|  |
| --- |
| Module: |
| Data Management in Healthcare |
| Coursework 2 |

|  |
| --- |
| Author  5/23/2024 |

Contents

[Part A 3](#_Toc167444606)

[Introduction 3](#_Toc167444607)

[Tool Selection 3](#_Toc167444608)

[Data Import and Manipulation 4](#_Toc167444609)

[Visualizations and Interpretation 5](#_Toc167444610)

[Code snippet 6](#_Toc167444611)

[1. Count Plot of Observations for Each Variable 6](#_Toc167444612)

[2. Box Plot of Systolic and Diastolic Blood Pressure 7](#_Toc167444613)

[3. Histogram of Systolic and Diastolic Blood Pressure 8](#_Toc167444614)

[4. Violin Plot of Systolic and Diastolic Blood Pressure 8](#_Toc167444615)

[Code snippet 9](#_Toc167444616)

[1. Line Plot of Blood Pressure Readings for Patient 12 10](#_Toc167444617)

[2. Box Plot of Systolic and Diastolic Blood Pressure Readings 10](#_Toc167444618)

[3. Histograms of Systolic and Diastolic Blood Pressure Readings 11](#_Toc167444619)

[Code snippet 12](#_Toc167444620)

[1. Scatter Plot with Regression Line 12](#_Toc167444621)

[2. Line Plots of Blood Pressure Readings Over Time for Different Patients 13](#_Toc167444622)

[3. Joint Plot of Systolic and Diastolic Blood Pressure Readings 16](#_Toc167444623)

[4. Hexbin Plot of Systolic vs. Diastolic Blood Pressure Readings 16](#_Toc167444624)

[Code snippet 17](#_Toc167444625)

[Strip Plot of Systolic Blood Pressure Readings by Patient ID 17](#_Toc167444626)

[Strip Plot of Diastolic Blood Pressure Readings by Patient ID 18](#_Toc167444627)

[Code snippet 19](#_Toc167444628)

[1.Average Systolic Blood Pressure per Patient 19](#_Toc167444629)

[Average Diastolic Blood Pressure per Patient 20](#_Toc167444630)

[Literature and Case Studies 20](#_Toc167444631)

[Conclusion 21](#_Toc167444632)

[Part B 0](#_Toc167444633)

[1](#_Toc167444634)[st](#_Toc167444634) [dataset(lung cancer): 0](#_Toc167444634)

[Introduction 0](#_Toc167444635)

[Methodology 0](#_Toc167444636)

[1. Loading the dataset: 0](#_Toc167444637)

[2. Data Preprocessing: 2](#_Toc167444638)

[3. Applying algorithm (Classification used): 2](#_Toc167444639)

[4. Results 2](#_Toc167444640)

[5. Confusion matrix 4](#_Toc167444641)

[6. Discussion 4](#_Toc167444642)

[7. Conclusion 4](#_Toc167444643)

[2nd dataset(heart disease): 4](#_Toc167444644)

[Introduction: 4](#_Toc167444645)

[Methodology: 5](#_Toc167444646)

[1. Loading the dataset: 5](#_Toc167444647)

[1. Data Preprocessing: 6](#_Toc167444648)

[2. Applying algorithm (Cluster used): 6](#_Toc167444649)

[3. Results 7](#_Toc167444650)

[4. Discussion 8](#_Toc167444651)

[5. Conclusion 8](#_Toc167444652)

[References 0](#_Toc167444653)

# Part A

# Introduction

latestgenerationofhealthcaremanagementisbasedondataandinformation,thatiscrucialindeterminingkeydecisions. New technologies of monitoring enable the continuous level measurement of patients’ conditions, creating a vast quantity of patient data on a daily basis, thus presenting a great opportunity to enhance the measurement of patient benefit. In this case, this report intends to show the effectiveness of representing health data with a view to analyzing the last thirty blood pressure readings for different patients.

It is common knowledge that blood pressure is one of the more important factors that determine the health of the cardiovascular system. Hypertension being a chronic condition, constant monitoring and evaluation assist in identifying trends, efficacy of treatment as well as assist in avoiding complications like heart attacks and some forms of strokes (Zakir, 2023). This project entails loading in a dataset of blood pressure levels, cleaning up this data for analytical purposes, and finally generating different graphical representations for use in understanding the trends within these levels. These visualizations are beneficial for healthcare providers to help assist patient care, detect possible harm or disease early, and incorporate a patient-specific approach as needed.

1. To read and process a given data set, which contains some blood pressure data information.
2. In order to generate and represent visuals that would be informative of the current trend analysis in the available data collected.
3. As a way of supporting the findings made in this research, the following literature review and case studies have been used.

Through the achievement of these objectives, this report will demonstrate the benefits of using visualisation in improving the quality of patient care in the current healthcare institutions and the significance of Big data analytics in the present day healthcare facilities.

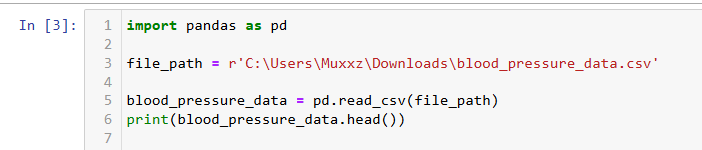
# Tool Selection

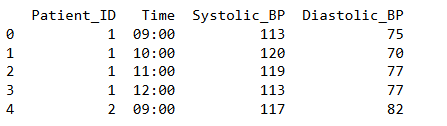
For this assignment, Python was selected as the primary tool for data analysis and visualization for several reasons;

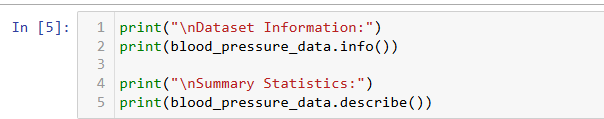
1. **Widely Used**: Python is widely used in data science and health care analytics because it offers vast libraries (Pandas, Matplotlib, Seaborn) necessary to aggregation, analyze, and visualize large data sets (Lavanya, 2023). It is beneficial when one wishes to practice in the future as a healthcare professional.
2. Rich Ecosystem: Python brings various high-level packages namely, Pandas for data analysis, Matplotlib and Seaborn for data visualization and Scikit-learn for machine learning (Bhardwaj 2023, November).
3. Ease of Use: It is quite easy to use, understand and explain the code for beginners, but still challenging enough for the professional programmers because of the wide variety of resources and support available on the Internet for Python.
4. Integration: Python has seamless compatibility with databases and various file formats, can work with web services and APIs making it useful throughout the data analysis process (Saabith, 2021).
5. These advantages make it very appropriate to use Python when dealing with blood pressure data import process, management and visualization to gain insight.

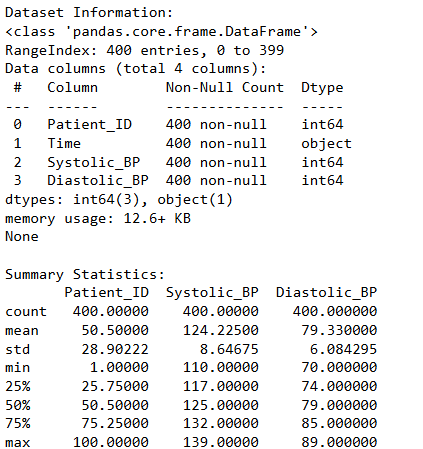
Data Import and Manipulation

The dataset was imported and processed using Pandas.









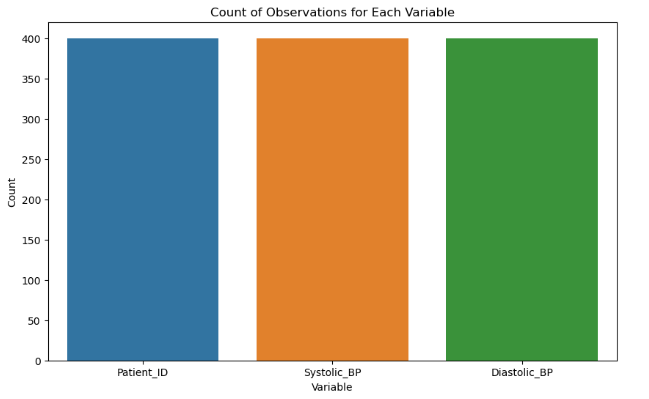
the output is showing information of the dataset, which is the number of columns , the datatypes in each column. the output indicates no missing values. The summary statistics indicate that systolic blood pressure values range from 110 to 139 and diastolic blood pressure values range from 70 to 89, with moderate variability within these measurements as evidenced by their standard deviations.

Visualizations and Interpretation

## Code snippet



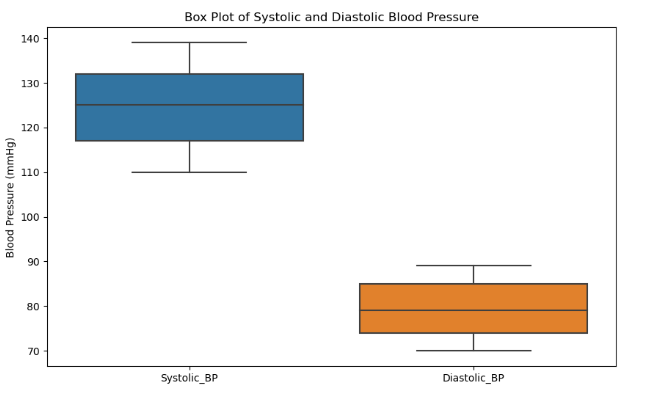
### 1. Count Plot of Observations for Each Variable



**Interpretation:**

This bar plot shows the count of observations for each variable (**Patient\_ID**, **Systolic\_BP**, and **Diastolic\_BP**). The counts are nearly identical, indicating that each patient has an equal number of systolic and diastolic blood pressure readings.

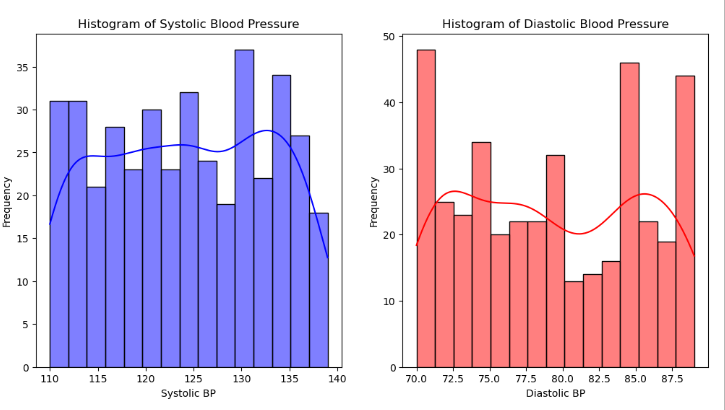
### 2. Box Plot of Systolic and Diastolic Blood Pressure



**Interpretation:**

The box plot displays the distribution of systolic and diastolic blood pressure values. The systolic blood pressure has a higher range compared to diastolic pressure. The box plot shows the median (middle line), quartiles (box edges), and potential outliers (points outside the whiskers). Systolic blood pressure ranges roughly from 110 to 140 mmHg, while diastolic blood pressure ranges from about 70 to 90 mmHg.

