JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF PREGNANCY HYPERTENSION USING BOOSTING ALGORITHMS

[1]Nima Nayak, [2]Prof. Maulik Trivedi

[1]M.E. Student, [2]Assistant Professor

[1]Computer Engineering,

[1]Darshan Institute of Engineering & Technology, Rajkot, India

Abstract : Pregnancy is a delicate stage in every woman's life. In pregnancy, normally the blood pressure of a mother goes down by 10 mmHg from her normal blood pressure. And that is good for fetus health. But during pregnancy, so many complications and risks can arise and blood pressure goes up. High blood pressure is a symptom of hypertension. And it creates a risk to maternal and fetus health at several stages of pregnancy. In today's era, machine learning has developed so many techniques for the prediction of disease. The main goal of this paper is to analyze and compare different machine learning algorithms used to predict hypertensive disorders. These classification methods help to decrease maternal and fetus mortality, which remains a large problem in many developing nations, especially in rural areas.

IndexTerms **- Data mining, Machine learning algorithms, Prediction, Pregnancy**

I. INTRODUCTION

Hypertension refers to high blood pressure during pregnancy. Hypertension can be categorized into 4 different types as shown in fig. 1. First is, Chronic hypertension means high blood pressure before conceiving which is not related to pregnancy or which occurs before 20 weeks of pregnancy. As this high blood pressure usually does not have any symptoms, it might be hard to determine when it began. Second is, Gestational hypertension means having high blood pressure after the 20th week of pregnancy and there is no excess protein in the urine or other signs of organ damage. Third is, Preeclampsia-eclampsia refers to high blood pressure after the 20th week of pregnancy with the signs of damage to other organ systems like kidneys, liver, blood, or brain. Untreated preeclampsia can lead to serious complications for women and fetuses including the development of seizures. Fourth is, Chronic hypertension with superimposed preeclampsia refers to the condition in which a woman already has chronic hypertension before pregnancy and develops worsening high blood pressure and protein in the urine or other blood pressure-related complications during pregnancy [9].

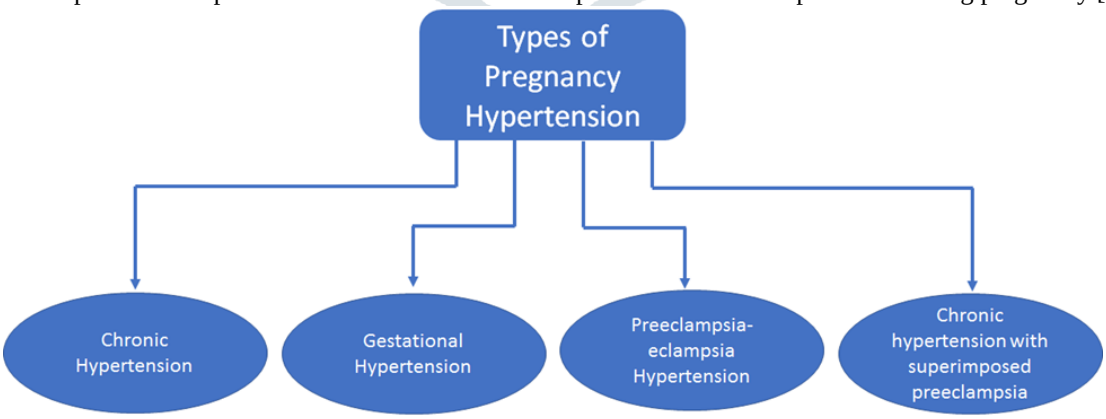

Fig.1 - types of pregnancy hypertension

II. LITERATURE SURVEY

Data mining technology is growing day by day in many fields rather than computer science. Nowadays, data mining techniques are used in the prediction of disease too. There is many research going on which uses different machine learning algorithms as well as data mining techniques to get more accurate result by making a best predictive model for pregnancy hypertension. Sirinat Wanriko [1] proposed a predictive model for the risk assessment of pregnancy-induced hypertension for eclampsia and pre-eclampsia which uses seven machine learning algorithms such as Logistic Regression (LR), K-nearest neighbor (KNN), decision tree (DT). Random forest (RF), multilayer perceptron neural network (MLP), support vector machine (SVM), and naïve Bayes (NB). The imbalanced

data was balanced using Synthetic Minority Over-Sampling Technique (SMOTE) technique and Principal Component Analysis (PCA) was used to extract 3 principal components on 17 attributes, imbalanced data and balanced data. As well as three data preprocessing methods like MinMaxScaler, StandarScaler and Normalizer were used. Random Forest (RF) had provided the highest accuracy at 89.62 percent.

