Rainfall Prediction Project

T-28 | CSE-343 | Monsoon 2021

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

- Humans need water for their survival, and Rainfall plays an important role to fulfill their needs.
- It also plays an important aspect in a Country's GDP.
- Although there are a plethora of Rainfall issues around the globe, floods is an important issue causing a lot of devastation of life and property.
- Saving human life and property becomes a challenging task at the time of floods and heavy rainfall.
- Unsuccessful rainfall predictions can lead to huge losses, hence, it becomes a challenging problem to solve.

Our motive is to try-out different ML models to make correct predictions.

Literature Review – I

Rainfall Prediction using ML and NN (link)

Methodology

- Used LASSO regression and ANN
- Pick better algorithm based on accuracy
- Various metrics and graphical analysis is presented
- LASSO regression is picked as the better model than ANN

Limitation

There was a high percentage of error while calculating the accuracy through different methods of evaluation(metrics). This system is very complex which makes the computational costs very high.

Literature Review – II

Prediction of Rainfall Using Machine Learning Techniques (link)

Methodology

- Use of ANN, classification of categories via Neuro-Fuzzy
- Statistical method using multiple regressions for value predictions using descriptive variables
- Linear relationship between descriptive variable and output values is observed.
- β_0 and β_p are the constant intercept and slope of the descriptive variable respectively.
- Proposed method predicts rainfall for Indian dataset using multiple linear regressions and provides improved results wrt accuracy, MSE and correlation

Limitation

Assumptions was made by the multiple linear regressions that there is a linear relationship between both the descriptive and independent variables. Also, the highly correlated variables are considered independent variables,

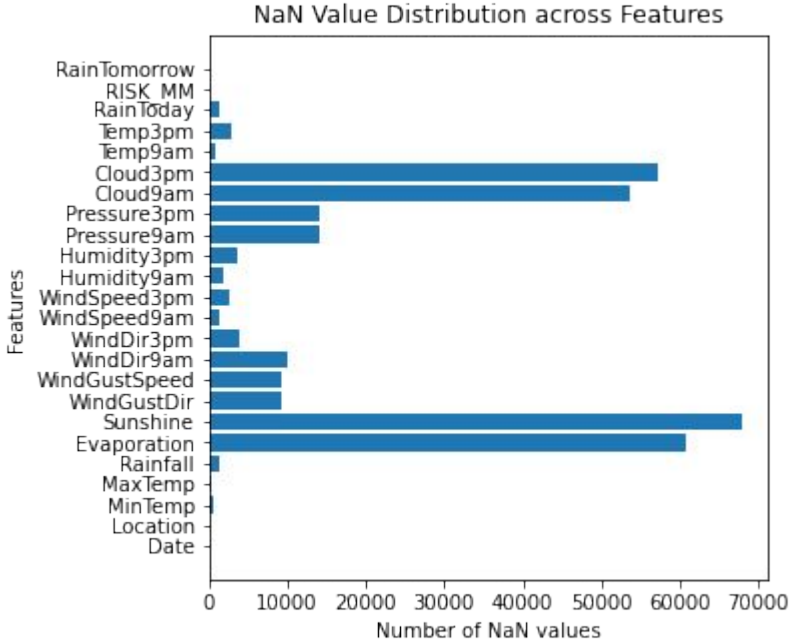
Dataset Description

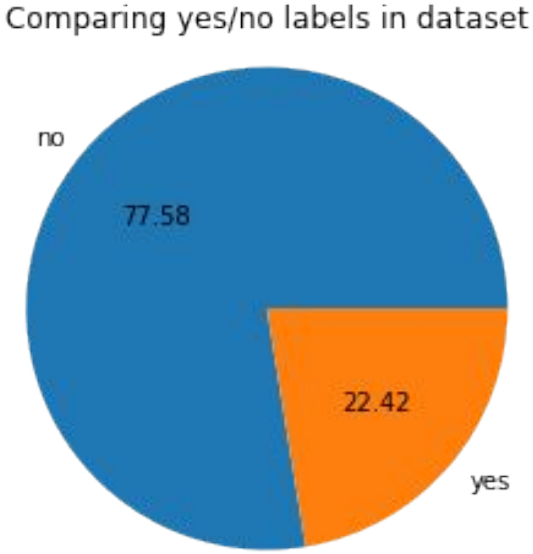
10 years of daily weather observations in locations across Australia

