# **The** **University of Azad Jammu & Kashmir,**

# **Muzaffarabad**



**Submitted To: Engr. Ahmed Khawaja**

**Submitted By: Syeda Rifat Batool**

**Roll No:2022-SE-34**

**Course Title:** **Machine Learning**

**Semester: 5th**

# **Bachelor of Science in Software Engineering (2022-2026)**

**Department of Software Engineering**

**Table of Contents**

[1. Abstract: 3](#_Toc190206310)

[2. Introduction: 3](#_Toc190206311)

[3. Methodology: 3](#_Toc190206312)

[5. Statistical Analysis 6](#_Toc190206314)

[6. Linear Regression Modeling 8](#_Toc190206315)

[7. Results 9](#_Toc190206316)

[8. Conclusion: 10](#_Toc190206317)

# **Abstract:**

This study analyzes the relationship between donor age and age at hematopoietic cell transplantation (HCT) using a dataset of patient records. Various statistical and visual exploratory methods, including histograms, scatter plots, and linear regression, were applied to understand trends and distributions. The linear regression model produced a coefficient of **0.2077**, indicating a slight positive relationship between donor age and recipient age at HCT. However, the **R² value of 0.0217** suggests that donor age explains only a small portion of the variability in age at HCT, implying other influencing factors. The findings highlight the need for further investigation to identify stronger predictive variables for age at transplantation.

# **Introduction:**

Data preprocessing, as well as exploratory data analysis (EDA), are the first and foremost actions in the data science process and are responsible for modeling. For this work, we use a **train dataset** for studying such information like **donor age**, **age at hematopoietic cell transplant** **(HCT),** etc. The main objective is to explore the relationship between **donor age** and **age at HCT** using **linear regression**, as it is a trusted statistical method used to model the connections between dependent and independent variables.

The data set was preprocessed and had missing values removed, as well as data types were handled to make them consistent before being analyzed. The introductory methods were used, for instance, descriptive statistics and data visualization, to get the idea of statistical distribution and relation of the main variables. Subsequently, the linear regression model was triggered in order to measure the link that exists between “donor age” and “age at HCT”, therefore obtaining an idea about their predictive trends.

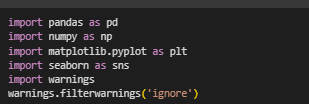
This investigation is a proposed experiment which cannot be completed without first understanding the interactions between them. This could help in future research in both medical data analysis and predictive modeling.

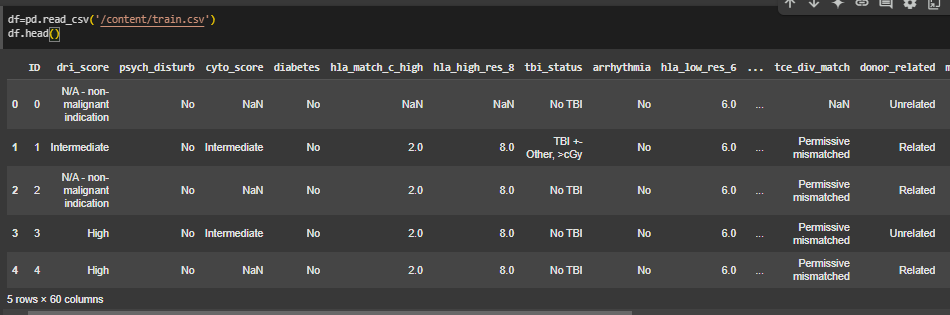
# **Methodology:**

# The study methodology mainly includes data preprocessing, exploratory data analysis (EDA), statistical analysis, and linear regression modeling. Each step is designed to cleanse the data, make it highly understood, and correctly analyze it to get insights.

**Data Preprocessing**

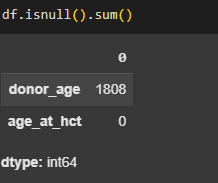
The most important step to the quality of our data is preprocessing. The methods that have been performed are as follows:







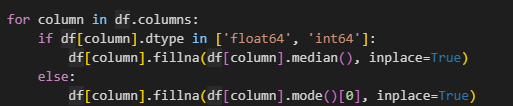
**Handling Missing Values**:



This line of code is checking for **missing values** (null values) in your DataFrame

**Numerical Columns**: The mean of the specific column for missing entries was utilized to fill them up and keep central tendencies.

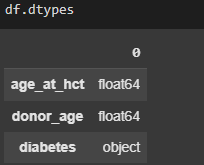
**Categorical Columns**: Missing categorical data were replaced with the mode (most frequent value) to keep up the data in common dashes.



