Assignment 2

Hadiyah Khan

Computer Engineering Technology -Computing Science

041049366

February 25, 2024

CST_8380 BI and Data Analytics

Contents

1.0 Introduction

We will be exploring the titanic provided in the samples section of RapidMiner.

2.0 Business Understanding

RMS Titanic was the largest ship afloat in its time and one of three Olympic-class Ocean Liners operated by the Water Star Line. During its maiden voyage from Southampton to New York City, on the fateful day of 15th of April 1912, the Ocean Liner sank into the North Atlantic Ocean during the early morning after colliding with an iceberg, killing more than 1,500 of its estimated 2,224 passengers and crew on board.

It was one of the deadliest commercial peacetime maritime disasters in modern history. (Douges, 2018)

3.0 Data Understanding

Use the data set called Titanic.

3.1 Collect Initial Data

The data set used in this report will be the Titanic data set provided in the samples section of Rapid Miner.

3.2 Descript Data

Attributes:

Attribute	Descriptions
Survival	Survival of the people aboard
Id	A unique id number to identify each passenger
Passenger Class	The ticket class of the passenger; there were several classes for the passengers on board the Titanic
Sex:	The sex of the passengers; male or female
Age:	Age measured in years
No of Siblings or Spouses on Board.	Number of siblings /spouses aboard the Titanic
No of Parents or Children on Board	Number of parents /children aboard the Titanic
Ticket Number	The ticket number of the passengers.
Passenger Fare	The fare of the ticket for the passengers
Cabin	The cabin number
Port of Embarkation	The name of the port which the passengers embarked from
Lifeboat	The identifying label of the lifeboat

3.3 Explore Data

Explore the data entries in Titanic data set.

In exploring data I have found that the more wealthy passengers had a higher chance of survival as seen in these graphs:

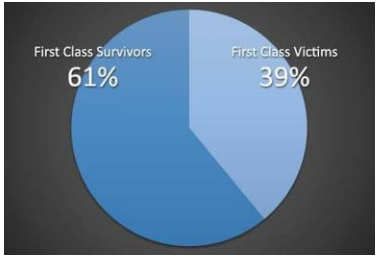


Figure 1 first class passengers.

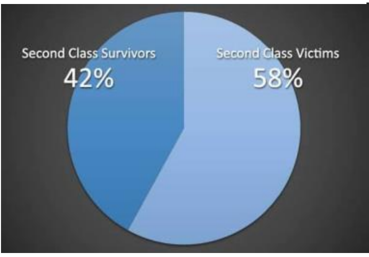


Figure 2 second class passengers.

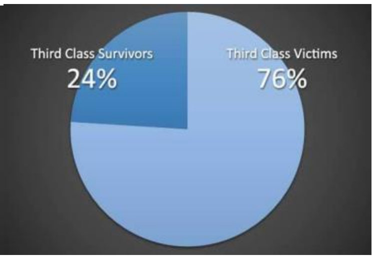


Figure 3 third class passengers.

This means that passenger class is important in this data set. (Titanic Facts, n.d.)

3.4 Verify Data Quality

There are a few missing data points:

- Port of embarkation has 2 missing values.
- Age has 263 missing values.
- Cabin has 1014 missing values.
- Lifeboat has 823 missing values.

4.0 Data Preparation

In this step, we start preparing the data to perform classification using Decision trees, clustering using KMeans and outlier detection using LOF and distance approaches.

4.1 Select Data

Load the sample Titanic data set into Rapid Miner.

4.2 Clean Data

Removed empty data using operator ‘Replace Missing Data

4.3 Construct Data

We generate attributes Age Group and Relative.

4.4 Integrate Data

I chose frequency binning because its better for the large amounts of missing data we have in the Titanic data set.

4.5 Format Data

For clustering, I select some attributes and turn sex and survival form nominal to binary.

