EmAIT

http://publisher.uthm.edu.my/ojs/index.php/emait
e-ISSN : 2773-5540

Emerging Advances in Integrated Technology

Modelling of Ladyfinger Plantations Using Open Data Malaysia: A Comparative Study of Machine Learning Techniques

Mohamad Farhan Mohamad Mohsin[1,2*], Mohamad Ghozali Hassan[2], Kamal Imran Mohd Sharif[2], Mohd Azril Bin Ismail[2], Syairah Aimi Shahron[2], Noor Ashri Ja'afar[2]

[1]School of Computing,
Universiti Utara Malaysia, Kedah, MALAYSIA

[2]School of Technology Management and Logistics,
Universiti Utara Malaysia, Kedah, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/emait.2023.04.02.005
Received 10 October 2023; Accepted 13 December 2023; Available online 31 December 2023

Abstract: This study investigates the potential of using open data from Open Data Malaysia to develop classification models for agricultural practices specifically focusing on ladyfinger plantations. The integration of climate data with agricultural data is performed to build predictive models for crop yield prediction for sustainable agriculture. Four machine learning models, namely Naïve Bayes, SVM, KNN, and decision tree, are evaluated based on various performance metrics. The ladyfinger dataset was obtained from Open Data Malaysia containing both climate and agricultural data. was preprocessed and mined. The results indicate that the Naïve Bayes model achieves the highest performance making it the most suitable model for predicting ladyfinger yield. The decision tree model performed poorly and may not be suitable for this type of classification task. This study highlights two important findings. Firtstly, the inclusion of climate data significantly improved the classification performance of the models. Secondly the limited size of the ladyfinger dataset emphasizes the need for larger and more diverse datasets to enhance the accuracy and generalizability of predictive models in agriculture. Open data initiatives are important for providing researchers with data however larger and more diverse datasets are needed to improve model accuracy. Future research could investigate machine learning models for predicting crop yields in different crops with various climate and agricultural data combinations.

Keywords: Agriculture, classification, data mining, machine learning, prediction

1. Introduction

The agriculture sector plays a crucial role in Malaysia's economy and there is increasing interest in utilizing data mining techniques to improve agricultural productivity [1], [2]. Open data has the potential to facilitate such efforts by providing researchers and practitioners with access to diverse datasets including from various organizations. Open Data Malaysia is a key platform for sharing open data across various sectors in Malaysia [3]. It provides a wide range of datasets including agriculture such data on crop production, land use, and weather patterns.

In recent years, the Malaysian government has been actively promoting the use of open data in research and decision-making, launching the Malaysian Public Sector Open Data Portal in 2015, which later became integrated into the Open Data Malaysia platform [4]. As a result, the platform has recorded a significant number in users reaching a total of 1156339 in September2023. Currently, there are more than 12,500 datasets (780 related to agriculture) and 403

*Corresponding author: farhan@uum.edu.my

data providers available on the platform [5]. This is an indicator of the growing availability of open data in Malaysia including agriculture and climate research fields. The success story of this services can be seen in various practical application developed using the shared data such Mobile Trainer, My Transplant Diary, Kitar, OurAuthority, and more [6].

In agriculture, climate variables such as temperature, rainfall, and humidity have a significant impact on agricultural outcomes [7]. Integrating climate data with agricultural data can provide valuable insights into predicting crop yields and informing decision-making in the agriculture sector [8], [9]. Moreover, climate variability can affect the entire agricultural supply chain, resulting in disruptions in the availability and pricing of agricultural commodities [10]. Extreme weather disaster like as droughts and floods lower yields, reduce quality, and increase prices. From that, it challenges to policymakers and stakeholders to face the resilience of the agriculture sector towards climate change.

