ACML Project 2020
Survival of Titanic Passengers

SM Curtis - 1657041

***Abstract*—This report looks at the prediction of whether a person on the RMS Titanic would have survived or not. Variables such as gender, location embarked from, if the person had a sibling or spouse on the ship and others influence if the person would have survived or not. The classification has two possible outcomes making it a binary classification. Using these variables, we can train and test three different binary classification models and determine which passengers would have survived. Three models are used to test their accuracy for this binary classification problem. The models used are logistic regression, K nearest neighbour and the perceptron. Each of these models is trained and tested on the Titanic dataset to predict which passengers survived and which did not. The model's accuracies are also found. It was found which passengers survived and that that K nearest neighbour has the highest accuracy. This shows that the K nearest neighbour model is better than the logistic regression and perceptron models for this binary classification problem. The results from the K nearest neighbour prediction of whether the passengers would have survived or not on the RMS Titanic are, therefore, the most accurate of the three different predictions.**

***Index Terms*—K nearest neighbour, perceptron, logistic regression, RMS Titanic, prediction, binary classification, survived**

I. Introduction

The sinking of the RMS Titanic is the most famous, and one of the deadliest, maritime disasters of the 20th century with 1501 people losing their lives [1]. This was more than two-thirds of the total amount of people aboard the ship [1]. There is much to learn from this disaster. It is interesting and beneficial to predict, based on training data, which of the passengers on the RMS Titanic would have survived based on variables such as gender, age, where the person embarked from and a few others. This prediction problem is, therefore, a supervised machine learning problem. It is specifically a binary classification problem as prediction has only two outcomes, survived or died. The implementation of a binary classification model with good accuracy for this problem can be useful in improving ship safety in current times. The most accurate model is best to help decide which passengers on ships today are most at risk in the event of an emergency based on the same variables in the dataset as they are still relevant today. Extra safety regulations can then be put in place for the different categories of people that are most at risk of dying in the case of an emergency. Using the most accurate model for prediction, it can be found which gender, average age, the average fee paid for a ticket and the ticket class of the passenger, is most at risk of death based on the RMS Titanic data. The three different models that will be implemented and compared are logistic regression, K nearest neighbour and the perceptron. Each of these models with all use the same two datasets as input. The first dataset is the training dataset and the second is the testing dataset. The output of the models is the binary classification of whether the passenger on the RMS Titanic would have survived or not. The accuracies of these models must also be found.

II. Related Work

According to [1], of all the females aboard the Titanic disaster, 72% them survived, however, of all the males, only 20.6% of them survived. It is also shown from [1] that of all first-class passengers, 61.7% of them survived, of all second-class passengers, 40.4% of them survived and of all third-class passengers, 25.3% of them survived. The mean age on the Titanic was 30 years old. It can, therefore, be seen that women and wealthier people aboard the Titanic had a higher chance of survival than men and less wealthy people.

A. Logistic Regression

Logistic regression is one of the most widely used methods in binary classification [2]. Most of the methods used in logistic regression are very similar to those used in linear regression [2]. The logistic function commonly used in a logistic regression model with the expected binary outcome is, $\frac{1}{1+e^{-x_i\beta}}$ where β is unknown parameters [2]. Logistic regression works by learning these parameters with a learning rate α to create a decision boundary between the binary classification targets.

B. Perceptron

The perceptron is a supervised learning algorithm that works best with linearly separable problems and not well with non-separable problems [4]. Data is separated using a weight vector, w, and a bias, b, to obtain the discrimination function $h(x^{(n)}) = w^T x^{(n)} + b$. After convergence occurs, $w = \sum_{n=1}^{N} \alpha_n x^{(n)}$ where α is the learning rate. The separating line is used to determine which binary classification label the test point will belong to using the weights and bias. This model can go from having an optimal set of weights to then having the worst possible weights in one iteration [4].

