Integrated PreNatal Care

Team Members:
Aamleen Ahmed (2020002)
Mohd. Sufyan Azam (2020312)
Mohd. Shariq (2020220)

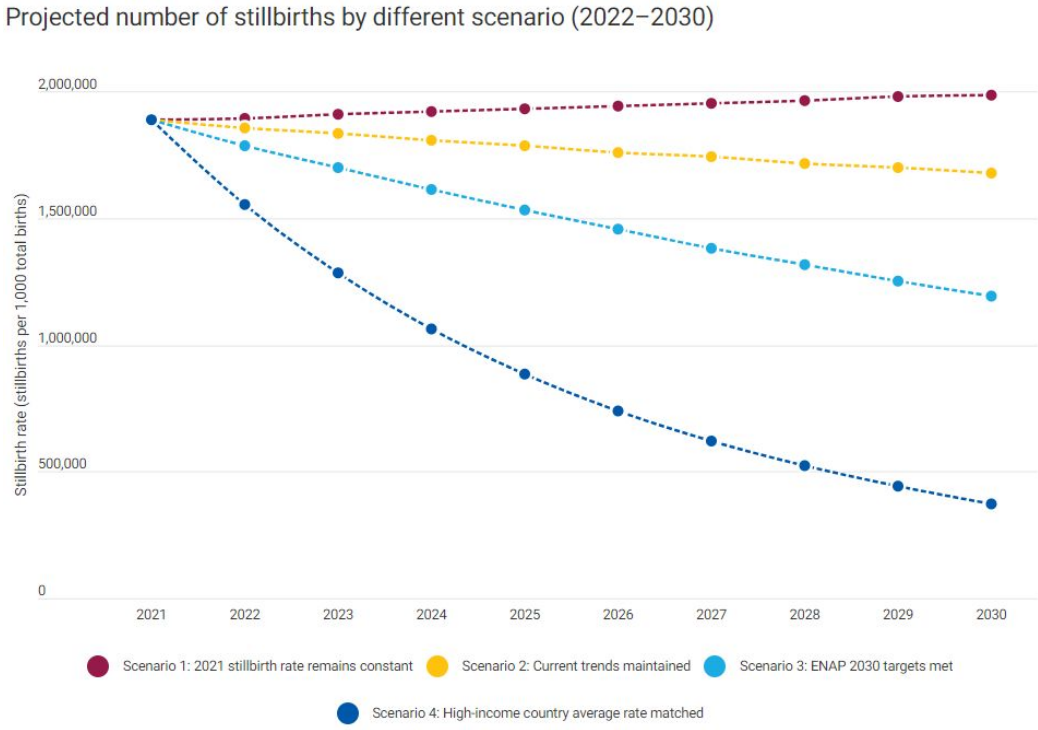
INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

The journey to parenthood is an extraordinary passage marked by joy and challenges. However, stillbirths, neonatal deaths, and preventable complications overshadow this miraculous process.

Our project is driven by an unwavering motivation to address this heart-wrenching issue. By leveraging the power of Machine Learning and analyzing a wealth of pre-natal data sources, we aim to provide healthcare professionals with a tool to predict potential health risks for mothers and fetuses.

With these insights, medical teams can make informed decisions, enabling timely interventions and a more coordinated approach to maternal and child health.



Through the fusion of data science and compassionate care, we envision a future where this integrated model significantly reduces the occurrence of tragic outcomes, paving the way for healthier beginnings and brighter tomorrows.

Literature Review

Research Papers Summary

Paper 1 A Machine Learning Approach for the Prediction of Fetal Health using CTG

Astik Kumar Pradhan; Jitendra Kumar Rout; Aurobinda Maharana; Bunil Kumar Balabantaray; Niranjan Kumar Ray et.al.

- Discussion on machine learning techniques for the prediction of fetal health based on cardiotocography (CTG) data
- CTG is a method for monitoring fetal heart rate and uterine contractions during pregnancy.
- classifiers such as Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting Machine (GBM)
- Model performance evaluation is conducted using metrics such as accuracy, precision, recall, and F1-score.
- Random Forest model emerges as the most effective, achieving the highest accuracy rate of 0.99.
- future research directions, the paper under-scores the potential for further enhancement in predictive model performance through the incorporation of additional preprocessing steps, such as linear discriminant analysis or principal component analysis.

Paper 2 Evaluation of support vector machines and random forest classifiers in a real-time fetal monitoring system based on cardiotocography data

Vinayaka Nagendra et.al.

- Evaluation techniques for the foetal state prediction based on Cardiotocography (CTG) data are compared.
- Extract features that potentially offer more details about the foetal status and assess the performance of these predictions in a real-time clinical decision support system.

- Takes into account all three foetal states (normal, suspicious, and abnormal).

- models like Support Vector Machines(SVM) and Random Forests (RF) are used

- SVM performed marginally better for suspicious instances, although both SVM and RF had above 96% accuracy

- Highlights importance of CTG data in forecasting the condition of the foetus during labour to show the danger of foetal acidosis (low blood pH from low oxygen levels)

Paper 3 Machine Learning Approaches for Early Diagnosis and Prediction of Fetal Abnormalities

R. Chinnaiyan and Stalin Alex

- explored machine learning approaches for early detection of prenatal anomalies.
- discussion on importance of early diagnosis of fetal anomalies, particularly in the first trimester,with the goal of lowering their occurrence rates and improving accuracy for ultrasound fetal imaging.
- It involves four key steps: segmentation, image enhancement, feature extraction, and image classification.
- Best model for our classification task is a Neuro Fuzzy Based Genetic Algorithm and the accuracy achieved through it was around 0.98

Paper 4 Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future

Muhammad Nazrul Islam et.al.

