Hyper Supra 1.O

Supra Without Brakes

Automated EDA and Machine Learning Framework

Developer:

TC.Antony - Data Scientist.

- Manual Book

Purpose:

- Automated Exploratory Data Analysis (EDA) with business insights.

- Machine learning capabilities for classification, regression, and clustering.

Features:

- Automatic Summary Statistics: Provides quick insights into dataset structure.
- Visualization Tools: Includes histograms, scatterplots, and boxplots for data exploration.
- PCA for Dimensionality Reduction: Reduces feature space for better analysis.
- Outlier Detection: Identifies anomalies in the dataset.
- Correlation Insights: Highlights relationships between variables.
- Supervised Learning: Supports classification and regression tasks.
- Unsupervised Learning: Includes clustering algorithms like KMeans, DBSCAN, and Agglomerative Clustering.

- Cross-Validation and Model Evaluation: Ensures robust model performance.
- Feature Engineering and Preprocessing: Handles missing values, encoding, and scaling.
- Business Insights: Provides actionable conclusions for decision-making.
- Downloadable Reports and Datasets: Allows users to export results and cleaned datasets.

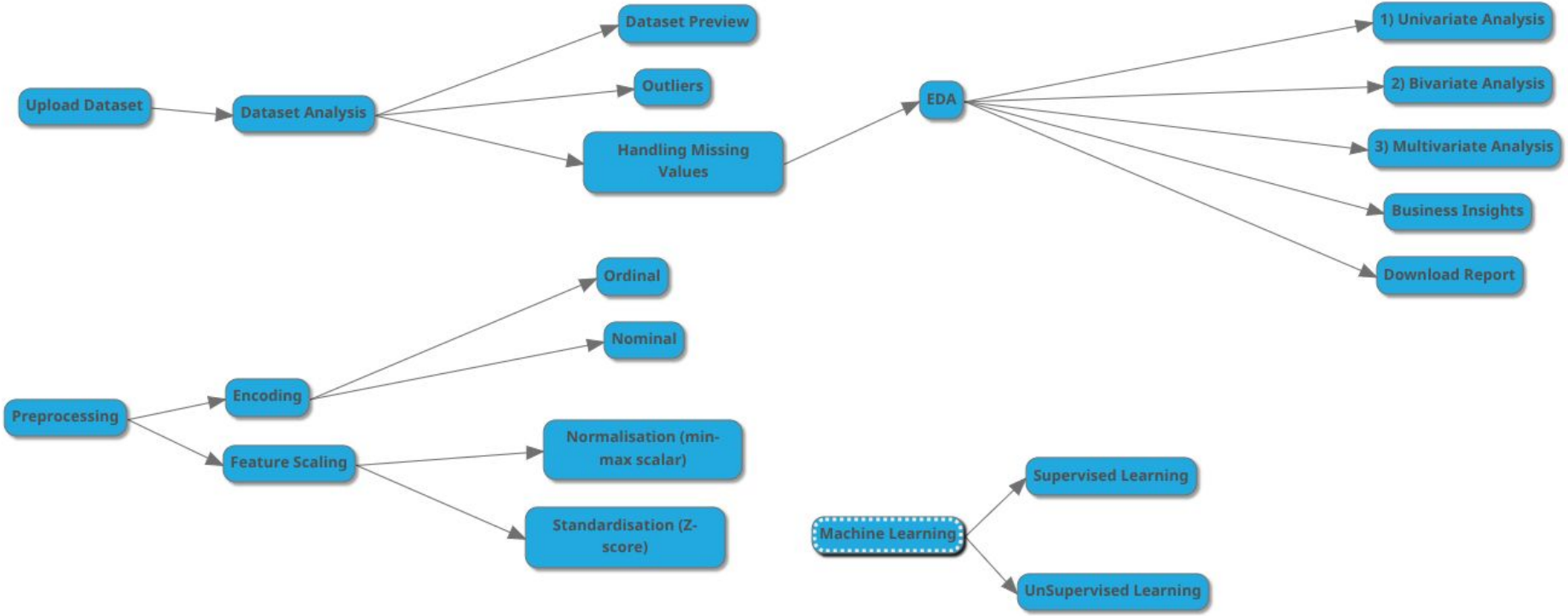
Instructions:

- Upload Dataset: Upload datasets in CSV, Excel, or JSON format.
- Dataset Analysis: Explore summary statistics, outliers, and missing values.
- EDA Process: Perform univariate, bivariate, and multivariate analysis.
- Preprocess: Handle missing values, encode categorical variables, and scale features.
- Machine Learning: Train and evaluate supervised and unsupervised models

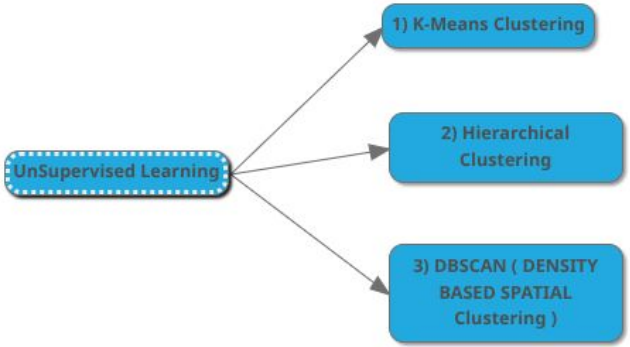
Dependencies:

Streamlit, Pandas, NumPy, Scikit-learn, Seaborn, Matplotlib, XGBoost, LightGBM, Pandas Profiling.

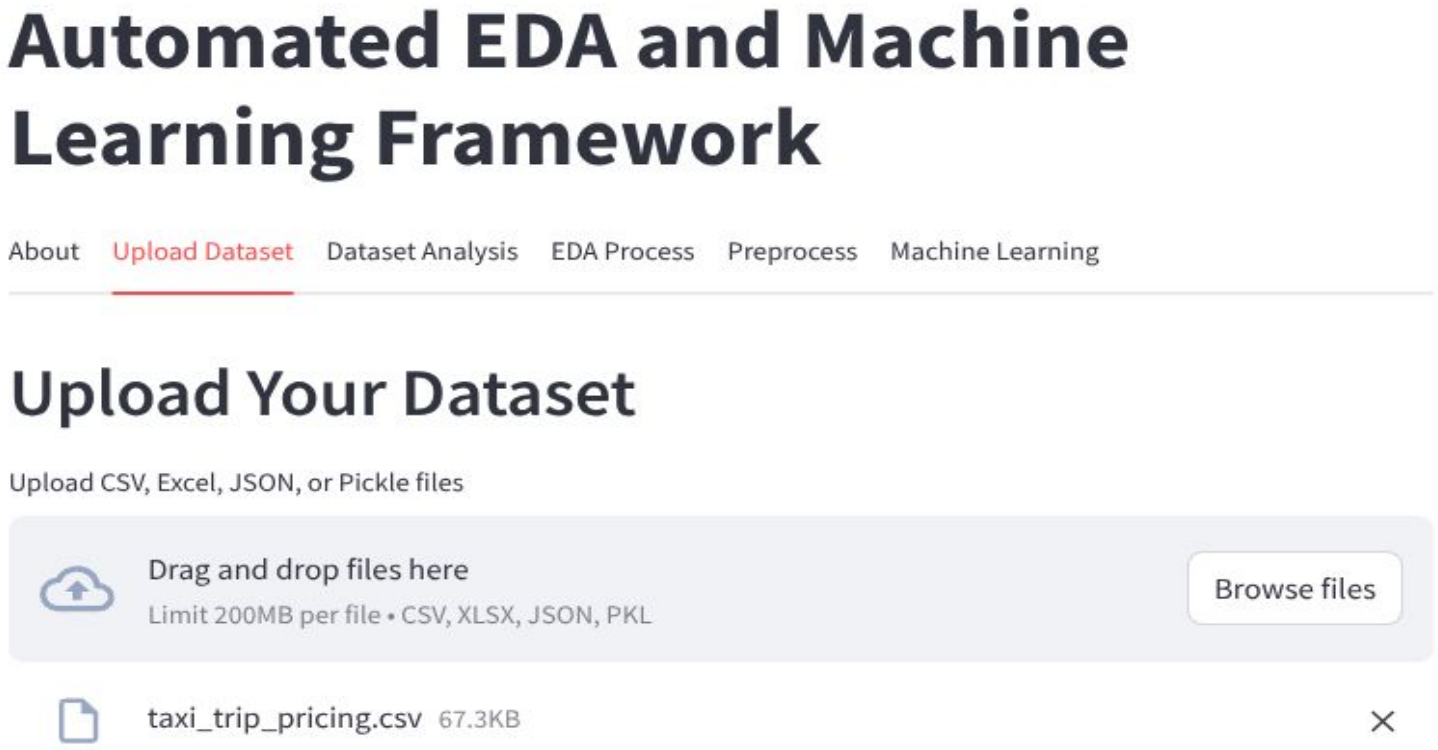
Mind Map:


Upload Dataset
Dataset Analysis
Dataset Preview
Outliers
Handling Missing Values
EDA
1) Univariate Analysis
2) Bivariate Analysis
3) Multivariate Analysis
Business Insights
Download Report
Preprocessing
Encoding
Ordinal
Nominal
Feature Scaling
Normalisation (min-max scalar)
Standardisation (Z-score)
Machine Learning
Supervised Learning
UnSupervised Learning


Supervised Learning
Regressor
Classification
Continuous
Discrete
1) Linear Regression
2) Ridge Regression L2
3) Lasso Regression L1
4) Elasticnet
1) Decision Tree Regressor
2) Random Forest Regressor
3) Boosting Regressor
4) KNN
5) SVM
Binary
Multiclass
1) Logistic Regression
2) SVM Classifier
1) Decision Tree Classifier
2) Random Forest Classifier
3) Boosting Classifier
4) KNN Classifier


UnSupervised Learning
1) K-Means Clustering
2) Hierarchical Clustering
3) DBSCAN (DENSITY BASED SPATIAL Clustering)

Step 1: Dataset Upload (CSV, EXCEL,JSON,PICKLE)



Step: 2 - Data Preview & Summary

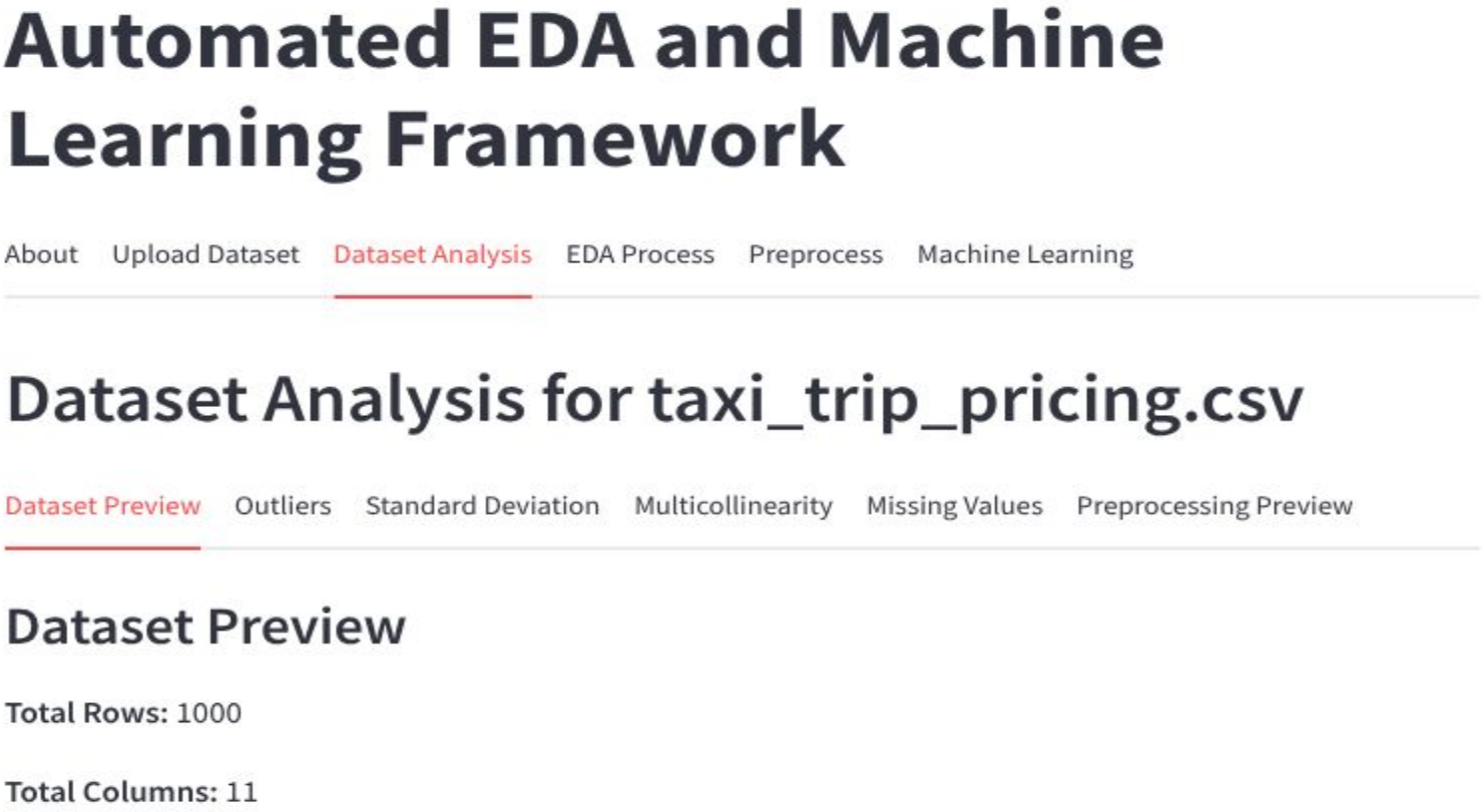
Data Preview

	Trip_Distance_km	Time_of_Day	Day_of_Week	Passenger_Count	Traffic_Conditions	Weather	Base
0	19.35	Morning	Weekday	3	Low	Clear	
1	47.59	Afternoon	Weekday	1	High	Clear	
2	36.87	Evening	Weekend	1	High	Clear	
3	30.33	Evening	Weekday	4	Low	None	
4	None	Evening	Weekday	3	High	Clear	