FEATURES

Location	WindDir9am (dir)	Date
MinTemp (°C)	WindDir3pm (dir)	Cloud9am (oktas)
MaxTemp (°C)	WindSpeed9am (k/h)	Cloud3pm (oktas)
Rainfall (in mm)	WindSpeed3pm (k/h)	Temp9am (°C)
Evaporation (in mm)	Humidity9am (%)	Temp3pm (°C)
Sunshine (hrs)	Humidity3pm (%)	RainToday (bool)
WindGustDir (dir)	Pressure9am (hpa)	Risk_MM (cont)
WindGustSpeed (k/h)	Pressure3pm (hPa)	RainTomorrow (bool)

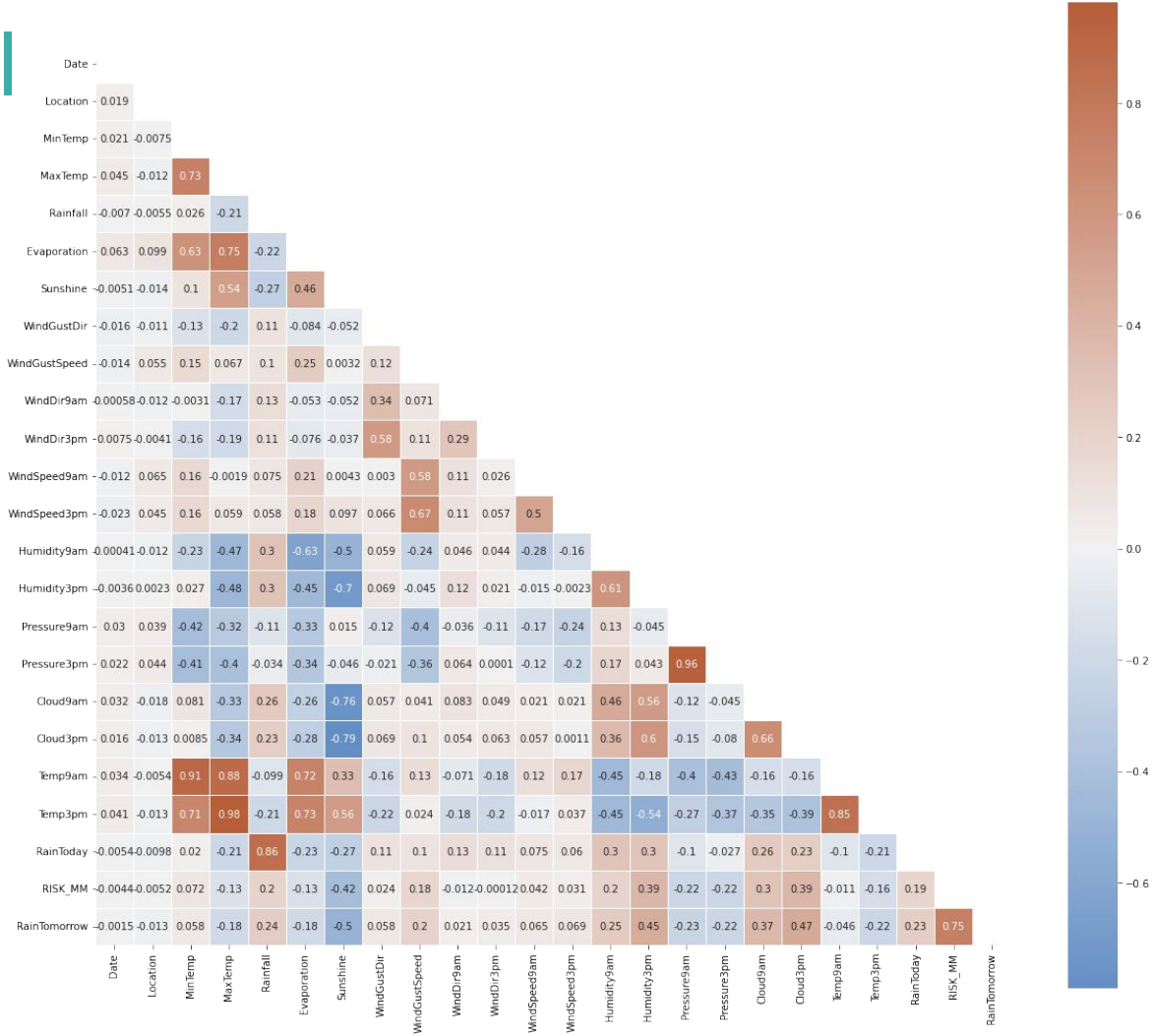
Visualizations-I



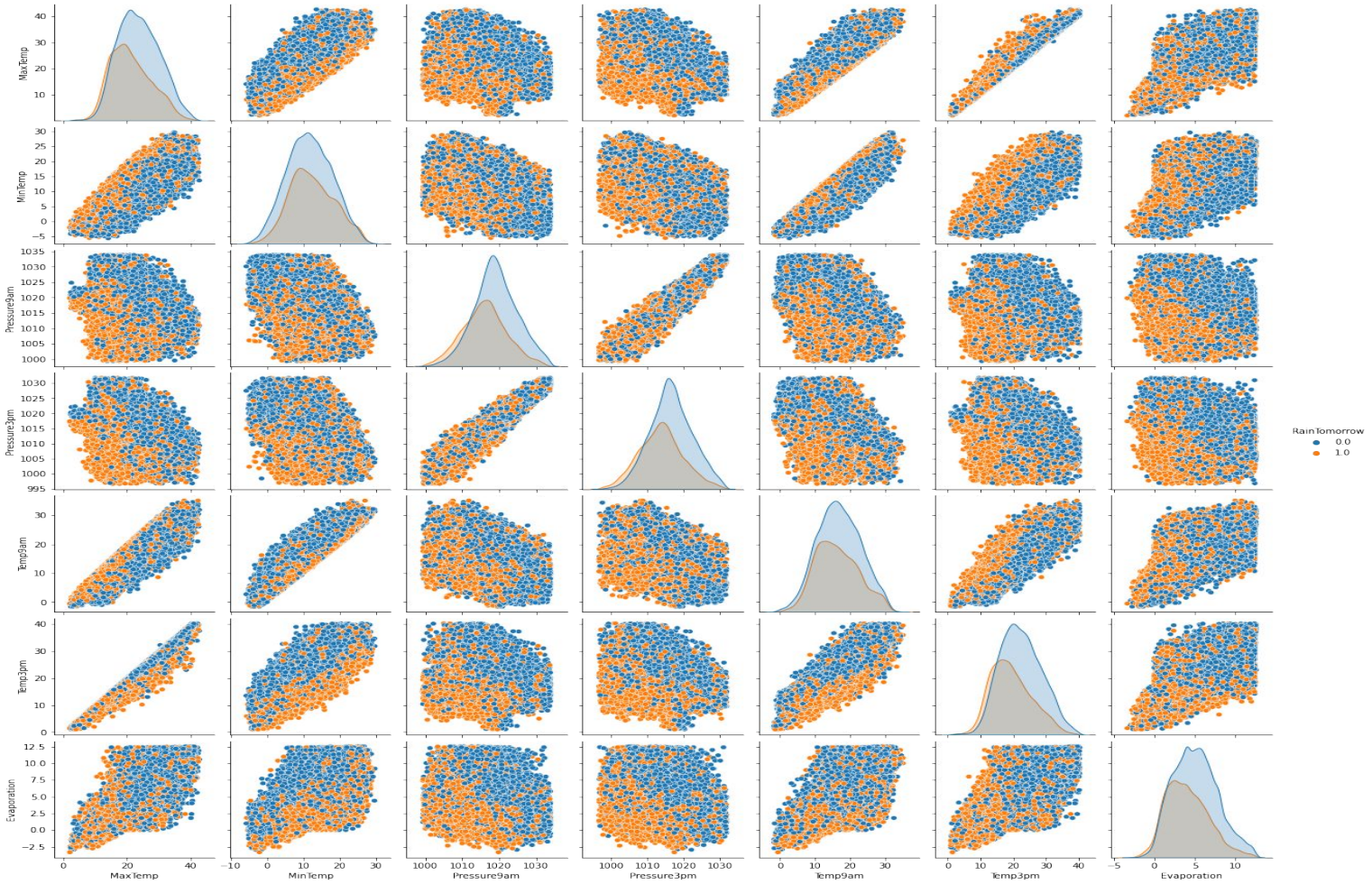


Visualizations-II

- Correlation matrix of all the features in input dataset
- Total **24 columns**, out of which **2 columns act as output label** and are correlated by a formula, hence only 1 of them is used
- **142193 (over)/34025 (under) data points**



