Machine Learning Prediction of Preterm Birth: An Analysis of Facility-Based Paper Health Records in Uganda

Shaheen Memon (shaheenmemon28@gmail.com)
University of Rwanda
Robert Wamala
Makerere University
Ignace Kabano
University of Rwanda

Research Article

Keywords: Preterm Birth, Machine Learning, Variable Importance, Paper Medical Records

Posted Date: July 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1877209/v1

Abstract

Purpose: Preterm Birth (PTB) is one of the leading causes of neonatal mortality in Uganda. Machine Learning (ML) can be used to identify women at risk of PTB in time for medical intervention and adequate preparation by mothers.

Methods: We utilized data from paper-based maternal health records at Kawempe National Referral Hospital, Uganda. A case-control method was employed, where for every woman who experienced a PTB, a woman without PTB and delivered in the same day was selected as a control. Treatment of missing data was done using Random Forest imputation. Variable Importance was analyzed using Random Forest. The following classification methods were applied in the prediction of PTB: Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes (NB). Performance of methods was investigated using prediction accuracy, sensitivity, and specificity.

Results: 1,540 women were included in the study, where 770 women had experienced PTB, and 770 women formed the controls. According to variable importance analysis, number of antenatal care visits had the biggest impact on PTB. SVM had the highest accuracy in predicting PTB at 64% (sensitivity 64% and sensitivity 63%).

Conclusions: Prediction of PTB using paper-based records in a developing country yielded similar results to studies done using electronic health records in developed countries. The predictive power could be low in this study due to fewer variables available from routinely collected ANC data. The inclusion of significant variables in the maternal records could potentially increase predictive power.

Introduction

Premature or Preterm Birth (PTB) is defined as birth occurring before 37 completed weeks (or less than 259 days) of pregnancy [1]. In 2014, an estimated 15 million babies were born prematurely worldwide, with the highest burden (80%) for preterm births being concentrated in sub-Saharan Africa and Asia [2]. In 2015, complications of PTB were responsible for approximately one-third of neonatal mortality globally [3]. PTB complications are the third leading cause of death in the first 28 days of life in sub-Saharan Africa, after severe neonatal infections and perinatal asphyxia [4].

PTB was the main reason for neonatal admission (37.7%) at Uganda's National Referral Hospital in 2014 [5]. In a study done in a tertiary hospital in Western Uganda, 31.6% of the admitted preterm neonates died [6]. At a rate of 13.6 preterm births per 1,000 live births, Uganda has the 28^{th} highest rate of PTB worldwide [7]. While these studies point to the high prevalence of PTB, its consequences and its risk factors, prediction of PTB remains largely unexplored, especially in developing countries.

Apart from increased risk of mortality, preterm babies face long-term neurodevelopmental impairment [8]–[10], and higher risk of having complex medical, psychological, educational, and socioeconomic needs [11], [12]. Preterm neonates who survive have increased vulnerability to diseases including retinopathy [13], pulmonary hypertension [14], and visual and hearing impairments [15]. PTB has negative effects on not only the neonate, but also causes anxiety and depression in postpartum women [16]. Most babies born prematurely are also of Low Birth Weight (LBW), which in turn is the biggest risk factor for more than 80% of infant deaths. LBW babies are at increased risk of poor birth weight recovery, post-neonatal mortality, growth failure, and adult-onset non-communicable diseases [17], [18].

The rate of neonatal mortality in Uganda has stagnated at a staggeringly high rate; 27 neonatal deaths per 1,000 live births, and PTB is directly responsible for 25% of these deaths [19], [20]. To attain the Sustainable Development Goals era target (SDG 3.2) of reducing the neonatal mortality rate (NMR) to 12 deaths/1000 live births or less by 2030 [21], [22], stakeholders, therefore, must scale up efforts at reducing PTB in Uganda.

Many recent studies have found Machine Learning (ML) processes an invaluable asset in the prediction of negative health outcomes, and are now becoming associated with more accurate prediction [23]–[28], and hence the same can be applied to the prediction of PTB. Whereas traditional statistical models depend on assumptions of linearity, ML methods do not make such assumptions. ML possesses the ability to process non-linear relationships and incorporate complex interactions without prior specifications, unlike statistical methods [29], [30]. Given these reasons, ML can be used to accurately identify women at risk of PTB in time for medical intervention, adequate preparation and care by mothers, families and healthcare organizations, for better obstetric outcomes [31], [32]. Early prediction of PTB can also allow healthcare facilities to be better prepared with neonatal care intensive units for preterm babies [33]. Many studies have been done on prediction of PTB using electronic health records in developed countries. However, hardly any studies have been done using paper-based health records in developing countries, particularly. This study attempts to close this gap in literature by utilizing paper-based maternal health records to predict PTB in Uganda.

Literature Review

Risk Factors of PTB

PTB has been associated with many risk factors including: inadequate antenatal care [34]–[37], antepartum hemorrhage [34], [38], [39], preeclampsia [34], [40], nulliparity [35], [37], [41] short interpregnancy interval [35], [42], [43], maternal age <20 years [35], [36], [44]–[46], advanced maternal age (≥ 35 years) [46], single status of mothers [35], [45], history of PTB [47]–[49], history of abortion [36], [45], [50], advanced maternal age [45], [46], [49] , pre-pregnancy hypertension [49], [51], history of fetal demise [45], underweight mothers [46], [52], [53], first antenatal visit after the first trimester [37], [54], lower education of mothers [41], smoking in pregnancy [55], prior cesarean delivery [42], [56], and pre-pregnancy diabetes [51].

Related Works

Many studies have been done on preterm birth in developing countries [34], [35], [57]–[61] . These studies have been done to determine risk factors of PTB, mainly using traditional statistical analysis. For instance, Bater et al. [35] did a study on the predictors of LBW and PTB in rural Uganda. They derived household, maternal, and infant characteristics data from a prospective birth cohort study from 2014 to 2016 in 12 districts. Stepwise Logistic regression was done using 3,841 (744 PTB) women to determine predictors of PTB. Ayebare, Ntuyo, Malande, and Nalwadda [34] did a study on maternal, reproductive and obstetric factors associated with preterm births in Kampala's National Referral Hospital. They also used Logistic Regression but on a smaller sample; 296 women (99 PTB). Other studies have similarly addressed determinants of PTB in developing countries without assessing for predictive power of the models [57]–[61].

In addition to utilizing statistical approaches, other studies have assessed machine learning methods by predictive power of PTB. However, many of these have been done in developed countries in the context of Electronic Health Records (EHR) [62]–[67]. For instance, Mercer et al. [62] developed a risk score-based system to predict PTB. They identified a number of risk factors, including fetal fibronectin, short cervix and history of preterm birth and used a sample of 2,929 women in the United States (US) to train a multivariate logistic regression. The model yielded a sensitivity of 24.2% (18.2%) and a specificity of 28.6% (33.3%) for nulliparous (multiparous) women. Using the same dataset, Vovsha et al. [63] compared Support Vector Machine (SVM) and Logistic and Lasso Regression with different model selection along with a model based on decision rules for the prediction of PTB. With linear SVM yielding 47% sensitivity and 57% specificity for predicting PTB at 28 weeks, they demonstrated an improvement over the sensitivity and specificity obtained by Mercer et al. Goodwin et al. [64] used data mining techniques and identified seven demographic variables that predict PTB. They used an ethnically diverse sample of 19,970 women in the US and obtained a 0.72 area under the receiver operator characteristic curves (AUCs). Weber et al. [65] utilized administrative data and extracted records for singleton pregnancies among nulliparous women in California from 2007 to 2011. The prediction of PTB was performed using K-nearest neighbors (KNN), lasso regression, and random forests (RF). They used demographic, maternal, and residency characteristics in a machine learning prediction model for PTB. The model yielded low AUC; 0.67. Koivu & Sairanen [66] used LR, ANN, gradient boosting decision tree, and ensemble models to construct individual classifiers to predict early stillbirth, late stillbirth and preterm birth pregnancies. They used pregnancy data provided by the Centers for Disease Control and Prevention (CDC), National Center of Health Statistics via their National Vital Statistics System in the US. They achieved a 0.64 AUC for PTB under the best performing model. Sun et al. [67] extracted data from EHR in a Beijing hospital. They used data based on physical examination, blood test, urine test strip, and gynecological examination. They compared six algorithms in the prediction pf PTB; Naive Bayesian (NBM), SVM, RF, artificial neural networks (ANN), K-means, and logistic regression. A total of 9550 pregnant women were included in the study, of which 4775 women had PTB. At 81.6%, the accuracy of the RF model was the highest compared to other algorithms.

