Assignment 6

For this assignment, I worked with the Adult Income dataset from the UCI ML Repo in order to predict whether or not an individual would earn more than $50k. I first looked over the dataset and found that missing values were coded as "?" instead of NaN. I replaced "?" with np.nan, dropped the missing rows, and reset the index. After that, I removed unnecessary or replicated columns (e.g., fnlwgt and education, as educational-num already provides numeric information) and encoded the target variable income as binary (1 if >50K and 0 otherwise). To make them comparable during training, I encoded categorical columns into numeric values while standardizing numerical features.

My first model implementation was Elastic-Net Logistic Regression. I used GridSearchCV with 5-fold cross-validation to tune the hyperparameters by searching over seven log-spaced values of C (from 10^{-4} to 100) and different l1_ratio values. C = 0.1 and l1_ratio = 0.8 was the best configuration, producing a mean CV accuracy of 0.8211. Using an 80/20 train–test split, I found that the model had a test accuracy of 0.824, but a lower F1 score of 0.557. This was because the model was unable to detect the >50K class (class 1 recall = 0.45). In cases of class imbalance, accuracy can be misleading, so I focused more on the F1 score.

Next, I trained a Random Forest because it can capture non-linear patterns and feature interactions better than a linear model. I tuned its hyperparameters with another GridSearchCV (5-fold CV) over n_estimators, max_depth, min_samples_split, and min_samples_leaf. The best model was 200 trees, max_depth = 20, min_samples_split = 5, min_samples_leaf = 2, with mean CV accuracy = 0.8624. On the test split, it achieved accuracy = 0.9002 and F1 = 0.7719, showing a strong improvement in class 1 performance (precision = 0.87, recall = 0.69). Additional cross-validation results were consistent (CV std ≈ 0.0035).

Finally, I tested ANN (MLPClassifier) models with 5, 6, and 7 hidden layers using 5-fold cross-validation. Because training was time-consuming, I compared a small set of architectures and found the 6-layer network performed best among them (mean CV accuracy ≈ 0.8443). I also evaluated it with F1 scoring, obtaining a mean CV F1 ≈ 0.6619.

As summary, Elastic-Net logistic regression generated 0.824 test accuracy, but a relatively low F1 score of 0.557, as it had difficulty detecting the >50K class. The most successful model was the tuned Random Forest, which achieved 0.900 test accuracy and a much higher F1 score of 0.772, demonstrating a significant improvement in class-1 performance. The six-layer ANN achieved mean cross-validation (CV) accuracy of approximately 0.844 and mean CV F1 of approximately 0.662. This was better than logistic regression in terms of F1, but still below the Random Forest results.

In conclusion, I obtained the best model for this data in Random Forest format.