# Final Project Progress

3D – DIGITAL SKOLA

# PROJECT GOAL

To predict tomorrow will be rain or not

# DATA UNDERSTANDING

- 49 different locations in Australia **from 1 November 2007 until 25 June 2017**
- 23 columns with approximately 145K rows
- The number of data collected is **not equal for all locations**
- **1 dependent variable ('RainTomorrow')** , which called target, and **22 independent variables**, which called features.

# DATA INFO

Data shape : 145460 rows and 23 columns

Top columns contain null:

1. Sunshine (69835)
2. Evaporation (62790)
3. Cloud3pm (59358)
4. Cloud9am (55888)

Duplicated data : not found

Total Location = 49 locations

Date = 3436 days

# EXPLORATORY DATA ANALYSIS

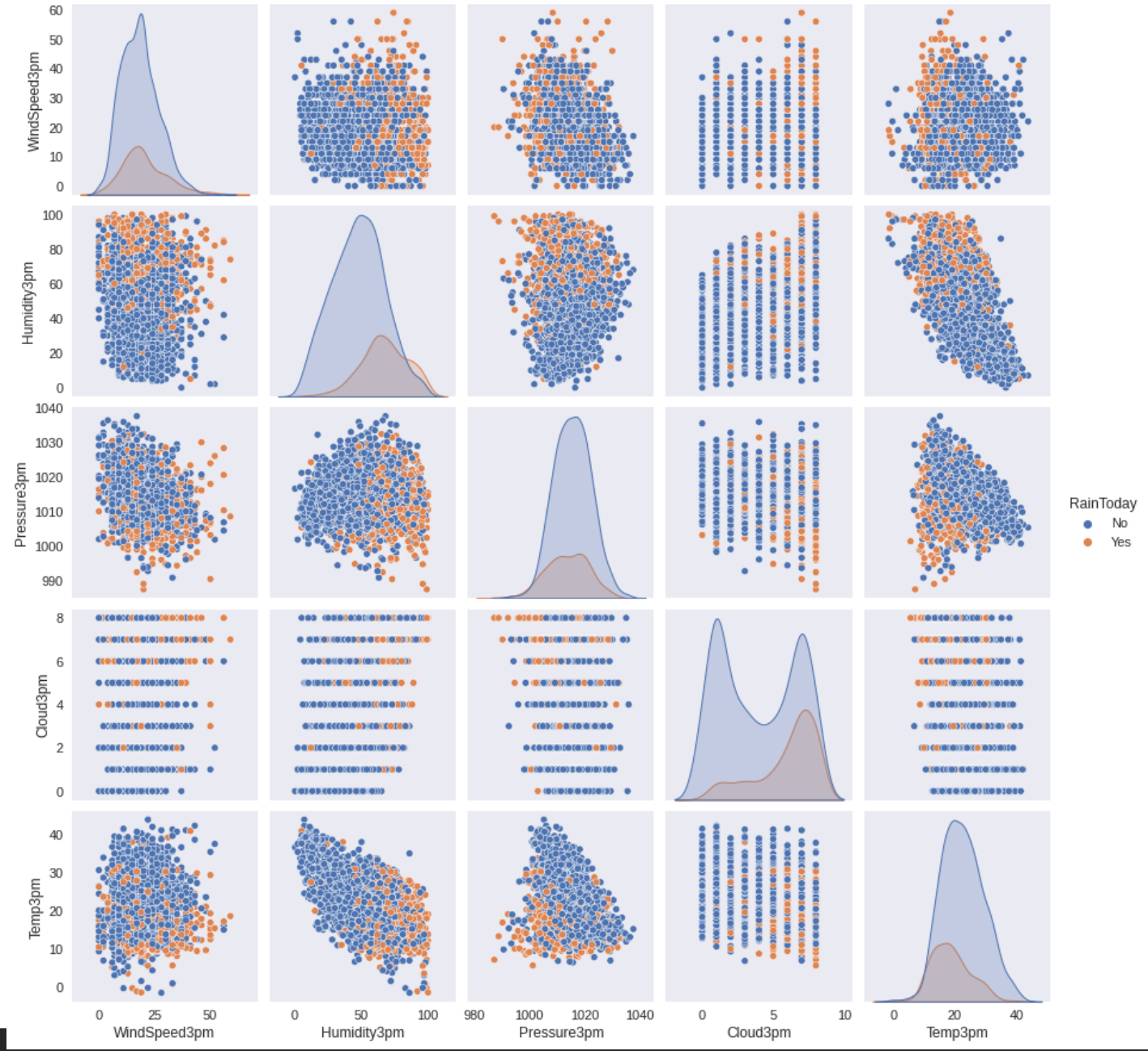
- 1. Berapa besar risiko turun hujan pada setiap tanggal dan setiap bulan sepanjang tahun?
- 2. Berapa besar risiko turun hujan pada setiap kota dalam satu tahun?
- 3. Bagaimana perubahan intensitas hujan dari tahun ke tahun
- 4. Bagaimana distribusi data dari masing-masing kolumn terhadap hujan/tidak hujan?
- 5. Bagaimana korelasi antara fitur numerik?
- 6. Berapa persen jumlah mising value dari masing-masing feature?
- 7. Bagaimana perbandingan antara RainTomorrow = Yes dan RainTomorrow = No?.

