Decision Support Tool for Uterine Fibroids Treatment with Machine Learning Algorithms – A Study

[1]Dr.V.Sumathy, [2]Dr. S.J.Rexline, [3]Ms.T.D. Gowri

[1]Asssistant Professor, Department of Data Science, Loyola College, Chennai.
[2]Assistant Professor, Department of Computer Science, Loyola College, Chennai
[3]PG Research Scholar, Department of Data Science, Loyola College, Chennai.

DOI: 10.29322/IJSRP.12.08.2022.p12853
http://dx.doi.org/10.29322/IJSRP.12.08.2022.p12853

Paper Received Date: 4th August 2022
Paper Acceptance Date: 20th August 2022
Paper Publication Date: 24th August 2022

***Abstract*-** Uterine fibroids are benign growth in the tissues of the uterus which gives discomforts in the form of symptoms like over bleeding, pain in lower abdomen, irregular periods, misconception, in conception etc., for which there are several possible treatment options. Patients and physicians generally approach the decision process based on a combination of the patient's degree of discomfort, patient preferences, and physician practice patterns. While there have been many successes in applying data mining technology to the improvement of diagnostic accuracy. In this paper the use of classification algorithms in combination with Machine learning algorithms as a decision support tool to facilitate more systematic fibroid treatment decisions is examined. Machine learning algorithms like Decision Tree Classifier, Gaussian Naive Bayes, Random Forest Classifier, K-Nearest Neighbours, Gradient Boosting Classifier and XG Boost Classifier algorithms results are used to decide the possible decision for treatment.

***Index Terms*-** Machine learning algorithms, uterine fibroids, Decision Tree Classifier, Gaussian Naive Bayes, Random Forest Classifier, K-Nearest Neighbours

I. INTRODUCTION

Uterine fibroids are benign tumours of the uterus which are formed by the excessive growth of smooth-muscle cells in the wall of the uterus. They tend to be round-shaped with well-defined boundaries and their diameter can range from 1 cm to more than 10 cm[2]. They will grow as one neoplasm, or there are several of them within the womb. There are three types of fibroids based on their position in the uterus namely Intramural, Submucosal and Subserosal. The treatment options namely hysterectomy, Myomectomy, Gonadotropin Releasing Hormone (GnRH) depends on the symptoms, age, marital status etc., Fibroids are commonly diagnosed in all women, but certain factors such as age, ethnicity and heredity increase the risk.

Computer aided approach to fibroid treatment decision is described in many papers which includes identifying an appropriate distribution to estimate the waiting time of the patient to get a treatment [6]. Studies on non-parametric approach are also done to estimate the waiting time of the patient to receive an appropriate treatment[7]. Simplified case-based reasoning software has been developed to assist clinicians in helping their patients in taking the right fibroid treatment decision. While there have been many successes in applying data mining technology to the improvement of diagnostic accuracy of the Uterine fibroids [3]. Data Mining is that the technique of extracting vital data from an over sized information set. The applying of knowledge mining in extremely noticeable fields like medical, health management [1]. The health care data are wealthy however information is extremely poor attributable to lack of facility. So, there's an absence of effective analysis tools to get hidden relationships in information [4].

II. METHODS

The primary objective of this research paper is to analyze the Uterine fibroids data using various Machine learning algorithms like Decision Tree Classifier, Gaussian Naive Bayes, Random Forest Classifier, K-Nearest Neighbours , Gradient Boosting Classifier and XG Boost Classifier algorithms and that could be a useful to the patients and physician to make possible decision for treatment . The pipeline of the procedure is shown in Figure 1.