- Focuses on ML techniques for predicting optimal childbirth modes and detecting complications.
- Explores algorithms, features, data sources, and their performance in pregnancy outcomes.
- highlights future research opportunities for reducing maternal complications and mortality rates, including unsupervised and deep learning algorithms, ML-based clinical decision support systems, dataset enhancement, and surgical robotic tools

DataSet

Description, Visualisation, PreProcessing

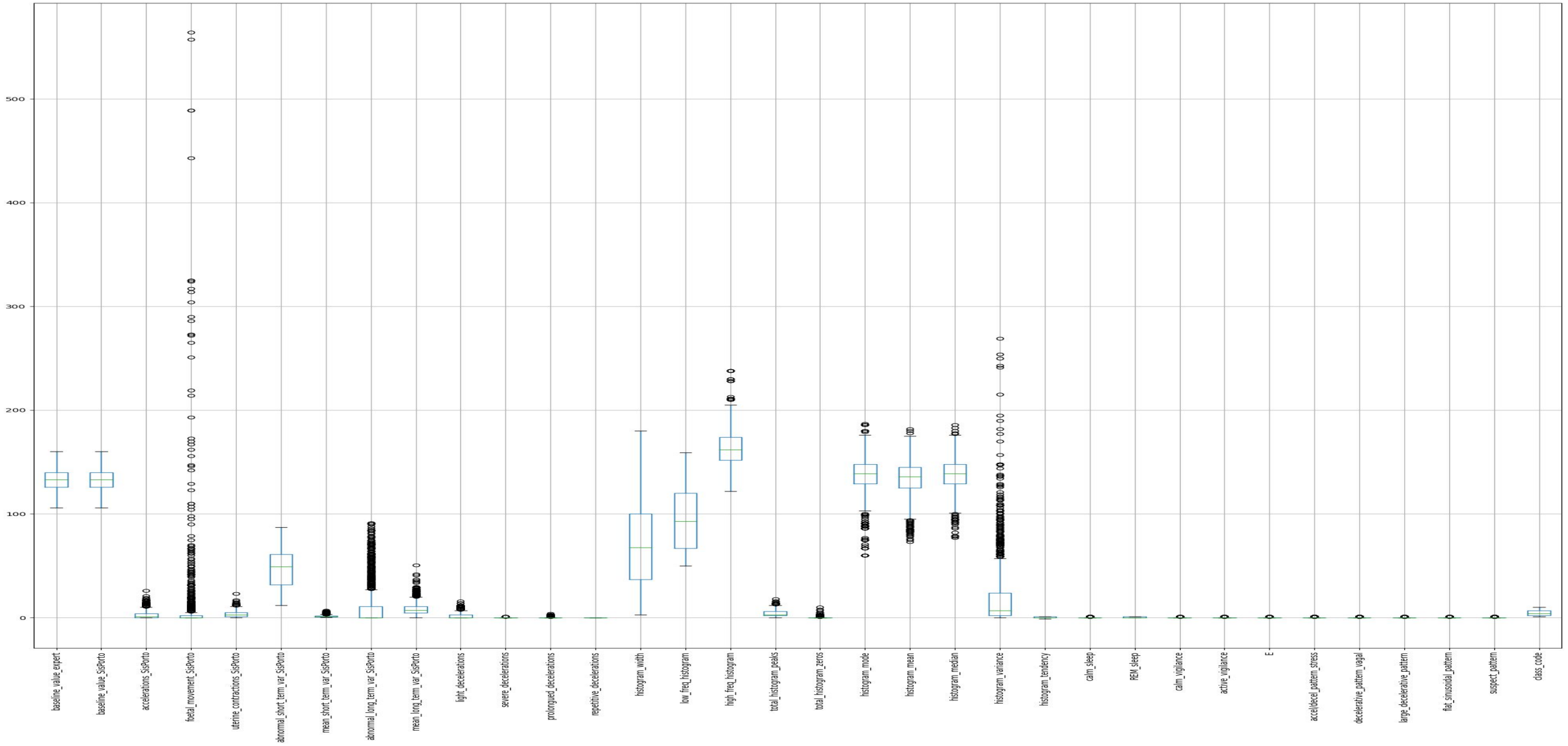
Description

- The Cardiotocography (CTG) dataset used in this project is taken from the UCI Machine Learning Repository. Link: Dataset
- It consists of measurements of features such as fetal heart rate (FHR) and uterine contraction (UC) on a device called cardiotocograph. Each CTG observation obtained is classified by expert obstetricians.
- This dataset is multivariate and contains 2126 instances along with 40 features that can be used for classification purposes.
- It can be used for classifying either the morphological pattern of the fetus (A, B, C, … columns) which is a 10-class classification problem, or the fetal state (N=Normal, S=Suspect, P=Pathological) which is a 3-class classification problem.

