**Abstract**

In this project, we explore how data@analytics and visualization can reveal crucial insights for healthcare. Part A focuses on using Python to visualize data from the "Heart.csv" dataset, helping us identify key health metrics and patterns related to cardiovascular diseases. We focused on bar plot, pair plot, and histogram to making trends and correlations more visible so that the stakeholder could better understand the factors that concern heart health.

In Part B, we switch to Weka, an open-source data mining tool, to analyze two different datasets: Studying the IMDB-F: dataset and Vertebral \_Column\_data The IMDB-F dataset@The research conducted with the IMDB-F dataset allowed to learn what has an impact on movie ratings and what possible psychological effects the motion picture may have, such as the impact of the genre@on mood and our psyche. In order to understand the aspects of biomechanics of vertebrae the Vertebral Column dataset served as the primary source of information, which is critical for examining the spine disorders and their treatment.

As noted in this report, we outline the methods, tools, and different forms of analysis applied through the study, and why@our choice of they were suitable for the healthcare field. The arguments provided are supported through the literature and case evidence that demonstrates just how important data visualization and data mining are in contributing to better healthcare. This project does not only develop our competencies on practical aspects of data administration and analysis but also raises the question of the applicability of data in decision-making within the realm of medicine.

Contents

[Part A 2](#_Toc167598918)

[**Introduction** 2](#_Toc167598919)

[**Methodology** 3](#_Toc167598920)

[**Results and Analysis** 7](#_Toc167598921)

[**Conclusion** 18](#_Toc167598922)

[Part B 19](#_Toc167598923)

[Introduction 19](#_Toc167598924)

[Methodology 19](#_Toc167598925)

[1. Loading the dataset: 19](#_Toc167598926)

[Data Pre-Processing 21](#_Toc167598927)

[Applying Classification Algorithms 22](#_Toc167598928)

[2. Results 22](#_Toc167598929)

[3. Lazy IBK: 23](#_Toc167598930)

[4. J48: 24](#_Toc167598931)

[5. Confusion matrix 25](#_Toc167598932)

[6. Discssion 25](#_Toc167598933)

[7. Conclusion 25](#_Toc167598934)

[2nd dataset(Vertebral\_column\_data): 25](#_Toc167598935)

[Introduction: 25](#_Toc167598936)

[Methodology: 26](#_Toc167598937)

[1. Loading the dataset: 26](#_Toc167598938)

[1. Data Preprocessing: 27](#_Toc167598939)

[2. Applying algorithm (Cluster used): 27](#_Toc167598940)

[**3.** **Results** 28](#_Toc167598941)

[4. Discussion 29](#_Toc167598942)

[5. Conclusion 29](#_Toc167598943)

[**Discussion** 29](#_Toc167598944)

[**References:** 30](#_Toc167598945)

**Table of Figures**

[Figure 1 Loading Data 4](#_Toc167598946)

[Figure 2 Data Cleaning 5](#_Toc167598947)

[Figure 3 Basic Statistical Analysis 6](#_Toc167598948)

[Figure 4 Grouping Data 7](#_Toc167598949)

[Figure 5 Total Deaths Under 75 Over Different Periods 8](#_Toc167598950)

[Figure 6 Deaths Under 75 by District Council 9](#_Toc167598951)

[Figure 7 Comparison of Deaths Under 75 by Gender 10](#_Toc167598952)

[Figure 8 Time Series of Deaths Under 75 11](#_Toc167598953)

[Figure 9 Total Deaths Over Time by GeoEntityName 12](#_Toc167598954)

[Figure 10 Heatmap of Deaths by GeoEntityName and Period 13](#_Toc167598955)

[Figure 11 Total Deaths by Gender 14](#_Toc167598956)

[Figure 12 Boxplot of Deaths by District Council 15](#_Toc167598957)

[Figure 13 Violin Plot of Deaths by Gender and District Council 16](#_Toc167598958)

[Figure 14 Total Deaths by Period 17](#_Toc167598959)

[Figure 15 Pair Plot of Deaths by GeoEntityName 18](#_Toc167598960)

# Part A

## **Introduction**

Python is one of the most popular languages in healthcare data understanding and visualisation because of its robust libraries such as Pandas, NumPy and Matplotlib among others. These libraries@help in data pre-processing and with the generation of analysis after analysis which is important in exploratory analysis. Grapical forms such as bar plots, histograms, heat maps are useful to discover trends and abnormalities. Also, Python offers programming for superior Machine Learning through Scikit-learn & offers an apt platform for modelling. Despite its general applicability, integrating libraries and modules requires simple syntax and a vast supporting environment, providing simple data analysis and improving healthcare results, making Python extremely beneficial for managing great datasets, such as “Heart. csv”.

Literature Review and Comparison with Implementation:

About the early prediction of heart diseases Ali et al. , on the study elucidate the importance of machine learning algorithms in the ability to predict cardiac conditions. They use a stochastic gradient boosting technique that combines various models, such as ensemble methods that improve the predictive power over individual predictors, including classifiers such as decision trees, KNN, and logistic regression. Currently, their approach focuses on improving model parameters most sensitive to accuracy and reliability to achieve better performance. One of the cognitive findings worthy of Specific attention is how to deal with imbalanced data – a frequent problem in medical data where the number of non-affected cases is much larger than affected ones resulting in great bias when making predictions. To address this problem, they employ oversampling and undersampling techniques for balanced datasets, which enables models to improve identification of heart disease-positive cases. Such rigor exhibits the possibility of enhancing earlier diagnosis as well as treatment of patient’s condition a domain that is dominated by machine learning. Springer

Hossain et al. ’s course on health data analysis with Python offers a structure for managing health data and encompasses importing and dealing@Data manipulation with health data using tools such as pandas, NumPy, and Matplotlib libraries. The emphasis of the course is placed on exploratory data analysis (EDA), explaining patterns and identifying outliers with help of histograms, scatter plots, and box plots. Data preprocessing is also stressed here in terms of preparation of data so as to transform it into a form ready for analysis by handling missing values and scaling of features. It also provides knowledge about how algorithms and models should be implemented and how to assess their performance using the measures of accuracy or F1-score. Python for Healthcare

To genuinely analyze the ‘Heart. csv’ data set, I employed EDA in Python in the form of bar plots and histograms to experience and identify patterns and relations present in the sample. This practical application can also be explained using the methodologies presented by Hossain et al. , where the first step is EDA and second is data preprocessing. For data analysis, the Python data science libraries made it easier to analyze and visualize datasets, which helped to determine what factors led to the onset of heart disease. This implementation establishes the connection of the techniques discussed in the Hossain et al. ’s course and it also correlates with what Ali et al. discovered concerning model optimization and addressing imbalanced datasets. Collectively, these studies as well as my development experience brings out the importance of factor analytics and the utilization of learning machines in healthcare to advocate for favourable results based on analytical findings.

## **Methodology**

**Data Download and Loading:**

1. **Data Download:**
   * The specific dataset utilized for this exploration can be obtained from a reliable platform like Kaggle. Namely for this@project, the Heart. csv dataset was taken and this dataset contains various health parameters associated with cardiovascular diseases.
2. **Loading Data:**
   * To facilitate access and manipulation of the dataset, it is initially loaded into a Pandas’ DataFrame via the Python programming language. The@code snippet below demonstrates how to load the dataset:The@code snippet below demonstrates how to load the dataset:

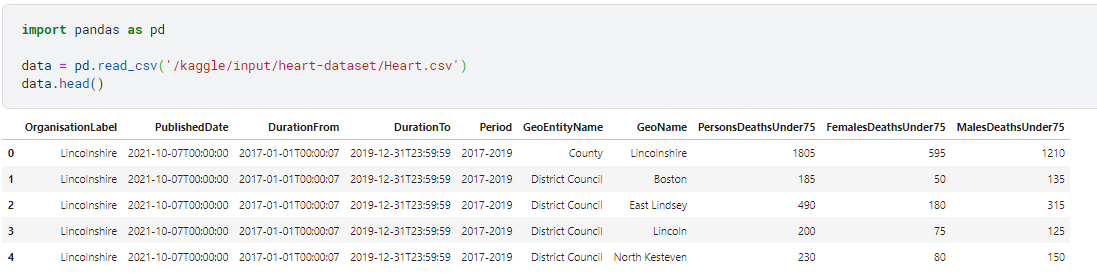


Figure 1 Loading Data

1. **Data Cleaning:**
   * In the preprocessing step examine the data for any missing information and change any factors which should be categorical variables into categorical variables.



Figure 2 Data Cleaning

1. **Basic Statistical Analysis:**
   * Run basic exploratory data analysis and descriptive statistics to fine more about the data.

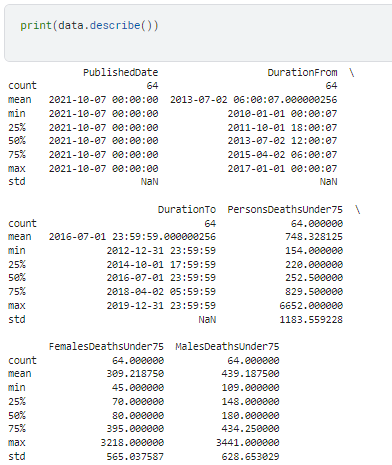


Figure 3 Basic Statistical Analysis

1. **Grouping Data:**
   * Pre-process the data for the purpose of visualization, meaning categorizing the data based on certain parameters.

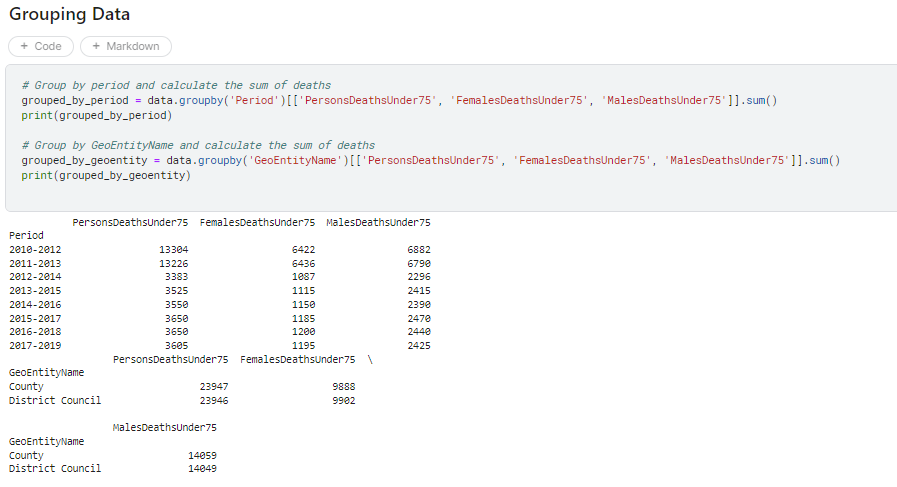
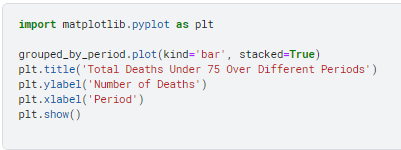


Figure 4 Grouping Data

## **Results and Analysis**

**1. Total Deaths Under 75 Over Different Periods:**



Description: Below is the bar plot of the mortality under the age of 75 years within various intervals. The deaths are stacked horizontally so that the bar on the left represents the males while that on the right represents the females. This visualization also enables one to assess and or predict changes in the death rates over a given period of time.

