*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 1, Ser. I (Jan - Feb 2020), PP 38-41*
*www.iosrjournals.org*

# Comparative Study on Predicting Gynecological Consequences using Data Mining Algorithms

Priyanka Mazumder[1], Dr. Siddhartha Baruah[2]
[1](MCA Department, Jorhat Engineering College, India)
[2](MCA Department, Jorhat Engineering College, India)

***Abstract:*** *Disease prediction is one of the critical tasks which always depend on relevant evidence with appropriate data. The paper deal with the comparison of supervised method with unsupervised method to understand the variance of accuracy. A gynecological dataset is used to see and compare the nature of data mining algorithm in providing outcomes with respect to pre processing steps. The KNN imputation is explained while dealing with missing value in raw dataset. The understanding of WEKA and Python in handling data pre processing step is also carried out in this paper.*

Date of Submission: 01-01-2020 Date of Acceptance: 16-01-2020

## I. Introduction

Data mining is the study that always extracts valuable information in every field. Data mining is best known for its improved algorithms and for the best knowledge providence. The data mining techniques provides mostly prediction and suggestion of improvement in the field of medical healthcare. Medical Healthcare is the biggest aspects where many progressive analysis and fruitful outcomes is given by Data Mining Techniques. Medical Healthcare System always constructs with huge data set and this data set if compiled can be found many useful pattern of recognizing various diseases. Data mining techniques are capable of handling and extracting knowledge from the large dataset. The algorithms are rich within it to gather information and give predictive analysis or outcomes. The algorithms are grouped into two types- Supervised and Unsupervised techniques, this paper handle both the types and explanation with comparison of which algorithm is better and more appropriate. Supervised algorithms or techniques include the most relevant information of the model and provides outcomes depending upon the previous acknowledged data. In simpler words the acceptance of training and test data by the model is used in supervised techniques. Unsupervised algorithm and techniques are those which don't include any information of the model and provide result depending upon the present scenario.

Now a day's diagnosis of disease is one of the toughest job for medical units and it's highly depends on the assumptions and predictions upon certain medical test and imaging reports. This paper considered two important algorithms from supervised and unsupervised techniques, Random Forest Algorithm which falls under supervised techniques and Expectation Maximization which is of types unsupervised. Random forest algorithm is the type of ensemble techniques which combine many weak learners to form a single strong learner. Random forest algorithm is the upgraded formation of Decision tree techniques where it constructs more than one decision tree with different structure of outcomes and test data will be the result with maximum similarity among the tree models. It is highly used and recommended algorithm as it is free from the problem of pruning and can handle the missing value in the dataset. It also provides better accuracy compare to other decision tree techniques. Expectation Maximization is the clustering techniques that help to build a practical model with maximum expectation result. It is a type of a statistical model that handles the unknown parameters. If any study wants to include additional attributes then the Expectation Maximization will provide accurate probability either to be included or rejected. It is one of the best likelihood finding using statistical data analysis. The two models supervised and unsupervised will be tested using the WEKA and the accuracy will be compared. Here in this paper the UCI data set will be used to get the accuracy of the algorithms.

The paper is organized as follows. In section 2, data preprocessing step will be explained in addition to KNN imputations. In Section 3, UCI Data set explanation will be provided. In section 4, comparison of algorithm will be shown and finally, in section 5 conclusions and consider future work will be justified.

## II. Data Pre-processing Steps using KNN Imputation

Data preprocessing step is one of the most considerable steps in data mining. If the data is accurate the algorithm works accurately and provides justified result. Data preprocessing step include many stages like cleaning, transformation and reductions. The most important aspects of the data preprocessing is the handling of missing value in the data set. Missing value must always be handled with best possible method and must not be

deleted as it is beyond the rules and ethnicity. There are various methods to handle the missing value but most appropriate method is the usage of the simple lazy learner algorithm K Nearest Neighbor (KNN) method. KNN provide best accuracy to fill the appropriate value in the data set that is missed out.

Missing value in a particular dataset can be of different types – Missing Completely at Random (MCAR), in this case the data is independent of observational data. Next is Missing at Random (MAR), in this case data is highly depend on the attributes outcomes and observational value. Finally, the last case is Missing not at Random (MNAR), this case covers those value which the participant usually hide or avoid to answer [ 2] . KNN algorithm here finds the most appropriate nearest value and handles the missing values.

The process by which KNN is used to handle missing data is called NN imputation techniques. NN imputation technique replace the value either by value that are correctly measured for another record in the dataset called 1NN or by assigning K value considering the average of the continuous dataset called KNN [3].

