ISSN (Online): 2230-7893
www.IJCEM.org

A Critical Study of Polycystic Ovarian Syndrome (PCOS) Classification Techniques

Vikas B. [1], B. S. Anuhya [2,] Manaswini Chilla[3] and Sipra Sarangi[4]

[1] Assistant Professor, [2,3,4]Research Scholar, Department of Computer Science and Engineering
GITAM Institute of Technology, GITAM
Visakhapatnam, India.

Abstract

Data mining, a field that can uncover patterns from large repositories, has numerous applications such as building predictive models which can be extremely beneficial in the healthcare industry. Polycystic Ovarian Syndrome is a hormonal disorder which is largely prevalent among women of reproductive age. In order to help in the diagnosis of PCOS, classification techniques such as Naïve Bayes, Decision Tree and Neural Networks can be applied to classify real time PCOS data based on an established training set. In this paper, an attempt has been made to compare the accuracies and other performance measures of the prior mentioned data mining techniques to predict whether a person is likely to have PCOS or not through.

Keywords: *Data Mining, Polycystic Ovarian Syndrome, Naïve Bayes Algorithm, Decision Trees, Artificial Neural Networks.*

1. Introduction

PCOS, Polycystic ovarian syndrome is a prevalent type heterogeneous endocrine disorder in women of child bearing age. Women associated with this abnormality are highly prone to infertility, anovulation, cardiovascular disease, type 2 diabetes, mellitus, obesity, gynecological cancer etc. Some of the common symptoms include oily skin and darkened acne marks, weight gain, hypertension, mood and anxiety disorder, irregularity in menstrual cycle etc. [1]. Nearly 5-10% of reproductive age women are affected by this ailment and some of the current techniques of medication include Metformin therapy for decrement of insulin, Progesterone based oral contraceptives, change in lifestyle, Lamotrigine for women with epilepsy, prevention screening and management of cardio metabolic features, coasting etc.

Despite the gravity of the circumstances and available options of medication, there are very little options for treatment. It is therefore of vital importance to sift through the patients at an early stage to prevent any consequences of this ovarian dysfunction. Hence in order to help better diagnosis of patients with improved accuracy, analysis of data collected from a survey of larger audience of women of age ranging from 15-45, has been done as series of steps. It involves grouping the data into groups onto which various data mining techniques such as association, classification and clustering is applied [2]. The data is made to go through various data pre-processing techniques to achieve optimum accuracy. Finally, the decision support system gives the prediction analysis as the output.

Data Mining has proved to a boon in the field of bioinformatics as it amalgamates the statistical science, machine learning and database systems into one thus simplifying the entire procedure of analytics and prediction of vast databases. In short, data mining is an interdisciplinary sub field of computer science which helps a great deal in gaining relevant data from a vast amount of incoherent and unrelated raw data.

In prior conducted research [3], an attempt had been made to recognize recurring patterns among symptoms of Polycystic Ovarian Syndrome patients using Frequent Itemset Mining. It also focused on Apriori Algorithm which had been used to predict those who are susceptible to the syndrome. By analyzing the experimental results, it was understood that some of the attributes have a significant influence in the prediction of PCOS. But these results did not provide with cent percent accuracy. Better diagnosis and prognosis of PCOS can be achieved by using a hybrid algorithm of classification to get results with improved accuracy which has been incorporated in this paper. Classification is a data mining technique which is used for extracting models describing important classes or to predict future data trends. In this data mining technique, a model or a classifier is constructed to predict the categorical labels. It is a data mining function that assigns items in a collection to target categories or classes. The main motive of classification is to precisely predict the target class for each case in the data. Classification has applications such as business modelling, customer segmentation, credit analysis and biomedical and drug response modeling, thus in this paper an attempt has been made to design a hybrid algorithm based on classification

ISSN (Online): 2230-7893
www.IJCEM.org

in order to predict attributes that are responsible for PCOS with better accuracy.

2. Dataset

The dataset for PCOS is a real-time data set that taken from a survey conducted among 119 women between the ages of 18 and 22 [4]. The dataset is primarily based on their lifestyle and food intake habits. The symptoms i.e. attributes are classified based on classification algorithms to predict whether the patient may have PCOS or not. The database consists of 119 samples with 18 attributes belonging to two different classes (maybe or maybe not). There are 14 binary attributes and 4 categorical attributes as shown in Table 1.

