Multi-output Classifier Chain applied to Cardiotocography Dataset

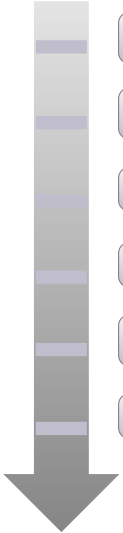
DS 397: Advanced Computational Methods in Data Science

Enrick Cavas Jose Aries De Los Santos Christian James Galotera Jayrah Bena Riñon

June 15, 2024

University of the Philippines Diliman

Presentation Outline



- Exploratory Data Analysis
- Dimensionality Reduction
- Process Flow
- Artificial Neural Network
- Light Gradient Boosting Machine
- Results and Discussion

Exploratory Data Analysis

Cardiotocography Dataset

Cardiotocography

Donated on 9/6/2010

The dataset consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features on cardiotocograms classified by expert obstetricians.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Health and Medicine	Classification
Feature Type	**# Instances**	**# Features**
Real	2126	21

21 Features of the Cardiotocography Dataset

- LB - FHR baseline (beats per minute)
- AC - of accelerations per second
- FM - of fetal movements per second
- UC - of uterine contractions per second
- DL - of light decelerations per second
- DS - of severe decelerations per second
- DP - of prolongued decelerations per second
- ASTV - percentage of time with abnormal short term variability
- MSTV - mean value of short term variability
- ALTV - percentage of time with abnormal long term variability
- MLTV - mean value of long term variability
- Width - width of FHR histogram
- Min - minimum of FHR histogram
- Max - Maximum of FHR histogram
- Nmax - of histogram peaks
- Nzeros - of histogram zeros
- Mode - histogram mode
- Mean - histogram mean
- Median - histogram median
- Variance - histogram variance
- Tendency - histogram tendency

2 Targets of the Cardiotocography Dataset

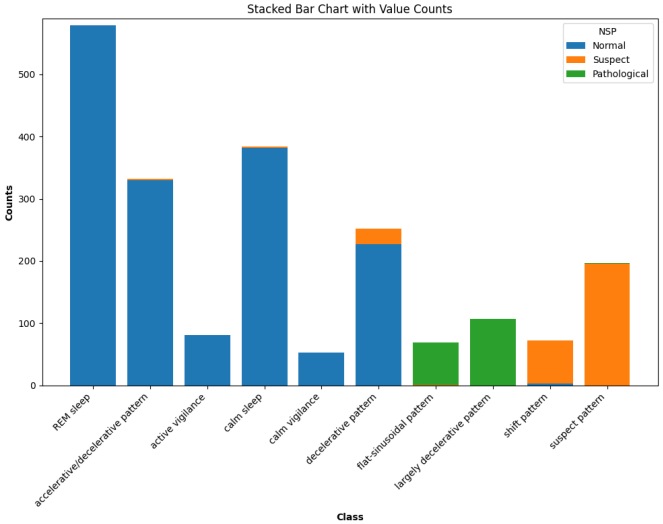
CLASS - FHR pattern class code (1 to 10)

- ✓ calm sleep
- ✓ REM sleep
- ✓ calm vigilance
- ✓ active vigilance
- ✓ shift pattern (A or Susp with shifts)
- ✓ accelerative/decelerative pattern (stress situation)
- ✓ decelerative pattern (vagal stimulation)
- ✓ largely decelerative pattern
- ✓ flat-sinusoidal pattern (pathological state)
- ✓ suspect pattern