Despite the potential benefits of open data for agricultural research in Malaysia, it is relatively underutilized. It emphasizes the importance of further study that employs data mining modelling techniques. The availability of open data in machine-readable forms and frequently on an aggregated or annual basis format. However, the degree of information in aggregated or summarised data sometime may be insufficient for certain data mining analytical activities such as predictive modelling. Beside that, based on observations of Open Data Malaysia, particularly in relation to agriculture it has been noted that the data volume is inadequate which may not be sufficient for developing a predictive model. Therefore, the aim of the study is to examine the feasibility of developing agricultural predictive models utilizing open data from Open Data Malaysia. This study examines at the integration of climatic data with agricultural data from ladyfinger plantations from Open Data Malaysia. Data mining modeling techniques, such as support vector machine (SVM), decision tree (DT), k-nearest neighbors (KNN), and Naïve Bayes (NB), were employed to develop a predictive model that can be used to inform decision-making in the agriculture sector.

The paper begins with an introduction section, followed by a discussion of related work in the second section, which will cover topics such as open data in Malaysia, data mining and open data in agriculture, and machine learning for classification. Section 3 will describe how this study was conducted, and the results will be reported and discussed in sections 4 and 5, respectively. Finally, section 6 will provide concluding remarks for the study.

2. Related Work

In this section, the related work related to open data Malaysia, data mining and open data Malaysia in agriculture, and machine learning for classification will be discussed.

2.1 Open Data Malaysia

Open Data Malaysia is a government-led initiative launched in 2015. The aim is to promote transparency, accountability, and innovation in Malaysia by providing open data. The platform provides access to a diverse range of data categories, including Demographic, Environment, Transportation, Health, Education, Economic, Geographic, Government data and statistics, Energy, Agriculture and forestry, Finance and banking, Tourism, Social welfare and development, Culture and heritage, and Science and technology data. With such an extensive collection of data categories, Open Data Malaysia is an invaluable resource for researchers, policymakers, and individuals seeking to explore the Malaysian data landscape.

The platform's data is contributed by various government agencies such as the Department of Statistics Malaysia, the Ministry of Health Malaysia, the Malaysian Meteorological Department, and the Ministry of Transport Malaysia, among others. The agencies provide data in multiple formats, such as CSV, XLS, and JSON, making it easy for researchers and practitioners to analyze and utilize the data.

The agriculture and forestry sector are a crucial component of Open Data Malaysia's dataset collection, with 236 datasets as of March 2023. These datasets are grouped into eight categories, including Crop Production, Land Use, Livestock & Fisheries, Forestry Management, Soil and Climate, Agribusiness, Agro Ecosystem, and Plant & Animal Diseases as shown detail in Table 1. The platform's vast collection of agricultural data makes Open Data Malaysia a valuable resource for stakeholders seeking to advance agricultural research and innovation in Malaysia.

Table 1 - Agriculture & forestry categories in Open Data Malaysia

No	Categories	Description	Example
1	Crop production	production of crops	crop yields, planting patterns, and harvesting methods.
2	Land use	the use of land for agriculture and forestry	land tenure, land cover, and land use changes over time.
3	Livestock & fisheries	livestock and fisheries production	fish catch rates, aquaculture production figures, and livestock population numbers.
4	Forestry management	forestry management and conservation efforts	deforestation rates, forest fire incidence, and forest conservation initiatives.

5	Soil and climate	soil quality and climate patterns	soil moisture content, soil fertility levels, and climate change trends.
6	Agribusiness	business aspects of agriculture	market trends, trade volumes, and supply chain management.
7	Agro ecosystem	ecological aspects of agriculture	biodiversity, soil conservation, and sustainable agriculture practices.
8	Plant & animal diseases	spread and management of plant and animal diseases	disease incidence rates, vaccination programs, and pest control measures

2.2 Data Mining and Open Data Malaysia in Agriculture

The application of data mining techniques in agriculture has gained increasing attention in recent years [11], [12]. Data mining refers to the process of extracting useful information and patterns from large datasets [13]. The integration of open data in data mining can provide researchers with access to large and diverse datasets for analysis, enabling the development of predictive models to improve agricultural practices and outcomes [14]. Contradict to [15], citizens may choose not to utilize open data even when it is accessible on the portal. This reluctance can be attributed to a lack of awareness or the potential absence of an openness culture. Nonetheless, a study on the usage of Open Data Malaysia among academicians indicates that they believe utilizing Open Government Data will enhance their job performance, thus making them more motivated to intend to use it in including the agricultural research endeavors [16].

