Titanic AI project

Session 2

Team:
Yuliia Nikolaenko
Eduardo Bonnefemne
Gabriel Pérez García

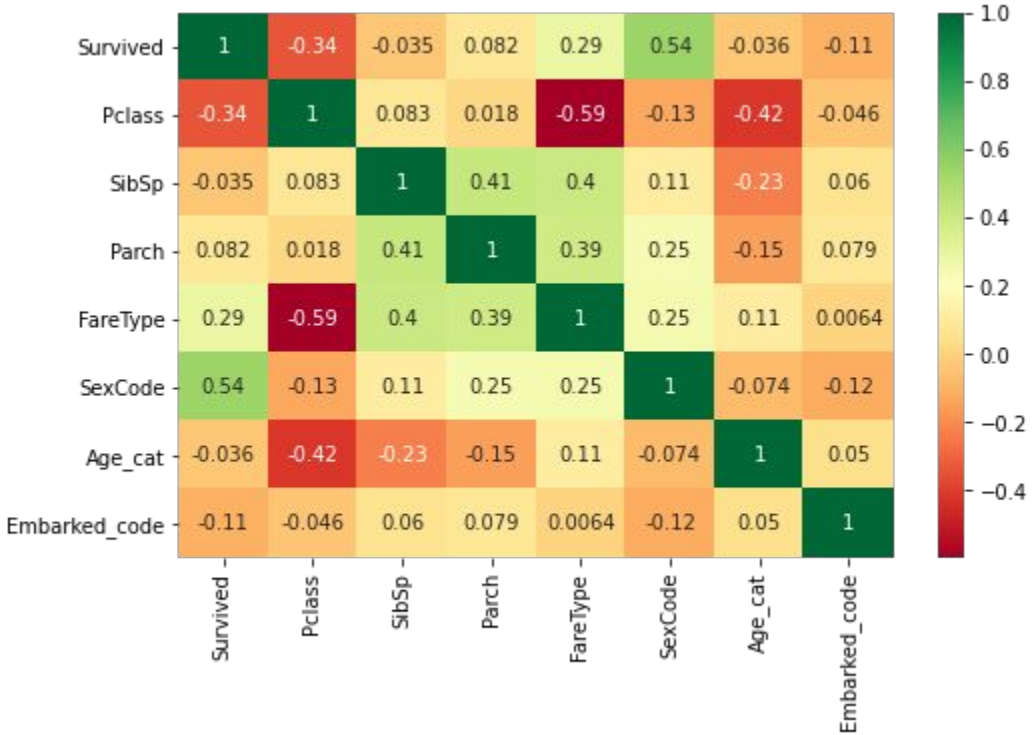
Questions to answer:

1. What error rate can we expect on the test set?
2. What is the best model?
3. What are the most important variables to predict the outcome?

Tasks for data pre-pocessing:

1. to deal with missing values;
2. to recode the variables into categories;
3. to define the features for the models what will show the highest accuracy scores.

Logistic Regression



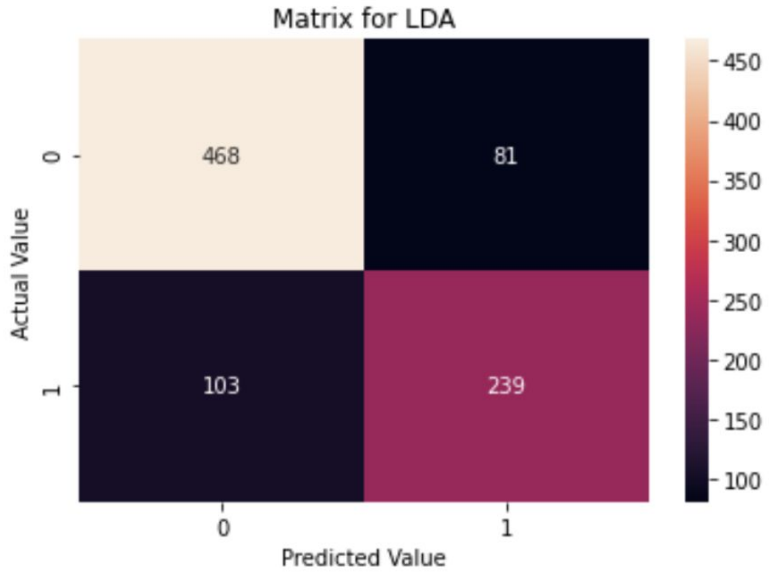
LDA

The Test set has a better accuracy → Maybe overfitting ?