C. K Nearest Neighbour

K nearest neighbour is a distance-based algorithm and is extremely popular. Distance-based algorithms are mostly used in classification problems [3]. K nearest neighbour measures the distances between the test data and the training data to determine the final classification output for the test sample based on the k nearest training data's labels. K nearest

neighbour can be used for categorical data but is mostly used for numerical data [3]. The two more popular distance formula used by K nearest neighbour are Euclidean and Manhattan [3]. Their formulas for calculating the distance between points (x_1, y_1) and (x_2, y_2) are as follows, for Euclidean, $\sqrt{(x_1-x_2)^2 + (y_1-y_2)^2}$, for Manhattan, $|x_1-x_2| + |y_1-y_2|$. The steps to performing K nearest neighbour are, to determine the number of neighbours (k), find the distance between the test data point and all training data, sort these distances and select the k smallest distances and finally predict the classification for the test data point based on the label from the k nearest training data that appears the most [3].

Using built-in methods in python, LogisticRegression, Perceptron and KNeighborsClassifier and can be used to create the logistic regression, perceptron and K nearest neighbour models respectively from sklearn to compare the accuracies of each model for the problem. We can find the accuracy, precision, recall and f-score values using the following formula, for accuracy, $\frac{TP+TN}{TP+TN+FP+FN}$, for precision, $\frac{TP}{TP+FP}$, for recall, $\frac{TP}{TP+FN}$, and for f-score, $2(\frac{Recall \times Precision}{Recall+Precision})$ where TP is true positive, TN is true negative, FP is false positive and FN is false negative [3].

III. Dataset and Variables

The dataset for this problem was obtained on 16 June 2020 from [5]. The dataset contains two csv files, one for training and one for testing. The training data originally has 891 observations and testing data has 418 observations. Both contain the following variables, “PassengerId”, “Pclass”, “Name”, “Sex”, “Age”, “SibSp”, “Parch”, “Ticket”, “Fare”, “Cabin” and “Embarked”. “PassengerId” is the passenger’s identification number starting at 1 in the training data and is used as the index for the data. “Pclass” is the passenger’s ticket class and can have the values 1, 2 or 3 which indicates either upper, middle or lower class. “Name” is the passenger’s name and their title. “Sex” is the passenger’s gender and can be either male or female. “Age” indicates the passenger’s age in years at the time of the RMS Titanic disaster. “SibSp” is the number of siblings and spouses that specific passenger has aboard the ship. “Parch” indicates the number of children or parents that specific passenger has aboard the ship. “Ticket” is the ticket number of the passenger. “Fare” is how much the passenger paid for their ticket. “Cabin” is the passenger’s cabin number. “Embarked” indicates which port the passenger embarked from and can contain the values C, Q or S which indicates that they embarked from either Cherbourg, Queenstown or Southampton. The training dataset also contains the “Survived” variable whereas the testing data does not. This variable is the response variable which shows if the person survived the RMS Titanic disaster or not. Any accuracy tests, therefore, need to be done by splitting the training data into training and testing so that the accuracy of the model that uses this data can be found based on the prediction and actual response labels. Survived is shown by the integer value 1 and not survived by 0 in the “Survived” variable. The first 10 samples of the training data can be seen in Figure 1 and testing data in Figure 2.

Preprocessing was applied to both datasets. They needed to be cleaned as they both contained empty values for some variables. The “Cabin” variable contains mostly empty values so was removed from both datasets. The “Ticket” and “Name” variables were also removed as they have no impact of whether or not a passenger survived as they contain only the ticket number and their name and “PassengerId” is still there to identify the passenger. This leaves the variables, “PassengerId”, “Survived”, “Pclass”, “Sex”, “Age”, “SibSp”, “Parch”, “Fare”, and “Embarked” for the training dataset and “PassengerId”, “Pclass”, “Sex”, “Age”, “SibSp”, “Parch”, “Fare”, and “Embarked” for the testing dataset. All observations that contained any missing values were also removed from both datasets to ensure that all three methods won’t have any issues with missing values. This leaves 712 training observations (about 68%) and 331 testing observations (about 32%). The cleaned training dataset can be seen in Figure 3 and the cleaned testing dataset can be seen in Figure 4. Once the data has been cleaned, normalisation was applied to both datasets to obtain the final format of the datasets using the formula $x = \frac{(x-\mu)}{(\sigma+1e-8)}$ where μ and σ is calculated for each variable/parameter. The validation data comes from the training data and has a split of 87.5/12.5 training to testing data as the validation contains eight folds.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

