Solent University

Faculty of Business, Law and Digital Technologies

**Data Analysis of Hepatitis C Virus (HCV) and Prediction by Machine Learning Techniques**

Author : **Q16367618 Ganiyu Idris**

Course Title : **COM725**

Module Leader : **Dr Drishty Sobnath**

Date : **August 30, 2023**

Table of Contents

[INTRODUCTION 4](#_Toc144275078)

[Aim 4](#_Toc144275079)

[Objectives 4](#_Toc144275080)

[Data Source 4](#_Toc144275081)

[Description Of Attribute 5](#_Toc144275082)

[Data Analysis and Visualization 6](#_Toc144275083)

[Data Cleaning 6](#_Toc144275084)

[Exploratory Data Analysis (EDA) & Visualization 6](#_Toc144275085)

[Data Overview and Summaries 6](#_Toc144275086)

[Missing Values 7](#_Toc144275087)

[Data Visualization 8](#_Toc144275088)

[Data Pre Processing 14](#_Toc144275089)

[Machine Learning Algorithm 15](#_Toc144275090)

[Excel 16](#_Toc144275091)

[Tableau 16](#_Toc144275092)

[Multi-Class Visualization 16](#_Toc144275093)

[Binary-Class Visualization 20](#_Toc144275094)

[Weka 20](#_Toc144275095)

[Workflow 20](#_Toc144275096)

[Multi Class Weka 21](#_Toc144275097)

[Binary Class Weka 23](#_Toc144275098)

[Conclusion 25](#_Toc144275099)

[Bibliography 25](#_Toc144275100)

Table of Figure

[**Figure 1 Descriptive statistics table** 6](#_Toc144274205)

[**Figure 2 Missing Values Table** 7](#_Toc144274206)

[**Figure 3 Gender Distribution** 8](#_Toc144274207)

[**Figure 4 BMI Distribution** 8](#_Toc144274208)

[**Figure 5 Age Distribution** 9](#_Toc144274209)

[**Figure 6 Age Distribution by Gender** 9](#_Toc144274210)

[**Figure 7 ALT Boxplot** 10](#_Toc144274211)

[**Figure 8 RNA Boxplot** 11](#_Toc144274212)

[**Figure 9 Pair Plots** 11](#_Toc144274213)

[**Figure 10 ALT improvement bar chart** 12](#_Toc144274214)

[**Figure 11 RNA Improvement bar chart** 12](#_Toc144274215)

[**Figure 12 Correlation Table** 13](#_Toc144274216)

[**Figure 13 Pearson Correlation** 14](#_Toc144274217)

[**Figure 14 Confusion Matrix** 15](#_Toc144274218)

[**Figure 15 Age- Multi Class Relationship** 17](#_Toc144274219)

[**Figure 16 Age against BMI** 18](#_Toc144274220)

[**Figure 17 Age-BMI by Gender Relationship** 18](#_Toc144274221)

[**Figure 18 RBC-WBC Boxplot** 19](#_Toc144274222)

[**Figure 19 Multi Class Dashboard** 19](#_Toc144274223)

[**Figure 20 Binary Class Dashboard** 20](#_Toc144274224)

[**Figure 21 Weka Multi Class** 21](#_Toc144274225)

[**Figure 22 Weka Multi Decision Stump** 22](#_Toc144274226)

[**Figure 23 Weka Multi Random Tree** 22](#_Toc144274227)

[**Figure 24 Weka Binary Class** 23](#_Toc144274228)

[**Figure 25 Weka Binary Decision Stump** 24](#_Toc144274229)

[**Figure 26 Weka Binnary Random Tree** 24](#_Toc144274230)

# **INTRODUCTION**

The principal blood-borne human pathogens known as Hepatitis C include the RNA virus known as Hepatitis C virus (HCV). HCV infection frequently shows no or very few symptoms throughout the infection stage. Most acute infections become chronic if untreated, eventually developing into liver disorders such cirrhosis and hepatocellular carcinoma. According to the World Health Organization (WHO), 3–4 million new cases of HCV are reported every year, making the global prevalence of the disease about 3%, or 120–130 million persons. The essential need for effective intervention programs with an emphasis on identification and treatment is highlighted by this startling prevalence.

Notably, there are 86 subtypes of HCV and 8 verified genotypes, with variations in distribution according on location. The country with the highest frequency, Egypt, stands out with about 22% of its population. This can be ascribed to a widespread intravenous antischistosomal injection treatment campaign that took place between 1950 and 1980 [4]. While Australia, North America, and Western Europe have lower incidence rates than other places, Asia and Africa have greater prevalence rates.

For both pediatric and adult patients, the conventional liver biopsy has proved an invaluable diagnostic tool for identifying and assessing liver fibrosis. This technique, however, can result in post-procedure hospitalization and is invasive and uncomfortable. Additionally, it has been linked to diagnostic or staging mistakes of about 20–30%. Due to these difficulties, alternative non-invasive diagnostic biomarkers have been investigated, including serum biochemical markers and several imaging modalities like ultrasound and magnetic resonance imaging. These methods aim to evaluate hepatic stiffness and may combine a number of non-invasive techniques to overcome the drawbacks of liver biopsy.

## **Aim**

This project has the aim of investigating the various contributing factors that can potentially influence the baseline histological staging (Class Label). Additionally, he project also seeks to conduct performance comparisons between two types of class labels: multi-class labels and binary class labels.

## **Objectives**

The objective was to predict the histological staging of the disease, which has four classes: F1 (portal fibrosis without septa), F2 (few septa), F3 (many septa without cirrhosis), and F4 (cirrhosis). Additionally, a binary class label was created by combining F1 and F2 as one class and F3 and F4 as another.

## **Data Source**

The publicly accessible [UCI Machine Learning](https://archive.ics.uci.edu/dataset/503/hepatitis+c+virus+hcv+for+egyptian+patients) repository is the source of the Hepatitis C Virus (HCV) dataset specific to Egyptian patients. There are 1385 cases in this dataset, and each one has 29 multivariate attributes. The multiclass label known as "Baseline histological staging," which is linked to different values and their related frequencies, provides the foundation for the dataset's classification. These values are F1 (336 occurrences of portal fibrosis without septa), F2 (332 instances), F3 (355 instances of numerous septa without cirrhosis), and F4 (362 instances of cirrhosis).

|  |  |
| --- | --- |
| Multi Class Label | Instances |
| Portal Fibrosis, F1 | 336 |
| Few Septa, F2 | 332 |
| Many Septa, F3 | 355 |
| Cirrhosis, F4 | 362 |

|  |  |
| --- | --- |
| Binary Class Label | Instances |
| Class = 0 (Mild to Moderate, F1-F2) | 668 |
| Class = 1 (Advanced cirrhosis, F3-F4) | 717 |

## **Description Of Attribute**

A total of 29 attributes make up the dataset, 28 of which are input attributes and 1 of which is a key attribute known as "staging" with multiclass labels (F1-F4) as stated in the data source.

|  |  |
| --- | --- |
| **Description** | **Description** |
| Age | ALT 1 (1week) - Alanine transaminase ratio 1 week |
| Gender | ALT 4 (4week) - Alanine transaminase ratio 4 weeks |
| BMI (Body Mass Index) | ALT 12 (12week) - Alanine transaminase ratio 12 weeks |
| Fever | ALT 36 (36week) - Alanine transaminase ratio 36 weeks |
| Nausea/Vomiting | ALT 48 (48week) - Alanine transaminase ratio 48 weeks |
| Headache | ALT 24 (24 week) - Alanine transaminase ratio 24 weeks |
| Diarrhea | ALT (after 24week) – Alanine transaminase ratio after 24 weeks |
| Fatigue & generalized bone ache | RNA Base |
| Jaundice | RNA 4 |
| Epigastric pain | RNA 12 |
| WBC (White blood cells) | RNA EOT (RNA End-of-Treatment) |
| RBC (Red blood cells) | RNA EF (RNA Elongation Factor) |
| HGB (Hemoglobin) | Baseline histological Grading |
| Plat (Platelets) | Baseline histological staging (Class labels) |
| AST 1 - Aspartate transaminase ratio |  |

# **Data Analysis and Visualization**

I will take you through the Data cleaning, Exploratory Data Analysis & Visualization, data preparation and feature selection in this section.

## **Data Cleaning**

The data cleaning process begins with the enhancement of column names by removing extraneous spaces and addressing typographical errors. This action is taken to ensure uniform and easily referenceable column names, thereby facilitating subsequent analyses.

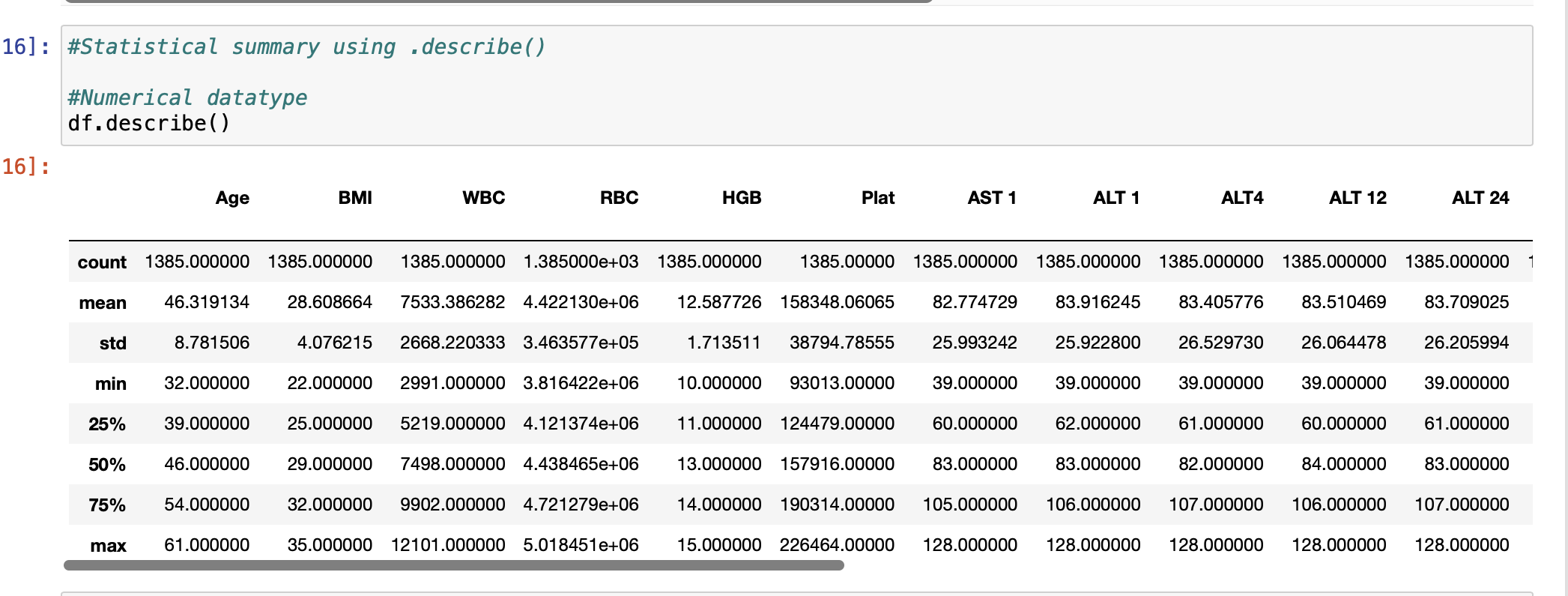
Furthermore, categorical value replacements were executed to enhance data clarity. Notably, the 'Gender' column underwent transformation, with numeric values 1 and 2 being replaced by the corresponding descriptive labels 'Male' and 'Female'. Similarly, for various other categorical columns, the numerical indicators 1 and 2 were converted to more intuitive 'Absent' and 'Present' labels. This refinement contributes to the improved interpretability of the dataset.

## **Exploratory Data Analysis (EDA) & Visualization**

The EDA process begins with a new column ‘Patient ID’ added to uniquely identify each patient.

## **Data Overview and Summaries**

Descriptive statistics and unique value counts were generated using .describe() and .nunique() functions. While duplicate values were identified using .duplicated().sum().