It is important to note that studies on PTB using machine learning have also been carried out in developing and semi-developed countries [68]–[70]. For example, Prema and Pushpalatha [68] used data from local hospitals of Mysuru, India, and compared SVM with linear and nonlinear kernels, and logistic regression. The risk factors they considered included age, number of times pregnant, diabetes, obesity, and hypertension. In the balanced dataset, SVM with linear kernel yielded accuracy of 76% (sensitivity 84% and specificity 73%) and Logistic Regression yielded accuracy of 75% (sensitivity 70% and specificity 80%). Raja, Mukherjee, and Sarkar [69] used data from community health centers in Jharkhand, India. They used a feature selection approach based on the notion of entropy and compared prediction accuracy of three different classifiers, namely, decision tree (DT), logistic regression, and SVM for PTB prediction. SVM classifier yielded an accuracy of 90.9%. However, their predictive accuracy is the highest so far reviewed in the literature. Batoul et al. [70] compared SVM and Logistic Regression for predicting and classifying factors affecting PTB in women from Tehran, Iran. The dataset they used includes demographic and pregnancy characteristics and achieved 57% and 67% accuracy in logistic regression and SVM, respectively.

The evidence shows an abundance of literature addressing PTB. The studies done in developing countries have mostly been done to determine the risk factors of PTB without assessing predictive power of the models. Further, studies done on prediction of PTB have mostly been carried out in the developed countries using EHR. Unlike paper-based health records, EHR have higher rates of completeness and are

easier to access when needed [71], [72]. While some studies have addressed PTB prediction in developing countries, the etiology of PTB depends on the geographical and demographic features of the population studied [73]. Therefore, the results of studies in the developed countries may not be applied entirely to the situation in the developing countries. This study, therefore, seeks to address the gap in literature by using data extracted from paper-based maternal health records in Uganda to train machine learning methods to predict PTB.

Materials And Methods

Study Design, Site and Population

This study utilized a facility-based retrospective case-control approach based on administrative records of women from Kawempe National Referral Hospital. The Hospital deals mainly in Maternal Health Services (MHS) including Antenatal Care (ANC), intrapartum care and postnatal care for both mothers and newborns. On an average, about 2,000 women give birth at the hospital every month. The Hospital provides free ANC and delivery services to any pregnant woman who seeks the services; further, the hospital accepts referrals from all parts of the Country. For the purpose of this study, only records of women who had given birth at the Hospital in the period January 2017 to January 2021, and who had at least one ANC visit were considered. Figure 1 displays the workflow adopted in this study.

Inclusion Criteria

We included only live preterm and term singleton births. Women who delivered before completed 37 weeks of gestation were considered as cases. Women who delivered at term (≥ 37 weeks of gestation) formed the controls. Only records with mostly complete ANC cards were captured because the cards had most of the required data.

Exclusion Criteria

Extreme preterm births (< 28 weeks), post term births (≥ 42 weeks), still births, and multiple pregnancies were excluded. Records with missing or mostly incomplete ANC cards were not considered.

Sampling Procedure

Starting with the most recent records, each maternal file was assessed for PTB. These were the cases. We used a case-control ratio of 1:1. For every file with PTB, a file without PTB in the same day was selected as a control. The process was repeated until all files in the time interval considered were exhausted. A total of 1,540 records were captured. Out of these, 770 women delivered prematurely, while 770 women gave birth at full term.

Data Management and Quality Control

An online data capturing tool was developed using Open Data Kit (ODK) to capture data from maternal records which contain ANC Cards and Maternal Delivery Notes (MDN). Six Research Assistants (RA) with medical backgrounds and experience in clinical research, worked on data extraction. Prior to the actual data collection, the RAs were trained on the tool and selection of records. Pre-testing was done to ensure adequacy of the tool and thereafter data collection began. The Principal Investigator (PI) worked closely with the RAs to ensure reliable data collection procedures. Additionally, the PI would regularly check if extracted data matched the data in the maternal files. Variables with missingness greater than 90% were dropped at the pretest stage.

Variables Adopted in the Study

The maternal records include the following: (i) ANC section which captures baseline data on the women's socio-demographic characteristics, chronic illnesses, surgical history, gynecological and obstetric history; (ii) delivery notes which capture aspects like type of delivery, weeks of gestation at delivery, issues pertaining to morbidities developed and care given; and (iii) Infant notes which captures the baby's weight, Apgar score and anomalies.

The following variables were captured from the respective sections: (a) **ANC section**: district, age, marital status, religion, occupation, education level, pre-delivery weight, HIV sero-status, STD, hypertension, gestational diabetes, gravidity, parity, pre-partum anemia, previous caesarean section (c/s), previous stillbirth, previous PTB, previous abortion, previous Early Neonatal Death (ENND), birth spacing, number of ANC visits, gestation age at first ANC; (b) **Delivery notes**: multiple pregnancy, preeclampsia, antepartum hemorrhage, mode of delivery, gestational age at birth, maternal death; (c) **Infant notes**: birth weight, Apgar score, ENND. The outcome variable was incidence of PTB. This was deduced from the gestational age at birth as indicated in the delivery notes. Births at less than 37 completed weeks of gestation were termed as prebirths.

Data Analysis

Data was exported from ODK to R Studio for further cleaning and analysis subsequently. Missing data was imputed using Random Forest imputation. Imputation was done because analysis on primary care data using only complete information reduces predictive power and produces biased estimates leading to invalid conclusions [74]–[76]. Random Forest imputation was used based on its ability to impute missing data in multiple categorical variables simultaneously [76], [77]. Next, a descriptive summary of the maternal records was done using frequency distribution. The purpose of the analysis was to provide a description of the mothers utilized in the study. Simple Logistic Regression was used to produce Crude Odds Ratios for each independent variable. Next, Importance Analysis of the RF model was done to find the variables with the greatest effect on PTB. Further, the Fisher's Exact test was done to check if PTB had a significant effect on mode of delivery, maternal death, low birth weight, low Apgar score, and early neonatal death. Thereafter, the complete dataset was split into training (75%) and validation sets (25%). The following classification methods were applied to the training set: Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes (NB).

Machine Learning Algorithms

Logistic Regression (LR) – It is used to determine the relationship between a categorical dependent variable and independent variable(s). LR is used when the dependent variable is binary in nature i.e., it has only two categories, for instance 0 and 1, yes and no, or true and false. Suppose the dependent variable takes on the values 1 and 0, then LR will be in the following form:

$$P(Y = 1) = \frac{1}{1 + e^{-x\beta}}$$

Where X is the set of independent variables, and β are their corresponding regression coefficients. LR models the probability of one event (out of two possible events) by using a logistic function to map the results of linear regression between 0 and 1. The log-odds (the logarithm of the odds) for the event is a linear combination of one or more independent variables [78]. Simple LR was used to produce crude odds ratios for each independent variable. Only variables with $p < 0.1$ in simple LR were used in the final LR model.