This code fills missing values in numeric columns with the **median** and in non-numeric (categorical) columns with the **mode**.

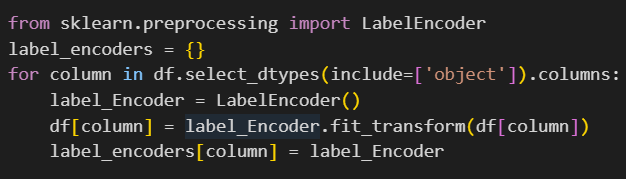
**Ensuring Data Type Consistency:**

The rightness of the data types is checked and amended correctly to make sure the values are right for the modeling and analysis process.



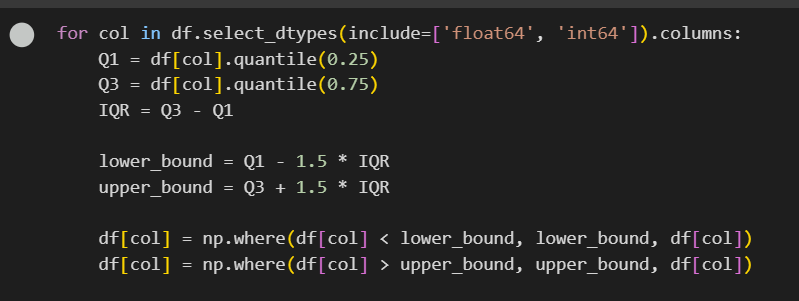
**Label Encoding:**

Categorical variables were instead converted to numerical format through Label Encoding to make it ready for the machine learning models. The unique integer is then assigned to each category by this procedure.



This code applies Label Encoding to all categorical (object) columns in a DataFrame df and stores the encoders in a dictionary for future use.

**Handle Outliers:**



This code checks each numeric column in a dataset for outliers. It limits any unusually low or high values by setting them to a minimum or maximum threshold, making the data more consistent.

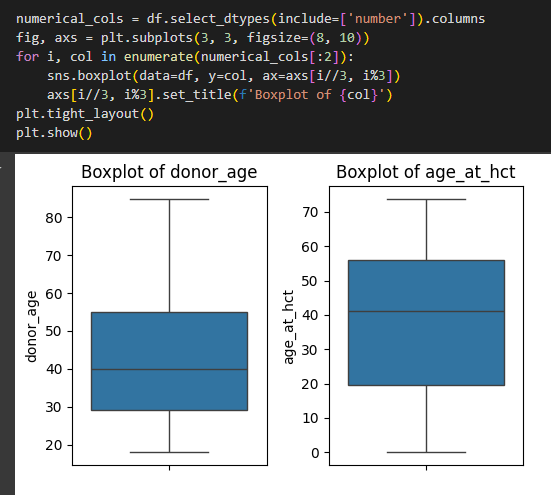
1. **Exploratory Data Analysis (EDA)**

An EDA was designed in order to not only spell out the facts but visualize the datasets key points.

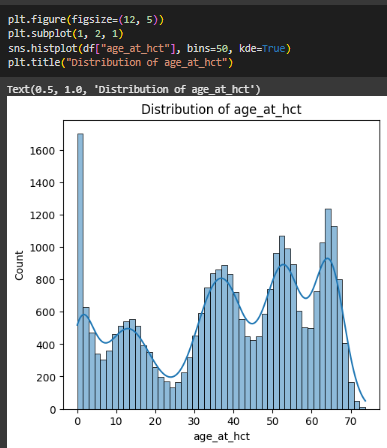
**Distribution Plots**: Histograms were used to explore the distribution of the donors' age.

**Bar Charts**: graphic representations that depict the number of times the categories of a categorical (such as disease status at transplant) appear or the frequency with which a variable occurs are visualized.