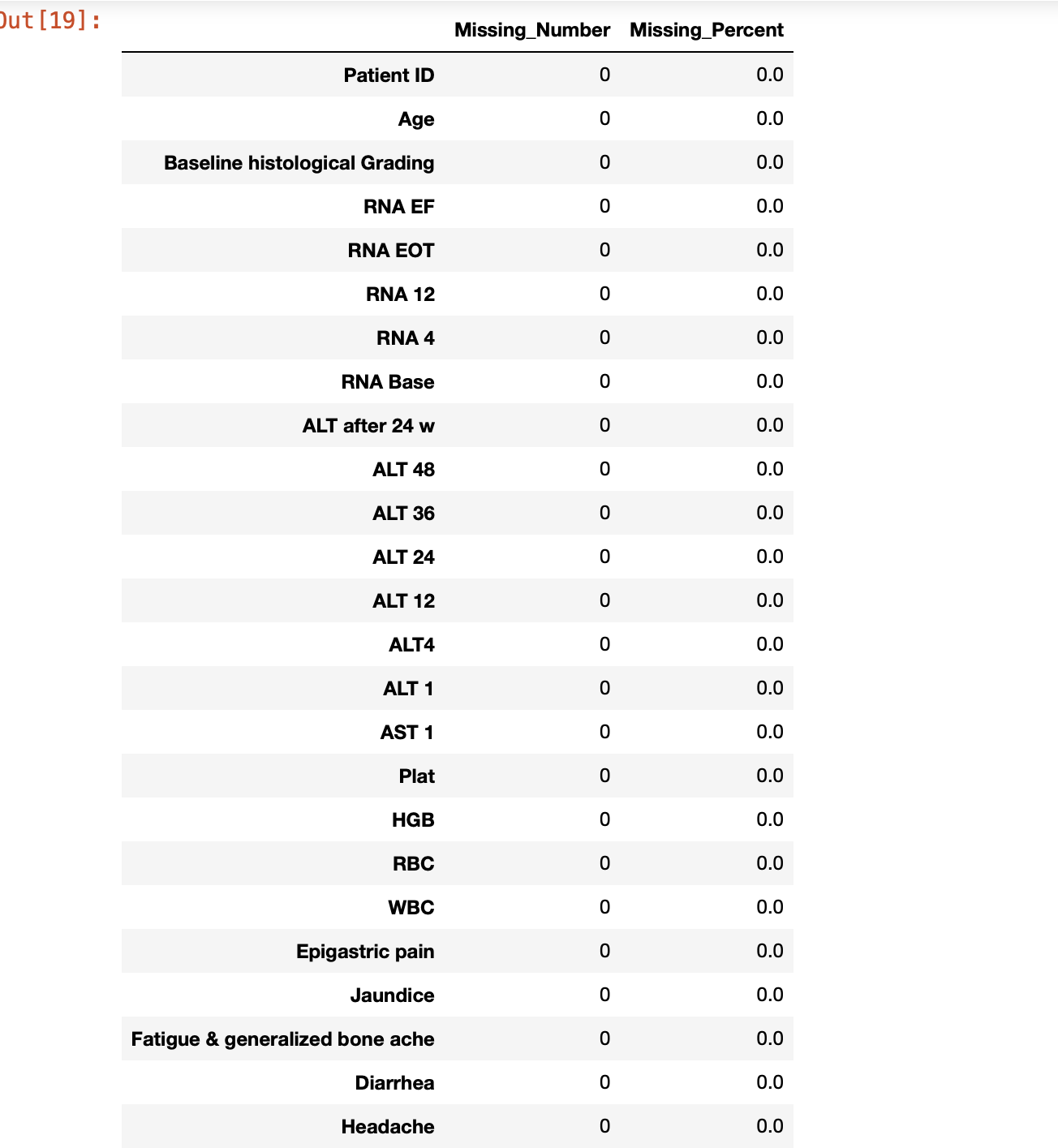


**Figure 1 Descriptive statistics table**

The above image shows the count, mean, standard deviation and other descriptive statistics of each continuous columns.

## **Missing Values**

A function named missing was created to determine the number and percentages of missing values for each attribute. Missing value statistics were presented for the dataset, assisting in further data quality assessments.

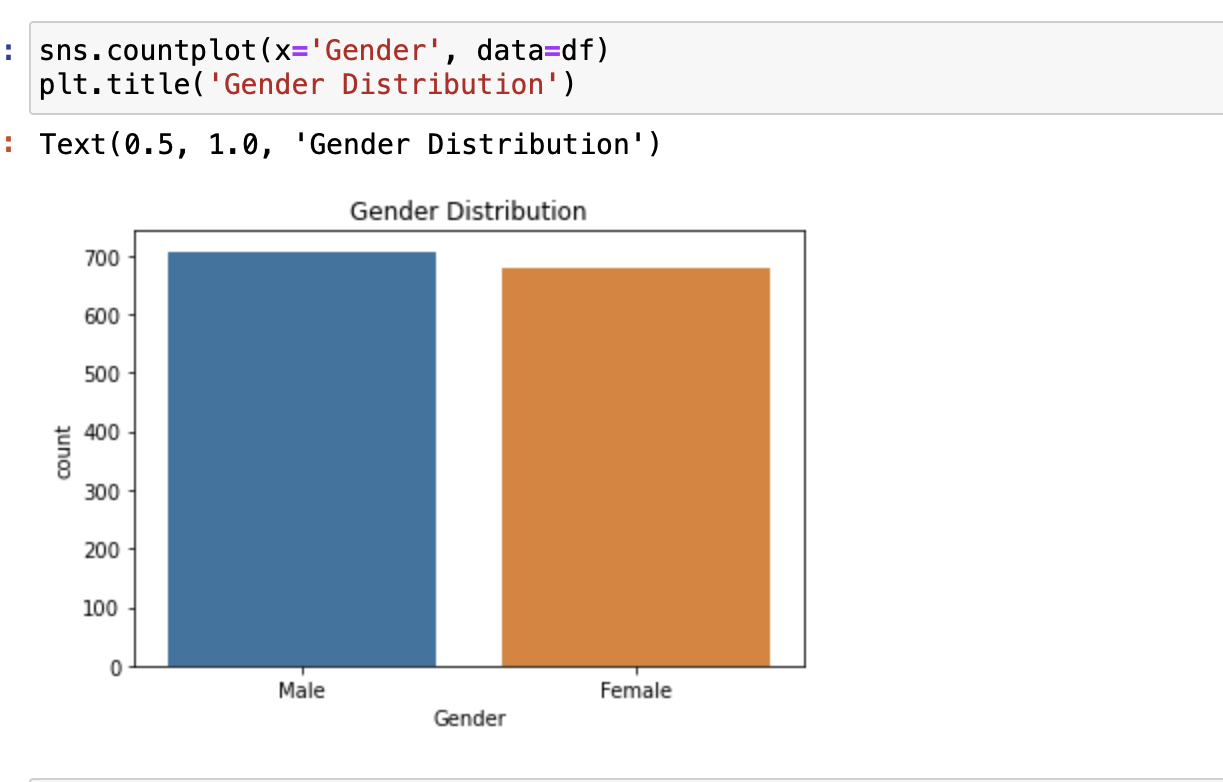


**Figure 2 Missing Values Table**

The above image shows there’s no missing value in my dataset.

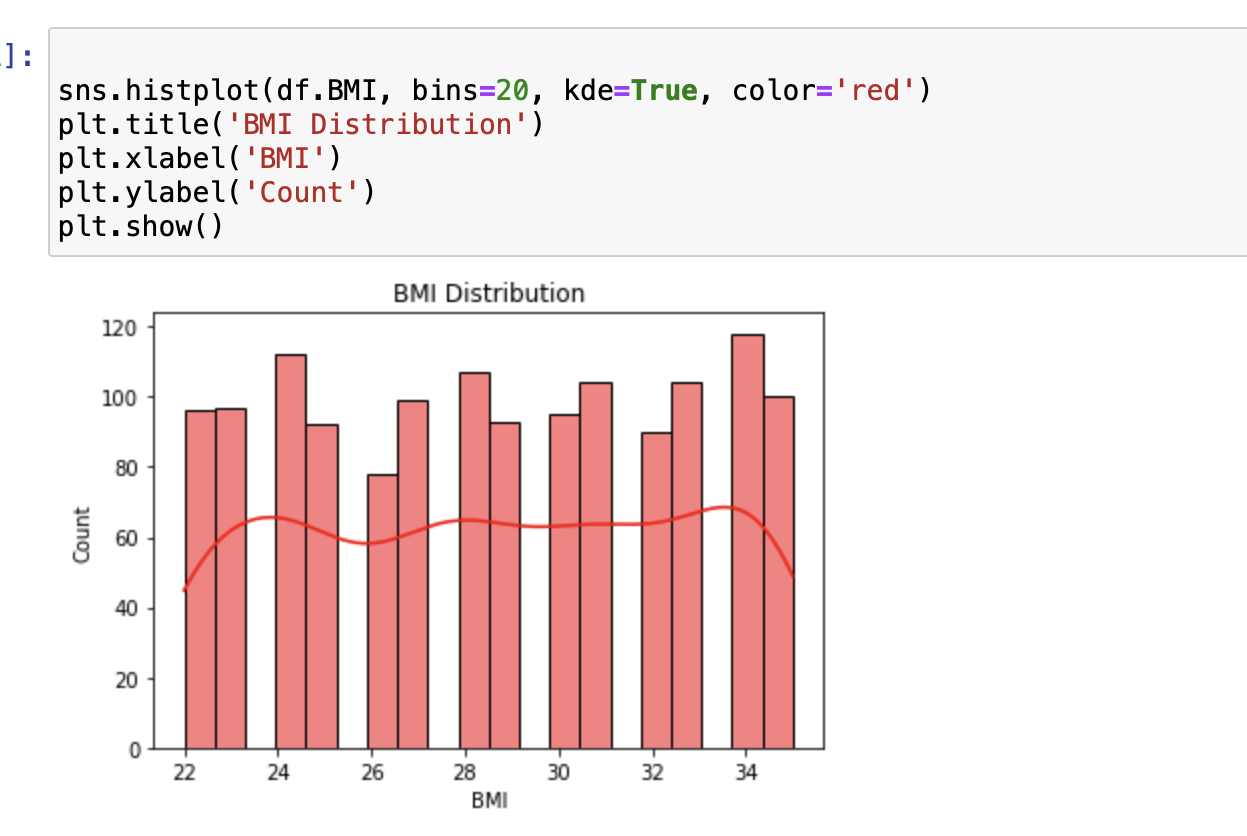
## **Data Visualization**

Gender Distribution: A count plot was used to visualize the distribution of genders among patients.



**Figure 3 Gender Distribution**

BMI Distribution: A histogram with KDE was employed to illustrate the distribution of BMI values.



**Figure 4 BMI Distribution**

Age Distribution: A histogram displayed the frequency distribution of patients' ages.



**Figure 5 Age Distribution**

Age Distribution by Gender: Separate histograms for male and female patients were plotted to compare age distribution.



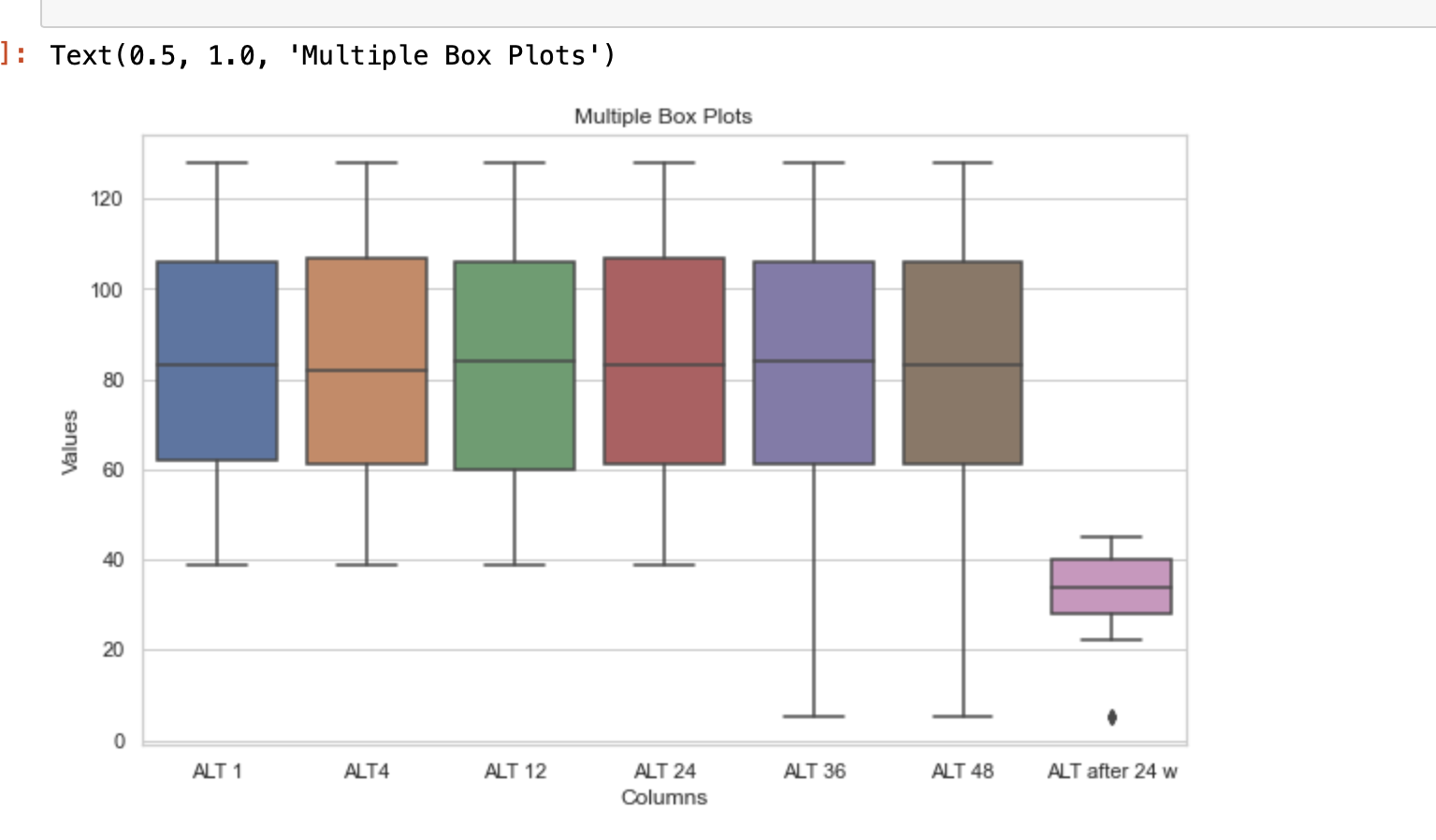
**Figure 6 Age Distribution by Gender**

Scatter Plots: Scatter plots were generated to explore relationships between selected columns.

