A PRELIMINARY MINI PROJECT REPORT ON

Smart Healthcare Expense & Disease Prediction System

SUBMITTED TOWARDS THE PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

BACHELOR OF TECHNOLOGY (Third Year B. Tech.)

Academic Year: 2025-26

By:

Piyush Ahirao – 123B1B077.

Aabha Chaudhari – 123B1B099.

Under The Guidance of

Dr. Sarika Deokate

DEPARTMENT OF COMPUTER ENGINEERING,

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

SECTOR 26, NIGDI, PRADHIKARAN

Abstract

The *Smart Healthcare Expense & Disease Prediction System* is an intelligent machine learning–based solution developed on Google Colab that predicts disease risks, estimates healthcare expenses, and suggests personalized lifestyle improvements. The system integrates multiple machine learning techniques — **classification** for disease prediction, **regression** for cost estimation, and **clustering** for patient grouping and health pattern analysis. By analyzing user inputs such as age, gender, symptoms, medical history, and lifestyle habits, the model generates real-time insights into potential health risks and their associated treatment costs. Additionally, it provides preventive lifestyle recommendations to help users maintain better health and reduce medical expenses. This system bridges the gap between predictive analytics and healthcare awareness, promoting early diagnosis, cost planning, and data-driven wellness management.

Index

Chapter 1: Introduction

1.1 Problem Statement

The healthcare industry suffers from **escalating costs** and a lack of tools for **proactive risk and expense assessment**. Patients lack a unified, personalized system to simultaneously predict their **disease risks** and the **associated treatment costs** based on individual data. This deficiency hinders early intervention and effective financial planning. Our project aims to solve this by developing the **Smart Healthcare Expense & Disease Prediction System**, which uses machine learning (Classification, Regression, and Clustering) to provide integrated health risk prediction, cost estimation, and personalized wellness recommendations. This intelligent solution promotes data-driven wellness management and cost transparency.

1.2 Objective

The project objectives are structured around the three core machine learning functionalities:

- **Classification (Disease/Risk Prediction):** To develop an accurate model (e.g., Random Forest Classifier) to categorize patients into health risk levels (**Low, Medium, or High Risk**).
- **Regression (Expense Estimation):** To build a robust model (e.g., ElasticNet Regression) to precisely **estimate potential medical charges** (expenses) using patient demographic and lifestyle data (Age, BMI, Smoker, etc.).
- **Clustering (Personalized Awareness):** To apply **K-Means Clustering** to identify distinct patient groups and uncover underlying health patterns for delivering **personalized lifestyle recommendations**.
- **Functional Goal:** To facilitate **early diagnosis** and **effective financial planning** by converting data into actionable awareness.

Chapter 2: Data Exploration

2.1 Data Sources & Features

The Smart Healthcare Expense & Disease Prediction System utilizes a single comprehensive dataset of patient health records for both predictive modeling tasks (Regression and Classification) and clustering analysis. (LINK TO DATASET-Github)

- **Dataset Shape:** (1338, 7)
- **Number of Records (Rows):** 1338 patient records.
- **Number of Features (Columns):** 7 key health and demographic features.
- **Purpose:** To analyze patient health data for predicting disease likelihood and estimating medical expenses, aiding in early diagnosis and financial planning.

Feature Name	Type	Description
Age	Numerical	The age of the primary beneficiary.
Sex	Categorical	The gender of the beneficiary (male or female).
BMI	Numerical	Body Mass Index, a measure of body fat based on height and weight.
Children	Numerical	The number of children covered by the health insurance / number of dependents.
Smoker	Categorical	Indicates whether the beneficiary is a smoker (yes or no).
Region	Categorical	The beneficiary's residential area in the US (e.g., northeast, southeast, southwest, northwest).
Charges	Numerical	The individual medical costs billed by health insurance (the expense).

Data Information & Statistical Summary-

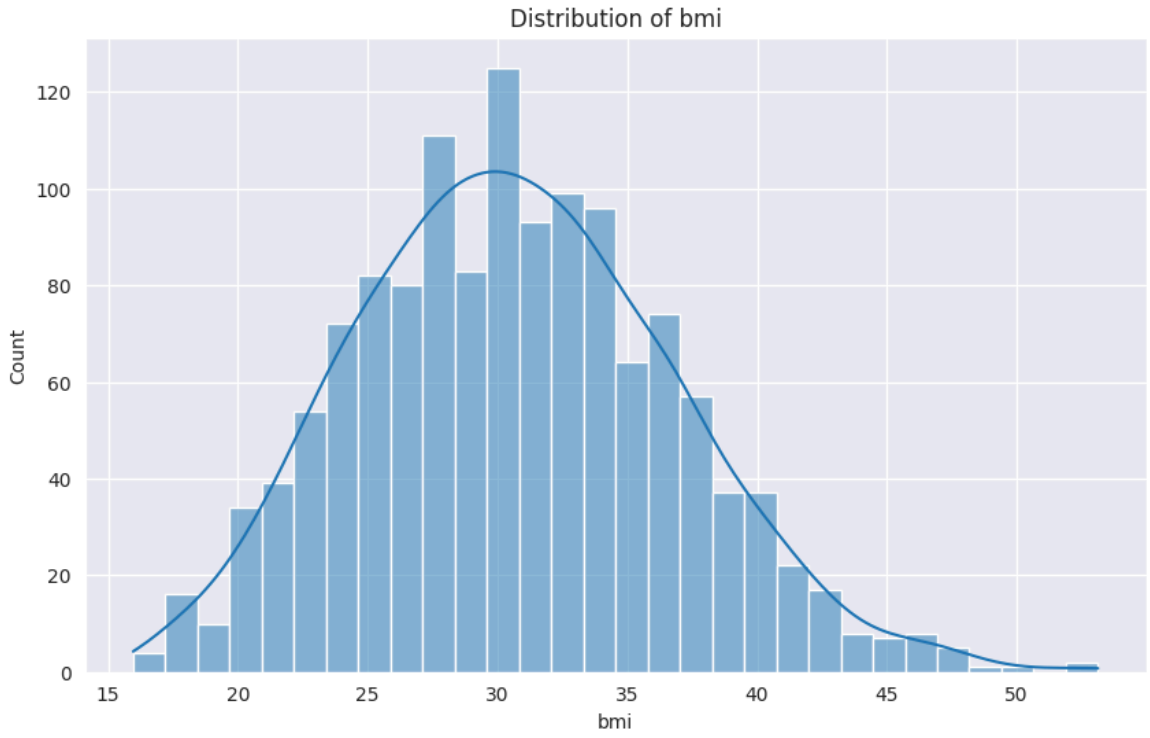
```
Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   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
```

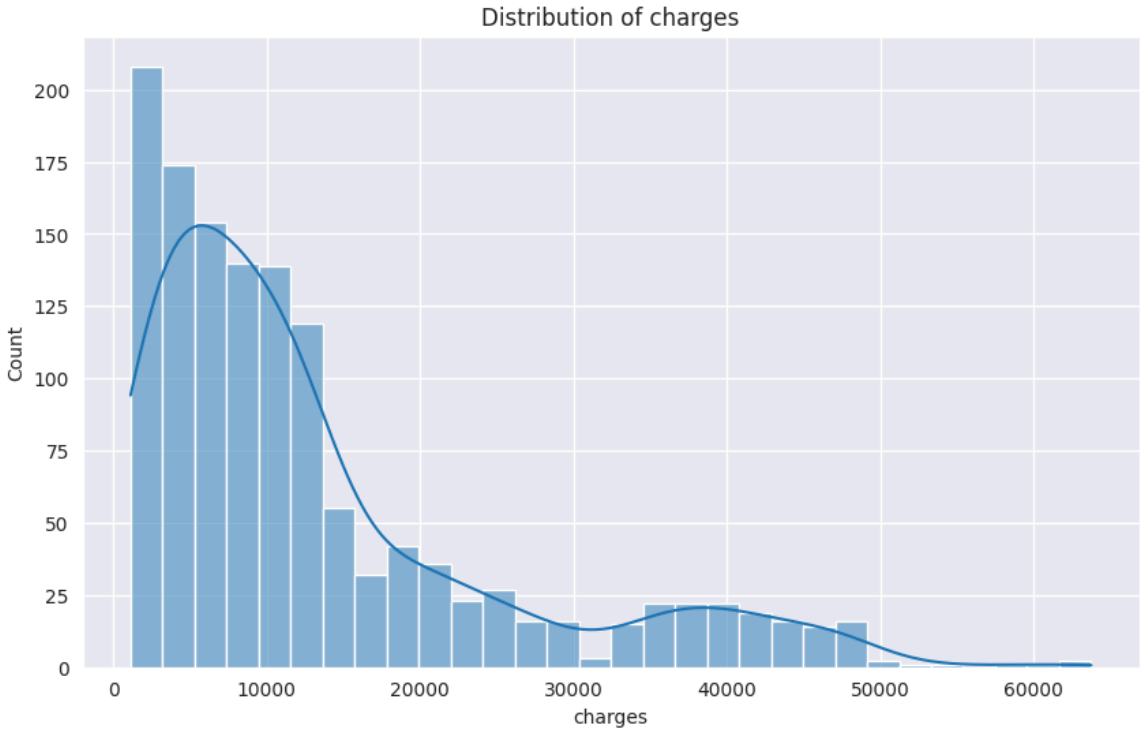
Statistical Summary:

	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

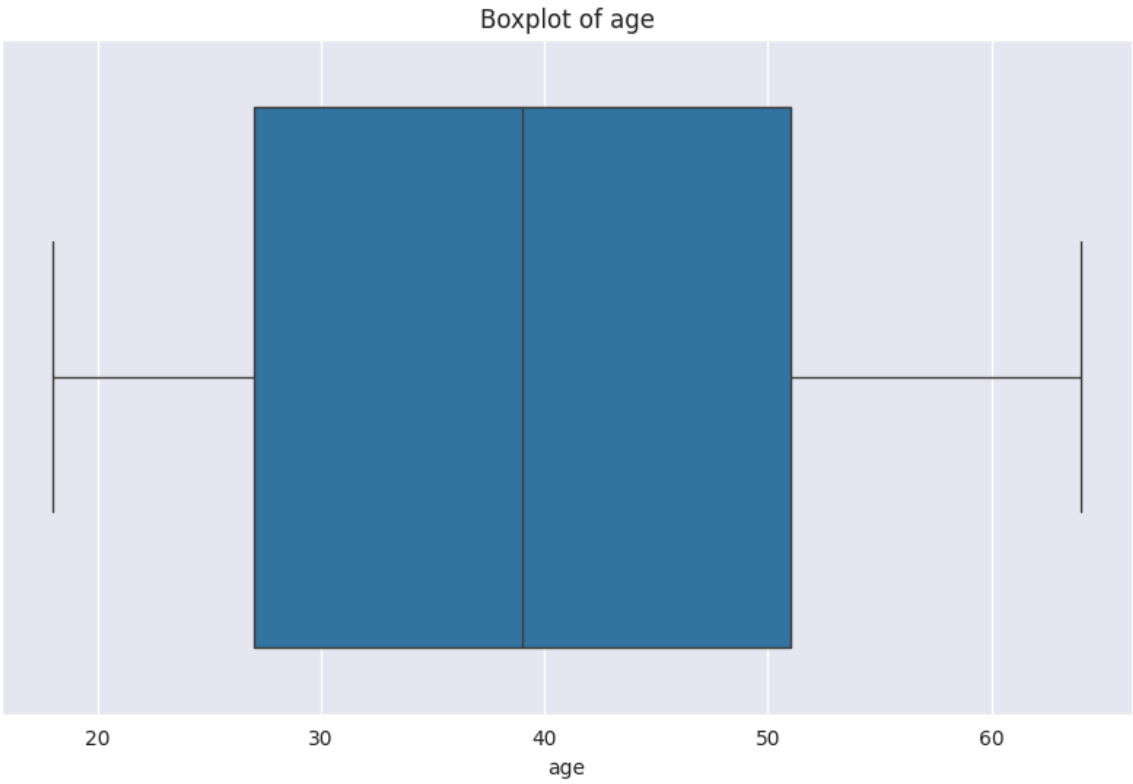
2.2 Exploratory Data Analysis (EDA)