Description – Header Details

FileName	of CTG examination	Width	histogram width
Date	of the examination	Min	low freq. of the histogram
b	start instant	Max	high freq. of the histogram
e	end instant	Nmax	number of histogram peaks
LBE	baseline value (medical expert)	Nzeros	number of histogram zeros
LB	baseline value (SisPorto)	Mode	histogram mode
AC	accelerations (SisPorto)	Mean	histogram mean
FM	foetal movement (SisPorto)	Median	histogram median
UC	uterine contractions (SisPorto)	Variance	histogram variance
ASTV	percentage of time with abnormal short term variability (SisPorto)	Tendency	histogram tendency: -1=left assymetric; 0=symmetric; 1=right assymetric
mSTV	mean value of short term variability (SisPorto)	A	calm sleep
ALTV	percentage of time with abnormal long term variability (SisPorto)	B	REM sleep
mLTV	mean value of long term variability (SisPorto)	C	calm vigilance
DL	light decelerations	D	active vigilance
DS	severe decelerations	SH	shift pattern (A or Susp with shifts)
DP	prolongued decelerations	AD	accelerative/decelerative pattern (stress situation)
DR	repetitive decelerations	DE	decelerative pattern (vagal stimulation)
CLASS	Class code (1 to 10) for classes A to SUSP	LD	largely decelerative pattern
NSP	Normal=1; Suspect=2; Pathologic=3	FS	flat-sinusoidal pattern (pathological state)
		SUSP	suspect pattern

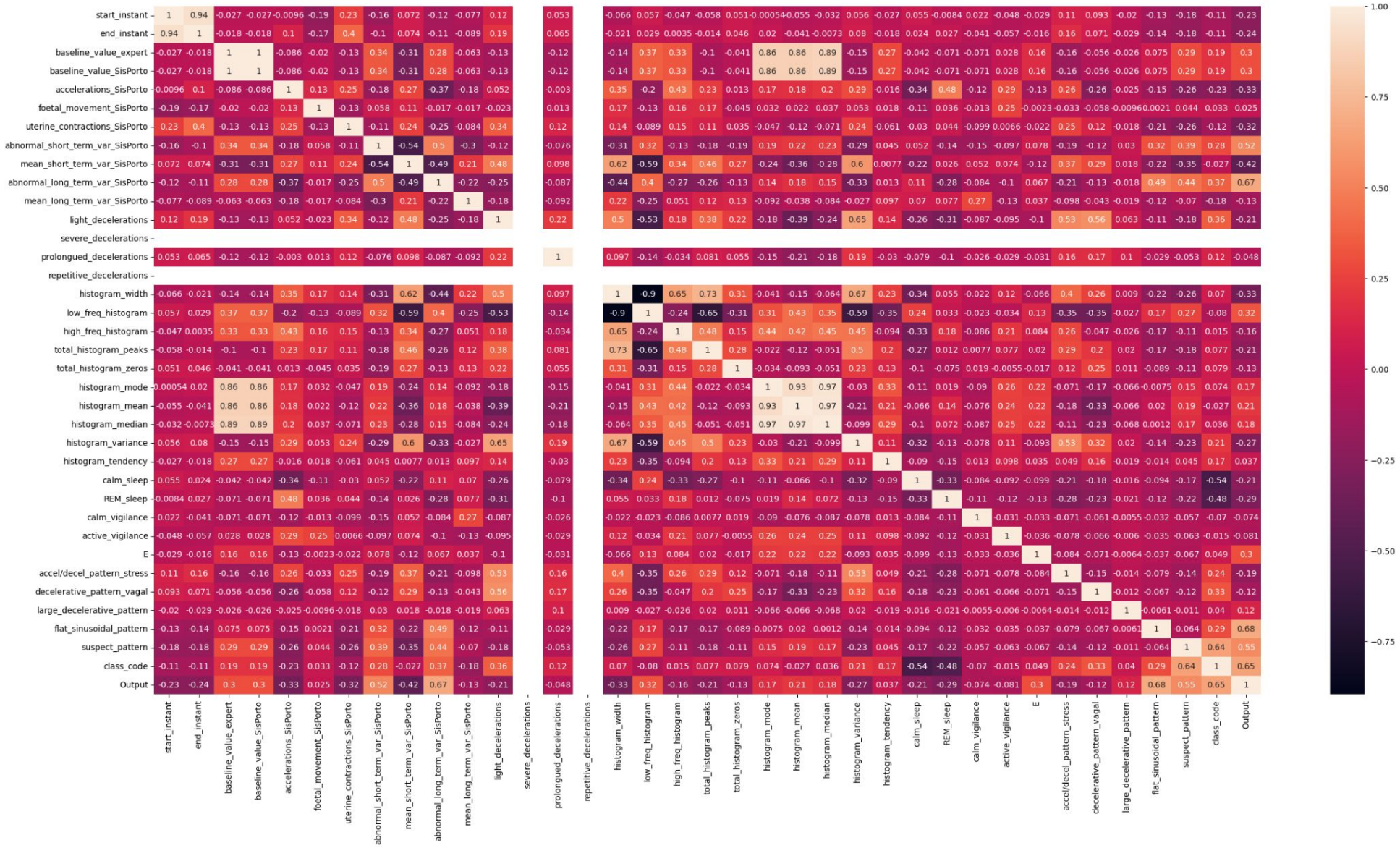
Visualization – Box Plot for Outliers



Preprocessing – Handling Outliers

- We checked for the presence of outliers using boxplots and removed them by using a variety of methods -
  - **Z-Score normalization**: Maps the data to a normal distribution with mean 0 and standard deviation 1;
  - **KNN Imputer**: Replaces outliers with the mean of their nearest neighbors;
  - **Inter-Quartile Range**: Maps the difference between quartiles and removes values outside a certain range.
  - **Robust Covariance:** Uses robust estimator for the covariance matrix and removes values above a threshold.
- The best performing method among them for our dataset was the Robust Covariance Method.

