Project 2 - Soldier_Race Report

Comparing the four models (Logistic Regression, SVM, Random Forest, and XGBoost) based on their evaluation scores, we can make the following observations:

Logistic regression model: The performance scores for this model were the best compared to other models we used, it performed will even before hyperparameter tuning, and it was not prone to overfit compared to RF and Xgboost. Logistic regression has relatively high accuracy and F1-scores, indicating good performance in classifying the different groups (Black, Hispanic, White). We managed to improve the scores for the Hispanic class after using over_sampling and under_sampling, but not very significantly.

SVM model: looking at the graphs we did we see SVM was the second-best model after logistic regression, its scores were very close. F1, recall and precision were good. the SVM model performs well on both the training and test sets, with relatively high accuracy and F1-scores, indicating good performance in classifying the different groups (Black, Hispanic, White). The results are consistent with those of the Logistic Regression model.

Random Forest model: From the performance graph, we see that RF performed poorly compared to logistic regression, RF was prone to overfitting, the train results were perfect while the test results were the opposite. It was expected to overfit due to the complexity of Random Forest, and another reason is the noise in the data (features that would not contribute very will to the Hispanic class for example).

Xgboost Model: XGBoost performs decently but has some challenges with classifying a specific class.

In conclusion, after training and testing different models (Logistic regression, SVM, Random Forest and Xgboost), the best results we were able to get were in the logistic regression. So, we chose logistic regression as our final model.