VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

RAINFALL PREDICTION

Name: Kush Desai, Raginee Titar, Raviraj Patil

Reg. no: 20MCB1002, 20MCB1004, 20MCB1016

Subject: Domain Specific Predictive Analytics(CSE6021)

OVERVIEW

- Objective
- Supervised learning
- Binary classification
- Methodology used for binary classification:
- Dataset description:
- Sample Dataset
- Procedure to predict the rain tomorrow
- Data preparation visualization
- Linear Regression to Rainfall Prediction
- Decision tree Regression to Rainfall Prediction
- Random Forest Regression to Rainfall Prediction
- Conclusion

OBJECTIVE:

Rainfall has a significant impact on society. Festivals, all sorts of activities, and sports matches are examples of events that could be heavily influenced by rainfall. Therefore, this topic is studied heavily. The goal of this analysis is to predict if it is going to rain the next day, based on the weather measures of the day before. This creates a binary classification problem (1 = Rain and 0 = No Rain). The dataset consists of weather data in Australia over 10 years (1 Nov 2007 to 25 Jun 2017).

SUPERVISED LEARNING

In supervised learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data by associating patterns to the unlabeled new data. Supervised learning can be divided into two categories: classification and regression.

Supervised learning can be divided into two categories: classification and regression.

Classification is a technique for determining which class the dependent belongs to based on one or more independent variables.

Logistic regression is kind of like linear regression, but is used when the dependent variable is not a number but something else (e.g., a "yes/no" response). It's called regression but performs classification based on the regression and it classifies the dependent variable into either of the classes.

BINARY CLASSIFICATION:

- Binary classification is the task of classifying the elements of a set into two groups on the basis of a classification rule. Accurate Sales Forecasts enable companies to make informed business decisions and predict short-term and long-term performance.
- Binary classification is dichotomization applied to a practical situation. In many practical binary classification problems, the two groups are not symmetric, and rather than overall accuracy, the relative proportion of different types of errors is of interest. For example, in medical testing, detecting a disease when it is not present (a false positive) is considered differently from not detecting a disease when it is present (a false negative).

METHODOLOGY USED FOR BINARY CLASSIFICATION:

The following are the algorithms that is be used binary classification;

- Decision Tree
- Logistic Regression
- Random Forest

DATASET DESCRIPTION:

Name: Rain in Australia

Size: 13.44 MB

Source: Kaggle

DATASET NAME:

weatherAUS.csv

ATTRIBUTES:

- **DATE**- The date represents each day Rainfall

- **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. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.

- **Cloud3pm-**Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values

- **Temp9am-**Temperature (degrees C) at 9am

- **Temp3pm-**Temperature (degrees C) at 3pm

- **RainToday-**Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0

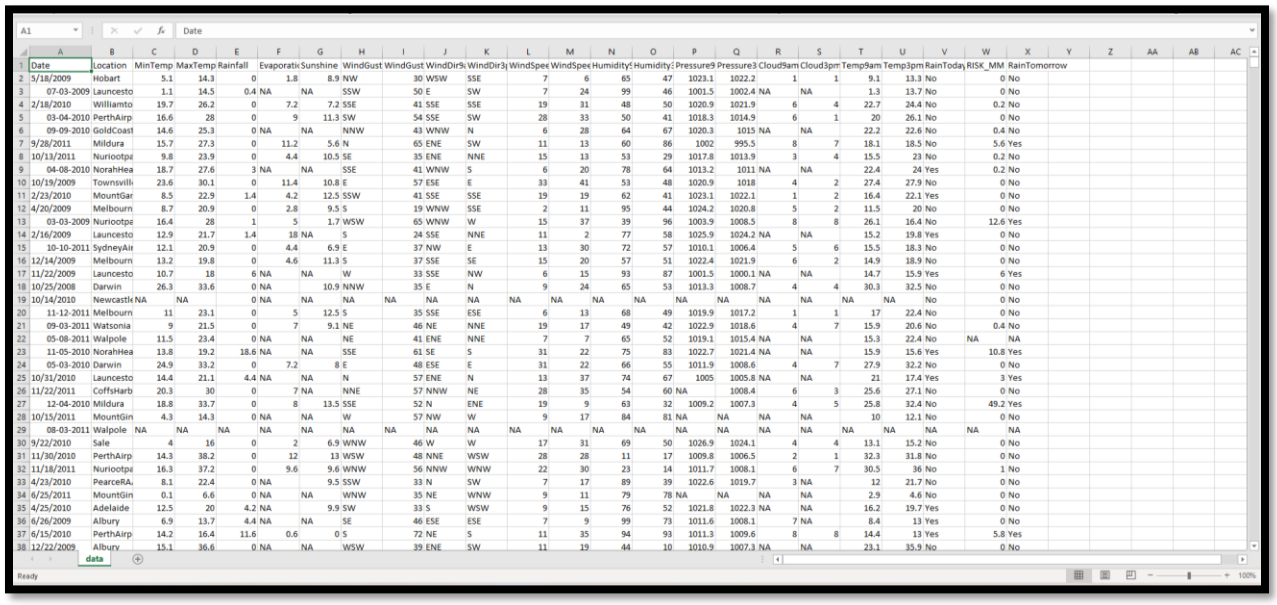
- **RainTomorrow-**The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".

