COMPARATIVE STUDY ON AUSTRALIA WEATHER CLASSIFICATION

Mohamed Moubarak Mohamed Misbahou Mkouboi (P139575)

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi Selangor, Malaysia
p139575@siswa.ukm.edu.my

Abstract. One type of data analysis that extracts models representing important data classes is classification. In this study, we develop a classification model from an Australian weather dataset using four different classifier algorithms: Decision Tree, Neural Network, Naïve Bayes, and Support Vector Machine. Building a predictive classifier to predict the next day's rain that is described on the target variable RainTomorrow is our objective. A variety of evaluation indicators were used to assess each classification model. We also explained the discoveries obtained from the top-performing model.

Keywords: Predictive model, rain tomorrow, machine learning.

1. Introduction

Weather is a natural phenomenon that impacts our daily activities and living conditions. The rain, wind, and temperature in the atmosphere above the earth, particularly when they occur over a specific region which is Australia in this study and at a specific time[1]. The temperature is one of the major factors of weather conditions whether it's by rain or snow and so on. Daily activities can be ruined because of the weather conditions and even accidents can be occurred because of it. Additionally, it is impacting other sectors as well such as the ecosystems and its biodiversity. Research by [2] reported that the water temperature change may alter the metabolism and physiology of aquatic animals thereby affecting the growth, fecundity, feeding behaviour, distribution, migration and abundance of fish as well as other aquatic animals. Some related work shows that using machine learning methods could reduce major incidents caused by the weather.

Weather systems can benefit greatly from data mining by using data and analytics to find future weather forecasts and best practices that reduce costs and enhance daily operations. The fundamental difficulty lies in properly applying data mining techniques to navigate through the vast amount of available data and find relevant and crucial information. It is crucial to the progress and creation of new techniques that handle the massive amounts of data in the climatic fields. The primary step in data

mining is data pre-processing, which includes cleaning and transforming the original data. The pre-processed data will then be used in the data modelling step to create a prediction model. The selected classifier method and the training data have a major impact on the model's efficiency and accuracy. Furthermore, better understanding and insights from the data are provided by a well-constructed model. Consequently, it is essential to make sure the data is properly pre-processed and that a suitable classifier algorithm is used.

The Australian weather dataset was evaluated with four classifier algorithms in this study: Decision Tree, Neural Network Classification, Naïve Bayes, and Support Vector Machine. The algorithms were evaluated using Cross-validation and Percentage Split techniques to compare the prediction models' accuracy. The objective of this research is to predict the next day's rain using classifier algorithms and assess how effective the selected methods are.

2. Related Work

Machine learning for climate analysis and weather forecasting has advanced significantly in recent years. Several approaches and strategies have been investigated to increase the precision and dependability of weather forecasts. Here, we go over several noteworthy studies that have advanced this area.

Using the Naive Bayes Classifier (NBC), Azmi et al. (2021) studied rainfall prediction in Banyuwangi, the largest district in East Java[3]. Because of the region's particular climate, which is shaped by its long coastline, precise predictions of the weather are essential for both aviation and agriculture. The authors selected NBC because it works well with big, irrelevant datasets and used rainfall data that was divided into three categories: light, normal, and heavy. According to their research, NBC was able to predict rainfall with an impressive 96% accuracy rate, highlighting the platform's potential for accurate weather forecasting in areas with limited meteorological data.

In 2021, Purwandari et al. studied the use of social media data for weather prediction[4]. They aimed to use text mining methods and machine learning algorithms, such as Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Logistic Regression (LR), to classify meteorological conditions by examining Twitter data. With a 93% accuracy rate, their tests showed that SVM performed better than alternative techniques, indicating its applicability for text classification. This creative method improves the precision and timeliness of weather forecasts by utilising real-time social media updates.

By using machine learning methods to predict climate change, Karthikeyan et al. (2021) addressed the shortcomings in the simulation models used by the Intergovernmental Panel on Climate Change (IPCC)[5]. They assessed several algorithms, such as Naive Bayes and SVM, to enhance their comprehension and prediction of simulation model failures. According to their research, SVM offered a strong categorization framework for estimating the likelihood of simulation crashes, which helped create climate models with higher levels of accuracy. This study emphasises how crucial machine learning is to improve the capacity for complex geoscientific models to make predictions.

The application of data mining classification algorithms for predicting the weather has been studied by Kareem et al. in 2021[6]. They classified meteorological conditions like rain, fog, and cloudy days using neural networks, Naive Bayes, random forests, and k-nearest neighbour algorithms. They discovered that the random forest algorithm, with an accuracy of 89%, surpassed other models using synoptic data from Kaggle. This study shows how different data mining approaches can be used to create accurate weather prediction systems, which are crucial for industries like aviation and agriculture.

A review of the implementation of deep neural networks (DNN) for rainfall prediction was conducted by Naik et al. (2020)[7]. The study highlighted how effective DNNs are at handling difficult prediction tasks compared to more conventional machine learning models like SVM, random forests,

and decision trees. Using techniques like Adam and gradient descent to optimise model parameters, DNNs demonstrated enhanced performance in rainfall prediction. This paper highlights how deep learning techniques can be used to provide weather forecasts that are more reliable and accurate, which is important for many daily activities and planning.

Together, these papers demonstrate the progress made in machine learning methods for predicting rainfall. With their advantages and uses, the Naive Bayes classifier, SVM, neural networks, and deep learning models have all demonstrated a significant amount of ability to raise the precision of weather predictions. Combining sophisticated machine learning algorithms with social media data opens up new possibilities for accurate and timely weather forecasting in real-time, which is essential for many socio-economic activities.

3. Classification Methods

In the supervised machine learning process of classification, the model attempts to predict the correct label of a given input set of data[8]. Before being used to make predictions on new, and unseen data, the model in classification is fully trained using the training set and then evaluated using test data.

