Comparative Analysis of Machine Learning Models for Imbalanced Prenatal Outcome Classification Using National Health Survey Data

Corresponding Author's Name: Dr Sandesh Paul
Corresponding Author's Designation: Assistant Professor
Corresponding Author's Institution, City, Country: Department of Mathematics and Statistics, Banasthali Vidyapith, Rajasthan, India

Additional Author's Name: Dr. Pragya Mishra
Additional Author's Designation: Assistant Professor
Additional Author's Institution: Department of Mathematics and Statistics, Integral University, Lucknow, Uttar Pradesh, India

Abstract

The integration of classification techniques in healthcare, particularly through machine learning (ML) is revolutionising patient data analysis and improving diagnostic accuracy. This study explores the application of various classification algorithms to prenatal care data, utilising a dataset from the National Sample Survey Office (NSSO) that captures household health consumption patterns in India. The dataset categorises outcomes into three groups: MABA (mother alive & live birth), MABD (mother alive & abortion/stillbirth), and OTHR (other outcomes). Through preprocessing methods, including tailored imputation strategies and the SMOTENC technique to address class imbalance, several classification models were evaluated. The models included Multiclass Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbours, Decision Trees, Random Forest, XGBoost, and others. Results indicated that ensemble models, particularly CatBoost and XGBoost, achieved the highest accuracy (0.91) and demonstrated superior performance in detecting minority classes compared to traditional statistical models. Despite improvements in recall for Class 2 due to SMOTENC, challenges persisted for Class 1, highlighting the limitations of linear models in capturing complex data relationships. This research underscores the transformative potential of advanced classification methods in enhancing maternal and fetal health outcomes, paving the way for more personalised and effective healthcare interventions.

Keywords: Personalised Healthcare, Prenatal Care, Machine Learning, Classification, Data Imbalance

Introduction

Classification problems in healthcare utilize algorithms and statistical models to categorize patient data, leading to improved diagnoses and personalized treatment plans. Advanced technologies like machine learning (ML) and artificial intelligence (AI) enable healthcare professionals to analyze vast amounts of medical data—such as symptoms, lab results, and imaging studies to predict disease outcomes and enhance patient care. This approach supports better decision-making, facilitates early detection of conditions, and optimizes resource allocation.

The importance of classification in healthcare lies in its ability to improve diagnostic accuracy and treatment effectiveness. Various methodologies, including logistic regression, decision trees, and support vector machines, convert raw data into actionable insights for early disease detection. Given the complexity of medical data, advanced techniques such as evolutionary-

fuzzy systems are employed to refine classification rules and enhance interpretability.

Overall, the integration of sophisticated classification techniques is vital for advancing healthcare outcomes, enabling timely interventions, and improving patient management. By leveraging ML and AI, healthcare providers can predict patient outcomes, tailor treatment plans, and enhance preventive care strategies. This data-centric approach is transforming healthcare delivery, allowing practitioners to make informed, data-driven decisions that boost patient engagement and satisfaction. It fosters a culture of continuous improvement, where real-time data analysis drives ongoing advancements in patient care practices.

Machine Learning (ML) is making waves in prenatal and perinatal healthcare, offering new ways to identify and predict various conditions early on. For instance, Muthureka et al. (2021) showed how Logistic Regression can effectively identify risk factors for Cerebral Palsy during pregnancy. Similarly, Arayeshgari et al. (2023) used the same method to predict low birth weight, emphasizing the crucial role of maternal health. Sufriyana et al. (2020) explored the complexities of pregnancy outcomes by comparing logistic regression with other ML algorithms. In terms of delivery methods, research by Samson Balogun et al. (2015) created a discriminant function to tailor care based on maternal and infant variables. Kamel et al. (2022) looked into linear discrimination analysis and programming for better classification accuracy. In fetal health monitoring, Chuatak et al. (2023) utilized decision trees to analyze cardiotocography (CTG) data, which is vital for timely medical interventions. Waynforth (2022) employed random forests to identify preterm birth risk factors using data from the UK Millennium Cohort. Advancements continue with Uddin et al. (2022), who developed an ensemble of nearest-neighbour classifiers for high-risk fetal predictions. Gondane and Susheela Devi (2015) introduced the Probabilistic Random Forest for better feature selection. Sari et al. (2023) took it further with a multi-branch random forest method that combines PCA and K-means for improved data classification. Recent studies have also tackled specific conditions. Al Duhayym et al. (2023) achieved an impressive 99% accuracy in classifying fetal health using CTG readings with XGBoost. Hu et al. (2023) created a prediction model for gestational diabetes that outperformed traditional methods. Koivu and Sairanen (2020) highlighted the promise of advanced ML techniques in predicting abnormal pregnancy outcomes. To address class imbalance, Iosifidis et al. (2021) proposed a cost-sensitive boosting method called AdaCC to improve performance for minority classes. Lastly, Hornyák and Iantovics (2023) introduced an enhanced AdaBoost algorithm to boost classification accuracy. Together, these studies illustrate how ML is transforming maternal and fetal health through innovative predictive modeling.