Data Summary

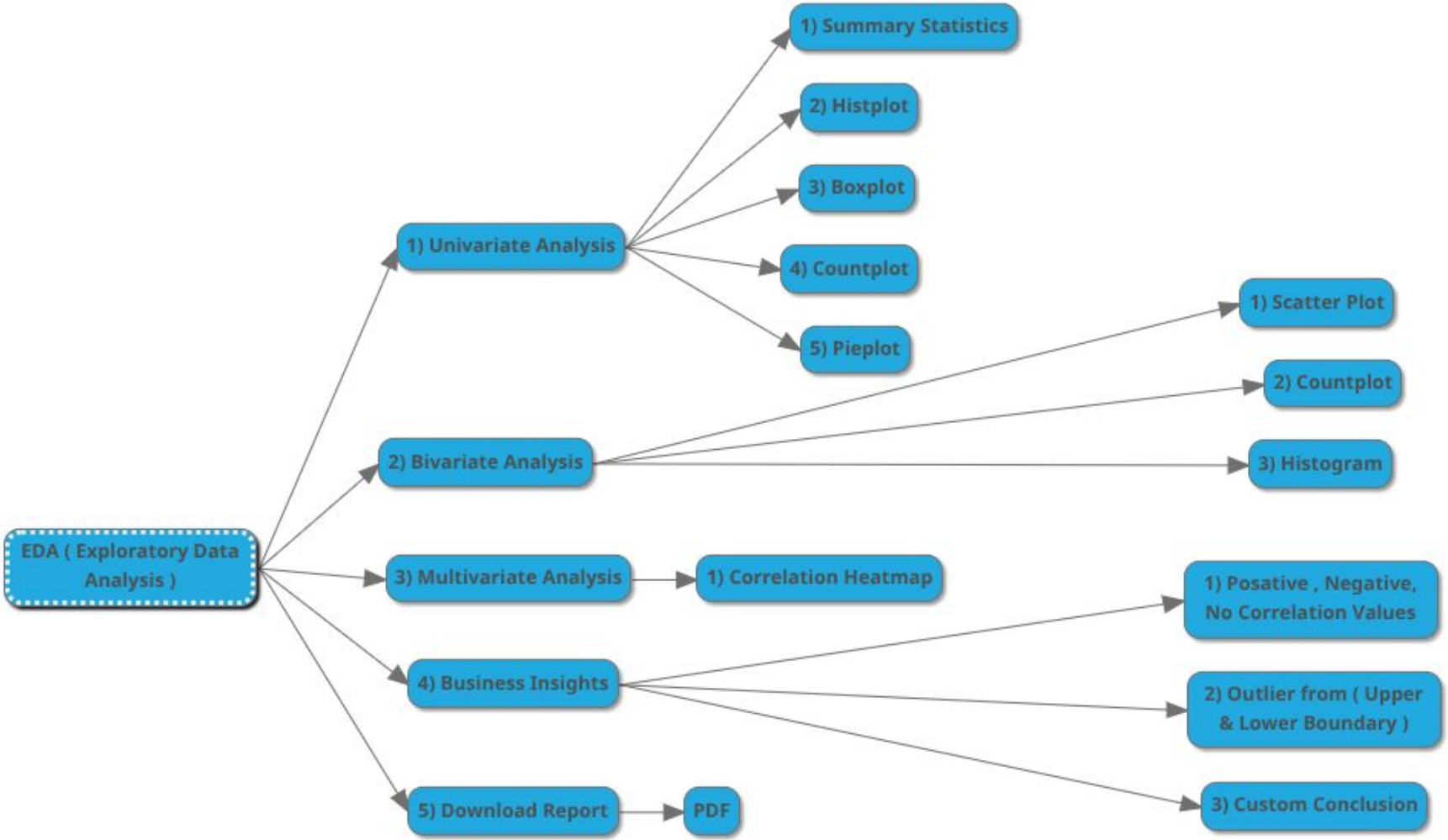
	Trip_Distance_km	Passenger_Count	Base_Fare	Per_Km_Rate	Per_Minute_Rate	Trip_Duration_Minu
count	950	950	950	950	950	
mean	27.0705	2.4768	3.503	1.2333	0.2929	62.1
std	19.9053	1.1022	0.8702	0.4298	0.1156	32.1
min	1.23	1	2.01	0.5	0.1	5
25%	12.6325	1.25	2.73	0.86	0.19	35.8
50%	25.83	2	3.52	1.22	0.29	61
75%	38.405	3	4.26	1.61	0.39	89.

Step 3: Dataset Analysis

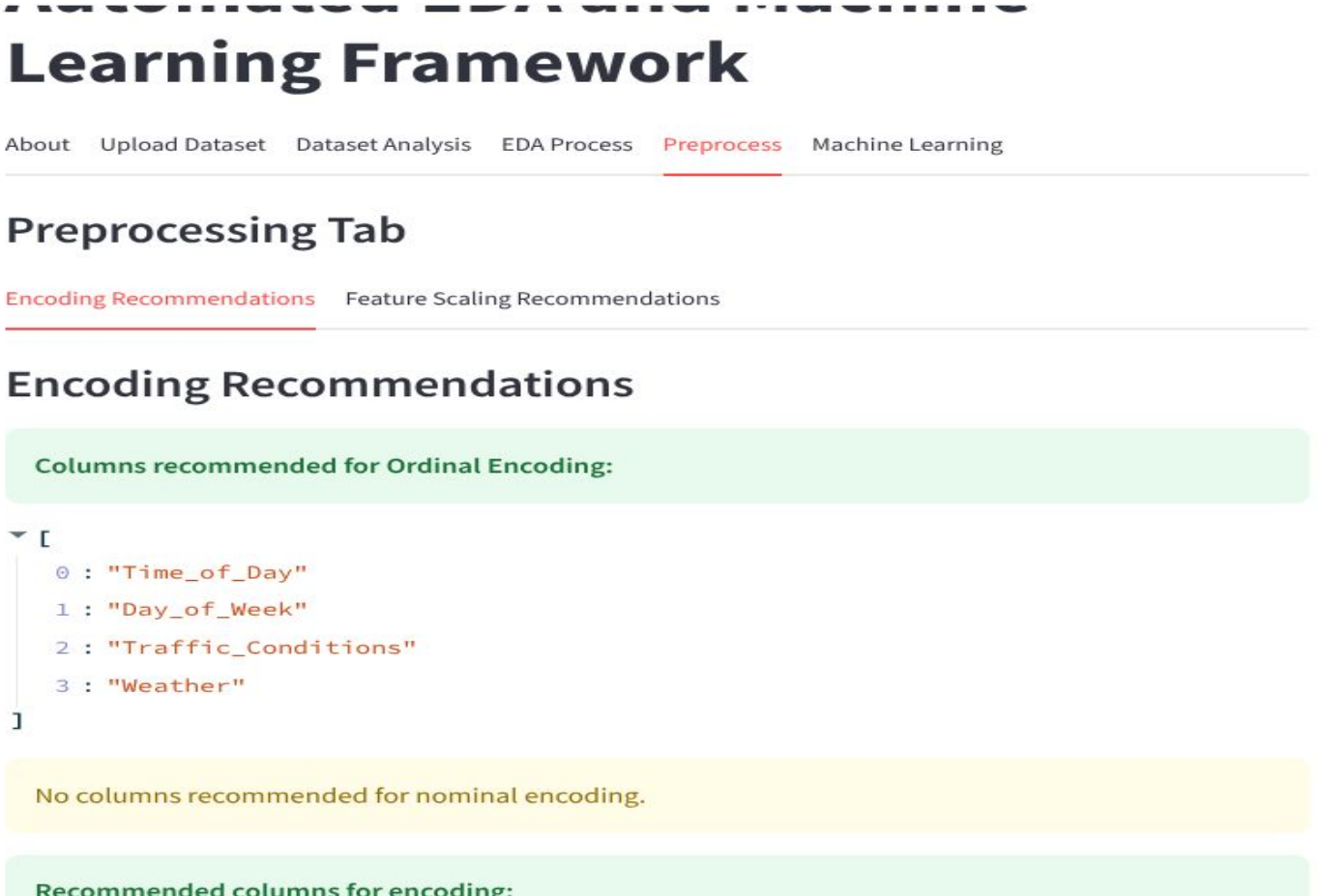


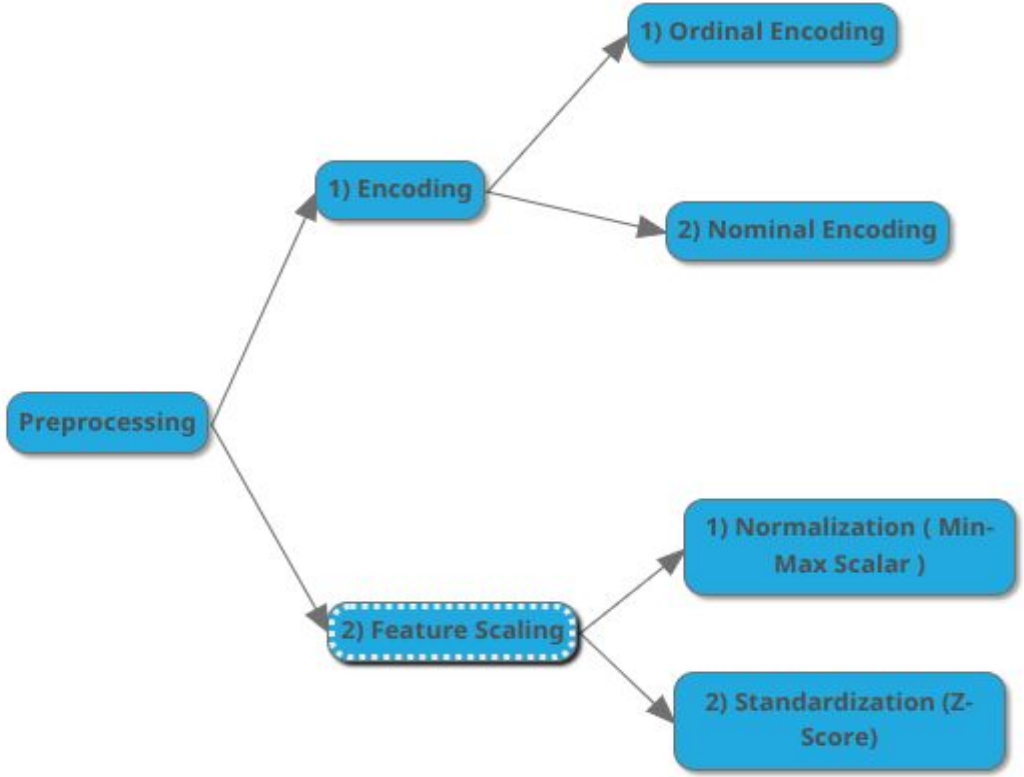

Dataset Analysis
Dataset Preview
Categorical Colum
numerical Column
Outliers
Upper Boundary
Lower Boundary
Standard Deviation
same SD values Multiple Columns will Delete to reduce Feature Selection
Multicollinearity
Missing Values
Mean, Median, Mode, Ffill, Bfill, Custom
Based on Client Opinion to fill
filled based on skewed or Outliers
Preprocessing Preview
Encoding Recommendation
Feature Scaling Recommendation

Step 4: EDA (Exploratory Data Analysis)

