Python Machine Learning and Explainable AI (XAI) Based Preterm Birth Prediction

Kazi Safkat Taa Seen
Department of Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85721
safkat@arizona.edu

Abstract

Infants born preterm are the most susceptible to neonatal mortality and morbidity and hence preterm births (PTBs) are immensely burdensome to impacted families, economy and education system. Due to a number of limitations, PTB screening tests that are currently used in practice lack the capability to accurately predict PTBs. Due to their ability to model complex non-linear relationships and make accurate predictions based on data, machine learning algorithms are being widely applied to medical research in recent times. In this study, we aim to predict PTB in the early stages of pregnancy by utilizing readily available maternal factors through machine learning techniques. A number of feature selection techniques have been implemented in this work to improve usability and computational efficiency of models. To accurately predict PTB, several machine learning classifier algorithms were explored. Random search, a hyperparameter tuning technique, has also been employed to optimize performance of models. It was found that the XGBoost model was the best performing with a test accuracy of 71.22%. In addition to making accurate predictions, we need to understand the contribution of PTB risk factors to effectively address PTB in its early stages and to make predictions reliable. For this, we have utilized Shapley Additive Explanations (SHAP) to provide us with reliable explanations that aided us in understanding which risk indicators or features contribute towards a classification. From SHAP, we conclude that health of the amniotic sac, prenatal care and household income are the major contributing factors towards a prediction. The novelty of this study lies in its attempt to improve the effectiveness of perinatal care by identifying the risk of PTB early in the gestation period.

1 Introduction

Preterm birth (PTB) is the primary reason for infant deaths and bodily defects worldwide Walani (2020). More than 10% of babies born in the year 2010 were preterm. According to a study, half of all perinatal deaths are caused by PTBs. PTB primarily refers to babies born before 37 weeks of gestation Goldenberg et al. (2008). According to its severity, PTB can be classified into 4 different categories Raja et al. (2021):

- Extreme preterm birth (28 weeks of gestation)
- Very preterm birth (28 to 32 weeks of gestation)
- Moderate preterm birth (32 to 34 weeks of gestation)
- Late preterm Birth (34 to 37 weeks of gestation)

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Health maintenance of preterm infants is usually costly. 10% of all babies are born preterm in the US, which costs the healthcare system at least $26 billion yearly Russell et al. (2007). Similarly, about 8% of all births in Canada are preterm and cost $580 million annually Shah et al. (2018). Postnatal hospitalization time for a preterm baby is 16 times higher than that of a normal baby Petrou et al. (2003). Also, women who experience preterm delivery are vulnerable to health concerns and might require costly treatments. Besides financial loss, PTB deaths are also emotionally burdensome for a family. Moreover, infants born preterm tend to have recurring complications that may lead to fatal health concerns and long-term disabilities. Such disabilities include mental retardation, vision impairment, chronic lung disease and cardiovascular diseases Greenough (2012). Studies show preterm babies tend to have neural and behavioral complications as well Fergus et al. (2013). People born preterm also tend to be academically inept.

Detecting PTBs in the early stages of gestation can be an effective solution to prevent or mitigate PTBs. Commonly used screening tests for PTBs fail to analyze the multiple factors and complexities associated with PTB detection Tran et al. (2016), Esplin et al. (2008), on Good Clinical Practice in Maternal-Fetal Medicine et al. (2019). These tests are also inaccurate and fail to provide proper understanding and explanation of the results Georgiou et al. (2015). Recently, machine learning approaches have been found to be effective in the field of health and medical science Ravindra et al. (2023), Kokkinidis et al. (2023). This is due to their ability to learn patterns in complex dimensions and make accurate predictions correctly Koivu and Sairanen (2020). Explainable artificial intelligence (XAI) extends the popularity of machine learning in medical science by providing accurate explanations (and hence proper understanding) behind model predictions and suggesting dominant features of a classification Kokkinidis et al. (2023).

In this paper, we aim to effectively predict premature birth using several important features correlated to preterm birth. We also look to explain features affecting predictions so that accurate steps can be taken to mitigate preterm birth. We utilize several feature selection techniques such as variance threshold, Pearson's correlation, and mutual information to reduce the number of required features to avoid redundancy and achieve computationally efficient models. State-of-the-art machine learning models such as K-NN (K-nearest neighbors), Random Forest, Decision Tree, XGBoost, and SVM are also put to use to learn patterns and accurately make predictions.

2 Related Works

Prior research has attempted to assess the clinical risk factors associated with preterm delivery as well as the prediction of PTB. Despite the abundance of studies attempting to develop prediction models to estimate the risk of PTB, only a few efficient, affordable, and low-risk interventions are available Jehan et al. (2020), Espinosa et al. (2021), Moufarrej et al. (2022). At present, PTB history is thought to be the biggest risk factor for preterm delivery Esplin et al. (2008). Additional risk factors include smoking, age, short inter-pregnancy intervals, previous cervical surgery, multiple gestations, and poor nutritional condition. Still today, anticipating PTB continues to be a difficult task. Only a few imaging tests for PTB prediction exist. Currently, transvaginal ultrasound is utilized for routine prenatal check-ups and a low cervical length on this examination between weeks 18 and 24 of present gestation is a key risk factor for preterm delivery Iams et al. (1996). Besides, it is known that fetal fibronectin (a biomarker) can predict preterm delivery in 7–14 days Hezelgrave et al. (2016).

Three primary sets of screening tests are available for PTB prediction: biochemical evaluation, risk factor assessment, and cervical measurement. However, not all methods have the potential to be applied safely and economically to the advantage of clinical predictions on Good Clinical Practice in Maternal-Fetal Medicine et al. (2019). Also, they might not be adequate to identify true-positive PTB cases. For instance, biochemical evaluation is an expensive process that might cause the expectant mother bodily and psychological stress. Another often employed strategy is risk factor assessment, a data-driven approach that is time-consuming, costly, and might potentially overlook several plausible risk variables that escaped the notice of researchers leading to the testing of hypotheses. Also, prior PTB history is a dominating risk factor with a relative risk of 13.56 and hence nulliparous women go unnoticed Tran et al. (2016), Esplin et al. (2008). These results demonstrate the ineffectiveness of the present approaches for anticipating high-risk pregnancies, particularly in first-time mothers.

