Rainfall Prediction Using Machine Learning

N.Bhanu Sravya[1],P.Rafiya[2],K.Abhigna[3],A.Thanuja[4]

[1,2,3]Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

[4]Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

nissankararaobhanusravya@gmail.com[1],rafiyapathan4085@gmail.com[2], karnatiabhignareddy@gmail.com[3],a.thanuja18@gmail.com[4]

ABSTRACT- "Rainfall Prediction Using Machine Learning" is the project's name. The dataset for this project is kept in Microsoft Excel, and the project is written in Python. Manymachine learning algorithms are used in this predictionto see which method makes the best accurate predictions. Forecasting rainfall is crucial in many areas of a nation and can aid in averting catastrophic natural catastrophes. Logistic Regression, Random Forest Classifier, Gradient Boosting, KNN, Decision Tree, Adaboost Classifier, and Catboost Classifier were allutilised to make this prediction. This project uses a total ofseven modules. The Australian rainfalldataset was utilised. The project's primarygoal is to evaluate different algorithms and identifythe top algorithmout ofthose algorithms. The farmers maygreatly benefit fromthis prediction by planting the appropriate crops based ontheir requirement for water.

KEYWORDS: Machine Learning, Logistic Regression,KNN, Random forest classifier, Gradient Boosting,Adaboost,Decision tree,Catboost.

1. INTRODUCTION

How to predict when it will rain is a topic that interests governments, corporations, risk management organisations, and the scientific community all at the same time. Rainfall is a climatic factor that affects a variety of human endeavours, such as tourism, forestry, construction, and agricultural production.

This project is used to predict the rainfall in the 49 cities of Australia.The prediction uses various algorithms. Forecasting rainfall is crucial in many areas of the nation and can aid in averting catastrophic natural catastrophes.

The goal of this study is to offer complete machine learning life cycle models. Here, we'll look at several model descriptions in more depth. The models in question are listed as follows:

1. DATA COLLECTION
2. DATA VISUALIZATION
3. DATA PREPROCESSING
4. MODEL SELECTION
5. PERFORMANCE EVALUATION.

The structure of the essay is as follows. In section 2, we first explain the dataset. Section 3 presents the strategies and approaches that were employed. In section 4, the outcomes are finally discussed.

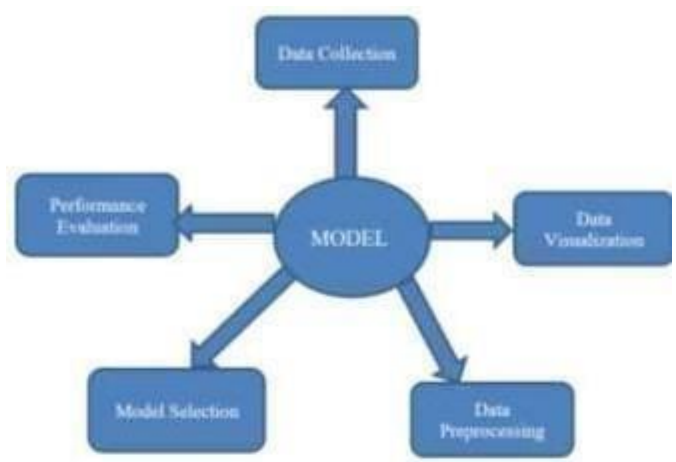

Fig.1(a) 5 steps involved in Model

2. DATASET

This section will cover every aspect of the dataset. We'll look at the location of the dataset, the properties it contains, and how each attribute is described in the dataset.

Several descriptions are provided for each characteristic.

The Attributes and datatype and no.of null values in each attribute of the dataset are described as shown below:

```
 0   Date           142193 non-null  object
 1   Location       142193 non-null  object
 2   MinTemp        141556 non-null  float64
 3   MaxTemp        141871 non-null  float64
 4   Rainfall       140787 non-null  float64
 5   Evaporation    81350 non-null   float64
 6   Sunshine       74377 non-null   float64
 7   WindGustDir    132863 non-null  object
 8   WindGustSpeed  132923 non-null  float64
 9   WindDir9am     132180 non-null  object
 10  WindDir3pm     138415 non-null  object
 11  WindSpeed9am   140845 non-null  float64
 12  WindSpeed3pm   139563 non-null  float64
 13  Humidity9am    140419 non-null  float64
 14  Humidity3pm    138583 non-null  float64
 15  Pressure9am    128179 non-null  float64
 16  Pressure3pm    128212 non-null  float64
 17  Cloud9am       88536 non-null   float64
 18  Cloud3pm       85099 non-null   float64
 19  Temp9am        141289 non-null  float64
 20  Temp3pm        139467 non-null  float64
 21  RainToday      140787 non-null  object
 22  RISK_MM        142193 non-null  float64
 23  RainTomorrow   142193 non-null  object
dtypes: float64(17), object(7)
memory usage: 26.0+ MB
```

In the above dataset total we are having 23 attributes in the above metioned attributes our main aim is to predict wheather there will be rain tomorrow or not the main attribute is used for this prediction is **"RAINTOMORROW"** .

