TDS3301 DATA MINING
GROUP PROJECT

Group 14

Chang See Jie	1181103362	0143490382
Loo Chen Zhi	1181103230	01116789079
Lim Wei Jie	1181103501	0125681547

TASK DISTRIBUTION

Chang See Jie: Data Preprocessing, Feature Selection, Classification Model, Slides

Loo Chen Zhi: External Dataset, Data Preprocessing, ARM, Clustering Analysis, Report

Lim Wei Jie: EDA, Feature Selection, Regression Model, Streamlit

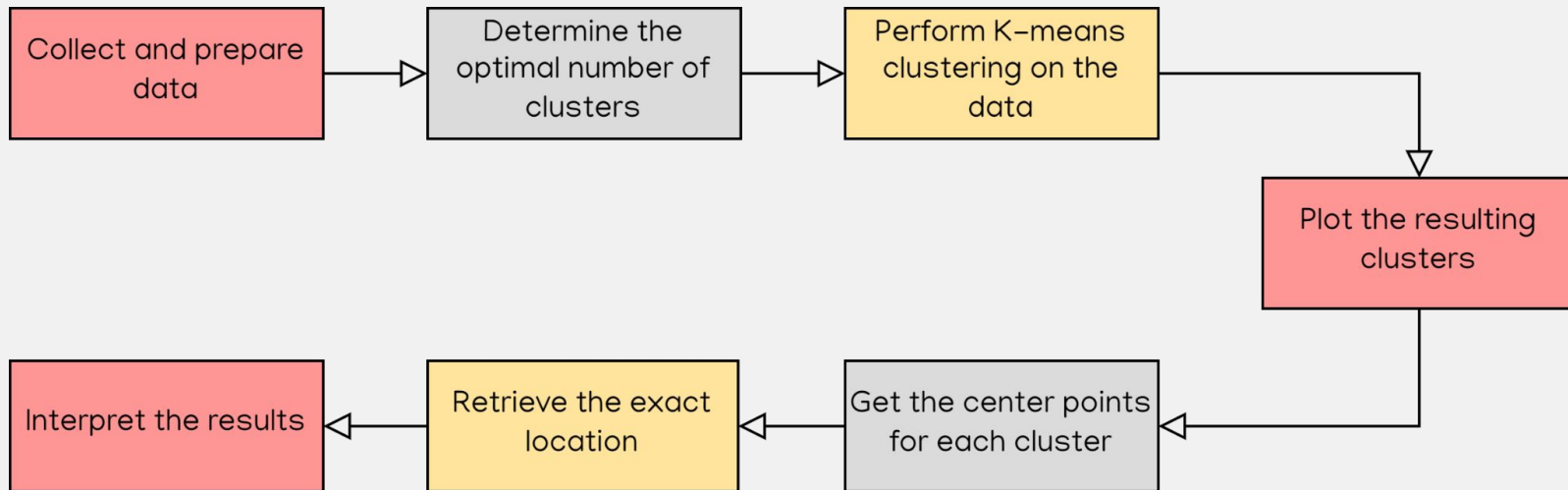# EXTERNAL DATASET - WEATHER

PROVIDED BY VISUAL CROSSING

https://www.visualcrossing.com/weather/weather-data-services/cyberjaya/metric/2015-10-01/2016-03-31

- Historical weather of Cyberjaya town from 2015-10-01 until 2016-03-31.

- tempmax
- tempmin
- temp
- feelslikemax
- feelslikemin
- feelslike
- humidity
- precip
- precipprob

