CS 4442 - Assignment 2 (Machine Learning)
Brandon Assing - 250787774
Kripal Patel - 250796697

1. Briefly discuss the problem and form your machine learning task
2. Get the dataset, describe the data structure with necessary output; visualize the data to discover more insights, apply at least one approach to analyze one perspective like the feature correlation
3. Prepare the data: apply at least one approach (data cleaning, or feature scaling etc.) to your data
4. Use these two models: **naive bayes classifier** and **Perceptron**, describe your evaluation metrics, train and fine-tune each model to get the promising one
5. Predict the test set, generate the required prediction format then submit to Kaggle, report your model's result in your assignment

Step 1
The Titanic is one of the most famous tragedies in human history. In this task, Titanic passenger survival data is used to build a machine learning model that estimates which passengers would survive the Titanic. The objective of the machine learning system will be to evaluate factors such as: passenger class, sex, age, number of siblings/spouses, number of parents/children, fare, and port embarked from to decide whether the specific passenger would survive the crash or not.

Step 2
The data was downloaded from the Kaggle site https://www.kaggle.com/c/titanic/data as CSV files and imported into the Jupyter notebook for analysis. Various steps taken (as seen in cells 3-5) were used to display the unaltered survivor data allowing us to determine what parameters could be useful in the data analytics. Certain columns were dropped (ticket and name) as they were deemed not useful in determining survival. In cell 13 we display survival in correlation to the port of embarkment. We did this as we were unsure as to whether the port had an impact on survival; however, we observed that the ratio of survived:not survived per port differed drastically, suggesting that port of embarkment has an effect on survival. This process was repeated with ticket class in cell 16. In cell 17 we displayed survival based on age to get further insight on what effect age might have, and in cell 22 and 23 this process was repeated for fare prices.

Step 3
First, in cell 11 we padded the null/empty values for port of embarkment with the most commonly occurring value (S for Southampton). Then cell 14 pads the null values of a age with a random age between (mean_age - age_standard_deviation) and (mean_age + age_standard_deviation). We originally tried to take the same approach as port of embarkment and pad the empty values with the most common age: 29; however, this resulted in a spike in that age

that did not mimic an accurate age distribution. This padding is verified in cell 15 to ensure there are no empty values. Lastly, the fare price is padded with a 0. After padding, the data that's inputted into the machine learning algorithm needs to be in the form of an integer. Cells 18-20 and 24 convert the data points age, port of embarkment, gender, and fare price to integers in preparation to train the machine learning model. Cell 25 then verifies that the data has all converted to integer values.

Step 4

Cell 26 is the last step in data preparation. It specifies *x* values as all the table values apart from survival and *y* as survived or not. Cell 27 uses a Gaussian Naive Bayes model for survival prediction with a value of 0.0001 for smoothing (var_smoothing), and cell 28 uses a Perceptron model with 2000 iterations (max_iter) and a stopping criteria (tol) of 0.00001. Different values were tried for var_smoothing, max_iter, and tol until a sufficient prediction accuracy was reached.

Step 5

Cells 27 and 28 also run the data through the different models in order to predict the survival of the passengers in the test data set. Cell 29 prints the prediction scores of the models and then saves the Perceptron results to a CSV file in order to submit to Kaggle. We chose the Perceptron model since it received a higher accuracy rate.