A Brief Methodology Report

1. Data Preprocessing Steps

1.1 Loaded Datasets: Train, Test, and Blinded data.

1.2 Missing Value Handling:

- Identified columns with nulls.
- **Low-variance columns** (≤3 unique values, negligible correlation with 'CLASS') were dropped.
- **High-variance columns** (~50% null, wide range) were retained.
- Imputed null values using **median grouped by 'CLASS'**.

1.3 Outlier Handling

- Capped extreme values at the **99th percentile**.
- ~5% outliers remained post-processing — negligible, hence not removed.

1.4 Feature Selection & Dimensionality Reduction

- Applied Mutual Information (MI) — dropped 1372 features with **MI = 0.0**.
- Used PCA (**98% variance**) to reduce features from 3120 to 91.

1.5 Class Imbalance

- Class 0: 191 samples | Class 1: 124 samples.
- Tried SMOTE, but performance dropped — not used.

2. Model Development and Evaluation

2.1 Logistic Regression

- **Parameters**: solver='saga', max_iter=1000, class_weight='balanced', Grid search on 'C': [0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2']
- **Best Params**: C=0.1, penalty='l1'
- **Cross-Validated AUROC**: 0.65
- **Threshold**: 0.5
- **Test Results**: **Accuracy**: 0.67, **AUROC**: 0.71, **Recall**: 0.73, **Specificity**: 0.62, **F1 Score**: 0.65

2.2 Random Forest Classification

- **Parameters:** class_weight='balanced', random_state=42
- **Random search on:** n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap
- **Best Params:** n_estimators=100, max_depth=5, min_samples_split=7, min_samples_leaf=1, max_features='sqrt', bootstrap=True
- **Cross-Validated AUROC:** 0.6457

- **Threshold:** 0.45
- **Test Results: Accuracy**: 0.63, **AUROC**: 0.662, **Recall**: 0.69, **Specificity**: 0.586, **F1 Score**: 0.61

2.3 Support Vector Machine (SVM)

- **Parameters:** probability=True, random_state=42, class_weight='balanced'
- **Grid search on:** C, kernel, gamma
- **Best Params**: C=1, kernel='linear', gamma='scale'
- **Cross-Validated AUROC**: 0.6709
- **Threshold:** 0.4
- **Test Results: Accuracy**: 0.60**, AUROC**: 0.7118**, Recall**: 0.762**, Specificity**: 0.483**, F1 Score**: 0.615

2.4 Stacking Classifier

- Combined Logistic Regression, Random Forest, and SVM.
- Logistic Regression used as final estimator.
- **Threshold**: 0.4
- **Test Results: Accuracy:** 0.67**, AUROC:** 0.7101**, Recall:** 0.762**, Specificity:** 0.603**, F1 Score:** 0.66

3. Model Strengths

- Consistent **AUROC ~0.71** across models - indicates good class separability.
- High recall/sensitivity - models detect positive class effectively **(important in medical/critical tasks**).
- Stacking classifier improved overall balance in metrics.
- Robustness across algorithms shows reproducibility of results.

4. Model Limitations

- Moderate accuracy and specificity - some difficulty in identifying the negative class.
- PCA and MI may have discarded non-linear interactions or important original features.
- SVM specificity was particularly low, despite high recall**.**

5. Improvement can be done

- Apply non-linear transformations to important features before PCA.
- Experiment with alternative resampling methods beyond SMOTE
- Add more labeled data to improve generalization.
- Exploring more granular feature engineering
- Experiment with other classification models