- precipcover
- preciptype
- windgust
- windspeed
- winddir
- cloudcover
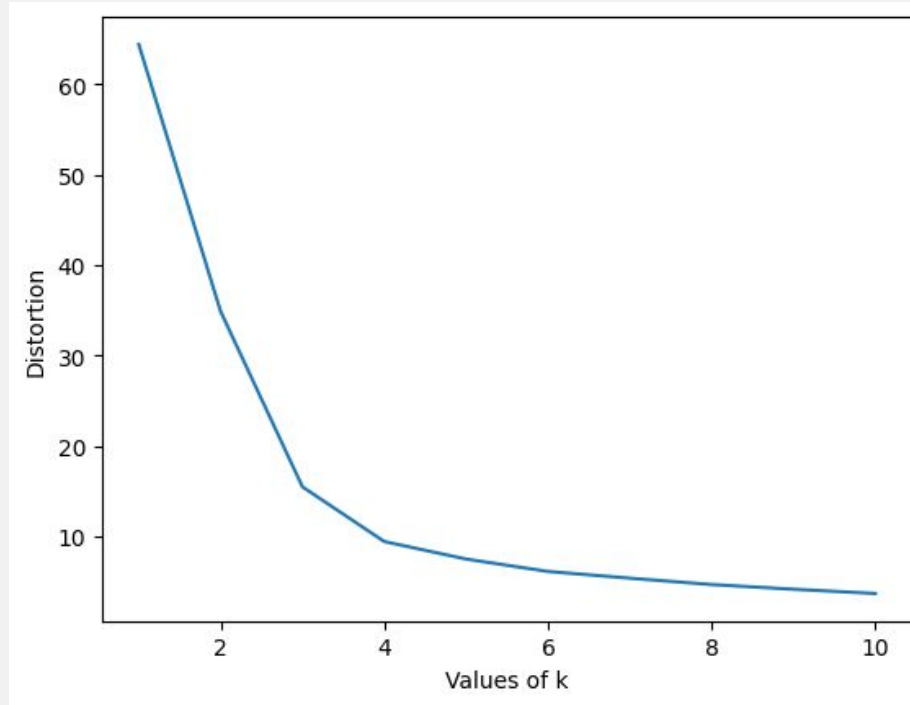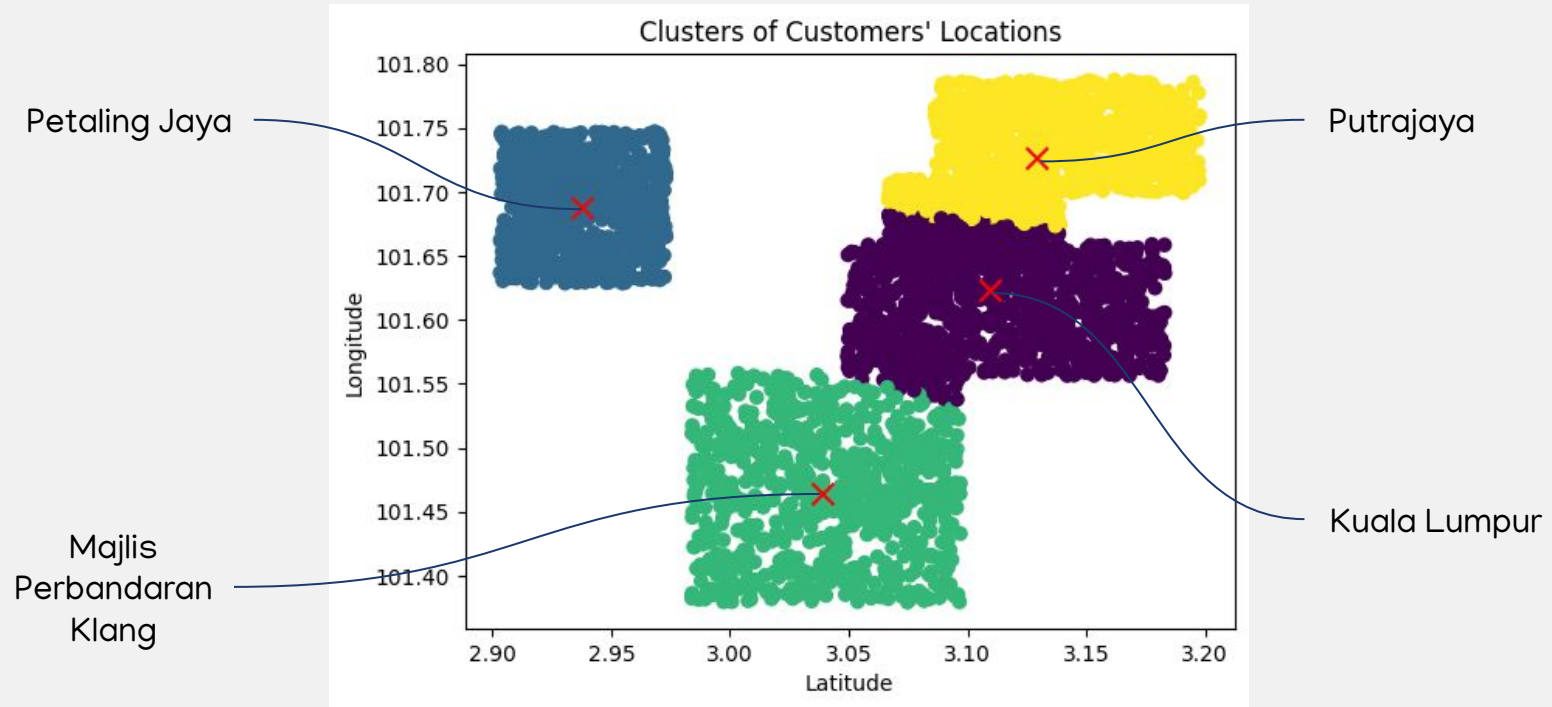- visibility
- severerisk
- conditions

