ANALYZING MACHINE LEARNING ALGORITHMS IN THE CONTEXT OF HEALTHCARE

Introduction

Machine Learning (ML), a subset of Artificial Intelligence, has grown in popularity over the years and has found its applications in multiple industries such as healthcare, robotics and more. Machine Learning focuses on the development of algorithms and statistical models, which enable computers to learn and improve their performance on a specific task without being explicitly programmed for that task. Algorithms analyze this data to identify patterns, relationships, and trends that can be used to make predictions or decisions in new, unseen data.

Machine learning has a vast number of applications in the healthcare domain, as it has shown promising results in diagnosis, personalized treatment plans, efficient data analysis, and improved patient outcomes.

Medical data is being generated every second, this data when processed and applied with algorithms can provide valuable insights into patient health, disease patterns, and treatment outcomes. By using Machine Learning algorithms, healthcare professionals can improve decision-making, overall enhancing patient health.

In this report, we will be discussing the different algorithms of Machine Learning used in healthcare, datasets employed, exploratory data analysis, the pros and cons of applying ML in the healthcare domain, and its promising future implications.

Exploratory Data Analysis (EDA) | Data Visualization:

For our report, we have used the **Heart Disease Cleveland UCI** dataset from Kaggle. This dataset can provide information such as the presence or absence of heart disease. It includes independent variables such as

- age
- sex (1 = male; 0 = female)
- cp chest pain type (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
- trestbps (resting blood pressure)
- chol (cholesterol level)
- fbs (fasting blood sugar > 120 mg/dl (1 = true; 0 = false))
- restecg (resting electrocardiographic results, 0: normal, 1: ST-T wave abnormality, 2: probable or definite left ventricular hypertrophy))

- thalach (maximum heart rate achieved)
- exang (exercise induced angina, (1 = yes; 0 = no))
- oldpeak (ST depression induced)
- slope (slope of the peak exercise ST segment, 1: upsloping, 2: flat, 3: downsloping))
- ca (number of major vessels 0-3)
- thal (thalassemia, 3 = normal; 6 = fixed defect; 7 = reversible defect))

and dependent/target variable

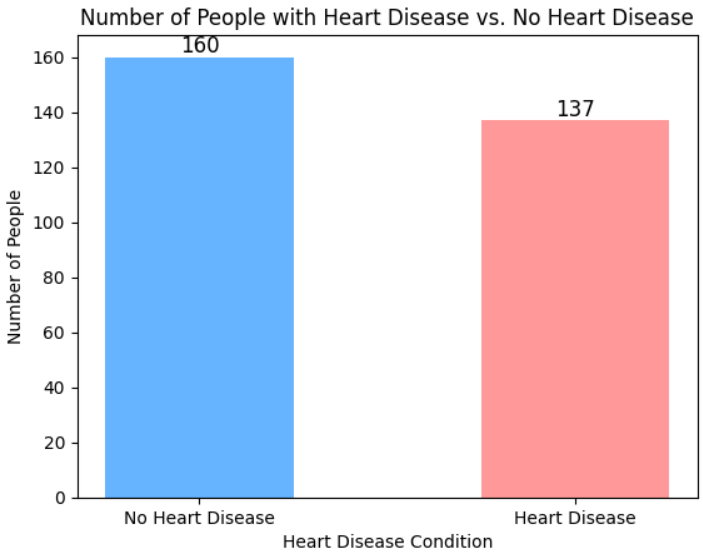
- condition (1: presence of heart disease, 0: absence of heart disease)

We first study the dataset and conduct EDA on the various features and their relationships in order to better understand the data before applying any machine learning algorithms to it.

We used a variety of data visualization charts, including scatter plots, correlation matrix heatmaps, box plots, stacked bar charts, and bar charts.