Table 1: PCOS Dataset

S.NO.	**ATTRIBUTE**	**VALUE**
1	CLASS LABEL	MAYBE, MAYBE NOT (mb, mb n)
2	REGULARITY OF MENSTRUAL PERIODS	Yes (y), Infrequent menses (im), Irregular bleeding (ib), Heavy bleeding (hb)
3	WEIGHT GAIN	Yes(y), No (n)
4	EXCESS FACIAL OR BODY HAIR.	Yes (y), No (n)
5	DARK AREAS ON SKIN	Yes (y), No (n)
6	PIMPLES	Yes(y), No (n)
7	DEPRESSION AND ANXIETY	Yes(y), No (n)
8	HISTORY OF DIABATES AND HYPER TENSION	Yes(y), No (n)
9	BODY WEIGHT MAINTAINENCE	Yes(y), No (n)
10	OILY SKIN	Yes(y), No (n)
11	LOSS OF HAIR	Yes(y), No (n)
12	FREQUENT EATING PLACES	Hostel mess(hm), Campus canteen (cc)
13	REGULAR EXERCISE	Yes(y), No (n)
14	MENTAL STRESS DUE TO NEW ADMISSION IN HOSTEL	Yes(y), No (n)
15	MENTAL STRESS DUE TO PERSONAL PROBLEMS	Yes(y), No (n)
16	MENTAL STRESS DUE TO PEER PRESSURE	Yes(y), No (n)
17	MENTAL STRESS DUE TO CHANGE IN DIETARY HABITS	Yes(y), No (n)
18	FAST FOOD INTAKE	Every day(ed), Once in a week(w), Once in a month(m), Once in a year (y)

3. Classification

The data mining function that deals with allocation of objects in a collection to classes is called as classification [5]. The main motive behind opting classification is to accurately predict the target class for each of the objects and the data in entirety. For example, a classification technique could be used to predict if a particular person has malignant tumor or benign tumor basing on the various characteristics the existing tumor possess. The classification of any dataset goes about in two main steps which are building up the classifier or model or target class and using this obtained classifier to do the classification of the entire data set. The classifier is extracted with the help of the training data set produced beforehand. Once the target class has been obtained, the test data sets can be classified with optimal accuracy.

There are many classification techniques and algorithms [6], out of which some have been listed below: -

1. Decision Tree Induction
2. Support Vector Machine (SVM) Algorithm
3. ID3 Algorithm
4. C4.5 Algorithm
5. Naïve Bayes Algorithm
6. Artificial Neural Networks (ANN) Algorithm
7. K nearest neighbour algorithm

3.1 Decision Tree Induction

The decision trees [7] have the sole purpose of generating rules. These rules are nothing but a conditional statements or parameters which can be easily interpreted by the users and in turn can be used within a dataset to recognize a set of records. A researcher under the name of J. Ross Quinlan evolved a decision tree algorithm in 1980 called as ID3 (Iterative Dichotomiser). Thereafter, he also developed C4.5 algorithm, which was the apparent inheritor of ID3. Both the algorithms make use of the greedy approach. Here no backtracking procedures have been inculcated as the trees are constructed following a top-down, recursive, divide-and-conquer mannerism.

3.2 Naive Bayes Algorithm

Naïve Bayes algorithm is a simple classifier based on probability [8]. The Bayesian principle is to allocate case to the class that has the highest posterior probability. By counting the frequency and the permutation of values, this classifier calculates a set of probabilities from the given data set. An assumption is made, such that all the fields are independent given the value of the class variable. This algorithm is called naïve because the above assumption rarely holds true in real life applications. Nevertheless, this algorithm's performance is good and also it has the ability learn quickly in supervised classification problems [9].

3.3 Artificial Neural Networks Algorithm

The idea of ANN is based on belief that working of human brain can be replicated silicon and wires as living neurons and dendrites. The neurons are organized in layers with one layer of input and output layer each and hidden layer(s). An ANN learns the relationships between the input and output data sets. During the training phase, training data is introduced into the neural network. The difference between the actual output values of the network and the training output values is then calculated. This difference is nothing but an error value which is decreased during the training by modifying the weight values of the connections. This is repeated until the error value has reached the predetermined training goal [10]. ANN's have the ability to learn, generalize and most importantly they place no restrictions on the input variables [11].