A total of 22 attributes and 154314 instances will be used for developing predictive models. And 21 attributes and 42677 instances will be used for testing the models. Table 1 shows the description of the final datasets used for this study.

Table 1: Description of the dataset

Attributes	Description	Data Type
Location	Name of the city from Australia	Nominal
MinTemp	The Minimum temperature during a particular day. (degree Celsius)	Numerical
MaxTemp	The maximum temperature during a particular day. (degree Celsius)	Numerical
Rainfall	Rainfall during a particular day. (millimeters)	Numerical
Evaporation	Evaporation during a particular day. (millimeters)	Numerical
Sunshine	Bright sunshine during a particular day. (hours)	Numerical
WindGusDir	The direction of the strongest gust during a particular day. (16 compass points)	Ordinal
WindGuSpeed	Speed of strongest gust during a particular day. (kilometers per hour)	Numerical
WindDir9am	The direction of the wind for 10 min prior to 9 am. (compass points)	Ordinal
WindDir3pm	The direction of the wind for 10 min prior to 3 pm. (compass points)	Ordinal
WindSpeed9am	Speed of the wind for 10 min prior to 9 am. (kilometers per hour)	Numerical
WindSpeed3pm	Speed of the wind for 10 min prior to 3 pm. (kilometers per hour)	Numerical
Humidity9am	The humidity of the wind at 9 am. (percent)	Numerical
Humidity3pm	The humidity of the wind at 3 pm. (percent)	Numerical
Pressure9am	Atmospheric pressure at 9 am. (hectopascals)	Numerical
Pressure3pm	Atmospheric pressure at 3 pm. (hectopascals)	Numerical
Cloud9am	Cloud-obscured portions of the sky at 9 am. (eighths)	Numerical
Cloud3pm	Cloud-obscured portions of the sky at 3 pm. (eighths)	Numerical
Temp9am	The temperature at 9 am. (degree Celsius)	Numerical

Temp3pm	The temperature at 3 pm. (degree Celsius)	Numerical
RainToday	If today is rainy then 'Yes'. If today is not rainy then 'No'	Binary
RainTomorrow	If tomorrow is rainy then 1 (Yes). If tomorrow is not rainy then 0 (No)	Binary

3.1. Decision Tree

A non-parametric supervised learning approach that is used for both regression and classification tasks is the decision tree[9]. With a root node, branches, internal nodes, and leaf nodes, it has a hierarchical tree structure.

3.2. Neural Network Classification

An artificial intelligence technique called a neural network trains a model to process information like that of the human brain[10]. Deep learning is a kind of machine learning technique that uses networked nodes or neurons arranged in a layered pattern to mimic the organisation of the human brain.

3.3. Naïve Bayes

The "naive" assumption of conditional independence between each pair of features, given the value of the class variable, is the foundation of supervised learning algorithms known as "naive Bayes" approaches[11]. According to Bayes' theorem, given a class variable (y) and a dependent feature vector x1 through xn, we can estimate P(y) and P(xi | y) using Maximum A Posteriori (MAP) estimation; the former represents the relative frequency of class (y) in the training set.

3.4. Support Vector Machine (SVM)

The support vector machine (SVM) is a machine learning algorithm that determines boundaries between data points based on predefined classes, labels, or outputs[12]. It uses supervised learning models to solve complex problems related to classification, regression, and outlier detection. SVMs are widely used in many sectors, including speech and image recognition, natural language processing, healthcare, and signal processing applications.

4. Modelling and Measurement Methods

4.1. Modelling Set-Up

To predict rainfall, we developed and evaluated several machine-learning models. The two primary methods used in the modelling setup were k-fold cross-validation and percentage split (train-test split).

1. Percentage Split (Train-Test Split):

   80% of the dataset was placed aside for training, and the remaining 20% was set aside for testing. This simple strategy ensures that the model is assessed on data that hasn't been seen before, giving a clear picture of how well it performs in real-life scenarios.

2. k-Fold Cross-Validation:

   10-fold cross-validation was used to verify the models further. With this method, the dataset is divided into 10 equal parts. The model is trained on 9 of the parts, and the remaining part is used for testing. Each portion is used as the test set exactly once during the 10 repetitions of this process. To produce a more reliable estimate of the model's performance, the data are then averaged. Because every data point is used for both training and testing, this approach reduces the chance of overfitting and offers a thorough assessment.

4.2. Measurement Metrics

To evaluate the models' performance, we applied multiple assessment indicators. These indicators offer several viewpoints regarding the forecast accuracy and efficacy of the models:

1. Accuracy:

The ratio of accurately predicted cases to total instances. It is the primary metric that provides a clear indication of the model's performance.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

2. Precision:

The ratio of accurately predicted positive instances to all predicted positive instances. It shows how many of the expected positive events were correct.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

3. Recall (Sensitivity):

The ratio of accurately predicted positive instances to actual positive instances in the dataset. It evaluates the model's ability to correctly identify positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

4. Confusion Matrix:

A table that illustrates the classification model's performance by presenting true positives, true negatives, false positives, and false negatives. It offers a complete overview of the model's performance and errors.

5. AUC (Area Under the Curve):

The area under the ROC curve is a single value that summarises the model's ability to differentiate between positive and negative cases. Higher AUC values indicate improved model performance.

6. ROC (Receiver Operating Characteristic) Curve:

A graphical illustration of the model's performance at various threshold values. It plots the true positive rate (recall) versus the false positive rate to demonstrate the relationship between sensitivity and specificity.

4.3. Evaluation Procedure

We did the following analysis for each model:

1. Training:

Train the model using the training dataset (80% for percentage split, 90% for cross-validation).

2. Prediction:

Use the trained model to predict the training data (for performance evaluation) and test data (for final evaluation).

