ENSF 611: FINAL PROJECT PRESENTATION

BY:

CHELSEA JOHNSON

TOYA OKEKE

STEPHANIE WALSH

INTRODUCTION

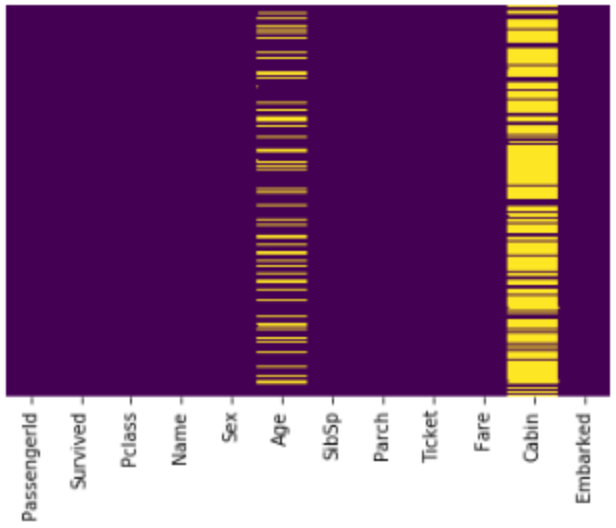
- The Kaggle **Titanic: Machine Learning from Disaster Project**
- Predicting whether a passenger survived or didn't survive the disaster

PROCESS

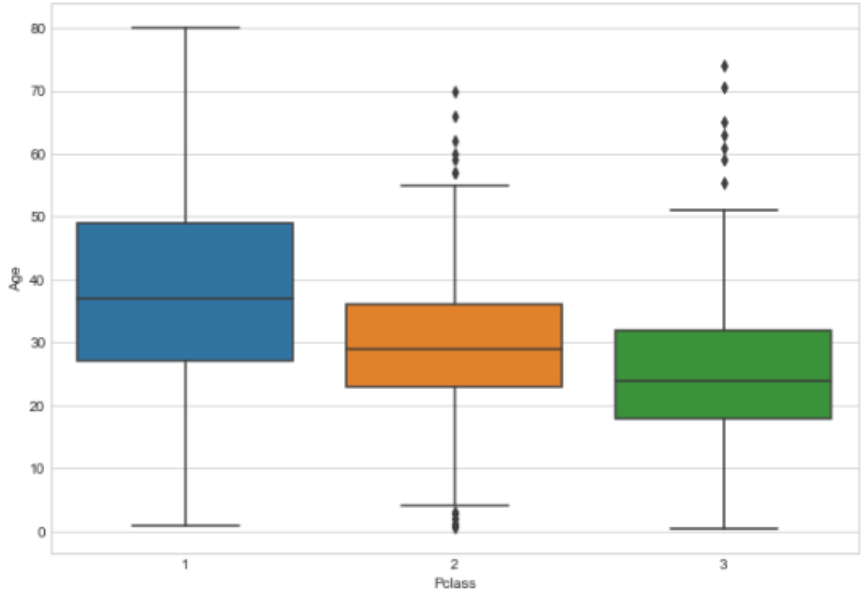
1. Visualize the data to see what is missing
2. Clean the data to fill in NaN values
3. Remove features that we deemed were not helpful
4. Data correlations and data transformation
5. Build models based on the models we learned in this course
6. Use our best hyperparameters for the Ensemble Method
7. Challenges and Conclusions

