RAINFALL PREDICTION USING MACHINE LEARNING

Abstract:

Rainfall prediction is the process of giving machine an ability to predict whether it will rain or not when given set of climate features that might affect rain. Given a machine the ability to predict the rain, it helps various people and business to plan their day ahead. Predicting rain is a heavily uncertain task. It depends on various factors relating climate, environment, geography and other. Forecasting rain is a difficult task in machine learning as it depends on various factors. Essentially, a machine learning algorithm is trained to understand the patterns in the time series history data and predict whether it is going to rain on a particular day or not based on weather conditions prevailing on that day. If someone could be able to know it’s going to rain tomorrow, they can be prepared for that. In addition, it will be of immense value to farmers if we could predict rain. Similarity, every business could benefit from rain fall prediction. To accomplish this, we have trained models using machine learning algorithms namely K Nearest Neighbour Classifier, Decision Tree Classifier and Random Forest Classifier. Exploratory data analysis and other feature engineering techniques have been used to prepare the data for algorithm training. By allowing us to learn more about the dataset using EDA, we can construct a more reliable model. Multiple plots were generated in EDA to aid in comprehension of the gathered information. Feature engineering included the use of methods like label encoding and Standard Scaling to improve the quality of the underlying data.

Keywords:

Rainfall prediction, Machine Learning, K Nearest Neighbour Classifier, Decision Tree Classifier, Random Forest Classifier.

Introduction:

Agriculture is the backbone of India's economy. There can be no agricultural success without rain. It's also useful for conserving water. Having access to historical rainfall data benefits the country's economy by allowing farmers to better manage their harvests. Rain forecasting helps to avoid flooding, which can save lives and property. The inconsistency in the occurrence and severity of rainstorms is a challenge for meteorologists trying to predict the weather. Predicting whether or not it will rain, given a collection of climate parameters that could affect rain, is known as rainfall prediction. Several industries, like aviation, agriculture, and travel, might benefit greatly from more accurate weather predictions, making this a fascinating area of study. Among the difficulties of weather forecasting include learning weather representation using a large dataset, and developing a reliable weather prediction model that makes use of latent structural patterns. When a machine can predict rain, it facilitates planning for a wide range of individuals and enterprises.

Predicting when and how much rain will fall is a highly uncertain task. Many factors, including climate, environment, location, and others, play a role. Due to the complexity of the problem, machine learning has not yet been successfully applied to the task of rain forecasting. In essence, a machine learning algorithm is taught to recognize patterns in time series history data to forecast whether or not it will rain on a certain day, given the specific climate conditions expected on that day. Knowing the weather in advance allows for better preparation. If you know it's going to rain tomorrow, you can make plans accordingly. In addition, if we can accurately forecast rainfall, it will be of great use to farmers. Similarly, just about any company would stand to gain from accurate weather forecasting. Predicting rainfall is a supervised classification task in machine learning. Supervised learning is a type of learning in which models are trained on both inputs and outputs. Classification is a type of problem in machine learning where the output is a discrete variable unlike continuous variable as you see in regression kind of tasks. In this project, we have trained machine learning algorithms namely K neighbours' classifier, Decision Tree Classifier and Random Forest Classifier. The algorithms were constructed and trained using Python as the programming language. For simpler and faster implementations, we relied on third-party modules like scikit-learn, pandas, and matplotlib. As we discovered in our review of the relevant literature, numerous studies have attempted to put into practice the same machine learning algorithms, but some of these studies have neglected to carry out essential preliminary steps such exploratory data analysis and feature encoding. In order to construct a reliable model, we will utilize all available machine learning methods in this project.

Here is how the rest of the paper is structured. The motivation for this work is discussed in Section 2. In the next paragraph, I'll briefly describe the project's primary outcomes and contributions. Following this is a synopsis of the literature survey, which provides a snapshot of the state of the art so far. In-depth information regarding the proposed framework is provided in Section 5. In Section 6, we provide a detailed description of the dataset. Results from the experiments are then presented. Together with these examples, a comparison of models is provided. At the end, you'll find a list of the references that were consulted during creating this project.

