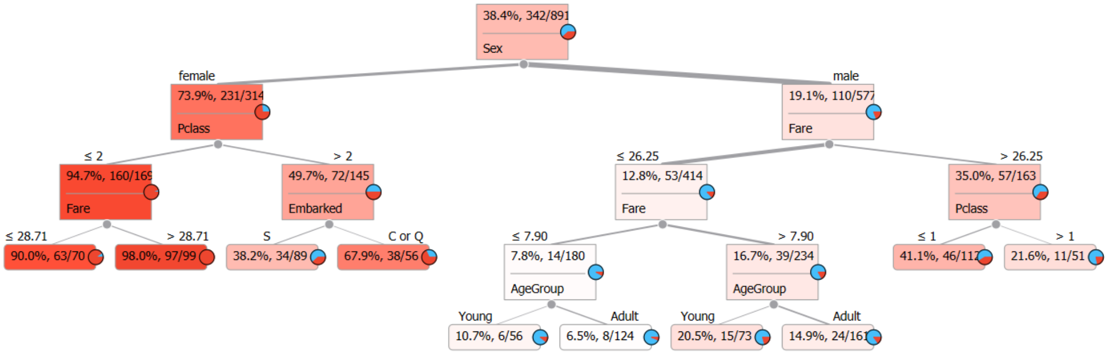
Dandy Wibowo (334541)

The Goals:

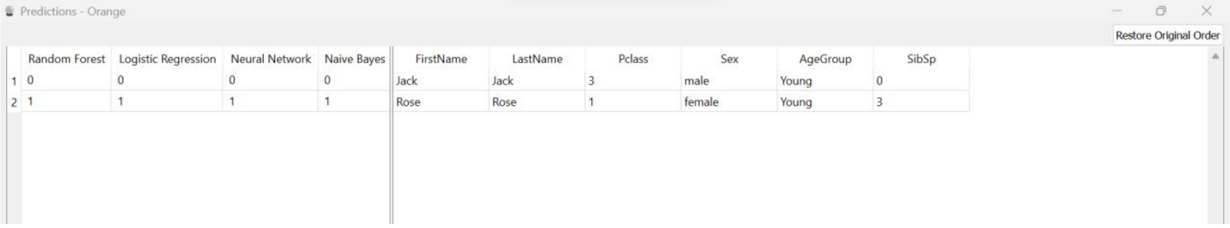
1. Find which factors are crucial to surviving. Your data mining findings should allow you to show the most important factors,
2. Answer the question - using your model (not what was in the movie: Who has a higher chance to survive: Jack, as a young male, traveling alone as a passenger of 3rd class,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

The Answer:

1. The highest factors that affect survival rates are gender. Women have a higher chance of survival than men. 73.9% of women survive, and men only 19.1%. The second factor is the class of passengers; 94.7% of women in the first and second classes stayed than those in the third. For men, the fare has the most significant impact after gender, there 35% of the men passengers survived if they paid > 26.25.

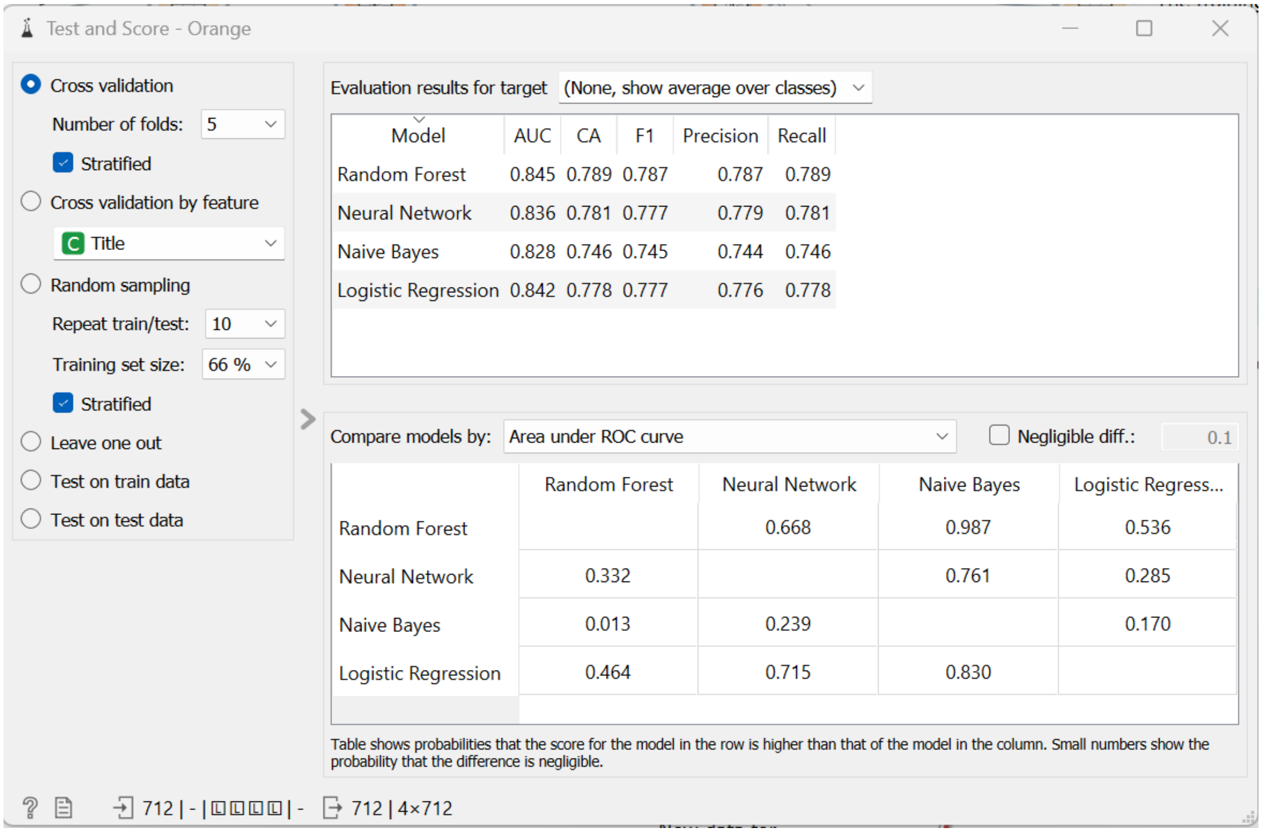


2. According to the model created by the author, Rose has a bigger chance of survival than Jack.



	Random Forest	Logistic Regression	Neural Network	Naive Bayes	FirstName	LastName	Pclass	Sex	AgeGroup	SibSp
1	0	0	0	0	Jack	Jack	3	male	Young	0
2	1	1	1	1	Rose	Rose	1	female	Young	3

The above picture shows Rose is predicted to survive the titanic accident based on the four models. For test and score results show in the following picture.



● Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

C Title

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.845	0.789	0.787	0.787	0.789
Neural Network	0.836	0.781	0.777	0.779	0.781
Naive Bayes	0.828	0.746	0.745	0.744	0.746
Logistic Regression	0.842	0.778	0.777	0.776	0.778

Compare models by: Area under ROC curve

□ Negligible diff.: 0.1

	Random Forest	Neural Network	Naive Bayes	Logistic Regress...
Random Forest		0.668	0.987	0.536
Neural Network	0.332		0.761	0.285
Naive Bayes	0.013	0.239		0.170
Logistic Regression	0.464	0.715	0.830	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.