European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 08, 2020

Classification Of Caesarian Data Using Machine Learning Models

Dr K Butchi Raju[1],[2] Prasanna Kumar Lakineni, [3]K sarada, [4]M.Kiran Kumar, [5]k saikumar

[1]*Department of CSE, GRIET, Hyderabad, Telangana, INDIA*
[2] *Associate Professor, Dadi Institute of Engineering and Technology, Visakhapatnam, A.P. INDIA*
[3]*Associate professor, Dept of EEE, KLEF, India, saradak@kluniversity.in*
[4]*Research Scholar, Sri Satya Sai University of Technology and Medical Sciences, Sehore.*
[5]*JRF, Dept of ECE, KLEF, India, saikumarkayam4@gmail.com*

Abstract –
There is a substantial rise in population but there are still a large number of villages where health facilities are not still they depend on the conventional methods during times of pregnancy which resulting in some times of deaths and at the same time due to the miscalculations of the doctors the unnecessary women are being Caesarian which eventually leading to many side effects for them. So, in this paper we are using different classifiers to predict Caesarian is required or not so the doctors can consult to have normal delivery or move to another high facilities hospitals.

Key words: Classification; Women; Caesarian; Hospitals; Pregnancy

1. INTRODUCTION

Classification has been taken as one basics of the machine learning models especially in the supervised learning. That classification can be broadly defined as the predicting a discrete class output or state. One of my primary areas of focus in this paper is health care , today health care systems has taken to such an extent for detecting and cure of diseases with the help of advanced machine learning algorithms describing by giving the features of a person the model should predict whether the disease is present or not and we have happened to see some datasets on predicting the parkinsons disease.Now our dataset is a Caesarian dataset and main motto is to predict whether the women requires Caesarian or not and this one of the dataset found in the UCI website and the research is being done by the many health institutes so they can easily predict the required outputs. This could be really useful for the doctors to improve their suggestions. Our motivation to this dataset because most women are necessarily or unnecessarily being done Caesarian which results in decreasing their strength.

2. LITERATURE SURVEY

Researchers of computer learning methods have demonstrated interest in the area of medicine in recent years. Biomedical, genomic and medicinal diagnosis analysis includes different methods for machine learning. These methods have been utilised to compare, forecast, and predict. We are addressing a variety of studies that are relevant to our work.

D Kavithaand T. Balasubramanian[1] Validate and recommend a model decision tree for forecasting the mode of delivery and the risk factors relevant to caesarean delivery.

Betrán AP et al., [2] Nationally representative data collected on CS concentrations and regional and subregional estimated weighted averages from 1990 and 2014. The data collected Authors also carried out a longitudinal study estimating the CS ranking variations as an absolute adjustment and as an annual average rate of increase (AARI).

Robson SJ et al., [3] Using global evidence from the Australian Children's Longitudinal Review, recognise previously unsuspected threats correlated with caesarean. Data is from the cohort of

birth, a long-term longitudinal analysis include some 5000 girls, which provides rich data on maternal health and pregnancy exposures. The contribution of a broad variety of deliveries, births and social variables to caesarean was investigated by the logistic regression.

Athukorala C et al., [4] Consider the incidence and effect on reproductive, peripartite and neonatal outcomes of mothers becoming overweight and obese during early to mid-pregnancia. A secondary study was conducted in the Australian collaborative test of antioxidant vitamins Vitamin C and Vitamin E to pregnant women for the prevention of the symptoms of pre-eclampsia on data gatherings of nulliparous women with singleton pregnancy (ACTS).Smith GC et al., [5] Its goal was: (1) to define the relationship between maternal age and labour outcomes; (2) to evaluate the share of increase in primary caesarean rates due to increases in maternal age distribution; and (3) to determine whether the uterine smooth muscle (myometrium) contractility was variable at maternal age.

Brennan DJ et al., [6] The goal of this research was to emphasise differences and behaviours in obstetric communities and to classify improvements in CS rates in various institutions. Data from 9 institutional cohort countries were analysing in 9 countries, centred on the 10-group classification scheme of 4 características in each pregnancy: single/multifarial, nulliparity/multiparity, CS scar multiparity, spontaneous/inductive work onset and gestation term (37 weeks). Data were obtained from 9 different countries and were analysed from 9 different institutions.Wiklund I et al., [7] In the absence of a medical indication, examine first time mothers performing caesarean section; their motivation for the request; self-estimated fitness, childbirth experience; and breastfeeding period. They also explored whether symptoms of postpartum depression are more prevalent in this population.

McCourt C et al., [8] Study the research paper published after Gamble and Creedy attacked it in 2000, on the female chosen or elective cessarean portion

3. METHODOLOGY

If historical data are available, predictive analysis can be made. Multiple sources can combine data. To be used with ML techniques, it should be washed and converted. Our BPC model architecture is seen in the Figure. 1. There are three stages of our design.In the first phase data remainsdownloaded from UCI repository. This knowledge is cleaned and converted in order to reach the refined data collection. The ML strategies for achieving the classification are implemented in the second step. Decision tree, Gaussian Process, Bernoulli NB, AdaBoost, SVM, K Neighbours, XG Boost and Gradient Boosting classifiers are used aimed at the classification. We use these two algorithms by way of they are the best classifying models. In the third phase we measure the performance of the classifiers with confusion matrix values and Accuracy. Precision, Recall and F1 measures.

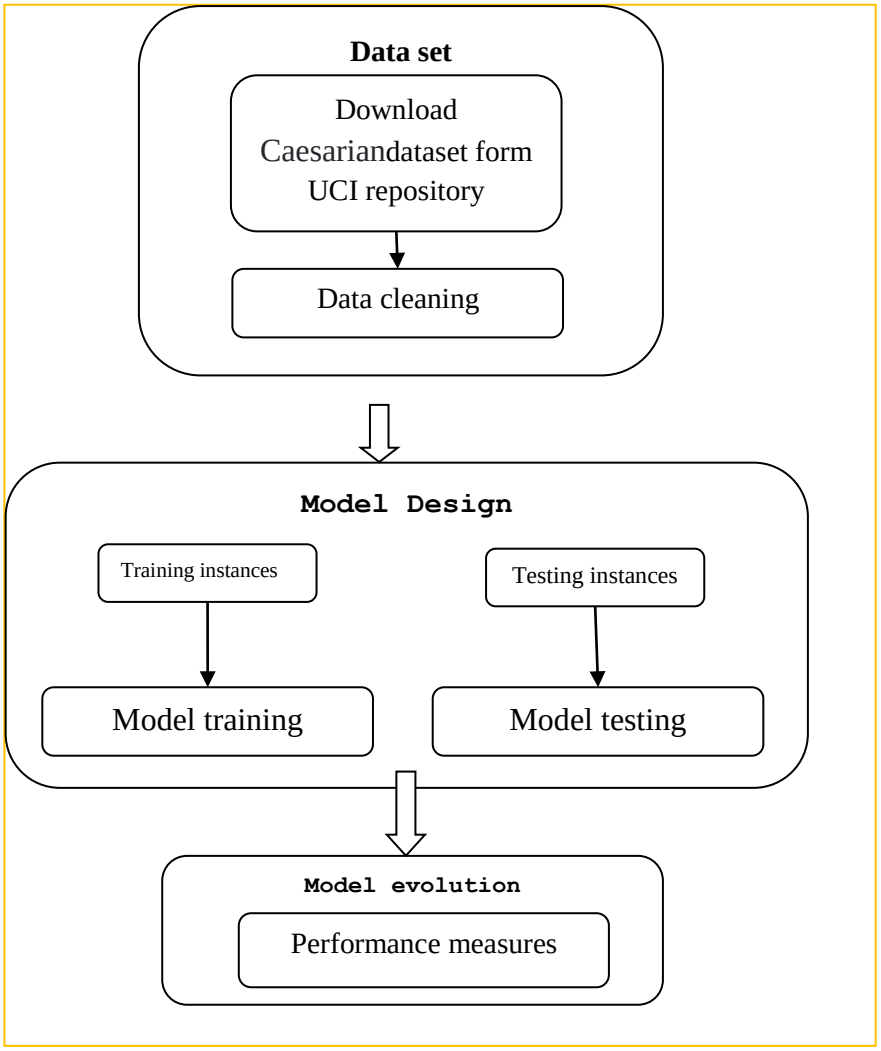

Fig 1 System Design

3.1. ML TECHNIQUES

In this section we described ML techniques which are used in this process is discussed.

Decision tree[9]: The first classificator we have used is the commonly recognised Decision Tree classification, which is why Decision Tree Classifier is an easy and frequently used grading technique. The solution of the classification problem is based on a basic premise. Tree Classifier Judgment provides a range of well-designed questions on the characteristics of the test record. It splits down a dataset into smaller and smaller sub-sets while concurrently creating an associated decision tree. The end product is a tree with leaf nodes and judgement nodes. There are two or three divisions of a judgement node (e. Outlook) (e.g., Sunny, Overcast and Rainy). Leaf node (e.g. Play) is the ranking or judgement. The highest judgement node in a tree that fits the strongest indicator defined as a root node. Both numerical and category data may be processed through decisions trees.

BernoulliNB[10]: The another that we are using is BernoulliNB the same steps will be taken for this too. This model can be good for the one with the outputs with only 0 or 1.This model is common for the classification of records, where binary term occurrence functions are used instead of term frequencies.

AdaBoost[11]: AdaBoost is a boosting type Ensembling method. The strengths of this model is it is very fast, one of the model for classsifer and it will have less tendency to overfit.Adaboost is a boosting type ensemble learner. This method works by combining multiple individual "weak" learning hypotheses to create one strong model. Each weak hypothesis used is better at classifying the data than random chance. However, it's the combination of all of these

independent weak learning hypotheses what makes the model more capable of predicting accurately on unseen data than each of the individual hypothesis would.
SupportVectorMachines[12]: This model is effective for the problems with complex domains but with a clear separation of the data.The weakness of this model is it take a huge time for the large datasets.This model is good candidate for this problem for the reason as some features in the dataset have a clear boundary**.**
KNN[13]: The KNN algorithm is a strong and robust classification scheme that often serves as a standard for complex classification systems. KNN can achieve more efficient classification systems considering its simplicity
XG Boost[14]: It is a model of tree ensemble that is a group of trees for classification and regression (CART). Trees are planted one by one and in subsequent iterations aim to reduce the misclassification rate. Each tree provides a particular prediction value based on the knowledge it sees and the values of each tree are summarized to produce the final score.
Gradient Boosting[15]:Gradient boosting identifying weak learnersthrough using gradients in the loss function. The loss function is a calculation of how well the coefficients of the model match into the base results. A rational awareness of loss function relies on what we want to maximize.

4. DATASET

This data collection provides details on the outcomes of the caesary segment of 961 pregnant women with the most severe characteristics of medical conditions. We chose age, distribution number, period of delivery, blood pressure and cardiac condition. Here every feature is important for the classification.
Our arrival period is graded as premature, timely and latecomer. In three mild, regular and high moods settings, blood pressure is treated like delivery period. Heart problem is deemed ideal and incompetent.
@attribute 'Age' { 22,26,28,27,32,36,33,23,20,29,25,37,24,18,30,40,31,19,21,35,17,38 }
@attribute 'Delivery number' { 1,2,3,4 }
@attribute 'Delivery time' { 0,1,2 } -> {0 = timely , 1 = premature , 2 = latecomer}
@attribute 'Blood of Pressure' { 2,1,0 } -> {0 = low , 1 = normal , 2 = high }
@attribute 'Heart Problem' { 1,0 } -> {0 = apt, 1 = inept }
@attribute Caesarian { 0,1 } -> {0 = No, 1 = Yes
The dataset is obtained from the reference
link**http://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification+Dataset**
The data exploration stage that every feature is important for the determination of the problem so Below we are writing a definition to obtain the barplots for the data to check with different parameters how do they get effect with the results of the caesarian.

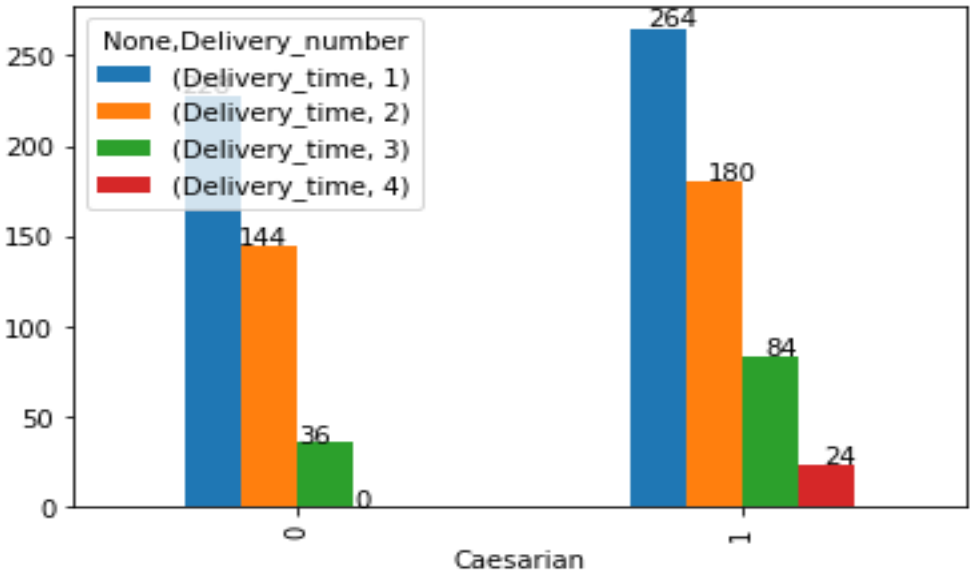

Fig 2 Bar plot for caesarian data with delivery number

The bar plot which is shown in Fig 2 for the dataset with delivery number and Caesarian the plots resemble the delivery time we can see that the when there is a delivery time 4 then the Caesarian becomes compulsory for the women this could be because of the they cannot with stand the pain of pregnancy for the 4TH time. A another bar plot is drawn between the Caesarian and blood pressure with respect to the delivery time this resembles that maintaining a normal blood reassure is good which is shown in Fig 3.

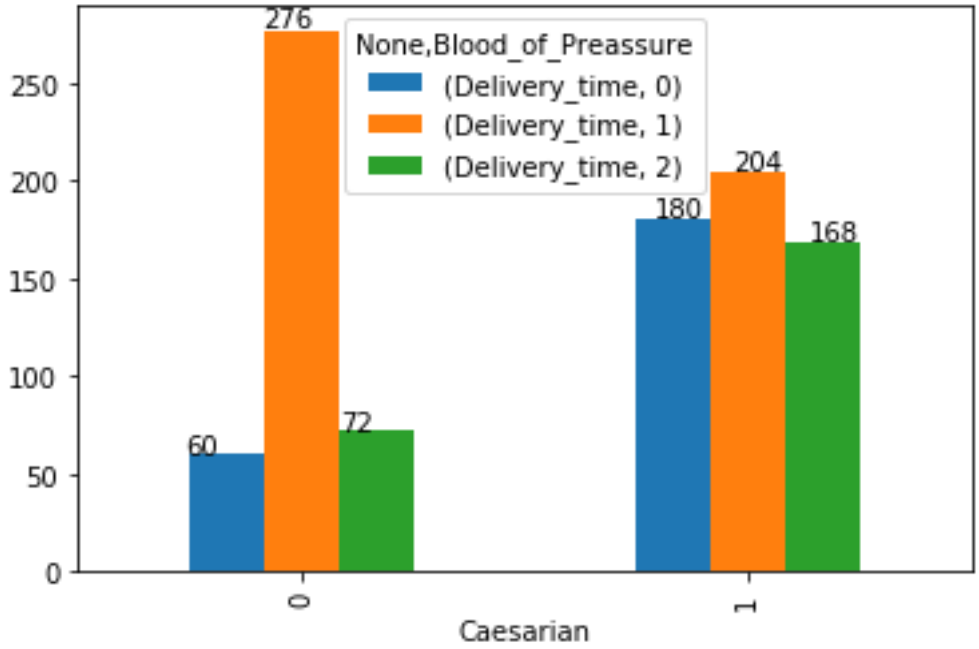

Fig 3 Caesarian and blood pressure with respect to the delivery time

The scatter mattrix is drawn for the features which is shown in Fig 4 and the skewness is checked if any skew property is present we can handle them in the next phase and we have found no sknewness in the scatter matrices and a correlation matrix is drawn and every features is important for the determination.

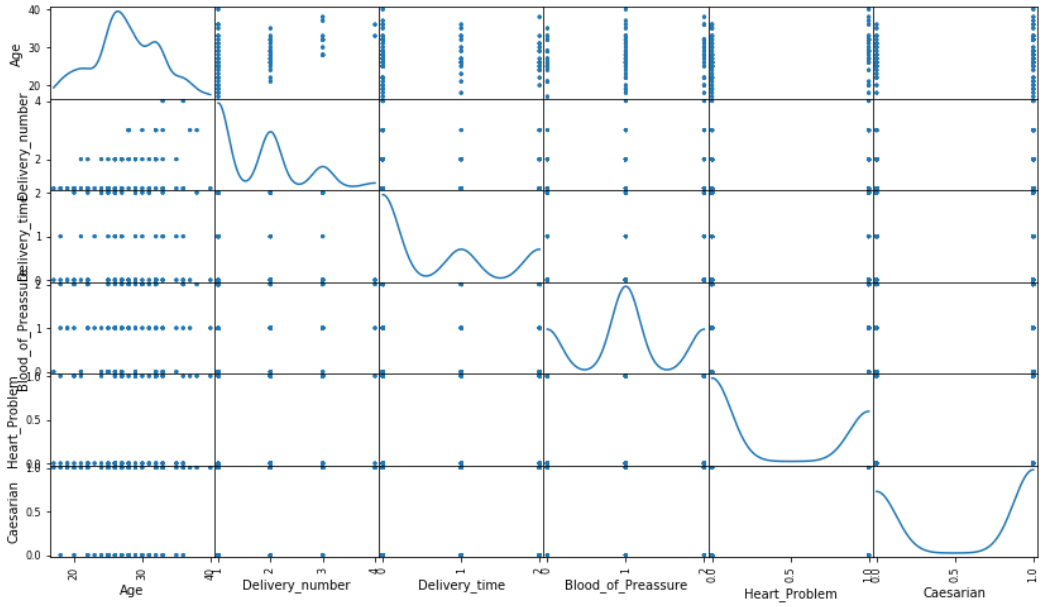

Fig 4 scatter mattrix for the features

4.1. IMPLEMENTATION

But before dive into the results of the machine learning algorithms on the dataset, we had to explore clean and rework the dataset. The reason we had to do this because the data is using a long memory and that empty columns slowed the training process of our model. So before entering the model we have to clean the dataset we have removed the missing values the dataset. After that visualization phase have been done by using the bar plots. A correlation matrix have been drawn using corr() which helped us know which feature is required for the classification.

	Age	Delivery_number	Delivery_time	Blood_of_Preassure	Heart_Problem	Caesarian
Age	1.000000	0.427160	-0.021857	0.074448	0.250485	0.077966
Delivery_number	0.427160	1.000000	-0.074017	0.134315	0.200267	0.144894
Delivery_time	-0.021857	-0.074017	1.000000	-0.087298	-0.003985	-0.166233
Blood_of_Preassure	0.074448	0.134315	-0.087298	1.000000	0.036515	-0.035760
Heart_Problem	0.250485	0.200267	-0.003985	0.036515	1.000000	0.352557
Caesarian	0.077966	0.144894	-0.166233	-0.035760	0.352557	1.000000

Fig 5 Correlation matrix

Here by observing the correlated values which are shown in Fig 5, we can consider that all the features of the dataset have a strong correlation with the Caesarian so every feature is necessary and cannot be ignored. From Fig 4, we can see that the data is continuous distribution and the Gaussian Bayes on the other hand is a method designed specifically to provide a method of normalizing continuous data into values we can apply Bayes' Theorem to. When we calculated probabilities before, we just used frequency. For continuous inputs, we find the mean and standard deviation of the input values (after removing outliers) to represent the distributions.KNeighbourstakes the large time in finding the best fit and as the parameters list increases the time of doing the job also increases. This could be the problem in this solution. The syntax that is used in the above is the utilization of the default parameters. We have applied the kfoldcross_validation both to the Decisiontree classifier and KNeighbors but they didn't gave the expected results like the normal cross_validation metric . Changing of random_state doesn't create a decrease in accuracy_score and if we add more values to the dataset it may not be affecting the dataset because if we observe the scatter matrix of the dataset every value is closely related like a point as most of the observations are similar.

5. RESULTS

Table 1 represents ML models' accuracy and these accuracies are represented as bar plots which are shown in Fig 6. By observing results shown in Table 1, thought of using the Adaboost classifier but it gave me with an accuracy of only 76% but this KNeighbours gave me an accuracy score of 95%. The evaluation metric that I have used here is the metric accuracy_score as the formula is mentioned in the above. The KNeighbours is the best to choose for the reason it

has given me the best accuracy but in the terms of time it is better to choose the XGBClassifier the new advanced boosting algorithm there is no such predominant change in accuracies between the two. we are using the evaluation metric accuracy_score for this model as this is a balanced dataset, we are in a no need to use the f1_score or recall score. My dataset will be spitted by the cross_validation.train-test-split which is goes better it will have k examples and runs on the different classifiers and the model that fits the best with the known classifier will be given under testing of the different hyper parameters using GridSearchCV. The other models that I have used for this classification are 'Decision Tree', 'GaussianProcess', 'BernoulliNB','AdaBoost','SVC','Kneighbours','XGBClassifier','GradientBoosting'.The model that gave me the greatest classifier is the KNeighbours and the greatest neighbors are 7 with a max_depth of 30 and the cv and we have applied is 10. I thought of using the Adaboost classifier but it gave me with an accuracy of only 76% but this KNeighbours gave me an accuracy score of 95%.The evaluation metric that we have used here is the metric accuracy_score as the formula is mentioned in the above. The KNeighbours is the best to choose for the reason it has given me the best accuracy but in the terms of time it is better to choose the XGBClassifier the new advanced boosting algorithm there is no such predominant change in accuracies between the two.

Table 1 Accuracy of ML models

Model Name	Accuracy
Decision Tree	0.93
Gaussian Process	0.92
Bernoulli NB	0.65
Ada Boost	0.76
SVC	0.72
KNN	0.95
XGB Classifier	0.93
Gradient Boosting	0.88

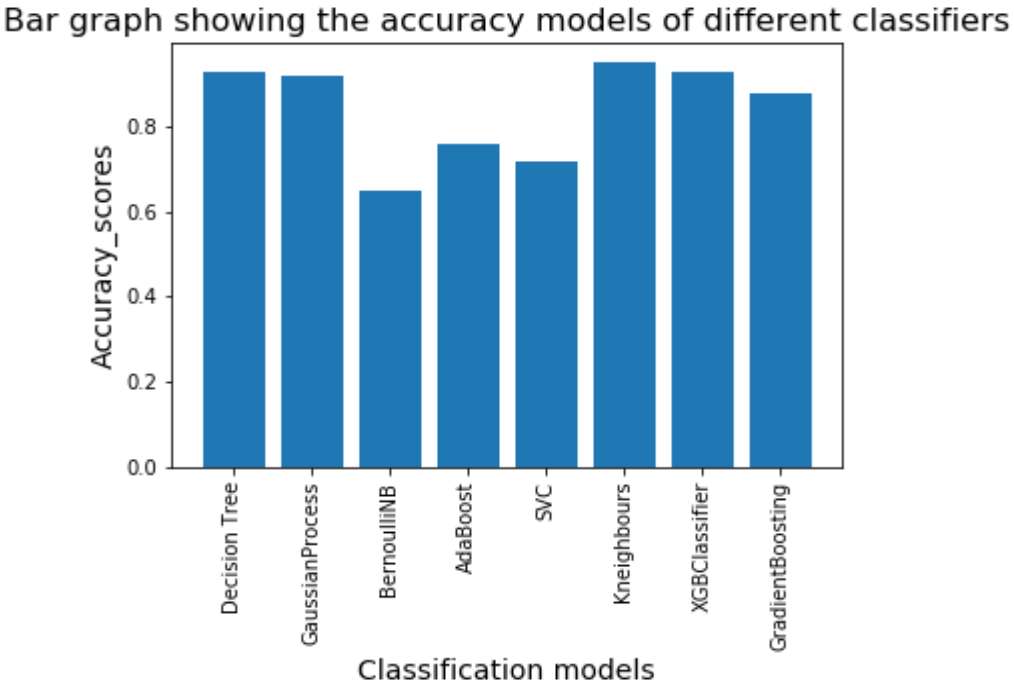

Fig 6 Accuracy comparison Bar plot

6. CONCLUSIONS

We also presented in this article the effective usage of machine learning methods in the medical field. Information and machine learning were used to derive trends of illness from the accessible medical evidence. For medical diagnosis, estimation and treatment, the patterns derived can be used. This paper has classified the caesarian data using different ML models. The classifier that best fits my model is a KNeighbors classifier. We even used the latest award-winning algorithm XGBClassifier which means extreme gradient boosting classifier and it's accuracy is nearer to the above model but KNeighbous is slow due to several computations which it takes the number of computations that it takes in my model is 202 computations. The results which are represented in this paper will useful to doctors while take the decision before doing the caesarian.

REFERENCES

[1] D Kavitha and T. Balasubramanian, "Predicting the mode of delivery and the risk factors associated with cesarean delivery using decision tree model", international journal of engineering sciences & research technology, 7(8): August, 2018
[2] Betrán AP, Ye J, Moller A-B, Zhang J, Gülmezoglu AM, Torloni MR (2016) The Increasing Trend in Caesarean Section Rates: Global, Regional and National Estimates: 1990-2014. PLoS ONE 11(2): e0148343. https://doi.org/10.1371/journal.pone.0148343
[3] Robson SJ, Vally H, Mohamed AL, Yu M, Westrupp EM. Perinatal and social factors predicting caesarean birth in a 2004 Australian birth cohort. Women and Birth. 2017 Dec 1;30(6):506-10.
[4] Athukorala C, Rumbold AR, Willson KJ, Crowther CA. The risk of adverse pregnancy outcomes in women who are overweight or obese. BMC pregnancy and childbirth. 2010 Dec;10(1):56.
[5] Smith GC, Cordeaux Y, White IR, Pasupathy D, Missfelder-Lobos H, Pell JP, Charnock-Jones DS, Fleming M. The effect of delaying childbirth on primary cesarean section rates. PLoS medicine. 2008 Jul 1;5(7):e144.
[6] Brennan DJ, Robson MS, Murphy M, O'Herlihy C. Comparative analysis of international cesarean delivery rates using 10-group classification identifies significant variation in spontaneous labor. American journal of obstetrics and gynecology. 2009 Sep 1;201(3):308-e1.
[7] Wiklund I, Edman G, Andolf E. Cesarean section on maternal request: reasons for the request, self- estimated health, expectations, experience of birth and signs of depression among first- time mothers. Actaobstetricia et gynecologicaScandinavica. 2007 Apr;86(4):451-6.
[8] McCourt C, Weaver J, Statham H, Beake S, Gamble J, Creedy DK. Elective cesarean section and decision making: a critical review of the literature. Birth. 2007 Mar;34(1):65-79.
[9] Rao H, Shi X, Rodrigue AK, Feng J, Xia Y, Elhoseny M, Yuan X, Gu L. Feature selection based on artificial bee colony and gradient boosting decision tree. Applied Soft Computing. 2019 Jan 1;74:634-42.
[10] Gao H, Zeng X, Yao C. Application of improved distributed naive Bayesian algorithms in text classification. The Journal of Supercomputing. 2019:1-7.
[11] Tavallali P, Yazdi M, Khosravi MR. Robust cascaded skin detector based on AdaBoost. Multimedia Tools and Applications. 2019 Jan 1;78(2):2599-620.
[12] Chen SG, Wu XJ. A new fuzzy twin support vector machine for pattern classification. International Journal of Machine Learning and Cybernetics. 2018 Sep 1;9(9):1553-64.
[13] Saxena R, Johri A, Deep V, Sharma P. Heart Diseases Prediction System Using CHC-TSS Evolutionary, KNN, and Decision Tree Classification Algorithm. InEmerging Technologies in Data Mining and Information Security 2019 (pp. 809-819). Springer.
[14] Li H, Pu B, Kang Y, Lu CY. Research on massive ECG data in XGBoost. Journal of Intelligent & Fuzzy Systems. 2019 Jan 1;36(2):1161-9.
[15] Biau G, Cadre B, Rouvière L. Accelerated gradient boosting. Machine Learning. 2019 Jun 15;108(6):971-92.