INFO 523 – Final Project

Link to Git repository where presentation and R-file are stored: <https://github.com/rayell86/Final-Project>

Link to data: <https://www.kaggle.com/datasets/vinicius150987/titanic3?select=titanic3.xls>

**Predicting the Probability of Survival on Titanic**

Titanic data has been around for a long time and has always picked the interest of data miner enthusiasts. This dataset has been selected for this analysis due to its accessibility and simplicity. The data was analyzed and picked apart carefully to gain a better understanding of the variables in the data. Next, the dataset was meticulously wrangled, and anomalies were excluded from the data, however, the integrity and consistency of the data was preserved. A logistic regression model as well as Random Forest and Decision Tree were used to train the data and identify the most relevant variables that impacted the probability of survival of the passengers that boarded Titanic. The findings of this analysis suggested that age, passengers’ ticket class and gender were among the most impactful variables that shaped the probability of survival on Titanic. The findings suggest that male passengers were more likely to survive the impact, if they were in a higher ticket class. Another trend that appeared in the findings was that passengers who were accompanied by a parent, sibling or a child had a higher chance of survival. This phenomenon could be directly related to passengers’ ticket class, for wealthier passengers tended to purchase more expensive tickets and were traveling with families, whereas the passengers in the 3rd class were mostly travelling alone. As for the limitations of this study, one could argue that the method of designating missing age could have been performed more sophisticatedly. The missing passenger age was filled based on the average age of the other passengers in the same class. However, this method does not include the passengers’ gender which is also an important factor. In addition, the number of significant variables that were included in the analysis were limited and this could potentially impact the data training step.