DATA VISUALIZATIONS

Feature Heatmap

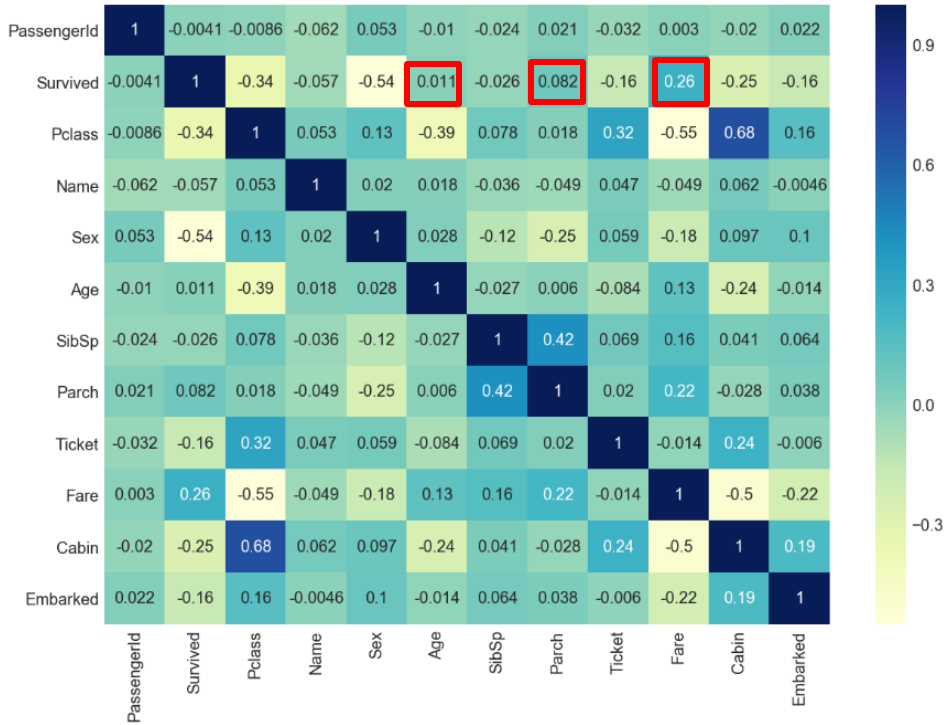


Age Range Boxplot

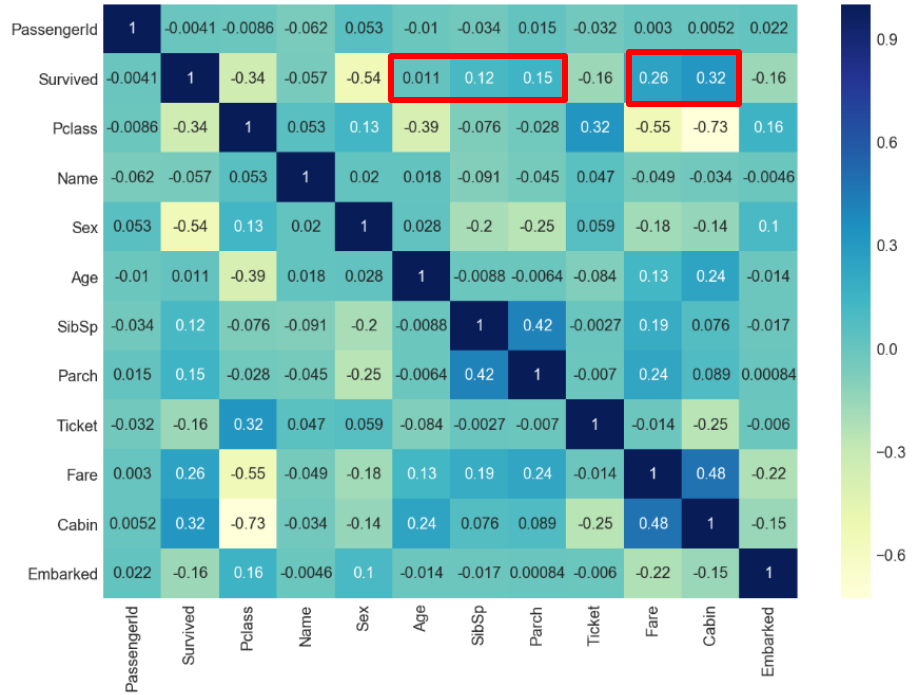


CORRELATION MATRICES

Raw Data Correlation Matrix



Transformed Data Correlation Matrix

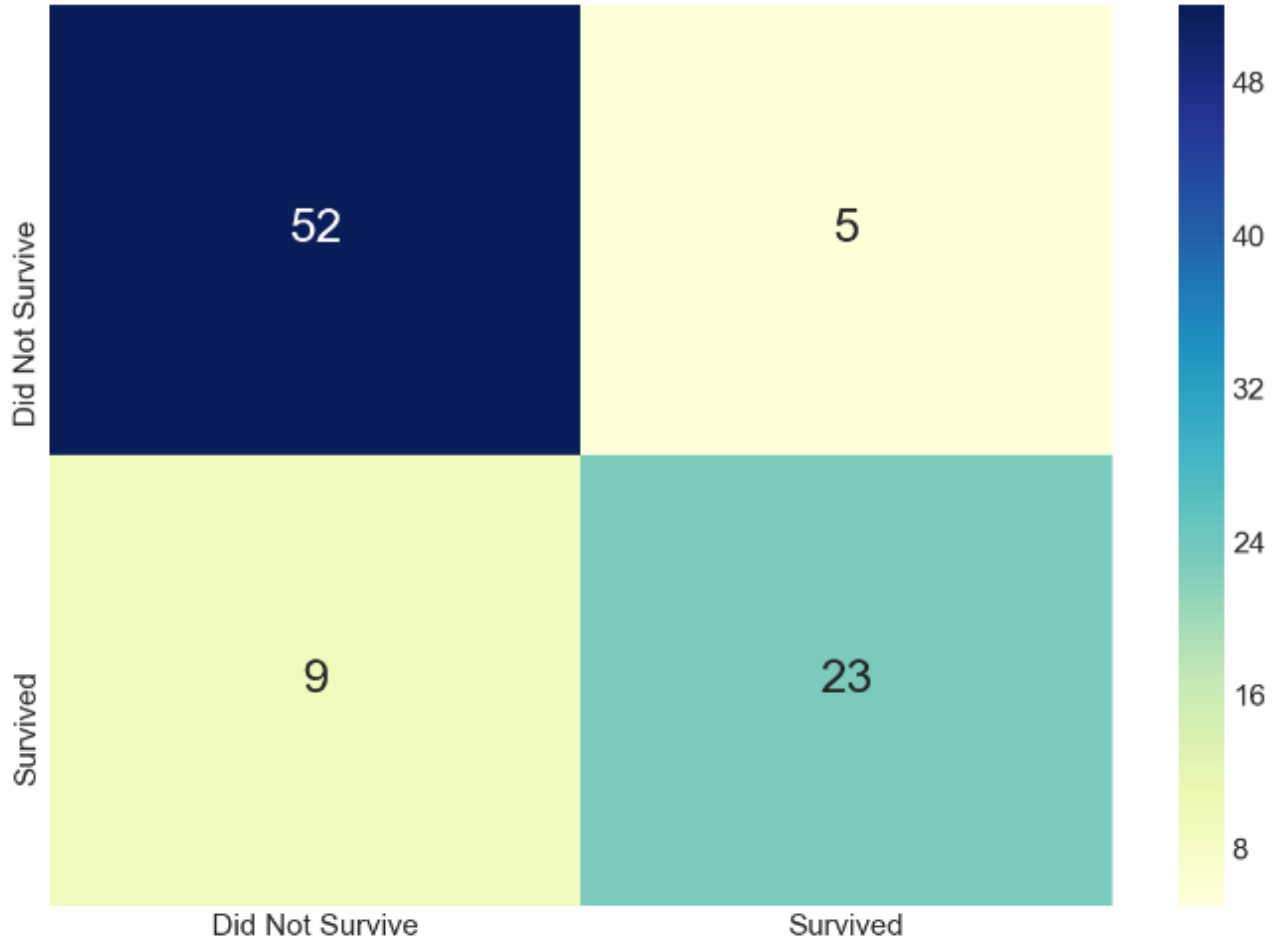


LOGISTIC REGRESSION MODEL

Best Hyperparameters
Min_sample_split: 50

Validation Performance

	precision	recall	f1-score	support
Did Not Survive	0.85	0.91	0.88	57
Survived	0.82	0.72	0.77	32
accuracy			0.84	89
macro avg	0.84	0.82	0.82	89
weighted avg	0.84	0.84	0.84	89



KAGGLE SCORE: 75.598%

SVM MODEL – TRIAL 1 (STANDARD NORMALIZATION)

```
best hyperparameters:    {'C': 1, 'degree': 2, 'kernel': 'poly'}
best mean cross-validation score:       0.803577929822359
```

