Rain Prediction Next Day

(Alekhya Devi Ranabothu)

Motivation:

Weather data is unstable in nature which makes forecasting weather with current measurements less accurate. However, with the help of machine learning techniques weather predictions can be achieved more accurately. The main goal of this project is to predict whether rain occurs the next day based on the data available on current day. The prediction of rain is very useful for planning our daily activities where we can take some preventive measures like taking an umbrella or a raincoat to avoid becoming wet, postponing few activities like hiking which will not be possible in case of rains. Scheduling the matches of outdoor sports require the accurate prediction of rain because postponing of matches in the last minute involves more cost. If the occurrence of rain is known before, preventive measures like evacuation can be taken beforehand in case of floods, to avoid the destruction. All these scenarios require accurate prediction of occurrence of rain (Positive) than prediction of non-occurrence of rain (Negative). This project uses methods such as Logistic Regression, Random Forest, Support Vector Classifier (SVC), Neural Networks and Recursive Feature Elimination with Cross Validation of machine learning to forecast weather achieving higher accuracy.

Exploratory Data Analysis:

We have taken the weather data set of different locations of Australia from Kaggle. Below are the variables in the dataset.

Variable	Type	Description
Date	Categorical	The date of observation in the format MM/DD/YYYY
Location	Categorical	The common name of the location of the weather station
MinTemp	Numeric	The minimum temperature in degrees Celsius
MaxTemp	Numeric	The maximum temperature in degrees Celsius
Rainfall	Numeric	The amount of rainfall recorded for the day in mm
Evaporation	Numeric	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine	Numeric	The number of hours of bright sunshine in the day.
WindGustDir	Categorical	The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed	Numeric	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am	Categorical	Direction of the wind at 9am
WindDir3pm	Categorical	Direction of the wind at 3pm
WindSpeed9am	Numeric	Wind speed (km/hr) averaged over 10 minutes prior to 9am
WindSpeed3pm	Numeric	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
Humidity9am	Numeric	Humidity (percent) at 9am
Humidity3pm	Numeric	Humidity (percent) at 3pm
Pressure9am	Numeric	Atmospheric pressure (hpa) reduced to mean sea level at 9am
Pressure3pm	Numeric	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
Cloud9am	Ordinal	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	Ordinal	Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values
Temp9am	Numeric	Temperature (degrees C) at 9am
Temp3pm	Numeric	Temperature (degrees C) at 3pm
RainToday	Binary (Categorical)	1 if rainfall (mm) exceeds 1mm, otherwise 0

RISK_MM	Numeric	The amount of next day rain in mm. Used as threshold to create response variable RainTomorrow. If the RISK_MM is less than 1, then RainTomorrow is 'No' and if the RISK_MM is greater than 1, then RainTomorrow is 'Yes'.
RainTomorrow	Binary (Categorical)	The target variable. Did it rain tomorrow? 1 if RISK_MM (mm) exceeds 1, otherwise 0

Table1: Variables in the dataset

As mentioned in the Table 1, there are 23 independent variables out of which 6 are categorical, 2 are ordinal and 15 are numeric variables. The target variable is binary. The variable 'RISK_MM' is used to set the threshold for the target variable. That means it directly contributes to the target variable and if we keep this variable while training a machine learning model, the model may leak the prediction. So, we remove the variable 'RISK_MM'.

As we have more than 140k samples, it is taking more than a day to train and fit the predictive models. So, we have selected the data from certain locations with high, medium and low average rainfall. By this we made sure that selected samples have wet and dry locations.

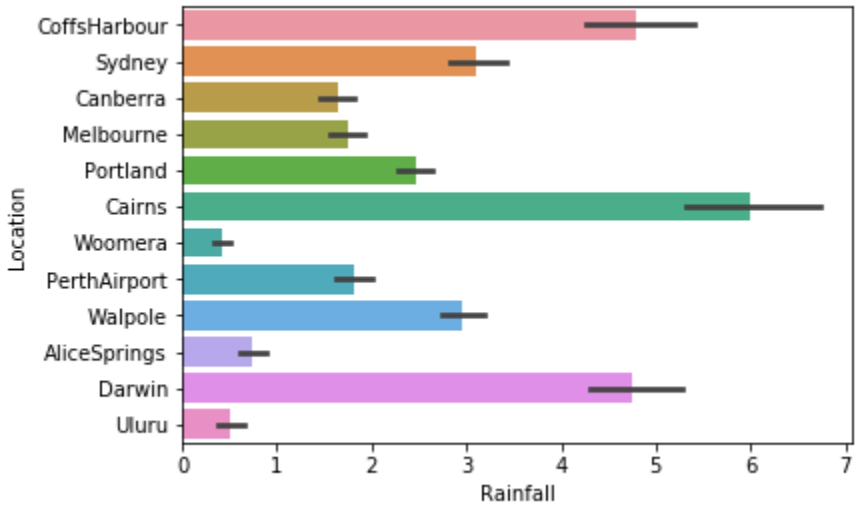


Figure 1: Average Rainfall w.r.t Location

It can be observed from Figure 1, that CoffsHarbour, Cairns and Darwin are the locations with high average rainfall. Woomera, Uluru and AliceSprings are dry locations with low average rainfall. Here, we considered the location with high average rainfall as wet location and the location with low average rainfall as dry location.

Now we have 41k (40988) samples.

Trend of Rainfall

We have analyzed the trend of rainfall in both wet and dry locations, in the years 2019,2018,2015 and 2014.

