# Project Title:

# Malaysia’s Job Market Trends & Salary insights: Analysis and Prediction

# Project Background

## Facts

**48.6%** of Malaysian **graduates** were **overqualified** for their current jobs, forcing them to opt for low-skilled jobs with lower starting salaries*

**50%** of skilled workers in **Kelantan** are **overqualified** (exceed national average of 36.9%) **

## Purpose

To build a **model** for salary **prediction** (i.e. job categories, location, job types)

## Target users

- Job seekers
- Employers
- Data analysts

*Report by Khazanah Research Institute (Seraj, 2024)
**Report by World Bank (The Sun, 2024)

# Problem Statement

There is a **need** to use data insights to address **salary gaps** and **job market trends**, and develop **predictive tools** to guide job seekers and employers on salary expectations.


| | |
|---|---|
| **Mismatch in education & job roles** | Graduates struggle to find jobs that match their skill sets |
| **Rise of informal employment** | 3 million Malaysians involved in gig/daily wage jobs<br>- lack stability & fair pay |
| **Underpaid** | Only 10.8% of graduates in 2021 earned >RM3,000 * |

# Project Objectives

| Domain: Human resources | |
|---|---|
| 1 | To analyse the key trends in job market by examining job categories, locations, and salary ranges |
| 2 | To predict the expected salary range for job postings based on the features |
| 3 | To recommend actionable insights for job seekers & employers |

# 6.0 Description of Methodology

**01** **Obtain**

kaggle

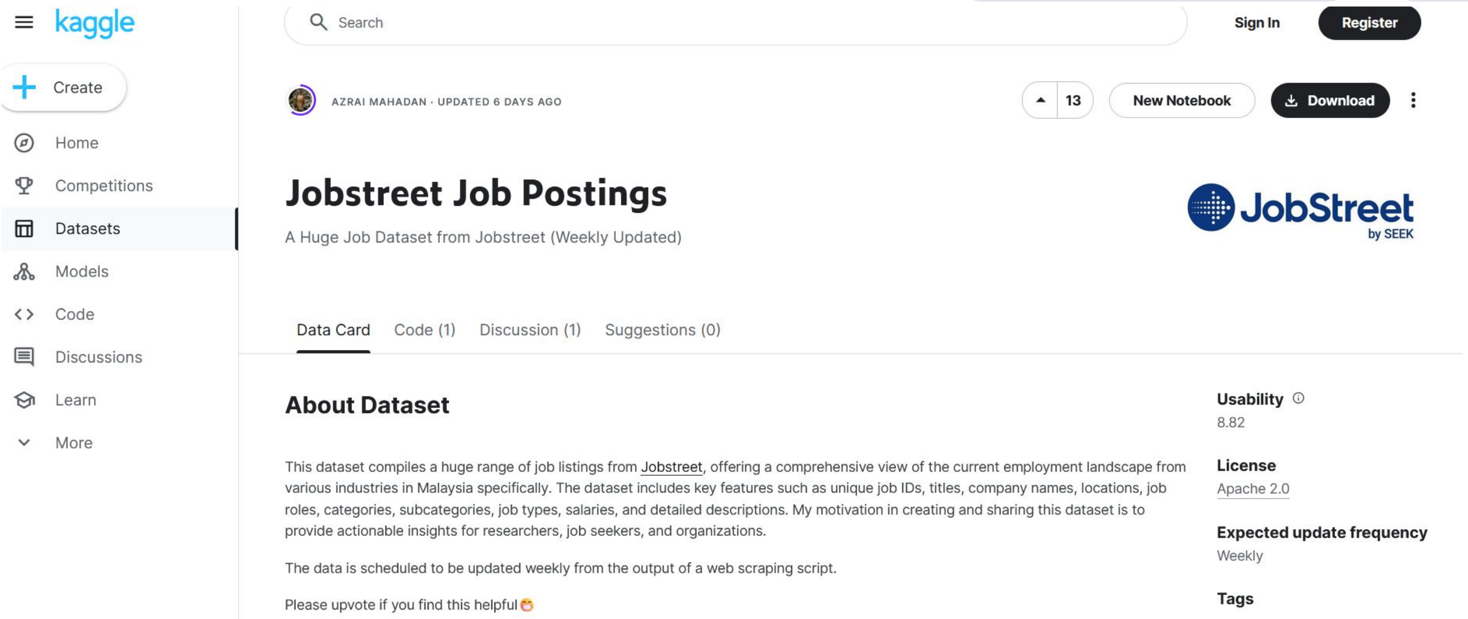
**Job postings** by states, job categories, types, and salaries from 23rd March 2023 - 19th June 2024

python™

Analytical Tool: Statistics and Modeling

Power BI

Visualization Tool: Dashboard



# 6.0 Description of Methodology

## 02 Scrub

### Basic Data Understanding

**Key Data Overview**

- Total rows : 59,306
- Total columns: 10

**Key variables:**

Job id, job title, company, descriptions, location, category, subcategory, role, type, salary and listingDate