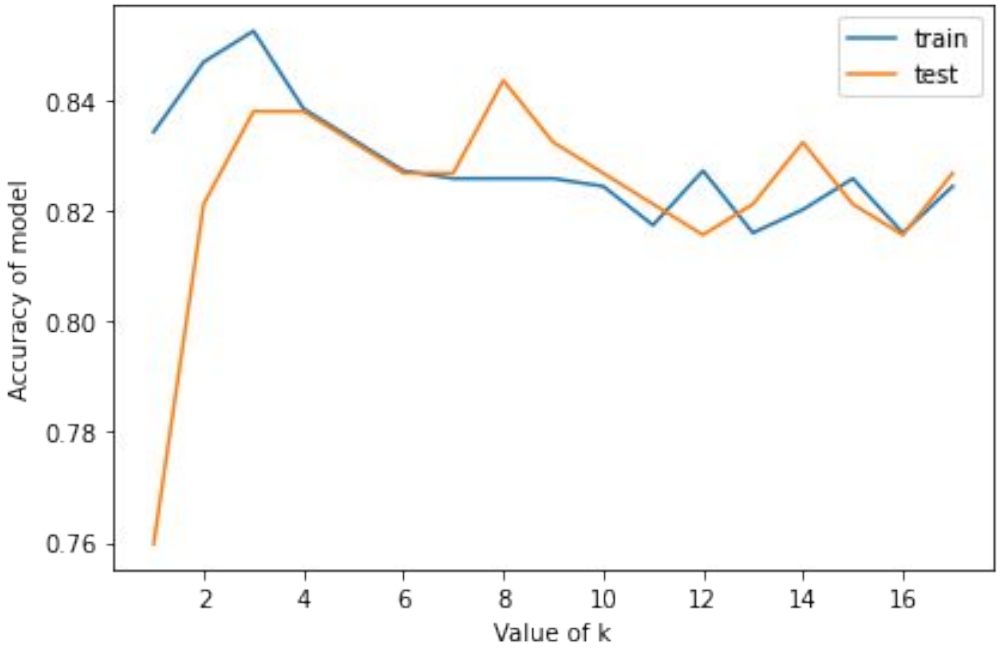
```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
ldaModel = LinearDiscriminantAnalysis().fit(X_train,y_train)
yhat = ldaModel.predict(X_test)

print("Train set Accuracy: ", metrics.accuracy_score(y_train, ldaModel.predict(X_train)))
print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat))
```

```
Train set Accuracy:  0.7879213483146067
Test set Accuracy:  0.8212290502793296
```

The confusion Matrix shows 81 FP and 103 FN



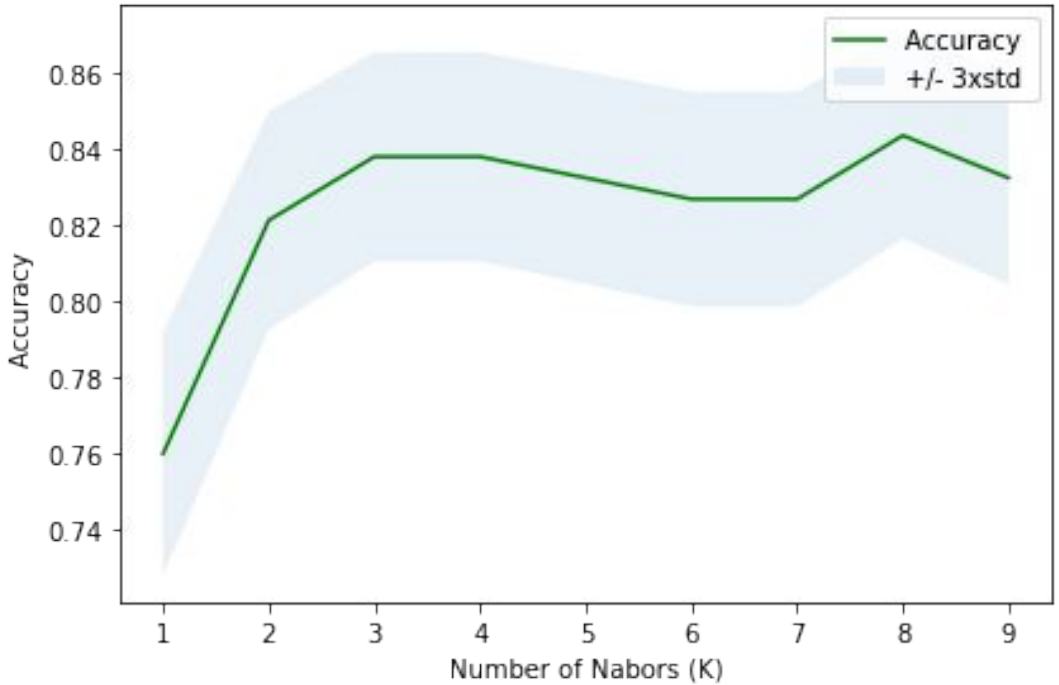
k-Nearest Neighbors (kNN)



We can calculate the accuracy of KNN for different Ks.