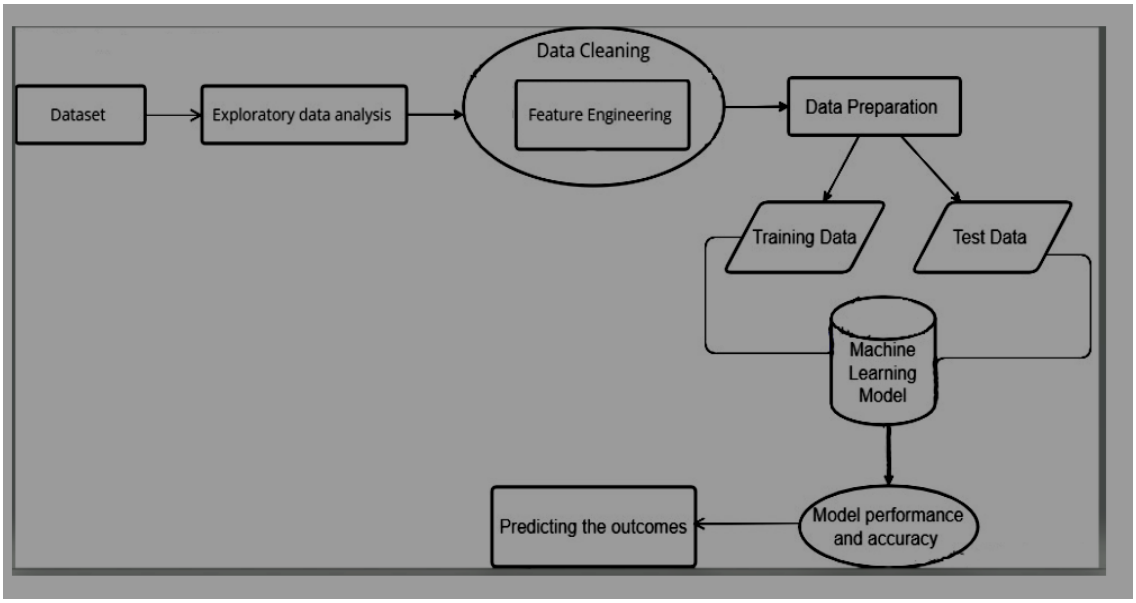

Figure 1. Pipeline of the Machine learning procedure

Data source is Kaggle.com and list of the features involved in the analyses are Age, Blood group, Age of menarche, Age of marriage, Age of menopause family income, place of residence, Location of fibroid, no., of fibroids, fibroid size, Symptom, no of children, month & year of diagnosis, month & year treatment, waiting time (in months), Type of Treatment taken.

The model is trained in such a way that it is capable of understanding the symptoms along with the issues related to fibroids to predict and choose the exact treatment accordingly. It is also an automated process where in the machine is trained only once and it gets easily capable to deal with the future data. Machine learning models used to analyze the information are Decision Tree Classifier, Gaussian Naive Bayes, Random Forest Classifier, K-Nearest Neighbours, Gradient Boosting Classifier and XG Boost Classifier algorithms. Type of Treatment is taken as the target variable and it has multiple classification levels.

3.1 Decision Tree Classifier

In Decision tree the given features are divided into Parent node (Root node) which forms the mountain top followed by the child node which takes the central part of the decision tree. At the end we could spot the Leaf node (End node) where in, it indicates the final decision. At the leaf node, the data could not be further split as it reaches the end point. Information gain measured from Entropy and Gini impurity are used to split the nodes.

Entropy:

It is a measure of impurity or randomness of each result after it is been split at every level. It is meant to be reduced as and when a new level is being introduced.

$$\boldsymbol{E(S) = \sum_{i=1}^{c} -p_i log_2 p_i} \quad \text{---(1)}$$

The value of Entropy always lies in between 0 and 1.

Gini Impurity:

Gini Impurity is same as entropy concept but the formula implemented is slightly different. In most of the cases, Gini Impurity is preferred over entropy as it is computationally more effective since it requires less time to execute due to not possessing a logarithmic function.

$$\boldsymbol{Gini = 1 - \sum_{i=1}^{c} (p_i)^2} \quad \text{---(2)}$$

The value of Gini impurity is lies between 0 to 0.5

Information Gain:

Information gain is the measure of amount of information that could be extracted at each level which helps in taking decision to further build the decision tree model. The main aim is to increase the information gain at each and every level. As the entropy decreases, the information gain increases.

$$\boldsymbol{Gain(S, A) = H(S) - \sum_{V \in Values(A)} \frac{|S_v|}{|S|} H(S_v)} \quad \text{---(3)}$$

3.2 Naive Bayes

In Machine Learning, Naïve Bayes falls under the category of Supervised Classification Algorithm. It is the probability of one event occurring given that the other event has already occurred. The formula is indicated as

$$P(A|B) = \frac{P(B|A)p(A)}{P(B)} \quad \text{---(4)}$$

The naïve Bayes formula is used to calculate the probability of each class and the class with highest probability is given as the outcome.