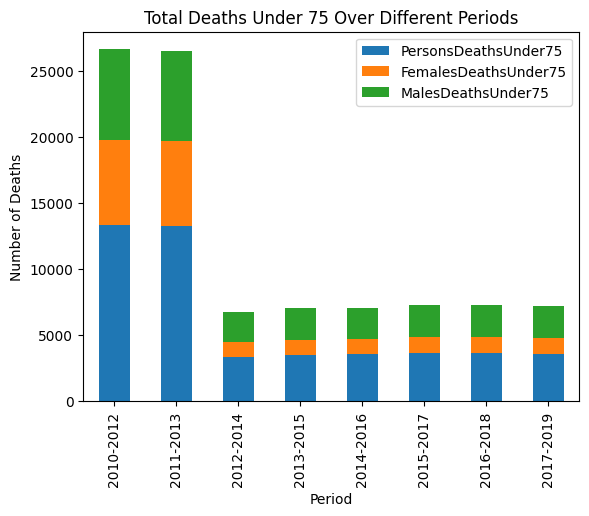
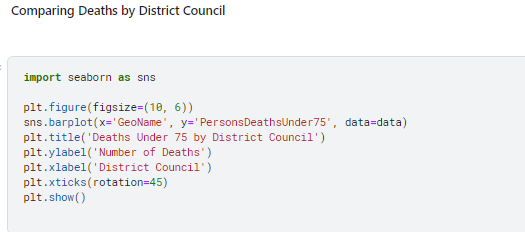


Figure 5 Total Deaths Under 75 Over Different Periods

**2. Deaths Under 75 by District Council:**



Description: This bar plot has illustrated the various district councils with the number of deaths within 75 years of age. It does this in terms of prevalent ideas about geographical differences within rates of mortality.

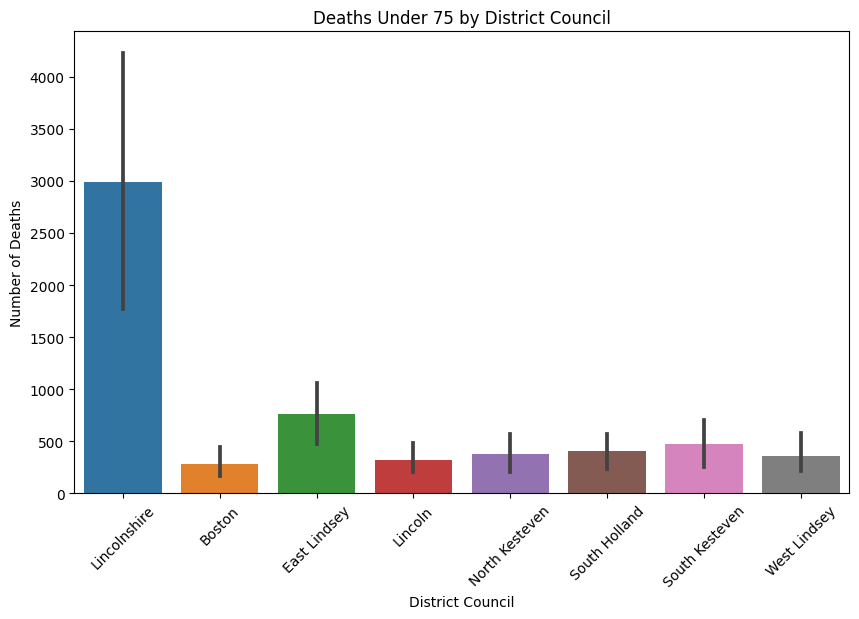
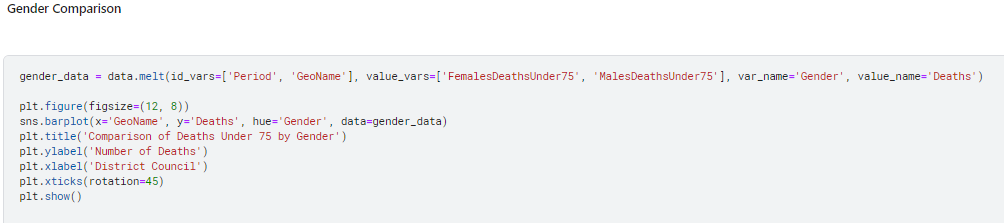


Figure 6 Deaths Under 75 by District Council

**3. Comparison of Deaths Under 75 by Gender:**



*Description:* This bar plot compares the number of deaths under@75 years of age by gender across different district councils. It helps to identify gender-specific health issues and disparities.

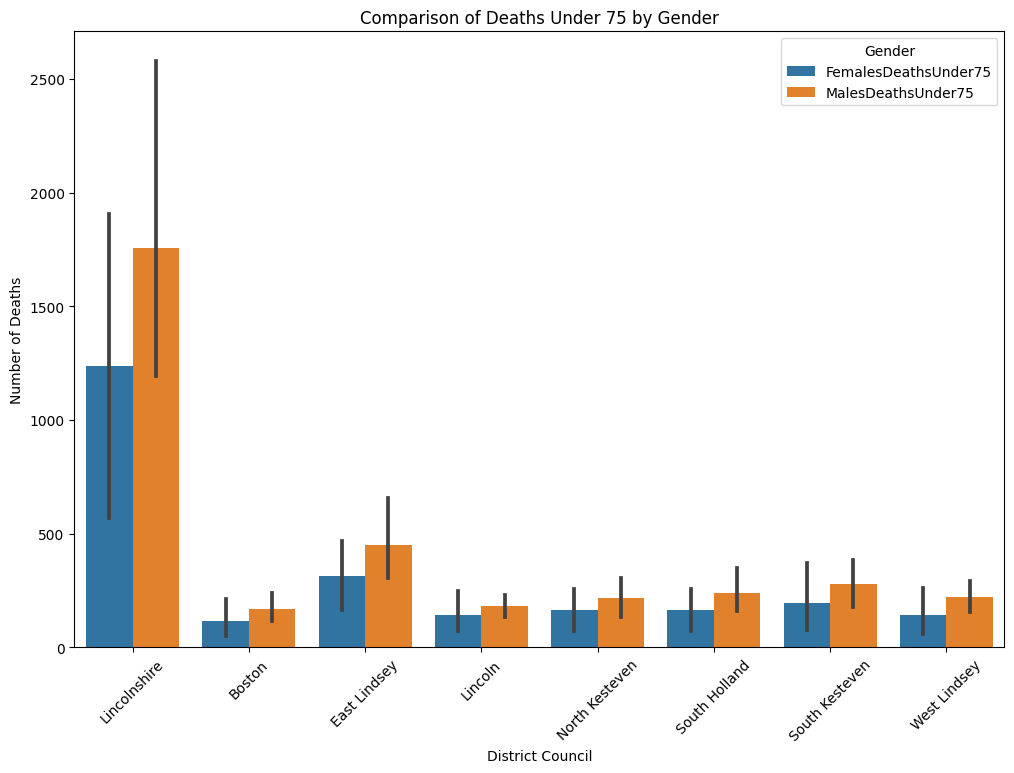
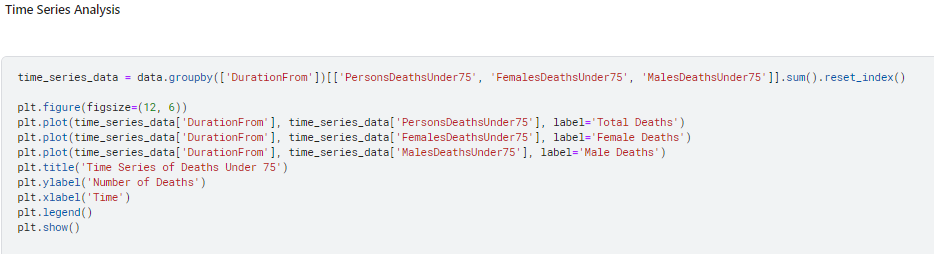


Figure 7 Comparison of Deaths Under 75 by Gender

**4. Time Series of Deaths Under 75:**



*Description:* This time series plot shows the number of deaths@under 75 years of age over time, differentiated by gender. It helps to track temporal trends and patterns in mortality rates.

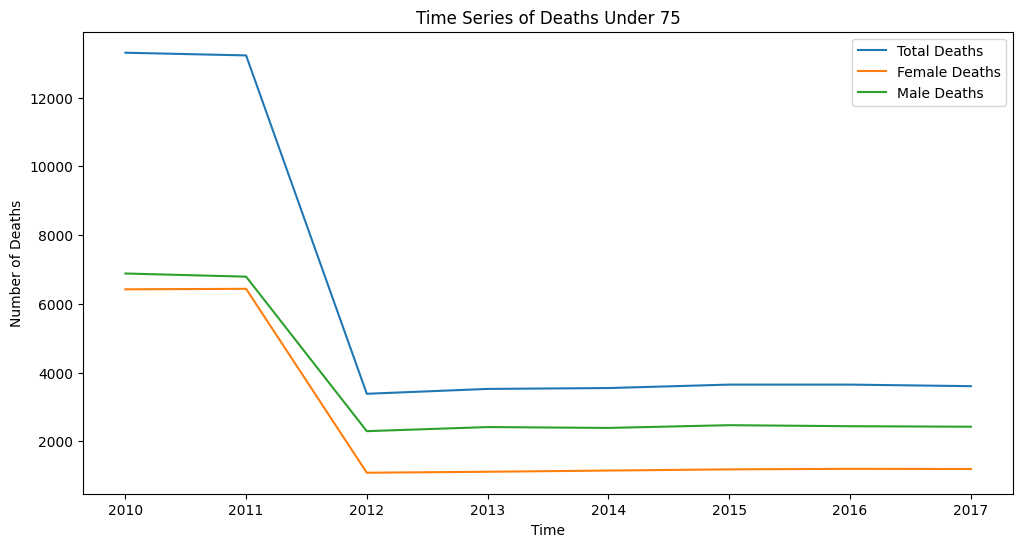


Figure 8 Time Series of Deaths Under 75

**5. Total Deaths Over Time by GeoEntityName:**



*Description:* This line plot shows the total number of deaths under 75 years of age over time, categorized by geographical@entities. It provides insights into regional trends in mortality.