2. METHODOLOGY AND ANALYSIS

2.1 DATASET

The data entitled ***"Household Social Consumption: Health, NSS 75th Round Schedule-25.0: July 2017-June 2018,"*** has been taken from National Sample Survey Office (NSSO) under the Ministry of Statistics & Program Implementation, Government of India. This dataset provides comprehensive insights into household health consumption patterns across India during the specified period. Sponsored by the Ministry of Statistics & Program Implementation, the data collection aimed to capture various health-related metrics, including prenatal care.

2.2 PREPROCESSING OF DATA

The target variable in the pre-natal dataset is categorized into three groups for classification: MABA (mother alive & live birth), MABD (mother alive & abortion/stillbirth), and OTHR (all other outcomes, including maternal death scenarios). This consolidation simplifies the dataset while maintaining essential distinctions for predictive analysis. Missing values are addressed using tailored imputation strategies: numerical variables are filled with the mean, while categorical variables are replaced with the mode, ensuring data quality for robust analysis.

The dataset is divided into training (80%) and testing (20%) sets. The training set allows the model to learn patterns, while the testing set evaluates the model's performance on unseen data, ensuring accurate generalization assessment. To tackle class imbalance, the SMOTENC method is applied, enhancing the representation of minority classes (MABD and OTHR) to align with the majority class (MABA), which has significantly more samples. A Column Transformer within a Pipeline applies specific transformations: One-Hot Encoding for categorical variables and StandardScaler for numerical variables. This ensures categorical data is encoded appropriately, and numerical features are standardized for effective model processing. Classification models are trained on the pre-processed training data, with a separate test set used for evaluation, ensuring the model's effectiveness on unseen data.

2.3 CLASSIFICATION MODELS USED FOR PRENATAL CARE DATA

Multiclass Logistic Regression: Extends binary logistic regression to K classes, predicting the class with the highest probability using a linear combination of features and the SoftMax function for probability calculation. **Linear Discriminant Analysis (LDA)**: A supervised technique that maximizes class separability by modeling relationships between features and class labels through mean vectors, scatter matrices, and optimization objectives based on the Fisher criterion. **K-Nearest Neighbors (KNN)**: A method that classifies data by identifying the K closest neighbors in feature space using Euclidean distance, predicting based on majority voting among neighbors. **Decision Tree Classifier**: Utilizes recursive binary splitting based on impurity measures (Entropy, Gini Index) to create subsets that improve classification purity. **Random Forest**: An ensemble method that combines multiple decision trees through bagging, training on different data subsets, and using random feature selection to enhance diversity and reduce overfitting. **XGBoost Classifier**: Applies gradient-boosted decision trees to optimize a differentiable loss function, minimizing residual errors through iterative updates based on second-order derivatives of the loss. **CatBoost Classifier**: Designed for categorical data, it builds decision trees sequentially to minimize errors while addressing overfitting and prediction bias. **AdaBoost Classifier**: Combines weak classifiers to form a strong model, adjusting sample weights iteratively to focus on misclassified instances. **Gradient Boosting Classifier**: Sequentially trains models to correct predecessors' errors by optimizing a loss function using gradient descent, combining contributions from all weak learners for final predictions. These models enhance classification accuracy in pre-natal datasets, improving predictions for maternal and fetal health outcomes.

3. RESULTS

3.1 The performance of each model applied in analyzing the prenatal dataset is shown in Table 3.1

Table 3.1 Model's Performance Table

Model	Accuracy	Class	Precision	Recall	F1-Score
Logistic Regression	**0.59**	**0**	***0.95***	*0.61*	***0.74***
		1	0.06	0.27	0.1
		2	0.13	***0.68***	0.21
KNN	**0.65**	**0**	***0.95***	0.68	***0.79***
		1	0.07	***0.34***	0.12
		2	0.14	0.48	0.22
Decision Tree	**0.74**	**0**	***0.95***	0.77	***0.85***
		1	0.09	0.19	0.13
		2	0.16	***0.67***	0.25
Random Forest	**0.89**	**0**	0.94	**0.95**	***0.95***
		1	***0.2***	0.13	0.16
		2	*0.37*	0.4	***0.39***
Linear Discriminant	**0.56**	**0**	***0.95***	0.56	***0.71***
		1	0.06	0.31	0.09
		2	0.13	***0.7***	0.23
XGBoost	**0.91**	**0**	***0.93***	***0.97***	***0.95***
		1	0.19	0.06	0.09
		2	***0.42***	0.39	***0.4***
CatBoost	**0.91**	**0**	***0.93***	***0.98***	***0.95***
		1	***0.32***	0.06	0.11
		2	0.41	0.37	0.39
AdaBoost	**0.77**	**0**	***0.95***	0.8	***0.87***
		1	0.07	0.16	0.1
		2	0.22	***0.66***	0.33
Gradient Boosting	**0.89**	**0**	0.94	***0.95***	***0.94***
		1	***0.27***	0.06	0.09
		2	0.31	***0.55***	***0.4***