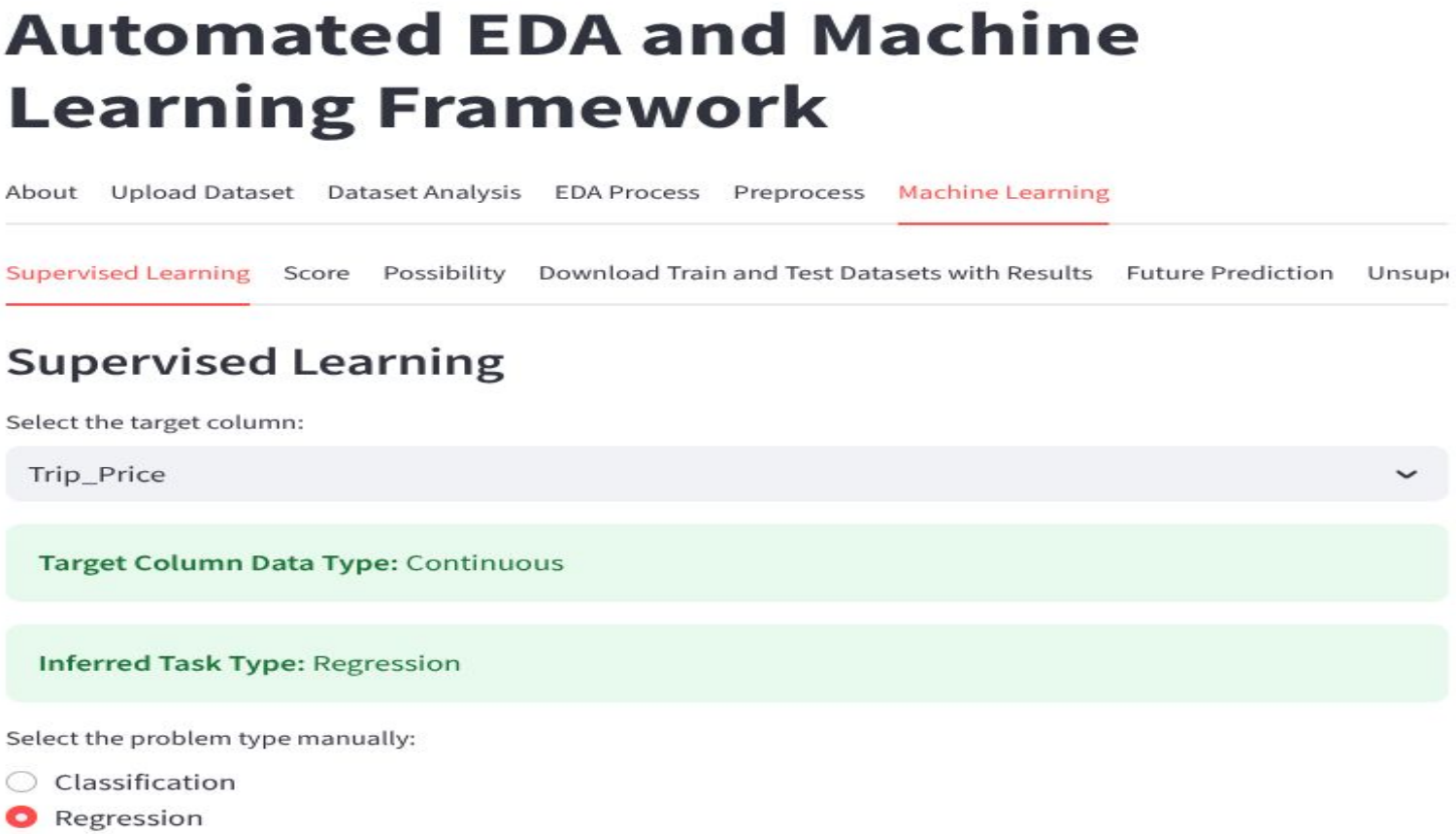

EDA (Exploratory Data Analysis)
1) Univariate Analysis
1) Summary Statistics
2) Histplot
3) Boxplot
4) Countplot
5) Pieplot
2) Bivariate Analysis
1) Scatter Plot
2) Countplot
3) Histogram
3) Multivariate Analysis
1) Correlation Heatmap
4) Business Insights
1) Posative , Negative, No Correlation Values
2) Outlier from (Upper & Lower Boundary)
3) Custom Conclusion
5) Download Report
PDF

Step 5: Preprocessing




Preprocessing
1) Encoding
1) Ordinal Encoding
2) Nominal Encoding
2) Feature Scaling
1) Normalization (Min-Max Scalar)
2) Standardization (Z-Score)

Step 6: Supervised Learning (Regressor)



Step 7: Algorithm Selection

Time_of_Day × Day_of_Week × Traffic_Conditions × Base_Fare × Weather ×

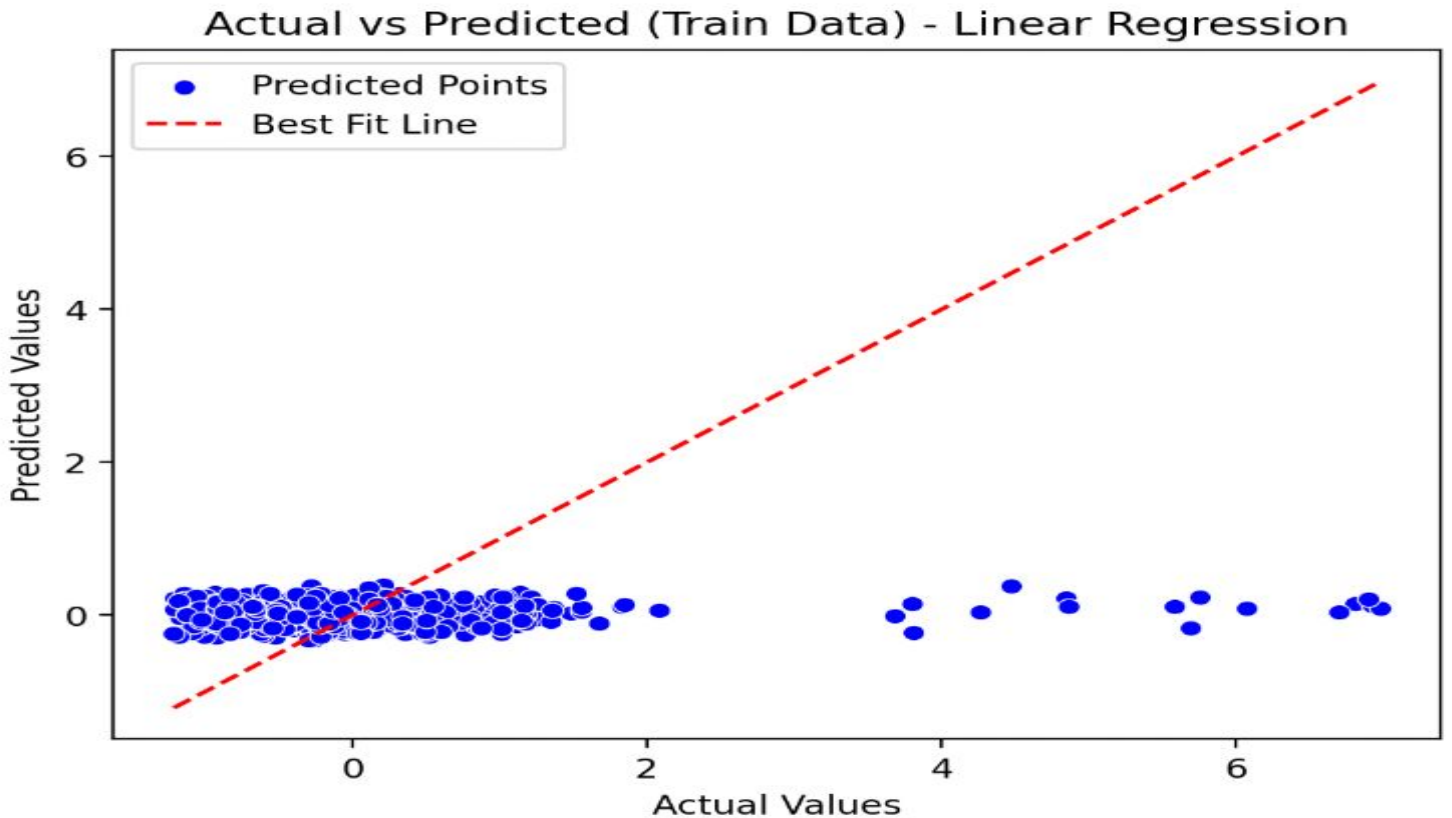
Select Algorithm

Linear Regression

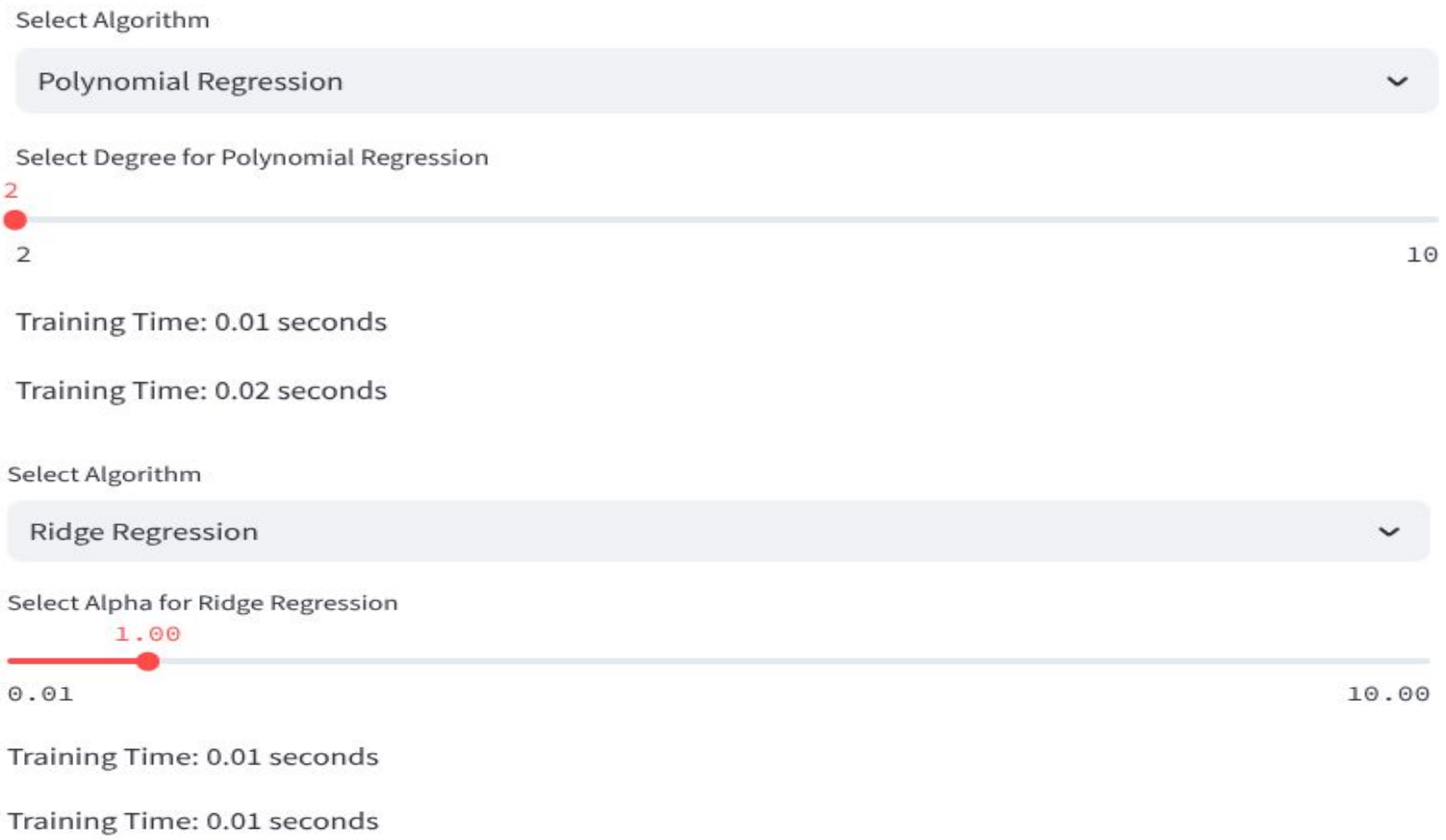
- Linear Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression
- SVM Regression
- Decision Tree Regressor
- Random Forest Regressor

PCA 2d Chart based on Algorithm (Target & Feature)

Data has more than 2 features. Reducing dimensions using PCA...

