CST8390_012 - Assignment 1

Titanic Dataset Analysis Report using kNN and Decision Trees

Author of the overall report and Workload

Shu Han Han #041-060-762	Business Understanding/Data Understanding/ Data Preparation/Modeling/Evaluation
Wan-Hsuan Lee #041-060-761	Introduction/Business Understanding/ Data Understanding/Discussion of Result/Conclusion

Computer Programming, Algonquin College

June 2, 2023

Table of Contents

Introduction

The sinking of the RMS Titanic on April 15, 1912, remains one of the most infamous maritime disasters in history. The tragic incident resulted in the loss of over 1,500 lives and sparked global interest in maritime safety and disaster response. The Titanic dataset, which contains information about the passengers onboard the ill-fated ship, provides a valuable resource for analyzing and understanding the factors that influenced survival outcomes. This report aims to explore and analyze the Titanic dataset using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology.

The goal of this analysis is to perform classification using two different machine learning algorithms: k-Nearest Neighbors (kNN) and Decision Trees. By leveraging these algorithms, we can predict the likelihood of survival for individual passengers based on their attributes. The analysis will involve several steps, including business understanding, data understanding, data preparation, modeling, evaluation, and a discussion of the results.

Through this analysis, we seek to gain insights into the factors that played a significant role in determining the survival outcomes of Titanic passengers. By utilizing the CRISP-DM methodology, we will follow a structured approach to understand the data, prepare it for analysis, build and evaluate models, and derive meaningful conclusions. The findings from this analysis can contribute to our understanding of the Titanic incident and provide valuable insights into the factors that influenced survival, potentially shedding light on broader patterns and principles related to maritime safety and disaster response.

In the following sections, we will delve into the details of each step in the analysis, including the data understanding, data preparation, modeling, evaluation, and discussion of results. By following the CRISP-DM methodology, we aim to provide a comprehensive and systematic exploration of the Titanic dataset, ultimately contributing to our understanding of this historic event and showcasing the potential of machine learning in analyzing complex datasets.

Business Understanding

1. Determine Business Objectives

This project is a research assignment, hence, no business perspective.

2. Assess Situation

Availability of Resources:

- Dataset Files:
  - Titanic_train.csv: The training set, used to build the machine learning models.
  - Titanic_test.csv: The test set, used to see how well the model performs on unseen data.
- Machine Learning Algorithm:
  - K-Nearest Neighbors (KNN)
  - Decision Tree
- Applicable Software:
  - Weka
- References:
  - http://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html
  - https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8
  - https://www.kaggle.com/c/titanic
  - http://csis.pace.edu/~ctappert/srd2014/d3.pdf
  - https://titanicfacts.net/titanic-survivors/

Assess Risks:

- Not able to finish in time

Contingency Plans for Risks:

- Put more effort into it

3. Description of the Incident – Titanic

The RMS (Royal Mail Ship) Titanic was a British passenger liner, the second of three Olympic-class ocean liners operated by the White Star Line. It was the largest ship afloat at the time it entered service in 1912.

On April 15, 1912, in the early morning, it embarked on its maiden voyage from Southampton, UK, subsequently stopped at the ports of Cherbourg, France, and Queenstown (now Cobh), Ireland, across the North Atlantic, headed to New York City, USA.

During its voyage, it collided with an iceberg, leading to the loss of 1,502 lives out of the total 2,224 passengers and crew members on board. The shortage of lifeboats on board was the prevailing belief that the ship was unsinkable at the time, resulting in insufficient accommodation for evacuation when the ship sank.

4. Determine Goals

Our goal is to develop a predictive model that can accurately forecast the likelihood of survival for different individuals in the scenario: 'Which types of people were more prone to surviving?”

5. Produce Project Pan

By following CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology and step guidelines in the assignment file (CST8390 Assignment 1), we make a work breakdown list to ensure the collaboration of teamwork.

Data Understanding

1. Collect Initial Data

- Titanic_train.csv (screenshot)

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	'Braund, Mr. Owen Harris'	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	'Cumings, Mrs. John Bradley (Florence Briggs Thayer)'	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	'Heikkinen, Miss. Laina'	female	26	0	0	STON/O2. 3101282	7.925		S
5	4	1	1	'Futrelle, Mrs. Jacques Heath (Lily May Peel)'	female	35	1	0	113803	53.1	C123	S
6	5	0	3	'Allen, Mr. William Henry'	male	35	0	0	373450	8.05		S

2. Describe Data

- Description of Data:
    - Instances: 889
    - Attributes: 12

No.	Attribute	Description	Note
1	PassengerId	Unique Id of a passenger	
2	Survived	Survival	0 = No, 1 = Yes
3	Pclass	Ticket class	1 = 1st (Upper), 2 = 2nd (Middle), 3 = 3rd (Lower)
4	Name	Name	Quoted with double quotation marks.
5	Sex	Gender	Male or female.
6	Age	Age in years	Age is fractional if less than 1. If the age is estimated, it is in the form of 'xx.5'.
7	SibSp	The number of siblings or spouses the passenger had aboard	Sibling = brother, sister, stepbrother, stepsister. Spouse = husband, wife (mistresses and fiancés were ignored).
8	Parch	The number of parents or children the passenger had aboard	Parent = mother, father. Child = daughter, son, stepdaughter, stepson. Children travelled only with a nanny has a parch=0 for them.
9	Ticket	Ticket number	
10	Fare	Passenger fare	
11	Cabin	Cabin number	
12	Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton.

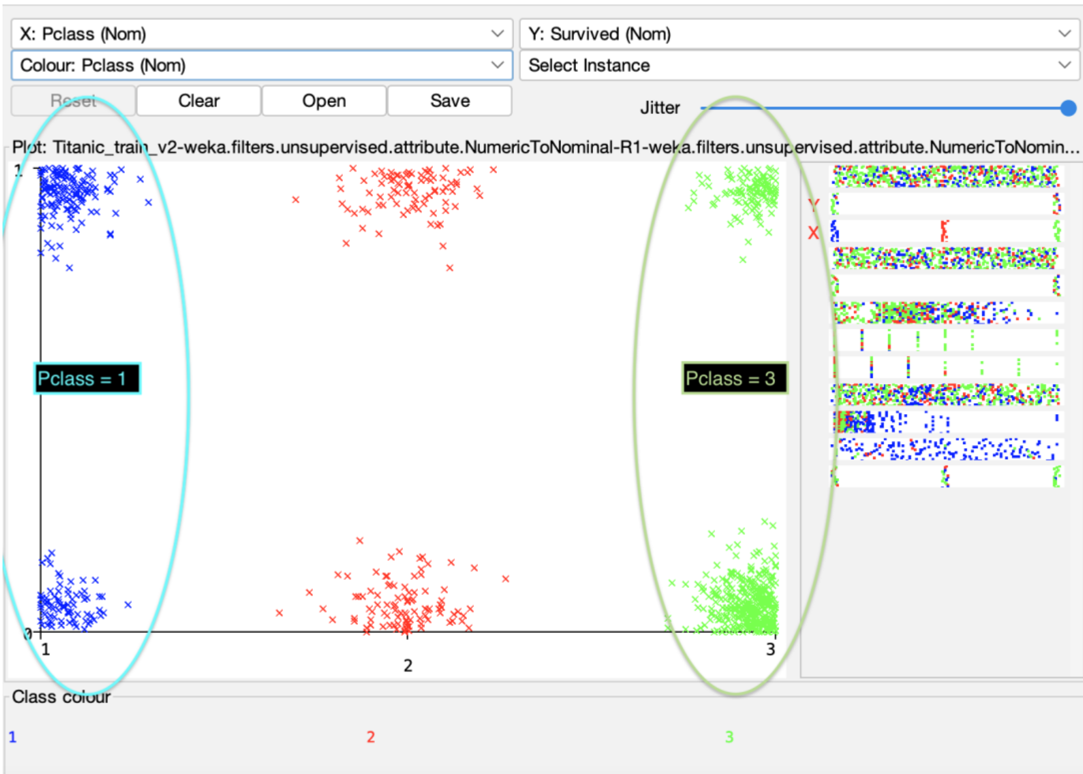
- Data Format:

Attribute	Original Format	Revised Format	Reasons
PassengerId	Numeric	String	Each passenger id represented an individual passenger onboard.
Survived	Numeric	Nominal { 0 = No, 1 = Yes }	Passengers either survived or died, there was nothing in between.
Pclass	Numeric	Nominal { 1 = 1st (Upper), 2 = 2nd (Middle), 3 = 3rd (Lower) }	Ticket class should be one of the three classes, there was nothing in between.
Name	Nominal	String	Each passenger had a name. There might be passengers with the same name.
Sex	Nominal	Nominal {male, female}	Suppose there were no genders other than male and female.
Age	Numeric	Numeric	Age can be a fraction of the year.
SibSp	Numeric	Numeric	People are countable.
Parch	Numeric	Numeric	People are countable.
Ticket	Nominal	String	Each ticket was an individual instance. There were identical tickets.
Fare	Numeric	Numeric	The money amount was countable.
Cabin	Nominal	String	Each cabin was an individual instance.
Embarked	Nominal	Nominal { C = Cherbourg, Q = Queenstown, S = Southampton }	Three were three different embarkation ports.

