PREDICTION OF CAESAREAN SECTION USING CARDIOTOCOGRAPHY

MINI PROJECT REPORT

Submitted by

Varshini S	**210701302**
Yamini H	**210701320**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY:: CHENNAI 600 025

APRIL 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**Prediction of caesarean section using cardiotocography**" is the bonafide work of **"Varshini S (210701302), Yamini H (21070320)"**who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Karthik. V
Assistant Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on ___________________

Internal Examiner **External Examiner**

ABSTRACT

In this study, we propose the utilisation of the Random Forest algorithm within the domain of machine learning to predict caesarean sections. caesarean sections, crucial in various fields ranging from materials science to nuclear physics, often require accurate prediction for efficient design and analysis. Leveraging the versatility and robustness of Random Forest, we aim to develop a predictive model capable of accurately estimating caesarean sections based on relevant input parameters. By harnessing this machine learning technique, we anticipate enhancing predictive capabilities, thereby facilitating advancements in fields reliant on precise caesarean section predictions. This research endeavour to contribute to the optimization of processes and designs across multiple disciplines through the application of advanced computational methodologies.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Karthick V** Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Varshini S-210701302
Yamini H-210701320

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

FHR	Fetal Heart Rate
SVM	Support Vector Machines
EFM	Electronic Fetal Monitoring
CTG	Cardiotocography

CHAPTER 1

INTRODUCTION

1.1 In recent years, the utilisation of machine learning algorithms has surged across various industries, offering powerful tools for predictive modelling and data analysis. One such application lies in predicting caesarean sections, a crucial aspect in fields ranging from materials science to aerospace engineering. The Random Forest algorithm, renowned for its versatility and robustness, emerges as a promising technique for this task. By leveraging ensemble learning and decision trees, Random Forest excels in handling complex datasets and capturing intricate relationships between variables. In this study, we delve into the realm of caesarean section prediction, aiming to harness the predictive capabilities of Random Forest to enhance our understanding of this phenomenon. Through a comprehensive analysis, we seek to uncover valuable insights that can inform decision- making processes and drive advancements in various domains reliant on accurate scission section predictions. Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialised algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labelling to learn data has to be labelled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data.

1.2 PROPOSED SYSTEM

In response to the need for accurate prediction of scissor lift sections, a proposed system utilising the Random Forest algorithm within machine learning techniques emerges as a robust solution. This system aims to revolutionise the precision and efficiency of caesarean section predictions by leveraging the power of ensemble learning offered by Random Forest. By employing a multitude of decision trees, Random Forest effectively captures the intricate relationships between various parameters influencing scissor lift sections, such as material properties, dimensions, and load capacities. Through extensive training on a comprehensive dataset comprising diverse caesarean lift designs and configurations, the proposed system can learn complex patterns and correlations, enabling it to make highly accurate predictions. Moreover, the versatility of Random Forest allows for the inclusion of both numerical and categorical features, facilitating a comprehensive analysis of caesarean lift sections across different contexts. The proposed system holds the potential to streamline the design and optimisation processes of caesarean lift sections, contributing to enhanced safety, efficiency, and performance in diverse industrial applications.

CHAPTER 2
LITERATURE SURVEY

The challenges of childbirth and maternal mortality are significant problems during childbearing. For instance, research proposes a model that implements feature selection to select relevant features and can provide improved performance predictions. The proposed feature selection techniques are Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Sequential Forward Floating Selection (SFFS), Sequential Backward Floating Selection (SBFS), and SelectKBest. A decision support system named Henry gas solubility optimization (HGSO)-based random forest (RF), with an improved objective function, called HGSORF, for the classification of CS and non-CS classes has been developed as a potential application to support clinical staffs . The proposed ROSE-PCA-RF model is a novel CDSS for caesarean section (CS) prediction using electronic health records of pregnant women. It uses the random oversampling example (ROSE) technique, principal component analysis for feature extraction, and a random forest (RF) model for classification, obtained 96.29% accuracy on training data, and improved the accuracy of 97.12% on testing data. 10 The classification algorithm used to classify is Naive Bayes with accuracy 65%. Authors have employed Gradient Boosting, KNN, XGBoost, Voting, Stacking, Gradient Boosting, LGBM, Random Forest, SVM and Logistic Regression algorithms with varying C-statistics values ranging from 0.59 to 0.7. Among these, logistic regression models demonstrated the best prediction performance. In another study, authors use two ML algorithms and different data balancing techniques to predict the C section form secondary dataset. They use LR and MLP as ML model in their study. Using oversampling data balancing technique SMOTE they get 93% accuracy, precision, recall and f1 score in LR and 95% accuracy, precision, recall and f1 score in MLP. They only focus the prediction of the C section, not the models inner story that how it predict the output. For the purpose of predicting birth mode in the context of Bangladesh, authors have developed a bagging ensemble strategy based on SVM, DT, KNN, and NB. According to the findings, bagging ensemble models outperformed conventional models. The study also determined the connection between significant factors and the prevalence of caesarean sections. A performance examination of classification methods using the birth dataset was carried out by Abbas et al. A caret package in R was used to complete this task. In this study, seven different methodologies are applied. With 0.94 and 0.95 precision, respectively, Adabag and BagFda fared better in their study. Three distinct ensemble prediction models were applied to the prediction of caesarean birth. Among them, XGBoost had the best predictive performance, coming in at 88.91%. According to the study, the most important factors include a medical indication, amniotic fluid, foetal intrapartum ph, number of prior caesareans, and pre-induction. A model was created by Carlos et al. to improve the Discriminatory Accuracy (DA) of emergency cesarean procedures. In the study, logistic regression and random forest models were employed. The overall model's LR+ was of a modest to moderate magnitude. The DA was in the range of 0.74 to.81. The highest among them is the DA General of the RF model [13]. Finding the optimal characteristics for predicting birthing modes is the goal of this study. In another study [14] authors use CTU-UHB dataset to classify the C section and perform a comparison among different classifiers. They maximum 87% AUC in RF out 11 of the seven machine learning algorithms they tested.

SUMMARY OF LITERATURE SURVEY

Year	Methods	Result
2022	HGSORF, Gaussian Naive Bayes (GNB). linear discriminant analysis (LDA), K-nearest neighbors (KNN), gradient	HGSORF achieved superior performance with an accuracy of 98.33%.
2022	Proposed ROSE-PCA-RF model	96.29% accuracy on training data and improved the accuracy of 97.12% on testing data
2022	Sequential Forward/Backward and	Maximum 65% accuracy

	Floating Selection, and Select KBest . Naive Bayes classifier.	
2022	KNN, Voting, XGBoost, Stacking, Gradient Boosting, Random Forest, LGBM, Logistic Regression & SVM	Accuracy 0.68, Specificity 0.83. Sensitivity 0.41.
2022	LR and MLP	Accuracy, precision, recall, and fl score is 0.95
2021	DT(Bagging), KNN(Bagging), NB(Bagging), SVM(Bagging), DTBagging	DTBagging accuracy 0.87
2020	KNN, NB, ANN, SVM, DT, XGB, BagFda, AdaBag BagFda	BagFda shows 93.44% accuracy

CHAPTER 3
SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

PROCESSOR	Intel Core i3
RAM	Minimum 2GB
HARD DISK	80GB

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre-installed and the languages needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

OPERATING SYSTEM	Windows 10 or later
TOOL	Anaconda with Jupiter Notebook

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

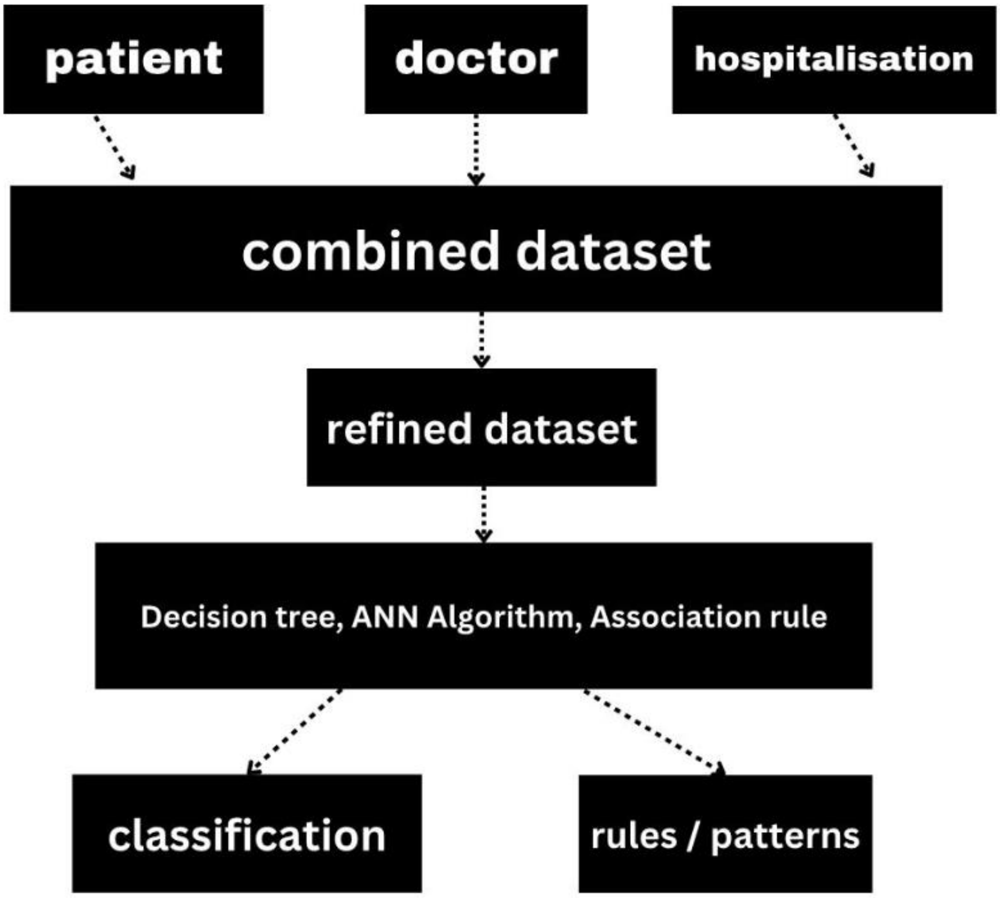


Fig 3.2.1 Architecture Diagram

.

CHAPTER 4
PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 DATA PRE-PROCESSING:

Data pre-processing is a crucial step in any data analysis or machine learning project. It involves cleaning and transforming raw data into a format that is suitable for analysis or model training. Common tasks include handling missing values, removing duplicates, scaling features, and encoding categorical variables

4.1.2 DATA ANALYSIS AND VISUALIZATION:

Data analysis involves exploring and summarising data to extract meaningful insights. Visualization plays a key role in understanding patterns and relationships within the data. Techniques such as charts, graphs, and plots are used to present data visually, making it easier to comprehend and interpret..

4.1.3 RANDOM FOREST:

Random Forest is an ensemble learning technique used in machine learning for both classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs the mode (for classification) or mean prediction (for regression) of the individual trees.

Here's a detailed description of Random Forest:

Ensemble Learning: Combining Multiple Models: Random Forest belongs to the ensemble learning family, where multiple models are combined to improve overall performance and accuracy.

Decision Trees: Base Model: The fundamental building block of a Random Forest is a decision tree. Each tree is constructed using a subset of the training data and a subset of the features. Predictive Decision Rules:

Decision trees make decisions by asking a series of questions based on features and constructing a tree-like structure of decisions.

4.1.4 RANDOMIZATION:

Subset of Features: For each decision tree, a random subset of features is considered at each split. This introduces diversity among the trees. - Bootstrap Aggregating (Bagging): The training data for each tree is created by randomly sampling, with replacement, from the original training dataset. This creates different datasets for each tree.

4.1.5 TRAINING PROCESS:

Create Subsets: Random subsets of the training data are created using both random sampling of instances (with replacement) and random subsets of features. Build Trees: Decision trees are built using the created subsets. Aggregate Predictions: For classification tasks, the final prediction is determined by a majority vote (mode) among the predictions of individual trees. For regression tasks, it's the average of individual tree predictions. In summary, Random Forest is a powerful and versatile machine learning algorithm that leverages the strength of multiple decision trees to make accurate and robust predictions in various types of tasks.

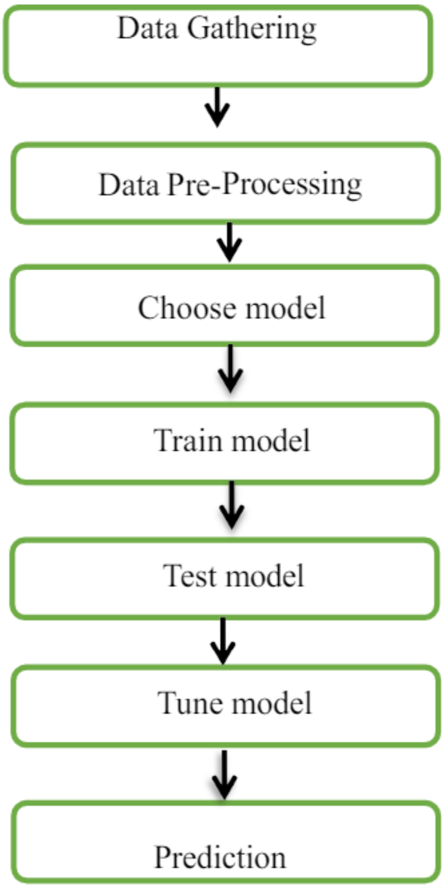
4.1.6 DEPLOYMENT:

Deployment involves making your machine learning model accessible for use in a realworld setting. This can include integrating the model into a web application, creating APIs, or deploying it on cloud platforms. It's crucial to consider scalability, performance, and monitoring when deploying machine learning models. For a comprehensive understanding of each topic, further study and practical application are recommended.

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION



It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualise the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection.

You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalise. A way to do this is to use different visualisation methods to show the average accuracy, variance and other properties of the distribution of model accuracies. In the next section you will 25 discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

Performance Metrics to calculate:

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive. False **Negatives (FN):** A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing. True **Negatives (TN):** A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

True Positive Rate (TPR) = TP / (TP + FN)

False Positive rate (FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation: Accuracy = (TP + TN) / (TP + TN + FP + FN) Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same

Precision: The proportion of positive predictions that are actually correct. **Precision = TP / (TP + FP)** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labelled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict) **Recall = TP / (TP + FN)** Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

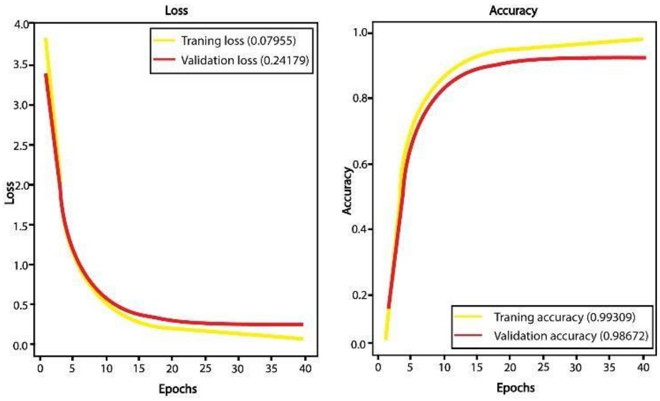
F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula: F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula: F1 Score = 2*(Recall * Precision) / (Recall + Precision)

5.2 OUTPUT SCREENSHOTS

Table 5.2.1 Accuracy Curve



(a) Loss curve

(b) Accuracy curve

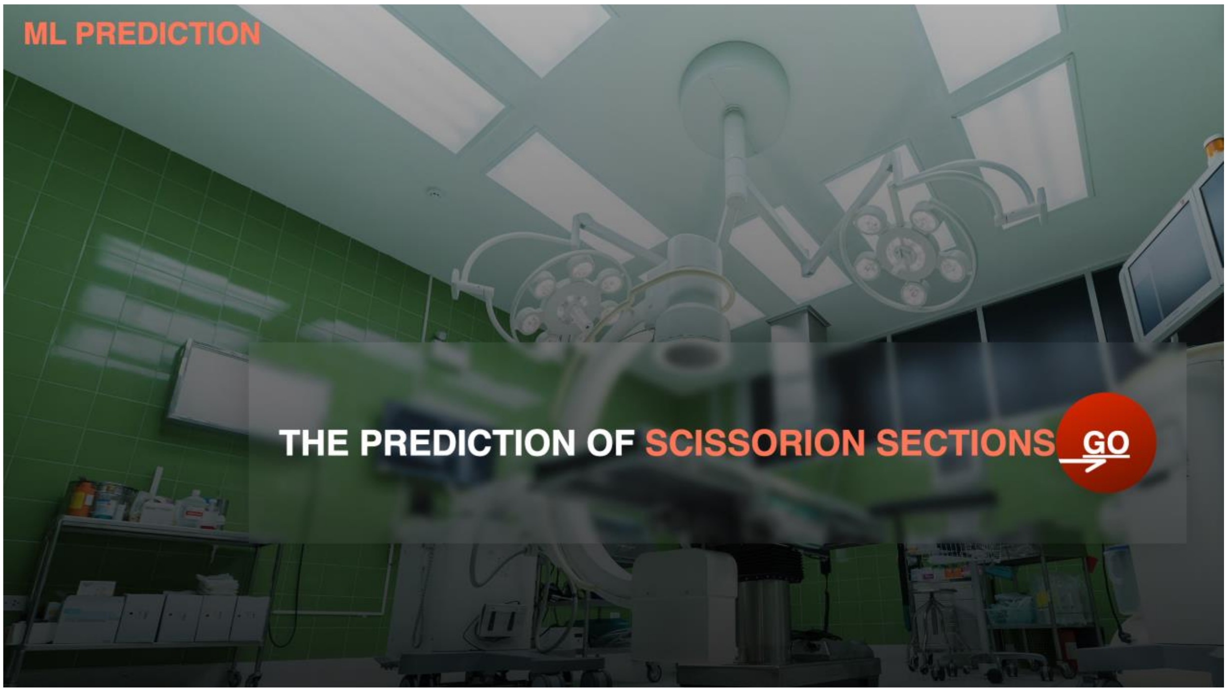
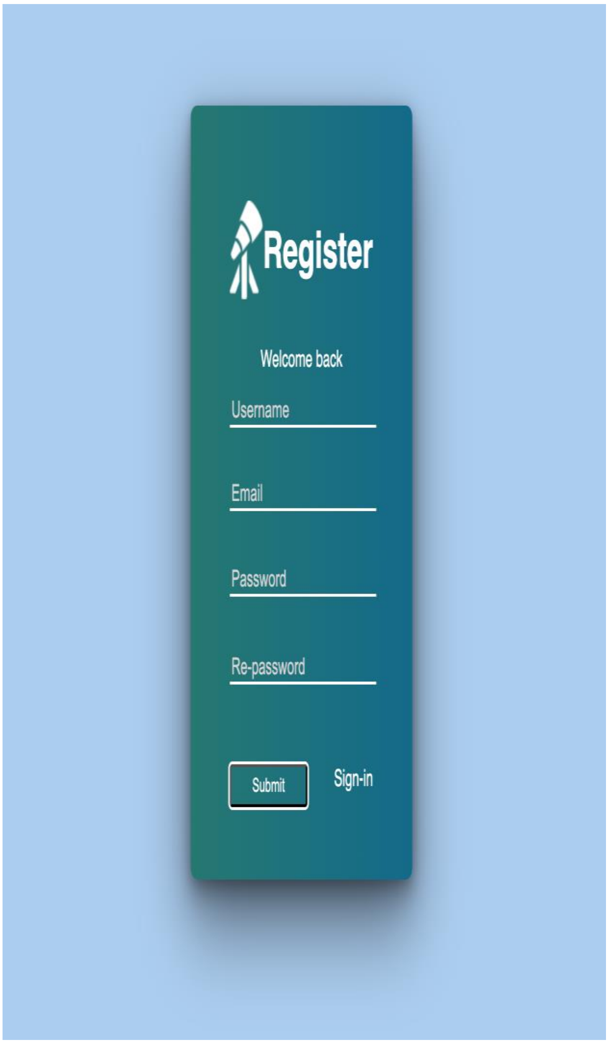
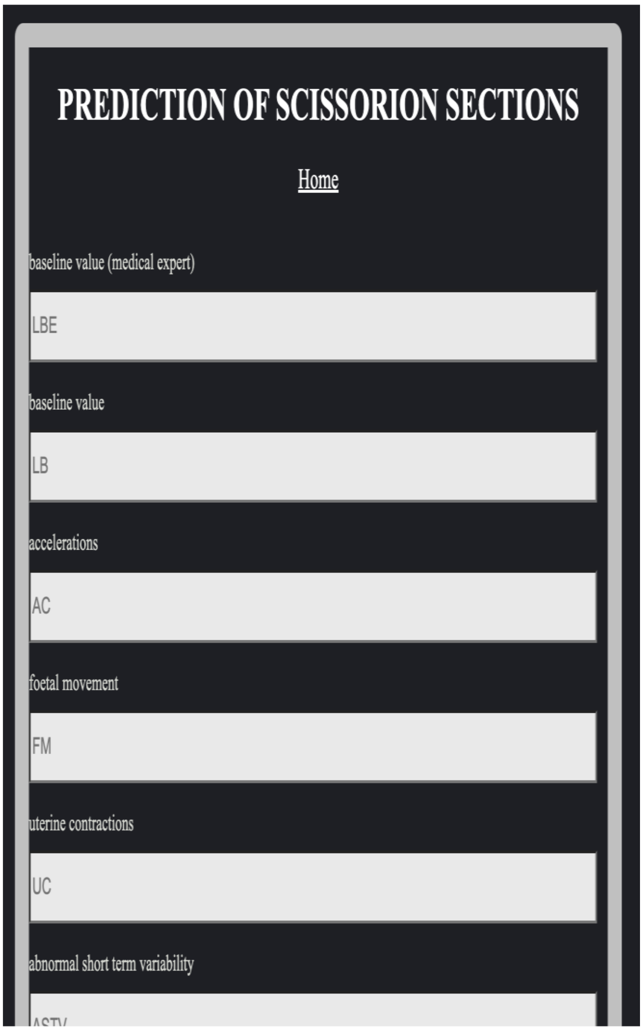


Fig 5.1 HOME PAGE




PREDICTION OF SCISSORION SECTIONS

Home

baseline value (medical expert)

LBE

baseline value

LB

accelerations

AC

foetal movement

FM

uterine contractions

UC

abnormal short term variability

ASTV

Fig 5.2 INPUT PAGE

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

In conclusion, the application of the Random Forest algorithm in predicting caesarean sections demonstrates its efficacy in addressing complex classification tasks within the realm of machine learning. Through the utilization of ensemble learning and decision tree- based models, Random Forest exhibits robustness in handling high-dimensional data and mitigating over fitting. The accuracy and efficiency of this approach in predicting caesarean sections pave the way for its potential utilization in various other fields, ranging from medical diagnostics to industrial quality control. Furthermore, by leveraging the insights derived from this study, future research endeavours can explore further refinements and adaptations of the Random Forest algorithm, ultimately contributing to advancements in both theoretical understanding and practical applications of machine learning techniques.

6.2 FUTURE ENHANCEMENTS

In future work, further enhancement and exploration of the Random Forest algorithm for predicting caesarean sections could be undertaken. This could involve several avenues of research and development:

1. **Feature Engineering:** Investigate and experiment with additional features that could potentially improve the predictive performance of the model. This might involve extracting more nuanced information from the caesarean sections data or incorporating external datasets that could provide complementary information.
2. **Algorithm Optimisation:** Explore techniques to optimise the Random Forest algorithm parameters such as the number of trees, maximum depth of trees, and minimum number of samples required to split a node. Fine-tuning these parameters could lead to improved model performance and generalisation ability.
3. **Ensemble Methods:** Investigate the effectiveness of ensemble methods that combine predictions from multiple Random Forest models. Techniques such as bagging, boosting, or stacking could be explored to leverage the strengths of different models and further enhance prediction accuracy.

REFERENCES

[1] T. Desyani, A. Saifudin, and Y. Yulianti, ‘‘Feature selection based on naive Bayes for caesarean section prediction,’’ IOP Conf. Ser., Mater. Sci. Eng., vol. 879, no. 1, Jul. 2020, Art. no. 012091.

[2] S. A. Abbas, A. U. Rehman, F. Majeed, A. Majid, M. S. A. Malik, Z. H. Kazmi, and S. Zafar, ‘‘Performance analysis of classification algorithms on birth dataset,’’ IEEE Access, vol. 8, pp. 102146–102154, 2020.

[3] N. I. Khan, T. Mahmud, M. N. Islam, and S. N. Mustafina, ‘‘Prediction of cesarean childbirth using ensemble machine learning methods,’’ in Proc. 22nd Int. Conf. Inf. Integr. Web-Based Appl. Services, Nov. 2020, pp. 331–339.

[4] D. M. W. Powers, ‘‘Evaluation: From precision, recall and Fmeasure to ROC, informedness, markedness and correlation,’’ 2020, arXiv:2010.16061

[5] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, ‘‘Data imbalance in classification: Experimental evaluation,’’ Inf. Sci., vol. 513, pp. 429–441, Mar.2020.

[6] L. N. Mahdy, K. A. Ezzat, H. H. Elmousalami, H. A. Ella, and A. E. Hassanien, ‘‘Automatic X-ray COVID-19 lung image classification system based on multi-level thresholding and support vector machine,’’ medRxiv, 2020.

[7] D. Chicco and G. Jurman, ‘‘The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation,’’ BMC Genomics, vol. 21, no. 1, pp. 1–13, Dec. 2020.

[8] D. Dave, H. Naik, S. Singhal, and P. Patel, ‘‘Explainable AI meets healthcare: A study on heart disease dataset,’’ 2020, arXiv:2011.03195

[9] M. S. Bin Alam, M. J. A. Patwary, and M. Hassan, ‘‘Birth mode prediction using bagging ensemble classifier: A case study of Bangladesh,’’ in Proc. Int. Conf. Inf. Commun. Technol. Sustain. Develop. (ICICT4SD), Feb. 2021, pp. 95–99

[10] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, ‘‘A survey on missing data in machine learning,’’ J. Big Data, vol. 8, no. 1, pp. 1–37, Oct. 2021

[11] M. S. Bin Alam, M. J. A. Patwary, and M. Hassan, ‘‘Birth mode prediction using bagging ensemble classifier: A case study of Bangladesh,’’ in Proc. Int. Conf. Inf. Commun. Technol. Sustain. Develop. (ICICT4SD), Feb. 2021, pp. 95–99

[12] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, ‘‘A survey on missing data in machine learning,’’ J. Big Data, vol. 8, no. 1, pp. 1–37, Oct. 2021