Data Preparation

I started by loading the dataset from a CSV file and checking its shape and structure. Duplicate rows were removed. I then replaced certain characters that I identified to be placeholder/empty values with NaN values, and replaced them with the mode of each column. I generated plots of the data for every column to review the distribution of values. Finally, I split the data into predictors and the target variable, and then encoded the categorical features using one-hot encoding (with the first category dropped).

INSIGHTS FROM DATA PREPARATION

I found numerous insufficiencies in the original data. Duplicates existed in the raw data and were removed. Missing values were present in some columns and were replaced with the mode. The distribution visualizations provided a clear view of how values were spread across features, which helped me make sure that there were no significant imbalances in the data.

Model Training Procedure

First, I split the encoded data into training and test sets using stratified sampling to maintain the same class proportions in both sets. I built the KNN classifier by training it with default parameters and then trained the second KNN model with the GridSearchCV approach, testing a range of neighbor values (from 1 to 10) with 5-fold cross-validation to find the best parameter. Lastly, a Logistic Regression model was trained with 1,000 maximum iterations. For each model, predictions were made on the test set and classification reports were generated to evaluate performance.

MODEL PERFORMANCE

Both the KNN and Logistic Regression models were evaluated using standard classification metrics generated by the classification_report function from sklearn.metrics. The initial KNN model and its GridSearchCV-tuned version provided data that supported their ability to correctly predict the classes. Similarly, the Logistic Regression model's performance was assessed to have high predictive capabilities. Overall, the performance reports suggest that the models were effective in distinguishing between the classes in the dataset.

CONFIDENCE IN THE MODEL

The model evaluation through cross-validation and hyperparameter tuning shows that all the models generalize well to unseen data. However, there is still room for improvement and the model could perhaps perform better with more complexity or higher n_neighbors.