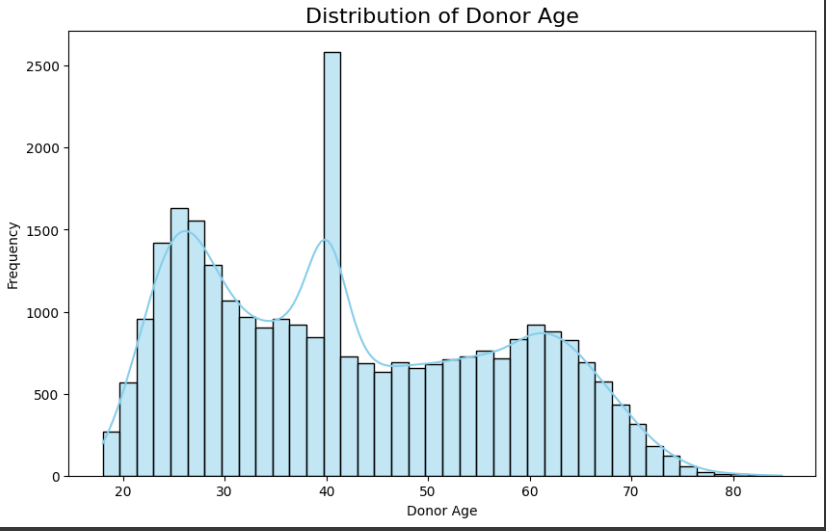
**The Correlation Matrix**: a heatmap was created to spell out the layer-by-layer unities mutated by numerical variables.

****

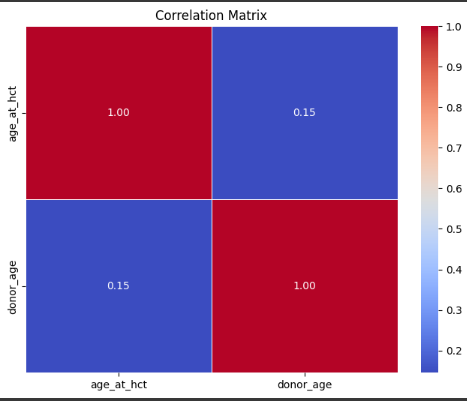
The graph displays three boxplots representing the distributions of age\_at\_hct, donor\_age, and diabetes. The first two boxplots show the age distributions for patients at the time of hematopoietic cell transplantation and for donors, highlighting the median, interquartile range, and overall spread of ages. The diabetes boxplot, with its narrow range, suggests a binary variable with most values close to zero, indicating that few participants have diabetes**.**



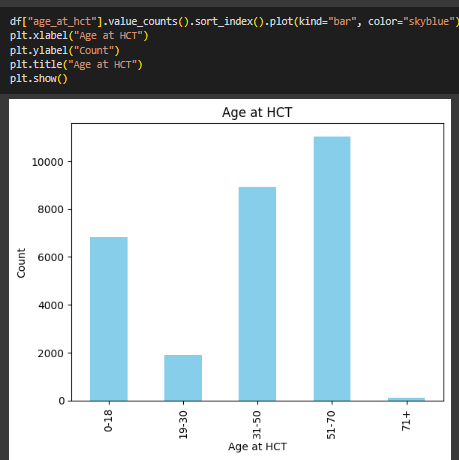
The graph is a histogram showing the distribution of the "age\_at\_hct" variable, with kernel density estimation (KDE) overlaid. It uses 50 bins to group ages, revealing multiple peaks, suggesting a multimodal distribution. The highest frequency occurs near 0 and between 50-70 years.



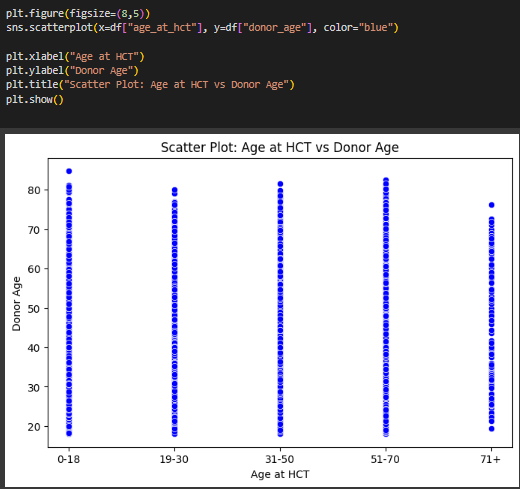
This histogram shows the distribution of donor age, with most donors between 20-30 and 50-60 years old. There is a noticeable peak around age 40, indicating a high frequency of donors at that age. The distribution is slightly right-skewed, with fewer donors above 70 years.