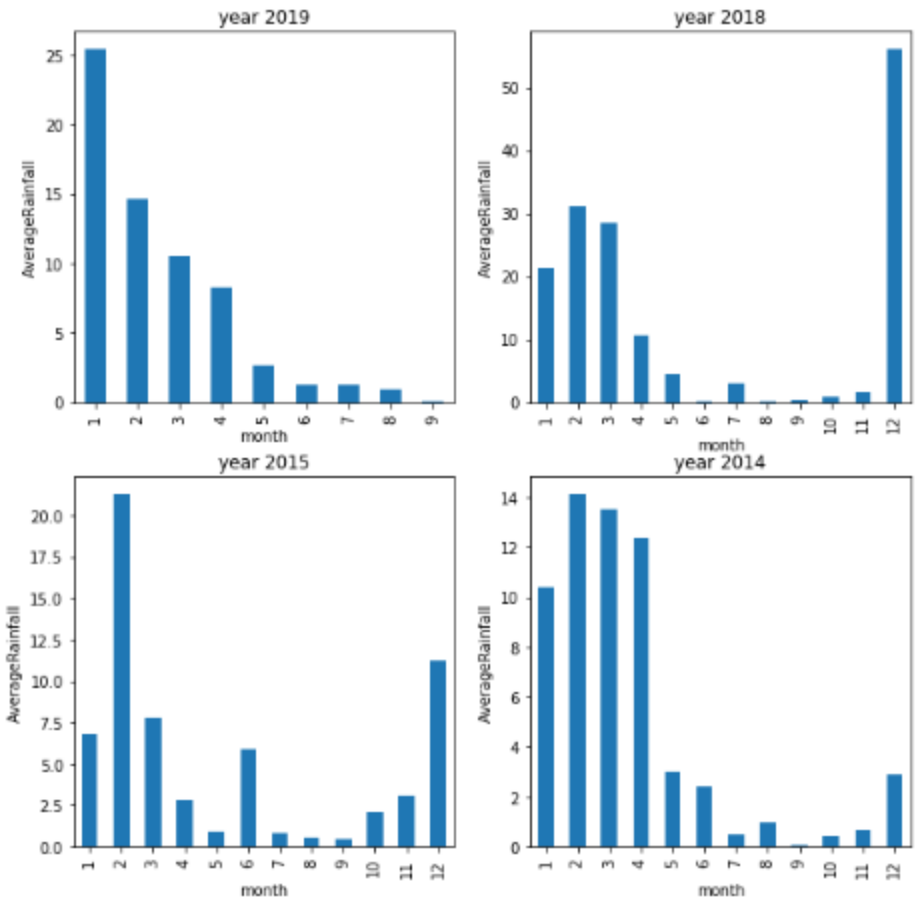


Figure 2: Trend of the rainfall in the high average rainfall location (Cairns) in different years.

It can be observed from Figure2, that there is a chance of heavy rainfalls in the months from December to March. Then we have analyzed the rainfall trend in the months with heavy average rainfall in the years 2018 and 2019.

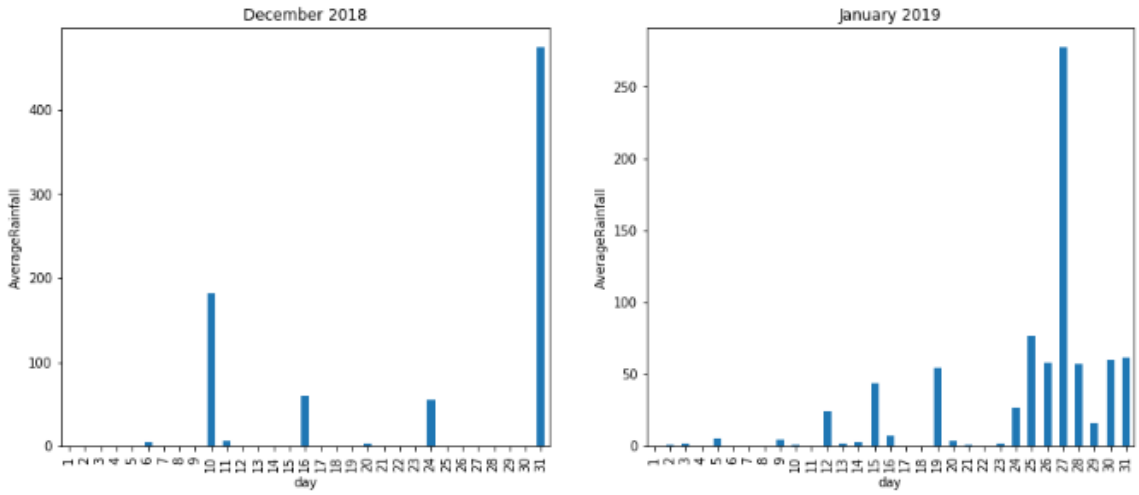


Figure 3: Daily Pattern of Rainfall in the Months of December and January

From the Figure 3, it can be observed that in the months of heavy rainfall, only few days have been recorded with heavy rainfall. It did not rain daily. And also, the highest amount of rainfall recorded is more than 400mm (in December 2018).

