Name : Bram Andika Ahmad Al Aziz
Student Number : 349166

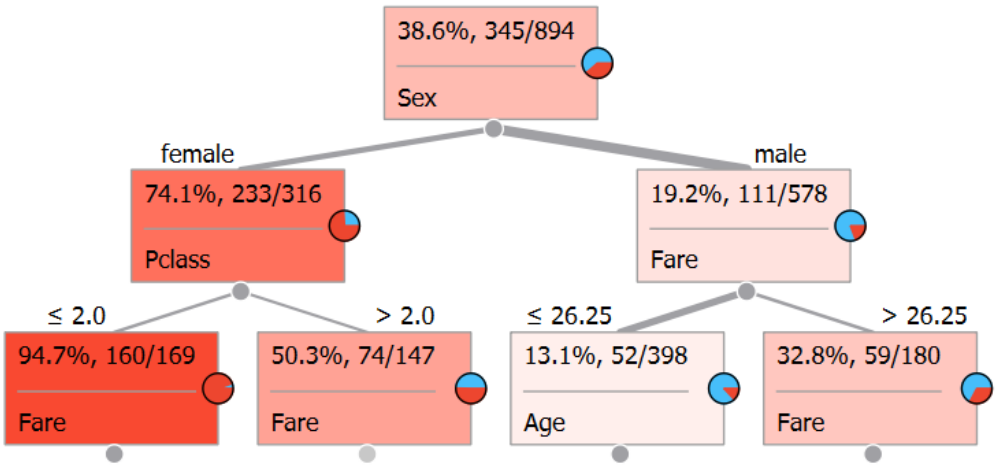
Final Project: Data Mining

Goals:

1. Find which factors are crucial to survive. Your data mining findings should allow you to show the most important factors,
2. Answer the question - using your model (not what was on movie: Who has a higher chance to survive: Jack, as a young male, traveling alone as a passenger of 3rd class, or Rose, as a young female, traveling with family and passenger of 1st class?

Answers:

1. The factor that greatly influences survival rate is gender, from the overall data the percentage of survival is 38.6% consisting of 74.1% for women and 19.2% for men. The second factor that affects women is the class of passengers, 94.7% for class 1 and 2 and 50.3% for class 3. Meanwhile for men the second factor that influences are based on fare, 13.1% if they paid <= 26.25 and 32.8% if they paid 26.25 %.



2. According to the prediction that was created with some models, We can conclude that Rose has bigger chance to survival than jack.

Predictions - Orange

Show probabilities for: Classes in data

	Logistic Regression	Random Forest	Naive Bayes	Gradient Boosting	Neural Network
1	0.90 : 0.10 → 0	0.57 : 0.43 → 0	0.86 : 0.14 …	0.88 : 0.12 → 0	0.86 : 0.14 → 0
2	0.09 : 0.91 → 1	0.20 : 0.80 → 1	0.14 : 0.86 …	0.16 : 0.84 → 1	0.01 : 0.99 → 1

Survived	Name	Pclass	Sex	Age	SibSp
0	Jack	3	male	25	0
1	Rose	1	female	25	3

CRISP-DM for Titanic Dataset

The **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*) has six phases:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

We will follow each of the phases for Titanic Dataset

1. Business understanding

We have some question regarding this dataset to answer:

a. Which factors are crucial to survive and show the most important factors?
b. Answer the question with model, Who has a higher chance to survive: Jack, as a young male, traveling alone as a passenger of 3rd class, or Rose, as a young female, traveling with family and passenger of 1st class?

2. Data understanding

The data that are used is titanic dataset with some column on it:

```
data = pd.read_csv("titanic.tsv", sep='\t')
data.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	ship
0	1	0	3.0	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7,25	NaN	S	Titanic
1	2	1	1.0	Cumings, Mrs. John Bradley (Florence Briggs Th…	female	38	1	0	PC 17599	71,2833	C85	C	Titanic
2	3	1	3.0	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7,925	NaN	S	Titanic
3	4	1	1.0	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53,1	C123	S	Titanic
4	5	0	3.0	Allen, Mr. William Henry	male	35	0	0	373450	8,05	NaN	S	Titanic

3. Data preparation

There are some problems with data like missing values, incorrect labeling, false data. So, in this step we clean and prepare data.

Data describe before cleaning:

```
data.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	894.000000	894.000000	894.000000	721.000000	894.000000	893.000000	892.000000
mean	446.030201	0.381432	2.305369	35.836019	0.604027	0.371781	32.075985
std	259.208003	0.508529	0.847653	164.927968	2.571231	0.768325	49.868844
min	-12.000000	-4.000000	-2.000000	-12.000000	0.000000	0.000000	-90.000000
25%	223.250000	0.000000	2.000000	20.000000	0.000000	0.000000	7.895800
50%	444.500000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.750000	1.000000	3.000000	38.000000	1.000000	0.000000	30.771850
max	1143.000000	1.000000	3.000000	4435.000000	70.000000	5.000000	512.329200