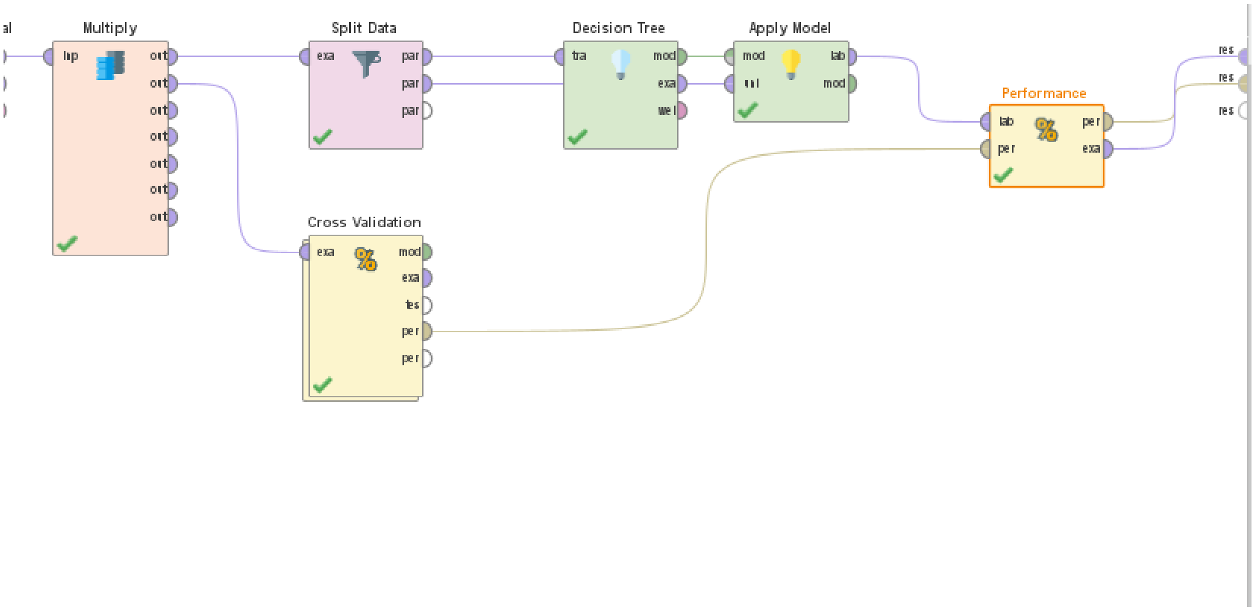
5.0 Modeling

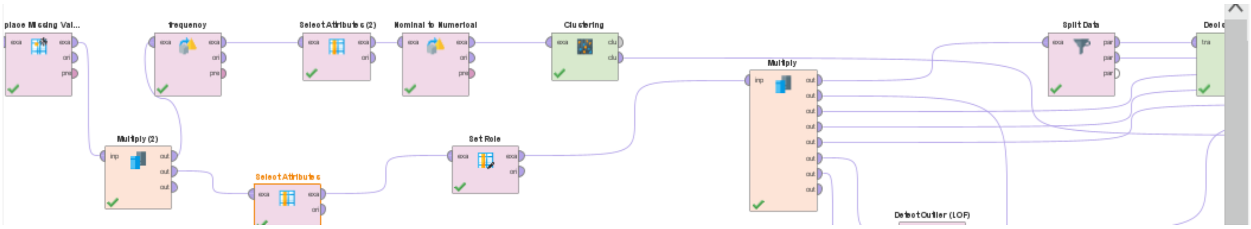
Now we start moddeling

We have to do decision trees, kmeans and outlier detection using LOF and distance approaches.

5.1 Decision Tree

For decision tree





What I did was: I put label as survival: the attributes of the tree are:

age

No of siblings or spouses on board

No of parents or children on board

Passenger class

Sex

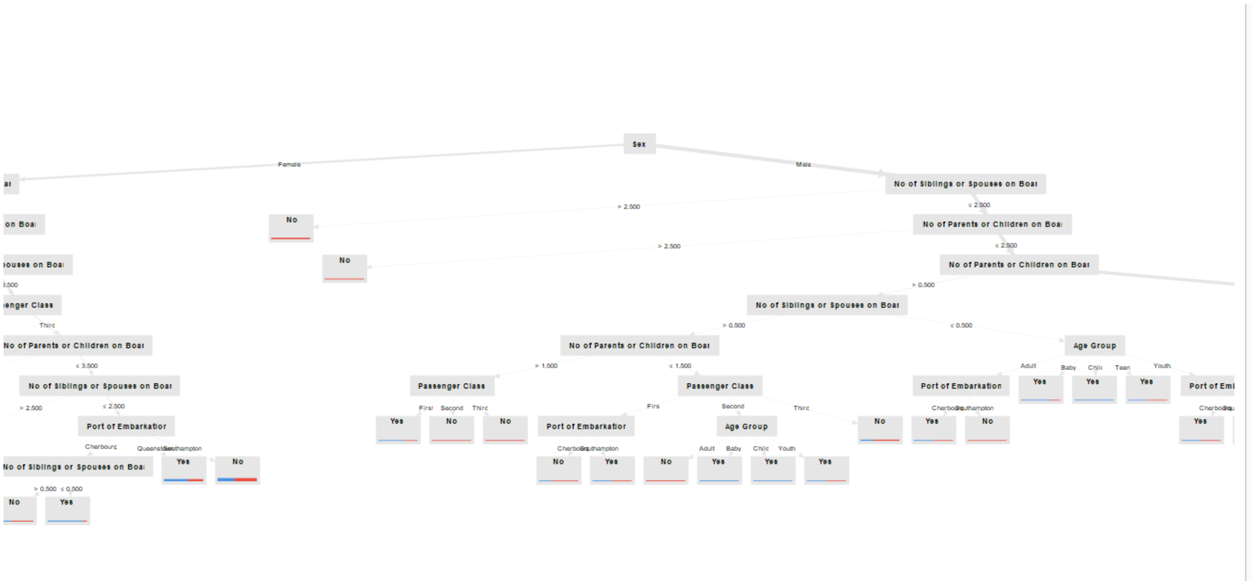
Survived

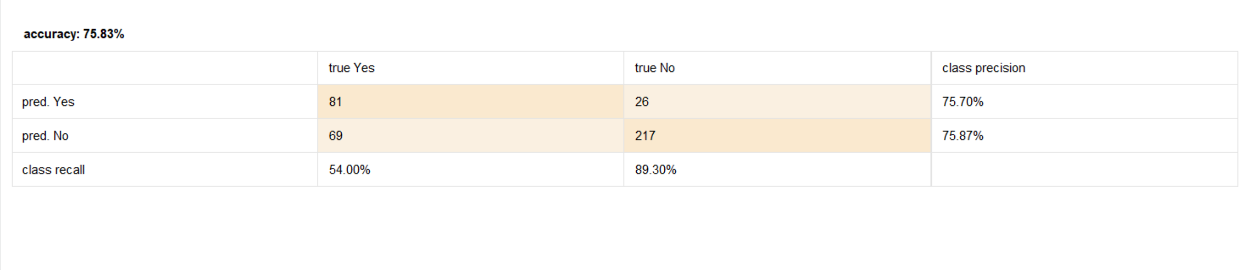
Age group

Port of embarkation: because the probability of survival is higher for both sexes depending on the port they embark from

I chose these attributes because

1) the gender mattered for survival as women and children were the fist to leave
2) being a parent increased odds of survival
3) the richer a person was increased the odds of survival and the passengers in class 1 had a higher chance of surviving. You can really see this in the decision tree below.