Motivation:

The use of science and technology to forecast rainfall is one way to determine how much rain is going to happen in each location. Accurate reporting is the most crucial factor in predicting rain. The goal of this project is to predict rainfall which in-turn would be helpful for farmers, in water resource planning and other agricultural uses. Farmers can better protect their crops and properties from heavy rains by using earlier rainfall information. Through effective rainfall information, farmers may better manage the country's economic growth. Precipitation forecasting is required to protect lives and property from flooding. Rainfall forecasting benefits residents of coastal areas by avoiding flooding. Rainfall Prediction is the application area of data science and machine learning to predict the state of the atmosphere. To produce crops efficiently and decrease flood-related and other rain-related disease-related mortality, it is crucial to forecast rainfall intensity. Above all, rainfall prediction could also help all commuters in planning their day. These problems will give us an opportunity to develop a machine learning model that could predict rainfall. Given, sufficient historic data describing different climate conditions on a day, machine learning algorithms are good enough to understand the patterns and predict whether it is going to rain or not tomorrow.

Main Contribution and Objectives:

- The goal of this survey is to create an ability to predict rainfall for the near future given climate conditions.
- Our project aims to implement algorithms namely K Nearest Neighbor Classifier, Decision Tree Classifier and Random Forest Classifier

- For nations heavily dependent on agriculture, predicting rainfall with heavy accuracy is extremely important for better organization and business.
- Our contribution through this project is we were able to produce a machine learning algorithm that predict rainfall given climate data.
- We were able to produce a machine learning model that produced 99% score on the test dataset.
- **Related Work:**
- Our literature review included a review of previous studies and other publications in the field. Multiple publications propose research that could be used to predict rain. Brief findings from our literature survey are presented in this section. A research paper proposed by Man Galih Salman et al. [6] has performed weather forecasting using deep learning techniques. Deep learning weather forecasting is studied. RNN, CRBM, and CN prediction performance are compared. These models are tested using meteorological datasets from BMKG (Indonesian Agency for Meteorology, Climatology, and Geophysics) from 1973 to 2009 and El-Nino Southern Oscillation (ENSO) data from international institutions such as the National Weather Service Center for Environmental Prediction Climate (NOAA). Frobenius norm measures forecasting accuracy. The problem with this paper is accuracy metrics were not properly defined and no proper comparison of models built were demonstrated in the study. Another research by R. Kingsy Grace and Suganya [7] has implemented machine learning based rainfall prediction. Multiple Linear regression was used to predict rainfall from meteorological parameters. Mean squared error and accuracy are the parameters used to validate the trained model. It is a regression problem where they are going to predict the amount of rainfall. But they haven't mentioned about the dataset. Aakash Parmar et al., [8] in their paper reviewed various papers published in predicting rainfall using machine learning algorithms. In their paper, they mentioned Rainfall estimation is important for water

management, human lives, and the environment. Because of geographical and regional changes and features, rainfall estimation can be erroneous or incomplete. The have provided all the existing studies with their algorithms implemented in the form of a table. From our literature survey, we have realized that most of the approaches have just implemented algorithms but not focussed on the data analysis part and feature engineering techniques. We consider these are the important prerequisites to be performed before training machine learning algorithms. These help us building better and robust models. In the next section, we will go through in-depth explanation of framework proposed through this project.

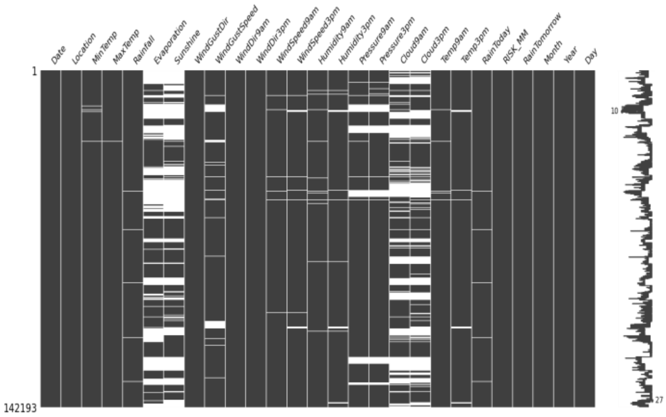
Proposed Framework:

This section is divided into three sections. They are:

