# Data Preparation:

1. The csv is loaded into a Pandas data-frame.
2. The pandas built in describe method is used to get a brief insight into the data show below.

A screen shot of a computer

Description automatically generated

The count refers to the number of non null values in the dataset, from this we can deduce that all features except from ‘Age’ and ‘Cabin’ have no null values. Removal of the null values in age will be necessary for analysis. Due to the high null rate in the ‘Cabin’ feature it will be excluded from analyis.

The variation in min max values in the data points show that the data will need to be normalised for the model.

PassengerId will be removed from the dataset as

1. Focusing on the Survived column of the aformentioned Phik matrix shows that the most important categorical features for predicting survival are Sex, Passenger class, Fare, and Embarked so these will be used to fit the model.

|  |  |
| --- | --- |
| Feature | Corelation coefficient (0<x<1), 1 is more corelated |
| Sex\_female 0.742633 | |
| Sex\_male 0.742633 | |
| Pclass\_1 0.443398 | |
| Pclass\_2 0.117118 | |
| Pclass\_3 0.496300 | |
| Embarked\_C 0.291650 | |
| Embarked\_Q 0.028397 | |
| Embarked\_S 0.235120 | |

# Data visualization:

1. Using the KDE curve of the ages of survivors(green) and all passengers(blue) together it can be seen age has an effect on survial. The effect age has on the probability of surviving does not increase continuously over the ages (0-18 more, 20-40 less), because of this the ages will be one-hotted based on age ranges. Chosen ranges {0-14,14-31,31-45,45-max} (Chosen based on when the graphs cross, where the age is shown to effect the survival rate in the opposite direction). A graph of a function

   Description automatically generated with medium confidence
2. Using the KDE of the fares of non-survivors(red) and everyone else(blue) it can be shown that there is a strong link between death and paying a low fare. This link does not increase continuously over all ages and hence the Fares payed will be split into groups similarly to age. Chosen fare bins {0-30,30-100,100-max}A graph of a person with a blue and red line

   Description automatically generated with medium confidence
3. Comparing Parch and SibSp

# Data Classification:

1. The data is split up based on findings from the analysis. To recap:

* Split age and fare into bins.
* Unwanted columns removed.
* Other transformations were already done previously.

1. The outcomes are separated from the dataset.
2. The dataset is split randomly into training and testing subsets using 95% of the dataset for training.
3. A logistic model is fitted adding a bias column. Explanation of how it’s fitted:

* Random starting weights are chosen.
* Derivatives of likelihood function with respect to features is calculated.
* Using the derivatives along with the Adam gradient descent optimisation algorithm the weights are updated.
* The weights are continually updated using this method until either the mean of the absolute values of the gradients is below a threshold (0.0000001) or if non convergence criteria is met (>100,000 iterations)

1. With the model fitted it is tested on the testing set, the resulting accuracy is calculated by (total predictions correct)/(total predictions).
2. To have confidence in the accuracy reported from the model I ran the program three times which makes different training/testing splits. Here are the accuracy results:

* 80.55555555555556%
* 83.33333333333334%
* 94.44444444444444% (Anomalous so excluded in average)

The average accuracy over three runs is 81.9444444444%

Please see this excel spreadsheet below for the predictions the model made on the full dataset, including the final representations of the features for the model. For full project see the [GitHub repository](https://github.com/jonafinmicclen/Titanic-Dataset-Analyis)