Climate data plays a critical role in agricultural practices and outcomes. Studies have shown that climate factors such as temperature, rainfall, and humidity have a significant impact on crop growth and yield [17], [18]. Therefore, integrating climate data from Open Data Malaysia with agricultural data can enable the development of predictive models for crop yield and growth. Moreover, the use of open data in agriculture can lead to more informed decision-making by policymakers and stakeholders where it enables more sustainable and effective agricultural practices [19]. The availability of open data on Open Data Malaysia has also led to increased transparency and accountability in the agricultural sector [4].

2.3 Machine Learning for Classification

Machine learning offers various classification models such as Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), and Naïve Bayes (NB.

2.3.1 Support Vector Machine (SVM)

SVM is a supervised learning approach for categorising data into two groups. To train the model that can predicts the category of a new data point, a set of data separated into two groups is used. SVMs are classified into two types: linear and non-linear. Linear SVM is appropriate for data that can be divided into two groups using a single straight line. The non-linear SVM is for data that requires a non-linear boundary.

2.3.2 Decision Tree (DT)

A DT is a supervised predictive modelling machine learning approach. It is a tree-like structure that predicts the category or value of the target variable. DT is like a flowchart with rectangles as primary nodes and ovals as terminal nodes. They can provide reliable forecasts with high-quality data for both continuous or regressive and categorical data.

2.3.3 K-Nearest Neighbors (KNN)

KNN is a widely used supervised machine learning technique for classification and regression. It calculates the chance of a data point belonging to a specific group by looking at the group of data points closest to it. KNN is a non-parametric method where it makes no assumptions about the underlying data distribution. KNN uses a voting mechanism to classify new observations with the class with the most votes among its nearest neighbours being chosen as the forecast.

2.3.4 Naive Bayes (NB)

NB is a classification approach that categorises things using Bayes' theorem. It assumes of strong or naive independence between data point properties. The Nave Bayes classifier assumes that all data point properties are independent of one another. Simple Bayes and independent Bayes classifiers are other names for NB classifiers. The term "naive Bayes classifier" refers to a class of machine learning algorithms that rely on statistical independence rather than a single methodology.

3. Methodology

This study employs data mining modelling techniques to investigate the potential of open data in agricultural research in Malaysia, specifically the integration of climate data from Open Data Malaysia with agricultural data related to ladyfinger plantations. As shown in Fig. 1, the methodology consists of four stages: data collection, data preparation for modelling, model development, and result analysis.

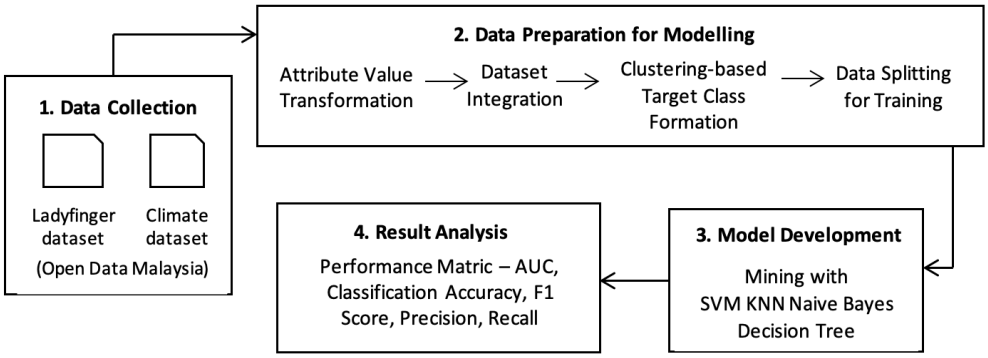

Fig. 1 - The four-stage methodology

3.1 Data Collection

Open Data Malaysia was used for this study particularly the agricultural and climate datasets. The agricultural dataset was contributed by the Ministry of Agriculture and Food Security (MAFS). This dataset contains a ten-year time series on ladyfinger plantations (2011-2020) including year, state, planting area (in hectares), harvested area (in hectares), and production (in metric tonnes). Meanwhile, the climate dataset from the National Hydraulic Research Institute of Malaysia (NAHRIM) represents daily mean rainfall (in millimeters) by state for six years (2014-2020). Fig. 2 illustrates sample data from both datasets.