3.3 Random Forest Classifier

Random forest which uses Bootstrap Aggregation is an ensemble technique which is combination of multiple machine learning models. It includes Bagging and Boosting Techniques

Bagging:

Random forest builds multiple decision trees where in each decision tree is instructed to work on its data separately. The dataset considered is split into many samples and is fed to the decision tree models which happens with replacement that is the rows could also be repeated. Each data sample is been processed by the respective models. For each and every decision tree model, it produces its own output which is then combined at last to produce a single output

3.1 K-Nearest Neighbor

The K value indicates the number of nearest datapoints that is being considered in terms of distance. The distance is calculated with the help of Euclidean or Manhattan. The K value considered should always be an odd number in order to neglect the same number of datapoints appearing in both the categories. A user defined loop is designed which starts from 1 till the preferred n. And the higher accuracy helps in choosing the right K value.

Gradient Boosting Algorithm

Gradient boosting is one of the categories of Boosting techniques and is a commonly used algorithm to solve real time big data. Here, Multiple models are combined together to form the final output. The main aim is to optimize the loss function which indicates that it tries to reduce the value of the error as much as it could.

3.3 XG Boost Classifier

Extreme Gradient boost is a boosting algorithm. It is an extension of Gradient Boosting machine learning algorithm. It is a regularized type of boosting which means that it automatically prevents over fitting. It is also designed in a way where in it deals with null values internally. It offers a good balance between bias and variance.

III. Performance Analysis

There are numerous ways to evaluate the performance of a classifier. Here Precision-Recall curves are used to evaluate the performance of the Decision Tree Classifier, Gaussian Naive Bayes, Random Forest Classifier, K-Nearest Neighbours , Gradient Boosting Classifier and XG Boost Classifier algorithms. Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. A PR curve is simply a graph with Precision values on the y-axis and Recall values on the x-axis. In other words, the PR curve contains TP/(TP+FN) on the y-axis and TP/(TP+FP) on the x-axis. It is important to note that Precision is also called the Positive Predictive Value (PPV). Recall is also called Sensitivity, Hit Rate or True Positive Rate (TPR).

To estimate performances in the models, we employed accuracy, sensitivity as a parameter.

The accuracy, sensitivity and specificity were calculated by TP, TN, FP and FN. Accuracy is the overall correctness of the model and Sensitivity is measure of the ability of a prediction model to select instances of a certain class from a data set. It is commonly called as recall and corresponds to the true positive rate.

Figure 2 shows the Average precision value of Decision Tree classifier. It gives the Average precision value as 0.58.
Figure 2.Average precision value of Decision Tree classifier
Figure 3 shows the Average precision value of Gaussian Naïve Bayes. It gives the Average precision value as 0.38.