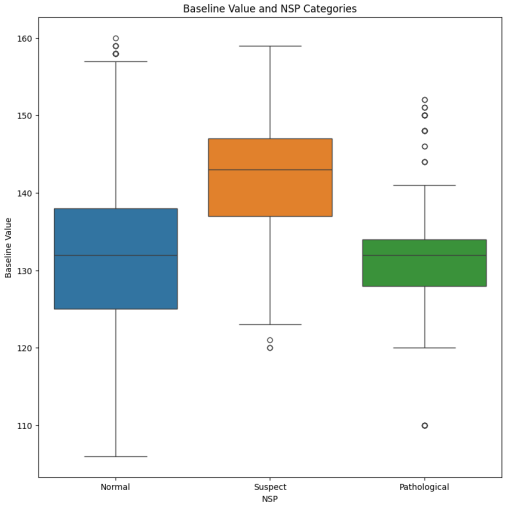
NSP - fetal state class code

- ✓ N = normal
- ✓ S = suspect
- ✓ P = pathological

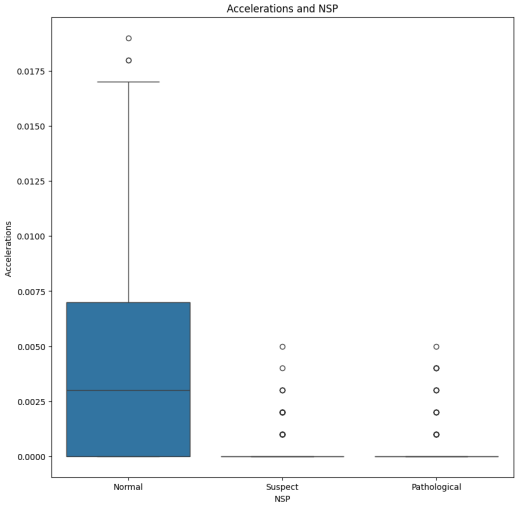
Exploratory Data Analysis



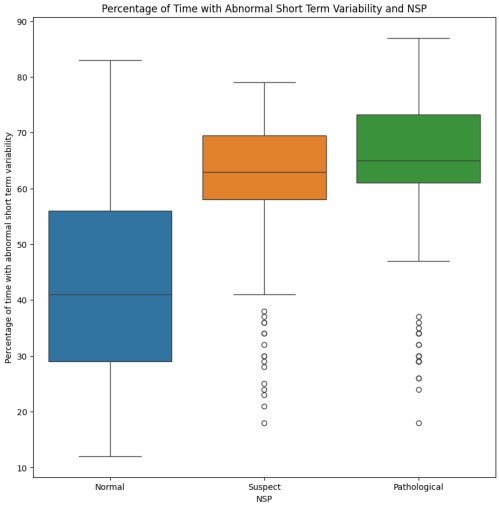
Exploratory Data Analysis



Exploratory Data Analysis

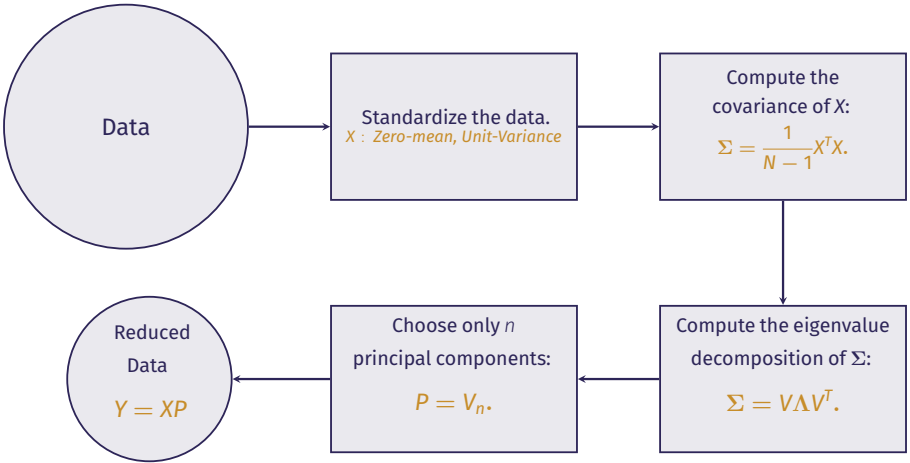


Exploratory Data Analysis