State	Year	Planting area (hectares)	Harvested area (hectares)	Production (metric tonnes)
Johor	2011	340.8	390.1	10630.7
Kedah	2011	202	204.9	1834.8
Kelantan	2011	160.4	165.8	1564.9
Melaka	2011	119	119	3633.8
Negeri Sembilan	2011	386.7	405.8	4733.1
Pahang	2011	101.1	101.1	1639.6
Perak	2011	199.3	274.1	14186.8
Perlis	2011	27.6	27.6	156.4
Pulau Pinang	2011	172.3	176	1827.3

(a) ladyfinger

State	Year	Month	Day	Rainfall (mm)
Johor	2014	1	1	4.4
Johor	2014	1	2	21.38
Johor	2014	1	3	8.55
Johor	2014	1	4	1.58
Johor	2014	1	5	3.16
Johor	2014	1	6	1.97
Johor	2014	1	7	0.14
Johor	2014	1	8	0.16
Johor	2014	1	9	0.09

(b) climate

Fig. 2 - Ladyfinger (a) and climate; (b) dataset

To enable analysis, the next step involved merging and pre-processing both datasets through thorough data cleaning and transformation. The goal was to ensure that the data is in a consistent and accurate format, free from errors or inconsistencies that could affect the analysis's quality. This pre-processing step included checking for missing or duplicate data, standardizing variables, and transforming the data into the appropriate units or formats. After the datasets underwent cleaning and transformation, they were ready for analysis.

3.2 Data Preparation for Modelling

The first step in the analysis involved transforming the climate dataset to align it with the ladyfinger dataset. We converted the daily rainfall values to yearly values to allow for easier comparison and merging between the datasets. Since the climate dataset only covers data from 2014-2020, we selected this range for merging, resulting in a total of 77 records in the merged dataset, referred to as the ladyfinger dataset.

To prepare the dataset for classification, we created a new target class called Production Cluster by clustering the data into two groups, C1 and C2, using the K-Means algorithm. Each record was assigned a cluster number representing its target class, with 30% of the dataset labelled as C1 and the remaining records labelled as C2. This clustering process enabled us to classify the ladyfinger plantations into two distinct production clusters based on their attributes.

The final step in this phase was to split the ladyfinger dataset into a 70% training set and a 30% testing set. The training set was used to train and optimize the classification model, while the testing set was used to evaluate the model's performance on unseen data. This approach ensured that the model's performance was not over-optimized to the training set and was generalizable to new data. The resulting pre-processed dataset is now ready for model training and evaluation, and the next phase of the analysis can proceed. Fig. 3 shows the resulting pre-processed dataset.

State	Year	Planting area (hectares)	Harvested area (hectares)	Production (metric tonnes)	Rainfall (mm)	Production Cluster
Johor	2011	340.8	390.1	10630.7	7.07282192	C1
Kedah	2011	202	204.9	1834.8	6.44739726	C1
Kelantan	2011	160.4	165.8	1564.9	6.72578082	C1
Melaka	2011	119	119	3633.8	5.75079452	C1
Negeri Sembilan	2011	386.7	405.8	4733.1	5.74454795	C2
Pahang	2011	101.1	101.1	1639.6	6.24068493	C1
Perak	2011	199.3	274.1	14186.8	6.48183562	C1
Perlis	2011	27.6	27.6	156.4	5.78961644	C1
Pulau Pinang	2011	172.3	176	1827.3	5.6459726	C2

Fig. 3 - The pre-processed dataset

3.3 Model Development

To evaluate the performance of the machine learning models, various performance metrics such as AUC, CA, precision, recall, and F1-score will be recorded. These metrics provide a quantitative assessment of the models' accuracy and effectiveness in predicting the target class [20].

The Area Under Curve (AUC) metric assesses a classifier's ability to distinguish between classes. It is sometimes referred to as the True Positive Rate (TPR) versus False Positive Rate (FPR) plot. AUC measures overall performance across all available categorization criteria. It can be thought of as the odds that the model will rate a random positive example higher than a random negative example.