Visualizations-III - PairPlot



Preprocessing

Over/Under sampling

Over/Under sampling data points to handle bias in data

Categorical Values

Convert all categorical values to continuous values using LabelEncoder

NaN Values

Replace NaN values with the mode and then use Iterative Imputer to fill them

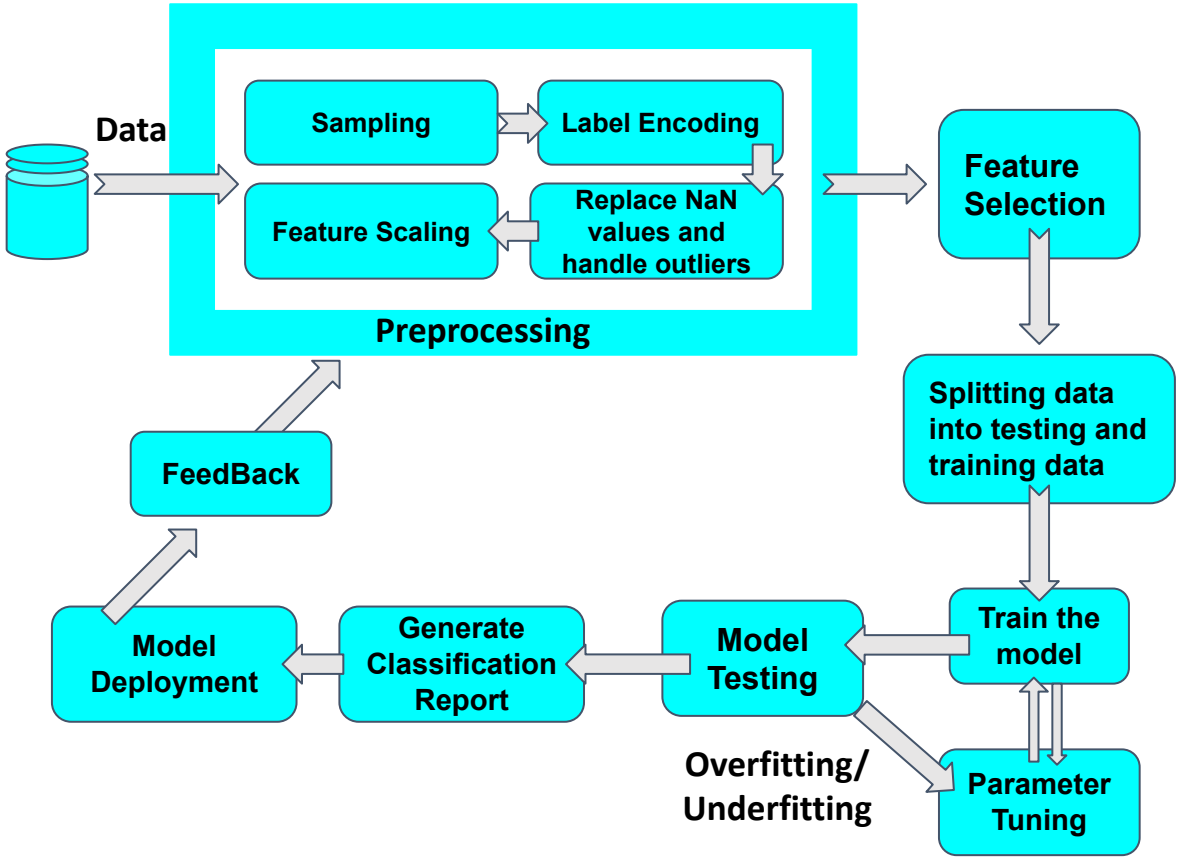
Handling Outliers

Calculate the IQR and removing values outside Q1 and Q3 beyond a threshold

Scaling

All data points are scaled using the Standard Scaler

Methodology

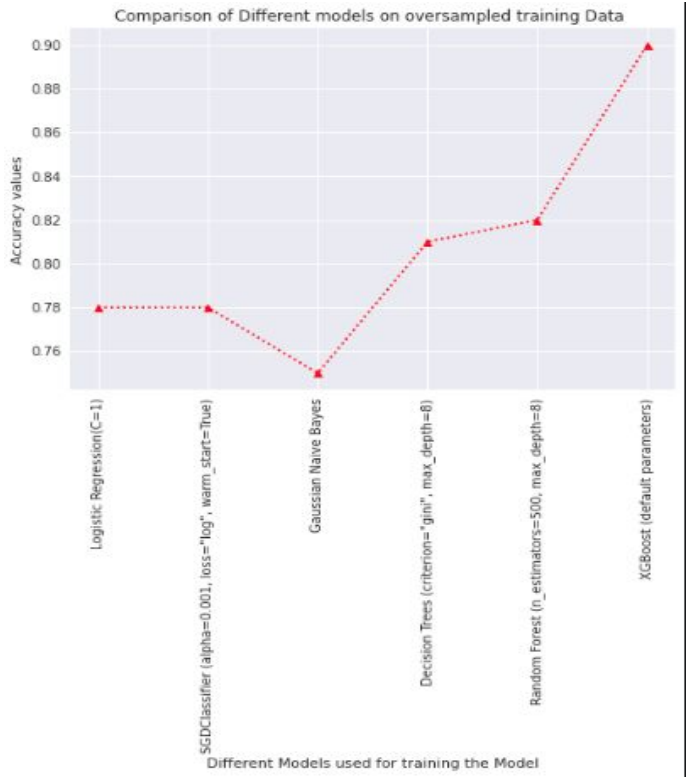


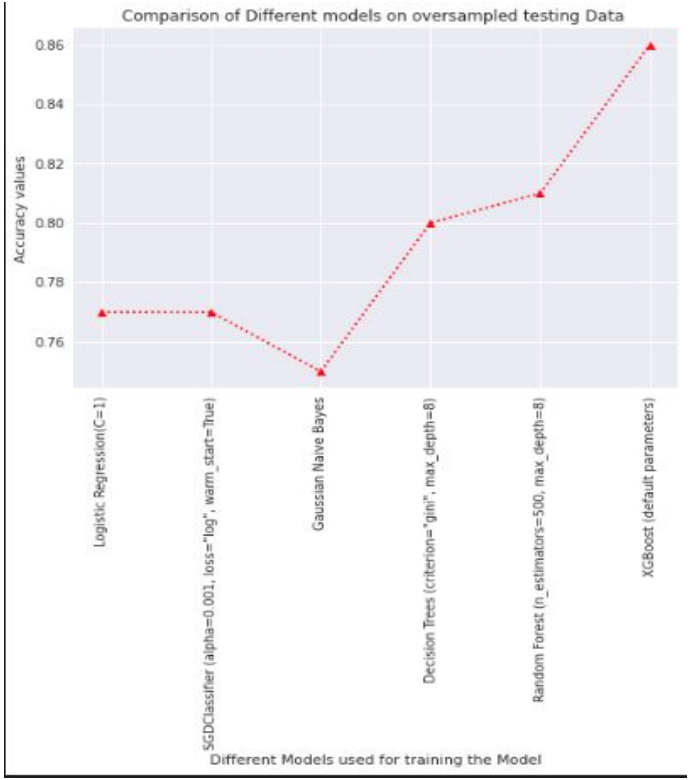
- Preprocessing
- select k best features
- train test split
- create model classifier
- apply grid search
- run on best parameters
- generate classification report

Results (Oversampled)

	Train Data			Test Data		
Model	**Accuracy**	**F1 Score**	**ROC AUC**	**Accuracy**	**F1 Score**	**ROC AUC**
Logistic Regression (C=0.1, solver="saga", penalty="l1")	0.78	0.77	0.86	0.78	0.77	0.86