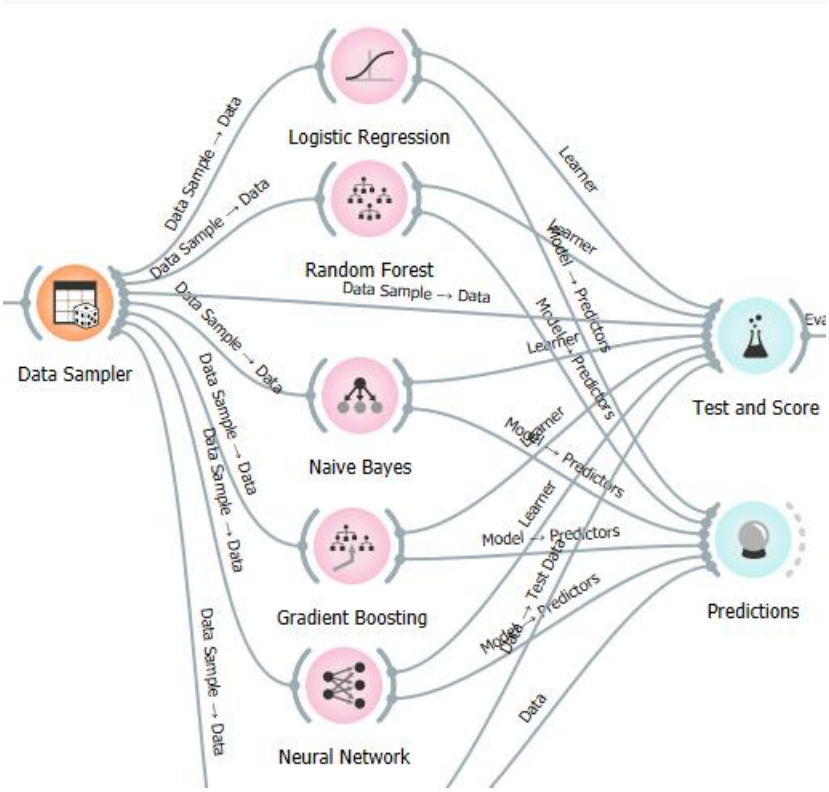
Data describe after cleaning:

```
data.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	894.000000	894.000000	894.000000	894.000000	894.000000	894.000000	894.000000
mean	446.030201	0.385906	2.309843	29.495526	0.463087	0.371781	32.750723
std	259.208003	0.487081	0.835370	12.829530	0.879200	0.767895	49.466207
min	-12.000000	0.000000	1.000000	1.000000	0.000000	0.000000	4.012500
25%	223.250000	0.000000	2.000000	22.000000	0.000000	0.000000	7.925000
50%	444.500000	0.000000	3.000000	28.000000	0.000000	0.000000	15.245800
75%	668.750000	1.000000	3.000000	35.000000	1.000000	0.000000	32.075985
max	1143.000000	1.000000	3.000000	80.000000	5.000000	5.000000	512.329200

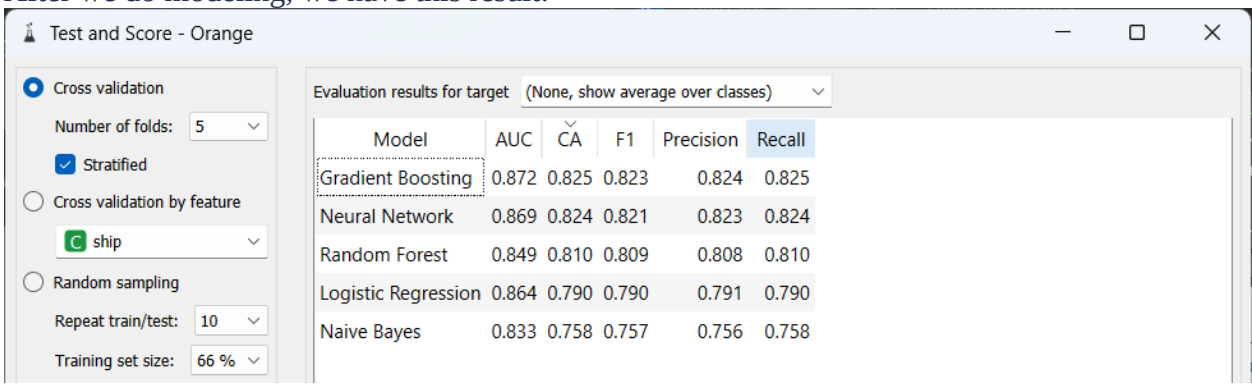
4. **Modeling**

In this step we use several algorithms: Logistic Regression, Random Forest, Naïve Bayes, Gradient Boosting, Neural Network to generate model and create a prediction about who has a high change of survival and determining factor that are crucial to survive.



5. **Evaluation**

After we do modeling, we have this result:



Model	AUC	CA	F1	Precision	Recall
Gradient Boosting	0.872	0.825	0.823	0.824	0.825
Neural Network	0.869	0.824	0.821	0.823	0.824
Random Forest	0.849	0.810	0.809	0.808	0.810
Logistic Regression	0.864	0.790	0.790	0.791	0.790
Naive Bayes	0.833	0.758	0.757	0.756	0.758

We can see from there, Gradient Boosting has higher accuracy for classification 82.5%. The second algorithms that has higher accuracy is Neural Network with percentage 82.4%.

6. **Deployment**

In this step we should make a deployment for the customer to access the result, but we will not do it now.