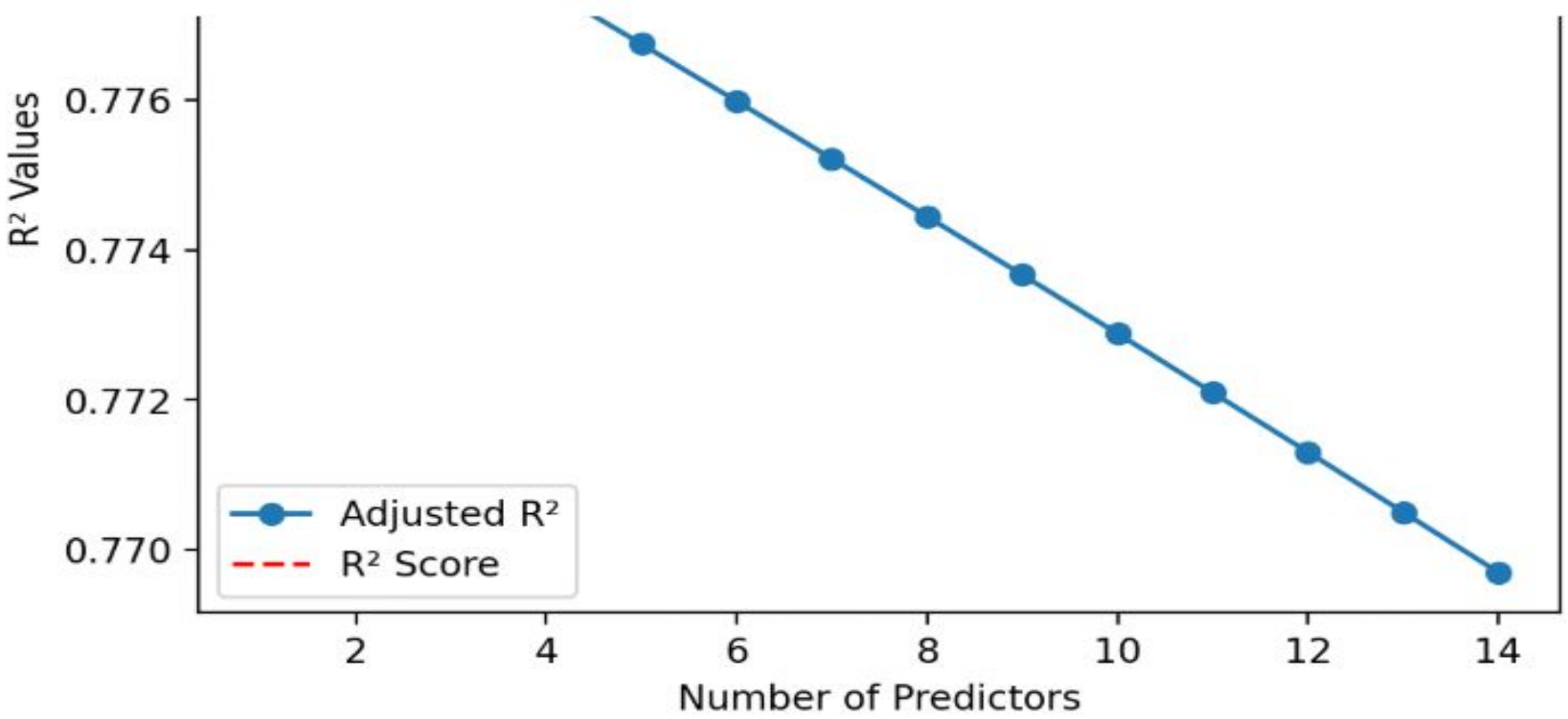


Polynomial , L1 & L2 (Hyperparameter Option)



Step 8 : Scores for Regressor (Both Train & Test Date) with Chart

1) MSE (Mean Square Error)
2) MAE (Mean Absolute Error)
3) R2
4) Adjusted R2



Adjusted R^2 Score on Training Data: 0.8512

Adjusted R^2 Score on Test Data: 0.7729

Adjusted R² Score on Test Data: 0.7729

Training vs Test Accuracy (R^2 Score)

Training R^2 Score: 0.8534

Test R^2 Score: 0.7805

The model may be overfitting, as the training accuracy is significantly higher than the test accuracy.

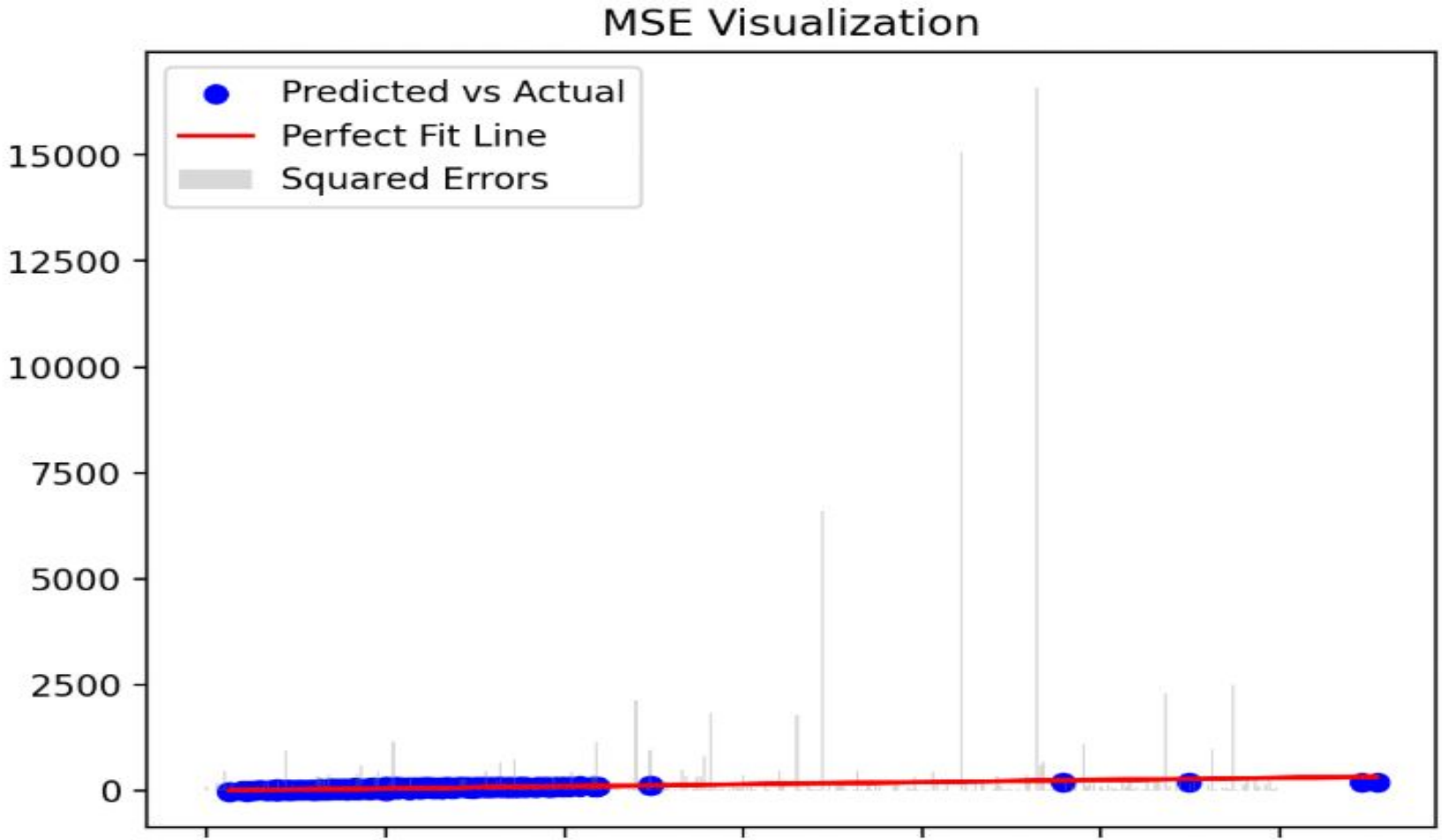
Training vs Test Accuracy (Adjusted R^2 Score)

Training Adjusted R^2 Score: 0.8512

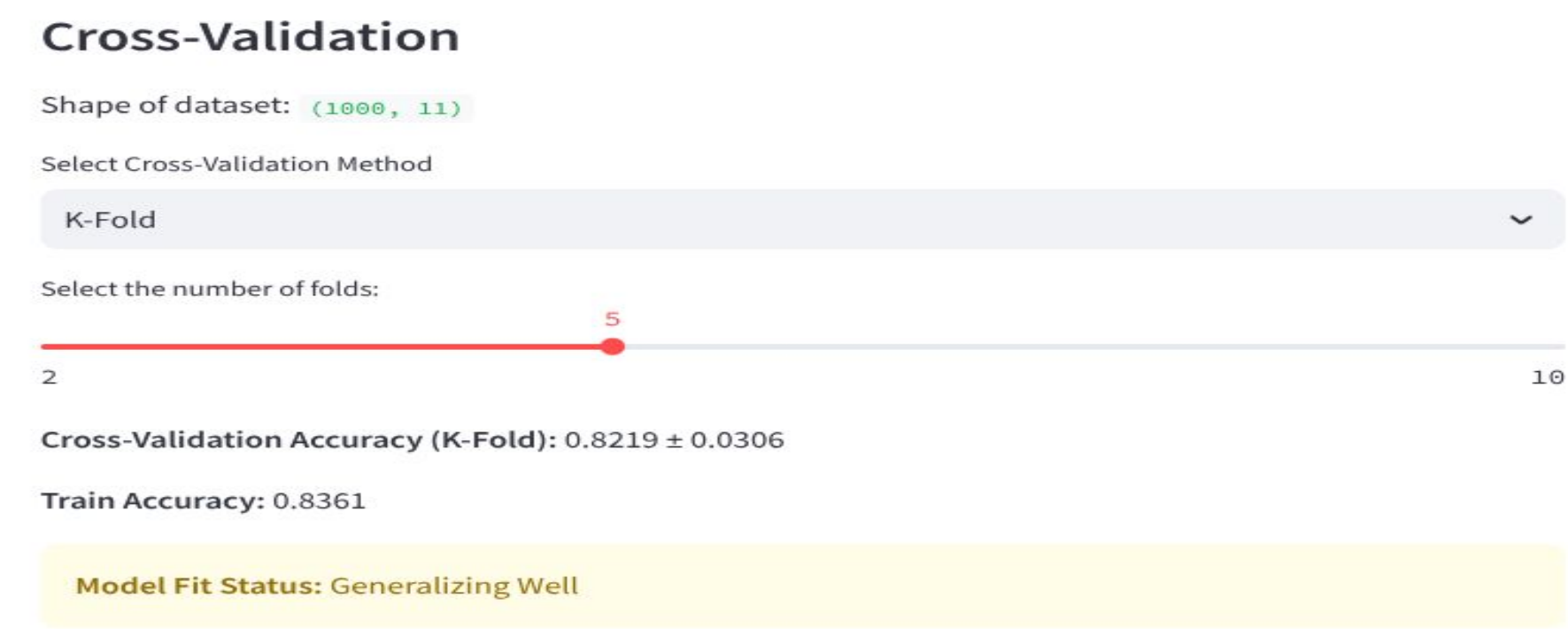
Test Adjusted R^2 Score: 0.7729

The model may be overfitting, as the training accuracy is significantly higher than the test accuracy.

Definition: Mean Squared Error (MSE) measures the average squared difference between actual and


MSE Visualization
Predicted vs Actual
Perfect Fit Line
Squared Errors
15000
12500
10000
7500
5000
2500
0

Step 9: Cross Validation with hyperparameter option



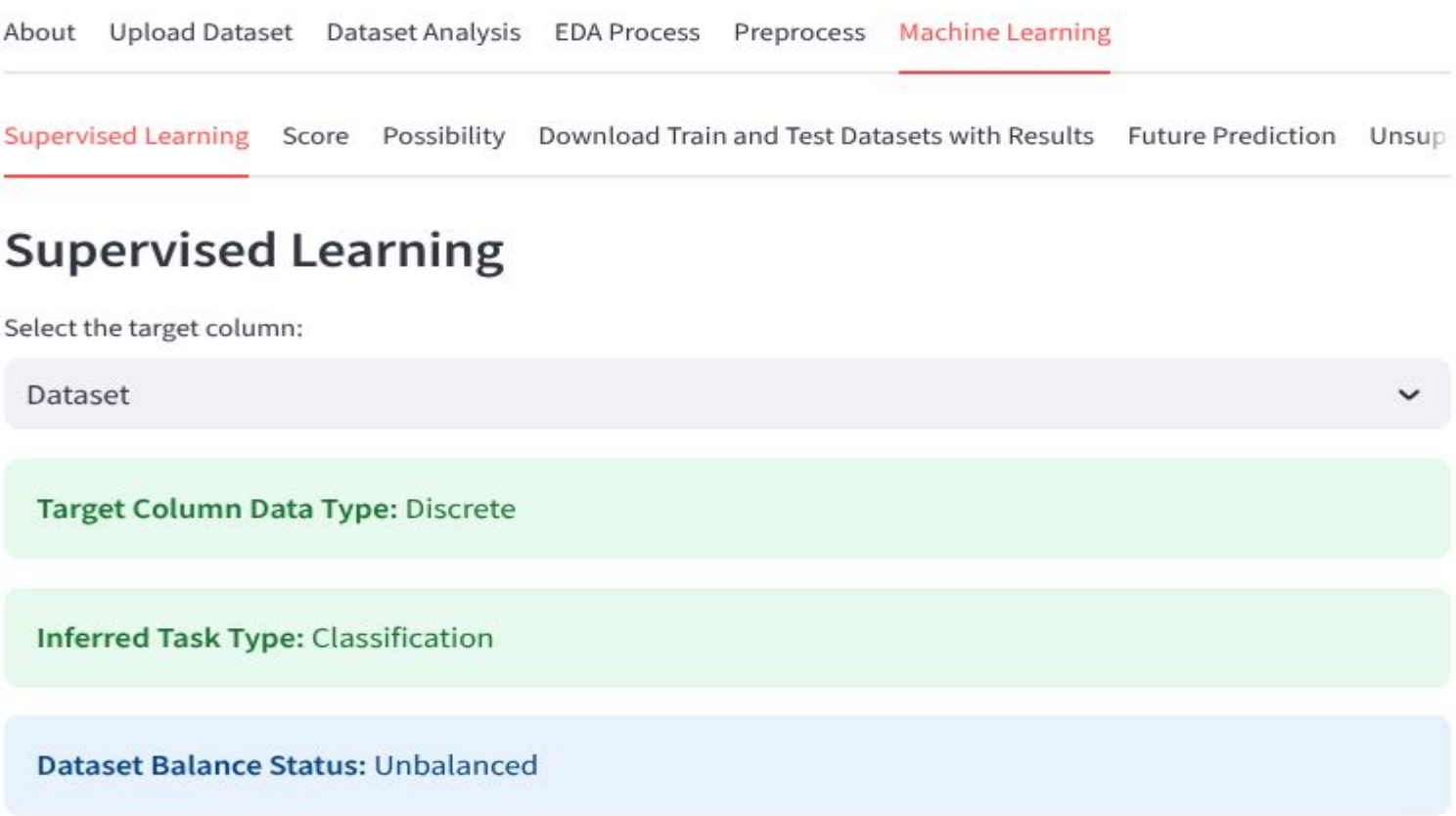
Recommendation

- **K-Fold:** Suitable for general cases with balanced datasets.
- **Stratified K-Fold:** Best for imbalanced classification problems.
- **Holdout:** Quick but less reliable for small datasets.
- **Leave-One-Out:** Best for very small datasets but computationally expensive.
- **Leave-P-Out:** Rarely used due to computational cost.

