Assignment 2

24W_CST8390 BI

Hasib (Khesrow) Yosufi

Submission Date Feb 23rd

Contents

1 Introduction

On April 10th, 1912, the RMS Titanic, the largest passenger liner ship ever set sail from Southampton England, bound for New York City (United States). The supposedly unsinkable Titanic carried an estimated 2224 passengers and crew. However, the voyage quickly turned into one of the world's most tragic accidents. After being at sea for five days, the RMS Titanic struck an iceberg in the cold water of Atlantic Ocean. The collision caused the supposedly unsinkable ship to sink, resulting the loss 1500 lives, making it one of the deadliest sinking of a ship. In this report we are going to use the data from Titanic passengers to perform machine learning algorithms such as Decision Tree, Clustering K-Mean and Outlier detection to predict the survival rete of RMS Titanic.

2 Business Understanding

The Titanic dataset comprises 12 attributes. These attributes include (passenger class, name, sex, age, number of siblings or spouse on board, number of parents or children on board, ticket number, passenger fare, cabin, port of embarkation, lifeboat, and survival status). Among these attributes, five contain missing values. The objective of this report is to utilize these attributes to predict the survival rate of the passengers. We will employ the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology to conduct our machine learning predictions. The Titanic data set has 1309 instances, out of which 809 of them did not survive and only 500 survived.

Survived	500
Death	809
Total instances 1309	

3 Data Understanding

3.1 Collect Data

In this report, we will perform Decision Tree (C4.5), Clustering K-Means, and Outlier detection using LOF and distance-based methods. Therefore, we will focus on selecting numeric and nominal attributes. Attributes such as Name, Ticket Number, Cabin, Port, Lifeboat are the irrelevant features in this context therefore, they will be ignored when building the model.

3.2 Describe Data

There are 12 attributes in Titanic data set, there are four numeric data types, two binominal and, six polynominal types. They are:

Attribute	Description	Type
Name	Full name of the RMS Titanic passengers.	Polynominal (String)
Passenger Class	Describe the socio-economic status of passengers. Similar to airplane class (First, Business, economy) this also describe the class for each passenger.	Nominal (nominal)
Sex	Gender of the passengers	Binominal
Age	Age of the passengers, from the they were born until the accident date.	Real (double)
No of Siblings or Spouses on Board	Number of siblings or spouses each passenger had accompanying them on the Titanic.	integer
No of Parents or Children on Board	Number of Children or parents each passenger had accompanying them on the Titanic	integer
Ticket Number	Unique number on each boarding ticket.	Polynominal (string)
Passenger Fare	Price of the ticket	numeric

Cabin	Describe cabin number or identifier for passengers. Provide specific information about or location to each passenger aboard.	Polynominal (string)
Port of Embarkation	Port of which each passenger boarded	Polynominal(string)
Lifeboat	Describe weather passengers were assigned to life boat or not	Polynominal (string)
Survived	Describe if the passenger survived or not.	Binominal

3.3 Explore Data

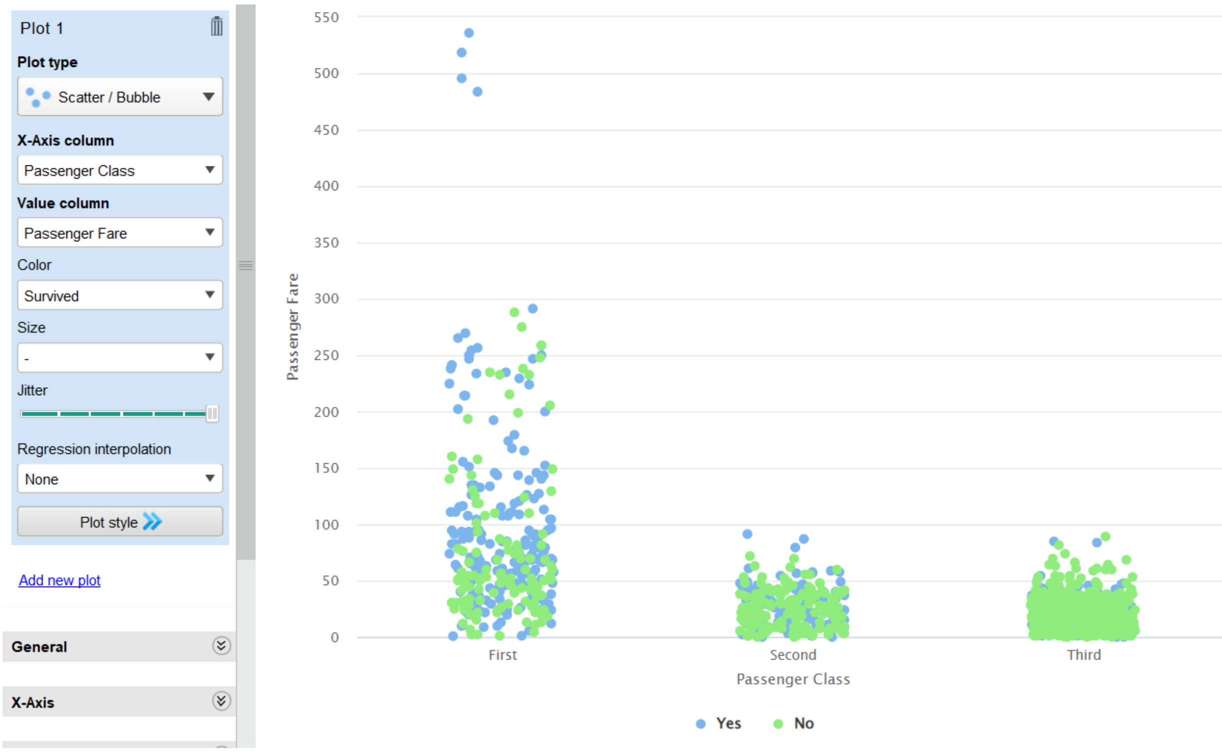
The Following figures shows the relationship between features in scatter graph.