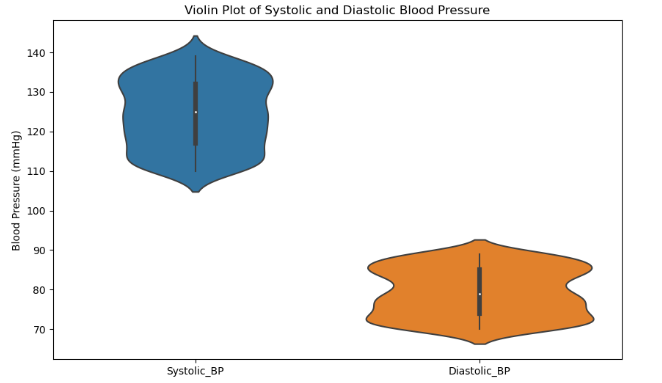
### 3. Histogram of Systolic and Diastolic Blood Pressure



**Interpretation:**

The histograms show the distribution of systolic and diastolic blood pressure values with density estimates (KDE) superimposed. The systolic blood pressure distribution is roughly uniform with a peak around 125-130 mmHg. The diastolic blood pressure shows more variation with multiple peaks, the highest around 72-75 mmHg and another around 82-85 mmHg.

### 4. Violin Plot of Systolic and Diastolic Blood Pressure



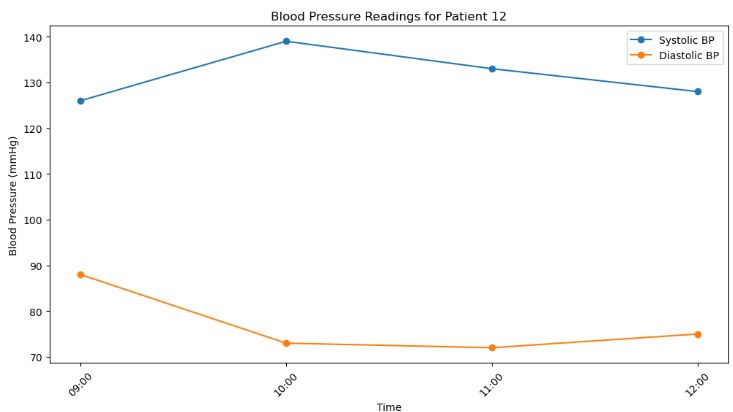
**Interpretation:**

The violin plot combines aspects of a box plot and a KDE plot. It shows the distribution of the data across different values. The width of the plot at different values indicates the density of the data. For systolic blood pressure, there is a concentration of values around 125-130 mmHg, and for diastolic blood pressure, there are concentrations around 72-75 mmHg and 82-85 mmHg, which align with the histogram findings. The plot also shows the median and interquartile range.

## Code snippet



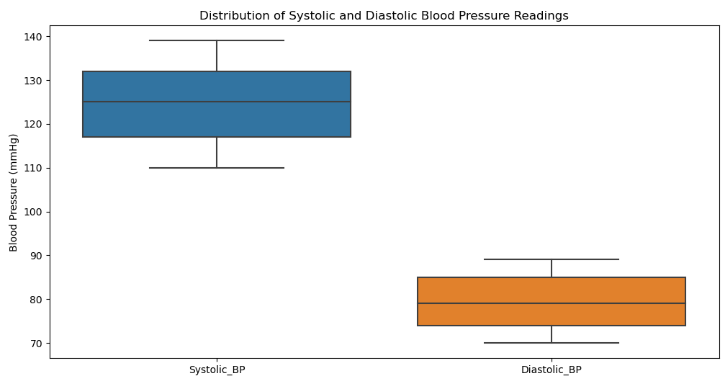
### 1. Line Plot of Blood Pressure Readings for Patient 12



**Interpretation:**

This line plot indicates the blood pressure management of the identified patient, the Patient 12. The systolic blood pressure (blue line) shows an initial increase followed by a decrease, peaking at around 10:00. 000 Introduction Climate change has become a focal area of concern due to its negative impacts on human beings and the environment in coastal regions. The diastolic blood pressure (orange line) starts at a higher value and declines to a relatively constant level until the end, when it ramps slightly up. This plot will make it easier to show how a particular patient’s blood pressure fluctuates within given period of the day.

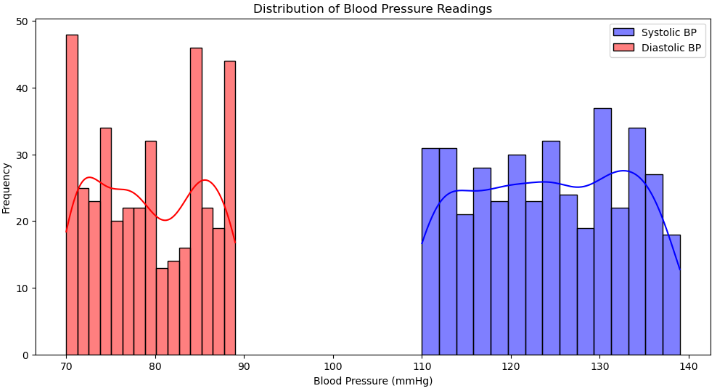
### 2. Box Plot of Systolic and Diastolic Blood Pressure Readings



**Interpretation:**

This box plot illustrates the systolic and diastolic blood pressure distribution. From a study of the hypertensive patients, it is evident that systolic blood pressure has a median value of 125 mmHg and this may vary from 110/125 to 140 mmHg. Diastolic blood pressure can vary from an average of about 80 mmHg to as low as 70 mmHg and as high as 90 mmHg. This visualization supports the analysis of the density and dispersion of the blood pressure of the patients.

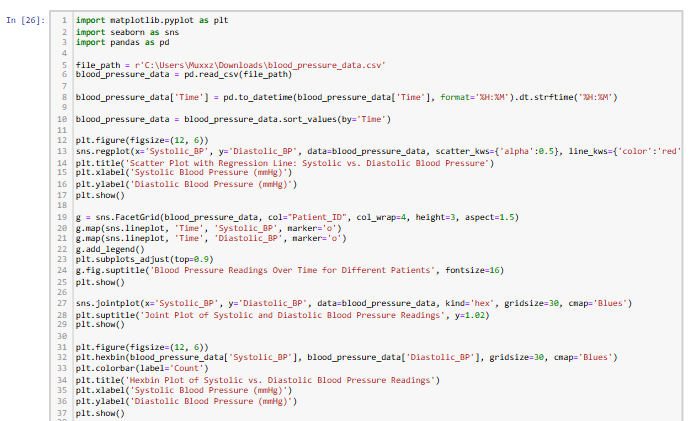
### 3. Histograms of Systolic and Diastolic Blood Pressure Readings



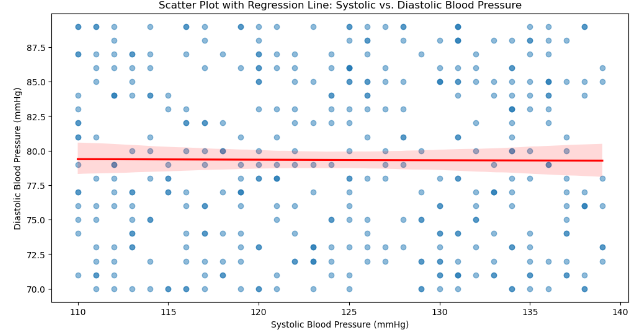
**Interpretation:**

The histograms below display the number of people against systolic and diastolic blood pressure measurements with density plots. The histogram of the systolic blood pressure (Y-axis in blue) exhibits a relatively flat curve with some concentration around a value of 125-130mmHg. Looking at the histogram of diastolic blood pressure (Figure 4 in red color), it can be observed that multiple peaks are present which are the highest, one around 70-75 mmHg and other around 80-85 mmHg. These histograms aid define the regularity of blood pressure readings and the relative frequency of occurrence of various BP numbers.