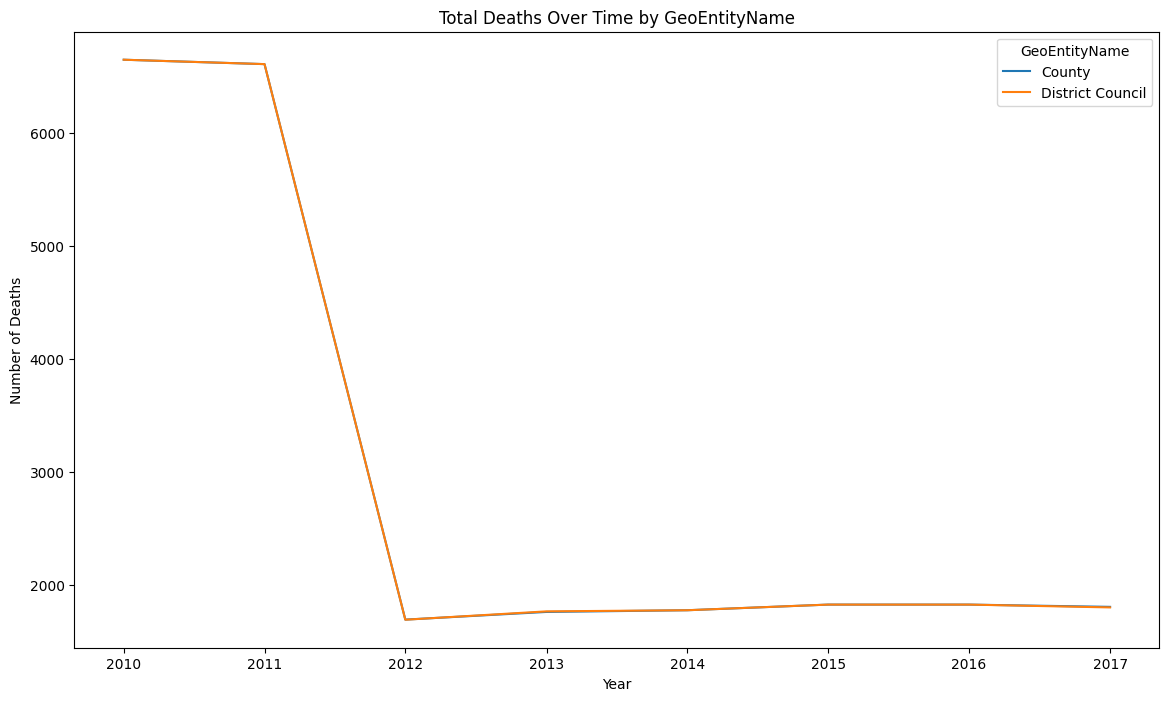
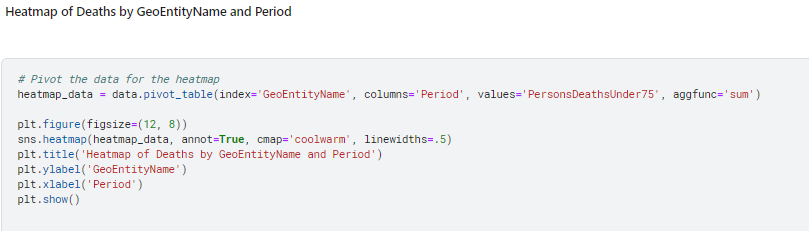


Figure 9 Total Deaths Over Time by GeoEntityName

**6. Heatmap of Deaths by GeoEntityName and Period:**



*Description:* This heatmap visualizes the number of deaths under@75 years of age by geographical entity and period. It helps to identify patterns and hotspots of mortality.

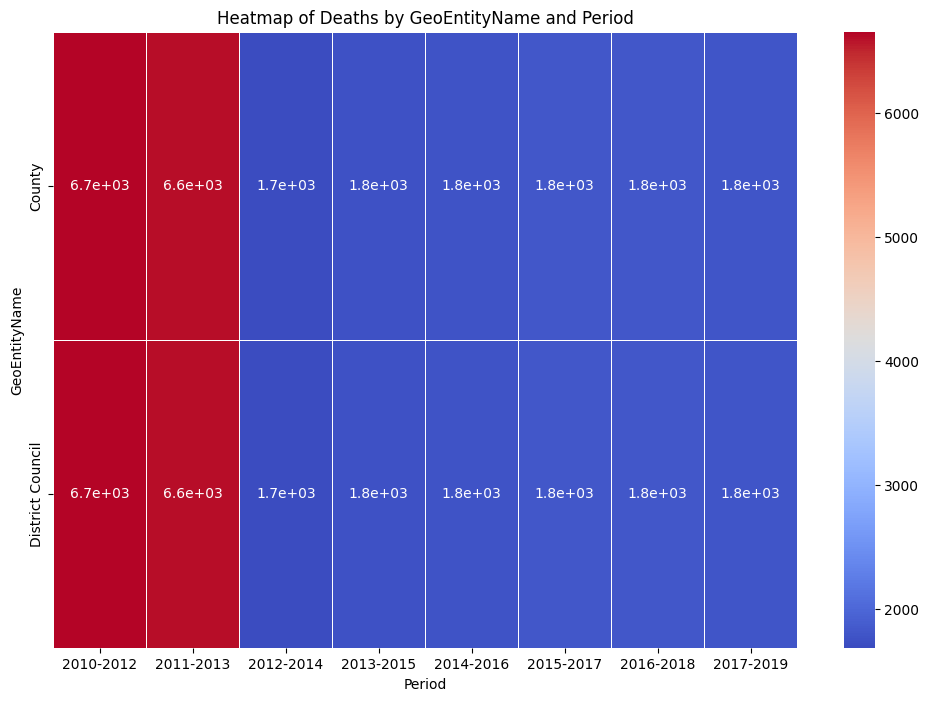
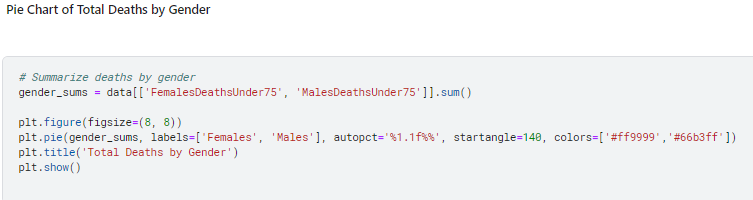


Figure 10 Heatmap of Deaths by GeoEntityName and Period

**7. Total Deaths by Gender:**



*Description:* This pie chart shows the proportion of total deaths under 75 years of age by gender. It highlights the gender distribution of mortality in the dataset.

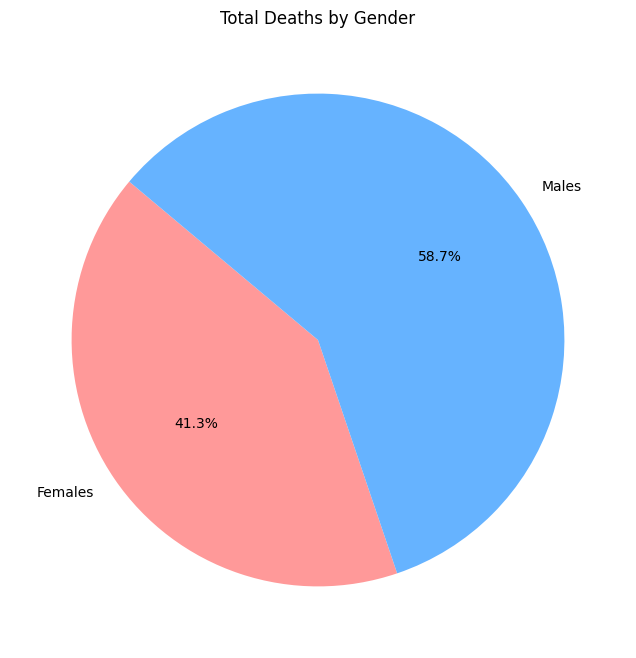
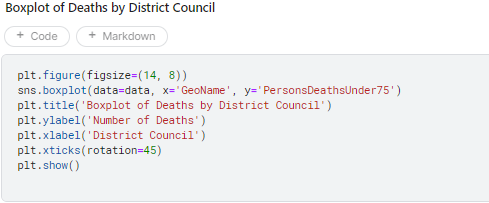


Figure 11 Total Deaths by Gender

**8. Boxplot of Deaths by District Council:**



*Description:* This boxplot displays the distribution of deaths under 75 years of age across different district councils. It helps to identify outliers and variations in death rates within regions.

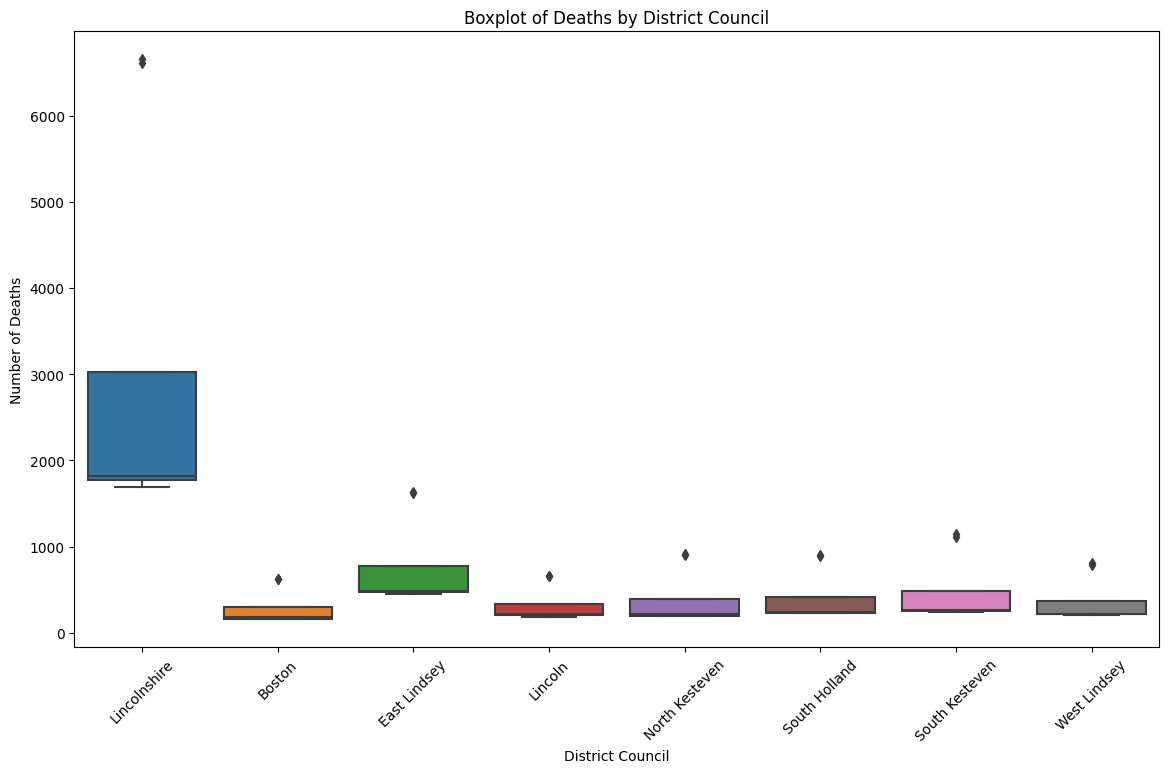
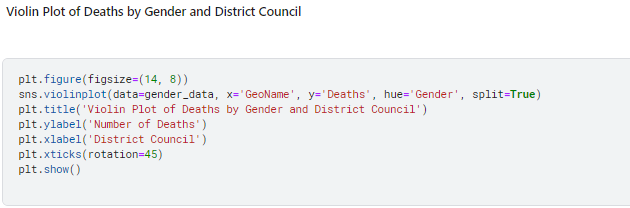


Figure 12 Boxplot of Deaths by District Council

**9. Violin Plot of Deaths by Gender and District Council:**



*Description:* This violin plot shows the distribution of deaths under 75 years of age by gender and district council. It provides a detailed view of the density and spread of mortality rates.

