Prediction of Disease through Data Mining

Alok Nagargoje, Anand Bohara, Vijay Kakade, Mrunal Kale and Prof.P.S.Hanwate
NBN Sinhgad School of Engg,Ambegaon (Bk),Pune.

Abstract

According to the World Bank collection of development indicators in 2018, shows that out of the total population of India, the Rural population was reported as 65.97%. Due to less availability of public heath-care, lack of awareness, inadequate facilities and less knowledge regarding various diseases are among the main causes for improper health issues. With increase in various types of technologies supporting in the field of health-care, it can be possible to help to reduce the number of people dying by various diseases in rural. It can be achieved through various types of data of symptoms regarding different diseases that are generated by various hospitals, NGO's and various government organizations. We propose to build our project with the main focus on creating an application that classifies and predicts diseases on basis of prior data-sets regarding various symptoms on diseases and the user's knowledge about their health and symptoms. It consists of three different algorithms - Navies Bayes, Random Forest and Decision Tree to classify and predict the diseases. Due to lack of adequate health facilities, often people are not aware of their disease until it's too late. The private health-care is too costly and located at far distances, while most often the public health-care doctors have knowledge regarding common viral's and influenza. The starting symptoms can be quite similar and this leads to wrong prediction of disease and treatment, and this causes late treatment of disease after non-effectiveness of treatment from prior. At times when particular disease is volatile, late treatment can lead to fatal death. With the help of data mining, data-sets of symptoms and algorithms we can achieve the predicting of disease before its too late and thus improve the future health conditions of people in rural as well as urban areas.

INTRODUCTION

India is one of the fastest growing economies in the world. Still the very basic components like proper nutrition, primary health-care, safe water, literacy, provision to knowledge of health issues, etc. are ignored. India stands on number 2 in the list of countries by population, that is equivalent to 17.7% of the total population in the world. The whole population of India is widely distributed with population density of 420 people per square kilometer. While this is not same overall, considering that India is very diverse country. The majority of the population of India still lives in rural areas. The population ratio as per the 2011 census shows that only 31.6% lives in urban areas, while 68.84% of the people still lives in rural areas. The condition of health of people in rural areas is far worse as compared to urban residents. The main reasons towards the low health ratio is due to non accessibility of health-care, lack of quality infrastructure, lack of awareness, less facilities, high illiteracy rate, high cost of treatment, etc. The Government of India have launched various Public Health-care Centers (PHCs), National Rural Health Mission, Rashtriya Bal Swasthya Karyakram (RBSK), National AIDS Control Organisation, etc. Inspite of all these efforts, due to improper implementation the right need is not been delivered to the rural people. The PHCs are very limited in the rural areas, while 8% of the centers do not have doctors or medical staff, 18% do not have pharmacist and 39% do not have lab technicians. According to the Economic Survey 2018-19, sixty percent of the PHCs have only one doctor to treat, while almost five percent have none. Only 20% of the PHCs are found to fit and follow the Indian Public Health Standards (IPHS). Even the required medicines and treatment are not fully available in the public health-care, while the Government of India spends only 1.4% of GDP on health. There are some private heath-care centers in rural areas, but they are located far away at distances and people prefer to visit them only when having major problem. People of the rural areas avoid private health, due to its remote location and high cost of treatment. The early symptoms of various diseases tend to be similar, and this can lead to confusion at the early stage of the treatment if not diagnosed properly. Due to inadequate availability of professional doctors and medical facilities in public health-care, much of the diseases are diagnosed too late when the disease have already start to affect the health. To overcome all these problems and to predict diseases prior to its increasing stage, so that it can help the patient to take better treatment from beginning if the disease is fatal. It saves the high expenditure on various tests and also saves time that is lost in wrong determination of diseases at early stage. The amounts of data gathered on various diseases and their symptoms are vast. With the help of data-set of various diseases and their symptoms, we can find new and interesting pattern in the data through data mining. The medical practitioners with the help of this technology can analyze the disease beforehand, give certainty about a disease, and helps to reduce the expensive early tests, medicines and treatment. This project proposes to build an application that can predict diseases with the help of data-sets, data mining and three different algorithm. To avoid confusion and guarantee accuracy, the three algorithms will be working parallel and the accuracy of the highest algorithm can lead to determining and predicting of diseases.