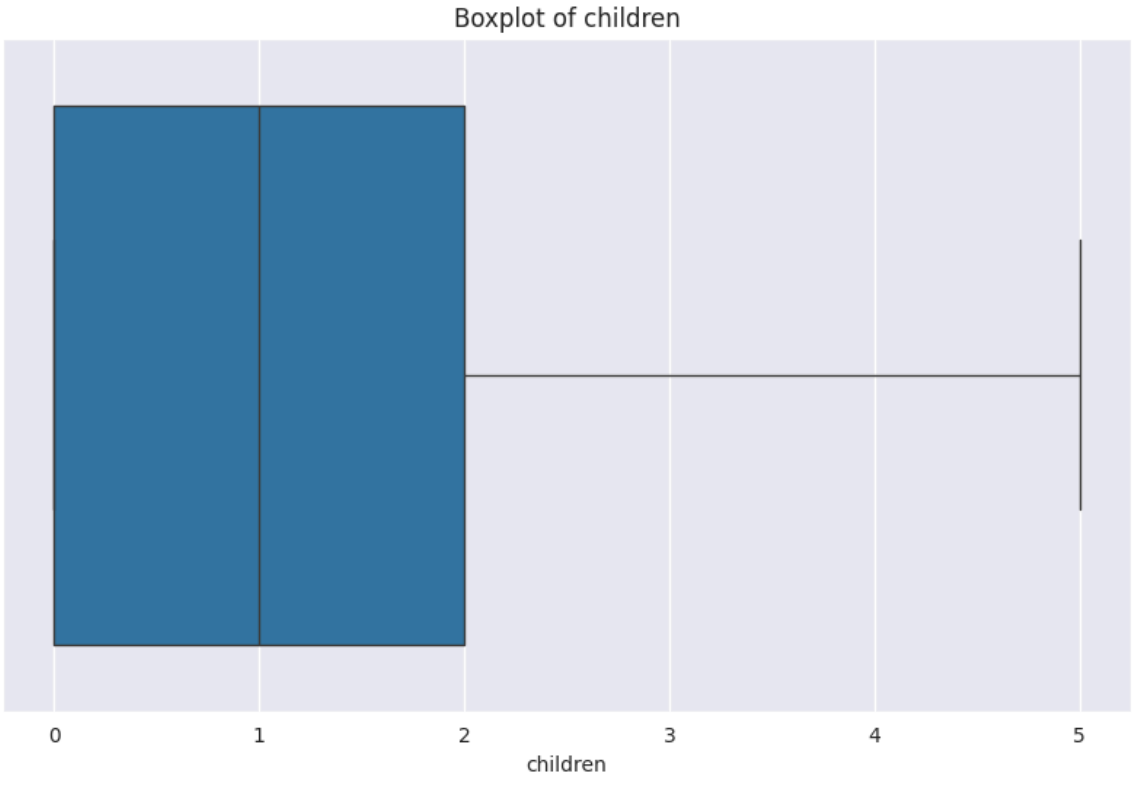
1) Histogram



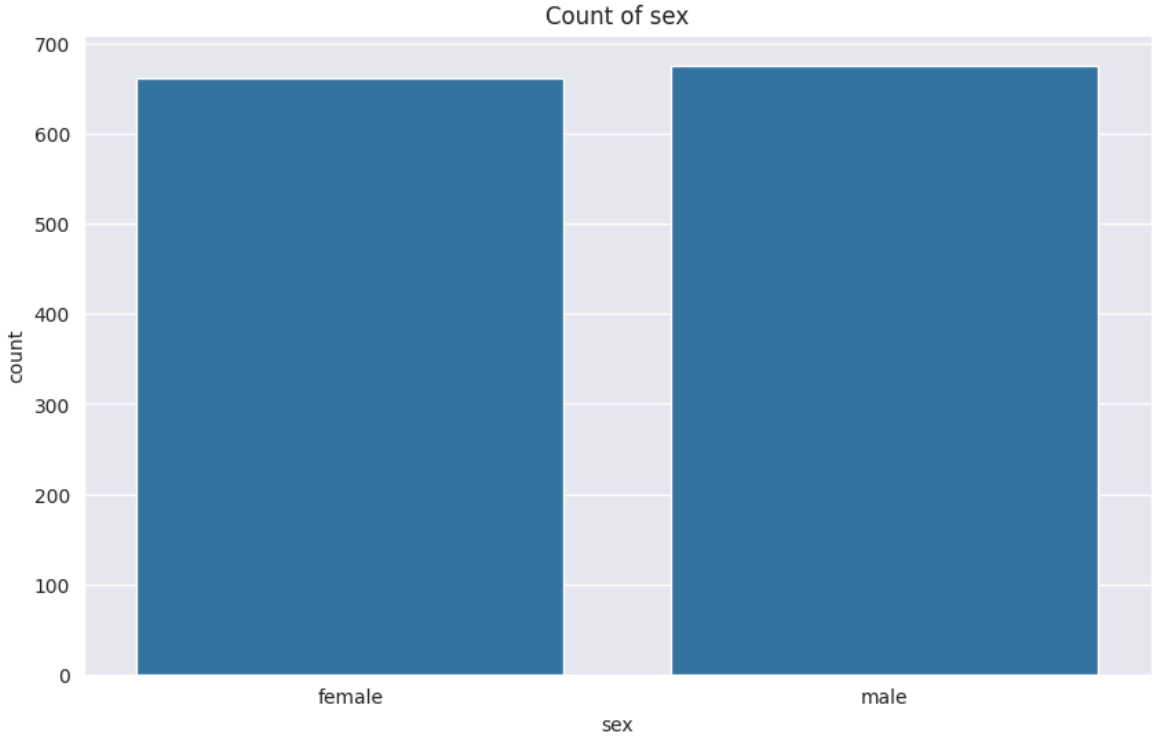


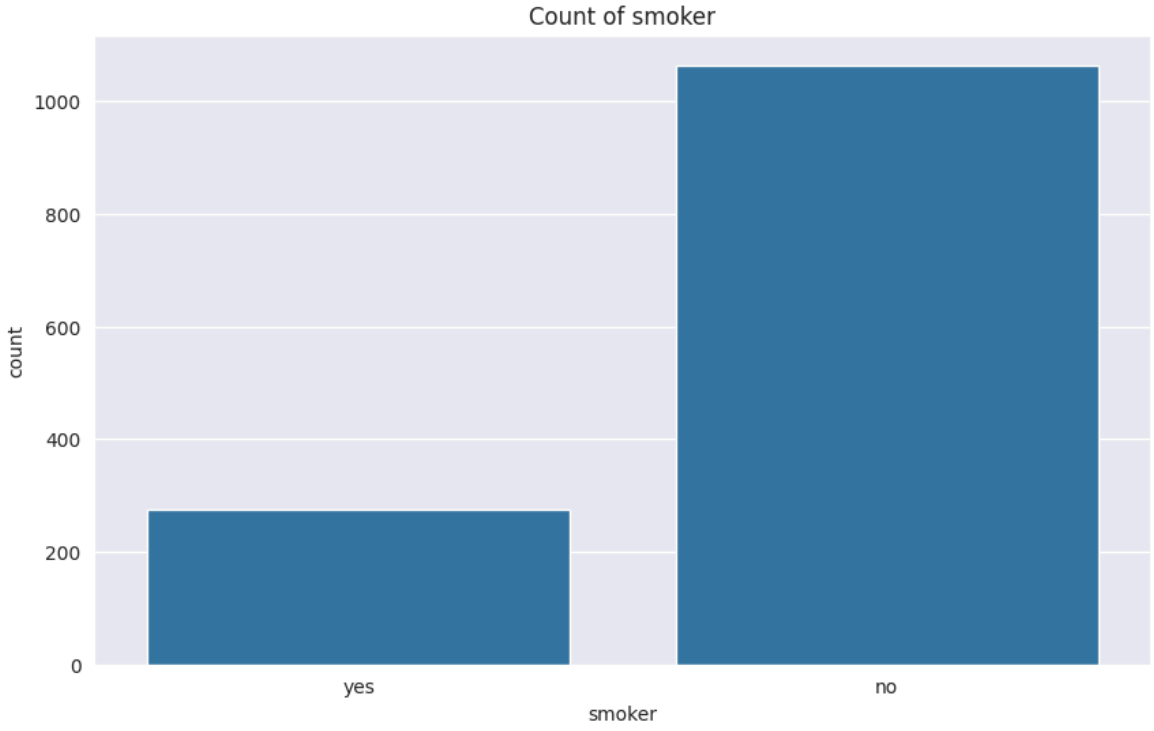
2) Boxplot


Boxplot of age
20
30
40
50
60
age

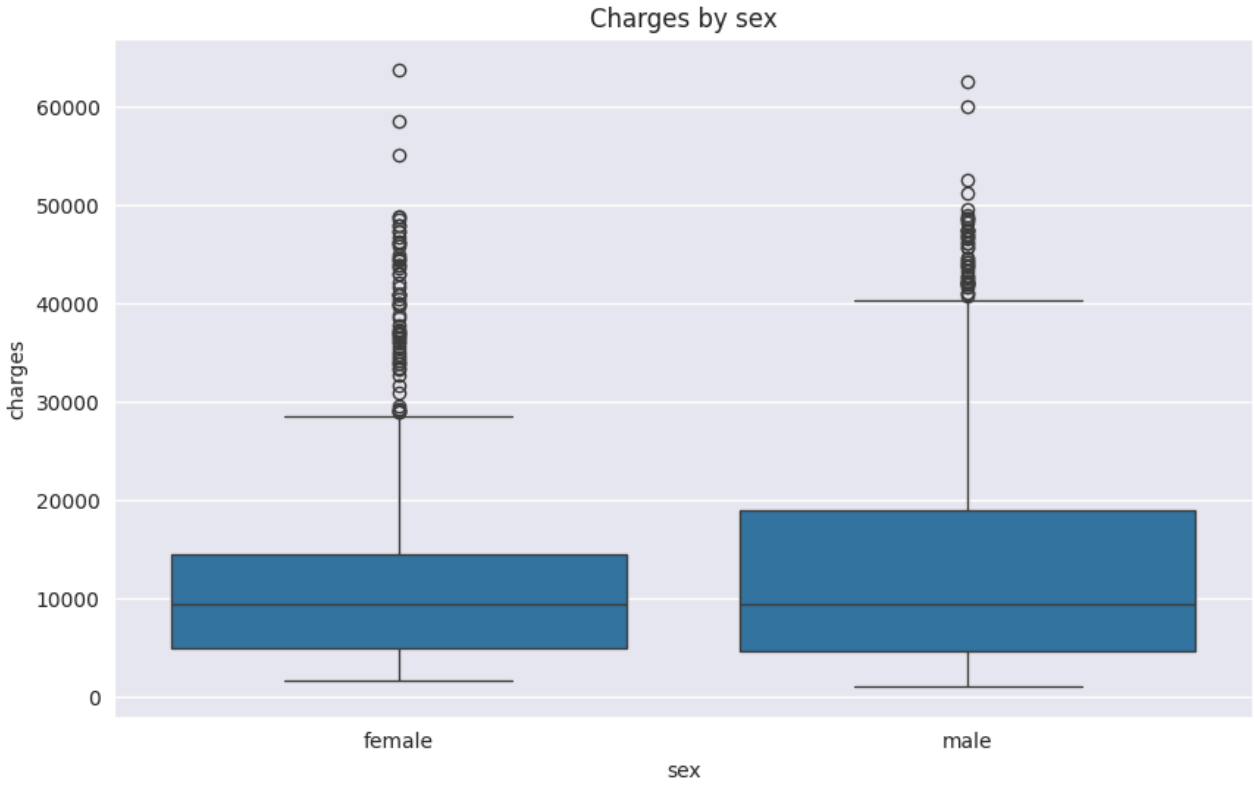

Boxplot of children
0
1
2
3
4
5
children

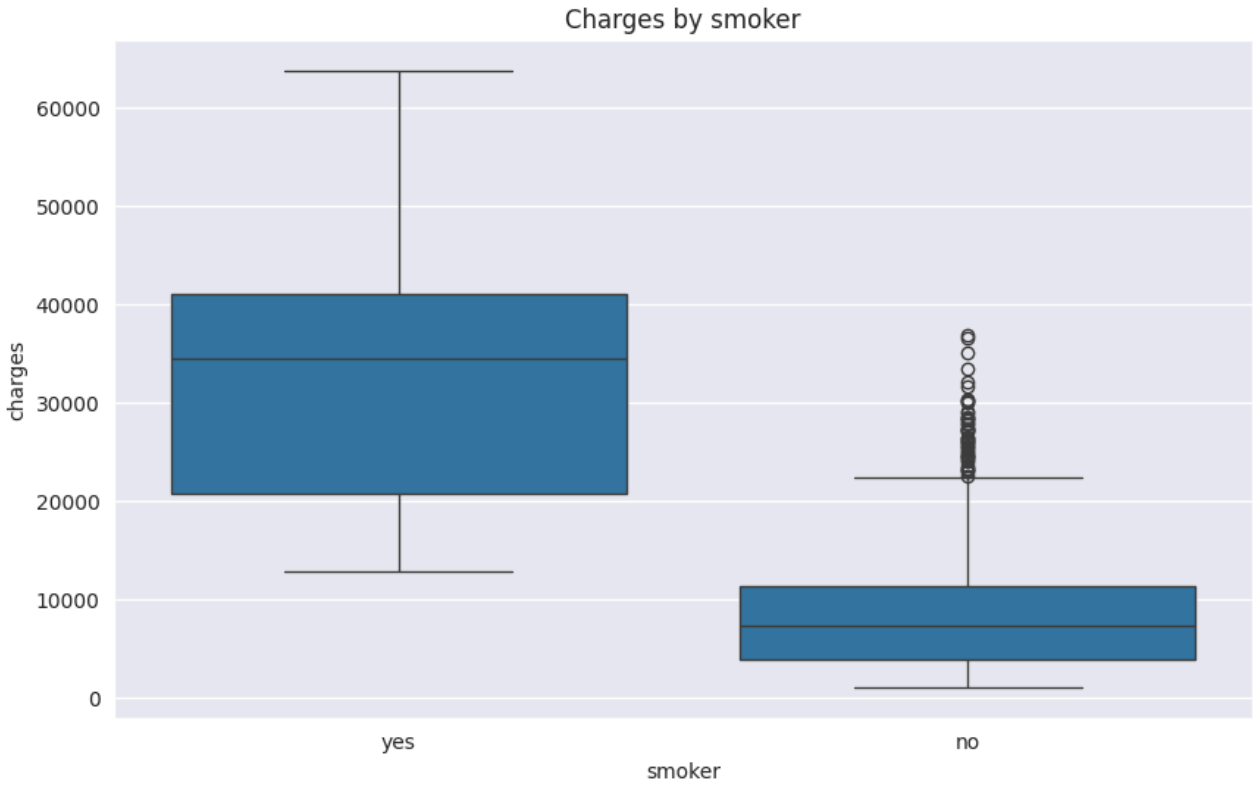
3) Count Plot


Count of sex
700
600
500
400
300
200
100
0
count
female
male
sex


Count of smoker
1000
800
600
400
200
0
count
yes
no
smoker

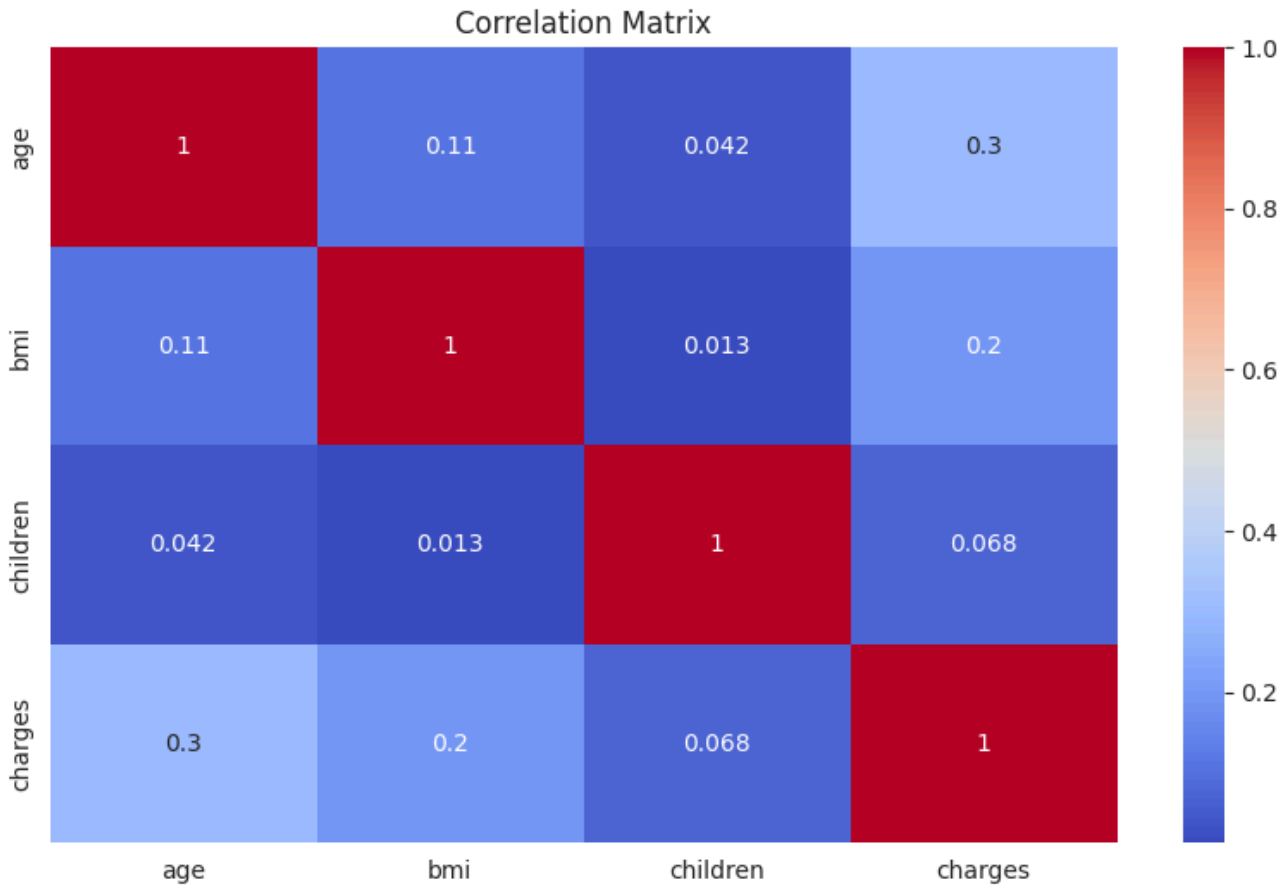
4) Outliers(Represented through Boxplot)






```
Number of outliers in charges: 139
Data shape after removing outliers: (1199, 8)
```