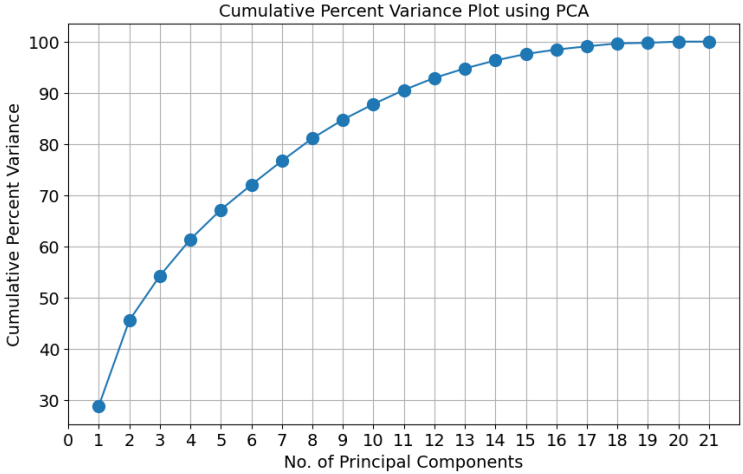


Dimensionality Reduction

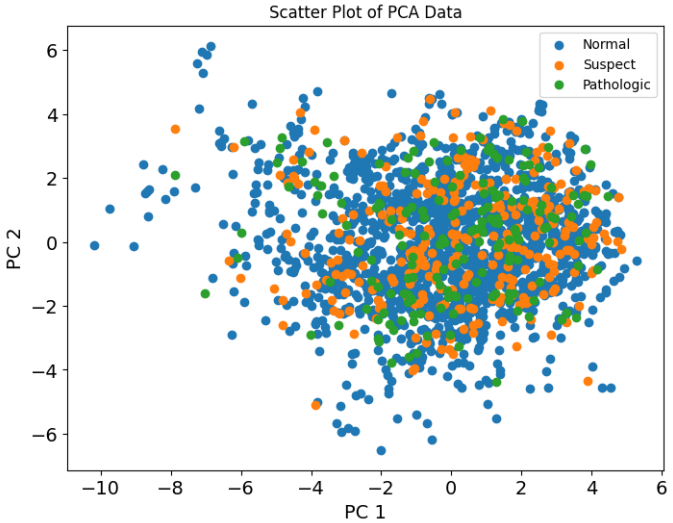
Principal Component Analysis (PCA)



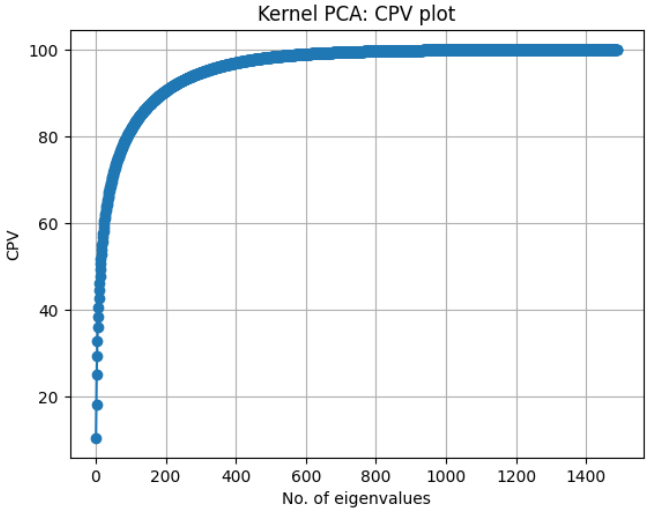
Principal Component Analysis (PCA)



Principal Component Analysis (PCA)

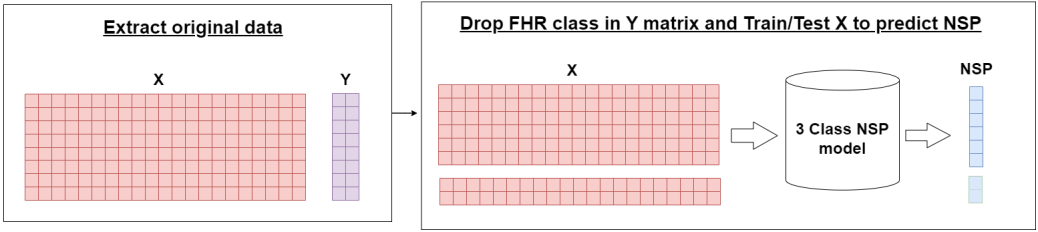


Kernel Principal Component Analysis (KPCA)