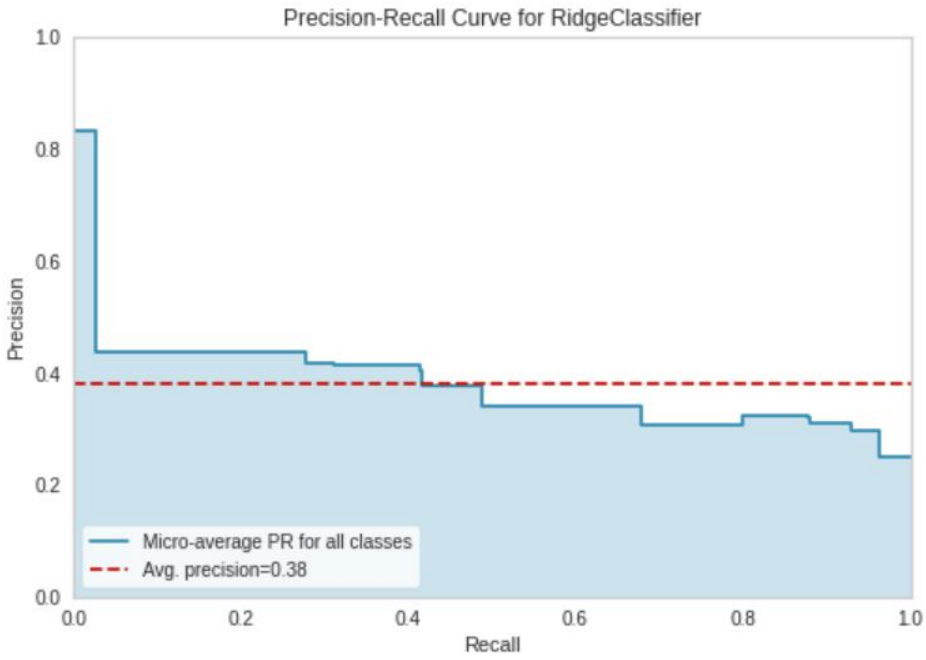

Figure 3.Average precision value of Gaussian Naïve Bayes

Figure 4 shows the Average precision value of Random Forest Classifier. It gives the Average precision value as 0.57.

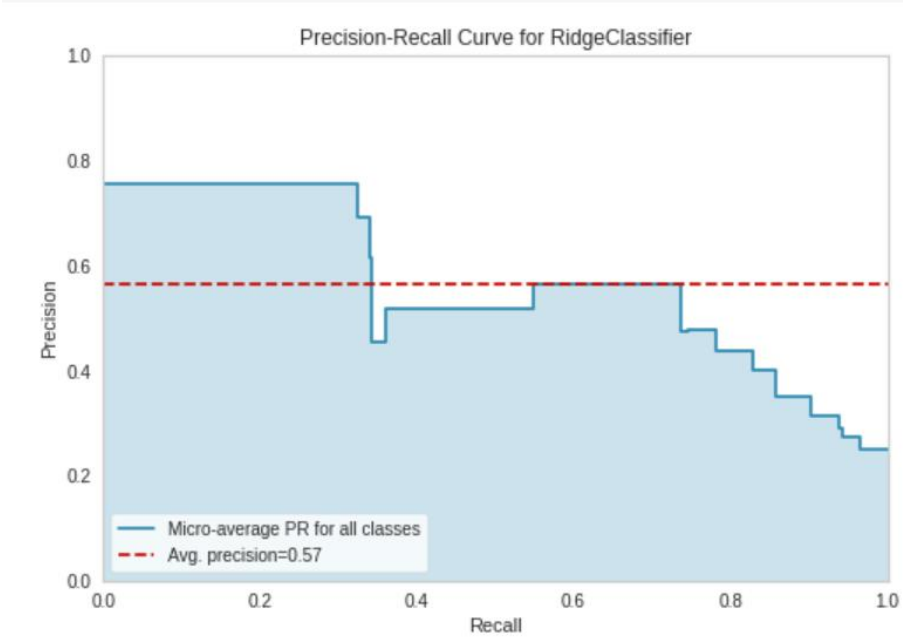

Figure 4.Average precision value of Random Forest Classifier

Figure 5 shows the Average precision value of K-Nearest Neighbours. It gives the Average precision value as 0.51.

