MAKERERE UNIVERSITY

SEMESTER ONE 2024/2025 ACADEMIC YEAR

COLLEGE OF COMPUTING AND INFORMATION SCIENCES

SCHOOL OF COMPUTING AND INFORMATICS TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

MASTER OF SCIENCE IN COMPUTER SCIENCE

COURSE NAME: MACHINE LEARNING

COURSE CODE: MCS 7103

LECTURER: DR JOYCE NAKATUMBA-NABENDE

DATE: Wednesday, 11 September 2024

COURSEWORK ONE (1)

STUDENT DETAILS

NAME	STUDENT NUMBER	REG NUMBER
ABILA Raphael	2400721912	2024/HD05/21912U

MACHINE LEARNING DATASET REPORT

DATASET: HIV PATIENT RISK ASSESSEMNT AND CLASSIFICATION

Choosing a dataset for my machine learning course involved careful consideration and reflection, leading to a series of important questions that guided my decision-making process. Selecting the right dataset is crucial, as it shapes the direction of the project and influences the insights and conclusions that can be drawn from the analysis.

Having worked in the health domain for several years, I am particularly interested in deriving solutions that benefit the healthcare sector and improve services offered to the population. This interest drove me to focus on HIV patients, especially in light of recent reports highlighting a significant increase in HIV cases across the country, with **Masaka** district experiencing the highest surge in patients over the past one to two years. While there is a generally adequate level of medication adherence, some patients struggle with adherence, which poses risks to effective HIV treatment.

Additionally, there has been a notable number of patients transitioning between different WHO regimen lines—such as moving from first-line to second-line treatments, or even from second-line to third-line treatments. This situation raises critical questions regarding the risks associated with patients changing treatment lines and the implications of non-adherence to their medication. Specifically, I have been contemplating the extent of these risks and how they might affect patient outcomes. This line of inquiry has led me to formulate several key questions:

Questions Before Dataset Selection

1. What are the potential risks faced by patients suffering from HIV?
   The patient's risk having an increase in viral load values and CD4 values that might cause them to die or be moved from one line of treatment to the other

2. How can these risks be categorized?
   Patients on treatment can be categorized as high-risk HIV patients and low risk HIV patients basing on the viral load and CD4 values

3. Is there a standardized method for determining whether an HIV patient is at risk?
   Medical personnel rely so much on viral load values and CD4 values to be able to qualify the risk category of a patient

4. What types of data are necessary to assess the risks associated with HIV patients?
   One needs viral load as the major one however CD4 values can also be used the two relate to each other

5. Where can I find reliable data to evaluate the risks faced by HIV patients?
   This data is collected from hospitals and stored in the National warehouse of MINISTRY OF HEALTH or DHIS2

6. What features in the dataset will allow me to gauge the risks in HIV patients?
   If viral load data is available, I will be able to gauge the risk level of a particular patient
7. How can I effectively categorize the risks associated with HIV patients?
   Low-risk and High-risk HIV patients however a consideration can also be made for medium

Data wrangling process explanation

After I loaded my data into google Collab, I had a total of 340,364 rows and 19 columns in my dataset. Out of the 19 columns I had 6 columns with float as their data types, whereas the remaining 13 were of type object

The major columns that I can use to determine patient risk levels will be latest_viral_load_copies and baseline_cd4, please note that either of the two can be used to determine risk level in HIV treatment

A total of 240,334 rows did not have viral load (latest_viral_load_copies) data forming 70.61% of the data with missing viral load data while a total of 277997 rows did not have CD4 (baseline_cd4) data forming a percentage of 81.68% of the data

These two columns being my major set of features to be considered while determining my risk category of a particular patient, I had to further determine if there are rows that had both values missing and discovered a total of 212142 of the rows with both values missing out of the grand total of 340364 rows

In conclusion the complete rows made a total of 128222 records making a 37.7 % of my total data meaning omitting the records I would then have lost 62.3 % of my data

The most practical way out of this is to fill the missing values with the median in order to have my dataset complete for the two core numerical dataset columns latest_viral_load_copies (viral load) and baseline_cd4 (CD4)

