PROJECT REPORT

Medical Cost Prediction using Machine Learning Models

Submitted by: Sachin S

Registration No: 12321957

Programme: B.Tech (CSE/IT)

Section: KM003

Course Code: INT234

Under the Guidance of

(Dr. Mrinalini Rana, UID: 22138)

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

DECLARATION

I Sachin s, student at Lovely professional university (B.Tech) under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 15-12-2025 Signature

Registration No: 12321957 Sachin

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my project guide **Dr. Mrinalini Rana** for continuous support, guidance, and encouragement throughout the project. I am also thankful to the faculty members of Lovely Professional University for providing the necessary resources and learning environment. Lastly, I would like to thank my friends and family for their constant motivation.

TABLE OF CONTENT

1. INTRODUCTION

With the rapid growth of the healthcare sector, predicting medical expenses has become crucial for insurance companies, hospitals, and policy makers. Machine Learning plays a vital role in analyzing historical healthcare data and predicting future medical costs based on patient attributes.

The objective of this project is to perform **Exploratory Data Analysis (EDA)** and apply multiple **machine learning algorithms** to predict annual medical costs and classify patients into high-cost and low-cost categories.

---

2. SOURCE OF DATASET

The dataset used in this project is the **Medical Cost Prediction Dataset**, obtained from an open-source platform.

https://www.kaggle.com/datasets/miadul/medical-cost-predication-dataset

- **Type:** Structured dataset
- **Number of Records:** 5000
- **Target Variable:** Annual Medical Cost
- **Features:** Age, Gender, BMI, Smoking Habit, Health Conditions, Insurance Type, City Type, and Previous Medical Cost

  The dataset contains both numerical and categorical attributes suitable for regression and classification tasks.

---

3. EXPLORATORY DATA ANALYSIS (EDA) PROCESS

The EDA process was performed to understand the dataset and identify patterns before applying machine learning models. The following steps were carried out:

- Viewing dataset structure using head() and info()
- Analyzing statistical summary using describe()
- Visualizing categorical features using bar charts
- Visualizing numerical features using histograms
- Studying relationships between age and medical cost using scatter plots
- Analyzing correlations using heatmaps

EDA helped in identifying feature importance, data distribution, and potential outliers.

---

4. ANALYSIS ON DATASET

I. General Description

The dataset was preprocessed and analyzed using the following steps:

- Handling missing values using forward fill
- Encoding categorical variables using Label Encoding
- Feature scaling using StandardScaler
- Splitting data into training and testing sets

Both regression and classification approaches were applied to the dataset.

II. Regression Analysis

Linear Regression was used to predict the Annual Medical Cost.

Evaluation Metrics Used:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

- R^2 Score

The regression model showed a high R^2 score, indicating good prediction accuracy and strong relationship between features and medical cost.

A scatter plot of actual vs predicted values was used to visualize model performance.

III. Classification Analysis

A new target variable High_Cost was created by dividing patients into:

- High medical cost
- Low medical cost

The following classification algorithms were applied:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- Naive Bayes

Evaluation Metrics Used:

- Accuracy Score
- Confusion Matrix

A bar graph was created to visually compare the accuracy of all classification models.

IV. Visualization

The following visualizations were included in the analysis:

- Bar charts for Gender and Smoking Status
- Histogram for Age distribution

- Histogram for Annual Medical Cost distribution
- Scatter plot for Actual vs Predicted cost
- Bar chart comparing model accuracy

These visualizations made the analysis more interpretable and insightful.

---

5. CONCLUSION

This project successfully demonstrated the use of Machine Learning techniques for predicting medical costs and classifying patients based on healthcare expenses. Data preprocessing and EDA played a crucial role in improving model performance. Among the classification models, Decision Tree and Logistic Regression showed reliable results.

The study highlights the importance of data-driven decision-making in the healthcare domain.