ABOUT DATASET:

This dataset contains about 10 years of daily weather observations from many locations across Australia.

RainTomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

SAMPLE DATASET:



Date	Location	MinTemp	MaxTemp	Rainfall	Evaporatio	Sunshine	WindGust	WindGust	WindDir9	WindDir3	WindSpee	WindSpee	Humidity9	Humidity3	Pressure9	Pressure3	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RISK_MM	RainTomorrow
5/18/2009	Hobart	5.1	14.3	0	1.8	8.9	NW	30	WSW	SSE	7	6	65	47	1023.1	1022.2	1	1	9.1	13.3	No	0	No
07-03-2009	Launcesto	1.1	14.5	0.4	NA	NA	SSW	50	E	SW	7	24	99	46	1001.5	1002.4	NA	NA	1.3	13.7	No	0	No
2/18/2010	Williamto	19.7	26.2	0	7.2	7.2	SSE	41	SSE	SSE	19	31	48	50	1020.9	1021.9	6	4	22.7	24.4	No	0.2	No
03-04-2010	PerthAirp	16.6	28	0	9	11.3	SW	54	SSE	SW	28	33	50	41	1018.3	1014.9	6	1	20	26.1	No	0	No
09-09-2010	GoldCoast	14.6	25.3	0	NA	NA	NNW	43	WNW	N	6	28	64	67	1020.3	1015	NA	NA	22.2	22.6	No	0.4	No
9/28/2011	Mildura	15.7	27.3	0	11.2	5.6	N	65	ENE	SW	11	13	60	86	1002	995.5	8	7	18.1	18.5	No	5.6	Yes
10/13/2011	Nuriootpa	9.8	23.9	0	4.4	10.5	SE	35	ENE	NNE	15	13	53	29	1017.8	1013.9	3	4	15.5	23	No	0.2	No
04-08-2010	NorahHea	18.7	27.6	3	NA	NA	SSE	41	WNW	S	6	20	78	64	1013.2	1011	NA	NA	22.4	24	Yes	0.2	No
10/19/2009	Townsvill	23.6	30.1	0	11.4	10.8	E	57	ESE	E	33	41	53	48	1020.9	1018	4	2	27.4	27.9	No	0	No
2/23/2010	MountGar	8.5	22.9	1.4	4.2	12.5	SSW	41	SSE	SSE	19	19	62	41	1023.1	1022.1	1	2	16.4	22.1	Yes	0	No
4/20/2009	Melbourn	8.7	20.9	0	2.8	9.5	S	19	WNW	SSE	2	11	95	44	1024.2	1020.8	5	2	11.5	20	No	0	No
03-03-2009	Nuriootpa	16.4	28	1	5	1.7	WSW	65	WNW	W	15	37	39	96	1003.9	1008.5	8	8	26.1	16.4	No	12.6	Yes
2/16/2009	Launcesto	12.9	21.7	1.4	18	NA	S	24	SSE	NNE	11	2	77	58	1025.9	1024.2	NA	NA	15.2	19.8	Yes	0	No
10-10-2011	SydneyAir	12.1	20.9	0	4.4	6.9	E	37	NW	E	13	30	72	57	1010.1	1006.4	5	6	15.5	18.3	No	0	No
12/14/2009	Melbourn	13.2	19.8	0	4.6	11.3	S	37	SSE	SE	15	20	57	51	1022.4	1021.9	6	2	14.9	18.9	No	0	No
11/22/2009	Launcesto	10.7	18	6	NA	NA	W	33	SSE	NW	6	15	93	87	1001.5	1000.1	NA	NA	14.7	15.9	Yes	6	Yes
