Task 6.2

Michael Christopher – s224830467

1a. missing values

- Categorical features -> imputed with mode
- Continuous features -> imputed with median

1b. Split

- Stratified train/test split with ratio 80/20 to preserve class proportions

1c. Feature types

- Continuous features: identified by numeric dtype not originating from encoded categorical
- Categorical features: original object/string columns

1d. Encoding

- Label encoding categorical features to continuous and map target to 0, 1, and 2
- Standardization applied to features

1e. Label distribution

- Imbalance

Q2. Modeling & Validation

Models trained on preprocessed train data:
• Logistic Regression (multinomial)
• Random Forest Classifier
• XGBoost (or Gradient Boosting fallback if XGBoost unavailable)

Validation method & metric (Q2a):
• Stratified 5-fold cross-validation to keep class ratios consistent.
• Scoring: weighted-F1 (balances per-class F1 by support, suitable for imbalanced multiclass targets) and accuracy as a secondary check.

Under/overfitting diagnostic:
• Compare mean CV F1 vs. test F1. Large positive gaps (CV≫Test) → overfit. Large negative gaps (Test≫CV) → underfit. Small gap → well-aligned.

Q2b. Design Decisions (Justified)

• Logistic Regression (LR):
  – Rationale: Simple, fast baseline; interpretable coefficients (log-odds); detects approximately linear signal; regularization controls variance.
  – Trade-offs: Limited with strong non-linear interactions unless engineered; may underperform if boundary is highly non-linear.

• Random Forest (RF):
  – Rationale: Non-parametric, captures non-linearities & interactions via many randomized trees; bagging reduces variance; handles mixed scales.
  – Trade-offs: Less interpretable than LR; can overfit if trees too deep; larger models cost more compute.

• XGBoost (XGB):
  – Rationale: Boosting with shrinkage & column/row subsampling; strong accuracy with built-in regularization; robust on tabular data.
  – Trade-offs: More hyperparameters; risk of overfitting if not tuned; compute cost higher than LR/RF.

Q2c. Hyperparameter Optimization

• LR grid: solver ∈ {lbfgs, saga}, C ∈ {0.01, 0.1, 1.0, 10.0}, penalty ∈ {l1, l2}
  – Why: Controls strength/type of regularization and solver suitability for multinomial problems; balances bias–variance and sparsity.

• RF grid: n_estimators ∈ {200, 400}, max_depth ∈ {None, 8, 12}, min_samples_split ∈ {2, 5}, min_samples_leaf ∈ {1, 2}, max_features ∈ {sqrt, log2, None}
  – Why: Depth/leaves control overfitting; number of trees reduces variance; feature subsampling trades bias for variance reduction.

• XGB grid: n_estimators ∈ {300, 500}, learning_rate ∈ {0.05, 0.1}, max_depth ∈ {3, 5, 7}, subsample ∈ {0.8, 1.0}, colsample_bytree ∈ {0.8, 1.0}
  – Why: Learning rate & trees set capacity; depth controls complexity; subsampling & colsample regularize and combat overfitting.

• Selection criterion: Best CV (Stratified 5-fold) weighted-F1.

Q2d. Imbalance Handling & Improvement Check

- Applied SMOTE on the training set to synthetically oversample minority classes; evaluated models again (same CV/Test protocol).
- Compare pre/post-SMOTE weighted-F1 on the same test set to judge improvement. [Insert your pre/post table]

Q2e. Recommendation

- Recommend the model/stage (Tuned vs Tuned+SMOTE) with the highest test weighted-F1 (with accuracy/recall as tie-breakers if needed).

Q3. Final Prediction on Preprocessed Test Set

- Train the best model on the chosen training data (with/without SMOTE), then predict on X_test.
- Report metrics:
  – Accuracy, weighted-F1, weighted-Precision, weighted-Recall
  – Multiclass ROC-AUC (OVR) if probabilities available
  – Confusion matrix and full classification report

Q4. Feature Importance – Two Approaches & Statistical Reasons

Approach A – Model-based importance:
- Trees: mean impurity reduction (Gini/gain) approximates each feature's contribution to predictive splits.
- Logistic Regression: |coefficients| reflect effect sizes on log-odds scale.
- Statistical reasoning: Larger absolute effects or impurity reductions imply features that more strongly separate target classes within the model class.

Approach B – Permutation importance (held-out test):
- Measures drop in weighted-F1 when a feature is permuted (breaks its association), providing a model-agnostic estimate of marginal utility.
- Statistical reasoning: If shuffling a feature reduces performance, the feature carries information about the target beyond noise.

Consistency & caveats:
- Agreement across methods ⇒ robust signal; discrepancies can arise from collinearity, interactions, or linear vs non-linear model forms.