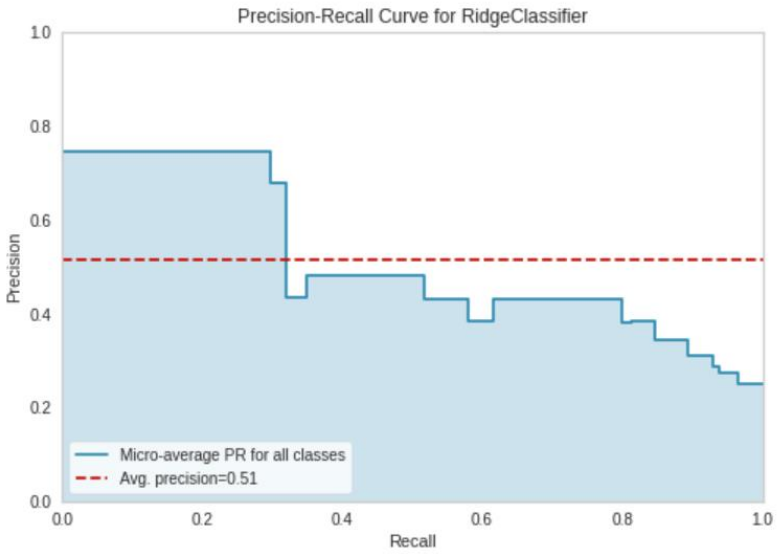

Figure 5.Average precision value of K-Nearest Neighbours

Figure 6 shows the Average precision value of Gradient Boosting Algorithm. It gives the Average precision value as 0.56.

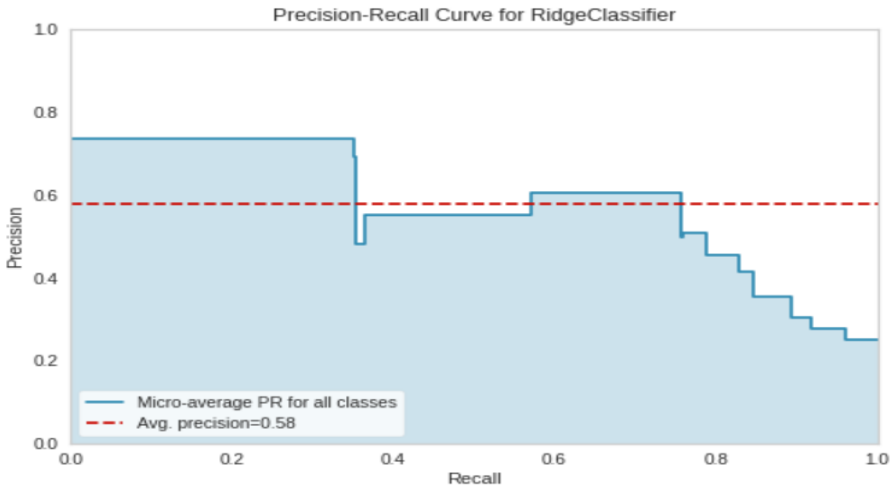

Figure 6.Average precision value of Gradient Boosting Algorithm

Figure 7 shows the Average precision value of XG Boost Algorithm. It gives the Average precision value as 0.51.