Fig. 1. Training Data Sample

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
897	3	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250	NaN	S
898	3	Connolly, Miss. Kate	female	30.0	0	0	330972	7.6292	NaN	Q
899	2	Caldwell, Mr. Albert Francis	male	26.0	1	1	248738	29.0000	NaN	S
900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.0	0	0	2657	7.2292	NaN	C
901	3	Davies, Mr. John Samuel	male	21.0	2	0	A/4 48871	24.1500	NaN	S

Fig. 2. Testing Data Sample

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
1	0	3	1	22.0	1	0	7.2500	3
2	1	1	2	38.0	1	0	71.2833	1
3	1	3	2	26.0	0	0	7.9250	3
4	1	1	2	35.0	1	0	53.1000	3
5	0	3	1	35.0	0	0	8.0500	3
7	0	1	1	54.0	0	0	51.8625	3
8	0	3	1	2.0	3	1	21.0750	3
9	1	3	2	27.0	0	2	11.1333	3
10	1	2	2	14.0	1	0	30.0708	1
11	1	3	2	4.0	1	1	16.7000	3

Fig. 3. Training Data Sample Cleaned

PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
892	3	1	34.5	0	0	7.8292	2
893	3	2	47.0	1	0	7.0000	3
894	2	1	62.0	0	0	9.6875	2
895	3	1	27.0	0	0	8.6625	3
896	3	2	22.0	1	1	12.2875	3
897	3	1	14.0	0	0	9.2250	3
898	3	2	30.0	0	0	7.6292	2
899	2	1	26.0	1	1	29.0000	3
900	3	2	18.0	0	0	7.2292	1
901	3	1	21.0	2	0	24.1500	3

Fig. 4. Testing Data Sample Cleaned

IV. Methods

To solve this machine learning problem, three different methods, namely logistic regression, perceptron and K nearest neighbour are implemented. From these three methods, we want to see which has the highest accuracy for the RMS Titanic passenger survival prediction.

For logistic regression, the LogisticRegression function from sklearn is used to create the model. The training data is then fit to the logistic regression model using the normalised training data and the target values from the "Survived" variable. This is done by using the "fit" method from sklearn. The data that is fit to the logistic regression model is then used to predict the classification of the test data using the normalised test data. This is done by using the "predict" method from sklearn. As stated in the previous section, the test data does not have any target values so cannot be used to find the accuracy of the model. The training data is instead used to find the accuracy using "score" method in sklearn. The validation of the data comes from the training data too and has a split of 87.5/12.5 training to testing data. Validation is done using the "cross_val_predict" method in sklearn. This validates the data by doing the prediction using logistic regression eight times with a different combination of training and testing data and obtaining an accuracy for each combination. The mean of these accuracies is then found to obtain the validation accuracy. The confusion matrix is found using the "confusion_matrix" method. The precision, recall and the f-score for this model are then found by using the "precision_score" and "recall_score" from sklearn and $2(\frac{Recall \times Precision}{Recall + Precision})$ for the f-score. The same steps are done for the perceptron and K nearest neighbours models but using the "Perceptron" method with max iterations of 20 and "KNeighborsClassifier" method with the number of neighbours being nine. To find the number of neighbours that should be used in the K nearest neighbours model, the elbow method is implemented by running the model with different K values from 1 to 28 and plotting the errors using 1- accuracy. It was found that the K nearest neighbours model has the highest accuracy for the RMS Titanic classification and so we implement our method for K nearest neighbours without using sklearn to create the model.

A. Own K Nearest Neighbour Model

A distance method was created to calculate the Euclidean distance between two different observations using the formula $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \dots}$ for each variable (column) in the observations. A neighbourhood method was created to find the distances between a single test datapoint and all training data and add these distances to a list. These distances are then sorted and the smallest nine distances are returned. A prediction method is created to predict whether the passenger survived or not by running the neighbourhood method to obtain the nine nearest neighbours and choose the classification that appears the most for the training datapoint prediction value. This method is then run for all testing data so that all the test passengers are either classified as survived or not survived. To find the accuracy of our model, we instead predict the training data, using the implemented prediction method, and count the number of datapoints where the predicted value matches the actual value. The accuracy is then $\frac{correct\ predictions}{total\ datapoints}$. Cross-validation is done by splitting the training data into eight folds and using the implemented prediction method on each fold. The confusion matrix is found by finding the true positive (actual = 1 and prediction = 1), false negative (actual = 1 and prediction = 0), true negative (actual = 0 and prediction = 0) and false positive (actual = 0 and prediction = 1). The confusion matrix values are then used to find the precision, recall and f-score using the formula mentioned in the "Related Work" section.