GOAL & OBJECTIVE

Goal - To predict diseases of patient at early stages.

Objective - To classify and predict diseases with the data provided in data-set of symptoms of various diseases with the assist of data mining and three different algorithms.

TECHNICAL KEYWORDS

- Classifier, Prediction, Data Mining, Naive Bayes, Decision Tree, Random Forest, Bagging, K-means

PROJECT IDEA

The application can help to classify and predict a particular disease with the data-set on various disease, it’s symptoms and knowledge about patient’s symptoms, achieving awareness, promoting good health and helping to save the overall expenditure.

MOTIVATION OF THE PROJECT

The lack of proper medical facilities, inadequate professional care, less knowledge about various diseases and their precautions lead to high negligence to health and often taking treatment too late when the disease have already started to influence the overall health of the patient.

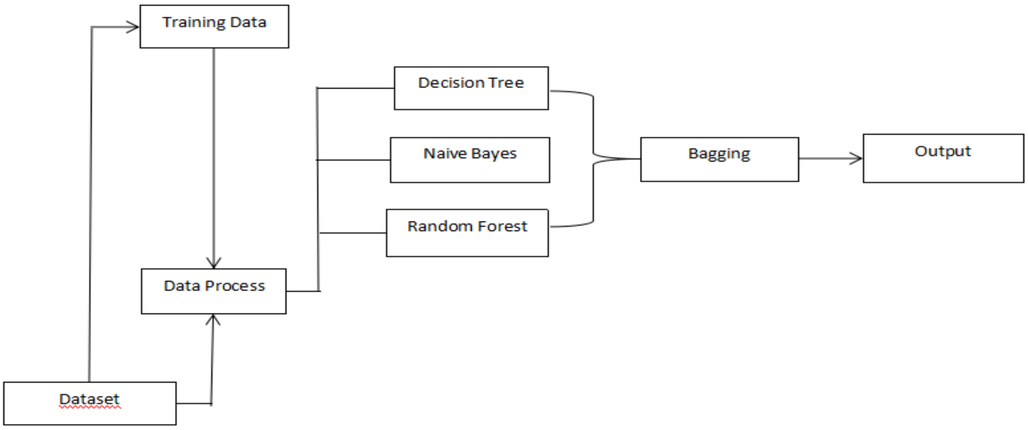
PROBLEM STATEMENT

Most of the population of India live in rural areas. Due to lack of public health-care system and also the high cost various tests to determine a particular disease in remote areas leads to the late diagnosis of a disease, which can be fatal. To ensure the classification and prediction of disease depending upon its symptoms and providing appropriate knowledge and further treatments.

PROJECT PLAN

Various types of algorithms and different techniques are used to study, classify and analyse data-sets to predict different types of diseases.The algorithms and techniques used in this project are - Decision Tree, Naive Bayes, Random Forest and Bagging. The data-sets consists of various different symptoms like itching, shivering, joint pain, anxiety, weight loss, dehydration, sunken eyes, headache etc. These symptoms are used to predict diseases with the help of the algorithms. Once all the algorithms have derived their respective results, than bagging is used to find the highest accurate algorithm and than present the output to the user.

SYSTEM ARCHITECTURE



METHODOLOGIES USED

1. Decision Tree:

Decision tree is a type of supervised learning data-mining algorithm. Decision tree is a popular algorithm used in data-mining due to its simple nature and easy to implement. Decision trees produce rules, which can be inferred by humans and used in knowledge system such as database[12]. The main intention to use decision tree is to develop training model, so that it can be used to forecast the class or value of target variable from training data[2].It is a decision based algorithm that uses different models of decisions and their outcomes, various consequences, to show a tree-like graphical representation of the output. The tree like structure is represented upside down with the root node at the top of the tree. Tree based supervised learning algorithm are considered as methods that predicts model with high precision, ease of interpretation and firmness. The “if-then-else” rules are used to build different rues to be used in the decision tree. Decision tree works for both categorical and continuous input and output variables, also they map non-linear relationship well. They follow Sum of Product (SOP) representation. They are most commonly used for classification and regression problems.

