RAINFALL PREDICTION USING MACHINE LEARNING MODEL

ABSTACT

Rainfall Prediction is one of the most important and challenging tasks. We are going to use the weather data of Australia to predict the possibility of the rain for tomorrow. The main objective of this study is to find relevant features present in the dataset and use them to predict the possibility of rainfall in Austraila tomorrow. For better data visualization we are going to divide this data in rainy season and unrainy season and apply the models for rainy datasets only to build more accurate models. First, we will do the data analysis on the dataset then based on the analysis we will perform the preprocessing on the dataset. Preprocessing includes oversampling the data, removing the outliers using IQR (Inter Quartile Range), Filling missing value using mode and median for object and numerical values respectively, Standardizing the data using min-max Scaler, selecting importance features using Chi-Square. Then we have applied the AutoML and pipeline has been used for splitting the data according to problem type (Binary Classification) and get approximate results. We have applied the best models for classifications like Logistic Regression, Random Forest Classifier, Decision Tree Classifier, XGBoost, CatBoost and find metrices like Accuracy, Recall, Confusion metrix, F1 Score and recall for evaluating the models.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 - INTRODUCTION

The rainfall prediction is one of the most difficult and challenging tasks nowadays due to the global warming and environmental pollution. The possibility of rain is getting unpredicted days by days so we need an algorithm to find the correct prediction of rain and that's where the use of machine learning comes in the picture. It helps us to find the hidden pattern which is normally impossible to find by humans in specific time.

Rainfall prediction is getting more and more demanding because of its use in many industries and sectors like Agricultural, Construction, investing in any stocks also requires an analysis of rain to predict the stock price. Planning a trip or even planning a daily schedule greatly depends on the rain. Since, Rain depends on many important sectors, we can also say that rain is one of the pillars of finance sector for any nation.

Rainfall prediction is essential since this variable is the one with the highest correlation with adverse natural events such as landslides, flooding, mass movements and avalanches. These incidents have affected society for years. Therefore, having an appropriate approach for rainfall prediction makes it possible to take preventive and mitigation measures for these natural phenomena.

This study presents a set of experiments which involve the use of prevalent machine learning techniques to build models to predict whether it is going to rain tomorrow or not based on weather data for that particular day in major cities of Australia. The wet rainy season starts in month October-November and last till April-May. We will make classification model which will make prediction in this rainy season months.

To solve this uncertainty, we used various machine learning techniques and models to make accurate and timely predictions. These paper aims to provide end to end machine learning life cycle right from Data preprocessing to implementing models to evaluating them. Data Preprocessing steps includes imputing missing values, feature transformation, encoding categorical features, feature scaling and feature selection. We have implemented classification models such as Logistic Regression, Decision Tree, Random Forest, XGBoost and CatBoost. For evaluation purpose, we used Accuracy, Precision, Recall, F-Score and Area Under Curve as evaluation metrics. We are expecting accuracy of the model to achieve 95 to 96 percent accuracy.

CHAPTER 2 - LITERATURE REVIEW

Here we have take the literature written by Ayisha Siddiqua L, Senthil kumar N C on rainfall prediction. In this survey various decision tree based classification algorithms are discussed and compared with the proposed work. This literature proposes a Data mining approach in order to predict rainfall based reflectivity of radar, This paper uses data mining algorithms for some conditions like, temperature, humidity, pressure and etc from data to predict rainfall. Pre-processing was performed on the dataset to fill the missing values and to remove the noise before classification. The mining techniques included Naïve Bayes, Neural Network, SVM, Decision Tree, and Random Forest.

https://www.researchgate.net/publication/354968068_RAINFALL_PREDICTION_USING_MACHINE_LEARNING_CLASSIFICATION_ALGORITHMS

According to R. Kingsy Grace et.al one of the most important techniques for predicting meteorological conditions in any country is rainfall prediction. For the Indian dataset, this paper proposes a rainfall prediction model based on Multiple Linear Regression (MLR). Multiple meteorological characteristics are included in the input data, allowing for a more accurate estimate of rainfall. The suggested model is validated using the Mean Square Error (MSE), accuracy, and correlation metrics. The proposed machine learning model outperforms all other techniques in the literature, according to the results.

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

According to Urmay Shah et.al estimating precipitation is one of the most important aspects of meteorological science. To forecast and estimate meteorological parameters, a couple of factual procedures and machine learning approaches are used to forecast and estimate precipitation. The purpose of this paper is to provide climatic insights to clients from diverse industries, such as agriculturists, researchers, and others, so that they may understand the relevance of changes in climate and atmosphere characteristics such as precipitation, temperature, and humidity.

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.

CHAPTER 3 - PROPOSED METHODOLOGY

In this paper, the overall architecture includes four major components: Data Exploration and Analysis, Data Pre-processing, Model Implementation, and Model Evaluation.