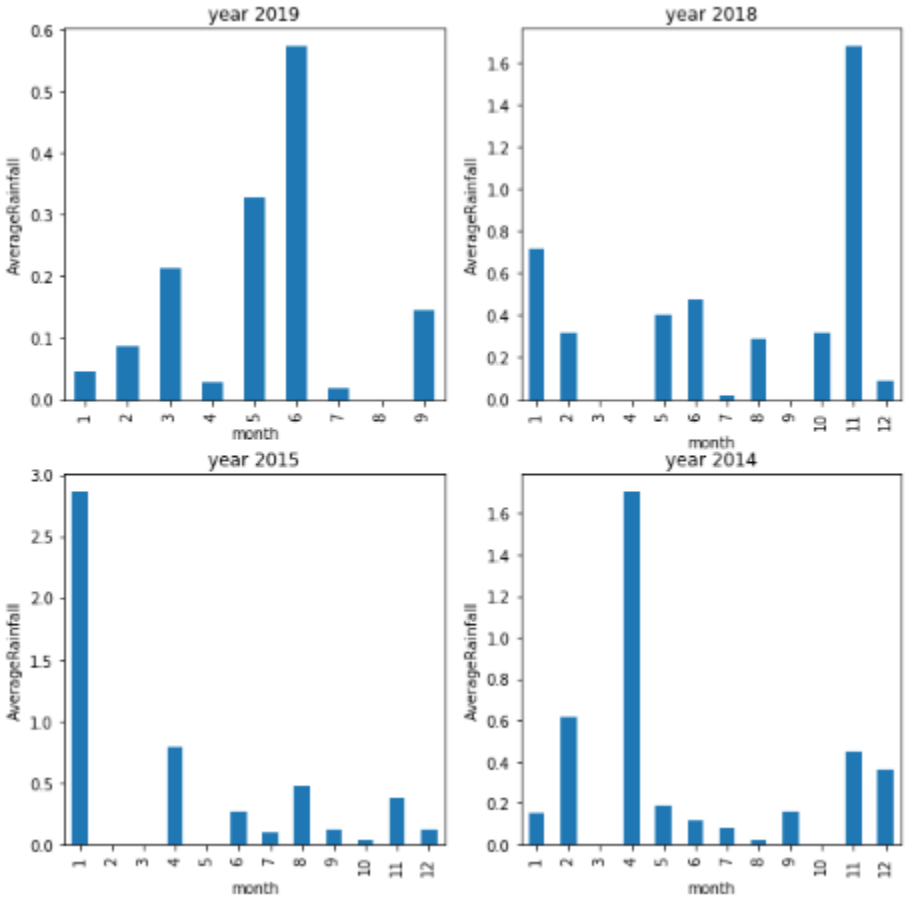


Figure 4: Trend of the rainfall in the one of the low average rainfall location (Woomera) in different years.

It can be observed from Figure 4 that the average rainfall recorded is less than 3mm.Then we have analyzed the rainfall trend in the months with heavy average rainfall which is in the years 2014(April) and 2015(January).

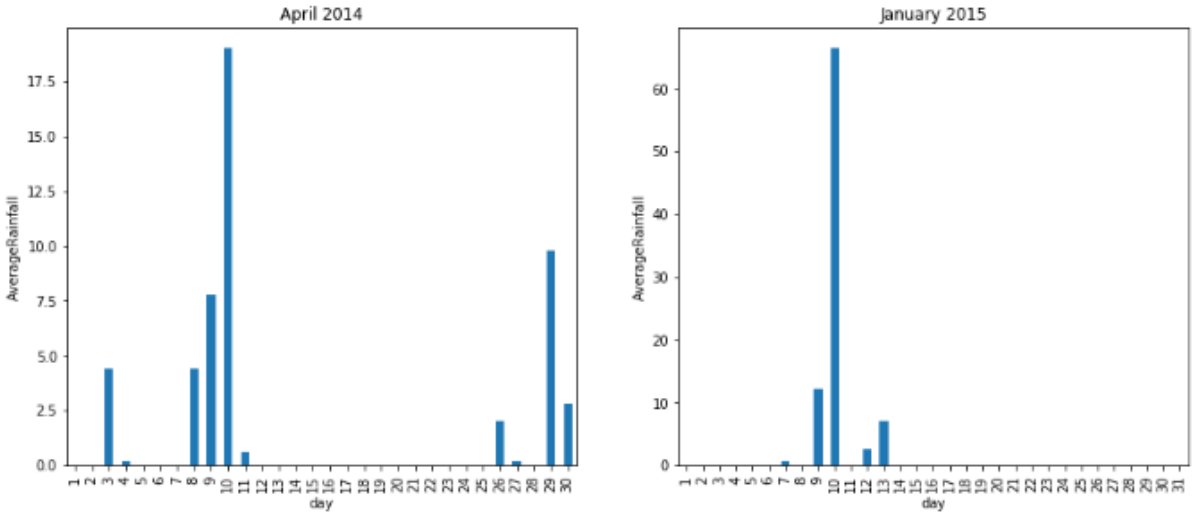


Figure 5: Daily pattern of Rainfall in the months of April and January

From the Figure 5, it can be observed that in the months of heavy rainfall, only few days have been recorded with heavy rainfall. It did not rain daily. And also, the highest amount of rainfall recorded is more than 60 and less than 80 (in January 2015).

After analyzing the trend of rainfall, we understood that the occurrence of rain does not change with time. So, we cannot perform time series analysis for the prediction of target variable RainTomorrow.

Target Variable

We have 78 % of negative class (class 0) samples and 22% of positive class samples (class 1). So, the data is highly imbalanced.