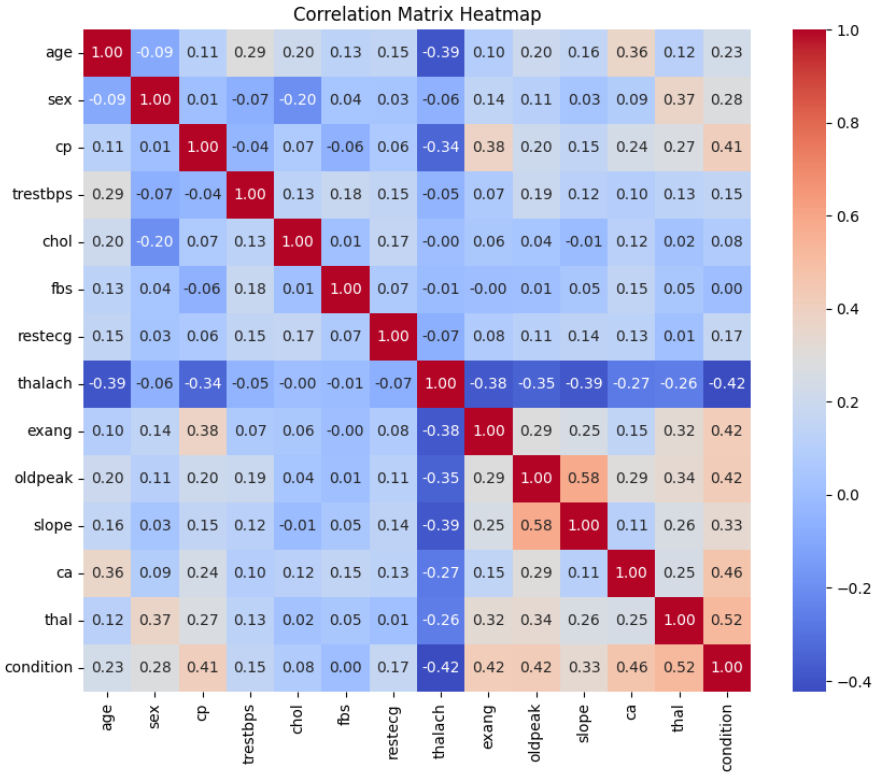
It must be noted that the dataset is quite old (1988).

