RESEARCH ARTICLE OPEN ACCESS

Data Mining in Health Centres with Tree Classifiers A Study

R.Shobana*, K.shanthashalini*,B.Madhumitha**, S.Sangeetha**, C.S.Pavithra**

*(Faculty, Department of CSE, AVIT, Chennai
Email: shobana.cse@avit.ac.in)
**(Student,Department of CSE, AVIT, Chennai)

--************************----------------------------------

Abstract:

This paper describes a decision making systembythe way of data mining to help doctors in decision making on a Caesarean Section during the healthcare. The dataset is cleaned and data mining is performed to ensure the robust data. This is life saving approach which could be used in labour ward. Higher rates of mother and baby survival is more in Caesarean section and is preferred across the globe.The proposed system would predict whether the surgery is needed. In this elaborate review,three modes of data mining such as ID3,Random Forest and CART are studied and out of these. Random Forest was clinched with an accuracy of 96.4% and less penalty cost.

Keywords **—Healthcare Decision Making, Decision Trees, Data Mining,Random Forest,CART**

--************************----------------------------------

I. INTRODUCTION

The present day healthcare needs effective methods and research methodologies to have successful deliveries and also reduce the cost of the healthcare services. With Ma-chine Learning emerging as the future booming technology, it is incorporated in healthcare for predicting the need for a Caesarean Section surgery [1]. Many techniques are used to perform child delivery in a Labour Ward. The delivery procedures are most commonly addressed as vaginal deliveries and caesarean sections. During complicated vaginal deliveries, vacuum extraction and obstetric forceps might be used[2]. When natural vaginal delivery and operative procedures are not feasible or risky for the patient, Caesarean Section is performed. An assertive decision needs to be taken as to whether or not a particular birth requires assis-tance,whether it would exacerbate the patients conditions. Predicting the appropriate delivery mode in advance would help in identifying which pregnant women actually need operative procedures or caesarean section. This way,it re-duces the proportion of unnecessarily assisted procedures which are used liberally with marginal medical benefit. An-other benefit of this system would be to avoid malpractice and negligence of doctors during labour, leading to better quality in labour ward [3]. The purpose of this study is to apprehend Data Mining in real-time applications to predict the type of delivery more compatible with the pregnancy characteristics of each patient. This article includes five sections. Following the introduction, the second section presents the review of literature. In section three, a brief description of the dataset is given. Section four describes the proposed models used. Finally, the results along with conclusions and future work comprise the last section [4].

II. DATAMINING

Data Mining(DM) refers to the process of extracting data, analyzing it from many dimensions or perspectives, then producing a summary of the information in a useful form that identifies relationships within the data. It is a sub-field of Computer Science which blends many techniques from Mathematics, Statistics, Data Science, Database and Ma-chine Learning. Using various models, after preprocessing and re-sampling, useful information is gained [5]. Knowledge Discovery in Databases(KDD) is considered as a programmed, exploratory analysis and modelling of vast data repositories. Data Mining is the root of the KDD procedure, including the inferring of algorithms that investigate the data, develop the model, and find previously unknown patterns. The model is used for extracting the knowledge from the data, analyze the data, and predict the data[6].

The different steps involved in performing Data Mining over a given dataset is given in figure 1 and Steps involved are :

Selection: In this step, application domain is explored for indepth insights. The end-user requirements are alsounderstood.

Target dataset creation: A dataset is selected or a subset of variables are chosen. Discovery is performed on thisdataset.

Data Cleaning/Preprocessing: It is a vital step of DM where 60% of the time taken for building the model is spent on preprocessing by the researchers and data scientists. Pre-processing techniques such as missing value replacement (imputation), outlier removal, scaling and feature selection are carried out to enhance the overall performance of the model. Various visualization techniques are used to make it user-friendly.

Dimensionality reduction: Dimensionality reduction is used to reduce the effective number of dimensions for better interpretation of input features and visualisation .This reduces memory complexity and time complexity.