5) Heatmap


Correlation Matrix
age
bmi
children
charges
1
0.11
0.042
0.3
0.11
1
0.013
0.2
0.042
0.013
1
0.068
0.3
0.2
0.068
1
age
bmi
children
charges
1.0
0.8
0.6
0.4
0.2

6) Pairplot


Pairplot Colored by Smoker Status
age
bmi
children
charges
smoker
yes
no

2.3 Preprocessing Pipeline

2.3.1. Handling Data Quality

Based on the initial data inspection (text{df.info()} output), the dataset was found to be clean regarding null values:

- **Missing Values:** All 7 columns across 1338 entries were reported as non-null, meaning no imputation strategy was required.
- **Outlier Analysis:** Outliers were observed primarily in the Charges column, particularly within the non-smoker group. While 139 outliers were identified in Charges, these were retained for the regression task to maintain the integrity of real-world high expense predictions.

2.3.2. Feature Engineering and Encoding

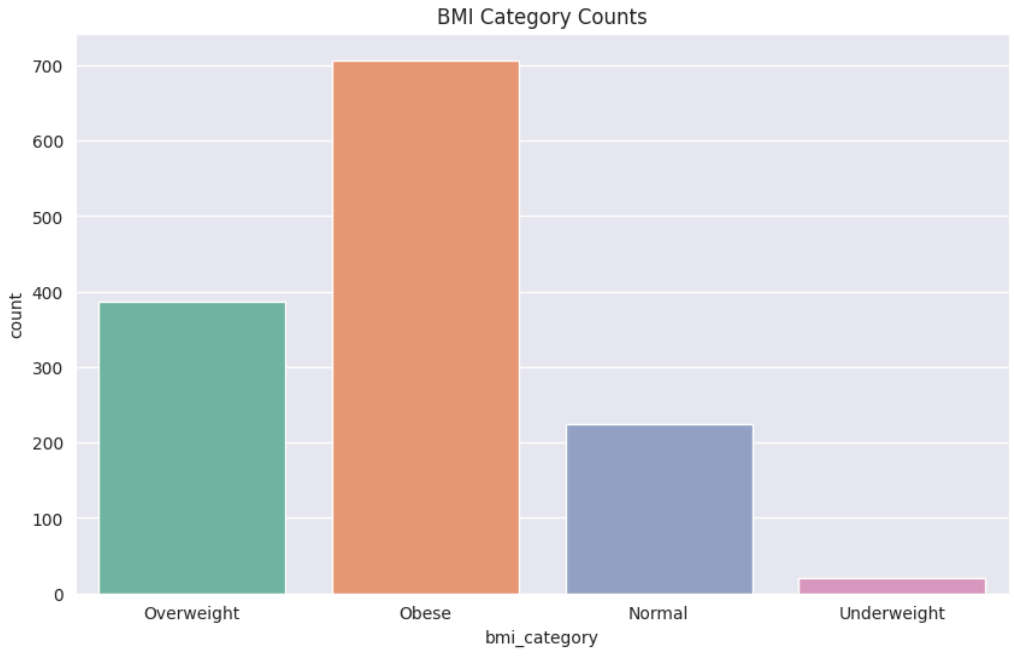
Two critical feature engineering steps were performed to create derived features essential for improving model accuracy and enabling the classification task.

A. Categorical Feature Encoding

- **Action:** Categorical features ({sex},{smoker}, and{region}) were converted into a numerical format.
- **Method: One-Hot Encoding** (or a similar technique like Label Encoding for binary features) was applied to features like ({sex},{smoker}, and{region}) to prevent the models from assuming ordinal relationships.

B. Feature Engineering: BMI Categories

- **Action:** A new feature, {bmi_category}, was created from the continuous {BMI} column.
- **Method:** {BMI} values were grouped into four classes: **Underweight, Normal, Overweight, and Obese**.
- **Encoding:** {bmi_category} was then subjected to **Label Encoding**.



C. Target Variable Transformation (Classification Only)

- **Action:** The continuous {charges} column (the regression target) was transformed into a categorical target variable for the classification task.
- **Method:** {Charges} were bucketed into three risk levels based on expense ranges: **Low Risk, Medium Risk, and High Risk**.
- **Encoding:** These three risk categories were then **Label Encoded** into numerical values for the classification models.

2.3.3. Scaling and Normalization

To ensure that no single feature dominates the model training solely due to its magnitude, numerical features were scaled:

- **Features Scaled:** {age}, {bmi},{children}, and the transformed {charges}.
- **Method:** A standard scaling technique (e.g., **StandardScaler** o**MinMaxScaler**) was applied to standardize the range of the independent numerical variables. This step is particularly important for distance-based algorithms like **Ridge/Lasso Regression** and **K-Means Clustering**.

Chapter 3: Methodology: Model Testing & Comparison

3.1 Data Splitting Strategy

The data splitting ensures that models are trained on one subset of the data and rigorously tested on a separate, unseen subset. This process allows for an unbiased evaluation of the models' generalization ability to new patient records.

A. Train-Test Split

The dataset was initially partitioned into two primary sets:

- Training Set: Used for model fitting, where the algorithms learn the relationships between the input features and the target variables (Charges for Regression; Risk Category for Classification).
- Test Set: An independent subset of data reserved exclusively for the final model evaluation to measure predictive performance.

B. Consideration for Target Variables

The split strategy addressed the specific nature of the two different target variables in the project:

1. Regression (Charges): A simple random split was sufficient for the continuous {Charges} target.
2. Classification (Risk Category): Given the highly imbalanced nature of the engineered risk categories (e.g., High Risk had only 23 total samples), a Stratified Sampling technique was crucial.

   Goal: To ensure that the test set maintained the same proportions of Low Risk, Medium Risk, and High Risk samples as the full dataset. This prevents the training set from being overly dominant by one class and ensures the test results are reliable for all risk levels.

3.2 Regression Models Tested (Cost Estimation)

Regression models were selected to evaluate different approaches (linear, regularized, and ensemble) for predicting the continuous **Charges** target variable.

- **Linear Regression:** Baseline model, assuming a linear relationship.
- **Ridge Regression:** Linear model with L2 regularization to prevent overfitting by penalizing large coefficients.
- **Lasso Regression:** Linear model with L1 regularization, useful for feature selection by shrinking some coefficients to zero.
- **ElasticNet Regression:** Combines L1 and L2 penalties, offering a balance between regularization and feature selection.
- **Random Forest Regressor:** An ensemble tree-based model that handles non-linearity and interactions well.

3.3 Classification Models Tested (Disease Prediction)

Classification models were chosen to predict the categorical target variable: **Low Risk, Medium Risk, or High Risk**.

- **Logistic Regression:** A fundamental linear model for binary/multinomial classification.
- **Decision Tree:** A non-linear, interpretable model based on sequential feature splits.
- **Random Forest:** An ensemble method known for high accuracy and stability, effective for handling complex interactions.
- **Support Vector Machine (SVM):** A versatile model that uses kernels to find the optimal hyperplane for classification.

3.4 Clustering Methods Tested (Patient Grouping)

- **Method: K-Means Clustering** was used to segment the data.
- **Optimal K Determination:** The **Elbow Method** was applied to the inertia (sum of squared distances) to determine the optimal number of clusters, which was found to be {k}=3.
- **Visualization: Principal Component Analysis (PCA)** was used to reduce the high-dimensional features into two components for visualization of the {k}=3 clusters.

Chapter 4: Regression Analysis and Results (Cost Estimation)

4.1.1 Linear Regression

Linear Regression is a foundational machine learning model used for predicting a continuous outcome (Charges) by modeling a straight-line relationship with the input features. The model aims to find the line that best minimizes the distance to all data points.