Cross-Validation Best Practices

	Method	Best for Classifier	Best for Regressor
0	K-Fold	Yes	Yes
1	Stratified K-Fold	Yes	No
2	Holdout (80/20 Split)	No	Yes
3	Leave-One-Out	Yes	Yes
4	Leave-P-Out	No	No

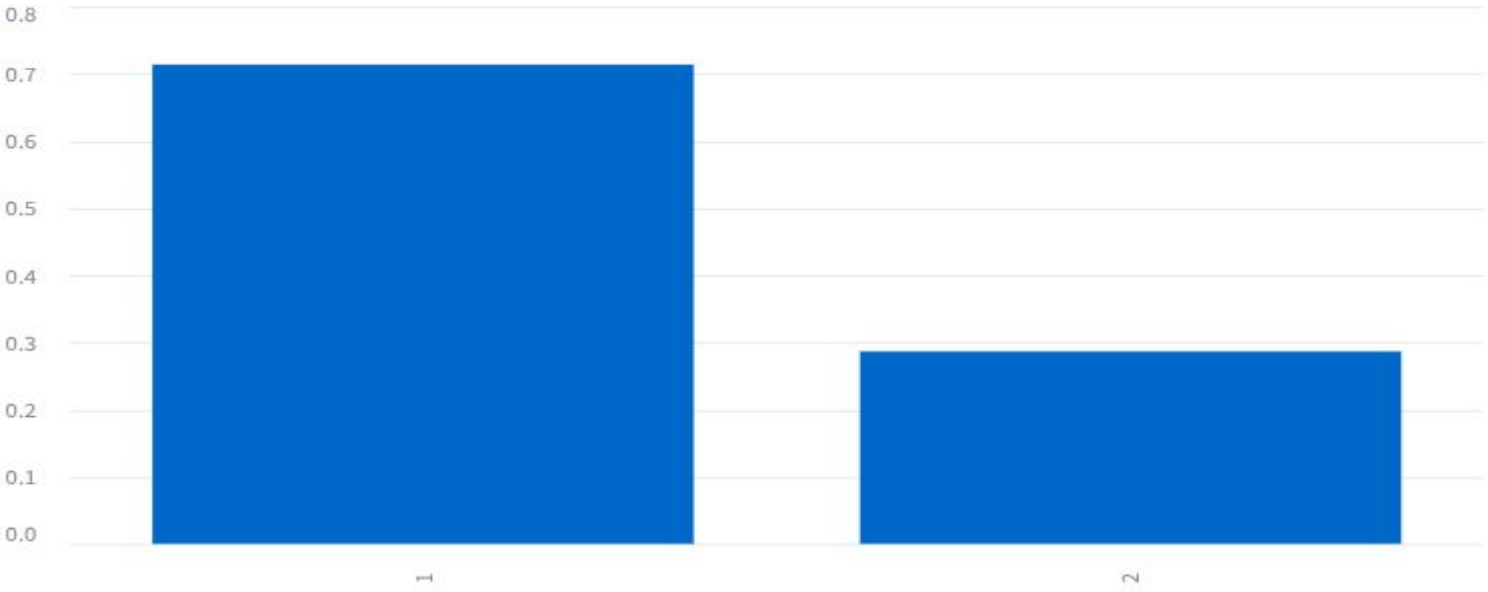
Step 10: Supervised Learning (Classification)



Class Distribution Status

Dataset Balance Status: Unbalanced

Class Distribution:



Suggested Binary Classification Algorithms:

- Logistic Regression
- SVM Classifier

Select feature columns

Age × Gender × Total_Bilirubin × Direct_Bilirubin ×

Select Algorithm

Logistic Regression

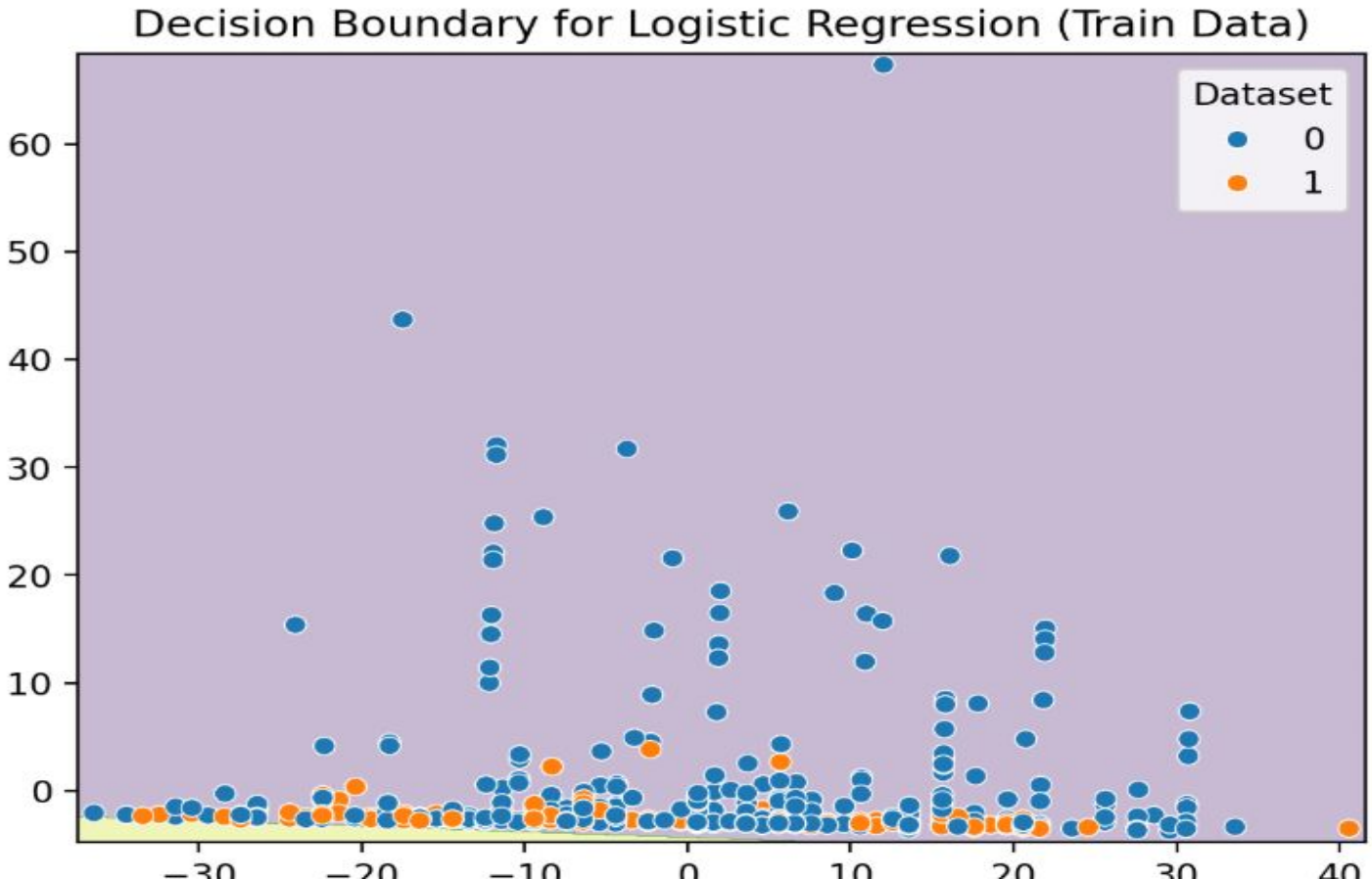
Training Time: 0.03 seconds

Training Time: 0.02 seconds

Select data for visualization:

- (•) Train Data
- ( ) Test Data

PCA Chart for Classification



Score for Train & Test Data

Score

Train Metrics

Accuracy: 0.6985

Precision: 0.4524

Recall: 0.4972

F1 Score: 0.4190

Test Metrics

Accuracy: 0.7257

Precision: 0.3649

Recall: 0.4961

F1 Score: 0.4205

Model Fit Status

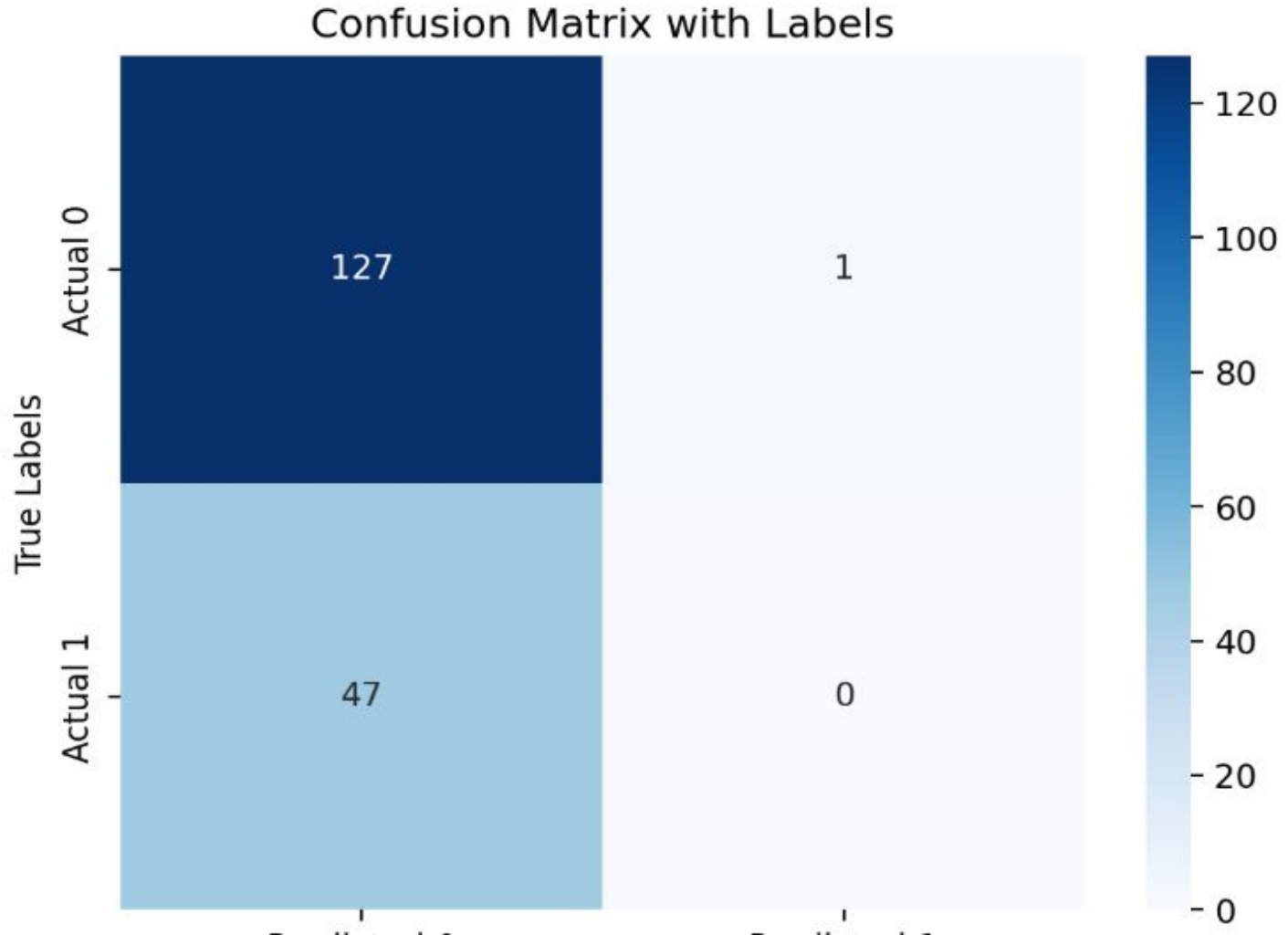
Status: Good Fit: Model generalizes well to new data.

