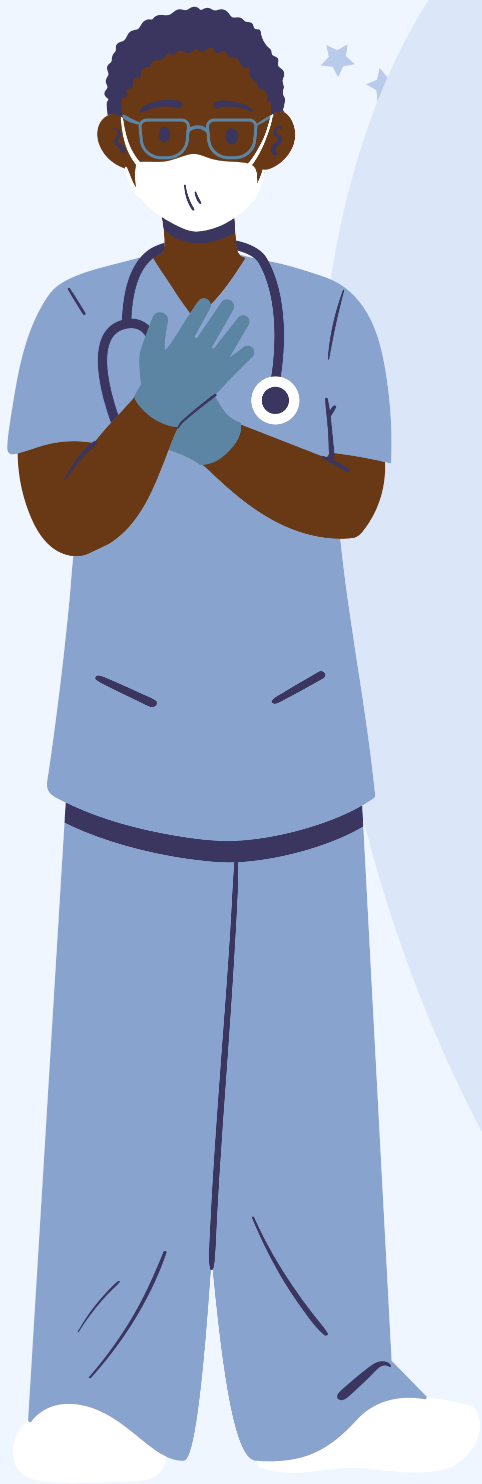
Smart Healthcare Expense & Disease Prediction System

Problem Statement

This project aims to build a Smart Healthcare Prediction System that can predict diseases and estimate healthcare expenses using patient data and improve lifestyle by giving suggestions.

Data Overview



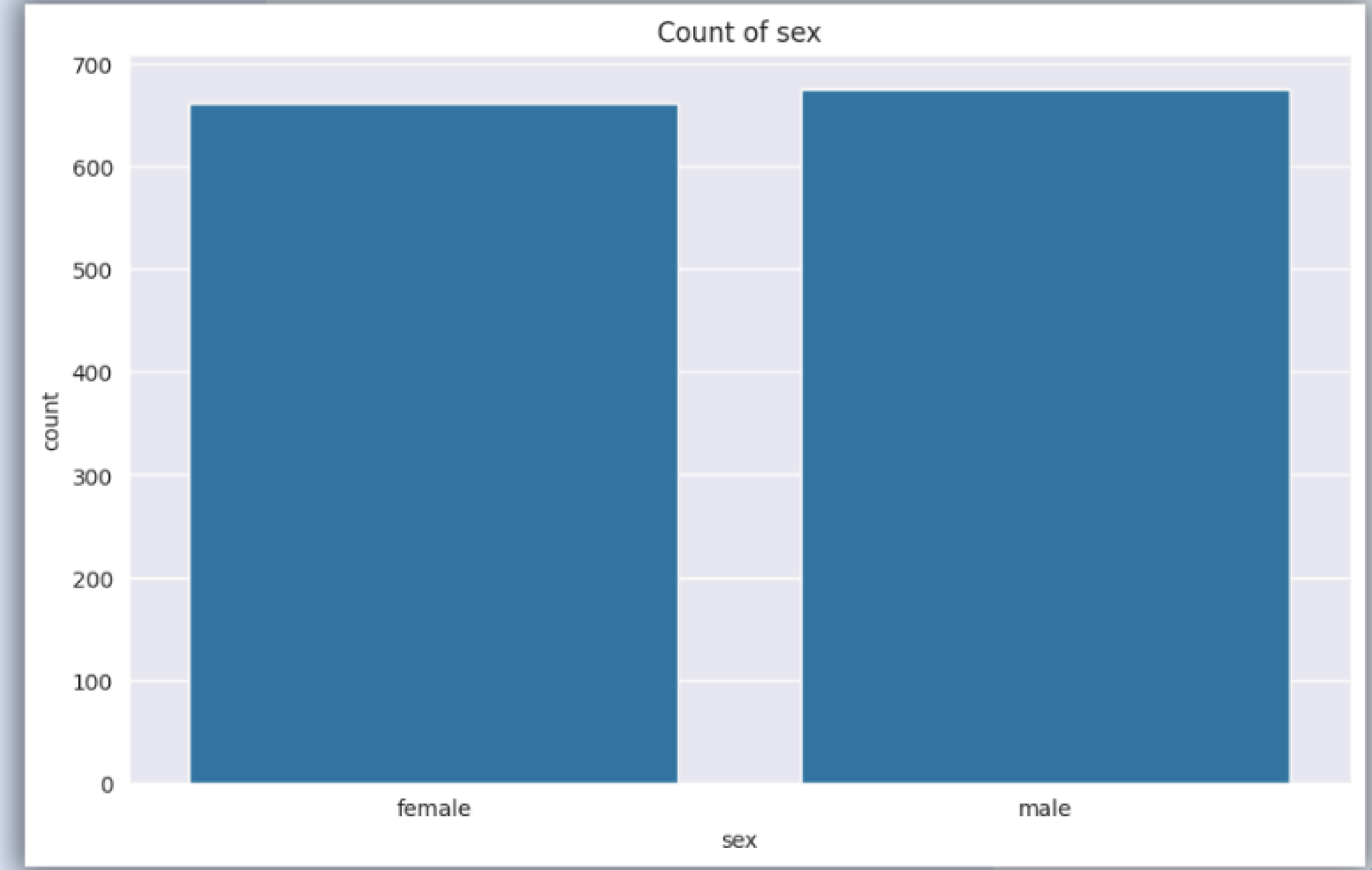
Dataset Shape: (1338, 7)
Rows: Number of patient records – 1338.

Columns:
Age, Sex, BMI, Children, Smoker, Region, Charges.

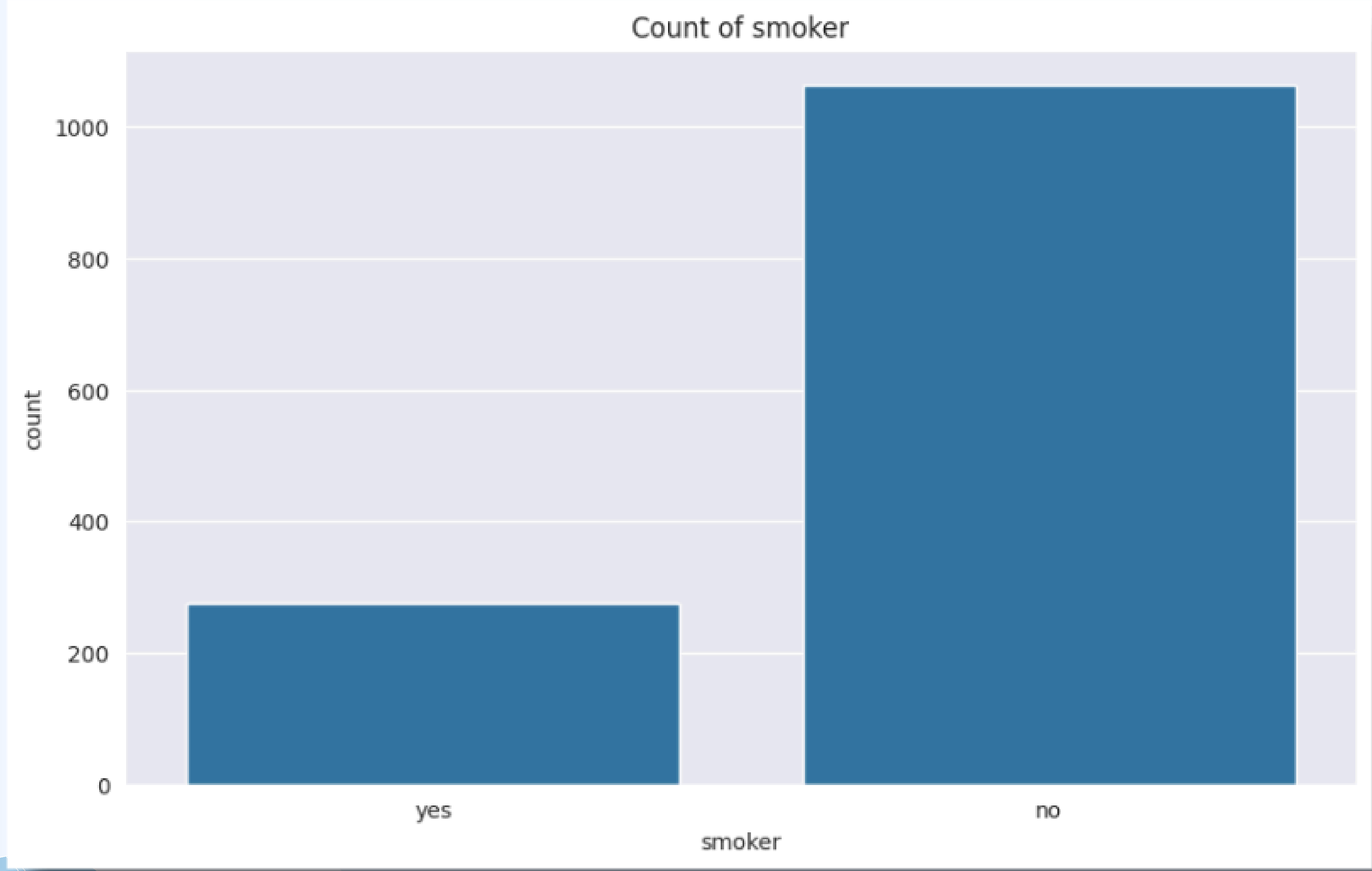
Purpose:
To analyze patient health data for predicting disease likelihood and estimating medical expenses, helping in early diagnosis and effective financial planning.