SGDClassifier (alpha=0.001, loss="hinge", warm_start=True)	0.78	0.77	0.86	0.79	0.78	0.86
Gaussian Naive Bayes	0.75	0.74	0.82	0.75	0.75	0.82
Decision Trees (criterion="entropy", max_depth=8)	0.81	0.81	0.89	0.80	0.80	0.88
Random Forest (n_estimators=500, max_depth=8)	0.82	0.81	0.90	0.81	0.80	0.89
XGBoost (default parameters)	**0.90**	**0.91**	**0.97**	**0.86**	**0.85**	**0.93**

Comparison of Models (Oversampled)





Results (Undersampled)

	Train Data			Test Data		
Model	**Accuracy**	**F1 Score**	**ROC AUC**	**Accuracy**	**F1 Score**	**ROC AUC**
Logistic Regression (C=1)	0.78	0.76	0.85	0.77	0.75	0.85
SGDClassifier (alpha=0.001, loss="log", warm_start=True)	0.78	0.76	0.85	0.77	0.75	0.85
Gaussian Naive Bayes	0.75	0.73	0.81	0.74	0.73	0.81
Decision Trees (criterion="gini", max_depth=8)	0.82	0.80	0.89	0.79	0.77	0.85
Random Forest (n_estimators=500, max_depth=8)	0.83	0.81	0.90	0.80	0.79	0.88
XGBoost (default parameters)	**0.88**	**0.93**	**0.93**	**0.86**	**0.92**	**0.87**