Classification accuracy (CA) is a statistic used to assess the performance of classification models. It is defined as the percentage of true predictions made by the model. A higher percentage of accuracy indicates a better model performance. However, accuracy is only a good indicator when the quantities of false positives and false negatives in the datasets are about equal in size. Accuracy is calculated by dividing the number of correct predictions by the total number of forecasts. It can be calculated in binary classification as (TP+TN)/(TP+TN+FP+FN), where TP stands for True Positive, TN stands for True Negative, FP stands for False Positive, and FN stands for False Negative.

Precision is a measure of a machine learning model's ability to properly predict positive observations. It is the ratio of correctly predicted positive observations to all positively anticipated observations. Precision can be estimated using the formula TP/(TP+FP).

The model's ability to detect positive samples is measured by recall or sensitivity. It is the proportion of correctly predicted positive observations to all actual class observations. The greater the recall, the greater the number of positive samples. Recall can be estimated using the formula TP/(TP+FN).

The F1-score combines a classifier's precision and recall into a single metric by computing their harmonic means. Its primary goal is to compare the efficacy of two classifiers. For example, if classifier A has higher precision and classifier B has higher recall, their F1 Scores can be compared to see which classifier produces better results. The F1-score is usually more advantageous than accuracy, especially if your class distribution is skewed. When the cost of false positive and false negative is significantly different, it is desirable to incorporate both precision and recall. F1-score is equal to (2*recall*precision)/(recall+precision).

4. Experiment Results

In this section, we analyze the results obtained from the application of four machine learning models (SVM, DT, KNN, and NB) to the ladyfinger data. The experiments were conducted in two phases. Firstl experiment was using only agricultural data without rainfall information. Second is included climatic factors. We evaluated the performance of

each model using various metrics such as AUC, accuracy, F1 score, precision, and recall to provide insight into each model's effectiveness and ability to predict and classify outcomes related to ladyfinger plantations. Table 2 displays the results obtained by applying the four machine learning models to the ladyfinger data without considering rainfall information.

Table 2 - Comparison of four machine learning models on Ladyfinger Data without Rainfall Information

Model	AUC	CA	F1	PRECISION	RECALL
SVM	74.5%	73.9%	62.8%	54.6%	73.9%
Decision Tree	48.0%	65.2%	65.2%	65.2%	65.2%
K-NN	73.0%	73.9%	75.0%	77.1%	73.9%
Naïve Bayes	79.4% win	78.3% win	79.4% win	83.1% win	78.3% win

The results presented in Table 2 indicate that the SVM and Naïve Bayes models performed the best achieving AUC values of 74.5% and 79.4%, respectively. The KNN model also achieved a decent AUC of 73.0% while the decision tree model underperformed with an AUC of only 48.0%. In terms of CA, SVM and KNN both achieved a CA of 73.9%, while the Naïve Bayes model was slightly lower at 78.3%. The decision tree model had the lowest CA at 65.2%.

The F1-score which balances precision and recall. It is a crucial measure for imbalanced datasets. In this case, the Naïve Bayes model achieved the highest F1-score of 79.4%, followed by KNN at 75.0%. SVM and the decision tree model both achieved an F1-score of 62.8% and 65.2%, respectively. Precision and recall measure how accurately each model predicted positive and negative samples, respectively. The Naïve Bayes model had the highest precision of 83.1%, indicating that the model accurately classified a high percentage of positive samples. The KNN model had the second-highest precision at 77.1%. The SVM and decision tree models had precision values of 54.6% and 65.2%, respectively. The recall values of the SVM and Naïve Bayes models were both high. This is an indicator that these models could detect most of the positive samples with recall values of 73.9% and 78.3%. KNN and the decision tree model had lower recall values of 73.9% and 65.2%

Overall, the results of the analysis suggest that Naïve Bayes is the most suitable model for predicting ladyfinger yield without rainfall information with the highest AUC, F1-score, and precision. KNN and SVM also performed relatively well with similar AUC and CA values but slightly lower precision and F1-score than the Naïve Bayes model. The decision tree model performed the worst and may not be suitable for this type of classification task.