EDA Visuals

Countplot- Gender Distribution

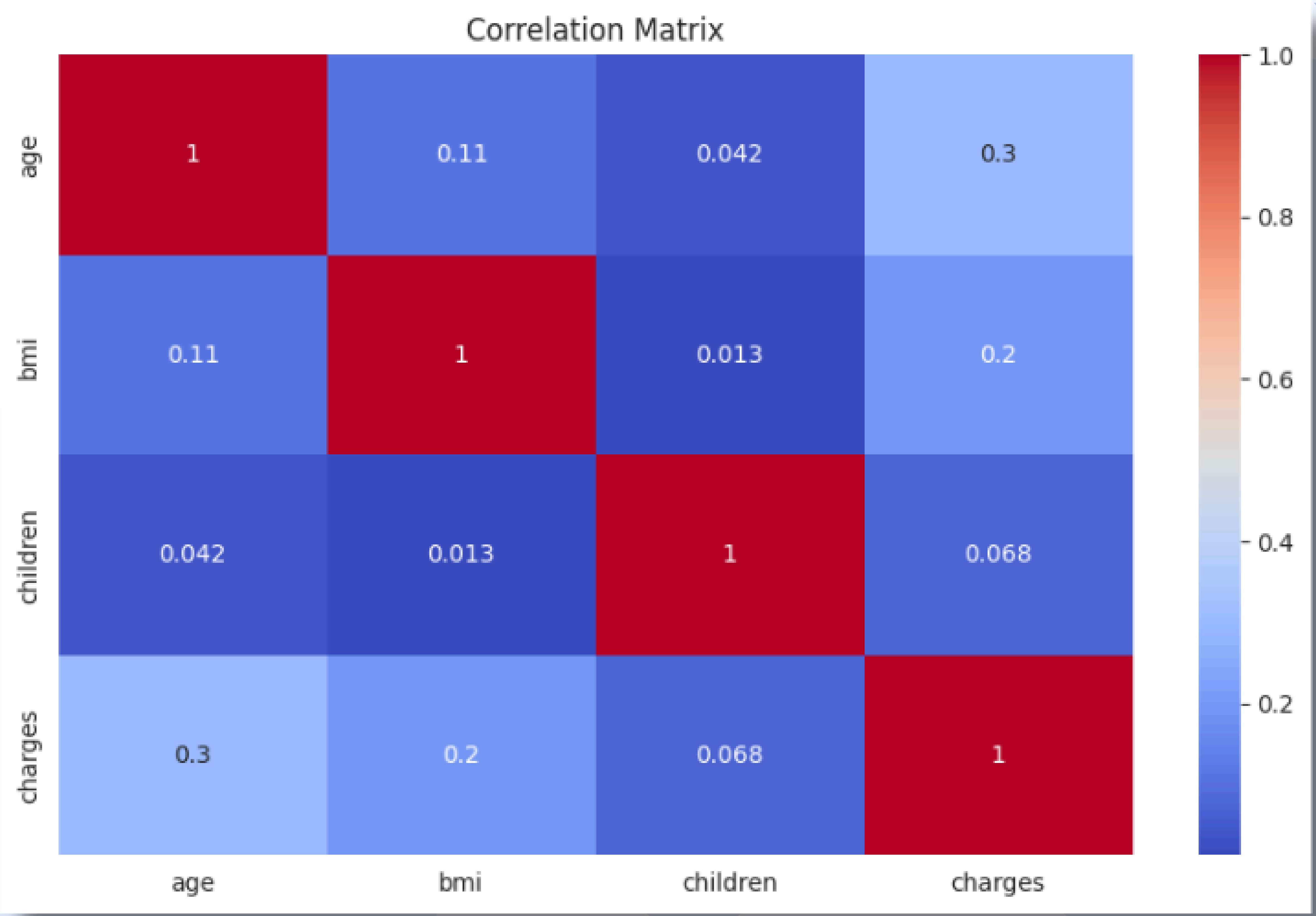


Countplot- Smokers Distribution

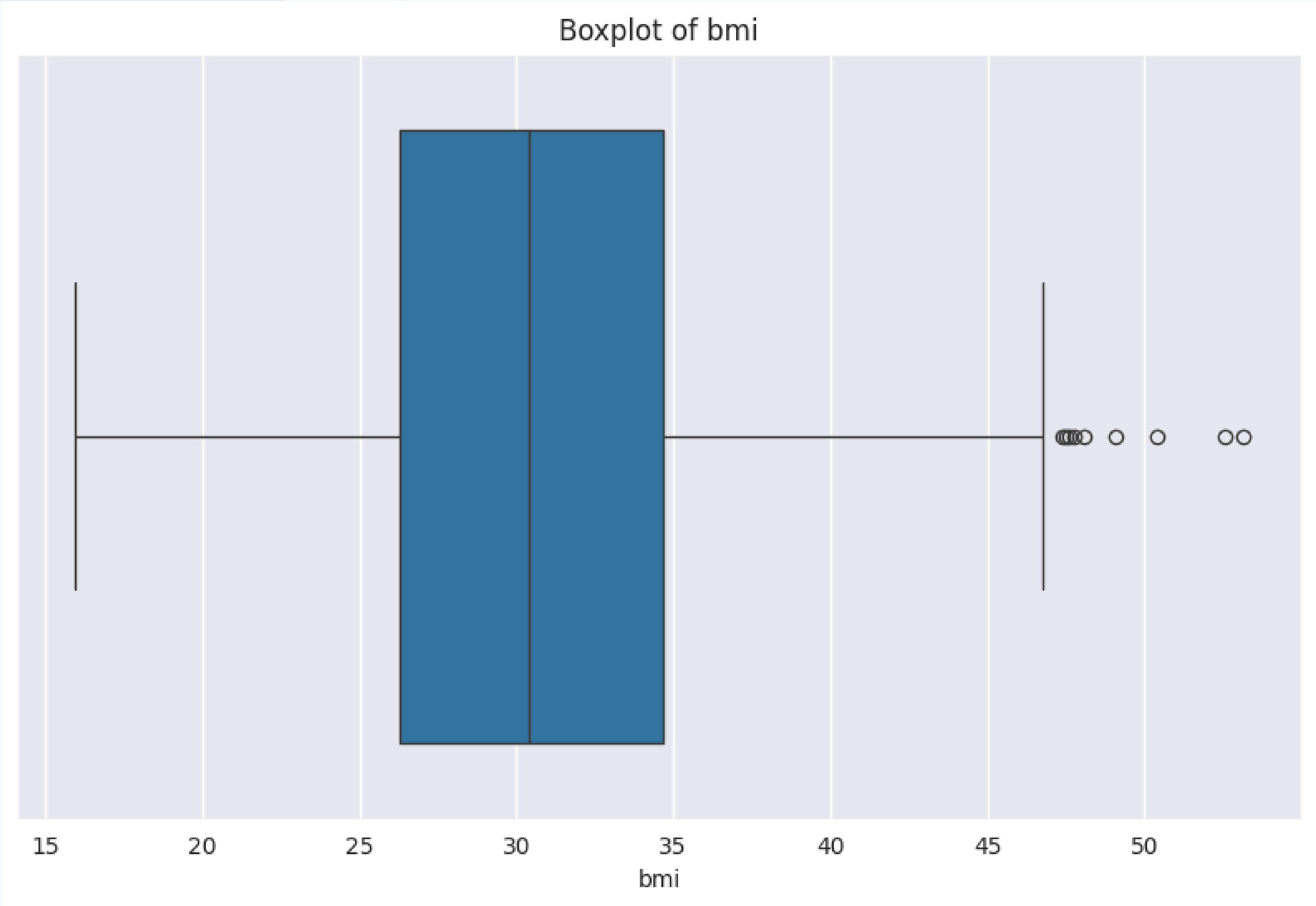


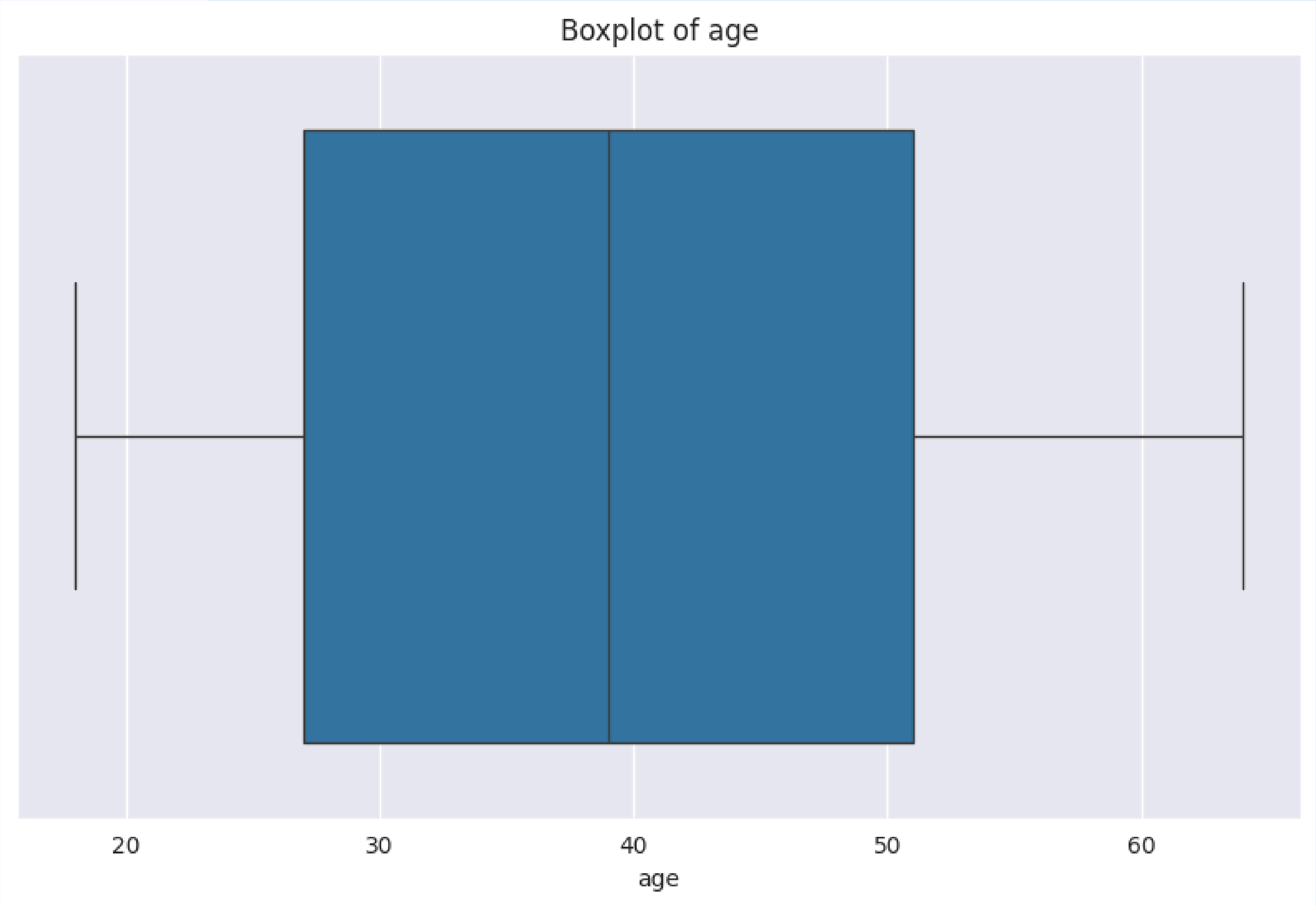
Pairplot : relations and clusters by smoker status