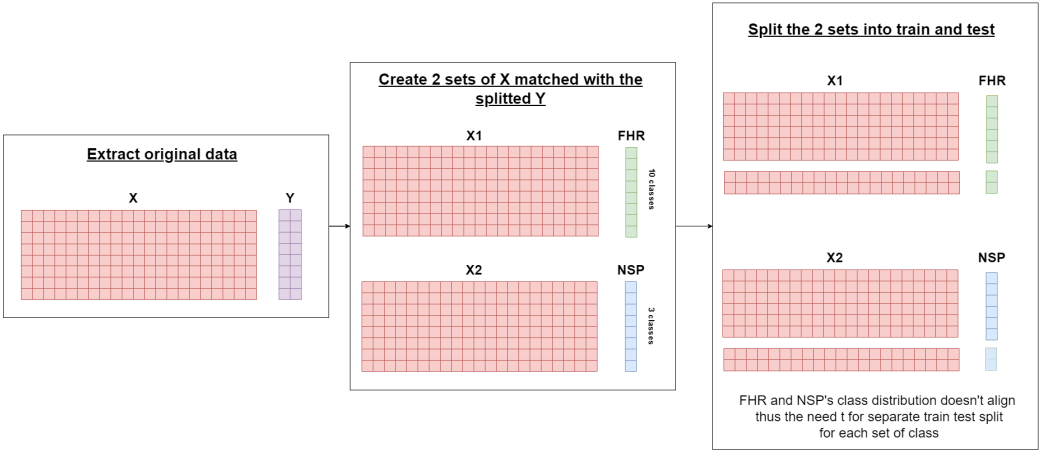


Process Flow

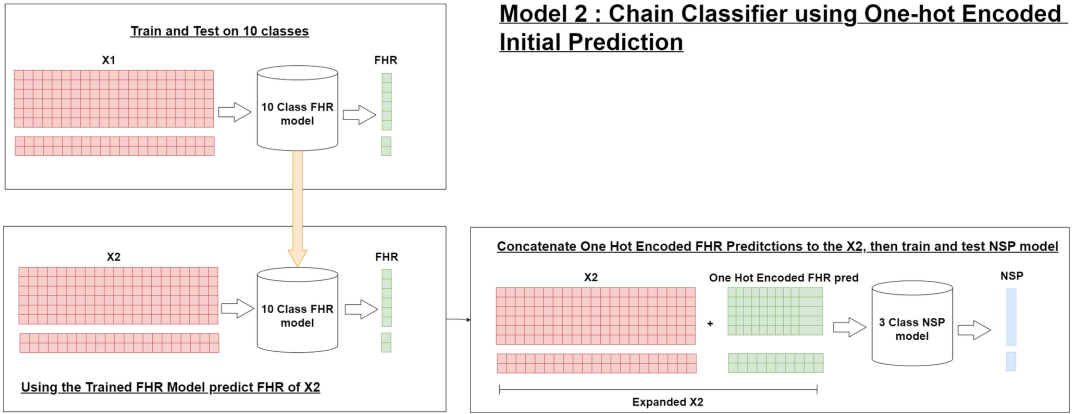
Model 1: Independent Classifier



Preprocessing step for Model 2 and 3

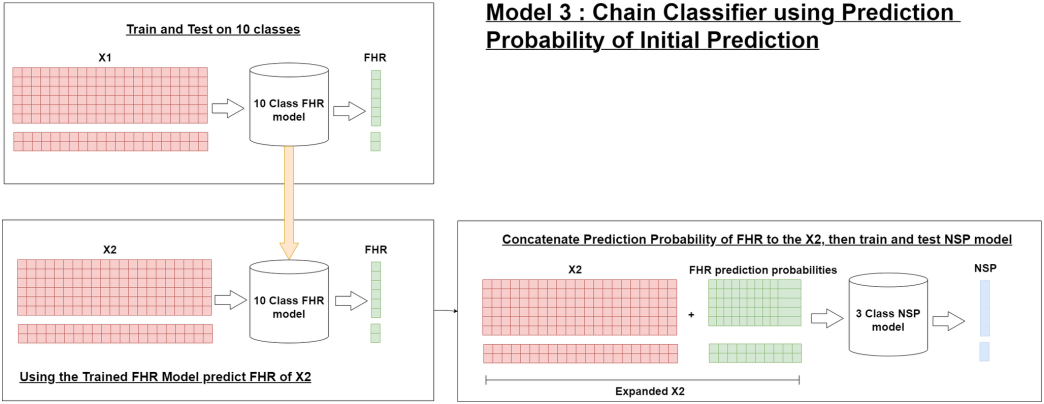


Process Flow



