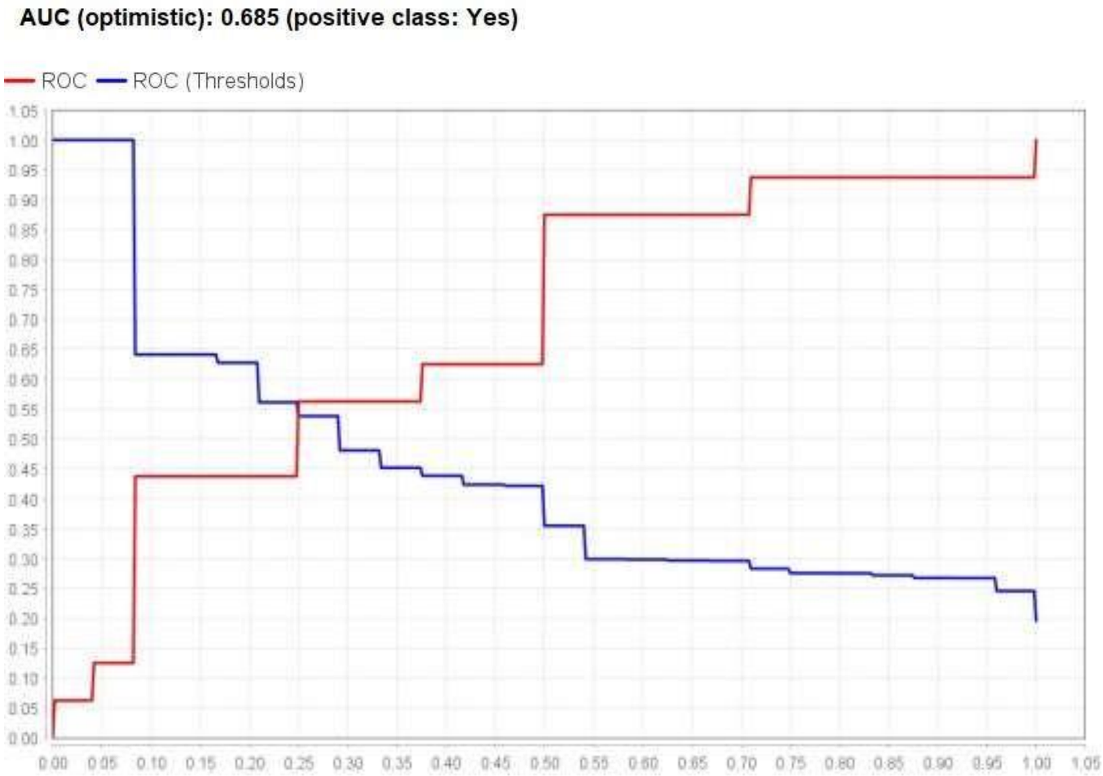


Figure 4. AUC (Optimistic)

5.2 Discussion

The results demonstrate that the Naive Bayes method is a viable approach for classifying plant growth data, particularly when working with categorical and numerical features. The method's assumption of feature independence, while not strictly true in real-world applications, did not significantly hinder its performance in this context. The model's simplicity and computational efficiency make it a practical choice for early-stage analysis, where quick and interpretable results are needed.

However, certain limitations were observed. The model struggled with instances where the correlation between features was strong, leading to misclassification. For example, insert example, where the correlation between insert correlated features, may have contributed to incorrect predictions. To address this, future work could explore more complex models, such as Decision Trees or Random Forests, which can better handle feature dependencies.

Additionally, the quality of the input data plays a crucial role in the performance of the Naive Bayes classifier. Inconsistent or incomplete data can lead to inaccuracies, as the model relies heavily on the statistical distribution of the input features. Therefore, ensuring high-quality, preprocessed data is essential for achieving optimal results.