10/25/2008	Darwin	26.3	33.6	0	NA	10.9	NNW	35	E	N	9	24	65	53	1013.3	1008.7	4	4	30.3	32.5	No	0	No
10/14/2010	Newcastle	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	No	0	No
11-12-2011	Melbourn	11	23.1	0	5	12.5	S	35	SSE	ESE	6	13	68	49	1019.9	1017.2	1	1	17	22.4	No	0	No
09-03-2011	Watsonia	9	21.5	0	7	9.1	NE	46	NE	NNE	19	17	49	42	1022.9	1018.6	4	7	15.9	20.6	No	0.4	No
05-08-2011	Walpole	11.5	23.4	0	NA	NA	NE	41	ENE	NNE	7	7	65	52	1019.1	1015.4	NA	NA	15.3	22.4	No	NA	NA
11-05-2010	NorahHea	13.8	19.2	18.6	NA	NA	SSE	61	SE	S	31	22	75	83	1022.7	1021.4	NA	NA	15.9	15.6	Yes	10.8	Yes
05-03-2010	Darwin	24.9	33.2	0	7.2	8	E	48	ESE	E	31	22	66	55	1011.9	1008.6	4	7	27.9	32.2	No	0	No
10/31/2010	Launcesto	14.4	21.1	4.4	NA	NA	N	57	ENE	N	13	37	74	67	1005	1005.8	NA	NA	21	17.4	Yes	3	Yes
11/22/2011	CoffsHarb	20.3	30	0	7	NA	NNE	57	NNW	NE	28	35	54	60	NA	1008.4	6	3	25.6	27.1	No	0	No
12-04-2010	Mildura	18.8	33.7	0	8	13.5	SSE	52	N	ENE	19	9	63	32	1009.2	1007.3	4	5	25.8	32.4	No	49.2	Yes
10/15/2011	MountGin	4.3	14.3	0	NA	NA	W	57	NW	W	9	17	84	81	NA	NA	NA	NA	10	12.1	No	0	No
08-03-2011	Walpole	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9/22/2010	Sale	4	16	0	2	6.9	WNW	46	W	W	17	31	69	50	1026.9	1024.1	4	4	13.1	15.2	No	0	No
11/30/2010	PerthAirp	14.3	38.2	0	12	13	WSW	48	NNE	WSW	28	28	11	17	1009.8	1006.5	2	1	32.3	31.8	No	0	No
11/18/2011	Nuriootpa	16.3	37.2	0	9.6	9.6	WNW	56	NNW	WNW	22	30	23	14	1011.7	1008.1	6	7	30.5	36	No	1	No
4/23/2010	PearceRA	8.1	22.4	0	NA	9.5	SSW	33	N	SW	7	17	89	39	1022.6	1019.7	3	NA	12	21.7	No	0	No
6/25/2011	MountGin	0.1	6.6	0	NA	NA	WNW	35	NE	WNW	9	11	79	78	NA	NA	NA	NA	2.9	4.6	No	0	No
4/25/2010	Adelaide	12.5	20	4.2	NA	9.9	SW	33	S	WSW	9	15	76	52	1021.8	1022.3	NA	NA	16.2	19.7	Yes	0	No
6/26/2009	Albury	6.9	13.7	4.4	NA	NA	SE	46	ESE	ESE	7	9	99	73	1011.6	1008.1	7	NA	8.4	13	Yes	0	No
6/15/2010	PerthAirp	14.2	16.4	11.6	0.6	0	S	72	NE	S	11	35	94	93	1011.3	1009.6	8	8	14.4	13	Yes	5.8	Yes
12/22/2009	Albury	15.1	36.6	0	NA	NA	WSW	39	ENE	SW	11	19	44	10	1010.9	1007.3	NA	NA	23.1	35.9	No	0	No

PROCEDURE TO PREDICT THE RAIN TOMORROW:

STEP 1: We imported the packages namely NumPy, pandas, matplot lib, seaborn, sklearn.