Heatmap : Feature Correlations



Boxplot: Outliers



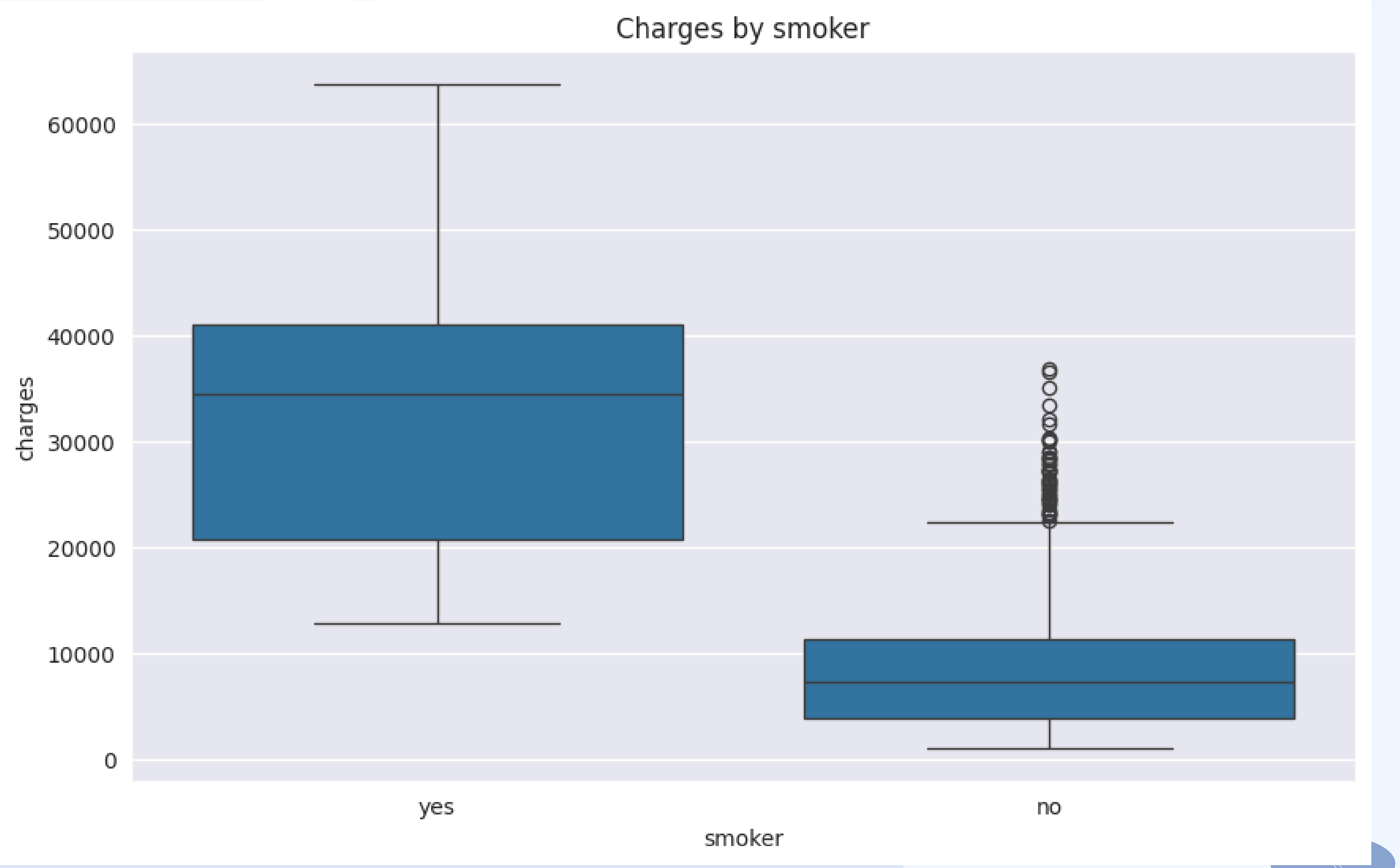


```
Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Statistical Summary:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

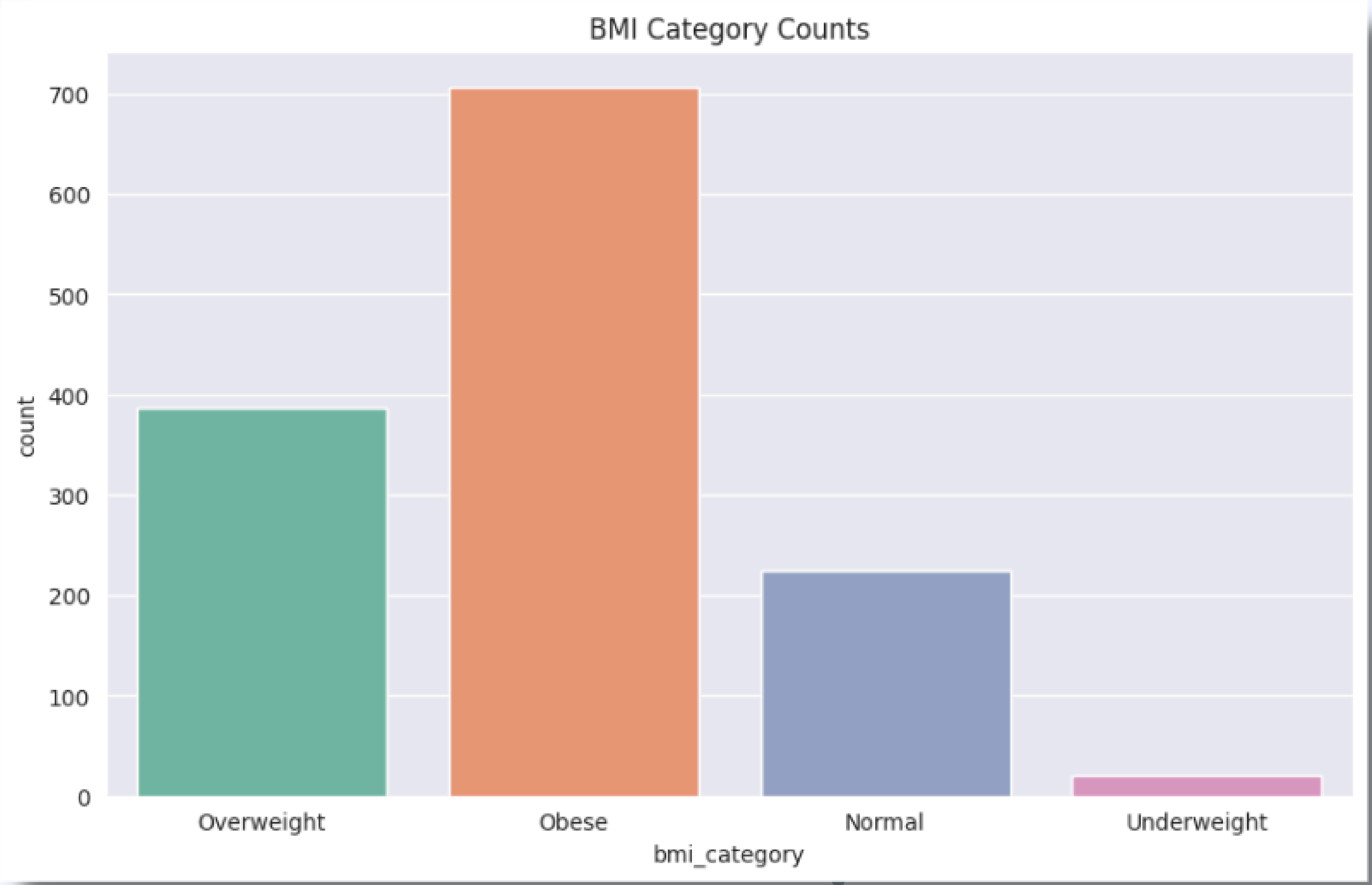
Boxplot: Outliers



Number of outliers in charges: **139**

Feature Engineering

- A new feature bmi_category was created to classify patients based on their Body Mass Index (BMI).
- BMI values were grouped into four categories – Underweight, Normal, Overweight, and Obese