4. Classification Performance Measures

The performance of the above-mentioned classification technique can be calculated by the following metrics. For each algorithm we have observed the sensitivity specificity precision and accuracy described as follows [12]:

4.1 Sensitivity

Sensitivity is nothing but, the true positive rate. We can also define it as the fraction of positive tuples that are correctly classified.

$$Sensitivity = \frac{frequency\ of\ true\ positives}{frequency\ of\ true\ positives + frequency\ of\ false\ negatives}$$

4.2 Specificity

The negative rate which is nothing but the fraction of negative tuples that are correctly classified is called specificity.

$$Specificity = \frac{frequency\ of\ true\ negatives}{frequency\ of\ false\ positives + frequency\ of\ true\ negatives}$$

4.3 Precision

It is the measure in which the fraction of true positives in contrary to all positive results is calculated.

$$Precision = \frac{frequency\ of\ true\ positives}{frequency\ of\ true\ positives + frequency\ of\ false\ positives}$$

4.4 Accuracy

The percentage of the test tuples that are properly classified by the classifiers is nothing but the accuracy of the particular algorithm in hand.

$$Accuracy = \frac{frequency\ of\ true\ positives + frequency\ of\ true\ negatives}{frequency\ of\ true\ positives + false\ negatives + false\ positives + true\ negatives}$$

5. Experimental Results

The experiment is done using IBM SPSS Modeler 16.0 tool. The complete dataset is fed into the mining tool and the proposed algorithms are applied to it. The following figures depict the results vividly.

Figure 1 clearly shows the decision tree we get by applying the C5.0 algorithm on the data set. Attribute selection is done and the appropriate results are displayed in the form of a tree.

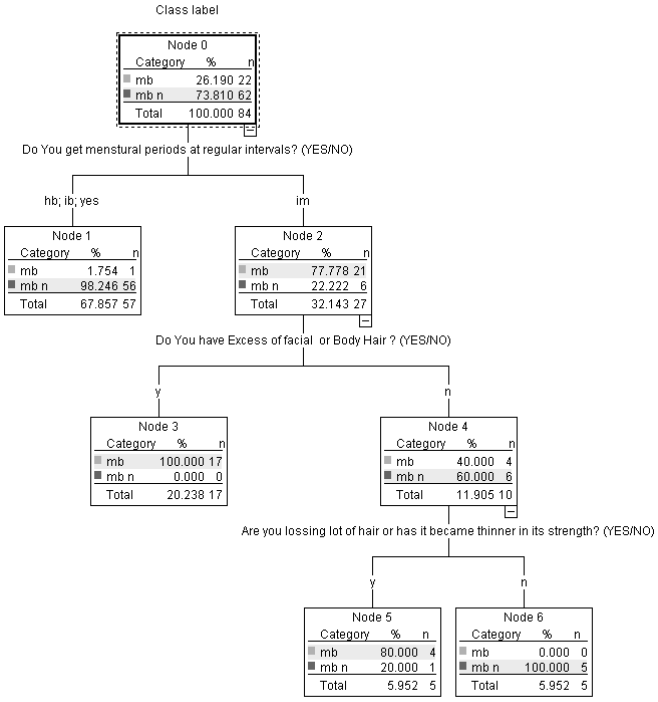

Fig. 1 Decision Tree obtained by applying C5.0 algorithm

The graph in figure 2 represents the affluent attributes that majorly contribute towards the prediction analysis. The algorithm applied here is the Artificial Neural Network(ANN) Backpropagation algorithm.

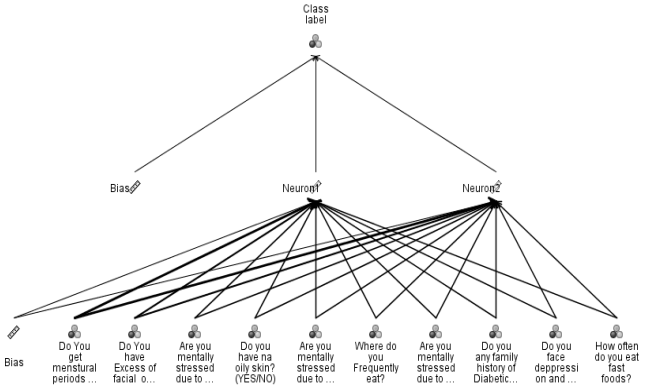

Fig. 2 Neural Network obtained by applying Backpropagation algorithm

The figure 3 depicts the results obtained by applying the Naïve Bayes Network on the entire dataset. We can see how the class label has been obtained by using different sets of attributes.