Precision = TP/(TP + FP)

Recall = TP/(TP + FN)

Classification Report (Test Data)

	precision	recall	f1-score	support
0	0.7299	0.9922	0.8411	128.0000
1	0.0000	0.0000	0.0000	47.0000
accuracy	0.7257	0.7257	0.7257	0.7257
macro avg	0.3649	0.4961	0.4205	175.0000
weighted avg	0.5339	0.7257	0.6152	175.0000


Confusion Matrix with Labels
True Labels
Actual 0
Actual 1
127
1
47
0
120
100
80
60
40
20
0

Confusion Matrix (Test Data)

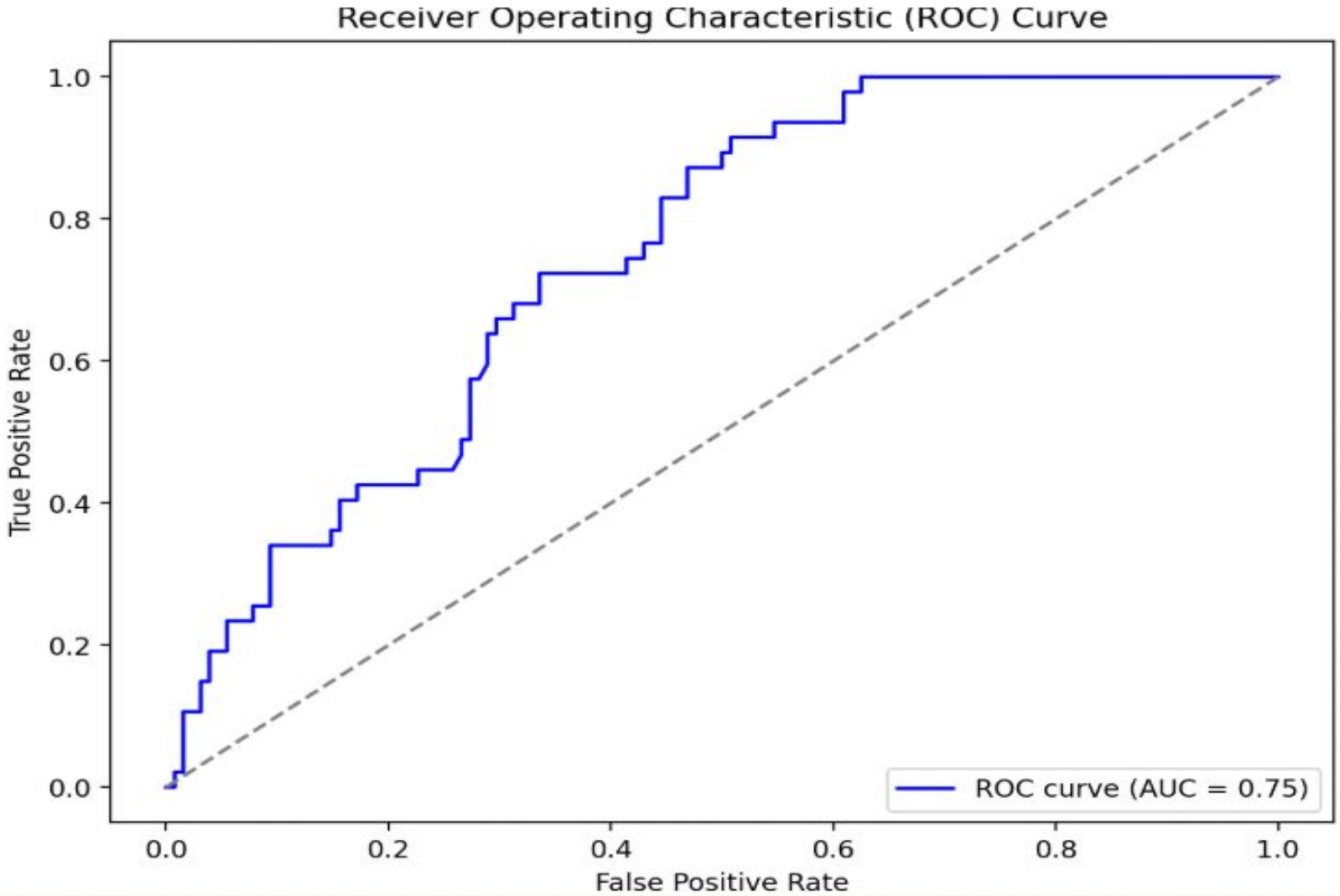
True Positive (TP): 0 - Predicted 1 and Actual 1

True Negative (TN): 127 - Predicted 0 and Actual 0

False Positive (FP): 1 - Predicted 1 but Actual 0

False Negative (FN): 47 - Predicted 0 but Actual 1

True Positive: The test correctly predicts a positive outcome when the actual outcome is positive.
True Negative: The test correctly predicts a negative outcome when the actual outcome is negative.
False Positive: The test incorrectly predicts a positive outcome when the actual outcome is negative.
False Negative: The test incorrectly predicts a negative outcome when the actual outcome is positive.


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.75)

Can Fix Threshold Option

- **Diagonal line (Random Classifier)**: The line represents the performance of a random classifier. A good classifier should have its ROC curve above this line.

AUC (Area Under the Curve): AUC quantifies the overall ability of the model to discriminate between positive and negative cases.

- An AUC of 0.5 means the model is no better than random guessing.
- An AUC of 1.0 indicates a perfect model.

Adjust Threshold for Classification

0.50

0.00 1.00

Adjusted Metrics at Threshold 0.5:

Accuracy: 0.7257

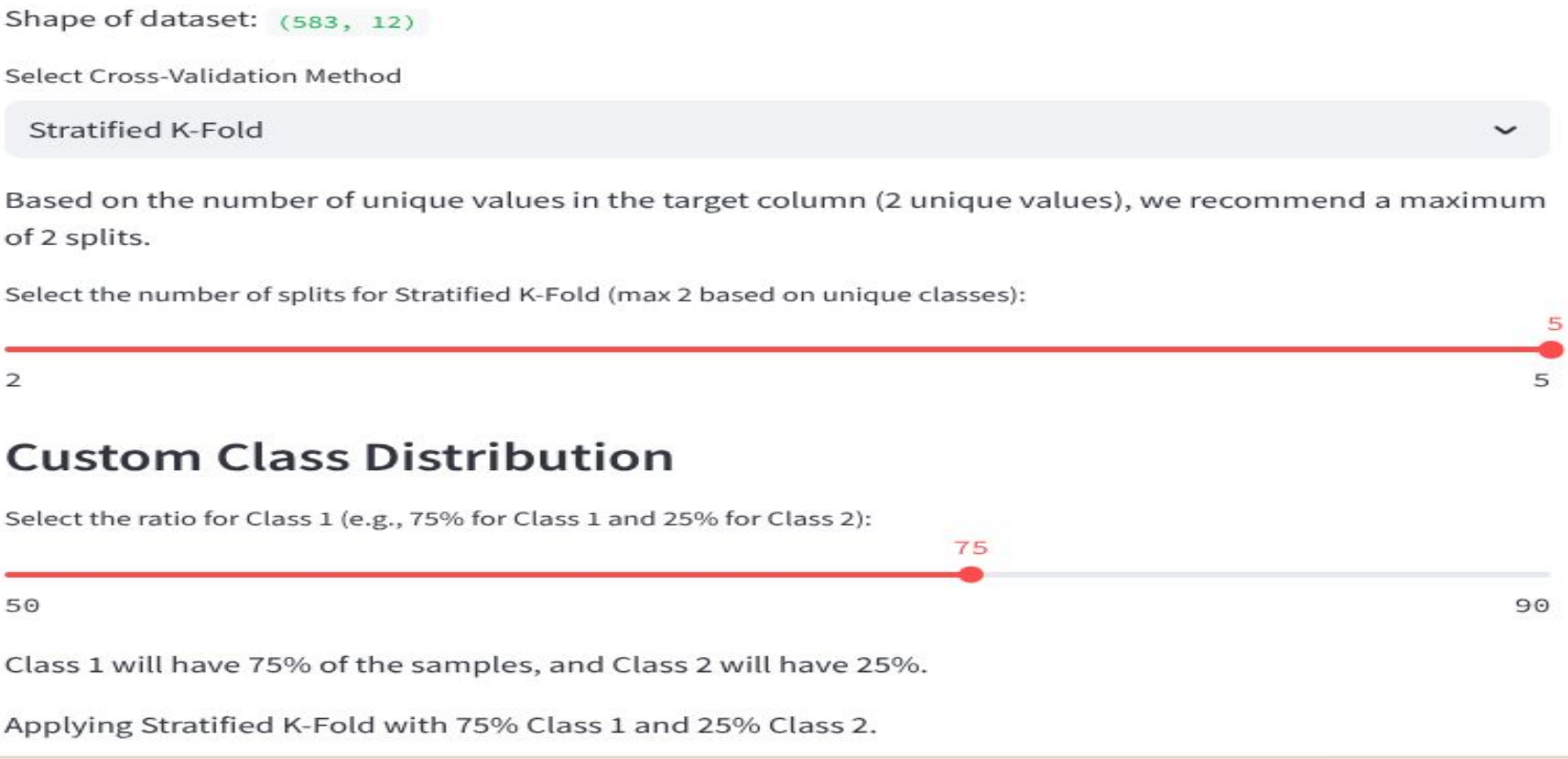
Precision: 0.3649

Recall: 0.4961

F1 Score: 0.4205

Cross Validation for Classification (Stratified K-Fold)

Reduce Bias - Hyperparameter Option Select value for unique classes



To get more Accuracy and prevent Overfit or Underfit

Class 1 will have 75% of the samples, and Class 2 will have 25%.

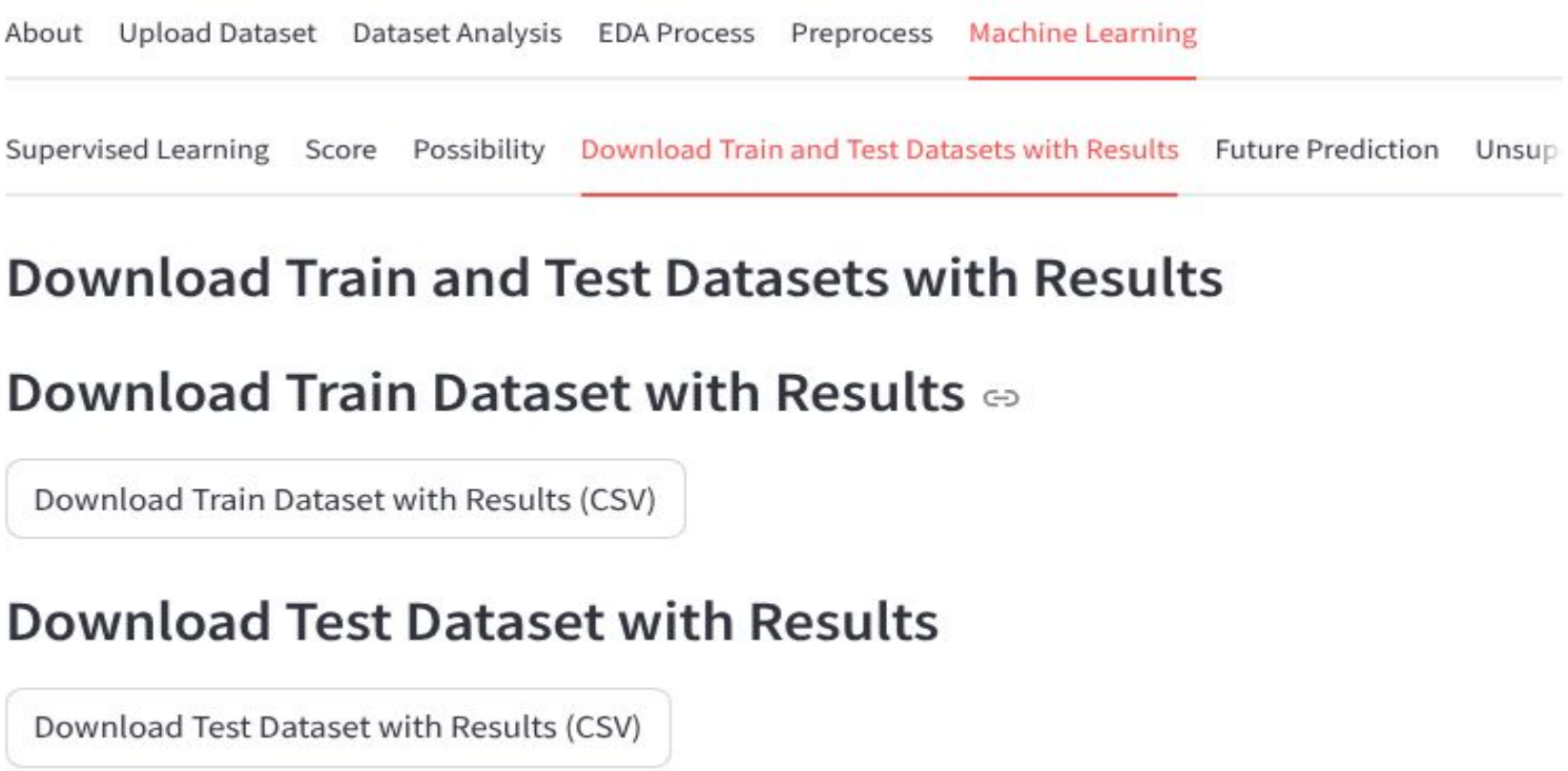
Applying Stratified K-Fold with 75% Class 1 and 25% Class 2.