A few predictive systems have also been explored utilizing a variety of data, such as maternal demographics, medical and obstetrical history, and well-known risk factors Mercer et al. (1996),

Lee et al. (2011). However, their predictive ability was limited. This constraint could arise from the fact that they frequently depend on basic linear statistical models that are unable to adequately represent intricate issues like PTB. It is known that risk factor evaluation using traditional methods is inadequate since it fails to detect more than 50% of PTB pregnancies Georgiou et al. (2015). Finding new screening measures to fill the void left by traditional prediction methods is therefore crucial, as it aids prenatal care by preparing for any early actions that could be needed in the event of a poor prognosis. In addition to avoiding needless and occasionally expensive interventions in patients at lower risk, identifying patients at higher risk for preterm birth would enable the development of effective interventions to prevent the negative perinatal outcomes associated with preterm birth Conde-Agudelo and Romero (2014). An accurate prognosis of preterm delivery may result in more thorough and effective prenatal care.

In the recent past, machine learning (ML) techniques have been employed to enhance individual risk prediction beyond conventional models. Since the etiology of spontaneous preterm birth is complicated and involves intricate interdependencies, good prediction requires a range of factors to be related in complex and non-linear ways. Complex nonlinear relationships between PTB risk factors and PTB may be modeled using a variety of ML techniques. Because of their unique ability to make accurate predictions based on data, ML algorithms do not need to be programmed explicitly to operate Koivu and Sairanen (2020). Hence, ML is a good choice for intricate tasks like PTB prediction. According to recent review articles Akazawa and Hashimoto (2022), Sharifi-Heris et al. (2022), ML models revealed potential in accurately predicting PTB. Nevertheless, a sizable amount of data (big data) is required for ML techniques to produce reliable models i.e., ones with high predictive accuracy. Hence, if applied appropriately, ML can improve the effectiveness of prenatal care.

Developing new strategies to lower the risk of PTB is one of the primary objectives of obstetric care. Achieving this objective may be aided by digital tools and initiatives. A work published in 2011 presented the development and validation of logistic regression models to predict PTB Allouche et al. (2011). More recently, the QUiPP application attempted to predict PTB within clinically meaningful time frames based on a pregnant woman's current pregnancy data, medical history, and predictive clinical testing Dehaene et al. (2022). QUiPP is an algorithm-based tool created in 2016 and it combines cervical length, quantitative fetal fibronectin, demographic data as well as obstetric history to make predictions about PTB Kuhrt et al. (2016). Based on data for 1249 pregnant women, QUiPP's algorithmic performance AUC ranged from 0.77 to 0.99. Another example is the PredictPTB model AlSaad et al. (2022) which is a deep learning model that uses data from electronic health records (EHRs) to predict PTB. Furthermore, a sophisticated system (built upon the SVM algorithm) with an ability to predict PTB has also been proposed Moreira et al. (2018). Latest research on PTB prediction incorporates explainable AI and combines predictions with reasoning to produce better performing models and to improve suitability for clinical use Ravindra et al. (2023), Kokkinidis et al. (2023). Amongst them, a particular work is similar to ours Kokkinidis et al. (2023). It attempts to utilize AI-based models to predict PTB although on a different set of data. Some of the similarities include the uses of maternal factors for model training and testing, coinciding ML classifier algorithms (SVM, Random Forest, XGBoost) to make accurate predictions, 5-fold cross-validation to prevent overfitting, hyperparameter tuning for model optimization, and XAI for interpretations.

3 Procedure

This section describes the complete methodology to develop appropriate models for predicting preterm birth. We used several procedures to train and explain models appropriately. Fig 1 visually represents the methodology employed in this work. Firstly, data is preprocessed to remove null, anomalous, and duplicate values. We further analyze the data to fill in missing values and drop features available late or after birth. We cross-validate the data using 5-fold cross-validation. We treat each train fold with 3 feature selection methods: variance threshold, Pearson's correlation, and mutual information. We remove redundant and unimportant features from each fold's train and test set. We train different models such as Decision Tree, k-NN, XGBoost etc. We tune our models using random search cross-validation. We evaluate the models on the test set using different performance metrics and find the best model in all folds. We use the best to explain predictions made by the model using the SHAP algorithm.

