Introduction This brief report outlines the research steps taken to analyze and predict the income status of US citizens based on various other variables such as gender, age, race, marital status, country of origin etc. The target variable was operationalized as binary: greater than $50,000 & less than or equal to $50,000. The initial dataset had 48842 individual records.

Data Preparation There were only missing values in 3 distinct variables: occupation, country of origin, and workclass (private, government, never worked etc.). Because handling missing values in string format is rather difficult and there were not many missing values (2799 in workclass, 2809 in occupation, 857 in country of origin), all rows with at least one missing value were dropped. This shrunk the dataset from 48842 individual records to 45222. Additionally, all unique values in each variable in the string format were assigned an arbitrary number to denote nominality among each variable itself using LabelEncoder to prevent any issues in the analysis procedure.

K-Nearest Neighbor Model The first model that was fitted was the k-nearest neighbor model. The optimization procedure found 14 to be the best number of neighbors according to the accuracy score obtained in the test phase after training the model. The macro average score of the f1-score[1] was .78. The mean accuracy score of the test trials in the cross-validation procedure was .849 (σ = .006).

Decision Tree Model The second model was a decision tree model. The optimization procedure found 14 to be the best depth.[2] The macro average score of the f1-score was .78. The mean accuracy score of the test trials in the cross-validation procedure was .854 (σ = .006).

Logistic Regression Model The final model was a logistic regression model. Penalizers (Lasso and Ridge) did not seem to change the model much. AUROC score was .816. The macro average score of the f1-score was .69. The mean accuracy score of the test trials in the cross-validation procedure was .801 (σ = .006).

Conclusion Among the three models, logistic regression was worse both in macro average of f1-score and accuracy. The k-nearest neighbor and decision tree model had the exact same macro average of f1-score but the decision tree model had a slightly larger accuracy percentage, $t(38) = 2.6352$, $p = .0121$ (2-tailed). Although the t-test yields a significant p-value with an alpha level of .95, an 84.9% and 85.4% accuracy rate is not necessarily meaningfully different in an intuitive sense. So, the guiding principle here should concern the model cost and interpretability. If there is no meaningful difference between those criteria as well, then in light of the t-test, the decision tree model should be preferred.

---

[1] Macro average of f1-score was chosen as the only statistical test output to be reported due to the limited scope of this report. F1-score includes both recall and precision for each class while macro average takes the unweighted mean of each f1-score.

[2] Other criteria are omitted.