Kernel Density Estimation: KDE plot was utilized to visualize the probability density of patients' ages.

Box Plots:

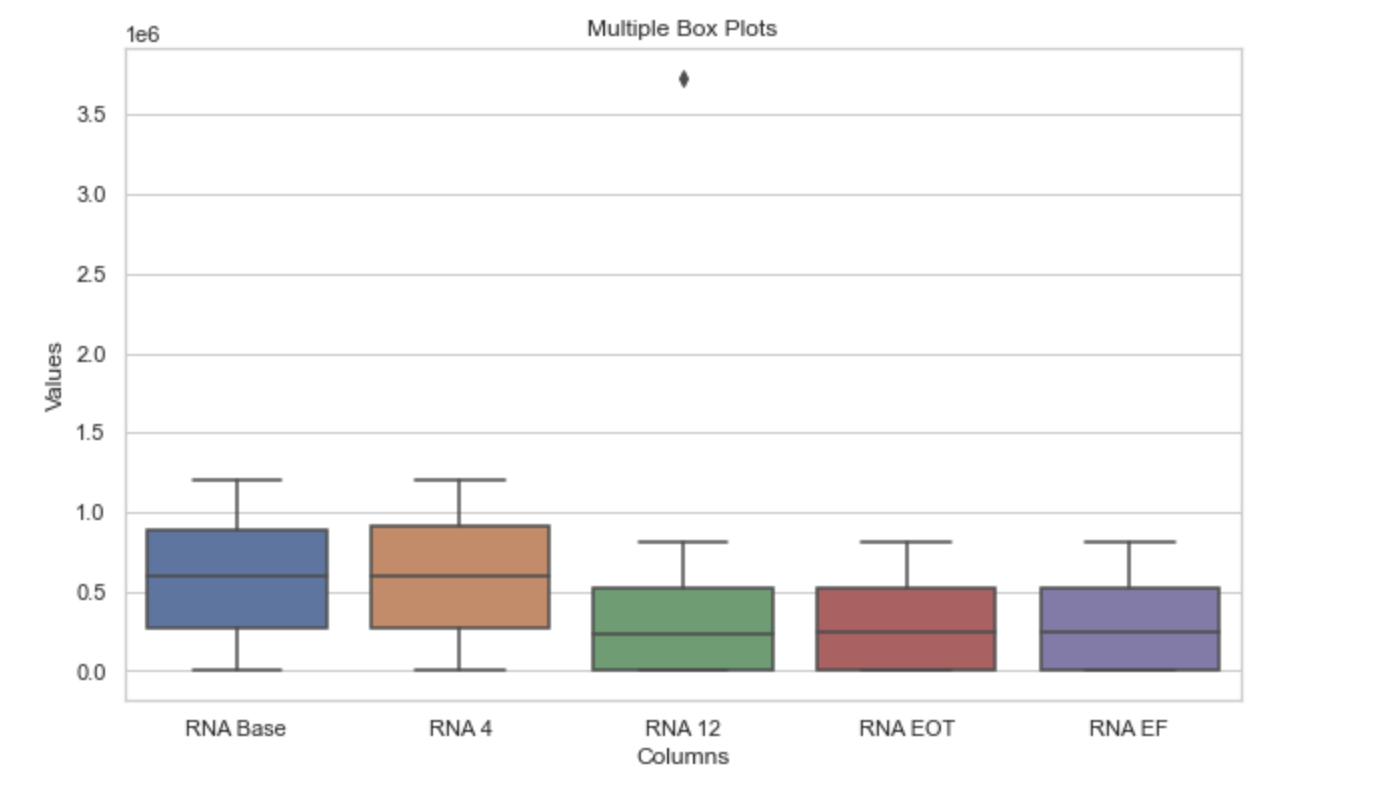
Multiple box plots were created for selected attributes to observe their distributions and identify potential outliers.



**Figure 7 ALT Boxplot**

From the above image there is a large difference between ‘ALT after 24 w’ and other ALT columns.

The selected attributes included 'ALT' measurements at different time points and 'RNA' measurements.



**Figure 8 RNA Boxplot**

The image shows there is significant improvement from the ‘RNA Base’ to ‘RNA EF’.

Pair Plots:

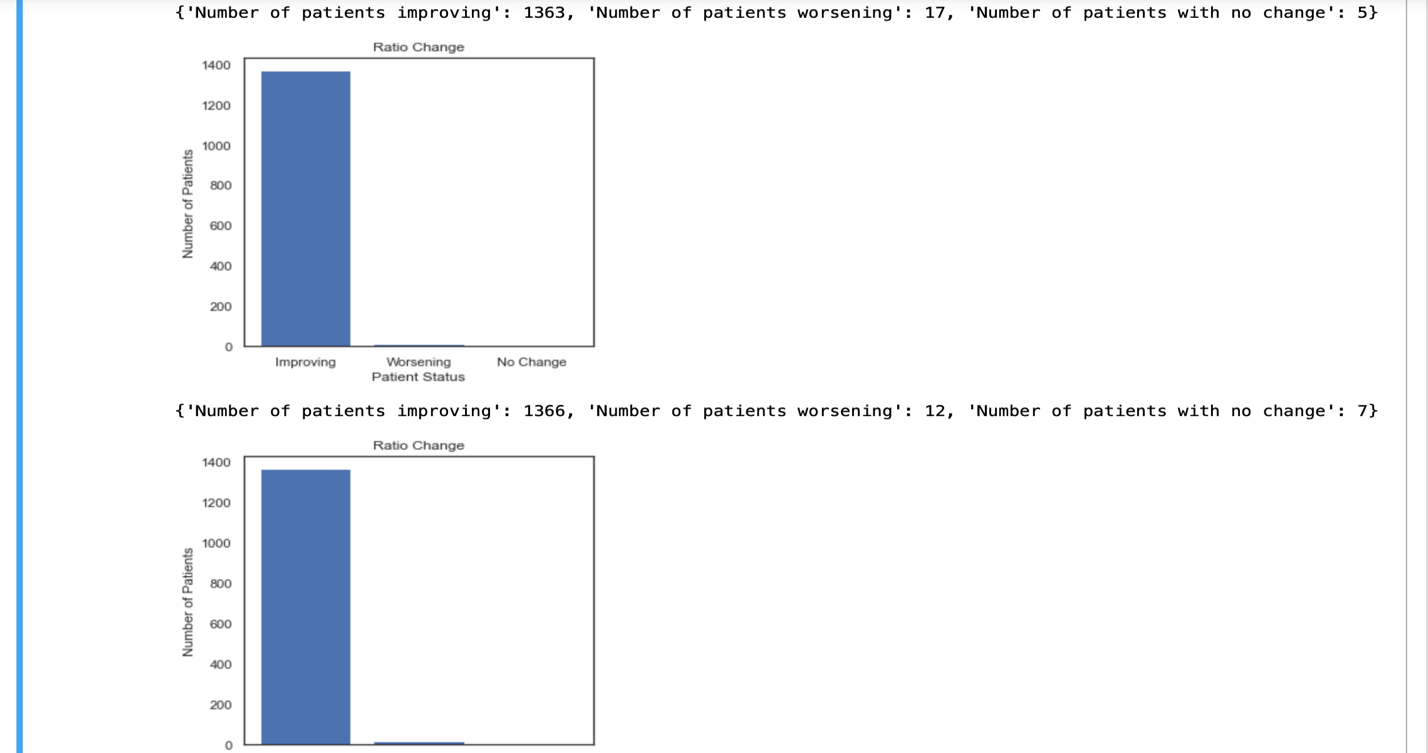
A pair plot was constructed to visualize relationships among 'Age', 'BMI', and 'WBC' attributes using kernel density estimates.



**Figure 9 Pair Plots**

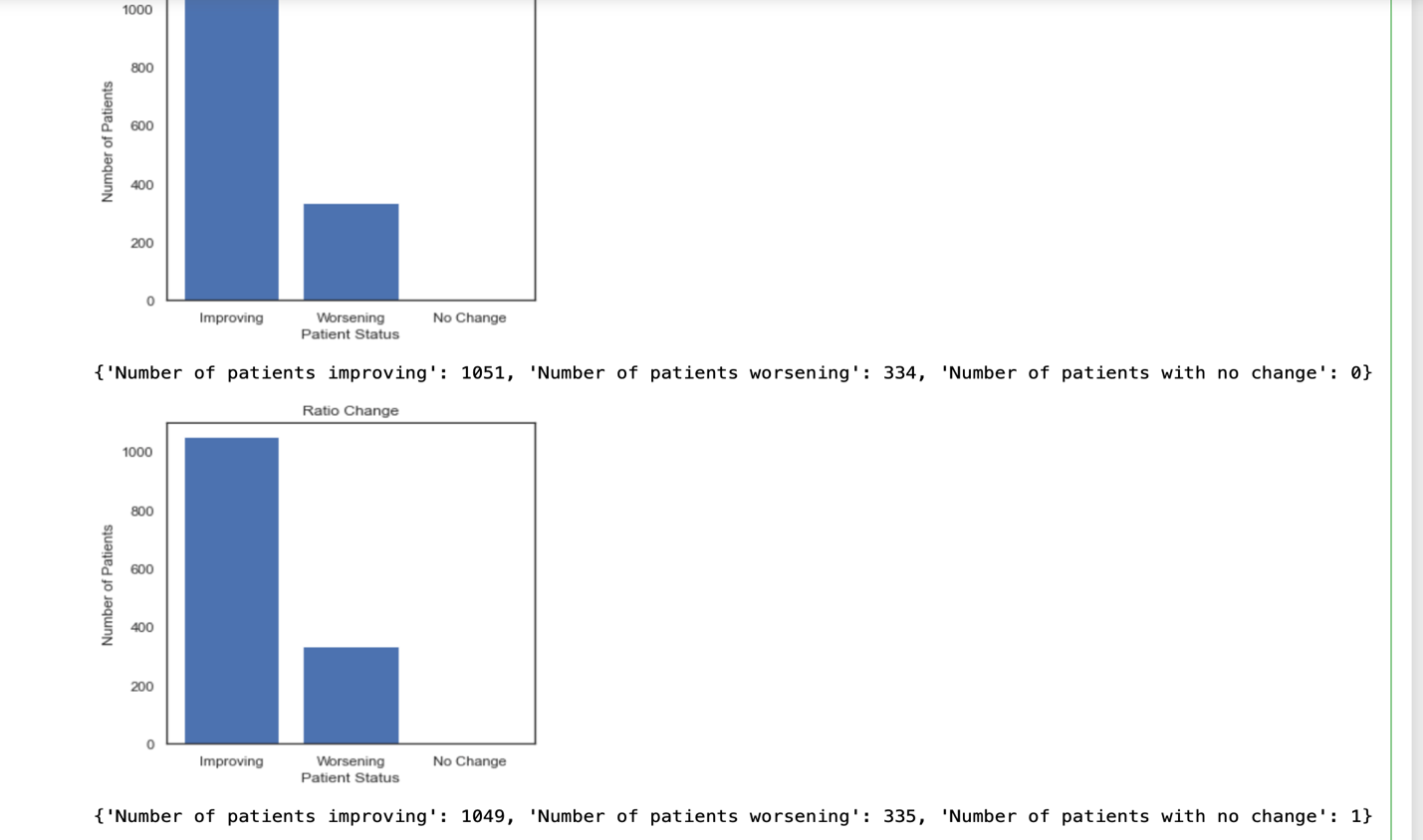
Patient Improvement Analysis:

A function calculate\_patient\_improvement was developed to assess the improvement or worsening of patients based on ALT and RNA measurements.