---

6. FUTURE SCOPE

- Applying advanced regression techniques for improved accuracy
- Hyperparameter tuning of models
- Using larger real-world healthcare datasets
- Integrating the model into healthcare insurance systems

---

7. REFERENCES

1. Scikit-learn Documentation
2. Pandas and NumPy Documentation
3. Kaggle Medical Cost Dataset
4. Machine Learning course material

CODE

File Edit Format Run Options Window Help

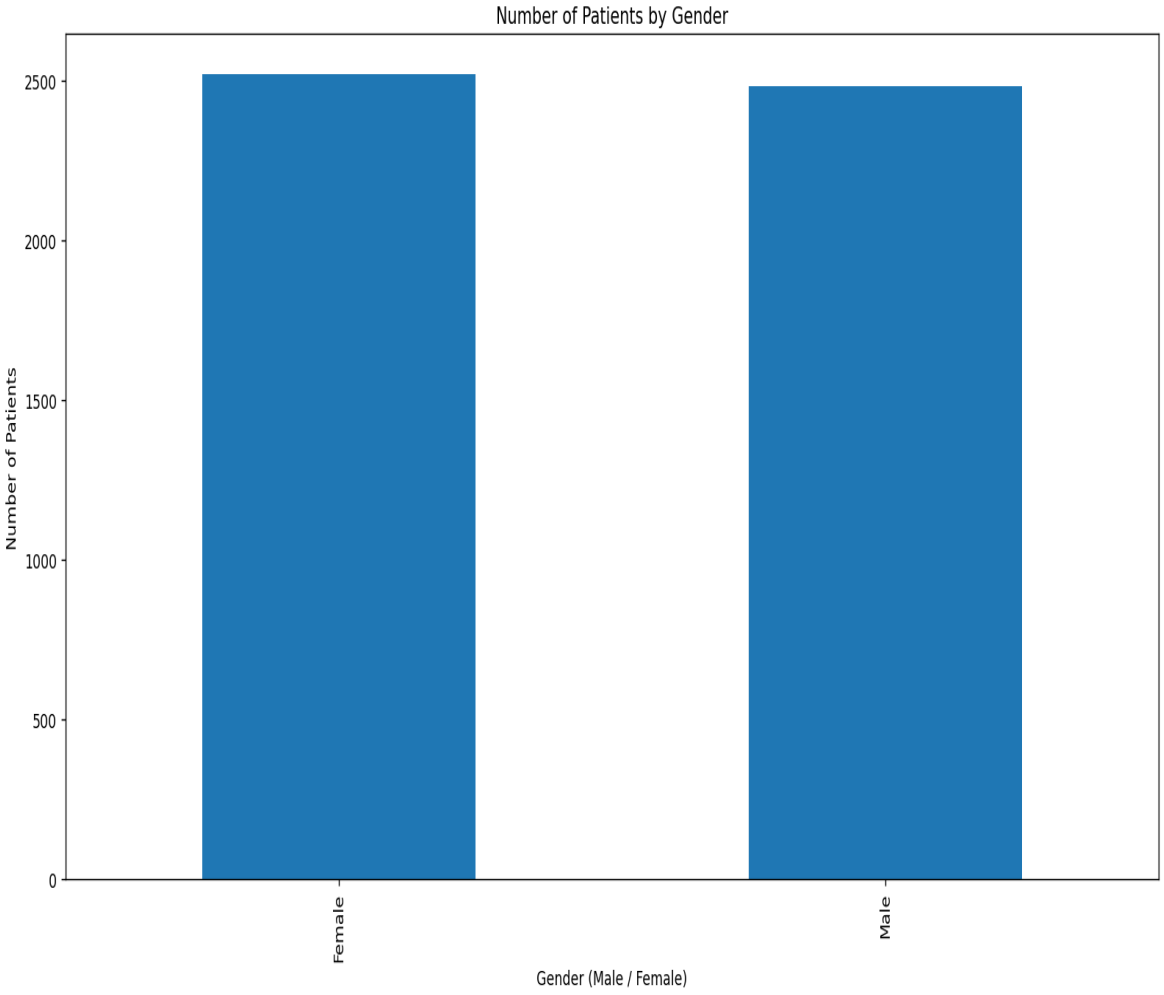
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB

from sklearn.metrics import (
    accuracy_score,
    confusion_matrix,
    mean_absolute_error,
    mean_squared_error,
    r2_score
)