1. Data Loading and Exploratory data Analysis
2. Data Pre-processing
3. Model training

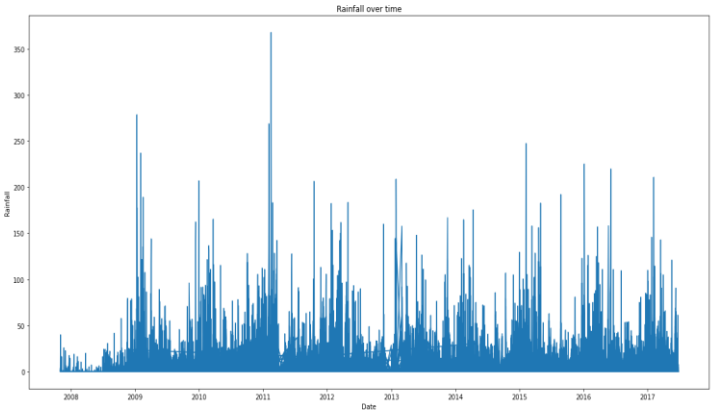
1. **Data Loading and Exploratory Data Analysis**

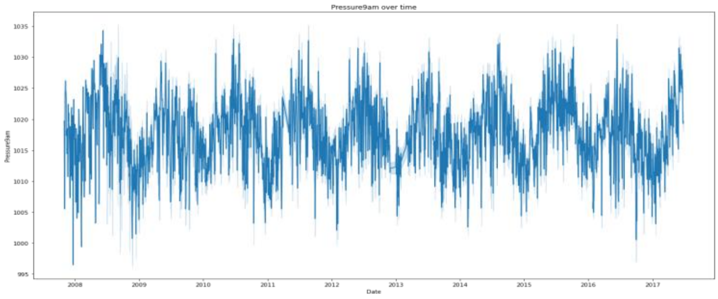
In this section, we will go through steps performed to load the data and to perform analysis. Once data is downloaded, it is made sure a python environment is created and all required external dependencies were installed. After setting up the environment, dataset is loaded using pandas. Thereafter, to start analysis, date column is split into day, month and year to see the trends in month and year. Dataset has been checked for null values. 4 columns in the dataset namely Sunshine, Evaporation, Cloud3pm and Cloud9pm has around 50% of the data. Figure [] visualized the missing values in each of the column in the dataset.

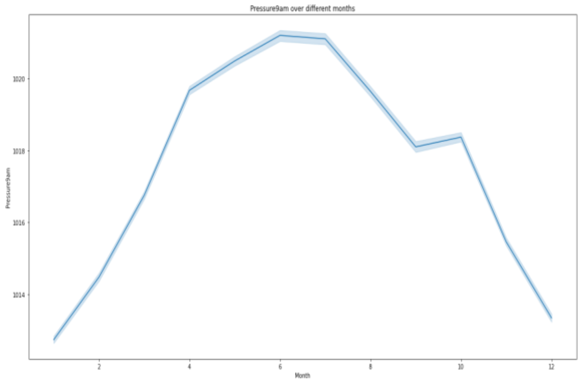


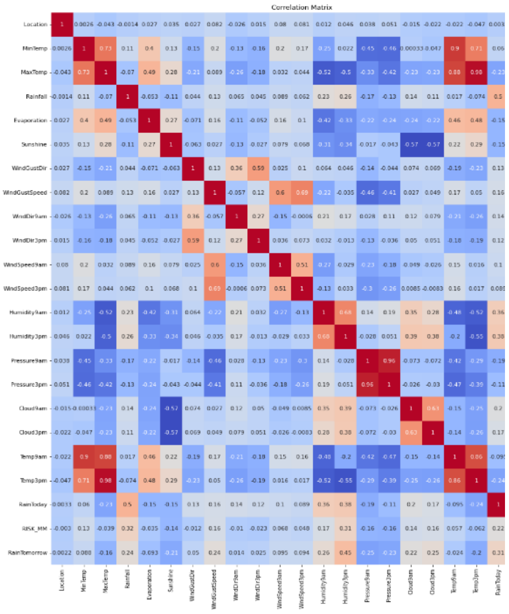
Target variable RainTomorrow is plotted to understand the distribution of labels as shown in figure []. It is found that dataset is imbalanced. Data belonging to no rain tomorrow are more in number than yes.