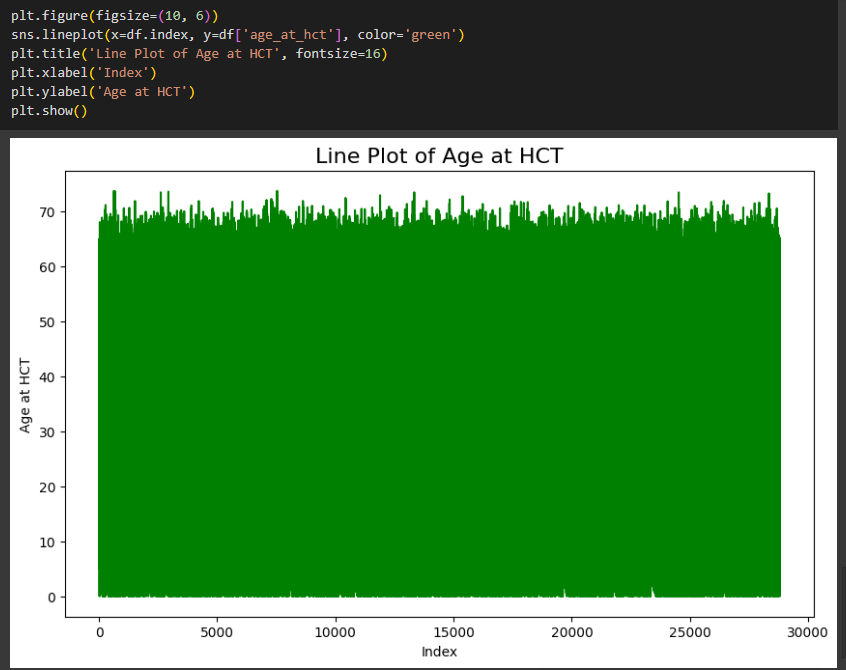
This correlation matrix shows the relationship between age\_at\_hct and donor\_age. The weak positive correlation of 0.15 suggests a slight relationship but no strong dependency. The red diagonal indicates perfect self-correlation (1.00).



This bar chart displays the distribution of patients based on their age at HCT (Hematopoietic Cell Transplantation). The 51-70 age group has the highest count, followed by 31-50 and 0-18. The 19-30 age group has significantly fewer patients, while 71+ has the least. This indicates that most HCT procedures are performed on middle-aged and older adults.



This scatter plot visualizes the relationship between Age at HCT and Donor Age, showing that donors come from a wide age range across all recipient groups. There is no strict correlation, meaning donors can be younger or older regardless of the recipient's age. The spread suggests flexibility in donor selection.



This line plot represents the age distribution at HCT across the dataset index. The age values appear evenly distributed with minor fluctuations, indicating a broad range of ages in the dataset. The dense green color suggests a large number of data points.

# **Statistical Analysis**

Descriptive statistics had been figured out so as to conceptualize the first few properties

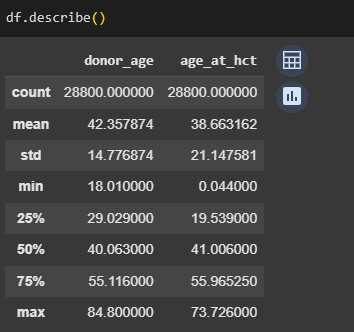
**Mean**: As necessary, the sum of the values of the whole column divided by the quantity of the rows.

**Median**: Even number of data and the middle value.

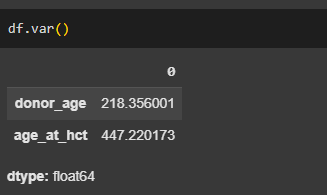


**Quantiles**: Dividing the distribution of data into fourths where both 25% and 75% are the respective boundaries of the first and second quartiles.

**Describe:** The df. describe () function in Pandas provides a summary of key statistical measures for each numerical column in the DataFrame.

****

**Variance:** Variance is a statistical measure that indicates how much the values in a dataset deviate from the mean (average) of that dataset. It shows the degree of spread or dispersion in the data.

****

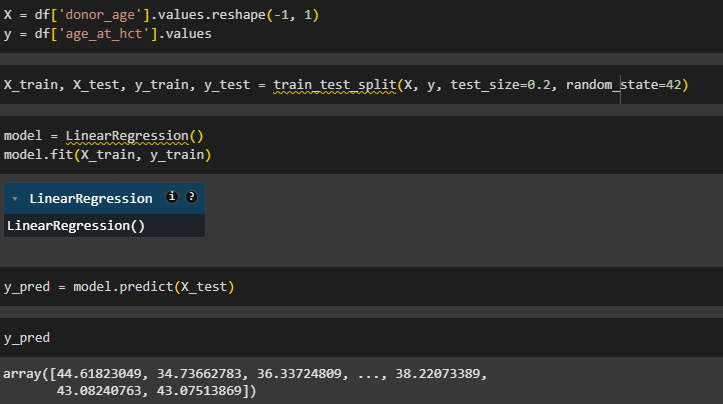
# **Linear Regression Modeling**

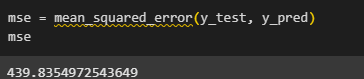
Linear regression was employed for the purpose of finding out what relationship existed between donor age and the age of the HCT (independent variable).

