Machine Learning: Methodology Report

1. Data Preprocessing & Feature Engineering

The dataset contained 3,238 features and 315 samples, with a binary target variable (`CLASS`). Given the high feature-to-sample ratio (>10), careful preprocessing and dimensionality reduction were essential to avoid overfitting and improve generalisation.

Steps performed:

- **Missing & Infinite Value Handling:** Replaced infinite values with NaNs, dropped columns with >10% missing, and imputed remaining with column means.
- **Low Variance Filtering:** Removed zero-variance and near-zero variance features (<0.01 variance), which accounted for over 28% of features.
- **Feature Scaling:** Applied both `StandardScaler` and `RobustScaler` (for sensitivity analysis) depending on the model's needs.
- **PCA:** Reduced dimensionality to 50 components, explaining 95% of variance. PCA was especially useful for linear models (Logistic Regression, SVM, Naive Bayes).
- **Model-Based Feature Selection:** Employed L1-regularized Logistic Regression to select the most informative features.
- **Threshold Tuning:** Used F1-score as the guiding metric to identify optimal decision thresholds post-prediction.

2. Model Architectures and Hyperparameters

a. Logistic Regression

class_weight='balanced', random_state=42, C=0.01, penalty='l1', solver='liblinear', max_iter=1000

b. Random Forest

class_weight='balanced', n_estimators=200, random_state=42, max_depth=None, min_samples_leaf=1, min_samples_split=2

c. XGBoost

eval_metric='logloss', random_state=42, learning_rate=0.1, n_estimators=200, max_depth=5, subsample=0.8, colsample_bytree=0.8, use_label_encoder=False, verbosity=0

d. SVM (Linear Kernel)

kernel='linear', class_weight='balanced', probability=True, random_state=42, C=0.1

Note: Chatgpt was used during experimentation and documentation

e. LightGBM

class_weight='balanced', random_state=42, learning_rate=0.1, max_depth=5, n_estimators=200, verbosity=-1

All models were evaluated using 5-fold stratified cross-validation and a test set.

3. Results Table (Test Set Performance)

Model	Accuracy	F1	Recall	Specificity	AUROC
SVM	0.6349	0.5840	0.6447	0.6286	0.6501
LogisticRegression	0.5700	0.5057	0.5238	0.6034	0.6043
RandomForest	0.6600	0.5143	0.4286	0.8276	0.7178
LightGBM	0.6700	0.5217	0.4286	0.8448	0.6539
XGBoost	0.6300	0.5316	0.5000	0.7241	0.6548

4. Discussion & Recommendations

Best Overall Model: SVM achieved the highest F1-score (0.5840) and accuracy (0.6349), with balanced sensitivity and specificity. Its ability to generalize to high-dimensional data made it ideal for this task.

Logistic Regression: Logistic Regression was a required baseline and performed reliably across metrics. While its overall accuracy (57%) was lower than SVM or LightGBM, it achieved a respectable F1-score of 0.5057 and maintained balanced sensitivity (0.5238) and specificity (0.6034). These metrics demonstrate that Logistic Regression effectively handles both classes under moderate imbalance and provides a robust baseline for comparison with more advanced models.

Observations:

- Tree-based models (Random Forest, XGBoost, LightGBM) achieved strong AUROC and specificity, but often sacrificed recall.
- Naive Bayes performed surprisingly well given its simplicity and assumptions.
- Threshold tuning based on F1-score consistently improved classification performance.

Improvements with More Time:

- Ensemble methods (e.g., voting or stacking)
- Advanced calibration techniques (Platt scaling, isotonic regression)
- Additional feature selection strategies (e.g., feature clustering, deep learning embeddings)