Decision Tree (DT) – A Decision Tree (DT) is made up of a root node, internal nodes, and leaf nodes. It is built by use of recursive binary splitting. The classification error rate (CE) is used as a criterion for making the binary splits. This is the fraction of the training observations in that region that do not belong to the most common class. The CE is measured for each independent variable. The variable that yields the smallest CE is selected to be the root node. CE is measured as follows:

$$CE = 1 - \max\left(\hat{p}_{mk}\right)$$

Where $\hat{p}_{mk}$is the proportion of training observations in the m^{th} region from the k^{th} class. Subsequent nodes are again determined using CE. However, if a node itself has a lower CE than a subsequent node, it is not split further and ends up as a leaf node. To prevent overfitting, pruning technique is used when the tree is fully grown. DTs have several advantages, including their ability to handle both categorical and continuous data, simplicity and comprehensibility. However, they are not very flexible and fail on test data; they suffer from inaccuracy [79]–[82]. DT classification was done using *rpart* package in RStudio.

Random Forest (RF) - Random Forests (RF) are built from decision trees but overcome their inaccuracy aspect. Based on the bagging idea of ensemble learning, they integrate multiple DTs. From the original dataset, a new bootstrapped dataset is created and used to build a DT using only a random subset of the independent variables at each step. The observations left out of the bootstrapped dataset form the Out-Of-Bag (OOB) dataset and are used to measure the accuracy of the trees to be built. This is repeated hundreds of times to yield a large variety of trees. It is this step that allows a random forest to overcome the drawback of the decision tree. To make a prediction for x_o, the data for x_o is run on each tree. The aggregate of the predicted class under each tree is computed and the class with the most votes is selected to be the predicted y for x_o. This process is known as bagging.

To measure accuracy, the OOB dataset is run through all the trees where it was not used to create. The class with the most values is selected. This is then compared to the actual class. The accuracy is measured by the proportion of OOB observations that were correctly classified by the random forest. The proportion of OOB observations that were misclassified form the OOB error. This process can also be used to determine the number of independent variables to consider at each step when building the trees [83]. RF classification was done using the *randomForest* package in RStudio.

Random forests can be used to rank the importance of variables in a regression or classification problem. This is done in the following steps:

1. The number of votes for the correct class in the OOB data is computed for each tree in the RF

2. The order of values in a predictor (say predictor m) are shuffled in the OOB data and the number of votes for correct class are computed.
3. The number of votes for the correct class in the shuffled m is subtracted from the number of votes for the correct class in the original OOB data.
4. The average of this number aggregated over all the trees in the RF forms the raw importance score for the m.
5. The variables with highest scores are ranked as the most important.
The *varImp* package was used to evaluate variable importance in RStudio.

K-Nearest Neighbors (KNN) - This is one of the simpler to understand classifier and closest to the gold-standard Bayes Classifier. It works in the following steps:

For a positive integer K and test observation x_0:

i) It 'looks' for K points in the training data that are nearest to x_0. These will be denoted by N_o
ii) Next, it estimates the conditional probability for class *j*:

$$pr\left(Y = j | X = x_0\right) = \frac{1}{K}\sum_{i \epsilon N_0} I(y_i = j)$$

iii) Then, it applies the Bayes Rule; it classifies {x}_{0} to the class with the largest probability.
The advantages of the KNN classifier are that it doesn't need to know the true distribution of Y given X. It still yields results akin to the optimal Bayes classifier [84], [85]. KNN classification was done using the *class* package in RStudio.

Support Vector Machine (SVM) - Support Vector Machines (SVM) are an extension of support vector classifiers, which are in turn an extension of maximal marginal classifiers. A maximal marginal classifier uses a threshold that gives the largest distance between the most extreme points in a class and the threshold itself. However, this classifier is super sensitive to outliers; the outliers can "pull" the threshold in a direction and this leads to poor performance in prediction.

The solution to this problem is a soft margin which allows some level of misclassification and hence reduce the sensitivity to outliers. Cross validation is used to determine how many observations and misclassifications to allow in the soft margin. This is now known as a support vector classifier. It also has a drawback; it fails when there is a lot of overlap in the observations. This is because wherever it places the threshold, there will still be a lot of misclassifications.

The solution to overlapping data is to move it to a higher dimension using a kernel and then putting a threshold through it that separates the data into two groups. This is now known a support vector machine [86], [87]. SVM classification was done using the *e1071* package in RStudio.

Naïve Bayes (NB) – Naïve Bayes (NB) is a probabilistic classifier and uses the Bayes Theorem for classification tasks [88], [89]. The Bayes Theorem is:

P\left(B|A\right)= \frac{P\left(A|B\right)P\left(B\right)}{P\left(A\right)}
Where P(B|A) is a posterior probability of class B, P(A) is the prior probability of predictor A and P(B) is the prior probability of class B. For *m* classes, $c_1, c_2, \ldots, c_m$, this classifier will assign test observation x_0 to the class with the largest posterior probability, i.e., x_o will belong to class c_i if and only if:

P\left({c}_{i}|{x}_{0}\right)>P\left({c}_{j}|{x}_{0}\right), j\ne i
NB classification was done using the *naivebayes* package in RStudio.

Performance Evaluation of Methods to Predict PTB

The validation set was used to determine final performance by assessing for accuracy, sensitivity and specificity for proper classification of PTB.

where:

Accuracy= \frac{tp+tn}{tp+tn+fn+fp}
where *tp* is the true positive, *tn* is the true negative, *fp* is the false positive, and *fn* is the false negative, and

Sensitivity= \frac{tp}{tp+fn}
Specificity= \frac{tn}{fp+tn}

In R Studio, “Yes” for PTB was selected as the positive response to ensure that sensitivity reflects accurate prediction of PTB

Ethical Considerations

Ethical approval for conducting the study was obtained from: (i) Uganda National Council of Science and Technology (UNCST) with Registration Number Ref: HS977ES, and (ii) Mulago Hospital Research and Ethics Committee (REC/IRB). Administrative clearance to accessing the administrative records was obtained from Kawempe National Referral Hospital. To ensure confidentiality of the women, the study was conducted according to the Helsinki Declaration (1975;2008) guidelines for medical research, where names of mothers were not captured. Only their file numbers were recorded. At data analysis, no individual records are reported – we utilize only aggregate findings. Passwords will be used for data archiving so that data will be accessed for the sole purpose of this study. Data was extracted from the records after discharge or death and hence study inclusion had no effect on the treatment. Because of the approach used in data collection, individual informed consent was not required.

Results

This section presents a description of the mothers in the study, the importance of the independent variables used, the performance of methods used to predict PTB, and the effects of PTB on mother and newborn.

Characteristics of Mothers

A total of 1,540 mothers (PTB: 770, control: 770) were included in our study. The number of women aged < 20 years and those aged ≥ 35 years were both higher in the PTB group than the control group, whereas the number of women aged between 20 and 34 years were less in the PTB group than in the control group (Table 1).

There were more cases of grand multigravida (women who have had ≥5 births (live or stillborn) at ≥20 weeks of gestation) in the PTB group as compared to the control group. Histories of still birth, early neonatal death and PTB were more prevalent in the PTB group than the control group (Table 2).

More women in the control group had attended ANC at least four times during their pregnancies as compared to women with PTB. More women who experienced PTB attended ANC only one or two times as compared to women in the control group. There were more cases of women with pre-delivery weight ≤ 55 kg in the PTB than control groups. There were more cases of chronic hypertension, preeclampsia and antepartum hemorrhage in women with PTB than those without. The median gestation age at birth of women who experienced PTB was 35 weeks, whereas the median gestation age for women who delivered at term was 38 weeks (Table 3).