QUESTION 1

Where are the customers located?

```
Collect and prepare data  →  Determine the optimal number of clusters  →  Perform K-means clustering on the data  →  Plot the resulting clusters

Interpret the results  ←  Retrieve the exact location  ←  Get the center points for each cluster  ←
```

ELBOW METHOD

K-MEANS CLUSTERING

Clusters of Customers' Locations

Petaling Jaya

Putrajaya

Kuala Lumpur

Majlis Perbandaran Klang
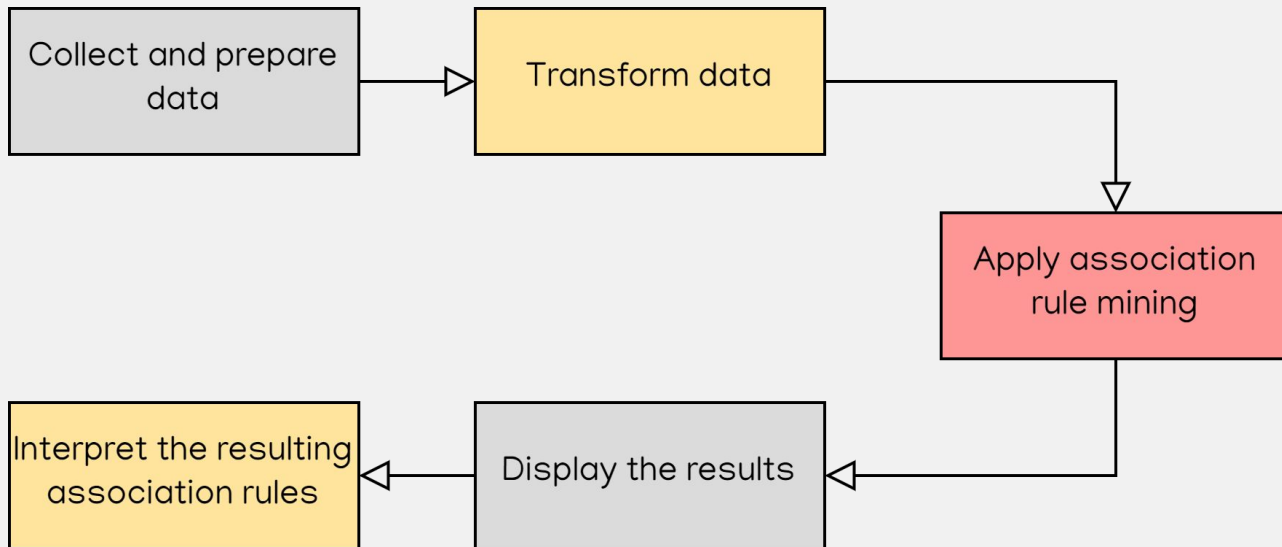
QUESTION 2

Are there any frequent patterns that occur between the various attributes of the customers and the laundry machine usage?
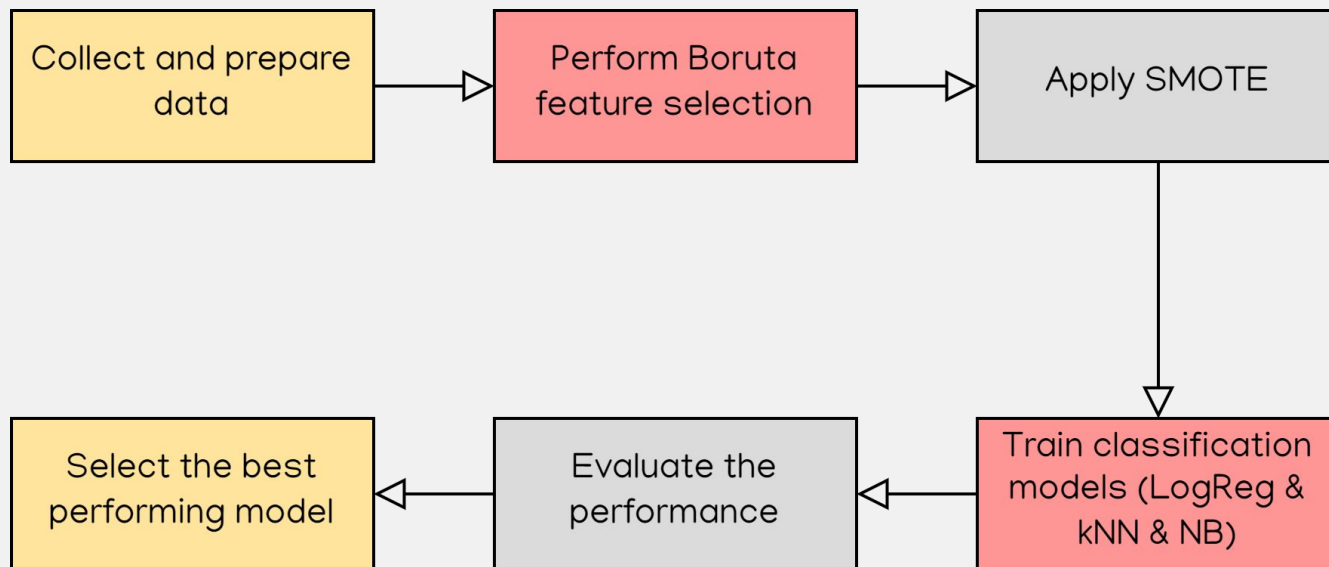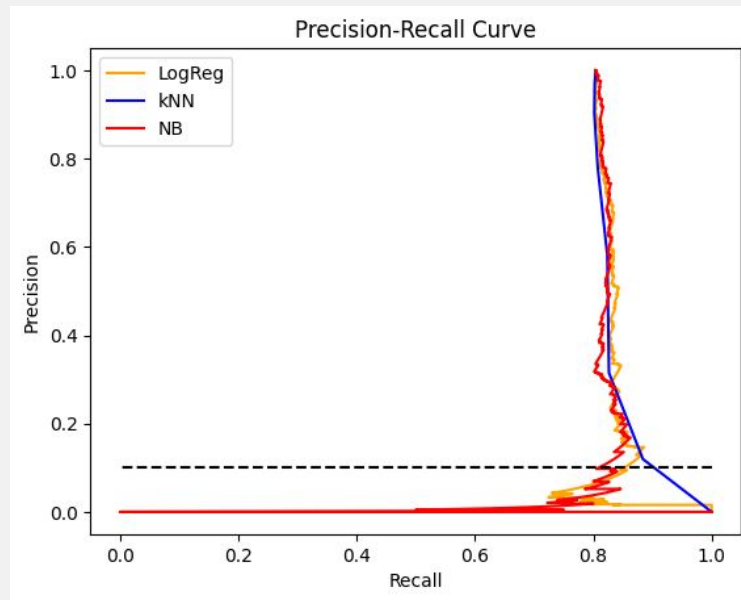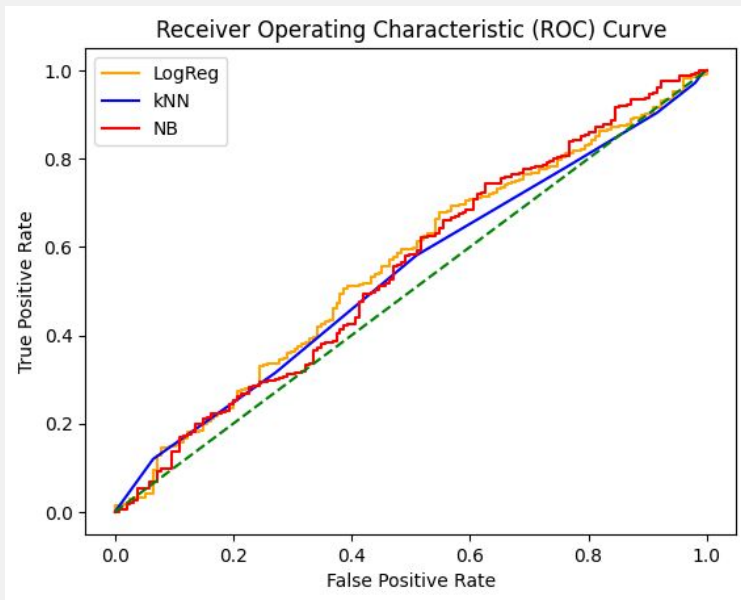
```
Collect and prepare data  →  Transform data  →  Apply association rule mining

Interpret the resulting association rules  ←  Display the results  ←
```