Table 3 - Performance Comparison of Four ML Models on Ladyfinger Data with Rainfall Information

Model	AUC	CA	F1	PRECISION	RECALL
SVM	84.3% win	73.9%	62.8%	54.6%	73.9%
Decision Tree	44.1%	65.2%	58.4%	52.8%	65.2%
K-NN	70.1%	60.9%	59.6%	58.6%	60.9%
Naïve Bayes	78.4%	78.3% win	79.4% win	83.1% win	78.3% win

Table 3 displays the results obtained after including rainfall information in the analysis. The performance of the models varied significantly with SVM and Naïve Bayes outperforming K-NN and decision tree. SVM achieved the highest AUC score of 84.3. The result indicates that it has good predictive power and is capable of distinguishing between the positive and negative classes. However, its F1 score remained the same as in Table 2 indicating that it had similar precision and recall rates. The decision tree had the lowest AUC score of 44.1%, indicating that it was the least effective in distinguishing between the two classes. Its F1 score was also lower than that of the other models.

K-NN's performance decreased after including rainfall information as seen from its lower AUC score of 70.1% and lower F1 score of 59.6%. This suggests that K-NN may not be the best choice for predicting ladyfinger yields when rainfall information is included in the analysis. Naïve Bayes achieved a high AUC score of 78.4% like the performance in Table 2. This indicates that it can effectively distinguish between the positive and negative classes. Its F1 score remained high, indicating that it had a good balance of precision and recall rates. Overall, these results suggest that including rainfall information improved the performance of some models (SVM and Naïve Bayes) but negatively affected others (K-NN and Decision Tree) in predicting ladyfinger yields.

5. Discussion of the Result

The results of our experiments indicate two important findings with regards to the predictor variable and number of datasets used in machine learning models for predicting ladyfinger plantations.

- Firstly, our analysis demonstrates that the inclusion of rainfall information had a significant effect on the classification performance of the machine learning models. The results from Table 2 show that the SVM and Naïve Bayes models performed better than the K-NN and Tree models when rainfall information was included.

Specifically, SVM showed the highest AUC of 84.3% and Naïve Bayes achieved the highest F1 score of 79.4%. However, it is worth noting that the overall classification accuracy decreased for all models after including rainfall information. This is a sign that this factor may introduce additional complexity to the prediction task. Therefore, the trade-off between increased accuracy and increased complexity must be considered when selecting the appropriate model for a given application. In conclusion, it is suggested that the performance of machine learning models for predicting ladyfinger plantations can be improved by including climate factors such as rainfall information. However, the inclusion of additional variables must be balanced against the resulting increase in model complexity.

- Secondly, our study revealed that the inclusion of climate data had a significant effect on the classification performance of machine learning models applied to the ladyfinger dataset. The accuracy of the models improved by up to 13.2% with the addition of rainfall information. However, the models' overall performance remained below the recommended threshold of 90% for real-world implementation, which could be attributed to the limited size of the dataset.

The ladyfinger dataset used in this study only consisted of 77 records. This highlights the need for larger and more diverse datasets to enhance the accuracy and generalizability of machine learning models in agriculture. Open data sources such as the datasets provided by Open Data Malaysia can be useful for researchers to develop predictive models for crop production. However, data availability in summarized format such as yearly data can limit the ability to perform in-depth data mining analysis. Therefore, it is necessary to ensure that data is collected in a format suitable for machine learning analysis to improve the accuracy and usefulness of predictive models.

Numerous studies have emphasized the importance of larger and more diverse datasets to improve the accuracy of machine learning models in agriculture [21], [22]. Furthermore, other studies have highlighted the need for more precise and detailed climate data to enhance the performance of predictive models for crop production [23], [24]. Our study supports these findings and emphasizes the need for continued efforts to collect and provide data suitable for machine learning analysis to improve crop yield prediction and management. In conclusion, the finding highlights the importance of climate data in predicting crop production and the need for larger and more diverse datasets to enhance the accuracy and generalizability of machine learning models in agriculture. Open data initiatives have a significant role in providing data for researchers however the collection of data in a format suitable for machine learning analysis is crucial to improving the accuracy and usefulness of predictive models.