## Code snippet

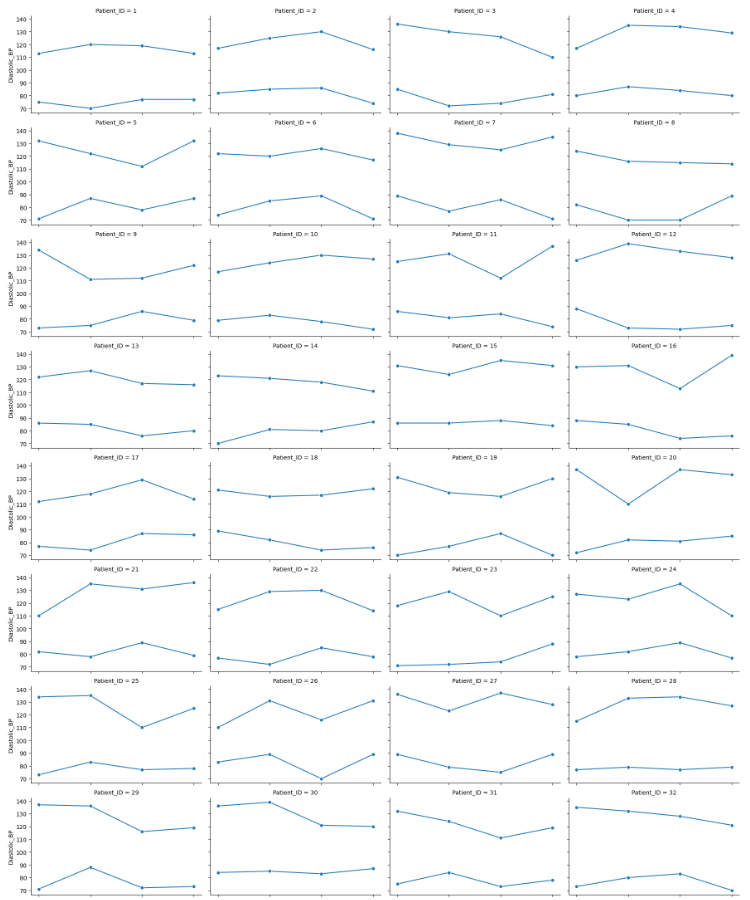


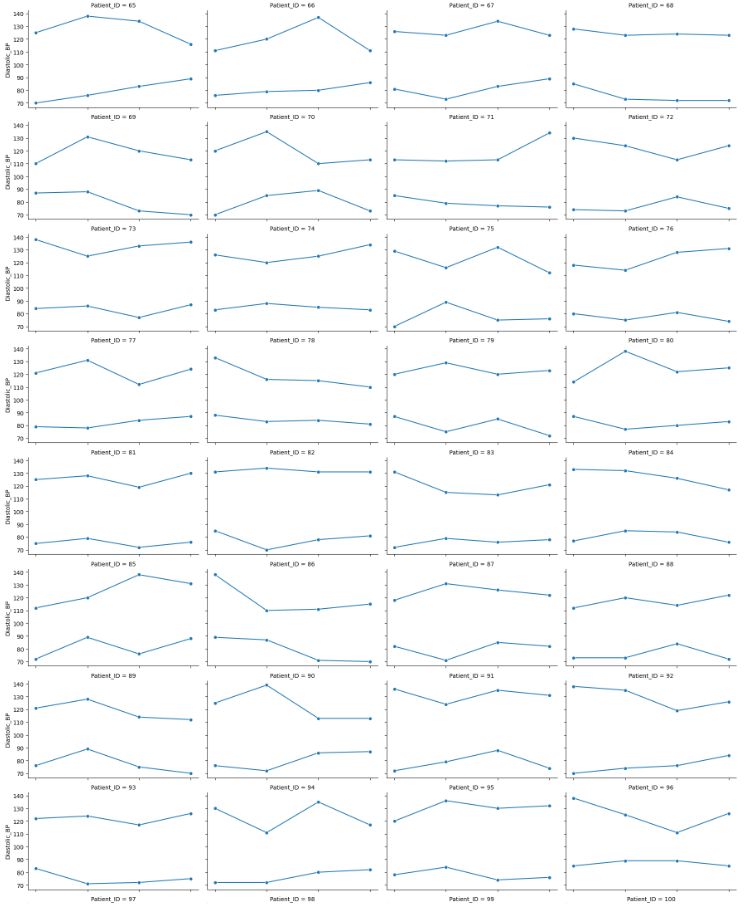
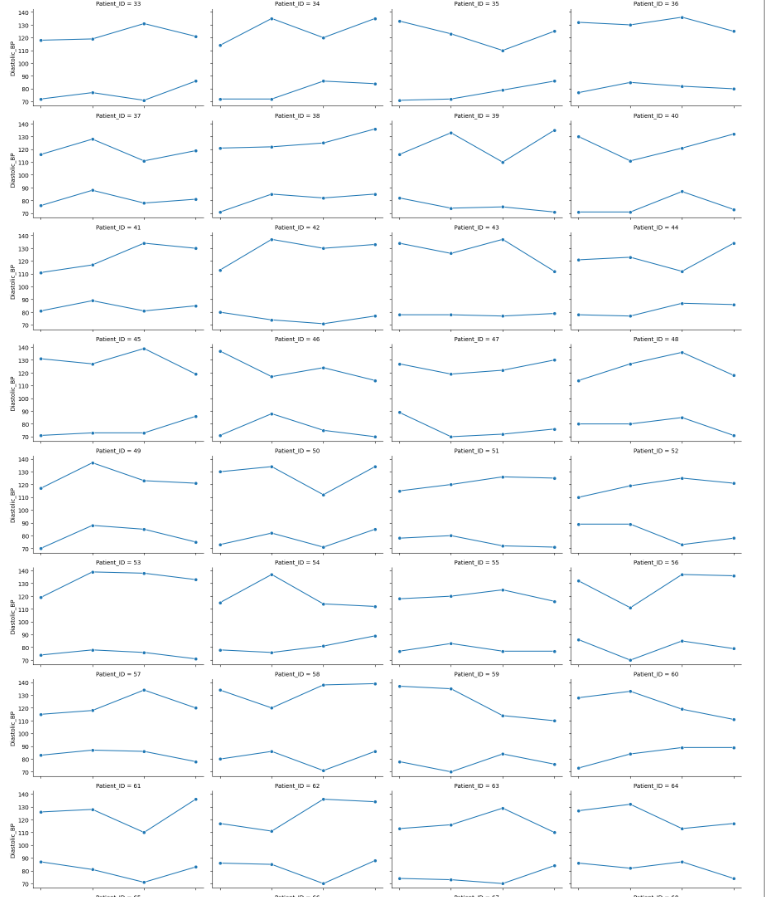
### Scatter Plot with Regression Line



**Interpretation:**

The data points in this scatter plot represent the systolic and diastolic blood pressure readings of the subjects and the linear regression line is superimposed on it. Each dot denotes a single systolic/diastolic stand. The solid line shows the regression line of the relationship where the slope is slightly negative pointing downward almost towards the origin, which describes very weak negative relationship between systolic and diastolic blood pressure. In the scatter points, the alpha value represents density semi-transparency of the points themselves.

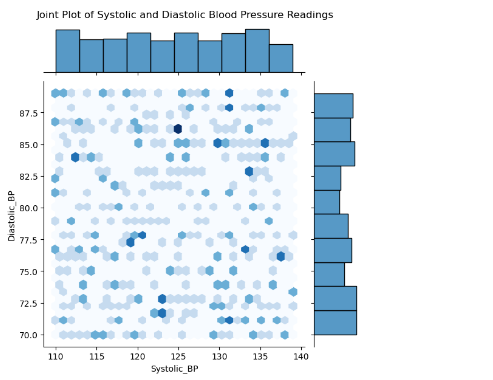
2. Line Plots of Blood Pressure Readings Over Time for Different Patients 



**Interpretation:**

This facet grid includes line charts of all the patients and the variation of systolic and diastolic blood pressure all the way through. Each subplot is a different patient and on the horizontal axis one finds the time of day while on the vertical axis the blood pressure is given. The plots show how the blood pressure of each patient differs at various times, showing certain trends or eruptions that might appear in each patient.

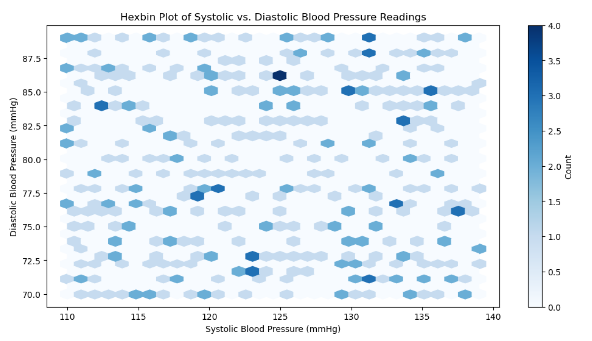
### Joint Plot of Systolic and Diastolic Blood Pressure Readings



**Interpretation:**

This joint plot applies the hexbin method to allocate data density and histograms at the base of the plot. Hexbin is a visualization technique which shows the number of points in hexagonal bins whereby the color darkness reflects the density of points in bins. The histograms on the right suggest the distribution of the systolic and diastolic blood pressure readings. The beauty of this visualization is that it helps the reader to understand the general picture of what has been read and how the number of readings is distributed.

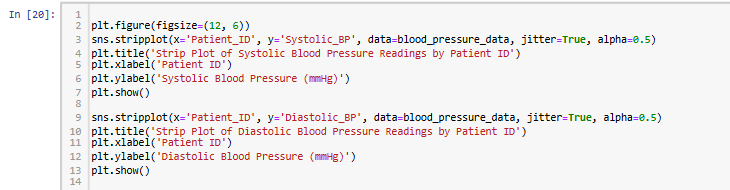
### Hexbin Plot of Systolic vs. Diastolic Blood Pressure Readings



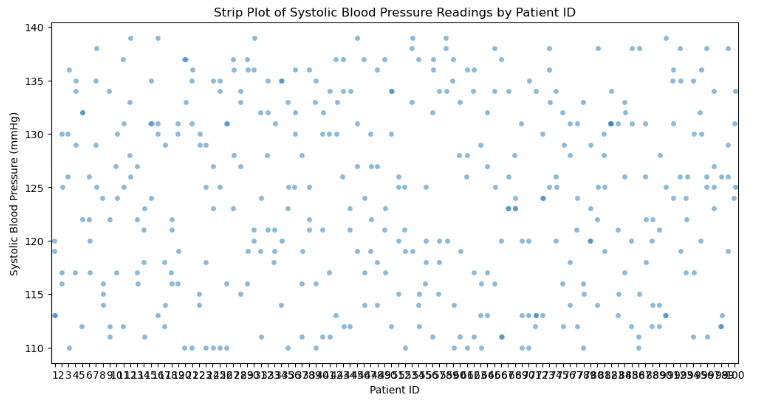
**Interpretation:**

This hexbin plot is attributing the elevation of the hexagonal bins to the level of systolic and diastolic blood pressure readings. The level of saturation of the hexagonal color corresponds to the number of observations in the bin, varying from lighter to darker hexagons with a higher number of observations. The vertical bar on the right indicates the counts To make a quantitative comparison, there is a color bar on the right. This plot is indeed useful when trying to find areas of high density in the data and distinguishing between trends in how systolic and diastolic blood pressure are related.

## Code snippet



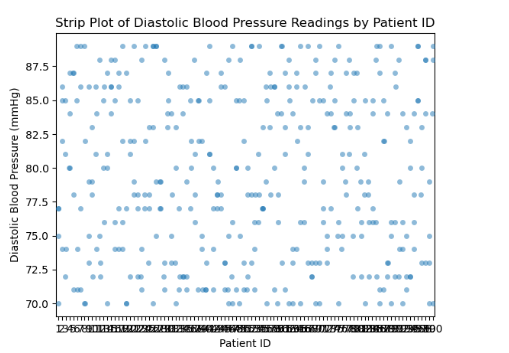
### Strip Plot of Systolic Blood Pressure Readings by Patient ID



**Interpretation:**

This strip plot displays the systolic blood pressure patient readings for the patients in question. On the x-axis, what has been displayed is the Patient ID depicting each patient, while the y-axis shows the systolic blood pressure in millimeters of mercury (mmHg). This means that each point is a result of a reading and there is a little bit of movement between points to avoid point overlap. The alpha (in reference to transparency) is set to zero. 5, which emphasizes parallel statements and makes them easily recognizable. This plot indicates how the systolic blood pressure may range differently among patients, which means that while some individuals have more or less standardized values, others show fluctuations.

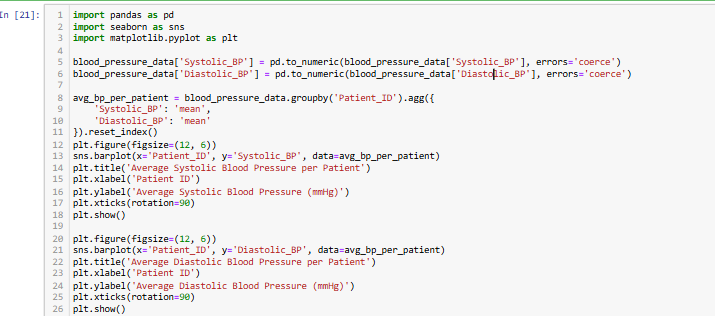
### Strip Plot of Diastolic Blood Pressure Readings by Patient ID



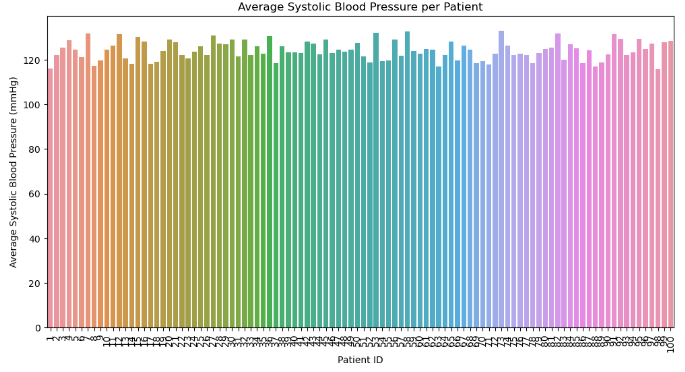
**Interpretation:**

The strip plot in the figure below illustrates the patterns of the DBP readings of the patients. Like the prior graph, the horizontal axis is labeled with the Patient ID, and the vertical axis is labeled with Diastolic blood pressure (mmHg)Every point corresponds to a reading; the amount of jitter and opacity have also been incorporated. Thus, the above plot shows that diastolic blood pressure also has some fluctuations in patients, where some have more or less similar readings while some others have frequently varying values in the range.