Training Performance

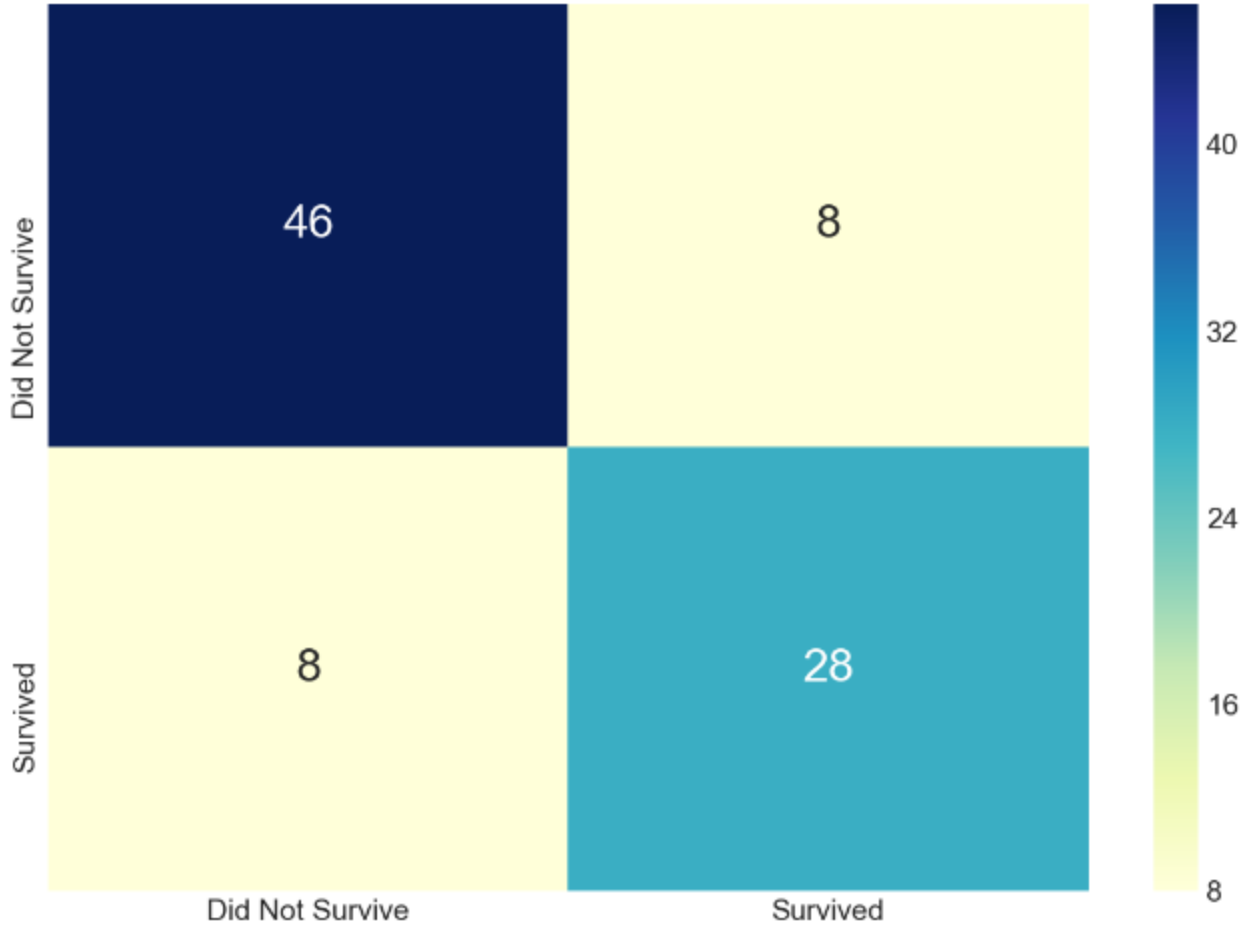
```
The Accuracy of the SVM model is:       80.65%
The Precision of the SVM model is:      79.84%
The Recall of the SVM model is: 66.01%
The F1-Score of the SVM model is:       72.27%
```

Validation Performance

```
Here is the classification report of the SVM Model:
                precision    recall  f1-score   support

Did Not Survive       0.85      0.85      0.85        54
       Survived       0.78      0.78      0.78        36

       accuracy                           0.82        90
      macro avg       0.81      0.81      0.81        90
   weighted avg       0.82      0.82      0.82        90
```



KAGGLE SCORE: 77.033%

SVM MODEL – TRIAL 2 (FEATURE TRANSFORMATION)

```
best hyperparameters:    {'C': 6, 'degree': 2, 'kernel': 'poly'}
best mean cross-validation score:       0.8069675475488044
```