Model 2 : Chain Classifier using One-hot Encoded Initial Prediction

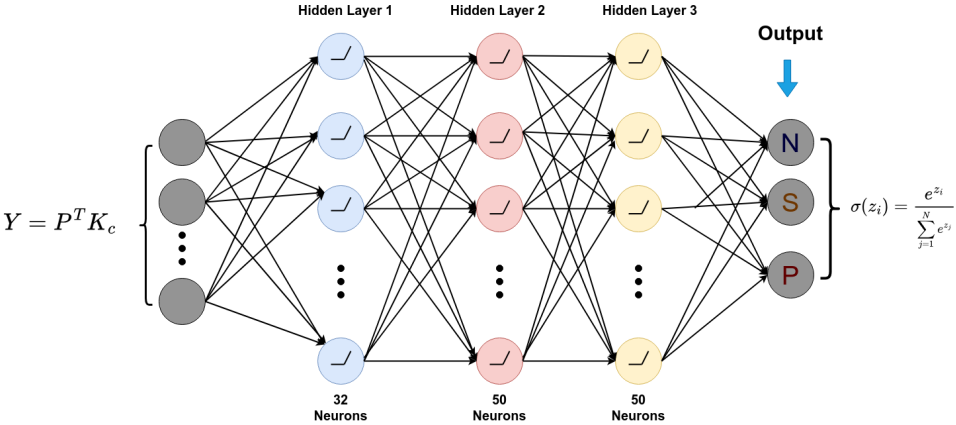
Process Flow



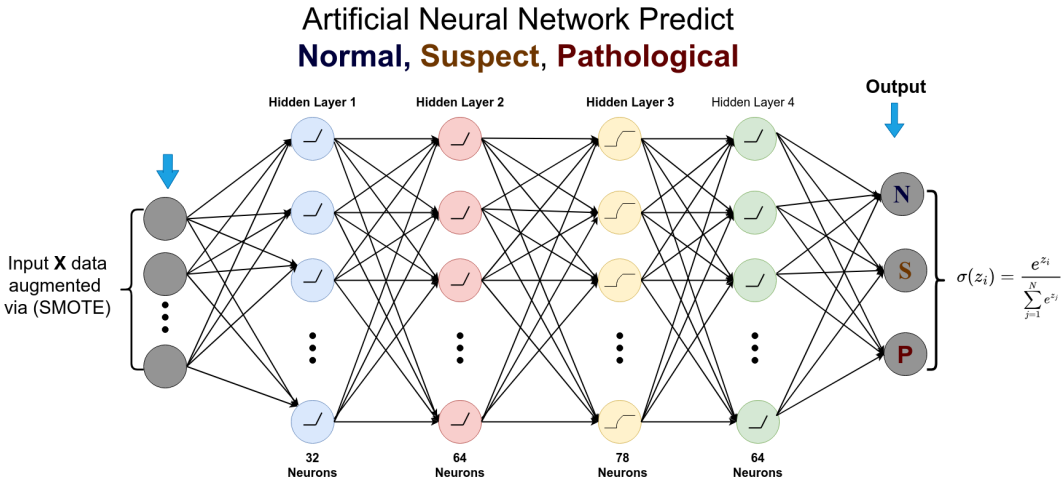
Model 3 : Chain Classifier using Prediction Probability of Initial Prediction