1) **No. of people with heart disease vs No. of people without heart disease:**



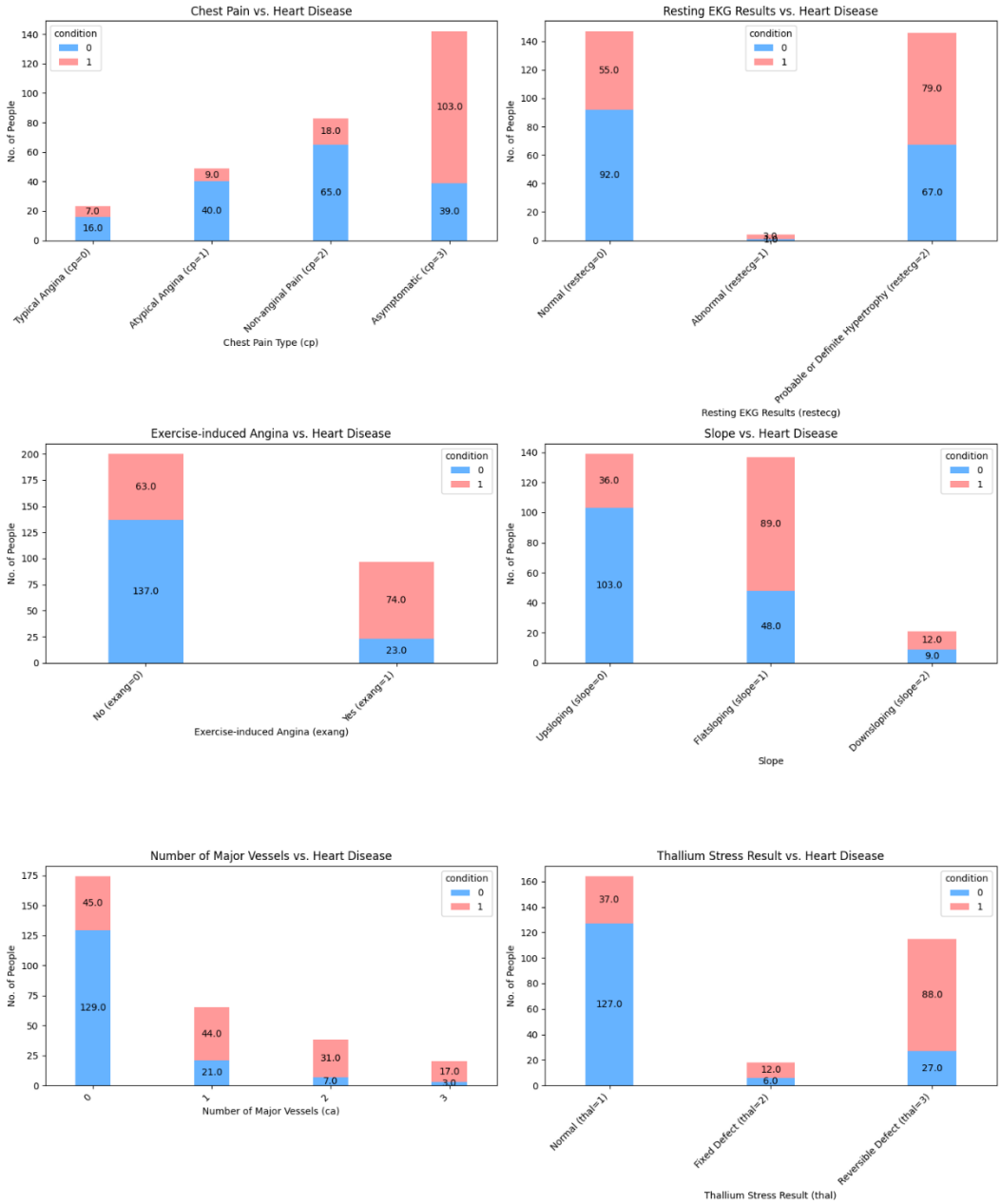
Our data is balanced because there are 137 individuals without heart disease and 160 individuals with heart disease.

2) **Correlation Matrix Heatmap:**



The target variable shows a weak correlation with fbs and chol, while all other variables exhibit a strong and significant correlation with the target.

3) Categorical Values Histogram:



a) Chest Pain (cp) vs. Heart Disease: People with chest pain types 1, 2, and 3 (Atypical Angina, Non-anginal Pain, and Asymptomatic) are more likely to have heart disease compared to those with chest pain type 0 (Typical Angina).

b) Resting EKG Results (restecg) vs. Heart Disease: People with a resting EKG result of 1 (Abnormal - reporting an abnormal heart rhythm) are more likely to have heart disease than those with a resting EKG result of 0 (Normal) or 2 (Probable or Definite Hypertrophy).