To ensure that my data types are appropriate for analysis, for the categorical variables in my dataset like gender, current_regimen, I converted their data types from object to category

Created new columns in my data for instance to determine the age from date of birth column, length of enrollment from the current timestamp and the date of enrollment

Removed all the records that have date of death values since am considering only active patients

After my data wrangling phase, a number of questions were answered as follows;

Questions during and after Data Wrangling phase

1. What is the primary purpose of my dataset?
   The primary purpose of my dataset is to classify HIV treatment patient risk levels into high risk and low risk patients

2. What variables are included in my dataset?
   A number of variables (19) are included in my dataset as listed below;

```
0   view_sentinel_events_id              340364 non-null  object
1   status_date                          340364 non-null  object
2   patient_clinic_no                    332399 non-null  object
3   gender                               340364 non-null  object
4   date_of_birth                        340364 non-null  object
5   diagnosis_date                       160971 non-null  object
6   art_enrollment_date                  161079 non-null  object
7   art_start_date                       173364 non-null  object
8   baseline_regimen                     168537 non-null  object
9   current_regimen                      0 non-null       float64
10  current_regimen_line                 0 non-null       float64
11  baseline_cd4                         60708 non-null   float64
12  latest_cd4_date                      62367 non-null   object
13  latest_cd4                           62367 non-null   float64
14  latest_viral_load_date               120539 non-null  object
15  latest_viral_load_copies             100030 non-null  float64
16  latest_viral_load_qualitative        103896 non-null  object
17  latest_viral_load_suppression_status 106500 non-null  float64
18  date_of_death                        13010 non-null   object
```

3. Does the dataset contain the relevant variables needed to assess risk levels in HIV patients?
   Yes, my dataset contains all the necessary variables to asses risk in HIV patients (`latest_viral_load_copies, latest_cd4`)

4. What does each column in my dataset represent?
   I was able to figure out what each column in my dataset represents being it is a categorical or numerical variable

5. How many variables/features are essential for determining HIV patient risk levels in my dataset?
   There are only two major variables to be used to determine patient risk levels however there is a high dependence on the other variables like gender among others

6. Could there be missing values in my dataset, considering the must-have columns/features?
   Yes, I found some missing values in my dataset

7. What should I do when only one feature is missing while the other has data for a particular patient, which is part of the essential features to gauge risks in HIV patients?
   A consideration can be made for one variable to assess risk levels in HIV patients however either viral load or CD4 can be used

8. What should I do if both features in my dataset have missing values, considering the core columns used to determine risks in HIV patients?
   I was able to figure out how to handle the missing values by filling them up with a median value for the numerical variables

9. Can I add more columns to my dataset to gain additional insights?
   I was able to add additional columns to my dataset in order to gain more insight into my data like Age, Length of treatment which were obtained from date of birth and enrollment dates respectively

10. Should I perform classification or regression analysis, considering that my dataset supports both?
    From my dataset, I was able to conclude that I will perform a classification algorithm to classify high risk and low risk HIV patients on treatment
11. Is there a relationship between the columns in my dataset, such as between viral load and other factors like gender?