In our dataset we are having the null values in each and every attribute so we have to remove those null values .In order to remove those null values we have the concept of data preprocessing.In this data preprocessing we will be using the data cleaing technique.The description of the attributes are shown in the below picture.

Feature	Description
Date	The date of observation
Location	The common name of the location of the weather station
MinTemp	The minimum temperature in degrees celsius
MaxTemp	The maximum temperature in degrees celsius
Rainfall	The amount of rainfall recorded for the day in mm
Evaporation	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine	The number of hours of bright sunshine in the day.
WindGustDir	The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am	Direction of the wind at 9am
WindDir3pm	Direction of the wind at 3pm
WindSpeed9am	Wind speed (km/hr) averaged over 10 minutes prior to 9am
WindSpeed3pm	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
Humidity9am	Humidity (percent) at 9am
Humidity3pm	Humidity (percent) at 3pm
Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9am
Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
Cloud9am	Fraction of sky obscured by cloud at 9am.
Cloud3pm	Fraction of sky obscured by cloud at 3pm.
Temp9am	Temperature (degrees C) at 9am
Temp3pm	Temperature (degrees C) at 3pm
RainToday	1 if precipitation exceeds 1mm, otherwise 0
RISK_MM	The amount of next day rain in mm.
RainTomorrow	The target variable. Did it rain tomorrow?

Fig 2.1.`weatherAUS.csv`

The total no.of rows in the dataset is 42191 rows for 24 columns.Sample data in the dataset is shown in the below picture format.

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	P
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WNW	20.0	24.0	71.0	22.0	
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WSW	4.0	22.0	44.0	25.0	
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WSW	19.0	26.0	38.0	30.0	
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	E	11.0	9.0	45.0	16.0	
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	NW	7.0	20.0	82.0	33.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
42186	2011-05-02	Wollongong	13.6	19.2	0.0	NaN	NaN	WSW	17.0	SW	NaN	11.0	0.0	75.0	69.0	
42187	2011-05-03	Wollongong	14.7	18.7	0.4	NaN	NaN	SSW	52.0	SW	SSW	9.0	30.0	88.0	68.0	
42188	2011-05-04	Wollongong	14.6	19.3	0.0	NaN	NaN	SSW	39.0	SSW	SSE	26.0	22.0	66.0	63.0	
42189	2011-05-05	Wollongong	13.3	17.3	0.0	NaN	NaN	S	56.0	SSW	S	22.0	33.0	57.0	60.0	
42190	2011-05-06	Wollongong	11.7	17.6	0.0	NaN	NaN	SE	33.0	SW	SE	17.0	24.0	56.0	57.0	

42191 rows × 24 columns

Fig 2.2.dataset

The irrelevant features in the above dataset is mentioned below:

1.Sunshine with 43% of null values.
2.Evaporation with 48% of null values.
3.cloud 3pm with 43% of null values.
4.cloud 9am with 38% of null values.

3.Methodology

Methodology is nothing but used methods and AI algorithms in our project here we are discussing the algorithms used in our project brefily.The algorithms used in our project are discussed below.Here we used the seven algorithms in order to predict the best one based on the accuracy percentage they got.the algorithms are:

1.KNN
2.Random Forest classifier
3.Logistic Regression
4.Gradient Boosting classifier
5.Adaboost
6.Decision Tree
7.Catboost

Now we will see about these algorithms one by one in detail brefily.

1.KNN:

The full form of KNN is K-Nearest Neighbour .This algorithm is one of the simplest Machine Learning algorithm.This comes under the Supervised Machine Learning Technique.This Algorithm can be used for both regression and classification.Among those two mostly this is used for classification problems.This algorithm can also be called as the **LAZY LEARNING** Algorithm.

2.Random Forest Classifier:

This is one of the popular Machine Learning Algorithm which comes under the Supervised Machine Learning Technique.In this Random Forest Classifier these will produce the more no.of tress among those trees we have to take the best tree that gives the more accuracy.

3.Logistic Regression:

This Logistic Regression is an example for the Supervised Machine Learning.This Algorithm mostly use to predict the probability for the occurring of binary event.There are three types of Logistic Regression.These are mentioned below:

1.Binary Logistic Regression
2.Multinomial Logistic Regression
3.Ordinal Logistic Regression

These are the three types of LOGISTIC REGRESSION.

4.Gradient Boosting Classifier:

This algorithm is a machine learning technique which is used in classification and regression.This Classifier is present in the ensemble model.This gives the outcome as the binary tree.Based on those we need to take the best part which we will get the less accuracy.In this Algorithm we will use the important parameter named **shrinkage.**

This Gradient Boosting Classifier is the Supervised Machine Learning Algorithm.