# INSIGHT FROM EXPLORATORY DATA ANALYSIS

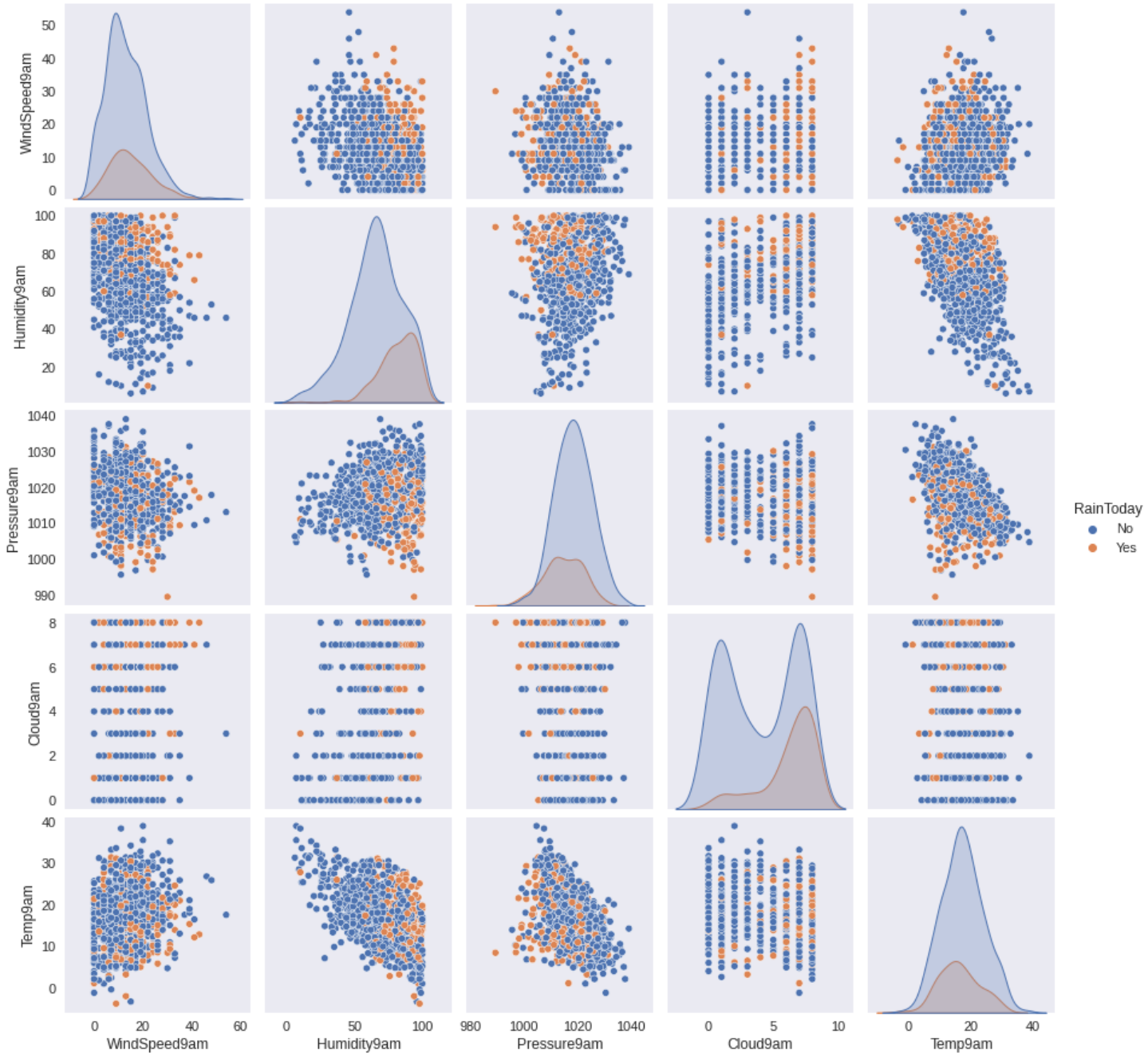
Insight:

1. Months with high rain probability are **June, July, and August**
2. Locations with high rain probability are **Portland (36.4%), Walpole (33.5%), and Cairns (31,7%)**.
3. Yearly rain probability is around **20%** and there is no significant change from **2009 to 2016**
4. Target class distribution is **77% is NO RAIN, 23% RAIN**.
5. Strong correlation between **Temp3pm and MaxTemp, Pressure3pm and Pressure9am, Temp9am and MinTemp, Temp9am and MaxTemp, Temp3pm and Temp9am**.

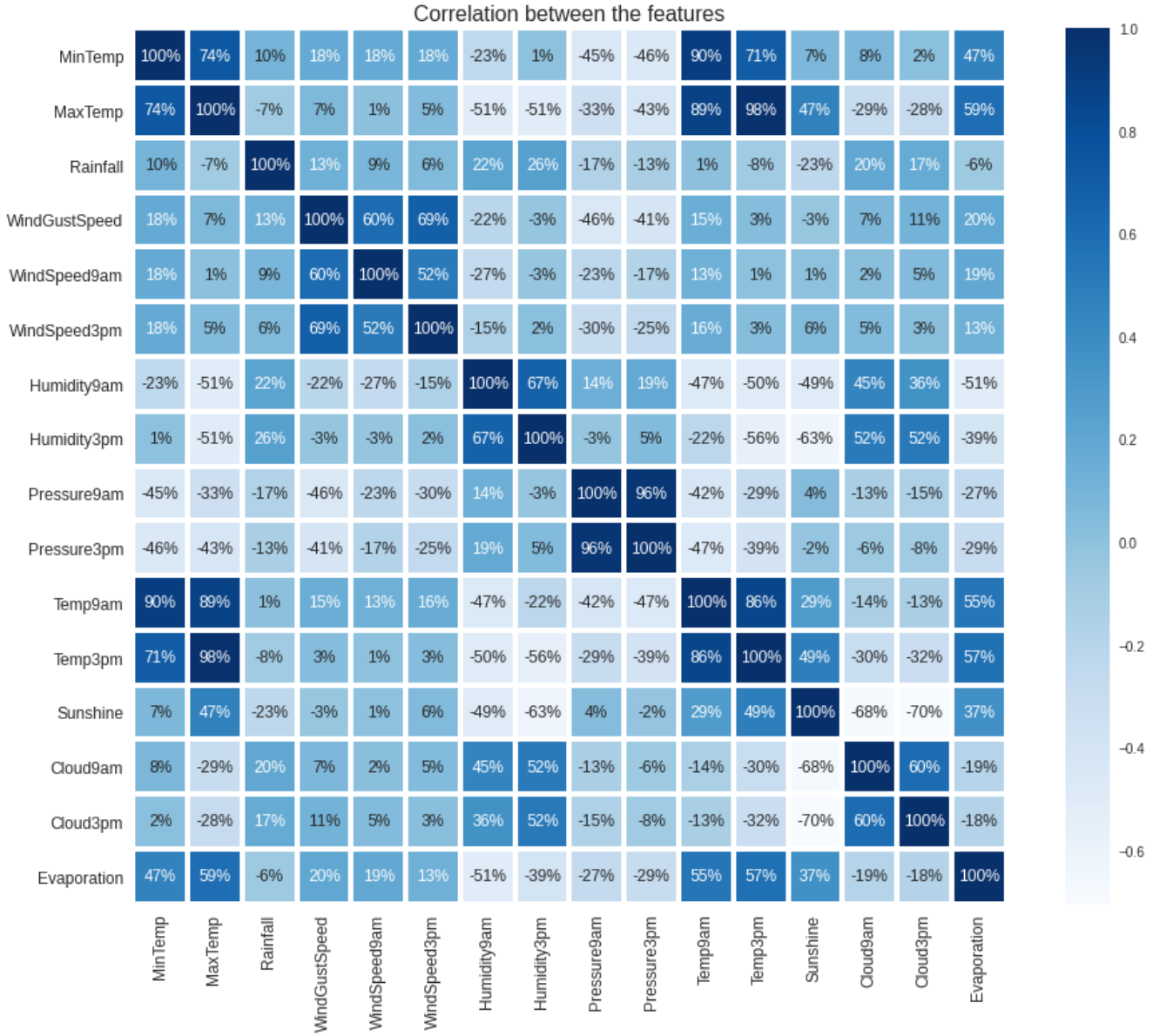

# DATA DISTRIBUTION FROM EACH FEATURE




# DATA DISTRIBUTION FROM EACH FEATURE




# DATA DISTRIBUTION FROM EACH FEATURE




# CORRELATION BETWEEN FEATURES

Removing Multicollinearity