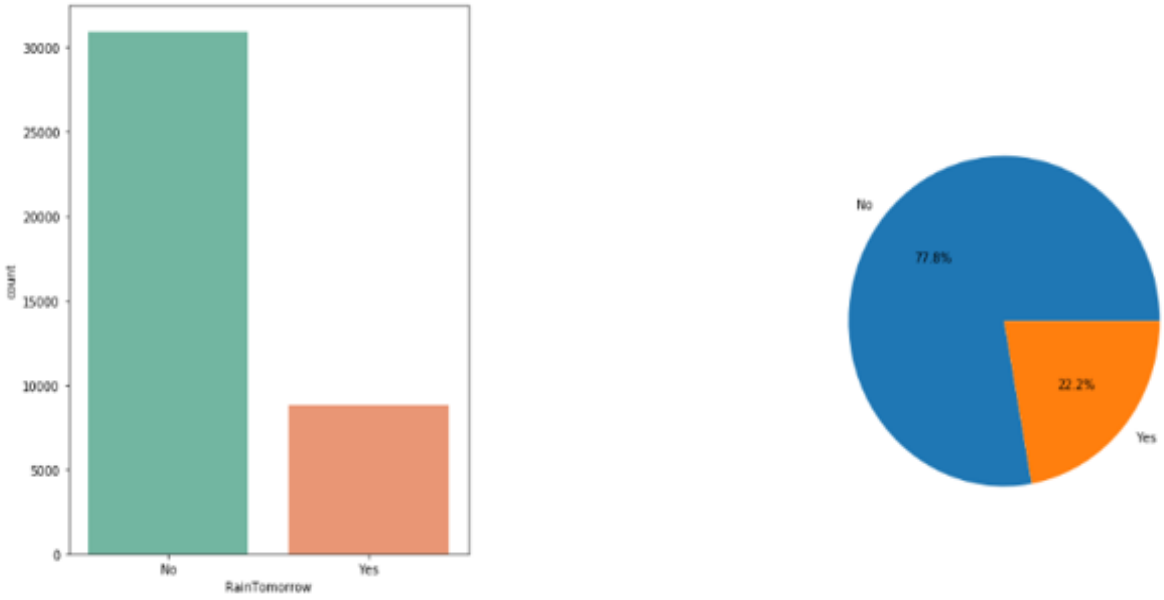


Figure 6: Distribution of positive and negative class samples

Independent Variables

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm
count	40474.000000	40497.000000	39754.000000	27898.000000	26702.000000	38778.000000	40333.000000	40350.000000	40359.000000	40427.000000
mean	13.709522	24.865155	2.652807	5.947412	7.833863	41.441539	15.744502	20.037621	64.014569	48.464393
std	6.841662	7.176279	10.424521	4.563610	3.704606	12.753264	8.563570	8.116095	20.601795	21.854618
min	-8.700000	4.100000	0.000000	0.000000	0.000000	11.000000	0.000000	0.000000	0.000000	0.000000
25%	8.800000	19.100000	0.000000	3.100000	5.200000	33.000000	9.000000	15.000000	52.000000	32.000000
50%	13.400000	24.500000	0.000000	5.200000	8.900000	39.000000	15.000000	19.000000	67.000000	51.000000
75%	19.000000	30.500000	0.600000	7.600000	10.800000	48.000000	20.000000	26.000000	79.000000	64.000000
max	31.800000	48.100000	474.000000	103.600000	14.300000	135.000000	67.000000	76.000000	100.000000	100.000000

Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
39978.000000	39983.000000	29467.000000	28987.000000	40470.000000	40514.000000
1017.326750	1014.717992	4.299352	4.304343	18.670665	23.353801
6.898118	6.922079	2.902862	2.795961	6.865173	7.070325
986.700000	985.100000	0.000000	0.000000	-2.100000	3.700000
1012.600000	1009.800000	1.000000	1.000000	13.300000	17.800000
1017.100000	1014.400000	5.000000	5.000000	18.200000	22.900000
1022.100000	1019.500000	7.000000	7.000000	24.100000	28.800000
1040.600000	1037.900000	9.000000	9.000000	39.400000	47.200000

Figure 7: Different attributes of independent numerical variables.

From the figure 7, it can be observed that the count of samples in each variable is not same and not equal to 40988(total samples selected). This is because of missing values in those variables.

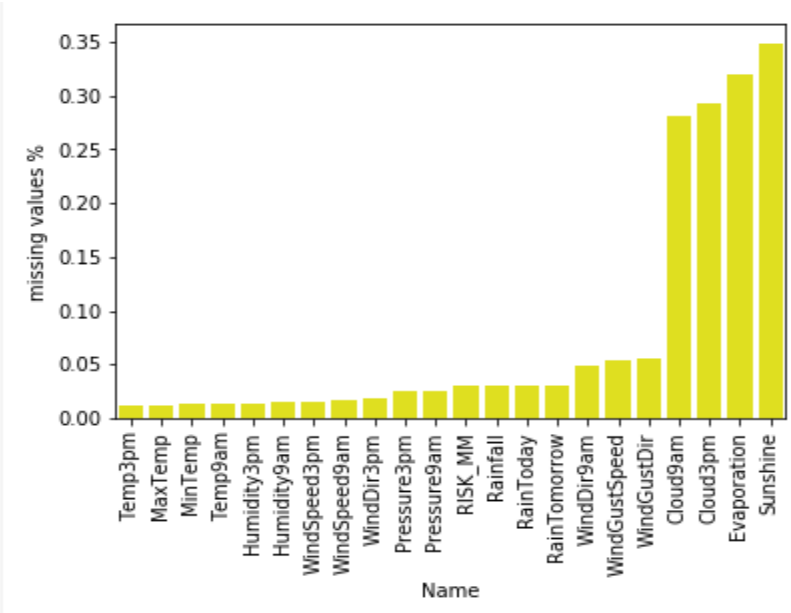


Figure 8: Percentage of missing values in each variable.

The variables Sunshine and Evaporation has highest number of missing values.

From the min and max values of each variable in figure 7, it can be observed that all variables are not of same scale and there is a need of feature scaling. 25%,50% and 75% are the first, second and third quartiles. We can calculate the Inter quartile range (IQR) as the difference of third and first quartiles. We have set a threshold of 1.5*IQR and the data points outside this range (both above and below) are considered as outliers. By comparing the max value and the threshold (1.5* IQR) for each variable it can be observed that many variables have outliers.

The variable Rainfall is used to set the threshold for Rain Today variable. We can use any of these two variables either Rain fall or Rain Today. Since, Rainfall is very much skewed as shown in figure 9, we are using Rain Today variable and removing Rainfall variable.

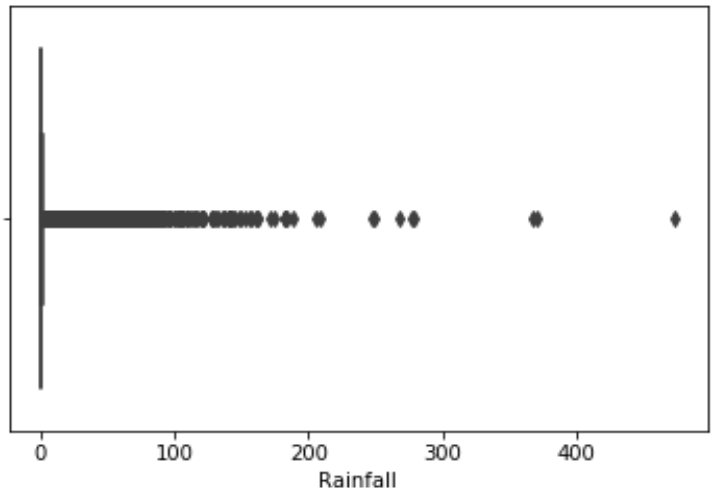


Figure 9: Skewness in Rainfall

We also used univariate method (using box plots) to detect the outliers.