3. Explore Data

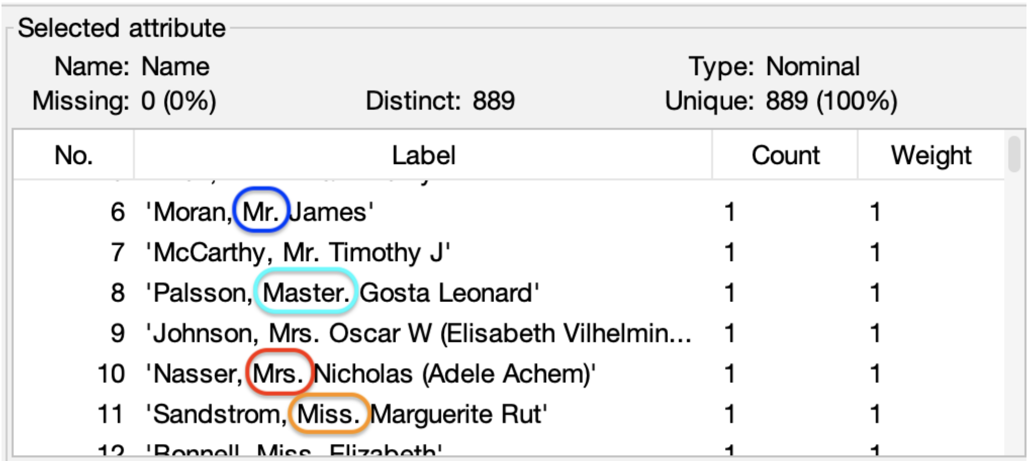
- **Y: Survived – X: Pclass relationship:**
    - Hypothesis: In reality, wealthy people generally have more resources (money, friends, and social influence). These factors make them easily receive priority or special treatment compared to the poor.

      It was possible that when the ship was sinking, during the evacuation, the passengers in the upper class received priority for boarding the limited lifeboats, which was not sufficient for everyone. Thus, they would have a better chance of survival.
    - Screenshot of the Visualized Relationship (Y: Survived – X: Pclass) in Weka:



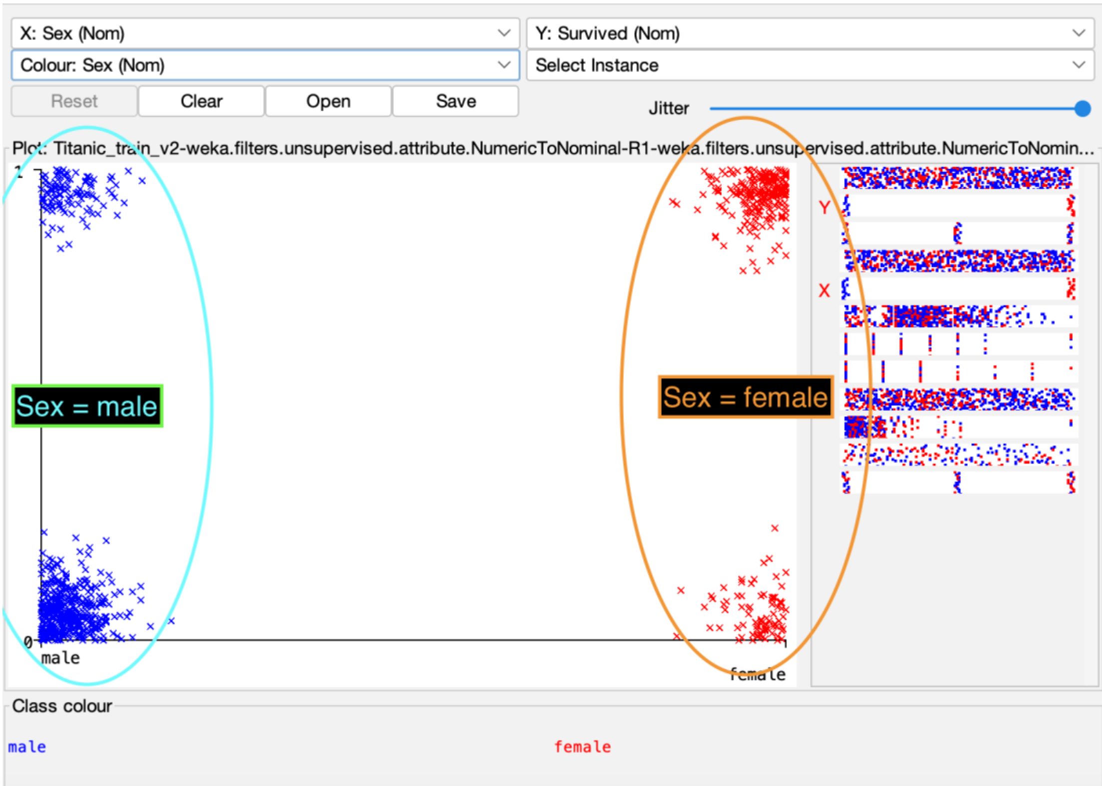
    - Observation: As we can see in the screenshot, passengers in Pclass=1 (upper class) had a higher probability of survival compared to the passengers in Pclass=3 (lower class).
    - Conclusion: Our hypothesis may be correct, and the Pclass (ticket class level) is an attribute correlated to the survival probability.

- **Explore the Name attribute:**
  - Hypothesis: In reality, a person's name cannot be a determinant of whether survived the shipwreck or not. However, the Name attribute contains "Mr., Master, Mrs., and Miss". Those are honorific titles which can refer to a passenger's gender and marital status (usually related to a person's age).
  - Screenshot of the Name attribute in Weka:



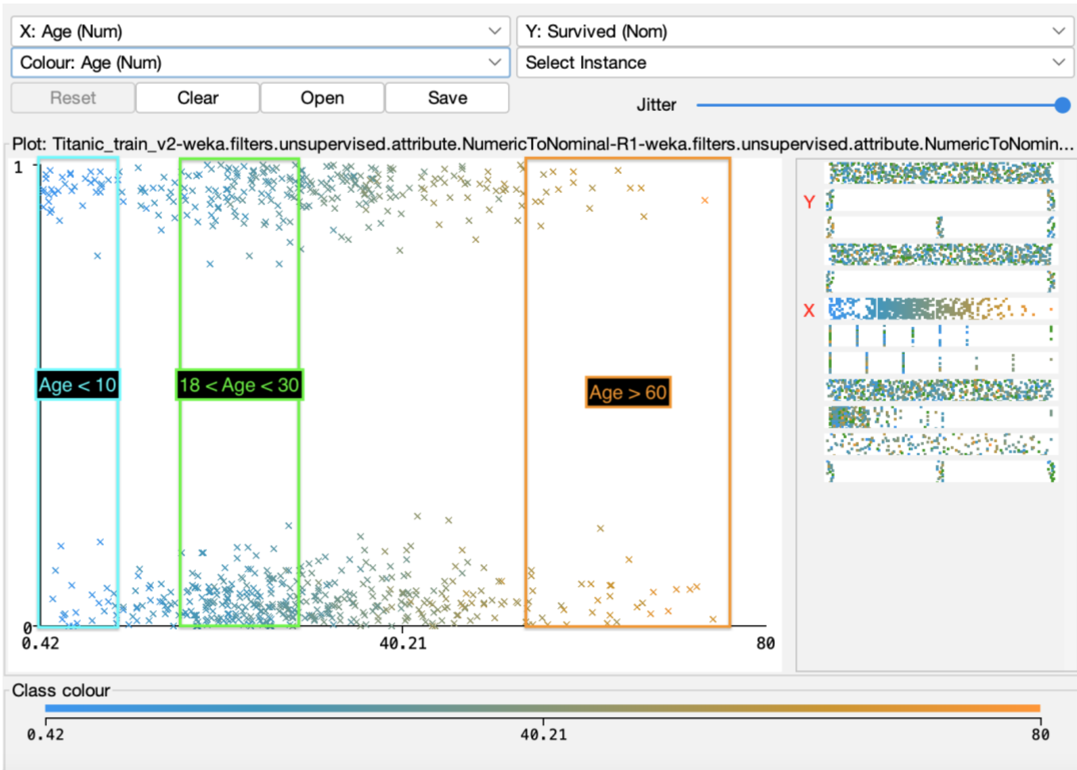
  - Observation: Almost every name contains an honorific title that can refer to a passenger's gender and age (deducted from marital status).
  - Conclusion: We already have complete gender data (Sex) but we are missing 177 (20%) of the age data (Age). Thus, we might need to analyze the relationship between the honorific title and the age of passengers whose age data is complete. And based on it, produce an age data generator, then use it to generate a reasonable value for those whose age is missing.

- **Y: Survived – X: Sex relationship:**
  - Hypothesis: In general, society usually follows the “ladies first” principle. We can expect that when the ship was sinking, the priority for evacuation was given to females, thus, they probably would have a greater chance of survival.
  - Screenshot of the Visualized Relationship (Y: Survived – X: Sex) in Weka:



  - Observation: As we can see from the instance density in the plot, male passengers had a higher probability to die whereas female passengers had a higher probability to survive.
  - Conclusion: The visualized result supports the hypothesis we made earlier.