3.2 Summary of the Analysis

(i) Overall Accuracy:
Gradient boosting models (XGBoost, CatBoost, GradientBoosting) perform the best in terms of accuracy (above 0.89), with CatBoost achieving the highest accuracy (0.91). Logistic Regression and Linear Discriminant Analysis show the lowest accuracy (0.59 and 0.56, respectively), likely due to their limitations in handling complex data structures.
(ii) Class 0 Dominance: All models achieve high precision, recall, and F1-scores for Class 0, as it is the majority class with the largest support (5911 samples). Tree-based models (e.g., Random Forest, CatBoost) perform particularly well for this class.
(iii) Challenges with Class 1: All models struggle with Class 1, showing low recall (maximum 0.34 for KNN) and poor precision. Even after SMOTENC, this class remains difficult to classify. CatBoost achieves slightly better precision (0.32) for Class 1, though recall remains low (0.06).
(iv) Improved Recall for Class 2: SMOTENC has helped models perform better in Class 2, as seen in the higher recall values (up to 0.70 for LDA and 0.67 for AdaBoost). Tree-based models like Random Forest and Gradient Boosting also achieve decent F1 scores for this class.
(v) Model Behaviour: Statistical Models show limited improvement for minority classes despite SMOTENC, as they struggle with non-linear decision boundaries. Tree-Based Models: Random Forest, AdaBoost, and Gradient Boosting handle imbalanced data better, particularly for minority classes, due to their ability to adapt to synthetic samples created by SMOTENC. Ensemble Models stand out for their balance of precision and recall, particularly for Class 2.
(vi) Trade-Off Between Precision and Recall: Precision for minority classes (Classes 1 and 2) is often lower due to the inclusion of synthetic samples, which may lead to false positives. However, recall improvements indicate models are detecting more minority class samples correctly.

4. CONCLUSIONS:

The linear nature of Statistical models limits their ability to effectively capture complex relationships within the data, leading to poorer performance. Despite the application of SMOTENC to address class imbalance, both models showed limited improvement in their performance, particularly for the minority classes. The application of SMOTENC has significantly improved recall for minority classes, particularly Class 2. However, challenges remain for Class 1, where both recall and precision are low across all models. Ensemble methods like CatBoost and XGBoost outperform other models, achieving the best balance between accuracy and minority class performance.

REFERENCES

Muthureka, K., Srinivasulu Reddy, U., & Janet, B. (2021). Implementation of Multivariate Logistic Regression Model for Cerebral Palsy Identification using Prenatal, Perinatal Risk Factors.

Arayeshgari, M., Najafi-Ghobadi, S., Tarhsaz, H., Parami, S., & Tapak, L. (2023). Machine Learning-based Classifiers for the Prediction of Low Birth Weight.

Samson Balogun, O., Janet Akingbade, T., & Oguntunde, P. (2015). An Assessment Of The Performance Of Discriminant Analysis And The Logistic Regression Methods In Classification

Of Mode Of Delivery Of An Expectant Mother.

Sufriyana, H., Husnayain, A., Chen, Y. L., Kuo, C. Y., Singh, O., Yeh, T. Y., Wu, Y. W., & Su, E. C. Y. (2020). Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: Systematic review and meta-analysis.

Kamel, maie, Salem, H., & Abdelgawad, W. A. (2022). An Application of Linear Programming Discriminated Analysis for Classification.

Chuatak, J. V. Y., Comentan, E. R. C., Moreno, R. L. H. G., Billones, R. K. C., Baldovino, R. G., & Puno, J. C. V. (2023). A decision tree-based classification of fetal health using cardiotocograms.

Waynforth, D. (2022). Identifying Risk Factors for Premature Birth in the UK Millennium Cohort Using a Random Forest Decision-Tree Approach.

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance Analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction.

Gondane, R., & Susheela Devi, V. (2015). Classification using probabilistic random forest.

Sari, J. N., Madona, P., Kusryanto, H., Zain, M. M., & Valzon, M. (2023). Electrocardiogram signals classification using random forest method for web-based smart healthcare.

Al Duhayyim, M., Abbas, S., Al Hejaili, A., Kryvinska, N., Almadhor, A., & Mughal, H.(2023). Ensemble Learning for Fetal Health Classification.

Hu, X., Hu, X., Yu, Y., & Wang, J. (2023). Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm.

Koivu, A., & Sairanen, M. (2020). Predicting risk of stillbirth and preterm pregnancies with machine learning.

Iosifidis, V., Zhang, W., & Ntoutsi, E. (2021). *Online Fairness-Aware Learning with Imbalanced Data Streams*.

Hornyák, O., & Iantovics, L. B. (2023). AdaBoost Algorithm Could Lead to Weak Results for Data with Certain Characteristics.