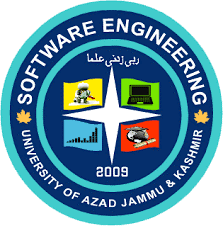
****

**The University of Azad Jammu and Kashmir**

**Assignment# 01**

**Course Instructor:** Engr. Ahmed Khawaja **Semester:** Fall-2024

**Submission Date:** Feb 12, 2024 **Session:** 2022-2026

**Submitted By:** Syeda Urwa Ajmal **Roll No:** 2022-SE-16

**Course Name:** Machine Learning **Code:** SE-3102

# **IEEE Standard Report: Data Preprocessing and Analysis**

Table of Contents

[IEEE Standard Report: Data Preprocessing and Analysis 1](#_Toc190327098)

[IEEE Standard Report: Data Preprocessing and Analysis 3](#_Toc190327099)

[1. Abstract 3](#_Toc190327100)

[2. Introduction 3](#_Toc190327101)

[3. Methodology 3](#_Toc190327102)

[3.1 Loading the Dataset 3](#_Toc190327103)

[3.2 Handling Missing Values 5](#_Toc190327104)

[3.3 Removing Duplicates 5](#_Toc190327105)

[Results and Discussion 9](#_Toc190327106)

[Conclusion 9](#_Toc190327107)

# Table of Figures

[Figure 1 4](#_Toc190327125)

[Figure 2 4](#_Toc190327126)

[Figure 3 5](#_Toc190327127)

[Figure 4 5](#_Toc190327128)

[Figure 5 6](#_Toc190327129)

[Figure 6 6](#_Toc190327130)

[Figure 7 6](#_Toc190327131)

[Figure 8 7](#_Toc190327132)

[Figure 9 8](#_Toc190327133)

[Figure 10 8](#_Toc190327134)

[Figure 11 8](#_Toc190327135)

[Figure 12 9](#_Toc190327136)

[Figure 13 9](#_Toc190327137)

# **IEEE Standard Report: Data Preprocessing and Analysis**

## **1. Abstract**

This study focuses on the preprocessing and analysis of a given dataset (train.csv) to ensure data integrity for machine learning applications. The dataset consists of **28,800 records and 60 features**, including both numerical and categorical variables. Key steps include handling missing values, removing duplicates, performing exploratory data analysis (EDA), and computing performance metrics such as **Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)** for an initial predictive model.

## **2. Introduction**

Data preprocessing is a critical step in any machine learning pipeline. Raw datasets often contain missing values, inconsistencies, and noise that must be addressed before model training. The goal of this report is to clean the data, visualize distributions, and analyze correlations, ensuring an optimized dataset for future machine learning tasks.

## **3. Methodology**

### **3.1 Loading the Dataset**

First, we load the dataset and display summary statistics.