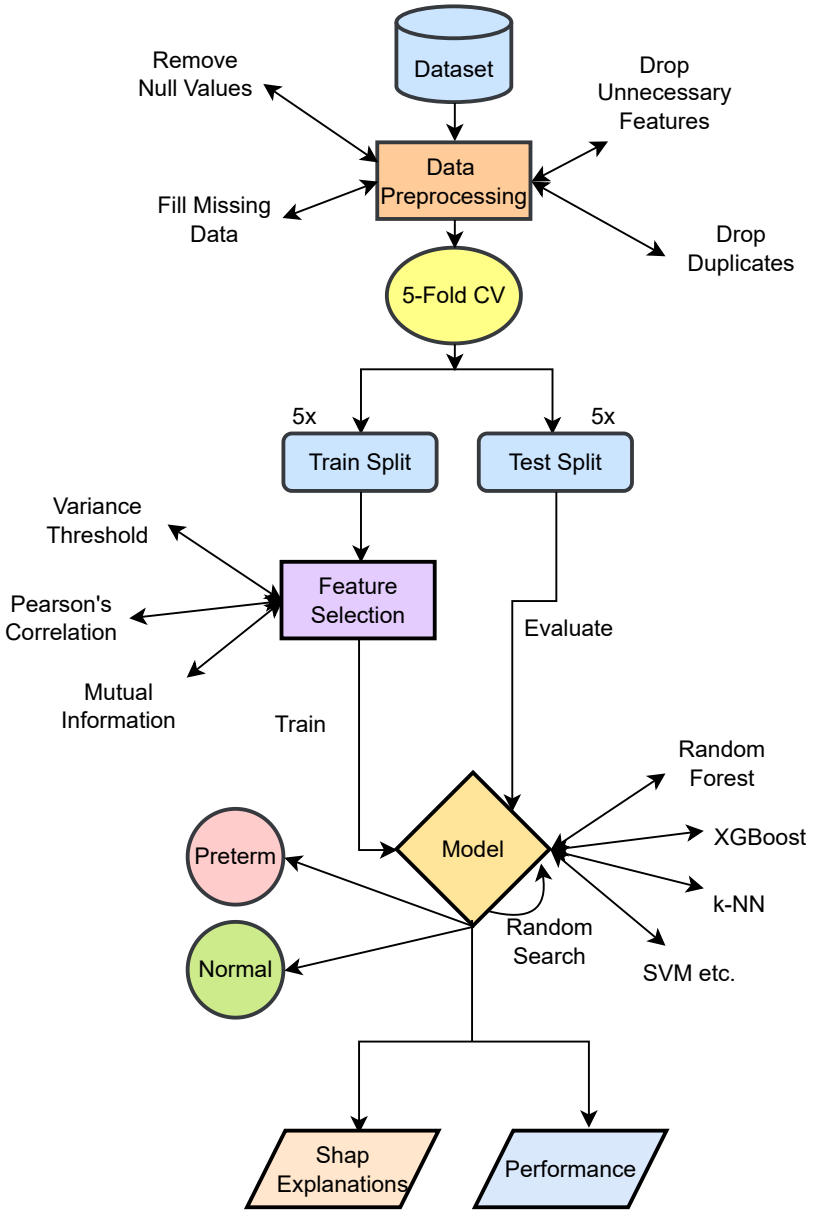

Figure 1: The complete methodology followed in the study to make predictions.

3.1 Data Preprocessing

The dataset Santos et al. (2020) we used for this study was based on the maternal factors of women living in the Brazilian Amazon. The dataset contains about 800 instances with balanced classes and various features (78 features). Data preprocessing was carried out fill ore remove missing values, drop features, and to analyze the dataset completely. Some features could only be collected after the completion of the gestational period. We removed every feature that might only be available after the completion of gestational age such as baby weight, Apgar 1, Apgar 5, delivery type, gestational age, and cause of prematurity. We removed features that do not contribute to a prediction such as the record identity. We also removed duplicate values. The dataset also consists of several null values. Removing all the instances with null values removes half of the dataset. So, we decided to analyze further and process the data. The feature that contained if the mother had a "previous cesarian birth" had the most null values. By analyzing we realized the missing values represented mothers who did not have any pregnancies previously. This can be understood by comparing the "premature child previous" feature with the "previous cesarian birth" feature. So, we filled the feature "previous cesarian birth" with another value that represents mothers who did not have a previous pregnancy. We filled the null values of the feature "BMI" and "household income" with the mean of all the average values. We removed other features that had too many null values. After this preprocessing, we were left with 696 instances and 61 features.

3.2 5-Fold Cross-Validation

Since limited data was available for analysis, the 5-Fold cross-validation method was chosen over the simple train-test split for robustness. Cross-validation offers a number of benefits Fushiki (2011). Firstly, cross-validation aids in the prevention of overfitting by offering a more reliable estimation of the model's performance on unseen data. While comparing various models, cross-validation enables the selection of the optimal model based on average performance. Besides, it facilitates the optimization of a model's hyperparameters by choosing the best-performing values on the validation set. Also, cross-validation is more data-efficient than traditional validation procedures since it uses all the available data for training and testing. In this work, we employed the k-fold cross-validation technique over Leave-one-out cross-validation (LOOCV) to evaluate model performance since LOOCV can be time-consuming. Based on the size of the dataset, a total of 5 folds were deemed appropriate.

3.3 Feature Selection

We carried out several feature selection techniques on all the features to remove redundant and unnecessary features. Feature selection is carried out improve performance and achieve computational efficiency Chandrashekar and Sahin (2014). We implement the method of feature selection only on the train set. This is because the test set should not have biases as it is to be used for evaluating the performance of the model. We describe below briefly the procedures used in this study.

3.3.1 Variance Threshold

Features with a high homogeneity do not contribute to predictions. Features with same values for almost every instance do not help the model classify or come to decisions. So a feature with only a singular value can be ignored. So the variance threshold Fida et al. (2021) removes any feature with high homogeneity. All the features with high homogeneity were removed from the dataset.

3.3.2 Pearson's Correlation Coefficient

The Pearson's Correlation Coefficient Cohen et al. (2009) is used to calculate the linear correlation between two values. The Pearson's Correlation Coefficient is usually between +1 and -1. A high coefficient means the values have a high positive correlation. A low coefficient signifies a high negative correlation. Values with a correlation coefficient of zero signify no correlation between values. We use Pearson's Correlation Coefficient to remove redundant features as it will help reduce computation. We choose a coefficient of 0.7 as a threshold to remove features. The coefficient is calculated between two features and if the coefficient is higher that 0.7 one of the features are

removed. The equation for the Pearson's Correlation is given below:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

Here, r is the correlation coefficient. x_i represents values of the the feature x in the sample. $\bar{x}$ represents the mean value of the feature x. y_i represents values of the the feature y in the sample. $\bar{y}$ represents the mean value of the feature y.

3.3.3 Mutual Information

Mutual information Ross (2014) is a statistical tool utilized for finding the relation between two groups of values. The mutual information finds out non-linear correlation between two variables, It calculates how understanding one variable X reduces the uncertainty of the output variable Y. It is usually calculated using the nearest-neighbor method Ross (2014). We calculated the mutual information of the related feature against the label and kept the top 20 features with the highest mutual information.

The model was finally trained on 20 distinct features. The features and their respective mutual information to the labels are illustrated in figure 2 for every fold. The data is now ready to be fed to the models.

