**Comparison of Classification Techniques**

Author: Jose Manuel Alvarez Gonzalez

Email Id: jmanuel.ag89@gmail.com

Date: 3/6/2022

Contents

[Introduction 3](#_Toc136769043)

[Project 1 Summary 3](#_Toc136769044)

[Project 2 Summary 3](#_Toc136769045)

[Project 3 Aims & Motivation 4](#_Toc136769046)

[Summary of Performance Indicators 5](#_Toc136769047)

[Ensemble Method 6](#_Toc136769048)

[ROC Curve 7](#_Toc136769049)

[Discussion 9](#_Toc136769050)

[References 10](#_Toc136769051)

# Introduction

In this Report, we build upon the previous work conducted in Project 1 and Project 2, where we explored and analysed a dataset containing daily weather observations from various Australian weather stations across 2019 and 2020. The data set has 25 features and 1 target variable which tells if it rained that day or not.

## Project 1 Summary

The aim of this project was to prepare the dataset for predictive modelling, by conducting extensive data cleaning, transformation, feature selection, and normalization processes. The goal was to ensure that the dataset was suitable for training and evaluating machine learning models. Additionally, we aimed to identify the most relevant features for predicting rainfall and gain a deep understanding of the dataset's characteristics.

During this project, we identified a mixture of quantitative and qualitative features, categorized as nominal and ordinal values. The dataset initially required cleaning, as there were missing values and inconsistent formats across the features. We began by conducting descriptive statistical analysis to understand the properties of each feature. Some features, such as "Cld\_9am" and "Cld\_3pm," had more than half of the values missing and were removed from the data. Additionally, "AVG\_Rain\_mm" had only one unique value and did not contribute useful information. Two weather stations with a high percentage of null values across multiple features were removed.

To handle the missing data, we implemented a strategy based on replacing null values with the median value for interval features, while for non-interval features, the mode was employed, the replacement values used were grouped for each state. We used the Min-Max Scaler to normalize the interval features in a range of 0 to 1. For categorical data we converted in to numerical, keeping ordinal data with a hierarchy.

For the feature selection, we employed the method SelectKbest, this approach allowed us to identify the most important features for predicting rainfall. The top 6 features in hierarchical order were as follow: 'Rain\_mm', 'RH\_9am', 'RH\_3pm', 'Temp\_3pm', 'Max\_Temp', 'Temp\_9am'. Visual analysis using scatter plots and heatmaps confirmed the importance of these selected features and correlation with target.

## Project 2 Summary

In this Project, we implemented and evaluated four different classification models: K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, and Support Vector Classifier (SVC). The objective was to predict rainfall based on the given weather observations. We performed model training, parameter tuning, and evaluation using various performance metrics such as precision, recall, F1-score, and balanced accuracy. The aim was to compare the performance of these models and identify the most effective approach for rainfall prediction.

After the data preparation, we proceeded to split the dataset into training, validation, and test sets. Each split followed a 25% partition, with the Test data used for the final performance indicators. Given the distribution of the target variable, with 70% of the data represents days without rainfall, we kept distribution on each split. This 70% is also used as a baseline for the accuracy of the models.

To identify the ideal number of features to use, we created a learning curve for each model, to see how the accuracy increases over each feature. We found that 6 features were the right balance before a reduction on accuracy gain.

For each model, we conducted further parameter tuning to optimize their performance. For the KNN model, we explored different values for the number of neighbours and distance measurement metrics. The Logistic Regression model involved fine-tuning the regularization parameter, penalty type, and solver algorithm. In the case of the Decision Tree model, we adjusted parameters related to the tree depth, maximum leaf nodes, and minimum samples per leaf. For the SVC model, we experimented with different kernel functions and gamma values.

Based on precision, recall, F1-score, and balanced accuracy metrics, we observed that the Decision Tree and KNN models were the top performers. These models delivered an accuracy of 98.58% and 97.67%, respectively, doing better than the baseline of 70%. The simplicity of the Decision Tree model gives it a major advantage, as it only relies on the feature "Rain\_mm" to predict whether it would rain or not, at the same time this model scored a 100% on predicting when it won’t rain and 100% on recall by no producing false negatives.