V. Experimental Results

The three different models mentioned above and our own K nearest neighbour method are all fit to the same training data for model learning. The training data is normalised and contains 712 observations. This data is split into a dataset of explanatory variables and a dataset of the response variable to train the models. Once the models have been trained, they are used to predict the response variable for the testing data which contains 331 observations. The perceptron model uses a max iterations value of 20 as anything less, such as 10, the maximum number of iterations is reached before convergence occurs which causes a large loss in the accuracy of the model and values larger than 20 do not change the model's accuracy. In the K nearest neighbour model and our model, nine neighbours are used as from the elbow method as seen in Figure 5, 14 is the best value to use for K as it has the smallest error however to avoid ties in prediction, an odd number is used so the next best value for K is selected which is nine. The Euclidean distance formula is used over other distance methods as it is simple, most commonly used and results in high accuracy for most binary classification problems as stated in [3]. The confusion matrix, accuracy, cross-validation, precision, recall and f-score are all calculated for each of the four different models.

Where 1 is survived (positive) and 0 is not survived (negative), True Positives are the number of outputs where the label of the data is 1 and the prediction is also 1. True Negatives are the number of outputs where the label of the data is 0 was and the prediction is also 0. False Positives are the number of outputs where the label of the data is 0 but was predicted to be 1. False Negatives are the number of outputs where the label of the data is 1 but was predicted to be 0. The confusion matrix stores the true positive, false positive, true negative and false negative values for a model. Accuracy is the ratio of the number of correctly predicted observations to the total number of observations. Cross-validation was done using eight folds and a split of 87.5/12.5 training to testing data. Eight was selected as it results in the highest accuracy for all the models after the cross-validation was fully performed. Precision is the ratio of correct positive predictions to the total predicted positive data observations. Recall the ratio of the number of correct positive predictions to the total number of observations that should be positive. F-score is the weighted average of precision and recall. The higher the percentage for accuracy, precision, recall, and f-score, the better the model performs.

The results from the three models that were created as well as our own K nearest neighbour model can be seen in Table I. The logistic regression model has a high accuracy which is only slightly lowered after performing cross-validation. Its confusion matrix is as follows $\begin{bmatrix} 363 & 61 \\ 86 & 202 \end{bmatrix}$. This shows that the number of true positives is 202, true negatives is 363, false positives is 61 and false negatives is 86. This leads to a fairly high precision value and a lower recall value which causes the f-score value to be relatively high. This shows that logistic regression is an adequate model to use for this binary classification problem due to the fairly high accuracy and fairly low misclassifications. A sample of the logistic regression prediction results can be seen in Figure 6.

The perceptron model has a low accuracy which is lowered further after performing cross-validation. Its confusion matrix is as follows $\begin{bmatrix} 291 & 133 \\ 75 & 213 \end{bmatrix}$. This shows that the number of true positives is 213, true negatives is 291, false positives is 133 and false negatives is 75. This creates a fairly low recall value and an even worse precision value which causes the f-score value to also be low. This shows that the perceptron is a poor model to use for this binary classification problem as it leads to many misclassifications. A sample of the perceptron prediction results can be seen in Figure 7.

The K nearest neighbour model has a high accuracy which is lowered by about 3% after performing cross-validation. Its confusion matrix is as follows $\begin{bmatrix} 372 & 52 \\ 82 & 206 \end{bmatrix}$. This shows that the number of true positives is 206, true negatives is 372, false positives is 52 and false negatives is 82. This leads to a high precision value but a lower recall value which results in the f-score value to be relatively high. This shows that the K nearest neighbour model is the best of the three models to use for this binary classification problem as it has the highest accuracy, cross-validation accuracy, precision, recall and f-score. We, therefore, create our model for the K nearest neighbour model. A sample of the K nearest neighbour prediction results can be seen in Figure 8.