- Label encoding bmi_category



Models Implemented

Linear Regression

Model Performance:

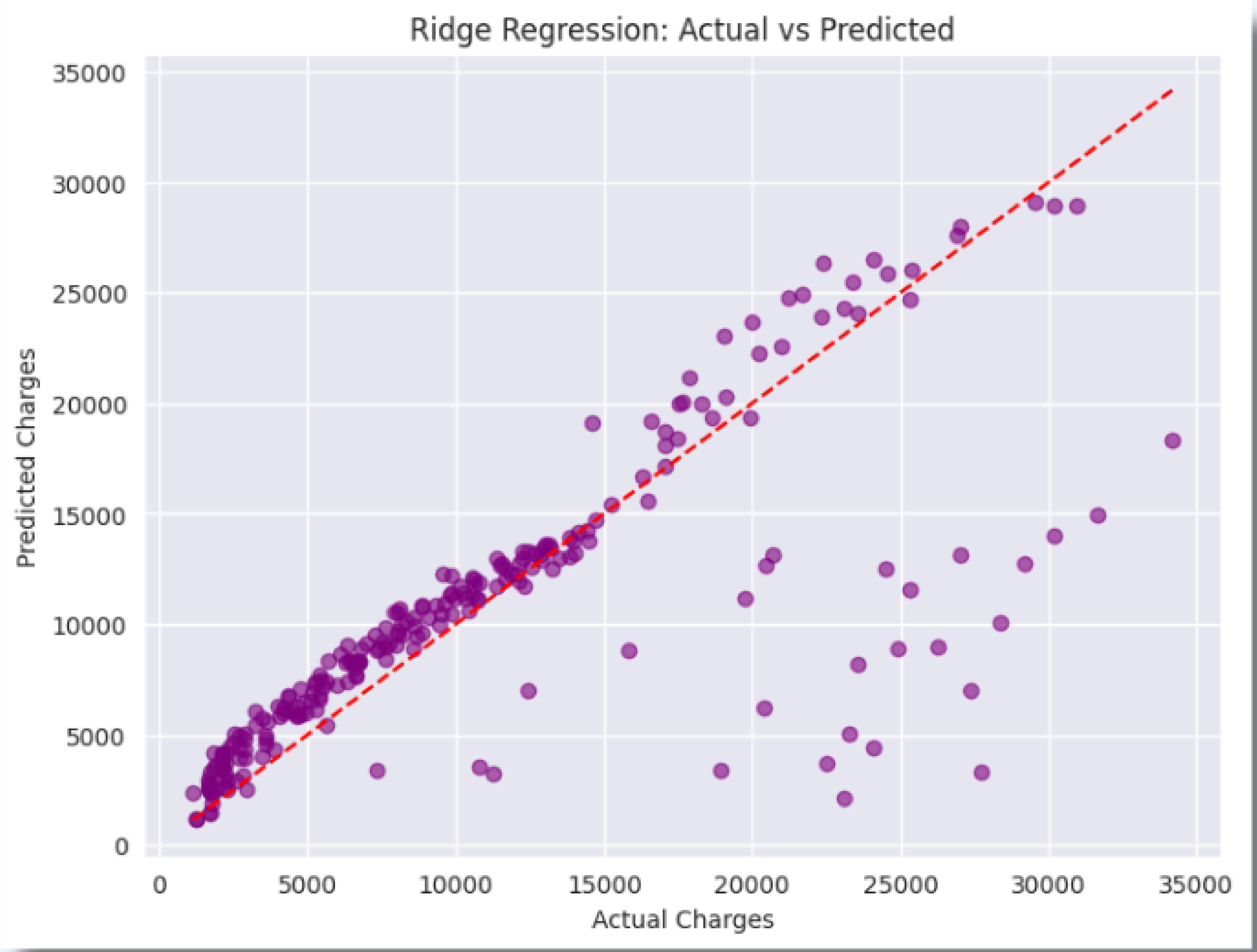
- R^2. Score: 0.5568.
- MSE: 27647351.6858.
- RMSE: 5258.0749.
- MAE: 2797.0320.

Models Implemented

Ridge Regression

Ridge Regression Performance:

- R^2 Score: 0.5579
- MSE: 27575547.8138
- RMSE: 5251.2425
- MAE: 2791.5766

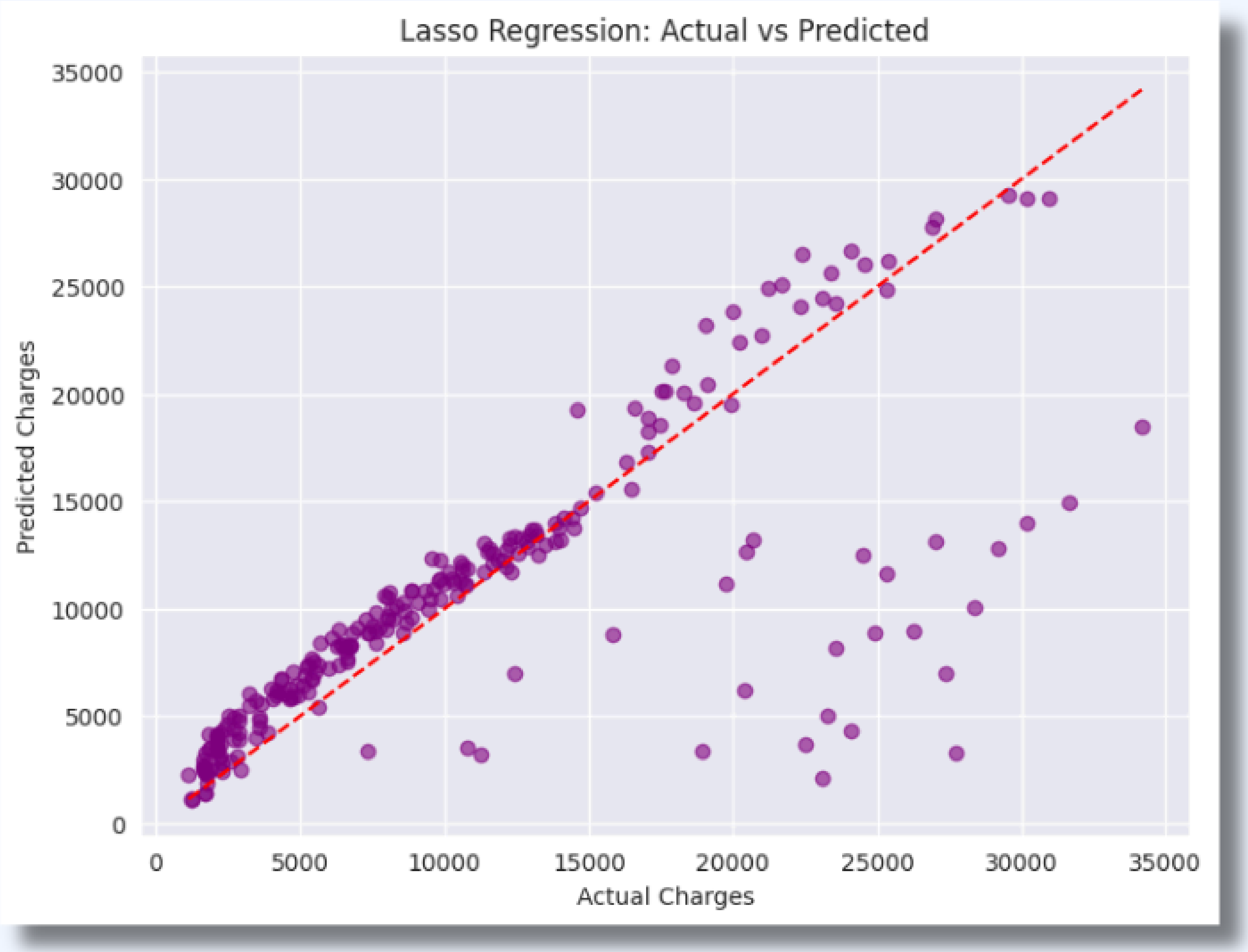


Models Implemented

Lasso Regression

Lasso Regression Performance:

- R^2 Score: 0.5568