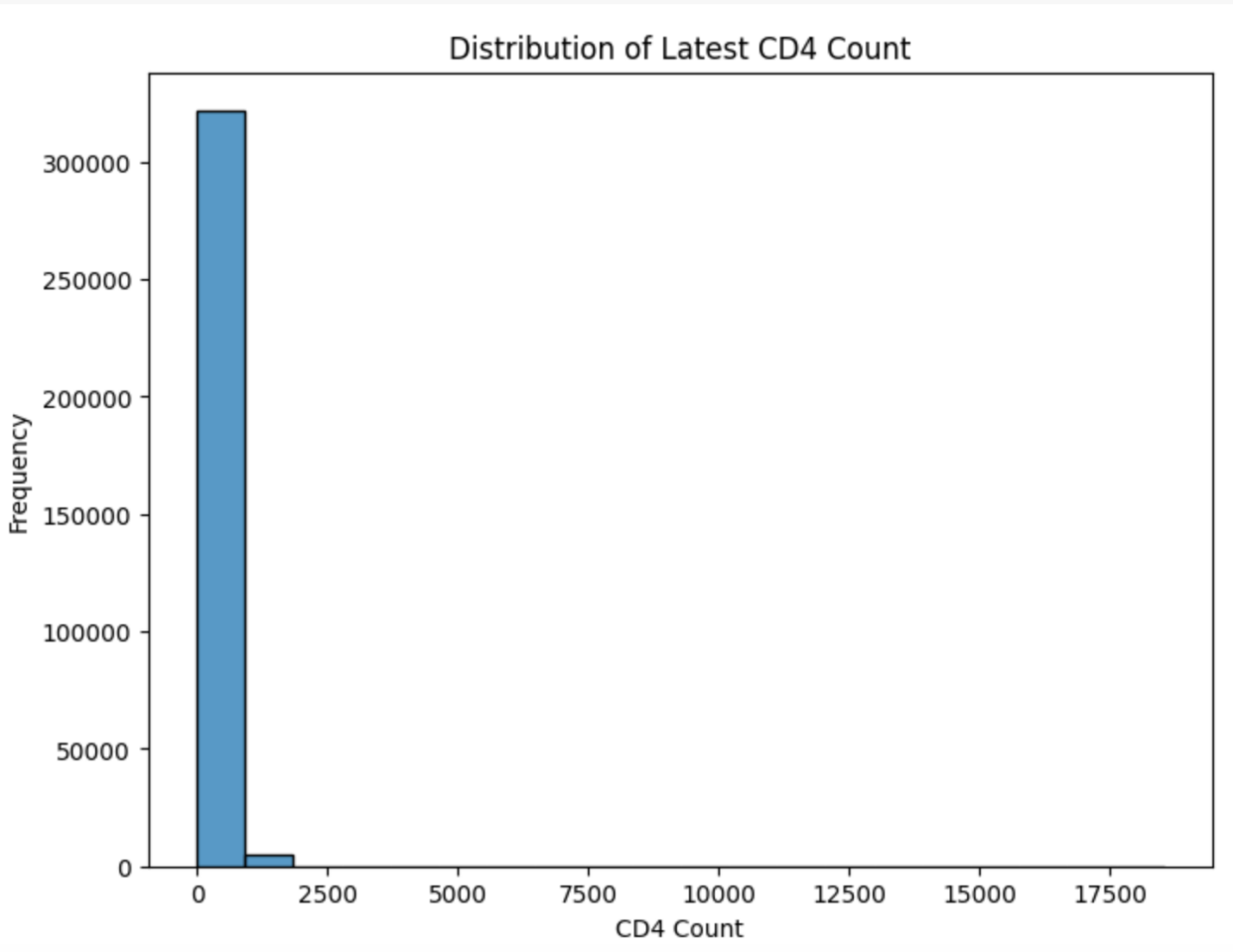
Yes, I was able to notice that females tend to have more risk when it comes to HIV viral loads and CD4 variables growing

EDA (Exploratory Data Analysis)

To gain more insights into my dataset, I attempted two major form of visualization techniques

