Isha Singh

Professor Irene Tsapara

MSDS 422 Practical Machine Learning

2 February 2025

Module 04: Titanic

In 1912, the Titanic shipwreck took the innocent lives of about 1,502 of the 2,224 passengers. The following analysis used Kaggle data to determine survival rates: age, gender, ticket price, and class. The three important machine learning models used in the research were K-Nearest Neighbors (KNN), Logistic Regression, and Linear Discriminant Analysis (LDA). These models were useful tools as they helped to identify trends in the data and make predictions.

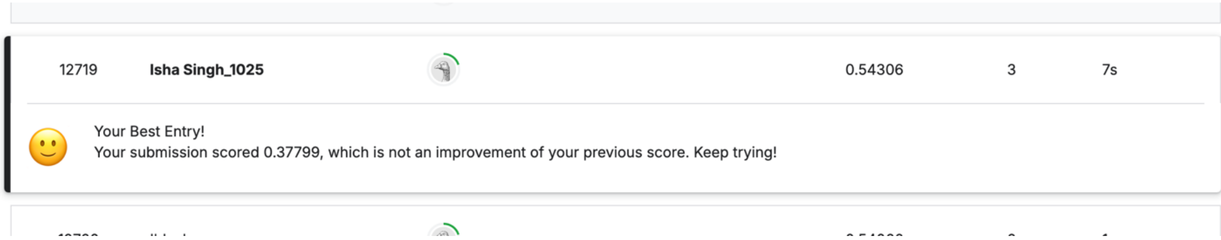
Prior to training the models, the data needed to be cleaned and improved so that it could be used for modeling. To reduce bias, missing ages were replaced with medians, and missing embarkation points were replaced with mode values. As we noticed from previous modules, machine learning models work more appropriately when the data is entirely numerical. This is why categorical data, such as gender and embarkation column, was converted to numerical data using One-Hot Encoding, which created new columns based on numerical values. With all this, we were able to understand everything in a better manner, and outliers were identified with the help of the IQR method to determine the extreme values in the data. Since KNN is dependent on distance calculations, numerical features were standardized using StandardScaler, which balances all values to be on the same level.

Each of the models was required to have assumptions done prior to the model formation. Firstly, KNN does not need any assumptions to be made as it is a non-parametric model, and it was ensured by standardization. Logistic regression required to be checked by the Variance Inflation Factor (VIF) test and showed no strong relationships, and tests revealed that some family-related features were very closely linked. LDA's requirement was homogeneity of variance and normality and worked best when the data is dependent on a normal distribution. However, Titanic's features did not follow that specific ordinary pattern.

The models were evaluated for each of their strengths: accuracy (how often they were correct), precision (how many predicted survivors actually survived), recall (how many actual survivors were correctly identified), and last but not least, F1-score (a balance between precision and recall). Overall, after testing all of them, we noticed that KNN performed in a sufficiently good way with an accuracy of 80.69 percent, a precision of 72.26 percent, a recall of 64.98 percent, and an F1-score of 66.98. Logistic Regression and LDA had accuracies of 78.74 and 80.05 percent, respectively.

The results overall showed that the factor of gender had the most significant value, where women were more likely to survive. Additionally, first-class passengers had higher survival rates. Higher ticket prices indicated better survival chances. Traveling alone reduced the likelihood of survival, suggesting that families may have helped each other during the evacuation scenario. To conclude, KNN performed better than the other models, but Logistic Regression and LDA would still be considered useful.

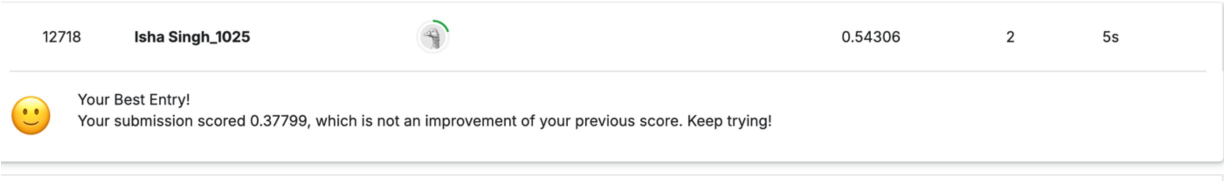
LDA



12719 **Isha Singh_1025** 0.54306 3 7s

🙂 Your Best Entry!
Your submission scored 0.37799, which is not an improvement of your previous score. Keep trying!

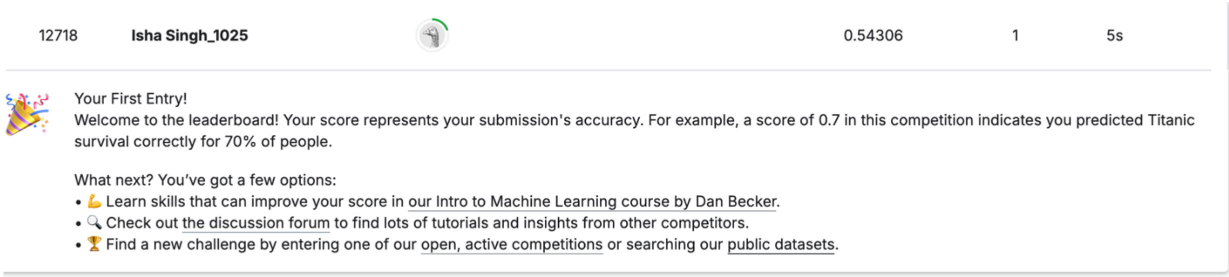
LOGISTIC REGRESSION



12718 **Isha Singh_1025** 0.54306 2 5s

🙂 Your Best Entry!
Your submission scored 0.37799, which is not an improvement of your previous score. Keep trying!

KNN



12718 **Isha Singh_1025** 0.54306 1 5s

🎉 Your First Entry!
Welcome to the leaderboard! Your score represents your submission's accuracy. For example, a score of 0.7 in this competition indicates you predicted Titanic survival correctly for 70% of people.

What next? You've got a few options:
- 💪 Learn skills that can improve your score in our Intro to Machine Learning course by Dan Becker.
- 🔍 Check out the discussion forum to find lots of tutorials and insights from other competitors.
- 🏆 Find a new challenge by entering one of our open, active competitions or searching our public datasets.