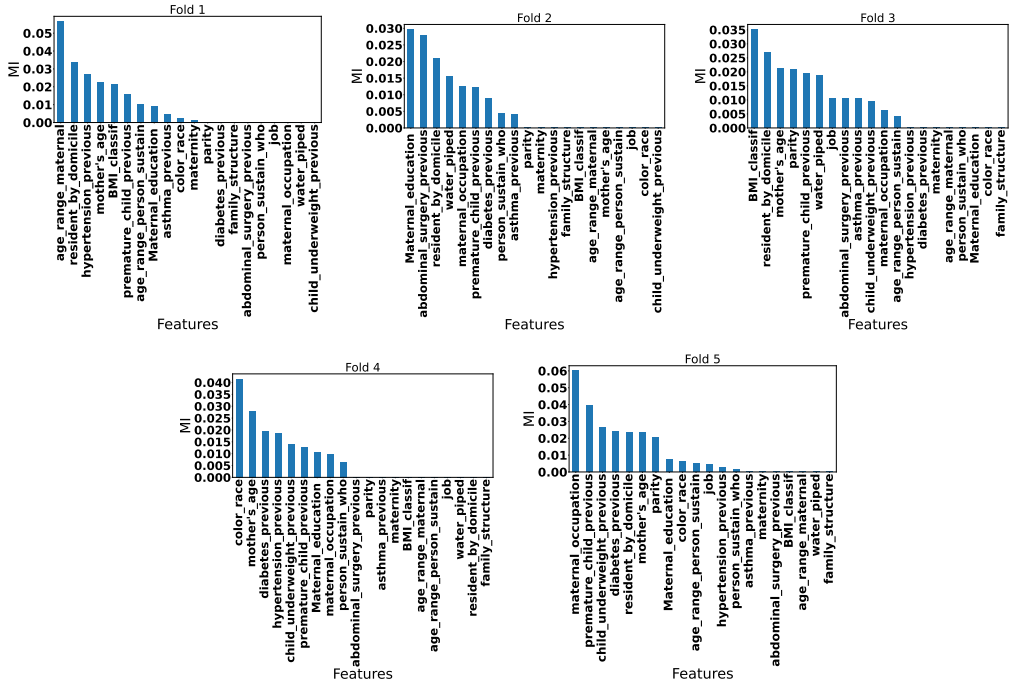

Figure 2: The figure illustrates the 20 features obtained after feature selection by mutual information on the x-axis and the mutual information score with the label for all the 20 features on the y-axis. All features with a mutual information of zero was not removed because mutual information is not able to capture all complex relations. So features other than those with mutual information higher than zero were randomly selected to make 20 features.

4 Algorithms

This section describes the algorithms/models used in this study and their use case.

- **Decision Tree:** A commonly used tree-based machine learning algorithm used mainly for supervised classification and regression. The decision tree algorithm Song and Ying (2015) learns a function that can be represented as a set of if-else-then statements. The decision

tree utilizes features to sort down and classify instances via the means of a tree Pranto et al. (2020). As we sort it down, the tree starts from a root node containing the whole dataset. Each node, as we go down, specifies a certain feature according to which data is to be split. The idea is to choose a feature that creates the most homogeneity thereby maximizing information gain3. Homogeneity refers to how pure the split is between the classes when considering a certain feature. We select a feature for each node using an information gain function, which is usually calculated by the entropy metric 2.

$$H(x) = -\sum_{i=1}^{n} p(x_i, k) \log_2 p(x_i, k) \tag{2}$$

Here, $H(x)$ represents the entropy where $p(x_i, k)$ is the proportion of data points that belong to class k from a certain number off data points. Using this entropy the information gain is calculated. The gain is represented as follows:

$$Gain(S, A) = S - \sum \frac{|S|}{|S_v|} S_v \tag{3}$$

Here A is the attribute $|S|$ is the entropy of the dataset sample. $|H_V|$ is the number of instances in the subset that have the value v for attribute A. The depth of the tree is an important hyperparameter in the decision tree that decides how many good features we look into. A higher depth might overfit and a lower depth may underfit. The decision tree algorithm is mainly used for solving decision-related problems. Utilization of multiple features to come to a decision makes the decision tree effective for complex problems. However, due to their simplicity they are prone to variance.

- **k-NN:** k-NN or k-Nearest Neighbours Peterson (2009) is a distance-based algorithm used mainly for supervised learning. The K-NN algorithm during training maps every training sample into an **n**-dimensional feature space. After receiving a query the K-NN algorithm calculates the distance of the query against all the samples in the feature space and considers the k nearest neighbors/samples to produce a prediction. The class of the query is decided using a majority voting algorithm. The algorithm computes the total number of instances of each class present in the k-nearest samples. It predicts query to be of the class with highest number of total instances present in the sample. k is a hyper-parameter for this algorithm. We choose the optimum k value by doing random search so that the algorithm does not overfit or underfit. The best k-value after random search was $k = 3$. k-NN algorithm usually uses the Euclidean Distance metric to calculate the distance between the query and samples in the vector space. The equation for the Euclidean distance is given below:

$$E_d\left(p, q\right) = \sqrt{\sum_{i=1}^{n} \left(q_i - p_i\right)^2} \tag{4}$$

Given two points p and q we calculate the Euclidean distance by taking the difference square of all the coordinates, n, of the two points. Then we sum the squared differences and find it's square root. The k-NN algorithm is best used when samples of similar classes form clusters in a given vector space.

- **SVM:** The SVM (Support Vector Machine) is a supervised classification algorithm. The SVM is trained to learn an optimal hyperplane often known as a linear decision boundary 5 to separate two or more classes in a feature space.

$$LDB = w^\top x_+ b \tag{5}$$

In the equation 5, x is the input feature, w is the normal to the boundary and b is the distance from the origin to the hyperplane. SVMs are used in tasks where the number of attributes are larger than the number of training instances.

- **Random Forest:** The Random Forest algorithm is an ensemble of decision trees that utilizes the bagging algorithm for training Géron (2022). The Random Forest classifier is robust and efficient. Its efficiency and robustness makes it particularly great for large datasets, imbalanced datasets, and even when a lot of data is missing. For the Random Forest Algorithm, a N number of trees have to chosen to be part of the ensemble. Similar to the

decision tree we also decide what number of features (depth) all the trees in the ensemble have to learn. For every tree N, a subset of size M is prepared by random sampling. Then the decision trees are built with the subset and F random features equal to the depth. The tree is then trained using information gain of the random subsets. A majority voting is done at the end to choose the correct prediction among all the trees in the ensemble.