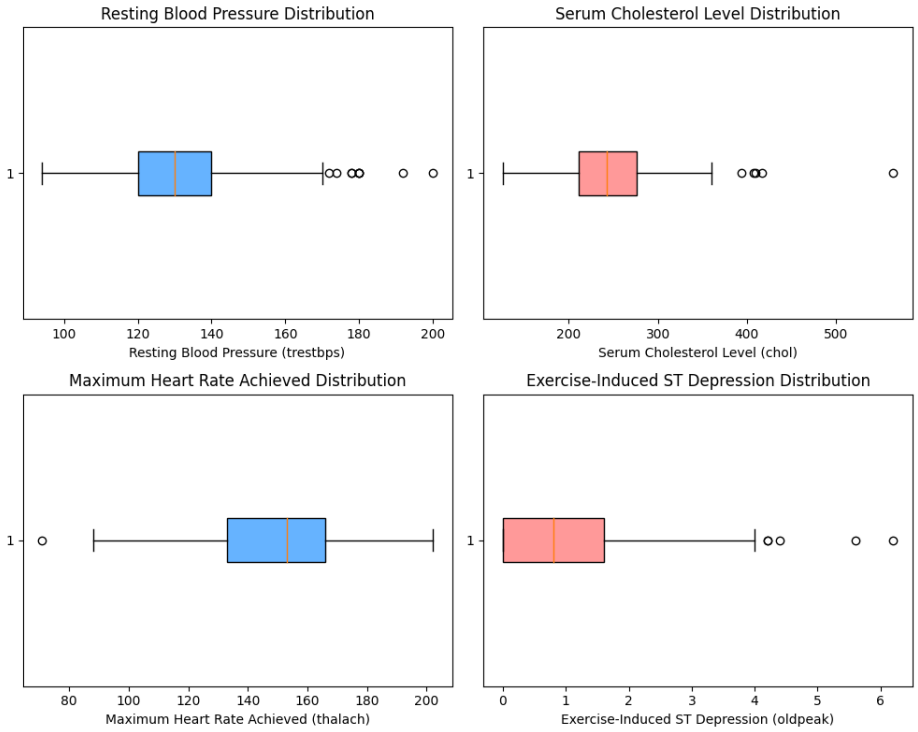
c) Exercise-induced Angina (exang) vs. Heart Disease: People with a value of 0 (no angina induced by exercise) have a higher likelihood of having heart disease compared to those with a value of 1 (angina induced by exercise).

d) Slope vs. Heart Disease: People with a slope value of 2 (Downsloping - signs of an unhealthy heart) are more likely to have heart disease than those with a slope value of 0 (Upsloping - best heart rate with exercise) or 1 (Flatsloping - minimal change, typical healthy heart).

e) Number of Major Vessels (ca) vs. Heart Disease: The more major blood vessels stained by fluoroscopy (ca), the less likely a person is to have heart disease. People with ca equal to 0 have a higher likelihood of having heart disease.

f) Thallium Stress Result (thal) vs. Heart Disease: People with a thallium stress result of 2 (Fixed Defect - once was a defect but now corrected) are more likely to have heart disease compared to those with a thallium stress result of 1 (Normal) or 3 (Reversible Defect).

4) Continuous Values Box Plot:



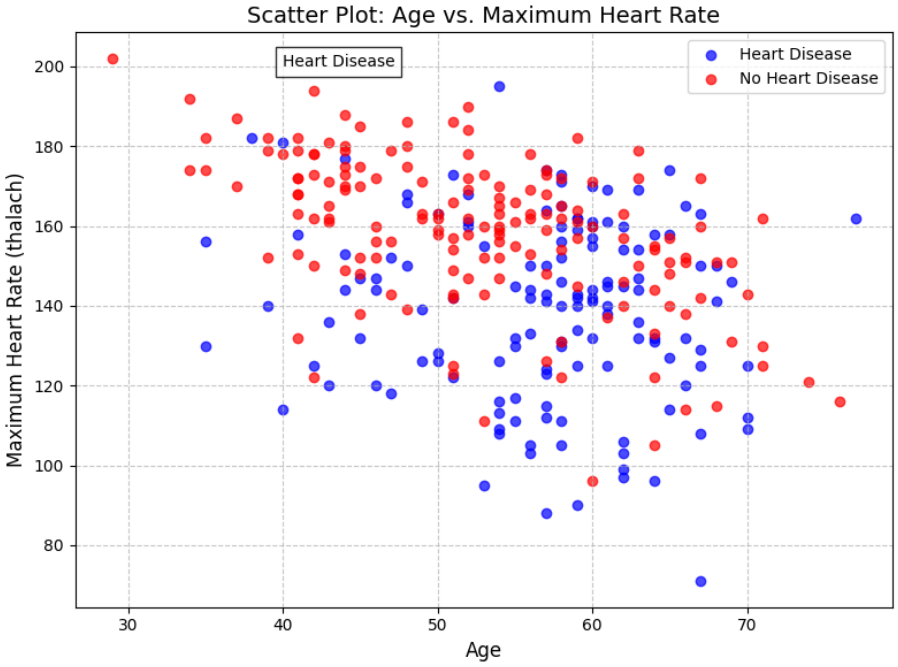
a) Resting Blood Pressure (trestbps): Anything above 120-140 is of concern.