**Figure 10 ALT improvement bar chart**

Various ALT and RNA measurement pairs were analyzed using the function, and results were presented through bar plots.



**Figure 11 RNA Improvement bar chart**

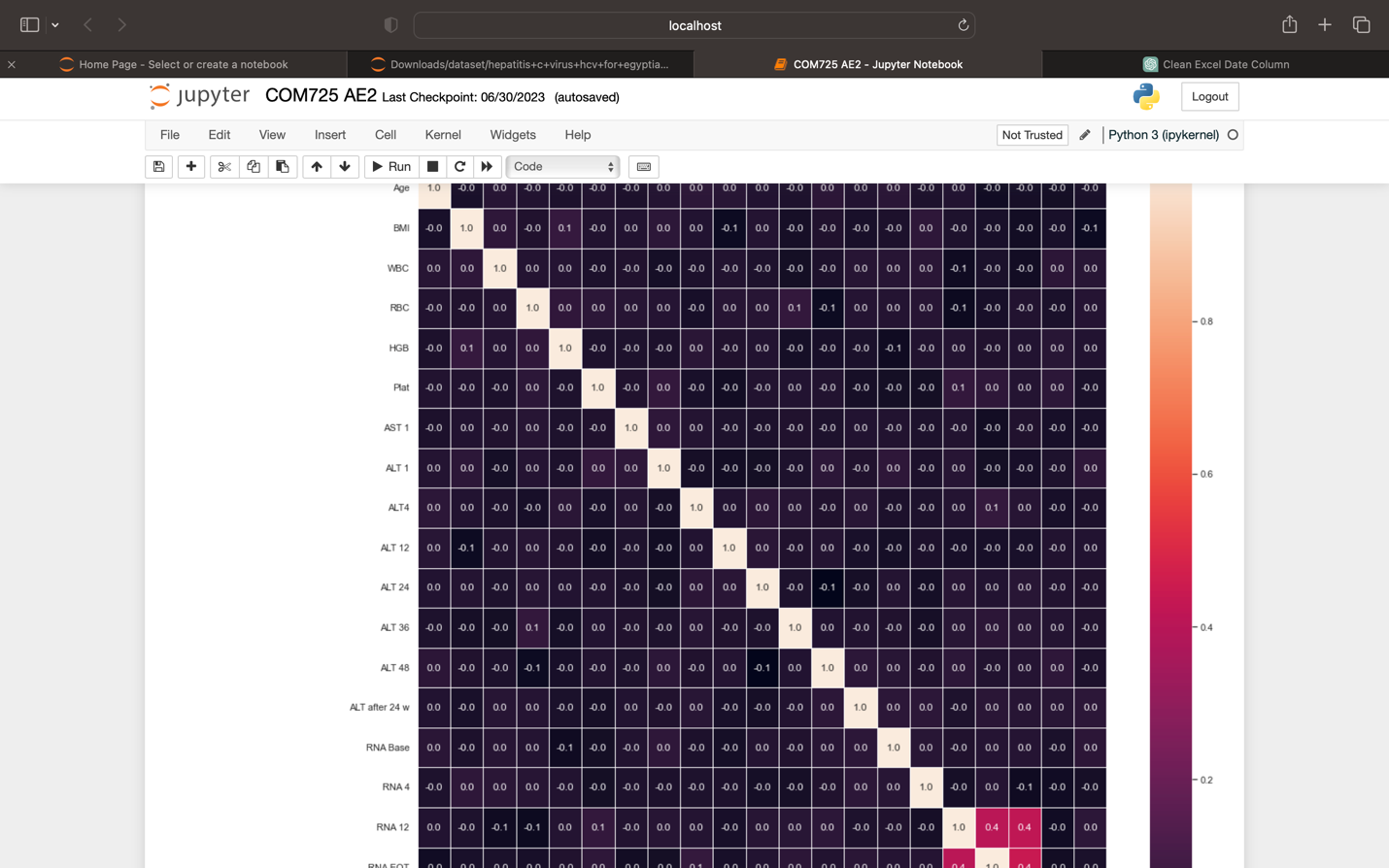
The above image shows RNA patients status how many are improving or getting worse or no change.

This exploratory data analysis has provided valuable insights into the dataset's characteristics, distribution of attributes, and relationships between variables. It serves as a foundation for further analyses and informed decision-making processes.

**Features Selection**

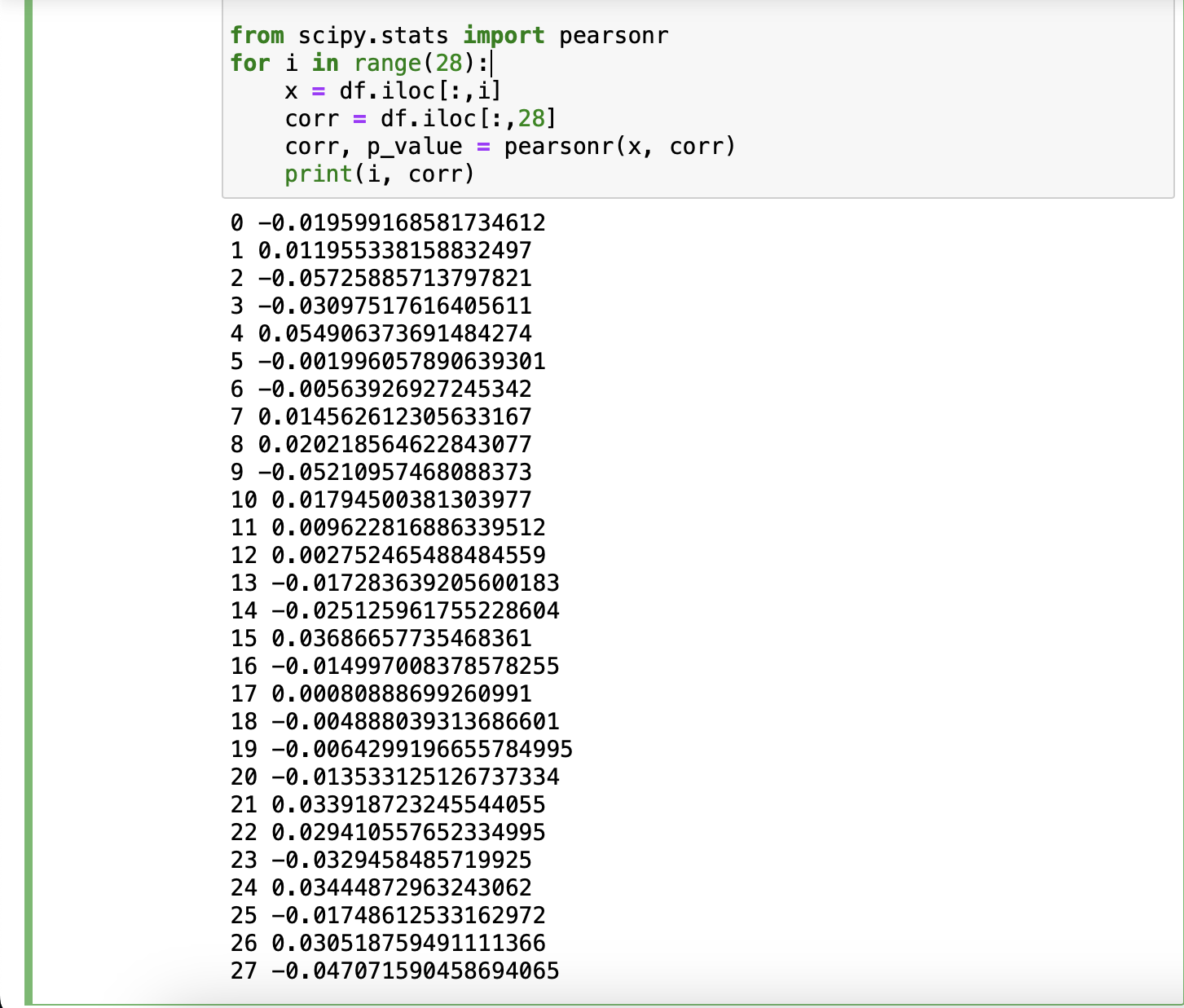
Feature selection is a crucial step in the data preprocessing phase, aiming to identify and retain relevant attributes while removing unnecessary or redundant ones. In this section, we explore the feature selection process conducted on the provided dataset.

A heatmap of the correlation matrix among numerical features was generated using the df\_num.corr() method. The heatmap visualizes the pairwise correlations between attributes, providing insights into potential linear relationships.



**Figure 12 Correlation Table**

I used pearson coefficient to select my features for a better accuracy. I selected all the columns with positive correlation.



**Figure 13 Pearson Correlation**

The positive correlation columns are the columns I chose to use for my machine learning problem. The feature selection process in this section aimed to identify and retain attributes that are most likely to influence the target variable, 'Baseline histological staging'. This selected subset of features can be further utilized in subsequent modeling and analysis steps to build predictive models effectively.

## **Data Pre Processing**

This process is a vital step in preparing the dataset for machine learning algorithms. It involves a series of transformations and manipulations that enhance the quality of the data and make it suitable for modeling.

The LabelEncoder from the sklearn.preprocessing module was used to transform categorical features into numerical labels. A loop iterated through each column in the dataset, and the values were encoded using LE.fit\_transform().

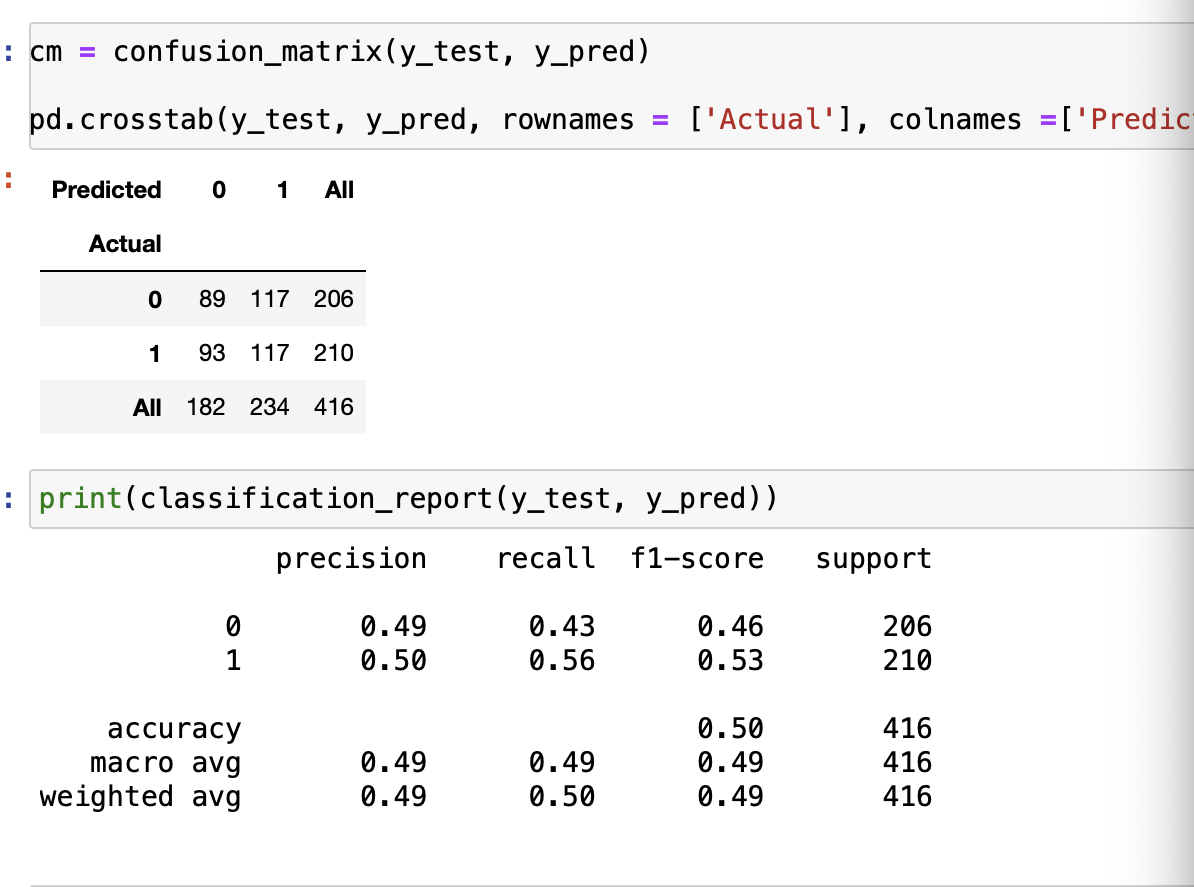
This step converted categorical labels into numerical representations, which is essential for certain algorithms that require numerical input. The original class labels for the target variable were encoded using the LabelEncoder.