Additionally, when primary or open data sources are limited or unavailable, the use of synthetic data can be a useful approach to overcome this limitation. Synthetic data refers to artificially generated data that resembles real-world data and can be used to supplement or replace real data in machine learning models. Synthetic data can be generated using various techniques such as generative adversarial networks (GANs), variational autoencoders (VAEs), or other statistical models. In the context of predicting ladyfinger plantations, the generation of synthetic data could potentially expand the size and diversity of the dataset, which could improve the accuracy and generalizability of machine learning models. However, it is important to note that the quality and relevance of the synthetic data depend on the quality of the underlying statistical model and the accuracy of the assumptions made during the data generation process.

Several studies have shown the potential benefits of using synthetic data in machine learning models for various applications, including agriculture [25], [26]. However, the use of synthetic data is still a relatively new and developing field, and further research is needed to explore its full potential and limitations in the context of predicting crop production. Therefore, the use of synthetic data could be considered as an alternative solution for researchers and organizations that have limited access to primary or open data sources. However, caution must be exercised when using synthetic data, and the quality and relevance of the generated data should be assessed carefully to ensure its suitability for machine learning models.

6. Conclusion

In conclusion, this study has demonstrated the potential of using open data from Open Data Malaysia to develop predictive or classification models for agricultural practices and outcomes. Specifically, we investigated the integration of climate data with agricultural data related to ladyfinger plantations using four machine learning models. Our results indicate that the Naïve Bayes model is the most suitable model for predicting ladyfinger yield without rainfall information, achieving the highest AUC, F1-score, and precision. SVM and KNN also performed relatively well, albeit with slightly lower precision and F1-score than the Naïve Bayes model. However, the decision tree model performed poorly and may not be suitable for this type of classification task.

Our study has also highlighted two important findings for predicting ladyfinger plantations using machine learning models. Firstly, the inclusion of climate data, specifically rainfall information, significantly improved the classification performance of the models. Secondly, the limited size of the ladyfinger dataset emphasizes the need for larger and more diverse datasets to enhance the accuracy and generalizability of predictive models in agriculture. Open data initiatives

can play a significant role in providing data for researchers; however, data collection in a suitable format for machine learning analysis is crucial to improving the accuracy and usefulness of predictive models.

In addition to the promising results of our study, it is important to note that the limited size of the ladyfinger dataset underscores the need for larger and more diverse datasets to improve the accuracy and generalizability of predictive models in agriculture. Open data initiatives can play a significant role in providing data for researchers, but efforts to collect data in a suitable format for machine learning analysis are crucial for advancing the field.

Our findings are consistent with previous research that emphasizes the importance of larger and more precise climate data and datasets to improve crop yield prediction and management. Ultimately, our study can guide researchers and practitioners in selecting appropriate models and data sources for crop production prediction and management. Future research could explore the effectiveness of machine learning models in predicting crop yields for other crops, using different combinations of climate and agricultural data.

Acknowledgement

The authors of this article gratefully acknowledge the support provided by the Ministry of Higher Education (MoHE) through the Fundamental Research Grant Scheme (Ref: FRGS/1/2021/SS02/UUM/02/1 (S/O Code: 20107)). However, the views expressed in this article are those of the authors alone and do not necessarily reflect the official position of the MoHE, Malaysia.

References

[1] B. B. Bakar, “The Malaysian agricultural industry in the new millennium: issues and challenges,” in *International Conference on Malaysia: Malaysia in Global Perspective*, 2009, pp. 337-356.

[2] O. A. Chukwukere and A. H. Baharuddin, “Risk and poverty in agriculture: expanding roles for agricultural cooperatives in Malaysia,” *Geogr. Malaysian J. Soc. Sp.*, vol. 8, no. 4, pp. 1-11, 2012.

[3] “MyGOV - Open Government Data | Policy, Strategy and Governance | Open Data Guidelines and Data Stewards,” *The Malaysian Administrative Modernsation and Management Planning Unit*. https://www.malaysia.gov.my/portal/content/30024

[4] A. A. Ahmad Shaharudin, “Open government data in Malaysia: Landscape, challenges and Aspirations,” 2021.

