RAIN PREDICTION IN AUSTRALIA

Machine Learning Project

Spetsiotis Dimitrios
October 2024

OUTLINE

- Executive Summary
- Table of Contents
- Introduction
- Methodology
- Results
- Discussion
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Data Collecting
- Data Wrangling

TABLE OF CONTENTS

INTRODUCTION

Weather Data from Australian Government's Bureau of Meteorology were used to create a rain prediction model for the city of Sydney using Machine Learning algorithms.

METHODOLOGY

- Data Collecting
  - Data Collection API
  - Data Collection with Web Scrapping
- Data Wrangling
  - Data filtering
  - Deal with missing values
- Exploratory Analysis
  - Using SQL
  - Using Pandas and Matplotlib
- Interactive Visual Analytics
  - Folium
  - Plotly Dash
- Predictive Analytics

DATA COLLECTING

Data Collecting was performed through AGBM’s website:

http://www.bom.gov.au/climate/duo/

where raw data in the form of csv files were downloaded.

DATA COLLECTING

- Date
- Min Temp
- Max Temp
- Rainfall
- Evaporation
- Sunshine
- WindGustDir
- WindGustSpeed
- WindDir9am
- WindDir3pm
- WindSpeed9am
- WindSpeed3pm
- Humidity9am
- Humidity3pm
- Pressure9am
- Pressure3pm
- Cloud9am
- Cloud3pm
- Temp9am
- Temp3pm
- Rain Today
- Rain Tomorrow

Feature Variables

- Rainfall (Regresion)
- Rain Tomorrow (Classification)

Target Variable

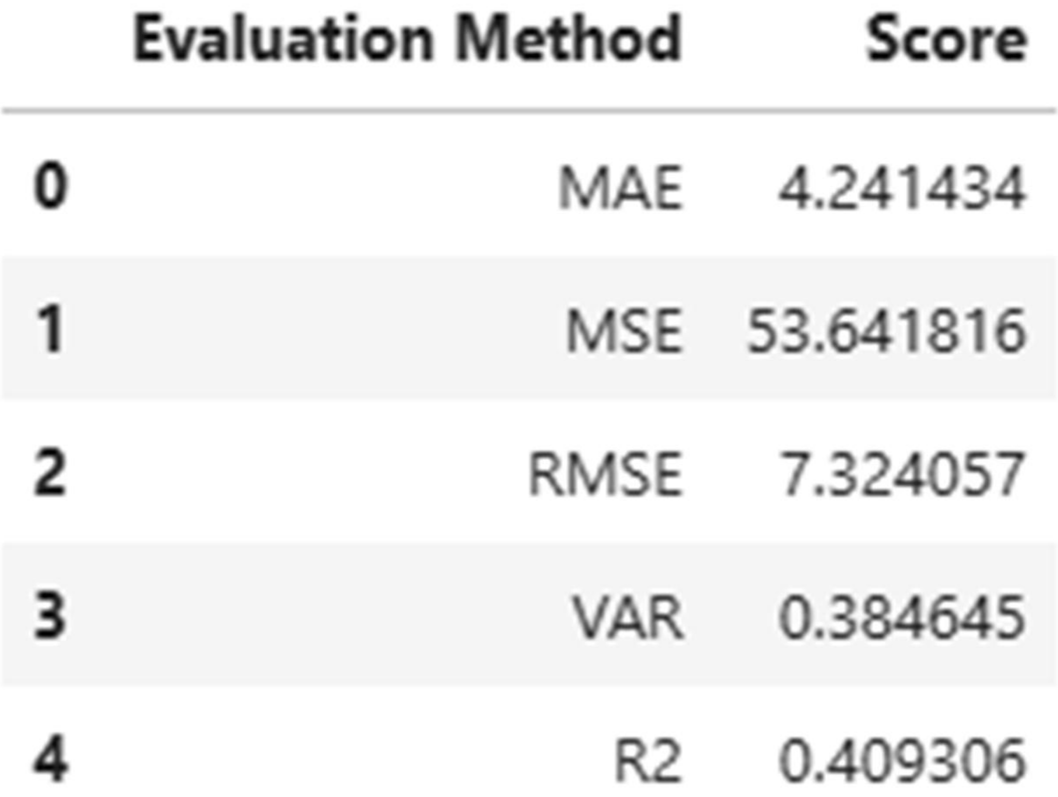
DATA PREPROCESSING

- One Hot Encoding was used to categorical variables to binary.
- Get Dummies method was used