Cross-Validation Accuracy (Stratified K-Fold): 0.7221 ± 0.0094

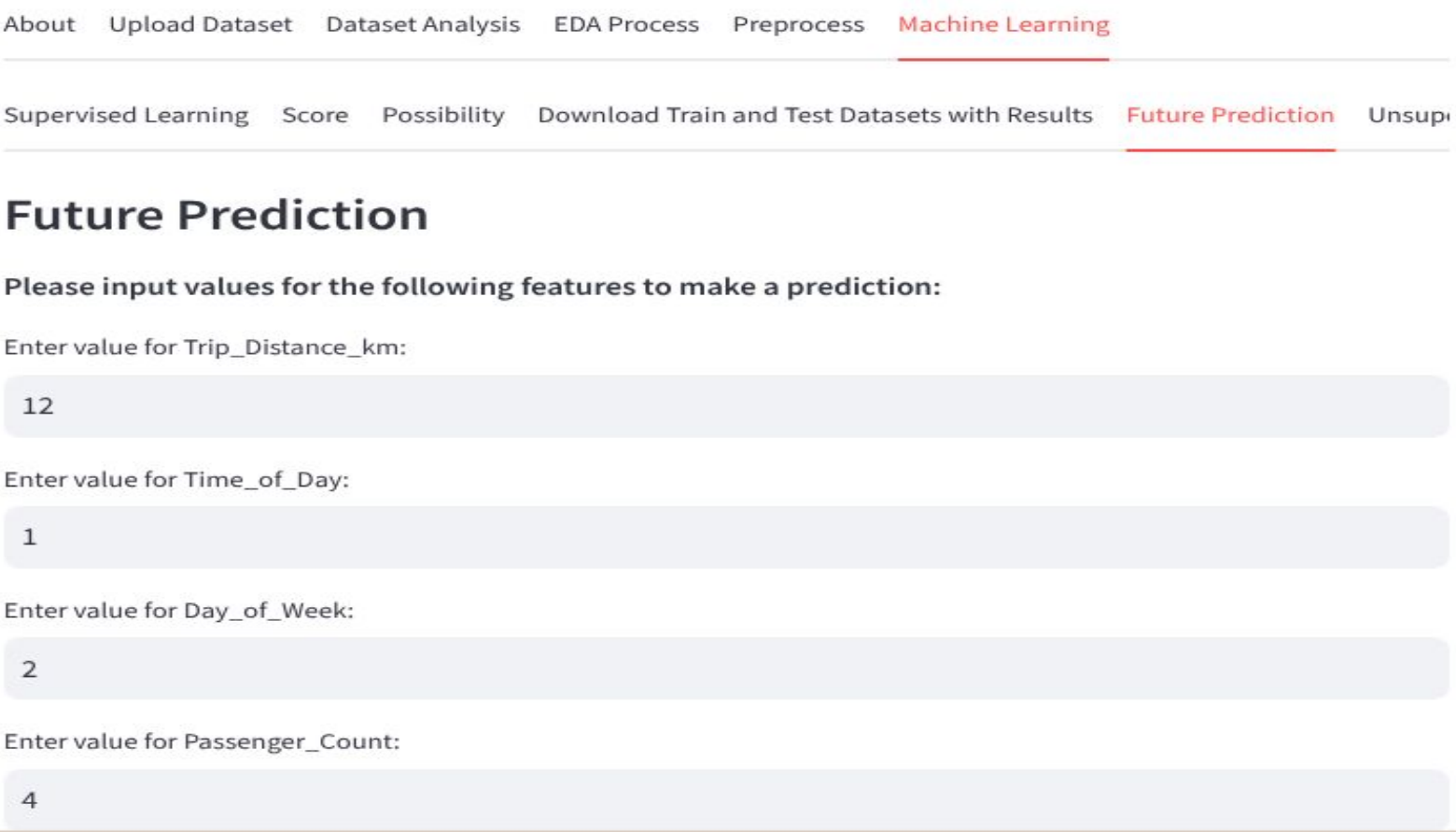
Train Accuracy: 0.7187

Model Fit Status: Generalizing Well

Download Train and Test Dataset



Step 11: Future Prediction



2

Enter value for Passenger_Count:

4

Enter value for Traffic_Conditions:

2

Enter value for Weather:

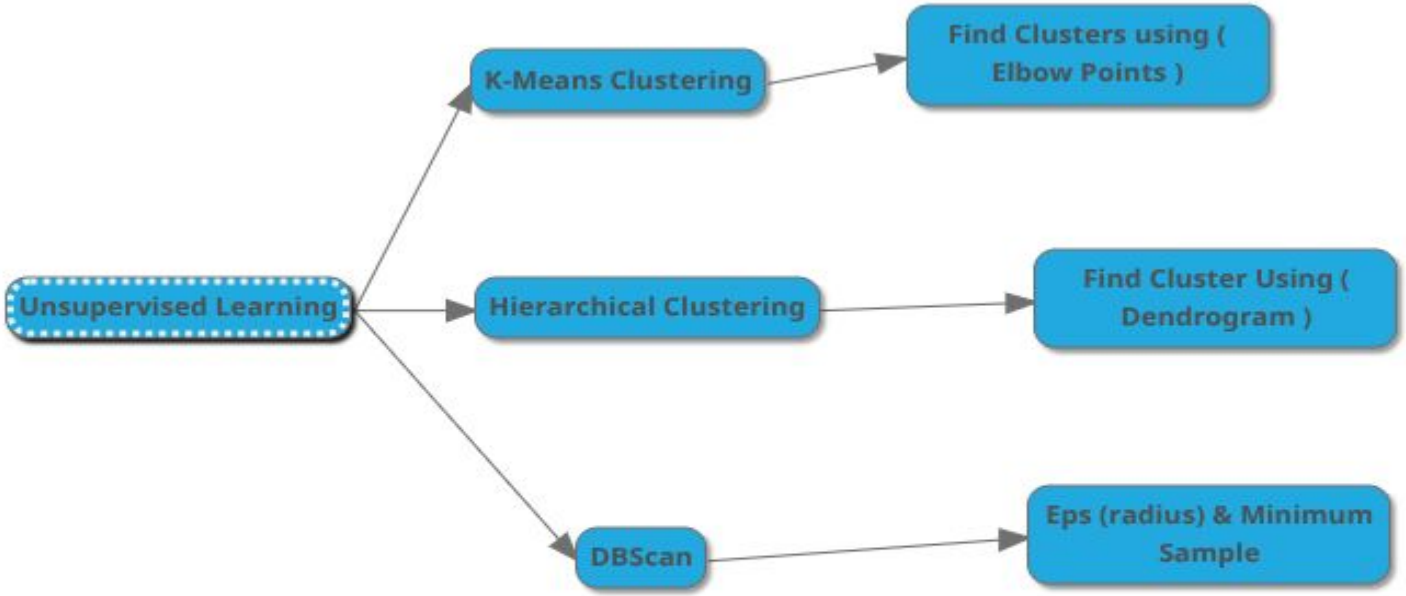
3

Enter value for Base_Fare:

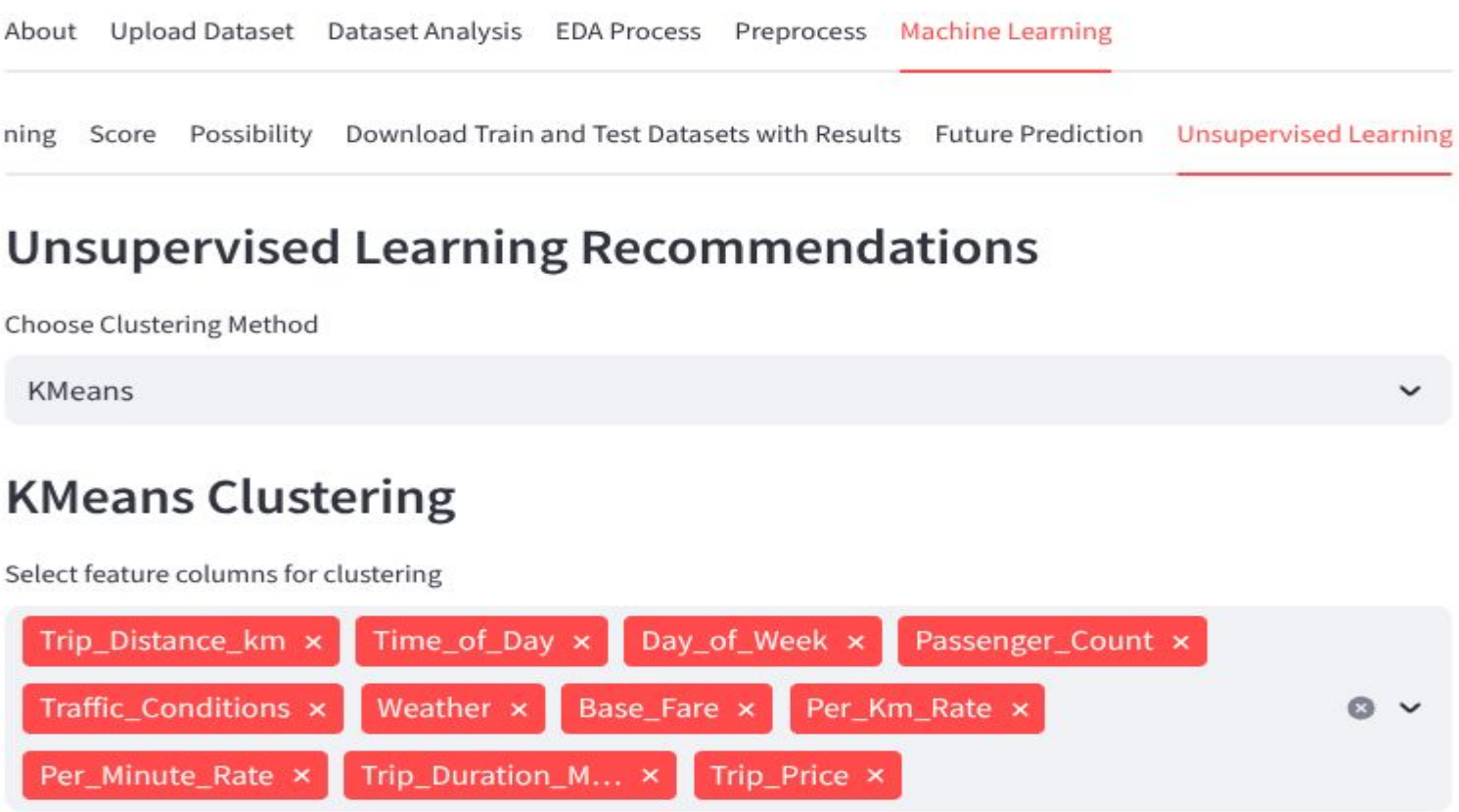
12

Make Prediction

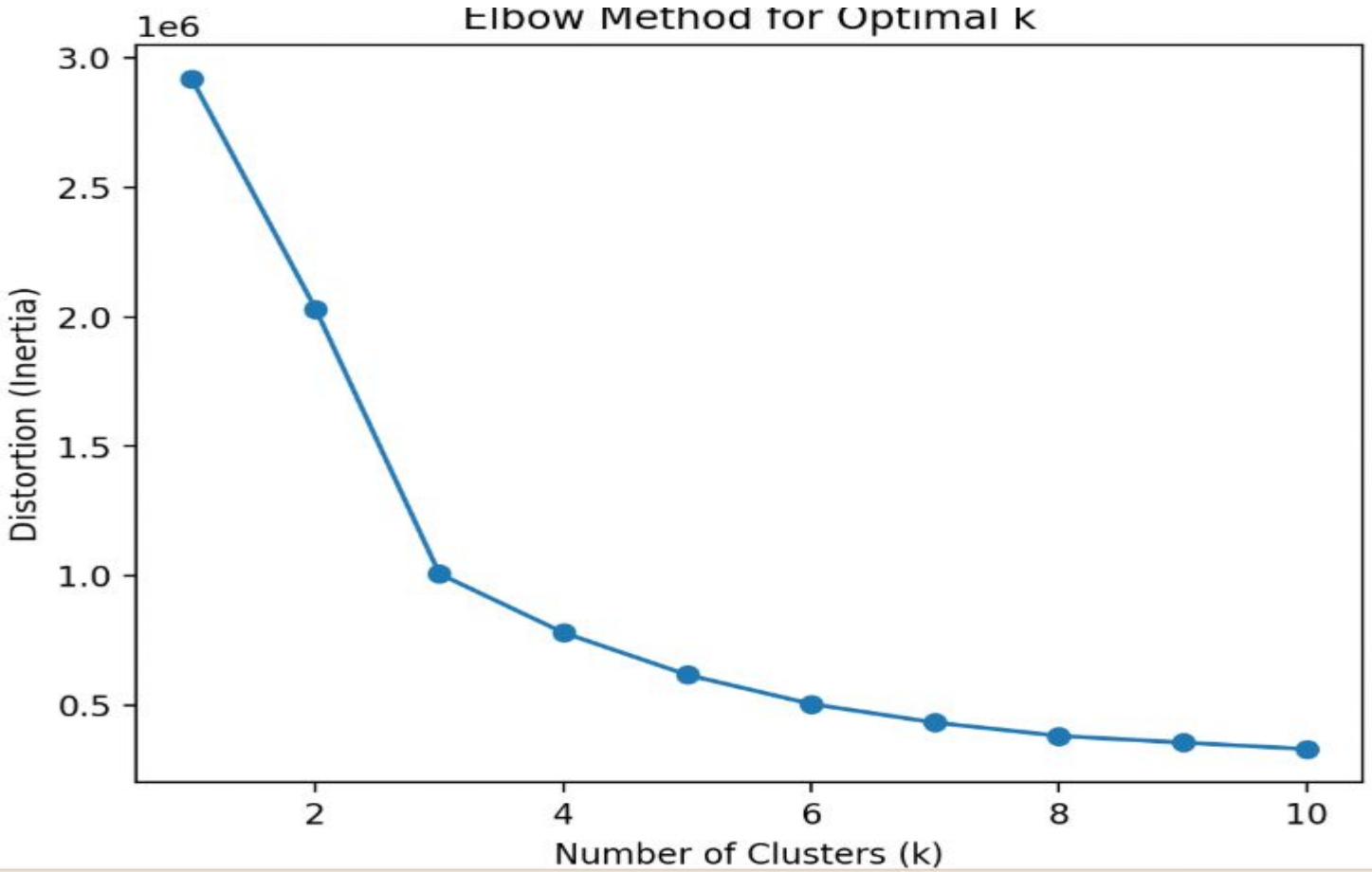
Prediction Result: 36.4206042870544


Unsupervised Learning
K-Means Clustering
Find Clusters using (Elbow Points)
Hierarchical Clustering
Find Cluster Using (Dendrogram)
DBScan
Eps (radius) & Minimum Sample