- **Naive Bayes:** The Naive Bayes model utilizes considered probabilities and probabilities obtained to make classifications. The probability of a child being born preterm given a set of attributes can be figured out using Bayes Theorem. The assumption we make is that the prediction and the attributes are independent, which is why it is called Naive.

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y), ..., P(x_n|y)P(y)}{P(x_1), ..., P(x_n)} \tag{6}$$

  Naive Bayes is a simple model so it is efficient and can handle large amount of data. The Naive Bayes also works good with data with many features.

- **XGBoost:** XGBoost Chen and Guestrin (2016) is an ensemble of decision tree that applies the boosting weak learners (CARTs or decision trees) using gradient descent. This technique of learning is also known as the gradient-boosting ensemble technique. In the gradient boosting technique for XGBoost, we take a weak learner decision tree, D_1 and try to make the best predictions by minimizing entropy to get the best splits. After this we calculate a differentiable loss, L_o of the tree. We add another decision model, D_2, sequentially to reduce this loss by gradient descent. This is done my parameterizing the tree D_2 and changing it in accordance to the gradient. Usually an L1 and L2 loss function is used for loss and gradient calculation. The XGBoost algorithm is very good in modelling with structured data such as tabular data. Parallelization in XGBoost allows it to work on large datasets with speed and efficiency. The XGBoost model is effective in learning missing data and is not affected by anomalies. So XGBoost is very good for real life data. Boosting in XGBoost prevents high variance and bias.

4.1 Hyperparameter Tuning using Random Search Cross Validation

The machine learning algorithms contain several hyperparameters that determine the performance of the model. Adjusting hyperparameters according to the data is required to get an optimal and efficient model. Hyperparameters are tuned through several trials. This requires heavy computational power. As a result, we look into utilizing Random Search Cross Validation Bergstra and Bengio (2012) for hyperparameter tuning. The random search cross validation randomly chooses some hyperparameters from a large search space. This does not always provide optimal models but provides good enough results given a very large search space.

4.2 SHAP (SHapley Additive exPlanations)

Interpretation in machine learning is important to understand the relationship between features and outputs. Interpretability can also help to understand what changes in features may lead to a change in prediction and which features are the most important. A common and effective explainable technique is SHAP Lundberg and Lee (2017). SHAP explains a prediction by calculating feature contribution to a prediction. The SHAP exlpanation techniques calculates this contribution by the means of Shapley values from coalition game theory. The features are taken as players of the coalition to make explanations. The Shapley values are used to fairly seperate the weight of prediction among features Hansen (2021). The explanation of SHAP can be presented by the equation below:

$$g(z^{'}) \ = \ \phi_0 + \sum_{j=1}^{M} \phi_j z_j^{'} \tag{7}$$

Here, the function $z^{'}$ is the simplified feature, ϕ_j is the feature attribution for any feature j, and M is the maximum size of the simplified feature. These make the explanation model g Hansen (2021).

5 Evaluation

To ensure a fair comparison, we evaluated every model for a particular fold on its respective test fold. The test set used for evaluation was not directly processed by feature selection to avoid data leaks. We ensured the data did not contain any features that may directly contribute to predictions such as type of prematurity. During evaluation, we considered several performance metrics to ensure the model was robust, such as accuracy, precision, recall, and F1-score.

We showcase the fold accuracy in Table 1 to show the best model obtained and the best model for each algorithm. We also compare the overall model performances in Table 2 to show which algorithm worked the best for the study. We discuss the results obtained in the next section.

6 Results

6.1 Model Performance

This section focuses on the performance achieved by the trained models in this study. Table 1 shows the test accuracy of each of the model used in this study for every fold (5 folds). The train set and test set for every fold is the same for all the models. By analyzing, we can understand the split in Fold 4 is optimal for learning correctly when considering tree-based models. As all tree-based models perform optimally in Fold 4. Splits at Fold 2 and 3 are the least optimal as all models are performing bad on those split. We decide to use the XGBoost model at Fold 4, which has a test accuracy of 71.22%, for inference and model explanations (discussed later).

Models	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Decision Tree	56.43	58.27	56.83	**65.47**	58.99
k-NN	**61.43**	54.67	51.79	56.11	50.35
Naive Bayes	**69.28**	51.80	56.83	66.90	65.47
SVM	59.29	43.88	56.12	63.31	52.51
Random Forest	61.42	56.11	60.43	**70.50**	63.30
XGBoost	65.71	60.43	59.71	**71.22**	64.03

Table 1: Fold accuracy for all the models used in this study. The bold cases refer to best performance of a model among all folds. Higher value means better performance.

Table 2 shows the average performance of the models used in this study. Among all the models, the k-NN algorithm only highly overfits the data as it has a high training accuracy and a low testing accuracy. The ensemble techniques (Random Forest and XGBoost) work better on the data as opposed to techniques that map data points to a feature space such as k-NN and SVM. Along with that, the performance of ensemble techniques shows that the data points are not together or in a cluster. As a result, calculations to make classes separate are not possible. The overall performance in all areas of ensemble techniques are better than the other algorithms used. Although the Naive Bayes and Random Forest models perform reasonably, by looking at the precision and recall, we can understand the models are only able to predict positive class. However, these models fail to work good for the negative class. We utilize the F1 metric to showcase the overall performance of the models as the F1 score demonstrates the ability of the model to predict both false positives and true positives. The F1 score of most of the models in the table 2 have an F1 score of 0.5 or less than that. This showcases that the model incorrectly predicts 50% of either false positves or true positives. However, the XGBoost model has a good overall average F1 score of 0.7%. This means the model is able to correctly classify more than 70% of both true and false positives.