This transformation is necessary to ensure consistency in labeling and compatibility with machine learning models. The MinMaxScaler from the sklearn.preprocessing module was applied to perform feature scaling on the dataset.

Feature scaling brings all features to a similar scale, which can improve the performance of some machine learning algorithms. The dataset was split into features (X) and the target variable (y) prior to scaling. The train\_test\_split function from the sklearn.model\_selection module was used to split the preprocessed data into training and testing sets.The scaled features (X\_scaled) and the target variable (y) were split into training and testing subsets, with a test size of 30%.

## **Machine Learning Algorithm**

I opted to employ the Random Forest classifier with a parameter setting of 'n\_estimators' as 100 for both the multi-class and binary classification scenarios. The accuracy achieved by the multi-class classifier stands at 26.9%, while the binary classifier demonstrates an accuracy of 51.2%. These results indicate that the binary classifier outperforms the multi-class classifier in terms of accuracy.



**Figure 14 Confusion Matrix**

The image above shows the precision , recall, f1-score, support and accuracy of the binary classifier.

For the decision tree classifier multi class the accuracy is 25.96%, Precision is 26%, recall 26%, f1-score also 26%. For the binary decision tree, it got an accuracy of 46.6%, precision of 47%, recall of 47% and f1-score of 47%.

# **Excel**

I duplicated my dataset to utilize it for both Tableau and Weka. Using Excel, I performed the same data cleaning and column adjustments that I executed in Jupyter using code. Additionally, I prefixed the class labels with 'F' to ensure that Tableau and Weka recognize them as part of a classification problem.

# **Tableau**

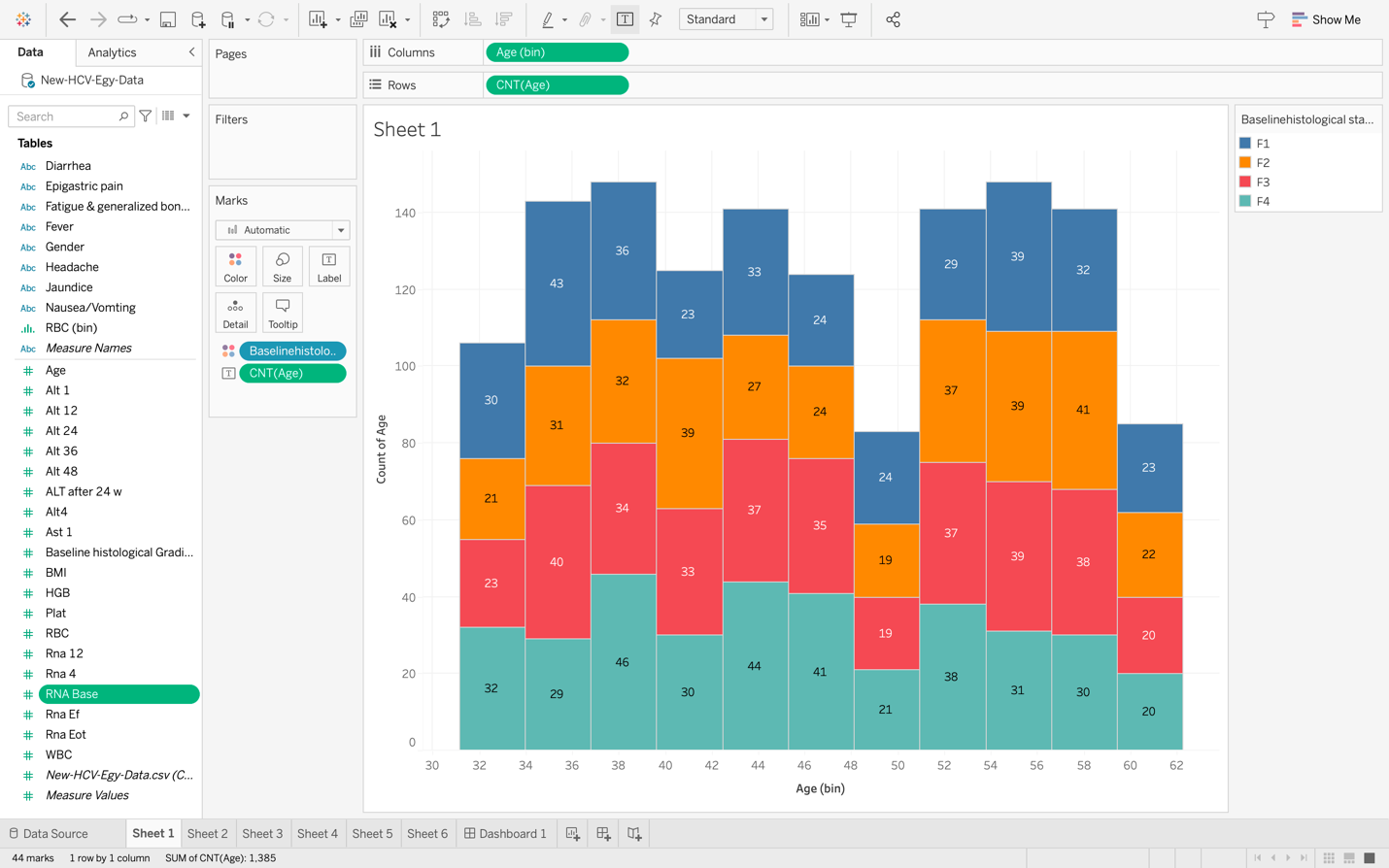
Tableau is a powerful and widely used data visualization tool that empowers individuals and organizations to explore, analyze, and present data in a visually appealing and insightful manner. It provides an intuitive platform for creating interactive and dynamic visualizations, dashboards, and reports, making complex data easy to understand for both technical and non-technical audiences.

In this project, I have undertaken the visualization of a dataset that contains classes divided into both multi-class and binary-class categories.

## **Multi-Class Visualization**

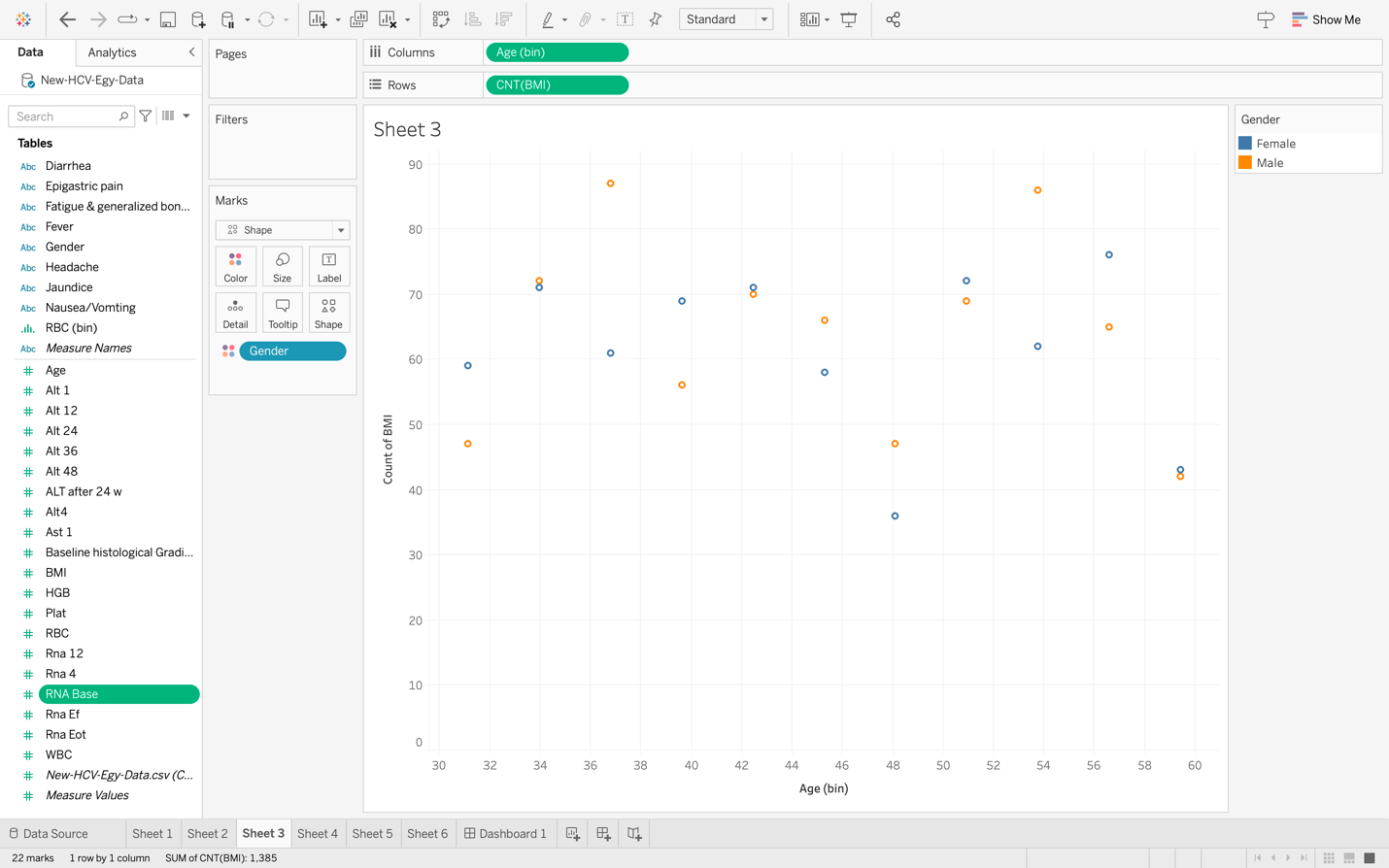
The multi-class visualization focuses on the classes labeled as F1 to F4. This visualization comprises six distinct sheets and one comprehensive dashboard.

The first sheet presents a visualization of age bins against the four classes, providing a breakdown of the age distribution within each class.



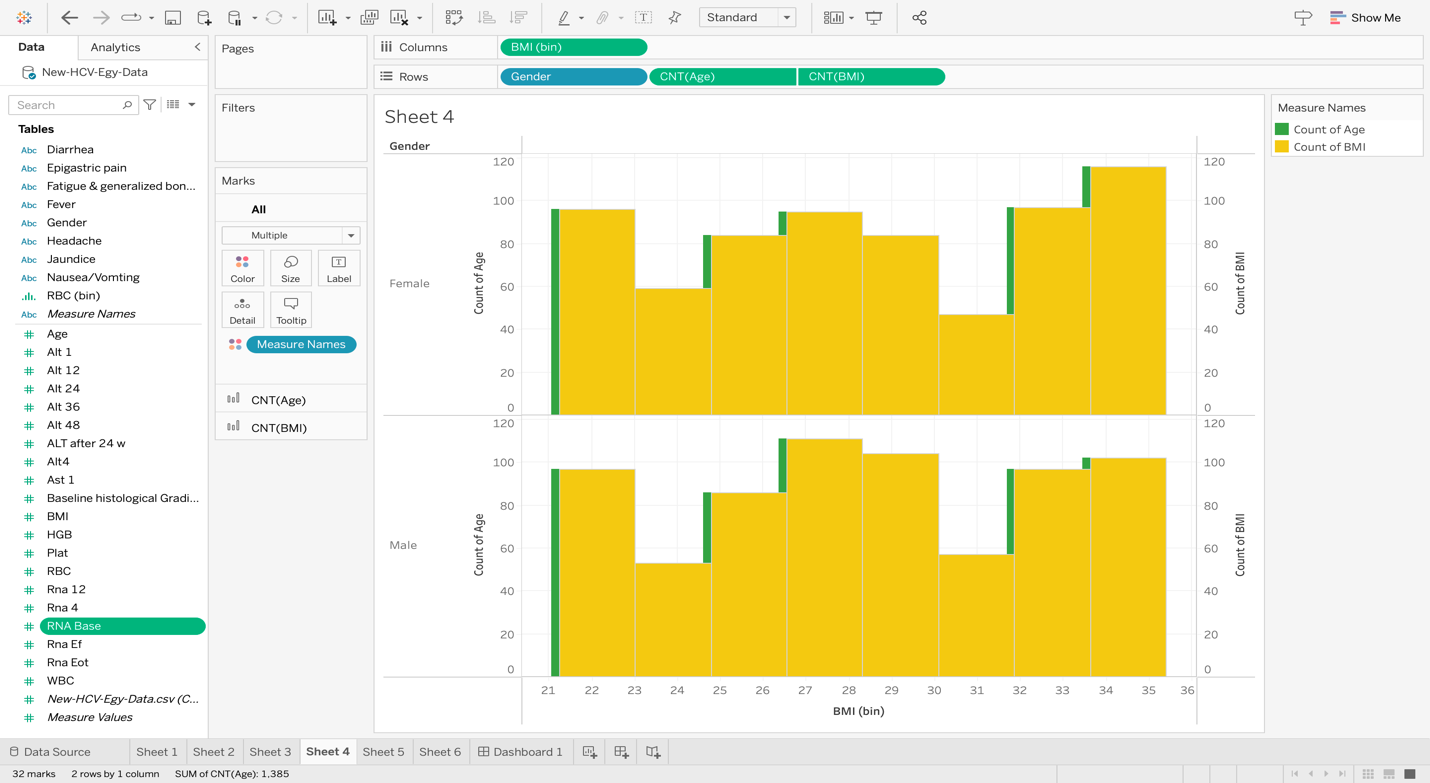
**Figure 15 Age- Multi Class Relationship**

The third sheet showcases the correlation between age (binned) and BMI.