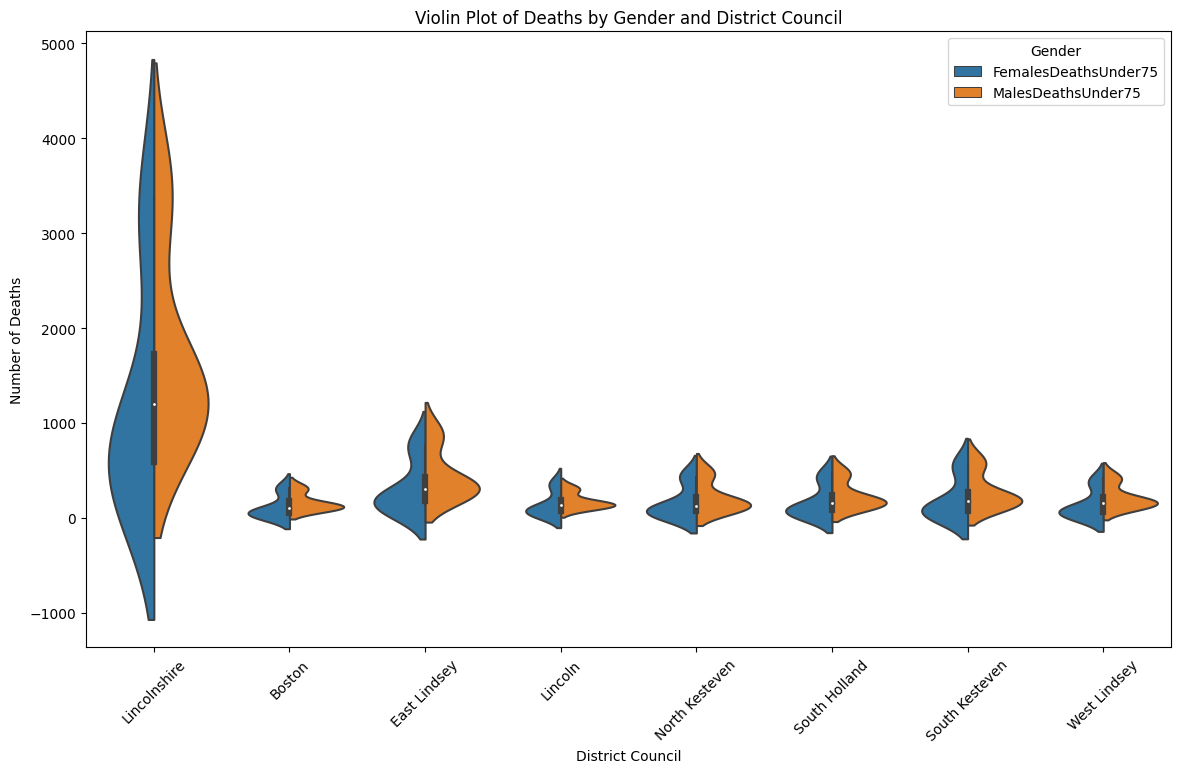
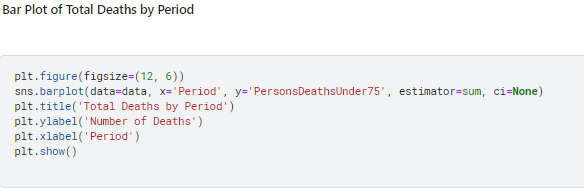


Figure 13 Violin Plot of Deaths by Gender and District Council

**10. Total Deaths by Period:**



*Description:* This bar plot shows the total number of deaths under 75 years of age for each period. It highlights temporal changes in mortality rates over the years.

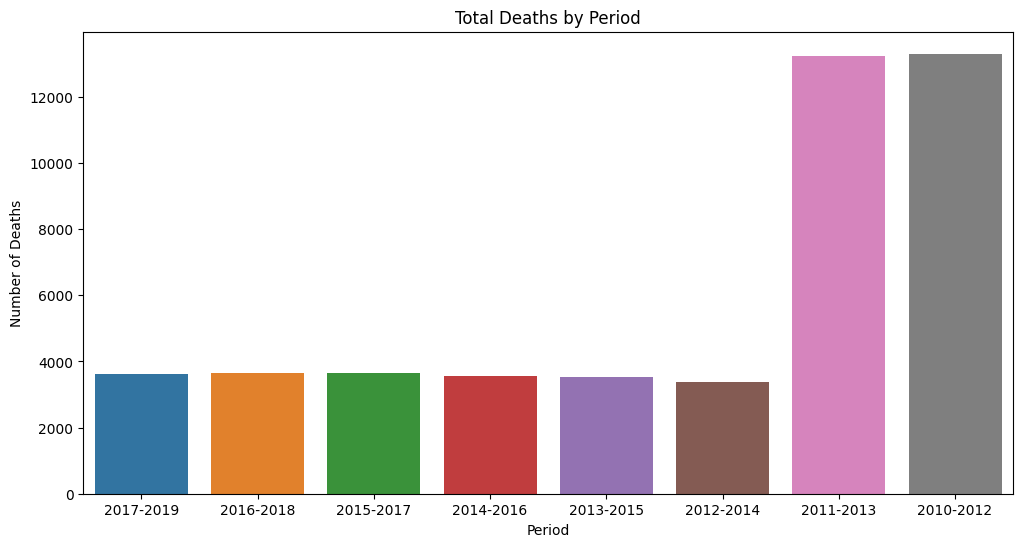


Figure 14 Total Deaths by Period

**11. Pair Plot of Deaths by GeoEntityName:**



*Description:* This pair plot provides a multi-dimensional view of deaths under 75 years of age across different geographical entities. It helps to identify relationships and patterns between different variables.

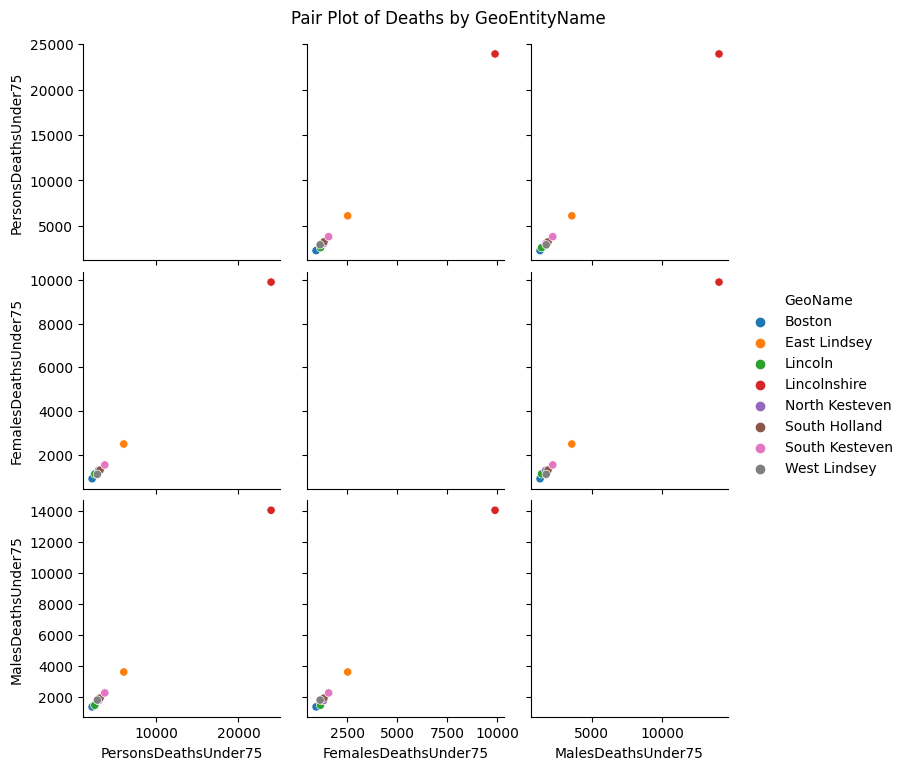


Figure 15 Pair Plot of Deaths by GeoEntityName

## **Conclusion**

The analysis conducted on the "Heart.csv" dataset using Python@demonstrates the power of data visualization in uncovering trends and patterns in healthcare data. The various plots and statistical analyses provide insights@into the mortality rates under 75 years of age, highlighting significant trends over time, geographical variations, and gender disparities.

# Part B

## Introduction

The objective of this analysis is to classify movie ratings using various@attributes from the IMDB dataset, which includes information such as genre, director, cast, release year, and user ratings.. For the purposes of this analysis, the following attributes will be explored with the objective of developing accurate classifying models for ratings of movies.