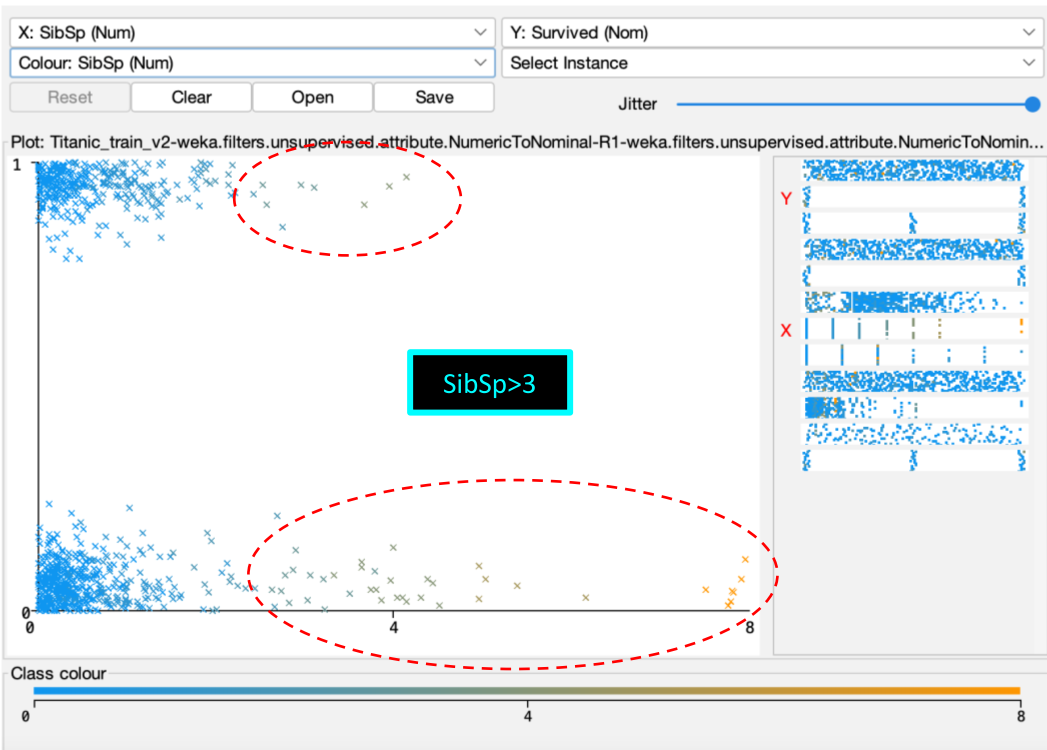
- **Y: Survived – X: Age relationship:**
  - Hypothesis: In traditional, society usually pays extra assistance to children and old people. It was possible that when the ship was sinking, those people received extra assistance or priority for evacuation on boarding the lifeboats, thus they would have a higher probability to survive.
  - Screenshot of the Visualized Relationship (Y: Survived – X: Age) in Weka:



  - Observation: As we can see from the visualized plot, children who were younger than 10 years old had a higher chance of survival, whereas adults aged between 18 and 30 had a higher probability of death.

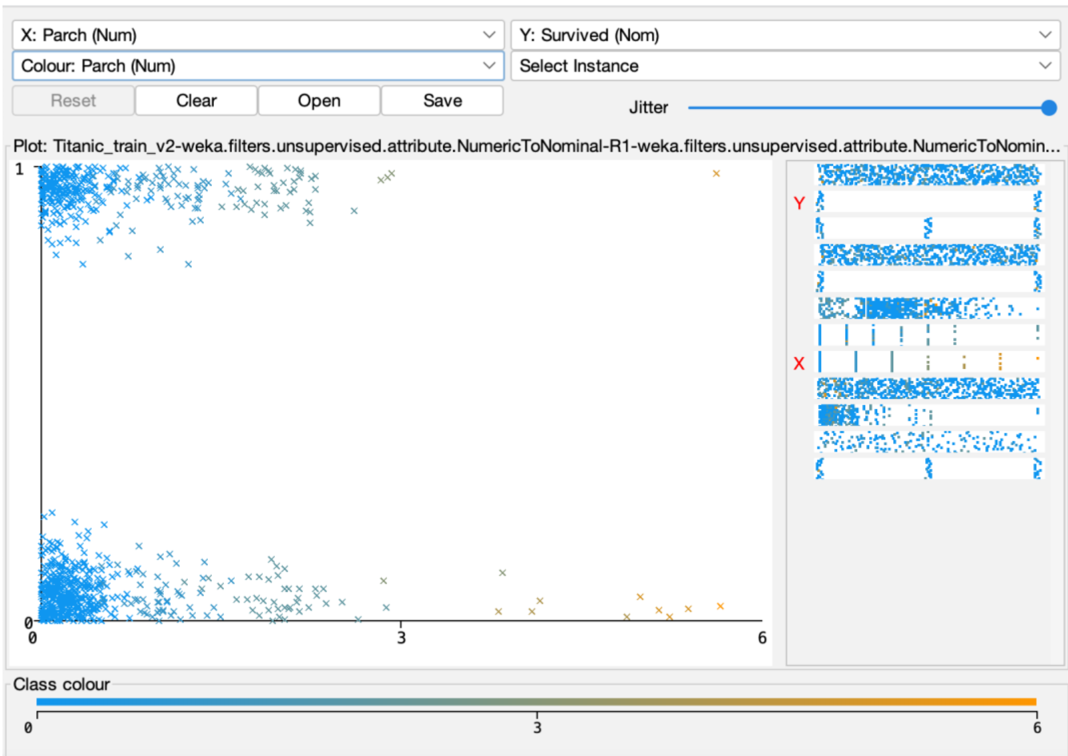
    However, there is no significant evidence showing that older people (age > 60) had a higher probability to survive.
  - Conclusion: The result may support the hypothesis that younger children received priority or extra assistance during the evacuation. However, adults who were aged between 18 and 30, were generally considered to be physically stronger, thus, they might have the least priority for evacuation, which lead to a higher probability of death.

- **Y: Survived – X: SibSp relationship:**
  - Hypothesis: In traditional, family members would stay together, especially during difficult times. It was likely that when the ship was sinking, families of siblings or couples stayed together to support each other.
    However, passengers with more siblings might face difficulty in finding sufficient seats on the same lifeboat. They might be forced to separate to evacuate to survive. Hence, we hypothesize passenger's sibling number is not a critical factor for survival.
    Although spouses usually have a stronger relationship than siblings. When the ship was sinking, a husband might have to let his wife board the lifeboat, with himself staying on the ship because of the insufficient lifeboats, and the "ladies and children first" principle. It was likely that a wife survived without her husband, which makes the spouse number not a critical factor in survival.
  - Screenshot of the Visualized Relationship (Y: Survived – X: SibSp) in Weka:



  - Observation: As we can see from the visualized plot, passengers with more than 3 siblings/spouse had a less chance of survival showing that the passenger's sibling or spouse number affected their chance of survival.
  - Conclusion: The SibSp attribute potentially be a factor that influences the survival of passengers.

- **Y: Survived – X: Parch relationship:**
    - Hypothesis: In traditional, people follow the "children and vulnerable people first" principle. It was likely that parents with younger children had the priority to board the lifeboats when the ship was sinking, which gave them a better chance of survival. However, the Parch attribute does not distinguish the children's age, i.e., an adult with older parents onboard is also in the group, which in this case does not contribute them a higher probability of survival. Thus, we hypothesize that the Parch attribute is not a critical factor in survival.
    - Screenshot of the Visualized Relationship (Y: Survived – X: Parch) in Weka:



    - Observation: As we can see from the visualized plot, there is no special trend showing that the passenger's parent or child number affected their chance of survival.
    - Conclusion: The Parch attribute is not a main factor in passenger survival.

- **Explore the Ticket attribute:**
  - Hypothesis: Traditionally, the ticket id can be categorized according to its class level. We may be able to categorize them by analyzing their number patterns with class level (Pclass) and price (Fare), but we can choose to use Pclass with Fare attribute instead.
  - Screenshot of the Ticket attribute in Excel:

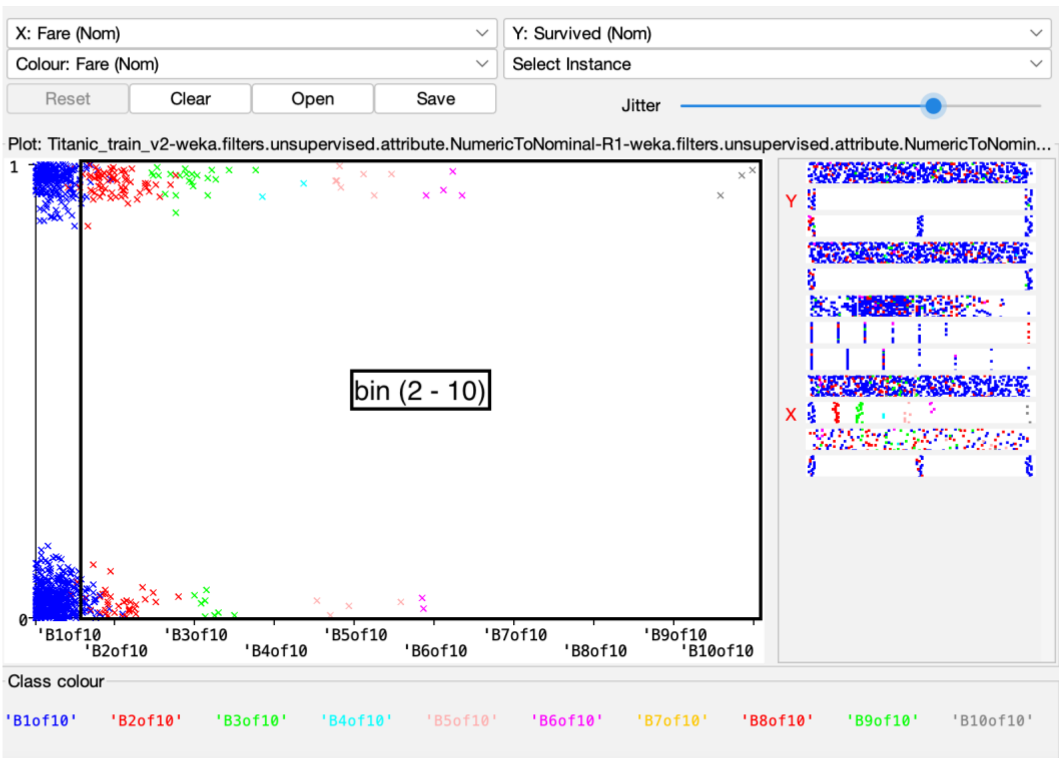
	A	B
1	Pclass	Ticket
2	3	A/5 21171
3	1	PC 17599
4	3	STON/O2. 3101282
5	1	113803
6	3	373450
7	3	330877
8	1	17463
9	3	349909
10	3	347742
11	2	237736
12	3	PP 9549
13	1	113783
14	3	A/5. 2151
15	3	347082
16	3	350406
17	2	248706
18	3	382652
19	2	244373
20	3	345763
21	3	2649