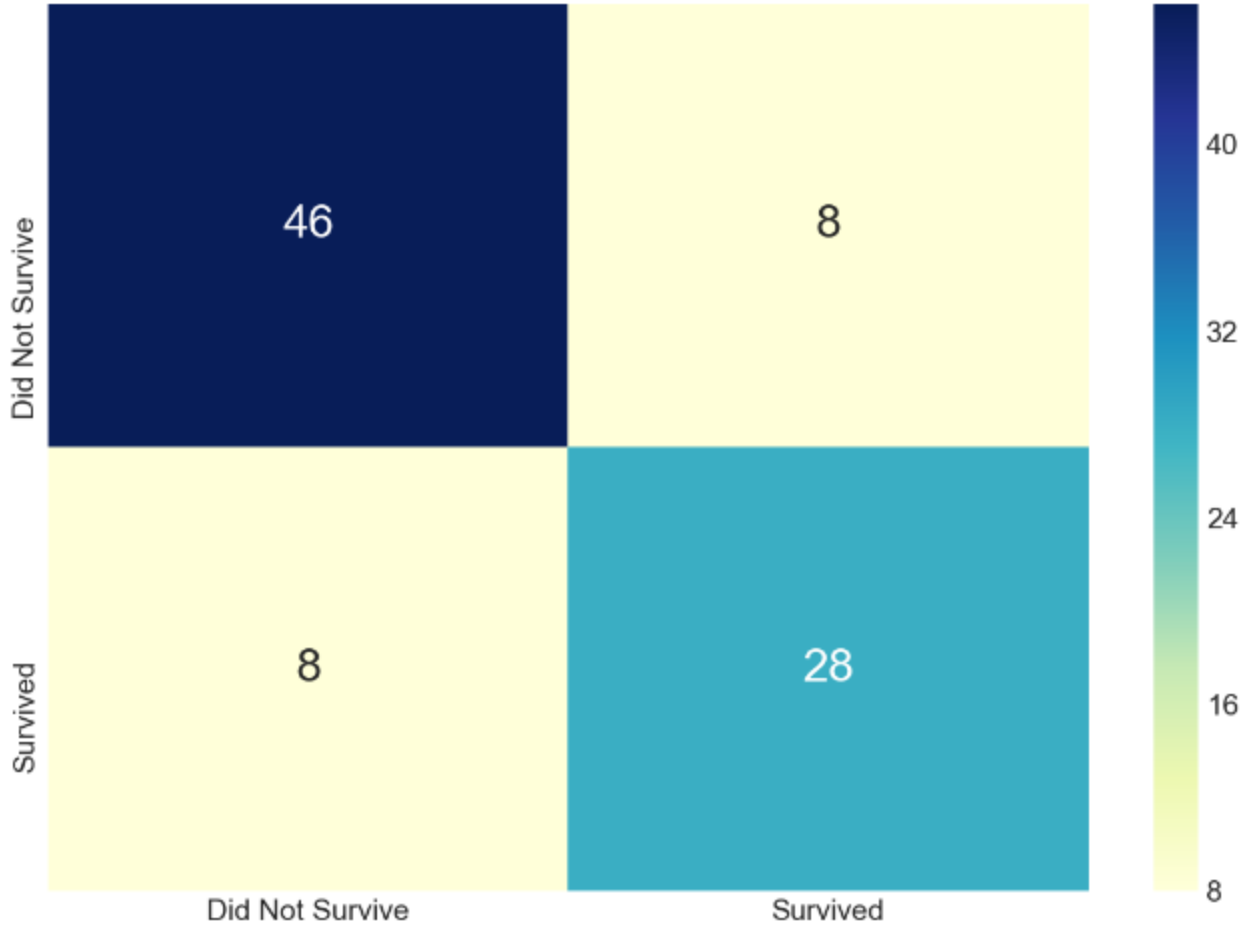
Training Performance

```
The Accuracy of the SVM model is:       81.52%
The Precision of the SVM model is:      80.62%
The Recall of the SVM model is: 67.97%
The F1-Score of the SVM model is:       73.76%
```

Validation Performance

Here is the classification report of the SVM Model:

	precision	recall	f1-score	support
Did Not Survive	0.85	0.85	0.85	54
Survived	0.78	0.78	0.78	36
accuracy			0.82	90
macro avg	0.81	0.81	0.81	90
