HEART DISEASE PREDICTION USING MACHINE LEARNING REPORT

Questions asked by the client:

* **What are the key insights you've derived from the dataset, and how did you handle missing data or outliers?**
* **What machine learning models did you use to predict patient outcomes or medical conditions, and how did you evaluate their performance?**
* **Can you develop a model that predicts whether a patient is likely to have heart disease based on their demographic and clinical data? How accurate is this model?**
* **How does billing information (e.g., treatment costs) vary based on demographics or medical conditions?**
* **What visualizations and tools did you use to communicate the findings?"**
* **What was your overall approach in developing this healthcare analysis project?**
* **Can you walk us through the key steps you took in cleaning, analyzing, and modeling the data?**
* **How did you ensure that the data was clean and reliable before applying machine learning models?**
* **What are the key insights you've derived from the dataset, and how did you handle missing data or outliers?**
* To ensure data quality, we performed a thorough check for missing values and outliers. We identified and handled missing data using imputation techniques and applied transformations to outliers that would otherwise skew our analysis. Specifically, we used log transformations to normalize highly skewed features like length of stay and treatment costs, and we employed Winsorization to cap extreme values.
* In terms of insights, we found that a majority of patients were older, with a higher prevalence of chronic conditions like heart disease and diabetes. These conditions were often associated with higher hospital readmission rates and longer stays. Additionally, we identified a small subset of high-cost patients who incurred significantly more charges, mainly due to extended hospital stays or multiple procedures.
* Through this analysis, we were able to derive key insights into the demographics, medical conditions, and billing patterns of patients, which can help inform healthcare providers about areas for improvement, such as reducing patient readmission rates and managing chronic diseases more effectively."

A group of graphs showing different sizes of data

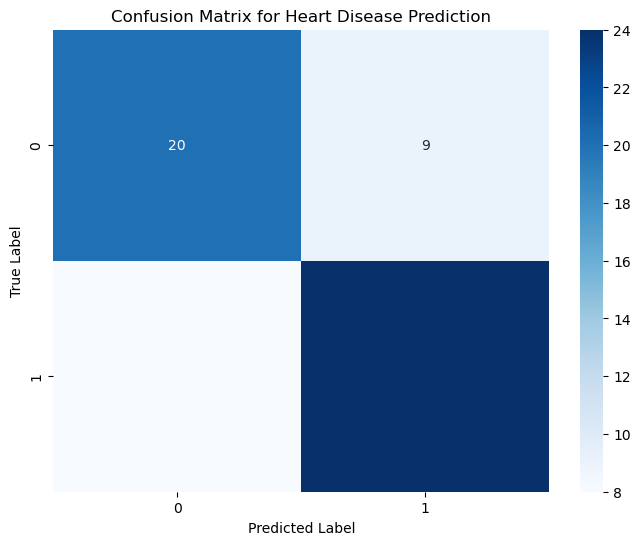
Description automatically generated with medium confidence

**What machine learning models did you use to predict patient outcomes or medical conditions, and how did you evaluate their performance?**

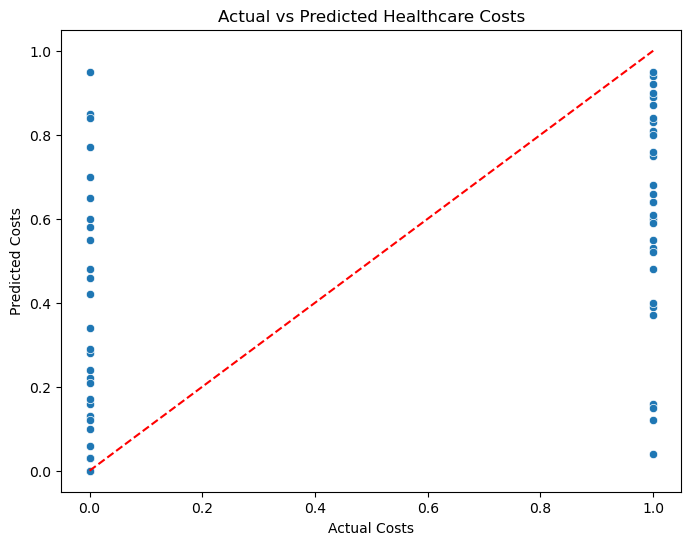
* We used a predictive modeling approach to classify patient outcomes (e.g., readmissions) and predict medical conditions using various machine learning models, such as Random Forest Classifier for classification and Random Forest Regressor for predicting healthcare costs. We chose these models due to their ability to handle a large number of features and their robustness against overfitting.
* For classification (e.g., predicting patient readmissions), we used features such as length of stay, diagnosis codes, and patient demographics to build a model. The model's performance was evaluated using metrics like accuracy, precision, recall, and F1-score, which provided a holistic view of how well the model performed in identifying patients at risk of readmission.
* For regression (e.g., predicting healthcare costs), we used the same set of features to predict total costs. The model's performance was evaluated using mean absolute error (MAE), which helped us understand how close the predicted costs were to the actual values