  - Observation: By comparing the ticket id with the class level (Pclass), we can conclude that, only for tickets with 6 digits, its first digit corresponds to the passenger's Pclass number. However, this information can be substituted by the Pclass attribute.

    There is no other trait that we found in this attribute, to allow us to categorize it into beneficial groups for machine learning.

    Conclusion: Although the Ticket attribute has a relationship with the Pclass, there are no further traits we could find in this attribute to allow us to categorize it into beneficial groups for machine learning.

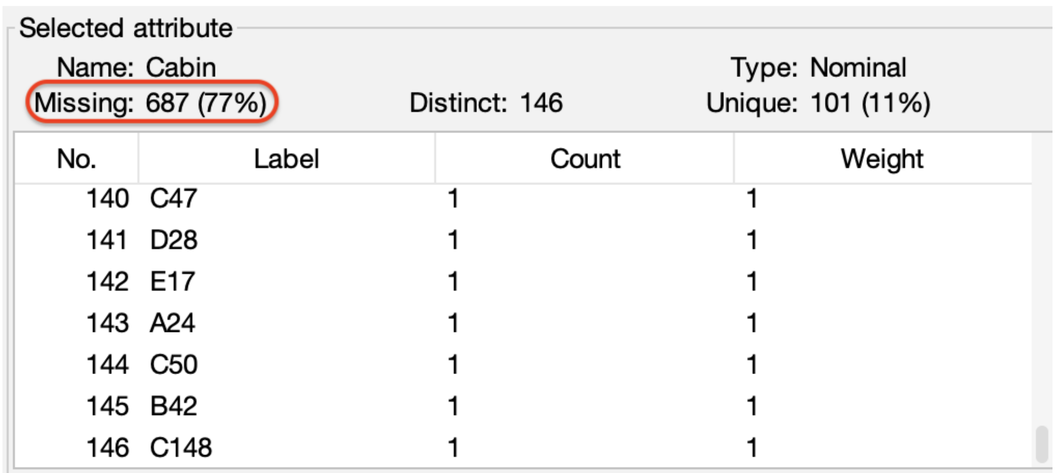
- **Y: Survived – X: Fare relationship:**
    - Hypothesis: In reality, wealthy people generally have more resources (money, friends, and social influence). These factors make them easily receive priority or special treatment compared to the poor.
      It was possible that when the ship was sinking, during the evacuation, the passengers who had paid a higher fare, received priority for boarding the limited lifeboats, which was not sufficient for everyone. Thus, they would have a better chance of survival.
    - Screenshot of the Visualized Relationship (Y: Survived – X: Fare, 10 equal-width bins) in Weka:



    - Observation: As we can see from the plot, passengers with a fare greater than 51.233 (bin-2 to bin-10 out of 10 bins) had a higher chance of survival. However, there is no trend displaying that as the fare increases the passenger had a higher probability to survive.
    - Conclusion: The Fare attribute is a factor in passenger survival, but it is not a main factor.

- **Explore the Cabin attribute:**
  - Hypothesis: In traditional, cabin id may be related to class level and the cabin location onboard, which determines the distance to its nearest lifeboats. As we can suppose, the closer the passenger was to the lifeboat, the higher possibility that he/she boarded a lifeboat. Thus, the Cabin attribute might be a factor in passenger survival.

    However, there is a high rate of missing data in this attribute, literally, 687 (77%). This can introduce bias and impact the overall performance of the model if we use it in machine learning.
  - Screenshot of the Cabin attribute in Weka:



Selected attribute

Name: Cabin | Type: Nominal

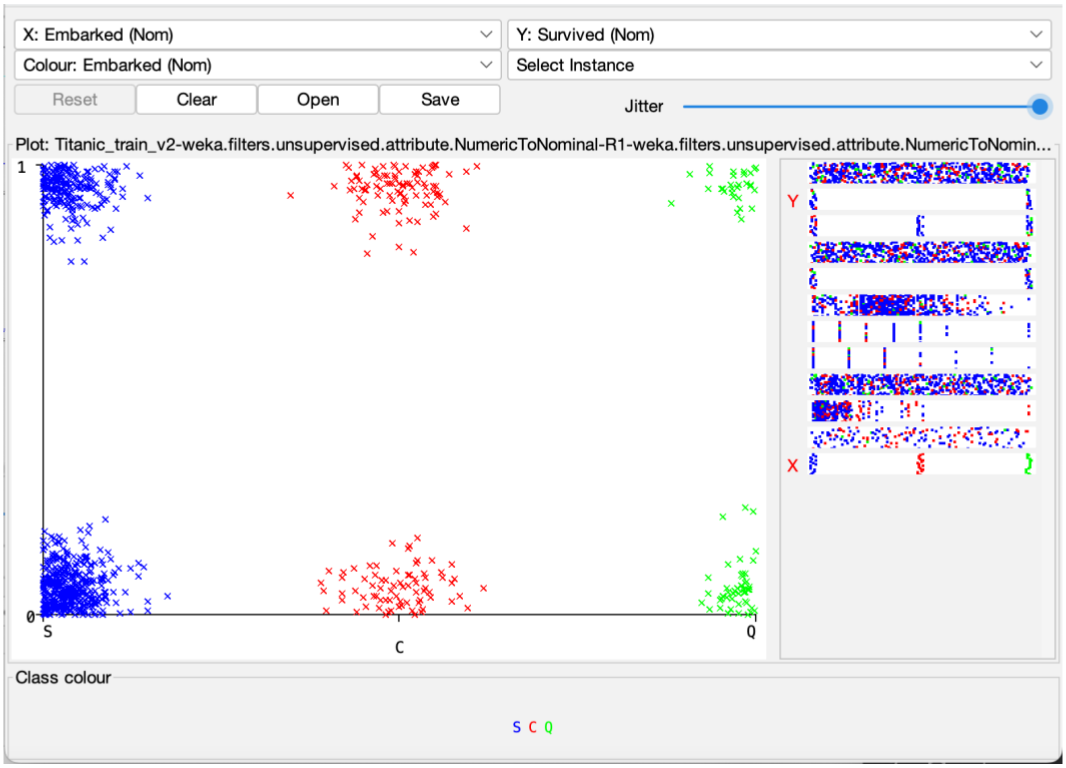
Missing: 687 (77%) | Distinct: 146 | Unique: 101 (11%)

No.	Label	Count	Weight
140	C47	1	1
141	D28	1	1
142	E17	1	1
143	A24	1	1
144	C50	1	1
145	B42	1	1
146	C148	1	1

  - Observation: As we can see in the screenshot, the attribute contains 687 (77%) missing data.
  - Conclusion: Even though the Cabin attribute might be a factor in passenger survival, we will probably drop it in the future due to its high missing data rate.

- **Y: Survived – X: Embarked relationship:**
  - Hypothesis: Passengers who boarded in different geographical places, might have differences in wealth, gender ratio, children ratio, customs or religion (which biases decision-making). These all are the factors in passenger survival as we discovered in previous sections.

    Thus, the Embarked attribute might be a factor in passenger survival.
  - Screenshot of the Visualized Relationship (Y: Survived – X: Embarked) in Weka:



  - Observation: As we can see from the visualized plot, there is no special trend showing that the port of embarkation affected their chance of survival.
  - Conclusion: The Embarked attribute is not a main factor in passenger survival.

4. Verify Data Quality

- Missing Data
  - Age: 177 (20%)
  - Cabin: 687 (77%)
- Error Data

Attribute	PassengerId	Cause	Fixed
Name	29, 102, 147, 148, 149, 157, 162, 166, 187, 188, 199, 200, 205, 209, 228, 238, 242, 275, 278, 291, 301, 305, 346, 349, 360, 382, 428, 437, 482, 490, 508, 519, 554, 557, 573, 600, 605, 625, 654, 655, 698, 706, 707, 710, 711, 718, 721, 743, 791, 863, 875, 888.	All of them contain extra double quotation marks inside the name quotation, which causes each instance's attributes separated improperly when loading them into Weka.	Delete them.