The worst-performing model was the SVC, which only managed an accuracy of 88.66% on the test data, this model scored well by predicting when it will rain, with a 95% precision, but it performed lower by creating many false negative values with a 66% recall.

## Project 3 Aims & Motivation

In this project, our primary aims are to apply a Random Forest method to the dataset, conduct formal evaluations of all models used, include the construction of a ROC curve, and analyse the advantages and disadvantages of each model over the dataset. By accomplishing these aims, we aim to further enhance our skills as professional Data Analysts, gain experience in comparing and justifying the selection of models and explore the potential of ensemble methods.

The motivation behind this project lies in the practical application of predictive modelling for rainfall prediction. By understanding and comparing the performance of different classification methods, we can identify the most accurate and reliable model for predicting rainfall based on the given weather observations. This knowledge can have significant implications in various fields, including agriculture, water resource management, and disaster preparedness. Through this project, we aim to contribute to the understanding and improvement of rainfall prediction models, leading to more informed decision-making and improved outcomes in weather scenarios.

# Summary of Performance Indicators



In summary, the Random Forest is the best performer, enhancing the performance of the Decision Tree model by reducing false positive predictions. The KNN model also performed well with its simplicity and balanced precision and recall scores. The Regression and SVC models had limitations in terms of recall and computational efficiency.

# Ensemble Method

* Random Forest is the method implemented, it uses the bagging method of training. Which means that trains the model with multiple data sets by randomly selecting samples with replacement.
* First we define the parameters to work:  
  - 'n\_estimators': the number of total trees to split data.  
  - 'max\_depth': the maximum depth of each tree.  
  - 'max\_leaf\_nodes': the maximum number of leaf nodes in each tree.  
  - 'min\_samples\_split': the minimum number of samples required to split an internal node.  
  - 'min\_samples\_leaf': the minimum number of samples required to be at a leaf node.   
  - 'criterion': the function to measure the gain of information per split.
* The value for some parameters is kept low to prevent overfitting.
* The search for parameters is with the help of GridSearchCV function, by looking at best performing combination.
* Once we find some parameters from the function, we test again using a Cross Validation on the test data.
* With the final parameters we predict on the Validation data and compare results with Test data, ideally, we want similar results as this means no overfitting.
* Finalise by performing a prediction on the training data, visualise results and collect performance indicators.

A picture containing text, screenshot, number, diagram

Description automatically generated

Figure , Confusion Matrix for Random Forest in Test Data

# ROC Curve

Now we are going to use a different method to compare the scores of the model, we apply the ROC curve and we obtain the following plot:

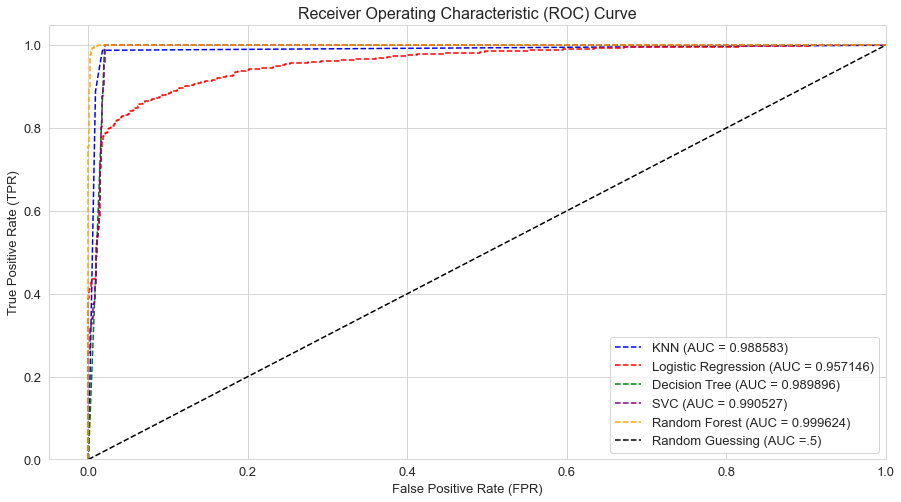


Figure , ROC curve for all models and 50% random guessing line.