STEP 2: Data Preparation

In Data Preparation we undergo following steps namely,

2.1 Load dataset

2.2 Dealing with missing value

2.3 Count missing values in each column

2.4 Replacing the missing values in features with mean, median and drop null value.

STEP 3: Feature Selection

Select Feature Importance using Filter Method (Chi-Square)

STEP 4: Exploratory Data Analytics

5.1 Correlation Heatmap

STEP 6: Rain tomorrow Prediction

Here we apply various classification techniques to predict the sales.

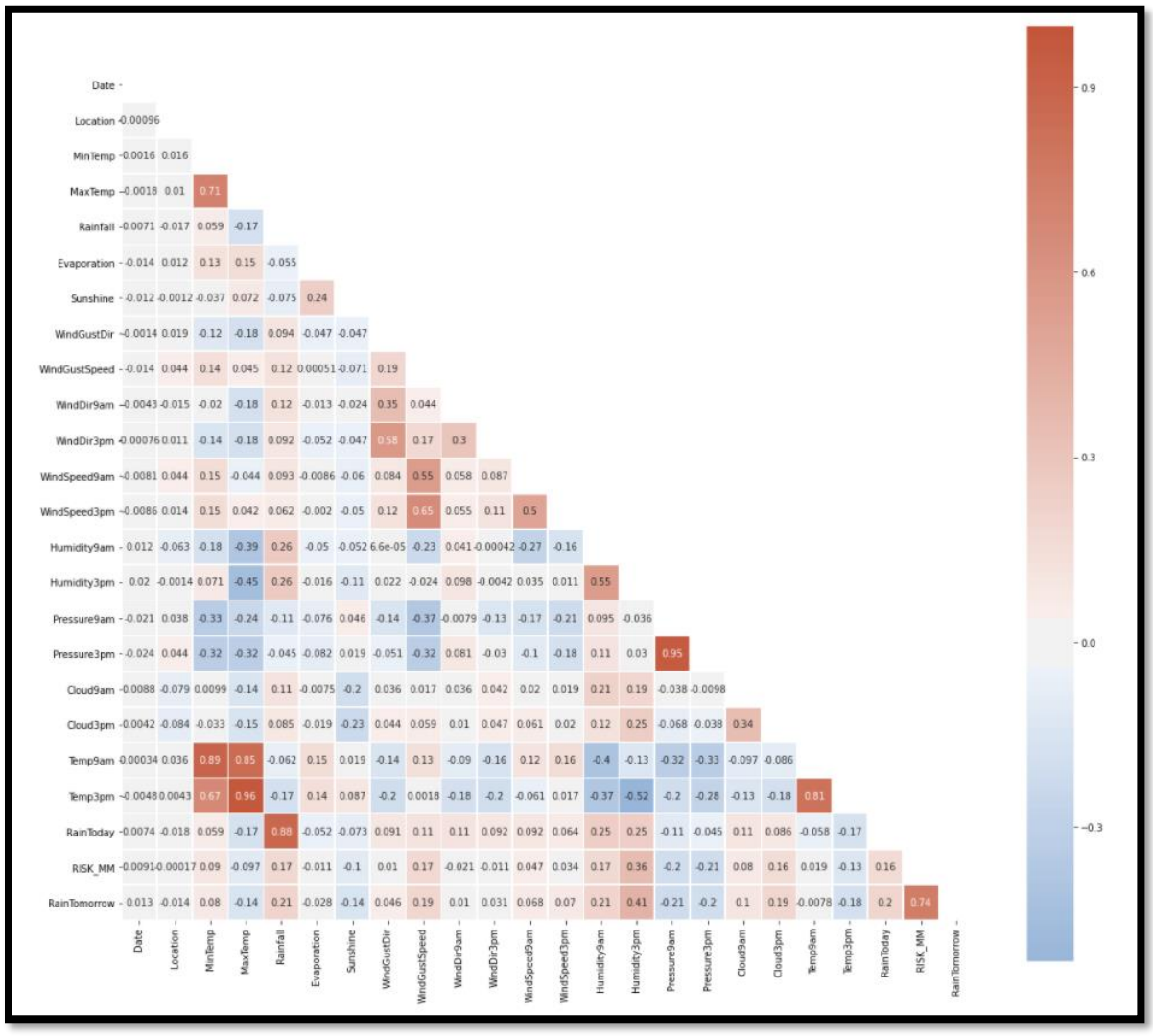
6.1 Logistic Regression

6.2 Decision Tree

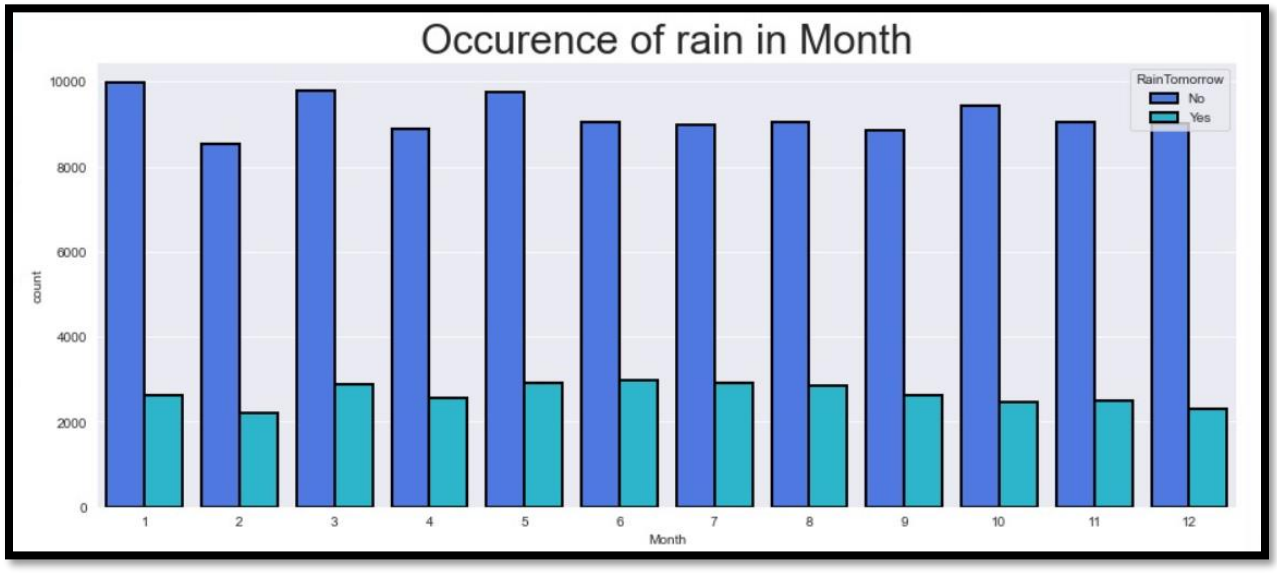
6.3 Random Forest

DATA PREPERATION VISUALIZATION

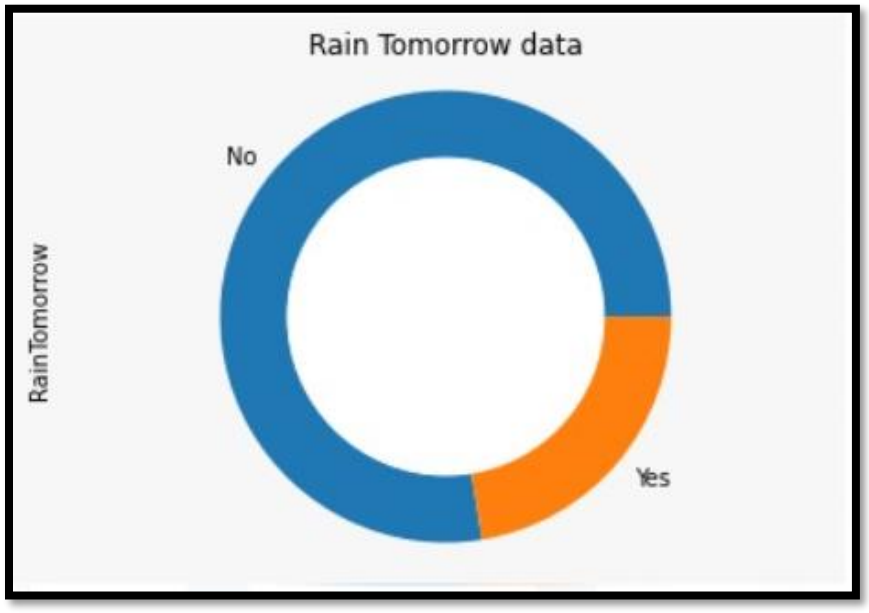
CORRELATION HEATMAP



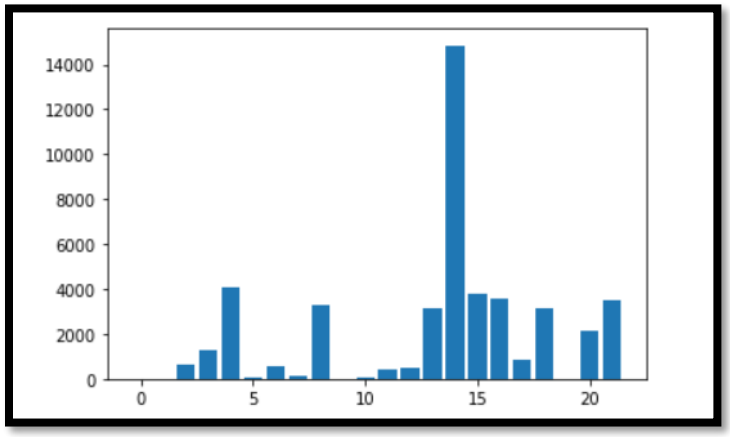
OCCURRENCE OF RAIN IN MONTH



CLASS DISTRIBUTION



FEATURE SELCION USING Chi-Square METHOD



LOGISTIC REGRESSION TO RAINFALL PREDICTION:

- To perform logistic regression, the sigmoid function. In a binary logistic regression model, the dependent variable has two levels (IN this case yes or no). Outputs with more than two values are modeled by multinomial logistic regression and, if the multiple categories are ordered, by ordinal logistic regression.