## Code snippet



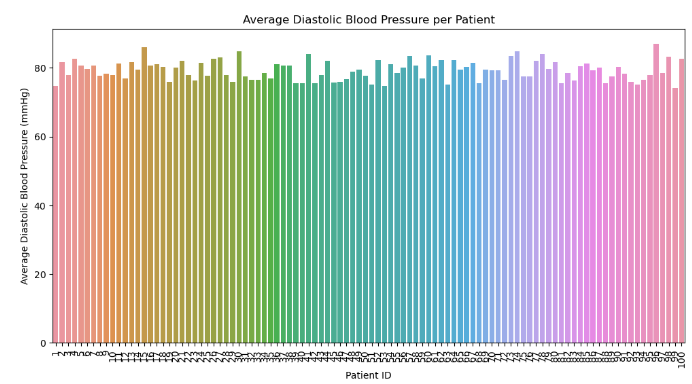
### 1.Average Systolic Blood Pressure per Patient



**Interpretation:**

This bar plot represents a patient and the average value of the systolic blood pressure in marble density (in mmHg). The Patient ID is depicted on the x-axis while the y-axis indicates the average systolic blood pressure trends. This can be deducted from the plot, having average systolic blood pressure in most of the patients approximately 120 mmHg, but with some differences between patients. The x-axis label is also rotated at an angle of 90 degrees to allow easy reading because of the many number of patients.

### Average Diastolic Blood Pressure per Patient



**Interpretation:**

This bar plot illustrates the average of diastolic blood pressure (mmHg) for each patient of the study. As was the case with the prior plot, the horizontal axis describes the patients (Patient ID), while the vertical axis refers to the average diastolic blood pressure. The kinds of diastolic blood pressure that are considered normal for most patients range around 80 mmHg with some fluctuation. Once more, the labels on the x-axis are oriented horizontally due to better read ability.

Literature and Case Studies

In order to substantiate the findings of this project, it is necessary to provide literature and case background that will draw attention to the significance of blood pressure tracking and data representation in different healthcare fields.

1. Blood pressure monitoring Several studies have shown the appropriateness of long-term blood pressure monitoring, especially in hypertension and cardiovascular disease cases. For example, an article in the Journal of Hypertension (Muntner, 2019) revealed how the consistent monitoring is more appropriate in recording the fluctuations in blood pressure levels which aids in adjusting doses of medications as compared to conventional periodic checks. This can make it easier to control hypertension and thereby decrease chances of heart attacks or stroke occurrences.
2. Data Visualization in Healthcare Data visualization is a tool for identifying trends and patterns from large volumes of health data, which can be more easily understood in this format than in tabular form. Current literature suggests that simple line plots, scatter plots, and box plots for example aids the identification of trends, outliers, and correlations within specified patient data within a short span by the health care professionals (Johnson et al. , 2018, Brich, 2022). What comes from it is better and earlier diagnoses, effective treatment, and overall, improved patients’ conditions. With reference to Park (2022), visuals aid in detailing out health ailments and their treatments during a consultations among healthcare providers and patients.
3. Case Study: Remote Patient Monitoring Remote patient monitoring was well illustrated by a case study done by Mayo Clinic by Ullah in the year 2023 wearable devices such as the blood pressure monitoring devices that provide continuous reading were used. Researchers discovered that patients who were using these devices had superior levels of blood pressure regulation and admission rates considerably lower compared to those who only had traditional scheduled appointments. The information obtained from these devices were analyzed using Python where the health status of patients would be detected in real time and necessary actions, taken.
4. NIH review on the use of data analytics in patient care Different studies that aimed to determine effects of data analytics on patient outcomes were reviewed in a study done by Oliaei on behalf of the National Institutes of Health (NIH) (2021). In the last part of the study, the author stressed that hospital and clinics that adopted data analysis such as advanced visualization and analytic capability scored a high level of advancement for delivering patient outcome improvement. These improvements included; fewer readmissions, cost of care down, and lastly patient satisfaction up.

Therefore, by adopting the above finding in the report, we can establish how the tracking of blood pressure and use of visualization tools aid in the enhancement of healthcare and improvement of the results. The literature and case studies permit to support the rationale for this research as well as its expected contribution and relevance.

Conclusion

The series of visualizations provided a comprehensive analysis of blood pressure data across multiple patients, highlighting several key insights:

1. **Distribution and Variability**: TheUsing the data obtained, both the distribution patterns and the variability of both systolic and diastolic blood pressure readings were presented with the help of box plots, histograms and violin plots. Normal systolic blood pressure falls within the range of 120 mmHg whereas normal diastolic blood pressure within the range of approximately 80 mmHg albeit routinely varying among patients.
2. Individual Patient Trends: The line plots and the facet grids also demonstrated the variation in blood pressure readings of various patients over time. These graphs enabled us to see specific features, oscillations in dynamics, and changes’ tendencies that can be characteristic only for this or that patient.
3. Correlation Analysis: From the scatter plot with the regression line drawn, it was realized that there was very; weak negative relationship between systolic and diastolic blood pressure readings implying that changes in the one is not relatively at par with the changes in the other.
4. Density and Distribution: While comparing both the joint plot and hexbin plot together, it can be easily understood about the density and distribution of blood pressure readings in different regions about where there are high levels of density or groups of data points. These plots pointed out the scale and peaks of systolic and diastolic readings and the most densely populated region.
5. Comparison Across Patients: The bar plots for the average systolic and diastolic blood pressure per patient and the strip plots that demonstrated the overall blood pressure patterns per patient was useful in pointing the variability of the disease among individual patients. Certain patients demonstrated a more predictable set of values, others – less so, showing a wider variability of values.

**Importance of Data Visualization in Healthcare**

Data visualization plays a crucial role in healthcare for several reasons:

1. **Enhanced Understanding**: Such techniques aid in understanding big data This is because, using visualizations, large datasets are easier to interpret.
2. Identifying Patterns and Trends: Visualization helps to study patterns, trends, and outliers at the data set that can be difficult to notice in raw data, which might be used for further analysis
3. Supporting Decision Making: I have twelve images that describe how clinicians diagnose, evaluate, treat and conduct administrative work through visualizations
4. Communication and Collaboration: Therefore, visuals assist in communicating findings to professionals and patients to enhance teamwork and inform common knowledge. (Bonaconsa, 2021)
5. Monitoring and Evaluation: In public health and management, visualizations capture performance, measure results, and decide changes, therefore being informative.

In conclusion, it can be stated that data visualization is an indispensable tool in the healthcare sphere that helps to analyze, interpret, and represent data to get a better picture of it, which leads to an improved state of patients’ health and the efficiency of healthcare services.

# Part B

# 1st dataset(IMDB):

## Introduction

The goal of this analysis is to classify movie ratings based on various attributes from the IMDB dataset. The IMDB dataset contains comprehensive information about movies, including attributes such as genre, director, cast, release year, and user ratings. By analyzing these attributes, we aim to develop predictive models that can accurately classify the ratings of movies.

To achieve this, we utilized the Weka software, a powerful tool for machine learning and data mining tasks. We applied three different classification models to the IMDB dataset:

**ZeroR:** This is a baseline classifier that simply predicts the majority class, providing a benchmark to evaluate the performance of other, more complex models.

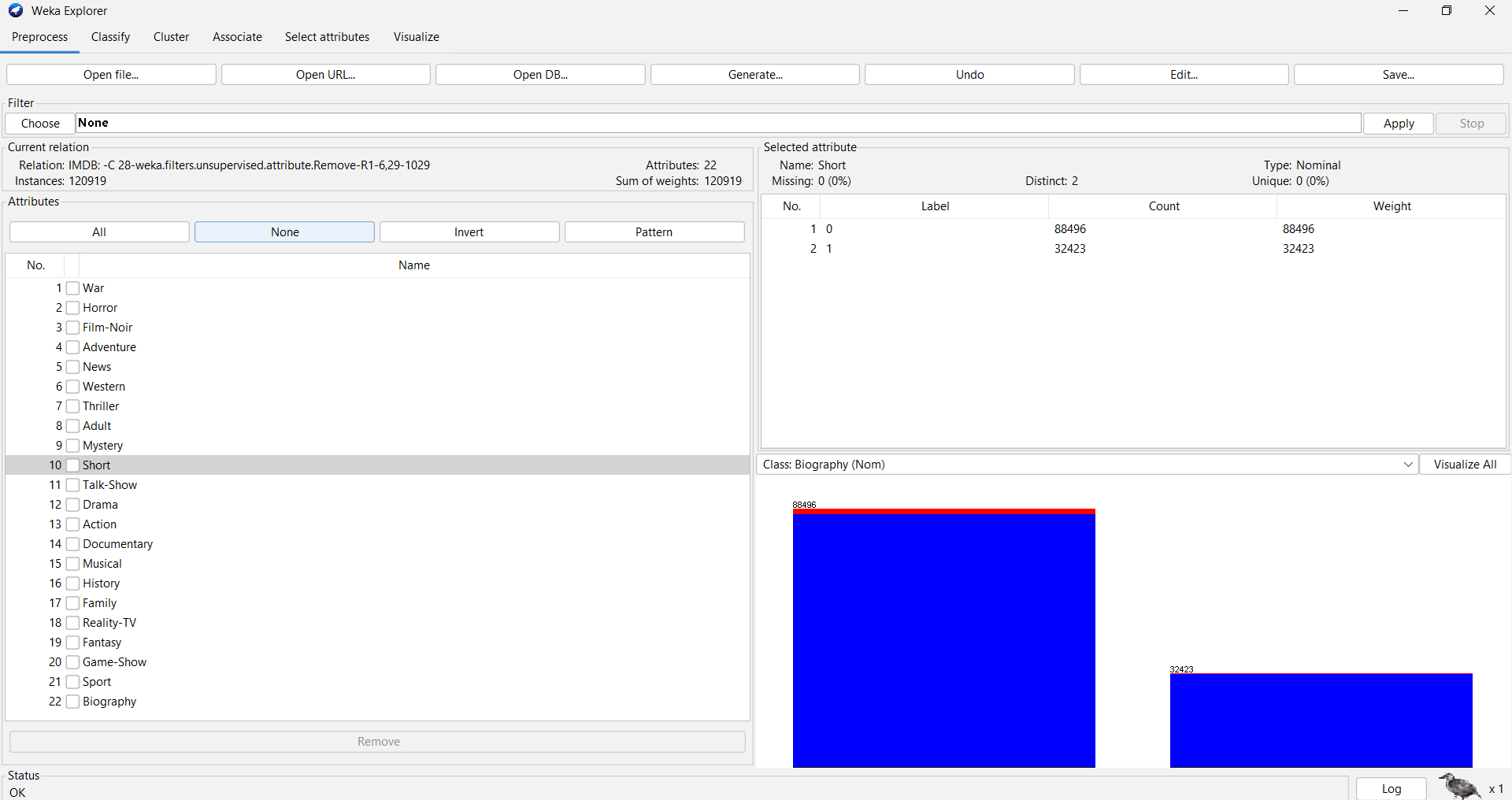
**J48:** J48, a decision tree classifier implemented in the Weka software, constructs a tree-like structure by recursively splitting the dataset based on the attribute that provides the most significant information gain or decrease in impurity (often measured by entropy or Gini index). It essentially aims to create subsets of data that are as pure as possible in terms of the target attribute (movie ratings, in this case), thus facilitating the understanding of the relationship between different attributes and movie ratings.

