Early Prediction of LBW Cases via Minimum Error Rate classifier: A Statistical Machine Learning Approach

***Abstract*— Low Birth weight (LBW) acts as an indicator of sickness in newborn babies. LBW is closely associated with infant mortality as well as various health outcomes later in life. Various studies show strong correlation between maternal health during pregnancy and the child's birth weight. This manuscript exploits machine learning techniques to gain useful information from health indicators of pregnant women for early detection of potential LBW cases. The forecasting problem has been reformulated as a classification problem between LBW and NOT-LBW classes using the Bayes' minimum error rate classifier rendering LBW detection as a binary machine classification problem. Expectedly, the proposed model achieved accuracy of 96.77%. Indian health care data was used to construct decision rules to be extrapolated to predictive health care in smart cities. A screening tool based on the decision model is developed to assist health care professionals in Obstetrics and Gynecology (OBG).**

***Index Terms*— Low Birth weight (LBW), Smart health informatics, Minimum error rate classifier, Predictive analytics, Machine Learning (ML), Feature Ranking.**

I. INTRODUCTION

Low birth weight is the term used to refer to babies born with a weight less than 2500gm [1]. Low birth weight(LBW) has been identified as a major public health problem around the world. LBW includes both pre-term babies as well as fully grown babies who are very small in size as a consequence of intra uterine growth retardation[2]. Birth weight is closely associated with neonatal and infant mortality,mortality rates being significantly higher in LBW babies when compared to the normal birth weight(NBW) babies[3]. This phenomenon is now of global concern in the view of serious short term and long term problems such as development disorders, neurosensory outcomes, health outcomes including Type 2 diabetes, cerebral stroke, hypertension and various other disorders that LBW babies are prone to. Studies in 2013 showed that out of the 22 million newborns about 16 percent were low birth weight cases globally[3]. This is a major problem in developing countries, especially in India which contributes to about 30 percent of the global LBW cases [3]. Innumerable studies around the world indicate strong between maternal health and impact on birth weight of babies[4]. Popular assumptions claim that LBW can be considerably reduced, with dedicated medical care during pregnancy. In our approach, the risk factors in pregnant women that can be easily assessed with basic methods are carefully examined throughout the gestation period and form the basis for predictions. Early detection can help in preventing the chances of LBW and also to put forward some recommendations under some intervention mechanisms.

The focus of this paper is on physical, obstetric and nutritional factors which was collected and measured by field workers on the health camp sites. Raw unstructured data is usually available as patient records from health camps. The objective is to mine important features from the raw data, examine the possible association of each of the features with the LBW issue, design a predictive model that will enable us to predict the outcomes as early as possible in pregnancy and thus recommend appropriate steps to be taken during the rest of the gestation period to reduce chances of LBW. The authors have used Bayesian model for classification purpose. A set of screening tools was developed based on the significant features in the form of a suitable scoring system, with an idea of providing special medical care to the women who come out to be predicted as at-risk patients.

II. LITERATURE SURVEY

Although a variety of factors have been identified as being predictors of LBW, only a few studies have combined these factors into a single scoring system. In [4], Kramer identified 43 features which contributed to LBW and suggested the importance of public health intervention to prevent LBW cases. The authors in [5] tried to critically examine the world data to estimate the proportion of LBW cases in developing countries. In [6], cluster sampling was used to sample mothers with children under the age of 5 years from the Block of Hoogly, West Bengal. Multiple Logistic regression model was used to show the relationship of socio-demographic factors as well as antenatal care related factors with LBW. In a study mentioned in [7], a screening tool was developed for pregnant women in West Bengal, using multivariate predictive discriminant analysis. A prospective validation reveals that the sensitivity ranges from 68.21% to 72.19% while probability of correct prediction using these tools varies from 60.3% to 65% only. The authors of [8], used two regression models to

Fig. 1: High level diagram of the smart screening tool for LBW prediction

determine the delivery weight based on both maternal and fetal measurable characteristics. The models could explain 62.9% and 59.4% of delivery weight variation for LBW babies. Sable in [9], used data from National Institute of Child Health and Human Development/Missouri Maternal and Infant Health Survey, and applied logistic regression analysis to determine that women who did not receive advice from Expert Panel were more likely to give birth to LBW babies. This formed the foundation for our aim to provide medical intervention to prevent LBW.