Age (as X) Survived (as Y)

Figure below shows relationship between Age and Survived features, the figure shows that most of the survivors were between age 10 and 40, out of which first and second class mostly survived.

Passenger Class (as X) Passenger Fare (as Y)

The following figure describes passengers who paid more for the ticket and were in more luxury class had higher chance of surviving.



Survived (as X) and Sex (as Y)

Figure below describes how females from higher class had high chance of survival.

3.4 Verify Data

Missing Values: Features Age and Passenger Fares contain missing values which will be handled through serval strategies, explained in detail in the Data Perparation section.

Duplicate Values: no duplicate instance found in the Titanic data set.

Invalid type and type conversion: no invalid value found in any instances, however we are going to normalize feature Passenger Fare to avoid putting too much weight on that feature.

4 Data Preparation

4.1 Select Data

In this report, we will perform Decision Tree (C4.5), Clustering K-Means, and Outlier detection using LOF and distance-based methods. Therefore, we will focus on selecting numeric and nominal attributes. Within this dataset, there are four numeric features (age, number of siblings, number of spouses, and passenger fare), all of which are relevant for Decision Tree classification, Clustering K-Means, and Outlier detection.

As for the nominal features, we will consider those that can contribute to building a precise model. These include 'survived' (the label feature for the decision tree), 'sex' (as most female passengers are relevant and carry weight in building a good model), and 'passenger class' (since most females from first class survived, it should be considered for prediction). The remaining five features (name, ticket number, cabin, port, lifeboat) are irrelevant for prediction and model building and should be discarded.

4.2 Clean Data

In our dataset, the attribute 'Age' contained the most missing values, with about 263 missing values out of 1309. To handle these missing values, we need to consider other attributes such as 'Passenger Fare,' 'Name,' 'Survived,' 'Cabin,' and 'Passenger Class.' These features can provide clues for filling in the missing values. For example, some missing values of the 'Age' attribute had a passenger fare of 7.225. By examining all passengers with a passenger fare of 7.225, we found that they were between the ages of 15 to 45. Additionally, those whose ticket numbers started with 267 were in their 20s and 30s.

We also analyzed the 'Survived' attribute. Since most survivors were women and young teenage males, we assumed that passengers with a passenger fare of 7.225, who were male and survived, might be between 10 to 18 years old. Otherwise, males with a ticket number starting with 267 might be between 20 to 30, while others could be between 30 to 45.

Another important consideration for filling in missing 'Age' attributes is the passenger's name, number of siblings or spouses on board, and number of parents or children on board. For instance, if the 'Name' attribute contained the prefix 'Master' or 'Miss' and the number of parents or children was one or more, the passenger might be between the ages of 9 to 16. Surnames also provided indications; for example, three missing age values with the surname 'Mobarek' had two males and one female. The female had the prefix 'Mrs.' in her name, and the males had the prefix 'Master.' Since the number of parents or children was one or more, we deduced the female might be the mother, aged between 25 to 45, and the males might be her children, aged between 9 to 15.

Considering surnames, we could also identify couples. For instance, if a couple shared the same surname, one was male and the other female, both were from the same passenger class, and their names had the prefixes 'Mr.' and 'Mrs.,' we concluded they were husband and wife. If the husband's age was missing, we assigned an age older than the wife's, and vice versa.

In cases where passengers shared the same passenger fare and ticket number, we took the mean of the age values and assigned it to the missing age values of those passengers. Additionally, in instances where none of the conditions applied, such as passengers with a passenger fare value of 7.75, we observed high entropy for age. In such cases, we took the mean age of those passengers and assigned it to the missing values.