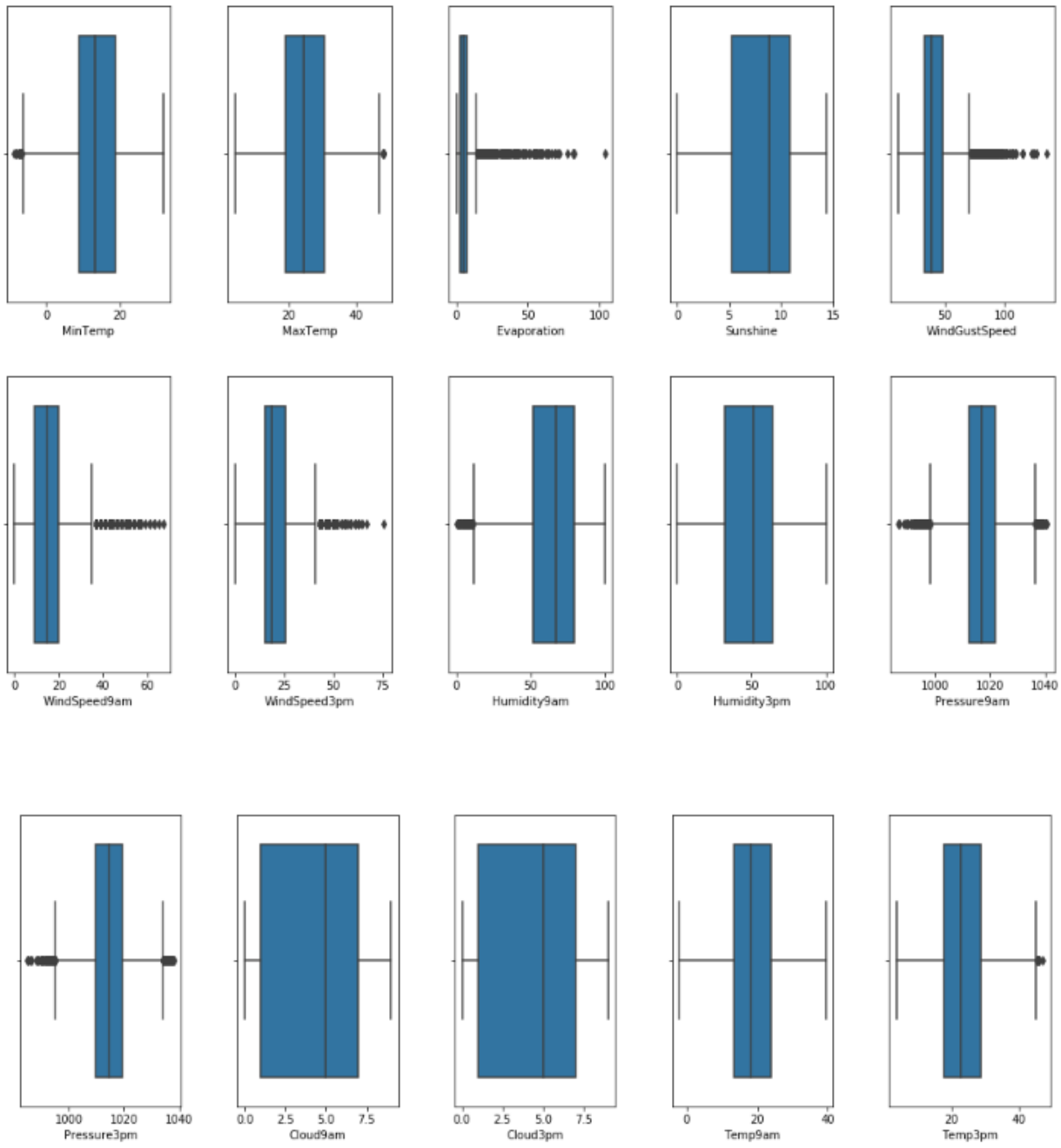


Figure 10: Outliers for all the variables

In the figure 10, the black dots before the first vertical line and after the last vertical line are considered as the outliers.

We analyzed the relationship between numerical variables using pair plot and correlation values.