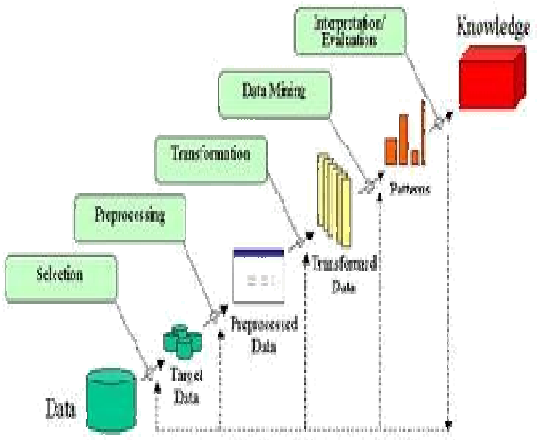

Fig. 1. KDD Process

Data Mining Algorithm: A suitable Data Mining Algo-rithm is chosen based on the goal of KDD process and on the application domain.

Data Mining: Patterns of a particular representation are found using the selected Algorithm. Characterization is used to decide the purpose of the model.

Interpretation of mined data: The inference is ob-tained on examining the patterns of mineddata.

Consolidation: The discovered knowledge is consol-idated and its impact is measured and reports are generated[7].

Data Mining Analysis is best suited for large volumes of data with maximum usage of data to arrive at reliable conclusions. When knowledge is acquired from a limited dataset, it can be expanded to bigger datasets assuming larger datasets to have a similar structure as the simple dataset.

A.Applications of DataMining

Data can be of types such as text, image, maps etc. Data Mining can be applied to any form of data. Some applications are illustrated below:

† Text Mining : Information extraction and text catego-rization using Natural Language Processing can make text documents structured. Tokenization is performed for better understanding of the document[8].

† Web Mining : Information is automatically discovered and extracted from web documents. Resource Finding and generalization is done to discover general patterns on websites [9].

† Medical Data Mining : This method is used for studying a patients vital signs to understand his illness and predict the future by analysing them. Medical dataset consists of mostly images, so image processing is used to detect anomalies in the image[10].

† Financial Data Mining : This technique is done by extracting hidden patterns and predicting future trends in financial markets. In many cases, neural networks are used for portfolio management, money laundering loan default prediction [11].

† Spatial Data Mining : This is a branch of data mining dealing with spatial or location data used for remote sensing and digital mapping. Here, the data is more complicated than classical data mining as they include objects like lines, points and polygons [12].

III. REVIEW OF LITERATURE

Data Mining has varied applications in domains includ-ing finance, cyber security and education. In an article titled Loan Approval Prediction based on Machine Learning Approach by K Arun et al. , Data Mining is applied on big data using 6 models including Decision Trees and Random Forests. It is proven to be a highly efficient component for accurate loan approval predictions[13]. In another article titled Credit Card fraud detection using machine learning as data mining technique by OS Yee et al, Bayesian Classifiers have given more accurate results when fed with filtered data [14]. In [15], movie reviews are analysed using Naive Bayes and K-NN Classifier. Naive Bayes gives higher accuracy than K-NN classifier. V.Mhetre et al. Have classified learners as slow, fast and average by applying Random Tree and Naive Bayes using Weka tool in the paper titled Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA [16]. In [17], Random Forest has been applied for landslide susceptibility mapping.

Data mining has a lot of use cases in the Medical Domain as well. Some examples are cited below:

† Cancer: In the article titled A Novel Approach for Breast Cancer Detection Using Data Mining Techniques, writ-ten by V Chaurasia et al. various data mining techniques have been used like Sequential Minimal Optimization and Bloom Filter Trees to obtain about 96% accuracy in Breast Cancer detection [18]. D Chauhan et al. have cited an efficient approach for lung cancer detection using PCA and LDA in their article titled An efficient data mining classification approach for detecting lung cancer disease. Feature Extraction is performed and system is made more user friendly[19].

† Stroke: M. Sheetal Singh et al. have predicted stroke occurence using decision trees and neural networks to get a very high accuracy [20]. A hybrid data pre-processing method is applied in [21] where boosting algorithms and decision tree are used to obtain the required accuracy.

† Diabetes Mellitus Disease: Das H. et al [22] performed classification and clustering algorithms to get the re-quired accuracy. Perveen et al. have used boosting algorithms and J48 decision tree to improve performance analysis of Diabetes prediction in their article titled Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Adaboost is found to perform better than other ensemble classifiers used [23].