III. PROPOSED METHODOLOGY

The objective of this paper was to take a closer look at the LBW scenario in Southern part of India. The work is broadly classified into the following stages: Data acquisition and preprocessing, identification of significant correlates of LBW, performing the classification with the minimum error rate classifier, testing and validating the classifier using the same set of features and developing a screening tool for predictions to assist health care professionals in screening patients. The methodology is described in Fig 1. The novelty is two-fold: transforming a forecasting problem to machine classification [10] and proposing Bayesian inference driven minimum error rate classifier.

A. Data Acquisition and Preprocessing

Data collection was a crucial and exhaustive step which was carried out with the help of field-workers at health camps, at the Kurnool district of Andhra Pradesh. Andhra Pradesh has many cities that have been listed in the recent survey of smart cities. Also the state presents a unique mix of heterogeneous population of urban and rural data. The data collected was in the form of hand written patient records during the whole gestation period. All of the data is from visits to medical camps in the period between July 2015- October 2016. The preprocessing of data required handling missing data and redundant data . Missing data was handled by first eliminating cases with significant missing values for most of the features. Instances with a few data features missing were filled with the most frequent occurring value for each feature.Finally the structured data was the resulting dataset containing 101 patient reports. The authors identified 18 demographic significant features of the pregnant women and categorized them as follows:

- Physical factors: These include age, maternal height and maternal weight during different stages of pregnancy etc.
- Social factors: These include education, caste, information about whether the patient is a resident of the village or not.
- Medical/Obstetric factors: These include blood pressure in first visit and blood pressure in second visit, parity, birth interval, complaints, uterus height and medical history.
- Nutritional factors and lab reports: These include intake of iron, folic acid and hemoglobin content.

All the features were grouped into a single table consisting of only numerical values.

B. Recursive Feature elimination

Recursive feature elimination is a feature selection method based on constructing a model repeatedly, selecting the best feature and assigning order of importance to it. This process is continued with the remaining features until all the features in the dataset are exhausted. It is a greedy algorithm whose aim is to find the most relevant features in each iteration.
Recursive feature elimination has been applied to bring into light the most important features. The basic aim is to boost the accuracy of our predictive models while training those to learn from small amounts of data.

C. Feature Ranking using Machine learning

Feature ranking is extensively used in a large number of machine learning problems. The resulting features after recursive feature elimination were ranked for importance by two tree based algorithms: Random Forest and XGBoost. Each feature was given a score, higher the score more importance is the feature. To prevent any incorrect conclusions the results from the above algorithms were verified with the results of the Principal Component Analysis. The results were quite consistent across the three methods used and so the data was interpreted in terms of these features. All of the models showed the pregnant woman's community, age, weight and intake of iron and folic acid as the most significant features towards her chances of being a LBW case. It must be noted that, different machine learning techniques are used to cross-validate the order of importance

TABLE I: Covariance matrix of the four most important features determined by FR and RFE.These form the feature vector $X = [x_i]$, $i = 1,2,3,4$

	community	age	weight	IFA
community	1.01	0.06	0.16	-0.11
age	0.06	1.01	0.07	0.02
weight	0.16	0.07	1.01	0.16
IFA	-0.11	0.02	0.16	1.01

of features only. These methods are not used to train the machine and construct decision rules. We reiterate that the focus on discriminating LBW from NBW is accomplished by the minimum error rate classifier.
Feature ranking (FR) is not only important in building a better predictive model, but also for analyzing the real cause LBW in a new patient. Thus, applicability of Machine Learning algorithms/PCA in FR and RFE has been established and consistency in major features across different methods is checked. Next, we proceed to verify the normality assumption of class conditional density.

D. *Chi square*

The Chi square test is a statistical hypothesis test frequently run to check the goodness of fit. It utilizes the null hypothesis principle to check if the samples considered come from a normally distributed population. Before applying the Bayesian classifier, normality of the features is checked via the Chi square test. Since normality of the features are established using chi-square test, multivariate normal density has been used as the class-conditional density in the Bayes' model. Covariance matrix is computed as shown in Table I, and for all practical purposes, the cross co-variances are close to zero. We may assume the features to be statistically independent and rewrite the multi variate Gaussian as a product of univariate Gaussian density functions. This is why the error is small since the errors due to each feature are multiplied in the product model and cumulative error becomes smaller.