- MSE: 27647317.0629
- RMSE: 5258.0716
- MAE: 2797.024

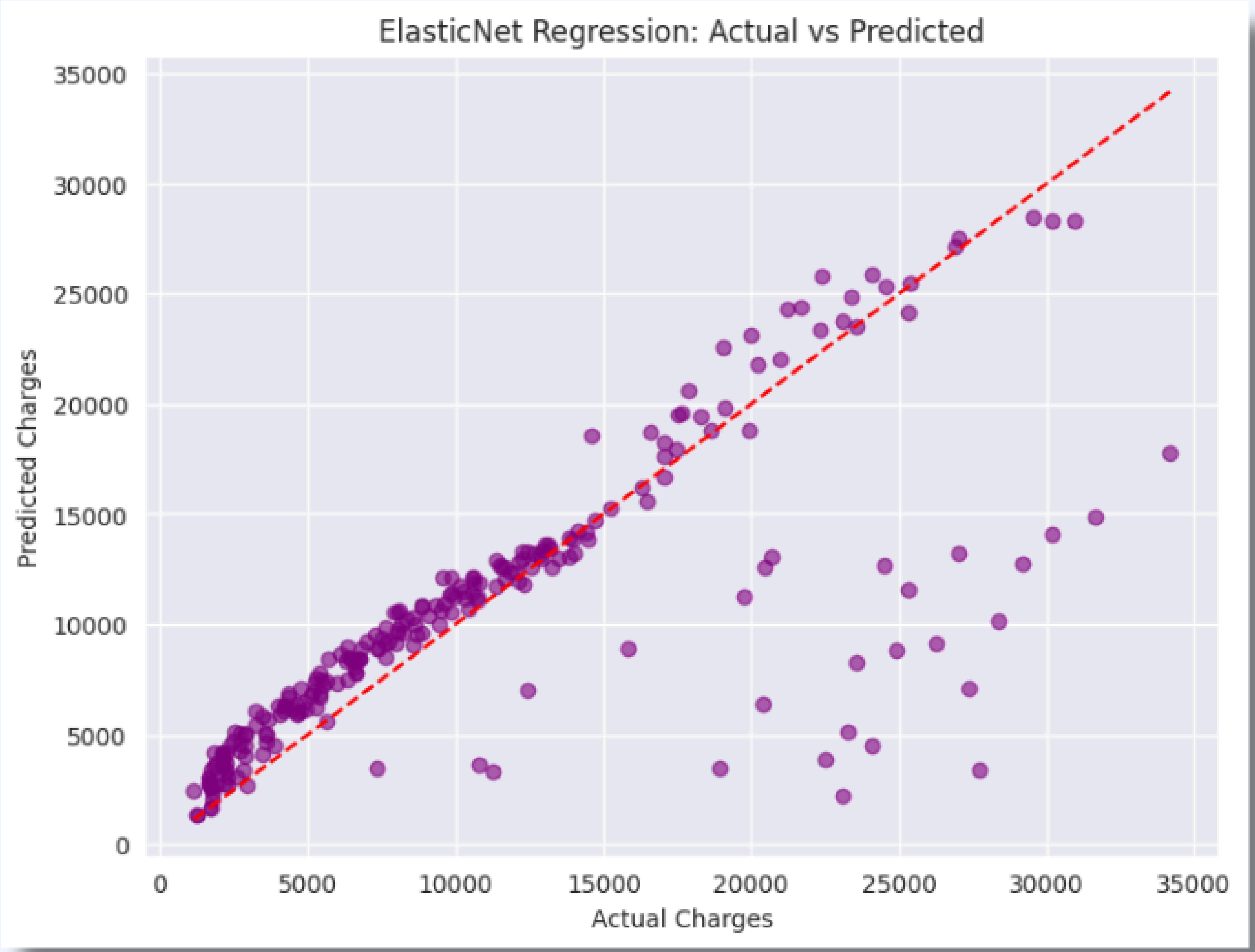


Models Implemented

ElasticNet Regression

ElasticNet Regression Performance:

- R^2 Score: 0.5611
- MSE: 27380338.4948
- RMSE: 5232.6225
- MAE: 2778.8538

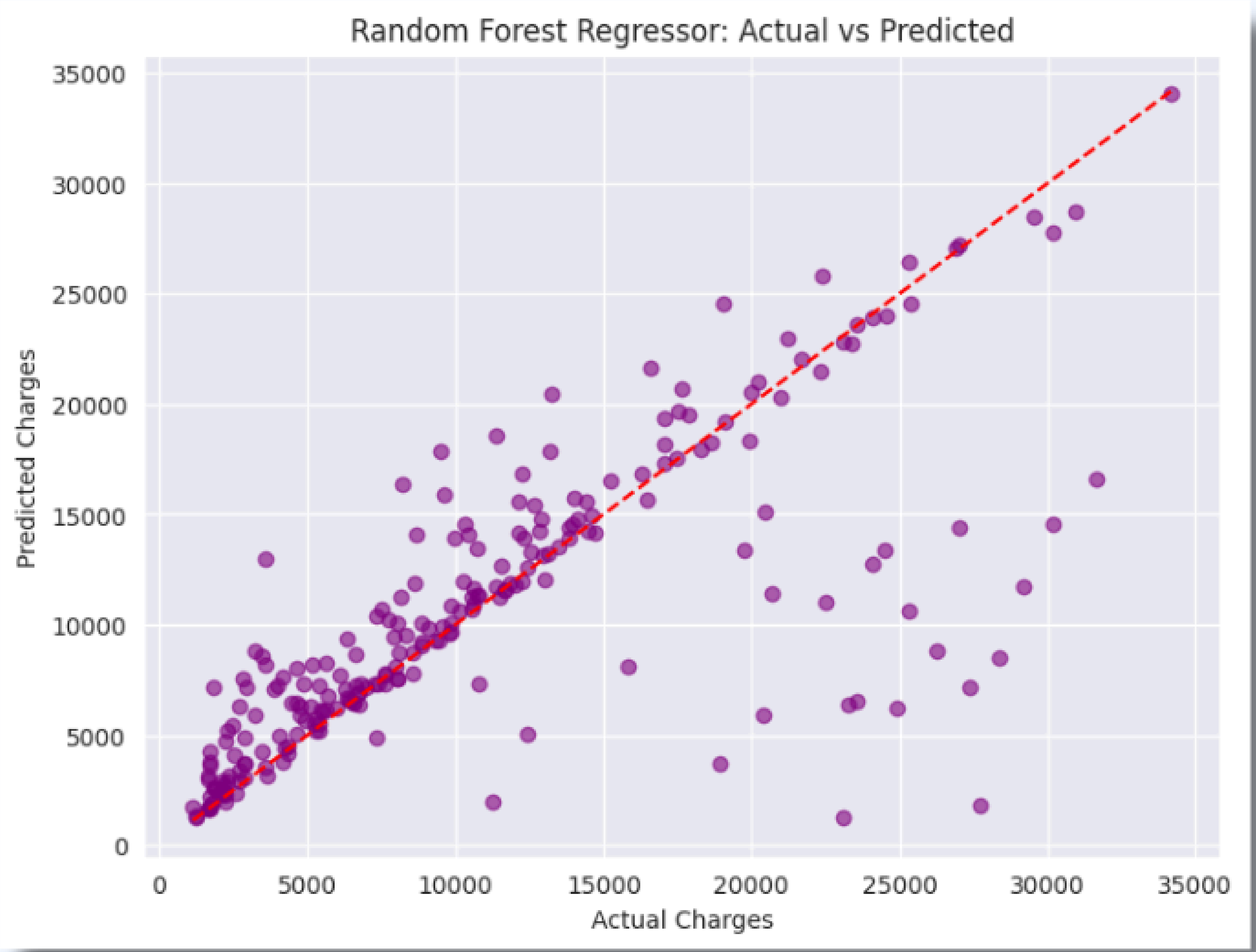


Models Implemented

Random Forest Regression

Random Forest Regressor Performance:

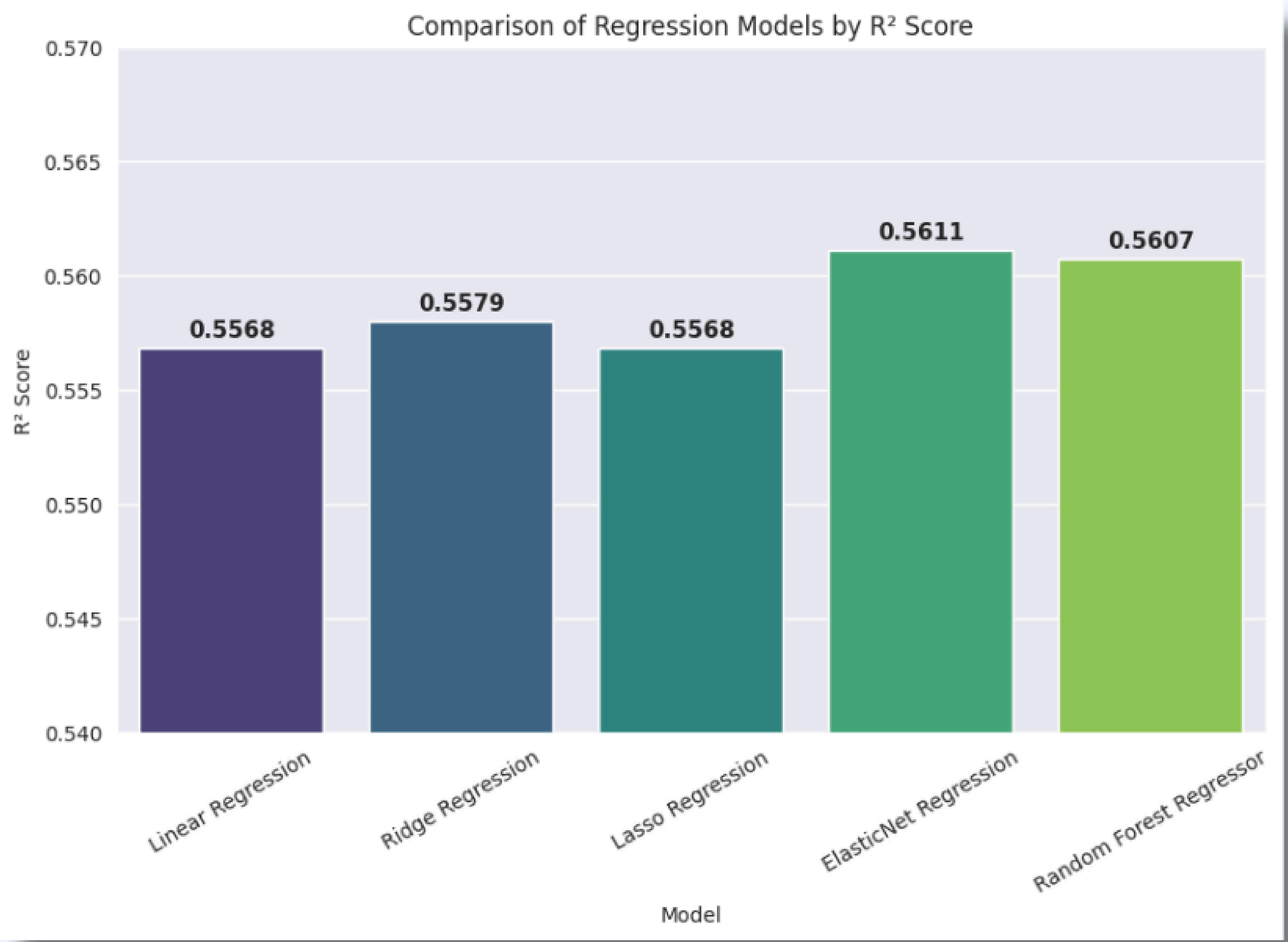
- R^2 Score: 0.5607
- MSE: 27403463.4422
- RMSE: 5234.8317
- MAE: 2689.3716



Model Comparison Table