Visualization - Correlation Heatmap



Preprocessing – Feature Engineering

- We achieved feature engineering by making strategic decisions to enhances model efficiency and interpretability through the help of the correlation heatmap.
- We identified certain columns that exhibited high correlation among each other, indicating redundancy in our feature set and removed them.
- We also eliminated columns with extremely low correlation with the output data, suggesting minimal relevance in predicting fetal health and prenatal complications.
- We also removed columns such as FileName, Date, etc that didn’t have an impact on the predictions.

Preprocessing – Output Class Imbalance

Context:

- Applied to 3-class classification due to class imbalance.
- Not extended to 10-class classification with nearly uniform data distribution.

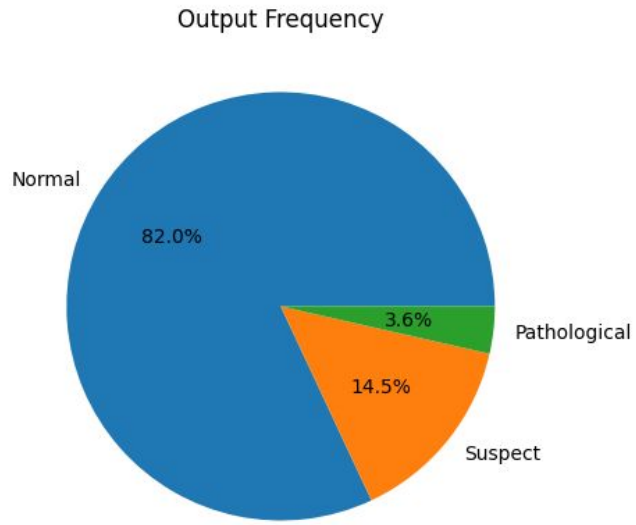
Imbalance Challenge:

- Imbalanced classes risk bias towards dominant class, resulting in inaccurate outcomes.

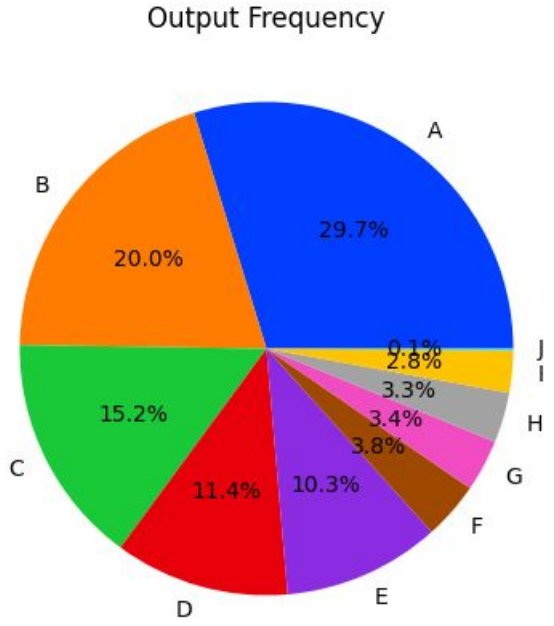
Resampling Techniques

- Utilized SMOTE and NEAR-MISS algorithms to address heavy imbalance.
- Specifically applied due to the dataset being skewed towards the normal class.

3-class



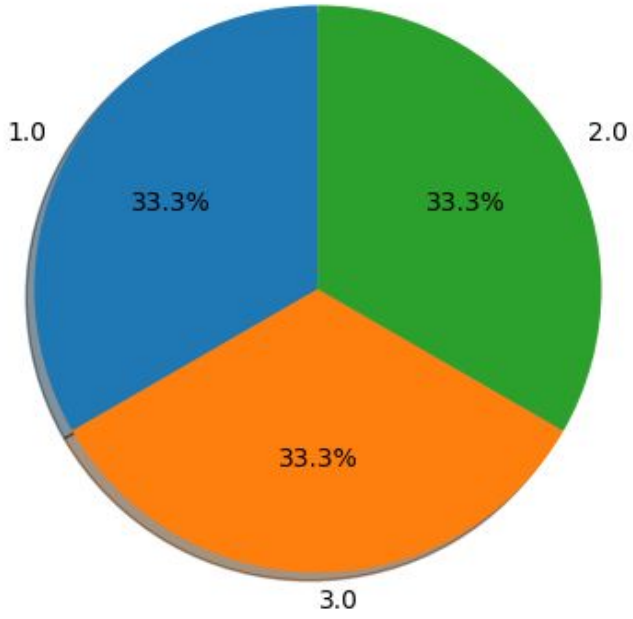
10-class



Preprocessing – Resampling Techniques

Oversampling with SMOTE:

- Utilizes synthetic sample creation to address imbalances.
- Selects minority instances and generates new samples along connecting lines.