- Temp9am and Temp3pm **will be dropped**
- MaxTemp and MinTemp **will be combined** into new variable named **AvgTemp** which equals to (MaxTemp+MinTemp)/2
- Pressure9am and Pressure3pm **will be combined** into new variable named **AvgPressure** which equals to (Pressure9am+Pressure3pm)/2

Handling Missing Values:

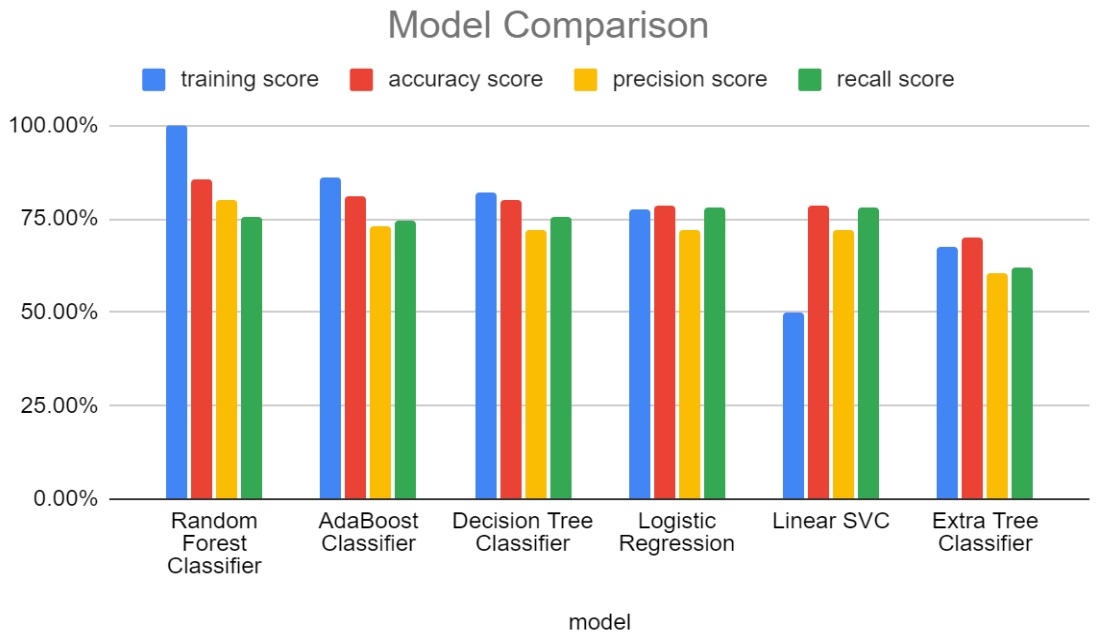
- Missing values in Rainfall, WindGustSpeed, WindSpeed9am, WindSpeed3pm, AvgPressure, AvgTemp, Evaporation, Humidity9am columns will be **imputed with median value**
- Missing values in Sunshine, Cloud9am, Cloud3pm, Humidity3pm columns will be **imputed with mean value**
- Missing values in WindDir9am, WindDir3pm, WindGustDir will be **imputed with mode value**