6.2 Model Explanation

We used the XGBoost model in Fold 4, addressed in Table 1 for model explanations as it is the best-performing model. Figure 3 shows the most contributing features or the most contributing feature for the whole dataset on the XGBoost model. Overall, we can understand women with abnormal amniotic sack is prone to preterm, family with less household income is prone to preterm and prenatal

Models	Train Accuracy	Test Accuracy	Precision	Recall	F1 score
Decision Tree	64.94 ± 2.64	59.20 ± 3.27	0.63 ± 0.08	0.40 ± 0.10	0.47 ± 0.05
k-NN	$\mathbf{70.22} \pm 1.33$	54.87 ± 3.86	0.53 ± 0.06	0.45 ± 0.04	0.49 ± 0.03
Naive Bayes	64.33 ± 2.12	62.06 ± 6.63	0.65 ± 0.04	0.44 ± 0.09	0.52 ± 0.08
SVM	55.93 ± 1.34	55.02 ± 6.61	0.56 ± 0.09	0.24 ± 0.07	0.33 ± 0.08
Random Forest	67.78 ± 2.65	62.35 ± 4.71	$\mathbf{0.66} \pm 0.05$	0.43 ± 0.03	0.52 ± 0.03
XGBoost	69.61 ± 4.41	$\mathbf{64.22} \pm 4.14$	0.63 ± 0.07	$\mathbf{0.74} \pm 0.03$	$\mathbf{0.68} \pm 0.05$

Table 2: Performance metrics obtained for the models in this study. Higher value means better performance.

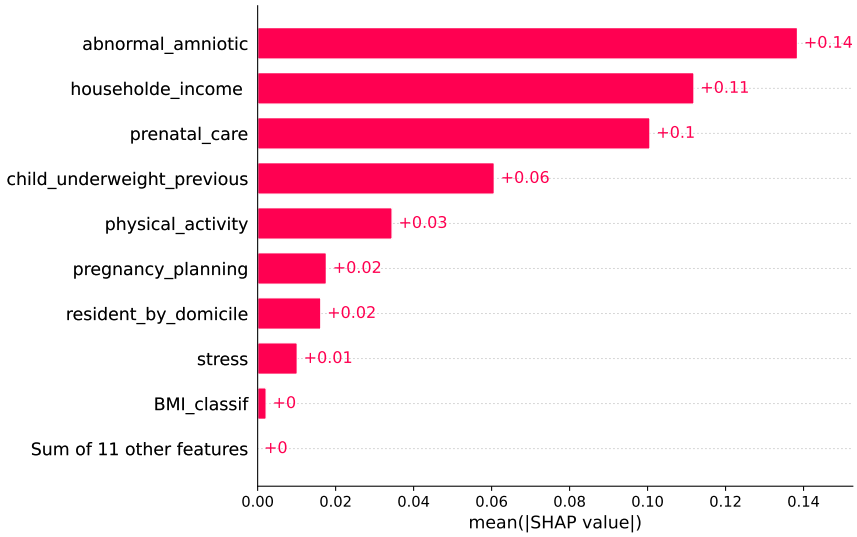

Figure 3: Global feature importance of a XGBoost model according SHAP analysis.

care is an important factor determining the status of prematurity of a child. We can also understand the Body Mass Index (BMI) does not play a part in prematurity. Stress can also affect prematurity but not very significantly.

Further we look into two singular predictions (a) A Normal Case (Figure 4 (b) Preterm Case 5

For the normal case in Figure 4 we can see the feature "abnormal amniotic" has a value of 2 which means that the mother does not have an abnormal amniotic sac. This moves the prediction (adds) towards a positive (normal) prediction. Similarly the "prenatal care" feature has a value of 2 meaning inadequate prenatal care was taken which pushes (subtracts) the output towards a negative (preterm) prediction. A high household income also provides towards a positive prediction.

For the premature case in Figure 5 we can see the feature "abnormal amniotic" has a value of 1 which refers to an abnormal amniotic fluid. This drastically moves the prediction to a negative (premature) prediction. Furthermore, a very inadequate prenatal care leads towards a premature prediction.

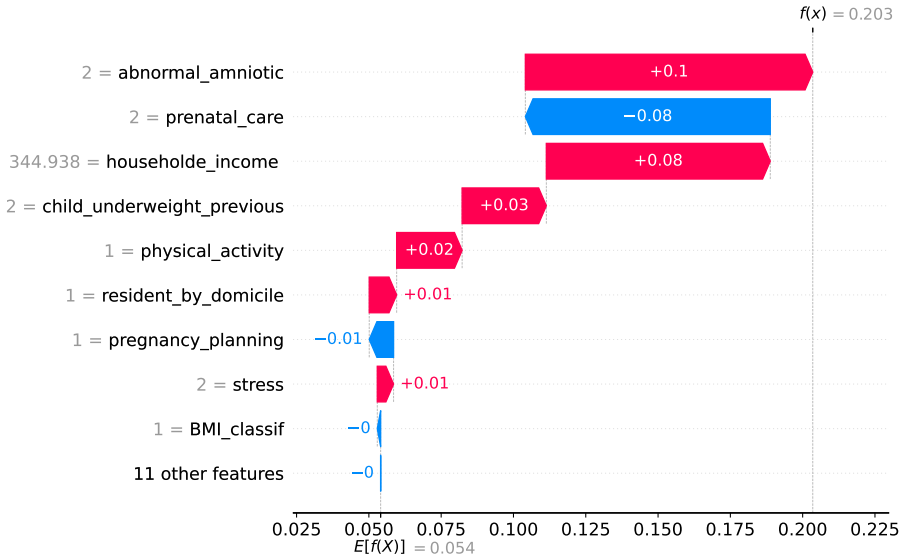

Figure 4: Contributing features towards a normal birth case. Features "abnormal amniotic sac", "household income", etc contribute towards a normal prediction. Feature "prenatal care" contributes against a normal prediction.