# DATA LOADING

df = pd.read_csv(r"C:\Users\SACHIN\Downloads\archive (7)\medical_cost_prediction_dataset.csv")

print(df.head())
print(df.info())


# BASIC VISUALIZATION

df['gender'].value_counts().plot(kind='bar')
plt.title("Number of Patients by Gender")
plt.xlabel("Gender (Male / Female)")
plt.ylabel("Number of Patients")
plt.show()

df['smoker'].value_counts().plot(kind='bar')
plt.title("Smoking Status of Patients")
plt.xlabel("Smoking Habit (Yes / No)")
plt.ylabel("Number of Patients")
plt.show()

plt.hist(df['age'], bins=20)
plt.title("Age Distribution of Patients")
plt.xlabel("Age (Years)")
plt.ylabel("Number of Patients")
plt.show()

plt.hist(df['annual_medical_cost'], bins=20)
plt.title("Distribution of Annual Medical Cost")
plt.xlabel("Annual Medical Cost")
plt.ylabel("Number of Patients")
plt.show()
```

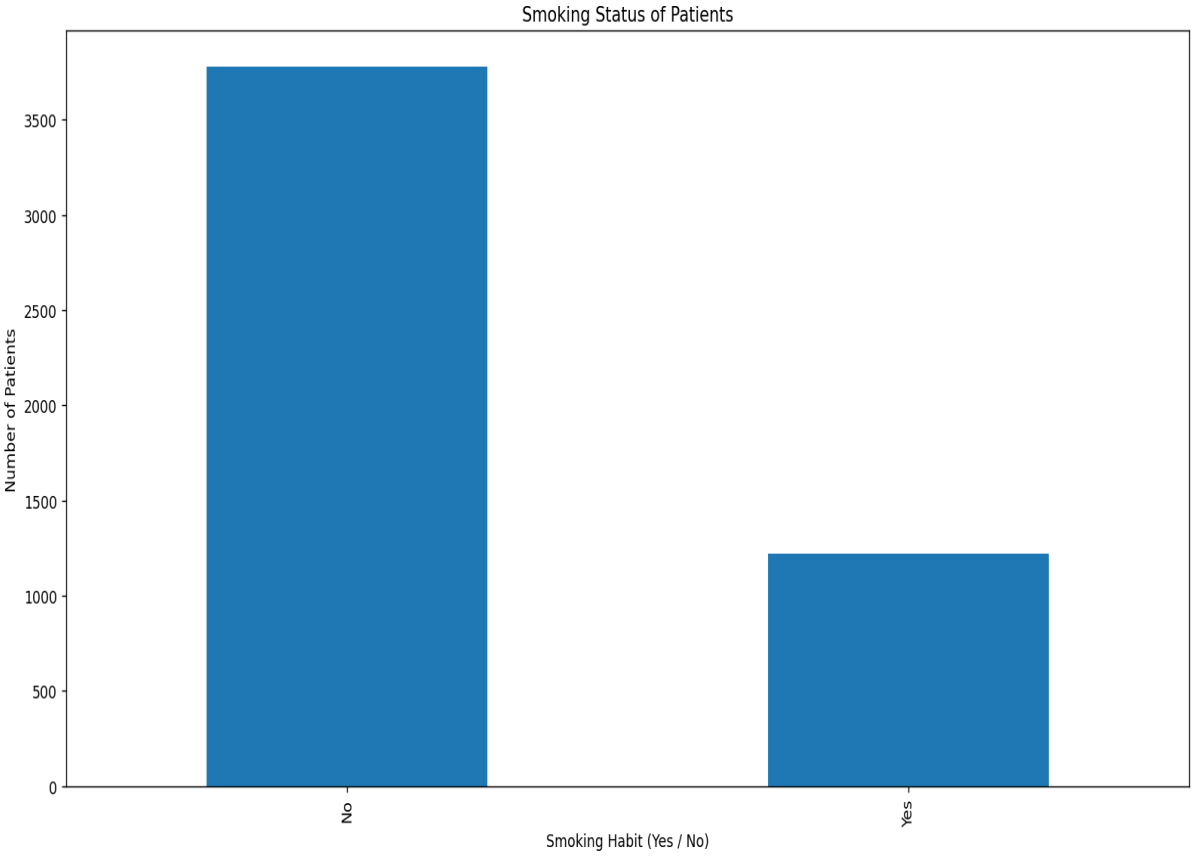
```
# PREPROCESSING

df.ffill(inplace=True)

le = LabelEncoder()
for col in df.select_dtypes(include='object').columns:
    df[col] = le.fit_transform(df[col])