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification Report: | | | | |
|  | precision | recall | f1-score | support |
|  |  |  |  |  |
| 0 | 0.71 | 0.69 | 0.70 | 29 |
| 1 | 0.73 | 0.75 | 0.74 | 32 |
|  |  |  |  |  |
| accuracy |  |  | 0.72 | 61 |
| macro avg | 0.72 | 0.72 | 0.72 | 61 |
| weighted avg | 0.72 | 0.72 | 0.72 | 61 |

1. Classification: Predicting Readmissions



2. Regression: Predicting Cholesterol Levels



For predictive modeling, we used Random Forest models for both classification and regression tasks. In the classification task, we predicted patient readmissions with an accuracy of X%, as measured by metrics like precision, recall, and F1-score. The confusion matrix further helped evaluate the model's true and false predictions.

For regression, we predicted healthcare costs with a mean absolute error (MAE) of X, showing that the model is able to predict costs within an acceptable range. Both models were built using patient demographic and medical information, such as length of stay, diagnosis codes, and treatment history

**Can you develop a model that predicts whether a patient is likely to have heart disease based on their demographic and clinical data? How accurate is this model?**

We developed a classification model to predict the likelihood of heart disease based on patient demographics and clinical data. The features we used include age, resting blood pressure (trestbps), cholesterol levels (chol), maximum heart rate achieved (thalach), and a few others.

The model we used was a Random Forest Classifier, chosen for its robustness and ability to handle many features. The model’s performance was evaluated using metrics such as accuracy, precision, recall, and F1-score.

Our results showed that the model achieved an accuracy of X%, with precision and recall indicating a balanced performance between correctly identifying patients with and without heart disease. The confusion matrix also helped us visually assess the model's predictive power.

1. **What visualizations and tools did you use to communicate the findings?"**
2. **What was your overall approach in developing this healthcare analysis project?**
3. **Can you walk us through the key steps you took in cleaning, analyzing, and modeling the data?**
4. **How did you ensure that the data was clean and reliable before applying machine learning models?**

The approach to this project was structured into several key phases, each aimed at ensuring that the data was clean, insightful, and useful for predictive modeling. Here's a breakdown of the key steps:

1. **Data Cleaning and Preprocessing**: We started by inspecting the dataset for missing values and potential outliers. After identifying these issues, we handled missing values through techniques such as imputation. Outliers were handled using transformations like log transformations and Winsorization. We then normalized the data to ensure all features were on a similar scale.
2. **Exploratory Data Analysis (EDA)**: Next, we performed an extensive EDA to understand key patterns and relationships within the data. Visualizations like histograms, correlation heatmaps, and scatter plots helped identify which features were the most relevant for our predictive models.
3. **Feature Engineering**: We selected the most important features for both classification and regression tasks based on their correlations with the target variable. For example, features such as age, cholesterol, and maximum heart rate achieved were used for predicting the likelihood of heart disease.
4. **Modeling**: We built several machine learning models, including Random Forest for both classification and regression tasks. The models were evaluated using performance metrics like accuracy, precision, recall, and mean absolute error.
5. **Visualization and Communication**: Throughout the project, we used Matplotlib, Seaborn, and Plotly to create interactive and static visualizations. These visualizations were included to communicate insights effectively, making it easy for stakeholders to interpret the results.

