TM-39 Academic Report

Group Members: Kelson Chua Wang Yan, Gabbie Clarissa Utama, Hong Keng Seng

I. Preprocessing and Cleaning of Data

We identified and cleaned the provided CTG.xls data by defining the dataframe. Unnecessary columns for classification are dropped from the dataframe. The columns dropped are: FileName, SegFile, Date, CLASS, A, B, C, D, E, AD, DE, LD, FS, SUSP, LBE, and DR.

II. Exploratory Data Analysis

The countplot of the NSP variable as shown in Diagram 2.1 shows that the dataset is imbalanced between classes, with the majority class being 1 (1=Normal, 2=Suspect, 3=Pathological). Feature correlation heatmap shows that DP has the highest correlation to NSP by magnitude as shown in Diagram 2.2.

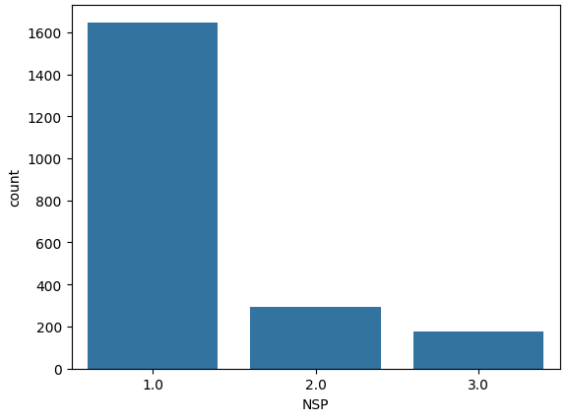

Diagram 2.1 Countplot of NSP

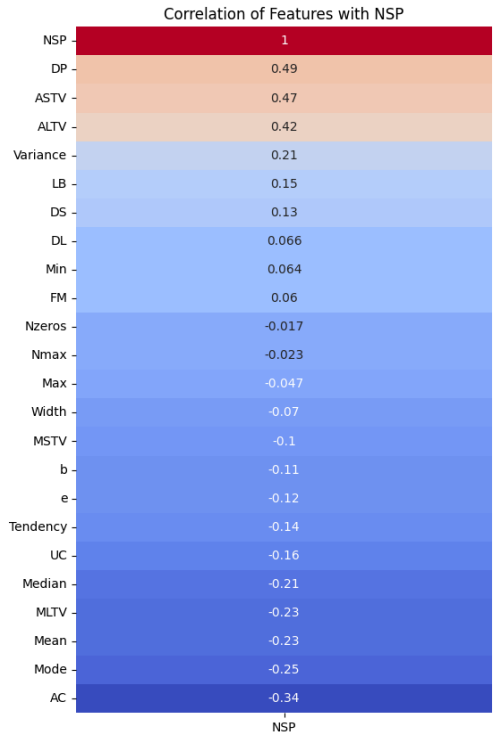

Diagram 2.2 Correlation Value of Features with NSP

III. Building The Classification Model

The classification model is built, the models used include Gradient Boosting, Random Forest, Support Vector Machine, Decision Tree, Logistic Regression, Neural Network and k-Nearest Neighbours.

IV. Performance Evaluation

The balanced accuracy and F1 Macro scores are shown in Table 4.1 and compared as shown in Diagram 4.2, with Gradient Boosting being the highest performing model. From the Gradient Boosting classification report as shown in Diagram 4.3, the model has the highest performance in precision with 96.2% and recall of 94.3% in predicting NSP = 3 (Pathologic). The high value of recall shows that the risk of missing the pathologic case is low.

Model	Balanced Accuracy	F1 Macro
Gradient Boosting	0.914192	0.916861
Random Forest	0.902703	0.902076
Support Vector Machine	0.872040	0.841064
Decision Tree	0.866831	0.837946
Logistic Regression	0.834386	0.790770
Neural Network (MLP)	0.755396	0.796023
k-Nearest Neighbors	0.717111	0.768323

Table 4.1 Model Accuracies

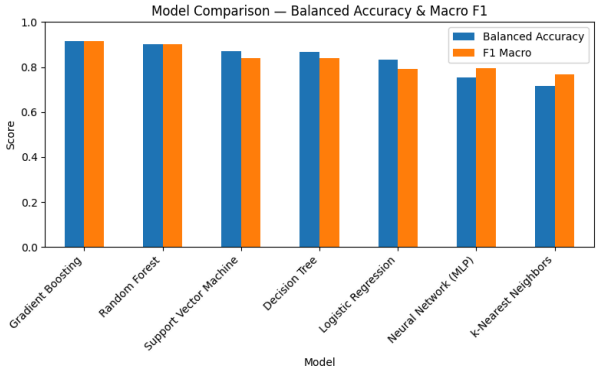

Diagram 4.2 Modern Comparison

Gradient Boosting — Balanced Accuracy: 0.914, Macro F1: 0.917

	precision	recall	f1-score	support
1	0.968	0.970	0.969	494
2	0.830	0.830	0.830	88
3	0.962	0.943	0.952	53
accuracy			0.948	635
macro avg	0.920	0.914	0.917	635
weighted avg	0.948	0.948	0.948	635

Diagram 4.3 Gradient Boosting Classification Report