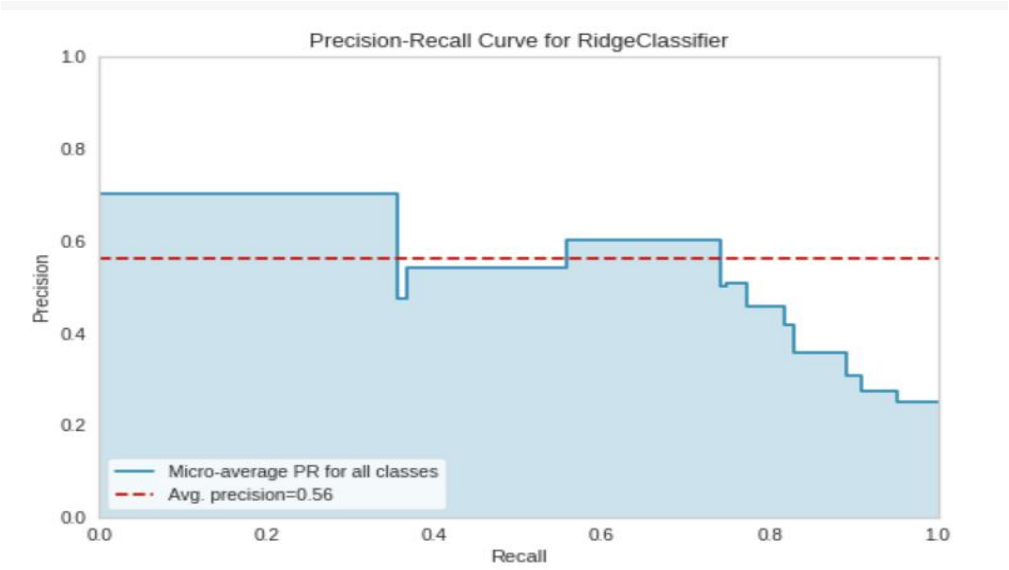

Figure 7. Average precision value of XG Boost Algorithm

Hereby the accuracy for the models is discussed in details. The Gradient Boosting Classifier is giving highest accuracy with 77%. The accuracy for the Decision Tree Classifier is 76% .And XG Boost Classifier gives the accuracy with 75%. Table 1 shows the accuracy value for all the Classification Algorithms. Figure 8 shows the model's performance based on its accuracy value for all the Classification Algorithms.

Table1. Accuracy value for all the Classification Algorithms

Classification Algorithms	**Accuracy Score**
Decision Tree Classifier	76%
Gaussian Naive Bayes	34%
Random Forest Classifier	73%
K-Nearest Neighbours	68%
Gradient Boosting Classifier	77%
XG Boost Classifier	75%