The decision tree consists of different parameters like -

Root Node - The root node in decision tree represents the overall sample from the data-set. It is than further divided into two or more homogeneous sets.

Internal Node - In decision tree, the internal node represents a parameter on an attribute. For example, a possibility whether a coin flip outcome is heads or tails.

Branch - The branch in decision tree represent the outcome of the test.

Leaf Node - In decision tree, each leaf node represents a class label. It consists of the decision taken after the computing of all the attributes.

The decision tree represents in a tree-like graph representation, here initially the tree decides on which types of features to choose and what type of conditions are to used for splitting. Depending upon the various features and categories the tree can grow more deeper for large data-sets. It often leads to over-fitting in the tree and causes problems. A decision tree can be used to classify an case by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance[14].

Splitting - Considering overall features and different attributes the splitting is done. The procedure consists of various attributes to do the splitting and calculate the accuracy of each split. The split having the least cost is chosen for next step. Pruning - Having large data-sets can sometimes leads to over-fitting the tree. To tackle this situation, the pruning method is used to increase the performance. Pruning involves removing the split branches that have feature with low importance. This helps to reduce the complexity as well as it also increases the probability of predicting right and reduces over-fitting.

2. Naive Bayes:

Naive Bayes is a type of supervised learning data-mining algorithm. It is a classification technique which is based upon Bayes' Theorem. It encompasses a family of simple "probabilistic classifiers" settled on applying Bayes theorem with strong straightforward independence expectation between the features[2]. Naive bayes assumes that all the attributes or features in the class are independent of each other. In probability theory, Bayes' law after relates the conditional and marginal probabilities of two random events[11].The presence of any feature or attribute in the class is unrelated to other features or attributes present in the class. One of the advantages of naïve bayes classifier is that it needs a small amount of training data to evaluate the parameters which is necessary for classification process[10]. This makes all the properties of the class independently of eachother and carries out their own assumptions and contributes to the probability[15]. It is an extension to Bayes algorithm that uses maximum posteriori decision rule in bayesian network, that makes classification more effective.

Bayes Equation -

Likelihood

Class Prior Probability

$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c\mid X)=P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

Where,

'c ' is target and 'x' is an attribute
P(c|x) - Posterior probability of class (c) given predictor (x)
P(c) - Prior probability of class

P(x|c) - Probability of predictor given class
P(x) - Prior probability of predictor

Assumption made while calculating Naive Bayes:

1. We presume that all the features present in the whole data-set are independent of eachother.
2. All the features in the whole data-set are considered as of same importance as others. This means that none of the attributes are irrelevant and it is assumed that each feature contributes equally to the output.

Naive Bayes is an extension to this Bayes Equation, where first the overall data-set is categorized in the form of frequency table. The data-set is partitioned into two divergent parts - Feature matrix and Response vector.

Feature Matrix - It is the combination of all the rows present in the overall data-set. Each vector consists of merit of the dependent features.

Response Vector - It contains the merit of class variable. The class variable is the prediction or the output that we are generating. These class variables are present for each row of feature matrix.

There are three kind of Naive Bayes model under the scikit-learn python library -

1. Gaussian Naive Bayes - It is associated with continuous values of each feature vector. It considers that the features follow a normal distribution.

2. Multinomial Naive Bayes - This is used for discrete counts and it is typically used for document classification.

3. Bernoulli Naive Bayes - This is mostly used where the features vectors are in binary form (1 and 0) instead of frequencies. Binary form determines whether a particular word occurs in a document or not.

3. Random Forest:

Random Forest is an add-on to the Decision Tree. The fundamental concept behind random forest is the wisdom of crowds. It is seen that large number of different unrelated trees (model) when operating together as a group will outrun any of the individual model. It is a type of ensemble learning algorithm that is used for classification and regression. It uses ensemble learning that solves a particular problem by creating multiple models for classification, regression and other tasks[2]. Random forest learning is affected by hyperparameters[16 17].Like its name suggest it consists large number of different decision trees that work as an ensemble, instead of simply averaging the prediction of trees this algorithm uses two important concepts to help it work -

1. While building trees erratic sampling of training data points must be done.
2. When splitting nodes erratic subsets of different features are considered.