Data Preparation

1. Select Data

Attribute	Included / Excluded	Reasons
PassengerId	Excluded	It is the passenger sequence number. Each of them is unique but without analytical meaning.
Survived	**Included**	It is the class attribute which can be used to categorize each instance's survival for machine learning.
Pclass	**Included**	It is a critical factor in passenger survival because wealthier people who could afford a higher ticket price usually had more resources (money, friends, social influence, extra treatment or assistance). These factors might give them priority or extra assistance to evacuate to the insufficient lifeboats over others. Another reason is, usually cabins with a higher class level are closer to the lifeboats, which might allow its residents to evacuate onto the insufficient lifeboats earlier than others.
Name	Excluded	The attribute has 889 distinct values which is as much as all the instances. We cannot use it as a determinant to build a predictive model unless we categorize it into relevant groups. Its implicit information of honorific titles can be substituted with Sex and Age attributes.
Sex	**Included**	Gender is a critical factor in passenger survival because people usually follow the "ladies first" protocol. Thus, they had a higher chance to board insufficient lifeboats to survive during evacuation.
Age	**Included**	Age is a critical factor in passenger survival for young children because people usually follow the "ladies and children first" protocol. Thus, they had a higher chance to board insufficient lifeboats to survive during evacuation. While this left adult males with a higher probability of death because of insufficient lifeboats.
SibSp	**Included**	There is trend showing that p ssengers' siblings and spouse number might play a factor in their survival. Refer to "Y: Survived – X: SibSp relationship" in "Explore Data" step of "Data Understanding".
Parch	Excluded	There is no trend showing that passenger's parent and children number plays a critical factor in their survival. Refer to "Y: Survived – X: Parch relationship" in "Explore Data" step of "Data Understanding".
Ticket	Excluded	By comparing the ticket id with the class level (Pclass), we can conclude that, only for tickets with 6 digits, its first digit

		corresponds to the passenger's Pclass number. However, this information can be substituted by the Pclass attribute. There is no other trait that we found in this attribute, to allow us to categorize it into beneficial groups for machine learning.
Fare	**Included**	It is a determinant factor in passenger survival. If we discrete it into 10 equal-width bins and visualized the data in a plot, there is a trend showing that, for bins between 2 and 10, passengers had a higher probability to survive than those in bin 1. Refer to "Y: Survived – X: Fare relationship" in "Explore Data" step of "Data Understanding".
Cabin	Excluded	This attribute has a very high rate of missing data, 680 (77%). We will not include it into our machine learning model, because it can introduce bias and impact the overall performance of the model.
Embarked	Excluded	There is no special trend showing that the port of embarkation affected their chance of survival.

2. Clean Data

- **KNN**

Attribute	Number of Missing Instances	Percentage of Missing Instances	Action	Reasons
Age	177	20%	Drop	The kNN algorithm relies on measuring distances between instances to determine the nearest neighbours. Missing values can result in inaccurate or inconsistent distance measurements. Dropping them avoids potential distortions in the distance calculations. Besides, any false or falsely generated data would bias the model. We only want accurate data based on the truth to develop the predictive model.

- **Decision Tree**

Attribute	Number of Missing Instances	Percentage of Missing Instances	Action	Reasons
Age	177	20%	Label as missing.	By labelling them as missing values, we can categorize them into separate

				categories. Thus, retain all the information data, and contribute to the decision tree training process.

3. Construct Data

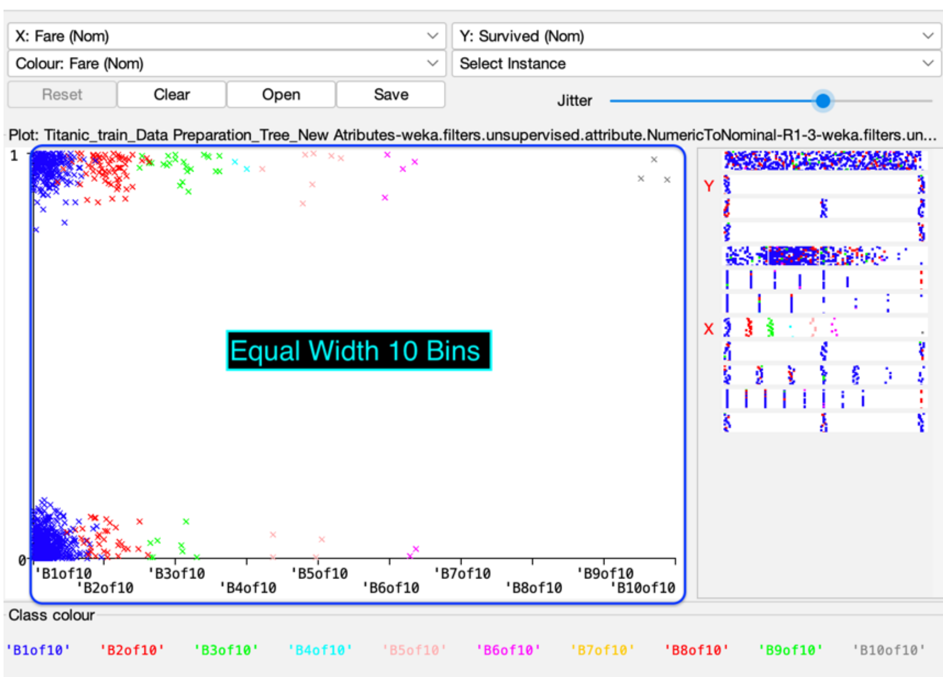
- **Decision Tree**
  - New Age Group Attribute:

Range	Label
Unknown	NK
$Age < 2$	Baby
$2 \leq Age < 12$	Child
$12 \leq Age < 18$	Teen
$18 \leq Age < 30$	Youth
$30 \leq Age \leq 65$	Adult
$Age > 65$	Senior

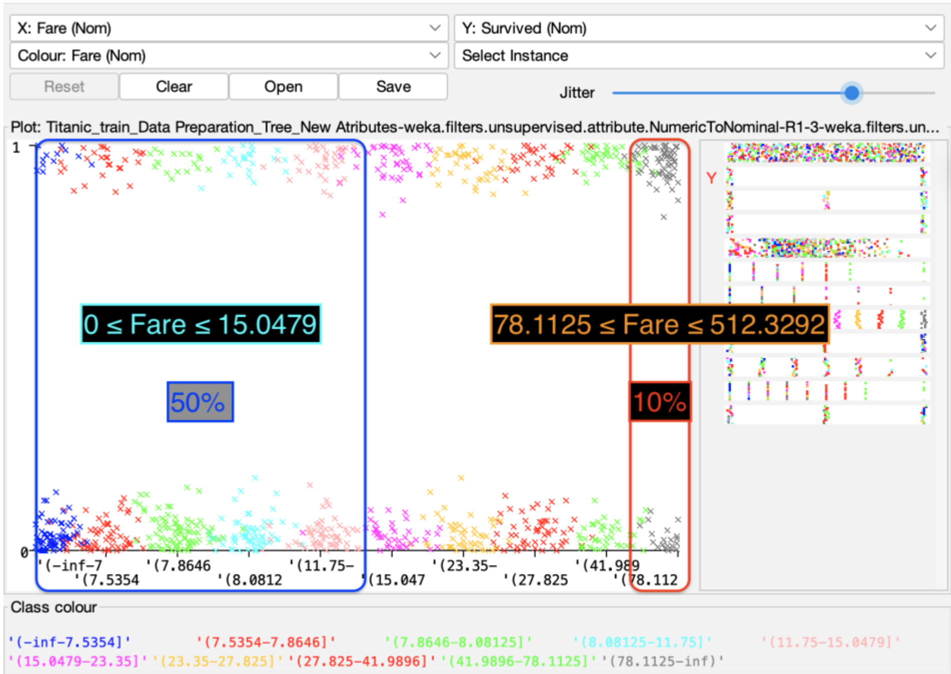
  - New Relatives Attribute:

Total Number of Relatives (siblings/spouse/parents/children)	Label
Relatives $= 0$	None
$0 <$ Relatives < 3	Few
Relatives ≥ 3	Many

  - Discretize Fare Attribute into 10 Bins:
    - Equal-width:



- Equal-frequency:



- Use Equal-width over Equal-frequency: The equal-width bins effectively discretize the Fare into 10 equal ranges. So, we can observe the passenger survival probability in different ranges by Fare ascending order.