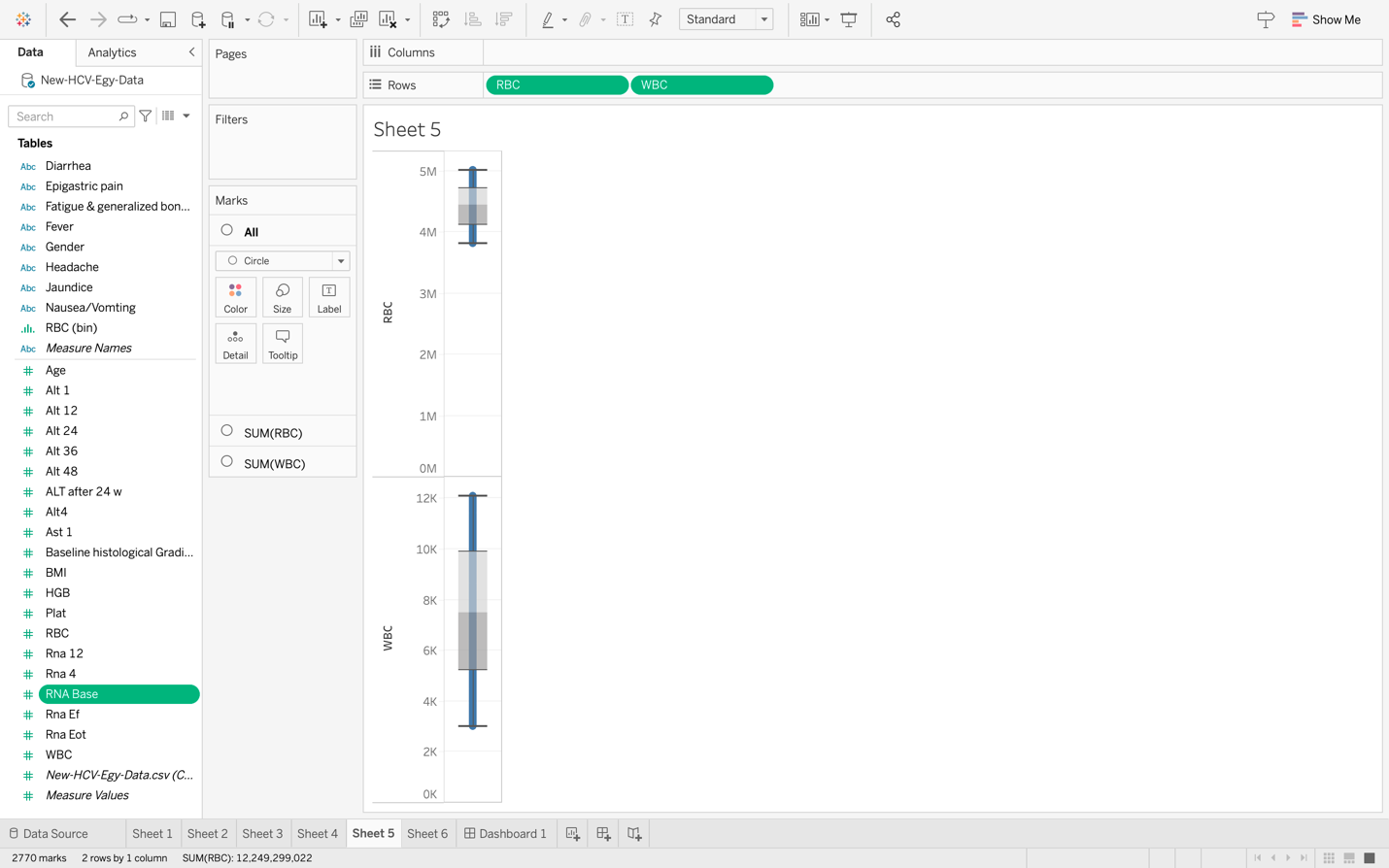
**Figure 16 Age against BMI**

The fourth sheet features a histogram depicting the count of Age and the count of BMI, distinguished by the Gender column.

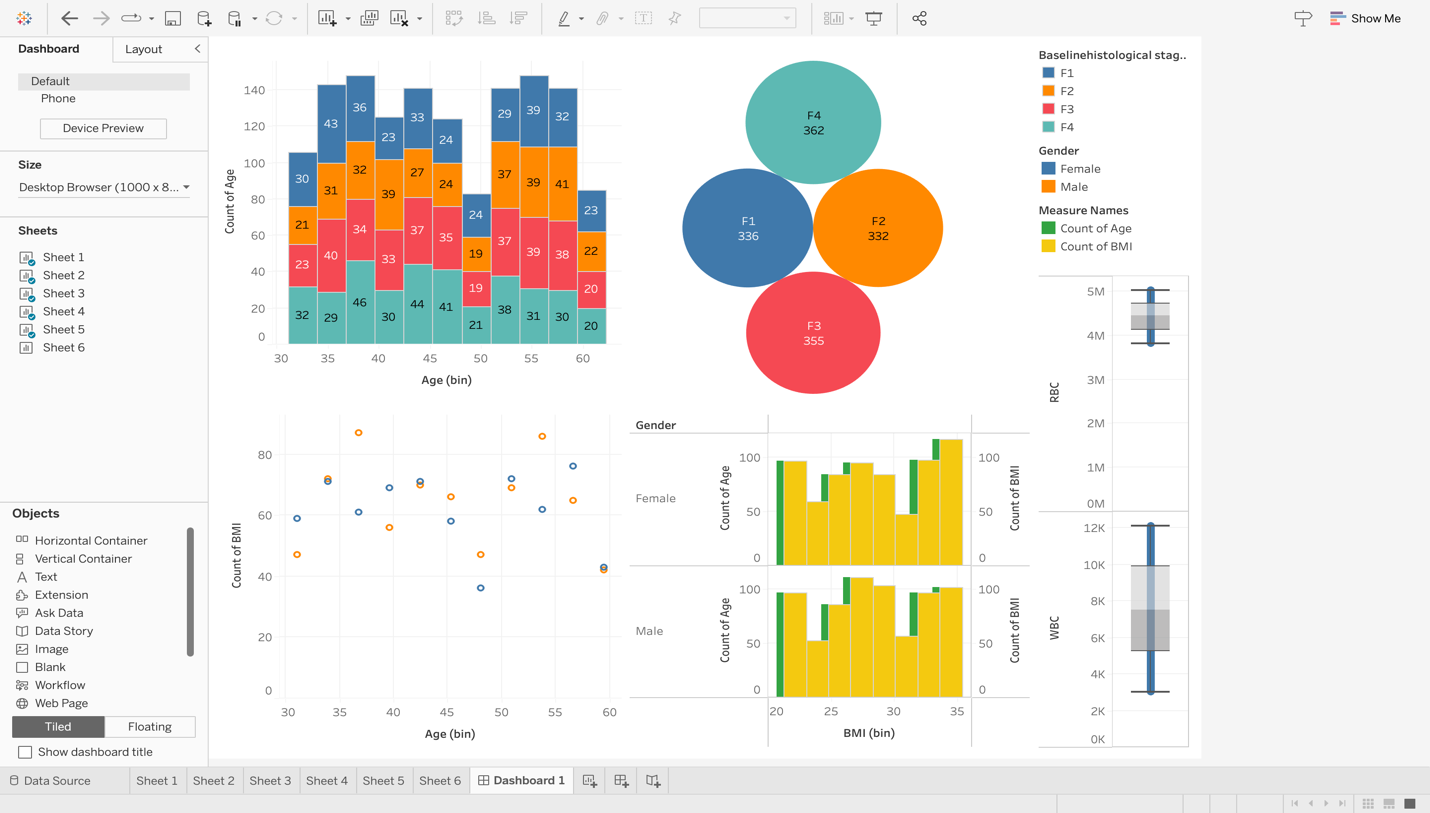


**Figure 17 Age-BMI by Gender Relationship**

Lastly, the fifth sheet employs a box plot to illustrate the summation of RBC and WBC, adding further insight into the dataset.



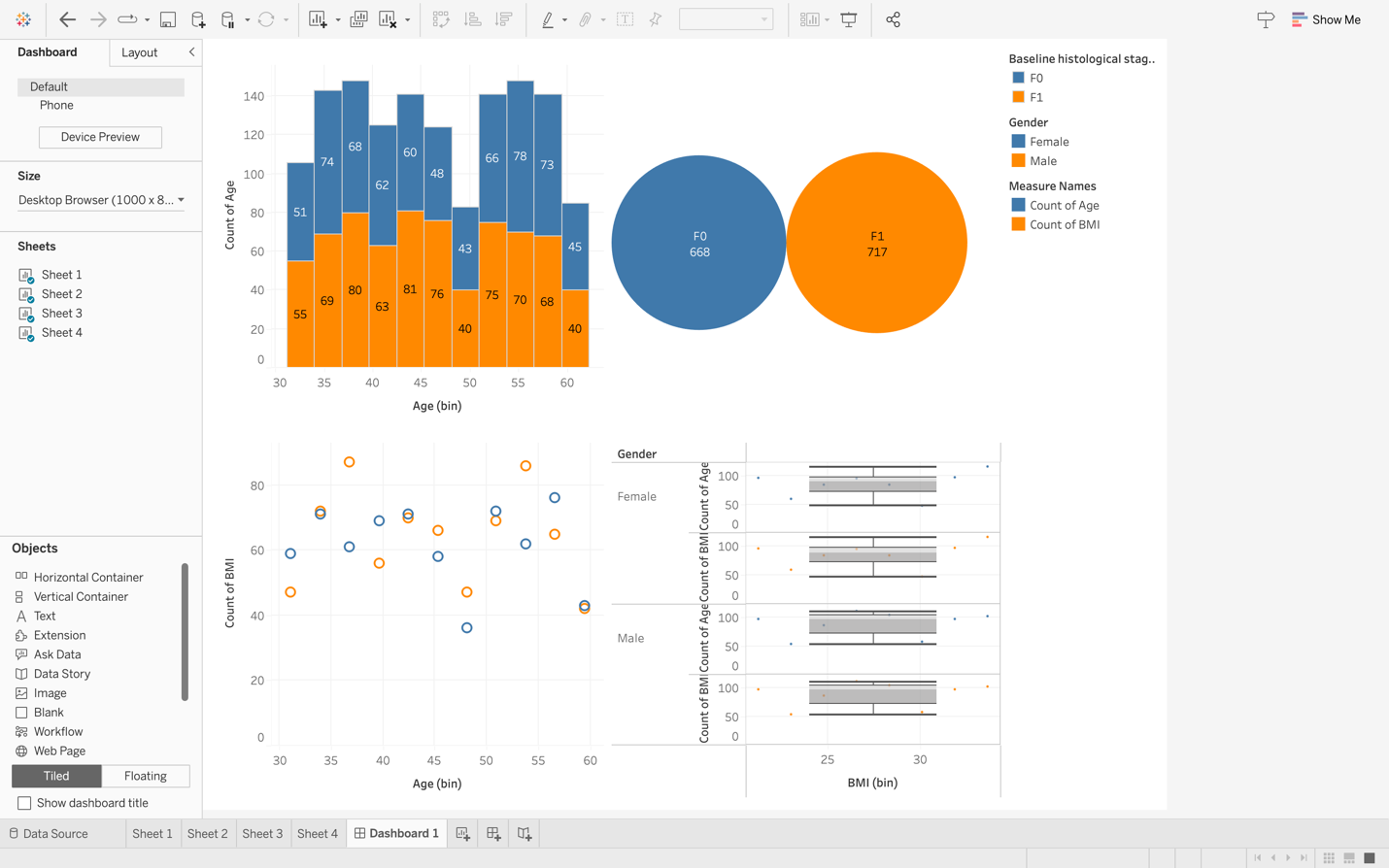
**Figure 18 RBC-WBC Boxplot**



**Figure 19 Multi Class Dashboard**

## **Binary-Class Visualization**

The binary class visualization has same with multiple class visualization expect that the class is only F0 and F1.