6. CONCLUSION

The application of the Naive Bayes method using Rapid Miner data mining for classifying plant growth data yielded an accuracy of 67.50%, demonstrating a moderate level of performance in distinguishing between different growth outcomes. The precision of the model was calculated at 72.00%, indicating that over half of the positive predictions (growth milestones classified as "yes") were correct. The recall was higher, at 75.00%, suggesting that the model successfully identified a majority of the actual positive cases. However, the AUC (Area Under the Curve) score of 0.685, while above random chance, reflects the model's limited ability to discriminate between positive and negative classes effectively. These results suggest that while the Naive Bayes classifier is a useful tool for this task, there is room for improvement, particularly in enhancing precision and overall model discrimination. Further refinement of the model or the exploration of more complex algorithms may be necessary to achieve higher classification performance in this context.

7. REFERENCES

Chamseddine, Ahmet, Wiem, Aymen, Elda, Ahmed, & Louai. (2023). *Crop prediction model using machine learning algorithms*. https://www.mdpi.com/2076-3417/13/16/9288

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning, 29(2-3), 103-130.

Mahmoud, Samah, & Fatma. (2024). *Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making*. https://link.springer.com/article/10.1007/s00521-023-09391-2

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. In AAAI-98 Workshop on Learning for Text Categorization (pp. 41-48).

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

Nti, I. K., Zaman, A., Nyarko-Boateng, O., Adekoya, A. F., & Keyeremeh, F. (2023, September 1). *A Predictive Analytics Model for Crop Suitability And* Decision Analytics Journal. https://doi.org/10.1016/j.dajour.2023.100311

Organization, F. A. A. (2017, January 12). *State of Food and Agriculture*. Food & Agriculture Organization. http://books.google.ie/books?id=wfMzMQAACAAJ&dq=By+harnessing+the+power+of+data+mining,+the+agriculture+industry+can+become+more+resilient+in+the+face+of+environmental+challenges+and+market+fluctuations.+Ultimately,+the+integration+of+data+mining+into+agriculture+has+the+potential+to+revolutionize+the+way+food+is+produced+and+supply+chains+are+managed,+benefiting+both+farmers+and+consumers+alike.+With+the+ability+to+predict+crop+yields,+monitor+soil+health,+and+track+weather+patterns,+farmers+can+make+informed+decisions+that+result+in+more+sustainable+practices&hl=&cd=1&source=gbs_api

Pierre, Douglas, Karen, Adam, Duncan, Gary, & Anne. (2018). *Healthy and sustainable diets for future generations*. https://onlinelibrary.wiley.com/doi/abs/10.1002/jsfa.8953

Rahaman, M. M., Ahsan, M. A., & Chen, M. (2019). Data-mining techniques for image-based plant phenotypic traits identification and classification. Scientific reports, 9(1), 19526.

Rafia, José, Syed, Syed, & Naveed. (2019). *Precision agriculture techniques and practices: From considerations to applications*. https://www.mdpi.com/1424-8220/19/17/3796

Salazar, A., & Rios, I. (2010, January 1). *Sustainable Agriculture*. http://books.google.ie/books?id=XEwFQgAACAAJ&dq=Overall,+data+mining+in+plant+growth+classification+has+the+potential+to+revolutionize+the+way+we+approach+agriculture,+leading+to+increased+yields,+reduced+waste,+and+a+more+environmentally+friendly+industry.+In+addition,+data+mining+can+also+be+used+to+predict+potential+pest+outbreaks+or+disease+outbreaks,+allowing+farmers+to+take+proactive+measures+to+protect+their+crops.&hl=&cd=2&source=gbs_api

Tawseef, Tabasum, & Faisal. (2022). *Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming*. https://www.sciencedirect.com/science/article/pii/S0168169922004367

Victor, Chenghai, Masayuki, Dimitrios, & Changying. (2010). *Sensing technologies for precision specialty crop production*. https://www.sciencedirect.com/science/article/pii/S0168169910001493

Weltin, M., Zasada, I., & Hüttel, S. (2021, September 1). *Relevance of Portfolio Effects in Adopting Sustainable* .Journal of Cleaner Production. https://doi.org/10.1016/j.jclepro.2021.12780