Evaluation Metrics:

R2 Score : 0.5568

MSA : 27647351.6858

RMSE : 5258.0749

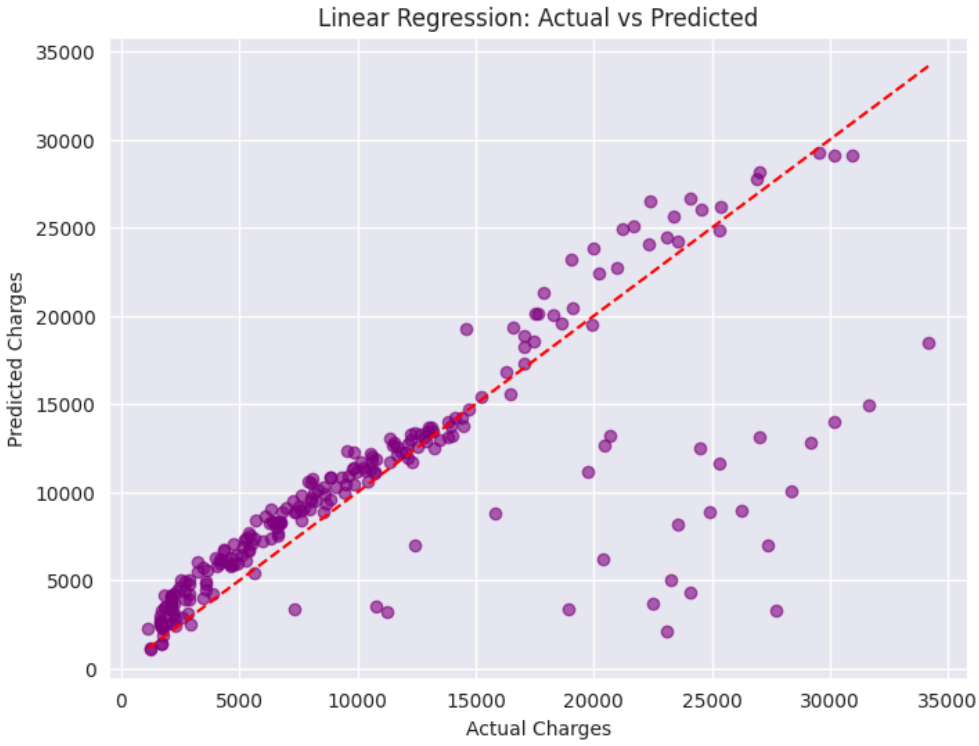
MAE : 2797.0320

Result Interpretation:

The R2 score suggests that approximately 55.68 of the variability in healthcare charges is explained by this model.

Actual vs. Predicted Plot:

The plot below visually confirms the model's fit, showing that while most predictions cluster along the ideal line, there is significant scatter, particularly for high actual charges, indicating potential limitations in its linearity for this dataset



4.1.2 Ridge Regression

This is a **regularized linear model** designed to prevent overfitting and improve prediction accuracy. It addresses multicollinearity by adding an **L2 penalty** (squared magnitude of coefficients) to the loss function. This penalty forces coefficients to be small, stabilizing the model and reducing its variance.

Evaluation Metrics:

R2 Score : 0.5579

MSA : 27575547.8138

RMSE : 5251.2425

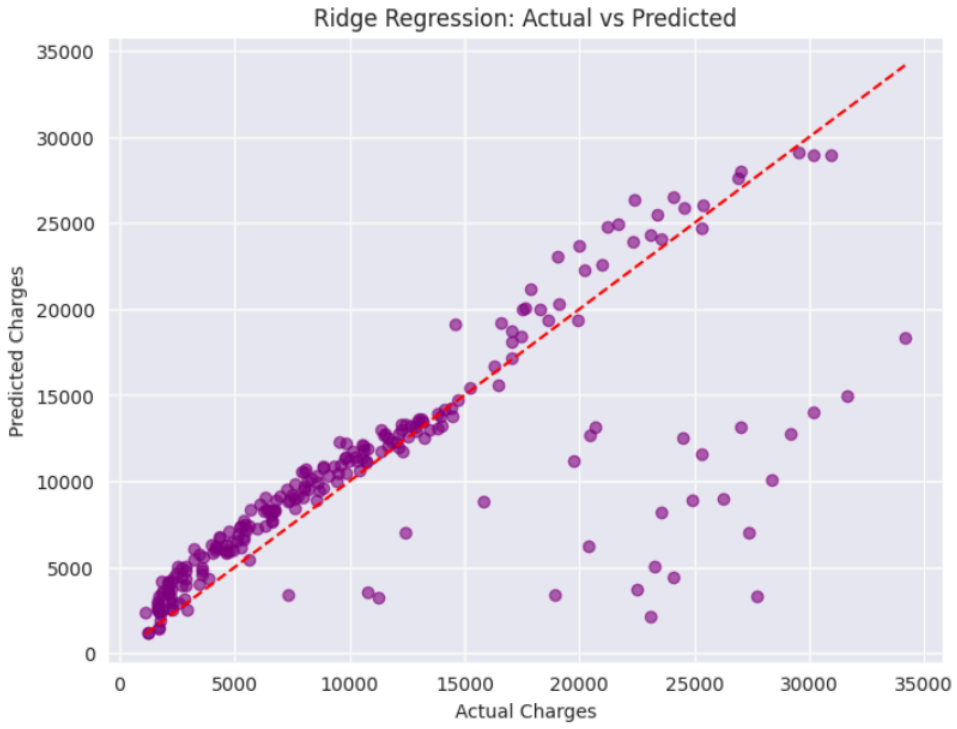
MAE : 2791.5766

Result Interpretation:

It performed slightly better than basic Linear Regression, achieving an R^{2} score of **0.5579**[9].

Actual vs. Predicted Plot:

Similar to Linear Regression, the plot indicates strong alignment at low charges and a **noticeable spread** from the ideal line at higher charges.



4.1.3 Lasso Regression

This **regularized linear model** uses an **L1 penalty** (absolute magnitude of coefficients) which is added to the loss function. The unique property of the L1 penalty is that it can drive the coefficients of less important features exactly to zero. This feature is useful for both regularization and automatic feature selection

Evaluation Metrics:

R2 Score : 0.5568

MSA : 27647317.0629

RMSE : 5258.0716

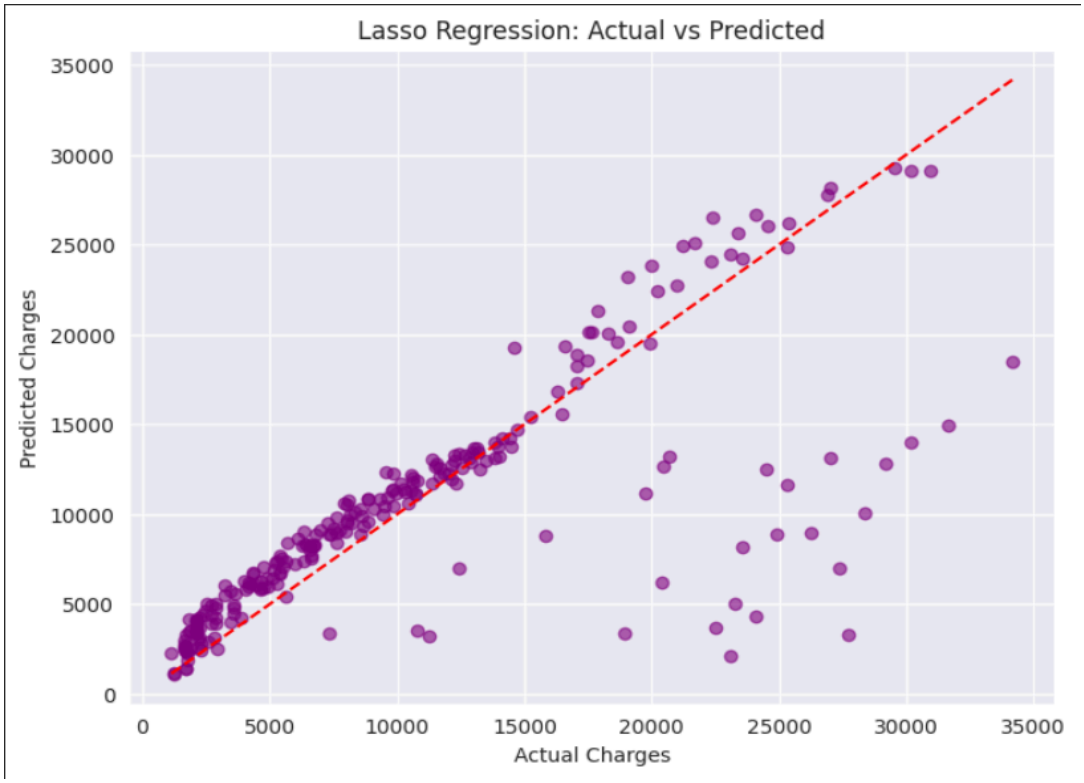
MAE : 2797.024

Result Interpretation:

The performance metrics were almost identical to those of the simple Linear Regression model, with an R2 score of **0.5568**.

Actual vs. Predicted Plot:

The plot is virtually the same as the Linear Regression plot, showing **scatter increase** as actual charges rise.



4.1.4 ElasticNet Regression

This is a **hybrid regularized linear model** that utilizes both **L1 (Lasso) and L2 (Ridge) penalties** in its loss function. It is particularly useful when dealing with a high number of features that are correlated with each other. It benefits from the coefficient shrinkage of Ridge and the feature selection of Lasso.

