Meta-learning assignment – ECG Signals

Madhav Bajaj 2019B3A70256G

QA)

1) Features TS_0 to TS_269 are Continuous variables, while the target variable CLASS is Categorical.

2) To analyze and detect the trends and patterns present among the features or time series independent variables, initially, the correlation factors for all time series variables with respect to the label encoded target variable were plotted, hypothesizing an increasing trend indicating more recent variables would display a more remarkable ability to predict the presence of Atrial Fibrillation, but no discernable trend was obtained.

The locality of correlation factors was observed after conducting a correlation analysis of time series variables. Time series variables are highly correlated with the variables close to them in time. A negative correlation was also observed between variables with a certain time difference.

3) Principal component Analysis (PCA) was conducted to reduce the dimensionality. 134 features were considered for the analysis, with an explained variance ratio of approximately 80%.

4) The Categorical target variable was converted to numerical to make the dataset suitable for modeling. CLASS 'A' is assigned 1 while 'N' is set to 0.

QB)

The following classification models were implemented:

1) Logistic Regression
2) K-Nearest Neighbor (KNN)
3) Support Vector Machine (SVM)
4) Decision Tree
5) Random Forest
6) Naïve Bayes

QC)

The model was compared on the following metrics

1) Precision
2) Recall
3) F1-Score
4) Accuracy

Index	Logistic Regression	KNN	SVM	Decision Tree	Random Forest	Gaussian NB
Accuracy	0.876429	0.877246	0.876973	0.769189	0.875068	0.360915
Precision	0.0	0.538462	0.0	0.141304	0.0	0.133127
Recall	0.0	0.015487	0.0	0.172566	0.0	0.761062
F1 Score	0.0	0.030108	0.0	0.155378	0.0	0.226614

Results for the imbalanced dataset

Given the dataset is heavily imbalanced (16030 instances of 'N' and 2340 instances of 'A'), the accuracy metric is unsuitable for comparing the models. Under such circumstances, the F1 score is the superior metric to differentiate the performance of multiple models.

QD)

To treat the imbalanced dataset, Synthetic Minority Oversampling Technique (SMOTE) has been utilized. SMOTE augments the minority class (the presence of Atrial Fibrillation (CLASS A) here) to overcome the imbalance.

Index	Logistic Regression	KNN	SVM	Decision Tree	Random Forest	Gaussian NB
Accuracy	0.529668	0.675014	0.722646	0.6957	0.83288	0.439031
Precision	0.118421	0.229197	0.18324	0.159509	0.19084	0.121412
Recall	0.438053	0.69469	0.362832	0.345133	0.110619	0.570796
F1 Score	0.186441	0.344676	0.243504	0.218182	0.140056	0.200233

Results for the balanced dataset after SMOTE

Models perform considerably better on the balanced dataset.