X = df.drop('annual_medical_cost', axis=1)
y_reg = df['annual_medical_cost']

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)


# LINEAR REGRESSION

X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y_reg, test_size=0.2, random_state=42
)

lr = LinearRegression()
lr.fit(X_train, y_train)

y_pred = lr.predict(X_test)

print("\n--- LINEAR REGRESSION RESULTS ---")
print("Mean Absolute Error (Average Error):", mean_absolute_error(y_test, y_pred))
print("Mean Squared Error:", mean_squared_error(y_test, y_pred))
print("Root Mean Squared Error:", np.sqrt(mean_squared_error(y_test, y_pred)))
print("R2 Score (Model Accuracy):", r2_score(y_test, y_pred))

plt.scatter(y_test, y_pred)
plt.plot(
    [y_test.min(), y_test.max()],
    [y_test.min(), y_test.max()]
)
plt.title("Actual Cost vs Predicted Cost")
plt.xlabel("Actual Annual Medical Cost")
plt.ylabel("Predicted Annual Medical Cost")
plt.show()
```

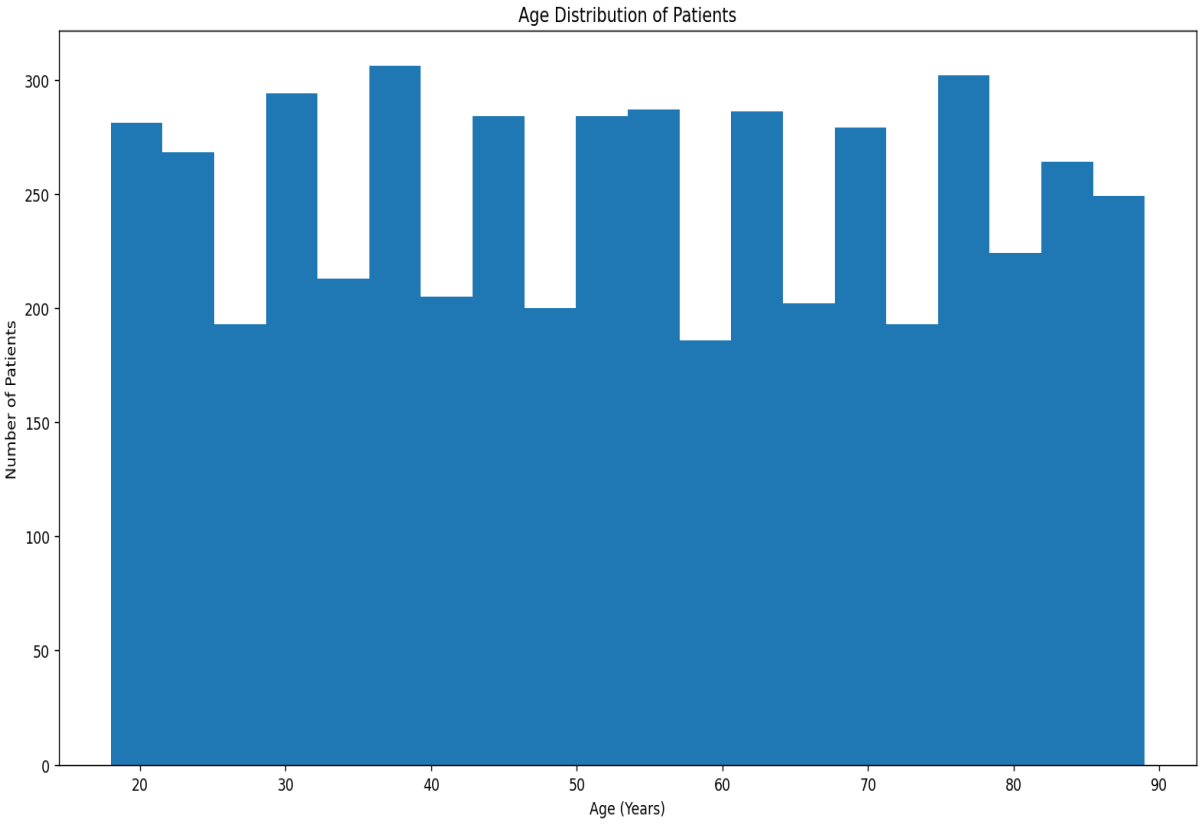
```
# CREATE CLASSIFICATION TARGET