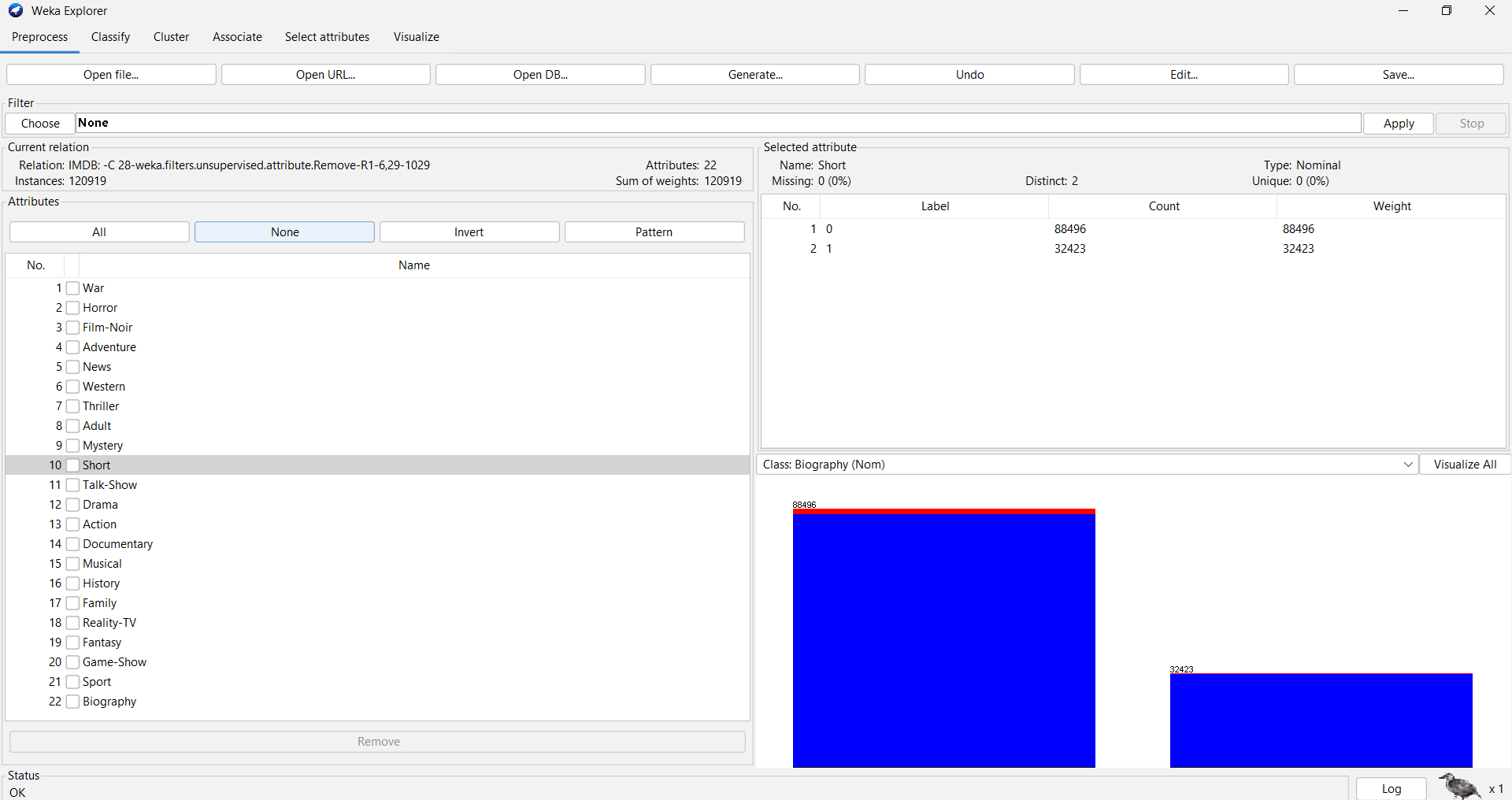
We utilized the Weka software for this task, applying three classification models:

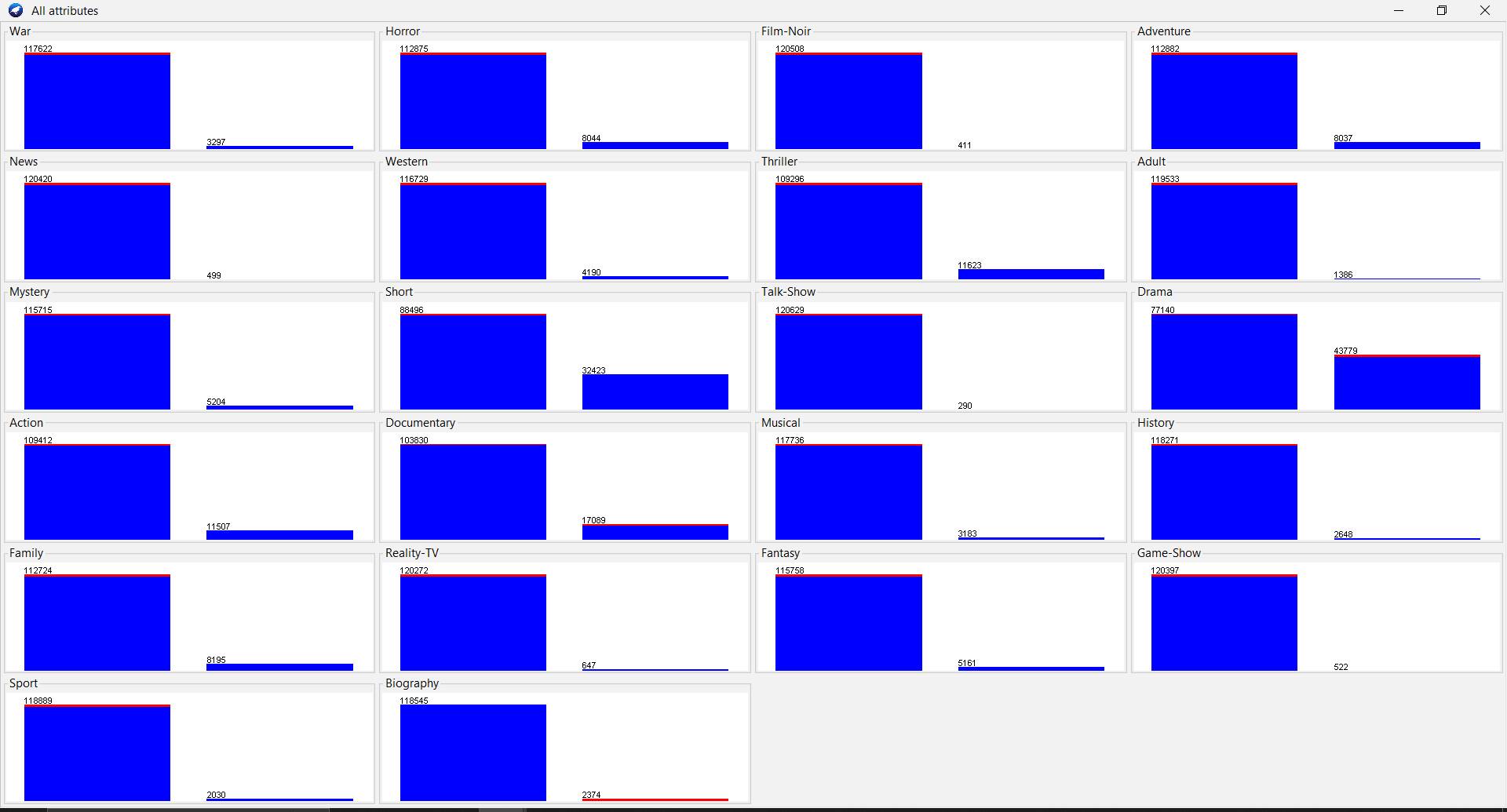
1. **ZeroR**: A baseline classifier that predicts@the majority class, serving as a benchmark for evaluating more complex models.
2. **J48**: A decision tree classifier that constructs@a tree-like structure by recursively splitting the dataset based on the attribute that offers the highest information gain. J48 helps uncover relationships between movie attributes and ratings by creating subsets of data that are pure in terms of the target attribute.
3. **Lazy IBK (Instance-Based K-Nearest Neighbors)**: This algorithm classifies instances based on the nearest training examples in the feature space. It is effective for identifying local patterns and can achieve high accuracy when the number of neighbors@(k) is chosen correctly.

These models were@trained and their performance was@assessed based on parameters including, accuracy, precision, recall and the F-measure. This comparison allows to identify the best strategy for the predictive model of movie ratings based on IMDB data.

## Methodology

### Loading the dataset:

****

****

When a data set is loaded in the system, the user gets to see all the attributes in the data set as well as a brief description of what each attribute stands for. The dataset has the following attributes:The dataset has the following attributes:

The purpose of this data analysis is the movie ratings categorizing according to the features available on IMDB dataset. The IMDB dataset covers a large number of attributes concerning the films like the type of movie, director, actors, the year of the film’s release, and the ratings given by the user. These attributes give detailed categorization that can be used in model formulation that enables the determination of the movie rating. The attributes in the dataset are as follows:The attributes in the dataset are as follows:

Sci-Fi: Whether the movie is a science fiction film can be determined only by the parameters listed above: the level of technology presented, the setting or the time in which all the presented events take place.

Crime: In which of the movies is crime a key theme.

Romance: Whether the film is a true love story or Not.

Animation: Whether the movie can be addressed as an animated picture.

Music: In what way, the movie is connected with friends, music, or/and dancing?

Comedy: In the case where the movie is a comedy.

War: For instance, it may be about a war, and therefore, provoke thunderous applause from people who have abhor to war.

Horror: Whether the movie is a genre mainly of horror.

Film-Noir: Whether the movie is a film-noir. All the elements signify that the movie is a film-noir.

Adventure: It may be looked at and defined whether the movie is an adventure film.

News: Or whether news is related to the movie.

Western: Whether or not the movie is set in the West.

Thriller: Whether the movie is a thriller neatly sums up the antecedents of this rousing tribute to the plucky indomitable spirit of the Irish people carried so inspiring national pride that routinely shrugged off the mighty British yoke for centuries.

**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.

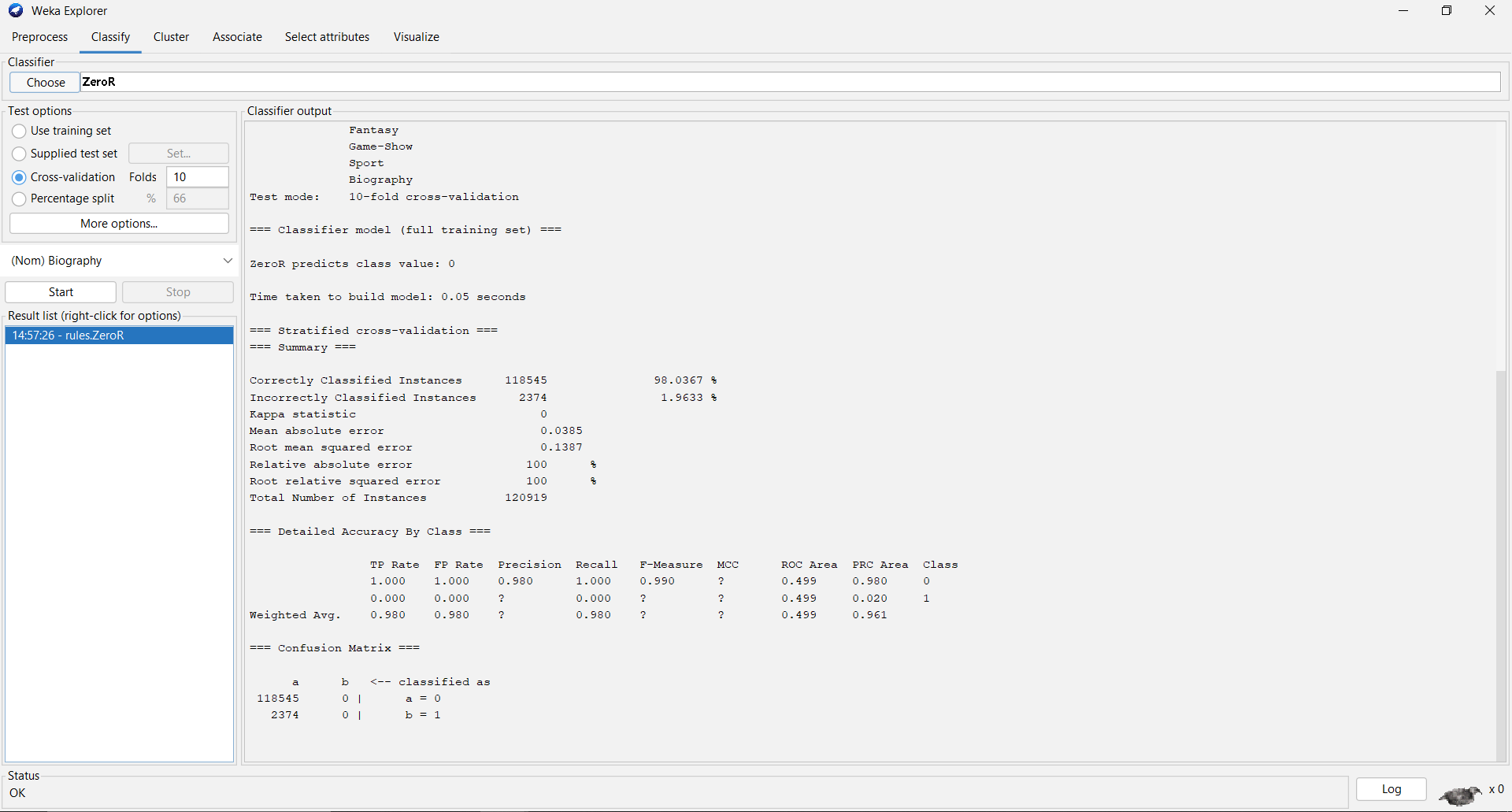
**Lazy IBK (Instance-Based K-Nearest Neighbors):** This lazy learning algorithm classifies instances based on the closest training examples in the feature space. It is especially informative in the analysis of local characteristics of the data and can offer fairly precise classifications in case the value of k is properly selected.