The variables with significant crude odds ratios have been bolded in Table 1, Table 2 and Table 3. These include age group, gravidity, history of still birth, history of early neonatal death, history of PTB, number of ANC visits, pre delivery weight, chronic hypertension, preeclampsia and antepartum hemorrhage. These were the variables used in the final LR classification model. It is important to note that parity was not used due to it being highly collinear with gravidity.

Variable Importance

The importance analysis of the RF model found that the top 10 most important variables (mean decrease accuracy (MDA) > 10 in RF model) include number of ANC visits, preeclampsia, age group, weight, gravidity, birth spacing, previous ENND, previous PTB, occupation, and district . The variable with the biggest effect on PTB is number of ANC visits (Figure 2).

Prediction of PTB

Table 4 shows that for classification of PTB, the SVM model yielded the highest accuracy at 0.64, whereas the KNN model yielded the lowest accuracy at 0.58. Even though DT gave the highest sensitivity (0.66), its specificity was the lowest at 0.55. KNN had the worst sensitivity at 0.42, followed by RF at 0.59. SVM yielded a sensitivity and specificity that were both above 0.63.

Table 4: Performance Evaluation of Methods to Predict Preterm Birth

Method	Accuracy	Sensitivity	Specificity
LR	0.6052	0.6207	0.5924
DT	0.6026	0.6609	0.5545
RF	0.6201	0.5882	0.6453
SVM	0.6364	0.6437	0.6303
KNN	0.5844	0.4195	0.7204
NB	0.6104	0.6092	0.6114

Effect of PTB on Mother and Newborn

The results from Table 5 show that PTB had a significant effect on mode of delivery and LBW. Women who experienced PTB were less likely to deliver by cesarean-section than women who delivered at term ($p < 0.05$). On the other hand, babies born prematurely were more likely to have low birth weight (< 2.5 kgs) than babies born at term ($p < 0.05$).

Table 5: Outcomes of Preterm Delivery on Mother and Newborn

Variable	Preterm n (%)	Control n (%)	Fisher's exact p-value
Mode of Delivery			
Normal	485 (63.0)	416 (54.0)	**0.00**
Cesarean-Section	285 (37.0)	354 (46.0)	
Maternal Death			
Yes	3 (0.4)	1 (0.1)	0.63
No	767 (99.6)	769 (99.9)	
Low Birth Weight			
Yes	495 (64.3)	127 (16.5)	**0.00**
No	275 (35.7)	643 (83.5)	
Low Apgar Score			
Yes	185 (24.0)	216 (28.1)	0.08
No	585 (76.0)	554 (71.9)	
ENND			
Yes	19 (2.5)	9 (1.2)	0.08
No	751 (97.5)	761 (98.8)	

Discussion

We utilized paper-based maternal health records to develop models to predict PTB in Uganda. Variable importance analysis in the Random Forest model revealed that the number of ANC visits had the biggest impact on PTB. Other important variables included preeclampsia, age, weight, gravidity and birth spacing. Six machine learning methods were compared by predictive power, namely: Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes (NB). We found that SVM had the highest predictive power, with an accuracy of 0.64, while KNN had the lowest at 0.58. Additionally, we found that PTB had a significant effect on low birth weight.

In the importance analysis using RF, we found that the top three variables in predicting PTB were inline with other studies elsewhere [34]–[36], [40], [46], [90]–[94]. For example, these studies have shown the importance of adequate antenatal care in averting the chances of PTB [34], [35], [90], [92], [93]. Adequate antenatal care helps identify risk factors in pregnant women, for instance history of PTB and hypertensive disorders, and hence early detection and management to prevent a PTB [90]. Likewise, studies have found that preeclamsia is a

major contributor to PTB [34], [40], [91] and hence a condition which must be targeted to reduce its prevalence. Additionally, both advanced maternal age (≥35 years) and young maternal age (<20 years) have been found to be risk factors for PTB [35], [36], [46], [94]. This evidence shows that antenatal care, preeclampsia and age are not unique to predicting PTB among women in the developed countries, but also apply to their counterparts in the developiong countries.

The highest predictive accuracy of 64% obtained in our study is higher than the accuracy in some studies carried out in the United States [62], [63]. On the other hand, our predictive accuracy is comparable to studies done in developed countries using EHR data [65], [66]. It is also important to note that the predictive accuracy was lower than other studies elsewhere [64], [67]–[69]. It is worth noting that Sun et al. [67] achieved an accuracy of 81.6% in PTB prediction. Their study utilized data based on physical examination, blood test, urine test strip, and gynecological examination. Other studies have found that variables like smoking [55], positive fetal fibronectin test [95] , cervical length [96], Body Mass Index (BMI) [69], Magnesium and Serum inorganic phosphorus in mother's blood, and mother's mean platelet volume [67] have significant impact on PTB. These variables are not routinely captured in public health facilities in Uganda. It is highly likely that the variables are not captured in other developing countries elsewhere. In other words, the women in our data were only tested for HIV, and hence we do not have other variables that are acquired from blood tests. Additionally, height was underreported (>90% missing) and hence BMI could also not be computed. This is one of the shortcomings of paper-based records; they have lower rates of completeness and are difficult to access when needed, as compared to electronic medical records [71], [72]. It is highly likely that the absence of these variables led to the underperformance of our model in predicting PTB. Nevertheless, these findings reveal that the performance of the model in predicting PTB is largely an aspect of variables utilized, rather than differences among women in the developed and developing countries. Further, our results also show a significant effect of PTB on LBW; 64.3% of mothers who experienced PTB gave birth to underweight babies, as compared to only 16.5% of mothers who delivered at term. This only solidifies the importance of predicting PTB and hence prevention and management in order to reduce the incidence of LBW.

FUTURE RESEARCH FOCUS

This study is restricted to one public health facility and may limit the use of the proposed model. Future studies need to utilize data from private health facilities where more variables could potentially be available. This could potentially increase the accuracy in predicting PTB.

Conclusion And Recommendations

The most important variables for predicting PTB include number of ANC visits, age, weight, gravidity and birth spacing. The best model for predicting PTB was SVM with a predictive accuracy of 0.64. Additionally, we found that PTB had a significant effect on low birth weight. The predictive power is largely attributed to limitations in data that is routinely captured in maternal health records in a public health facility in Uganda. Nevertheless, the evidence shows that the performance of the model in predicting PTB is largely an aspect of variables utilized, rather than differences among women in the developed and developing countries. Therefore, the scope of routinely collected data in public hospitals needs to be increased to improve the ability to accurately predict PTB.

Statements And Declarations

Funding and Competing Interests

Partial funding to support data collection was obtained from the African Centre of Excellence in Data Science, University of Rwanda.

The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval

Ethical approval for conducting the study was obtained from Uganda National Council of Science and Technology (Ref: HS977ES), and Mulago Hospital Research and Ethics Committee.