median_cost = df['annual_medical_cost'].median()
df['High_Cost'] = (df['annual_medical_cost'] > median_cost).astype(int)

y_class = df['High_Cost']

X_train, X_test, y_train, y_test = train_test_split(
    X_scaled,
    y_class,
    test_size=0.2,
    random_state=42,
    stratify=y_class
)


#  LOGISTIC REGRESSION

log_reg = LogisticRegression(max_iter=1000)
log_reg.fit(X_train, y_train)
log_pred = log_reg.predict(X_test)

log_acc = accuracy_score(y_test, log_pred)
print("\nLogistic Regression Accuracy (High vs Low Cost):", log_acc)



# KNN

knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
knn_pred = knn.predict(X_test)

knn_acc = accuracy_score(y_test, knn_pred)
print("\nKNN Accuracy (High vs Low Cost):", knn_acc)
print("KNN Confusion Matrix:")
print(confusion_matrix(y_test, knn_pred))
```

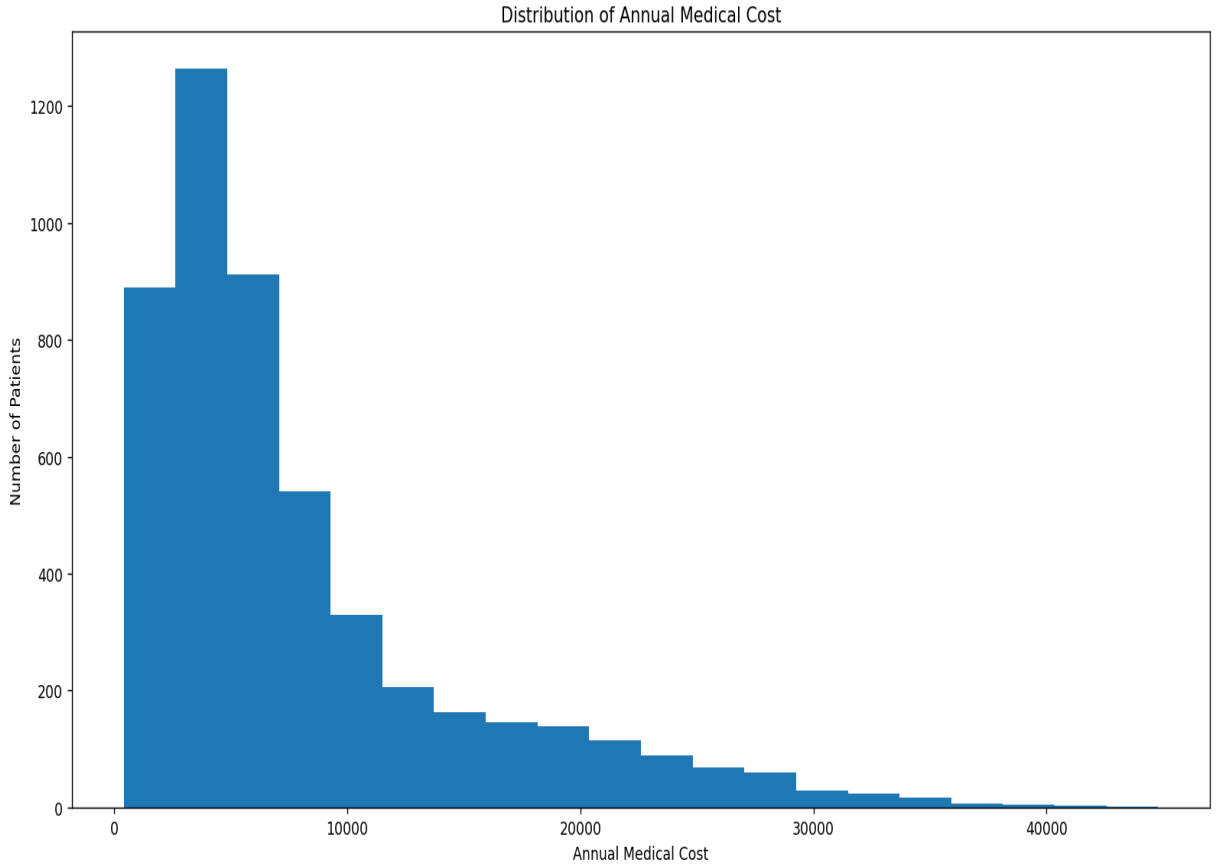
```
# DECISION TREE

