**Assignment no 1**

**For**

**Machine Learning**

**2022-2026**

**Submit To: Engr. Ahmed Khawaja**

**Submit By: 2022-SE-06**



**The University of Azad Jammu and Kashmir Muzaffarabad**

**Department of Software Engineering**

**Fifth Semester**

[Figure 1 4](#_Toc190279117)

[Figure 2 4](#_Toc190279118)

[Figure 3 4](#_Toc190279119)

[Figure 4 5](#_Toc190279120)

[Figure 5 5](#_Toc190279121)

[Figure 6 5](#_Toc190279122)

[Figure 7 6](#_Toc190279123)

[Figure 8 6](#_Toc190279124)

[Figure 9 7](#_Toc190279125)

[Figure 10 8](#_Toc190279126)

[Figure 11 8](#_Toc190279127)

[Figure 12 9](#_Toc190279128)

[Figure 13 9](#_Toc190279129)

**Data Preprocessing and Analysis Report**

**Abstract**

This report presents the preprocessing and analysis of a given dataset. The primary objective is to clean the data, handle missing values, and perform exploratory data analysis (EDA) with visualizations. The dataset is preprocessed to remove inconsistencies, fill missing values, and prepare it for further modeling. The results demonstrate the impact of preprocessing on dataset quality and integrity.

**Introduction**

Data preprocessing is a critical step in any data science workflow. Raw data often contains missing values, inconsistencies, and noise that must be addressed before analysis. This report focuses on data cleaning, handling missing values, and conducting exploratory data analysis (EDA) on the provided dataset. Various visualization techniques are employed to interpret the dataset effectively.

**Methodology**

The following steps were followed to preprocess the dataset:

1. **Loading Data:** Read the dataset using pandas.
2. **Handling Missing Values:** Identify missing values and apply appropriate techniques (e.g., dropping high-missing-value columns, filling with median/mode).
3. **Categorical Data Handling:** Convert categorical data where necessary.
4. **Duplicate Removal:** Identify and remove duplicate records.
5. **Exploratory Data Analysis (EDA):** Perform visual analysis using histograms, boxplots, and correlation heatmaps.

**Step 1: Loading the Dataset**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset

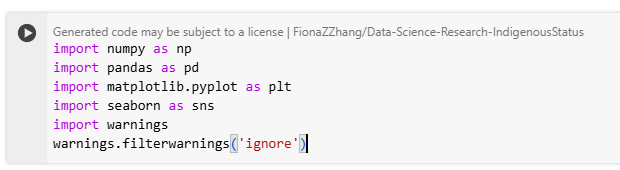
file\_path = "train.csv"

df = pd.read\_csv(file\_path)

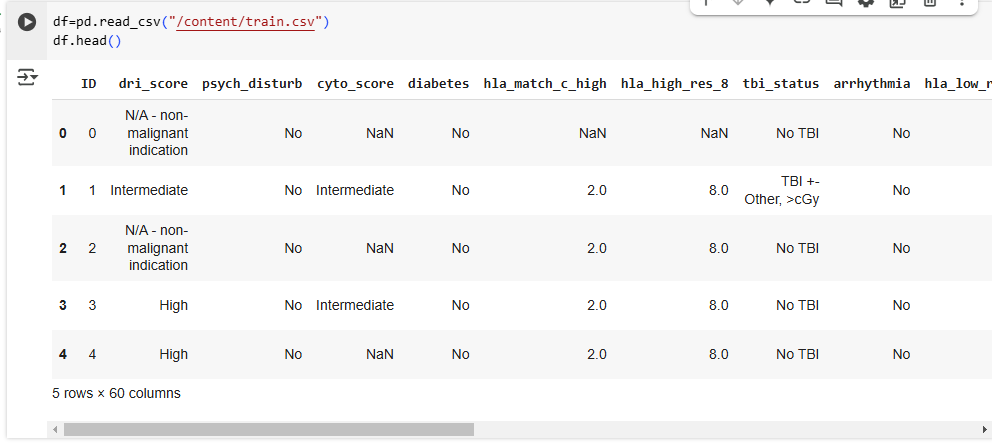
# Display basic information

df.head()

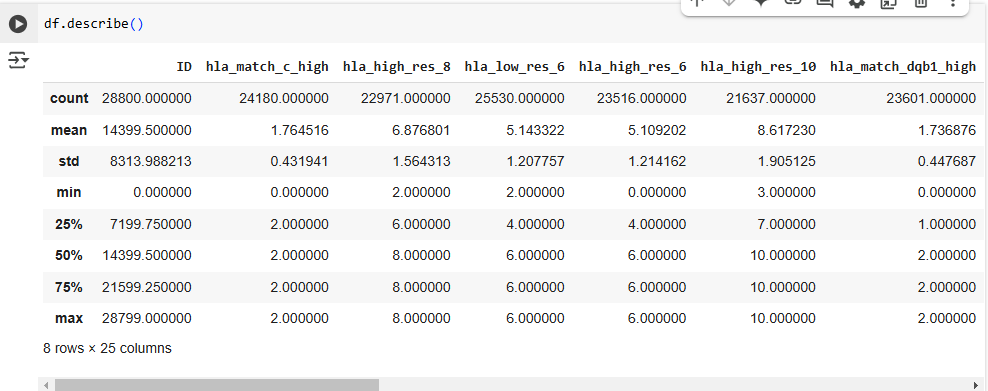
df.describe()