3. Performance Metrics Calculation:

Calculate the accuracy, precision, and recall, then plot the confusion matrix, ROC curve, and AUC score.

4. Model Comparison:

Using the obtained metrics, evaluate the performance of various models and training approaches (% split vs. cross-validation).

Using these modelling and measuring methodologies, we ensured a complete evaluation of each model's performance, providing insights into its strengths and weaknesses in predicting rainfall.

5. Results and Discussion

This section includes the experimental findings and analyses of the many classification models used to predict rainfall. The following models are compared: Neural Network, Naive Bayes, Support Vector Machine (SVM), and Decision Tree. Accuracy, precision, recall, AUC, and ROC curves were used as evaluation measures. The appendices provide detailed results, but this section presents a summary of significant findings along with relevant tables and figures.

5.1. Results Visualisation

5.1.1. Comparison of Classification Methods

The performance of each model was evaluated using the percentage split (train-test split) and k-fold cross-validation methods. Table 2 summarises the accuracy results for all models:

Table 2: Results of all models

Model	Accuracy (Percentage Split)	Accuracy (Cross-Validation)
Neural Network	0.8841	0.4425
Naive Bayes	0.6759	0.1181
SVM	0.8637	0.7636
Decision Tree	0.9999	0.5466

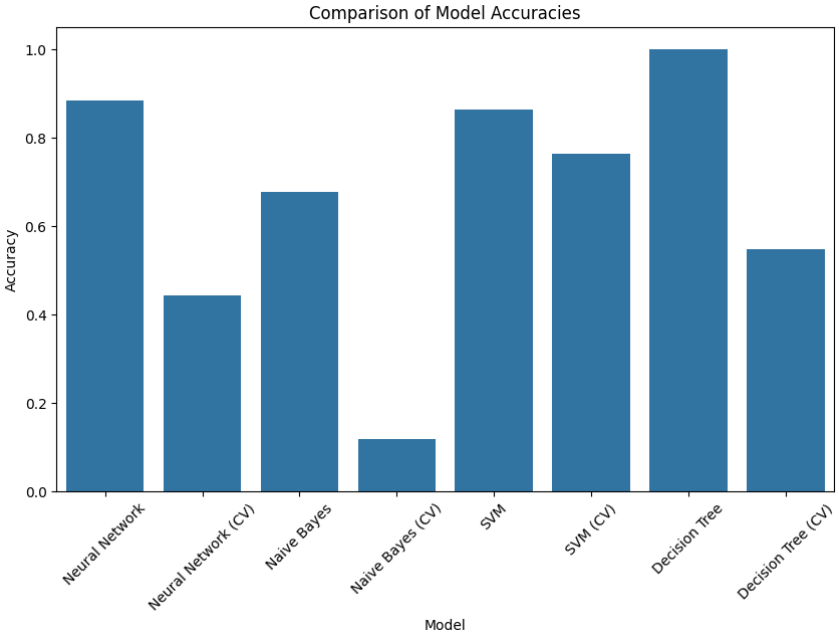


Figure 1. Plot of Model Accuracies

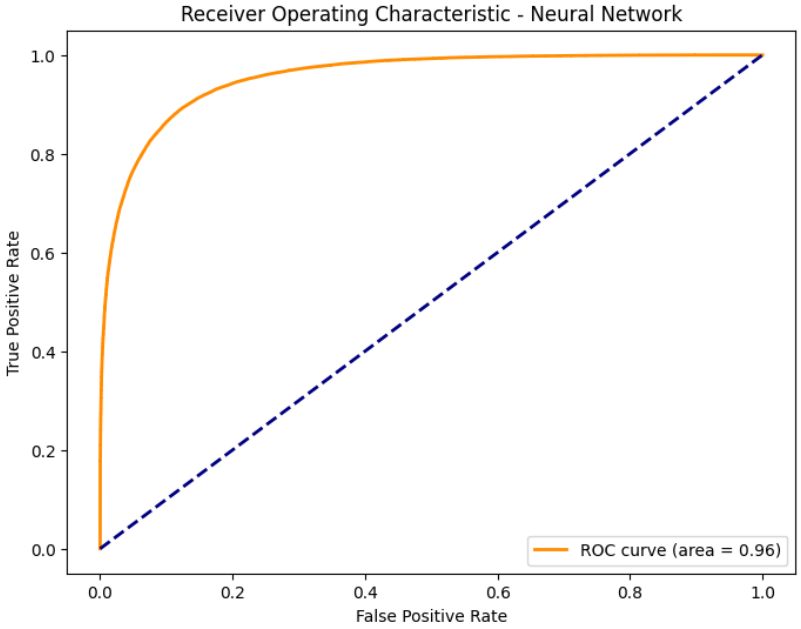


Figure 2. Plot of ROC and AUC of Neural Network (Percentage Split)

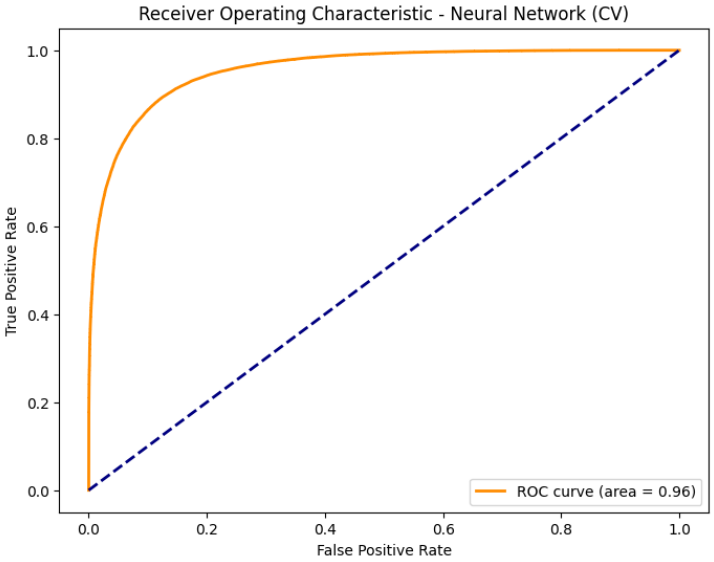


Figure 3. Plot of ROC and AUC of Neural Network (Cross-Validation)