4.3 Construct Data

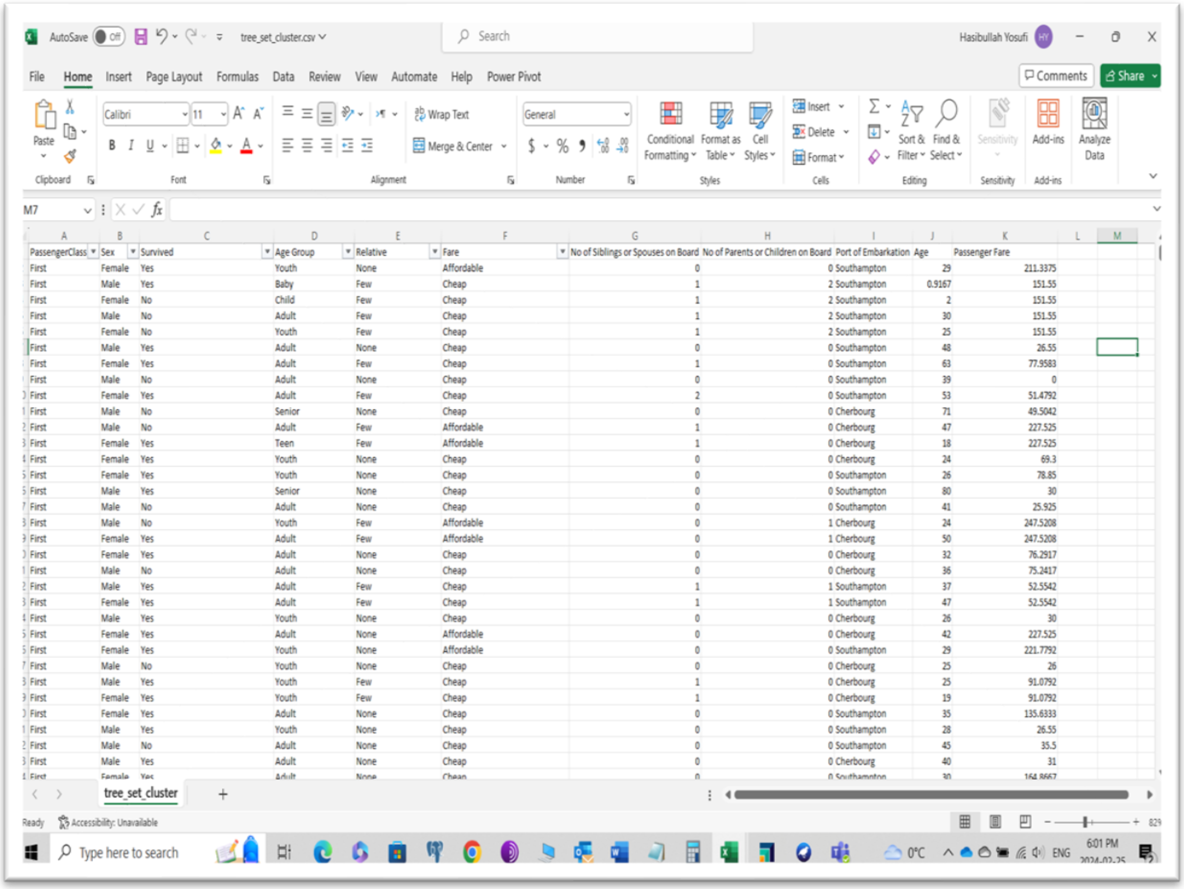
To enhance the accuracy and performance of the C4.5 (Decision Tree) algorithm, we have devised three new categorical attributes derived from numerical attributes. These attributes include:

1. Age Group: This attribute categorizes individuals into specific age groups such as Baby, Child, Teen, Adult, and Senior. It is determined by discretizing the continuous values of the Age attribute.

2. Relatives: This attribute categorizes individuals based on the total number of relatives accompanying them on board. It classifies passengers into categories of None, Few, or Many relatives, calculated by summing the attributes No of Parents or Children on Board and No of Siblings or Spouses on Board.

3. Fare: Utilizing equal frequency binning, this attribute assigns passengers to fare categories such as Cheap, Affordable, or Expensive, derived from the continuous values of the Passenger Fare attribute. We used frequency binning as to make decision tree more efficient for prediction. We know that most of Titanic survivors were from first and second class so it makes sense for the first and second class to have more expensive fare than the third class passengers.

A new 'Id' attribute is also created in RapidMiner process for outlier detection and to match the results of both the LOF and distance-based methods of outlier detection. These newly created categorical attributes aim to improve the interpretability and predictive performance of the C4.5 algorithm by capturing meaningful patterns in the data.