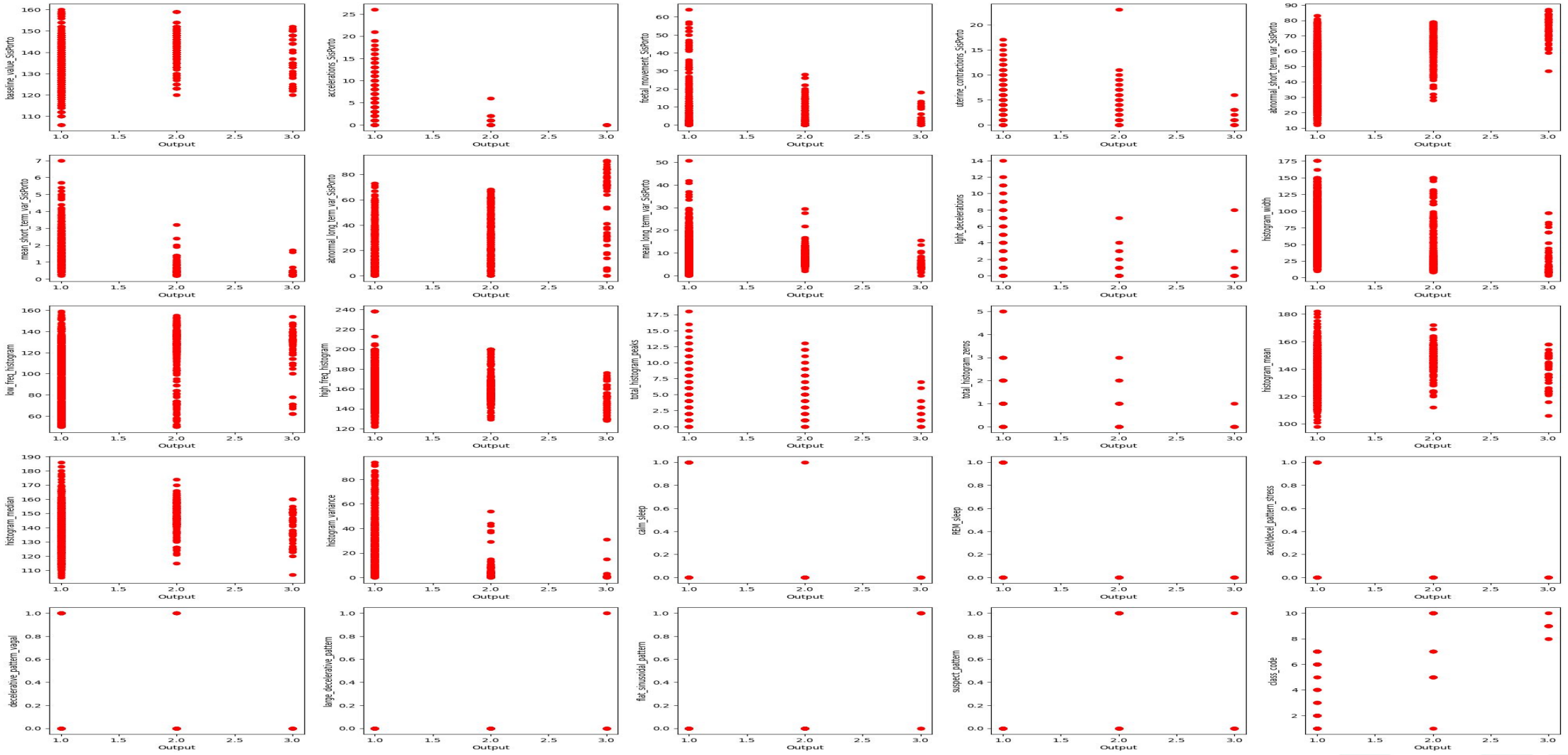


Undersampling with NEAR-MISS:

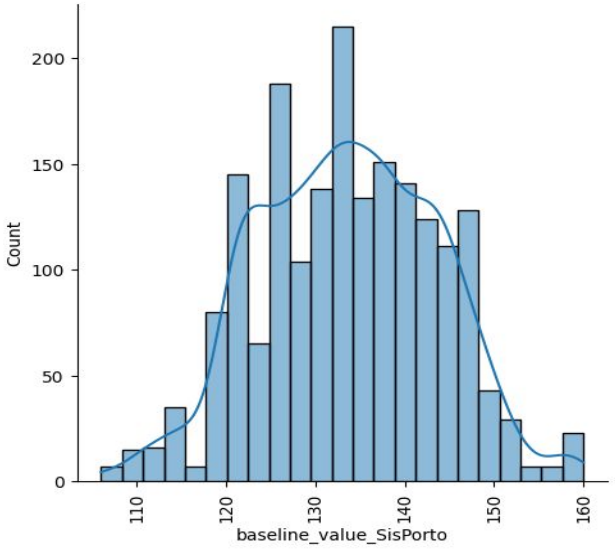
- Randomly removes data from the majority class for balance.
- Uses the NEAR-MISS algorithm to enhance classification but may lead to underfitting due to reduced total records (1000).

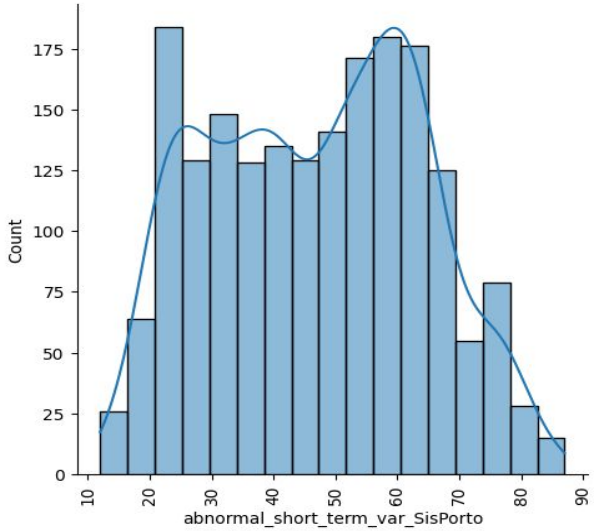
Did not use Resampling in 10 class classification due to data scarcity per-class, and data was more or less equally distributed.