Dataset Description:

Figure 3.1:

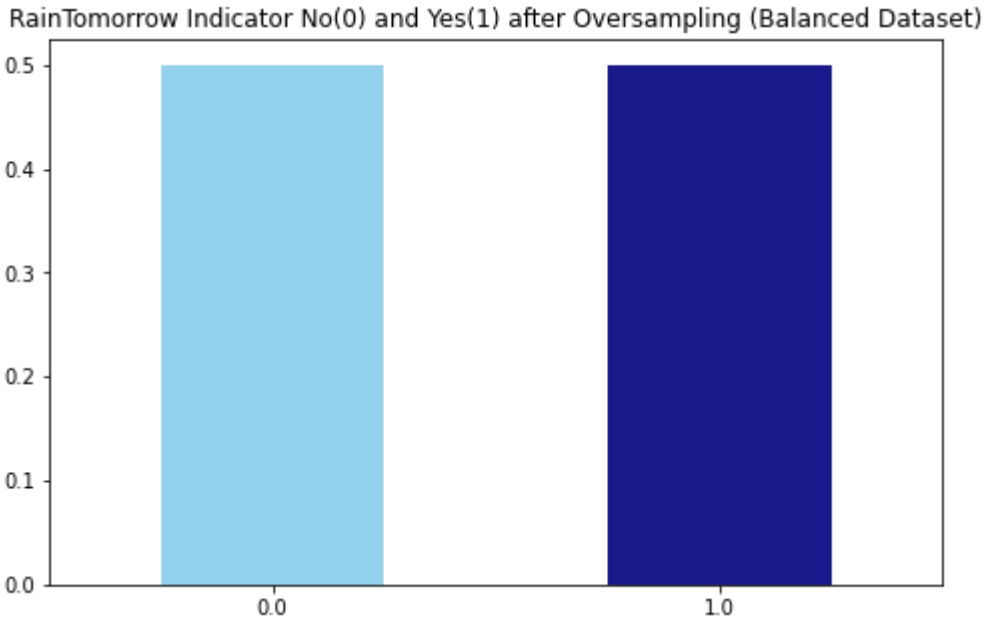
Table 1. Descriptive table of the dataset.

Feature Name	Description	Missing Values	Available Data	Type
Date	Day on which the measurement is carried out	0	142,193	string/date
Location	Station location name meteorological.	0	142,193	string
MinTemp	Minimum temperature in degrees Celsius.	637	141,556	float
MaxTemp	Maximum temperature in degrees Celsius.	322	141,871	float
Rainfall	Amount of rain recorded during the day in mm.	1406	140,787	float
Evaporation	"Class A pan evaporation" (mm) in 24 h until 9 a.m.	60,843	81,350	float
Sunshine	Number of hours of radiant sun during the day.	67,816	74,377	float
WindGustDir	Direction of the strongest wind gust in the 24 h to midnight.	9330	132,863	string
WindGustSpeed	Speed (km/h) of the strongest wind gust in the 24 h to midnight.	9270	132,923	float
WindDir9am	Wind direction at 9 a.m.	10,013	132,180	string
WindDir3pm	Wind direction at 3 p.m.	3778	138,415	string
WindSpeed9am	Average wind speed (km/h) in the 10 min before 9 a.m.	1348	140,845	float
WindSpeed3pm	Average wind speed (km/h) in the 10 min before 3 p.m.	2630	139,563	float
Humidity9am	Humidity (%) at 9 a.m.	1774	140,419	float
Humidity3pm	Humidity (%) at 3 p.m.	3610	138,583	float
Pressure9am	Atmospheric pressure (hpa) at the level of evil, at 9 a.m.	14,014	128,179	float
Pressure3pm	Atmospheric pressure (hpa) at the level of evil, at 3 p.m.	13,981	128,212	float
Cloud9am	Fraction of sky obscured by clouds at 9 a.m. The unit of measurement is "oktas", which is equal to a unit of eighths. It refers to how many eighths of the sky are obscured by clouds. A value of 0 indicates a completely clear sky, while a value of 8 indicates that it is completely obscured. Temperature at 3 p.m., in degrees Celsius.	53,657	88,536	float
Cloud3pm	Fraction of sky obscured by clouds at 3 p.m. The unit of measurements is the same as in Cloud9am measurements.	57,094	85,099	float
Temp9am	Temperature at 9 a.m., in degrees Celsius.	904	141,289	float
Temp3pm	Temperature at 3 p.m., in degrees Celsius.	2726	139,467	float
RainToday	Boolean: 1 if precipitation exceeds 1 mm in the 24 h to 9 a.m., if not 0.	1406	140,787	string
RISK_MM	The amount of rain for the next day in mm.	0	142,193	float
RainTomorrow	Variable created from variable RISK_MM. A type of risk measure.	0	142,193	String

Pre-processing:

- Performed oversampling of the data.

Figure 3.2:



Size of the data after oversampling: (110846, 23)

- Filling Missing categorical values with their mode. Filling missing numerical values with median of the column in value.
- Applying label encoder to each column with categorical data. We have also done standardization of the data.
- Standardization the Dataset using MinMaxSacler.
  - MinMaxScaler is used where the upper and lower boundaries are well known from domain knowledge. This helps in scaling all the data features in the range [0, 1] or else in the range [-1, 1] if there are negative values in the dataset.This scaling compresses all the inliers in the narrow range.
- Using Chi-Square for best feature selection.
- After that, Applying Automl and pipeline for splitting the data according to the problem type(Binary Classification) and get approximate results.