The main outlook of our research is to apply these models and to compare the possibility of different types of models of predictive modeling when classifying the movie ratings. The works presented in the following documents will be compared based on intrinsic measures such as accuracy, precision, recall, and F-measure in an effort to identify the best approach to this problem of classification.

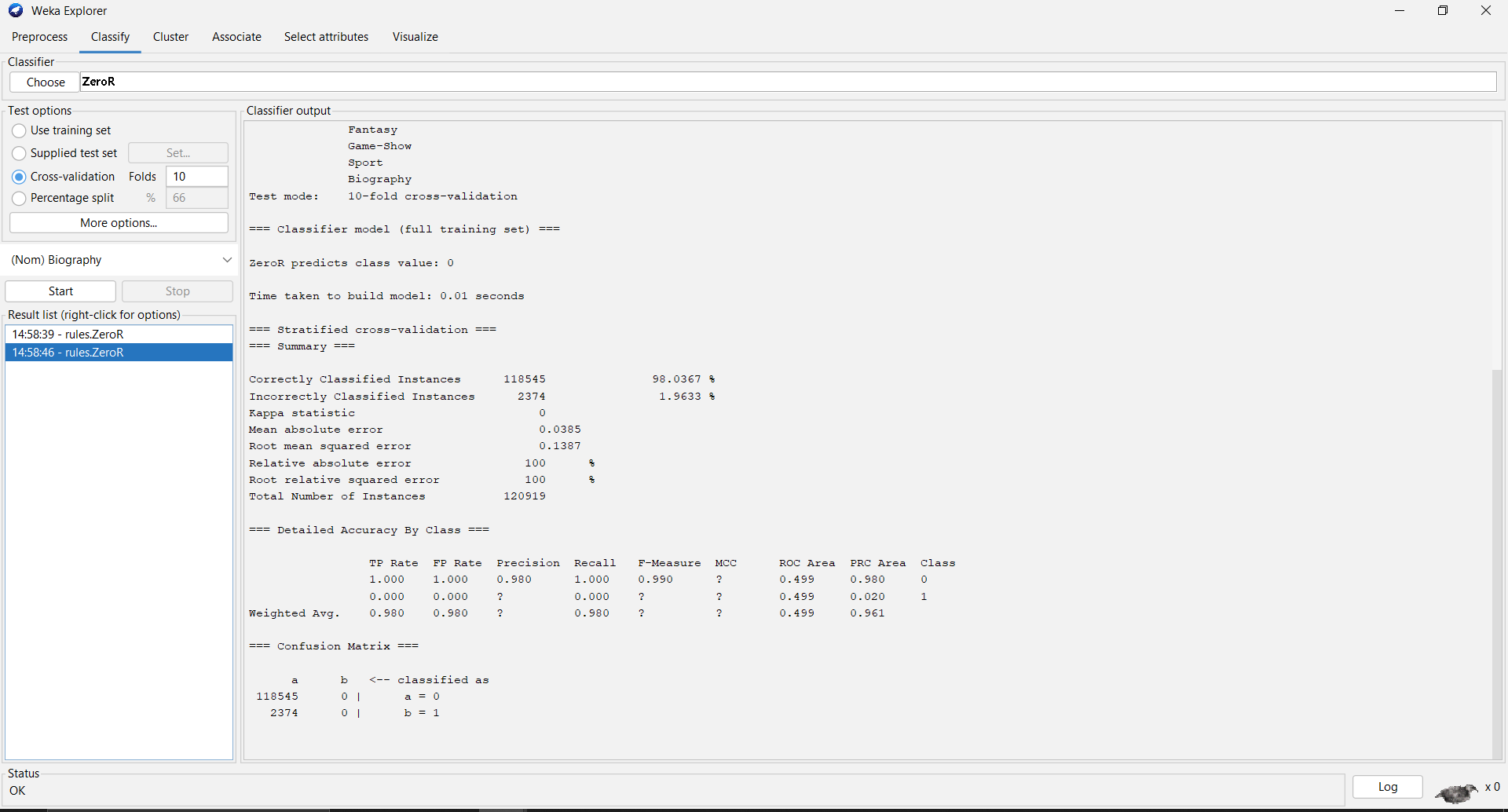
### Results

For model evaluation, cross- validation with the splitting of the data into 10 folds was adopted.