Data Availability

Data can be availed at reasonable request

References

1. World Health Organization, "Born too soon: the global action report on preterm birth," p. 112, 2012, Accessed: Apr. 27, 2022. [Online]. Available: https://apps.who.int/iris/handle/10665/44864

2. S. Chawanpaiboon *et al.*, “Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis,” The Lancet Global Health, vol. 7, no. 1, pp. e37–e46, Jan. 2019, doi: 10.1016/S2214-109X(18)30451-0.
3. L. Liu *et al.*, “Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals,” The Lancet, vol. 388, no. 10063, pp. 3027–3035, Dec. 2016, doi: 10.1016/S0140-6736(16)31593-8.
4. I. Ahmed *et al.*, “Population-based rates, timing, and causes of maternal deaths, stillbirths, and neonatal deaths in south Asia and sub-Saharan Africa: a multi-country prospective cohort study,” The Lancet Global Health, vol. 6, no. 12, pp. e1297–e1308, Dec. 2018, doi: 10.1016/S2214-109X(18)30385-1.
5. Y. Abdallah, F. Namiiro, J. Mugalu, J. Nankunda, Y. Vaucher, and D. McMillan, “Is facility based neonatal care in low resource setting keeping pace? A glance at Uganda’s National Referral Hospital,” *African Health Sciences*, vol. 16, no. 2, Art. no. 2, Jul. 2016, doi: 10.4314/ahs.v16i2.2.
6. W. I. Egesa *et al.*, “Preterm Neonatal Mortality and Its Determinants at a Tertiary Hospital in Western Uganda: A Prospective Cohort Study,” *Pediatric Health Med Ther*, vol. 11, pp. 409–420, Oct. 2020, doi: 10.2147/PHMT.S266675.
7. H. Blencowe *et al.*, “National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications,” *Lancet*, vol. 379, no. 9832, Art. no. 9832, Jun. 2012, doi: 10.1016/S0140-6736(12)60820-4.
8. H. Blencowe *et al.*, “Preterm birth–associated neurodevelopmental impairment estimates at regional and global levels for 2010,” *Pediatr Res*, vol. 74, no. 1, Art. no. 1, Dec. 2013, doi: 10.1038/pr.2013.204.
9. G. Namazzi *et al.*, “Neurodevelopmental outcomes of preterm babies during infancy in Eastern Uganda: a prospective cohort study,” Global Health Action, vol. 13, no. 1, p. 1820714, Dec. 2020, doi: 10.1080/16549716.2020.1820714.
10. C. H. T. Do *et al.*, “Neurodevelopment at 2 years corrected age among Vietnamese preterm infants,” Archives of Disease in Childhood, vol. 105, no. 2, pp. 134–140, Feb. 2020, doi: 10.1136/archdischild-2019-316967.
11. S. Petrou, O. Eddama, and L. Mangham, “A structured review of the recent literature on the economic consequences of preterm birth,” Archives of Disease in Childhood - Fetal and Neonatal Edition, vol. 96, no. 3, pp. F225–F232, May 2011, doi: 10.1136/adc.2009.161117.
12. J. G. Berry *et al.*, “Trends in Resource Utilization by Children with Neurological Impairment in the United States Inpatient Health Care System: A Repeat Cross-Sectional Study,” PLOS Medicine, vol. 9, no. 1, p. e1001158, Jan. 2012, doi: 10.1371/journal.pmed.1001158.
13. A. M. Lynch *et al.*, “The relationship of the subtypes of preterm birth with retinopathy of prematurity,” *American Journal of Obstetrics and Gynecology*, vol. 217, no. 3, p. 354.e1-354.e8, Sep. 2017, doi: 10.1016/j.ajog.2017.05.029.
14. E. Naumburg and L. Söderström, “Increased risk of pulmonary hypertension following premature birth,” BMC Pediatrics, vol. 19, no. 1, p. 288, Aug. 2019, doi: 10.1186/s12887-019-1665-6.
15. M. Hirvonen *et al.*, “Visual and Hearing Impairments After Preterm Birth,” Pediatrics, vol. 142, no. 2, Aug. 2018, doi: 10.1542/peds.2017-3888.
16. L. T. Singer, A. Salvator, S. Guo, M. Collin, L. Lilien, and J. Baley, “Maternal Psychological Distress and Parenting Stress After the Birth of a Very Low-Birth-Weight Infant,” *JAMA*, vol. 281, no. 9, pp. 799–805, Mar. 1999, doi: 10.1001/jama.281.9.799.
17. J. E. Lawn *et al.*, “Every Newborn: progress, priorities, and potential beyond survival,” The Lancet, vol. 384, no. 9938, pp. 189–205, Jul. 2014, doi: 10.1016/S0140-6736(14)60496-7.
18. F. B. Namiiro, J. Mugalu, R. M. McAdams, and G. Ndeezi, “Poor birth weight recovery among low birth weight/preterm infants following hospital discharge in Kampala, Uganda,” BMC Pregnancy and Childbirth, vol. 12, no. 1, p. 1, Jan. 2012, doi: 10.1186/1471-2393-12-1.
19. UBOS and ICF International, “Uganda Demographic and Health Survey 2016,” The DHS Program ICF Rockville, Maryland, USA, Kampala, Uganda, Jan. 2018. [Online]. Available: https://dhsprogram.com/pubs/pdf/FR333/FR333.pdf
20. J. Jitta and D. Kyaddondo, “Situation analysis of newborn health in Uganda,” Ministry of Health, The Republic of Uganda, Kampala, Uganda, 2008.
21. World Health Organization, “World health statistics 2018: monitoring health for the SDGs, sustainable development goals,” World Health Organization, 2018. [Online]. Available: http://www.who.int/iris/handle/10665/272596
22. United Nations Children’s Fund, “Every child alive, the urgent need to end newborn deaths,” United Nations, Switzerland, 2018. [Online]. Available: https://data.unicef.org/resources//every_child_alive_the_urgent_need_to_end_newborn_deaths
23. K. Buchan, M. Filannino, and Ö. Uzuner, “Automatic prediction of coronary artery disease from clinical narratives,” Journal of biomedical informatics, vol. 72, pp. 23–32, 2017.
24. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities,” Ieee Access, vol. 5, pp. 8869–8879, 2017.