[5] Malaysia Open Data Portal, “Set Data - MAMPU,” 2018. https://archive.data.gov.my/data/ms_MY/dataset

[6] Malaysia Open Data Portal, “Produk Data Terbuka,” 2018. https://archive.data.gov.my/app/galeri

[7] J. Ochieng *et al.*, “Effects of climate variability and change on agricultural production: The case of small scale farmers in Kenya,” *NJAS - Wageningen J. Life Sci.*, vol. 77, no. 2016, pp. 71-78, 2016.

[8] Y. Zhang *et al.*, “Integrating climate prediction and regionalization into an agro-economic model to guide agricultural planning,” *Clim. Change*, vol. 158, pp. 435-451, Feb. 2020.

[9] M. U. Azizan and K. Hussin, “Understanding the pressure on agriculture land as a safeguard for food security in Malaysia,” *Int. J. Built Environ. Sustain.*, vol. 2, no. 4, pp. 278-283, 2015.

[10] C. M. Godde *et al.*, “Impacts of climate change on the livestock food supply chain; a review of the evidence,” *Glob. Food Sec.*, vol. 28, p. 100488, Mar. 2021.

[11] K. Aishwarya and M. Jabbar, “Data Mining Analysis for Precision Agriculture: A Comprehensive Survey,” *ECS Trans.*, vol. 107, no. 1, pp. 17769-17781, Apr. 2022.

[12] J. Majumdar *et al.*, “Analysis of agriculture data using data mining techniques: application of big data,” *J. Big Data*, vol. 4, no. 20, Dec. 2017.

[13] J. Ha *et al.*, *Data Mining: Concepts and Techniques*, 3rd Ed. Elsevier Inc., 2011.

[14] C. L. Philip Chen and C. Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” *Inf. Sci. (Ny).*, vol. 275, pp. 314-347, 2014.

[15] P. Bhunia, “Brief Look at Open Government Data in 6 ASEAN Countries,” *9th Annual Singapore OpenGov Leadership Forum 2024*, 2017. https://opengovasia.com/brief-look-at-open-government-data-in-6-asean-countries.

[16] N. Zainal *et al.*, “Intention to Use Open Government Data among Academics - Empirical Findings,” *Glob. Bus. Manag. Res. An Int. J.*, vol. 14, no. 1, pp. 185-193, 2022.

[17] Arifah *et al.*, “Climate change impacts and the rice farmers’ responses at irrigated upstream and downstream in Indonesia,” *Heliyon*, vol. 8, no. 12, p. e11923, 2022.

[18] C. A. Rama Rao *et al.*, “Impact of climate change on productivity of food crops: a sub-national level assessment for India,” *Environ. Res. Commun.*, vol. 4, no. 9, p. 095001, Sep. 2022.

[19] A. Joshi and V. Kaushika, “Big Data and Its Analytics in Agriculture,” in *Bioinformatics for agriculture: High-throughput approaches*, A. K. Upadhyay, R. Sowdhamini, and V. U. Patil, Eds. Springer, Singapore, 2021, pp. 71-83.

[20] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” *Int. J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37-63, Oct. 2011.

[21] B. I. Evstatiev and K. G. Gabrovska-Evstatieva, "A review on the methods for big data analysis in agriculture," in *Conference of Communications, Information, Electronic and Energy System (CIEES)*, 2021, vol. 1032, no. 1, p. 012053.
[22] A. Monteiro *et al.*, "Precision Agriculture for Crop and Livestock Farming-Brief Review," *Animals*, vol. 11, no. 8, p. 2345, 2021.
[23] W. Mupangwa *et al.*, "Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in Eastern and Southern Africa," *SN Appl. Sci.*, vol. 2, p. 952, 2020.
[24] E. Harsányi *et al.*, "Data Mining and Machine Learning Algorithms for Optimizing Maize Yield Forecasting in Central Europe," *Agronomy*, vol. 13, p. 1297, 2023.
[25] Y. Lu *et al.*, "Machine Learning for Synthetic Data Generation: A Review," *J. Latex Cl. Files*, vol. 14, no. 8, pp. 1-19, 2023.
[26] Nitin *et al.*, "Developing precision agriculture using data augmentation framework for automatic identification of castor insect pests," *Front. Plant Sci.*, vol. 14, p. 1101943, 2023.