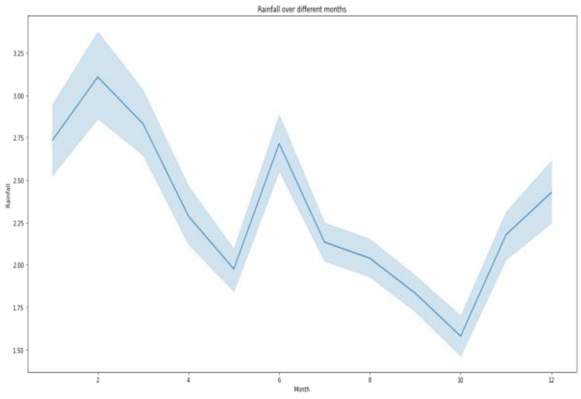
We will apply SMOTE technique on the data in the next step to make it balanced. Correlation matrix as shown in figure [] is visualized using seaborn library. Correlation matrix helps us to understand the correlation between pairs of variables. Followed by to understand about patterns in data, visualizations are done using pressure, rainfall over years as shown in figure [] and figure []. Similarly, both pressure and rainfall are visualized over month wise as well to see yearly trends as shown in figure [] and figure [].



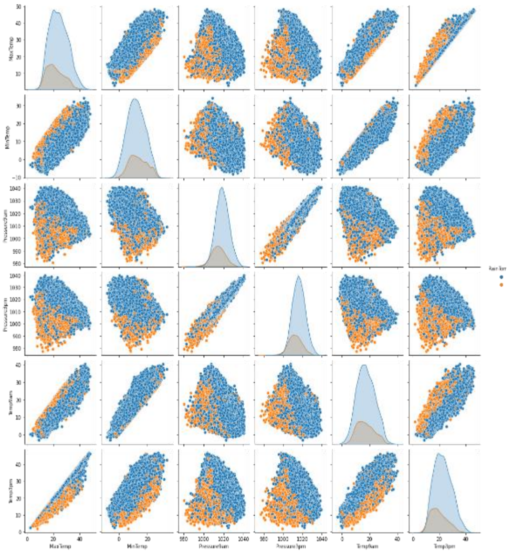








Finally, a pair plot has also been visualized to understand the relationship between numeric variables. Figure [] represents the pair plot of all numeric variables in the dataset.



2. Data pre-processing

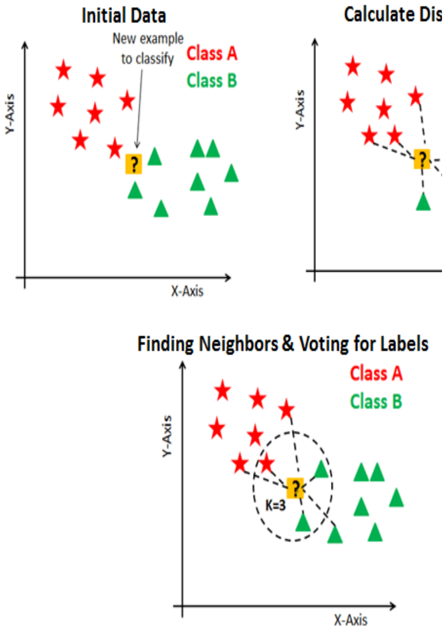
Once data is well loaded and analysed thoroughly using relevant visualizations, it is time to do pre-processing of the data. From our analysis, we could see that the data is imbalanced and there are null values in the dataset. These two challenges must be resolved using data engineering techniques before we start training our model. To eliminate null values, all of the null values in each of sunshine, evaporation, cloud3pm and cloud9am are filled with mode of the sample. Then, remaining null values are removed from the dataset. Now dataset is being split into training and testing ratios in the ratio of 80:20. Once the data is split, SMOTE technique is applied on train part of the data to handle data imbalance. After SMOTE technique, data gets balanced. Label Encoding is applied on all categorical values. Label Encoding is the process of converting categorical values into numerical values. Once label encoding is done, models are trained on training data and evaluated on test data.