b) Serum Cholesterol Level (chol): Greater than 200 is of concern.

c) Maximum Heart Rate Achieved (thalach): People with a maximum heart rate over 140 are more likely to have heart disease.

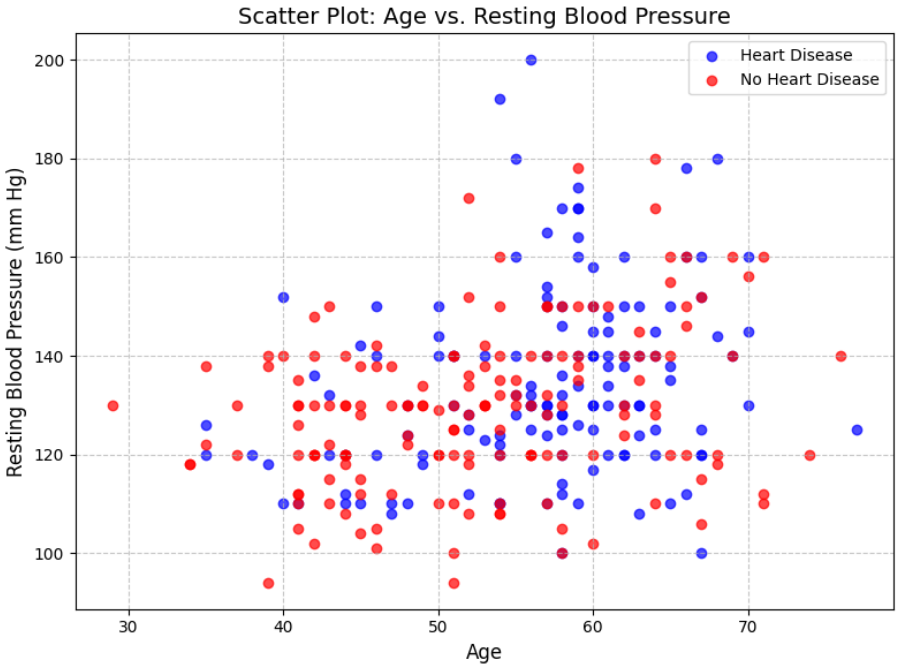
d) Exercise-Induced ST Depression (oldpeak): An unhealthy heart will stress more during exercise.

5) Scatter Plot (Age and Max Heart Rate Relation):