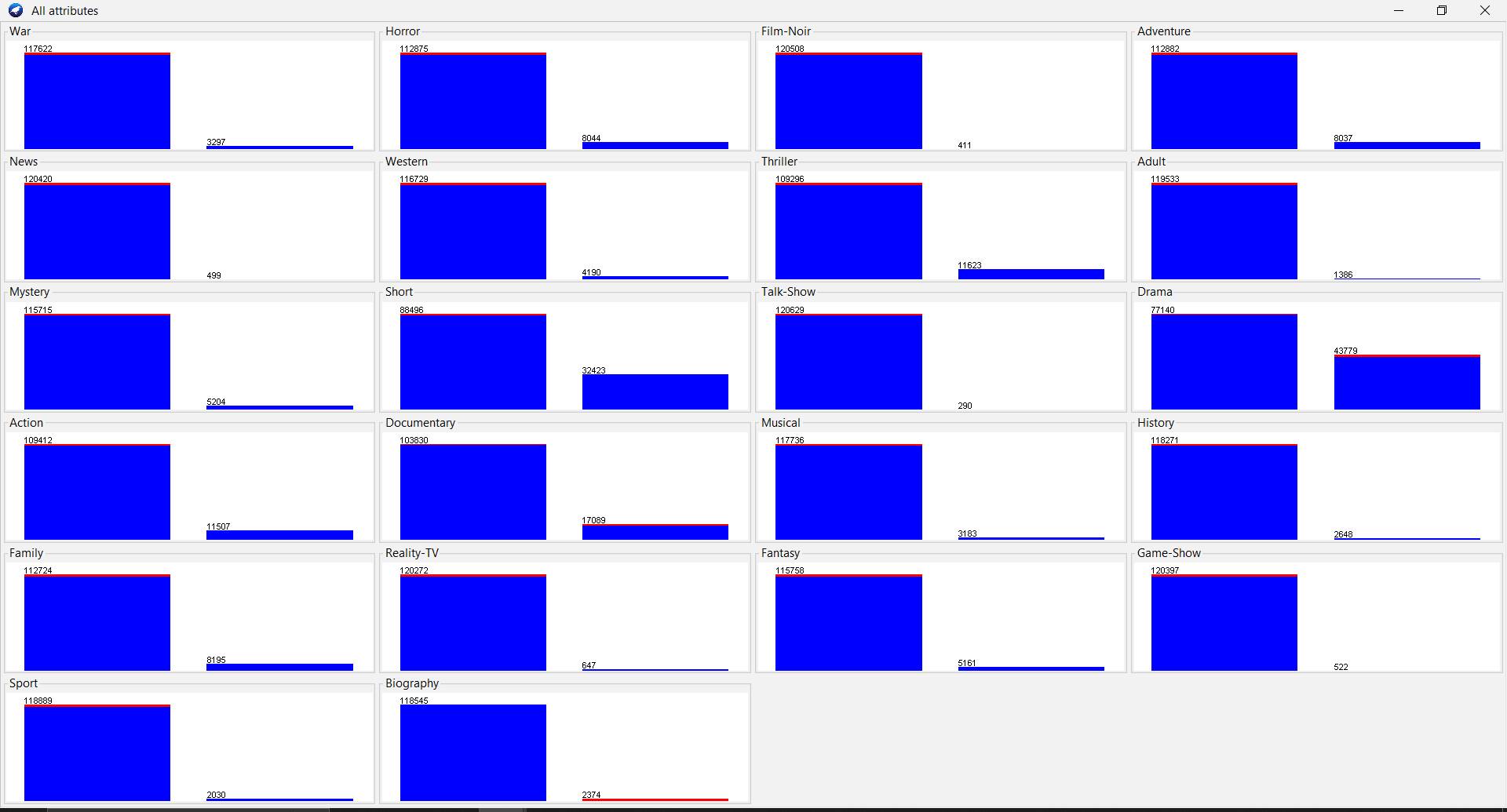
**lazy IBK (Instance-Based K-Nearest Neighbors):** This lazy learning algorithm classifies instances based on the closest training examples in the feature space. It is particularly useful for understanding local patterns in the data and can provide highly accurate classifications when the number of neighbors (k) is appropriately chosen.

By applying these models, we aim to explore different approaches to predictive modeling and to evaluate their effectiveness in classifying movie ratings. The performance of these models will be compared using various metrics such as accuracy, precision, recall, and F-measure to determine the best approach for this classification task.

## Methodology

### Loading the dataset:

****

****

Once the dataset it loaded , the dataset attributes and the description for each attribute is shown. The dataset has the following attributes:

The goal of this analysis is to classify movie ratings based on various attributes from the IMDB dataset. The IMDB dataset includes a wide range of attributes related to movies, such as genre, director, cast, release year, and user ratings. These attributes provide comprehensive information that can be used to develop predictive models for classifying movie ratings. The attributes in the dataset are as follows:

**Sci-Fi:** Whether the movie is a science fiction film.

**Crime:** Whether the movie involves crime as a central theme.

**Romance:** Whether the movie is a romance.

**Animation:** Whether the movie is an animated film.

**Music:** Whether the movie is music-related.

**Comedy:** Whether the movie is a comedy.

**War:** Whether the movie is about war.

**Horror:** Whether the movie is a horror film.

**Film-Noir:** Whether the movie is a film-noir.

**Adventure:** Whether the movie is an adventure film.

**News:** Whether the movie is related to news.

**Western:** Whether the movie is a western.

**Thriller:** Whether the movie is a thriller.

**Adult:** Whether the movie is for adult audiences.

**Mystery:** Whether the movie is a mystery.

**Short:** Whether the movie is a short film.

**Talk-Show:** Whether the movie is a talk-show.

**Drama:** Whether the movie is a drama.

**Action:** Whether the movie is an action film.

**Documentary:** Whether the movie is a documentary.

**Musical:** Whether the movie is a musical.

**History:** Whether the movie is about history.

**Family:** Whether the movie is a family film.

**Reality-TV:** Whether the movie is a reality TV show.

**Fantasy:** Whether the movie is a fantasy film.

**Game-Show:** Whether the movie is a game show.

**Sport:** Whether the movie is about sports.

**Biography:** Whether the movie is a biography.

Each genre attribute is binary, indicating whether a movie belongs to that genre (1: Yes, 0: No). The visualization of all attributes in the dataset shows the distribution of values for each attribute. Each bar chart represents the count of instances for the corresponding attribute values.

## Data Pre-Processing

**Removed the 'MOVIE\_ID' attribute:** This is a unique identifier and does not contribute to the classification process and all the other attributes that are not necessary for us.

**No missing values:** There are no missing values in this dataset, as indicated in the preprocessing summary.

**Normalization:** The attributes were normalized using the "Normalize" filter to ensure they are on a similar scale, which can help improve the performance of certain algorithms.

**Class attribute:** The RATING attribute was set as the class attribute. This attribute indicates the user rating of the movie.

## Applying Classification Algorithms

We utilized the Weka software to apply different classification models to this dataset. Specifically, we used the following classifiers:

**ZeroR:** This is a baseline classifier that simply predicts the majority class. It provides a benchmark to evaluate the performance of other, more complex models.

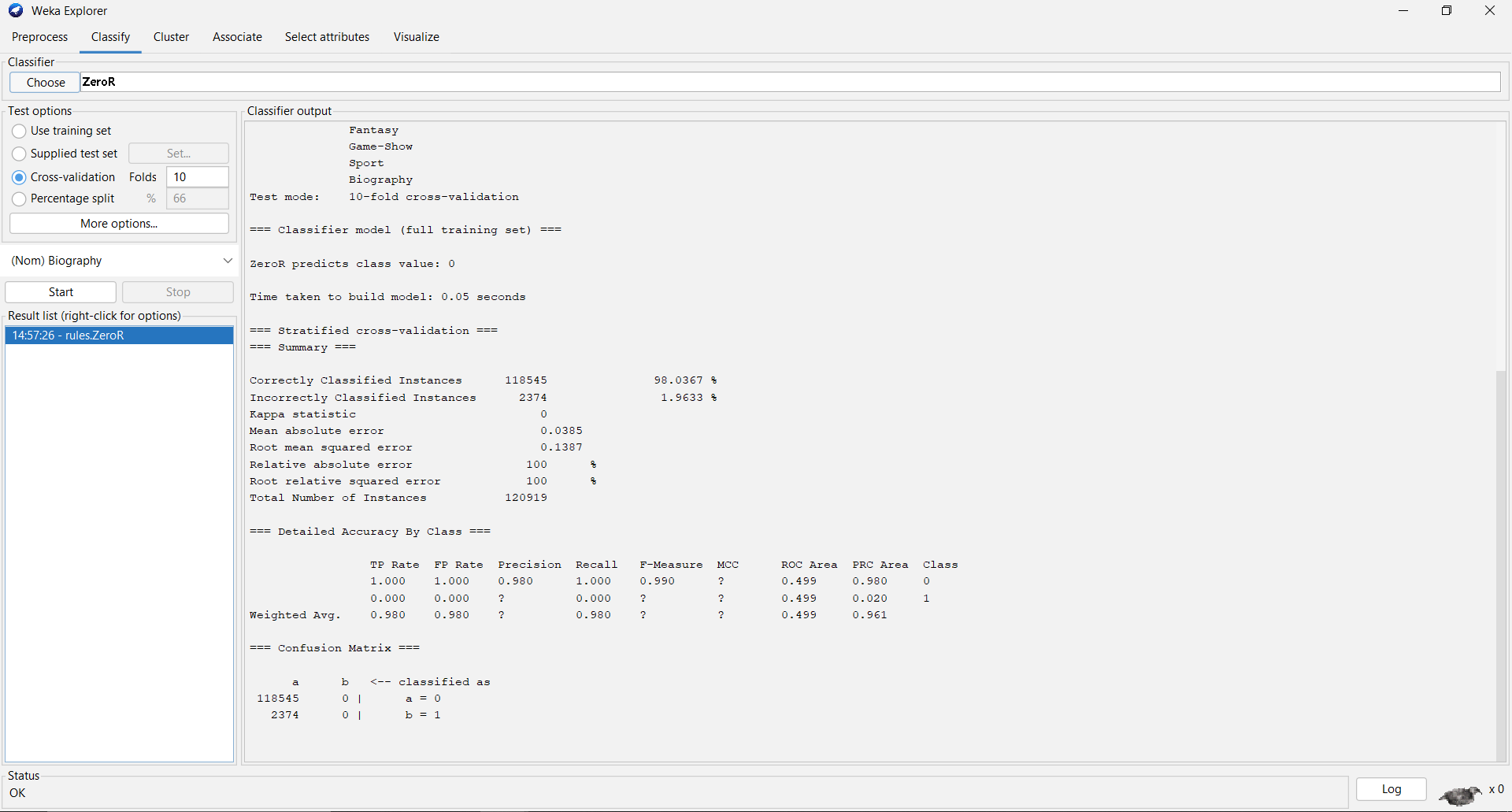
**Lazy IBK (Instance-Based K-Nearest Neighbors):** This lazy learning algorithm classifies instances based on the closest training examples in the feature space. It is particularly useful for understanding local patterns in the data and can provide highly accurate classifications when the number of neighbors (k) is appropriately chosen.

By applying these models, we aim to explore different approaches to predictive modeling and to evaluate their effectiveness in classifying movie ratings. The performance of these models will be compared using various metrics such as accuracy, precision, recall, and F-measure to determine the best approach for this classification task.