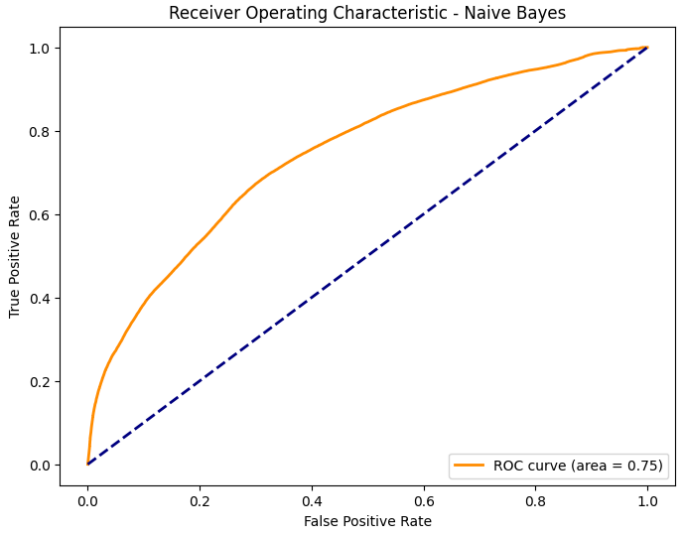


Figure 5. Plot of ROC and AUC of Naïve Bayes (Percentage Split)

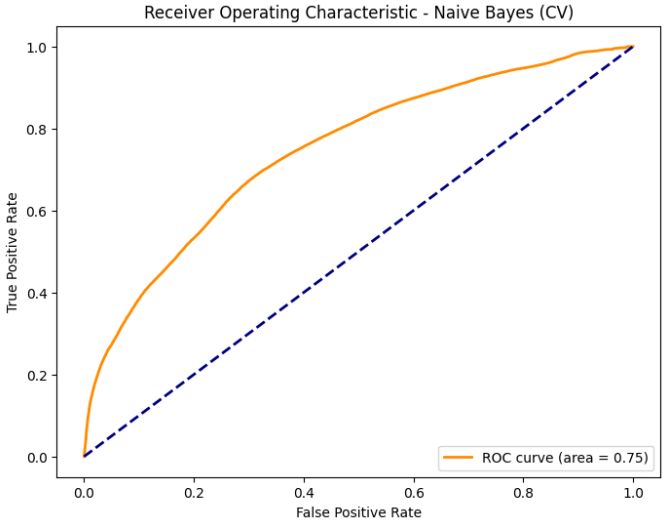


Figure 6. Plot of ROC and AUC of Naïve Bayes (Cross-Validation)

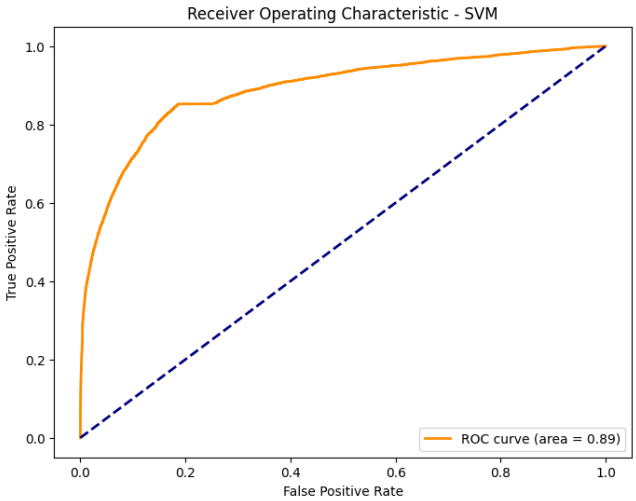


Figure 7. Plot of ROC and AUC of SVM (Percentage Split)

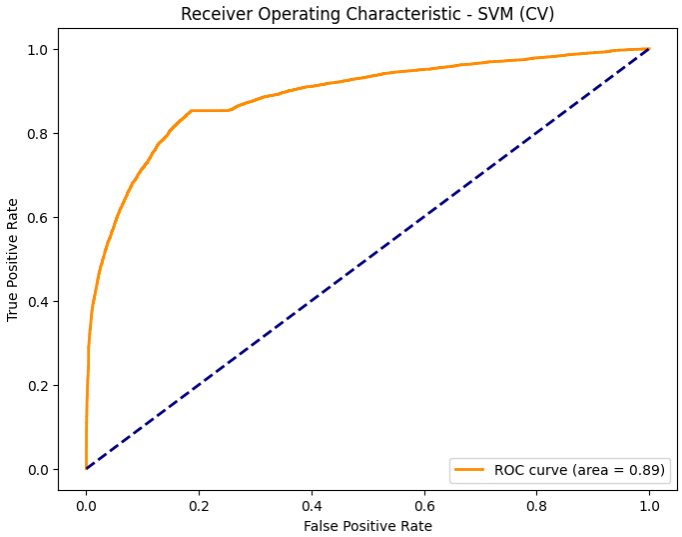


Figure 8. Plot of ROC and AUC of SVM (Cross-Validation)