† Alzheimers Disease: Singh M has used Random Tree and CART to diagnose Alzheimer Disease in early stages in [24]. Decision Trees and Generalized Linear Model are among the six machine learning techniques used by Shahbaz M. et al. with 88% accuracy in[25].

† Cataract: Random Forest Algorithm is used extensively by Nair M.S. et al[26] to build a model to predict Cataract occurrence. Zhang, K. et al. have used Ran-dom Forest and Naive Bayes Algorithms for complica-tion prediction. About 65% accuracy is obtained using both of them[27].

† Caesarean Section: In [1] various classification models such as KNN, Naive Bayes, SVM and Logistic Regres-sion were applied using WEKA tool and all accuracies were cited. Among the cited models,KNN and Random Forest produced an accuracy of about 95% of which KNN is a statistical classifier and RF is a Decision Tree classifier. Also, the focus of [1] was only on accuracy and other measures like Recall, Type II error(since it is a medical dataset) were not considered. In [2], among all the classifiers employed, J48 Decision Tree classifier yielded a maximum accuracy of around 65%.

The proposed work aims to make use of only the applica-tions of Decision Tree classifiers on the Caesarean Dataset. Among the decision tree classifiers,ID3,CART and RF are chosen for this study.

IV. DATASET DESCRIPTION

This dataset is taken from UCI Machine Learning repository which is wildly used by researchers to carry out their tasks related to the chosen domain. There are 507 datasets in UCI. Under medical domain, 116 datasets are available for exploration. In this study, the dataset named as Caesarean Section Classification Dataset has been explored containing pregnancy details of 80 women with 5 input features and 1 output feature. The snapshot of the dataset is shown in the following table I: The attributes of dataset mentioned in table I is detailed below:

1)Age : This parameter refers to the ages of women admitted in the labour ward. It has a range from17-40

2)Delivery Number : This parameter refers to the number of deliveries undergone by the patient. It ranges from1-4

3) Delivery Time : This feature refers to the time of the delivery,i.e. whether it is premature or on time. It ranges from 0-2.

TABLE I
SNAPSHOT OF CAESAREANDATASET

Age	Delivery Number	DeliveryTime	Blood Pressue	HeartProblem	Caesarean Section
22	1	0	2	0	0
26	2	0	1	0	1
26	2	1	1	0	0
28	1	0	2	0	0
22	2	0	1	0	1

4)Blood Pressure : This characteristic determines whether the blood pressure of patient is high or not. It ranges from0-2

5)Heart Problem : This attribute determines whether the person has heart problem or not. It has values 0 or1.

6) Caesarean Section : This output feature determines whether a Caesarean section is going to take place or not. It has values 0 No Caesarean, 1- Caesarean[1].

V.PROPOSED MODEL

The primary objective of this research work is to increase the recall [one of the performance metrics of Data Mining]and to keep the type II error(false negative count) as minimum as possible since this is a medical dataset. The reason being when actually surgeries are required but model predicts as no surgery, it results in life danger. Therefore,the focus of this article is to maintain the false negative count(Type II Error) at a reduced level and a penalty cost of say, Rs. 10000 for each count is imposed as a punishment on the model. Having examined the dataset, we found there are no missing values or outliers etc. RF as a feature selector has been used and based on trial and error decided on the threshold value as 0.12(feature selector importance).three features and the performance of the model drastically reduced than with all the input features can be obtained.Therefore all the input features(5) are significant.

Majority of the attributes are categorical except one. It has been proposed to build models to predict Caesarean Operation using decision tree classifiers such as CART,ID3 and Random Forest.The proposed models are shown in the following figures: The flowchart of this article is given in figure 2.

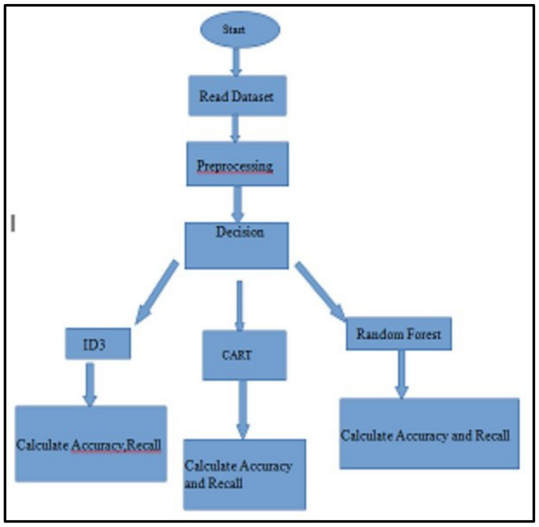
Start
Read Dataset
Preprocessing
Decision
ID3
CART
Random Forest
Calculate Accuracy,Recall
Calculate Accuracy and Recall
Calculate Accuracy and Recall