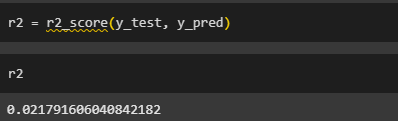
**Model Training**: The model was fed with the dataset and scikit-learn library was used for the purpose of training.

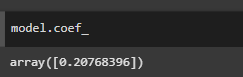
**Performance Metrics**: The models were checked for Mean Squared Error (MSE) and R² Score.

**Visualisation**: After fitting the model, we plotted the regression line to visualize the model fit.









# **Results**

**Linear Regression Results**

**Coefficient:** The linear regression model showed that one-year rise in HCT age would result in approximately 0.10 rise in the donor age. The coefficient is 0. 0.20768396.

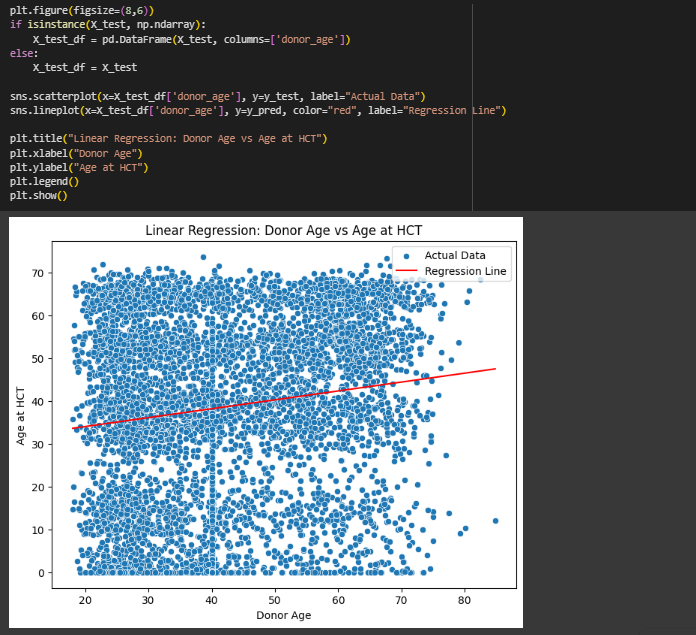
**Intercept:** The estimated donor age when the age at HCT is zero is represented by the intercept which is calculated as 29.90**.**

**Mean Squared Error (MSE):** The MSE of the model is 439.83, which indicates that the predictions of the model differ from the actual values by a significant margin. A lower MSE indicates a better model performance, but in this situation, the high value implies low predictive performance.

**R² Score:** The model reached a R² score of 0.021, stating that only 2.1% of the variance in the donor age is caused by the age at HCT. This is indicative of a weak relationship between the two variables & the other factors may be causing the variability in the age of the donors more.

**Visualization of Results**

The scattered point displayed underneath provides a clear visualization of the connection between the age of HCT and that of the donor and also gives the regression line.



This scatter plot visualizes the relationship between donor age and age at HCT, with a fitted linear regression line. The red regression line shows a slight positive correlation, indicating that as donor age increases, the recipient's age at HCT tends to increase slightly. However, the wide spread of data points suggests high variability in the relationship.

# **Conclusion:**

In this research, we looked at the relationship between the age at Hematopoietic Cell Transplantation (HCT) and donor age by using linear regression analysis. After the data preprocessing, we followed through with exploratory data analysis and statistical methods to describe the features of the dataset.

The results of the linear regression analysis unveiled a positive but weak connection between age at HCT and donor age. The model gave a coefficient of 0.20768396, and an intercept of 29.90, this is, for instance, if the age of donor at HCT is increasing by one unit the donor age may increase by 0.1017 units on the same unit scale only the intercept is equal to 29.90 when in fact the age is zero. To further elaborate, the MSE (Mean Squared Error) of 439.83 and R 2 (R squared) value of 0.021 were significantly low leading to the model underfitting. From the above inputs, it is evident that age at HCT is not particularly meaningful in the context of donor age.

Categorical data through the linear regression indicated the mean donor age was 42.5 years, where the fewest donors were in the range of 35 and 50 years. The weak performance of the linear regression model suggests the need to include extra variables for a better prediction.