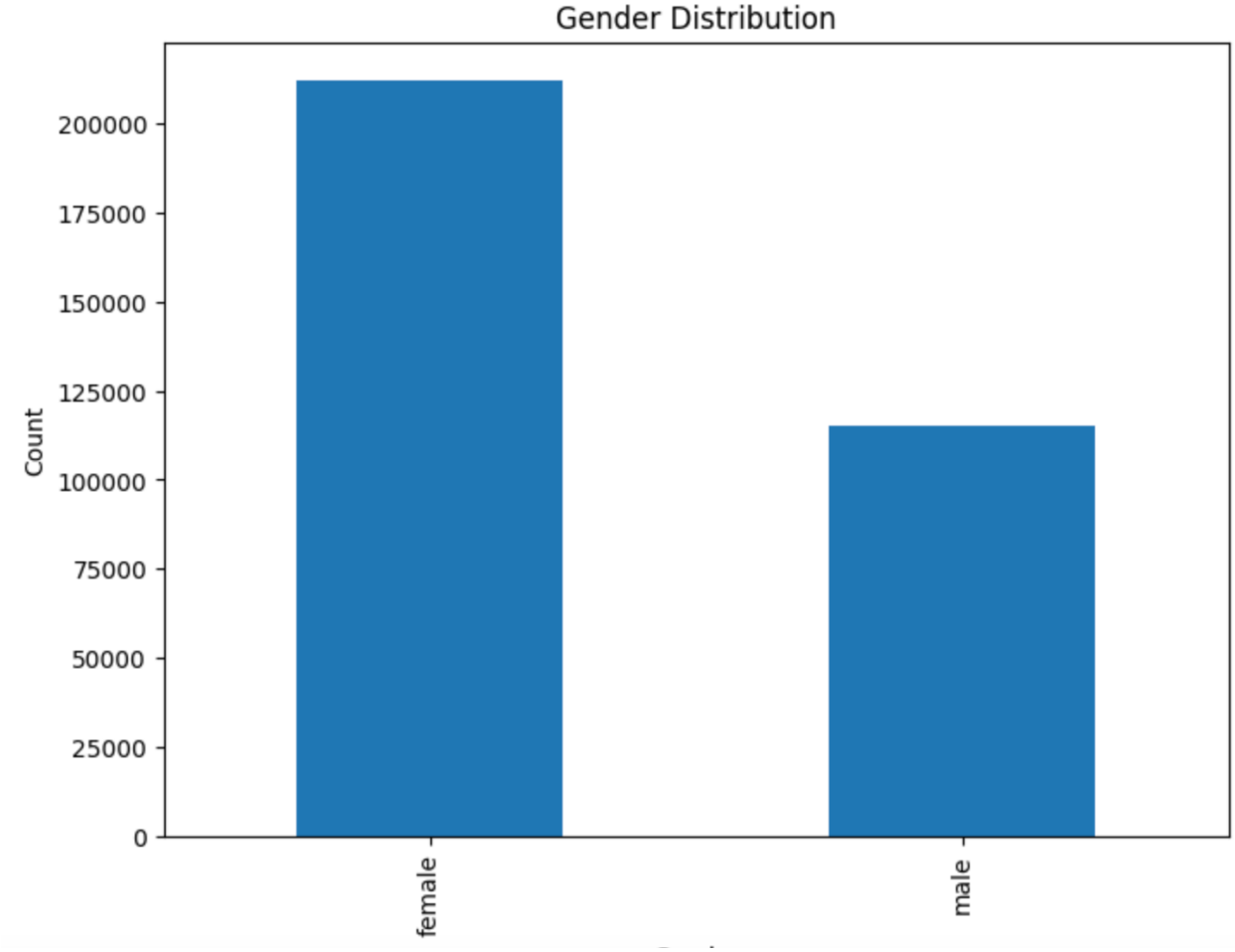
1. Univariate analysis:

Histogram: By plotting the histogram for my numerical variables like latest_cd4, latest_viral_load_copies and age, I was able to observe there were some skewness, outliers and the overall spread of data was fairly good