Fig 2 process flow of the proposed model

The dataset is split into training set(65%) and test set (35%) in this work. The models used in this work are explained in the following subsections:

A.Random ForestAlgorithm(RF)

Random Forest is a supervised classification algorithm. It is an extension of Bagging . Bagging (Bootstrap Aggregation) is an ensemble method which is used to reduce variance of a decision tree. It is a type of decision tree where finding root node and splitting of feature nodes are performed randomly. It is preferred over classifiers because of its capability to be used for regression and classification tasks. Over-fitting does not happen if there are enough trees in the forest. This classifier has the capability to handle null values and categorical variables.[28] The framework of this algorithm is given in figure 3.

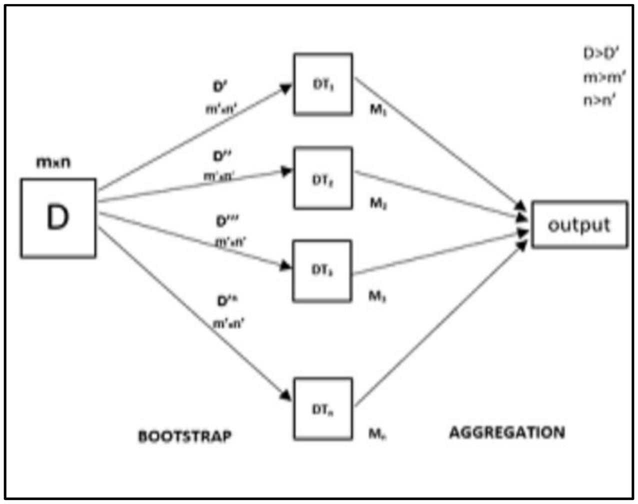

Fig 3 Process of Random Forest Method [28]

Step 1: Random selection of k features from m features.

Step 2: Anode d is computed among k features using best split method

Step 3: Node d is split into daughter nodes using best split

Step 4: The steps 1 to 3 are repeated till l nodes are obtained

Step 5:Steps 1 to 4 are repeated n times to create a forest with n trees

Step 6: Test features are taken and each tree predicts an outcome

Step 7: Votes are given for every outcome and the one with thehighest number of votes is the final prediction of Random Forest Algorithm[29]

B.Decision TreeAlgorithm(ID3)

Decision Tree is a supervised learning algorithm which uses a tree representation for solving problems. Any discrete attribute can be represented this way. Lower entropy or ran-domness is obtained after data is split. There is an information gain after splitting of data. It has the ability to predict class of target variable by learning decision rules inferred from prior data. Implementing a decision tree is relatively easier than other classification algorithms. Interpretation is easier to visualise when compared to other implementations[30].

Working of Decision Trees:

[1]Best attribute of the dataset is selected as root of the tree.

[2]Datasetissplitintotrainingandtestingwithsame valuesin both for an attribute.

Steps1 and 2 are repeated till leaf nodes are found for all branches of the tree.

C.Classification and Regression Trees(CART)

CART is a type of decision tree where the algorithm is structured as questions and the nodes form the answers. This results in a tree like structure. It is non parametric and does not rely on data from a particular distribution. There is more accurate fitting of data as it incorporates testing with test data and cross validation also. This can remove interdependencies among variables. Same variables are used repeatedly in different parts of the tree.

Classification Tree: This tree is used when dataset needs to be split into classes belonging to response variables. It is mostly of the form YES or NO.

Working of CART:

Step 1: Selection of input variables andsplitting points on those variables till a tree is constructed.

Step 2: Selection of which input variable to use and specific split is given by Greedy algorithm.