Encoding Categorical Data:

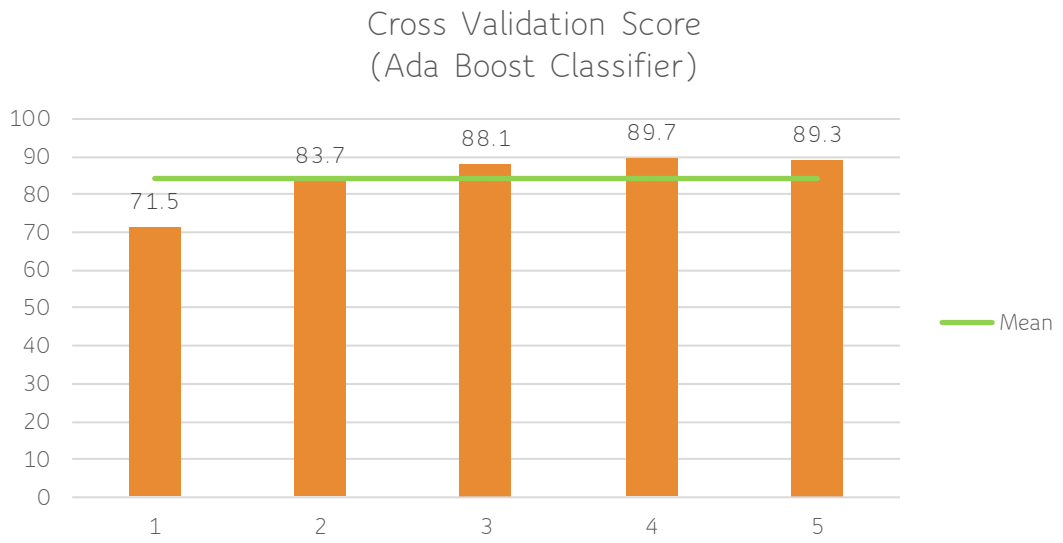
- For location and date will be encoded using pivot table method.
- For wind direction, encoded using one-hot encoding method

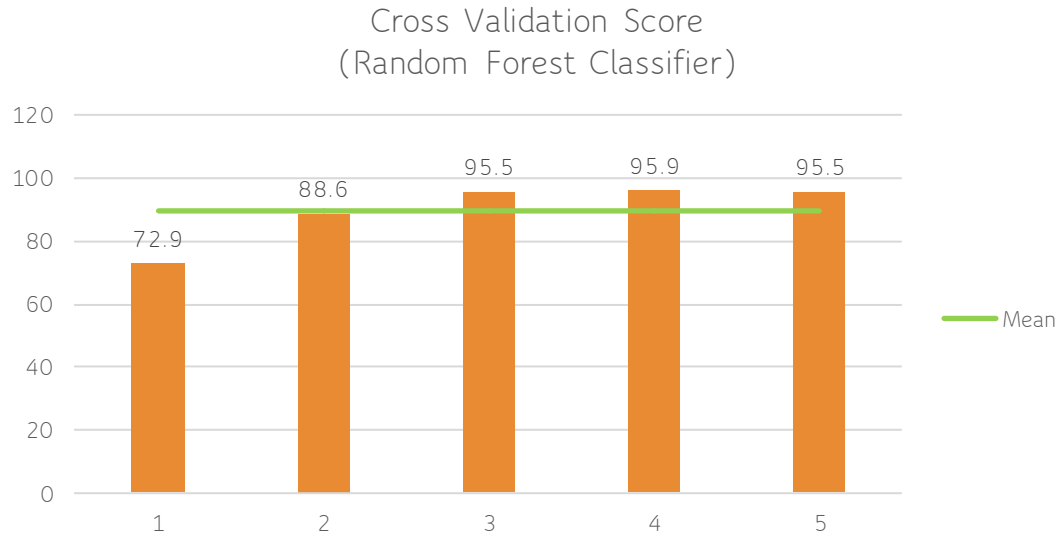
# DATA PREPROCESSING

| Model | Training_Score | Accuracy_Score | Precision_Score | Recall_Score |
|---|---|---|---|---|
| **Random Forest Classifier** | 100.00 | 0.856111 | 0.803630 | 0.758062 |
| **AdaBoostClassifier** | 85.76 | 0.811993 | 0.730967 | 0.744607 |
| **Decision Tree Classifier** | 82.55 | 0.795959 | 0.717664 | 0.756292 |
| **Logistic Regression** | 77.62 | 0.784191 | 0.715988 | 0.776283 |
| **Linear SVC** | 50.00 | 0.783933 | 0.718054 | 0.782123 |