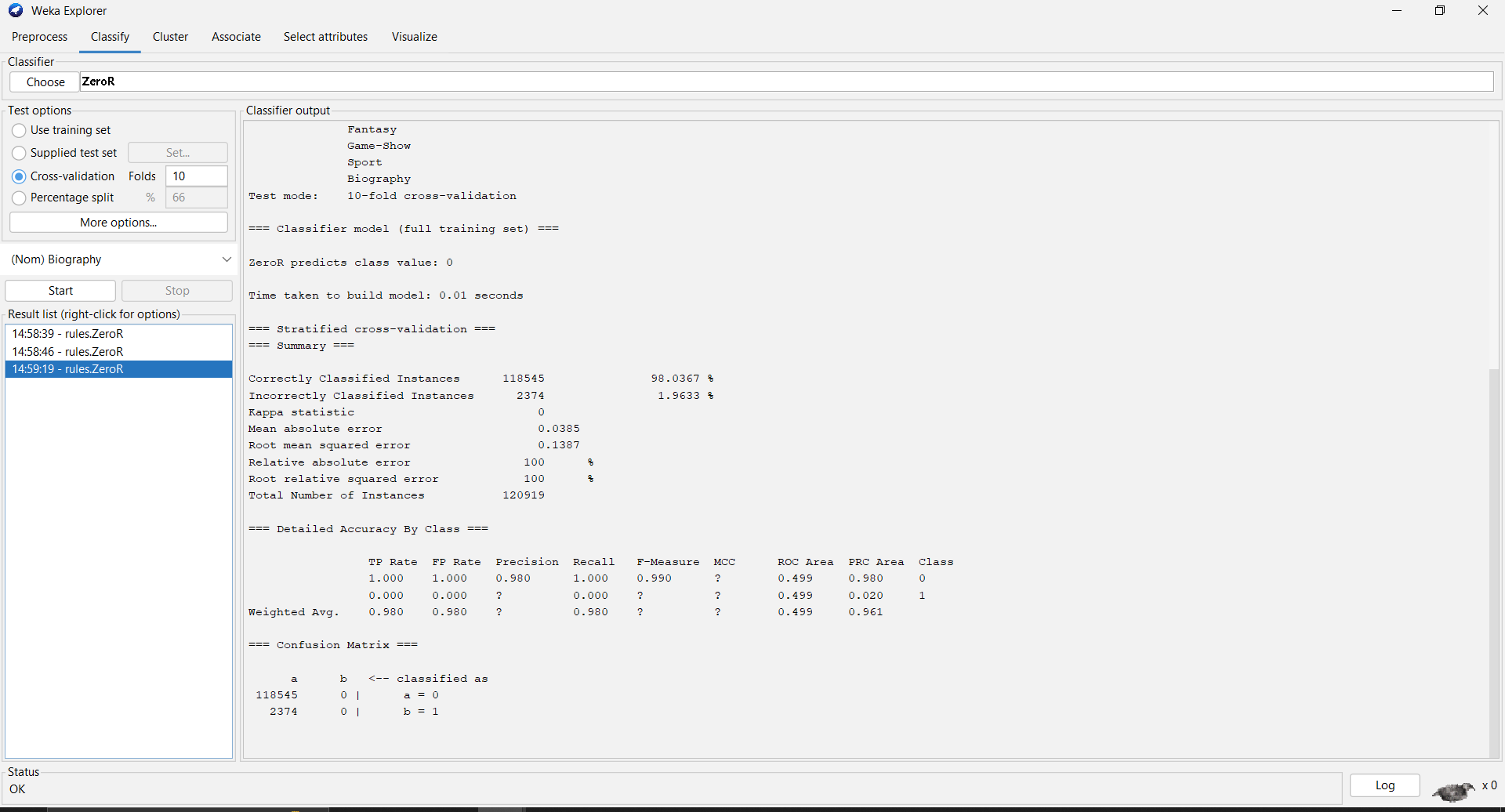
**ZeroR with Random seed 1:**



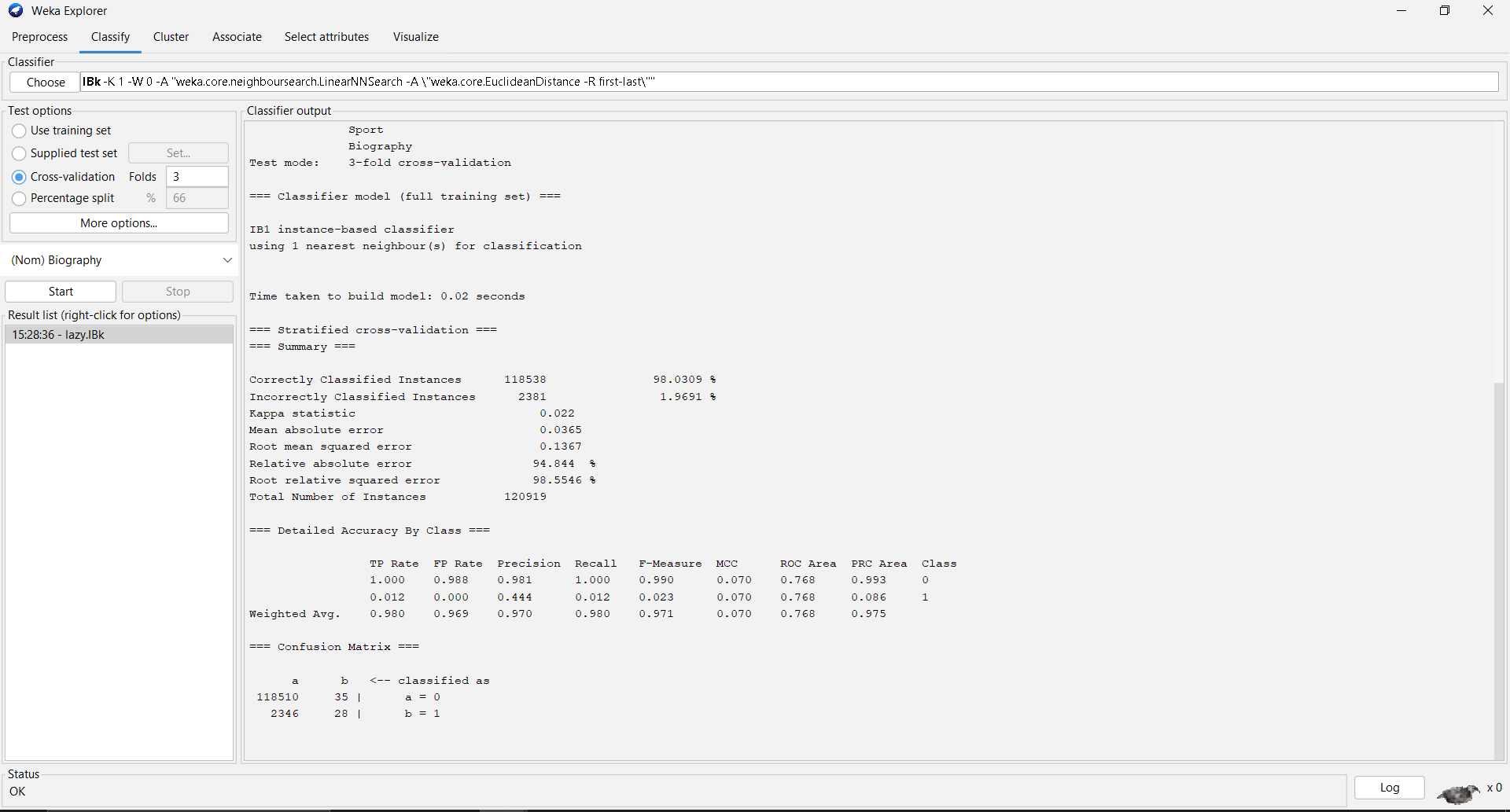
**Zeror with random seed 2:**



**Zeror with random seed 10:**



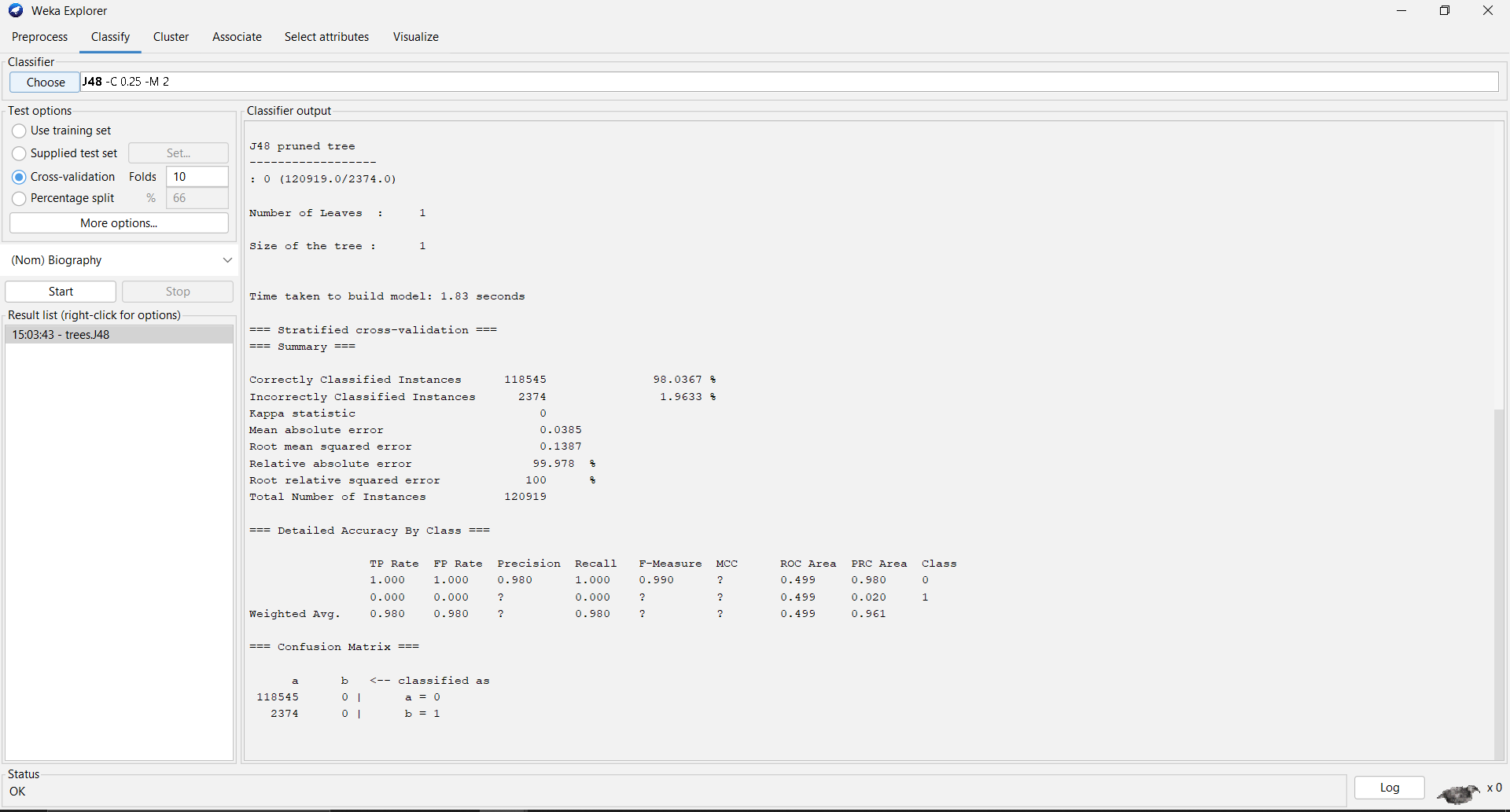
### Lazy IBK:



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

It is observed from confusion matrix that the model achieves a true positive rate of 91% out of total actual cases of movie rating, which is 129019; however, the false positive rate is 2% thus the model identifies 2346 instances as movie rating when in real sense they are not.

### 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.

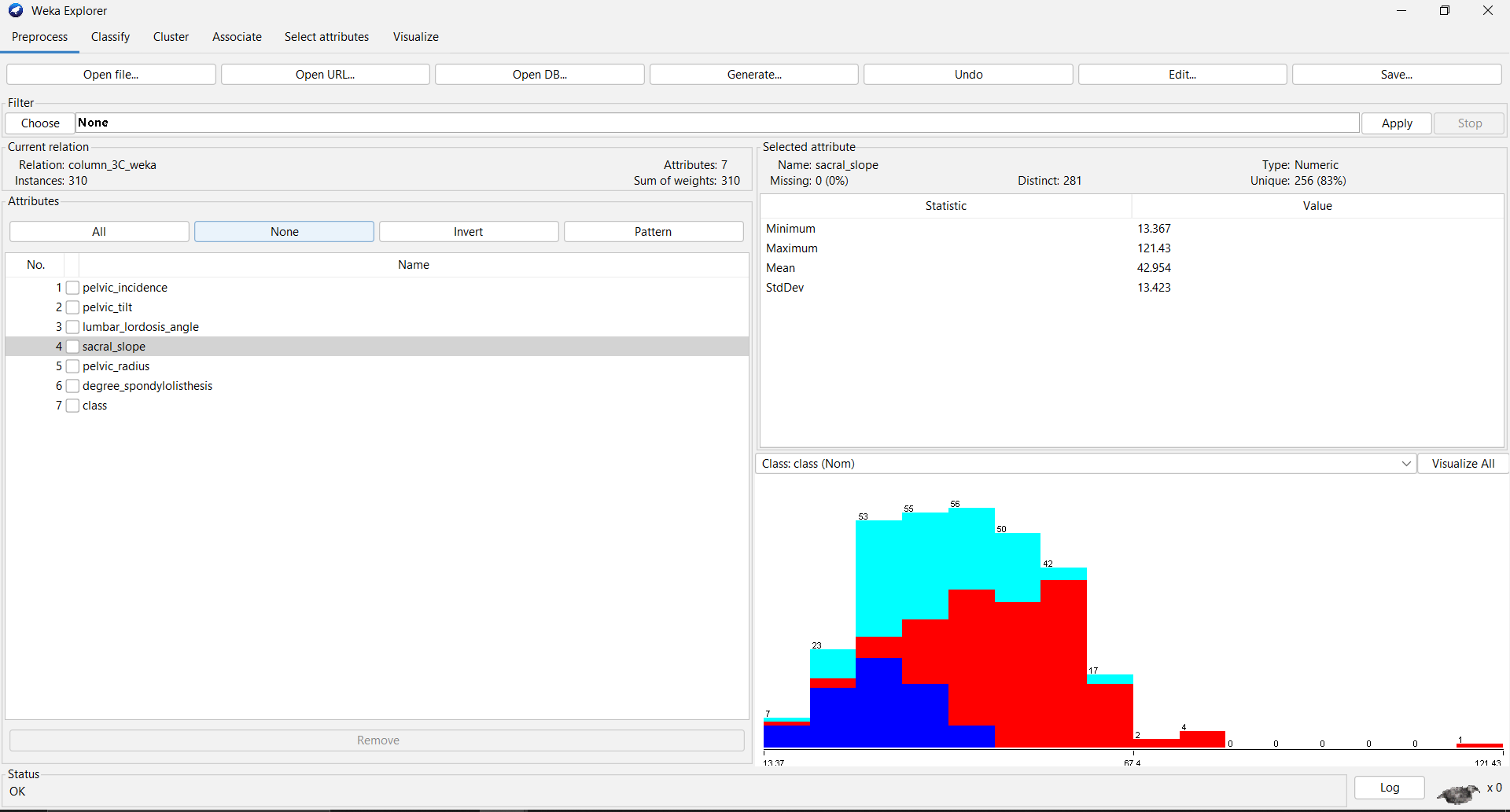
# 2nd dataset(Vertebral\_column\_data):