Evaluation Metrics:

R2 Score : 0.5611

MSA : 27380338.4948

RMSE : 5232.6225

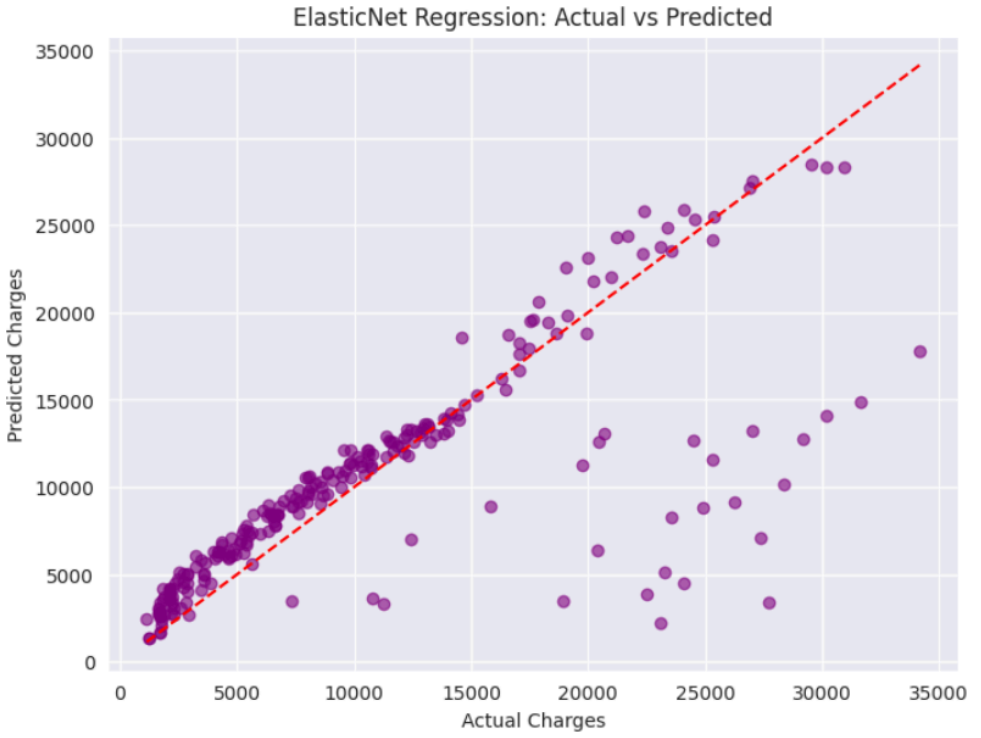
MAE : 2778.8538

Result Interpretation:

This model achieved the **highest R2 score (0.5611)** and the lowest MSE among all regression models, making it the top performer by these metrics.

Actual vs. Predicted Plot:

It displays the tightest clustering around the ideal prediction line, particularly for lower and moderate charge values.



4.1.5 Random Forest Regression

This is an **ensemble learning method** that fits multiple independent decision tree models on various sub-samples of the dataset. It reduces the risk of overfitting inherent in single decision trees by taking the **average** of the predictions from all individual trees to arrive at the final expense prediction.

Evaluation Metrics:

R2 Score : 0.5607

MSA : 27403463.4422

RMSE : 5234.8317

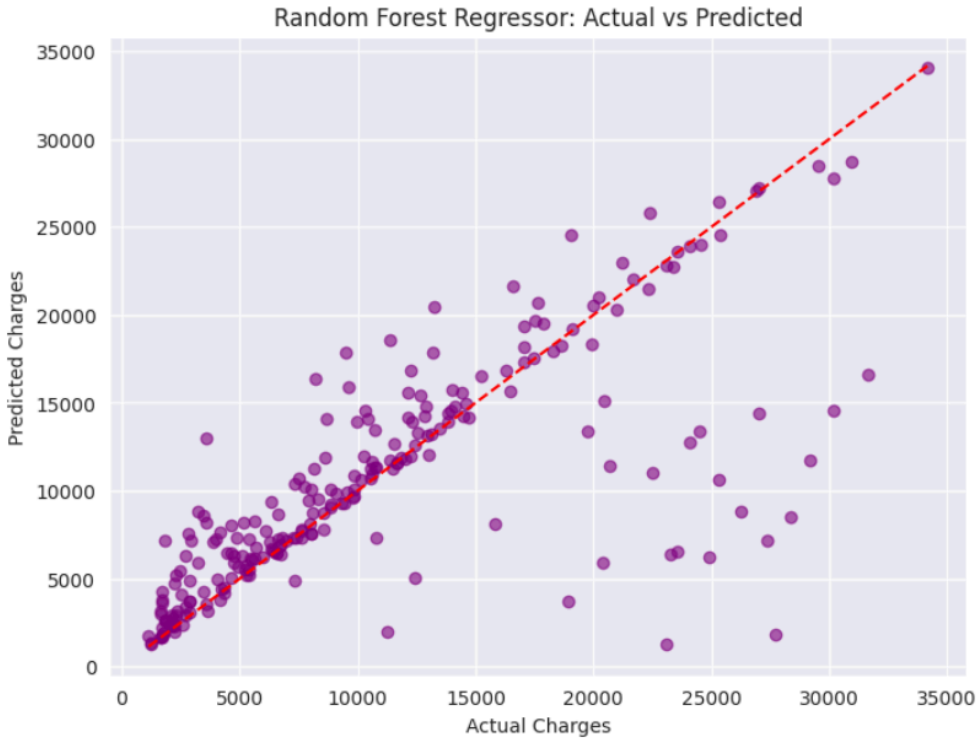
MAE : 2689.3716

Result Interpretation:

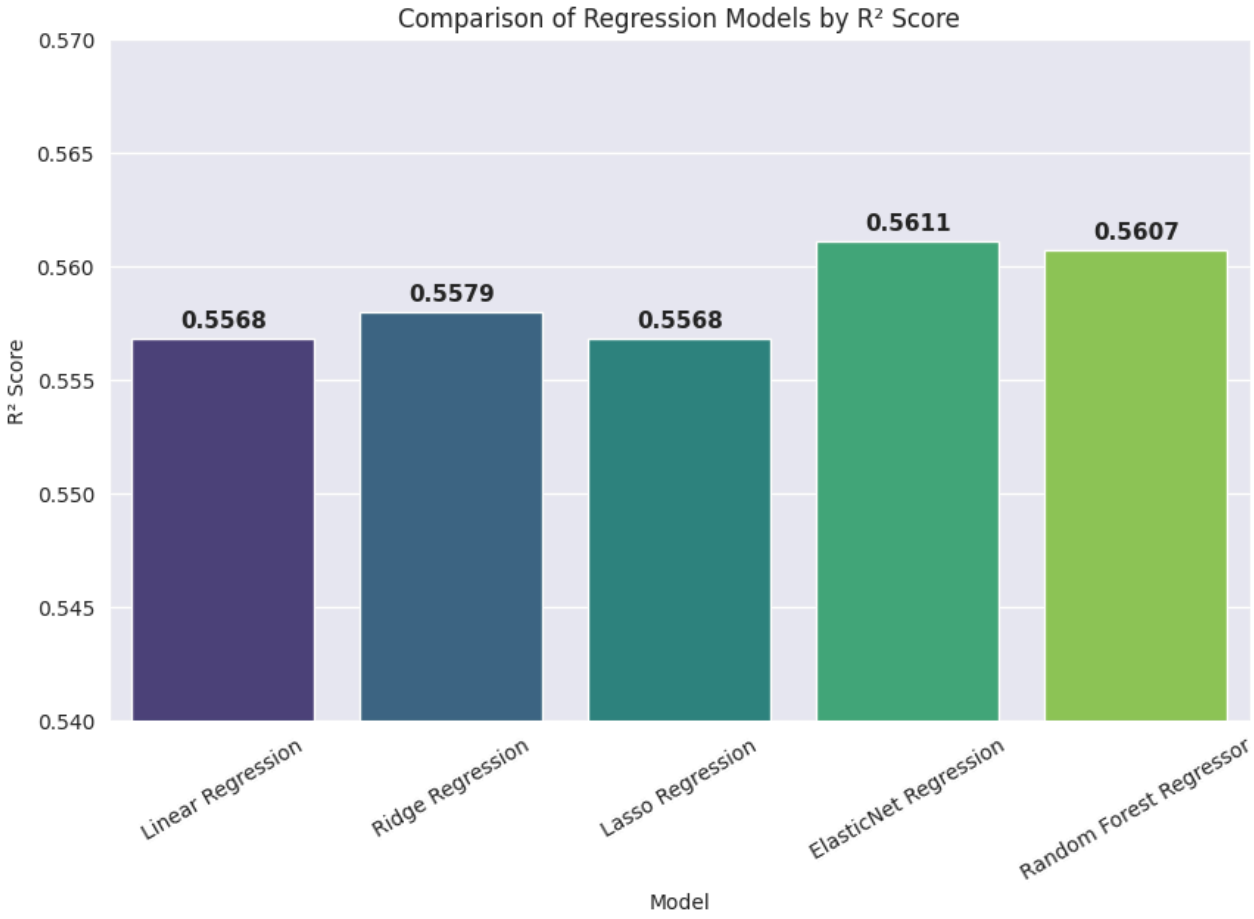
It performed very strongly, with an R2 score of 0.5607 25, and achieved the lowest Mean Absolute Error (MAE) of all models.

Actual vs. Predicted Plot:

The plot shows a concentration of points along the ideal line, demonstrating its predictive power, but it also has **outlier predictions** for high-charge instances.