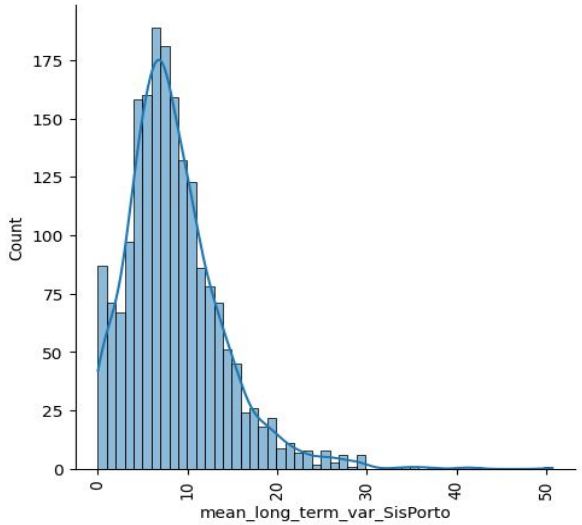
Visualization – Scatter Plots

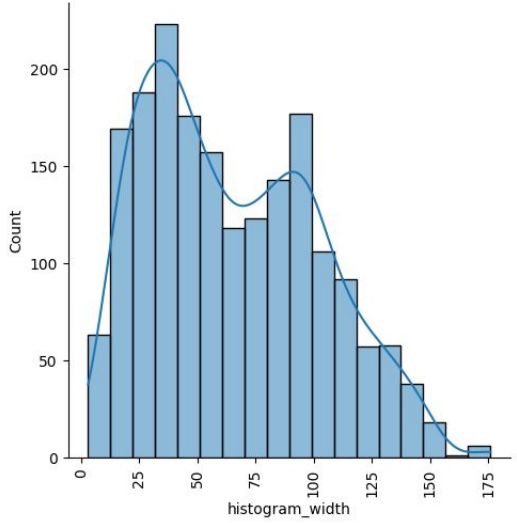


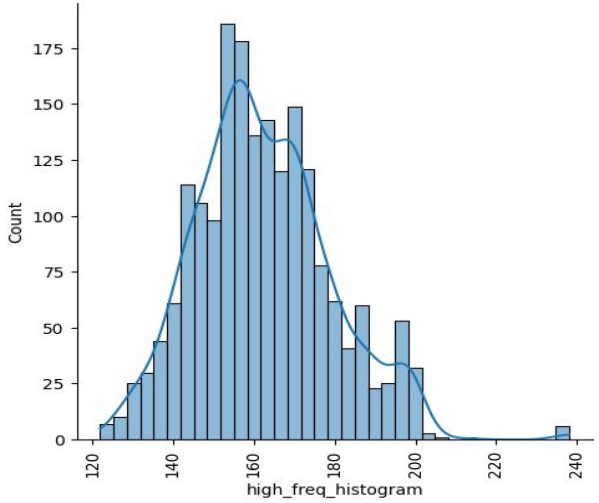
Visualization – Distribution Plots

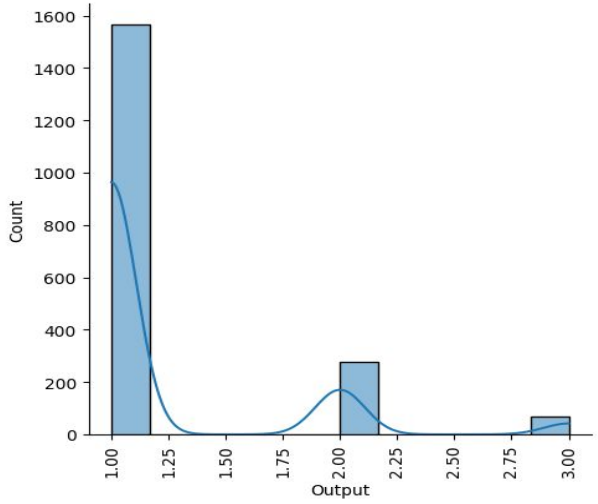












Visualization – Pair Plots

Methodology

Models Used, Overfitting Avoidance

Overfitting

Due to lack of a large dataset, our model was overfitting, as we were getting very high training scores ~100%, but testing scores were limited to 80-90%. So we explored the following methodologies:

- **PCA:** Principal Component Analysis, employed for dimensionality reduction, significantly improved our model's generalization capabilities, enhancing accuracy and mitigating overfitting.

- **Regularization (L1 and L2 using liblinear):** The application of L1 and L2 regularization methods with the liblinear solver helped control overfitting by optimizing feature coefficients, ultimately leading to more stable and reliable models.

- **K-Fold Cross-Validation (K=10, 3, 5):** We used 10-fold, 3-fold, and 5-fold cross-validation techniques to rigorously evaluate our models, ensuring robustness and accuracy across different validation scenarios.