**Figure 20 Binary Class Dashboard**

This is the dashboard for the binary class having done same for multi class.

# **Weka**

Weka is a versatile and widely used open-source machine learning software suite that provides a comprehensive set of tools for data preprocessing, classification, regression, clustering, association rules, and more.

## **Workflow**

Data Loading and Preprocessing: I Imported my dataset into Weka. Utilize the data preprocessing tools to clean, transform, and preprocess the data.

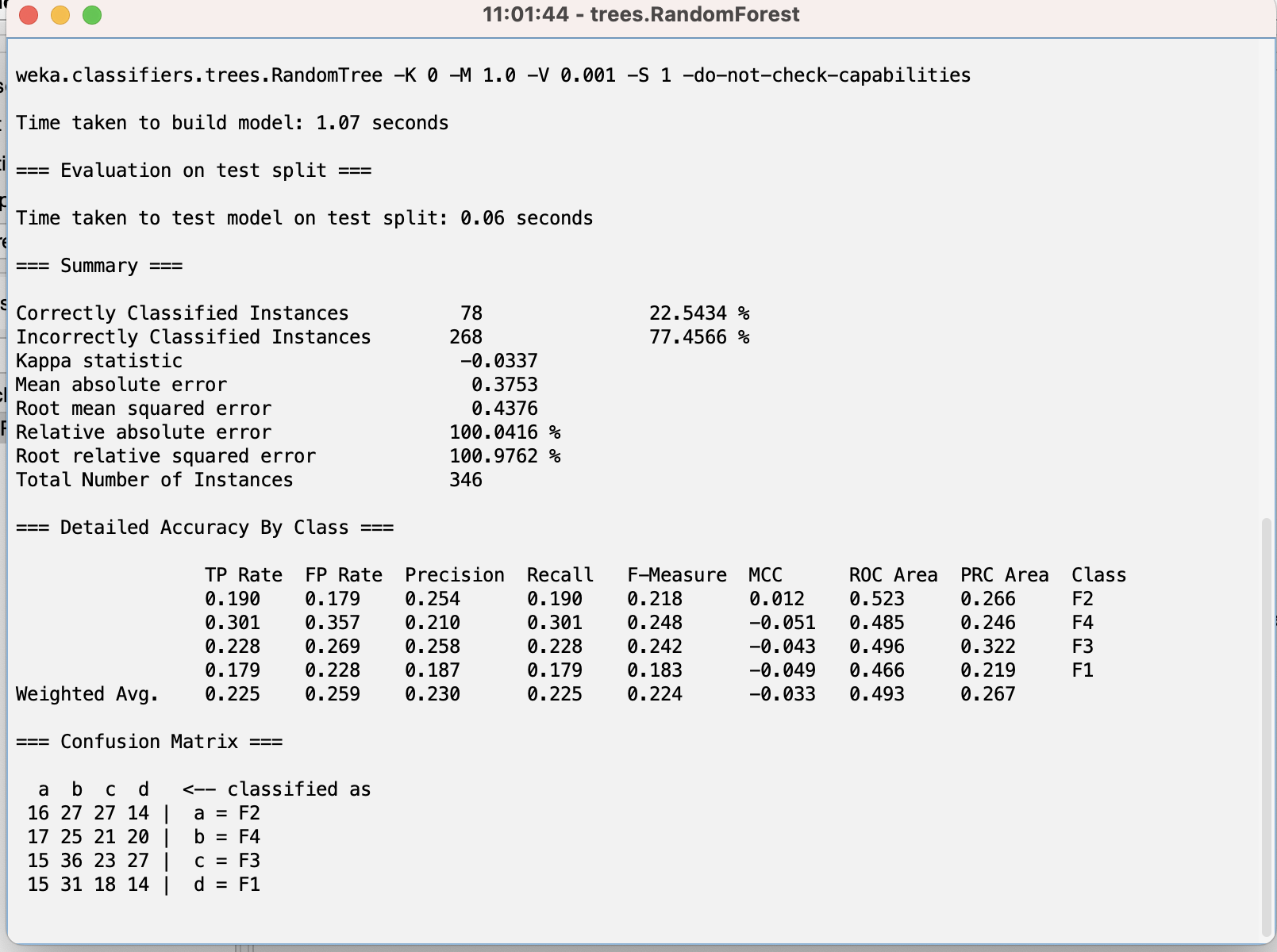
Algorithm Selection: I Choose a suitable algorithm (RandomForest) based on my task.

Model Building: I trained the selected algorithm using the preprocessed data.

Model Evaluation: I Employed cross-validation or separate training and testing datasets to evaluate your model's performance. Weka provides various metrics like accuracy, precision, recall, F1-score, etc.

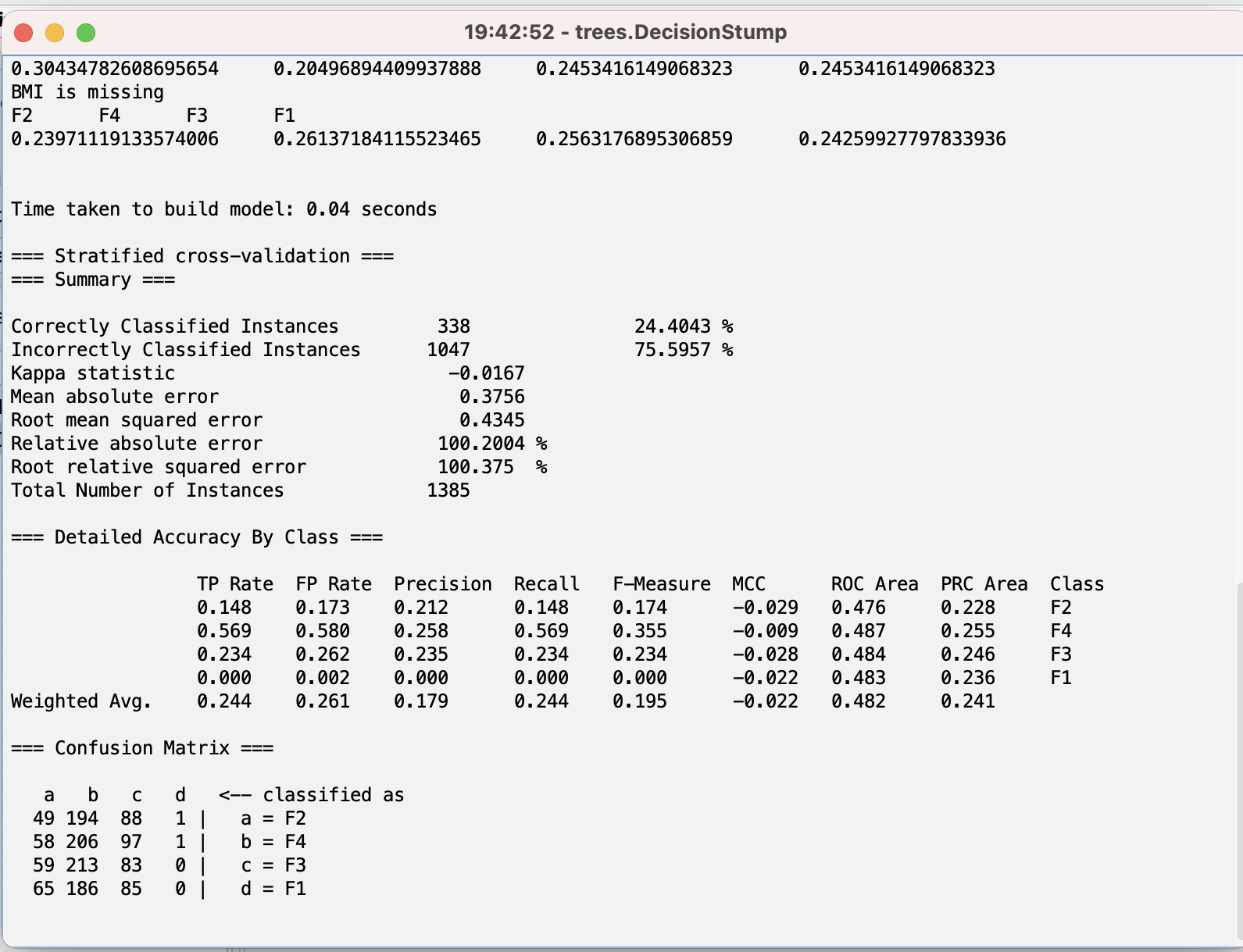
## **Multi Class Weka**

For this multi class I used Random forest algorithm to evaluate my model performance and below is my result. I used 75% of my dataset to train this model. The accuracy is 22.5%, Precision 23%, Recall 22.5% and F-1 measure is 22.4%



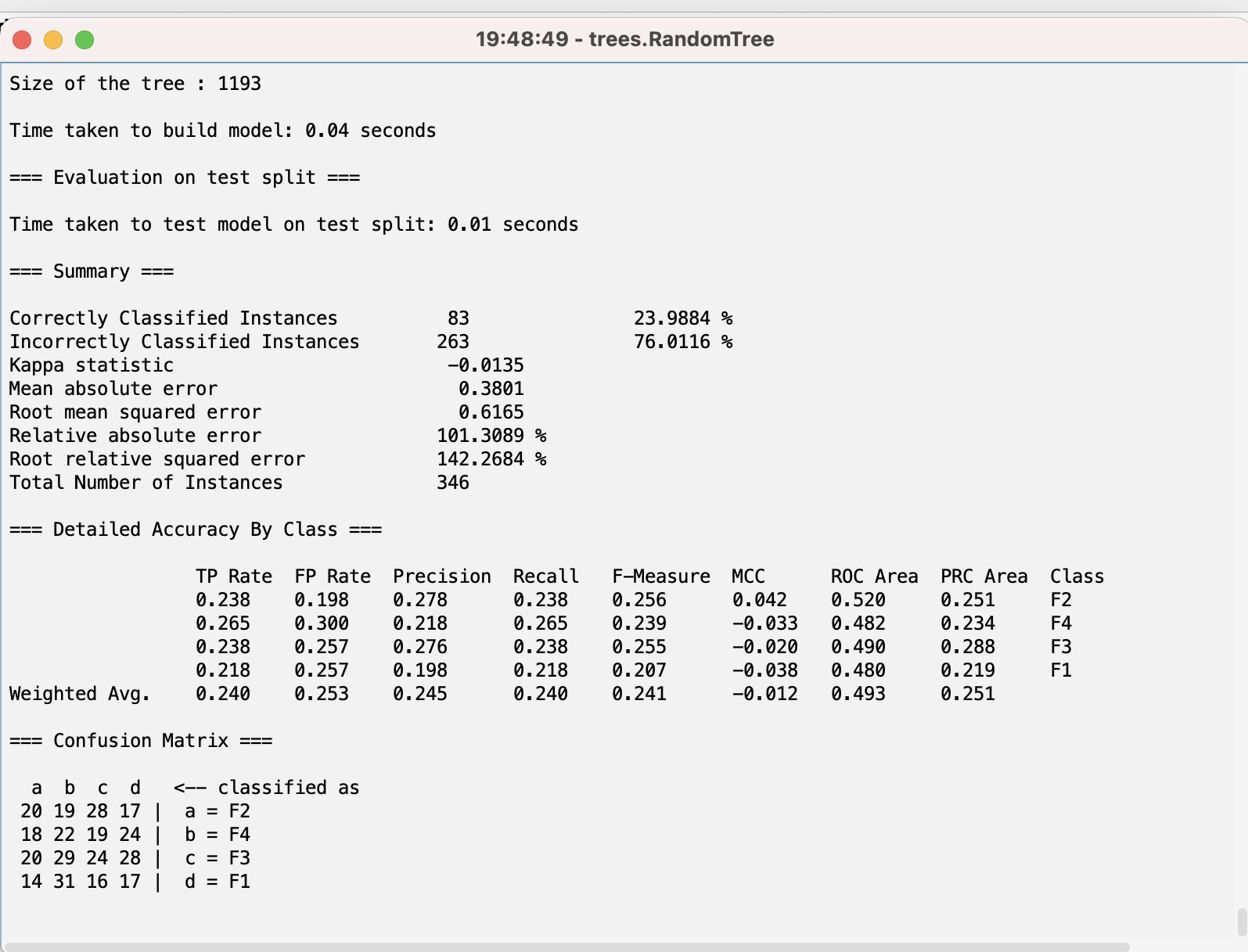
**Figure 21 Weka Multi Class**

For this multi class I used Decision Stump to evaluate my model performance and the results are 24.4% accuracy, 17.9% Precision, 24.4% recall and F1-Measure is 19.5%.