25. A. Mouelhi, H. Rmili, J. B. Ali, M. Sayadi, R. Doghri, and K. Mrad, “Fast unsupervised nuclear segmentation and classification scheme for automatic allred cancer scoring in immunohistochemical breast tissue images,” *Computer methods and programs in biomedicine*, vol. 165, pp. 37–51, 2018.
26. H. Yang and Y.-P. P. Chen, “Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information,” Expert Systems with Applications, vol. 42, no. 15–16, pp. 6168–6176, 2015.
27. X. Yuan, L. Xie, and M. Abouelenien, “A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data,” Pattern Recognition, vol. 77, pp. 160–172, 2018.
28. O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, “State-of-the-art in artificial neural network applications: A survey,” Heliyon, vol. 4, no. 11, p. e00938, Nov. 2018, doi: 10.1016/j.heliyon.2018.e00938.
29. K. K. Venkatesh *et al.*, “Machine Learning and Statistical Models to Predict Postpartum Hemorrhage,” *Obstet Gynecol*, vol. 135, no. 4, pp. 935–944, Apr. 2020, doi: 10.1097/AOG.0000000000003759.
30. T. Goto, C. A. Camargo, M. K. Faridi, R. J. Freishtat, and K. Hasegawa, “Machine Learning–Based Prediction of Clinical Outcomes for Children During Emergency Department Triage,” JAMA Netw Open, vol. 2, no. 1, p. e186937, Jan. 2019, doi: 10.1001/jamanetworkopen.2018.6937.
31. G. J. Escobar, N. R. Gupta, E. M. Walsh, L. Soltesz, S. M. Terry, and P. Kipnis, “Automated early detection of obstetric complications: theoretic and methodologic considerations,” *American Journal of Obstetrics and Gynecology*, vol. 220, no. 4, pp. 297–307, Apr. 2019, doi: 10.1016/j.ajog.2019.01.208.
32. K. Y. Ngiam and I. W. Khor, “Big data and machine learning algorithms for health-care delivery,” The Lancet Oncology, vol. 20, no. 5, pp. e262–e273, May 2019, doi: 10.1016/S1470-2045(19)30149-4.
33. C. Gao, S. Osmundson, D. R. Velez Edwards, G. P. Jackson, B. A. Malin, and Y. Chen, “Deep learning predicts extreme preterm birth from electronic health records,” Journal of Biomedical Informatics, vol. 100, p. 103334, Dec. 2019, doi: 10.1016/j.jbi.2019.103334.
34. E. Ayebare, P. Ntuyo, O. O. Malande, and G. Nalwadda, “Maternal, reproductive and obstetric factors associated with preterm births in Mulago Hospital, Kampala, Uganda: a case control study,” Pan Afr Med J, vol. 30, p. 272, Aug. 2018, doi: 10.11604/pamj.2018.30.272.13531.
35. J. Bater *et al.*, “Predictors of low birth weight and preterm birth in rural Uganda: Findings from a birth cohort study,” PLOS ONE, vol. 15, no. 7, p. e0235626, Jul. 2020, doi: 10.1371/journal.pone.0235626.
36. M. Jiang, M. M. Mishu, D. Lu, and X. Yin, “A case control study of risk factors and neonatal outcomes of preterm birth,” *Taiwanese Journal of Obstetrics and Gynecology*, vol. 57, no. 6, pp. 814–818, Dec. 2018, doi: 10.1016/j.tjog.2018.10.008.
37. A. Gurung *et al.*, “Incidence, risk factors and consequences of preterm birth – findings from a multi-centric observational study for 14 months in Nepal,” Arch Public Health, vol. 78, no. 1, p. 64, Jul. 2020, doi: 10.1186/s13690-020-00446-7.
38. S. A. Feresu, S. D. Harlow, and G. B. Woelk, “Risk factors for prematurity at Harare Maternity Hospital, Zimbabwe,” *International Journal of Epidemiology*, vol. 33, no. 6, pp. 1194–1201, Dec. 2004, doi: 10.1093/ije/dyh120.
39. J. A. Lykke, K. L. Dideriksen, Ø. Lidegaard, and J. Langhoff-Roos, “First-trimester vaginal bleeding and complications later in pregnancy,” Obstet Gynecol, vol. 115, no. 5, pp. 935–944, May 2010, doi: 10.1097/AOG.0b013e3181da8d38.
40. R. Alijahan, S. Hazrati, M. Mirzarahimi, F. Pourfarzi, and P. Ahmadi Hadi, “Prevalence and risk factors associated with preterm birth in Ardabil, Iran,” *Iran J Reprod Med*, vol. 12, no. 1, pp. 47–56, Jan. 2014, Accessed: Apr. 29, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4009588/
41. P. Astolfi and L. A. Zonta, “Risks of preterm delivery and association with maternal age, birth order, and fetal gender,” *Human Reproduction*, vol. 14, no. 11, pp. 2891–2894, Nov. 1999, doi: 10.1093/humrep/14.11.2891.
42. L. F. Wong, J. Wilkes, K. Korgenski, M. W. Varner, and T. A. Manuck, “Risk factors associated with preterm birth after a prior term delivery,” BJOG, vol. 123, no. 11, pp. 1772–1778, Oct. 2016, doi: 10.1111/1471-0528.13683.
43. T. Rodrigues and H. Barros, “Short interpregnancy interval and risk of spontaneous preterm delivery,” Eur J Obstet Gynecol Reprod Biol, vol. 136, no. 2, pp. 184–188, Feb. 2008, doi: 10.1016/j.ejogrb.2007.03.014.
44. A. M. Fraser, J. E. Brockert, and R. H. Ward, “Association of young maternal age with adverse reproductive outcomes,” *N Engl J Med*, vol. 332, no. 17, pp. 1113–1117, Apr. 1995, doi: 10.1056/NEJM199504273321701.
45. C.-C. Lo, J.-J. Hsu, C.-C. Hsieh, T.-T. Hsieh, and T.-H. Hung, “Risk Factors for Spontaneous Preterm Delivery Before 34 Weeks of Gestation Among Taiwanese Women,” *Taiwanese Journal of Obstetrics and Gynecology*, vol. 46, no. 4, pp. 389–394, Dec. 2007, doi: 10.1016/S1028-4559(08)60008-X.
46. M. Ip, E. Peyman, V. Lohsoonthorn, and M. A. Williams, “A case–control study of preterm delivery risk factors according to clinical subtypes and severity,” Journal of Obstetrics and Gynaecology Research, vol. 36, no. 1, pp. 34–44, 2010, doi: 10.1111/j.1447-

0756.2009.01087.x.