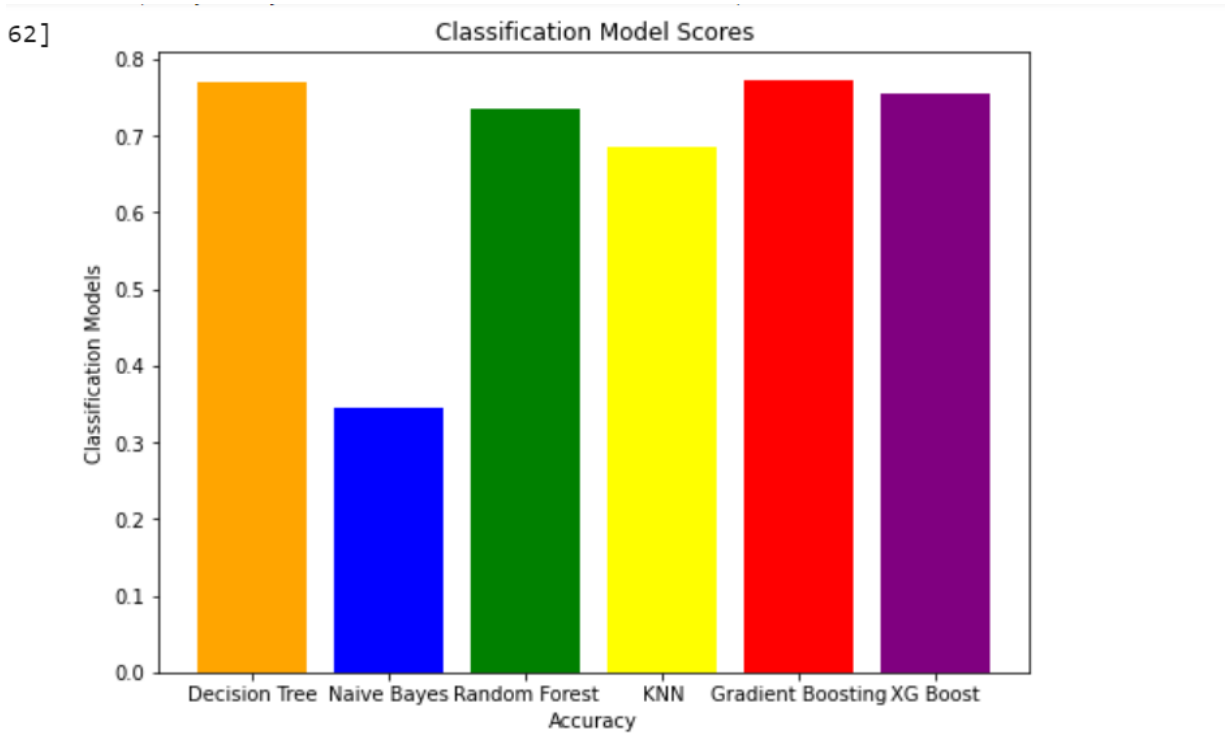

Figure 8. Accuracy value for all the Classification Algorithms

IV. Conclusion

In this paper, we have demonstrated machine learning approach to decision making for uterine fibroid treatments. We have used several data preprocessing strategies in order to optimize results from six well established classifying algorithms. The goal of classification is to accurately predict the target class for each case in the data. In order to believe any predictive model, the accuracy of the model must be estimated. We have noticed that decision tree model showed the highest accuracy, highest sensitivity and highest specificity for test options. By analyzing all these data, it is concluded that the excessive menstrual bleeding and pain in abdomen be the symptoms for fibroid.

References

[1] Aftarczuk, K.: Evaluation of Selected Data mining Algorithms Implemented in Medical Decision Support Systems. Blekinge Institute of Technology School of Engineering: Blekinge. (2007)

[2] Aziz R Samadi, Risk Factors for Self-Reported Uterine Fibroids: A Case-Control Study, Vol. 86, No. 6,American journal of Public health, June 1996.

[3] A.P. Binitie, Applying Case Based Reasoning System in the Treatment Decision of Gynecological Disorders: Fibroid, The International Journal Of Engineering And Science (IJES),Volume 2 ,Issue10 ,Pages 51-59 ,2013.

[4] Girija D.K, Comparison of Fibroid Syndrome Using Data Mining Techniques, International Journal of Emerging Technology and Advanced Engineering , Volume 2, Issue 11, November 2012.

[5] Girija D.K ,Classification of Women Health Disease (Fibroid)Using Decision Tree algorithm, International Journal of Computer Applications in Engineering Sciences,Vol Ii, Issue Iii, September 2012.

[6] J. Jothikumar & V. Sumathy, A Study on the Analysis of Fibroid Data Based on Parametric Distributions. International Journal of Current Medical And Applied Sciences, vol.6. Issue 2, April: 2015. PP: 99-103.

[7] J. Jothikumar & V. Sumathy, Non Parametric Method of Estimation of Survival and Hazard Functions for Patients with Uterine Fibroid, International Journal of Scientific and Research Publications, Volume 5, Issue 6, June 2015.

[8] Lisa S. Callegari, Associations between Race/Ethnicity, Uterine Fibroids, and Minimally Invasive Hysterectomy in the VA Healthcare System,Women health Issuses, Elsevier ,pp 1-8, 2018.

[9] E. Somigliana et al, Fibroids and female reproduction: a critical analysis of the evidence, Human Reproduction Update, Vol.13, No.5 pp. 465–476, 2007 ,Advance Access publication June 21, 2007

[10] Vineeta,Asha S Manek, Pranay Mishra, UFMDRA: Uterine Fibroid Medicinal Drugs Review Analysis, IOP Conference Series: Materials Science and Engineering,2021.

[11] Witten, I.H., Frank E.: Data Mining, practical Machine Learning Tools and Techniques. San Fancisco: 2nd Elsevier. (2005)

[12] Xuan, J., Deng, G., Liu, R., Chen, X. and Zheng, Y., 2020. Analysis of medication data of women with uterine fibroids based on data mining technology. Journal of infection and public health, 13(10), pp.1513-1516.

AUTHORS

First Author – Dr.V.Sumathy, Asssistant Professor, Department of Data Science, Loyola College, Chennai., sumathy@loyolacollege.edu
Second Author – Dr. S.J.Rexline, Assistant Professor, Department of Computer Science, Loyola College, Chennai, rexlinesj@loyolacollege.edu
Third Author – Ms.T.D. Gowri, PG Research Scholar, Department of Data Science, Loyola College, Chennai., 21pds011@loyolacollege.edu