3. Model training

To start training, three machine learning algorithms are selected. They are

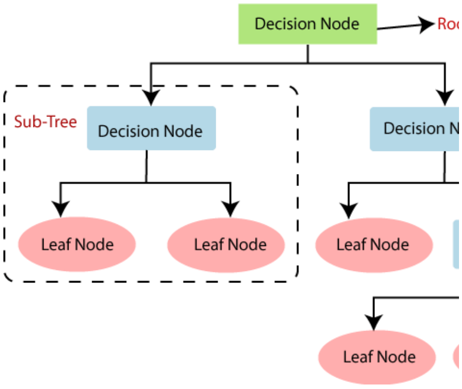
1. K Nearest Neighbor Classifier

   KNN Classifier is a simple and lazy machine learning algorithm. Given number of neighbors k, the algorithm tries to classify the new data point to one of the neighbors based on the distance metric like Euclidean Distance. The algorithm is implemented in Scikit-learn. Below is the code snippet to train and test the algorithm.



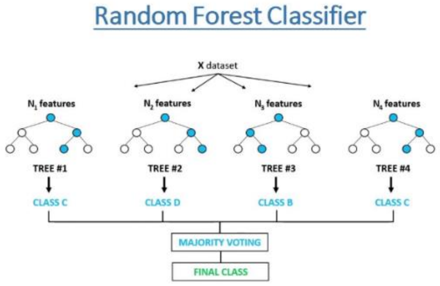
2. Decision Tree Classifier

   A decision tree is a type of supervised machine learning method that uses repeated questioning to classify data. The potential outcomes are shown down in a diagrammatic "decision tree" below.



3. Random Forest Classifier

   Like many other supervised machine learning algorithms, Random Forest Classifier is an ensemble of several decision trees. The decision trees in this case serve as independent estimators. This method of classification is superior to using just one decision tree.



Once models are trained, all of the models are evaluated using scikit learn metrics. Metrics used for evaluation accuracy, precision, recall and f1-score. Detailed classification report was generated for each of the model trained along with confusion matrix. Scores are presented in results section of the report.

Data Description:

The dataset has been extracted from Kaggle. The dataset contains around 142k samples of data. It can be downloaded from [1]. Data has been collected from various sources in Australia. The aim of the dataset is to predict whether is it going to rain tomorrow or not. There are 15 features in the dataset. They are as listed below:

1. Date – Date of the recorded observation
2. Location – Location of the sample recorded
3. MinTemp-Minimum Temperature in the 24-hour duration. In some cases, nearest whole degree is recorded. Units are degrees Celsius.
4. MaxTemp – Maximum Temperature in the 24-hour duration. In some cases, nearest whole degree is recorded. Units are degrees Celsius.
5. Rainfall – Rainfall in the 24 hours. In some cases, nearest whole degree is recorded. It is recorded in millimeters.
6. Evaporation – "Class A" pan evaporation in the 24 hours to 9 am. It is also recorded in millimeters.
7. Sunshine – Number of hours bright sunshine is present on the recorded day.
8. WindGustDir – Direction of the strongest Gust. Units of the measurement are 16 compass points.
9. WindGustSpeed – Speed of the strongest gust. Units of the measurement are kilometers per hour.

10.WindDir9am,3pm – Wind Direction over 10 mins prior. Units of measurement is compass points.

11. WindSpeed9am,3pm-Wind speed. Speed is recorded in kilometers per hour.

12. Humidity9am,3pm-Relative Humidity in the atmosphere on that day. It is presented in the dataset as a percentage.

13. Pressure9am,3pm – Pressure reduced to mean sea level. Units of measurement is hectopascals.

14. Cloud9am,3pm – Fraction of Sky Obscured by cloud. Units of measurement is eights.

15. Temp – Temperature recorded in degrees Celsius.

16. RainToday -Whether it rained on the day of observation

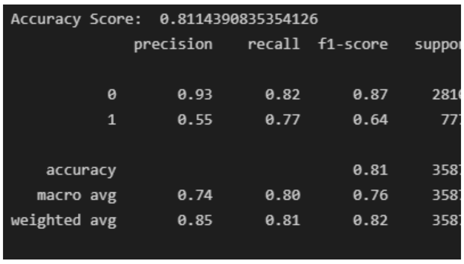
17. RainTomorrow – Whether it rained on the next day of observation or not.