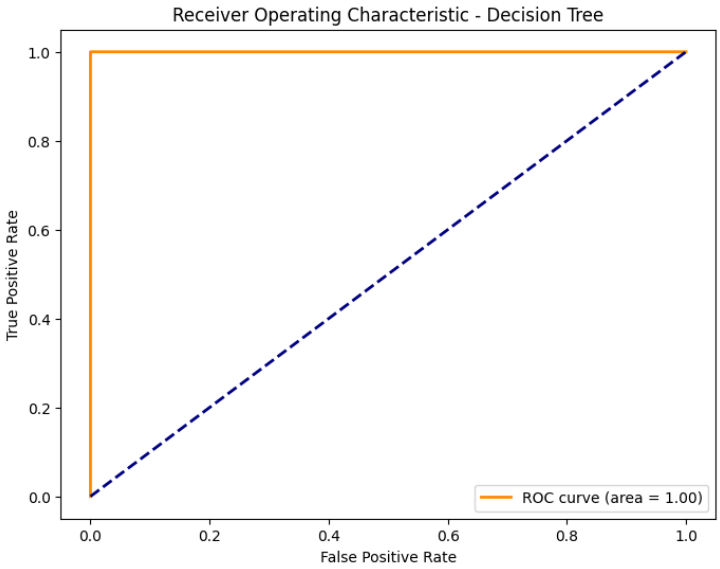


Figure 9. Plot of ROC and AUC Decision Tree (Percentage Split)

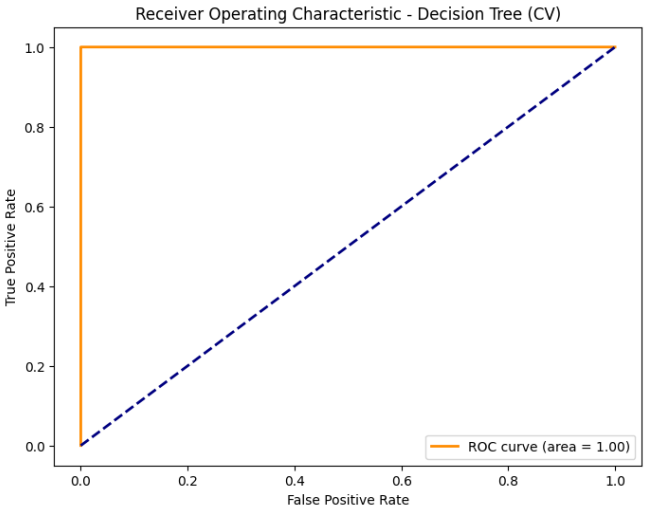


Figure 10. Plot of ROC and AUC Decision Tree (Cross-Validation)

5.1.2. Analysis and Best Technique

The performance of the models was determined by evaluating each metric to determine the most effective model for predicting rainfall. Based on the observations below, SVM appeared as the best model:

- Accuracy: SVM achieved great accuracy in both percentage split (0.8637) and cross-validation (0.7636), indicating good generalizability.
- Robustness: Consistent performance across multiple evaluation methodologies demonstrated SVM's robustness.
- Interestingness: Cross-validation resulted in a considerable decline in performance for neural networks and Naive Bayes, demonstrating the necessity for strict evaluation methods to detect overfitting.
- Error Analysis: Decision Tree had minimal misclassifications in the training set but suffered with generalisation, indicating overfitting. SVM balanced true positive and true negative rates.
- Speed and Scalability: Neural Networks and Naive Bayes were fairly quick. However, SVM training took longer due to their complexity, and we had to compress the dataset by 20% due to the large dataset we used. With huge datasets, decision trees struggled to scale.
- Interpretability: Decision Trees gave good interpretability, SVM delivered strong performance but had lower interpretability, Neural Networks were generally less interpretable, and Naive Bayes lacked the accuracy required for this task.

Therefore, SVM demonstrated to be the best model for rainfall prediction, balancing high accuracy, robustness, and generalizability.

5.2. Knowledge Analysis

This section provides a detailed study of the insights derived from the classification models' performance:

- Feature Importance: Decision Trees contributed to identifying the most relevant features influencing rainfall prediction, which can then be utilised to modify the dataset and enhance model performance.

- Model Generalizability: The cross-validation findings demonstrated the need to assess models on various data splits to minimise overfitting and ensure generalizability.
- Model Selection: SVM was identified as the most dependable model due to its consistent performance across many evaluation methodologies, emphasising the importance of endurance and accuracy in model selection.
- Practical Implications: SVM and Decision Trees' high accuracy suggests that they have real-world applications in rainfall prediction, which can help with agriculture, emergency management, and everyday planning.
- Areas for Improvement: The considerable decline in the performance of Neural Networks and Naive Bayes during cross-validation suggests that more tuning is required, as well as potentially more complex feature engineering or selection approaches.

Overall, the results show that, while multiple models can attain high accuracy, robustness and generalizability are critical for practical applications. SVM is the best overall model for this problem since it balances accuracy, robustness, and practical applicability. This investigation's results can help drive future attempts to improve rainfall prediction models and use them in real-world scenarios.

Conclusion and Suggestion

In conclusion, our study studied the value of various machine learning models for rainfall prediction, and Support Vector Machine (SVM) appeared as the most promising technique due to its high accuracy, robustness, and generalizability. Despite SVM's dominance, our study revealed crucial insights and opportunities for development. These include data limits that affect model performance, issues related to model complexity and interpretability trade-offs, and worries about overfitting. To solve these challenges, future research could concentrate on improved data-collecting methods, better feature engineering techniques, the study of ensemble learning methodologies, and the creation of hybrid models that combine the characteristics of various algorithms. Overall, while our work provides useful insights, more research and refinement are required to address the uncovered errors and improve the effectiveness of rainfall prediction systems.

References

[1] "WEATHER | English meaning - Cambridge Dictionary." Accessed: May 24, 2024. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/weather

[2] S. Prakash, "IMPACT OF CLIMATE CHANGE ON AQUATIC ECOSYSTEM AND ITS BIODIVERSITY: AN OVERVIEW," *International Journal Biological Innovations*, vol. 03, no. 02, 2021, doi: 10.46505/ijbi.2021.3210.