PassengerClass	Sex	Survived	Age Group	Relative	Fare	No of Siblings or Spouses on Board	No of Parents or Children on Board	Port of Embarkation	Age	Passenger Fare
First	Female	Yes	Youth	None	Affordable	0	0	Southampton	29	211.3375
First	Male	Yes	Baby	Few	Cheap	1	2	Southampton	0.9167	151.55
First	Female	No	Child	Few	Cheap	1	2	Southampton	2	151.55
First	Male	No	Adult	Few	Cheap	1	2	Southampton	30	151.55
First	Female	No	Youth	Few	Cheap	1	2	Southampton	25	151.55
First	Male	Yes	Adult	None	Cheap	0	0	Southampton	48	26.55
First	Female	Yes	Adult	Few	Cheap	1	0	Southampton	63	77.9583
First	Male	No	Adult	None	Cheap	0	0	Southampton	39	0
First	Female	Yes	Adult	Few	Cheap	2	0	Southampton	53	51.4792
First	Male	No	Senior	None	Cheap	0	0	Cherbourg	71	49.5042
First	Male	No	Adult	Few	Affordable	1	0	Cherbourg	47	227.525
First	Female	Yes	Teen	Few	Affordable	1	0	Cherbourg	18	227.525
First	Female	Yes	Youth	None	Cheap	0	0	Cherbourg	24	69.3
First	Female	Yes	Youth	None	Cheap	0	0	Southampton	26	78.85
First	Male	Yes	Senior	None	Cheap	0	0	Southampton	80	30
First	Male	No	Adult	None	Cheap	0	0	Southampton	41	25.925
First	Male	No	Youth	Few	Affordable	0	1	Cherbourg	24	247.5208
First	Female	Yes	Adult	Few	Affordable	0	1	Cherbourg	50	247.5208
First	Female	Yes	Adult	None	Cheap	0	0	Cherbourg	32	76.2917
First	Male	No	Adult	None	Cheap	0	0	Cherbourg	36	75.2417
First	Male	Yes	Adult	Few	Cheap	1	1	Southampton	37	52.5542
First	Female	Yes	Adult	Few	Cheap	1	1	Southampton	47	52.5542
First	Male	Yes	Youth	None	Cheap	0	0	Cherbourg	26	30
First	Female	Yes	Adult	None	Affordable	0	0	Cherbourg	42	227.525
First	Female	Yes	Youth	None	Affordable	0	0	Southampton	29	221.7792
First	Male	No	Youth	None	Cheap	0	0	Cherbourg	25	26
First	Male	Yes	Youth	Few	Cheap	1	0	Cherbourg	25	91.0792
First	Female	Yes	Youth	Few	Cheap	1	0	Cherbourg	19	91.0792
First	Female	Yes	Adult	None	Cheap	0	0	Southampton	35	135.6333
First	Male	Yes	Youth	None	Cheap	0	0	Southampton	28	26.55
First	Male	No	Adult	None	Cheap	0	0	Southampton	45	35.5
First	Male	Yes	Adult	None	Cheap	0	0	Cherbourg	40	31
First	Female	Yes	Adult	None	Cheap	0	0	Southampton	30	164.8667

4.4 Integrate Data

All newly integrated categorical data is seamlessly incorporated into the original dataset. Subsequently, we refine our dataset by selectively excluding any unnecessary attributes using the 'Select Attributes' process in RapidMiner.

4.5 Format Data

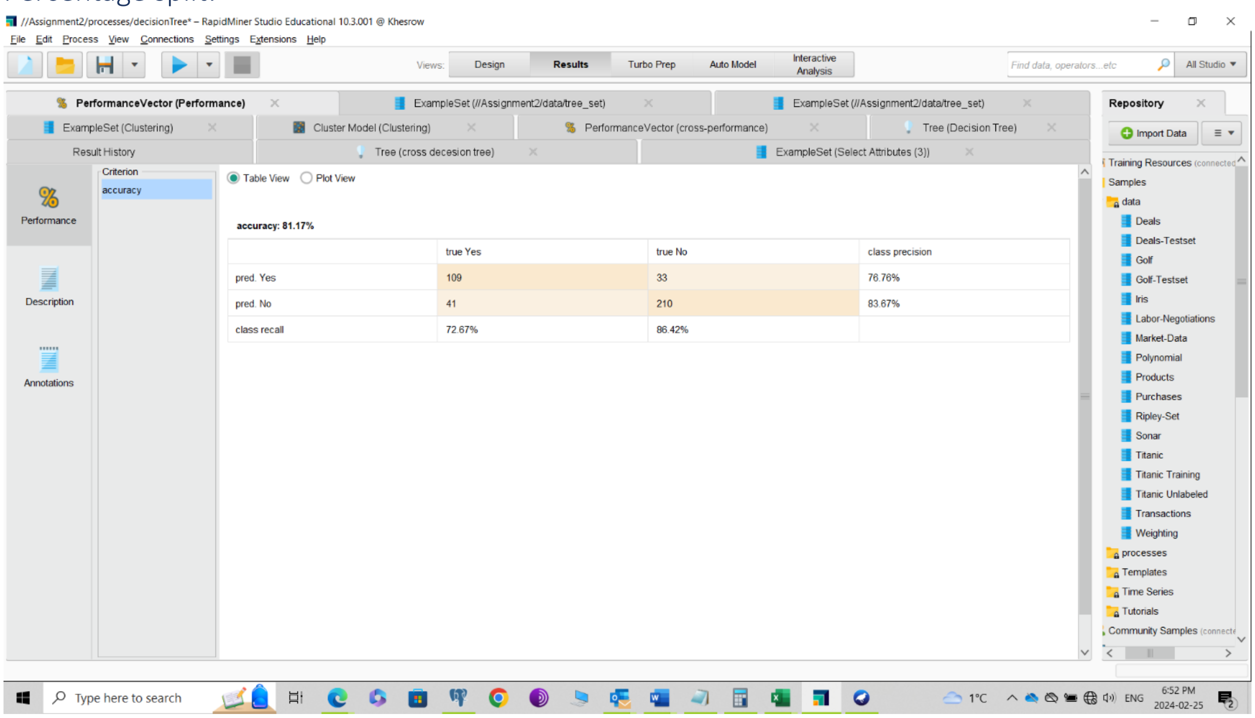
Attributes were converted from numerical to nominal, or vice versa, to enhance compatibility with specific algorithms. For example, categorical types such as Sex, Port of Embarkation, and Passenger Class were converted to numerical types for K-Means clustering and outlier detection.