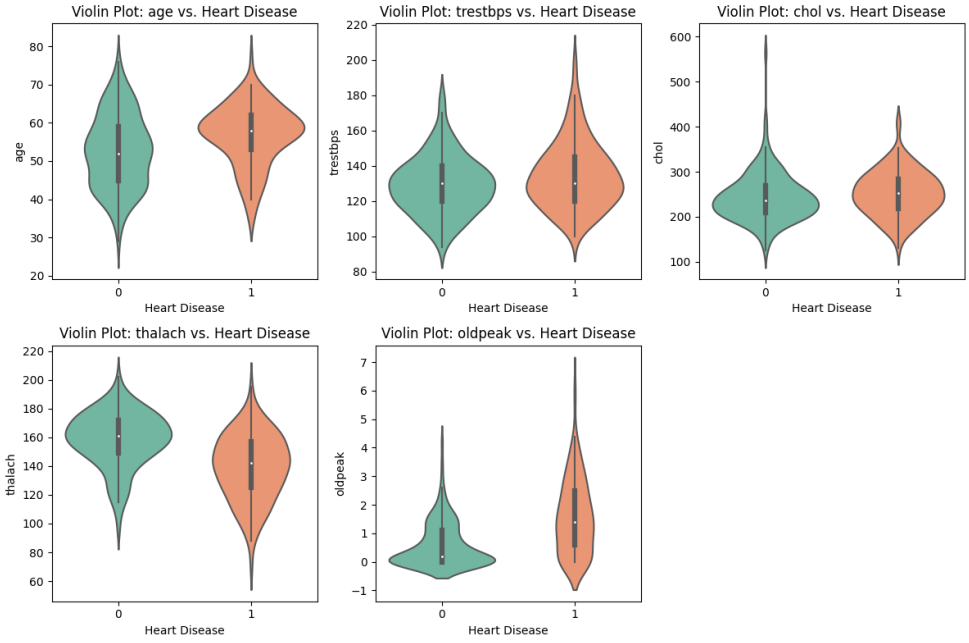
The blue data points represent individuals with heart disease (condition=1), while the red data points represent individuals without heart disease (condition=0). This scatter plot also helps us identify the outliers.

6) Scatter Plot (Age and Resting Blood Pressure Relation):



Resting Blood Pressure exhibits a positive correlation with the target variable, though the correlation strength is not notably strong. Additionally, the Resting Blood Pressure variable displays a few outliers that deviate from the main distribution of data points.

7) Violin Plots



a) Age vs. Heart Disease: The distribution of ages is relatively similar for individuals with and without heart disease. The highest concentration of heart disease cases appears to be among individuals aged 50 to 70, but heart disease cases are still present across different age groups.

b) Resting Blood Pressure vs. Heart Disease: The median resting blood pressure is slightly higher for individuals with heart disease compared to those without. The distribution of resting blood pressure for heart disease cases has a slightly wider spread, indicating increased variability compared to non-heart disease cases. Few outliers are observed in the higher blood pressure range for heart disease cases, suggesting some individuals with heart disease may have elevated resting blood pressure.

c) Serum Cholesterol vs. Heart Disease: Individuals with heart disease tend to have higher serum cholesterol levels compared to those without heart disease. The distribution of serum cholesterol for heart disease cases has a more extended tail, indicating some individuals with heart disease have significantly elevated cholesterol levels.

d) Maximum Heart Rate vs. Heart Disease: Individuals without heart disease tend to have slightly higher maximum heart rates compared to those with heart disease. The distribution of maximum heart rates for heart disease cases is narrower, indicating that individuals with heart disease may have lower maximum heart rates on average.

e) ST Depression vs. Heart Disease: ST depression induced by exercise relative to rest appears to be higher for individuals with heart disease. The distribution of ST depression is wider for heart disease cases, indicating greater variability in ST depression values among individuals with heart disease.

Machine Learning Algorithms in Healthcare:

Many ML algorithms are used in healthcare, some of these include:

a) **Supervised Learning Algorithms:** Supervised learning algorithms are trained on labeled datasets, meaning that the dataset contains input features and their corresponding output labels, and the overall goal is to learn to map input features to output labels, so that the algorithm can make accurate predictions on new, unknown data.

Examples of supervised learning algorithms are linear regression, logistic regression, decision trees, random forests, and neural networks.

Linear Regression Algorithm: Linear regression is one of the most commonly used machine learning algorithms used to predict numerical values. This model is considered linear because it assumes that the relationship between the dependent and independent variables is a straight line.