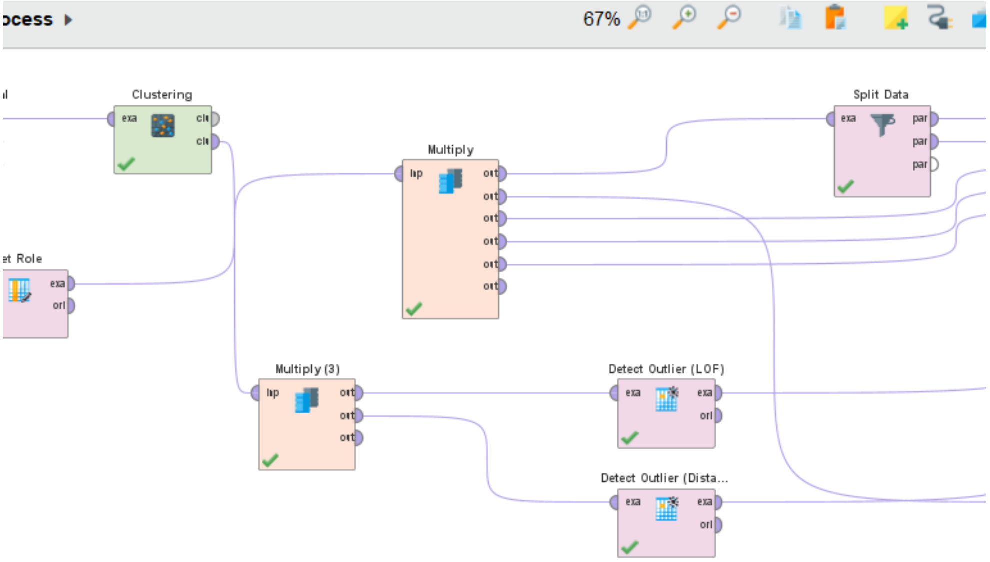




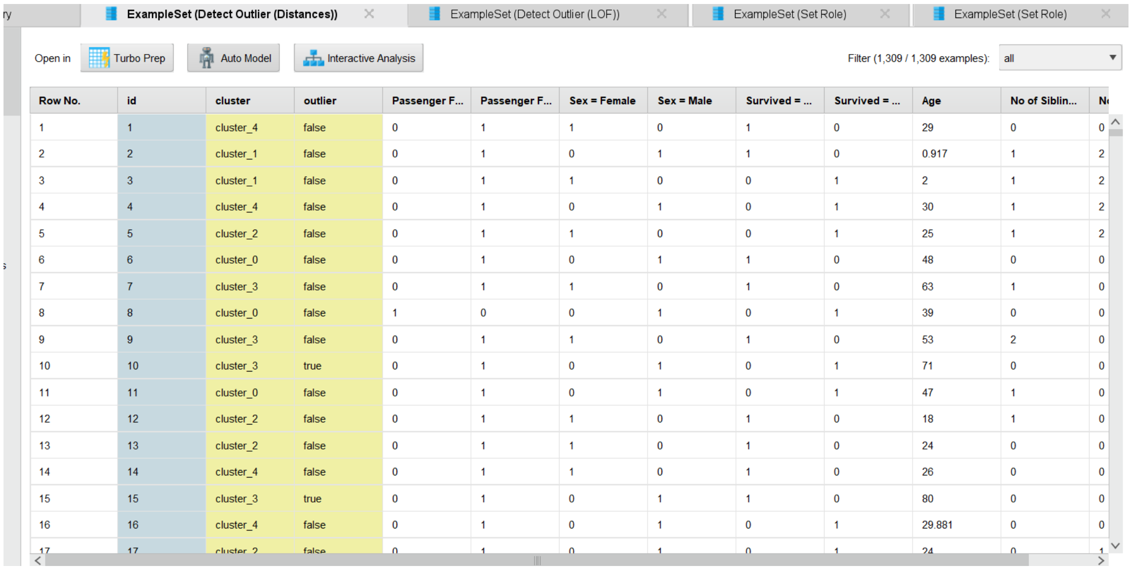
accuracy: 75.83%

	true Yes	true No	class precision
pred. Yes	81	26	75.70%
pred. No	69	217	75.87%
class recall	54.00%	89.30%	

5.2 LOF and distance

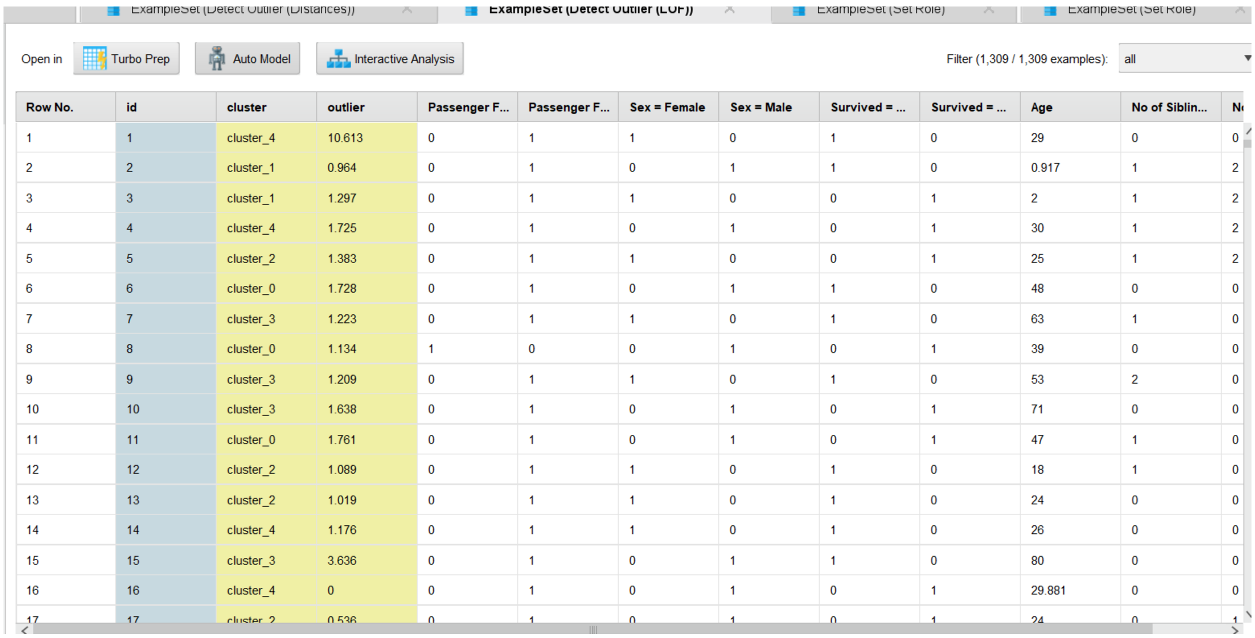


Distance:


ExampleSet (Detect Outlier (Distances))
ExampleSet (Detect Outlier (LOF))
ExampleSet (Set Role)
ExampleSet (Set Role)
Open in
Turbo Prep
Auto Model
Interactive Analysis
Filter (1,309 / 1,309 examples):
all

Row No.	id	cluster	outlier	Passenger F...	Passenger F...	Sex = Female	Sex = Male	Survived = ...	Survived = ...	Age	No of Siblin...	N
1	1	cluster_4	false	0	1	1	0	1	0	29	0	0
2	2	cluster_1	false	0	1	0	1	1	0	0.917	1	2
3	3	cluster_1	false	0	1	1	0	0	1	2	1	2
4	4	cluster_4	false	0	1	0	1	0	1	30	1	2
5	5	cluster_2	false	0	1	1	0	0	1	25	1	2
6	6	cluster_0	false	0	1	0	1	1	0	48	0	0
7	7	cluster_3	false	0	1	1	0	1	0	63	1	0
8	8	cluster_0	false	1	0	0	1	0	1	39	0	0
9	9	cluster_3	false	0	1	1	0	1	0	53	2	0
10	10	cluster_3	true	0	1	0	1	0	1	71	0	0
11	11	cluster_0	false	0	1	0	1	0	1	47	1	0
12	12	cluster_2	false	0	1	1	0	1	0	18	1	0
13	13	cluster_2	false	0	1	1	0	1	0	24	0	0
14	14	cluster_4	false	0	1	1	0	1	0	26	0	0
15	15	cluster_3	true	0	1	0	1	1	0	80	0	0
16	16	cluster_4	false	0	1	0	1	0	1	29.881	0	0
17	17	cluster_2	false	0	1	0	1	0	1	24	0	1

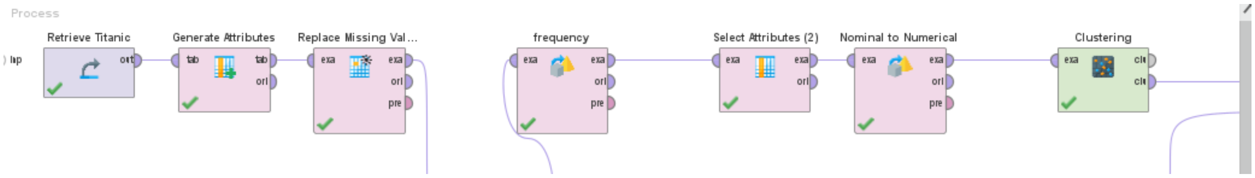
LOF:

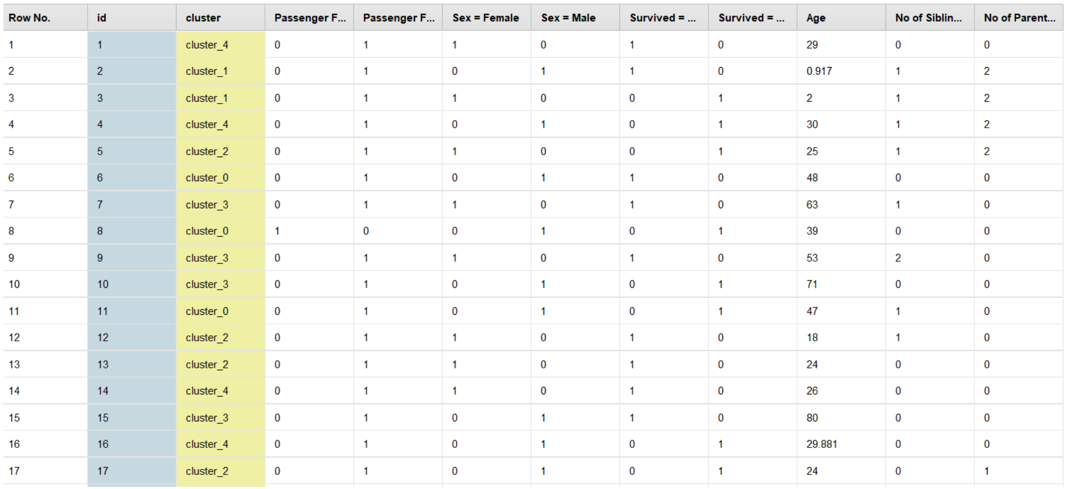


Filter (1,309 / 1,309 examples): all

Row No.	id	cluster	outlier	Passenger F...	Passenger F...	Sex = Female	Sex = Male	Survived = ...	Survived = ...	Age	No of Siblin...	N
1	1	cluster_4	10.613	0	1	1	0	1	0	29	0	0
2	2	cluster_1	0.964	0	1	0	1	1	0	0.917	1	2
3	3	cluster_1	1.297	0	1	1	0	0	1	2	1	2
4	4	cluster_4	1.725	0	1	0	1	0	1	30	1	2
5	5	cluster_2	1.383	0	1	1	0	0	1	25	1	2
6	6	cluster_0	1.728	0	1	0	1	1	0	48	0	0
7	7	cluster_3	1.223	0	1	1	0	1	0	63	1	0
8	8	cluster_0	1.134	1	0	0	1	0	1	39	0	0
9	9	cluster_3	1.209	0	1	1	0	1	0	53	2	0
10	10	cluster_3	1.638	0	1	0	1	0	1	71	0	0
11	11	cluster_0	1.761	0	1	0	1	0	1	47	1	0
12	12	cluster_2	1.089	0	1	1	0	1	0	18	1	0
13	13	cluster_2	1.019	0	1	1	0	1	0	24	0	0
14	14	cluster_4	1.176	0	1	1	0	1	0	26	0	0
15	15	cluster_3	3.636	0	1	0	1	1	0	80	0	0
16	16	cluster_4	0	0	1	0	1	0	1	29.881	0	0
17	17	cluster_2	0.536	0	1	0	1	0	1	24	0	1

5.3 Clustering





Row No.	id	cluster	Passenger F...	Passenger F...	Sex = Female	Sex = Male	Survived = ...	Survived = ...	Age	No of Siblin...	No of Parent...
1	1	cluster_4	0	1	1	0	1	0	29	0	0
2	2	cluster_1	0	1	0	1	1	0	0.917	1	2
3	3	cluster_1	0	1	1	0	0	1	2	1	2
4	4	cluster_4	0	1	0	1	0	1	30	1	2
5	5	cluster_2	0	1	1	0	0	1	25	1	2
6	6	cluster_0	0	1	0	1	1	0	48	0	0
7	7	cluster_3	0	1	1	0	1	0	63	1	0
8	8	cluster_0	1	0	0	1	0	1	39	0	0
9	9	cluster_3	0	1	1	0	1	0	53	2	0
10	10	cluster_3	0	1	0	1	0	1	71	0	0
11	11	cluster_0	0	1	0	1	0	1	47	1	0
12	12	cluster_2	0	1	1	0	1	0	18	1	0
13	13	cluster_2	0	1	1	0	1	0	24	0	0
14	14	cluster_4	0	1	1	0	1	0	26	0	0
15	15	cluster_3	0	1	0	1	1	0	80	0	0
16	16	cluster_4	0	1	0	1	0	1	29.881	0	0
17	17	cluster_2	0	1	0	1	0	1	24	0	1

Clustering was done with k = 5

Clustering was done on numeric attributes, and some had to turned from nominal to numeric like sex and survival.

- The attributes used in clustering and therefore in the outlier detection was:
- Age
- No. of parents or children on Board
- No. of siblings or spouses on board.
- Passenger class
- Passenger fare
- Sex
- Survived
- Port of embarkation

6.0 Conclusion

In conclusion, we have preformed clustering through KMeans, decision tree and outliers with LOF and distance in rapid miner to visualize the survival rate of the passengers aboard the titanic.

References

Douges, N. (2018, May 14). *MediumPredicting the Survival of Titanic Passengers*. Retrieved from Medium: https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8

Kaggle. (n.d.). *Titanic - Machine Learning from Disaster*. Retrieved from Kaggle: https://www.kaggle.com/c/titanic

stanford.edu. (n.d.). *A Titanic Probability*. Retrieved from web.stanford.edu: https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html

Titanic Facts. (n.d.). *Titanic Survivors*. Retrieved from TItanic Facts: https://titanicfacts.net/titanic-survivors/