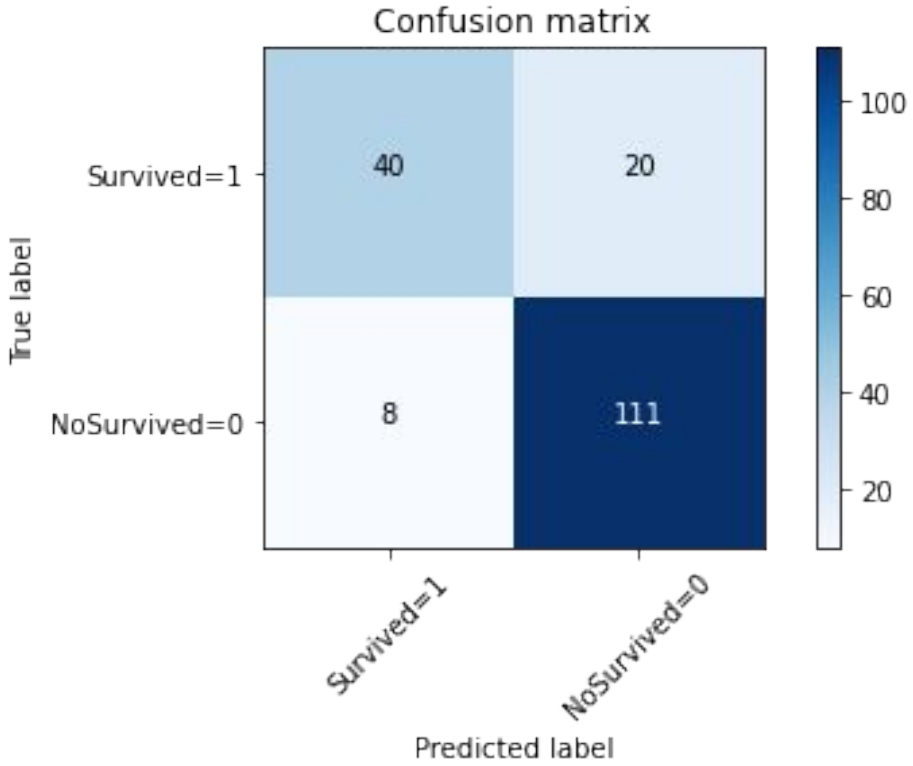
Accuracy, calculates how closely the actual labels and predicted labels are matched in the test set.

k-Nearest Neighbors (kNN)



The best accuracy was 0.8435 with k= 8

k-Nearest Neighbors (kNN)

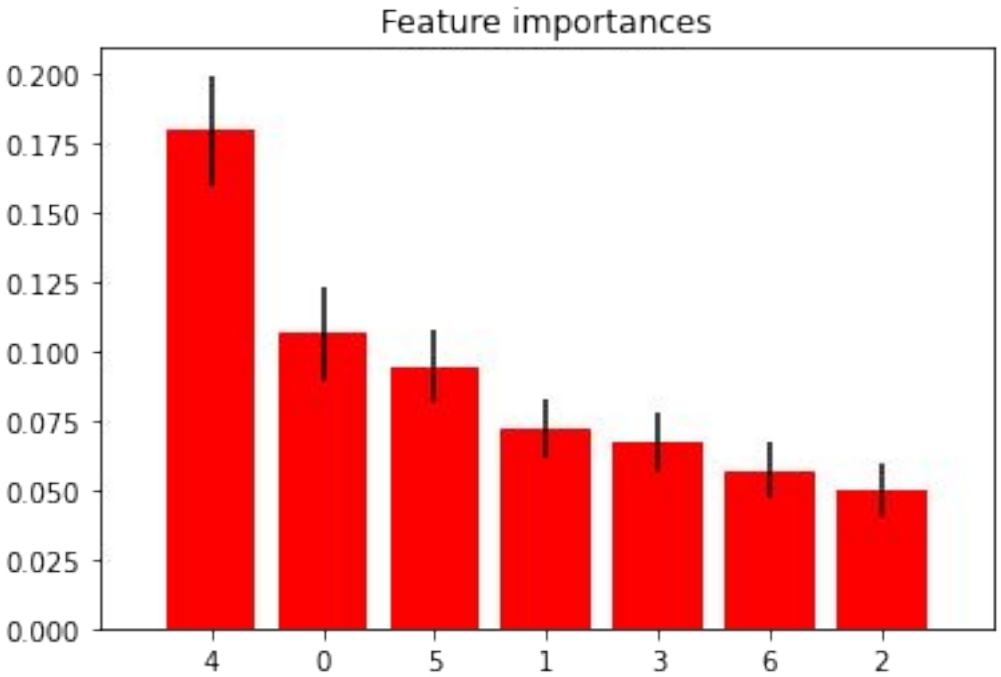


The confusion Matrix shows 8 FP and 20 FN.

Methods comparison

Model	**Accuracy score**	**Error rate**
Logistic Regression	85%	15%
LDA	82%	18%
QDA	83%	17%
KNN	84%	16%

The most important variables to predict the outcome



Feature name	Mean Importance	Std Importance
Sex	0.18	0.02
Pclass	0.11	0.02
Age	0.09	0.01
Fare	0.07	0.01
SibSp	0.07	0.01
Embarked	0.06	0.01
Parch	0.05	0.01