1. # Data preprocessing and feature engineering steps

❖ Preprocessing

- First load all available data such as train, test and Blinded test dataset and Separate features and target (CLASS) for training set and testing data.

❖ **Feature engineering steps**.

- Feature engineering make a model smarter preparing and improving data

a) Drop non-informative columns (ID) – Drop the ID column which is not useful for modeling.
b) Check & handle missing values –check for missing or infinite values in both train and test dataset and replace infinite value and fill NaNs(Not a number)if the value is missing.
c) Scale numeric features using StandardScaler
d) Feature selection- select those feature that are important .We use SelectKBest method to reduce dimensionality and keep only the most relevant features

2. # Model architectures / key hyper-parameters

- We used 3 models and they are logistic regression, Random Forest and Support Vector machine (SVM)
- We are using **GridSearchCV** for tuning model to find the best hyperparameters for these 3 models.
- Grid Search CV is a tool that helps to find the best combination of model settings (called hyperparameters) to make our model perform its best.
- We used GridSearchCV instead of RandomizedSearchCV because it gives more accurate value whereas RandomizedSearchCV might miss some combination.

- **Cross-validation** involves splitting the dataset into multiple parts (folds), training the model on some parts, and testing it on the remaining part(s). This process is repeated several times to get a better estimate of model performance.

- We used K-Fold Cross-Validation i.e k=5.The dataset is split into 5 equal parts (folds).The model is trained on 4 folds and tested on the remaining fold. This is repeated 5 times, each time using a different fold as the test set and their average is the final evaluation.

❖ **Model with their parameter**

a. Logistic Regression

Parameter	Description
C	Avoids overfitting or underfitting
Penalty(l2)	Shrinks all weights gradually → keeps all features, but makes them smaller

b. Randomforest

Parameter	Description
n_estimators	This is the number of decision trees the random forest will create. More trees = better performance
max_depth	Maximum Depth of Each Tree

c. Support Vector machine

Parameter	Description
gamma	defines how far the influence of a single training example reaches.

3. Results table and Discussion

Metrices\models	Logistic Regression	Random Forest	SVM
Accuracy	0.6300	0.6300	0.6100
AUROC	0.6576	0.6539	0.3087
Sensitivity(Recall)	0.4762	0.3333	0.8810
specificity	0.7414	0.7414	0.7414
F1-Score	0.5195	0.4308	0.6549

❖ **Discussion**

Strengths

- Multiple Model Comparison: our project includes logistic regression, random forest, and SVM, allowing a robust comparison across diverse classifiers.
- Feature Selection: Use of SelectKBest and f_classif enhances model performance by reducing dimensionality.
- Hyperparameter Tuning: Incorporating GridSearchCV ensures optimal parameter selection.
- Data Preprocessing: Standard scaling and checking for missing values are well-handled.

Limitations

- Limited Exploration of Feature Engineering: Focuses on SelectKBest but doesn't explore domain-specific feature creation or transformation.

Possible Improvements with More Time

- Advanced Feature Engineering: We can use other advance feature engineering to explore hidden structures.
- More hyperparameter tuning can be done for more accurate result.
- Also we can implement other model for better accuracy.