Step 3: The tree is constructed until a stopping criteron is obtained.[33]

D.PERFORMANCE METRICS:

Though there are various performance metrics such as accuracy,sensitivity,specificity,recall,f-measure,kappa static and AUC, this study mainly focuses on the following met-rics. They are outlined below:

ConfusionMatrix:

It is a performance measurement technique used in Machine Learning for knowing the correctness of an algorithm. A table like format has 4 values computed using Actual and predicted values.

Description of Confusion Matrix:

Actual Values and Predicted Values: 00 represents True Negative , 01 represents False Negative, 10 represents False Positive, 11 represents TruePositive.

† True Negative: When the actual and predicted values are negative.

† True Positive: When the actual and predicted values are positive.

† False Negative: When the actual value is Positive but the predicted value is negative. It is also referred to as TYPE 2 ERROR

† False Positive: When the actual value is Negative but the predicted value is positive. It is also referred to as TYPE 1 ERROR

Since, the research work is performed on a medical dataset, the scope of the research focuses on reducing the False Negative, which is very dangerous for patients.

Eg: When a patient needs to undergo Cesarian, and the model predicts as not necessary it can be fatal.

A. Precision:

Ratio of correctly predicted positive classes among all pos-itive classes.

P r eci si on ˘ T P/(T P ¯ FP) (1

B. Recall:

Ratio of correctly predicted classes with all correctly pre-dicted classes. It needs to be higher in medical dataset predictions [35].

Rec al l ˘ T P/(T P ¯ F N)

4. Accuracy:

Ratio of correctly predicted classes among all classes. It needs to be as high as possible.

Accur ac y ˘(T P ¯T N)/(T P ¯T N ¯F N ¯F P)

5. F-measure:

A performance metric used to compare 2 models with varying recall and precision. It penalises extreme values.

F ¡measur e ˘(2⁄Rec al l⁄P r eci si on)/(Rec al l ¯P r eci si on)

The models proposed in this study are implemented in Python using scikit learn package.

VI. EXPERIMENTAL R ESULTS

Among the two errors : Type I and Type II, Type I error is the prediction of normal delivery as Caesarean, which will not result in endanger to life, whereas Type II error can be fatal since when Caesarean is required, the model suggests normal delivery , which can be fatal. Therefore, Type II error is costlier than Type I error.

Hence, a penalty cost of say, Rs. 10000 is imposed on the model as punishment with respect to the Type II error. As a result, a model with the minimum penalty cost is preferred over the other performance metrics(Accuracy, Recall, Precision,etc.). Since the dataset has only 80 instances, to minimise the variability of the model, the experimental process is repeated 5 times and average obtained is reported in the following tables:

The experimental results(65:35 split) of the 3 models along with confusion matrix are given below:

A. The results of ID3 and its confusion matrix is displayed below(in table II and III):

TABLE II

DECISION TREE -ID3 RESULTS

Metrics	Accuracy	Precision	Recall	f1-score	f1- score
Values	75.0 %	0.857	0.7058	0.77	0.77

TABLE III

CONFUSION MATRIX OF ID3

12	2
5	9

The precision and recall are about 75% in this model and moreover, false negative count is 5 which is on the higher side. Therefore, the models performance is considered as satisfactory.

B. The results of CART and its confusion matrix is displayed below(in table IV and V):

CART results are considered to be good since the recall percentage is above 90% and the false negative count is 2, which is desirable.

TABLE IV

DECISION TREE CART RESULTS

Metrics	Accuracy	Precision	Recall	f1-score	f1-score
Values	92.85 %	1	0.875	0.93	0.93

TABLE V

CONFUSION MATRIX OF CART

14	0
2	12

C. The results of Random forest and its confusion matrix is displayed below(in table VI and VII):

From the table VIII , it is very evident that Random Forests performance is better in all aspects(more precisely w.r.t. Recall and False negative Count) when compared to other models and it results in minimum penalty cost. From the figure 5, it is very clear that with regard to the false negative count ,which is one of the performance metrics of the study, random forest has yielded minimum count. Also, the performance of this research work has been compared with the work reported in the literature and it is shown in the table IX: It is clear from table IX that Random Forest used in this study has performed better than the model reported in the literature[36]. Also, it is necessary to include the performance metrics like the count of False Negative and Recall in the medical domain in selecting the model, which is absent in the other research study considered. From tables VIII and IX, we can easily conclude that Random Forest is best suited for this particular dataset. It can ably assist the doctors with regard to taking suitable decisions in performing the surgery.