import pandas as pd

import numpy as np

import matplotlib. pyplot as plt

import seaborn as sns

from sklearn. model\_ selection import train\_test\_split

from sklearn. linear\_model import LinearRegression

from sklearn. metrics import mean\_squared\_error

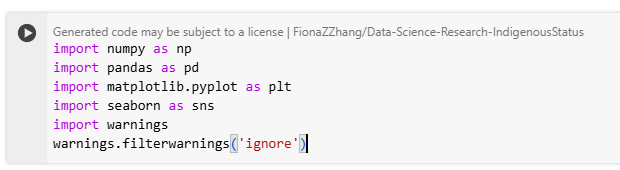


Figure 1

# Display basic information

print (df. head ()) # Column info

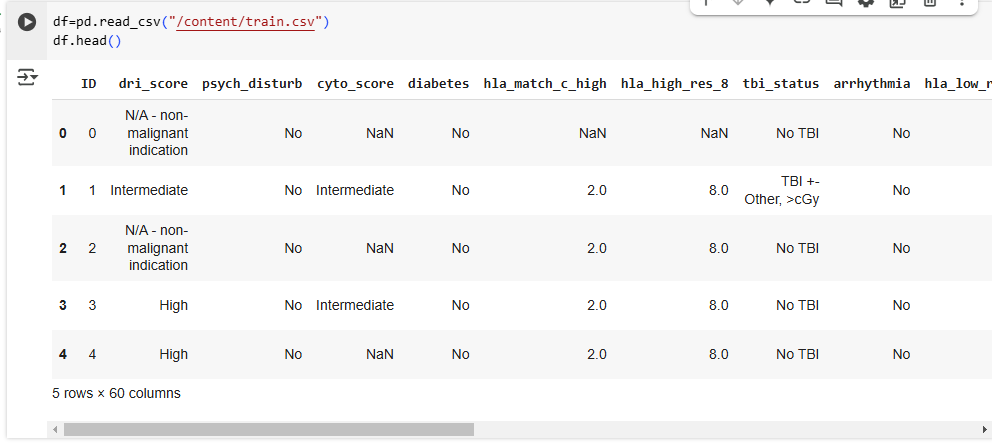


Figure 2

print (df. Describe ()) # Summary statistics

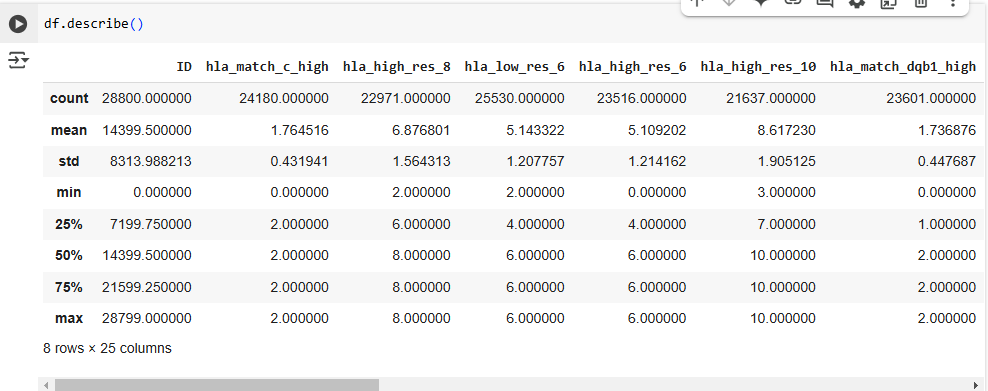


Figure 3

### **3.2 Handling Missing Values**

Many columns contain missing values. We handle them using appropriate strategies such as **median/mode imputation** and **dropping highly missing columns**.