weighted avg	0.82	0.82	0.82	90



KAGGLE SCORE: 75.598%

SVM MODEL – TRIAL 3 (BALANCING TRAINING DATA)

```
best hyperparameters:    {'C': 31, 'degree': 2, 'kernel': 'rbf'}
best mean cross-validation score:       0.8077943451499292
```

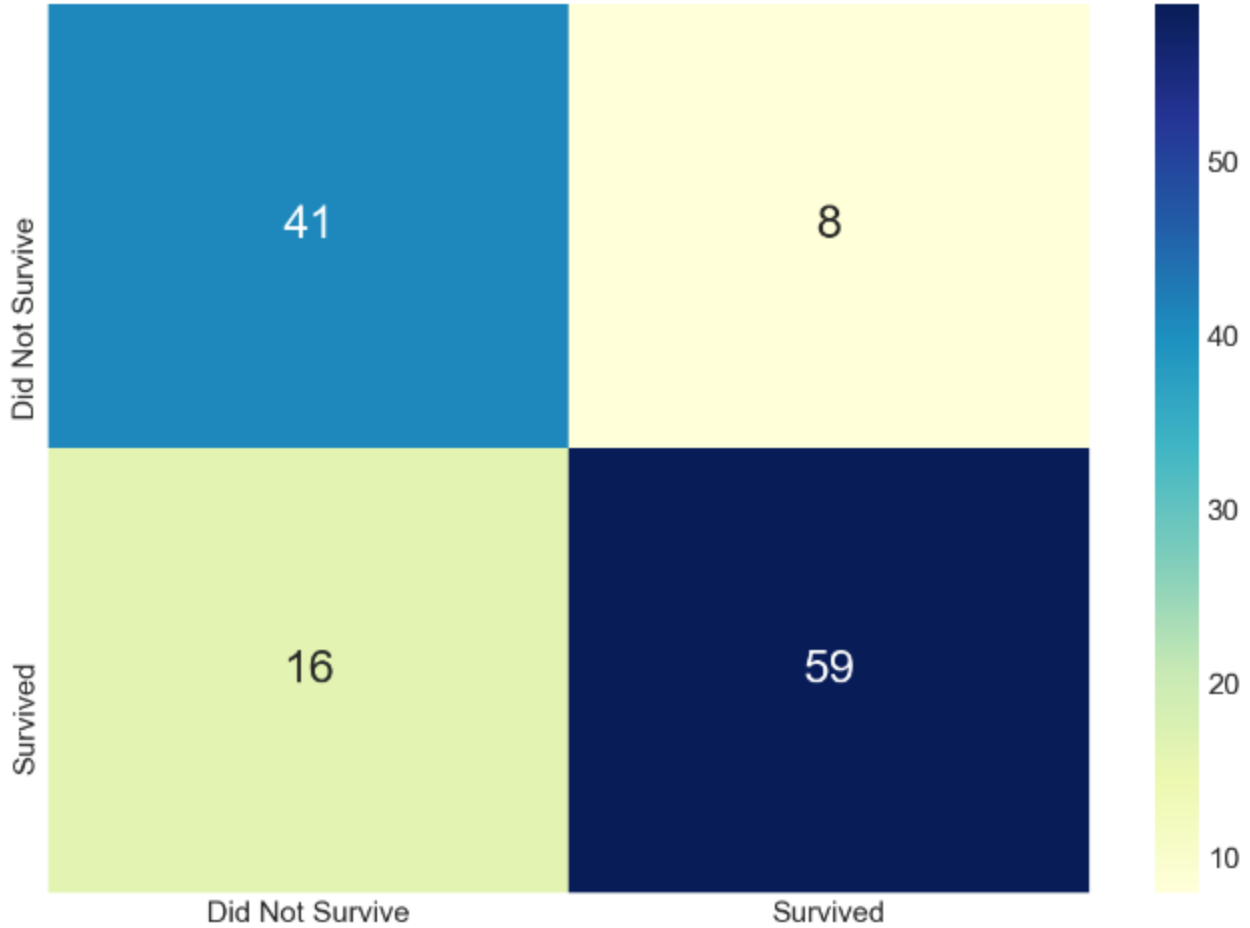
Training Performance

```
The Accuracy of the SVM model is:       84.22%
The Precision of the SVM model is:      87.16%
The Recall of the SVM model is: 83.58%
The F1-Score of the SVM model is:       85.33%
```