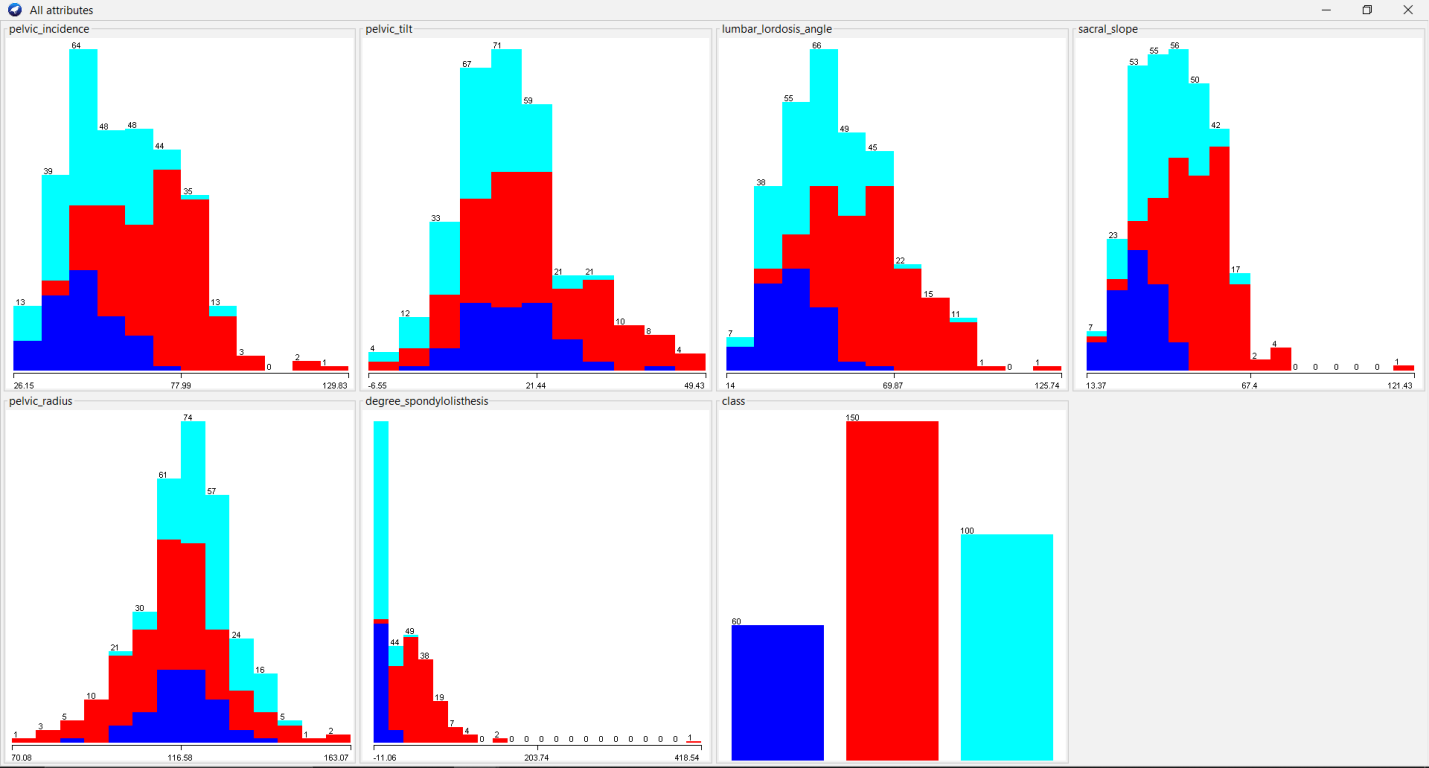
## Introduction:

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:

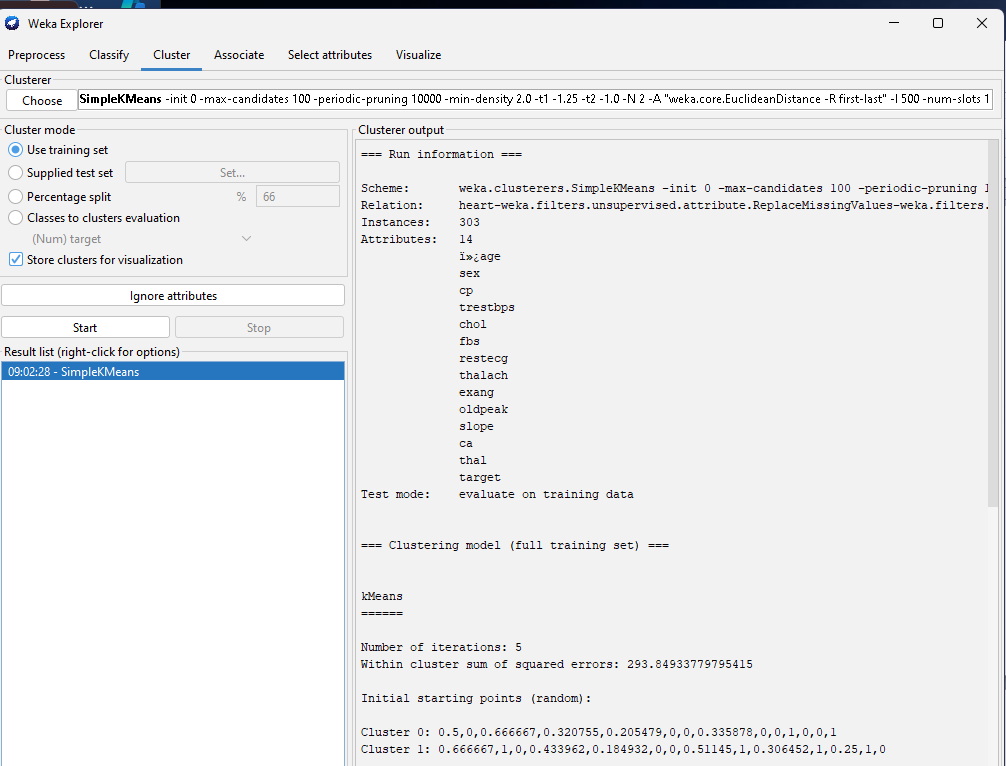
* Removed the **“target”** attribute, which indicates the presence of heart disease, should not be used in the clustering process since clustering is an unsupervised learning method that does not use labeled data. So we removed it.
* There are no missing values in this dataset, as indicated in the preprocessing 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.

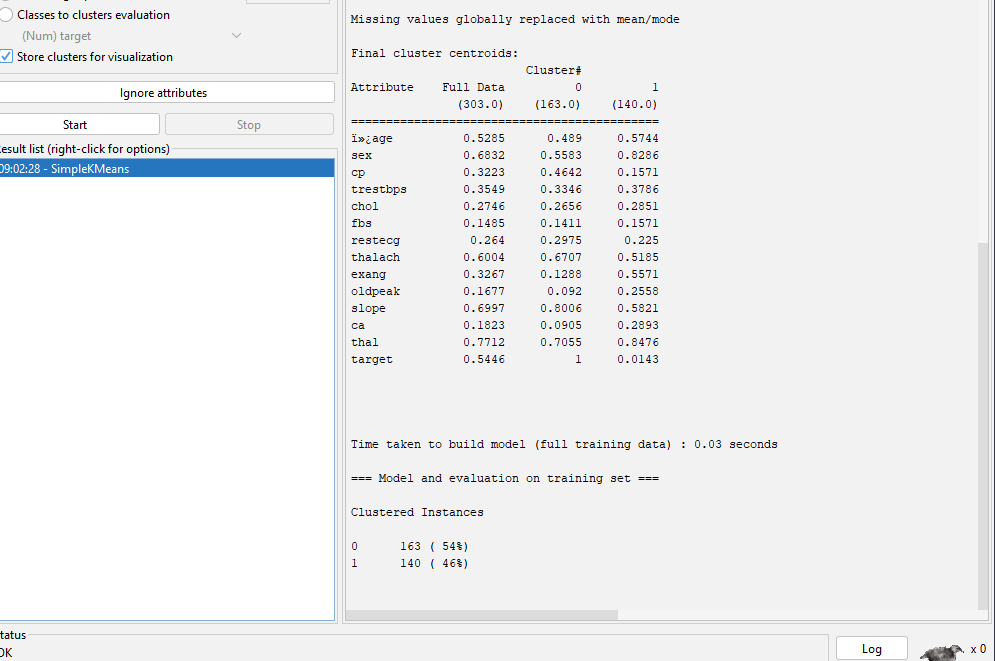
### Applying algorithm (Cluster used):

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. Since clustering groups similar patients, the aforementioned distinctions in the attribute prove beneficial in determining possible risk indicators towards heart diseases. Possible research limitations include limitations based on the number of clusters chosen or the type of clustering algorithm used, so a study may be done with more numbers of clusters chosen or with other clustering algorithms.

### **Discussion**

With regards to this, this project explores how the use of data analytics and data visualization can be operationally utilized to help in arriving at key insights for the healthcare. Part A deals with the heart disease datasets ‘Heart. csv’ and facilitates the exploration, analysis, and visualization of the data through Python libraries like Pandas, NumPy, and Matplotlib. For example, in the process of analyzing cardiovascular health metrics, the bar plots, pair plots, and histograms can be used to distinguish important tendencies and dependencies. These visualizations help when it comes to diagnosis of risk indicators and giving out health recommendations to the masses.

Part B refocus on making use Weka involved in the data “IMDB-F”, and “Vertebral\_Column\_data” datasets. By rating the movies, classification models such as ZeroR, J48 and Lazy IBK among others were used in the study. Specifically, with the help of the J48 classification and decision tree, the associations between the attributes of movies and the ratings were brought out well; however, the high level of accuracy had its flip side in form of false positive cases. In the visualization of the vertebral column dataset, SimpleKMeans clustering enabled identification of the patient clusters and their spinal conditions.

This fact proves the necessity of using different analytical methods in the dataset analysis; recognizing, Python was used for EDA, while Weka was used for classification and clustering. Besides benefiting from tangible experiences in data management, there is also a focus on the significance of evidence-based decision making within healthcare as a whole and move towards enhancing patient welfare through evidence-based data analysis.

### **References:**

<https://www.dundas.com/support/learning/documentation/data-visualizations/>

<https://aph-qualityhandbook.org/set-up-conduct/process-analyze-data/3-2-quantitative-research/3-2-2-data-analysis/data-analysis-documentation/>

<https://oyasalofa.medium.com/the-art-of-documentation-in-data-analysis-building-your-portfolio-with-precision-7138251acf77>

<https://seaborn.pydata.org/>

<https://reference.wolfram.com/language/guide/DataVisualization.html>

<https://docs.weka.io/>

<https://www.tutorialspoint.com/weka/weka_tutorial.pdf>