Linear regression is defined mathematically by the following equation:

$$y = mx + b$$

where,

- y is the dependent variable
- x is the independent variable
- m is the slope of the line
- b is the y-intercept

The slope of the line, m, tells us how much the dependent variable changes when the independent variable changes. The y-intercept, b, tells us the value of the dependent variable when the independent variable is 0. In healthcare applications, linear regression can be used for early identification of diseases and personalized treatment plans.

b) **Unsupervised Learning Algorithms**: In unsupervised learning, the models are trained on unlabeled datasets meaning there are no corresponding output labels involved. The overall goal here is to find patterns and structures in the data without any explicit guidance or supervision.

Applying ML algorithm to Healthcare Data:

We used linear regression to predict the presence or absence of heart disease, and explored which features had a greater influence on predicting the presence of heart disease.

Code for applying linear regression on the "Heart Disease UCI" dataset:

Step 1: Import necessary libraries.

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
```

Step 2: Load the dataset.

```
# Load the dataset
df = pd.read_csv('heart.csv')
```

Step 3: Split the data into features (X) and target (y)

```
# Split the data into features (X) and target (y)
X = df.drop('condition', axis=1)
y = df['condition']  # Target variable (heart disease presence: 0 or 1)
```

Step 4: Split the data into a training set and a test set.

```
# Split the data into a training set and a test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Step 5: Create a linear regression model.

```
# Create a linear regression model
model = LinearRegression()
```

Step 6: Fit the model to the training data.

```
# Fit the model to the training data
model.fit(X_train, y_train)
```

Step 7: Make predictions on the test set.

```
# Make predictions on the test set
y_pred = model.predict(X_test)
```

Step 8: Calculate Root Mean Squared Error and R-squared score

```
# Calculate the Root Mean Squared Error (RMSE)
rmse = mean_squared_error(y_test, y_pred, squared=False)

# Calculate the R-squared score
r2 = r2_score(y_test, y_pred)
```

Step 9: Print RMSE and R-squared score

```
# Print RMSE and R-squared separately in the output
print("Root Mean Squared Error (RMSE):", rmse)
print("R-squared Score:", r2)
```