dt = DecisionTreeClassifier(random_state=42)
dt.fit(X_train, y_train)
dt_pred = dt.predict(X_test)

dt_acc = accuracy_score(y_test, dt_pred)
print("\nDecision Tree Accuracy (High vs Low Cost):", dt_acc)
print("Decision Tree Confusion Matrix:")
print(confusion_matrix(y_test, dt_pred))



#  NAIVE BAYES

nb = GaussianNB()
nb.fit(X_train, y_train)
nb_pred = nb.predict(X_test)

nb_acc = accuracy_score(y_test, nb_pred)
print("\nNaive Bayes Accuracy (High vs Low Cost):", nb_acc)
print("Naive Bayes Confusion Matrix:")
print(confusion_matrix(y_test, nb_pred))



# ACCURACY COMPARISON BAR GRAPH

models = [
    'Logistic Regression',
    'KNN',
    'Decision Tree',
    'Naive Bayes'
]

accuracies = [
    log_acc,
    knn_acc,
    dt_acc,
    nb_acc
]

plt.bar(models, accuracies)
plt.title("Comparison of Classification Model Accuracy")
plt.xlabel("Machine Learning Models")
plt.ylabel("Accuracy (Higher Value = Better Model)")
plt.ylim(0, 1)

for i, acc in enumerate(accuracies):
    plt.text(i, acc + 0.01, f"{acc*100:.2f}%", ha='center')