○ Screenshot of Titanic_train_DT.csv:

	A	B	C	D	E	F
1	Survived	Pclass	Sex	Fare	'Age Group'	Relatives
2	0	3	male	'\'B1of10\''	Youth	Few
3	1	1	female	'\'B2of10\''	Adult	Few
4	1	3	female	'\'B1of10\''	Youth	None
5	1	1	female	'\'B2of10\''	Adult	Few
6	0	3	male	'\'B1of10\''	Adult	None
7	0	3	male	'\'B1of10\''	NK	None
8	0	1	male	'\'B2of10\''	Adult	None
9	0	3	male	'\'B1of10\''	Child	Many

- **KNN**
  - One-hot Encoding:

Attribute	Original Format	Revised Format
Pclass	Nominal { 1 = 1st (Upper), 2 = 2nd (Middle), 3 = 3rd (Lower) }	Numeric (Pclass=1) Numeric (Pclass=2) Numeric (Pclass=3)
Sex	Nominal { male, female }	Numeric (Sex=female)

  - Screenshot of Titanic_train_kNN.csv:

	A	B	C	D	E	F	G	H
1	Survived	Pclass=1	Pclass=2	Pclass=3	Sex=female	Age	SibSp	Fare
2	0	0	0	1	0	22	1	7.25
3	1	1	0	0	1	38	1	71.2833
4	1	0	0	1	1	26	0	7.925
5	1	1	0	0	1	35	1	53.1
6	0	0	0	1	0	35	0	8.05
7	0	1	0	0	0	54	0	51.8625
8	0	0	0	1	0	2	3	21.075
9	1	0	0	1	1	27	0	11.1333
10	1	0	1	0	1	14	1	30.0708
11	1	0	0	1	1	4	1	16.7

4. Integrate Data

- **No Other Dataset Sources**

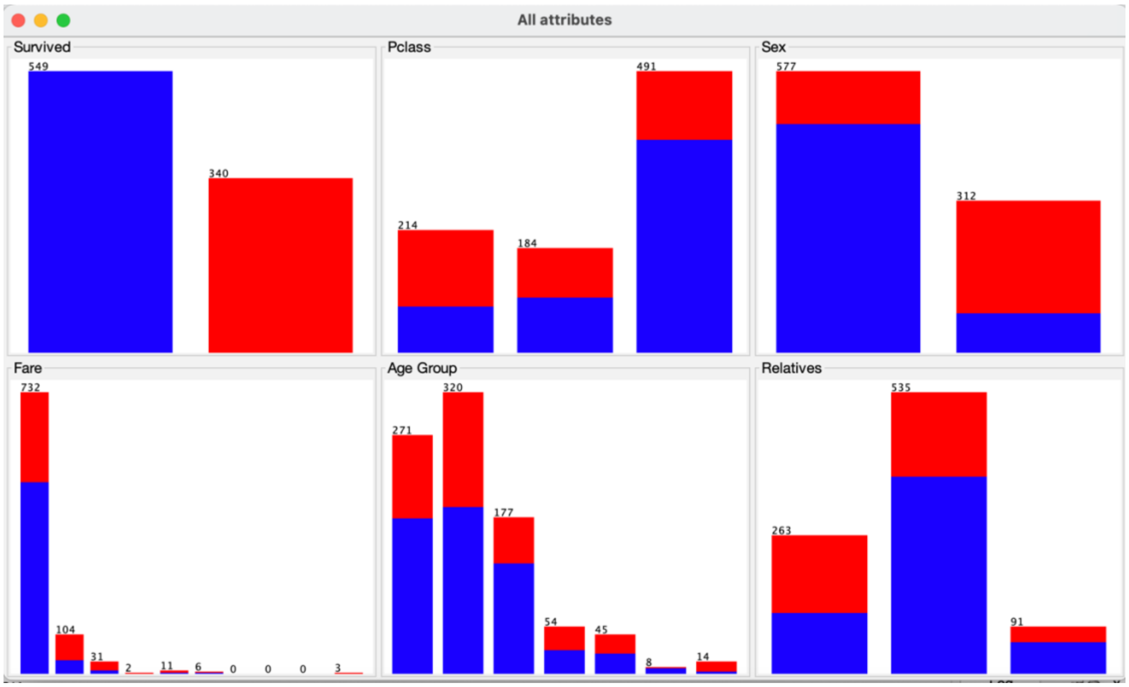
5. Format Data

- **All Well Formatted**

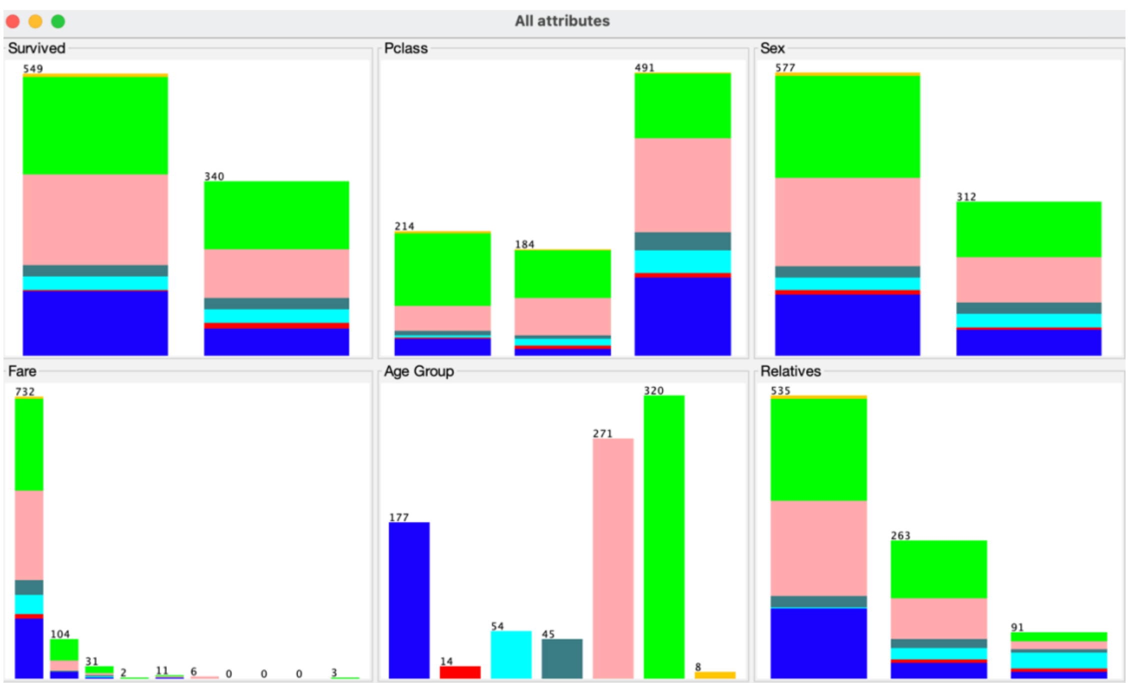
Screenshots

1. Attributes Distribution in Decision Tree Training Dataset:

- **Class Attribute (Survived) Distribution in All Attributes:**



- **Age Group Attribute Distribution in All Attributes:**



2. Files (Header and Instance) in Notepad++

- **titanic_train_DT.arff:**

```
@relation Titanic_train_DT

@attribute Survived {0,1}
@attribute Pclass {1,2,3}
@attribute Sex {male,female}
@attribute Fare {'\'B1of10\'','\'B2of10\'','\'B3of10\'','\'B4of10\'','\'B5of10\
@attribute 'Age Group' {NK,Baby,Child,Teen,Youth,Adult,Senior}
@attribute Relatives {None,Few,Many}

@data
0,3,male,'\'B1of10\'',Youth,Few
1,1,female,'\'B2of10\'',Adult,Few
1,3,female,'\'B1of10\'',Youth,None
1,1,female,'\'B2of10\'',Adult,Few
0,3,male,'\'B1of10\'',Adult,None
0,3,male,'\'B1of10\'',NK,None
0,1,male,'\'B2of10\'',Adult,None
```

- **titanic_train_kNN.arff:**

```
@relation Titanic_train_kNN

@attribute Survived {0,1}
@attribute Pclass=1 numeric
@attribute Pclass=2 numeric
@attribute Pclass=3 numeric
@attribute Sex=female numeric
@attribute Age numeric
@attribute SibSp numeric
@attribute Fare numeric

@data
0,0,0,1,0,22,1,7.25
1,1,0,0,1,38,1,71.2833
1,0,0,1,1,26,0,7.925
1,1,0,0,1,35,1,53.1
0,0,0,1,0,35,0,8.05
0,1,0,0,0,54,0,51.8625
```

Modeling

1. Select Modeling Technique:

- **kNN**
- **Decision Tree**

2. Generate Test Design:

- **kNN:**
  - 10-fold Cross-validation (k = 5)
  - Supplied Test Set (k = 5)
- **Decision Tree:**
  - 10-fold Cross-validation
  - Supplied Test Set

3. Build Model:

- **kNN:**
  - 10-fold Cross-validation (k = 5):
    - Result Summary:

```
=== Summary ===

Correctly Classified Instances         575               80.7584 %
Incorrectly Classified Instances       137               19.2416 %
Kappa statistic                          0.5963
Mean absolute error                      0.2494
Root mean squared error                  0.3818
Relative absolute error                 51.7517 %
Root relative squared error             77.7903 %
Total Number of Instances              712
```

    - Detailed Accuracy by Class:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.861	0.271	0.824	0.861	0.842	0.597	0.853	0.856	0
	0.729	0.139	0.781	0.729	0.754	0.597	0.853	0.800	1
Weighted Avg.	0.808	0.218	0.806	0.808	0.806	0.597	0.853	0.833	

- Confusion Matrix:

```
=== Confusion Matrix ===

   a   b   <-- classified as
 365  59 |   a = 0
  78 210 |   b = 1
```

- Supplied Test Set (k = 5):
  - Screenshot of res_kNN.arff in Notepad++:

```
@relation res_kNN.arff

@attribute 'prediction margin' numeric
@attribute 'predicted Survived' {0,1}
@attribute Survived {0,1}
@attribute Pclass=1 numeric
@attribute Pclass=2 numeric
@attribute Pclass=3 numeric
@attribute Sex=female numeric
@attribute Age numeric
@attribute SibSp numeric
@attribute Fare numeric

@data
0.999532,0,?,0,0,1,0,34.5,0,7.8292
0.999439,0,?,0,0,1,1,47,1,7
0.599663,0,?,0,1,0,0,62,0,9.6875
-0.199888,1,?,0,0,1,0,27,0,8.6625
0.199888,0,?,0,0,1,1,22,1,12.2875
0.599663,0,?,0,0,1,0,14,0,9.225
```