- The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification, though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.

DECISION TREE REGRESSION TO RAINFALL PREDICTION:

- Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.
- A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

RANDOM FOREST REGRESSION TO RAINFALL PREDICTION:

- The random forest is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a “forest”)
- The random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features. The final predictions of the random forest are made by averaging the predictions of each individual tree.

SAMPLE CODE:

LOGISTICREGRESSION

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.metrics import precision_recall_fscore_support
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

X_train, X_test, y_train, y_test =
train_test_split(features,target,test_size=0.1,random_state=1)
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
y_pred= logreg.predict(X_test)


accuracy = accuracy_score(y_test, y_pred)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
print("Precision Score:-", metrics.precision_score(y_test, y_pred))
print("Recall Score:-", metrics.recall_score(y_test, y_pred))
print("F1 Score:-", metrics.f1_score(y_test, y_pred))
#print("Average Precision Score:-",
metrics.average_precision_score(y_test, predictions))
#print("Log Loss:-", metrics.log_loss(y_test, predictions))
print("ROC-AUC Score:-", metrics.roc_auc_score(y_test, y_pred))
#print(precision_recall_fscore_support(y_test, predictions,
average='binary'))
```

Decision Tree:

```
from sklearn import metrics
from sklearn.metrics import precision_recall_fscore_support
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
X_train, X_test, y_train, y_test =
train_test_split(features,target,test_size=0.1,random_state=1)

dtree=DecisionTreeClassifier()
dtree.fit(X_train,y_train)
y_pred = dtree.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
print("Precision Score:-", metrics.precision_score(y_test, y_pred))
print("Recall Score:-", metrics.recall_score(y_test, y_pred))
print("F1 Score:-", metrics.f1_score(y_test, y_pred))
#print("Average Precision Score:-",
metrics.average_precision_score(y_test, predictions))
#print("Log Loss:-", metrics.log_loss(y_test, predictions))
print("ROC-AUC Score:-", metrics.roc_auc_score(y_test, y_pred))
```