E. *Classification*

The objective of classification is to develop a model that can be utilized to decide the nature of patients as belonging to either the LBW or NOT-LBW class. This is posed nicely as a binary classification problem. If the classes are linearly separable, standard methods such as SVM or LDA may be applied. However, we wanted to design a machine learning scheme independent of linear separability criterion. Since feature importance and normality of class conditional density have been determined, we shall present our model using major features for dimensionally reduced fast and reliable computation, since the front end of the screening tool is a GUI/mobile application. This will be demonstrated later.

The whole data was partitioned into training and testing samples in the ratio of 70:30 and supplied to the minimum error rate classifier. Each of the samples has a set of selected features that give good discrimination between the two classes. It needs to be mentioned that that testing and training were done in batches in such a way that the data was balanced across the two classes. The performance of the classifier is critically dependent on the data balance.

1) Minimum Error Classifier: A set of samples from both the classes is used to train the model and is known as training set.The class labels of each sample are know a-priori. Let X be a feature vector. The selected features mentioned above are typical of the LBW problem.

We take some samples from every class π_i and measure the value of X for every class and we find probability density function (pdf) of X for every class γ_i i.e p(X | γ_i) where i varies from 1 to c. $c = 2$ is the total number of classes. Now the problem is if any unknown sample comes we have to measure the feature X for that sample and take the decision in favor of either class or not belong to any class. By probability theory,
$p(\gamma_i, X) = p(\gamma_i \mid X)p(X)$ $\quad -(1)$
or $p(\gamma_i, X) = p(X \mid \gamma_i)P(\gamma_i)$ $\quad -(2)$
where X=$[x_1 \; x_2. \; . \; . \; . \; x_d]^t$; d is dimension of feature vector; P(γ_i |X) is the probability that class is γ_i given the feature x and P(X| γ_i) is the class conditional density function for class γ_i. Using equations (1) and (2), we obtain, $P(\gamma_i \mid X)p(X) = p(X \mid \gamma_i)P(\gamma_i)$ i.e. $P(\gamma_i \mid X) = p(X \mid \gamma_i)P(\gamma_i)/p(X)$ where $P(\gamma_i \mid X)$ = posteriori probability; $P(\gamma_i)$ = prior probability; $p(X) = \sum_{j=1}^{c} p(x \mid \gamma_i)P(\gamma_i)$ and c is the total number of classes. *Under this setting,* ***Bayes decision rule*** *states that we construct a decision in favor of that class which has maximum posteriori probability*. If we consider two class case (binary), then, $P(\gamma_1 \mid X) > P(\gamma_2 \mid X) \implies \gamma_1$ else $P(\gamma_1 \mid X) < P(\gamma_2 \mid X) \implies \gamma_2$.

If we assign any new unknown sample to any one class then there is finite probability of its association to other class. This indicates the measure of error in the *Bayes Decision Rule*. It is also common knowledge that error in classification will crop up whenever features overlap. Therefore the classification risk, P(error) indicating the likelihood of an incorrect decision, is introduced. If for any unknown sample X, P($\gamma_2 \mid X$) > P($\gamma_1 \mid X$), then the classifier may decide in favor of class γ_2 producing error of the form P($\gamma_1 \mid X$). OTOH, for any unknown sample X, P($\gamma_2 \mid X$) < P($\gamma_1 \mid X$) then the classifier takes decision in favor of class γ_1 and P($\gamma_2 \mid X$) is the error. We construct a Bayesian decision rule that ensures the error be minimized. This is known as the ***Minimum Error Classifier***. $P(error) = \int min[P(\gamma_1 \mid X),[P(\gamma_2 \mid X)]p(X)dX$. If $P(\gamma_1 \mid X) = P(\gamma_2 \mid X)$, the classifier fails to arrive at a decision.

2) Minimum Risk Classifier: Some loss is always incurred if classifier takes some wrong decision regarding the classification task. We define this loss by formulating a loss function. Loss function is more general compared to the probability of error. It quantifies the loss incurred in taking an action α_i when the action (or class) is γ_j. The action α_i implies assigning unknown samples to one of the classes γ_i among all the c classes where c is the total number of classes. **We may represent loss function** $\lambda(\alpha_i \mid \gamma_j)$ as λ_{ij}. *Expected Loss can be defined as*: $R(\alpha_i \mid X) = \sum_{j=1}^{c} \lambda_{ij} P(\gamma_j \mid X)$ where γ_j (the true class is unknown) is the true class, R(α_i |X) is the expected loss or risk function known as conditional risk, α_i is the action in favor of γ_i, X is the feature vector of dimension d, $P(\gamma_j \mid X)$ is the posterior probability and c is the total number of classes. λ_{ij} should be equal to zero when $i = j$ because we take the action α_j(i.e we assign unknown sample to class γ_j) and the correct class is γ_j. However, there exists some non zero probability of taking action in favor of the other class in a binary setting. Let us consider a binary classification problem i.e $c = 2$ and γ_1 and γ_2 are the two classes

action		**state/class**		**associated probability**
α_1	$\Longrightarrow$	γ_1	$\Longrightarrow$	P($\alpha_1 \mid \gamma_1$)
α_2	$\Longrightarrow$	γ_2	$\Longrightarrow$	P($\alpha_2 \mid \gamma_2$)

$R(\alpha_1 \mid X) = \lambda_{11}P(\gamma_1 \mid X) + \lambda_{12}P(\gamma_2 \mid X)$ & $R(\alpha_2 \mid X) = \lambda_{21}P(\gamma_1 \mid X) + \lambda_{22}P(\gamma_2 \mid X)$. If $R(\alpha_1 \mid X) > R(\alpha_2 \mid X)$, we assign the unknown sample to the class γ_2 else if $R(\alpha_1 \mid X) < R(\alpha_2 \mid X)$, we assign the unknown sample to the class γ_1. This minimizes the risk in the process of taking the decision. *This type of classifier is known as **Minimum Risk Classifier***.

Let us consider the case when $R(\alpha_1 \mid X) > R(\alpha_2 \mid X)$. A modified decision rule is derived as follows:

$$\lambda_{11}P(\gamma_1 \mid X) + \lambda_{12}P(\gamma_2 \mid X) > \lambda_{21}P(\gamma_1 \mid X) + \lambda_{22}P(\gamma_2 \mid X)$$
$$(\lambda_{12} - \lambda_{22})P(\gamma_2 \mid X) > (\lambda_{21} - \lambda_{11})P(\gamma_1 \mid X)$$

where $(\lambda_{12} - \lambda_{22}) > 0$ and $(\lambda_{21} - \lambda_{11}) > 0$. This decision rule sounds accurate since the loss incurred in taking the wrong decision is always greater than the loss incurred in taking the right decision. Similarly, a decision rule for the case may be constructed where $R(\alpha_1 \mid X) < R(\alpha_2 \mid X)$

$$\lambda_{11}P(\gamma_1 \mid X) + \lambda_{12}P(\gamma_2 \mid X) < \lambda_{21}P(\gamma_1 \mid X) + \lambda_{22}P(\gamma_2 \mid X)$$
$$(\lambda_{12} - \lambda_{22})P(\gamma_2 \mid X) < (\lambda_{21} - \lambda_{11})P(\gamma_1 \mid X)$$

3) Minimum Error Rate Classifier: Define the zero-one loss function as: $\lambda_{ij} = 1 \quad when \quad i \neq j$ else $\lambda_{ij} = 0 \quad when \quad i = j$, $1 \leqslant i \leqslant c$ and $1 \leqslant j \leqslant c$ where c is the total number of classes. The expressions for risk are as follows: $R(\alpha_i \mid X) = \sum_{j=1}^{c} P(\gamma_j \mid X)$ i.e $R(\alpha_i \mid X) = 1 - P(\gamma_i \mid X)$. Clearly, $P(\gamma_i \mid X)$ has to be maximized in order to minimize $R(\alpha_i \mid X)$. The posterior probability corresponding to the class γ_i shall determine the outcome of the classifier in favor of that class.

4) Bound Under Min Max Criterion: We now proceed to analyze the bounds of risk incurred during the classification task. Define R_1 as the region in the feature space where the classifier decides γ_1 and R_2 as the region in the feature space where the classifier decides γ_2.

$$R < \int_{R_1+R_2} (P(\gamma_1)p(X \mid \gamma_1) + (1 - P(\gamma_1))p(X \mid \gamma_2))dX$$

This is the bound on risk under ***Min-Max Criterion***.

Error probabilities and the classification problem: Since we are dealing with a binary problem, we try to find the source of errors for the two class problem. This is equivalent to partitioning the decision space into two regions R_1 and R_2 and investigating the affiliation of an observation point, x to R_2 when the true class is γ_1, or $X \in R_1$ and the true class being γ_2. These are mutually exclusive and exhaustive. Note,
$P(error) = \int_{R_2} p(X \mid \gamma_1)P(\gamma_1)dx + \int_{R_1} p(X \mid \gamma_2)P(\gamma_2)dx$.
If $P(X \mid \gamma_1)P(\gamma_1) > P(X \mid \gamma_2)P(\gamma_2)$, classifying $X \in R_1$ is better since the smaller quantity will contribute to the error integral. The decision rule guarantees the lowest average error rate. In the two-class case, the general error integral may be approximated analytically to provide an upper bound. We present the result for error in the Gaussian case (two category, multi dimensional data for the LBW problem). The general error integral may be approximated by an analytical upper bound. The following corollary verifies this claim.

Corollary: P(error), probability of misclassification in the binary classification problem is bounded above by

$$P_{(\gamma_1)}^{\beta} P_{(\gamma_2)}^{1-\beta} \int p^{\beta}(x \mid \gamma_1) p^{1-\beta}(x \mid \gamma_2)dx, \quad 0 \leq \beta \leq 1$$

5) Discriminant Function: The evaluation of discriminant function for every class is the next task that needs to be accomplished. This is necessary since classifier assigns unknown sample to the class having maximum value for the discriminant function.

$R(\alpha_i \mid X)$ has to be minimized while maximizing the discriminant function in order to design the minimum risk classifier. This implies, $g_i(X) = -R(\alpha_i \mid X)$. It is known that, minimum error rate classifier requires $P(\gamma_i \mid X)$ to be maximized while discriminant function needs to be maximized. This yields the expression for the discriminant function,

$g(X) = P(\gamma_1 \mid X) - P(\gamma_2 \mid X)$ &

$$g(X) = \ln \frac{p(X \mid \gamma_1)}{p(X \mid \gamma_2)} + \ln \frac{P(\gamma_1)}{P(\gamma_2)} \qquad -(3)$$

To check normality of features we perform Chi square test. This justifies using a multivariate normal probability density function when the feature vector is multidimensional. The covariance matrix of those features are computed and the cross covariances are zero implying that the features are statistically independent. Thus, the multivariate normal can be written as a product of univariate normals. The class conditional density for class γ_i with mean μ_i and covariance matrix $\sum_i$ can now be defined as $p(X \mid \gamma_i) = \frac{1}{(2\pi)^{d/2}|\sum_i|^{1/2}} \exp[\frac{-1}{2}((X-\mu_i)^t \sum_i{}^{-1}(X-\mu_i)], d = 4$. Consequently, we obtain the expression for discriminant function of class γ_i using the class conditional density of class γ_i i.e $p(X \mid \gamma_i)$ as $g_i(X) = \ln p(X \mid \gamma_i) + \ln P(\gamma_i)$. In other words, $g_i(X) = \frac{-1}{2}[(X-\mu_i)^t \sum_i{}^{-1}(X-\mu_i)] - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln \mid \sum_i \mid + \ln P(\gamma_i)$. This is the discriminant function expression of class γ_i. It is quadratic due to the term $\frac{-1}{2}[(X-\mu_i)^t \sum_i (X-\mu_i)]$. This implies the Bayes' classifier can take care of linearly non separable classes. Moreover, the multivariate density is rewritten as the product making the decision region discernible. Therefore,

$$p(X \mid \gamma_i) = \prod_{i=1}^{4} \frac{1}{(2\pi)^{1/2}\sigma_i{}^{1/2}} \sum_{i=1}^{4} \exp[\frac{-1}{2\sigma_i{}^2}((X-\mu_i)^2]$$

IV. The Loss function and Risk bound in LBW classification

Losses which may occur during the classification are described in the following manner: λ_{11} = loss occurred when we take action for a journal in favor of class LBW and the true class is LBW, λ_{12} = loss occurred when we take action for a journal in favor of class NOT-LBW and the true class is LBW, λ_{21} = loss occurred when we take action for a journal in favor of class LBW and the true class is NOT-LBW and λ_{22} = loss occurred when we take action for a journal in favor of class NOT-LBW and the true class is NOT-LBW. Intuitively, $\lambda_{11} < \lambda_{12}$ and $\lambda_{21} > \lambda_{22}$ because the loss occurred in making a wrong decision is obviously greater than the loss occurred in making the correct one. If a patient is classified as LBW the expected risk associated with this action is defined as: $R(\alpha_1 \mid X) = \lambda_{11}P(\gamma_1 \mid X) + \lambda_{12}P(\gamma_2 \mid X)$ where α_1 is the action taken in favor of class LBW for a patient and X is a d dimensional feature vector. OTOH, if we classify a patient as NOT-LBW, the expected risk due to this action is formulated as: $R(\alpha_2 \mid X) = \lambda_{21}P(\gamma_1 \mid X) + \lambda_{22}P(\gamma_2 \mid X)$ where α_2 is the action taken in the favor of class NOT-LBW for a patient. This yields the following decision rule: If $R(\alpha_1 \mid X) < R(\alpha_2 \mid X)$ then $1-P(\gamma_1 \mid X) < 1-P(\gamma_2 \mid X)$, *implying* $P(\gamma_1 \mid X) > P(\gamma_2 \mid X)$; *assign patient to class LBW.* However, if $R(\alpha_1 \mid X) > R(\alpha_2 \mid X)$ then $1-P(\gamma_1 \mid X) > 1-P(\gamma_2 \mid X)$ *implying* $P(\gamma_1 \mid X) < P(\gamma_2 \mid X)$; *assign patient to class NOT-LBW*. The rule is translated in the form of an algorithm described below.

Algorithm 1 Minimum Error Rate Classifier for LBW prediction

Input: d dimensional feature Vector X
Output: Class Labels, **LBW** and **NBW**
Train classifier with training samples
for i from 1 to c **do** ▷ two class problem and c is the no. of classes
  calculate mean vector μ_i , covariance matrix$\sum_i$
  calculate class conditional density (assume independence) ▷ multivariate normal density, rewritten as product of univariate normal density.
  calculate prior probability $P(\gamma_i)$
  calculate posteriori probability $P(\gamma_i \mid X)$==$\frac{p(X \mid \gamma_i)P(\gamma_i)}{p(X)}$
  calculate expected loss or risk function $R(\alpha_i \mid X)$=1-$P(\gamma_i \mid X)$
  calculate discriminant function $g_i(X)$=$P(\gamma_i \mid X)$
end for
calculate error bound
$\int P(\gamma_1 \mid x)P(\gamma_2 \mid x)p(x)dx \leqslant$ P(error) $\leqslant \int 2P(\gamma_1 \mid x)P(\gamma_2 \mid x)p(x)dx$ ▷ two class problem
if $g_1(X) > g_2(X)$ **then**
  signify that $P(\gamma_1 \mid X) > P(\gamma_2 \mid X) \implies R(\alpha_i \mid X) < R(\alpha_i \mid X)$
  Assign the sample to NOT-LBW class
else
  signify that $P(\gamma_1 \mid X) < P(\gamma_2 \mid X) \implies R(\alpha_i \mid X) > R(\alpha_i \mid X)$
  Assign the sample to LBW class
end if
return class label **LBW** or **NBW**.

V. Results of minimum error rate classifier

The proposed model allows us to classify patients as LBW or NOT-LBW (NBW). We use a 10-fold cross validation method for the classification. The original training data is split into 10 equal subsets. A single subset is retained as the validation data for testing the model. The remaining 9 subsets are used as training set. The cross-validation process is then repeated 10 times, with each of the subsets used exactly once as the validation set. The results are then averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. The results of the proposed classifier are reported in the form of frequently used metrics such as accuracy, Sensitivity, Specificity, Precision, Recall, F-score, and ROC.

VI. Conclusion

Low birth weight is a major concern in developing countries of South Asia. Some of the maternal features are uncontrollable like the woman's height or community, while some factors can be eliminated or reduced with appropriate intervention mechanisms. Targeted and

TABLE II: Results obtained from minimum error rate classifier. The high F-score is representative of the accuracy of the designed classifier where dataset exhibits data bias towards one class, NBW. The classifier is able to achieve high precision and high recall simultaneously.

Accuracy	Sensitivity	Specificity	F-score
0.9677	1.0	0.8571	0.9795

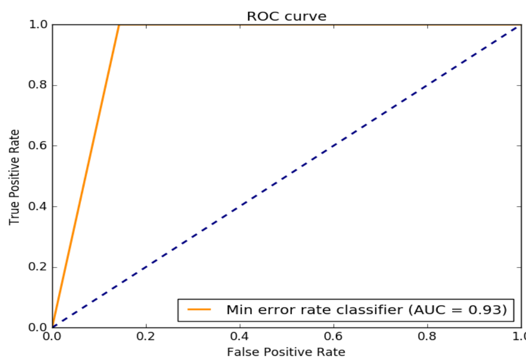

Fig. 2: True positive rate vs. False positive rate: The true positive rate (sensitivity) is plotted in a function of the false positive rate (specificity) for various threshold values. Area under the curve (AUC) is 0.93 validating the claims of good accuracy and reliable prediction. The curve is closer to the upper left corner indicating significantly high accuracy of the classifier despite small size of the data set.

personalized medical care given in the right direction within the early stages of pregnancy can go a long way in preventing LBW. The predictive analytics tool built on the principles of ML achieve this remarkably well. Standard training procedures in ML adopt a train:test ratio of $80:20$ or $90:10$ to ensure decent accuracy. We achieved remarkable accuracy with a lower ratio of $70:30$. This indicates significantly greater reliability and robustness of the learning model and should dispel concerns about using the outcome to urban case studies.

The indicative study may be extended further to smart health care in urban areas as well. It is observed that if the community factor is removed, the other factors are still relevant in the educated upwardly mobile population in Urban areas and smart cities. We may infer intuitively, based on the outcome of the model, that the order of feature importance shall be limited among the rest of the features since the generic list is common among urban and rural population, according to medical practitioners. The inference seems correct, as found in [11]. The discovery from data is exhaustive and rich. The relative interchange among the features to be displayed in the screening tool is expected to be cosmetic and subject to minor modifications in the back end implementation, without incurring additional overhead. The data discovery toward mapping Urban populace for LBW is optimistic but the authors believe, strongly, that it depends on the availability of data (new data) and the quality of the model will not be impacted negatively. The novelty in posing the problem as machine classification via minimum error rate classifier and tight theoretical bounds ensure robustness and reliability of LBW prediction. It is a conundrum that several pilot studies and published literature failed to identify a suitable machine learning approach. The authors are hopeful that the manuscript shall serve the purpose of providing the momentum in this direction and outperform the team of human experts.

REFERENCES

[1] World Health Organization-1992,International statistical classification of diseases and related health problems,Tenth revision, Geneva, World Health Organization.

[2] World Health Organization Maternal Health and Safe Motherhood Programme-1992,Low Birth Weight: A tabulation of available information, Geneva, WHO.

[3] United nations Children's Fund and World Health Organization-2004,Low Birth Weight: Country, regional and global estimates, New York, UNICEF.

[4] Kramer M.S-1987, Determinants of Low Birth Weight: Methodological assessment and meta-analysis, Bulletin of the World Health Organization, Volume 65, Number 5, Page 663-737.

[5] Ann K. Blanc and Tessa Wardlaw-2005, Monitoring low birth weight: an evaluation of international estimates and an updated estimation procedure, Bulletin of World Health Organization, Volume 83, Number 3, Page 178-185.

[6] Aparajita Dasgupta and Rivu Basu-2011, Determinants of low birth weight in a block of hooghly, west bengal:a multivariate analysis, International Journal of Biological & Medical Research Volume 2, Number 4, Page 838-842.

[7] Biswanath Das,2001-2004,Development of a Screening Tool for prediction of Low Birth weight babies by community Health Workers, sponsored by the Dept. of Family Welfare, Ministry of Health and Family Welfare, Govt. of India .

[8] Mali Abdollahian and Nadeera Gunaratne-2015, Low Birth weight prediction based on maternal and fetal characteristics, 12th International Conference on Information Technology-New Generations, Page 646-650.

[9] Sable M.R and Herman A.A-1997, The relationship between prenatal health behavior advice and low birth weight, Public Health Rep, Volume 112, Number 4, Page 332-339.

[10] L. Khaidem, S. Saha and S. Roy Dey-2016, Predicting the direction of stock market prices using random forest, arXiv:1605.00003 [cs.LG], Page 1-20.

[11] J.S. Deshmukh, D.D. Motghare, S.P. Zodpey and S.K. Wadhva-1998,Low Birth Weight And Associated Maternal Factors in an Urban Area, Indian Pediatrics,Volume 35, Page 33-36.