Galih Malela Damaraji [2] proposed a review of an expert system for the identification of various risks on pregnancy which identifies early symptoms encountered by pregnant women. The result of an expert system is grouped based on an expert system such as rule-based and fuzzy expert system and artificial neural networks and other machine learning techniques and other data mining methods. This expert system analyzed hypertension, premature birth, pregnancy abnormalities, ectopic pregnancy.

V. Madhusri [3] stated that stress is one of the main reasons for hypertension as it affects the physical as well as mental health of pregnant women and their fetuses during pregnancy and at delivery time. The proposed stress prediction model uses four machine learning algorithms like Support Vector Machine (SVM), Naïve Bayes (NB), K – Nearest Neighbors (KNN), and Decision Tree (DT) which have been used to predict the complication during labor and for baby in the later part of the life. Here, Naïve Bayes had achieved an accuracy of 90% rather than other algorithms.

Wei Wu [4] stated that hypertension and its complications during pregnancy is the main reason which affects maternal mortality. This paper proposed a classification model of pregnancy-induced hypertension based on random forest, XgBoost algorithm, and a fusion of both models. For this fusion model, they used Stacking's classifier combination strategy which integrates multiple classifications or regression models with integrated learning techniques through a meta-classifier. Here, the fusion model had two layers: the first layer used random forest and XgBoost algorithms, and the second layer used the logistic regression model. The fusion model had higher accuracy of 83.68% which is better than the single random forest and XgBoost model.

Xinke Lan [5] applied data mining technology on various models like logistic regression, support vector machine, and random forest for gestational hypertension analysis with different methods and techniques. They also showed that high body weight, edema, and low calcium have a higher connection with hypertension during pregnancy. As a result, the random forest showed 83% accuracy compared to the other two models.

Muhlis [6] said that preeclampsia and eclampsia were one of the combinations of pregnancy caused by the pregnancy itself. This study used Neural Network to classify preeclampsia data with the comparison of other algorithms like Naïve Bayes, K-Nearest Neighbors, Linear Regression, Logistic Regression, and Support Vector Machine. Some validation tests like split data, 10-folds cross-validation and Leave One Out (LOO) are applied to these algorithms. And from that, the Neural Network algorithm achieved the best accuracy than others.

The following table 1 shows comparative analysis of Literature Surveys from different papers.

Table.1 - comparative analysis of literature survey

Sr. No.	Paper Title	Algorithms / System used	Data mining Techniques used	Outcome of the paper
1	Risk Assessment of Pregnancy-induced Hypertension Using a Machine Learning Approach [IEEE 2021]	Logistic Regression (LR), K-nearest neighbor (KNN), decision tree (DT). Random forest (RF), Multilayer Perceptron Neural Network (MLP), Support Vector Machine (SVM), Naïve Bayes (NB).	SMOTE for imbalanced data, PCA for to extract 3 principal components, Data preprocessing methods - MinMaxScaler, StandardScaler and Normalizer	Random Forest (RF) had provided the highest accuracy at 89.62 percent.
2	A Review of Expert System for Identification Various Risk in Pregnancy [IEEE 2020]	Review of an expert system for the identification of various risks on pregnancy	-	This expert system analyzed hypertension, premature birth, pregnancy abnormalities, ectopic pregnancy.
3	Performance comparison of machine learning algorithms to predict labor complications and birth defect based on stress [IEEE 2019]	Support Vector Machine (SVM), Naïve Bayes (NB), K – Nearest Neighbors (KNN), Decision Tree (DT)	Data pre-processing techniques: Rescale data algorithm, Binary Data algorithm, Standardize data algorithm	Naïve Bayes had achieved an accuracy of 90% rather than other algorithms.
4	Classification of hypertension in pregnancy based on	random forest, XgBoost algorithm,	Stacking's classifier combination strategy	The fusion model had higher accuracy of 83.68% which is better