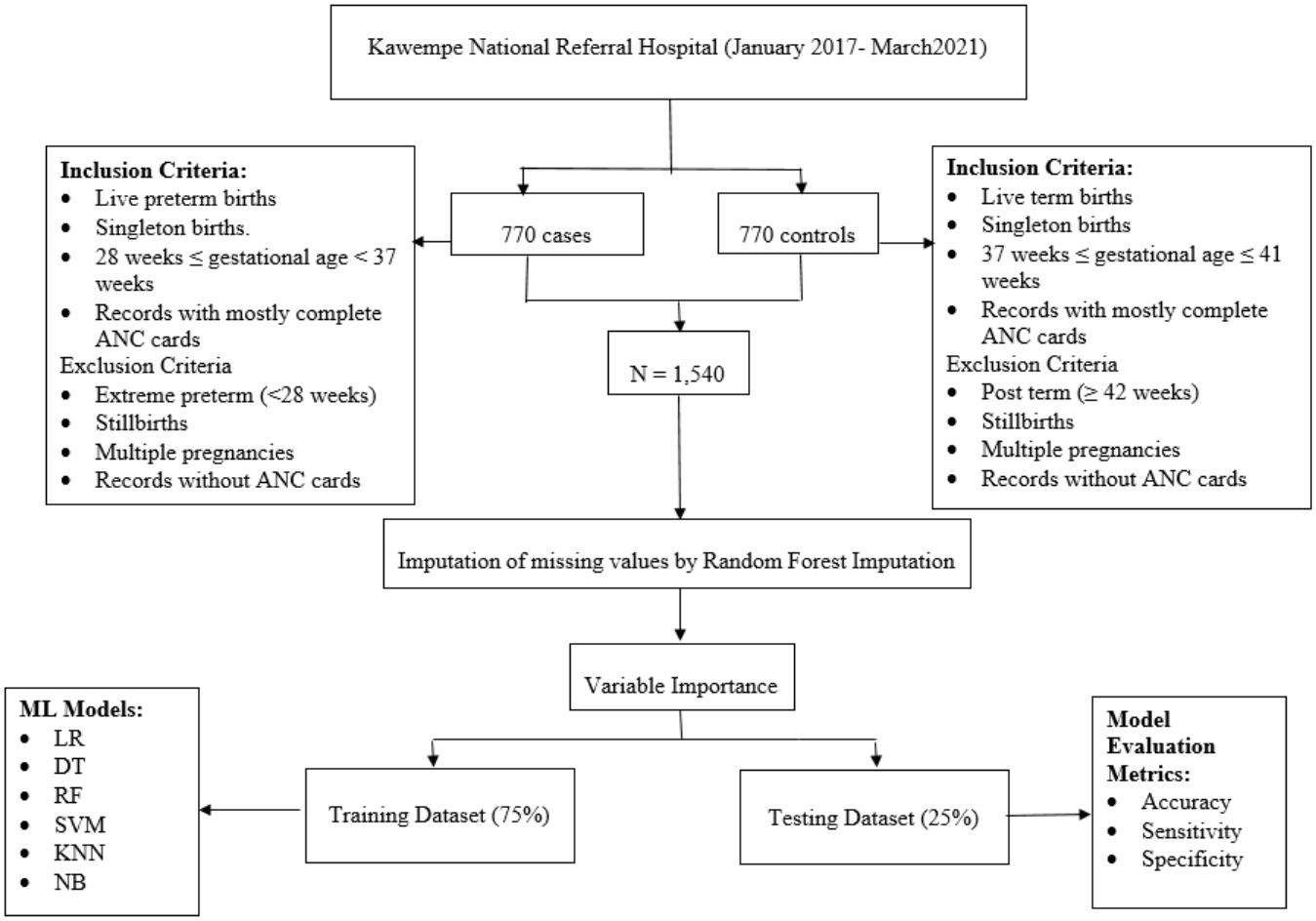
47. M. S. Esplin *et al.*, "Estimating recurrence of spontaneous preterm delivery," Obstet Gynecol, vol. 112, no. 3, pp. 516–523, Sep. 2008, doi: 10.1097/AOG.0b013e318184181a.
48. S. L. Bloom, N. P. Yost, D. D. McIntire, and K. J. Leveno, "Recurrence of preterm birth in singleton and twin pregnancies," Obstet Gynecol, vol. 98, no. 3, pp. 379–385, Sep. 2001, doi: 10.1016/s0029-7844(01)01466-1.
49. W. Yuan, A. M. Duffner, L. Chen, L. P. Hunt, S. M. Sellers, and A. L. Bernal, "Analysis of preterm deliveries below 35 weeks' gestation in a tertiary referral hospital in the UK. A case-control survey," BMC Research Notes, vol. 3, no. 1, p. 119, Apr. 2010, doi: 10.1186/1756-0500-3-119.
50. G. Saccone, L. Perriera, and V. Berghella, "Prior uterine evacuation of pregnancy as independent risk factor for preterm birth: a systematic review and metaanalysis," Am J Obstet Gynecol, vol. 214, no. 5, pp. 572–591, May 2016, doi: 10.1016/j.ajog.2015.12.044.
51. H. Berger *et al.*, "Impact of diabetes, obesity and hypertension on preterm birth: Population-based study," *PLOS ONE*, vol. 15, no. 3, p. e0228743, Mar. 2020, doi: 10.1371/journal.pone.0228743.
52. Z. Han, S. Mulla, J. Beyene, G. Liao, S. D. McDonald, and Knowledge Synthesis Group, "Maternal underweight and the risk of preterm birth and low birth weight: a systematic review and meta-analyses," *Int J Epidemiol*, vol. 40, no. 1, pp. 65–101, Feb. 2011, doi: 10.1093/ije/dyq195.
53. A. I. Girsen *et al.*, "Women's prepregnancy underweight as a risk factor for preterm birth: a retrospective study," *BJOG*, vol. 123, no. 12, pp. 2001–2007, Nov. 2016, doi: 10.1111/1471-0528.14027.
54. W. Nicholson, M. Croughan-Minihane, S. Posner, A. E. Washington, and S. K. Kilpatrick, "Preterm delivery in patients admitted with preterm labor: a prediction study," *J Matern Fetal Med*, vol. 10, no. 2, pp. 102–106, Apr. 2001, doi: 10.1080/714052726.
55. A. Burguet *et al.*, "The complex relationship between smoking in pregnancy and very preterm delivery," BJOG: An International Journal of Obstetrics & Gynaecology, vol. 111, no. 3, pp. 258–265, 2004, doi: 10.1046/j.1471-0528.2003.00037.x.
56. G. C. Di Renzo, I. Giardina, A. Rosati, G. Clerici, M. Torricelli, and F. Petraglia, "Maternal risk factors for preterm birth: a country-based population analysis," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 159, no. 2, pp. 342–346, Dec. 2011, doi: 10.1016/j.ejogrb.2011.09.024.
57. O. T. Okube and L. M. Sambu, "Determinants of preterm birth at the postnatal ward of Kenyatta National Hospital, Nairobi, Kenya," Open Journal of Obstetrics and Gynecology, vol. 7, no. 09, p. 973, 2017.
58. P. Wagura, A. Wasunna, A. Laving, D. Wamalwa, and P. Ng'ang'a, "Prevalence and factors associated with preterm birth at kenyatta national hospital," BMC pregnancy and childbirth, vol. 18, no. 1, pp. 1–8, 2018.
59. A. Muhihi *et al.*, "Risk factors for small-for-gestational-age and preterm births among 19,269 Tanzanian newborns," *BMC Pregnancy Childbirth*, vol. 16, no. 1, Art. no. 1, Dec. 2016, doi: 10.1186/s12884-016-0900-5.
60. G. Fetene, T. Tesfaye, Y. Negesse, and D. Dulla, "Factors associated with preterm birth among mothers who gave birth at public Hospitals in Sidama regional state, Southeast Ethiopia: Unmatched case-control study," PLOS ONE, vol. 17, no. 4, p. e0265594, Apr. 2022, doi: 10.1371/journal.pone.0265594.
61. W. K. Axame, F. N. Binka, and M. Kweku, "Prevalence and Factors Associated with Low Birth Weight and Preterm Delivery in the Ho Municipality of Ghana," *Advances in Public Health*, vol. 2022, p. e3955869, Feb. 2022, doi: 10.1155/2022/3955869.
62. R. L. Goldenberg *et al.*, "The preterm prediction study: the value of new vs standard risk factors in predicting early and all spontaneous preterm births. NICHD MFMU Network.," *Am J Public Health*, vol. 88, no. 2, pp. 233–238, Feb. 1998, doi: 10.2105/AJPH.88.2.233.
63. I. Vovsha *et al.*, "Using kernel methods and model selection for prediction of preterm birth," in *Machine Learning for Healthcare Conference*, 2016, pp. 55–72.
64. L. K. Goodwin, M. A. Iannacchione, W. E. Hammond, P. Crockett, S. Maher, and K. Schlitz, "Data Mining Methods Find Demographic Predictors of Preterm Birth," *Nursing Research*, vol. 50, no. 6, pp. 340–345, Dec. 2001, Accessed: Mar. 28, 2022. [Online]. Available: https://journals.lww.com/nursingresearchonline/Abstract/2001/11000/Data_Mining_Methods_Find_Demographic_Predictors_of.3.aspx
65. A. Weber *et al.*, "Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women," Annals of Epidemiology, vol. 28, no. 11, pp. 783–789, Nov. 2018, doi: 10.1016/j.annepidem.2018.08.008.
66. A. Koivu and M. Sairanen, "Predicting risk of stillbirth and preterm pregnancies with machine learning," Health Inf Sci Syst, vol. 8, no. 1, p. 14, Mar. 2020, doi: 10.1007/s13755-020-00105-9.
67. Q. Sun *et al.*, "Machine Learning-Based Prediction Model of Preterm Birth Using Electronic Health Record," *Journal of Healthcare Engineering*, vol. 2022, p. e9635526, Apr. 2022, doi: 10.1155/2022/9635526.
68. N. S. Prema and M. P. Pushpalatha, "Machine Learning Approach for Preterm Birth Prediction Based on Maternal Chronic Conditions," in *Emerging Research in Electronics, Computer Science and Technology*, vol. 545, V. Sridhar, M. C. Padma, and K. A. R. Rao, Eds. Singapore:

Springer Singapore, 2019, pp. 581–588. doi: 10.1007/978-981-13-5802-9_52.