Step 10: Plot the actual target values vs. the predicted values.

```
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, color='blue', alpha=0.7)
plt.plot([0, 1], [0, 1], color='red', linestyle='--', linewidth=2)
plt.xlabel('Actual Target Values', fontsize=12)
plt.ylabel('Predicted Values', fontsize=12)
plt.title('Linear Regression: Actual vs. Predicted Values', fontsize=14)
```

```
plt.text(0.1, 0.9, f'RMSE: {rmse:.3f}', fontsize=12, transform=plt.gca().transAxes)
plt.text(0.1, 0.85, f'R-squared: {r2:.3f}', fontsize=12, transform=plt.gca().transAxes)

plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

Results

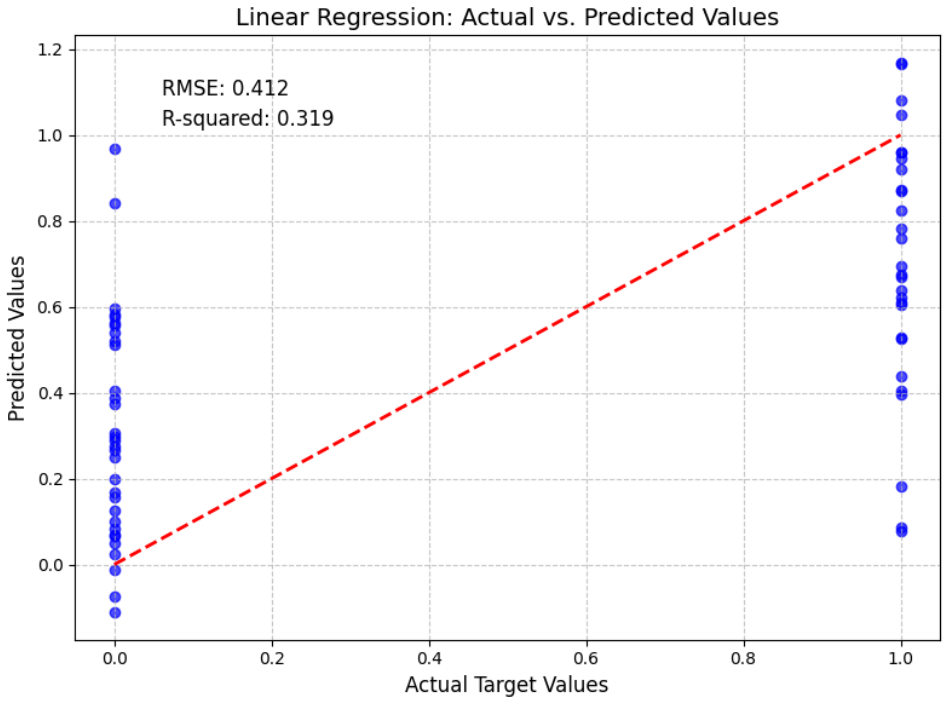
By applying the linear regression algorithm to our model, we were able to predict the presence or absence of heart disease.

The following results were obtained:

- Root Mean Squared Error (RMSE): 0.41173351381910284
- R-squared Score: 0.31887483142123774

Root Mean Squared Error (RMSE) tells us the average prediction error between the model's predicted values and the actual values. In our case, the RMSE was found to be 0.411, which shows that the values predicted by the model are close to the true values.

The R-squared Score tells us how well the model's predictions explain the variance in the target variable. The R-squared value ranges between 0 and 1, where 1 indicates a perfect fit. In our case, the R-squared Score came out to be 0.318 which means that using the features of this model we can explain 31.8% of the variation in the heart disease condition.



We also obtained a scatter plot for visualizing the relationship between the actual and predicted values from the model. The x-axis shows the actual values (condition: 0 or 1), and the y-axis shows the predicted values.

The red diagonal line represents the ideal condition where actual and predicted values are perfectly aligned. The blue data points show the actual vs. predicted values for each data point. The closer the data points are to the diagonal line, the better the model's prediction in alignment with the actual values, this shows higher accuracy of the model. Whereas the further away the values are from the diagonal line, the lower the accuracy of the model.

Pros and Cons of ML in Healthcare

Following are some of the pros and cons of Machine Learning in the context of healthcare applications:

Pros:

- ✓ Enhanced Diagnosis and Prediction
- ✓ Personalized Treatment Plans
- ✓ Efficient Drug Discovery
- ✓ Data-Driven Decision Making

Cons:

- ✗ Data Privacy and Security Concerns
- ✗ Data Quality and Bias
- ✗ Ethical Concerns
- ✗ Human-Machine Interaction

Future Innovations in Healthcare

With the rise in technology, machine learning is expected to further grow over the years enhancing healthcare applications. It has great potential to improve patient outcomes with the help of early disease detection, precision medicine, ML integration with IoT devices, and with the help of AI powered healthcare virtual assistants.

Conclusion

To sum up, ML has brought significant advancements in the field of healthcare. Machine Learning algorithms have proved their efficacy and potential in different healthcare

applications ranging from disease diagnosis and personalized treatment plans to drug discovery.

ML in healthcare has a promising future and is likely to open doors to newer opportunities in research. Moreover, it will automate processes improving overall efficiency in healthcare administration as well. At the same time, it is important to face the challenges posed by machine learning and come up with solutions accordingly.

Finally, as ML in the domain of healthcare continues to evolve, we will in no time see the healthcare landscape be reshaped with the help of powerful tools to deliver more precise, personalized, and efficient care.

Sources

Dataset: https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci

Linear Regression Code: https://colab.research.google.com/drive/12uZir_Fy4NLVqSDQrcmHoPujLsAJLTR-?usp=sharing

Codes for Data Viz: https://colab.research.google.com/drive/1lPyadaF5AN5UIKIJvTkGjEEODFmzbCgC?usp=sharing