Models

1. Logistic Regression (LR): Foundational understanding with F1 scores of 0.72 (3-class) and 0.68 (10-class).

2. K-Nearest Neighbors (KNN): Solid performance with F1 scores of 0.82 (3-class) and 0.78 (10-class).

3. Support Vector Machines (SVM): High F1 score of 0.87 (3-class) in distinguishing fetal states.

4. Naive-Bayes: Limitations in capturing dependencies with an F1 score of 0.66 (3-class). ***Baseline***

5. Decision Trees: Prone to overfitting, concerns about performance on unseen data, especially in the 3-class scenario.

Models

6. Random Forests: Effective in mitigating overfitting, achieving F1 scores of 0.79 (3-class) and 0.76 (10-class).

7. XGBoost: Demonstrated solid performance with an F1 score of 0.85 (3-class) and 0.8 (10-class).

8. CatBoost: Improved F1 score to 0.8+ in the 10-class scenario, showcasing enhanced suitability for datasets with categorical features.

9. Artificial Neural Network (ANN): Stood out as the most powerful model, achieving the highest F1 score of 0.92 (3-class) and 0.88 (10-class) for predicting intricate patterns in prenatal health data.

HyperParameter Tuning

- In addition to model evaluation, we performed hyperparameter tuning using GridSearchCV for models like SVM, XGBoost, etc., to obtain the best configurations for improved performance.
- *Optimization Parameters: max_iters, max_depth, learning_rate, solver, loss, etc.*

- For the Artificial Neural Network (ANN), we utilized Keras Tuner to determine the optimal layer size and number, striking a balance between computational complexity and achieving a high F1 score in our 3-class classification model.

- Similar tuning was done for 10-class classification model, but on a smaller scale around the tuned parameters of 3-class model.

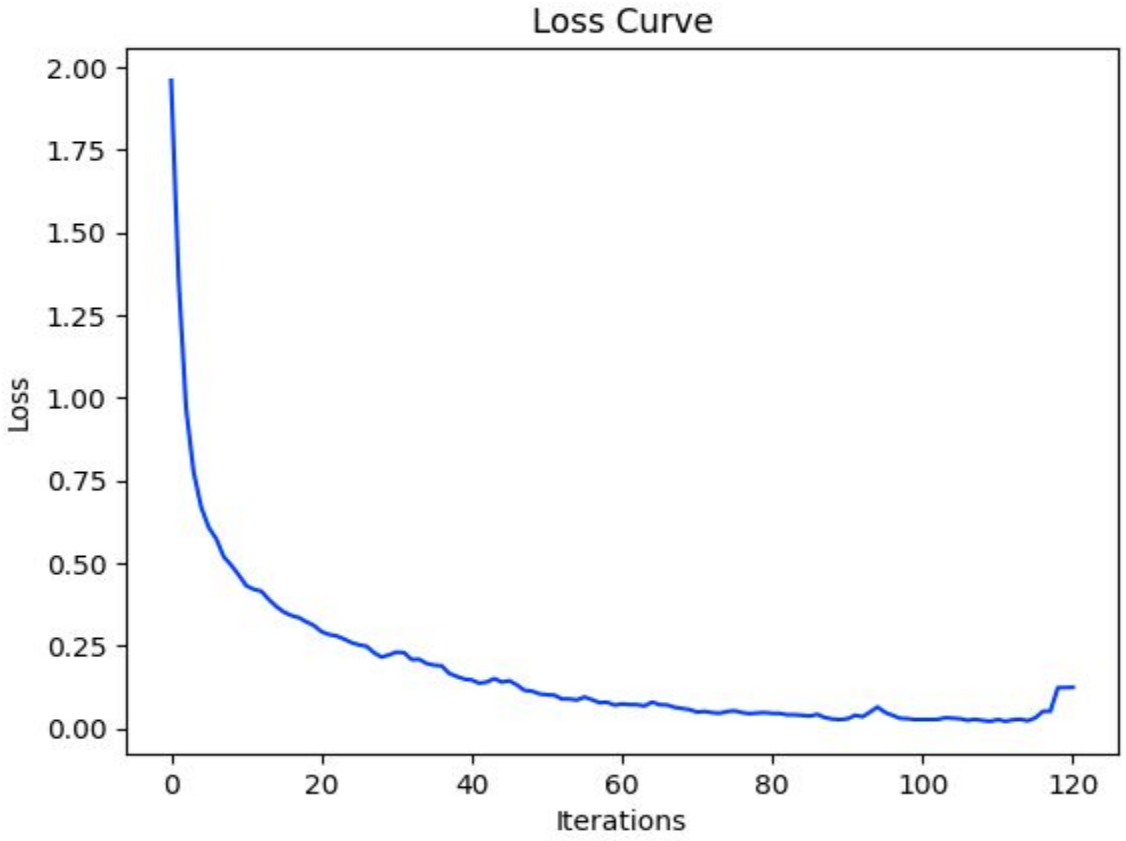
Results (3-Class Foetal Risk)

Model	Accuracy	F1 Score	Recall	Precision
Logistic Regression	79%	0.7 - 0.86	0.67 - 0.86	0.72 - 0.85
Naive Bayes	72%	0.63 - 0.78	0.63 - 0.81	0.63 - 0.79
K-Nearest Neighbors	88%	0.87 - 0.9	0.88 - 0.9	0.87 - 0.91
SVM	**93%**	**0.92 - 0.96**	**0.91 - 0.97**	**0.9 - 0.96**
Decision Trees	100% Tr | 84% Te	N/A	N/A	N/A
XGBoost	88%	0.84-0.9	0.87-0.9	0.85-0.88
Random Forests	85%	0.8-0.86	0.82-0.87	0.82-0.87
ANN	**97%**	**0.96-0.99**	**0.97-0.99**	**0.97-0.99**