	random forest and Xgboost fusion model [BIBE 2019]	fusion of both models.		than the single random forest and XgBoost model.
5	Research on Pregnancy Hypertension Based on Data Mining [IEEE 2018]	logistic regression, support vector machine, random forest	cross validation optimization method loss function	Random forest showed 83% accuracy.
6	Neural Networks Algorithms to Inquire Previous Preeclampsia Factors in Women with Chronic Hypertension During pregnancy in Childbirth Process [IEEE 2018]	Neural Network Naïve Bayes, K-Nearest Neighbors, Linear Regression, Logistic Regression, Support Vector Machine.	Validation tests- Split data, 10-folds cross-validation, Leave One Out (LOO),	Neural Network algorithm achieved the best accuracy than others.

The following table 2 shows different machine learning algorithms are used in different papers. It also refers to the dataset and future work of different papers.

Table.2 - summary of literature survey with dataset and future work

Year	Reference Number	Machine Learning Algorithms	Dataset	Future Work
2021	[1]	Logistic Regression K-Nearest Neighbor Decision Tree Random Forest Multilayer Perception Neural Network Support Vector Machine Naïve Bayes	Public dataset of Logan (2020)	Predictive model can be applied on the dataset from the Chaophraya Abhaibhubejhr Hospital, Prachin Buri
2020	[2]	Rule-based system Artificial Neural Network Expert system		Compare several methods with same dataset
2019	[3]	Support Vector Machine Naïve Bayes K-Nearest Neighbor Decision Tree	Personalized interview-based survey	Create several other models which predict the exact complications for baby
2019	[4]	Random Forest Xgboost Fusion model of Random Forest and Xgboost	Zhongda Hospital, Jiagnsu Province	Large size dataset is required
2018	[5]	Logistic Regression Support Vector Machine Random Forest	Zhongda Hospital, Jiagnsu Province	
2018	[6]	Neural network Naïve Bayes K-Nearest Neighbor Linear Regression Logistic Regression Support Vector Machine	Haji General Hospital Surabaya	Increase the quality of dataset and improve the learning process