Key concepts in Random forest-

1. Decision tree - Model that makes decisions upon various different feature values, having low bias and high variance often leads to overfitting.
2. Gini Impurity - It represents the probability such that the randomly selected sample from the node will be classified incorrectly from the distribution of samples in the node. This measure is used to reduce the splitting in each node.

3. Bootstrapping - It is the sampling of different sets of observation with replacement.
4. Random subsets of features - This is used for selecting a erratic set of features according to the breaks for each node in the decision tree.

During the time of training, each different tree in the random forest learns from random sample from the data-sets. Sometime the same samples will be used multiple times, as there is random selection of features. It is seen that the grater number of trees, the more precise the result. In decision tree we mostly face the problem of overfitting due to not specifying of the maximum depth of the tree. This is overcome through random forest as it generates many different trees and all the probabilities can be covered. The process of finding the root node and splitting of the feature nodes will be randomly selected and processed.

There are 2 stages while computing Random forest-

1. The creation of different trees in random forest
2. The prediction among these random forest created at first stage

Stage 1- The creation of random forest :

1. Erratic select 'k' features from total of 'm' features, such that k << m.
2. From these 'k' features, we need to determine node 'd' that makes the best split process.
3. Splitting of the child nodes using best split.
4. Repeat steps 1-3 until 'l' number of nodes have been found.
5. Building forests by repetitive steps 1-4 until 'n' number of times to create 'n' number of trees.

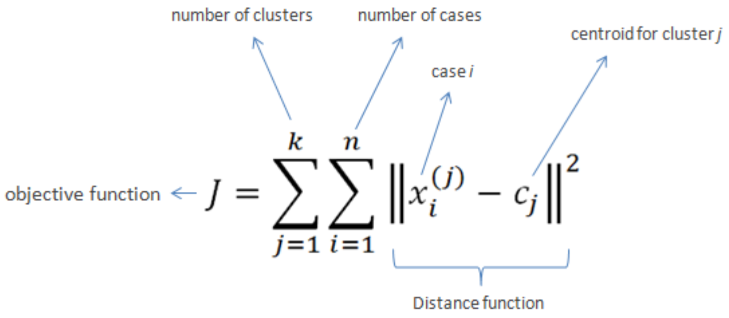
Stage 2- The prediction among random forest :

1. Makes the test features and uses the rules of each randomly created decision tree to predict the consequences and the predicted outcome is stored.
2. Calculate the votes for each predicted outcome.
3. The consequence with the soaring vote is considered as the final prediction from the random forest.

4. K-Means:

K-means is a type of iterative unsupervised algorithm that is mostly used for the purpose of clustering in problems. The k-means algorithm is an evolutionary algorithm that gains its name from its method of operations[12].Clustering has the job to identify and to distinguish various subgroups in the overall data-set. K-Means algorithm is a disruptive, non-hierarchical method of defining clusters[12]. They are assembled according to their various attributes that are similar in nature and are together grouped into clusters. While samples having different categorical are kept apart from each other. The clustering of different samples is always based on their features that are common between different samples. K-means clusters samples in such a way that each data point is a member of only one group. It focuses on deriving less variation within similar clusters so that we can achieve immense homogeneous data. K-means works with the aim to group numerous samples into small number of clusters such that each sample in the overall data-set belongs to anyone of the clusters depending upon its nearest mean.

K-means uses squared error function to minimize intra-cluster variance by the below formula:



Where,

|| xi-cj || - is the Euclidean distance between xi and cj
n - is the number of data points in ‘i’ cluster
k - is the number of cluster centers

The principle idea behind working of k-means is such that we need to define ‘k’ clusters having centers for each of them such that they are placed far away from each other. After the clustering of various groups it is time for each data point present in the data-set to be allocated to the nearest center.

