**Tanner Young**

**Midterm**

**Question 1A - Screen shot of data set initial lines**

A screenshot of a computer

Description automatically generated

**Question 1B - Discuss distributions of features.**

* Passenger ID is distributed “symmetric”.
* Age is distributed “normal”.
* Fare is distributed “skewed right”.
* SibSp is distributed “skewed right”.
* Parch is distributed “skewed right”.
* Survived is distributed “bimodal”.

**Question 1C - Discussion Kaggle feature selections.**

Age is an important feature because it is likely to be a strong predictor of survival on the Titanic. It's reasonable to assume that children and elderly passengers had a higher likelihood of being prioritized for lifeboats, and this can be an important factor in predicting survival.

The cabin feature might be used because the location of a passenger's cabin on the ship could be indicative of their proximity to lifeboats, exits, or other critical areas. It could provide valuable information about survival chances.

The features boat, body, and home.dest were likely discarded because they are unrelated to survival. They were collected after knowing the outcome of survival.

**Question 2A - Desired feature and story**

One feature that could potentially improve predictions for survival on the Titanic is a "Cabin Deck" feature, which represents the deck level on which a passenger's cabin was located.

Story:

Passengers with cabins located on higher decks (e.g., A, B, C) may have had better access to lifeboats and emergency evacuation routes, which could increase their chances of survival.

Passengers with cabins located on lower decks (e.g., F, G) may have been farther from lifeboats and had more obstacles to overcome, potentially reducing their chances of survival.

**Question 2B - Screen shot age vs survival.**

A screenshot of a computer

Description automatically generated

**Question 2C - Discussion of your two features.**

Age and Fare seem to be two very important features to the survival.

**Question 3A - Screen shot data cleaning code**

A screenshot of a computer

Description automatically generated

**Question 3B - Screen shot model training**

A screenshot of a computer program

Description automatically generated

**Question 3C - Screen shot confusion matrix and metrics.**

A screenshot of a computer screen

Description automatically generated

**Question 3D - Discussion of performance on training set.**

The model has an accuracy of 69%, which means it correctly predicts survival or non-survival in about 69% of the cases. Precision for both classes (survived and not survived) is reasonable, indicating that the model makes accurate positive predictions for both classes. However, recall for the "Survived" class is relatively low (41%). This means the model is less effective at identifying actual survivors. The F1-score for survivors is also lower (0.52), indicating that there is room for improvement in balancing precision and recall for survivors.