Figure

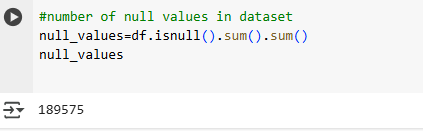


Figure

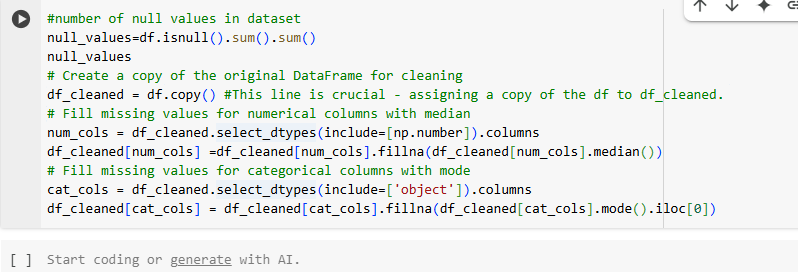


Figure

**Step 2: Handling Missing Values**



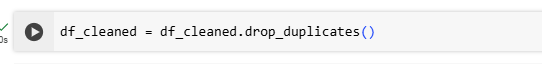
Figure



Figure

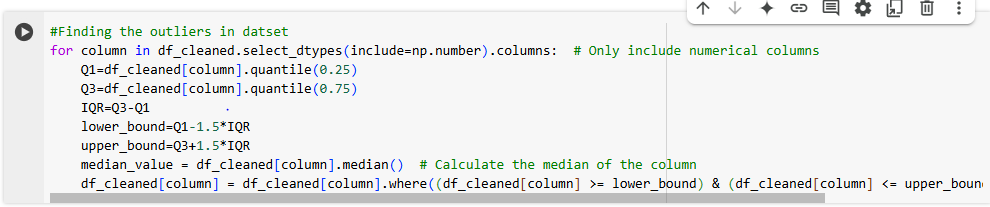
**Step 3: Duplicate Removal**

df\_cleaned = df\_cleaned.drop\_duplicates()

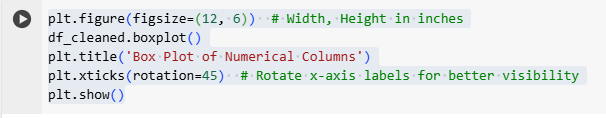


Figure

**Step 4: Outlier and there treatment**



Figure



Figure

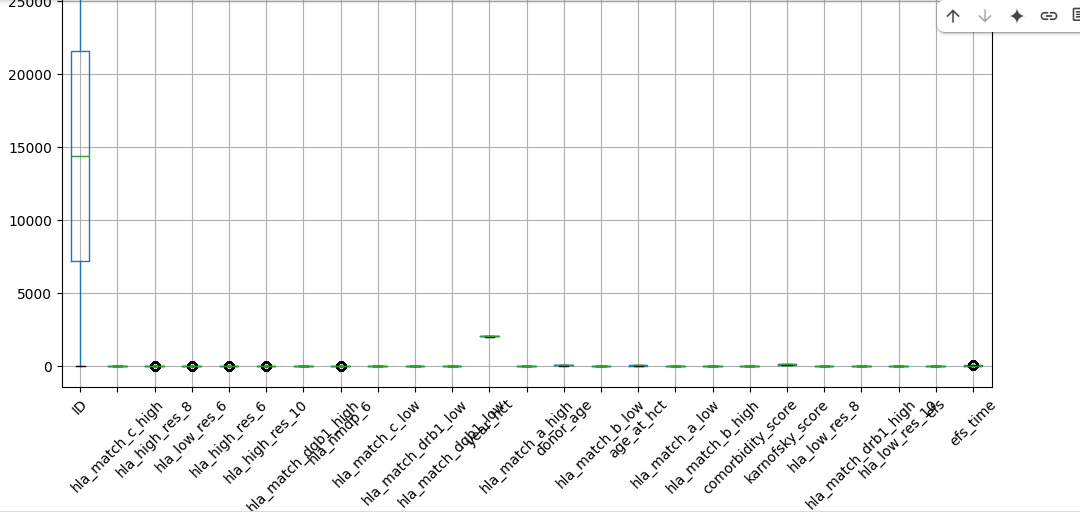
plt.figure(figsize=(12, 6))  # Width, Height in inches

df\_cleaned.boxplot()

plt.title('Box Plot of Numerical Columns')

plt.xticks(rotation=45)  # Rotate x-axis labels for better visibility

plt.show()



Figure

# Plot bar plots for categorical features

cat\_cols = df\_cleaned.select\_dtypes(include=['object']).columns

for col in cat\_cols:

    plt.figure(figsize=(8, 4))

    df\_cleaned[col].value\_counts().plot(kind='bar', color='skyblue', edgecolor='black')