[3] A. U. Azmi, A. F. Hadi, D. Anggraeni, and A. Riski, "Naive bayes methods for rainfall prediction classification in Banyuwangi," *J Phys Conf Ser*, vol. 1872, no. 1, p. 012028, May 2021, doi: 10.1088/1742-6596/1872/1/012028.

[4] K. Purwandari, J. W. C. Sigalingging, T. W. Cenggoro, and B. Pardamean, "Multi-class Weather Forecasting from Twitter Using Machine Learning Aprroaches," *Procedia Comput Sci*, vol. 179, pp. 47–54, Jan. 2021, doi: 10.1016/J.PROCS.2020.12.006.

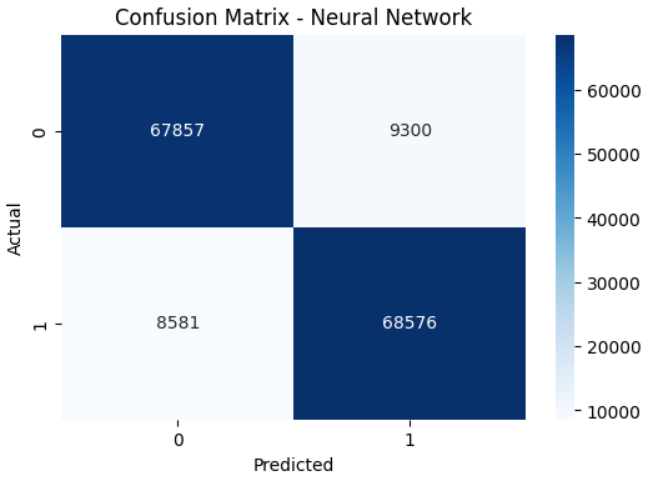
[5] C. Karthikeyan, G. Sunitha, J. Avanija, K. Reddy Madhavi, and E. S. Madhan, "Prediction of Climate Change using SVM and Naïve Bayes Machine Learning Algorithms," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 2, pp. 2134-2139–2134 – 2139, Apr. 2021, Accessed: Jun. 07, 2024. [Online]. Available: https://www.turcomat.org/index.php/turkbilmat/article/view/1856

[6] F. Q. Kareem, A. M. Abdulazeez, and D. A. Hasan, "Predicting Weather Forecasting State Based on Data Mining Classification Algorithms," *Asian Journal of Research in Computer Science*, pp. 13–24, Jun. 2021, doi: 10.9734/ajrcos/2021/v9i330222.

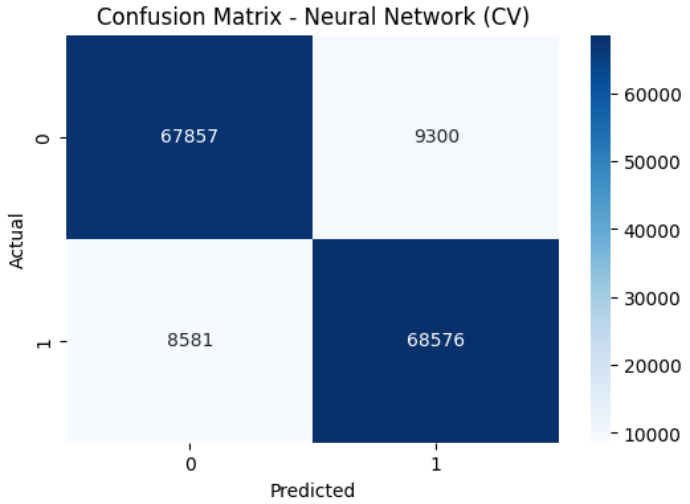
[7] A. R. Naik, A. V. Deorankar, and P. B. Ambhore, “Rainfall Prediction based on Deep Neural Network: A Review,” *2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 - Conference Proceedings*, pp. 98–101, Mar. 2020, doi: 10.1109/ICIMIA48430.2020.9074892.
[8] “Classification in Machine Learning: A Guide for Beginners | DataCamp.” Accessed: Jun. 08, 2024. [Online]. Available: https://www.datacamp.com/blog/classification-machine-learning
[9] “What is a Decision Tree? | IBM.” Accessed: Jun. 08, 2024. [Online]. Available: https://www.ibm.com/topics/decision-trees
[10] “What is a Neural Network? - Artificial Neural Network Explained - AWS.” Accessed: Jun. 08, 2024. [Online]. Available: https://aws.amazon.com/what-is/neural-network/
[11] “1.9. Naive Bayes — scikit-learn 1.5.0 documentation.” Accessed: Jun. 08, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html
[12] “All You Need to Know About Support Vector Machines.” Accessed: Jun. 08, 2024. [Online]. Available: https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/

Appendix: Detailed Results

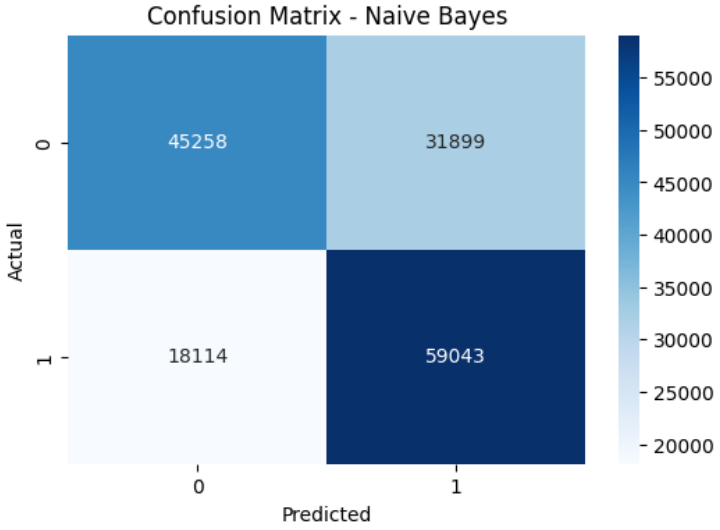
1. Confusion matrices for each model:



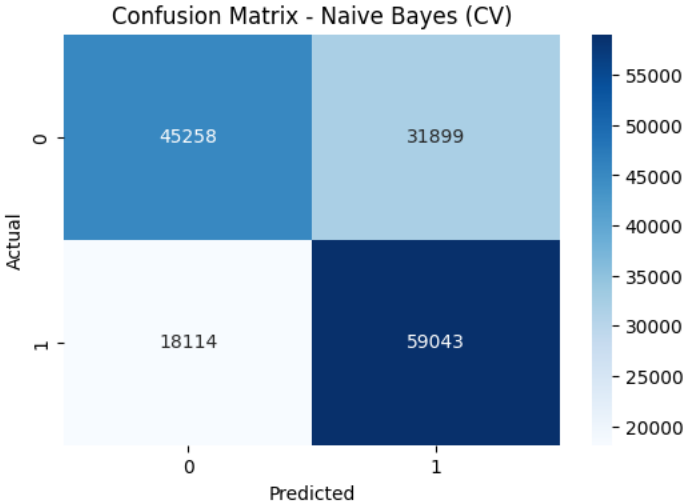
Neural Network



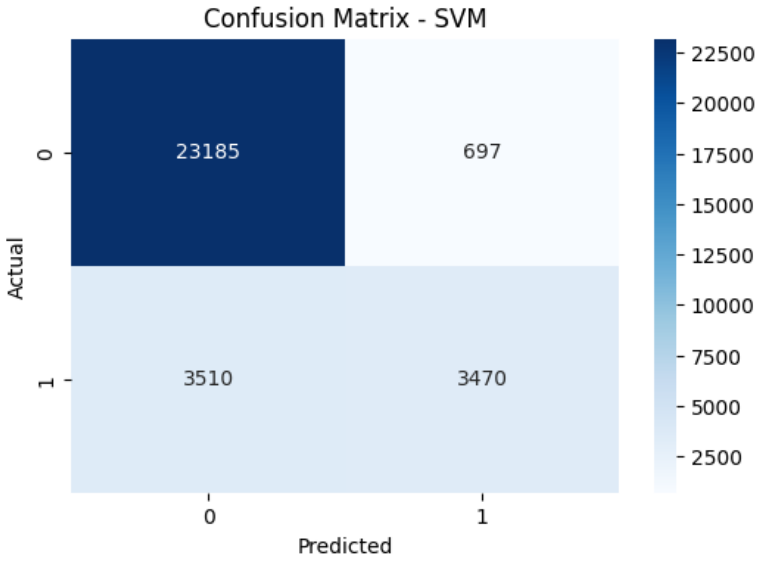
Neural Network (CV)



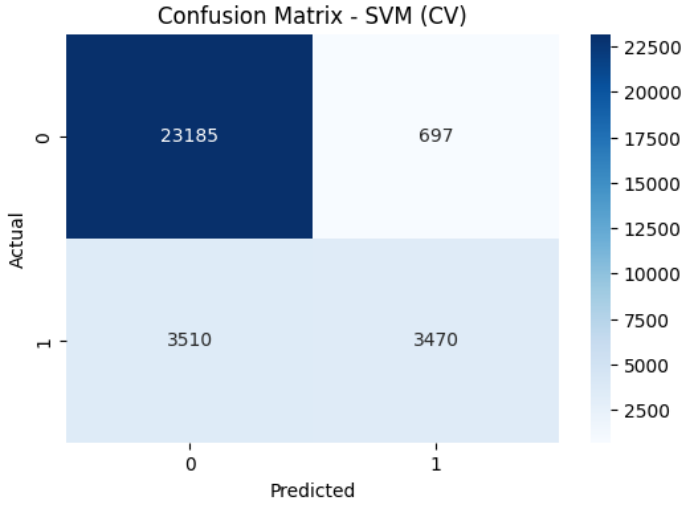
Naive Bayes



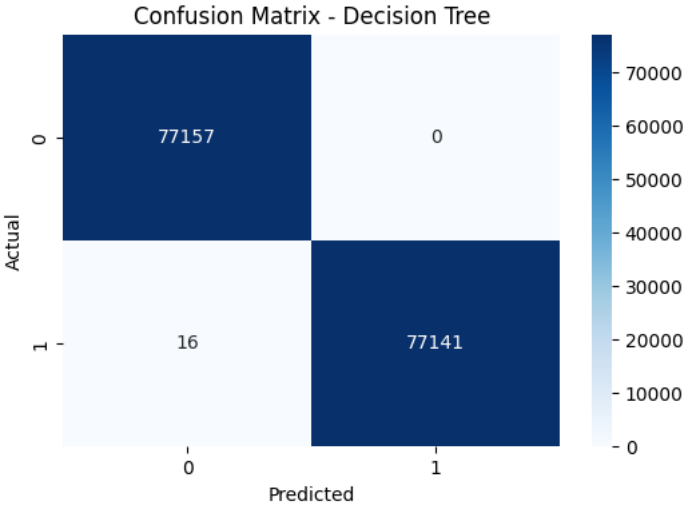
Naive Bayes (CV)