  - Table of Prediction:

Total instances in the test file	331
Number of persons predicted to survive (1)	135
Number of persons predicted not to survive (0)	196
Percentage of predicted survival	40.79 %

- **Decision Tree**
  - 10-fold Cross-validation:
    - Result Summary:

```
=== Summary ===

Correctly Classified Instances         710               79.865 %
Incorrectly Classified Instances       179               20.135 %
Kappa statistic                          0.5556
Mean absolute error                      0.2907
Root mean squared error                  0.3859
Relative absolute error                 61.5409 %
Root relative squared error             79.4152 %
Total Number of Instances              889
```

    - Detailed Accuracy by Class:

=== Detailed Accuracy By Class ===

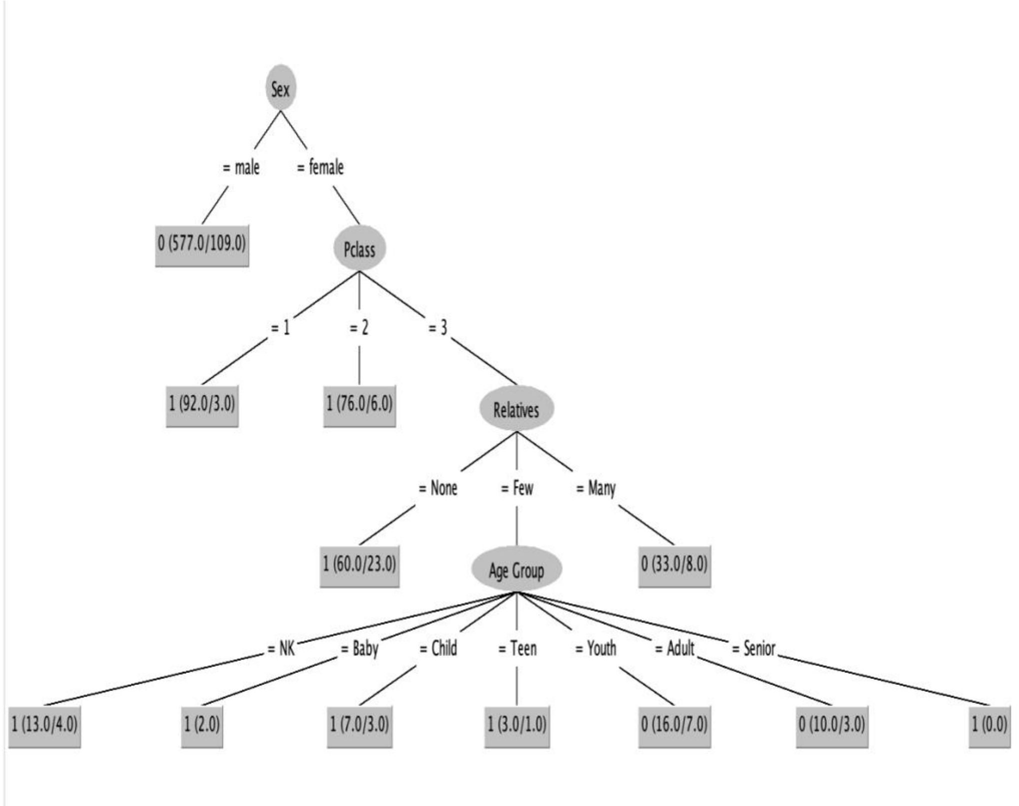
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.903	0.371	0.797	0.903	0.847	0.565	0.788	0.795	0
	0.629	0.097	0.801	0.629	0.705	0.565	0.788	0.769	1
Weighted Avg.	0.799	0.266	0.799	0.799	0.793	0.565	0.788	0.785	

    - Confusion Matrix:

```
=== Confusion Matrix ===

   a   b   <-- classified as
 496  53 |   a = 0
 126 214 |   b = 1
```

- Tree Visualization:



- Supplied Test Set:
  - Screenshot of res_DT.arff in Notepad++:

```
@relation res_DT.arff

@attribute 'prediction margin' numeric
@attribute 'predicted Survived' {0,1}
@attribute Survived {0,1}
@attribute Pclass {1,2,3}
@attribute Sex {male,female}
@attribute Fare {'\'B1of10\'','\'B2of10\'','\'B3of10\'','\'B4of10\'','\'B5of10\'',
@attribute 'Age Group' {NK,Baby,Child,Teen,Youth,Adult,Senior}
@attribute Relatives {None,Few,Many}

@data
-0.934783,1,?,1,female,'\'B2of10\'',Adult,Few
-0.934783,1,?,1,female,'\'B1of10\'',Adult,Few
-0.842105,1,?,2,female,'\'B1of10\'',Youth,Few
-0.233333,1,?,3,female,'\'B1of10\'',Youth,None
0.622184,0,?,3,male,'\'B1of10\'',Youth,None
```

- Table of Prediction:

Total instances in the test file	417
Number of persons predicted to survive (1)	131
Number of persons predicted not to survive (0)	286
Percentage of predicted survival	31.41 %

4. Assess Model:

- **kNN:**
  - 10-fold Cross-validation (k = 5):

		Prediction	
		a = 0 (non-survived)	b = 1 (survived)
Actual	a = 0 (non-survived)	TP = 0.861	FP = 0.139
	b = 1 (survived)	FP = 0.271	TP = 0.729

- Observation: TP (true-positive) of a = 0 (0.861) is higher than that of b = 1 (0.729) because the model tends to predict the result as non-survived. This can also be seen in FP (false-positive) cases, the model tends to predict the result as non-survived, which causes b = 1 (survived) a lower FP value (0.139), and a = 0 (non-survived) a higher FP value (0.271).

  - Supplied Test Set (k = 5):

File	Titanic_train_kNN.arff	res_kNN.arff
Total Number of Instance	712	331
Survived	288	135
Non-survived	424	196
Survival Rate	40.45 %	40.79 %
Survival Rate Difference	0.24 %	

- Observation: The predicted survival rate of the supplied test set is very close to the actual survival rate of the training dataset. Suppose that the two datasets are normally sampled from the same population, then their survival rates must be close to each other. This is convincing that our model is accurate.

- **Decision Tree:**
    - 10-fold Cross-validation:

		Prediction	
		a = 0 (non-survived)	b = 1 (survived)
Actual	a = 0 (non-survived)	TP = 0.903	FP = 0.097
	b = 1 (survived)	FP = 0.371	TP = 0.629

        - Observation: Compared with the training result of kNN algorithm, the TP (true-positive) of a = 0 (0.903) is even higher than that of 0.861 previously. And the FP (false-positive) of b = 1 (0.371) is also higher than that of 0.271. This implies that the decision tree model tends to predict the result as non-survived even more seriously. This could be a sign of the model is biased. This might be caused by using wrongly categorized groups to train the model.

        - Explanation of the Tree Visualization: The algorithm predicts the passenger will die if he is a male. If the passenger is a female, she will live if she is in ticket class 1 (High) or 2 (Medium). If she is in class 3 (Low), she will live if she does not have any relatives, or die if she has many relatives (3 or more). If she has few relatives (1 or 2), it depends on her age to survive or not. If she is a baby ($age < 2$), child ($2 \leq age < 12$), teen ($12 \leq age < 18$), senior ($age > 65$), or NK (age unknown), then she will live. Otherwise, if she is a youth ($18 \leq age < 30$) or adult ($30 \leq age \leq 65$), she will die.

    - Supplied Test Set:

File	Titanic_train_DT.arff	res_DT.arff
Total Number of Instance	889	417
Survived	340	131
Non-survived	549	286
Survival Rate	38.25 %	31.41 %
Survival Rate Difference	6.84 %	

- Observation: The predicted survival rate, in this case, is roughly 10 % lower than the survival rate of 40 % in kNN (actual and predicted) and 6.84 % lower than the actual survival rate in the decision tree dataset. This could happen because of wrongly categorizing the factors into inappropriate groups, which introduces bias into our model.

Evaluation

1. Evaluate Results:

- **kNN**
  - Comparison of the Survival Rates:

Data Source	Training Dataset	Supplied Test Set
Total Number of Instance	712	331
Survived	288	135
Non-survived	424	196
Survival Rate	40.45 %	40.79 %
Survival Rate Difference	0.24 %	

- Comments: The test result's predicted survival rate is close to the actual one in the training dataset. If the two datasets are sampled from the same population without bias, it is reasonable that the two figures should be close. Hence, we are satisfied with the model.

- **Decision Tree**
  - Comparison of the Survival Rates:

Data Source	Training Dataset	Supplied Test Set
Total Number of Instance	889	417
Survived	340	131
Non-survived	549	286
Survival Rate	38.25 %	31.41 %
Survival Rate Difference	6.84 %	

- Comments: The difference between the actual survival rate and the predicted survival rate is 6.84%, which is larger than the one (0.24%) of the previous model (kNN). We presume the reason is that the attributes are not well categorized into proper groups.

- **Approval: kNN Algorithm**

2. Review Process:

- **Decision Tree Bias:**
  - Sex: The decision tree model predicts all male passengers will die, which is contrary to the actual result, where only $\frac{468}{577}$ (81.11%) were dead, which reduces the overall survival rate. According to the male proportion of the actual training dataset, which is $\frac{577}{889}(64.90\%)$, the bias will reduce the overall survival rate by $(100 - 81.11)\% \times 64.90\% = 12.26\%$.