5.Adaboost Classifier:

This Adaboost Algorithm is a Boosting Technique this can be find in the Ensemble Method in Machine Learning.This Adaboost can be called as the Adaptive Boosting Algorithm.This Algorithm is First Successful boosting algorithm.This algorithm is developed for binary classification purpose.This is very important boosting technique.this converts the multiple "weak classifiers" into single "strong classifier".

6.Decision Tree Classifier:

This Decision Tree Algorithm is a Supervised Machine Learning Algorithm.This can be used for both Classification and Regression.Mostly we use this for Classification problems.The format for this is tree-structured format.There will be two kinds of nodes.These are:

1.Decision Node

2.Leaf Node

In the Decision Node ther will be extention of tree,where as for the Leaf Node there will be no extention.This will consider as the final output.

7.Catboost Classifier:

This catboost Classifier is an open-source library.This Algorithm comes under the gradient Boosting classifer.where we can use the decision tree.this algorithm is developed by **YANDEX RESEARCHERS AND ENGINEERS**.This catboost classifier algorithm can be used easily.

DATA PREPROCESSING

The data preprocessing is nothing but which is used to convert the raw data into the clean dataset.For example rawdata is nothing but having the null values.The machine Lerning Algorithm can not understand those null values our aim is to remove those null values.For this process of removing null values we will use the data cleaning step in the data preprocessing steps.The data preprocessing can be applied to the dataset before we use this dataset in our algorithm.Like wise also the Ranforest Algorithm can not perform analysis if the dataset contains the null values.The data preprocessing can also be used in order to format our dataset in particular way.The steps involved in the data preprocessing are mentioned as shown below.

1. Having Dataset
2. Import Required Libraries
3. Loading Dataset
4. Identifying Missing Data
5. Encoding Categorical Data
6. Splitting Dataset into Train and Test Datasets.
7. Feature Scaling.

These are seven steps involved in the **data preprocessing** process.After completion of these seven steps we call this dataset as the clean dataset.Now this dataset can used inour required Machine Learning Algorithms.

4.Experiments And Results

In this final step we are going to evaluate the accuracy for the Australian Dataset by using the different machine learning algorithms.The Algorithms we used are KNN,Random forest,Decision Tree,Catboost,Adaboost,Gradient Boosting,Logistic Regression.

Before this we need to do the Data Preprocessing step.we need to train and test our dataset set to get the accurate results.In this step we will find which algorithm is best to use in our project based upon the accuracy score we get for different machine learning algorithms.

Now,we will observe code for the Catboost algorithm and the same code is used for all the algorithms but,we need to change the importing statements.The sample code is provided below:

```
From sklearn.ensemble import CatBoostClassifier
model =CatBoostClassifier(iterations=2000, eval_metric =
"AUC")
model.fit(x_train, y_train)

y_pred= model.predict(x_test)

from sklearn.metrics import accuracy_score
ac = accuracy_score(y_pred, y_test)
#output - 0.86
```

Now,we will observe Accuracy for all the algorithms.

ALGORITHM	ACCURACY
Logistic Regression	79%
Decision Tree	73%
Random Forest	81%
KNN	80%
Gradient Boosting	81%
AdaBoost	80%
CatBoost	86%

By observing above table comparing algorithms we observe that catBoost classifier has highest accuracy and Decision tree has least accuracy.So for our project we took CatBoost classifier Algorithm.

CONCLUSION

In this work, we explored and applied many preprocessing techniques to find out how they impacted the overall performance of our classifiers. We also compared every classifier using different inputs, making note of how the entering data can affect the predictions made by the model.

We can infer that Australian weather is erratic and that there is no connection between rainfall and a certain location or time. We found a number of links and trends in the data, allowing us to pinpoint important traits.

Because of the large quantity of data we have, we may employ Deep Learning models like Multilayer Perceptrons, Convolutional Neural Networks (CNN), and others. It would be great to compare Deep Learning models and Machine Learning classifiers.

REFERENCES

1. World Health Organization: Climate Change and Human Health: Risks and Responses. World Health Organization, January 2003
2. Alcntara-Ayala, I.: Geomorphology, natural hazards, vulnerability and prevention of natural disasters in developing countries. Geomorphology 47(24), 107124 (2002)
3. Nicholls, N.: Atmospheric and climatic hazards: Improved monitoring and prediction for disaster mitigation. Natural Hazards 23(23), 137155 (2001)
4. [Online] InDataLabs, Exploratory Data Analysis: the Best way to Start a Data Science Project. Available: https://medium.com/@InDataLabs/ why-start-a-data-science-project-with-exploratory-data-analysis-f90c0efcbe49
5. [Online] Pandas Documentation. Available: https://pandas.pydata.org/ pandas-docs/stable/reference/api/pandas.get_dummies.html
6. [Online] Sckit-Learn Documentation Available: https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.FeatureHasher. html
7. [Online] Sckit-Learn Documentation Available: https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html