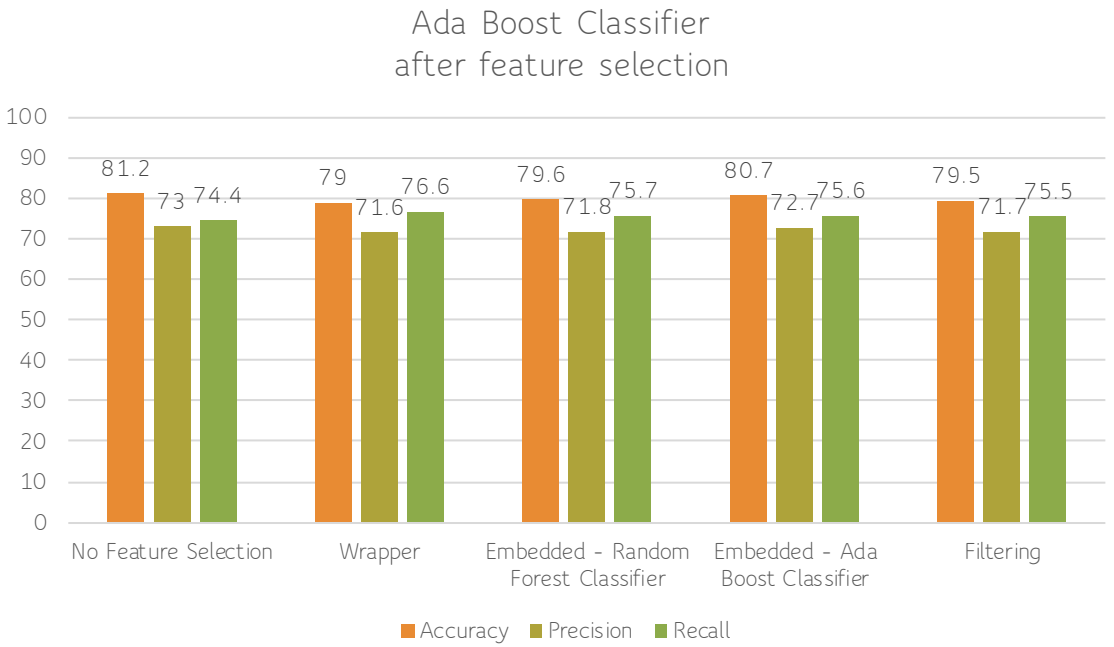


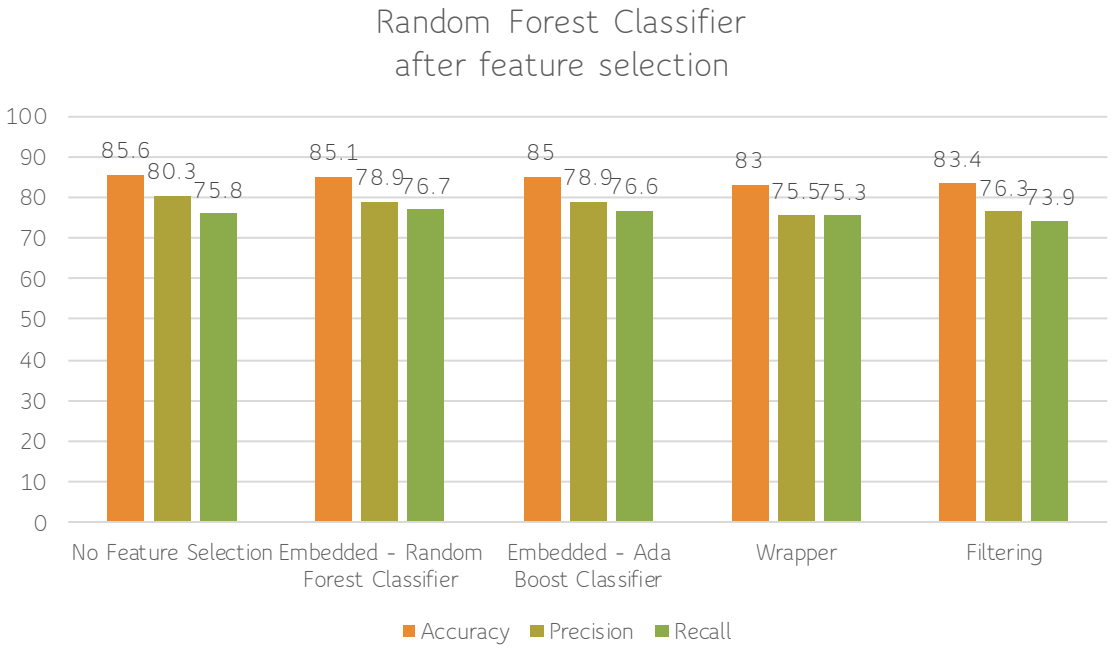

# MODELS DEVELOPMENT




# MODELS
# CROSS VALIDATION




# FEATURE SELECTION

# HYPER PARAMETER TUNING

## Random Forest Classifier

### Grid Search CV

The best parameters are:

- bootstrap : True
- max_depth : 4
- max_features : auto
- min_samples_leaf : 1
- min_samples_split : 2
- n_estimators : 72

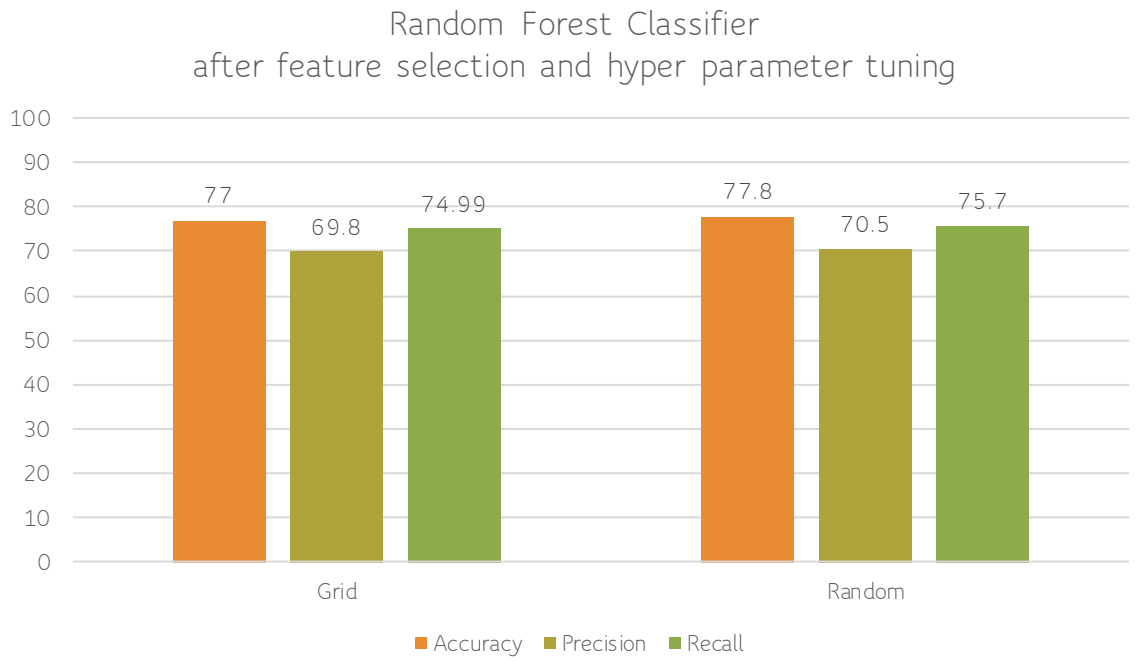
### Random Search CV

The best parameters are:

- bootstrap : False
- max_depth : 4
- max_features : auto
- min_samples_leaf : 1
- min_samples_split : 5
- n_estimators : 33

**Feature selection:**

Embedded Method – Random Forest Classifier




# RANDOM FOREST CLASSIFIER

## after feature selection and hyper parameter tuning

# Conclusion

The best model to predict tomorrow will be rain or not is Random Forest Classifier after feature selecting using embedded method with accuracy, precision, and recall is **85.1%, 78.9%, 76.7%** respectively.