Validation Performance

Here is the classification report of the SVM Model:

	precision	recall	f1-score	support
Did Not Survive	0.72	0.84	0.77	49
Survived	0.88	0.79	0.83	75
accuracy			0.81	124
macro avg	0.80	0.81	0.80	124
weighted avg	0.82	0.81	0.81	124



KAGGLE SCORE: 70.334%

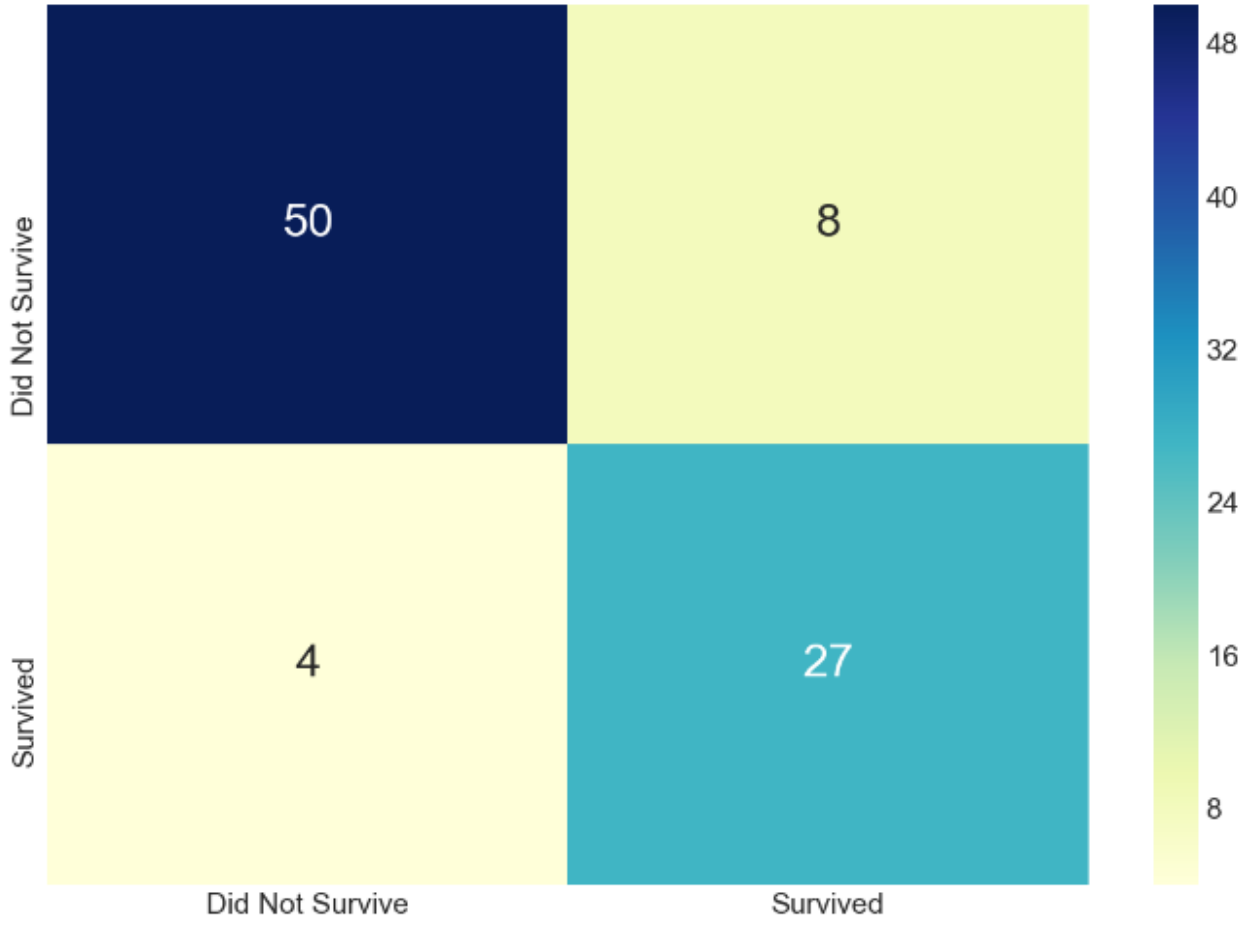
DECISION TREE MODEL

Best Hyperparameters
Min_sample_split: 50

Validation Performance

Here is the classification report of the Decision Tree Model:

	precision	recall	f1-score	support
Did Not Survive	0.93	0.86	0.89	58
Survived	0.77	0.87	0.82	31
accuracy			0.87	89
macro avg	0.85	0.87	0.86	89
weighted avg	0.87	0.87	0.87	89



KAGGLE SCORE: 57.894%

KNN MODEL

Best Hyperparameters

n_neighbors = 5
weights = ‘distance’
algorithm = ‘auto’