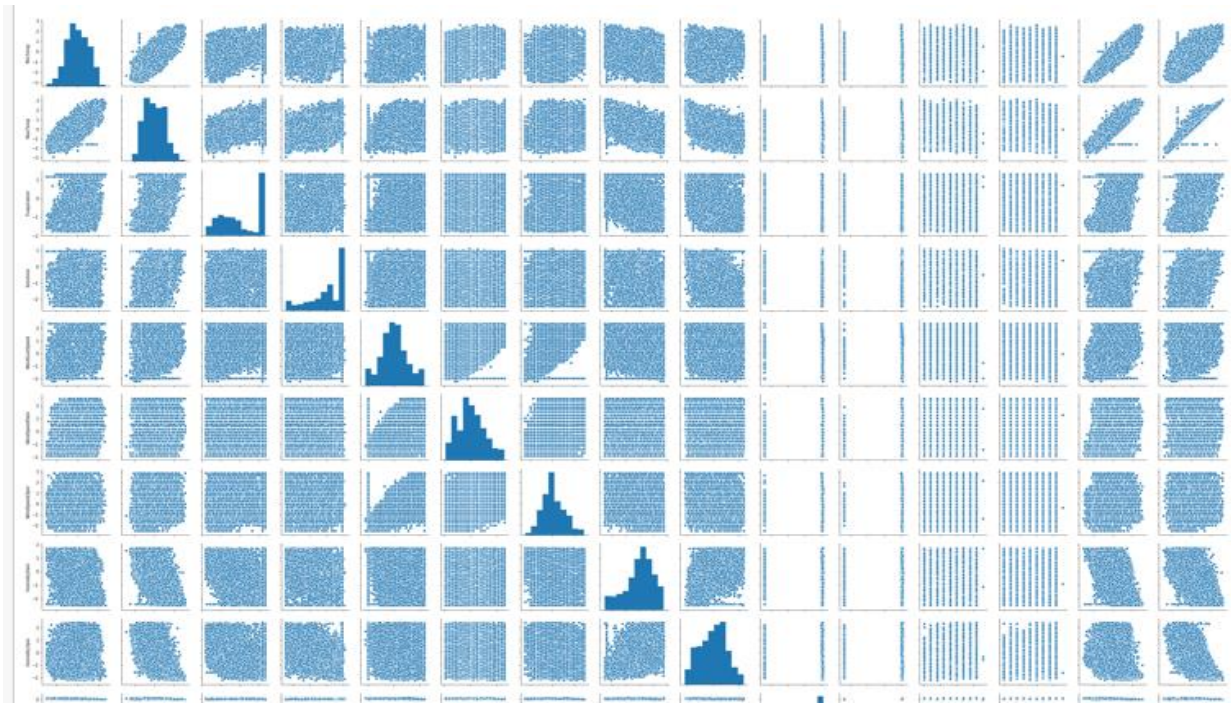


Figure 11: Pair plot which showing the relationship between some of the variables.

It can be observed that most of the variables are nonlinear and very few variables are linearly related.

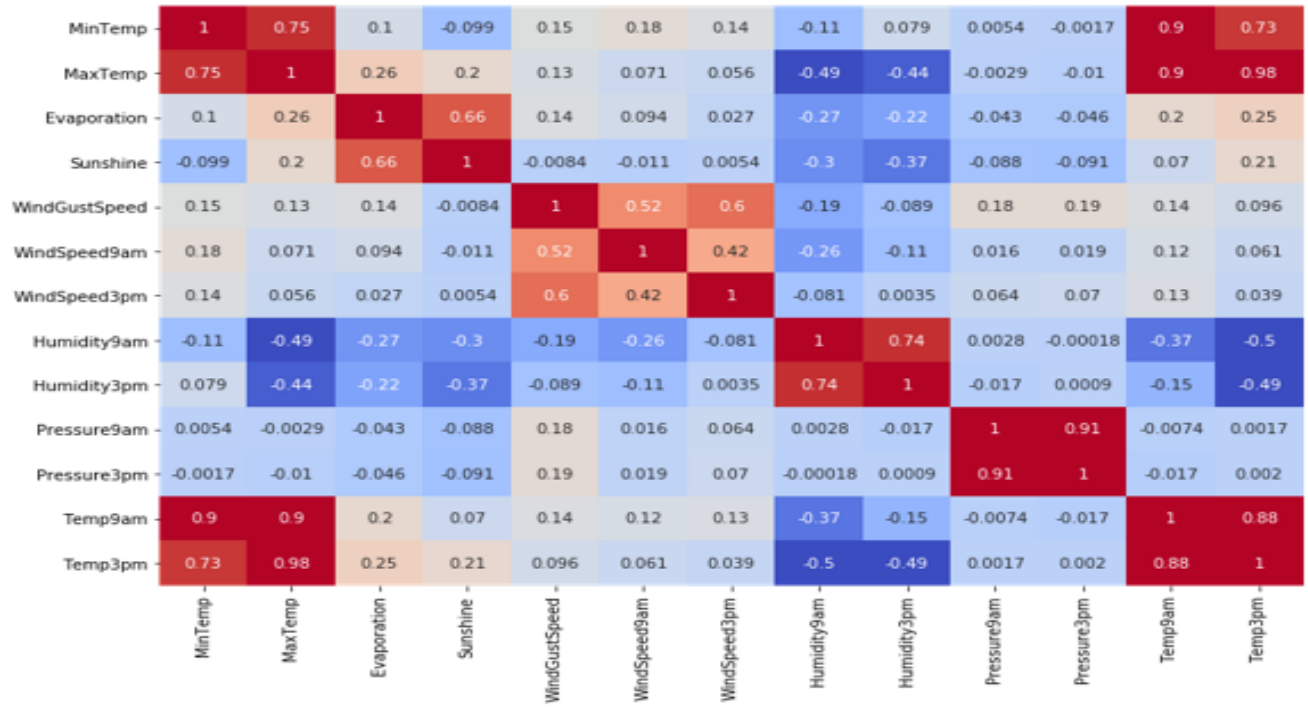


Figure 12: Correlation matrix showing the linear relation of variables.

It can be observed that few variables are highly correlated and most of the correlation values are very less. From this we can say that the data is nonlinear.

Data Preprocessing

We have removed the variables RISK_MM and Rainfall as they are used to set the threshold for RainTomorrow and RainToday, that means these variables directly contribute to RainTomorrow and RainToday.

We have converted the date format in Date column to year, month and day columns and considered these 3 columns as categorical variables.

We have calculated the Inter Quartile Range (IQR=(Q3-Q1)) and set the threshold as 1.5*IQR. So, maximum threshold is Q3+threshold and minimum threshold is Q1-threshold. The data points below the minimum threshold are replaced by minimum threshold and the data points above the maximum threshold are replaced by maximum threshold.

We have imputed the missing values of numerical data with mean, missing values of categorical variables with mode and missing values of ordinal variables with median of corresponding variables.

We have used label encoding and one hot encoding for converting categorical variable to numbers. After One hot encoding we got 134 variable columns.

We have done feature scaling to convert all the variables to same scale.

Predictive Models

We have selected both linear and nonlinear models to train the data and evaluate the performance. We have selected below algorithms.

1. Logistic Regression
2. Random Forest
3. Support Vector Classifier
4. Artificial Neural Networks

We have split the data into training (80%) and test data (20%). We further split this training data to training data and validation data using 5-fold cross validation and 10 different train and validation splits with F1 score as the scoring metric (Refer the Evaluation metrics section to know why F1-score is selected as better metric). We have taken the mean of all the obtained scores as the final F1-score.

1. **Logistic Regression**

   Below are the important parameters used in Logistic Regression.
   Regularization parameter C: This is Inverse of regularization strength lambda (we used ridge regularization)
   Solver: This is the algorithm used for optimization
   Max_iter: Maximum iterations taken to converge, by the optimization algorithm.