AdaBoost(n_estimators=500)	0.79	0.72	0.77	0.78	0.70	0.76
SVM(kernel = “linear”, gamma =”auto”, C = 2)	0.90	0.89	0.88	0.86	0.91	0.86
K-nearest Neighbours(leaf_size=20, n_neighbors=10, p=1, weights='distance')				0.80	0.74	0.79

Results (Undersampled)

	Train Data			Test Data		
Model	**Accuracy**	**F1 Score**	**ROC AUC**	**Accuracy**	**F1 Score**	**ROC AUC**
Neural Network (solver=”adam”,activation=”relu”) [32,16,8]	0.84	0.83	0.93	0.84	0.83	0.92
Neural Network (solver=”adam”,activation=”tanh”) [32,16,8]	**0.84**	**0.83**	**0.92**	**0.84**	**0.83**	**0.92**
Neural Network (solver=”sgd”,activation=”relu”) [32,16,8]	0.82	0.80	0.90	0.83	0.81	0.90
Neural Network (solver=”sgd”,activation=”tanh”) [32,16,8]	0.82	0.81	0.90	0.83	0.82	0.90
Neural Network (solver=”adam”,activation=”tanh”) [64,32,16,8]	**0.87**	**0.86**	**0.94**	**0.84**	**0.83**	**0.92**
Neural Network (solver=”adam”,activation=”tanh”) [4,2]	0.81	0.79	0.89	0.82	0.81	0.89