## III. Dataset Information

In this paper UCI data set is considered (https://archive.ics.uci.edu/ml/datasets/Caesarian+Section+Classification+Dataset) where it was explained that depending on what factor the pregnant women goes for Caesarian delivery instead of normal. The dataset contain 6 attributes. Attribute age is the real calculation of age of the pregnant women, next is delivery number as name suggest it defines who many time the women went for previous caesarian. Next is the Delivery time which explains when the baby was born. Attribute blood of pressure is the measuring of blood pressure. Attribute heart problem shows either the patient has heart problem or not. The attributes a type is explained value along with their value.

1. 'Age'{22,26,28,27,32,36,33,23,20,29,25,37,24,18,30,40,31,19,21,35,17,38}
2. 'Delivery number'{1, 2, 3, 4}
3. 'Delivery time' { 0,1,2 } {0 = timely , 1 = premature , 2 = latecomer}
4. 'Blood of Pressure' { 2,1,0 } {0 = low , 1 = normal , 2 = high }
5. 'Heart Problem' { 1,0 } {0 = apt, 1 = inept }
6. Caesarian { 0,1 } -> {0 = No, 1 = Yes }

## IV. Data Pre - Processing using WEKA

WEKA is the data mining tool develop by The University of Waikato where many data mining algorithms are predefined to test upon the real data. WEKA tool include the pre processing techniques into two categorical types under supervised and unsupervised. Both supervised and unsupervised is again sub divided into instances and attributes. Here instances are the value of the dataset and attributes are the variables. WEKA also has three possibilities of handling missing value. First possibility is replacing the missing value with mean and modes in the training set, next possibility is user replacement with constant value and last is to include the missing value. Figure 1 and Figure 2 shows the picture where missing value is replaced by the mean and mode of the training dataset.

There are many options to filter or pre process the data using WEKA, the advantage of pre processing using WEKA is that it gives easy understanding provides no time filter, works depend upon the user configuration and very much user friendly tool. WEKA pre processing help to increase the accuracy and decrease the error rate in the dataset which is very important to get best outcomes.

**Table 1: Shows Entry in Missing Value Data**

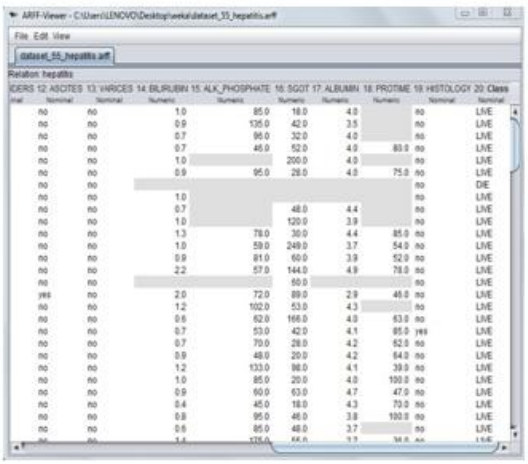

**Figure 1**
**Data set with Missing Value**

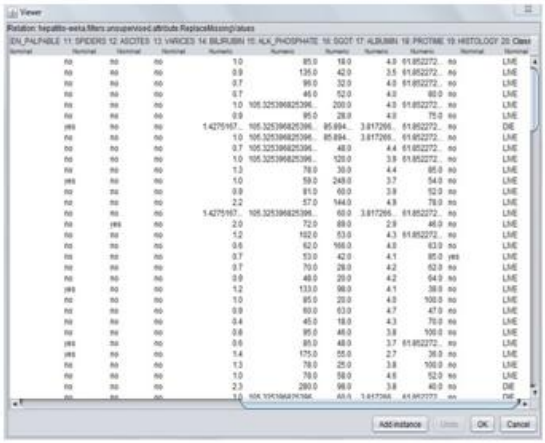

**Figure 2**
**Missing value is replaced with mean**

## V. Data Pre – Processing using Python

It is the core programming configuration to apply pre processing step in the dataset. Python is most usable and one of the best interpretable programming language which gives less coding guide with useful computation. There is almost every possibility to handle the pre processing steps but we will carry out the understanding on missing value replacement. Python is rich because of its libraries while handling the missing value in a dataset we first need to import libraries. Libraries will help to call the functions that will work upon the missing value. One of the famous libraries python has is Numpy, Numpy contain scientific computing and next is Pandas, it is basically used for data manipulation and handling. Python is capable of handling almost every type of data like categorical, binary, standardized data etc. We also need to import dataset in the python it support all extension to process the dataset. There are two ways to handle missing value in python - first is deleting the missing value row which is not applicable if suppose 50% dataset contain missing value. Here if we delete the row entire dataset will be affected and outcomes will not be liable and accurate while predicting. Second is either to replace with mean, median and mode value or finding value nearest to its neighbor. Second option is most useful to handle the missing value in python.