2. **Random Forest**

   We have trained the data containing all the features, with Random Forest model using default parameters. Below are the important parameters used in Random Forest.

n_estimators: Number of decision trees to be used in the random forest. Using a greater number of trees reduces the overfitting but slows down the learning process. So, we need to find optimum number of trees.
max_depth: Depth of each tree in the forest. The information about data can be learnt more accurately when the depth is more because as the depth increases, number of splits also increases.
min_samples_leaf: Minimum samples at the leaf node.
max_features: The number of features to be considered for the best split. When None all the features are considered and when 'sqrt' square root of total features is used at each spilt.

3. **Support Vector Classifier (SVC)**

We used the default parameters. Below are the important parameters used in SVC.
C: Regularization parameter
kernel: Type of kernel
gamma: Kernel coefficient

4. **Neural Networks (MLP)**

Below are parameters used in MLP.
Alpha: Learning rate
Hidden layers: Number of hidden layers
Hidden nodes: Number of hidden nodes
Solver – Optimization algorithm
Activation – Activation function used in the hidden layers

Experiments and Results:

Evaluation Metrics:

The data is imbalanced with majority of negative class and a smaller number of positive class. The main goal of this project is to measure how accurately the model predicts the occurrence of rain next day (i.e. positive class).

Precision is the measure of true positives among the total predicted positives. Recall is the measure of true positives among the actual positives. So, to get the good performance of positive class, both False positives and False Negatives should be less. That means both precision and recall should be more. Hence, we need to consider the score which gives the trade off between precision and recall. So, we considered F1- score as the better metric for this data.

With all the features and default parameters:

First, we have trained the data containing all the features, with Logistic Regression, Random Forest, SVC and MLP using default parameters. Below are the parameter values.

Logistic Regression: C =0.1, Solver ='liblinear', max_iter = 1000

Random Forest: n_estimators =10, max_depth =None, min_samples_leaf = 1, max_features= None

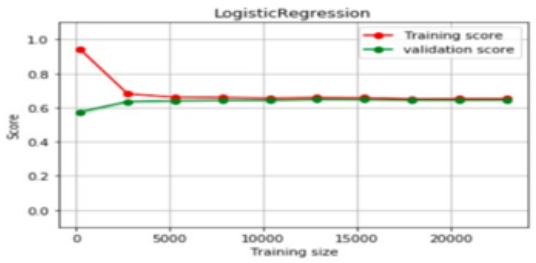
SVC: Kernel: rbf, C=1.0, Gamma=1/no of features

MLP: Hidden_layer_sizes=100, activation='relu', solver='adam', alpha=0.0001

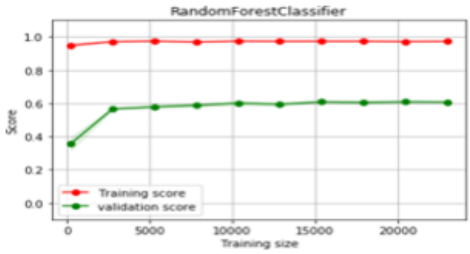
Learning curves:

We have used learning curves to analyze the performance of training data and validation data. We used f1-score as the scoring parameter. We used 5 -fold CV with 10 different training and validation data splits and took the mean of all the scores as the final score. We plotted the learning curves with training size and score.

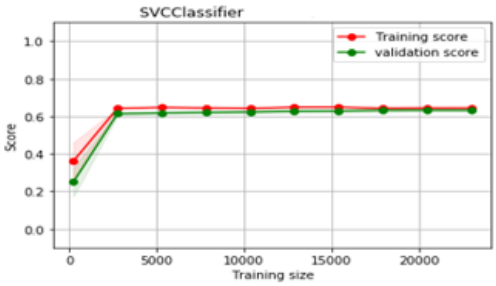
- Logistic Regression



- Random Forest



- SVC



- Neural Networks

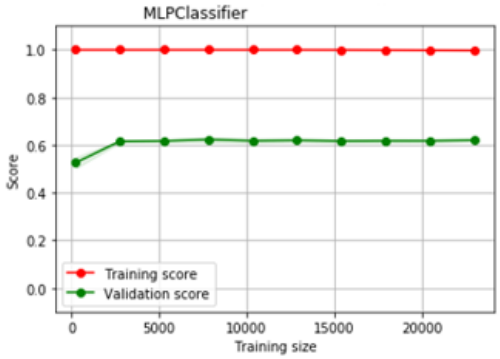


Figure 13: Learning curves before Feature Selection

We have analyzed the learning curves in the Figure13 and found the below observations.

Logistic Regression: Both training and validation performance are low. The model is underfit as logistic regression is simple and linear model and we are tying to predict the nonlinear data using linear model. In this case changing the training size will not help to increase the performance. We can try with feature selection by selecting the important features by reducing the noise. We can also try tuning the parameters listed in predictive models' section.

Random Forest: Training Performance is high and validation performance is low. So, the model is overfitting the data. Increasing the training size will not help to decrease the overfitting because, the performance is constant from training size of 15000 and decreasing the training size more may make the model underfit. The performance on validation data can be improved by feature selection and parameter tuning and select the best parameters that perform well on validation data.

SVC: Both training and validation performance are low. The model is underfits the data because we are using default parameters with kernel 'rbf' and very less gamma (1/ # features). These parameters make the model simple almost like linear kernel. Hence, performance is poor on nonlinear data. Changing the training size will not help in this case. The performance on validation data can be improved by feature selection and parameter tuning.

MLP: Training Performance is high and validation performance is low. So, the model is overfitting the data because the default parameters use 100 hidden nodes which makes the model very complex than required. Increasing the training size will not help to decrease the overfitting because, the performance is constant from training size of 10000 and decreasing the training size more may make the model underfit because neural networks is complex model and

require more training samples. The performance on validation data can be improved by feature selection and parameter tuning.