```
Model Comparison Table:
```

	Model	R2 Score	MSE	RMSE	MAE
3	ElasticNet Regression	0.561074	2.738034e+07	5232.622526	2778.853821
4	Random Forest Regressor	0.560703	2.740346e+07	5234.831749	2689.371620
1	Ridge Regression	0.557945	2.757555e+07	5251.242502	2791.576617
2	Lasso Regression	0.556794	2.764732e+07	5258.071611	2797.024703
0	Linear Regression	0.556794	2.764735e+07	5258.074903	2797.031953


Comparison of Regression Models by R² Score
0.570
0.565
0.560
0.555
0.550
0.545
0.540
R² Score
0.5568
0.5579
0.5568
0.5611
0.5607
Linear Regression
Ridge Regression
Lasso Regression
ElasticNet Regression
Random Forest Regressor
Model

Feature Engineering

- The charges column was divided into three categories — Low Risk, Medium Risk, and High Risk — based on medical expense ranges.

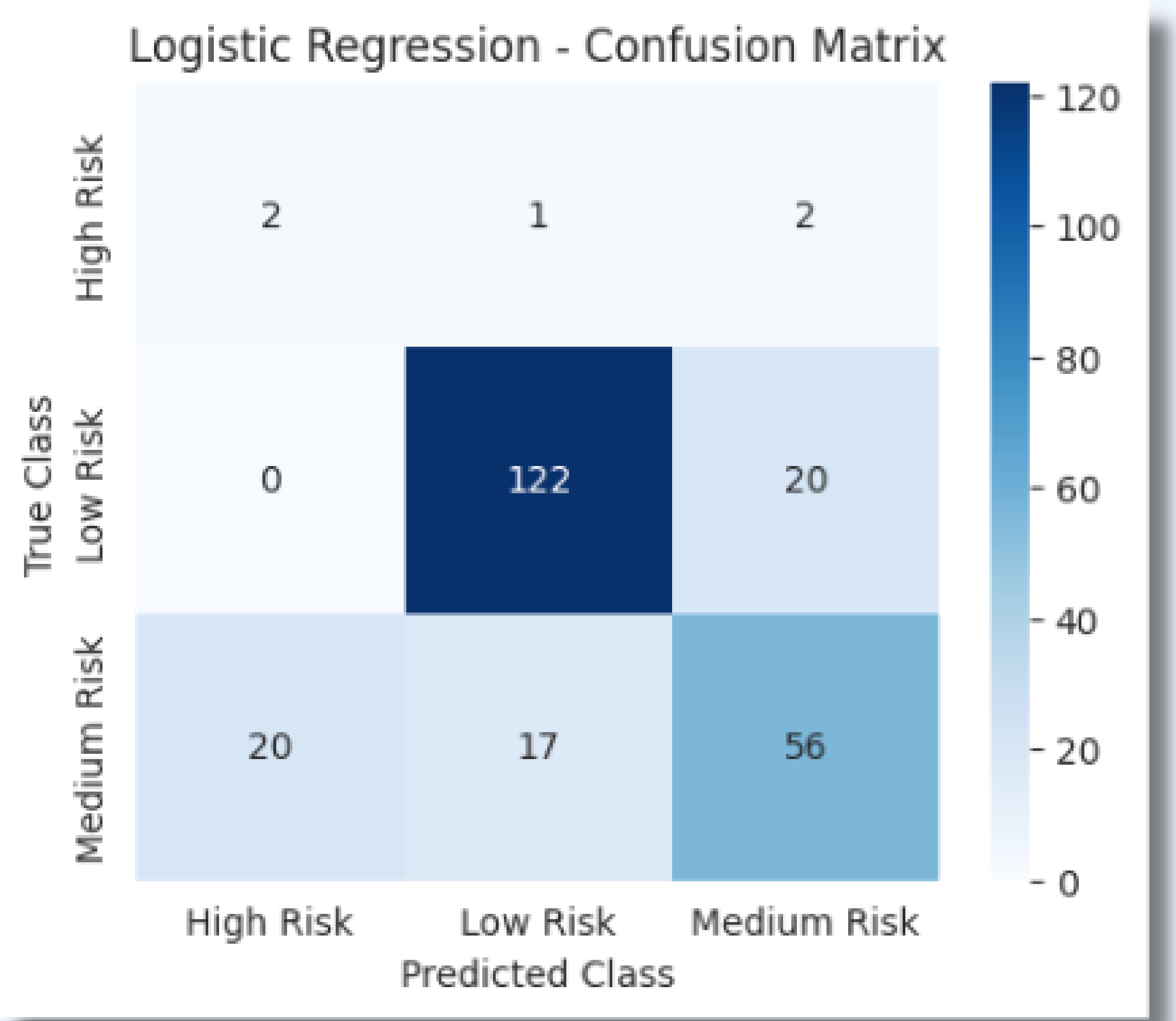
- These categories were label encoded into numbers so the model can process them easily.