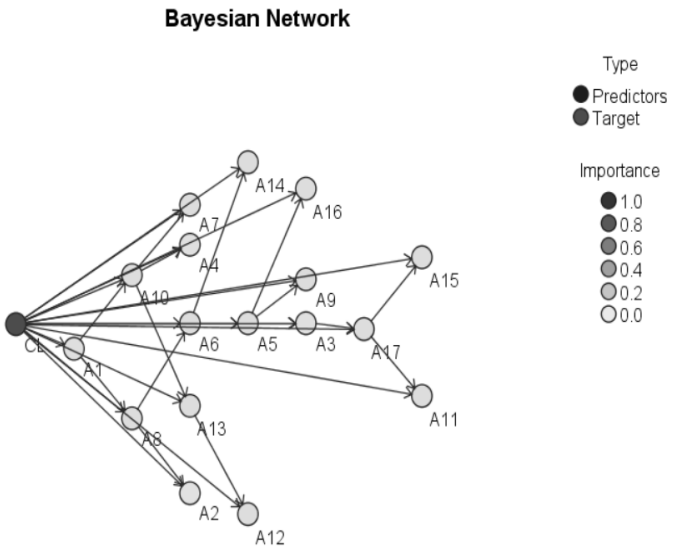

Fig. 3 Naïve Bayes algorithm applied on the dataset.

In figure 3 the attributes are denoted by notations given below:

A1 =Regularity of Menstrual Periods
A2 =Weight Gain
A3 =Excess Facial or Body Hair
A4 = Dark Areas on Skin
A5 = Pimples
A6 = Depression and Anxiety
A7 = History of Diabetes and Hyper Tension
A8 = Body Weight Maintenance
A9 = Oily Skin
A10 = Loss of Hair
A11 = Frequent Eating Places
A12 = Regular Exercise
A13 = Mental Stress Due to New Admission in Hostel
A14 = Mental Stress Due to Personal Problems
A15 = Mental Stress Due to Peer Pressure
A16 = Mental Stress Due to Change in Dietary Habits
A17 = Fast Food Intake
CL =Class label

6. Result Analysis

The paper deals with the application of three classification algorithms on the acquired data set and then drawing out a comparison of the results to one another. The results from the three selected algorithms, namely, Bayesian Network,

ISSN (Online): 2230-7893
www.IJCEM.org

C 5.0 Decision Tree and Artificial Neural Network backpropagation, were compared and tabulated. According to the outputs derived with the help of SPSS data mining tool, we have come to the conclusion that Naïve Bayes algorithm provides us with optimum accuracy of 97.65%, followed by ANN backpropagation algorithm with 96.27% and then by decision tree C 5.0 algorithm with 96.24%. In order to calculate the accuracy and find out their performance, a confusion matrix was constructed at first. From that matrix, the true positives, true negatives, sensitivity, specificity and precision was calculated using their specific formula. These parameters were then used to calculate the final accuracy.

Table 2: Performance analysis after applying Decision Tree C 5.0 Algorithm on dataset

Performance Metrics	***Classification Algorithm: Decision Tree C 5.0***
Sensitivity	0.95
Specificity	0.98
Precision	0.95
Accuracy	0.9624 (96.24%)

Figure 4 graphically represents the Performance analysis after applying Decision Tree C 5.0 Algorithm on the dataset.

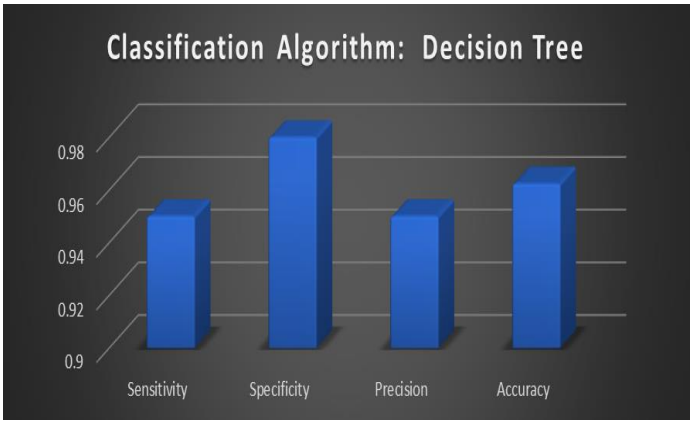

Fig. 4 Graphical representation of the Performance analysis after applying Decision Tree C 5.0 Algorithm on dataset

Table 3: Performance analysis after applying Neural Networks Backpropagation Algorithm on dataset

Performance Metrics	***Classification Algorithm: Neural Network Backpropagation***
Sensitivity	0.95
Specificity	0.98
Precision	0.95
Accuracy	0.9627 (96.27%)

Figure 5 graphically represents the Performance analysis after applying Neural Networks Backpropagation Algorithm on the dataset.