It can be seen that our K nearest neighbour model slightly improves upon the K nearest neighbour model which uses sklearn methods to create the model. It has the highest accuracy which isn't lowered substantially after performing cross-validation. Its confusion matrix is as follows $\begin{bmatrix} 371 & 53 \\ 79 & 209 \end{bmatrix}$. This shows that the number of true positives is 209, true negatives is 371, false positives is 53 and false negatives is 79. This leads to the almost highest precision (0.07% lower than the KNN using sklearn) and the highest f-score values, however, the perceptron has a higher recall value. This shows that our K nearest neighbour model is the best model to use for this binary classification problem out of the 4 models that were tested. A sample of our K nearest neighbour prediction results can be seen in Figure 9.

We also find the percentage of people who survived the RMS Titanic sinking from the training data as well as the testing data based on the predictions using each of the three built-in methods and the own K nearest neighbour method using the formula, $\frac{survived\ people}{total\ people} \times 100$. Using a combined dataset

of the training data and the test data with the predictions from the own K nearest neighbour method, we also calculate the male survival percentage ($\frac{males\ survived}{total\ males} \times 100$), female survival percentage ($\frac{females\ survived}{total\ females} \times 100$), the average price paid for a ticket that survived, the average price paid for a ticket that did not survive, average age survived, the average age that did not survive, and the percentage survival of each class ($\frac{class\ 1\ survived}{total\ class\ 1} \times 100$). These results are shown in Table II. This shows how the model can be used in current times to help identify which passengers are more at risk in the case of an emergency. These results are similar to those that were found in [1] as stated in the "Related Work" section.

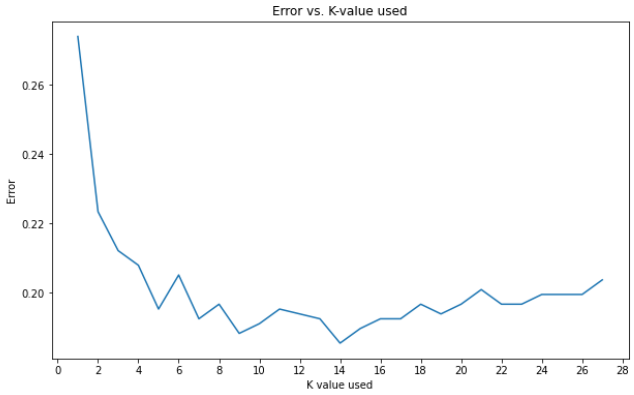

Fig. 5. Elbow Method

PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Survived
892	3	1	34.5	0	0	7.8292	2	No
893	3	2	47.0	1	0	7.0000	3	No
894	2	1	62.0	0	0	9.6875	2	No
895	3	1	27.0	0	0	8.6625	3	No
896	3	2	22.0	1	1	12.2875	3	Yes
897	3	1	14.0	0	0	9.2250	3	No
898	3	2	30.0	0	0	7.6292	2	Yes
899	2	1	26.0	1	1	29.0000	3	No
900	3	2	18.0	0	0	7.2292	1	Yes
901	3	1	21.0	2	0	24.1500	3	No

Fig. 6. Logistic Regression Prediction Results Sample

PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Survived
892	3	1	34.5	0	0	7.8292	2	No
893	3	2	47.0	1	0	7.0000	3	No
894	2	1	62.0	0	0	9.6875	2	No
895	3	1	27.0	0	0	8.6625	3	No
896	3	2	22.0	1	1	12.2875	3	No
897	3	1	14.0	0	0	9.2250	3	No
898	3	2	30.0	0	0	7.6292	2	Yes
899	2	1	26.0	1	1	29.0000	3	No
900	3	2	18.0	0	0	7.2292	1	Yes
901	3	1	21.0	2	0	24.1500	3	No