plt.show()
```

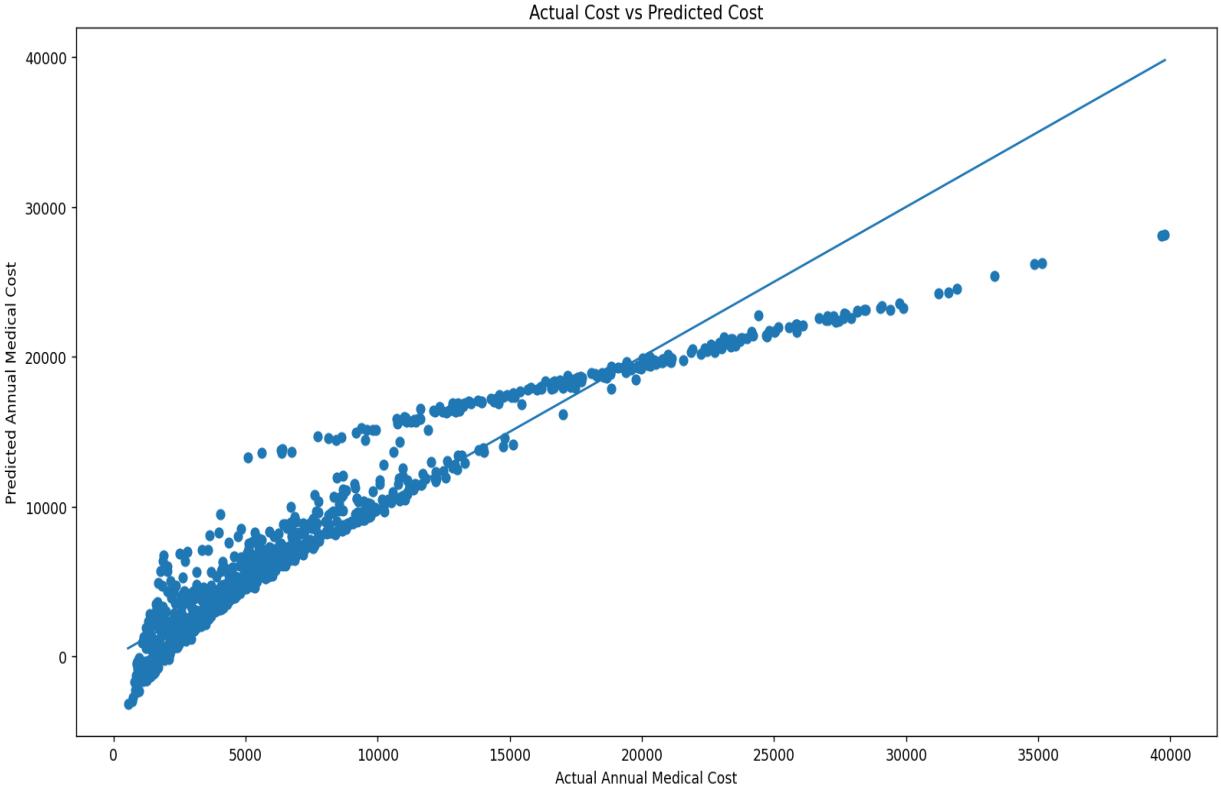
OUT PUT

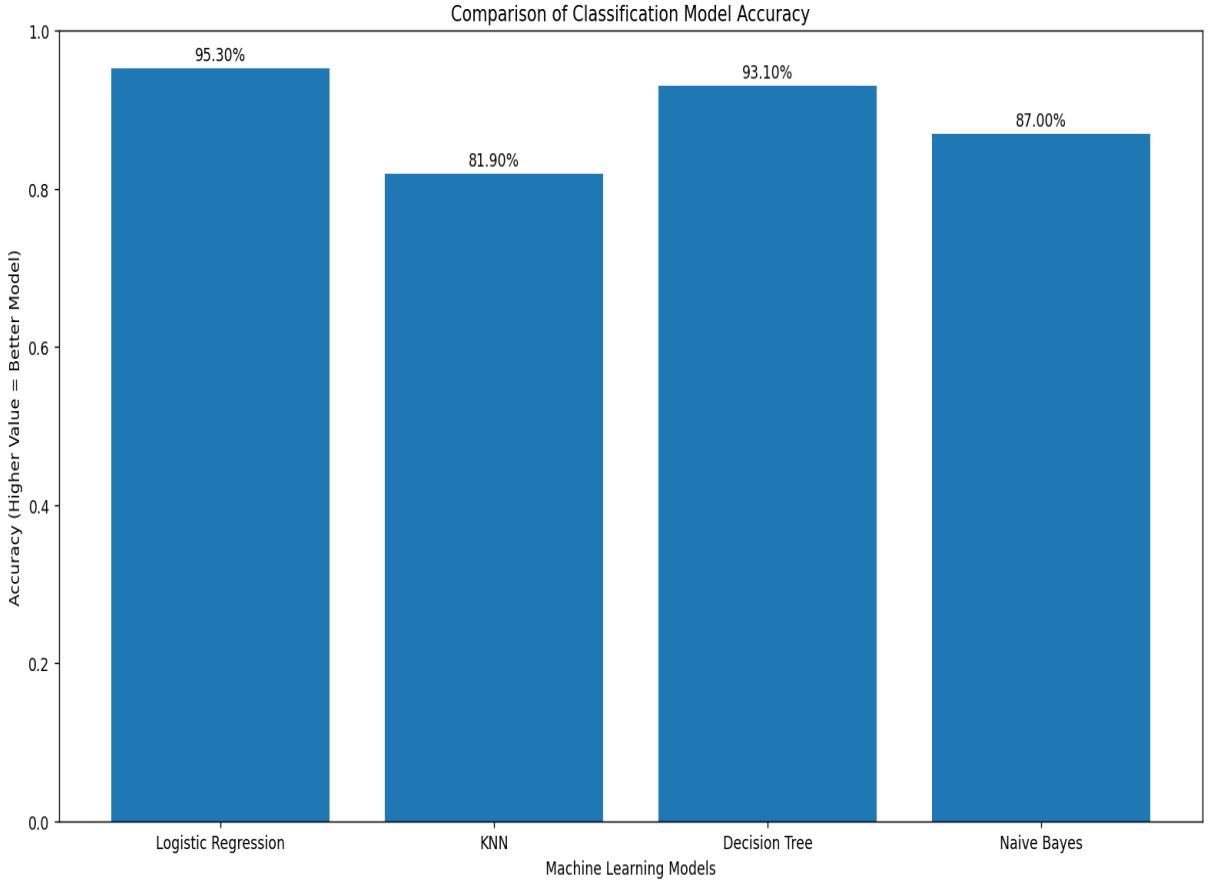

Number of Patients by Gender
Number of Patients
2500
2000
1500
1000
500
0
Female
Male
Gender (Male / Female)


Smoking Status of Patients
Number of Patients
0
500
1000
1500
2000
2500
3000
3500
No
Yes
Smoking Habit (Yes / No)


Age Distribution of Patients
Number of Patients
0
50
100
150
200
250
300
20
30
40
50
60
70
80
90
Age (Years)


Distribution of Annual Medical Cost
Number of Patients
Annual Medical Cost
0
200
400
600
800
1000
1200
0
10000
20000
30000
40000


Actual Cost vs Predicted Cost
Predicted Annual Medical Cost
Actual Annual Medical Cost
0
10000
20000
30000
40000
0
5000
10000
15000
20000
25000
30000
35000
40000


Comparison of Classification Model Accuracy
Accuracy (Higher Value = Better Model)
1.0
0.8
0.6
0.4
0.2
0.0
95.30%
81.90%
93.10%
87.00%
Logistic Regression
KNN
Decision Tree
Naive Bayes
Machine Learning Models