Analysis(Oversampled)

saga solver with **L1 loss** is used for logistic regression due to **larger dataset size** and better convergence
hinge loss (max(0,1-t.y)) is used for **SGDClassifier** to penalize misclassified points more
Decision Trees and **Random Forests** work well with **8 as the max depth** to ensure **lower variance** than prior linear models
Boosting algorithm **XGBoost** performs the best so far due to boosting characteristics.

Top 10 features (by ANOVA-F values obtained using sklearn's feature_selection method)

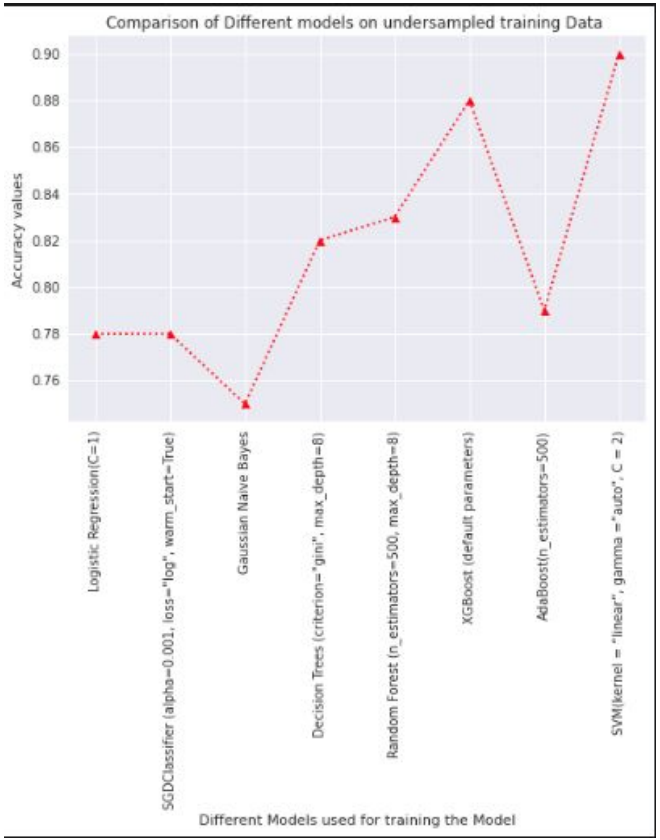
Rainfall	Sunshine
Humidity9am	Humidity3pm
Pressure9am	Pressure3pm
Cloud9am	Cloud3pm
RainToday	Temp3pm

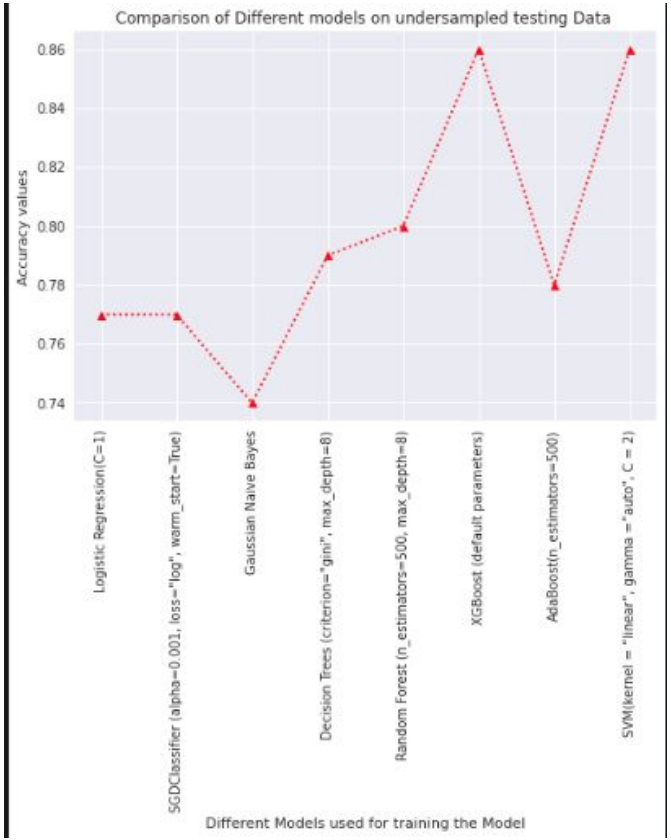
Analysis (Undersampled)