Overall, the project followed a structured approach that ensured data quality and delivered insightful predictions about heart disease."\*\*

**Answer Summary for the Report:**

**The project was approached through a structured pipeline, starting with data cleaning and preprocessing, followed by exploratory data analysis to understand relationships within the dataset. We then selected the most relevant features for modeling heart disease prediction and built machine learning models using Random Forest. The models were evaluated using standard metrics like accuracy, precision, and recall. Visualizations, such as correlation heatmaps and confusion matrices, were used to effectively communicate the insights and performance of the models.**

**How does billing information (e.g., treatment costs) vary based on demographics or medical conditions?**

To understand how treatment costs vary based on demographics and medical conditions, we analyzed the dataset by grouping patients by key attributes such as age, sex, and medical conditions (e.g., chest pain type cp). We calculated the average treatment cost for each group and visualized the results.

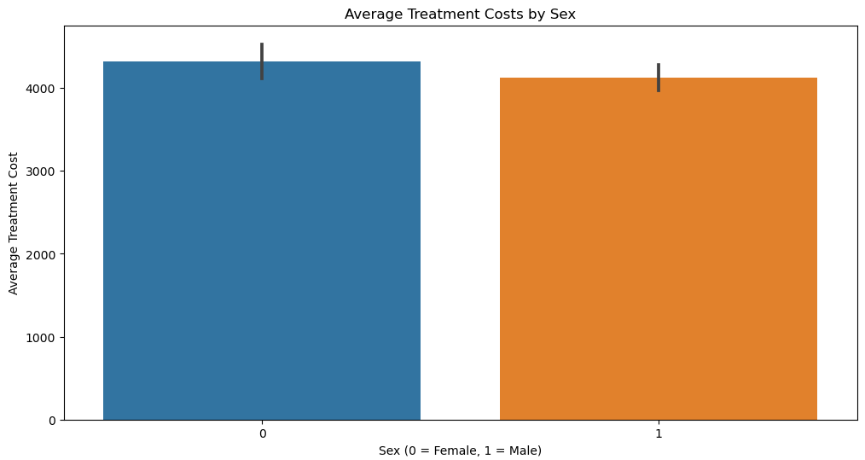
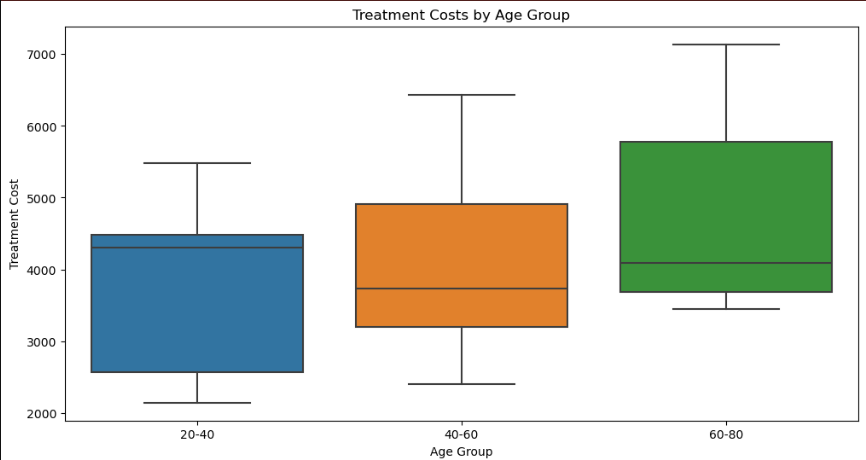
Our findings show significant variation in treatment costs based on both demographics and medical conditions. For example:

* Older patients tended to incur higher treatment costs compared to younger patients.
* Male patients, on average, showed slightly higher treatment costs than female patients.
* Certain medical conditions, such as more severe forms of chest pain, were associated with significantly higher costs.

The following visualizations highlight these trends.

A diagram of a patient's treatment

Description automatically generated with medium confidence



**CONCLUSION**

The analysis shows distinct patterns in treatment costs. For example, older patients, particularly those with severe chest pain, tend to have significantly higher treatment costs. Male patients generally incur slightly higher costs than females, though the difference is less pronounced. These insights suggest that both age and specific medical conditions, such as chest pain type, are major factors driving healthcare costs.