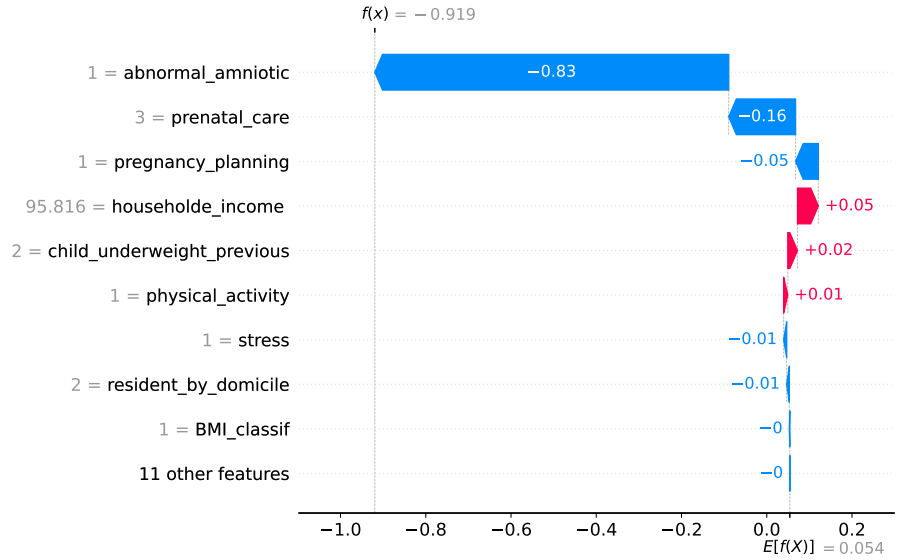

Figure 5: Contributing features towards a premature case. A bad amniotic sac and very inadequate prenatal care lead to a premature (negative) prediction.

7 Conclusion

Premature birth cases cause significant financial, economic, and social issues. Hence, improvement in detecting and mitigating preterm births can provide significant aid to the society. In this study, we worked with a a dataset of about 800 instances and the best test accuracy (71.22%) was achieved with the XGBoost model. We also interpreted predictions with SHAP explanations and conclude that health of the amniotic sac, prenatal care and household income are the major contributing factors towards a prediction. As future work, we look towards utilizing a larger dataset as well as neural network models.

References

Walani, S. R. Global burden of preterm birth. *International Journal of Gynecology & Obstetrics* **2020**, *150*, 31–33.

Goldenberg, R. L.; Culhane, J. F.; Iams, J. D.; Romero, R. Epidemiology and causes of preterm birth. *The lancet* **2008**, *371*, 75–84.

Raja, R.; Mukherjee, I.; Sarkar, B. K. A Machine Learning-Based Prediction Model for Preterm Birth in Rural India. *Journal of Healthcare Engineering* **2021**, *2021*.

Russell, R. B.; Green, N. S.; Steiner, C. A.; Meikle, S.; Howse, J. L.; Poschman, K.; Dias, T.; Potetz, L.; Davidoff, M. J.; Damus, K.; others Cost of hospitalization for preterm and low birth weight infants in the United States. *Pediatrics* **2007**, *120*, e1–e9.

Shah, P. S.; McDonald, S. D.; Barrett, J.; Synnes, A.; Robson, K.; Foster, J.; Pasquier, J.-C.; Joseph, K.; Piedboeuf, B.; Lacaze-Masmonteil, T.; others The Canadian Preterm Birth Network: a study protocol for improving outcomes for preterm infants and their families. *Canadian Medical Association Open Access Journal* **2018**, *6*, E44–E49.

Petrou, S.; Mehta, Z.; Hockley, C.; Cook-Mozaffari, P.; Henderson, J.; Goldacre, M. The impact of preterm birth on hospital inpatient admissions and costs during the first 5 years of life. *Pediatrics* **2003**, *112*, 1290–1297.

Greenough, A. Long term respiratory outcomes of very premature birth (< 32 weeks). Seminars in Fetal and Neonatal Medicine. 2012; pp 73–76.

Fergus, P.; Cheung, P.; Hussain, A.; Al-Jumeily, D.; Dobbins, C.; Iram, S. Prediction of preterm deliveries from EHG signals using machine learning. *PloS one* **2013**, *8*, e77154.

Tran, T.; Luo, W.; Phung, D.; Morris, J.; Rickard, K.; Venkatesh, S. Preterm birth prediction: Deriving stable and interpretable rules from high dimensional data. *arXiv preprint arXiv:1607.08310* **2016**,

Esplin, M. S.; O'Brien, E.; Fraser, A.; Kerber, R. A.; Clark, E.; Simonsen, S. E.; Holmgren, C.; Mineau, G. P.; Varner, M. W. Estimating recurrence of spontaneous preterm delivery. *Obstetrics & Gynecology* **2008**, *112*, 516–523.

on Good Clinical Practice in Maternal-Fetal Medicine, F. W. G.; Di Renzo, G. C.; Fonseca, E.; Gratacos, E.; Hassan, S.; Kurtser, M.; Malone, F.; Nambiar, S.; Nicolaides, K.; Sierra, N.; others Good clinical practice advice: Prediction of preterm labor and preterm premature rupture of membranes. *International Journal of Gynecology & Obstetrics* **2019**, *144*, 340–346.

Georgiou, H. M.; Di Quinzio, M. K.; Permezel, M.; Brennecke, S. P.; others Predicting preterm labour: current status and future prospects. *Disease markers* **2015**, *2015*.

Ravindra, N. G.; Espinosa, C.; Berson, E.; Phongpreecha, T.; Zhao, P.; Becker, M.; Chang, A. L.; Shome, S.; Marić, I.; De Francesco, D.; others Deep representation learning identifies associations between physical activity and sleep patterns during pregnancy and prematurity. *npj Digital Medicine* **2023**, *6*, 171.

Kokkinidis, I. K.; Logaras, E.; Rigas, E. S.; Tsakiridis, I.; Dagklis, T.; Billis, A.; Bamidis, P. D. Towards an Explainable AI-Based Tool to Predict Preterm Birth. *CARING IS SHARING–EXPLOITING THE VALUE IN DATA FOR HEALTH AND INNOVATION* **2023**, 571.

Koivu, A.; Sairanen, M. Predicting risk of stillbirth and preterm pregnancies with machine learning. *Health information science and systems* **2020**, *8*, 1–12.