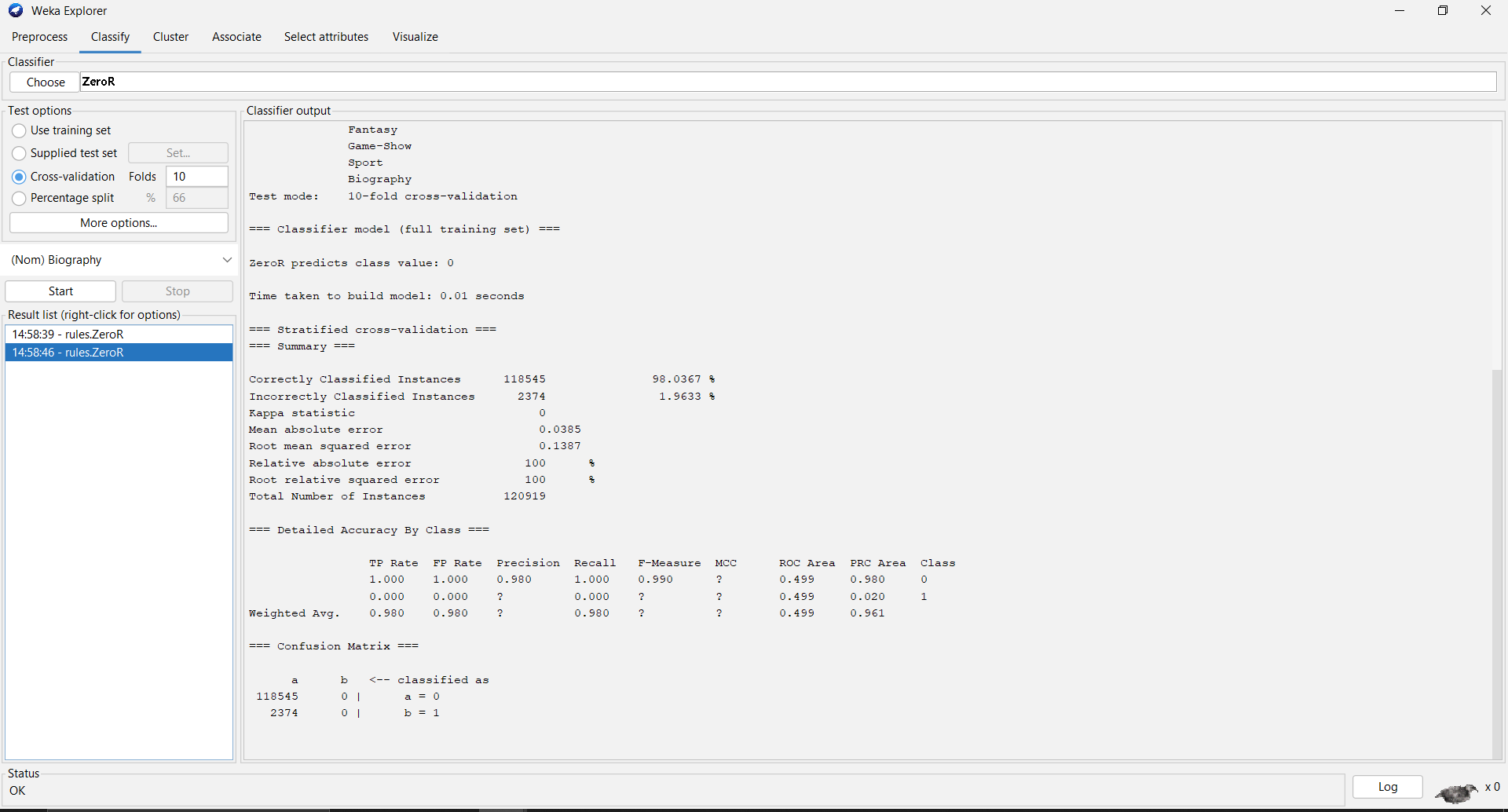
### Results

Cross-validation with 10 folds was used to evaluate the model's performance.