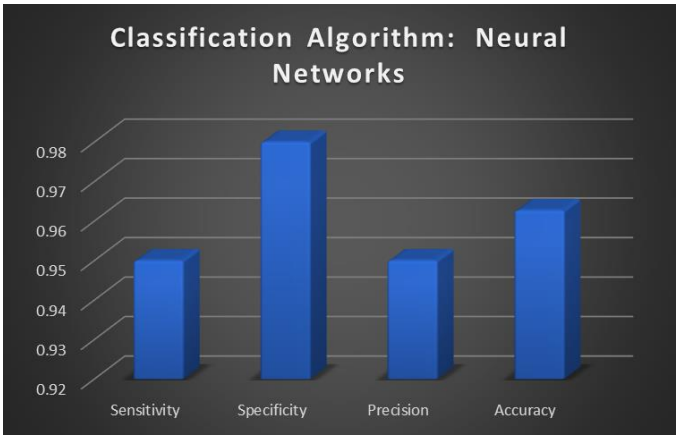

Fig. 5 Graphical representation of the Performance analysis after applying Neural Networks Backpropagation Algorithm on dataset

Table 4: Performance analysis after applying Naïve Bayes Algorithm on dataset

Performance Metrics	***Classification Algorithm: Naïve Bayes***
Sensitivity	0.95
Specificity	1
Precision	0.95
Accuracy	0.9765 (97.65%)

Figure 6 graphically represents the Performance analysis after applying Naïve Bayes Algorithm on the dataset.

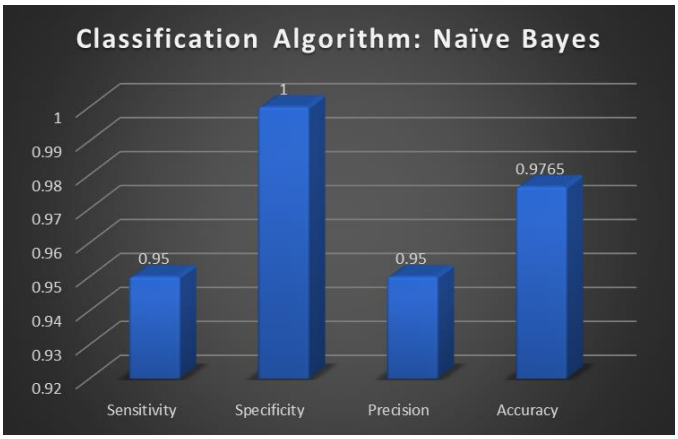

Fig. 6 Graphical representation of the Performance analysis after applying Naïve Bayes Algorithm on dataset

Figure 7 graphically represents the Performance analysis of all three algorithms.

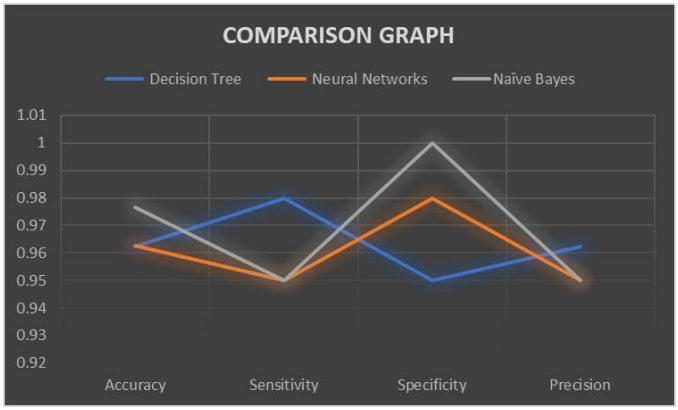

Fig. 7 Graphical representation of the Performance analysis of all three algorithms