5 Modeling

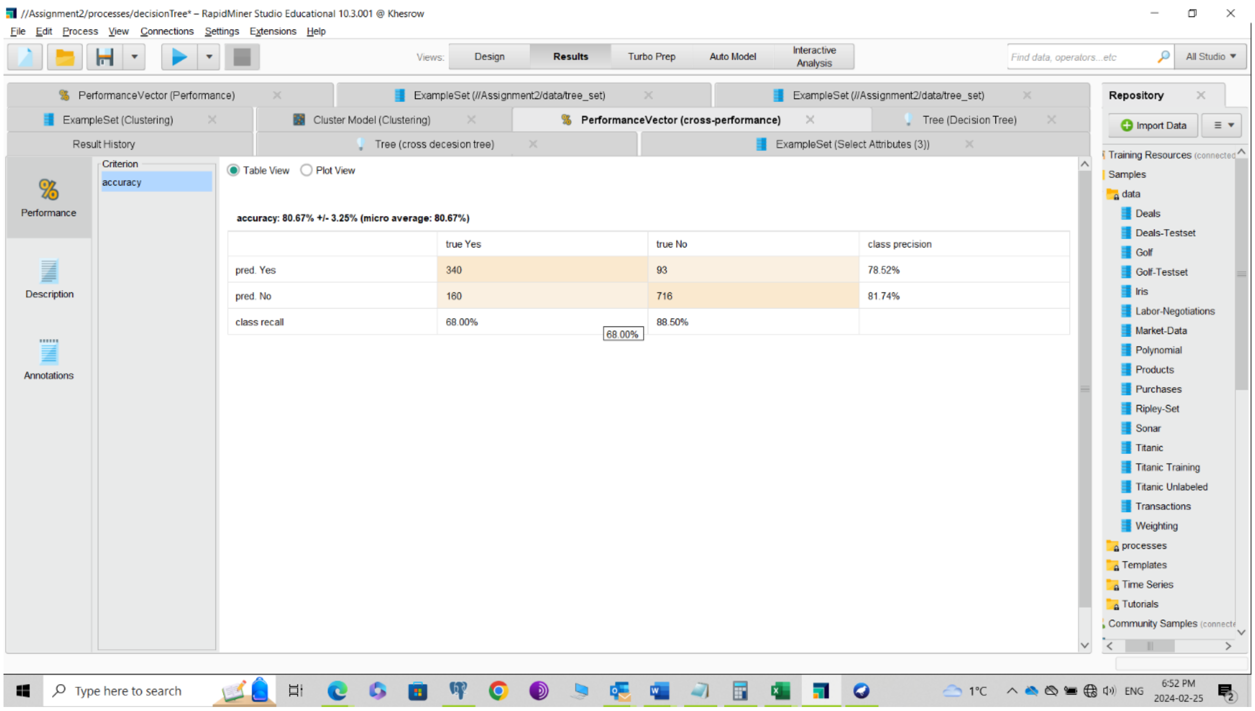
5.1 Decision Tree (C4.5)

We have used two approach for decision tree one with cross validation and the other percentage split. We can get the accuracy up to 80.67 % by using cross-validation classification operator and accuracy up to 81.17% with percentage split.

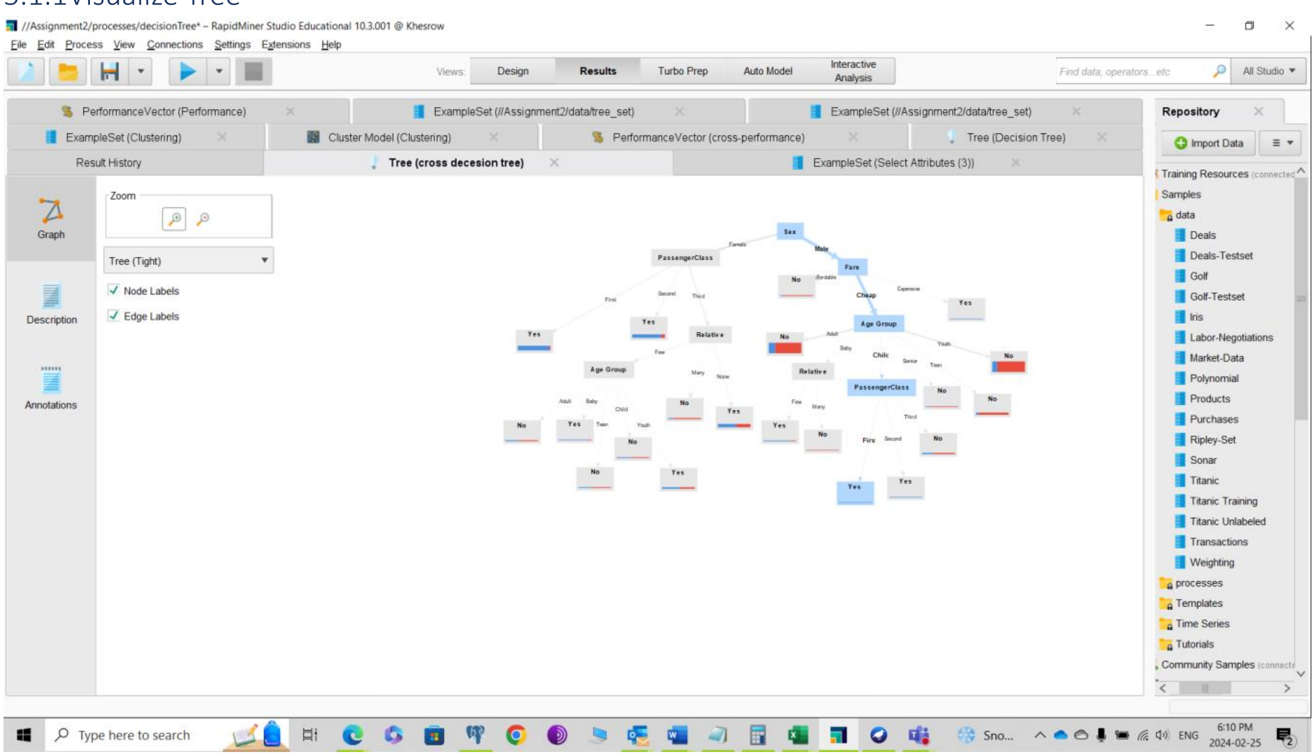
Percentage split:



5.1.0 Cross Validation:



5.1.1Visualize Tree

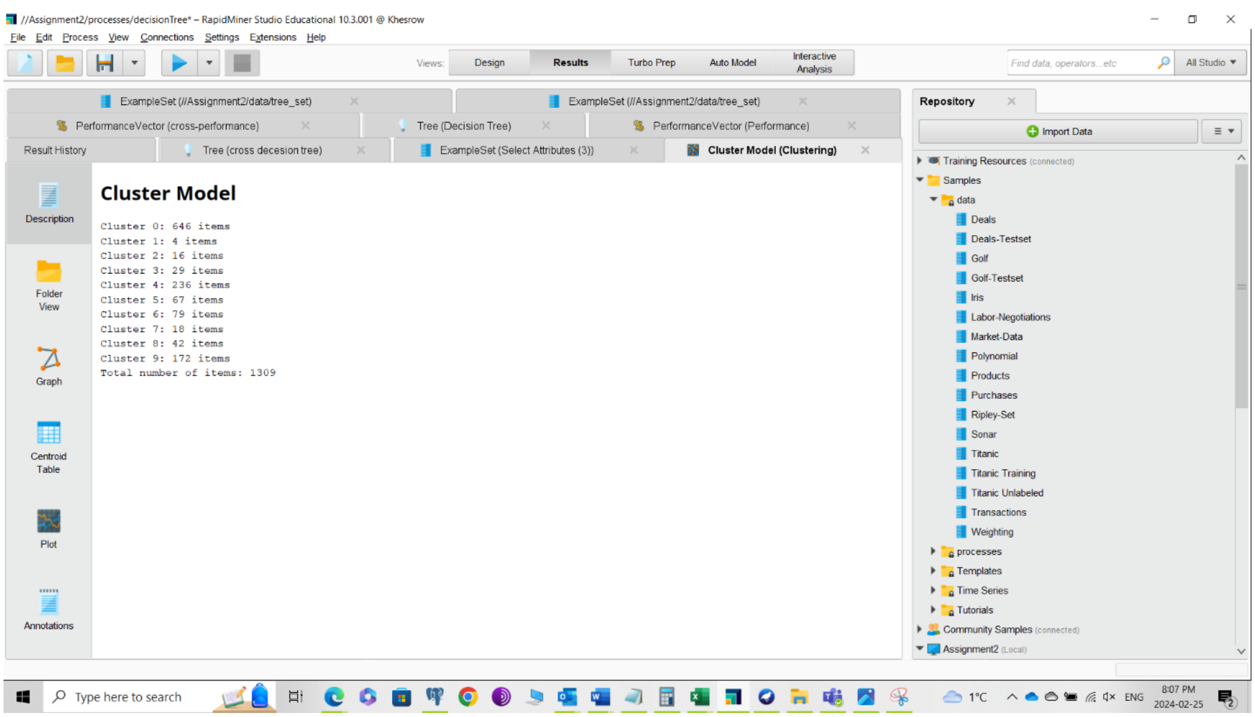


5.1.2 Tree Observation

From the above tree, we can observe that surviving the RMS Titanic accident was not entirely random but followed certain patterns and rules. Female passengers from the first and second classes had a significantly higher chance of survival. Third-class females without relatives were luckier than those with many or few relatives. Even among those with relatives, having a baby increased the chances of survival compared to having teenage or adult relatives. For first or second-class youth or teenage males with fewer relatives, the likelihood of survival was higher than for other males.