INTERPRET RESULTS

|	Left Hand Side	Right Hand Side	Support(%)	Confidence(%)	Lift	Rules
0	Basket_Size:unknown	With_Kids:no	2.0	66.96	2.069	Basket_Size:unknown -> With_Kids:no
1	Race:unknown	With_Kids:no	1.79999999999998	65.74	2.0314	Race:unknown -> With_Kids:no
2	Race:unknown	Basket_Size:big	1.4000000000000001	44.0	2.0212	Race:unknown -> Basket_Size:big
3	With_Kids:no	Dryer_No:10	1.7000000000000002	22.869999999999997	2.0461	With_Kids:no -> Dryer_No:10
4	With_Kids:no	Dryer_No:10	1.3	25.64	2.0276	With_Kids:no -> Dryer_No:10
5	With_Kids:no	Dryer_No:7	1.6	28.57	2.1153	With_Kids:no -> Dryer_No:7
6	With_Kids:no	Basket_Size:big	1.4000000000000001	24.89	2.2268	With_Kids:no -> Basket_Size:big
7	With_Kids:no	Basket_Size:big	1.0	65.0	2.1608	With_Kids:no -> Basket_Size:big

QUESTION 3

Will a customer purchase drinks in the laundry shop?

```
┌─────────────────┐      ┌─────────────────┐      ┌─────────────────┐
│ Collect and     │ ───▷ │ Perform Boruta  │ ───▷ │                 │
│ prepare data    │      │ feature         │      │ Apply SMOTE     │
│                 │      │ selection       │      │                 │
└─────────────────┘      └─────────────────┘      └─────────────────┘
                                                            │
                                                            ▽
┌─────────────────┐      ┌─────────────────┐      ┌─────────────────┐
│ Select the best │ ◁─── │ Evaluate the    │ ◁─── │ Train           │
│ performing      │      │ performance     │      │ classification  │
│ model           │      │                 │      │ models (LogReg  │
│                 │      │                 │      │ & kNN & NB)     │
└─────────────────┘      └─────────────────┘      └─────────────────┘
```

MODEL EVALUATION - W/O SMOTE

MODEL EVALUATION - W/ SMOTE
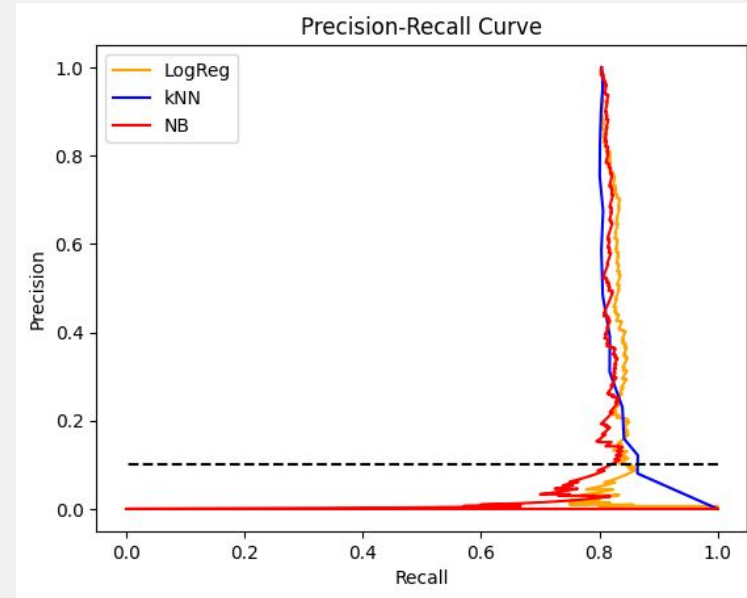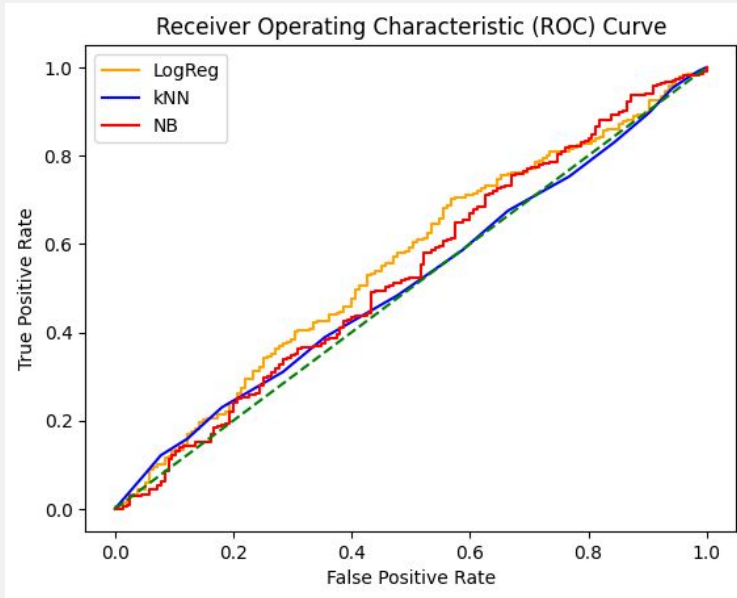
LOGISTIC REGRESSION