Fig. 7. Perceptron Prediction Results Sample

PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Survived
892	3	1	34.5	0	0	7.8292	2	No
893	3	2	47.0	1	0	7.0000	3	No
894	2	1	62.0	0	0	9.6875	2	No
895	3	1	27.0	0	0	8.6625	3	No
896	3	2	22.0	1	1	12.2875	3	No
897	3	1	14.0	0	0	9.2250	3	No
898	3	2	30.0	0	0	7.6292	2	Yes
899	2	1	26.0	1	1	29.0000	3	No
900	3	2	18.0	0	0	7.2292	1	Yes
901	3	1	21.0	2	0	24.1500	3	No

Fig. 8. K nearest neighbour Prediction Results Sample

PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Survived
892	3	1	34.5	0	0	7.8292	2	No
893	3	2	47.0	1	0	7.0000	3	No
894	2	1	62.0	0	0	9.6875	2	No
895	3	1	27.0	0	0	8.6625	3	No
896	3	2	22.0	1	1	12.2875	3	No
897	3	1	14.0	0	0	9.2250	3	No
898	3	2	30.0	0	0	7.6292	2	Yes
899	2	1	26.0	1	1	29.0000	3	No
900	3	2	18.0	0	0	7.2292	1	Yes
901	3	1	21.0	2	0	24.1500	3	No

Fig. 9. Own K nearest neighbour Prediction Results Sample

TABLE I
TABLE OF MAIN EXPERIMENT RESULTS

Model	***Accuracy***	***Coss-Val Accuracy***	***Precision***	***Recall***	***F-score***
Logistic Regression	80.20%	79.35%	76.81%	70.14%	73.32%
Perceptron	72.05%	70.79%	61.56%	73.96%	67.19%
KNN	84.41%	81.18%	79.84%	71.53%	75.46%
Own KNN	84.69%	81.46%	79.77%	72.57%	76.00%

TABLE II
TABLE OF ADDITIONAL EXPERIMENT RESULTS

Experiment	***Result Value***
Male percentage survived	18.11%
Female percentage survived	78.24%
Mean price paid for ticket that survived	55.80
Mean price paid for ticket that died	23.61
Mean age	30
Mean age survived	29
Mean age died	31
Ticket class 1 percentage that survived	66.31%
Ticket class 2 percentage that survived	43.68%
Ticket class 3 percentage that survived	24.00%

VI. CONCLUSION AND FUTURE WORK

The RMS Titanic disaster was one of the worst maritime disasters of the 20th century. Many rules and regulations were put in place to avoid any more devastating incidents. It is beneficial to investigate which passengers on the RMS Titanic were most ask risk of dying and which passengers had a high chance of survival as the same features in the training and testing data can be applied to passengers today. To do this, three binary classification models, namely logistic regression, perceptron and K nearest neighbour, were implemented to learn the training data and predict whether or not a passenger would have survived based on certain features. These models were then compared with each other by comparing their accuracy, cross-validation accuracy, precision, recall and f-score values. It was found that the K nearest neighbour model was best for this binary classification problem and so we created our own K nearest neighbour model. Our K nearest neighbour model slightly improved upon the K nearest neighbour model using built-in methods from sklearn. This is, therefore, the best model to use for this binary classification problem out of the four different models that were tested. This model can be used for data obtained today to find which passengers are more at risk in the case of an emergency as was done in the additional experiment results in Table II. Extra measures can be put in place for these "at-risk" passengers. Future work may include testing many more different binary classification methods to find the most accurate model from a larger group of models as well as using a combination of data from more recent ship sinkings.

REFERENCES

[1] B. Frey, B. Torgler, and D. Savage, "Behavior under Extreme Conditions: The Titanic Disaster" in Journal of Economic Perspectives, vol. 25, pp. 209–222, Jan 2011.

[2] M. Maalouf, "Logistic regression in data analysis: An overview" in International Journal of Data Analysis Techniques and Strategies, vol. 3, pp. 281–299, Jul 2011.

[3] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets" in SN Applied Sciences, vol. 1, Dec 2019.

[4] S. I. Gallant, "Perceptron-based learning algorithms" in IEEETransactions on Neural Networks, vol. 1, no. 2, pp.179-191, Jun 1990.

[5] Kaggle, "Titanic: Machine Learning from Disaster" dataset retrieved from Kaggle Online; accessed 16 Jun 2020 https://www.kaggle.com/c/titanic/data