**Titanic Data Machine Learning Writeup**

**Phivian Phun**

Given Titantic Data from Kaggle, the goal is to build an optimal predictive model that can determine whether a passenger will survive or not. This is a binary classification problem and there are well-known models that can solve this problem.

Before preprocessing the data, data visualization is needed for better understanding. I noticed that females in young adulthood were more likely to survive and males in young adulthood were more likely to die. There were also significant survival differences between classes and where they were embarked. We can assume that higher class, younger women were most likely the survivors.

Feature extraction was based on a tutorial provided by Kaggle. Categorical columns were transformed into multiple binary columns. Missing categorical values were replaced with ‘Unknown’ type. Missing continuous values like age were imputed with the median value.

The pipeline was used for imputation, standard scaling, hyperparameter tuning, and model evaluation. This is great practice for scoring your models so that you are not leaking information in every cross-validation loop. It is best practice to have all preprocessing steps occur internally with cross-validation.

Based on the scores, Random Forest performed the best with Logistic Regression as second best. This is not surprising because Logistic Regression is a simple model for binary classification problems. The data was completely transformed into binary values for all categorical features. Random Forests is the most robust and computationally heaviest compared to other models. It is known for feature selection and ensemble learning where it averages results from randomized decision trees.

Based on the ROC curve, both Logistic Regression and Random Forests are pretty similar in their performance. With more research and understanding of the Random Forests, there can be improvements to beat the Logistic Regression. Prepruning and feature expansion might help with the model size. For example, I noticed that passengers’ nobility, military, and religious statuses can be extracted from their honorifics. If there are more features added to the model, Random Forest would emphasize feature selection to determine how much to weigh certain features in the model. Also, coefficients of Logistic Regression can be analyzed to determine which feature has the most and least impact on the model.