**Data quality:**

Missing Data on “Salary” and “Role” columns

**Missing Data Percentage:**

- Salary : 55%
- Role : 3%

# 6.0 Description of Methodology

## 02 Scrub

### Data Cleaning: Handle Missing Values

**Drop** the rows with missing values in “Salary” as “Salary” is our target attribute

### Data Cleaning: Inconsistencies

**Standardize listingDate to date format: ‘yyyy-mm-dd-hh-mm’**

**Standardize salary into integer format**
Preliminary data: RM 3,200 - RM 4,000 per month
**Steps:**

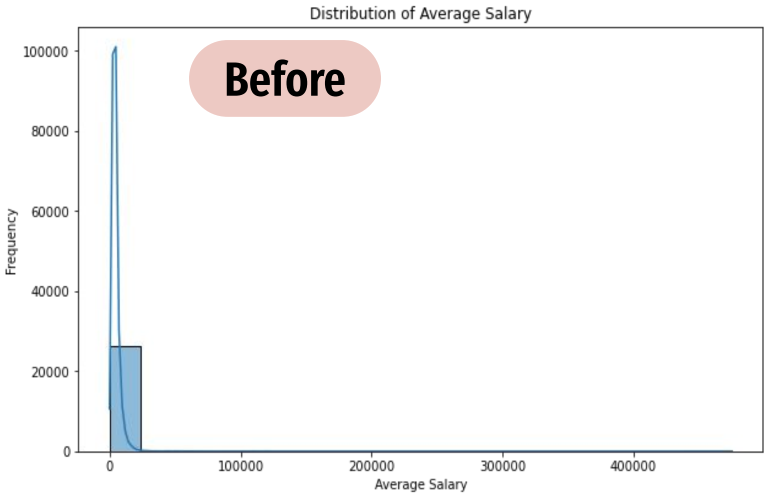
1. Split the salary range into two columns: Lower Bound and Upper Salary using delimiter ‘-’
2. Remove unwanted text and symbols likes “RM”, “MYR”, “$”, “per month” and “p.m.”
3. Convert the Lower Bound and Upper Bound Salary into numeric
4. For salary only contain Lower Bound Salary (**eg: RM3400+), fill the Upper Bound Salary with the Lower Bound value.
5. Create Average Salary column using the formula below:
   Average Salary = (Lower Bound Salary) + (Upper Bound Salary) / 2

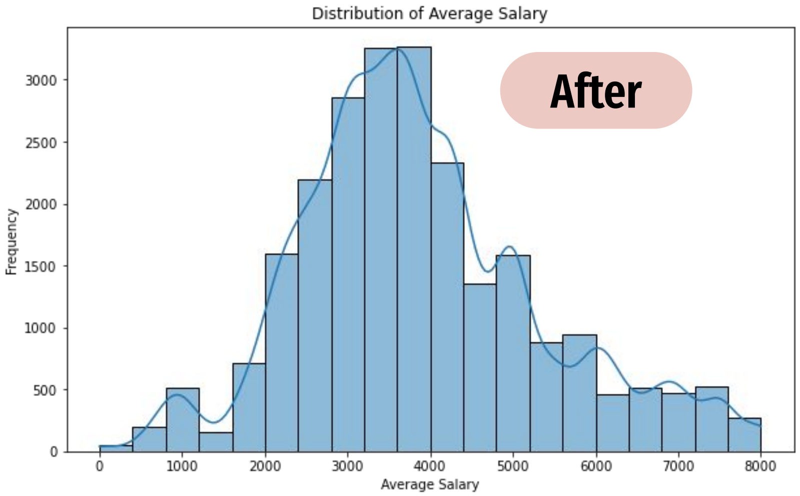
# 6.0 Description of Methodology

## 02 Scrub

### Data Cleaning: Remove Outliers

Remove outliers using IQR
Remove data below Q1 - 1.5 x IQR
Remove data above Q3 + 1.5 x IQR



# 6.0 Description of Methodology

## 02 Scrub

### Data Preprocessing: Handle Categorical Variables

**Challenges:**

- Dealing categorical variables with high cardinality (**a lot of unique values*), such as job location, category, sub-category and roles.
- Direct application of **One-Hot Encoding** will significantly increase the data dimensionality, causing issues like computational inefficiency and overfitting.

**Approach: Map and Group into Larger Categories**

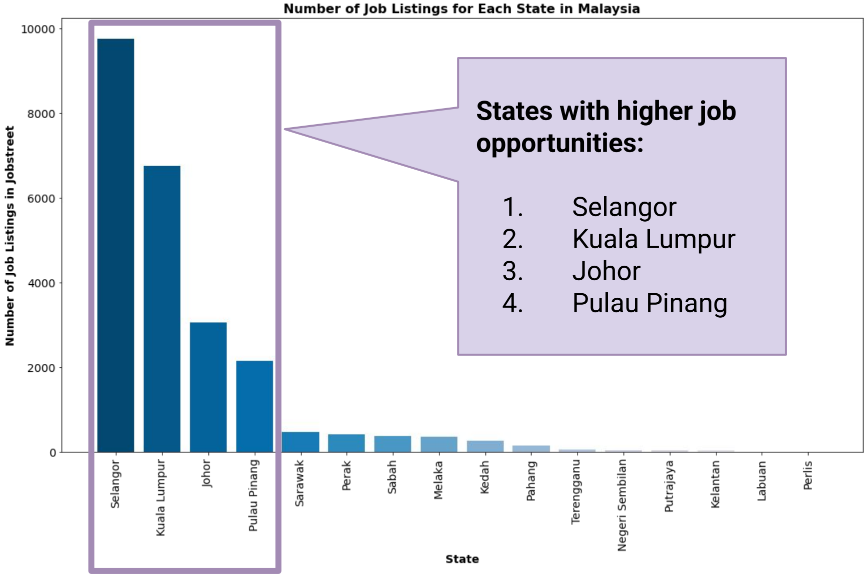
1. **Group the locations by state**
   E.g.: Petaling, Kajang, Klang, Shah Alam to 'Selangor'
1. **Leverage domain knowledge to group similar jobs into broader categories.**
   E.g.: Account, Business to Accounting/Finance

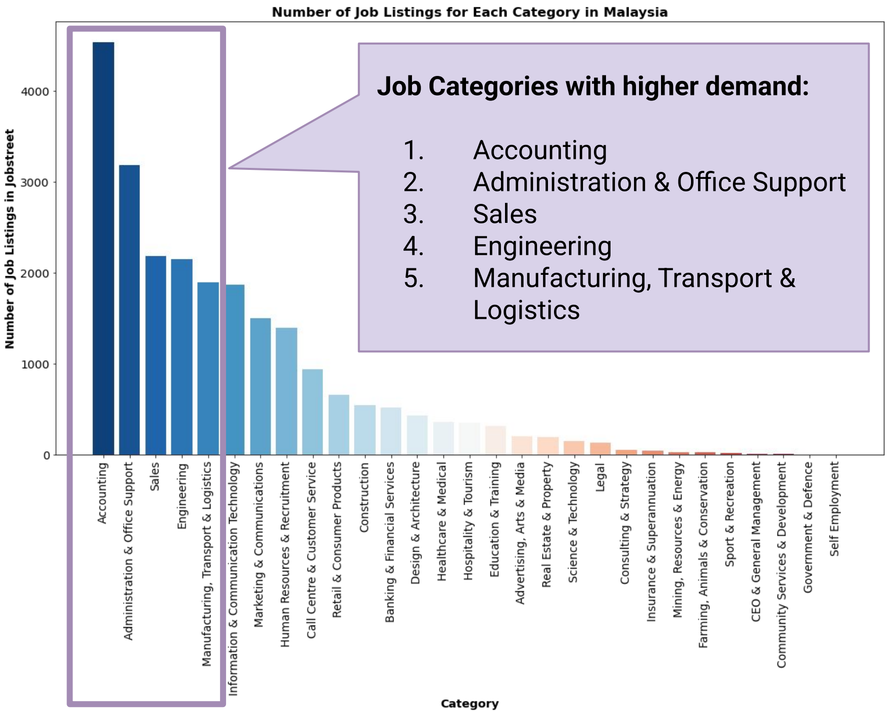
**Outcome:**

After grouping the data, we will apply **One-Hot Encoding** to the simplified categories. This reduced risk of high-dimensionality while retaining information for modelling.

# Exploratory Data Analysis (EDA)

## Univariate analysis



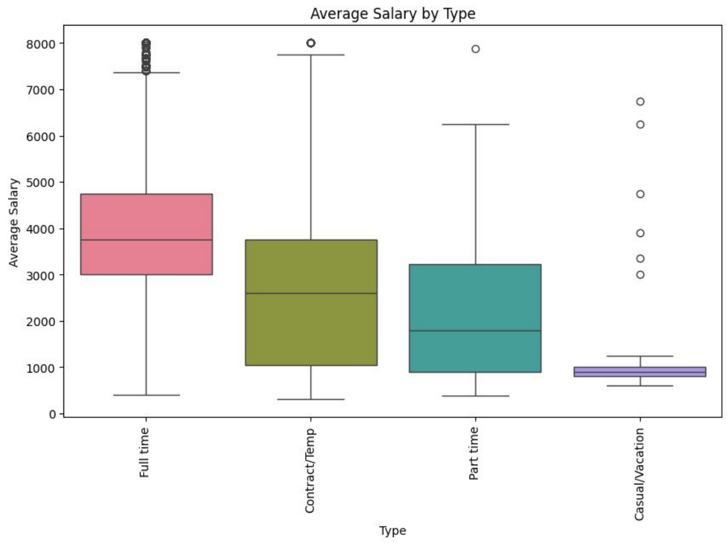

Box plot displayed the frequency distribution of **job location (state)**

Box plot displayed the frequency distribution of **job category**

# Exploratory Data Analysis (EDA)

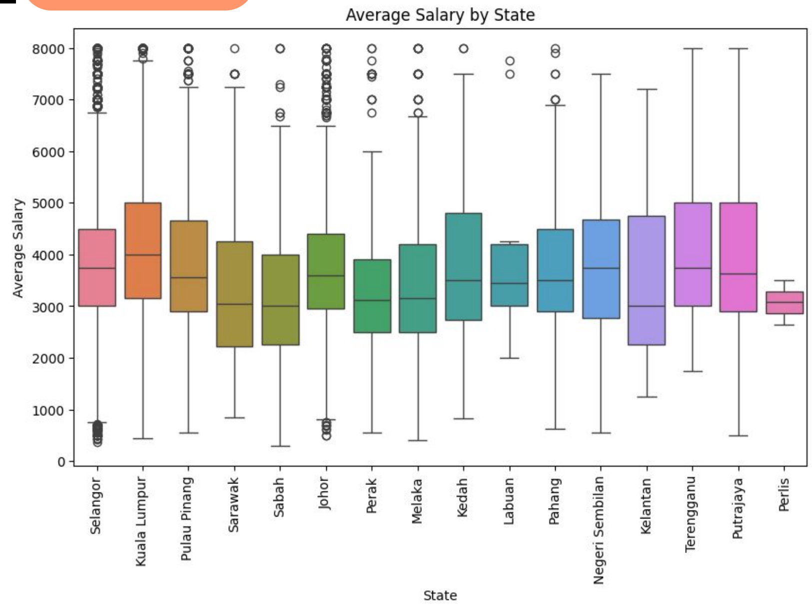
| **Bivariate analysis** | To determine the relationship and statistical association exists between salary average and other variables |
|---|---|

## 01 Job Type




**Full-time jobs** - Highest salary
**Casual /vacation jobs** - Lowest salary
Strong correlation between job type and salary.

## 02 State



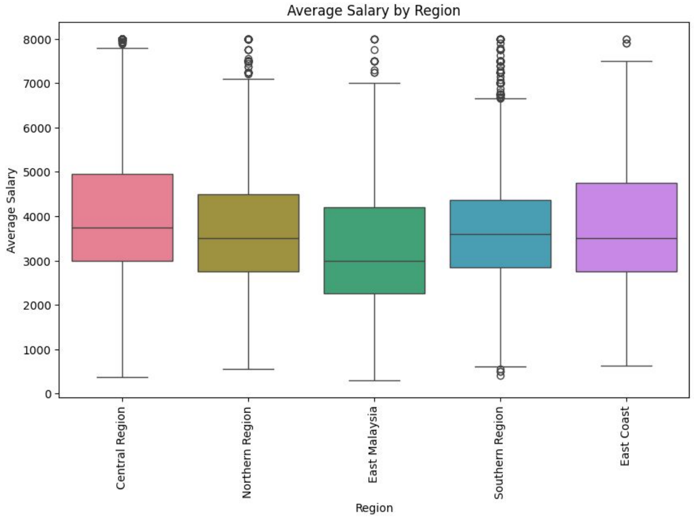

**Kuala Lumpur & Putrajaya** - Highest salary
**Perlis & Sabah** - Lowest salary
Geographic disparities in economic opportunities and wages.

# Exploratory Data Analysis (EDA)

**Bivariate analysis**

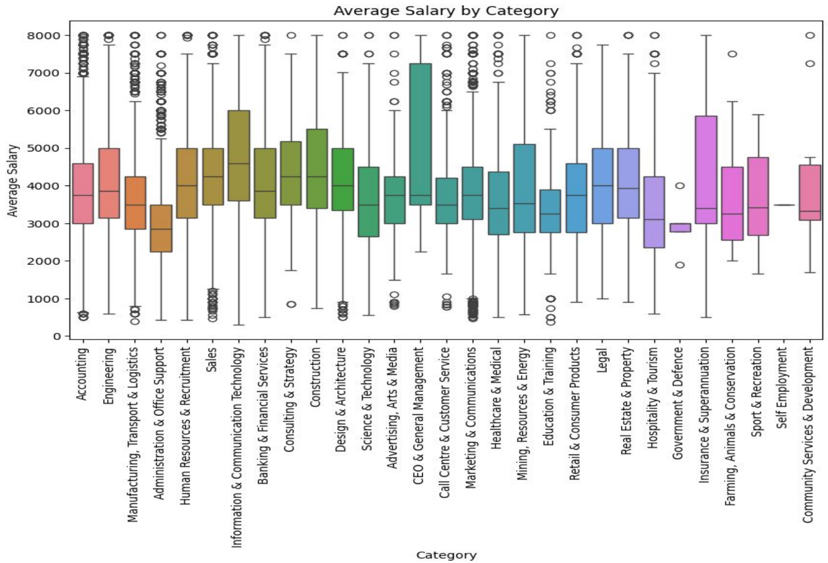
To determine the relationship and statistical association exists between salary average and other variables

## 03 Region




**Central Region (KL, Selangor & Putrajaya)** - Highest salary
**East Malaysia (Sarawak, Sabah & Labuan)** - Lowest salary
Salary differences in geographic and economic.

## 04 Job Category

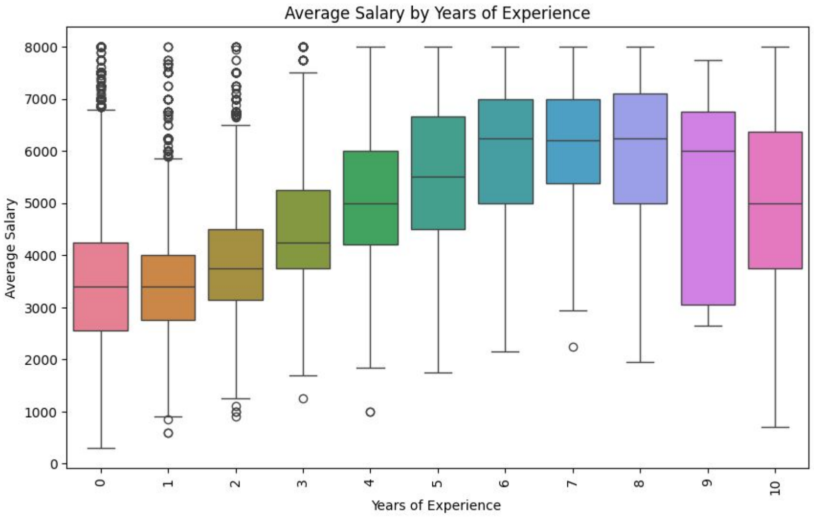



**CEO, ICT, Construction** - Highest salary
**Government, Office Support** - Lowest salary

# Exploratory Data Analysis (EDA)

**Bivariate analysis** — To determine the relationship and statistical association exists between salary average and other variables

## 05 Years of Experience




**6 to 8 years** - Highest salary
**0 to 2 years** - Lowest salary
Salaries grow steadily with experience.

## 06 Management Role


**Management Role** - Highest salary
**Non-Management Role** - Lowest salary
Management role require expertise and carry big responsibility.

# Modelling

**Objective: To build a machine learning model to accurately predict salary based on job details**

## Feature Engineering

### 01 Years of Experience

Extract "**Years of Experience**" from the "Description" column.

### 02 Management Role

Identify and extract "**Management Role**" details from the "Description" column

### 03 Salary category

Categorize salary into 3 groups, following government ranges

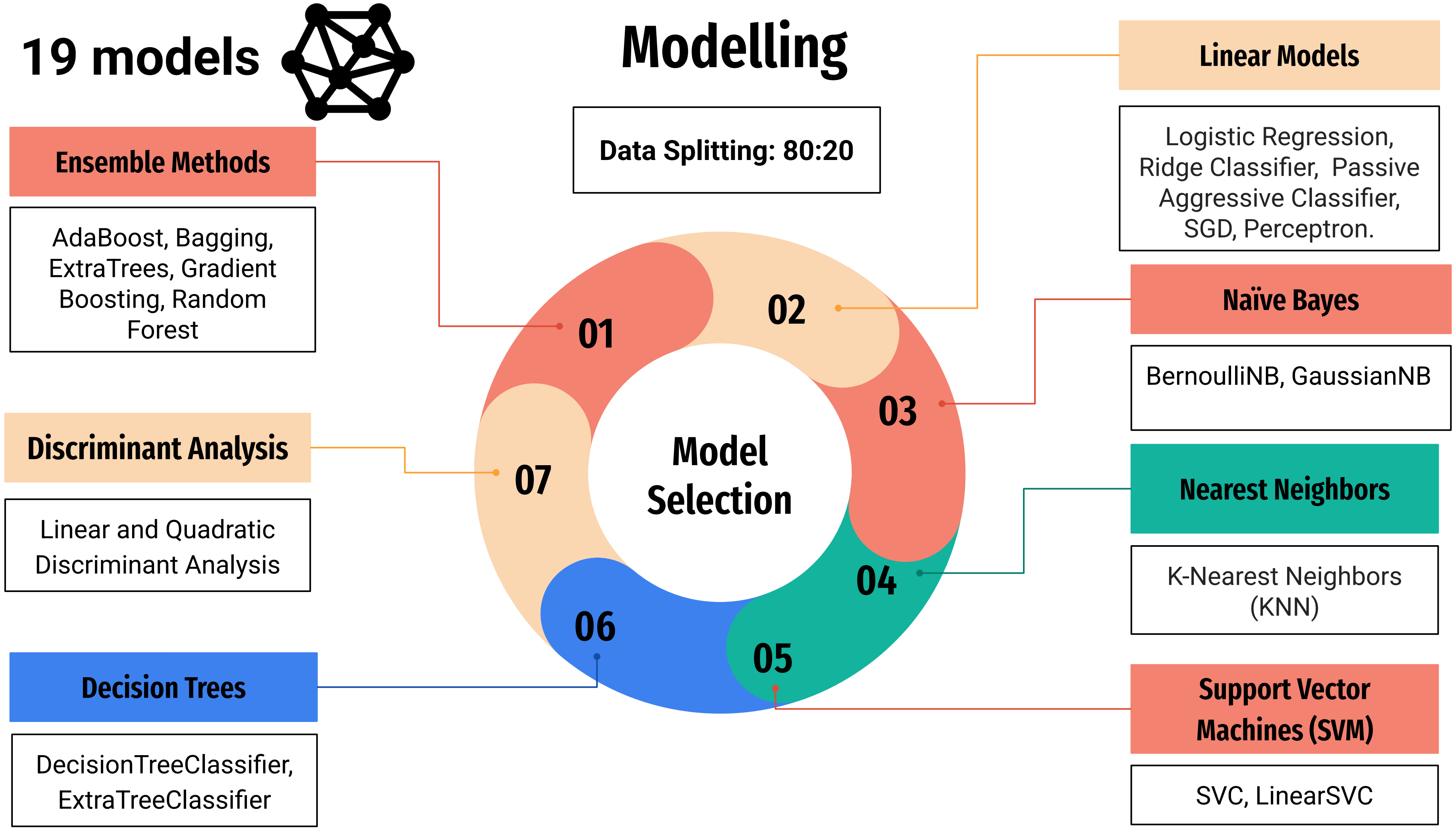
**B40**: < RM5,249

**M40**: RM5,250 to RM11,819

**T20**: > RM11,819

### 04 One-Hot Encoding

Apply "**One-Hot Encoding**" to Categorical Variable

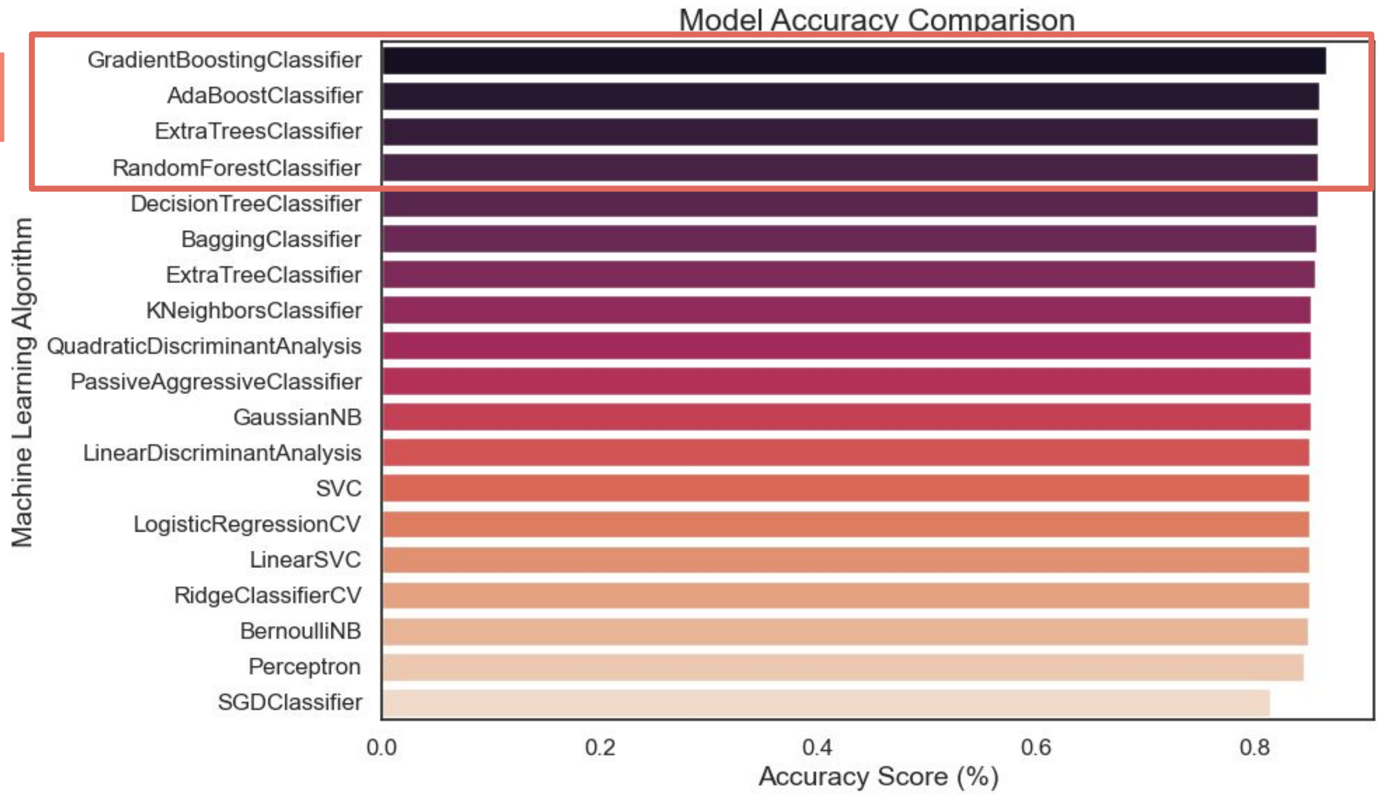

19 models
Modelling
Data Splitting: 80:20
Ensemble Methods
AdaBoost, Bagging, ExtraTrees, Gradient Boosting, Random Forest
Linear Models
Logistic Regression, Ridge Classifier, Passive Aggressive Classifier, SGD, Perceptron.
01
02
03
04
05
06
07
Model Selection
Naïve Bayes
BernoulliNB, GaussianNB
Discriminant Analysis
Linear and Quadratic Discriminant Analysis
Nearest Neighbors
K-Nearest Neighbors (KNN)
Decision Trees
DecisionTreeClassifier, ExtraTreeClassifier
Support Vector Machines (SVM)
SVC, LinearSVC

# Model Evaluation and Data Interpretation

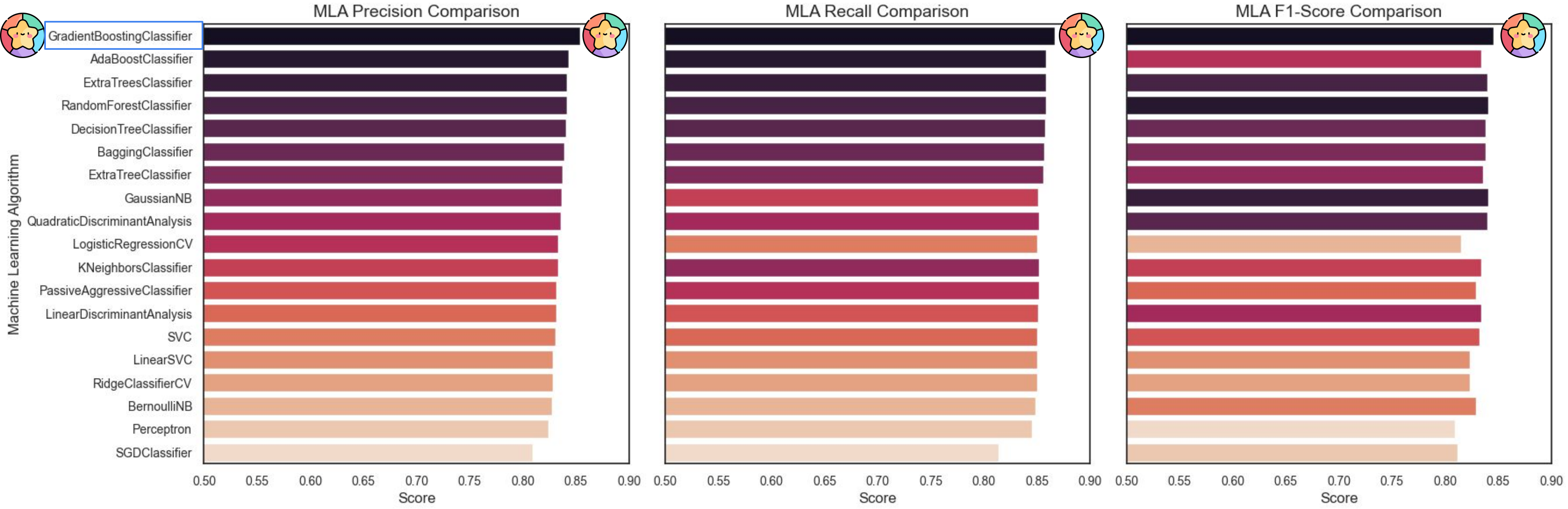
**Cross validation**

**10-fold ShuffleSplit**

**Ensemble**



# Model Evaluation and Data Interpretation




Gradient Boosting computing time = 0.97 s

**Gradient Boosting was the top-performing model.**

# Hyperparameter Tuning

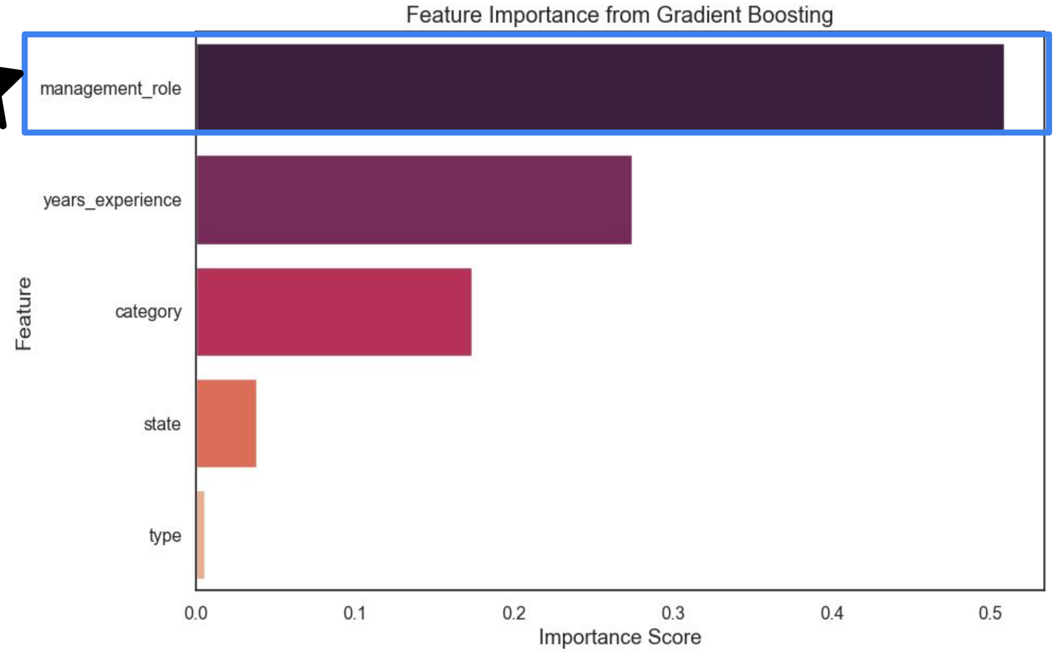
## Hyperparameter Tuning

Performed **Grid Search CV** to fine-tune the model.

### Optimum parameter

| | |
|---|---|
| Max depth | 3 |
| Min Sample Leaf | 2 |
| Min Sample Split | 2 |
| Number of Estimators | 100 |

## Feature Importance Analysis



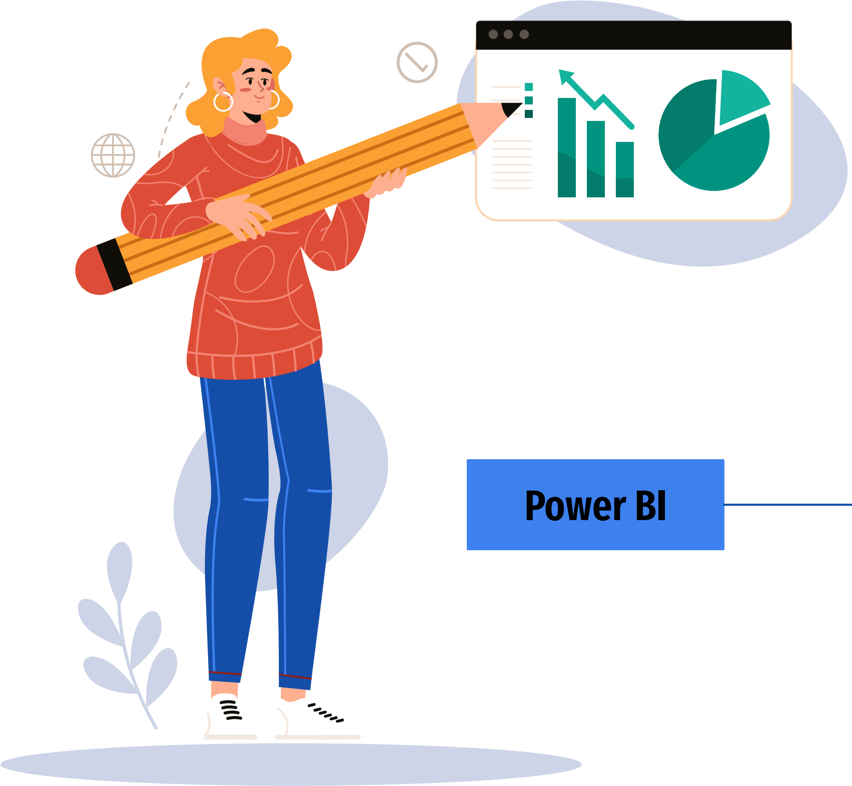

**Best cross validation accuracy: 0.87**

# Reproducible Research

**All steps fully documented and openly shared, allowed others can independently verify and replicate the results.**

| Data Folder Structure | Comments and Explanations | Control Randomness | Version Control | Automate Data Processing |
|---|---|---|---|---|
| Clear folder structure, project objectives and steps to run the project | Added detailed comments explaining the steps in the code | Set random seeds for reproducibility to ensure results are consistent across different runs. | Version history in Google Colab able to highlight the changes, that everyone can track clearly. | Python scripts to automate repetitive tasks like data cleaning and transformation. |

# Deployment of Data Product



**Power BI**

### Job Categories Overview

Identifying Trends in Job Categories and Hiring Companies

### Salary Insights and Trends

Analysing Salary Trends Over Job Categories

### Job Availability and Type Analysis

Understanding Job Availability by Type and Experience

# Insights and Conclusion

## Achievements and Limitations

### Objectives Achieved

- ➔ Key Trends in the Job Markets
- ➔ Expected Salary Range Prediction
- ➔ Actionable Insights

### Limitations in Predictions & Insights

- ➔ Data Coverage
  - ◆ 6 months of job postings
- ➔ Job Distribution
  - ◆ Majority of records are from the Accounting field

### Future Improvements

- ➔ Access to dataset more than a year
- ➔ Train the model with more jobs records
- ➔ Provide even more accurate predictions