Step 12 : Unsupervised Learning



K-Means Algorithm (Use Clusters based on Elbow Points)



Enter the optimal k value based on the Elbow method

3 − +

Silhouette Score: 0.38

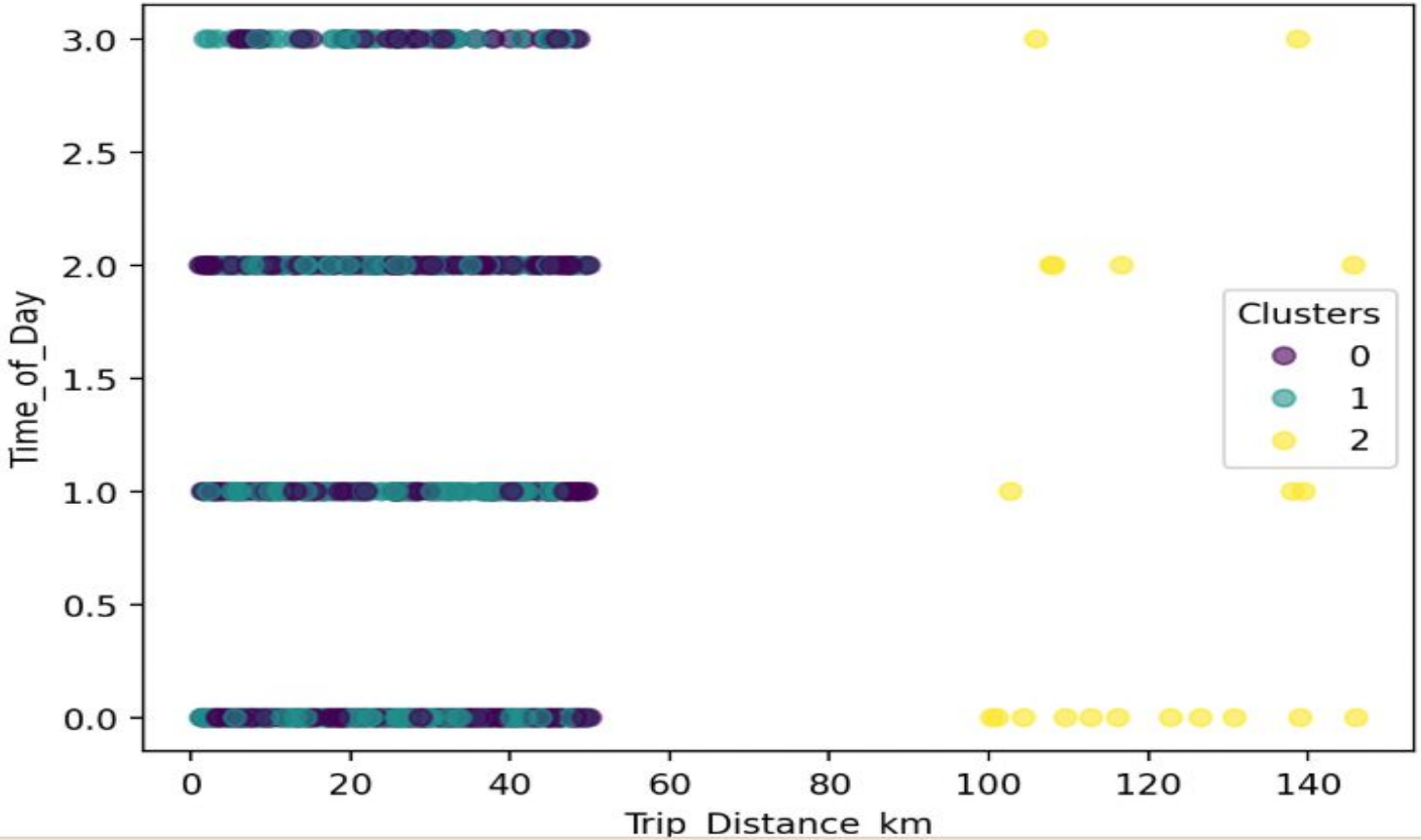
KMeans Clusters

Cluster 0: 467 points

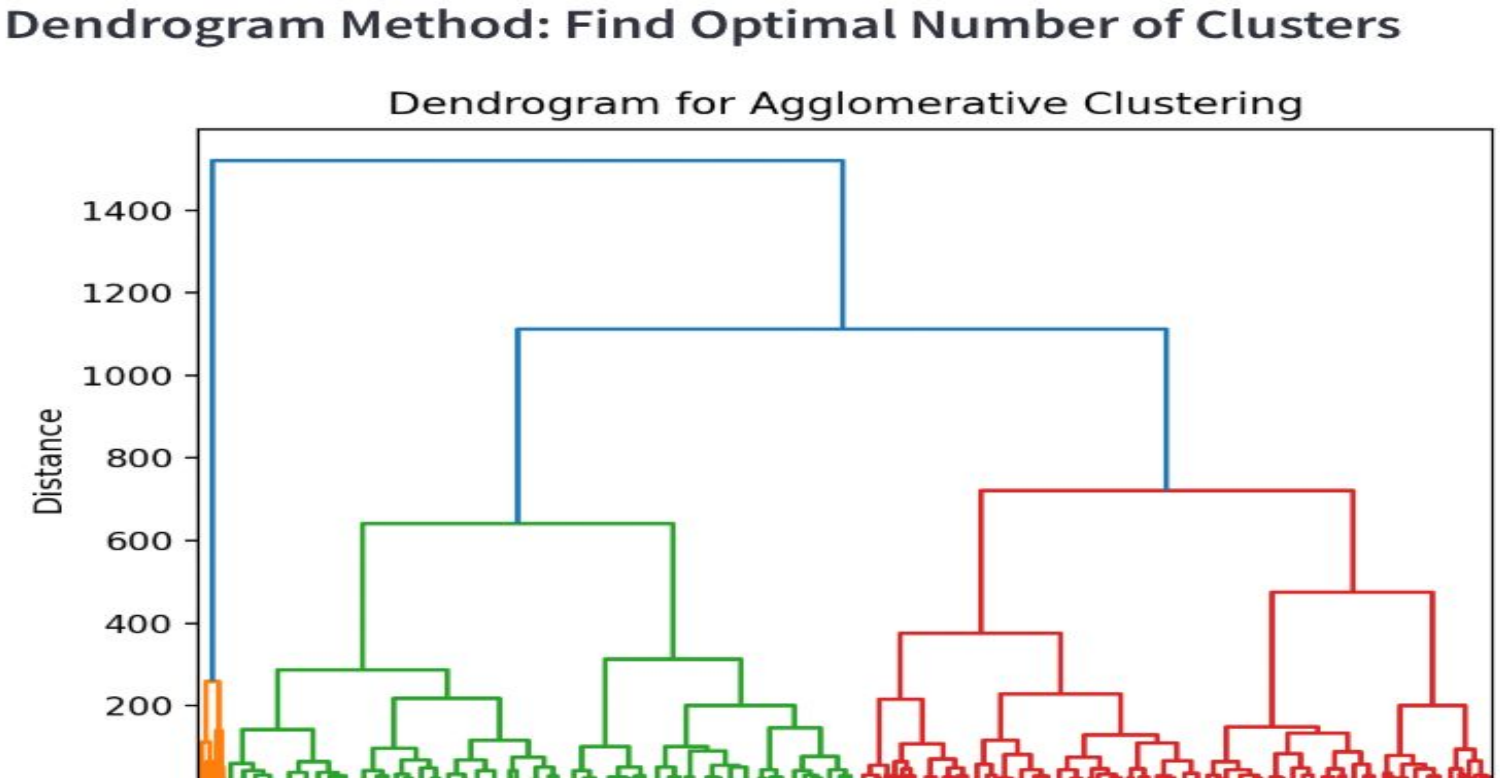
Cluster 1: 513 points

Cluster 2: 20 points

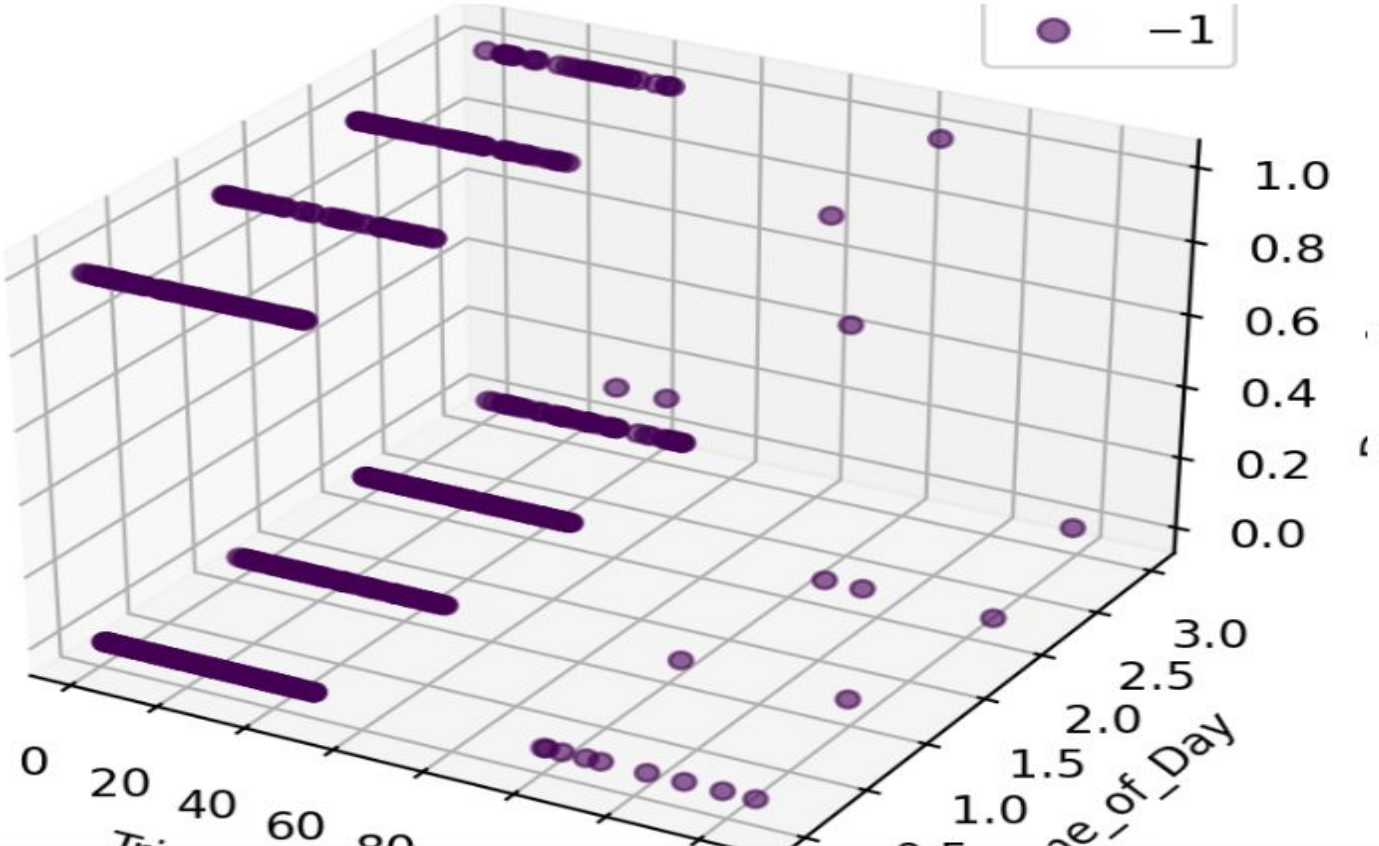
This is not actual Cluster its Just Example to Show



Hierarchical Clustering (Dendrogram)



This is not Actual Clustering Point Chart for only Understanding



Thank You

Developer:

TC Antony - [Data Scientist]