**Academic Report - Team TM 213: Jambulingam Madhumita, Li Qiyue, Soon Si Qi, Kelly**

We developed three models - Logistic Regression, XGBoost, and Random Forest — to evaluate how different preprocessing and cleaning techniques affect predictive performance. By systematically testing these variations, we identified which preprocessing approaches best aligned with each algorithm's assumptions and produced the strongest results on our dataset.

| Model Type | Specific Data Cleaning + Preprocessing | Common Data Cleaning + Preprocessing steps |
|---|---|---|
| Logistic Regression<br><br>**Accuracy: 78%**<br>**F1-score: 79%** | **Regularisation:** Prevented overfitting in the final LogisticRegression model by applying L2 regularization(C=0.1) to penalize model complexity and promote better generalization on unseen data. | 1.Dropped unnamed columns<br><br>2. Dropped the 'CLASS' feature (since our target variable is NSP)<br><br>3. No missing values present to be handled<br><br>4. Dropped duplicated rows |
| | **Handled Outliers:**Stabilized the model by applying IQR-based capping only to statistical features, preserving the integrity of critical clinical signals like DP and DS. | |
| | **Feature Selection**:Reduced model complexity by using SelectFromModel with an L1 penalty to automatically select the 16 most impactful features from the original 34. | |
| | **Data Transformation:** Corrected skewed features using a Yeo-Johnson transform for model stability. | |
| Random Forest<br><br>**Accuracy: 96%**<br>**F1-score: 98%** | **Regularisation:** Controlled overfitting by limiting the depth of trees (max_depth), increasing the minimum number of samples per leaf (min_samples_leaf), and setting the minimum number of samples required to split a node (min_samples_split). | |
| | **Handled Outliers:** Used Winsorization to handle extreme outliers by setting them to a predefined percentile threshold. | |
| XGBoost<br><br>**Accuracy: 93%**<br>**F1-score: 90%** | **Handled Outliers**: We only have 2000 records, removing outliers using algorithms (Boxplots) will lead to a significant number of records being removed. Hence, we used clinical domain knowledge to understand the plausible range of values for each feature. | |
| | **Feature Selection**: Removed redundant features (E.g. Mean, Median and Mode) were highly correlated with each other indicating multicollinearity. Hence, we removed 2 out of the three features. (Determined this using the correlation plot) | |
| | **Data Transformation:** Corrected the skew features using yeo-johnson power transformers. | |
| | **Feature Engineering**: Used feature important analysis and engineered new features for the most important features that impact the target variable | |

We evaluated and benchmarked the models above. Here are the tradeoffs we considered while deciding on our chosen model:

| Model | Pros | Cons |
|---|---|---|
| Logistic regression | **Average Performance:**<br>Macro F1-Score: 79%<br>Balanced Accuracy: 78% | **Unstable Performance:** The learning curve showed high variance, indicating inconsistent performance, which is a risk in a medical context. |
| | **Efficient:** Automatically reduced the number of features from 35 to 21 using L1 regularization. | **Sensitive to Class Imbalance:** Initial analysis showed the data is heavily skewed towards the 'Normal' class, which can challenge the model. |
| | **Highly Interpretable Model:** There is a clear relationship between inputs and outputs.The model learns a coefficient for each feature which shows how | - |

| | | |
|---|---|---|
| | each feature affects the outcome. | |
| XGBoost | **High Performance:**<br>Accuracy: 93%<br>Macro F1 - score: 96% | **Less interpretable**: Individual trees are easy to read but hundreds of trees cannot. |
| | **Efficient:** XGBoost is less sensitive to skewed features and includes an inbuilt L1 and L2 regularization | **Memory and Computationally intensive:** Training can be slower and more memory hungry compared to simpler models like Logistic Regression. |
| RandomForest | **High performance:**<br>Macro F1-Score: 98%<br>Balanced Accuracy: 96% | **Slight overfitting:** The cross validation log loss is slightly higher than training log loss, which signifies mild overfitting. |

Random Forest has achieved the highest accuracy and F1-score among the tested models but shows slight signs of overfitting, as its performance gap between training and test data is significant. XGBoost, on the other hand, produced slightly lower accuracy and F1-scores but demonstrated better generalization with minimal overfitting, likely due to its built-in regularization. Logistic Regression offered high interpretability of the model but it was limited in capturing nonlinear patterns, leading to comparatively lower accuracy.

Based on this benchmarking, we finalised on XGBoost as our solution due the balance between its high accuracy and generalization on new data.