→**Exploratory Data Analysis:** The code begins with exploratory data analysis. It includes visualizations of categorical variables distributions, box plots for numerical variables, and a heatmap for correlation analysis.

→**Data Preprocessing:** Next, the code is based on data preprocessing. This involves converting categorical variables into numerical using label encoding.

→**Data Splitting:** The data is split into training and testing sets using the "train_test_split" function. This allows you to evaluate the model's performance on unseen data.

→Outliers in numerical columns are handled by removing them based on the IQR (Interquartile Range) method. This can help improve the model's robustness to outliers.

→The Decision Tree model is trained on the preprocessed data. The accuracy score for this model on the test data is approximately 71.4%. The F1 score, precision, recall, and Jaccard score are also calculated.

→SHAP values are computed to explain the output of the Decision Tree model. The summary plot illustrates the impact of individual features on model predictions. This is valuable for understanding why the model makes certain predictions.

→The confusion matrix is plotted to show the distribution of actual and predicted labels. This helps in assessing the model's performance across different classes.

→**ROC Curve and AUC:** The ROC curve is plotted to assess the model's trade-off between true positive rate and false positive rate. The AUC score of 0.7510 indicates the model's ability to distinguish between classes.

→**Random Forest Classifier:** The Random Forest model is trained and evaluated. It achieves a higher accuracy score of approximately 91.9%, outperforming the Decision Tree model.

→**Multinomial Logistic Regression:** Multinomial Logistic Regression is applied, resulting in an accuracy score of 91.9%. This approach is suitable for this dataset due to its multi-class classification nature.

→**Cross-Validation:** Cross-validation is performed on the Decision Tree and Random Forest models. The Decision Tree achieves a mean accuracy score of around 71.2%, and the Random Forest achieves a mean accuracy score of approximately 91.4%. This indicates that the Random Forest model's performance is relatively consistent across different folds.

→**Hyperparameter Tuning:** Grid search is used to find the best hyperparameters for the Decision Tree and Random Forest models. The Best Decision Tree model achieves an accuracy score of approximately 91.9%.

→**XGBoost and SVM:** XGBoost and SVM models are trained and evaluated. Both achieved a high accuracy scores of approximately 91.9%.

Conclusion:

→In this project, exploratory data analysis, data preprocessing, modeling with Decision Trees, Random Forests, Logistic Regression, XGBoost, and SVM, and evaluated the models using various metrics, including accuracy, F1 score, precision, recall, and AUC are performed.

→Further, visualized feature importance and used SHAP plots for interpretation.

→Overall, the Random Forest and XGBoost models provided the highest accuracy on the test data.