REGRESSION

- Multi-Linear Regression is used to predict amount of rainfall(mm).
- Rainfall is set as target variable while all the remaining are set as features.
- Data divided into train set (80%) and test set (20%)
- Linear Model from Sklearn is used for the MLR

REGRESSION

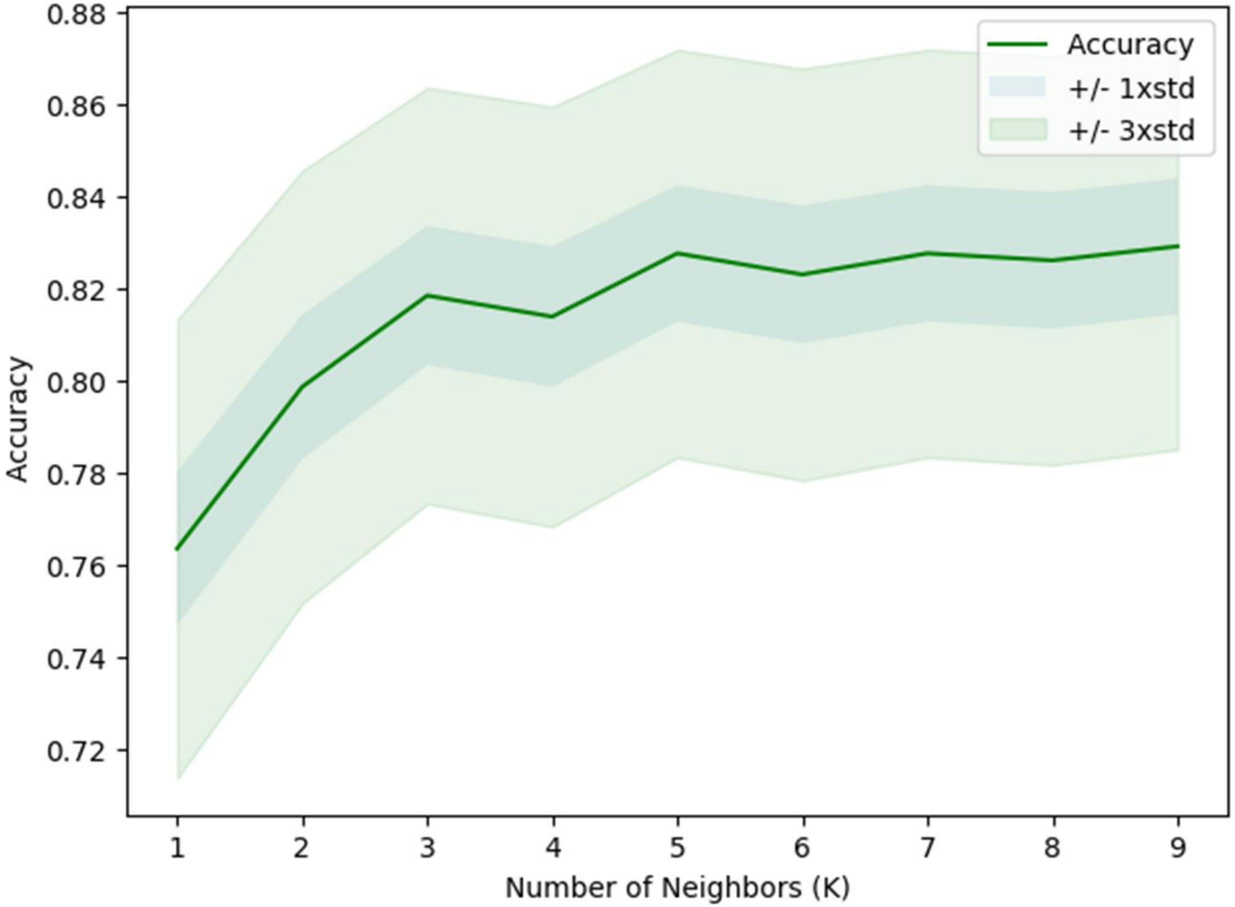


	Evaluation Method	Score
0	MAE	4.241434
1	MSE	53.641816
2	RMSE	7.324057
3	VAR	0.384645
4	R2	0.409306