CONCLUSION

This research work has focused on the use of decision tree classifiers like ID3, Random Forest and CART for classification. In this research study, the performance of the proposed models have been compared with the models reported in the literature. Among the models, Random Forest seems to perform better than the other Decision Classifiers used.

TABLE VI

RANDOM FOREST RESULTS

Metrics	Accuracy	Precision	Recall	f1score	f1-score
Values	96.42%	1	0.93	0.96	0.96

TABLE VII
CONFUSION MATRIX OF RANDOM FOREST

14	0
1	13

TABLE VIII
COMPARISON OF DECISION TREE MODELS

Model	Accuracy	Precision	Recall	f1-score	count
ID3	75.0 %	0.857	0.7058	0.77	5
CART	92.85 %	1	0.875	0.93	2
RF	96.42%	1	0.93	0.96	1

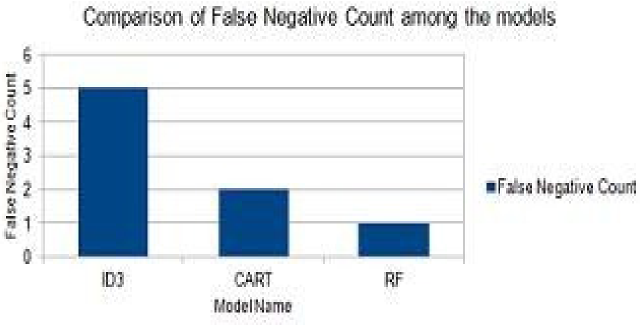

Fig. 5 Comparison of false negative count

TABLE IX: COMPARISON OF PROPOSED SYSTEM WITH OTHER MODELS

Performance metrics	Research study [36]	Proposed system
Accuracy	65%(J48)	964%(RF)
Recall	-	93.3%
False negative count	-	1(RF)

The scope of this work can be extended further by exploring the possibilities of other classifiers in the field of Statistics, Mathematics and Artificial Neural Networks(ANN). Also, bagging and boosting techniques can be attempted. Apriori Algorithm can be explored to find interesting pat-terns in the Caesarean dataset. C4.5,another Decision Tree classifier can also be experimented.

The performance of the model is also constrained by the number of samples available in the dataset. If the dataset is larger, then the behaviour and sustainability of the model can be very well understood. It will become more reliable to assist the doctors to take correct decisions in the medical domain.

REFERENCES

[1]Amin, Muhammad & Ali, Amir. (2017). Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions. 10.13140/RG.2.2.26371.25127.

[2] Demissie, K., Rhoads, G. G., Smulian, J. C., Balasubramanian, B. A., Gandhi, K., Joseph, K. S., & Kramer, M. Operative vaginal delivery and neonatal and infant adverse outcomes: population based retro-spective analysis. Bmj, 329(7456), 24. (2004)

[3] Simpson, K. R., and Thorman, K. E. Obstetric conveniences: elective induction of labor, cesarean birth on demand, and other potentially unnecessary interventions. The Journal of erinatal and neonatal nursing,19(2), 134-144. (2005)

[4] Pereira, S., Portela, F., Santos, M.F., Machado, J., & Abelha, A. (2015). Predicting type of delivery by identification of obstetric risk factors through data mining. Procedia Computer Science, 64, 601-609.

[5] Bharati, M. & Ramageri, Bharati. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering.

[6] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996

[7] M. Goebel and L. Gruenwald, A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations Newsletter, vol. 1, no. 1, pp. 2033, Jun. 1999.

[8] Salloum, Said & Al-Emran, Mostafa & Monem, Azza & Shaalan, Khaled. (2018). Using Text Mining Techniques for Extracting Infor-mation from Research Articles. 10.1007/978-3-319-67056-0_18.

[9] Raymond Kosala and Hendrik Blockeel. 2000. Web mining re-search: a survey. SIGKDD Explor. Newsl. 2, 1 (June, 2000), 115. DOI:https://doi.org/10.1145/360402.360406

[10] Antonie, Luiza & Zaïane, Osmar & Coman, Alexandru. (2001). Application of Data Mining Techniques for Medical Image Classification. In Proceedings of MDM/KDD. 94-101.