III. PROPOSED SYSTEM MODEL

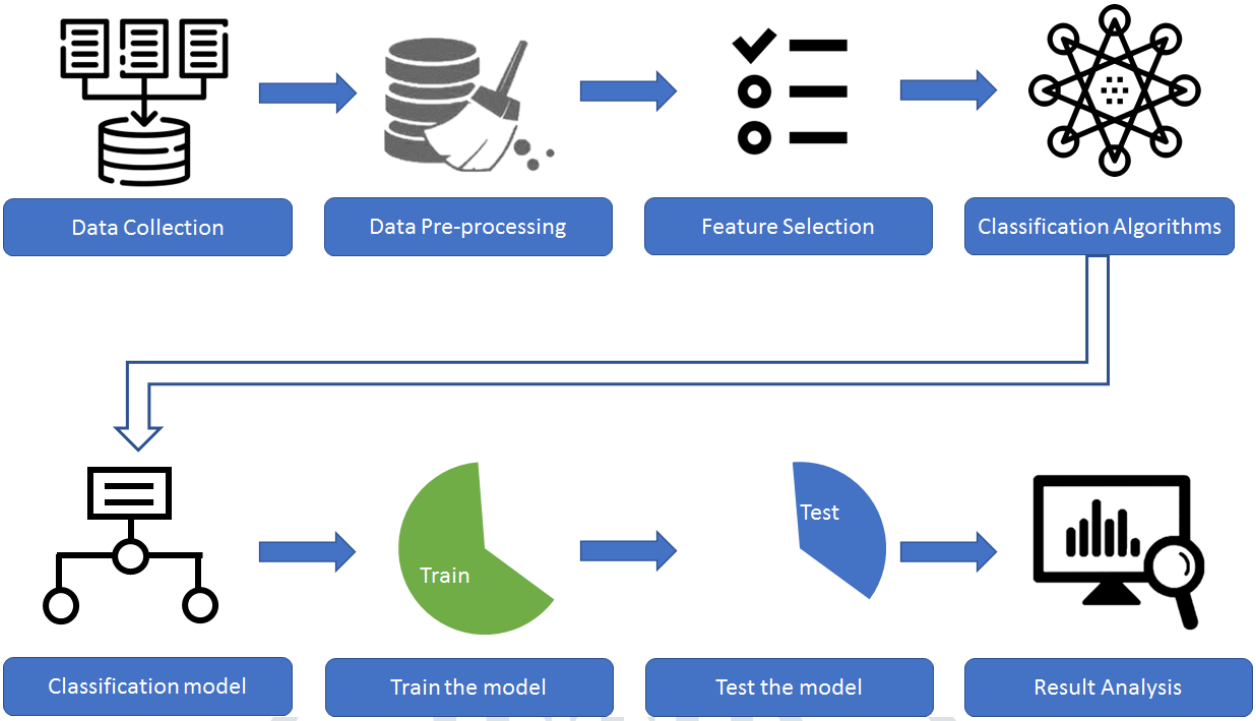

Fig.2 - flowchart of proposed method

The above fig. 2 shows the flowchart of proposed method. In this, first data is collected and then data pre-processing and feature selection will be done on that data. After that, classification algorithms are applied to test and train the model for best accuracy.

Proposed Algorithm:

BEGIN

Step-1. Collect all the data from resources

Step-2. Take input from database

Step-3. Apply data pre-processing methods on dataset

Step-4. Do feature selection on pre-processed data on dataset

Step-5. Use classification algorithms to train classification model

Step-6. Test the model, after training the model

Step-7. Analyze the result

Step-8. Result

End

IV. PROPOSED SYSTEM FLOW

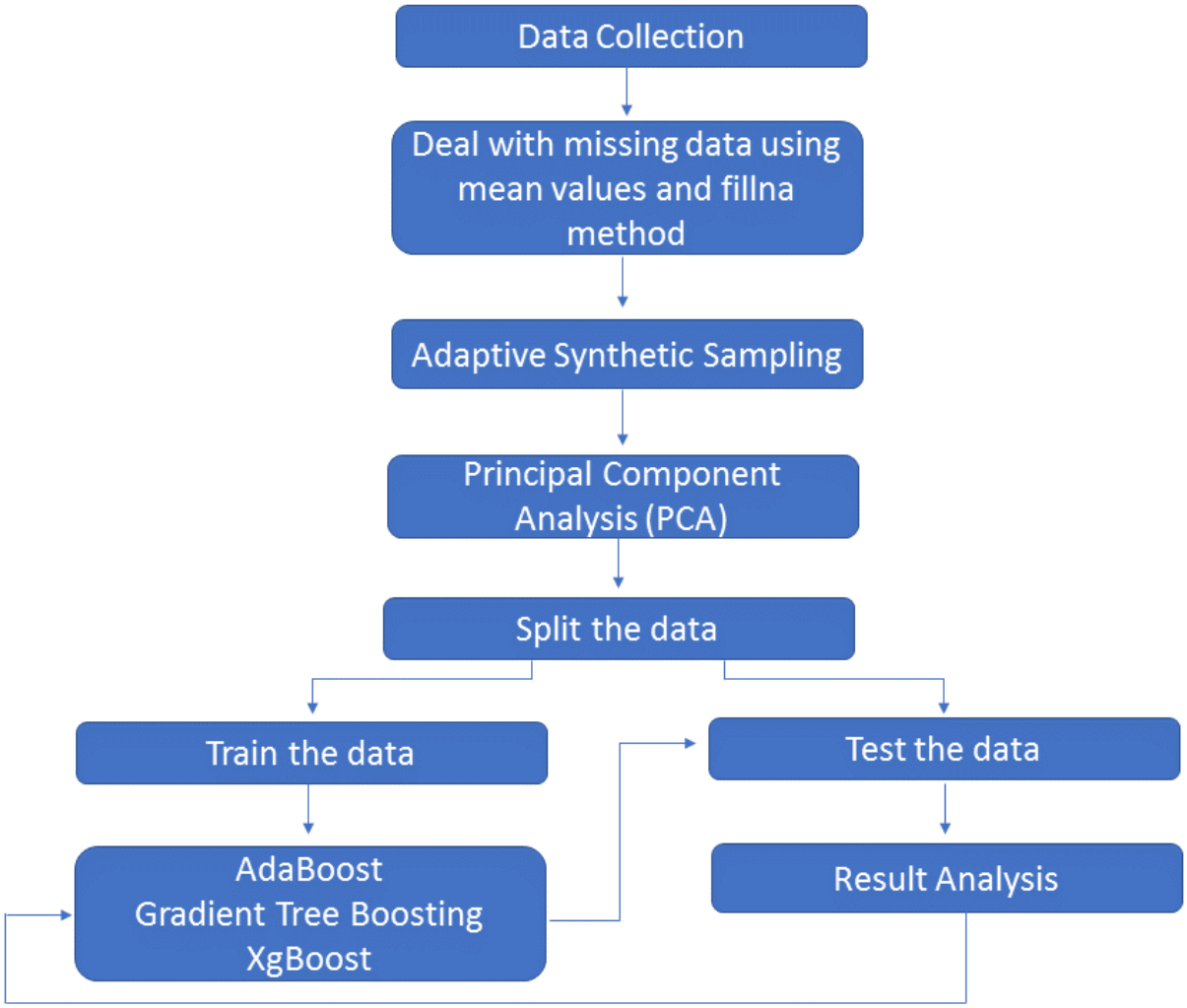

Fig.3 - proposed system flow

Data Collection

The main and basic step for data analysis is to collect data. Nowadays, datasets are easily available on websites, GitHub, etc. in CSV, xlsx form. The dataset used for this research contains 440 samples of data pregnant women [7]. From which 433 women are having normal pregnancy and 7 women are having preeclampsia.

Deal with missing data using mean values and fillna method

Sometimes, it may happen that the downloaded dataset is not 100% data. Some fields, rows, or columns are blank and the machine cannot be learned with missing data. So, there are many techniques to deal with missing data. Here, I have used fillna method to fill missing data with the help of a particular column's mean value. Fig. 4 shows data with missing values and fig. 5 shows data without missing values.