Feature Selection

We used correlation matrix to remove highly correlated (correlation value > 88) features. We removed the features 'MaxTemp', 'Temp9am' and 'Pressure9am'. As the data is nonlinear and we also have ordinal, categorical variables along with numerical variables we used Recursive Feature Elimination (RFE) with 5-fold CV with scoring metric as f1-score. We used RFE with CV for each model and selected the best features for logistic regression and random forest models.

We used ExtraTreeClassifier for feature selection for SVC and MLP because these models will not be supported by RFECV.

We got best number of features as 87,54,77 and 82 out of 134 features in logistic, random forest, svc and mlp respectively.

Parameter Tuning

We have tried different values for each parameter and used GridSearchCV with cv value 5 and scoring parameter as f1-score to select the best combination of features that performs well on validation data. Below are best parameters selected for each model.

Logistic Regression: C =0.1, Solver ='sag', max_iter = 1000

Random Forest: n_estimators =500, max_depth =50, min_samples_leaf = 1, max_features= 'sqrt'

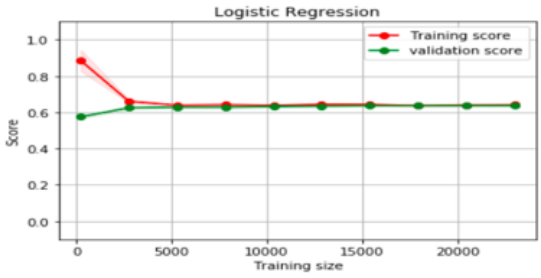
SVC: C = 1.0, gamma = 0.1 and kernel = 'rbf'

MLP: activation = 'logistic', solver = 'sgd', alpha = 0.01 and hidden layers = 1, number of hidden nodes=10

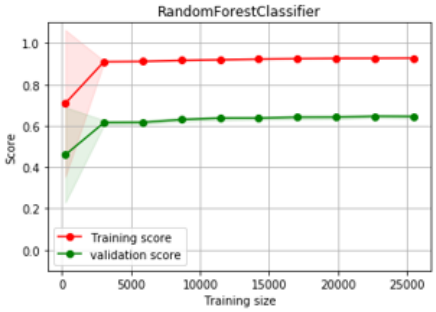
Figure 14 shows the learning curves after feature selection and parameter tuning with best parameters.

For logistic regression, the performance on validation data is improved but the model is still underfit because it is linear model and trying to predict the nonlinear data. For Random Forest, SVC and MLP models the performance on validation data is increased. Overfitting is reduced and the gap between training and validation performance is decreased in Random forest because parameter tuning as increasing the number of trees will reduce the overfitting as described in predictive models' section. Underfitting of the SVC model is handled by increasing the gamma which increases the complexity of the model. For MLP, the training performance is reduced because the number of hidden layers selected in best parameters is 1 which reduces the complexity of the model and makes the model simple.

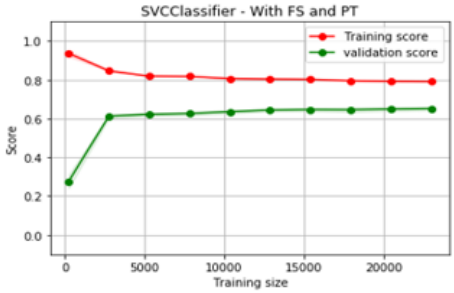
- Logistic Regression



- Random Forest



- SVC



- Neural Networks

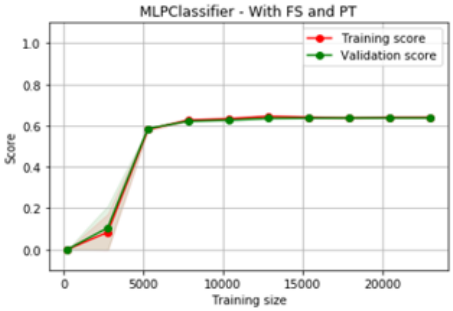


Figure 14: Learning Curves after Feature Selection and Parameter Tuning

Comparison of results on Test data

Algorithm	Without feature selection and tuning		Feature selection and parameter tuning	
	0	1	0	1
Logistic Regression	0.91	0.62	0.91	0.64
SVC	0.91	0.62	0.91	0.65
Random Forest	0.91	0.61	0.91	0.66
Neural Networks	0.89	0.62	0.91	0.64

Figure 15: Comparison results

The trained predictive models are used to predict the test data and figure 15 shows the results of different models on test data before and after feature selection and parameter tuning. It can be observed that all the models predicted the majority negative class with high f1 score, but the nonlinear models SVC and Random Forest performed well by predicting the minority positive class more accurately compared to Logistic and Neural networks (MLP). We believe this is because the data is nonlinear.

Limitations

- The data is highly imbalanced. Methods like oversampling and undersampling that are used to handle the imbalanced data are not explored. We did not use oversampling because we thought this may make the models rigid because of redundancy of samples of minority positive class and under sampling of majority class results in information loss.

Future Work

- Exploring different methods to convert imbalanced data set to balanced data set.
- Different combinations of features and creating new features to improve the model performance.
- In this project we are just predicting if the rain will occur the next day or not by setting the threshold for RainTomorrow(target) as 1 by using RISK_MM. If RISK_MM > 1 RainTomorrow is 'Yes' and if RISK_MM < 1 RainTomorrow is 'No'. We can predict heavy rains by increasing the threshold as required.
- We can forecast other variables like temperature, strength of the wind by forecasting the Windspeed and wind gust speed.

Individual Contribution

Alekhya Devi Ranabothu – Motivation, Exploratory Data Analysis, Data Preprocessing, Random Forest model (Feature Selection, Parameter tuning), evaluation metrics (choosing best metric), analyzing learning curves, comparison of results, Future work

Aruna Kamireddy – Exploratory Data Analysis, Logistic Regression (Feature selection, Parameter tuning), Limitations, Future work

Sonali Ghate – SVC (feature selection, parameter tuning), MLP (Feature selection, Parameter tuning)

References

https://scikit-learn.org/

https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba

https://medium.com/all-things-ai/in-depth-parameter-tuning-for-random-forest-d67bb7e920d

https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769