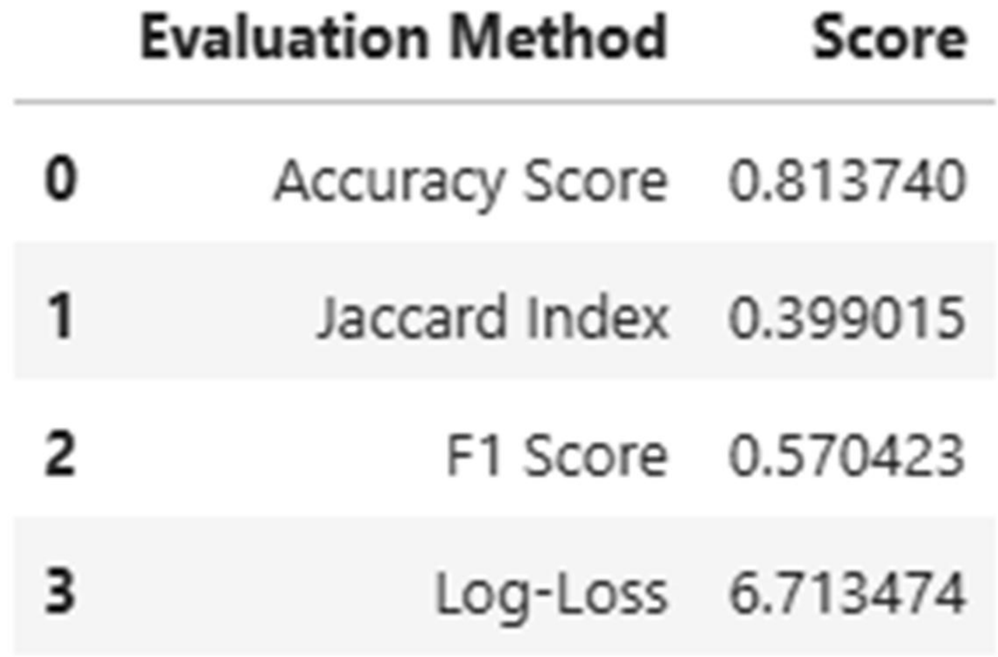
CLASSIFICATION

- Classification is used to predict Rainy Days
- Classification Methods
  - KNN
  - Decision Trees
  - Logistic Regression
  - SVM
- Rain Tomorrow is set as target variable while the remaining are set as features
- Data divided into train set (80%) and test set (20%)