lbfgs solver with **L2 loss** is used for logistic regression due to **smaller dataset size** and better convergence.
log loss is used for **SGDClassifier** to penalize misclassified points more.
Decision Trees and **Random Forests** work well with **8 as the max depth** to ensure **lower variance** than prior linear models. Here, gini criteria is used for feature selection.
Boosting algorithm **XGBoost** performs the best so far due to boosting characteristics.

SVM makes correct predictions for all training points unlike the testing points. This is is happening because of overfitting of the model.
For **Neural Networks,** it is observed that **adam** is the better optimizer compared to **sgd** due to exponentially weighted averages.
For **Neural Networks,** increase in depth of the network results in better metrics, however, the training time also grows, therefore, this is a tradeoff which needs to be considered.

Comparison Of Models (Undersampled)





Conclusion

Based on the data analysis and application of various models, we come to the conclusion that **ensemble Machine Learning models** such as boosting perform well on the given dataset. This is because it works on improving the weak components of the ensemble model to improve the bias and variance. **Neural Networks** slightly outperform ML models but they still remain comparable.

Out of all given features, factors such as pressure, humidity, cloud cover, sunshine, temperature and rainfall today play the most important role in determining if it will rain tomorrow or not. The undersampling and/or oversampling datasets show similar results.

For further work, some of the correlated features can be dropped further and the accuracy can be compared. Also, a ratio can be devised using these parameters which can be checked for a strong correlation with rainfall for the next day. We can also extend this model to a multiclass classification with levels indicating the severity of rainfall.

Learnings

→ We learned about how we can solve real life problems using various machine learning techniques that help reduce the human effort and bringing to the world correct rainfall predictions.

→ From the technical perspective, we learnt how we can pre-process our data, analyse and make it more cleaner so that it can later be used to train different ML models.

→ We learnt about various ML models, how they work for both classification problems and regression problems, how we hypertune the parameters to get great results. It helped us get deep insights into the subject as well as get a firm grip over the subject.

Individual Contribution

Aniket (2019233)	Hardik (2019040)	Shabeg (2019388)	Rahul (2019266)
Gaussian Naive Bayes, Random Forest, Data preprocessing(correlation matrix, feature selection and scaling, working with Nan Values), Building an ML pipeline, Analysis, Motivation, SVM, XGBoost.	Decision Trees, XGBoost, Dataset Description, Visualizations, Preprocessing (NaN values, oversampling, outliers), Analysis, Conclusion, Parts of report, Literature Review-1, Neural Network	Logistic Regression, Gaussian Naive Bayes, Latex Report coding, Data Preprocessing (handling NaN values, Label Encoding, correlation matrix), Analysis,Visualisation, Results, SVM, Adaboost.	SGD Classifier, Decision Trees, Data Preprocessing (Categorical Values, NaN values, Handling Outliers), EDA(Correlation Matrix, Graph Visualization), Literature Review-2, Report, ML Pipeline, Conclusion, Results, AdaBoost, SVM, KNN

THANK YOU