Upon observing the dataset, it is found that 4 features namely Sunshine, Evaporation, Cloud3pm, Cloud9am has 50% of the data as null. In the dataset, RainTomorrow is the variable which is to be predicted. Yes value indicates it rained on the next day and No indicates it didn’t rained the next day.

Results/Experimentation & Conclusion:

Following are the results of various experiments done.

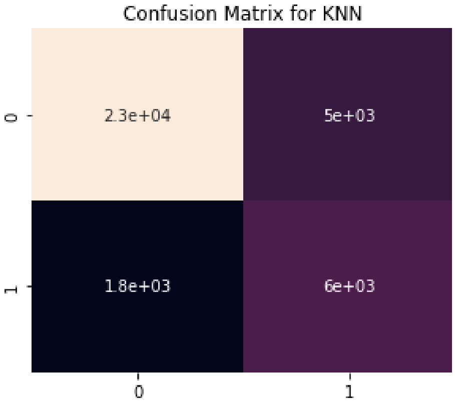
K Nearest Neighbour Classifier:



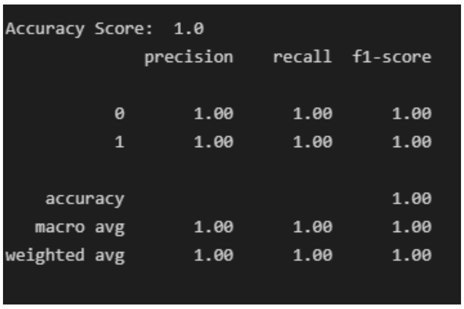
```
Accuracy Score:  0.8114390835354126
              precision    recall  f1-score   suppo

           0       0.93      0.82      0.87     281
           1       0.55      0.77      0.64      77

    accuracy                           0.81     358
   macro avg       0.74      0.80      0.76     358
weighted avg       0.85      0.81      0.82     358
```



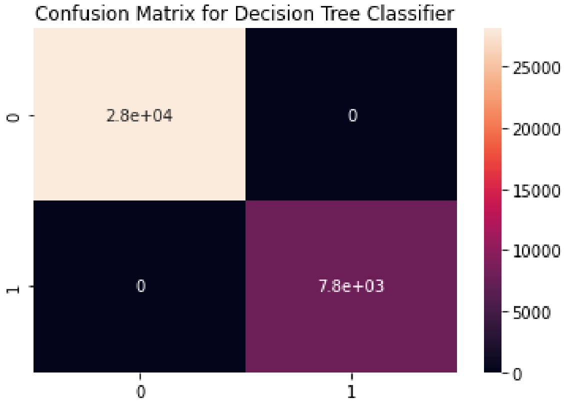
Decision Tree Classifier:



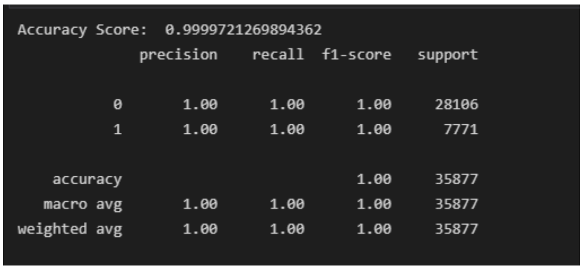
```
Accuracy Score:  1.0
              precision    recall  f1-score

           0       1.00      1.00      1.00
           1       1.00      1.00      1.00

    accuracy                           1.00
   macro avg       1.00      1.00      1.00
weighted avg       1.00      1.00      1.00
```



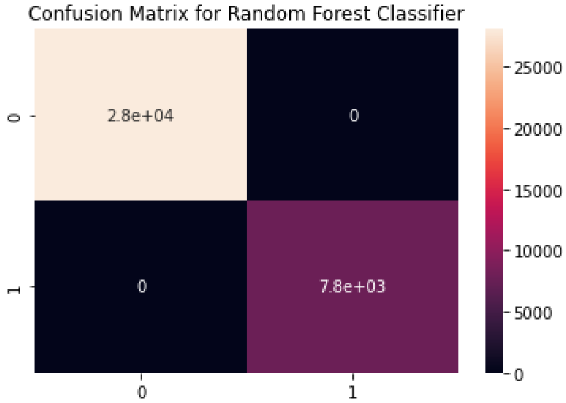
Random Forest Classifier:

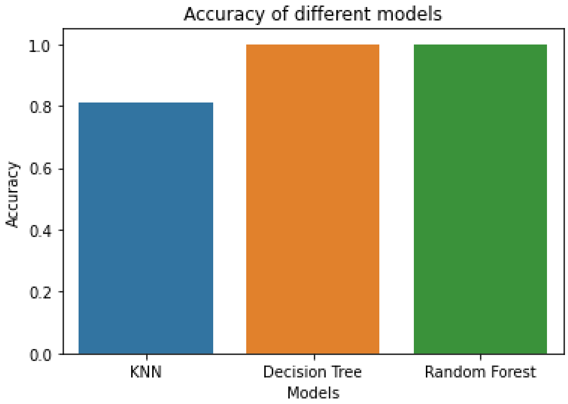


```
Accuracy Score:  0.9999721269894362
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28106
           1       1.00      1.00      1.00      7771

    accuracy                           1.00     35877
   macro avg       1.00      1.00      1.00     35877
weighted avg       1.00      1.00      1.00     35877
```





Decision tree and Random Forest Classifier performed better than K Nearest Neighbour Classifier. SMOTE technique for handling class imbalance has prven to improve the robustness of model.

References:

[1] https://www.kaggle.com/datasets/gauravduttakiit/weather-in-aus

[2] http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

[3] http://www.bom.gov.au/climate/data/

[4] Hernández, E., Sanchez-Anguix, V., Julian, V., Palanca, J., & Duque, N. (2016, April). Rainfall prediction: A deep learning approach. In International conference on hybrid artificial intelligence systems (pp. 151-162). Springer, Cham.

[5] Basha, C. Z., Bhavana, N., Bhavya, P., & Sowmya, V. (2020, July). Rainfall prediction using machine learning & deep learning techniques. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 92-97). IEEE.

[6] Salman, A. G., Kanigoro, B., & Heryadi, Y. (2015, October). Weather forecasting using deep learning techniques. In 2015 international conference on advanced computer science and information systems (ICACSIS) (pp. 281-285). Ieee.

[7] Grace, R. K., & Suganya, B. (2020, March). Machine learning based rainfall prediction. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 227-229). IEEE.

[8] Parmar, A., Mistree, K., & Sompura, M. (2017, March). Machine learning techniques for rainfall prediction: A review. In International Conference on Innovations in information Embedded and Communication Systems (Vol. 3).

[9] Refonaa, J., Lakshmi, M., Dhamodaran, S., Teja, S., & Pradeep, T. N. M. (2019). Machine learning techniques for rainfall prediction using

neural network. Journal of Computational and Theoretical Nanoscience, 16(8), 3319-3323.

[10] Barrera-Animas, A. Y., Oyedele, L. O., Bilal, M., Akinosho, T. D., Delgado, J. M. D., & Akanbi, L. A. (2022). Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. Machine Learning with Applications, 7, 100204.

[11] Just a moment. . . (n.d.). https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[12] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[13] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[14] 1.10. Decision Trees. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/tree.html

[15] Oswal, N. (2019). Predicting rainfall using machine learning techniques. arXiv preprint arXiv:1910.13827.

[16] Ridwan, W. M., Sapitang, M., Aziz, A., Kushiar, K. F., Ahmed, A. N., & El-Shafie, A. (2021). Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. Ain Shams Engineering Journal, 12(2), 1651-1663.

[17] Sumi, S. M., Zaman, M., & Hirose, H. (2012). A rainfall forecasting method using machine learning models and its application to the Fukuoka city case. International Journal of Applied Mathematics and Computer Science, 22(4), 841-854.

[18] Appiah-Badu, N. K. A., Missah, Y. M., Amekudzi, L. K., Ussiph, N., Frimpong, T., & Ahene, E. (2021). Rainfall Prediction Using Machine Learning Algorithms for the Various Ecological Zones of

Ghana. IEEE Access, 10, 5069-5082.

[19] Shah, U., Garg, S., Sisodiya, N., Dube, N., & Sharma, S. (2018, December). Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 776-782). IEEE.

[20] https://raw.githubusercontent.com/amankharwal/Website-data/master/weatherAUS.csv.