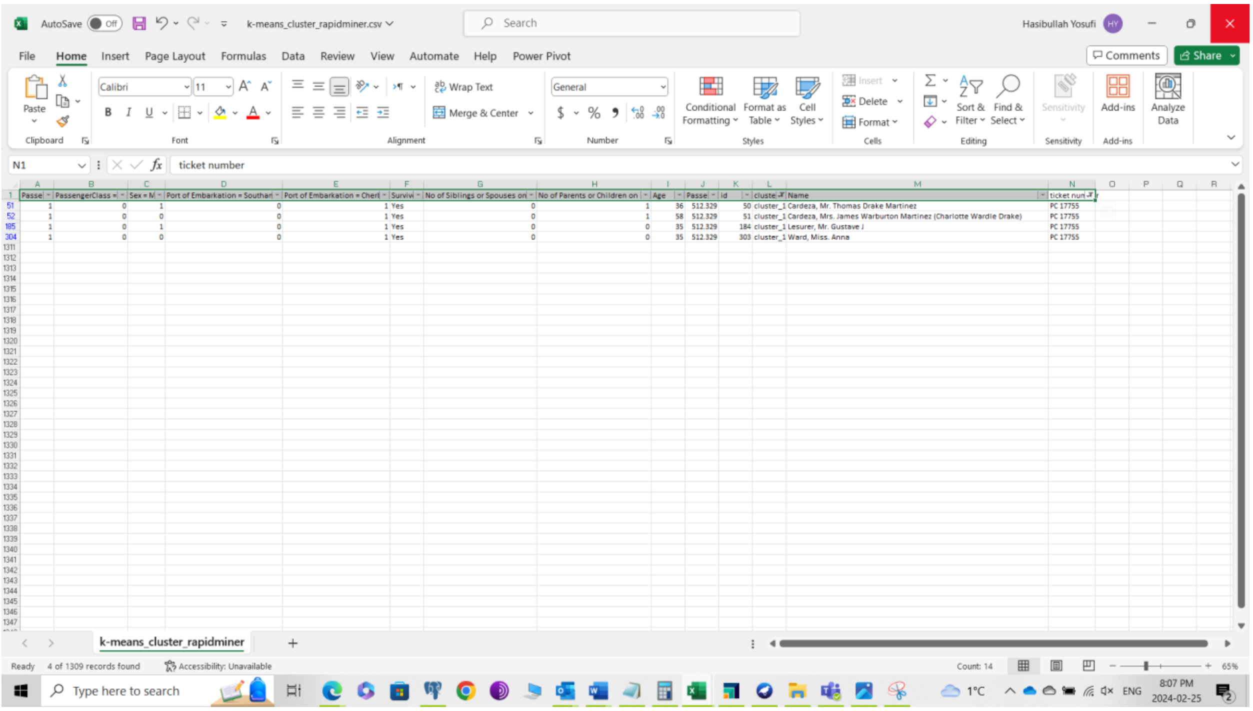
5.2 K-Means clustering

For the K-Mean I put the number of K = 10 and selected attributes Age, No of Parents or Children on Board, No of Siblings or Spouses on Board, Passenger Fare, Passenger Class, Port of Embarkation and Sex. Nominal attributes such as Passenger Class, Port of Embarkation and Sex were converted to numeric using one-hot encoding attribute.



5.2.0 Observation K-Means

We can see in the figure above that cluster one has only four items which can be an indication of outlier or noise in our Titanic data set. The following figure is snapshot of cluster one items in csv file which shows all of them have the same ticket number.



	Passe	PassengerClass =	Sex = M	Port of Embarkation = Southan	Port of Embarkation = Cherl	Surviv	No of Siblings or Spouses on	No of Parents or Children on	Age	Passe	id	cluste	Name	ticket num
51	1	0	1	0	1	Yes	0	1	36	512.329	50	cluster_1	Cardeza, Mr. Thomas Drake Martinez	PC 17755
52	1	0	0	0	1	Yes	0	1	58	512.329	51	cluster_1	Cardeza, Mrs. James Warburton Martinez (Charlotte Wardle Drake)	PC 17755
185	1	0	1	0	1	Yes	0	0	35	512.329	184	cluster_1	Lesurer, Mr. Gustave J	PC 17755
304	1	0	0	0	1	Yes	0	0	35	512.329	303	cluster_1	Ward, Miss. Anna	PC 17755

5.3 Outlier (LOF and Distance Based)

We used two method of outlier detection LOF and Distance Based method and combined the results of both method into one file (csv) to compare the results of both methods for better outlier detection.