```
   age  gender   bmi  ...   city_type  previous_year_cost  annual_medical_cost
0   69    Male  29.4  ...  Semi-Urban               10885              2645.50
1   32  Female  22.9  ...  Semi-Urban               18722             10959.70
2   89    Male  25.7  ...       Urban                4196              8409.80
3   78    Male  31.9  ...       Urban               11128              7996.62
4   38    Male  27.7  ...       Urban               15110              3202.52

[5 rows x 20 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   age                     5000 non-null   int64
 1   gender                  5000 non-null   object
 2   bmi                     5000 non-null   float64
 3   smoker                  5000 non-null   object
 4   diabetes                5000 non-null   int64
 5   hypertension            5000 non-null   int64
 6   heart_disease           5000 non-null   int64
 7   asthma                  5000 non-null   int64
 8   physical_activity_level 5000 non-null   object
 9   daily_steps             5000 non-null   int64
 10  sleep_hours             5000 non-null   float64
 11  stress_level            5000 non-null   int64
 12  doctor_visits_per_year  5000 non-null   int64
 13  hospital_admissions     5000 non-null   int64
 14  medication_count        5000 non-null   int64
 15  insurance_type          3952 non-null   object
 16  insurance_coverage_pct  5000 non-null   int64
 17  city_type               5000 non-null   object
 18  previous_year_cost      5000 non-null   int64
 19  annual_medical_cost     5000 non-null   float64
dtypes: float64(3), int64(12), object(5)
memory usage: 781.4+ KB
None
```

```
--- LINEAR REGRESSION RESULTS ---
Mean Absolute Error (Average Error): 1386.1068882341842
Mean Squared Error: 4302052.38904888
Root Mean Squared Error: 2074.1389512394967
R2 Score (Model Accuracy): 0.9118006231098668

Logistic Regression Accuracy (High vs Low Cost): 0.953

KNN Accuracy (High vs Low Cost): 0.819
KNN Confusion Matrix:
[[428  72]
 [109 391]]

Decision Tree Accuracy (High vs Low Cost): 0.931
Decision Tree Confusion Matrix:
[[465  35]
 [ 34 466]]

Naive Bayes Accuracy (High vs Low Cost): 0.87
Naive Bayes Confusion Matrix:
[[450  50]
 [ 80 420]]
```