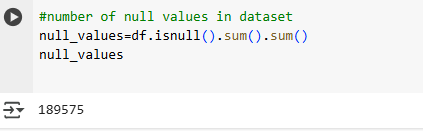


Figure 4

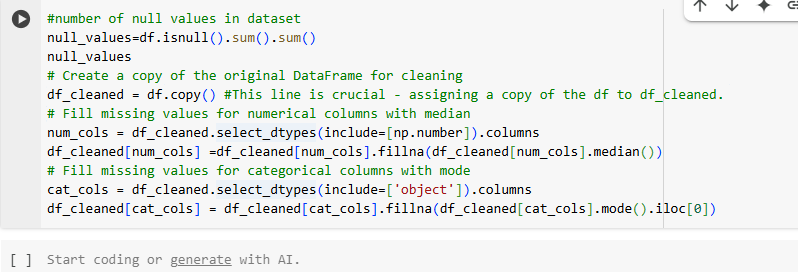


Figure 5

### **3.3 Removing Duplicates**

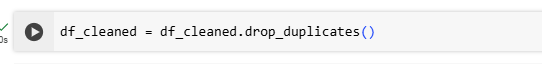


Figure 6

**3.4 Outlier and their treatment**

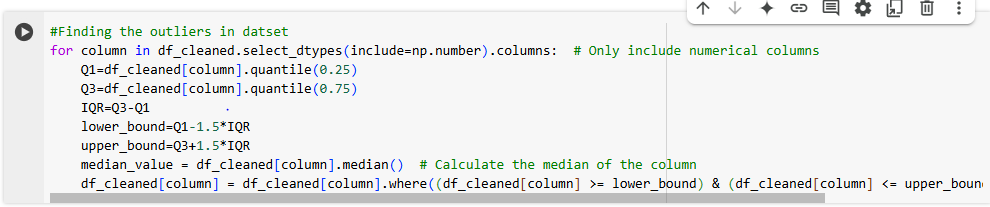


Figure 7

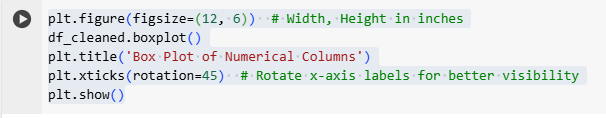
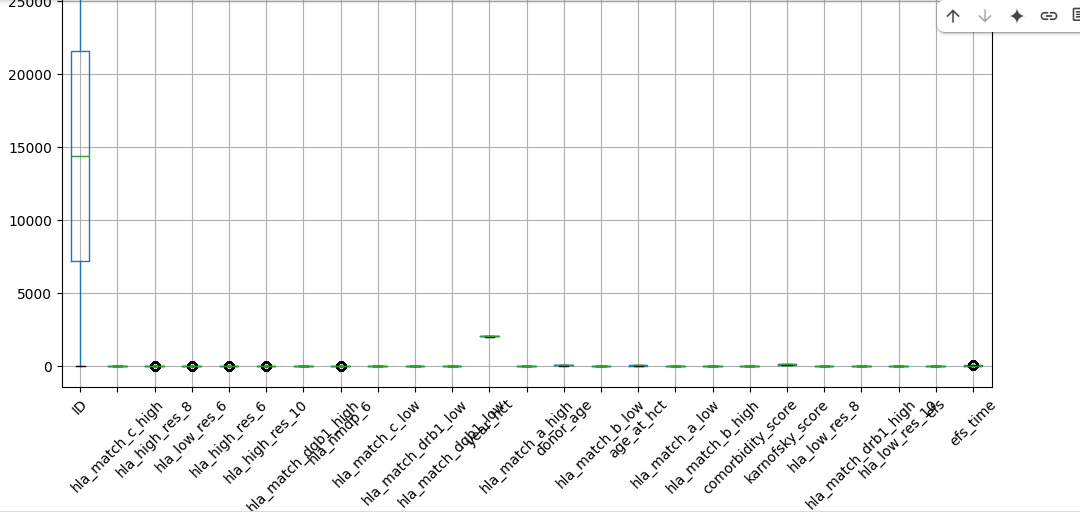