CLASSIFICATION - KNN

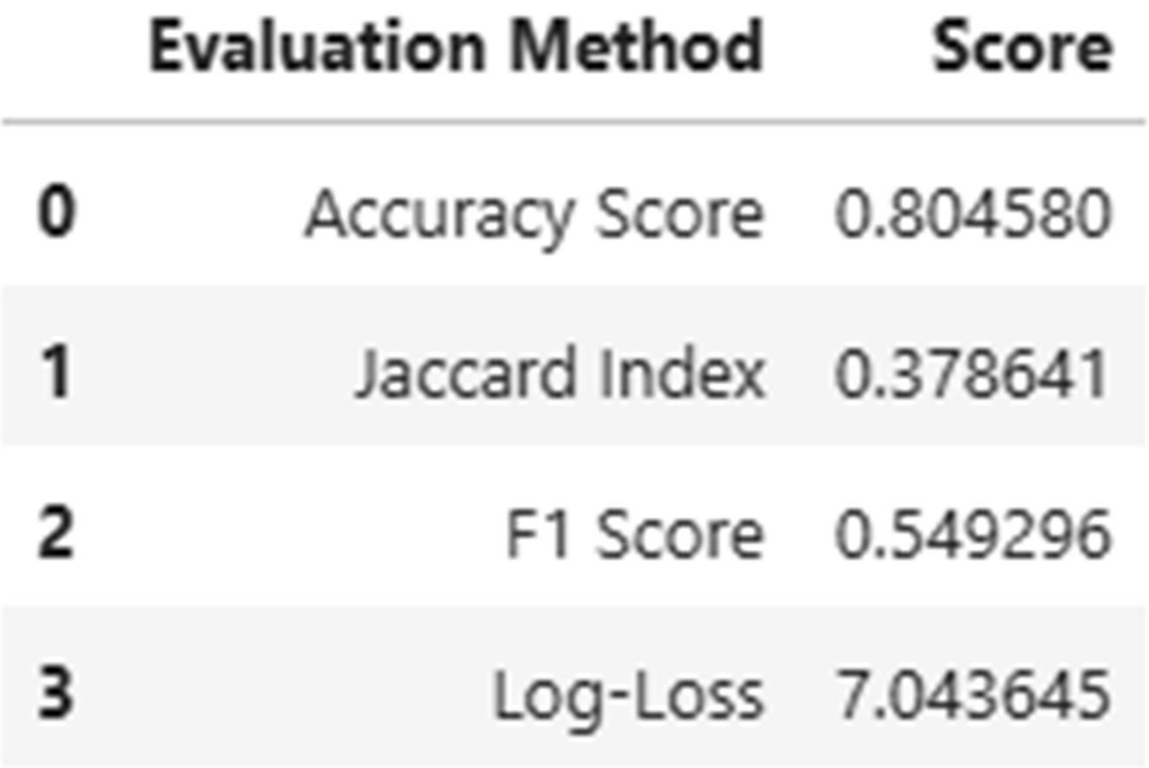


CLASSIFICATION - KNN



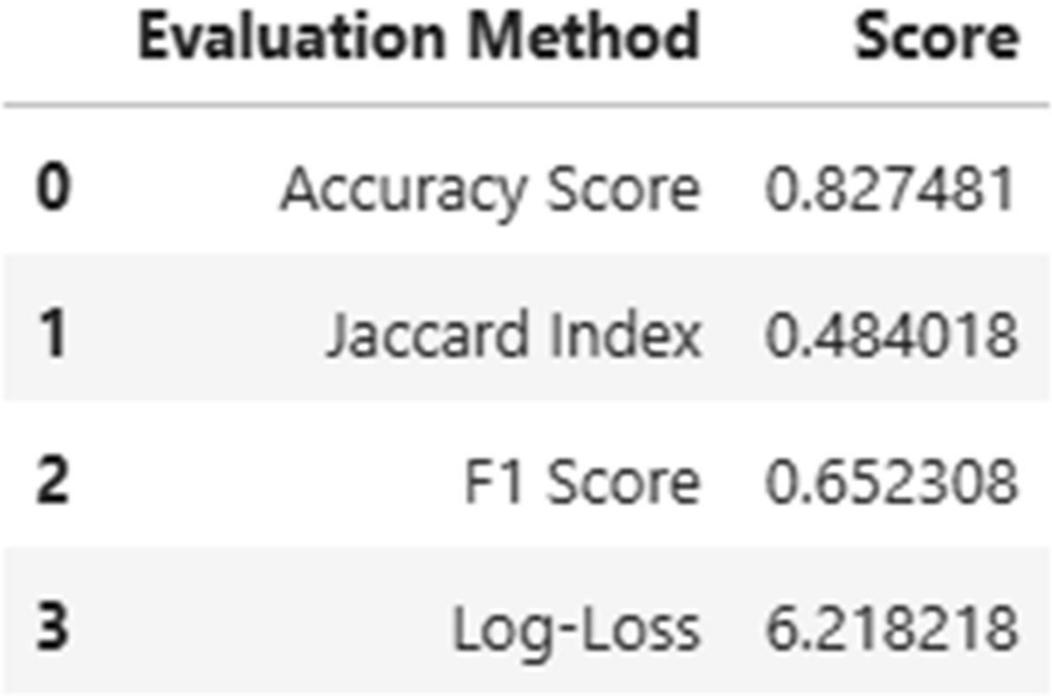
	Evaluation Method	Score
0	Accuracy Score	0.813740
1	Jaccard Index	0.399015
2	F1 Score	0.570423
3	Log-Loss	6.713474