```
Class distribution in full data:
risk
Low Risk       712
Medium Risk    464
High Risk       23
Name: count, dtype: int64
Class distribution in test set:
risk_label
1    142
2     93
0      5
Name: count, dtype: int64
```

Models Implemented

Logistic Regression

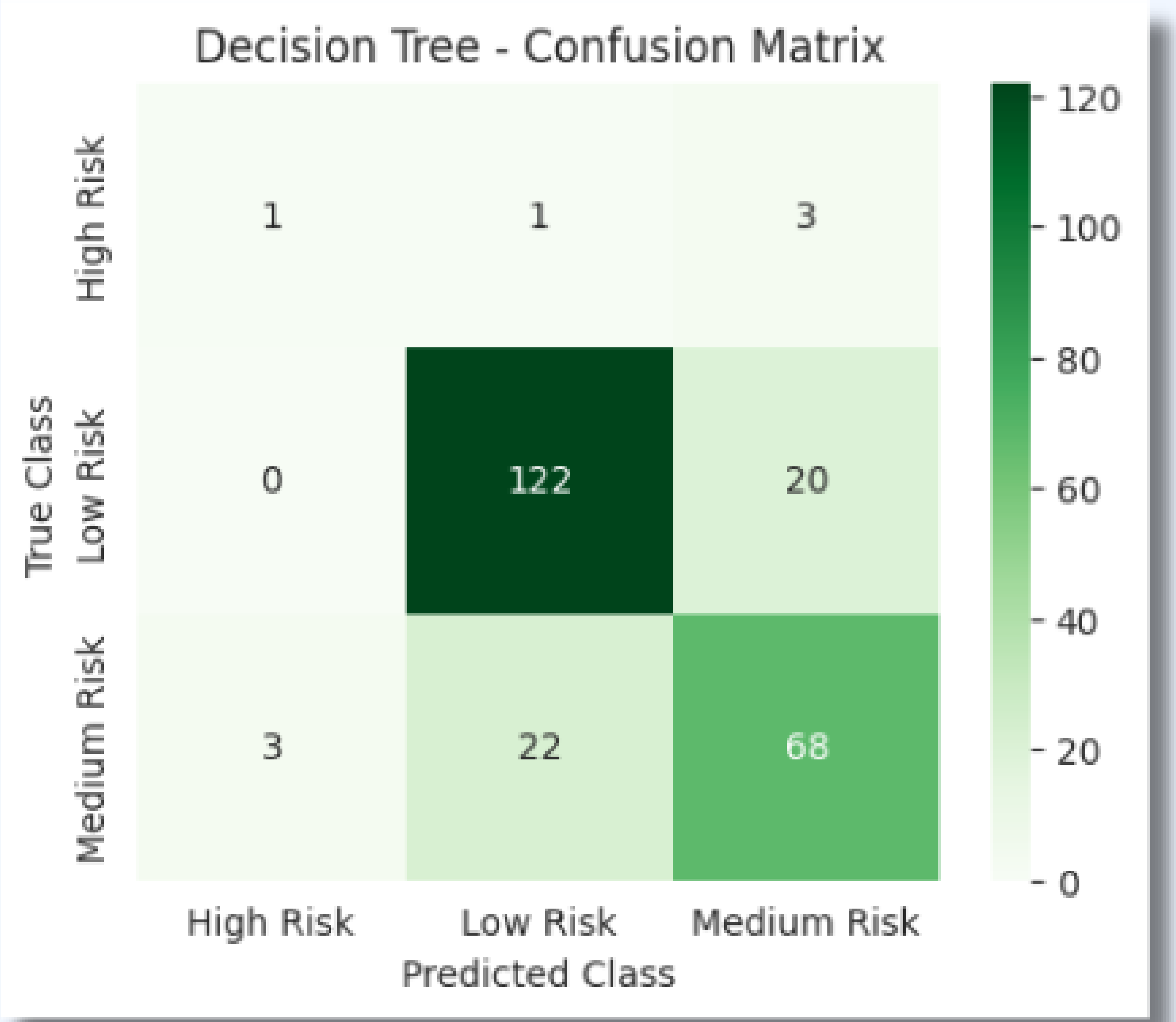
Accuracy: 0.7500,
Precision: 0.7957



Models Implemented

Decision Tree

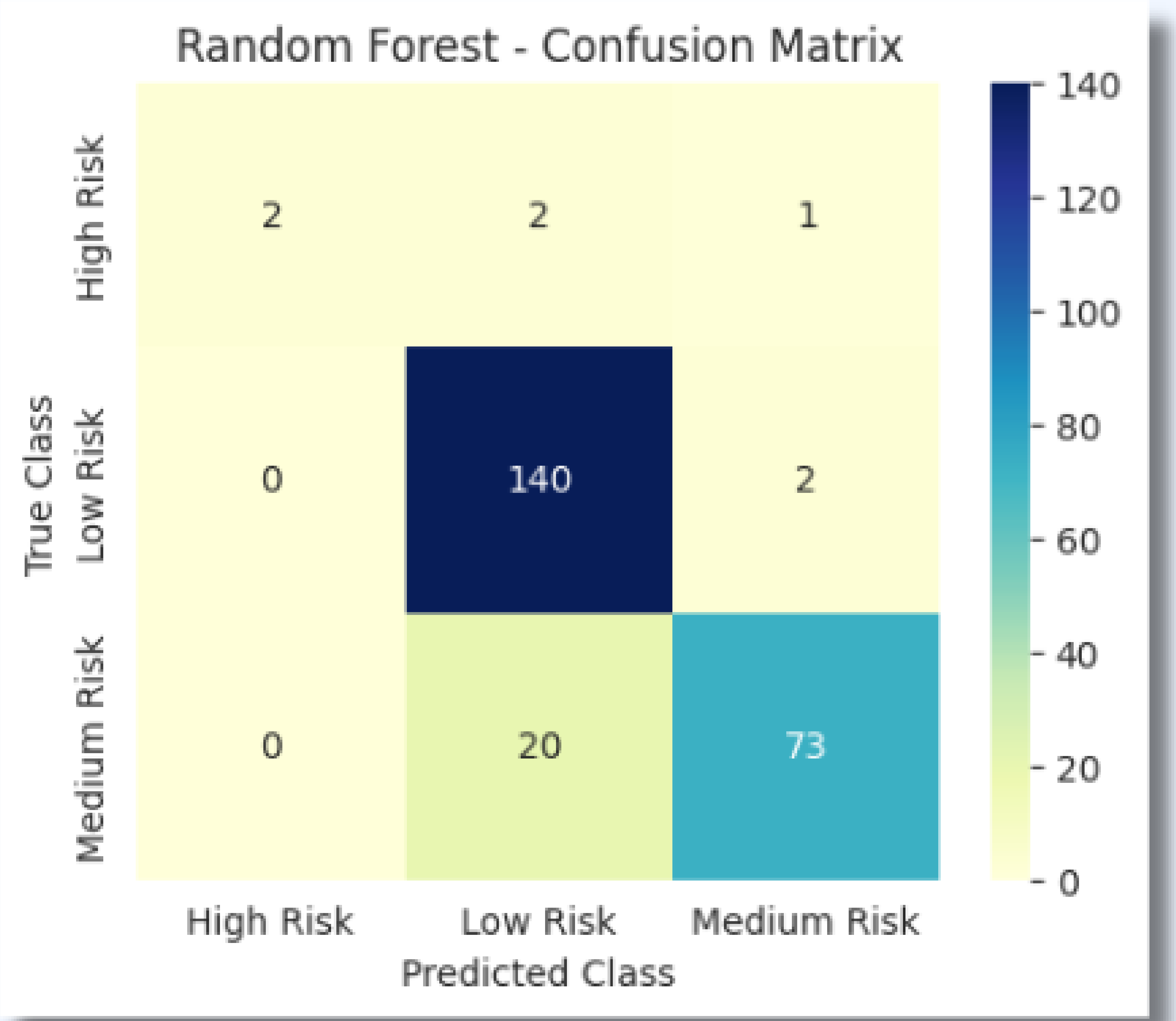
Accuracy: 0.7958,
Precision: 0.7926



Models Implemented

Random Forest

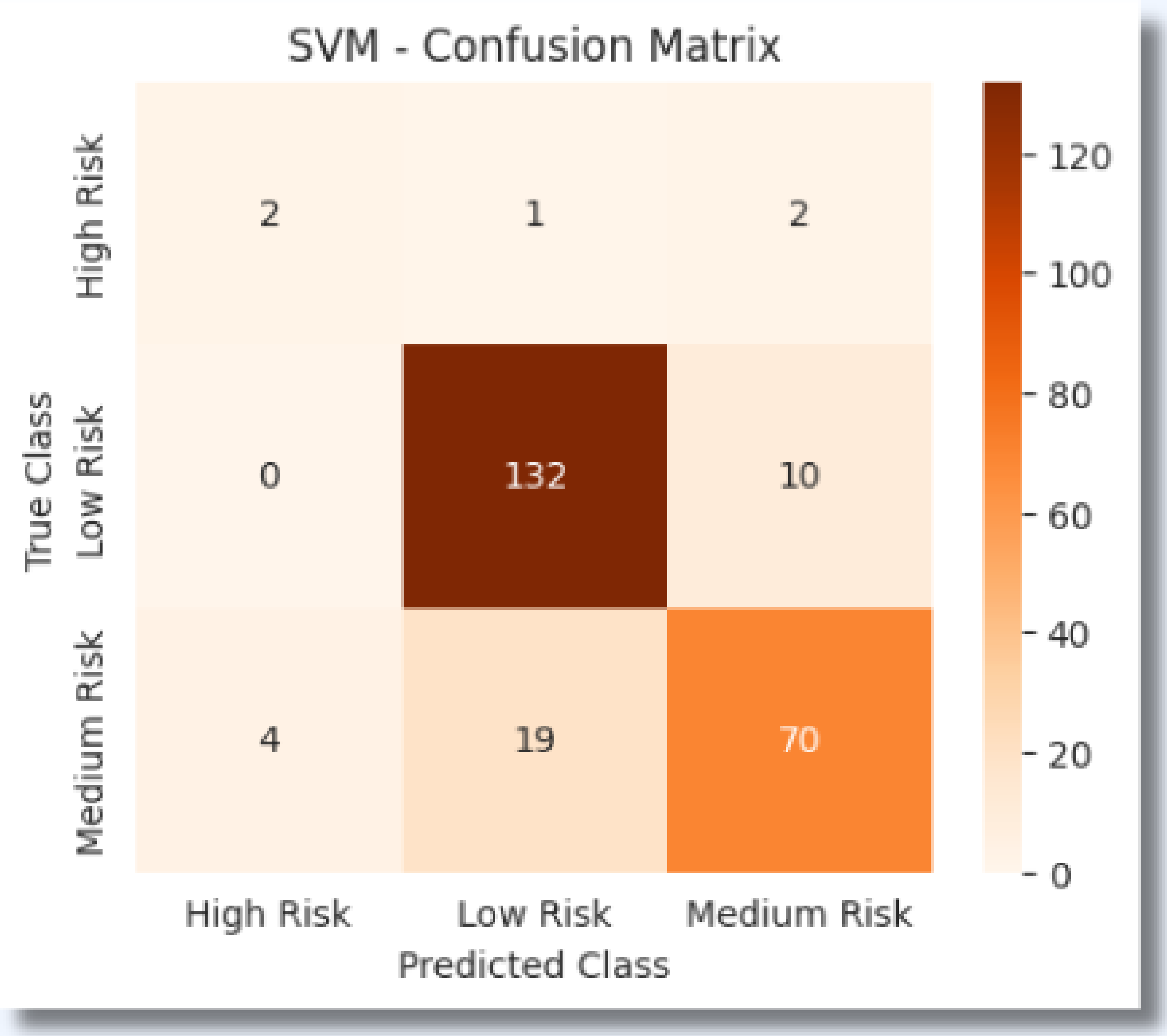
Accuracy: 0.8958,
Precision: 0.9044



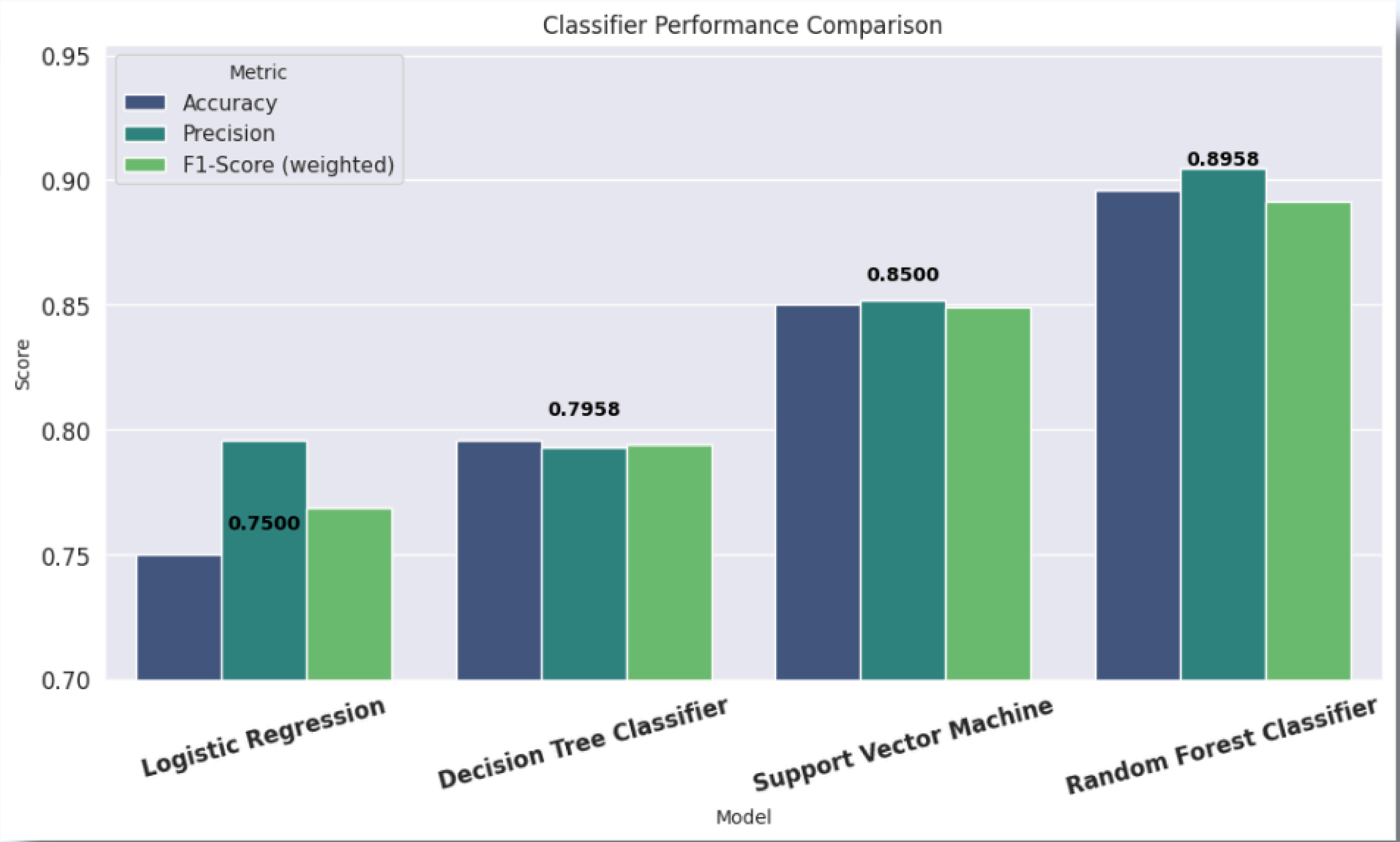
Models Implemented

SVM

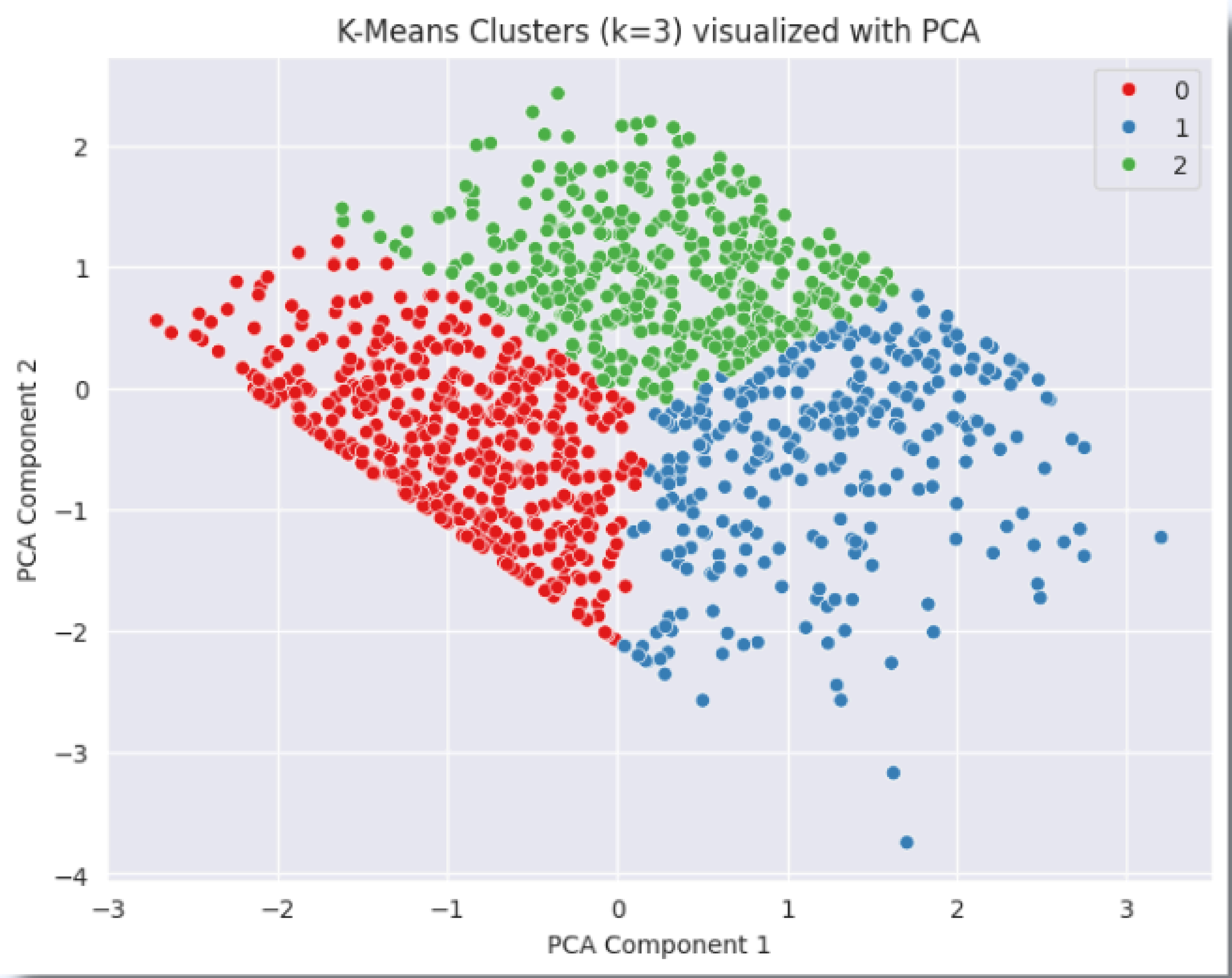
Accuracy: 0.8500
Precision: 0.8516



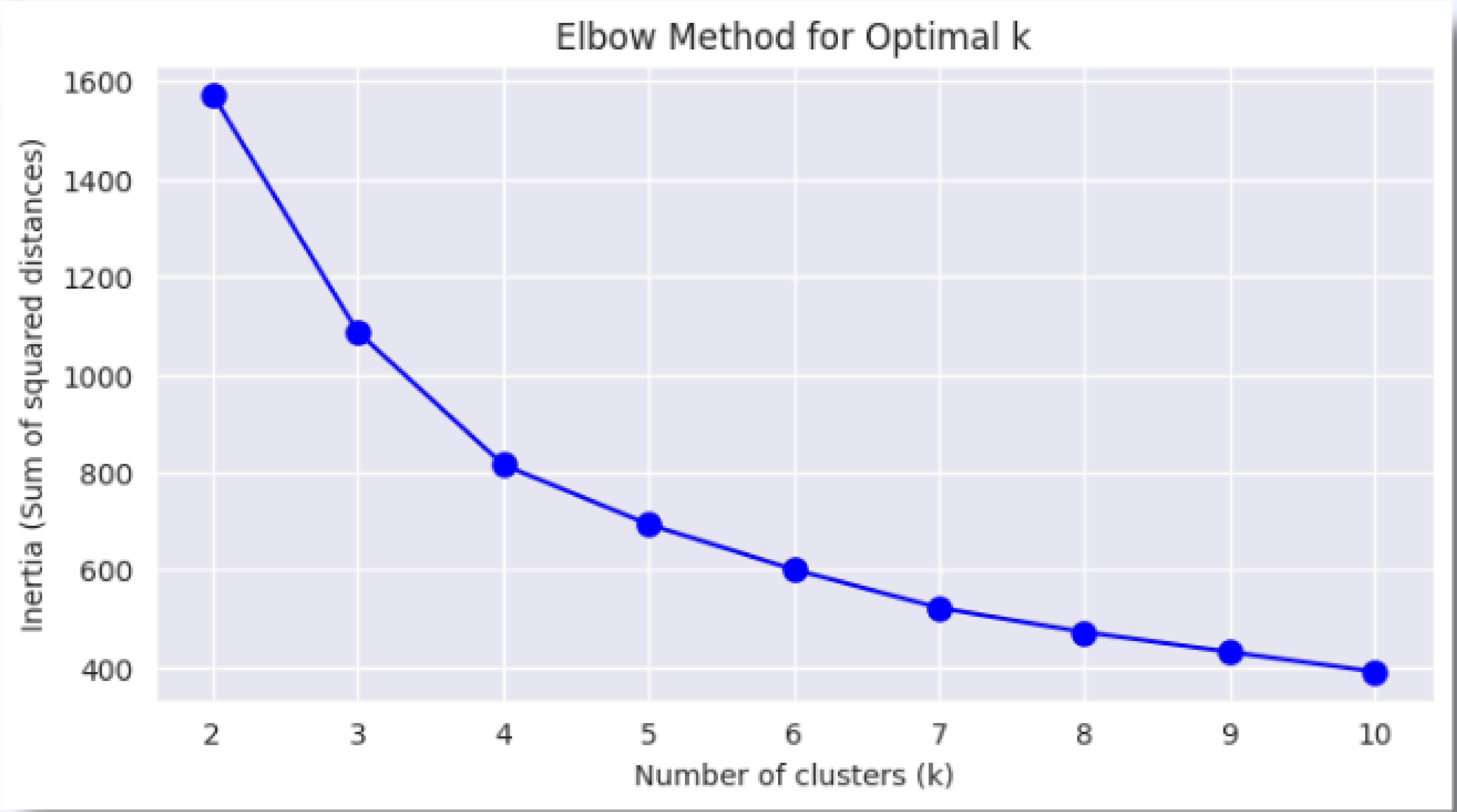
Model Comparison Table