[11] Kovalerchuk B., Vityaev E. (2005) Data Mining for Financial Applications.In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA

[12] Shekhar S., Zhang P., Huang Y. (2009) Spatial Data Mining. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA

[13] Arun, K., Ishan, G. and Sanmeet, K., 2016. Loan Approval Prediction based on Machine Learning Approach. IOSR J. Comput. Eng, 18(3), pp.18-21.

[14] Yee, O.S., Sagadevan, S. and Malim, N.H.A.H., 2018. Credit card fraud detection using machine learning as data mining technique. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(1-4),pp.23-27.

[15] Dey, Lopamudra, et al. "Sentiment analysis of review datasets using naive bayes and k-nn classifier." arXiv preprint arXiv:1610.09982 (2016).

[16] V. Mhetre and M. Nagar, "Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA,"2017 International Conference on Computing Methodolo-gies and Communication (ICCMC), Erode, 2017, pp. 475-479, doi: 10.1109/ICCMC.2017.8282735.

[17] Chen, W., Zhang, S., Li, R. and Shahabi, H., 2018. Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. Science of the total environment, 644, pp.1006-1018.

[18] Chaurasia, Vikas and Pal, Saurabh, A Novel Approach for Breast Cancer Detection Using Data Mining Techniques (June 29, 2017). International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 1, January 2014.

[19] D. Chauhan and V. Jaiswal, "An efficient data mining classification approach for detecting lung cancer disease," 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2016, pp. 1-8, doi: 10.1109/CESYS.2016.7889872.

[20] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), Bangkok, 2017, pp. 158-161, doi: 10.1109/IEMECON.2017.8079581.

[21] Ou-Yang C., Rieza M., Wang HC., Juan YC., Huang CT. (2013) Applying a Hybrid Data Preprocessing Methods in Stroke Prediction. In: Lin YK., Tsao YC., Lin SW. (eds) Proceedings of the Institute of Industrial Engineers Asian Conference 2013. Springer, Singapore

[22] Das H., Naik B., Behera H.S. (2018) Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach. In: Pattnaik P., Rautaray S., Das H., Nayak J. (eds) Progress in Computing, Analytics and Networking. Advances in Intelligent Systems and Computing, vol 710. Springer, Singapore

[23] Perveen, Sajida, et al. "Performance analysis of data mining classi-fication techniques to predict diabetes." Procedia Computer Science 82 (2016): 115-121.

[24] Singh, M. "Classification system via data mining algorithm: New tool to diagnose Alzheimer's disease." Journal of the Neurological Sciences 405 (2019): 161-162.

[25] Shahbaz, M., Ali, S., Guergachi, A., Niazi, A., & Umer, A. (2019). Classification of Alzheimer's Disease using Machine Learning Techniques. InDATA (pp. 296-303).

[26] Nair M.S., Pandey U.K. (2020) Rule Generation of Cataract Patient Data Using Random Forest Algorithm. In: Tiwary U., Chaudhury S. (eds) Intelligent Human Computer Interaction. IHCI 2019. Lecture Notes in Computer Science, vol 11886. Springer, Cham

[27] Zhang, K., Liu, X., Jiang, J., Li, W., Wang, S., Liu, L., ... & Wang, L. (2019). Prediction of postoperative complications of pediatric cataract patients using data mining. Journal of translational medicine, 17(1), 2.

[28] https://www.geeksforgeeks.org/random-forest-regression-in-python/

[29] https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674

[30] https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

[31]
https://www.google.co.in/url?sa=i&url=https%3A%2F%2Ftowards
datascience.com%2Fdecision-tree-algorithm-explained
83beb6e78ef4&psig = AovVaw3WA6eKwqK9E55Lved
=0CA0QjhxqFwo TCJD79biLtOoCFQAAAAAdAAAAABAD

[32] Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J., 1984. Classification and Regression Tree Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California

[33] https://machinelearningmastery.com/classification-and-regression-trees-for-machine-
learning/#:˜:text=Creating%20a%20CART%20model%20involves

[34] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[35] https://medium.com/@erika.dauria/accuracy-recall-precision-80a5b6cbd28d

[36] Ayyappan, G. "Various classifications for caesarean section classification dataset data set." Indian Journal of Computer Science and Engineering (IJCE) 9.6 (2018): 145-147.