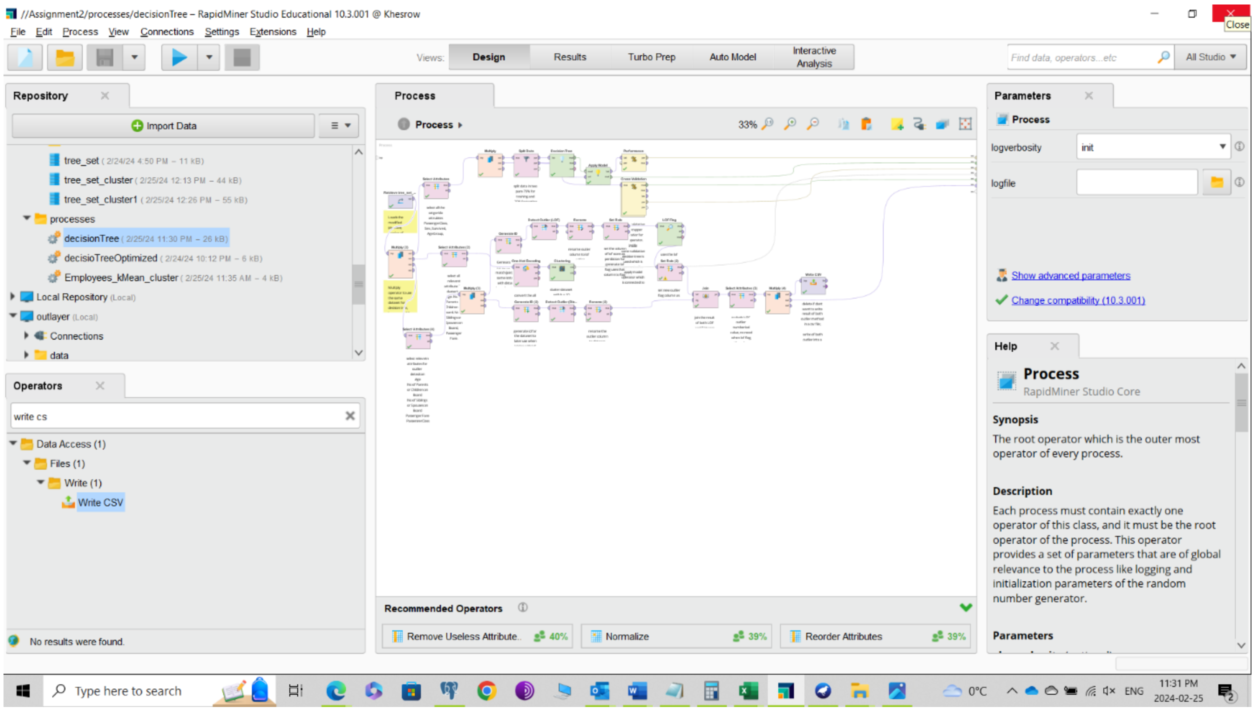
Open in Turbo Prep Auto Model Interactive Analysis Filter (1,309 / 1,309 examples): all

Row No.	id	Survived	outlier_flag	Distance Outlier	PassengerC...	Sex	No of Siblin...	No of Parent...	Port of Emb...	Age	Passenger F...
1	1	Yes	No Outlier	false	First	Female	0	0	Southampton	29	211.338
2	2	Yes	No Outlier	true	First	Male	1	2	Southampton	0.917	151.550
3	3	No	No Outlier	true	First	Female	1	2	Southampton	2	151.550
4	4	No	No Outlier	false	First	Male	1	2	Southampton	30	151.550
5	5	No	No Outlier	false	First	Female	1	2	Southampton	25	151.550
6	6	Yes	No Outlier	false	First	Male	0	0	Southampton	48	26.550
7	7	Yes	No Outlier	false	First	Female	1	0	Southampton	63	77.958
8	8	No	No Outlier	false	First	Male	0	0	Southampton	39	0
9	9	Yes	No Outlier	false	First	Female	2	0	Southampton	53	51.479
10	10	No	No Outlier	false	First	Male	0	0	Cherbourg	71	49.504
11	11	No	No Outlier	false	First	Male	1	0	Cherbourg	47	227.525
12	12	Yes	No Outlier	false	First	Female	1	0	Cherbourg	18	227.525
13	13	Yes	No Outlier	false	First	Female	0	0	Cherbourg	24	69.300
14	14	Yes	No Outlier	false	First	Female	0	0	Southampton	26	78.850
15	15	Yes	No Outlier	false	First	Male	0	0	Southampton	80	30
16	16	No	No Outlier	false	First	Male	0	0	Southampton	41	25.925
17	17	No	No Outlier	false	First	Male	0	1	Cherbourg	24	247.521

5.3.0 Outlier Observation:

In the figure above, we can see that for most instances, the predictions of both methods match (not outliers, false). However, for instances two and three, the distance-based method flags them as outliers. If we examine the other attributes of these instances, such as passenger fare, number of parents, and number of siblings, we find that both instances are identical. Additionally, both instances involve babies. According to our prediction, most females from the first class survived. However, in these cases, the female babies did not survive. Considering that the distance-based method assesses the number of neighbors, it makes sense to flag these instances as outliers.

6.0 RapidMiner Process



7 Conclusion

This report utilizes CRISP-DM (Cross-Industry Standard Process for Data Mining) to conduct supervised learning using Decision Trees to predict the likelihood of survival among RMS Titanic passengers. Additionally, it employs unsupervised learning techniques including K-Means clustering and outlier detection using the LOF (Local Outlier Factor) and Distance-based methods.

Before modeling, the Titanic dataset underwent cleaning to address missing values and to create new nominal columns aimed at enhancing the performance of the Decision Tree algorithm. Upon modeling the data with Decision Trees, it becomes evident that female passengers from the first and second classes had a higher chance of survival, whereas third-class females were less fortunate, particularly if they had numerous family members. Furthermore, most young males from the first and second classes had a better chance of survival. Numeric attributes from the Titanic dataset were also utilized for clustering and outlier detection. K-Means clustering grouped similar instances into clusters, facilitating the identification of unusual instances. Notably, one cluster contained instances with identical ticket numbers, which could be further investigated for anomalies.