Conclusion

In this paper, popular Classification algorithms were opted to evaluate the classification technique performance in the terms of performance measures which are accuracy, precision, sensitivity and specificity to classify if the particular patient may have PCOS or not. Classification technique was considered in the study as it enables us to predict if the patient has Polycystic Ovarian Syndrome or not based on the syndromes provided by the doctor or medical Centre. The dataset used up in this study has been collected via a survey conducted based on the lifestyle of the women. It has also been noticed that all the opted algorithms under the classification technique show near-about accuracy percentages, permitting the user to adopt any of the procedures. The accuracy of the prediction can be further enhanced by inculcating the data obtained from the clinical reports of the patient. Another aspect of predicting PCOS could be with the help of ultrasound images. By adding up the information to these classification techniques, doctors can foresee if the particular patient is susceptible to the syndrome, thus, help in curbing PCOS by seeking medical help and switching to a healthier lifestyle.

Acknowledgments

We would like to pay our heartfelt gratitude to Dr. Vijaya Lakshmi Chandrasekhar, Department of Obst. & Gyn., GIMSR, GITAM (Deemed to be University), Visakhapatnam for her generosity in accumulating the data and apprising us about the relationships among various symptoms. We would like to take this opportunity to thank GITAM (Deemed to be University) Girls Hostel Kokila Sadan and the Chief warden Prof. T. Sita Mahalakshmi for allowing us to conduct the survey and permitting us to gather data for the survey. We would also like to thank Prof. Y. Radhika, GITAM Institute of Technology, GITAM (Deemed to be University) for guiding us throughout the project and providing valuable resources.

References

[1] Vikas B, Sipra Sarangi, Manaswini Chilla, K Santosh Bhargav, B S Anuhya. (2017). A Literature Review on The Rising Phenomenon PCOS. *International Journal of Advances in Engineering & Technology,2(10), 216-224.*

[2] Mehrotra, Palak, et al. (2011). Automated Screening of Polycystic Ovary Syndrome Using Machine Learning Techniques. *IEEE India Conference (INDICON).*

[3] Vikas B, B.S.Anuhya, K Santosh Bhargav, Sipra Sarangi, Manaswini Chilla. (2017, June). Application of the Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome (PCOS). *4th International Conference on Information System Design And Intelligent Applications, 2017.*

[4] *PCOS-Survey/PCOSData*. (2017). *GitHub*. Retrieved 30 November 2017, from https://github.com/PCOS-Survey/PCOSData

[5] Oracle Help Centre. (2017). *Classification*. Retrieved from https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#CHDHBEGJ.

[6] Sagar S., Nikam.*A.* (2015). Comparative Study of Classification Techniques in Data Mining Algorithms. *Oriental Journal Of Computer Science & Technology,*1(8), 13-19.

[7] Dharm Singh, Naveen Choudhary, Jully Samota.(2013). Analysis of Data Mining Classification with Decision tree Technique. *Global Journal of Computer Science and Technology Software & Data Engineering*,13(13), 1-5.

[8] Durgesh K. Srivastava, Lekha Bhambhu.(2005). Data Classification Using Support Vector Machine. *Journal of Theoretical and Applied Information Technology*.

[9] Tina R. Patil, Mrs. S. S. Sherekar Sant Gadgebaba. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications,*2(6), 256-261.

[10] Korany, M., Mahgoub, H., Fahmy, O., & Maher, H. (2012). Application of artificial neural networks for response surface modelling in HPLC method development. *Journal Of Advanced Research*, *3*(1), 53-63.

[11] Anon, (2017). *Introduction to Neural Networks, Advantages and Applications*. Retrieved from https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207.

[12] Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu. (2011) A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. *International Journal of Database Management Systems,*2(3),101-114.

Vikas B received the Bachelor in IT degree from the JNTUH, Hyderabad, in 2010 and the Master in Bioinformatics degree from the JNTUH, Hyderabad, in 2012. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, GITAM (Deemed to be University), Visakhapatnam.

His research interests include Datamining, Bioinformatics, Information Security, and Data Sciences.

B S Anuhya is currently pursuing her Bachelor degree with the Department of Computer Science Engineering from GITAM (Deemed to be University), Visakhapatnam. Her research interests include data science, machine learning and cyber security.

Manaswini Chilla is currently pursuing her Bachelor degree with the Department of Computer Science Engineering from GITAM (Deemed to be University), Visakhapatnam. Her research interests include big data, network security and data mining.

Sipra Sarangi is currently pursuing her Bachelor degree with the Department of Computer Science Engineering from GITAM (Deemed to be University), Visakhapatnam. Her research interests include data mining, big data and language processors.