Jehan, F.; Sazawal, S.; Baqui, A. H.; Nisar, M. I.; Dhingra, U.; Khanam, R.; Ilyas, M.; Dutta, A.; Mitra, D. K.; Mehmood, U.; others Multiomics characterization of preterm birth in low-and middle-income countries. *JAMA network open* **2020**, *3*, e2029655–e2029655.

Espinosa, C.; Becker, M.; Marić, I.; Wong, R. J.; Shaw, G. M.; Gaudilliere, B.; Aghaeepour, N.; Stevenson, D. K.; Stelzer, I. A.; Peterson, L. S.; others Data-driven modeling of pregnancy-related complications. *Trends in molecular medicine* **2021**, *27*, 762–776.

Moufarrej, M. N.; Vorperian, S. K.; Wong, R. J.; Campos, A. A.; Quaintance, C. C.; Sit, R. V.; Tan, M.; Detweiler, A. M.; Mekonen, H.; Neff, N. F.; others Early prediction of preeclampsia in pregnancy with cell-free RNA. *Nature* **2022**, *602*, 689–694.

Iams, J. D.; Goldenberg, R. L.; Meis, P. J.; Mercer, B. M.; Moawad, A.; Das, A.; Thom, E.; McNellis, D.; Copper, R. L.; Johnson, F.; others The length of the cervix and the risk of spontaneous premature delivery. *New England Journal of Medicine* **1996**, *334*, 567–573.

Hezelgrave, N. L.; Abbott, D. S.; Radford, S. K.; Seed, P. T.; Girling, J. C.; Filmer, J.; Tribe, R. M.; Shennan, A. H. Quantitative fetal fibronectin at 18 weeks of gestation to predict preterm birth in asymptomatic high-risk women. *Obstetrics & Gynecology* **2016**, *127*, 255–263.

Mercer, B.; Goldenberg, R.; Das, A.; Moawad, A.; Iams, J.; Meis, P.; Copper, R.; Johnson, F.; Thom, E.; McNellis, D.; others The preterm prediction study: a clinical risk assessment system. *American journal of obstetrics and gynecology* **1996**, *174*, 1885–1895.

Lee, K. A.; Chang, M. H.; Park, M.-H.; Park, H.; Ha, E. H.; Park, E. A.; Kim, Y. J. A model for prediction of spontaneous preterm birth in asymptomatic women. *Journal of Women's Health* **2011**, *20*, 1825–1831.

Conde-Agudelo, A.; Romero, R. Prediction of preterm birth in twin gestations using biophysical and biochemical tests. *American journal of obstetrics and gynecology* **2014**, *211*, 583–595.

Akazawa, M.; Hashimoto, K. Prediction of preterm birth using artificial intelligence: a systematic review. *Journal of Obstetrics and Gynaecology* **2022**, *42*, 1662–1668.

Sharifi-Heris, Z.; Laitala, J.; Airola, A.; Rahmani, A. M.; Bender, M.; others Machine learning approach for preterm birth prediction using health records: systematic review. *JMIR Medical Informatics* **2022**, *10*, e33875.

Allouche, M.; Huissoud, C.; Guyard-Boileau, B.; Rouzier, R.; Parant, O. Development and validation of nomograms for predicting preterm delivery. *American journal of obstetrics and gynecology* **2011**, *204*, 242–e1.

Dehaene, I.; Steen, J.; Vandewiele, G.; Roelens, K.; Decruyenaere, J. The web-based application "QUiPP v. 2" for the prediction of preterm birth in symptomatic women is not yet ready for worldwide clinical use: ten reflections on development, validation and use. *Archives of Gynecology and Obstetrics* **2022**, *306*, 571–575.

Kuhrt, K.; Smout, E.; Hezelgrave, N.; Seed, P.; Carter, J.; Shennan, A. Development and validation of a tool incorporating cervical length and quantitative fetal fibronectin to predict spontaneous preterm birth in asymptomatic high-risk women. *Ultrasound in Obstetrics & Gynecology* **2016**, *47*, 104–109.

AlSaad, R.; Malluhi, Q.; Boughorbel, S. PredictPTB: an interpretable preterm birth prediction model using attention-based recurrent neural networks. *BioData Mining* **2022**, *15*, 6.

Moreira, M. W.; Rodrigues, J. J.; Marcondes, G. A.; Neto, A. J. V.; Kumar, N.; Diez, I. D. L. T. A preterm birth risk prediction system for mobile health applications based on the support vector machine algorithm. 2018 IEEE International Conference on Communications (ICC). 2018; pp 1–5.

Santos, C. L.; de Mendonça Costa, K. M.; Dourado, J. E. C.; de Lima, S. B. G.; Dotto, L. M. G.; Schirmer, J. Maternal factors associated with prematurity in public maternity hospitals at the Brazilian Western Amazon. *Midwifery* **2020**, *85*, 102670.

Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing* **2011**, *21*, 137–146.

Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* **2014**, *40*, 16–28.

Fida, M. A. F. A.; Ahmad, T.; Ntahobari, M. Variance threshold as early screening to Boruta feature selection for intrusion detection system. 2021 13th International Conference on Information & Communication Technology and System (ICTS). 2021; pp 46–50.

Cohen, I.; Huang, Y.; Chen, J.; Benesty, J.; Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. *Noise reduction in speech processing* **2009**, 1–4.

Ross, B. C. Mutual information between discrete and continuous data sets. *PloS one* **2014**, *9*, e87357.

Song, Y.-Y.; Ying, L. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry* **2015**, *27*, 130.

Pranto, B.; Mehnaz, S. M.; Mahid, E. B.; Sadman, I. M.; Rahman, A.; Momen, S. Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information* **2020**, *11*, 374.

Peterson, L. E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883.

Géron, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*; " O'Reilly Media, Inc.", 2022.

Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016; pp 785–794.

Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *Journal of machine learning research* **2012**, *13*.

Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.

Hansen, J. V. Coalition Feature Interpretation and Attribution in Algorithmic Trading Models. *Computational Economics* **2021**, *58*, 849–866.