Artificial Neural Network

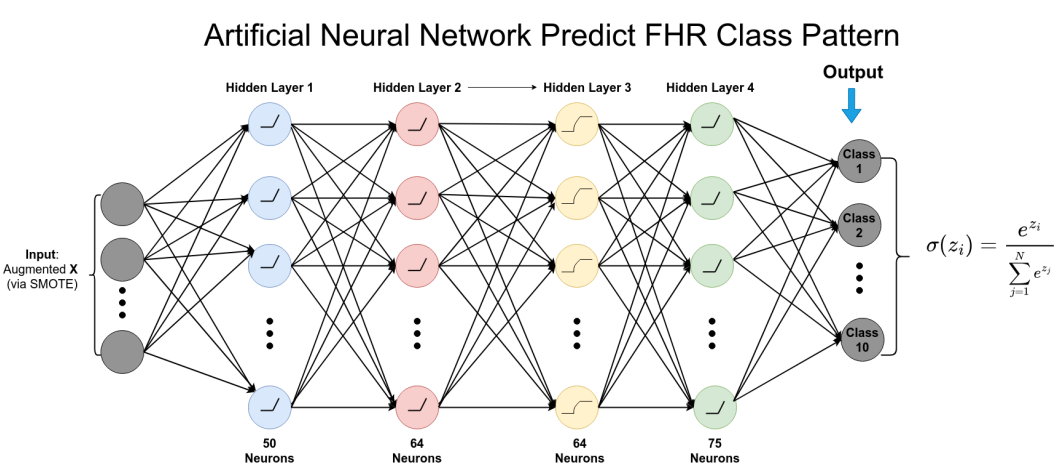
Artificial Neural Network with KPCA input layer
Predict **Normal**, **Suspect**, **Pathological**