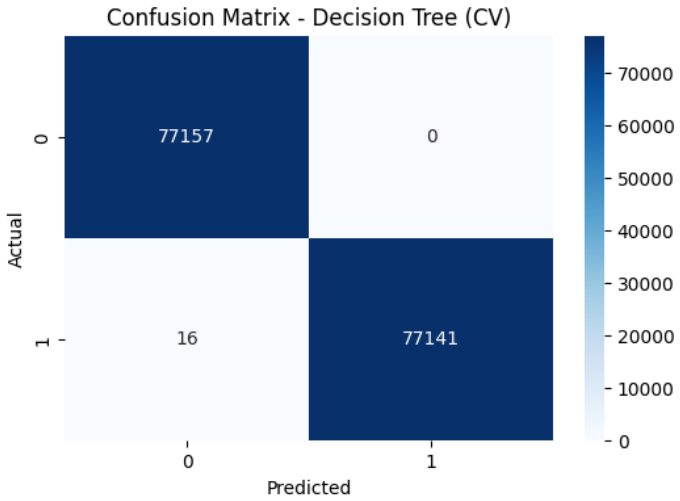
SVM



SVM (CV)

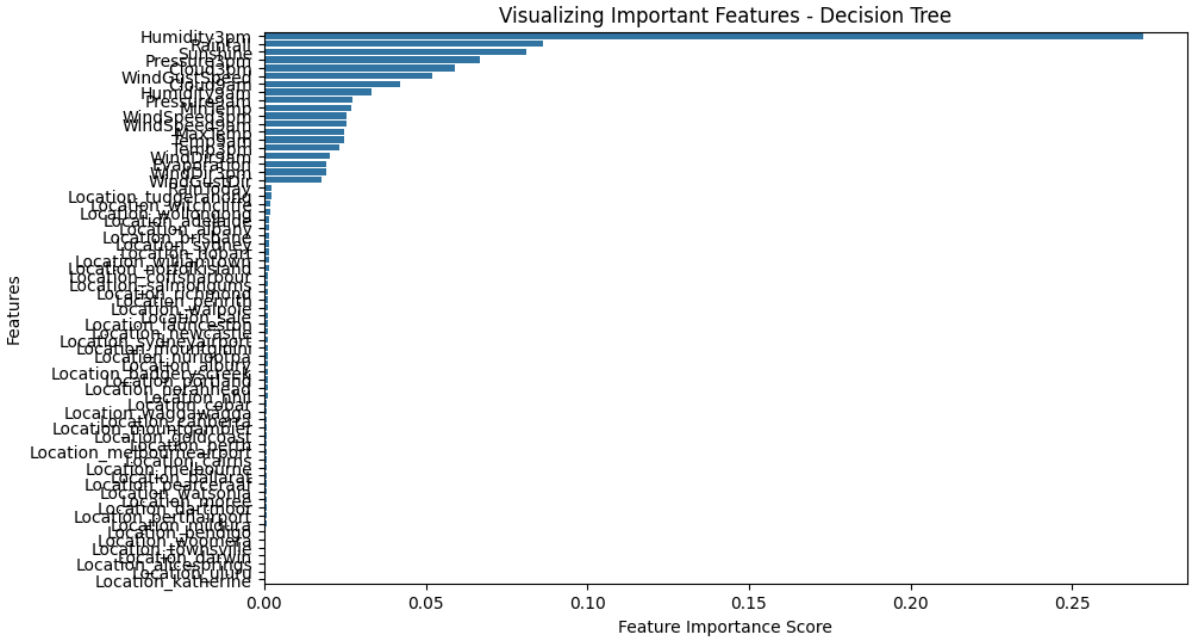


Decision Tree

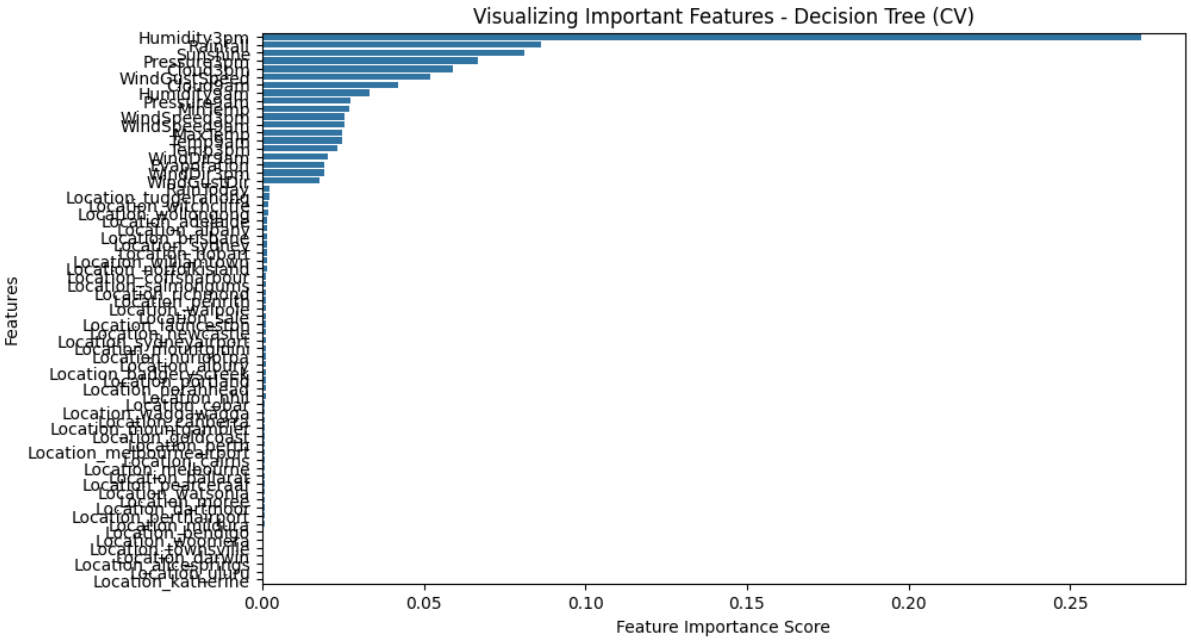


Decision Tree (CV)

2. Feature importance analysis:



Decision Tree



Decision Tree (CV)

3. Models set up: