SECOND
HAND
MARKET

Wheels of Fortune: Unveiling the Pakistani Used Bike Market

–Term Project:

December 2023

Group Members:

Name: M. Ashar Khalid And Ahmed Pervez Tara
Section: (A) Seat Nos: 19122176 and 19122165

Submitted to: Miss Uzma

Objective:

I have collected a dataset that includes data on motorcycles, specifically targeting models such as the Honda CD 70 and Honda 125. The purpose of this project is to develop a model that can eliminate ambiguity in the second-hand bike market and help maintain a standard price model to ensure consistent pricing standards.

The price of a used bikes based on various features such as:

1. year.
2. milaeage.
3. body type.
4. Model.
5. Price.
6. Stroke.

Introduction and Background:

Used motorcycle pricing is a complex puzzle pieced together by various factors like model, year, mileage, and engine stroke. This project harnesses the power of machine learning, blending supervised and unsupervised techniques, to refine the prediction game. By crunching historical data, the model strives to deliver trustworthy estimates, navigating the ever-shifting terrain of the used motorcycle market.

Data Collection:

Leveraging the rich trove of data on PakWheels, a leading online platform renowned for its detailed listings of used vehicles, this project employed web scraping to gather a robust dataset for motorcycle pricing analysis.

Data preprocessing:

The dataset underwent thorough preprocessing using both Excel and SPSS.

- Excel
  - Used to fill missing values, ensuring a complete dataset.
  - Defined specific ranges for price and mileage.

- SPSS
  - Utilized to eliminate duplicates, ensuring the dataset's integrity.
  - Visualized and inspected values to identify and handle any out-of-domain entries

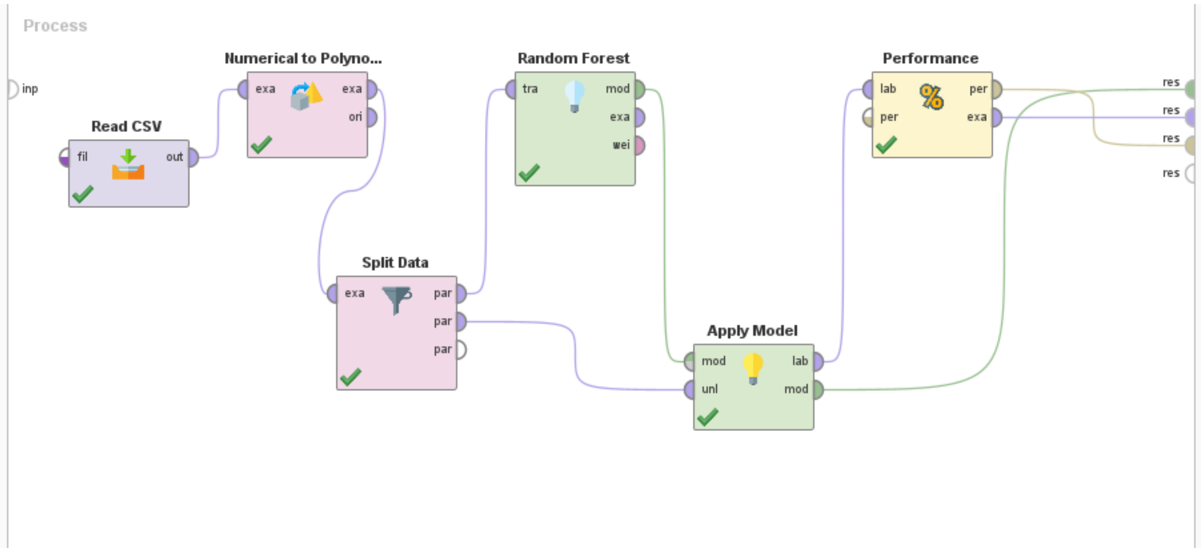
Modelling:

Two primary algorithms, Random Forest and K-Means clustering, were employed to develop and assess the predictive model.

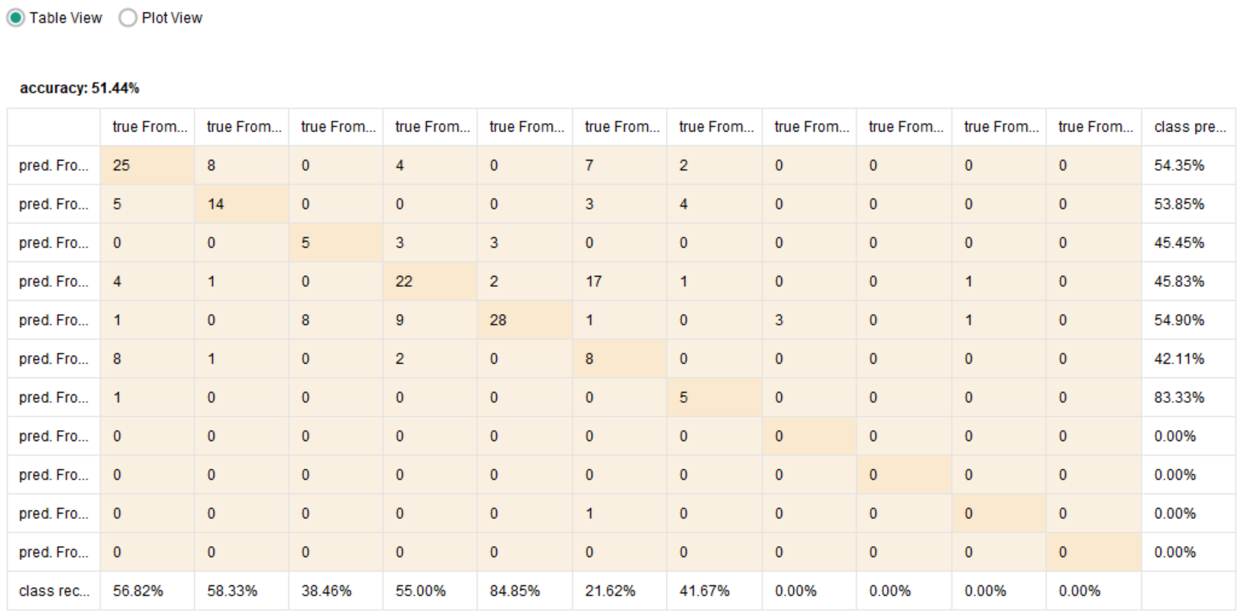
- **Random Forest:**

Utilized the Random Forest algorithm to create a model for predicting used car prices. This ensemble learning method leverages multiple decision trees to enhance prediction accuracy.The dataset was split into training and testing sets to train the model and evaluate its performance accurately.

Design:



Performance Metrics:



● Table View ○ Plot View

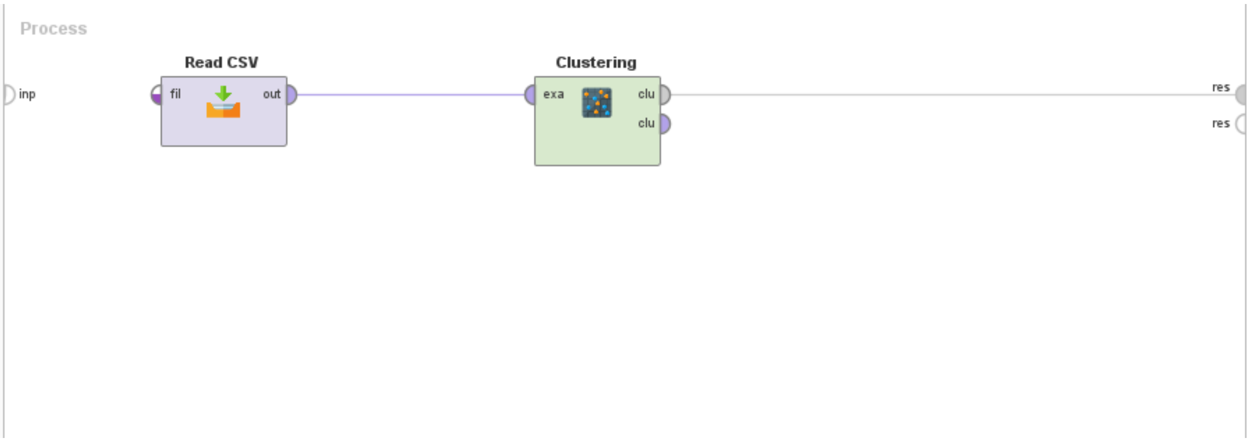
accuracy: 51.44%

	true From...	true From...	true From...	true From...	true From...	true From...	true From...	true From...	true From...	true From...	true From...	class pre...
pred. Fro...	25	8	0	4	0	7	2	0	0	0	0	54.35%
pred. Fro...	5	14	0	0	0	3	4	0	0	0	0	53.85%
pred. Fro...	0	0	5	3	3	0	0	0	0	0	0	45.45%
pred. Fro...	4	1	0	22	2	17	1	0	0	1	0	45.83%
pred. Fro...	1	0	8	9	28	1	0	3	0	1	0	54.90%
pred. Fro...	8	1	0	2	0	8	0	0	0	0	0	42.11%
pred. Fro...	1	0	0	0	0	0	5	0	0	0	0	83.33%
pred. Fro...	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. Fro...	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. Fro...	0	0	0	0	0	1	0	0	0	0	0	0.00%
pred. Fro...	0	0	0	0	0	0	0	0	0	0	0	0.00%
class rec...	56.82%	58.33%	38.46%	55.00%	84.85%	21.62%	41.67%	0.00%	0.00%	0.00%	0.00%	

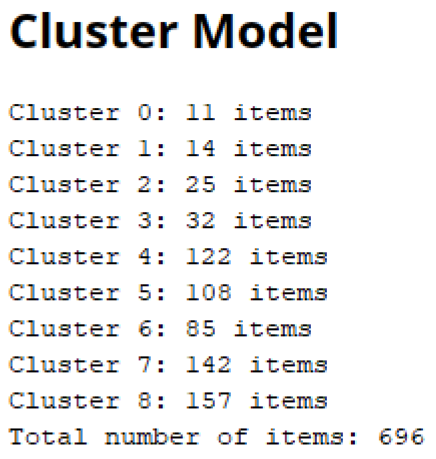
K-Means Clustering:

Implemented the K-Means clustering algorithm to identify patterns and group used cars based on similar features.Determined the optimal number of clusters to maximize the effectiveness of grouping. Used k value 9 and measure type is MixedMeasures.

Design:



Clusters:



Cluster Model

```
Cluster 0: 11 items
Cluster 1: 14 items
Cluster 2: 25 items
Cluster 3: 32 items
Cluster 4: 122 items
Cluster 5: 108 items
Cluster 6: 85 items
Cluster 7: 142 items
Cluster 8: 157 items
Total number of items: 696
```

Conclusion:

The Random Forest model demonstrates promising potential with a 50% accuracy rate in predicting used motorcycle prices. By further refining the model and data, we expect to significantly improve its performance.

The K-means clustering algorithm has effectively segmented the second-hand bike market data into 9 clusters, with sizes ranging from 11 to 157 items. The largest cluster may represent the most common attributes of bikes in the market, while the smallest cluster could indicate less common, niche market characteristics. With a total of 696 items categorized, this clustering provides a solid foundation for developing a standardized pricing model for motorcycles like the Honda CD 70 and Honda 125, aiming to reduce pricing ambiguity in the market.

The dataset can be find on below github repository:
https://github.com/AsharKhalidMehar/output-bike-detils.csv