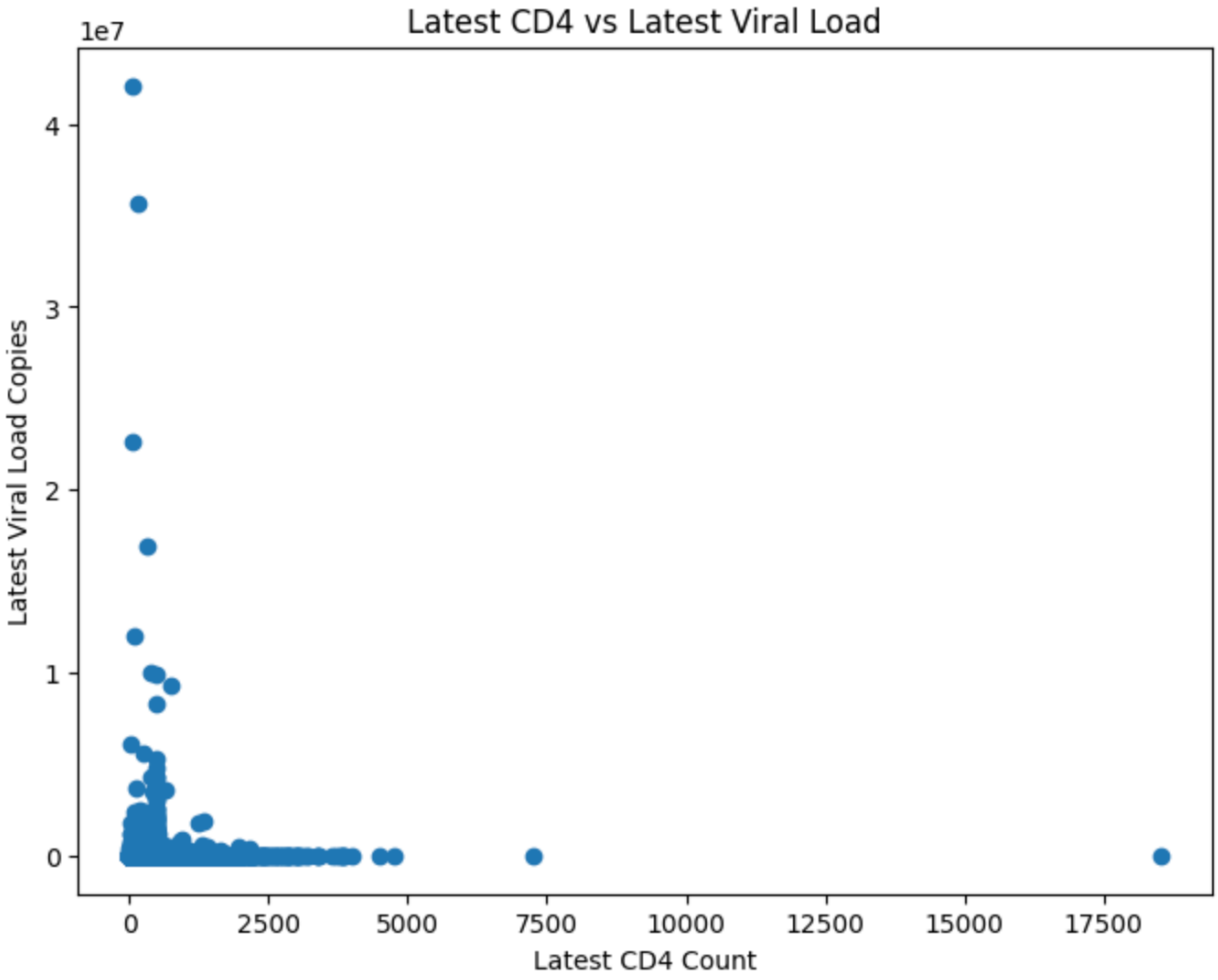
Bar plots: I was able to visualize categorical data like gender, viral_load_status using bar plots and observed the frequency of each

category which helped me understand the demographic distribution and treatment outcomes