CLASSIFICATION - DT



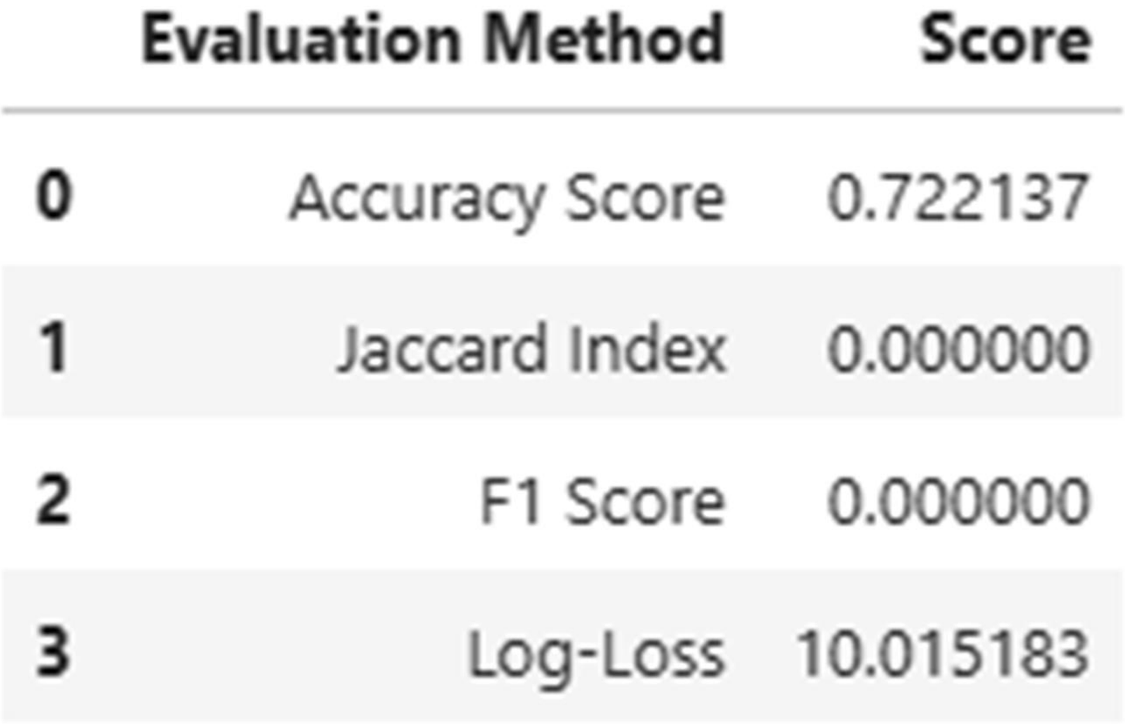
	Evaluation Method	Score
0	Accuracy Score	0.804580
1	Jaccard Index	0.378641
2	F1 Score	0.549296
3	Log-Loss	7.043645

CLASSIFICATION - LR



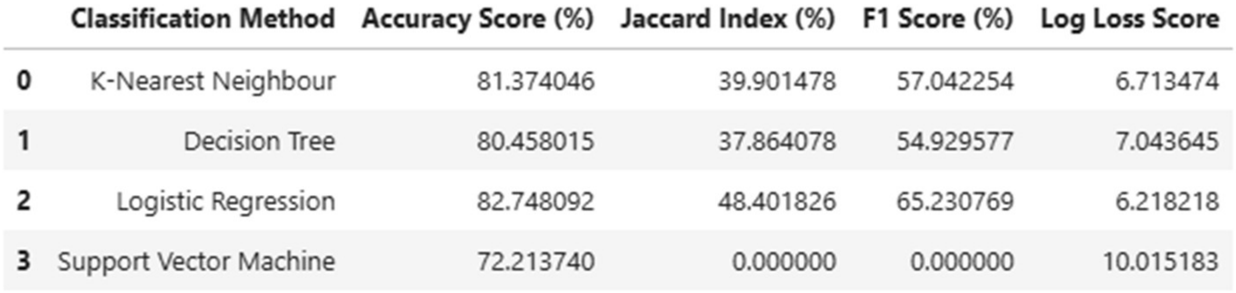
	Evaluation Method	Score
0	Accuracy Score	0.827481
1	Jaccard Index	0.484018
2	F1 Score	0.652308
3	Log-Loss	6.218218

CLASSIFICATION - SVM



	Evaluation Method	Score
0	Accuracy Score	0.722137
1	Jaccard Index	0.000000
2	F1 Score	0.000000
3	Log-Loss	10.015183

MODEL COMPARISON



	Classification Method	Accuracy Score (%)	Jaccard Index (%)	F1 Score (%)	Log Loss Score
0	K-Nearest Neighbour	81.374046	39.901478	57.042254	6.713474
1	Decision Tree	80.458015	37.864078	54.929577	7.043645
2	Logistic Regression	82.748092	48.401826	65.230769	6.218218
3	Support Vector Machine	72.213740	0.000000	0.000000	10.015183

CONCLUSIONS

- Linear Regression Model has a moderate performance with an R2 score 0,41
- Log Regression has the highest Accuracy Score, Jaccard Index and F1 Score while SVM has the lowest