```
#print(precision_recall_fscore_support(y_test, predictions,
average='binary'))
```

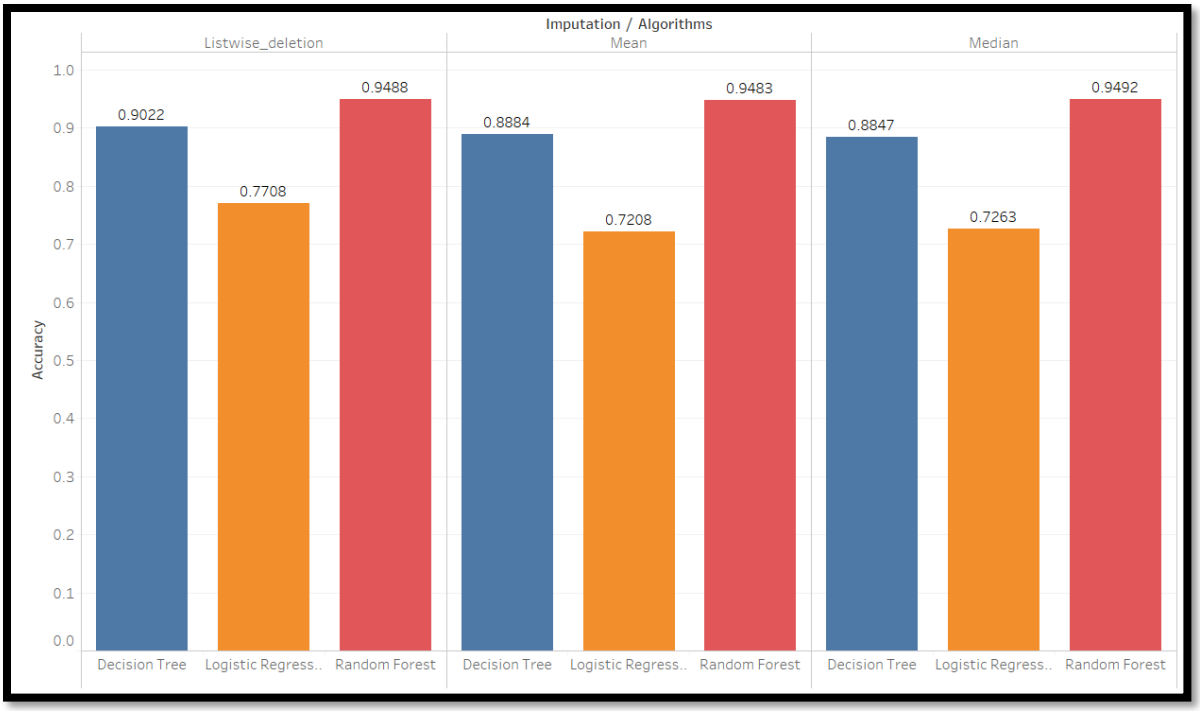
RANDOM FOREST

```
from sklearn import metrics
from sklearn.metrics import precision_recall_fscore_support
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score

x_train, x_test, y_train, y_test= train_test_split(features, target,
test_size= 0.1, random_state=1)
#class_weight={0:1,1:3}
#feature Scaling
st_x= StandardScaler()
x_train= st_x.fit_transform(x_train)
x_test= st_x.transform(x_test)

classifier= RandomForestClassifier(n_estimators= 100,
criterion="entropy",class_weight='balanced')
classifier.fit(x_train, y_train)
y_pred= classifier.predict(x_test)