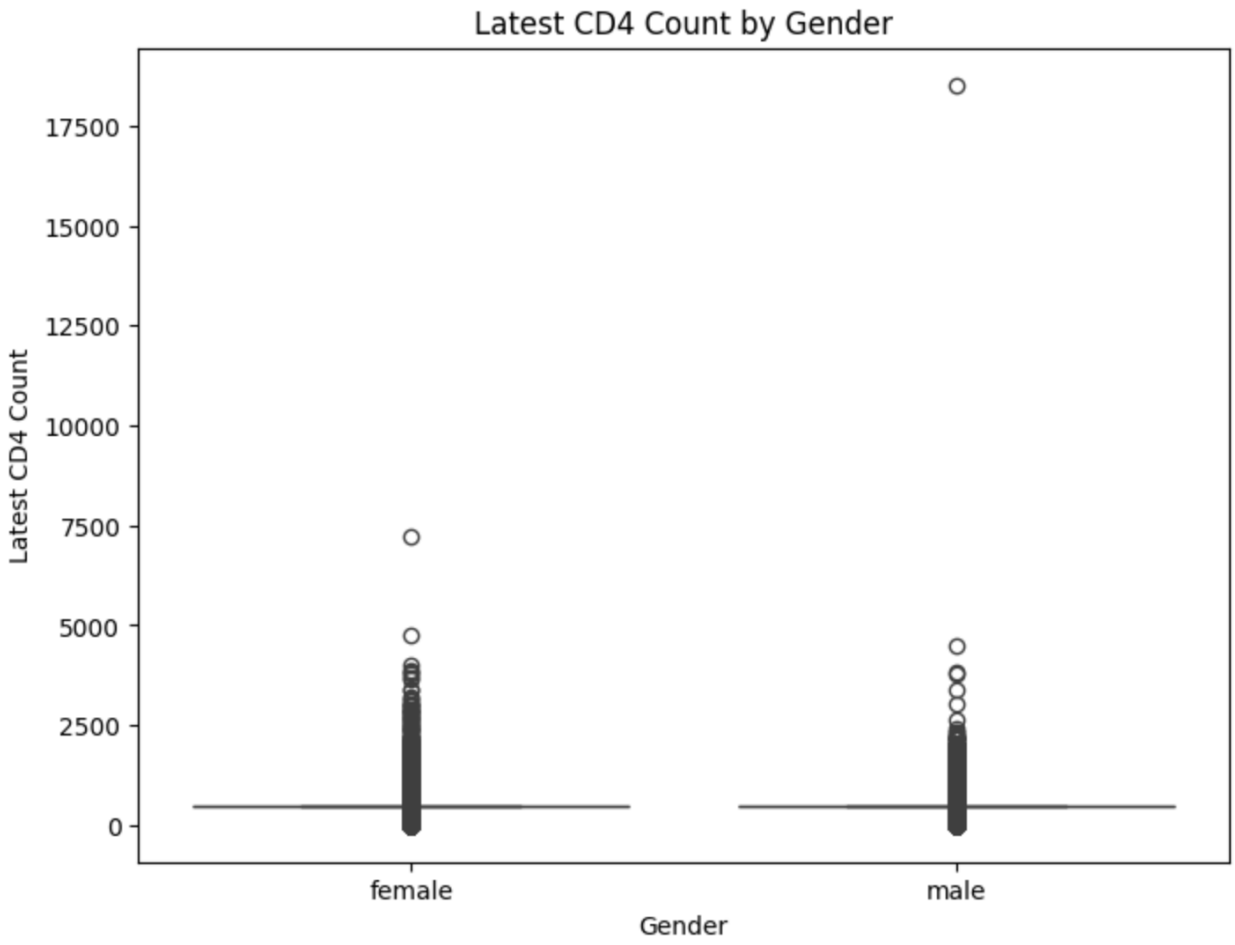


2. Bivariate Analysis

Scatter Plots: Using scatter plots to plot my latest_cd4 and latest_viral_load_copies could reveal that correlations or trends and relationships between my variables where I could conclude that higher CD4 values were greatly associated with lower viral load copies, which is crucial for assessing risks in treatment together with its effectiveness



Box plots: Through this visualization, I was able to compare the distribution of latest_cd4 across different categories like gender, and was able to identify the difference in treatment outcomes between groups and highlighted disparities in healthcare access or effectiveness



Conclusions made

- Use latest_viral_load_copies and baseline_cd4 columns as must have to determine the level of risk
- I will use classification algorithm to determine patients that are of high risk and low risk during their HIV treatment
- I will fill data rows that do not have either latest_viral_load_copies or baseline_cd4 values with median values since these are the only two features, I can use to gauge a patient
- Patients who are deceased will not be considered in my dataset, since I want to classify patients under going treatment within high or low risk categories
- There is a high relationship between cd4 copies and viral loads since both depend on each other

- Classifying my datasets or categorizing the dataset through variables like gender could clearly indicate that females have the highest risk of contraction of HIV and thus probably having the highest risk during HIV treatments.