Comparison of Regression Models by R2 Score-




```
Linear Regression: R2=0.5568
Ridge Regression: R2=0.5579
Lasso Regression: R2=0.5568
ElasticNet Regression: R2=0.5611
Random Forest Regressor: R2=0.5607

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

Chapter 6: Classification Models

6.1.1. Logistic Regression

Despite its name, Logistic Regression is a statistical model used for **classification tasks**. It models the probability of a binary outcome using the logistic function, but it is extended for **multi-class prediction** (like the three risk levels in this study) using methods such as one-vs-rest.

Evaluation Metrics:

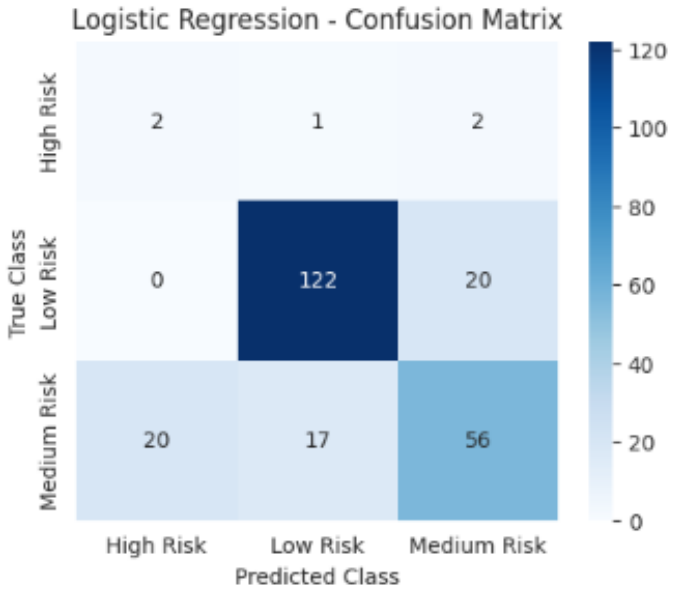
Metric	Value
Accuracy	0.7500 4444
Precision	0.7957 [cite: 356]

Result Interpretation:

Achieved an accuracy of **0.7500** , indicating it correctly predicted the risk category 75% of the time.

Confusion Matrix Plot

The matrix shows a high number of true positives for "Low Risk" (122) but notably confuses Medium Risk and Low Risk cases.



6.1.2. Decision Tree

A non-parametric supervised learning method that uses a tree-like structure for classification. It segments the data space into regions and assigns a class to each region, with internal nodes representing feature tests and leaf nodes representing the class label.

Evaluation Metrics:

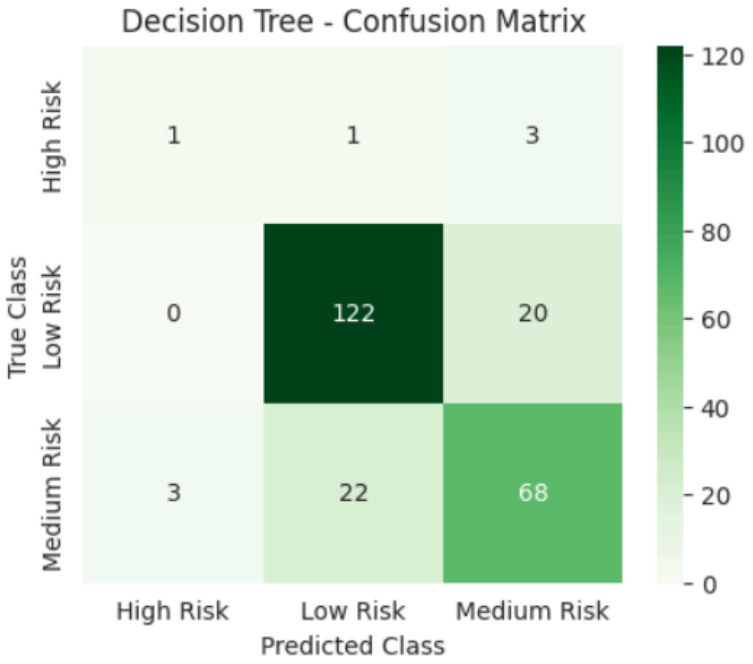
Metric	Value
Accuracy	0.7958
Precision	0.7926

Result Interpretation:

The model demonstrated improved accuracy over Logistic Regression at **0.7958** , indicating better performance in classifying risk.

Confusion Matrix Plot

The matrix shows an improvement in correctly classifying "Medium Risk" (68) compared to Logistic Regression, but it still has significant misclassification into "Low Risk" (22).



6.1.3. Random Forest

An **ensemble learning method** that constructs multiple independent decision trees during training. For classification, the final output is the class that is the **mode** (most frequent class) predicted by all individual trees, which helps to mitigate overfitting and improve stability and accuracy.

Evaluation Metrics:

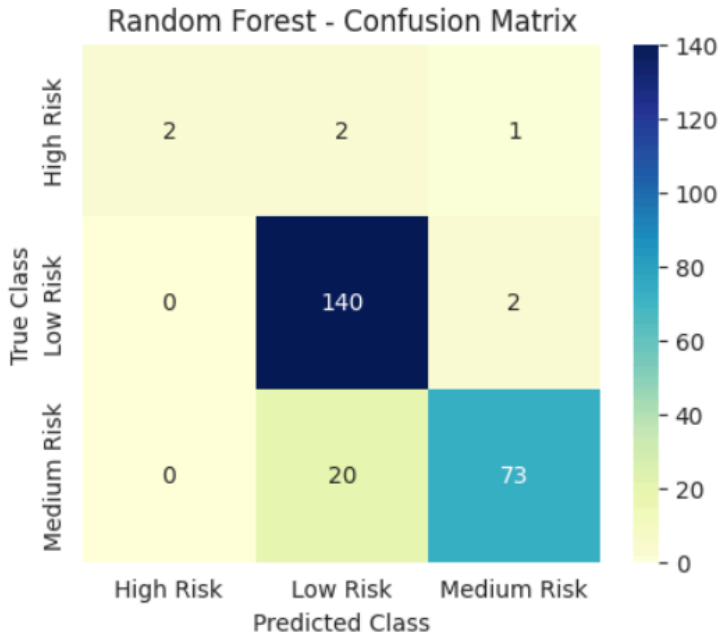
Metric	Value
Accuracy	**0.8958**
Precision	**0.9044**

Result Interpretation:

The **Random Forest Classifier** was the best-performing model, achieving the **highest accuracy (0.8958)** and precision.

Confusion Matrix Plot

This model shows the cleanest confusion matrix, with the highest counts of correct predictions in "Low Risk" (140) and "Medium Risk" (73) and very few misclassifications.



6.1.4. Support Vector Machine (SVM)

A powerful supervised learning model that works by finding the optimal **hyperplane** that distinctly separates the data points into classes. SVM is highly effective in high-dimensional spaces and in scenarios where the number of dimensions is greater than the number of samples.

Evaluation Metrics:

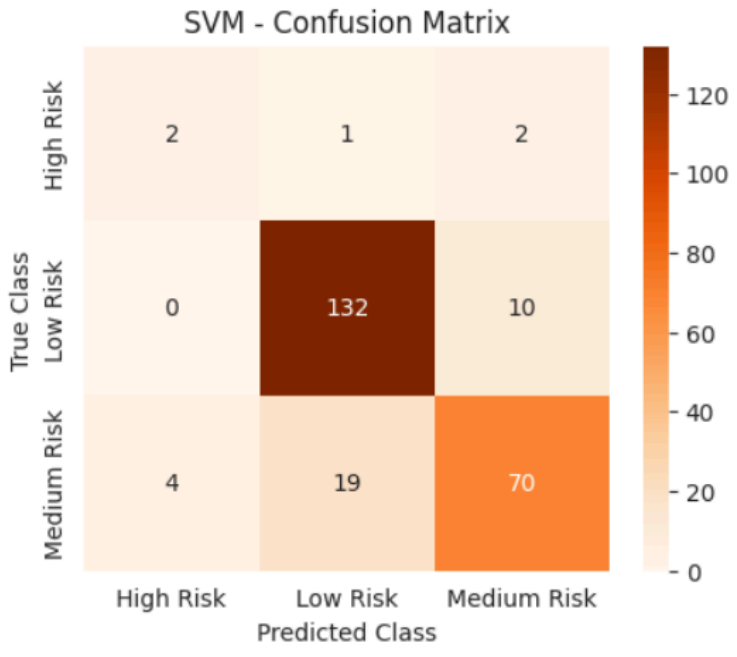
Metric	Value
Accuracy	0.8500
Precision	0.8516

Result Interpretation:

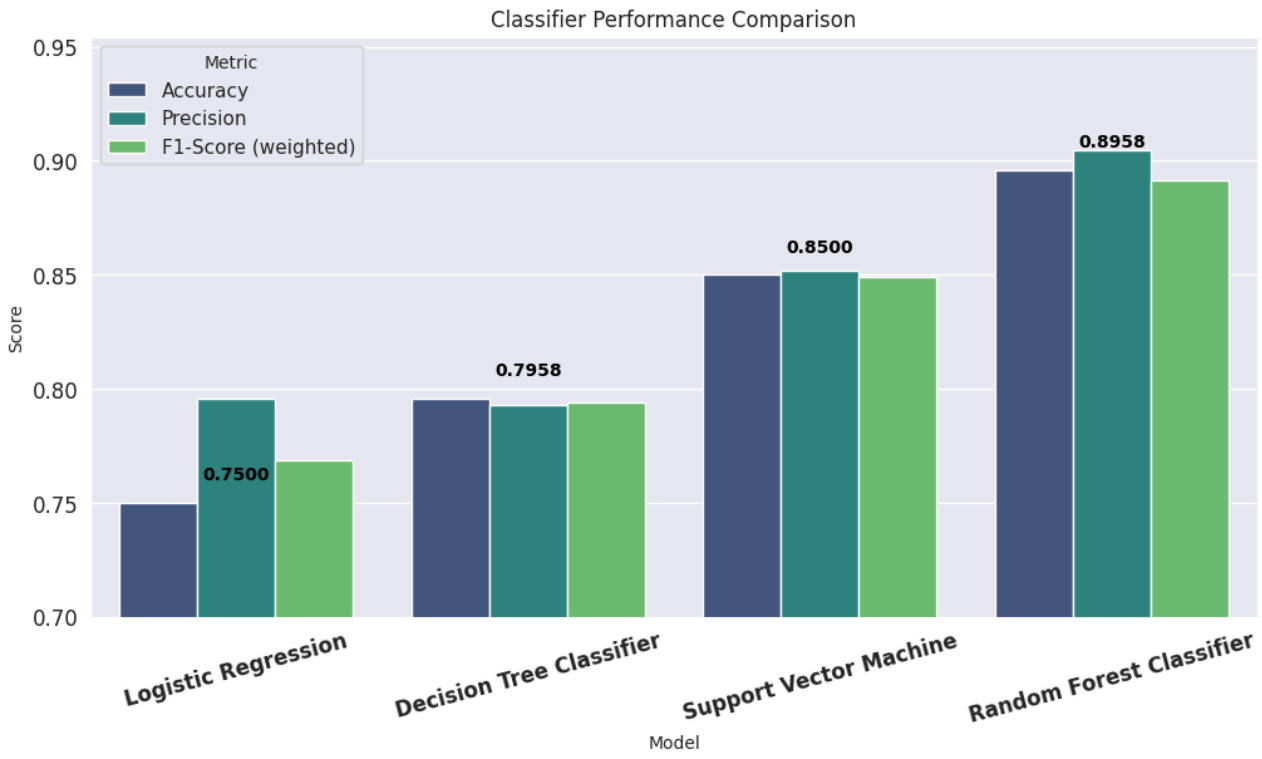
SVM secured the second-highest accuracy at **0.8500** , proving to be a highly effective model for the risk classification task.

Confusion Matrix Plot

The matrix displays strong performance, correctly classifying "Low Risk" (132) and "Medium Risk" (70) with better results than both Logistic Regression and Decision Tree.



Classifier Performance Comparison-