The results of the Decision Tree and SVC are overlapping, in the following zoomed plot, we can find the difference between them:

A picture containing text, screenshot, line, plot

Description automatically generated

Figure , Zoomed ROC curve.

The ROC curve measures the trade off between the True Positive Rate and the False Positive Rate, the curve represents different thresholds for classifying the positive and negative samples. The curve visualises how the performance changes as the threshold for classifying a sample as positive or negative varies.

The main measurement to consider on the ROC curve, is the Area Under the Curve (AUC) this area quantifies the overall performance of the model. Represents the probability that a randomly chosen positive sample will be ranked higher than a randomly chosen negative sample by the model.

Using this method to evaluate the models, we obtain the following order of results from best to worst:

1. Random Forest (AUC = 0.999624)
2. SVC (AUC = 0.990527)
3. Decision Tree (AUC = 0.989896)
4. KNN (AUC = 0.988583)
5. Logistic Regression (AUC = 0.957146)

If we compare the last results with the previous key performance indicators for the models, we can identify that in both cases the Random Forest is the best performing model. For the worst performing model in the ROC curve, we have the Logistic Regression which had the second worst scores by the performance indicators.   
  
The model with the main difference from the ROC curve and performance indicators, is the SVC. In this case, it has the second highest AUC value, while comparing with the other performance indicators it did the worst. This can be explained as the ROC curve is biased towards the positive class, as we see on the X-axis False Positive Rate (FPR) and Y-axis True Positive Rate (TPR). The SVC archives a good performance on the positive class with only 29 samples as false positive (Just 7 more than the Random Forest), but this was at the cost of high false negative classifications with 283 classifications on the Test data. While the AUC fails at recognising this, the low recall score of 66% explains this classification error and lowers the total BCR score of the SVC to 82%.

# Discussion

In this project, we applied various classification models to predict rainfall based on weather observations from an Australian dataset. The best performing model was the Random Forest, followed by the Decision Tree and the KNN. The worst performing model was the SVC.

One of the highlights of the Random Forest, is its ability to enhance the performance of the Decision Tree model by reducing false positive predictions. It achieved the highest accuracy on the Test data and achieved a balanced precision and recall score, with the highest AUC area. The Random Forest model is reliable when high accuracy is required, making it the most reliable model for rainfall prediction.

**Advantages of the Random Forest:**

1. High Accuracy: The Random Forest model achieved the highest accuracy among all the models, with a 100% accuracy at predicting when it won’t rain and the lowest level of false positive predictions for predicting rain.
2. Feature Importance: The Random Forest model provided insights into feature importance, allowing us to identify the most significant variables for rainfall prediction.
3. Reduced Overfitting: By aggregating predictions from multiple decision trees, the Random Forest model mitigated the risk of overfitting and improved generalization. At the same time, the model allows to adjust parameters and prevent the overfitting of the model to the data.

**Disadvantages of the Random Forest:**

1. Computationally Intensive: The Random Forest model requires more computational resources compared to simpler models due to its ensemble nature and multiple decision trees.
2. Interpretability: While the Random Forest model offers high accuracy, its complexity makes it more challenging to interpret and understand the decision process. Particularly if we compare to the KNN or Decision Tree models.

The SVC model demonstrated high computational complexity, taking the longest to instantiate with its parameters. While it scored well in precision, it had a high number of false negatives, leading to a lower recall score. Overall, the SVC model had the lowest performance among the models. The SVC model had the second highest AUC value, but its performance indicators revealed limitations, particularly in recall with a high number of false negative samples.

**Advantages of the SVC:**

1. Flexibility in Non-Linear Data: The SVC model can effectively handle non-linear data by using different kernel functions, allowing it to capture complex relationships between features.
2. Good Precision: The SVC model achieved a high precision score, indicating its ability to accurately predict rainfall when it does occur.

**Disadvantages of the SVC:**

1. Low Recall: The SVC model had a relatively low recall score, meaning it struggled to correctly identify instances of rainfall. This limitation indicates a higher number of false negatives in predicting rainfall.
2. Computational Intensity: The SVC model is computationally expensive, especially when dealing with large datasets. Training and optimizing the SVC model can be time consuming and expensive if applied in large scale projects.

# References

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. Machine Learning, 42(3), 203-231.