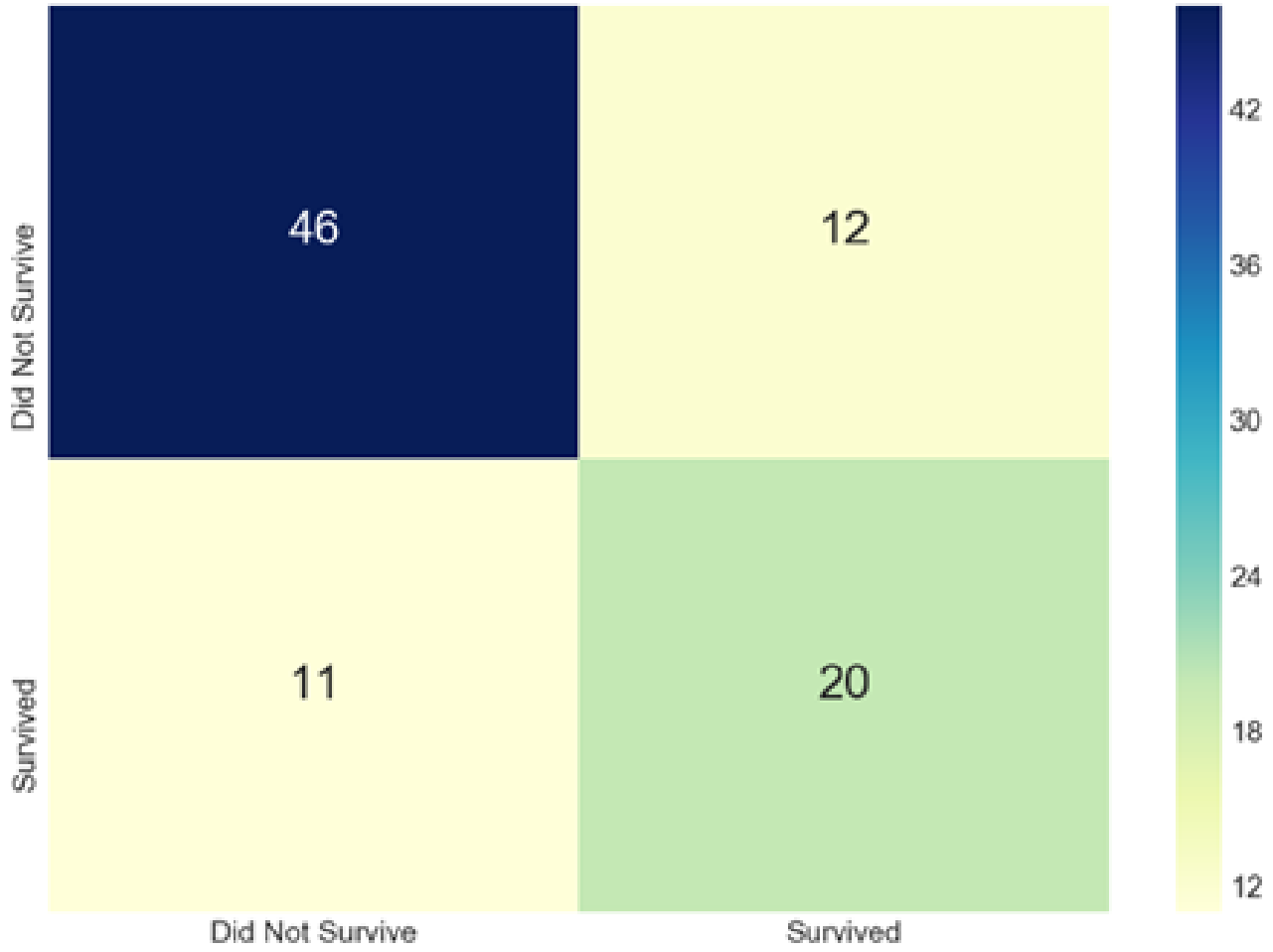
Features Excluded

Cabin
Name
Ticket
Embarked

Validation Performance

Here is the classification report of the KNN Model

	precision	recall	f1-score	support
Did Not Survive	0.81	0.79	0.80	58
Survived	0.62	0.65	0.63	31
accuracy			0.74	89
macro avg	0.72	0.72	0.72	89
weighted avg	0.74	0.74	0.74	89

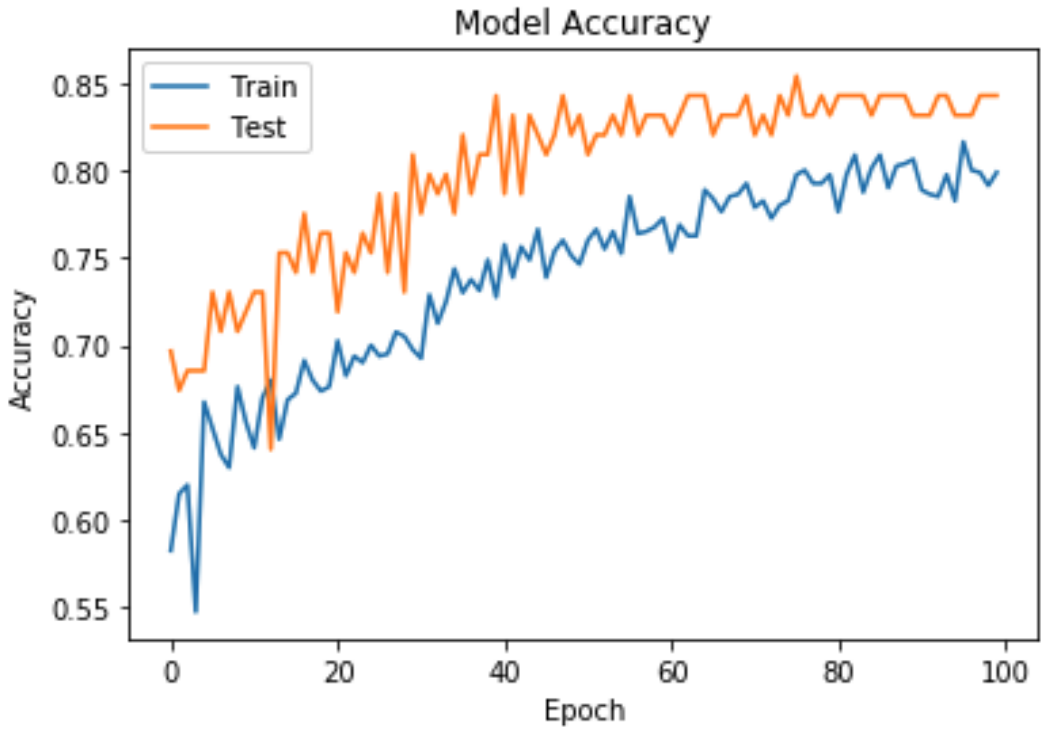


KAGGLE SCORE: 64.11%

NEURAL NETWORK (BASED ON LINEAR REGRESSION)

Best Hyperparameters

- Activation of 'relu' and 'sigmoid'
- Optimizer of 'adam'
- Epochs of 100
- Loss category of 'binary_crossentropy"



KAGGLE SCORE: 77.033%

CHALLENGES

- More training data for developing models
- F1-score was critical due to imbalanced training data (survived vs. not survived)
- No strong correlations between features and labels
- Missing data in training and test dataset

CONCLUSION

- Best model (based on Kaggle): SVM, Neural Network
- Overall drop of accuracy from training to the test data
- Was a fun project to get our feet wet with using different machine learning models
- Got to learn a bit more about selecting features