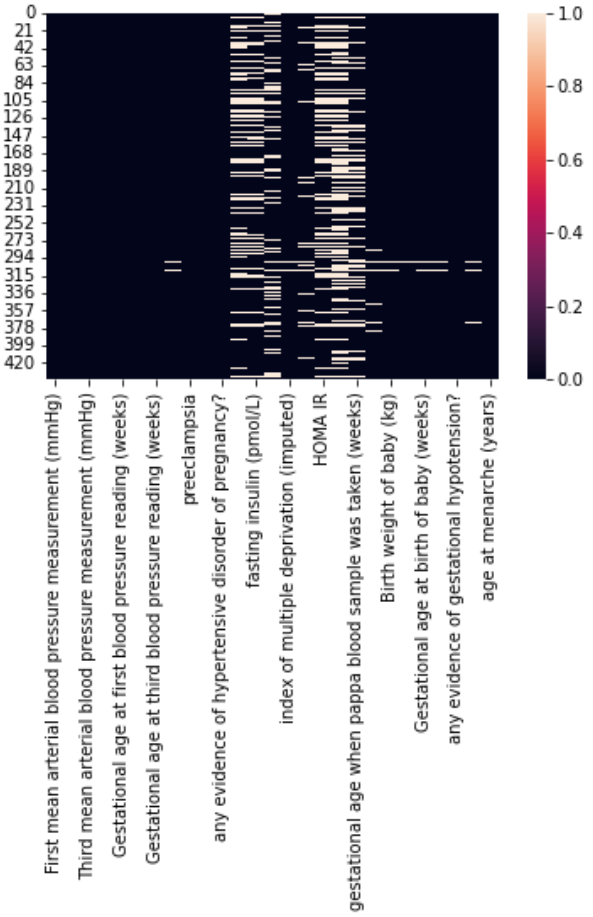

Fig.4 - heatmap with missing data

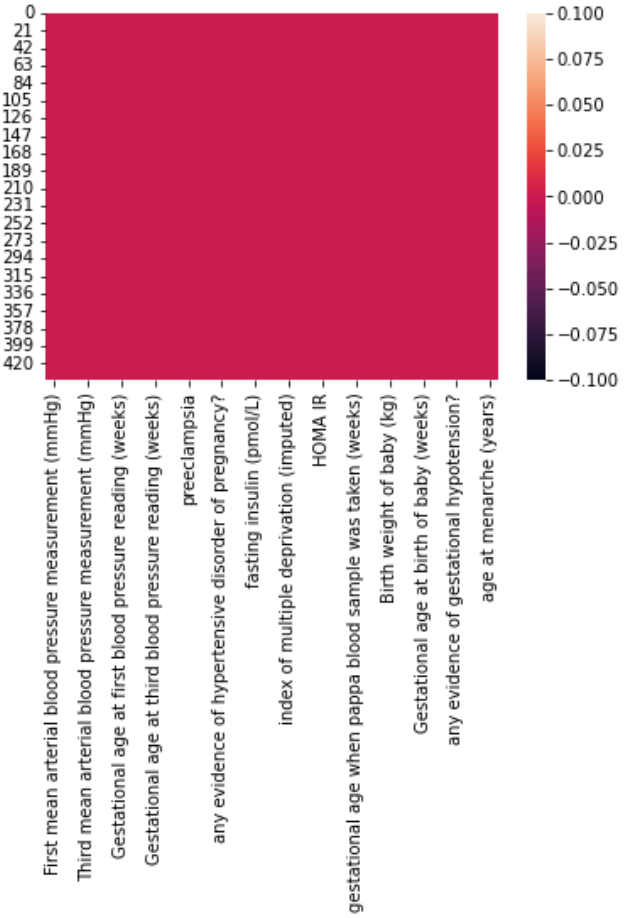

Fig.5 - heatmap without missing data

Adaptive Synthetic Sampling

ADASYN (Adaptive Synthetic) is an algorithm that generates synthetic data. This technique is used to balance the imbalanced data. Imbalanced data means the target column has an uneven distribution of observations and this kind of distribution does not provide accuracy properly. So, to get better accuracy, data must be balanced and observations should be equal to or near to the majority observation of the data. Fig. 6 and 7 show imbalanced and balanced data respectively.

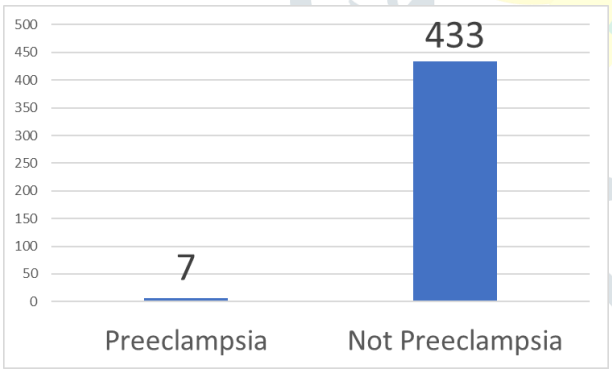

Fig.6 - imbalanced data

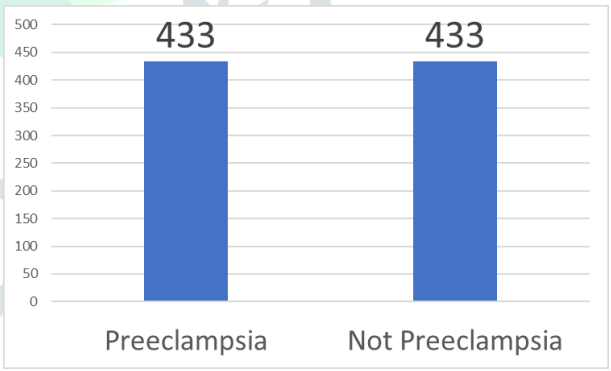

Fig.7 - balanced data

Principal Component Analysis (PCA)

PCA is one of the dimensionality reduction techniques for machine learning. Dimensionality reduction means reducing the number of input variables for the dataset. In the excel sheet, columns represent dimensions. However, the large number of dimensions acquire more space while running machine learning algorithms. To overcome this problem and to improve execution speed, dimensionality reduction techniques are used. Here, I have taken 3 main components using PCA for dimensionality reduction [10].