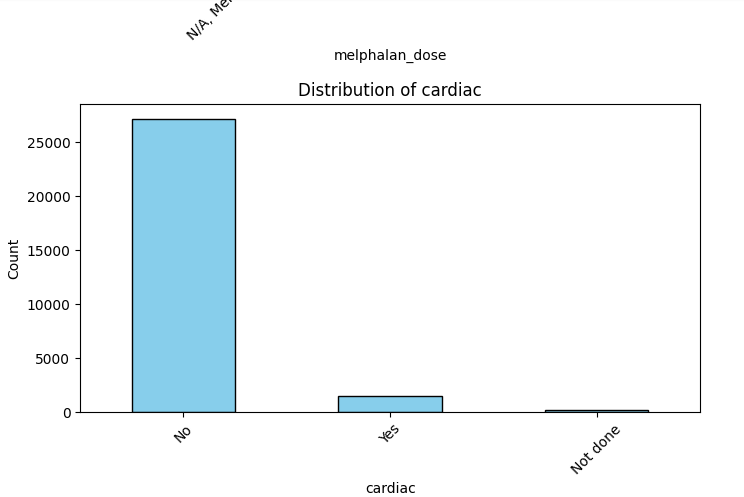
    plt.title(f"Distribution of {col}")

    plt.xlabel(col)

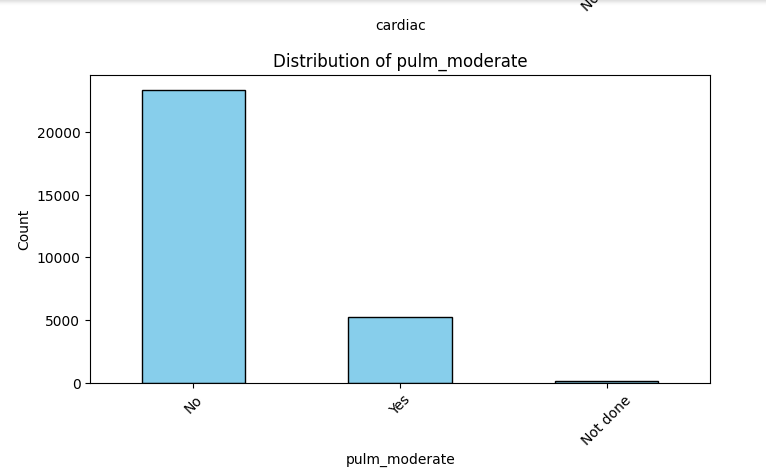
    plt.ylabel("Count")

    plt.xticks(rotation=45)

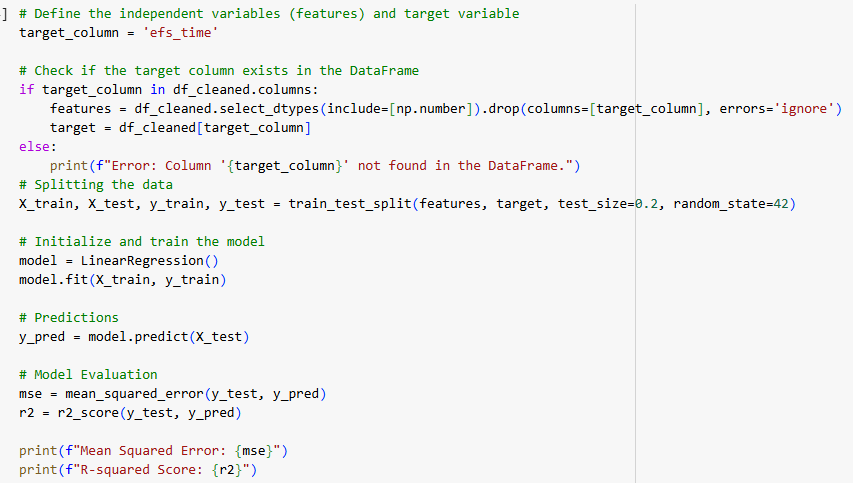
    plt.show()



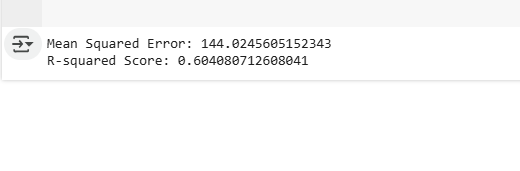
Figure



Figure



Figure



Figure

**Results and Discussion**

After preprocessing, the dataset was cleaned, missing values were handled appropriately, and duplicates were removed. The visual analysis revealed key insights about the distribution of variables and correlations between them. The linear regression model provided a predictive framework for the dataset, with performance metrics indicating its effectiveness.

**Conclusion**

This report demonstrated the importance of data preprocessing in machine learning and data analysis. By handling missing values, removing duplicates, and visualizing key aspects, the dataset was prepared for further modeling. Future steps may include feature engineering and model optimization to improve predictive performance.