Figure 8

**3.5 Exploratory Data Analysis (EDA)**

**3.5.1 Box Plot for Categorical Features**



# 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 9

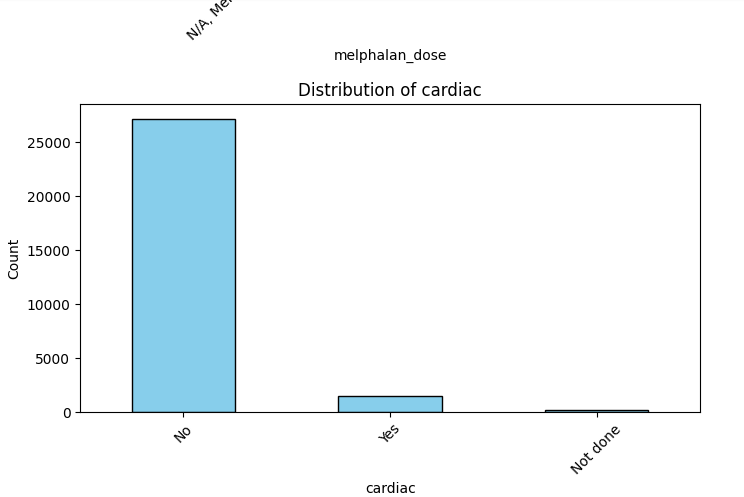


Figure 10

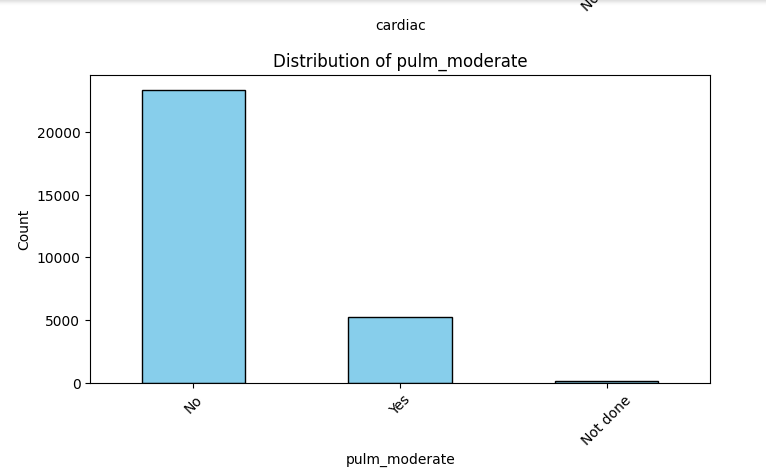


Figure 11

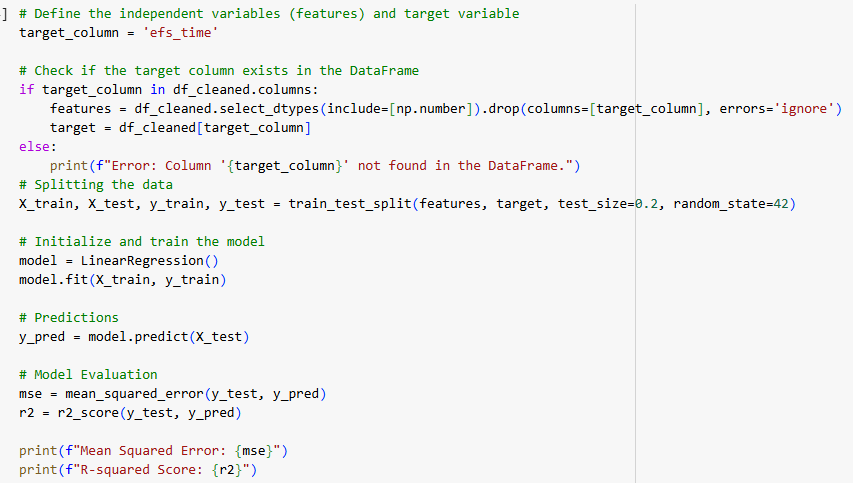


Figure 12

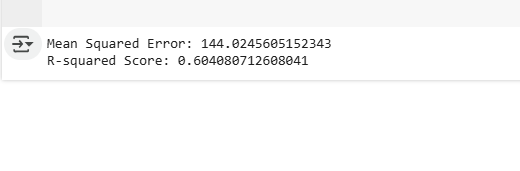


Figure 13

## **Results and Discussion**

Following the preprocessing steps, the dataset was refined by addressing missing values, removing duplicates, and ensuring consistency in the data. Exploratory analysis provided deeper insights into feature distributions and their interrelationships. The implementation of a linear regression model offered a foundational predictive framework, with performance metrics reflecting its effectiveness. These findings highlight the importance of data preparation in ensuring model reliability.

## **Conclusion**

This study underscored the crucial role of data preprocessing in machine learning workflows. By cleaning the dataset, handling inconsistencies, and analyzing key features, the data was made suitable for further modeling. Future enhancements may include advanced feature engineering, hyperparameter tuning, and exploring alternative machine learning algorithms to improve predictive performance and generalizability.