```
Logistic Regression: Accuracy=0.7500, Precision=0.7957
Decision Tree Classifier: Accuracy=0.7958, Precision=0.7926
Support Vector Machine: Accuracy=0.8500, Precision=0.8516
Random Forest Classifier: Accuracy=0.8958, Precision=0.9044

Classifier Performance Comparison Table:
```

	Model	Accuracy	Precision	F1-Score (weighted)
3	Random Forest Classifier	0.895833	0.904354	0.891624
2	Support Vector Machine	0.850000	0.851553	0.848868
1	Decision Tree Classifier	0.795833	0.792585	0.794062
0	Logistic Regression	0.750000	0.795694	0.768826

Chapter 6: Clustering Analysis and Results (Cost Estimation)

4.1.1 K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm used to partition n observations into k clusters. The algorithm aims to minimize the variance within each cluster, effectively grouping similar data points together. In this project, the analysis focused on partitioning the patient data into **k=3** groups, and the optimal number of clusters (k) was determined using the **Elbow Method**.

Evaluation Metrics:

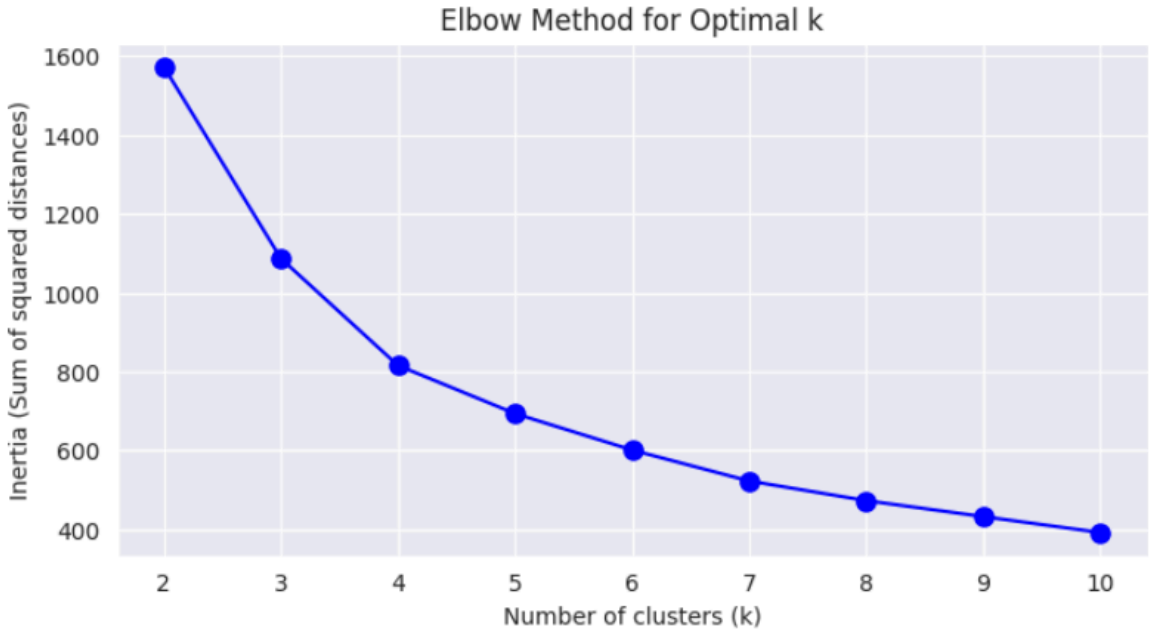
Technique Detail	Summary
Model Used	K-Means Clustering (k=3)
Optimization Method	Elbow Method for Optimal k
Evaluation Method	Silhouette Score

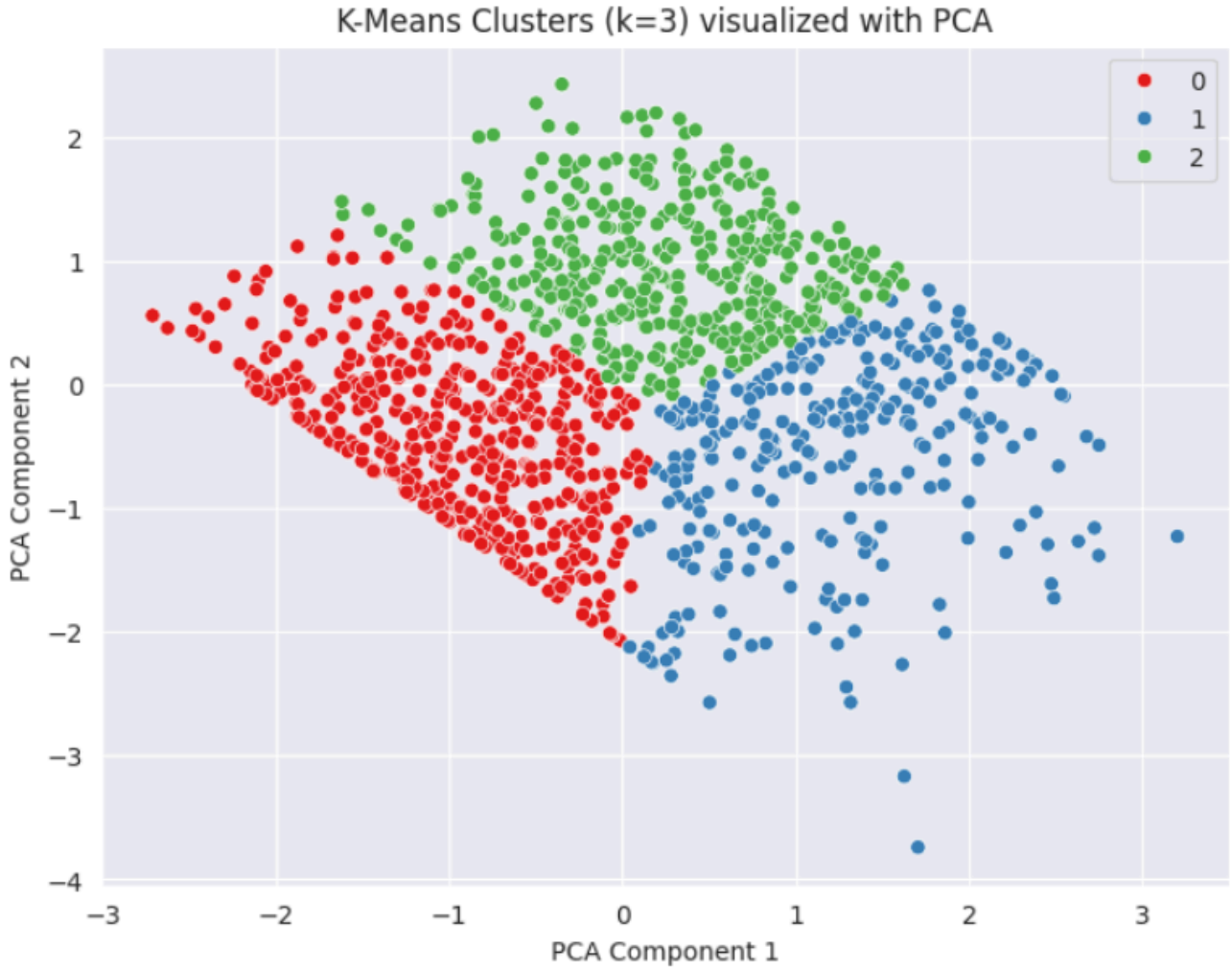
Result Interpretation:

The K-Means clustering algorithm successfully **grouped individuals into meaningful lifestyle or risk-based categories**. The quality and separation of these clusters were **confirmed using the Silhouette Score**.

Actual vs. Predicted Plot:

The **K-Means Clusters (k=3) visualized with PCA** plot shows distinct separation between the three groups (0, 1, 2). The **Elbow Method for Optimal k** plot displays the decrease in inertia as k increases, helping to identify the most suitable number of clusters.


Elbow Method for Optimal k
Inertia (Sum of squared distances)
Number of clusters (k)
1600
1400
1200
1000
800
600
400
2
3
4
5
6
7
8
9
10


K-Means Clusters (k=3) visualized with PCA
PCA Component 2
PCA Component 1
0
1
2
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3

Chapter 7: Integrated Findings and Future Work

7.1 Summary of Best Models

- Cost Estimation (Regression): ElasticNet Regression (Best R^2 Score: 0.5611).
- Risk Prediction (Classification): Random Forest Classifier (Best Accuracy: 0.9058).
- Patient Grouping (Clustering): K-Means ({k}=3) successfully identified distinct risk groups.

7.2 Recommendations

Based on the high correlation of Smoker status with high costs and risk, the primary recommendation is:

- Targeted Intervention: Focus on developing personalized support programs and policy incentives for smoking cessation, particularly for individuals categorized in the "High Risk" cluster.

7.3 Conclusion & Limitations

- Conclusion: The system successfully leverages integrated ML techniques to simultaneously predict health risk and associated costs, thereby promoting proactive awareness and data-driven decision-making in healthcare.
- Limitations: The system's generalizability is limited by the regional and specific nature of the training data. The R2 score (0.5611) suggests that of the variance in charges remains unexplained by the current feature set.

7.4 Future Directions

- Deployment: Implement the model via a user-friendly web application for public accessibility.
- Data Enrichment: Integrate additional features such as lab results, genetic data, or time-series symptom tracking for higher predictive accuracy.
- Model Enhancement: Explore advanced Deep Learning models (e.g., Neural Networks) for potential gains in prediction accuracy.