69. R. Raja, I. Mukherjee, and B. K. Sarkar, "A Machine Learning-Based Prediction Model for Preterm Birth in Rural India," *J Healthc Eng*, vol. 2021, p. 6665573, Jun. 2021, doi: 10.1155/2021/6665573.
70. A. Batoul *et al.*, "Using Support Vector Machines In Predicting And Classifying Factors Affecting Preterm Delivery," vol. 7, no. 3, pp. 37–42, Jan. 2016, Accessed: Apr. 29, 2022. [Online]. Available: https://www.sid.ir/en/Journal/ViewPaper.aspx?ID=511030
71. J. Tsai and G. Bond, "A comparison of electronic records to paper records in mental health centers," *International Journal for Quality in Health Care*, vol. 20, no. 2, pp. 136–143, Apr. 2008, doi: 10.1093/intqhc/mzm064.
72. N. Menachemi, C. Saunders, A. Chukmaitov, M. C. Matthews, and R. G. Brooks, "Hospital adoption of information technologies and improved patient safety: A study of 98 hospitals in Florida.," Journal of Healthcare Management, vol. 52, no. 6, 2007.
73. S. Beck *et al.*, "The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity," Bulletin of the World Health Organization, vol. 88, pp. 31–38, 2010.
74. I. Petersen *et al.*, "Health indicator recording in UK primary care electronic health records: key implications for handling missing data," Clin Epidemiol, vol. 11, pp. 157–167, Feb. 2019, doi: 10.2147/CLEP.S191437.
75. H. Kang, "The prevention and handling of the missing data," Korean J Anesthesiol, vol. 64, no. 5, pp. 402–406, May 2013, doi: 10.4097/kjae.2013.64.5.402.
76. S. M. Z. Memon, R. Wamala, and I. H. Kabano, "Missing Data Analysis Using Statistical and Machine Learning Methods in Facility-Based Maternal Health Records," SN COMPUT. SCI., vol. 3, no. 5, p. 355, Jul. 2022, doi: 10.1007/s42979-022-01249-z.
77. D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," Bioinformatics, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/bioinformatics/btr597.
78. J. Tolles and W. J. Meurer, "Logistic Regression: Relating Patient Characteristics to Outcomes," JAMA, vol. 316, no. 5, pp. 533–534, Aug. 2016, doi: 10.1001/jama.2016.7653.
79. R. Kohavi and R. Quinlan, "Decision Tree Discovery Handbook of Data Mining and Knowledge Discovery." Oxford University Press, Oxford, 2002.
80. A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, no. 6, pp. 275–285, 2004.
81. J. R. Quinlan, "Learning decision tree classifiers," ACM Computing Surveys (CSUR), vol. 28, no. 1, pp. 71–72, 1996.
82. C. H. Gladwin, *Ethnographic decision tree modeling*, vol. 19. Sage, 1989.
83. G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.
84. N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, Aug. 1992, doi: 10.1080/00031305.1992.10475879.
85. O. Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm," *Medium*, Jul. 14, 2019. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (accessed Jun. 22, 2022).
86. W. S. Noble, "What is a support vector machine?," *Nat Biotechnol*, vol. 24, no. 12, Art. no. 12, Dec. 2006, doi: 10.1038/nbt1206-1565.
87. S. Suthaharan, "Support Vector Machine," in *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, S. Suthaharan, Ed. Boston, MA: Springer US, 2016, pp. 207–235. doi: 10.1007/978-1-4899-7641-3_9.
88. P. A. Flach and N. Lachiche, "Naive Bayesian classification of structured data," Machine learning, vol. 57, no. 3, pp. 233–269, 2004.
89. J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng, and D. Cheung, "Naive bayes classification of uncertain data," in *2009 Ninth IEEE international conference on data mining*, 2009, pp. 944–949.
90. J. Pervin, S. M. Rahman, M. Rahman, S. Aktar, and A. Rahman, "Association between antenatal care visit and preterm birth: a cohort study in rural Bangladesh," *BMJ Open*, vol. 10, no. 7, p. e036699, Jul. 2020, doi: 10.1136/bmjopen-2019-036699.
91. E. L. Davies, J. S. Bell, and S. Bhattacharya, "Preeclampsia and preterm delivery: A population-based case–control study," Hypertension in Pregnancy, vol. 35, no. 4, pp. 510–519, Oct. 2016, doi: 10.1080/10641955.2016.1190846.
92. A. M. Vintzileos, C. V. Ananth, J. C. Smulian, W. E. Scorza, and R. A. Knuppel, "The impact of prenatal care in the United States on preterm births in the presence and absence of antenatal high-risk conditions," *American Journal of Obstetrics and Gynecology*, vol. 187, no. 5, pp. 1254–1257, Nov. 2002, doi: 10.1067/mob.2002.127140.
93. R. Ratzon, E. Sheiner, and I. Shoham-Vardi, "The role of prenatal care in recurrent preterm birth," European Journal of Obstetrics & *Gynecology and Reproductive Biology*, vol. 154, no. 1, pp. 40–44, Jan. 2011, doi: 10.1016/j.ejogrb.2010.08.011.

94. A. H. Schempf, A. M. Branum, S. L. Lukacs, and K. C. Schoendorf, “Maternal age and parity-associated risks of preterm birth: differences by race/ethnicity,” Paediatric and Perinatal Epidemiology, vol. 21, no. 1, pp. 34–43, 2007, doi: 10.1111/j.1365-3016.2007.00785.x.
95. H. Honest, L. M. Bachmann, J. K. Gupta, J. Kleijnen, and K. S. Khan, “Accuracy of cervicovaginal fetal fibronectin test in predicting risk of spontaneous preterm birth: systematic review,” *BMJ*, vol. 325, no. 7359, p. 301, Aug. 2002, doi: 10.1136/bmj.325.7359.301.
96. V. Berghella, A. Roman, C. Daskalakis, A. Ness, and J. K. Baxter, “Gestational Age at Cervical Length Measurement and Incidence of Preterm Birth,” *Obstetrics & Gynecology*, vol. 110, no. 2 Part 1, pp. 311–317, Aug. 2007, doi: 10.1097/01.AOG.0000270112.05025.1d.

Figures



9

Figure 1

Study Workflow

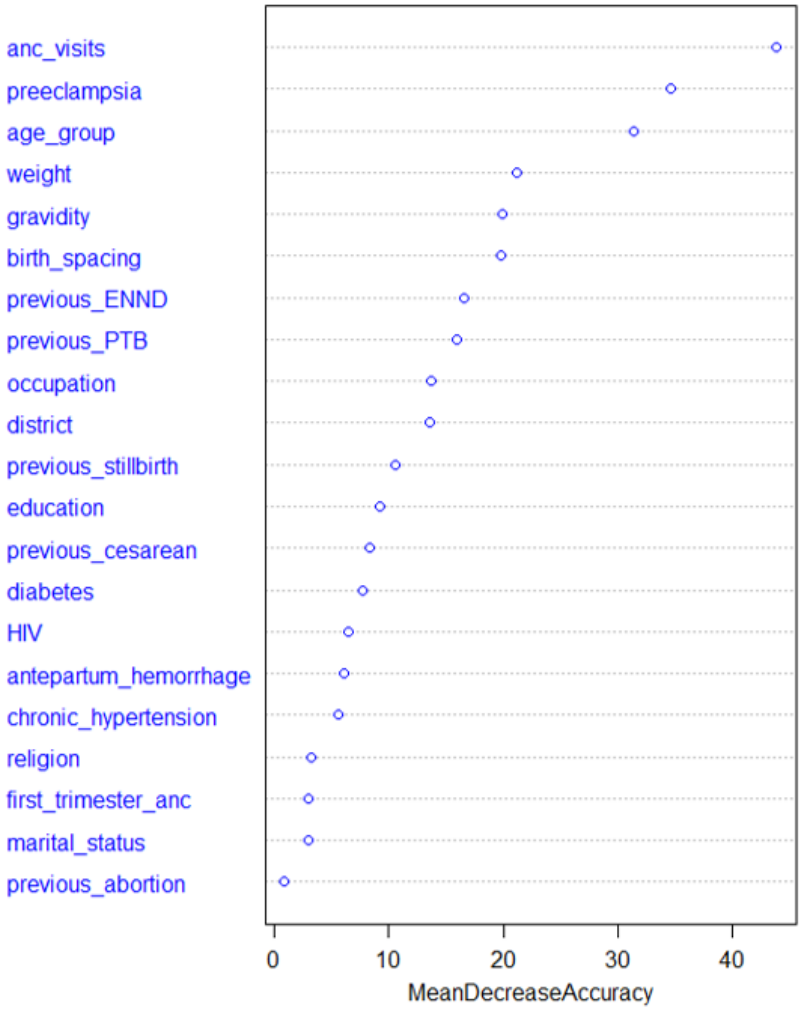


Figure 2

Variable Importance in Random Forest Model