**Table 2: Shows Comparison of different algorithm with same dataset**

| | Algorithms | Accuracy |
|---|---|---|
| 1. | J48 | 51% |
| 2. | Random Forest | 62% |
| 3. | Simple K Mean | 53% |
| 4. | Expectation Maximization | 57% |

## VI. Comparison on Data Mining Algorithms

Data mining algorithms are considered in WEKA taking the mentioned dataset from UCI to find the accuracy. In classification categorical algorithm, decision tree is considered to be used as it is most accurate in providing outcomes, medical healthcare need to be more perfect in finding the outcomes which is found in decision tree approaches. There are many decision tree types like J48, Random Forest, Random tree etc. Among all we choose to work using the Random Forest Algorithm because it provide optional tree structure which help to analysis data more accurately. It also handles the unexpected configured test data and can handle the missing value. Random forest considers all the attributes to build decision tree, in other algorithm techniques some attributes are missed out while creating structural decision tree or model. Random Forest also handles large dataset with most appropriate decision outcomes. One of the best knowledge that Random Forest gives is that each individual decision tree model protect each other from their individual error.

The mention dataset is used in the WEKA considering the Random forest algorithm and the result is shown below in Table 2. Random Forest provides high accuracy and run efficiently compare to other techniques in the decision tree algorithm. It overcomes the problem of over fitting and time computation. The disadvantage

of random forest is that it is not good for the regression techniques and provides very little control on what the model does.

The clustering algorithm Expectation maximization is based on techniques which involve dataset of same classes. Here E and M is consider as a set where E stands for the using the current theory and M stands for generating the best theory using current configuration of the data. E is most expected model and M is the justification to be better in the techniques. It is completely a statistical model and most widely used unsupervised clustering algorithm. The mentioned data set is also tested using an Expectation maximization technique which is shown in Table 2.

## VII. Conclusion

The analysis of particular algorithm in data mining is highly depending on the quality of dataset. The medical data set is configured with missing value which can be handled using different techniques. The missing value that dataset contain must be analyzed and not deleted. The paper contains detail study of handling missing value which is most important to consider every aspect of the data frame. Ensemble algorithm random forest is used to test data set, now a day's ensemble algorithm techniques are more approachable than any other techniques because of its principal to handle individual error. In future, medical data testing can be studies using deep learning technique for finding the root cause of the disease.

## References


[1]. https://en.wikipedia.org/wiki/Random-Forest.
[2]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4959387/.
[3]. https://towardsdatascience.com/the-use-of-knn-for-mining-value-cf33d935c637
[4]. H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005).
[5]. R. Kandwal, P. K. Garg and R. D. Garg, "Health GIS and HIV/AIDS studies: Perspective and retrospective", Journal of Biomedical Informatics, vol. 42, (2009), pp. 748-755.
[6]. U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD process of extracting useful knowledge form volumes of data communication.", ACM, vol. 39, no. 11, (1996), pp. 27-34.
[7]. J. Han and M. Kamber, "Data mining: concepts and techniques", 2nd ed. The Morgan Kaufmann Series, (2006).
[8]. Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh, "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods.", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 12, December 2016.
[9]. Pratibha Devishri S, Ragin O R, Anisha G S, "Comparative Study of Classification Algorithms in Chronic Kidney Disease",International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019.
[10]. Sonu Kumari, Archana Singh, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus".
[11]. Vivek Agarwal "Research on Data Preprocessing and Categorization Techniques for Smartphone Review Analysis" International Journal of Computer Applications(0975-8887) Volume 131-No. 4, December 2015
[12]. C. Sowmiya, Dr. P. Sumithra, "A Comparative Study of heart disease prediction using Data Mining Techniques",International Joournal of Scientific & Engineering Research, Volume 7, Issue 12, December 2016.
[13]. Divya Tomar, Sonali Agarwal "A survey on Data Mining Approaches for Healthcare", International Journal of Bio Science and Bio Technology, Vol 5, No.5(2013), pp.241-266.
[14]. Hariharan K, Vigneshwar W. S., Sivaramakrishnan N, Subramaniyaswamy V, "A Comparative Study on Heart Disease Analysis using Classification Techniques", International Journal of Pure and Applied Mathematics, Volume 119, No 12 2018,13357-13366.
[15]. I Ketut Agung Enriko, "Comparative Study of Heart Disease Diagnosis Using Top Ten Data Mining Classification Algorithms", ResearchGate


Priyanka Mazumder. "Comparative Study on Predicting Gynecological Consequences using Data Mining Algorithms." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22.1 (2020), pp. 38-41.