K-Means Clustering



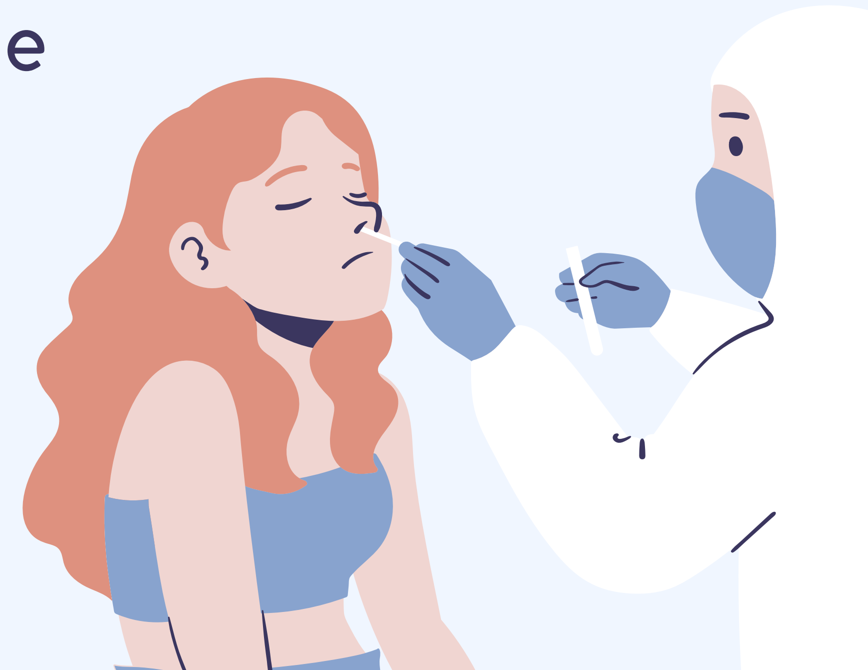
K-Means Clustering



- The K-Means clustering algorithm grouped individuals into meaningful lifestyle or risk-based categories.
- Evaluation using Silhouette Score confirmed good cluster quality and separation.
- Overall, this approach provides valuable insights for health risk prediction and personalized insurance planning.

Conclusion

- For classification, the Random Forest Classifier gave the highest accuracy and stability among all models.
- For regression, the Linear Regression model achieved the best R^2 score and lowest mean squared error (MSE).
- Hence, Random Forest is best suited for disease risk prediction, while Linear Regression is most effective for medical expense prediction.



Thank You