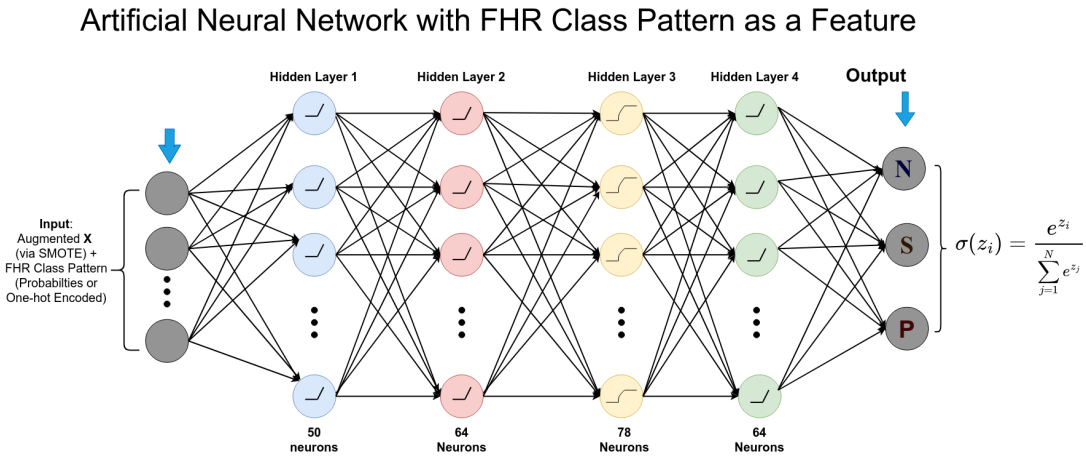


Artificial Neural Network

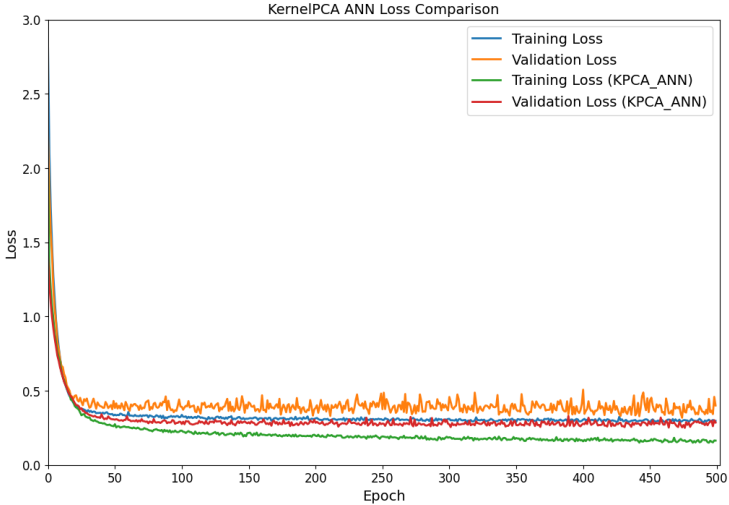


Artificial Neural Network




Artificial Neural Network with FHR Class Pattern as a Feature
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer 4
Output
Input:
Augmented X
(via SMOTE) +
FHR Class Pattern
(Probabilties or
One-hot Encoded)
N
S
P
50
neurons
64
Neurons
78
Neurons
64
Neurons
$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}$

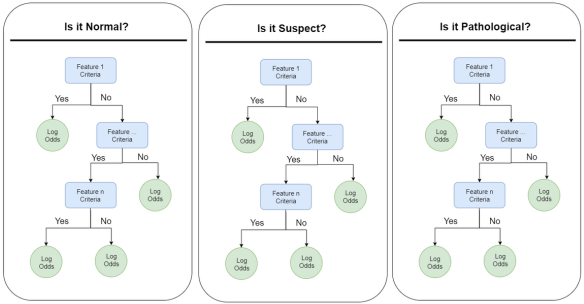
Artificial Neural Network



Light Gradient Boosting Machine

Light Gradient Boosting Machine

N number of trees will be created, representing the number of classes in the y vector.
These trees will be boosted multiple times as defined in the parameters



Specific steps taken

1. Pipeline is created with a standard scaler and lightGBM.
.
2. For the variable parameters (e.g. n_estimators, learning_rate, num_leaves, max_depth, min_child_samples, etc.) pipeline was optimized using **Optuna** to get the best input parameters.

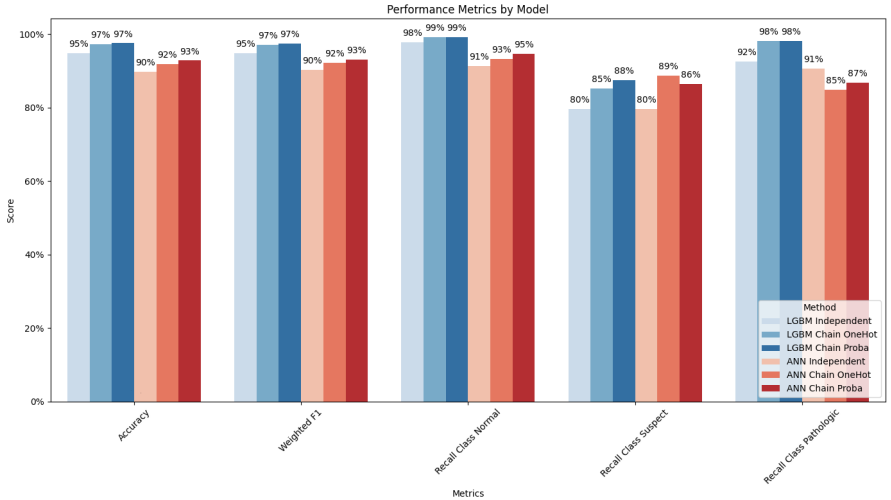
3. To address class imbalance, the is_unbalance is set to True

4. To ensure the model doesn't overfit, the following parameters were fixed:
   - boostin ='dart'
   - number_of boosting_round = 100
   - early stopping_round = 10