  - Pclass: The model predicts the females who have ticket class 1 (High) and 2 (Medium) will survive, which contradicts the actual cases in the training dataset, where $\frac{89}{92}$ (96.74 %) and $\frac{70}{76}$ (92.11 %) of the female in class 1 and 2 survived separately. This contributes $\frac{312}{889}(35.10\%) \times \frac{92}{312}(29.48\%) \times (1 - 0.9674) + \frac{312}{889}(35.10\%) \times \frac{76}{312}(24.36\%) \times (1 - 0.9211)$ =1.01% to the overall predicted survival rate in the test dataset.

  - Others: As the tree branches to lower levels, the remaining subgroup proportion to the whole group is getting lower, thus the subsequentially produced bias becomes minor, which can be ignored.

  - Conclusion: The major bias produced by the decision tree model is because of the inaccurate prediction of the male's survival (0%), which tremendously reduces its accuracy compared to the kNN model.

3. Determine Next Steps:

- **Consideration:**
  - kNN: In summary, the kNN model shows a smaller difference between the actual and predicted survival rates compared to the Decision Tree model. This indicates that the kNN model is performing better in predicting survival outcomes.

  - Decision Tree: The model introduces bias by only using its categories to make the decision progress, which biases the model by ignoring the factors of continuous data value. However, the tree model can help humans observe the crucial factor in a dataset in categorical groups, and watch how different categories tend to have a certain outcome.

- **Decision: Proceed**
  - Reasons: The survival rate difference between the training and test datasets in kNN is really small (0.24%). Though the survival rate difference between the training and test datasets in Decision Tree Model is 6.84%, apparently larger than that in kNN. We discovered that the bias is produced by using categorical groups to make the result decision, which ignores the continuous data value and produces biases in prediction. This can be unavoidable in using a tree-based training model. Since everything is within our expectations, we decide to proceed.

Discussion of Results

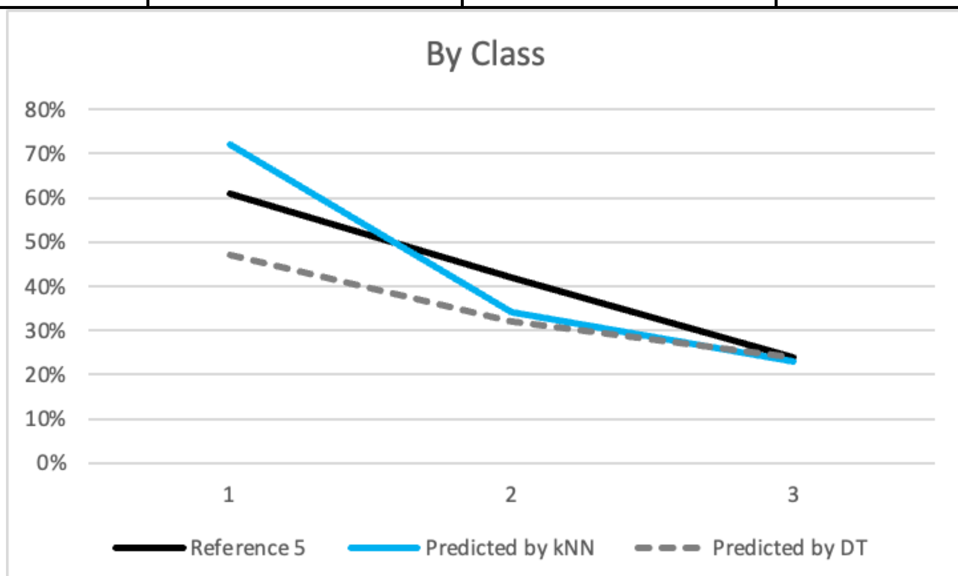
In this section, we compare the survival rates of passengers based on different attributes, specifically ticket class, and gender. We compare the actual survival rates from a reference source (Reference 5) with the predicted survival rates obtained from the kNN and Decision Tree models.

1. Titanic Survivors

	Reference 5	Predicted by kNN	Predicted by DT
Survivor %	37%	41%	31%

2. Titanic Survivors by Class

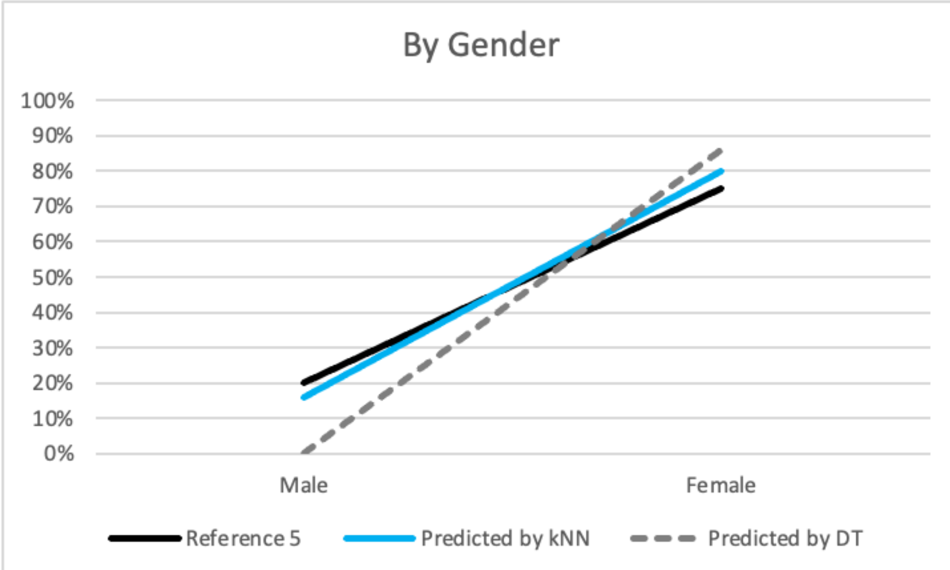
Pclass	Reference 5	Predicted by kNN	Predicted by DT
1	61%	72%	47%
2	42%	34%	32%
3	24%	23%	24%



Discussion: The reference source indicates that first-class passengers had the highest survival rate (61%), followed by second-class passengers (42%) and third-class passengers (24%). The kNN model predicts higher survival rates for first-class passengers (72%) compared to the reference, which aligns with the expectation of first-class passengers having a better chance of survival. However, the Decision Tree model predicts a lower survival rate for first-class passengers (47%), which suggests different decision rules in the model that might be capturing other factors influencing survival.

3. Titanic Survivors by Gender

Gender	Reference 5	Predicted by kNN	Predicted by DT
Male	20%	16%	0%
Female	75%	80%	86%



Discussion: The reference source indicates a significantly higher survival rate for females (75%) compared to males (20%). Both the kNN and Decision Tree models predict higher survival rates for females, with the kNN model predicting 80% and the Decision Tree model predicting 86%. The models correctly capture the importance of gender in predicting survival, indicating that being female increases the chances of survival.

Overall: The predictions of the kNN and Decision Tree models are consistent with the reference source for the attribute of gender. However, there are some variations in the predictions for ticket class, with the Decision Tree model differing from the reference source and the kNN model. These variations might be attributed to different decision rules and splits used by the Decision Tree model, which could capture unique patterns in the data.

Conclusion

In conclusion, this analysis of the Titanic dataset using kNN and Decision Tree models provides valuable insights into the factors that influenced the survival outcomes of passengers on the Titanic. The kNN model exhibited better accuracy in predicting survival rates compared to the Decision Tree model. It closely matched the actual survival rates, indicating its effectiveness in classification. The Decision Tree model, although less accurate, still provided useful information for understanding the relationships between attributes and survival outcomes.

The analysis revealed that passenger class and gender were significant factors in determining survival on the Titanic. First-class passengers had a higher likelihood of survival compared to those in lower classes, which aligns with historical data. Moreover, gender played a crucial role, with females having a significantly higher survival rate than males. These findings validate the widely known "women and children first" protocol followed during the Titanic tragedy.

Moving forward, there are several avenues for improvement. Further analysis could focus on exploring misclassified instances to understand the reasons for prediction errors and refine the models accordingly. Feature engineering techniques can be employed to create more informative attributes and enhance the models' performance. Additionally, alternative algorithms beyond kNN and Decision Trees could be considered to see if they offer better accuracy or interpretability. Ultimately, the models should be validated on an independent test set to ensure their generalizability and reliability.

Overall, this analysis demonstrates the potential of machine learning techniques in analyzing historical datasets and predicting outcomes. It provides valuable insights into the Titanic incident and offers a foundation for further research in the field. By leveraging the power of data and machine learning, we can gain a deeper understanding of complex events and contribute to improved decision-making in various domains.

References

[1] “CS109,” web.stanford.edu. http://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html

[2] “Medium,” Medium, 2022. https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers- (accessed Jun. 02, 2023).

[3] “Titanic: Machine Learning from Disaster,” kaggle.com. https://www.kaggle.com/c/titanic

[4] S. Cicoria, J. Sherlock, M. Muniswamaiah, and L. Clarke, “Classification of Titanic Passenger Data and Chances of Surviving the Disaster Data Mining with Weka and Kaggle Competition Data,” 2014. Available: http://csis.pace.edu/~ctappert/srd2014/d3.pdf

[5] “Titanic Survivors • Titanic Facts,” Titanic Facts, 2018. https://titanicfacts.net/titanic-survivors/