	W/O SMOTE	W/ SMOTE
ACCURACY	0.804	0.570
CONFUSION MATRIX	TN = 0 FP = 155 FN = 0 TP = 635	TN = 78 FP = 77 FN = 263 TP = 372
AUC	0.56	0.56
PRECISION-RECALL	0.83	0.83

QUESTION 4

What is the relationship between the weather conditions and the number of customers at the laundry shop?

Collect and prepare data → Perform RFE feature selection → Perform hyperparameter tuning using GridSearch → Train regression models (LinReg & SVR) → Perform hyperparameter tuning using GridSearch → Evaluate the performance → Select the best performing model
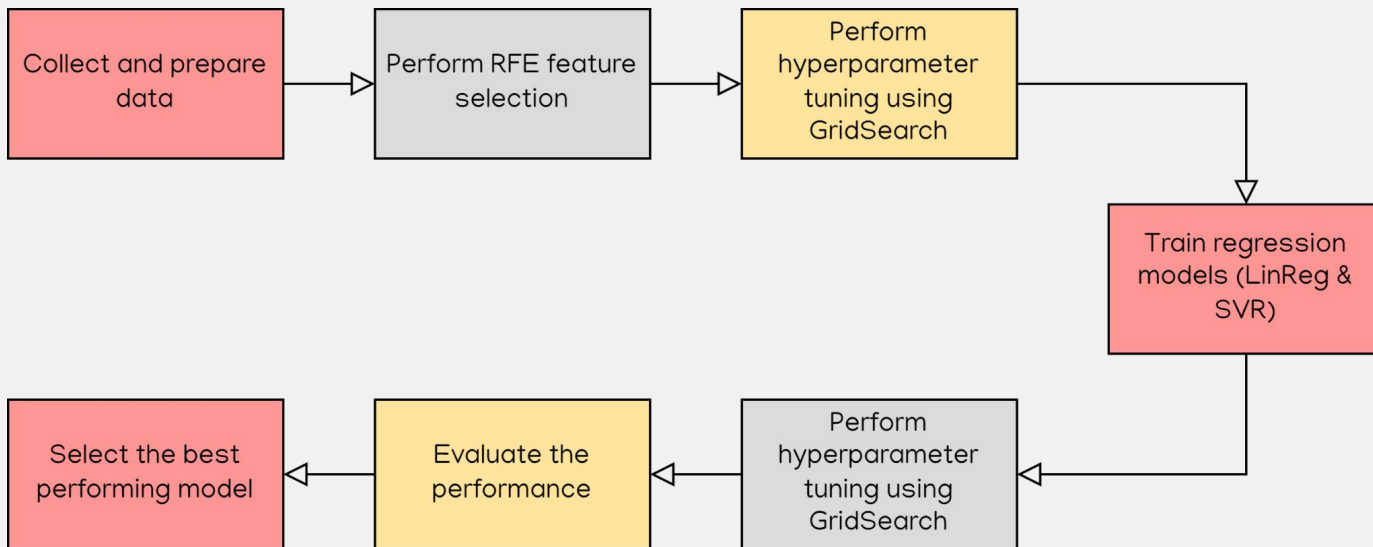
HYPERPARAMETER TUNING

LINEAR REGRESSION

	BEFORE	AFTER	DIFFERENT
MAE	9.149	9.348	△ 0.23
R-SQUARED	0.082	0.014	▽ 0.012

SUPPORT VECTOR REGRESSION

	BEFORE	AFTER	DIFFERENT
MAE	9.384	9.385	△ 0.001
R-SQUARED	−0.024	−0.006	▽ 0.018

THANK YOU!