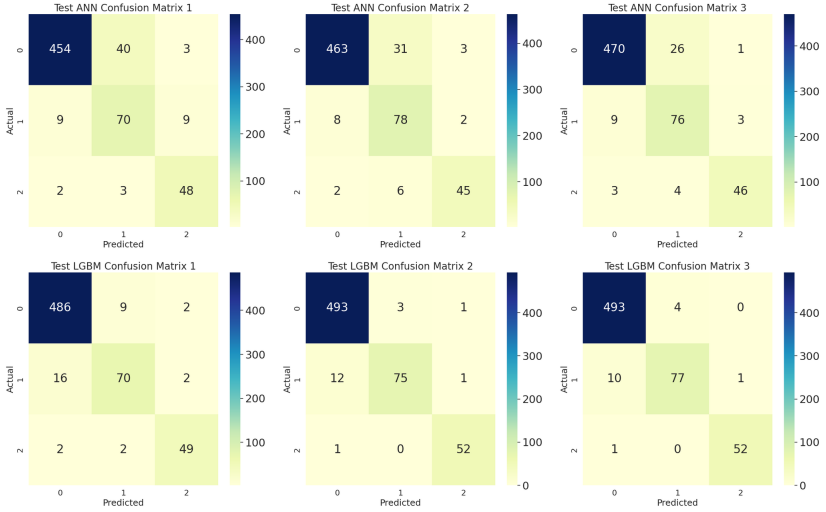
* Same steps and parameters were applied both for FHR and NSP models

Results and Discussion

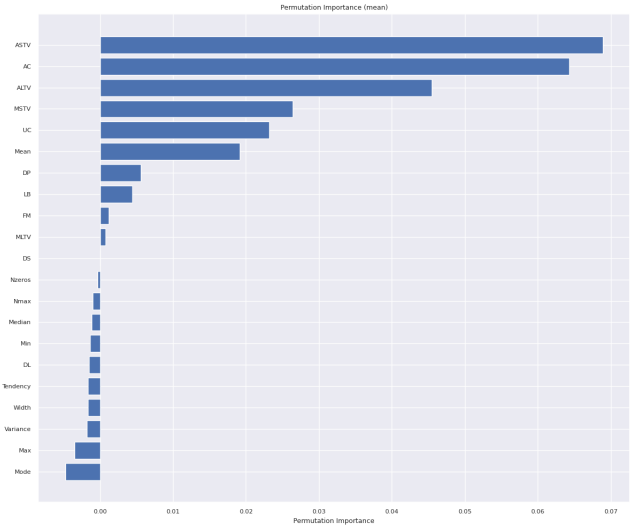
Performance Metrics by Model



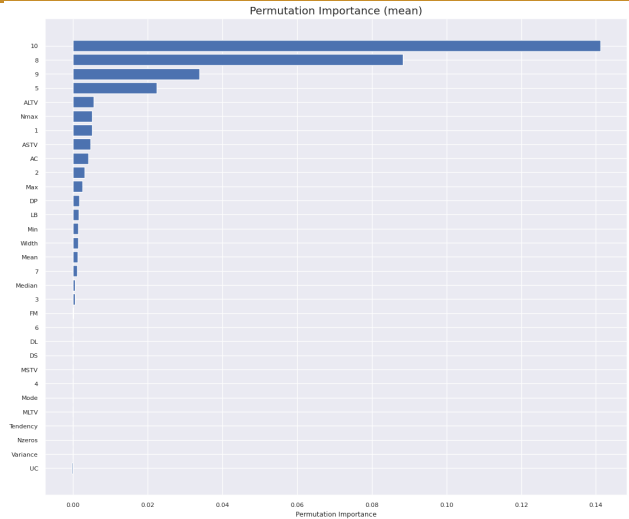
Confusion Matrix



Feature Importance



Feature Importance



Simulations

https://colab.research.google.com/drive/1oIAjgEyDyylk7fmIsE6sPd9qNrEUXkaG?usp=sharing

Thank you very much for listening!