**Figure 22 Weka Multi Decision Stump**

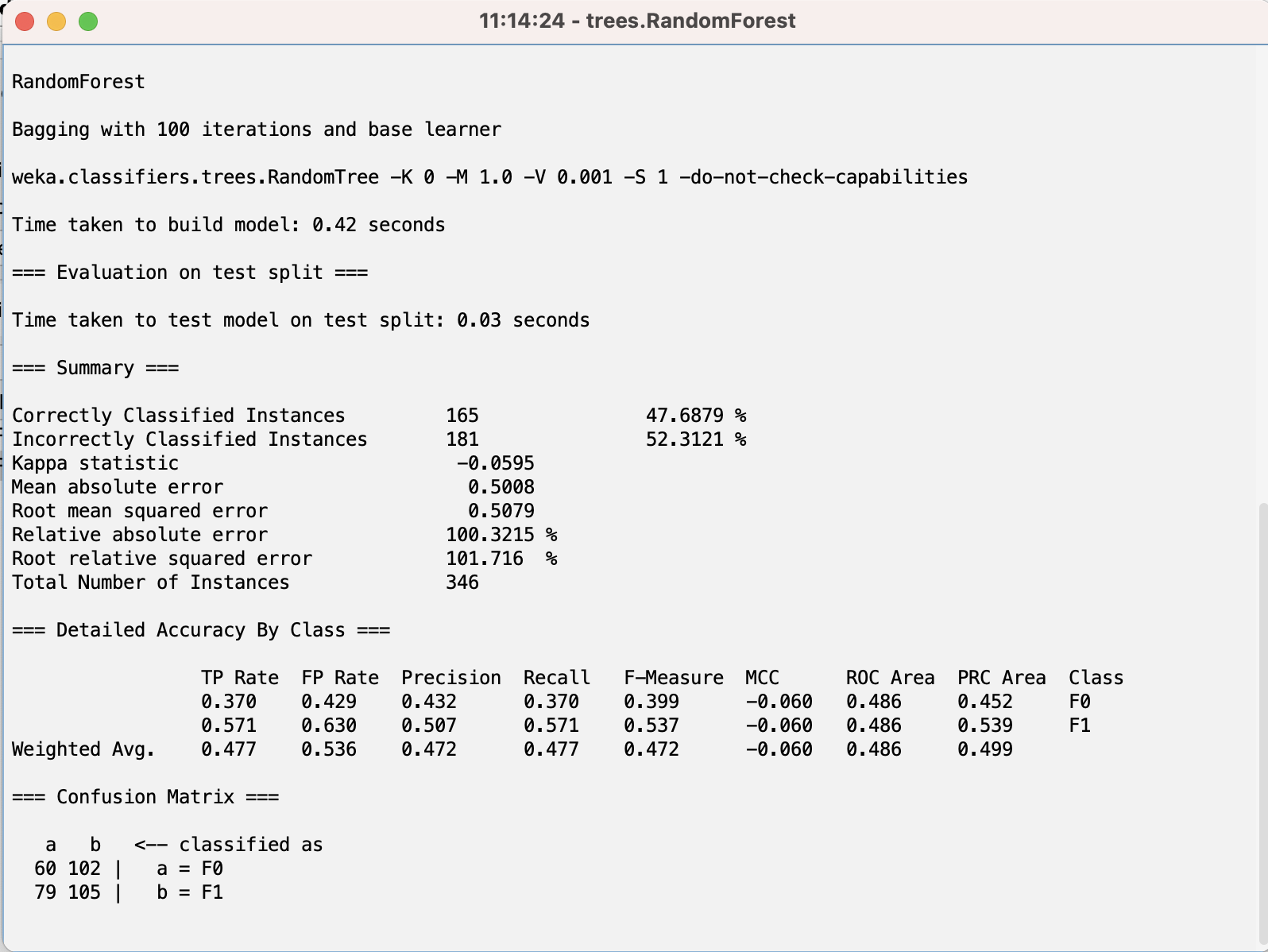
For this multi class I used Random Tree to evaluate my model evaluate and the result are 23.9% accuracy, 24.5% Precision, 24% Recall and 24.1% F1-Measure.



**Figure 23 Weka Multi Random Tree**

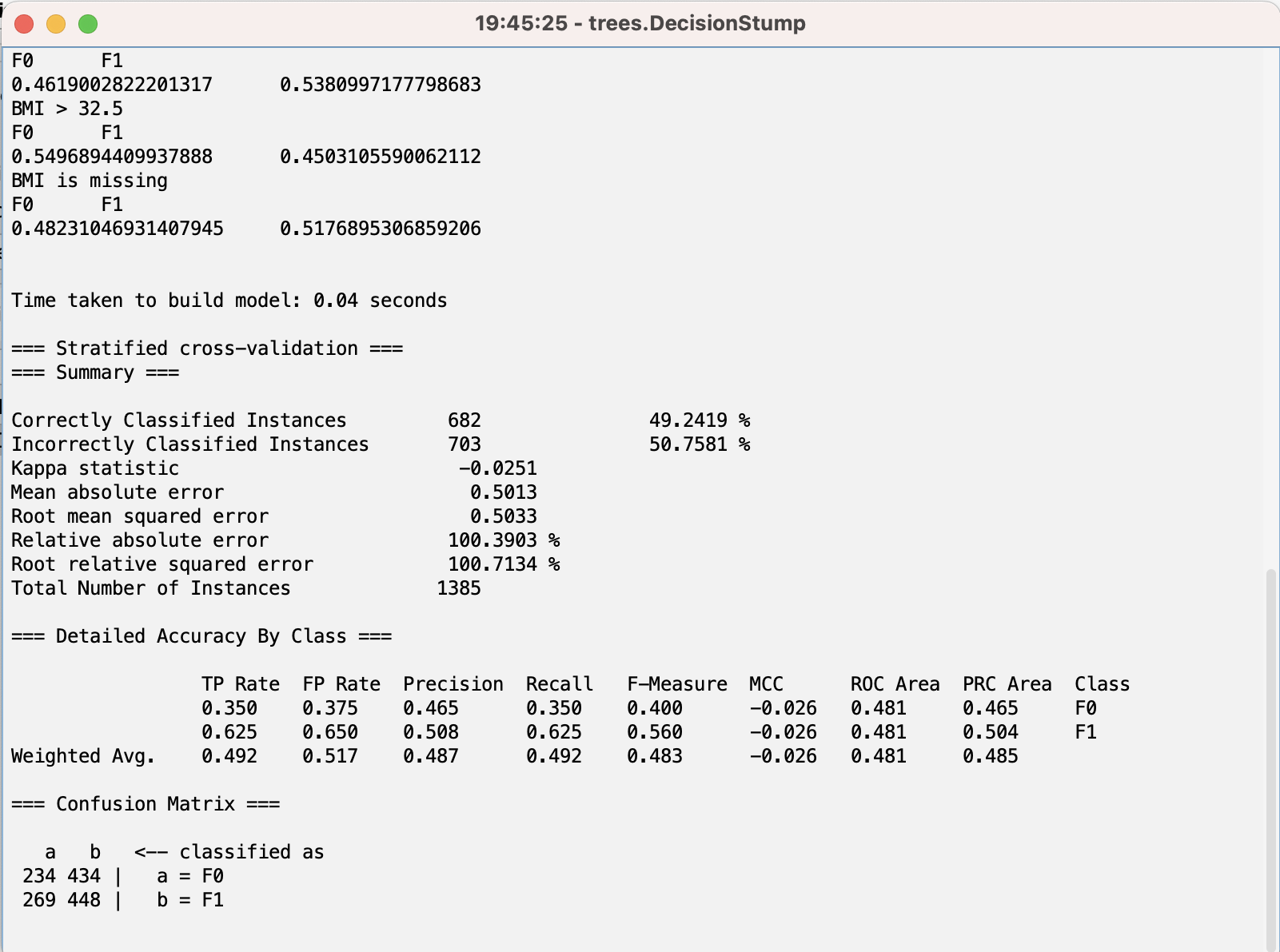
## **Binary Class Weka**

For this binary class I used Random forest algorithm to evaluate my model performance and below is my result. I used 75% of my dataset to train this model. The accuracy is 47.6%, Precision 47.2%, Recall 47.7% and F-1 measure is 47.2%



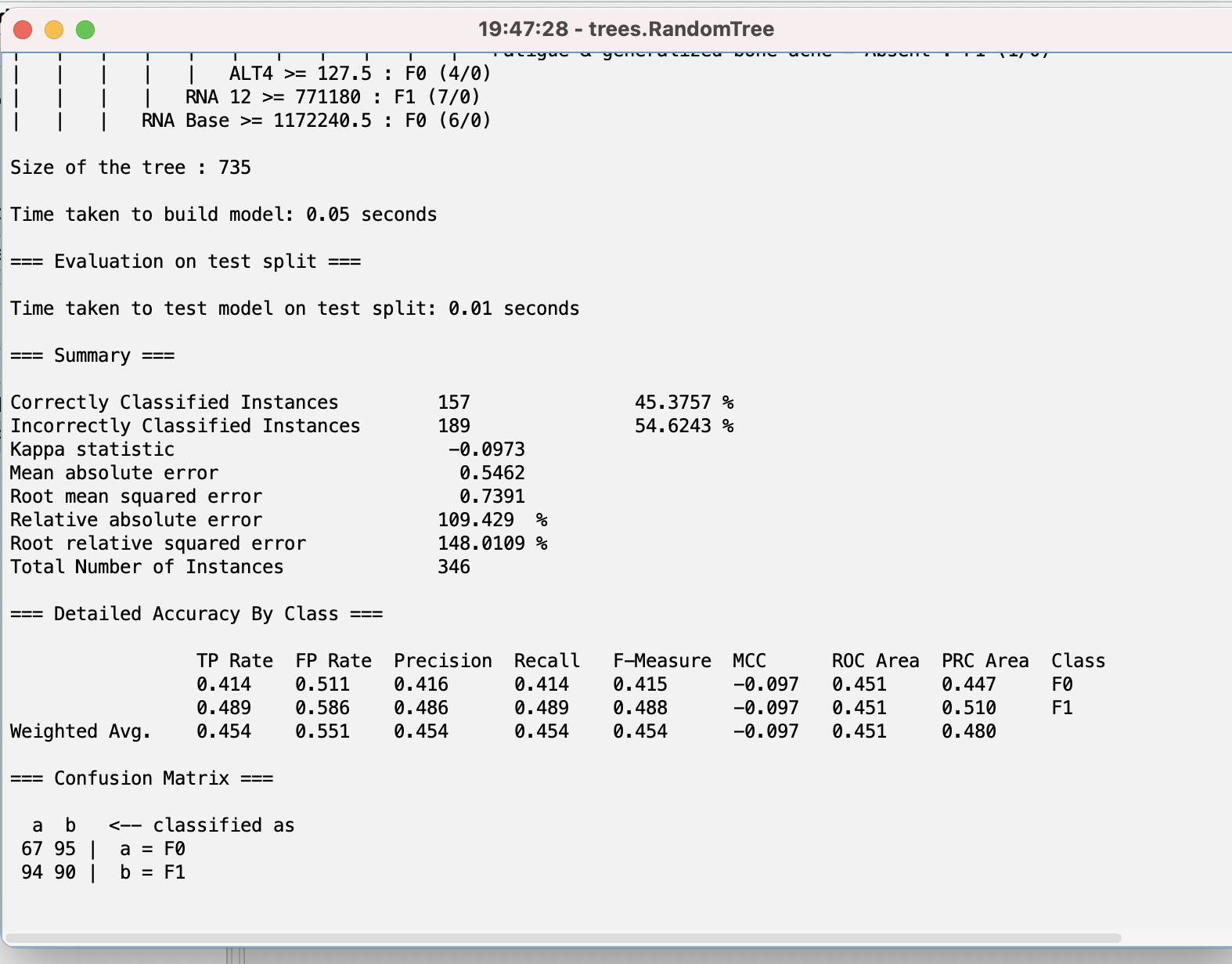
**Figure 24 Weka Binary Class**

For this binary class I used Decision Stump to evaluate my model performance and the results are 49.2% accuracy, 48.7% Precision, 49.2% recall and F1-Measure is 48.3%.



**Figure 25 Weka Binary Decision Stump**

For this binary class I used Random Tree to evaluate my model evaluate and the result are 45.3% accuracy, 45.4% Precision, 45.4% Recall and 45.4% F1-Measure.

****

**Figure 26 Weka Binnary Random Tree**

# **Conclusion**

In conclusion, this project successfully explored and analyzed the Hepatitis C Virus dataset, provided valuable insights into the data, and applied machine learning techniques for classification tasks. The project showcases the importance of data preprocessing, visualization, feature selection, and model evaluation in achieving accurate predictions and generating meaningful insights from complex datasets. The analysis also reveals a notable similarity between the accuracy obtained in Weka for both multi-class and binary classifications and the accuracy achieved in the Jupyter notebook analysis for different models’ algorithm.

# Bibliography

Satish CR Nandipati, C. X. (2020). Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques. *AQII Publication* , 89-100.