Split the data

After completion of data pre-processing, data is divided into two parts: to train the data and to test the data. Data size to split data into train and test can be in the ratio of 70:30 where 70% refers to training data and 30% refers to testing data from the dataset. This split ratio can be changed as per the requirement to get better results.

Model Evaluation

The training sample of data is applied to 3 different boosting machine learning algorithms i.e., AdaBoost, Gradient Tree Boosting and XgBoost. Then one evaluation model is made. Now, this model is applied to the sample of test data to check model is working properly or not.

V. RESULT

When evaluation model is applied to the test data, accuracy, precision, recall, F1 score values of Adaboost, Gradient Tree Boosting and XgBoost algorithms can be compared with each other. Here, fig. 8 shows that Gradient Tree Boosting and XgBoost algorithms have same accuracies than AdaBoost. Fig. 9 shows that Precision value of all the algorithms is same. Fig. 10 shows that XgBoost is having highest Recall values than AdaBoost and Gradient Tree Boosting and fig. 11 represents F1 score value, again XgBoost is having higher value than other 2 algorithms.

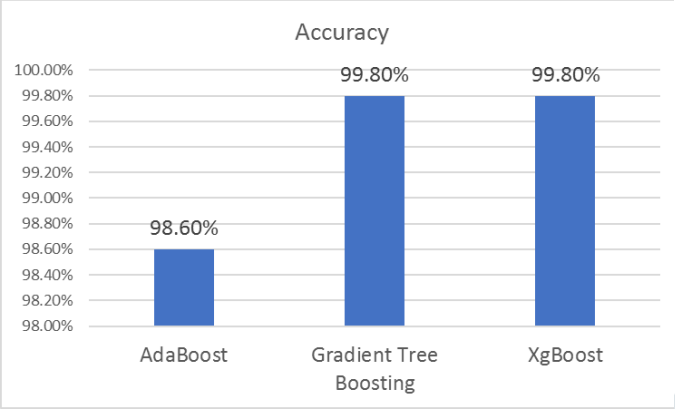

Fig.8 - accuracy

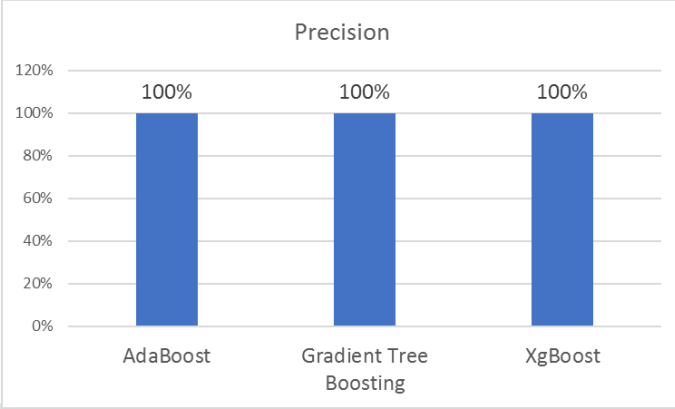

Fig.9 – precision

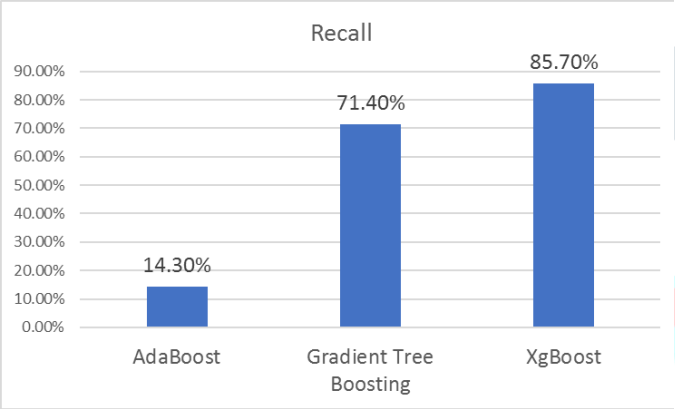

Fig.10 - recall

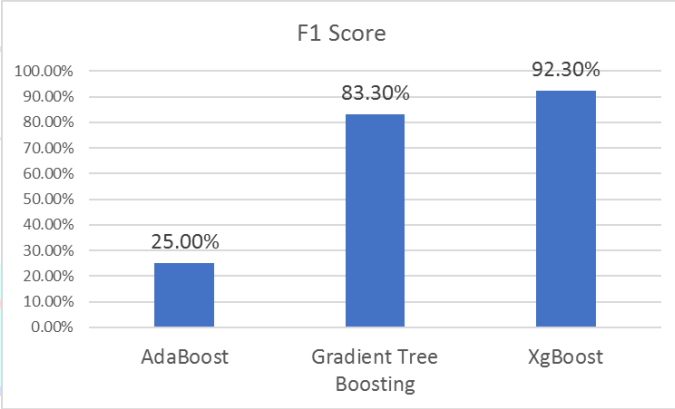

Fig.11 - F1 score

VI. CONCLUSION AND FUTURE WORK

Pregnancy hypertension has become a major concern for maternal and fetal health. It can increase both maternal and fetal mortality ratios if it is not detected and treated at an early stage. Data mining techniques are useful in the medical field for the prediction of diseases. Here, I have applied 3 boosting algorithms Adaboost, Gradient Tree Boosting, and XgBoost whose accuracies are 98.60%, 99.80%, and 99.80% respectively which shows Gradient Tree Boosting and XgBoost are having same accuracies as compared Adaboost.

As we know, Machine Learning algorithms work well with a huge amount of data and provide more and perfect accuracy. My try in the future will be to do an analysis with large size of dataset.

REFERENCES

[1] Sirinat Wanriko, Narit Hnoohom, Konklakom Wongpatikaseree, Anuchit Jipattanakul, Olarik Musigavong, "Risk Assessment of Pregnancy-induced Hypertension Using a Machine Learning Approach" 6th International Conference on Digital Arts, Media and Technology (DAMT) and 4th ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering (NCON), IEEE 2021.

[2] Galih Malela Damaraji, Adhistya Erna Permanasari, Indriana Hidayah, "A Review of Expert System for Indentification Various Risk in Pregnancy", 3rd International Conference on Information and Communications Technology (ICOIACT), IEEE 2020.

[3] V. Madhusri, G. Kesavkrishna, Dr Ramalatha Marimuthu, Sathyanarayanan R, "PERFORMANE COMPARISION OF MACHINE LEARNING ALGORITHMS TO PREDICT LABOR COMPLICATIONS AND BIRTH DEFECT BASED ON STRESS", 10th International Conference on Awareness Science and Technology (iCAST), IEEE 2019.

[4] Wei Wu, Danhong Peng, Tian Xu, Jun Wang, Gongdao Wang, Fengzhen Hou, Xinke Lan, "Classification of hypertension in pregnancy based on random forest and Xgboost fusion model", The Third International Conference on Biological Information and Biomedical Engineering, BIBE 2019.

[5] Xinke Lan, Wei Wu, Danhong Peng, Tian Xu, Jun Wang, Gongdao Wang, Fengzhen Hou, "Research on Pregnancy Hypertension Based on Data Mining", 11th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics, IEEE 2018.

[6] Muhlis Tahir, Tessy Badriyah, Iwan Syarif, "Neural Networks Algorithms to Inquire Previous Preeclampsia Factors in Women with Chronic Hypertension During pregnancy in Childbirth Process", International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), IEEE 2018.

[7] https://www.repository.cam.ac.uk/handle/1810/288081

[8] https://www.youtube.com/watch?v=oxzKg28mjtU
[9] https://www.youtube.com/watch?v=_p0W6g1h_dE
[10] https://machinelearningmastery.com/principal-components-analysis-for-dimensionality-reduction-in-python