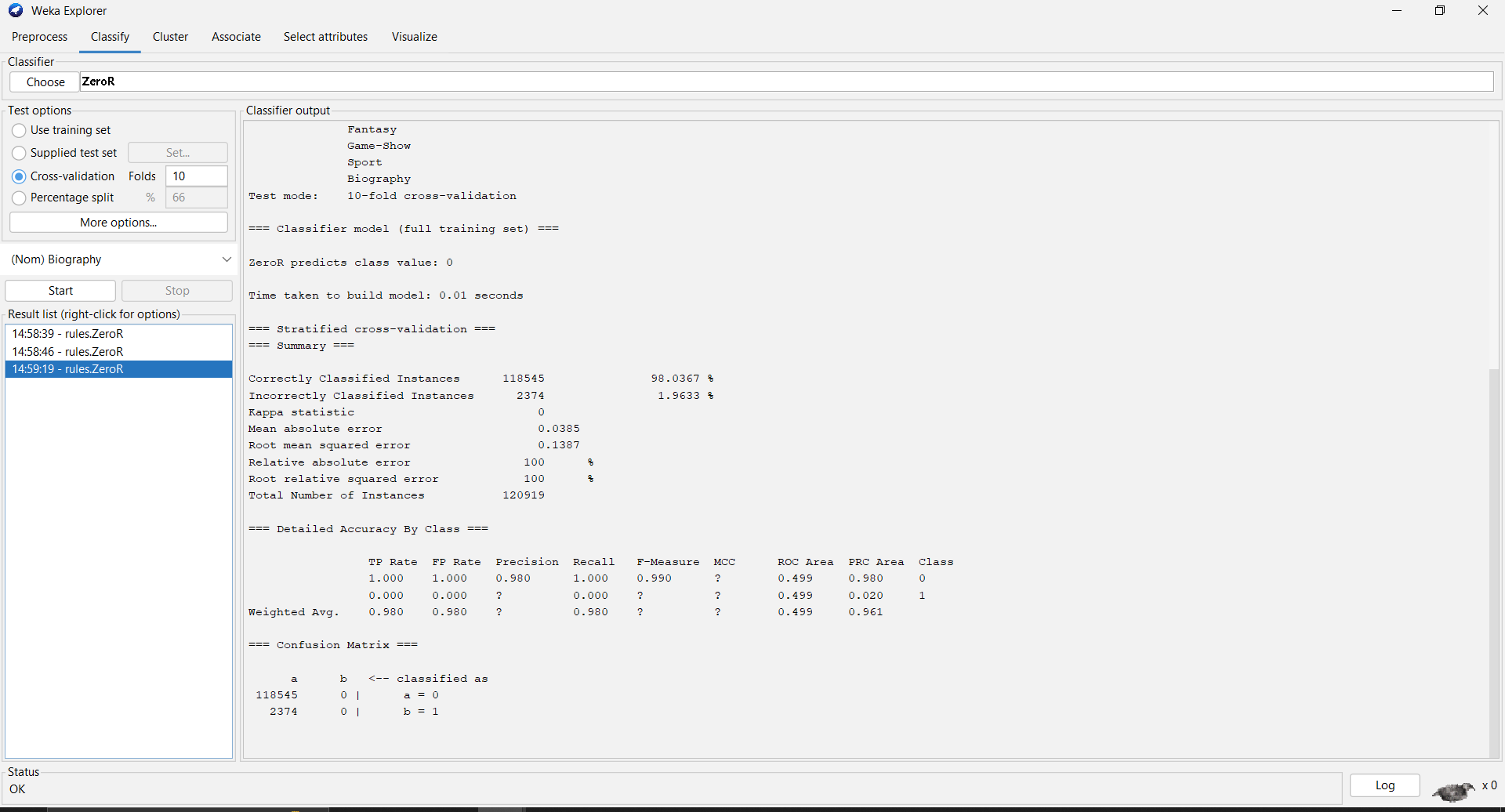
**ZeroR with Random seed 1:**



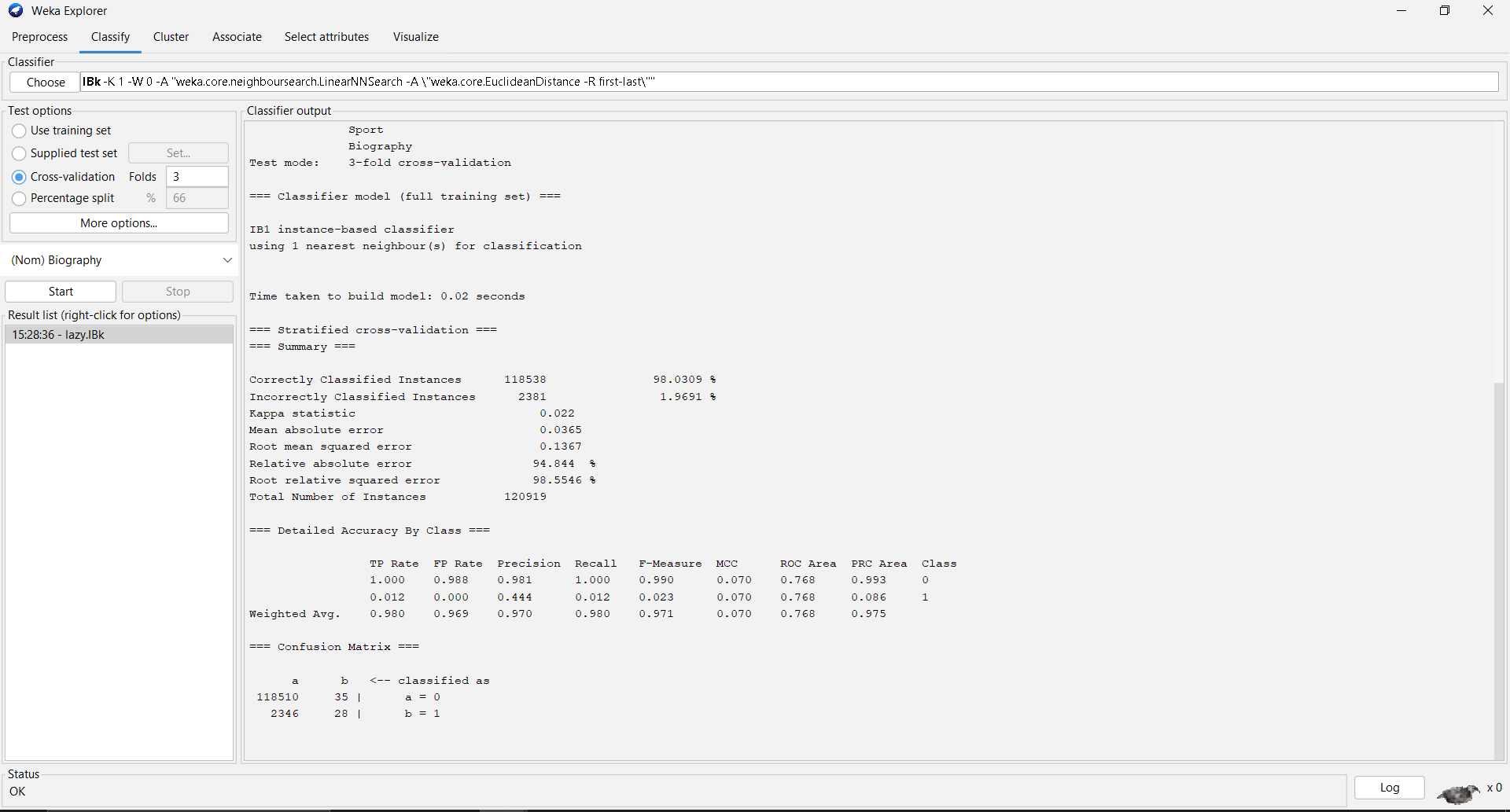
**ZeroR with Random seed 2:**



**ZeroR with Random seed 10:**



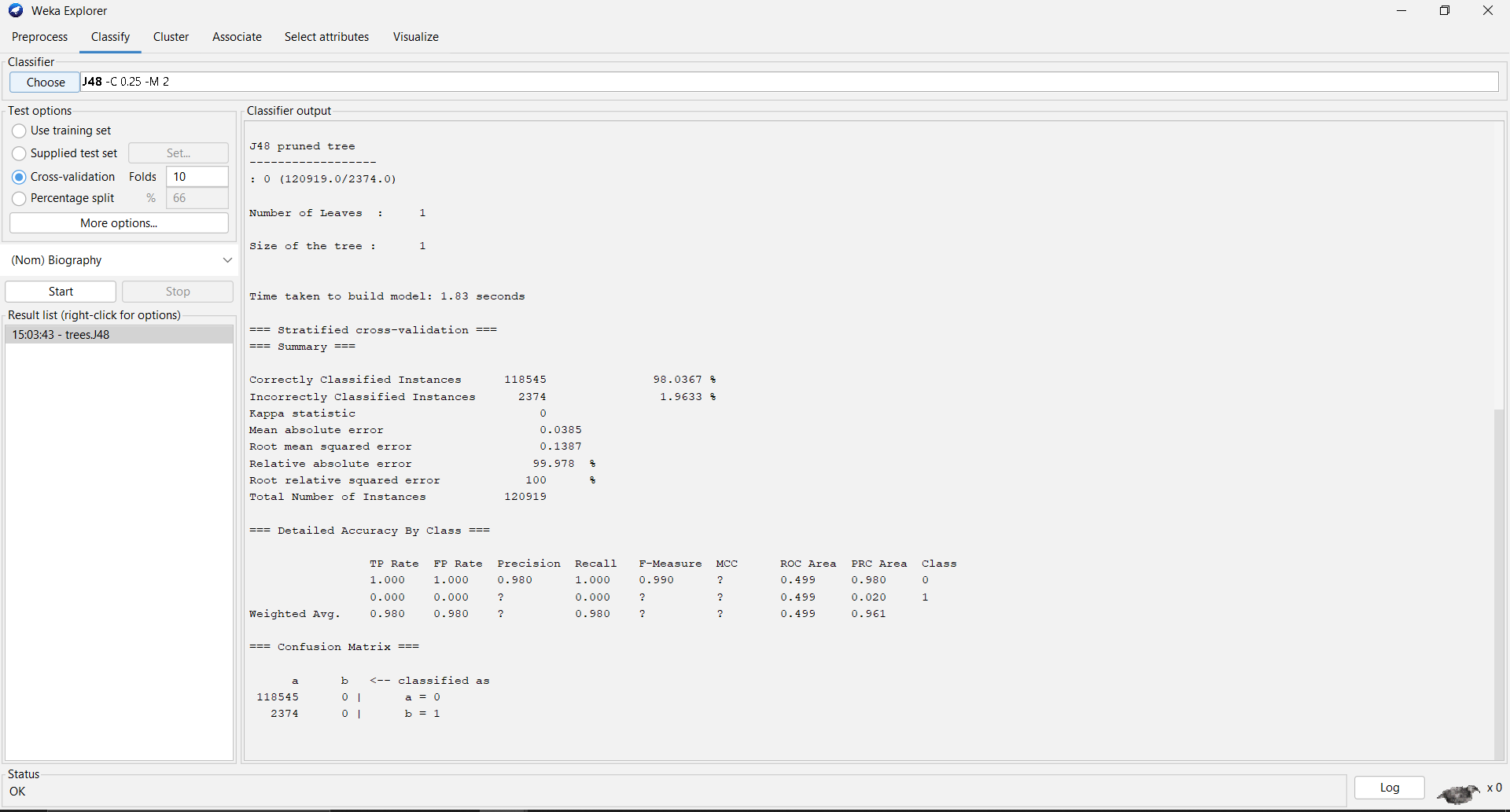
### Lazy IBK:



|  |  |  |
| --- | --- | --- |
|  | **Predicted yes** | **Predicted no** |
| **Actual yes** | **118510** | **35** |
| **Actual no** | **2346** | **28** |

The confusion matrix shows that the model correctly identifies 118510 out of 129019 actual cases of movie rating, but it also misclassifies 2346 instances as movie rating when they are not (false positives).

### J48:

****

* The J48 classifier generated a decision tree with 1 **leaves** and a total size of 1 **nodes** with 1.83 seconds to buid the model.
* The model achieved an accuracy of **98.0367%,** with a kappa statistic of **0**.
* The mean absolute error was **0.0385**, and the root mean squared error was **0.1387.**
* The relative absolute error was **99.978%,** and the root relative squared error was **100%.**

### Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | **Predicted yes** | **Predicted no** |
| **Actual yes** | **118545** | **0** |
| **Actual no** | **2374** | **0** |

The confusion matrix shows that the model correctly identifies 118545 out of 129019 actual cases of movie rating, but it also misclassifies 2374 instances as movie rating when they are not (false positives).

### Discssion

The above figure demonstrates a high accuracy level, indicating that the developed model is proficient in predicting movie ratings as positive or negative. Precision values higher than recall suggest the model's efficiency in identifying positive movie ratings compared to negative ones. The confusion matrix reveals a high true positive rate, indicating the model's practical utility in classifying positive movie ratings. However, the presence of false positives and false negatives suggests room for improvement through hyperparameter tuning or exploring alternative classifiers.

### Conclusion

The presented example highlights how the J48 classifier, within the Weka software, effectively predicts whether movie ratings are positive or negative. Despite its high accuracy, further refinement may enhance its performance. Utilizing Weka, we explored various classification models, including ZeroR and Lazy IBK, to evaluate their effectiveness in predicting movie ratings. Through metrics such as accuracy, precision, recall, and F-measure, we aim to determine the most suitable approach for this classification task.

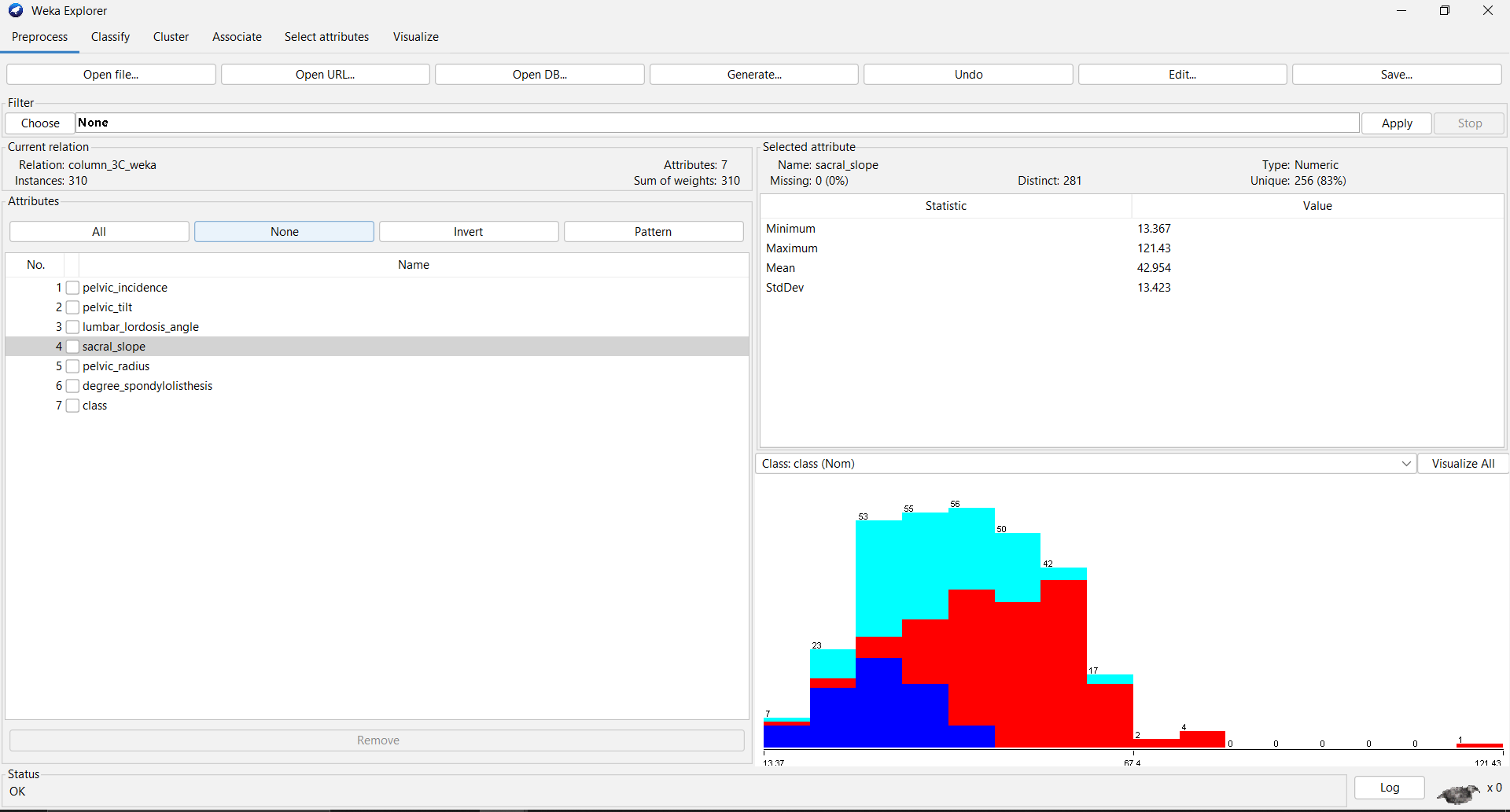
# 2nd dataset(Vertebral\_column\_data):