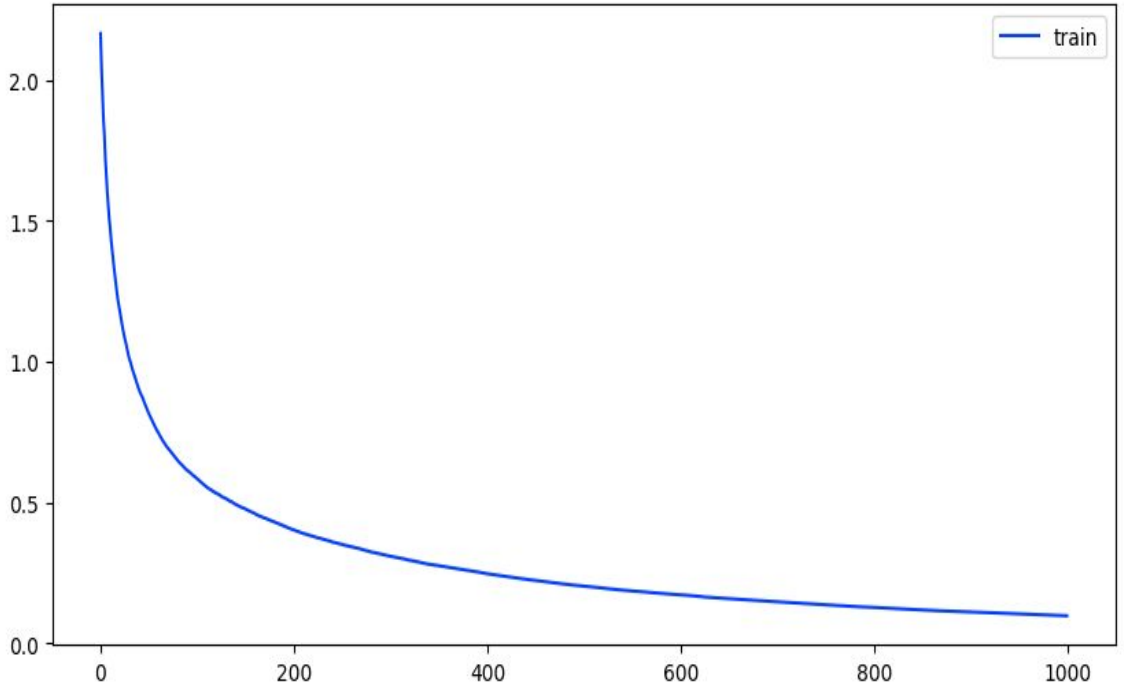
Results (10-Class Morphological Pattern)

Model	Accuracy	F1 Score	Recall	Precision
Logistic Regression	68%	0.64 - 0.66	0.67 - 0.69	0.65 - 0.66
Naive Bayes	57%	0.56 - 0.58	0.57 - 0.59	0.58 - 0.61
K-Nearest Neighbors	76%	0.74 - 0.78	0.75 - 0.77	0.73 - 0.76
SVM	**79%**	**0.79 - 0.83**	**0.8 - 0.81**	**0.78 - 0.8**
Decision Trees	95% Tr | 67% Te	N/A	N/A	N/A
Random Forest	70%	0.69-0.73	0.7-0.72	0.68-0.72
XGBoost	73%	0.72-0.74	0.71-0.73	0.73-0.75
CatBoost	**81%**	**0.81-0.82**	**0.81-0.83**	**0.82-0.84**
ANN	**87%**	**0.88-0.9**	**0.86-0.89**	**0.88-0.92**

Loss Curves



ANN Loss Curve



CatBoost Loss Curve

Note: We could not add an image here due to size and formatting (Click here). Rest of the plots are on the notebook.

Inferences

1. Data Scarcity: Overfitting issues were more prevalent in the 10-class model due to limited data.

2. Deep Learning: ANN's dominance underscores the importance of deep learning in capturing complex health patterns.

3. Data Characteristics: Differences in regularization, PCA usage, and clustering highlight the impact of data distribution and scarcity on performance.

4. Optimized Configurations: Consistent patterns in tuned hyperparameters suggest certain configurations are universally effective, improving accuracy and generalization.

5. Insights Transferability: Insights from both 3-class and 10-class challenges can enhance understanding of prenatal healthcare and predictive modeling.

TimeLine

Past Timeline

Week	Work Done
2	Project Ideation, Literature Review
3	Dataset: *Understanding & Visualisation*
4	Dataset: *PreProcessing*
5	Preprocessing (contd.), Logistic Regression, K-fold CV, PCA
6	Naive Bayes, k-NN, SVM
7	Decision Trees, Pruning

Future Timeline *(After Midsem)*

Week	Work Done
8	Random Forests
9	Boosting, ANN
10	HyperParameter Tuning, GridSearchCV
11	Preparing Another 10-class Model on the basis of Mother's health
12	Analyzing & Integrating Models *(Tentative)*
13	Final Report & Presentation

Contributions

Work	Contributors
Ideation	Aamleen, Sufyan
Literature Review	Mohd. Shariq
Dataset: understanding & analysis, preprocessing, visualisation.	Mohd. Sufyan
Methodology: Models implementation, Optimizing parameters. Future possibilites	Aamleen Ahmed, Mohammad Shariq, Mohd. Sufyan
Analysis of results	Mohd. Sufyan, Aamleen Ahmed
Report formation	Mohd. Shariq

Appendix

- Link for the Code

- Link for the Report