accuracy = accuracy_score(y_test, y_pred)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
print("Precision Score:-", metrics.precision_score(y_test, y_pred))
print("Recall Score:-", metrics.recall_score(y_test, y_pred))
print("F1 Score:-", metrics.f1_score(y_test, y_pred))
#print("Average Precision Score:-",
metrics.average_precision_score(y_test, predictions))
#print("Log Loss:-", metrics.log_loss(y_test, predictions))
print("ROC-AUC Score:-", metrics.roc_auc_score(y_test, y_pred))
#print(precision_recall_fscore_support(y_test, predictions,
average='binary'))
```

OUTPUT

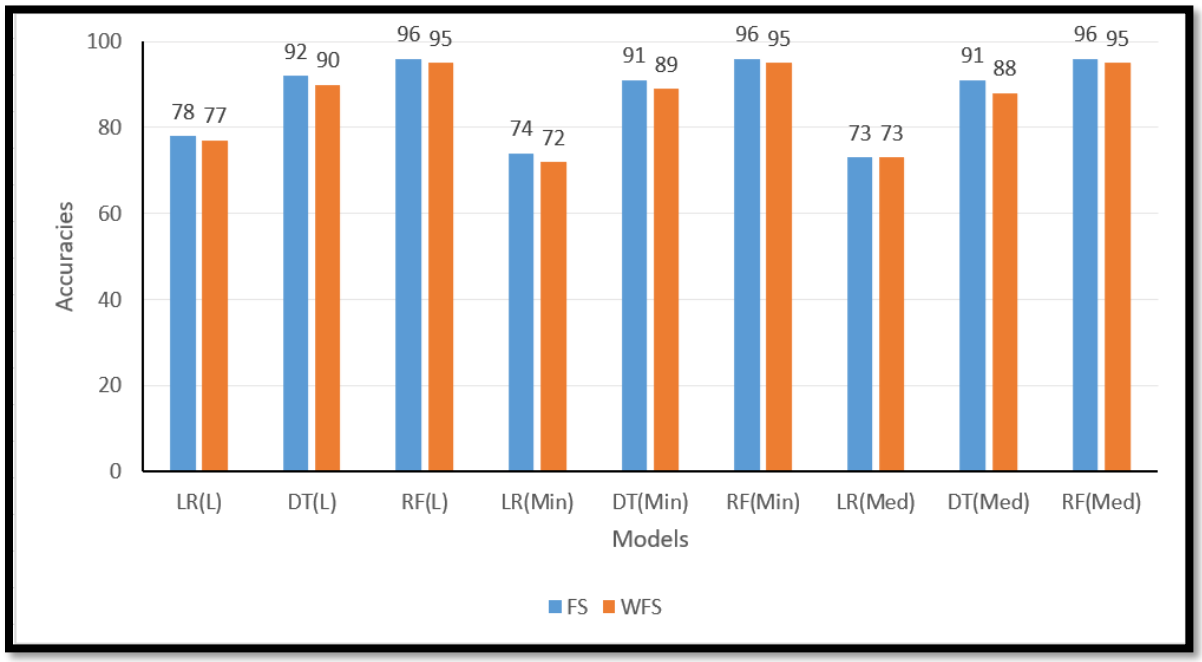
1) Without Feature Selection

Algorithms	Accuracy			Precision score			Recall score			F1 Score		
Model	Mean	Median	Drop null	Mean	Median	Drop null	Mean	Median	Drop null	Mean	Median	Drop null
Logistic Regression	72	73	77	0.70	0.73	0.77	0.55	0.55	0.64	0.62	0.62	0.98
Decision Tree	89	88	90	0.81	0.88	0.90	0.95	0.95	0.96	0.87	0.87	0.88
Random Forest	95	95	95	0.92	0.95	0.95	0.95	0.95	0.96	0.94	0.94	0.94



2)With Feature Selection

Algorithms	Accuracy			Precision score			Recall score			F1 Score		
Model	Mean	Median	Drop null	Mean	Median	Drop null	Mean	Median	Drop null	Mean	Median	Drop null
Logistic Regression	74	73	78	0.71	0.69	0.74	0.59	0.58	0.66	0.64	0.63	0.70
Decision Tree	91	91	92	0.83	0.83	0.84	0.97	0.97	0.98	0.90	0.89	0.90
Random Forest	96	96	96	0.94	0.93	0.93	0.97	0.97	0.98	0.95	0.95	0.95



LR: Logistic Regression
DT: Decision Trees
RF: Random Forest
(L): Missing Value imputation using Listwise deletion
(Min): Missing Value imputation using Mean
(Med): Missing Value imputation using Median
FD: With feature selection
WFS: Without Feature Selection

Inference: Random Forest gives best accuracy for all types of missing value imputation.

CONCLUSION:

In this project, handled missing values by using the mean, listwise deletion, and median imputation. Over-sampling method was used to create “synthetic” data to balance dataset. For feature selection Chi-Square method is used to select best top 15 columns to get better accuracy. We consider different binary classification machine learning approach for tomorrow rain fall prediction. From the all the three model Random Forest has got the best accuracy i.e., 96% accuracy.