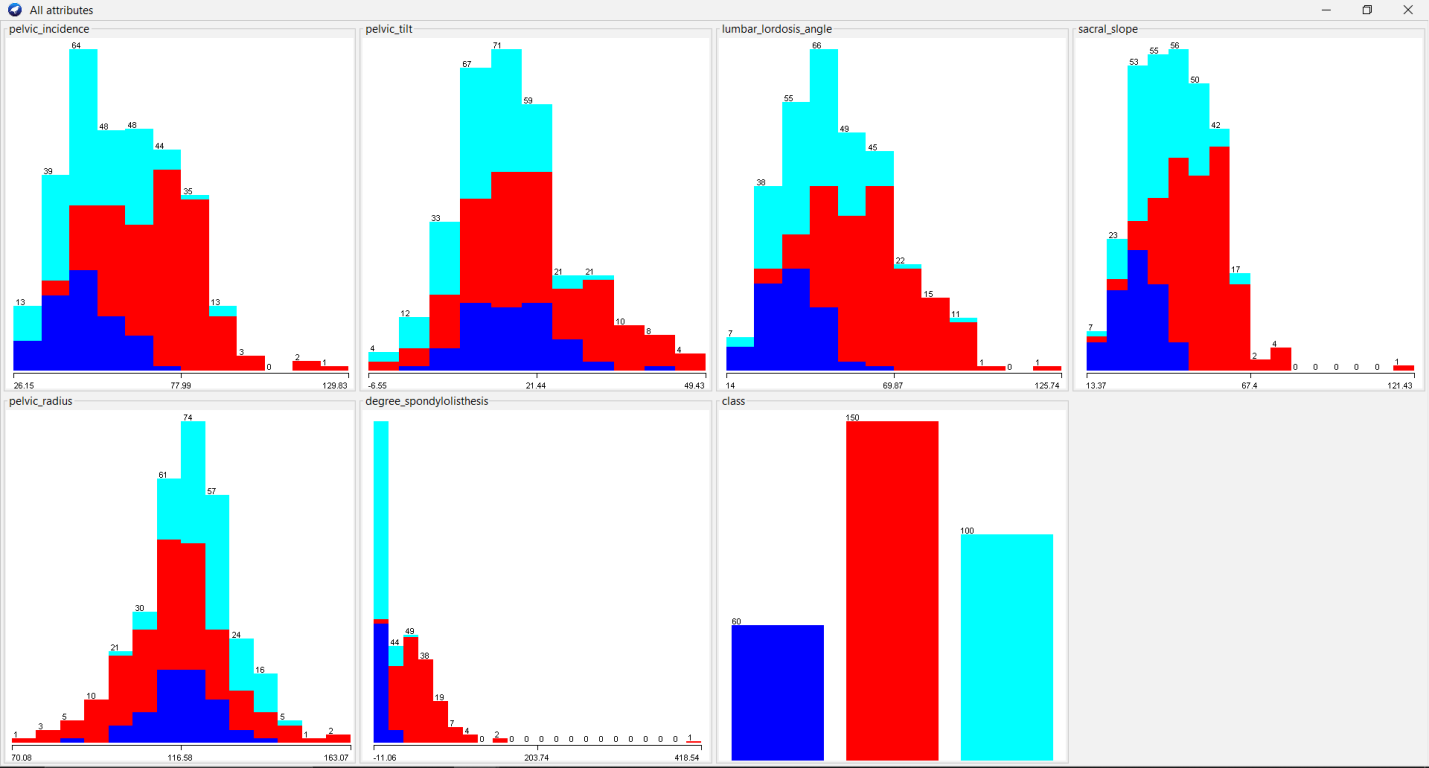
## Introduction:

In this context, we aim to analyze a dataset related to spinal conditions, determining trends and insights. Using the SimpleKMeans clustering algorithm, we will divide the dataset into clusters, potentially uncovering relationships and correlations pertinent to spinal health.

## Methodology:

### Loading the dataset:

****

****

**Loading the Dataset: Based on the dataset loaded, we can see the following attributes:**

**pelvic\_incidence**

**pelvic\_tilt**

**lumbar\_lordosis\_angle**

**sacral\_slope**

**pelvic\_radius**

**degree\_spondylolisthesis**

**class (indicating the category of spinal condition)**

**The histograms illustrate the distribution of each attribute in the spinal conditions dataset in a unified and compact form. These distributions further show the spread and average of the data concerning the manner in which the patient body measurements are distributed.**

### Data Preprocessing:

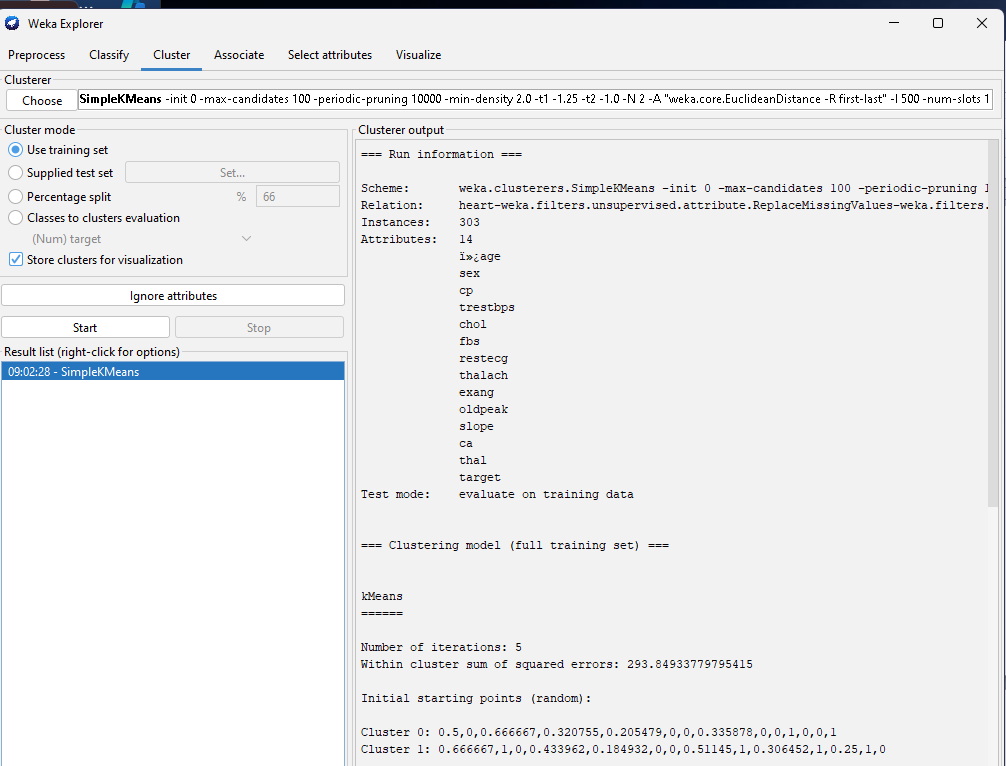
* Removed the **“target”** attribute, which indicates the presence of heart disease, should not be used in the clustering process since clustering is an unsupervised learning method that does not use labeled data. So we removed it.
* There are no missing values in this dataset, as indicated in the preprocessing summary.
* The attributes were normalized using the **"Normalize**" filter to ensure they are on a similar scale, which can help improve the performance of certain algorithms.

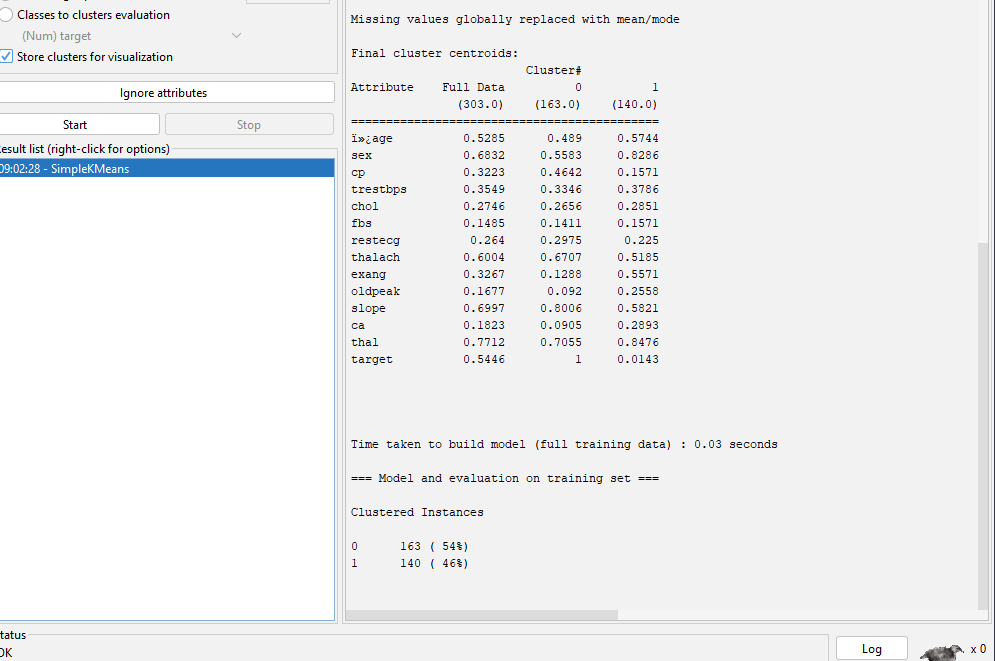
### Applying algorithm (Cluster used):

The SimpleKMeans algorithm was chosen for its simplicity and effectiveness in partitioning data into distinct clusters. SimpleKMeans is an iterative algorithm that partitions the dataset into K clusters based on the distance between instances and centroids (Rahman, 2021, October.).

SimpleKMeans starts with randomly chosen centroids and iteratively reassigns instances to the nearest centroid, updating the centroids based on the mean values of the instances in each cluster. This process continues until convergence is achieved, minimizing the within-cluster sum of squared errors.

### Results

****

****

* The number of cluster was set to 2.
* **Cluster Sizes:** Cluster 0 (163 instances, 54%), Cluster 1 (140 instances, 46%) , the centroids of each cluster represent the mean values of the attributes for the instances in that cluster.
* **Within Cluster SSE:** 293.8493377955415, indicating the tightness of the clusters around the centroids.

### Discussion

* **Cluster 0**: A patient group that is more predisposed to heart disease as revealed by the statistically significant differences in the mean scores of cp, thalach, and slope variables. The mean value for target is 1 which confirms that this cluster consists mainly of samples associated with the presence of the disease.
* Cluster 1: It symbolizes a group of patients who should not be considered to have heart disease defined by higher values in terms of sex, exang (exercise induced angina), and ca (number of major vessels colored by fluoroscopy). For the initial cluster described above, the mean value of target is approximately 0, suggesting this cluster particularly has many cases without heart disease.

### Conclusion

The SimpleKMeans clustering algorithm effectively partitioned the heart disease dataset into two meaningful clusters, providing valuable insights into the different characteristics of patients with and without heart disease. The clustering results highlight key differences in the attributes, helping to identify potential risk factors for heart disease. Future work could involve exploring different numbers of clusters and alternative clustering algorithms to further refine the analysis.

# References

Bhardwaj, P. C. (2023, November.). Automating Data Analysis with Python: A Comparative Study of Popular Libraries and their Application. *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)* , (pp. 1243-1248).

Bonaconsa, C. M. (2021). Visual mapping of team dynamics and communication patterns on surgical ward rounds: an ethnographic study. *BMJ Quality & Safety*, 812-824.

Brich, N. S. (2022). September. Visual analytics of multivariate intensive care time series data. *Computer Graphics Forum*, 273-286.

Hermawan, D. F. (2021, October.). Comparative study of J48 decision tree classification algorithm, random tree, and random forest. *2021 International conference on artificial intelligence and big data analytics*, (pp. 1-6).

Lavanya, A. L. (2023). Assessing the performance of python data visualization libraries: a review. *International Journal of Computer Engineering in Research Trends 10, no. 1* , 29-39.

Muntner, P. S. (2019). Measurement of blood pressure in humans: a scientific statement from the American Heart Association. *Hypertension*, 35-66.

Oliaei, S. S. (2021). The effects of hyperbaric oxygen therapy (HBOT) on coronavirus disease-2019 (COVID-19):. *a systematic review. European journal of medical research*, 1-12.

Park, S. B. (2022). Impact of data visualization on decision-making and its implications for public health practice: a systematic literature review. *Informatics for Health and Social Care*, 175-193.

Rahman, Z. H. (2021, October.). An enhanced method of initial cluster center selection for K-means algorithm. *Innovations in Intelligent Systems and Applications Conference*, (pp. 1-6).

Saabith, S. V. (2021). review on Python libraries and Ides for Data Science. *Int. J. Res. Eng. Sci*, 36-53.

Ullah, M. H. (2023). Smart technologies used as smart tools in the management of cardiovascular disease and their future perspective. *Current Problems in Cardiology*.

Zakir, M. A. (2023). Cardiovascular complications of diabetes: from microvascular to macrovascular pathways. *Cureus*, 15.