Evaluation:

Models Used for evaluation:

- XGBoost : XGBoost is a distributed gradient boosting library that has been developed to be very effective, adaptable, and portable. A parallel tree boosting method called XGBoost (also known as GBDT or GBM) is available to quickly and accurately address a variety of data science issues.
- Logistic regression : Logistic regression is used for prediction and classification problems. Some of the classification problems are Fraud Detection(logistic regression helps in detecting data anomalies and which are predictive to fraud). So we can also evaluate our model using logistic regresssion as it is a binary classification model.
- Decision Tree Classifier : The non-parametric supervised learning approach used for classification and regression applications is the decision tree. It is organised hierarchically and has a root node, branches, internal nodes, and leaf nodes.
- Random Forest Classifier : A random forest is a meta estimator that employs averaging to increase predicted accuracy and reduce overfitting after fitting several decision tree classifiers to distinct dataset subsamples.
- Catboost : Without doing any explicit pre-processing, we may utilize CatBoost to transform categories into numbers. CatBoost uses various statistics on categorical feature combinations and categorical and numerical feature combinations to translate categorical values into numbers.

CHAPTER 4 – RESULT ANALYSIS

Here the result got from all models are shown in table form:

Table 4.1:

Model Name	Class Label	Testing Evaluation			
		P	**R**	**F1**	**A**
Logistic Regression	No	0.77	0.82	0.79	0.76
	Yes	0.74	0.67	0.70	
Random Forest Classifier	No	0.94	0.91	0.92	0.91
	Yes	0.89	0.92	0.90	
Decision Tree	No	0.87	0.83	0.85	0.83
	Yes	0.79	0.84	0.81	
CatBoost	No	0.96	0.91	0.94	0.93
	Yes	0.89	0.96	0.92	
XGBoost	No	**0.98**	**0.94**	**0.96**	**0.955**
	Yes	**0.92**	**0.98**	**0.95**	

As we can see that we have got the best evaluation metrics in XGBoost model compare to other models and least performing model as Logistic Regression. Logistitc Regression underperforms than other models with accuracy of 0.76 and XGboost outperforms other models with accuracy of 0.955 which is quite good and other parameters are also near 0.96. so, we can clearly see that XGBoost is the best model for Rainfall prediction.

CHAPTER 5 - CONCLUSION & FUTURE WORK

In this research, it was examined whether machine learning techniques could be used to solve the issue of rainfall forecasting in the particular situation of Australia. Previous works exist. Who have utilised this method in places other than Australia and with various many monthly, yearly, and other time period dataset types. The places that were investigated. Victoria and Sydney were the locations, and neural models were typically used. Random Forest and networks. In accordance with these earlier investigations, a collection of was obtained from 49 distinct Australian cities' meteorological data. There is one in the set a variable (RainToday) that represents whether it rained on the day of the measurement. Besides the sample, there are more factors that exhibit climatic characteristics. The sample day, such as the presence or absence of clouds, wind, sunlight, humidity, pressure, or temperature. The preprocessed dataset was used to generate numerous machine learning prediction models. To predict rainfall, learning techniques were used (Knn, Decision Tree, Random Forest, neural networks, etc.). Therefore, it was determined that the most accurate model to represent this type neural networks are a phenomenon. The models' suitability for use in various cities was examined separately. It was shown that in this instance, the effectiveness of the higher algorithms. The potential for results to be enhanced by changing the amount of data and actual values utilised to conduct the training were examined, but without advancement from earlier analyses. Consequently, this latest study enabled.

I'll bring up various points now. The potential of machine learning, on the one hand, approaches as substitutes for conventional rain prediction techniques they also have some advantages over traditional forecasting techniques, including the ability to estimate the the accuracy of the outcomes utilising the indicators, the performance key, or the potential for modification the algorithms' performance by changing their input parameters, which enables they to be modified for certain situations). Additionally, it is clear that algorithms based on the other hand, neural networks simulate nonlinear natural processes fairly well. Since each set of data can be analysed separately, the phenomenon's location can be determined. The algorithms function and are more effective by city. There are numerous ways to continue the task. Therefore, it would be interesting to examine the outcomes of the study of data from various nations as well as the meteorological data from 2019 to the present. With the latter, I could determine whether the efficiency was transferable to other regions and whether the outcomes were independent of location. Last but not least, forecasting which models are the most intriguing or how many days in advance is ideal would be a very fascinating future study relating to the one just mentioned in order to predict.

REFERENCES

- https://www.researchgate.net/publication/336914968_Predicting_Rainfall_using_Machine_Learning_Techniques

- https://www.researchgate.net/publication/359513858_Prediction_of_Rainfall_in_Australia_Using_Machine_Learning

- https://www.researchgate.net/publication/354968068_RAINFALL_PREDICTION_USING_MACHINE_LEARNING_CLASSIFICATION_ALGORITHMS

- 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.

- 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