Mathematical model of K-means:

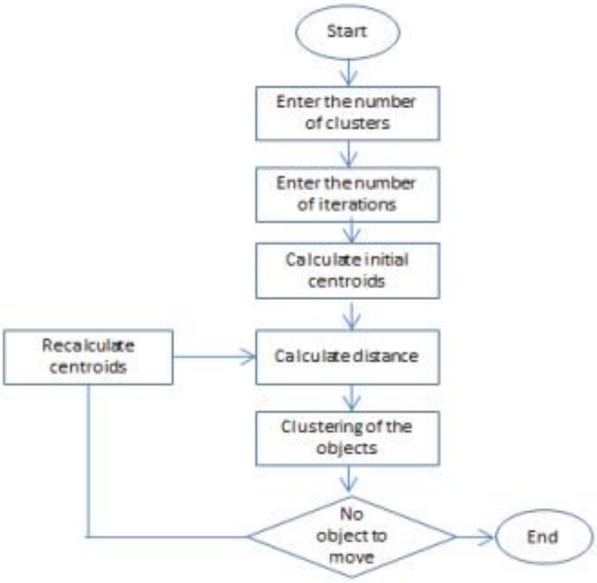

Start
Enter the number of clusters
Enter the number of iterations
Calculate initial centroids
Recalculate centroids
Calculate distance
Clustering of the objects
No object to move
End

Fig : Flowchart of K-Mean

Working of Diabetes Application via K-means:

ISSN: 2233-7857 IJFGCN

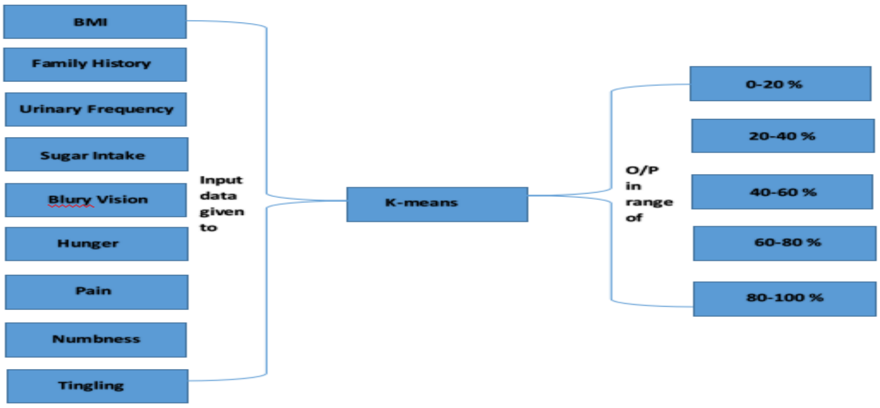


Fig: Working of Diabetes Application

Algorithm for K-means:

1. Decide the number of clusters to be created.
2. Initialize centroids for each cluster by shuffling the data-set and erratic selecting of data points for centroids without replacement.
3. Keep repeating until there is no change in the formation of centroids.
4. Calculate the addition of the squared distances between various data points and all the centroids.
5. Allot each data point to the closest center.
6. Calculate the centers for all the clusters by taking the average for all the samples in the data-set belonging to different clusters.

5 . Bagging:

Bagging is a type of machine learning ensemble algorithm. It is used to increase the reliability and precision of Machine learning algorithms that are been used for various different reasons like regression and classification. It is based upon the idea that by combining multiple algorithm models together we can derive or predict the best of them and produce the most powerful model. It often considers homogeneous weak learners, gains from them independently from each other in parallel and than it merge them into deterministic averaging process. It helps to make sure that there is reduction in variance and also it helps to avoid the problem of overfitting that arises. It is mostly applied on the decision tree methods. The main focus of bagging is to find a model that has less variance and compute accordingly. Bagging improves the classification by merging classification of randomly generated training sets. In different parallel methods the bagging considers by keeping the learners independent from each other and train them concurrently, this makes the model more robust and accurate as it considers and works on all the learners independently. The basic concept behind bagging is simple as " Learning from several independent models and finding an average from their prediction so that it can obtain a model that has the most accuracy of prediction and select the model with low variance." The problem arises when using it in practical, as there are a large data-set and numerous data is required. To cope with this situation multiple samples are created,which will than act as independent data-set from the true samples. This way it is possible to find out the samples that have less variance by adding a weak learner for each of the sample and than averaging all the samples such that we get average of their outputs. The standard technique produces 'm' training sets of 'D' of size ‘n’ by sampling ‘D’ evenly with replacement. It is a method that leads to the improvement of models that are unstable and to improve the overall efficiency and accuracy of the model.

Bagging also used in random forest to produce an output with lower variance. The biggest advantage of bagging method is that it can be computed to work parallel. This helps as various different models are used in such a way that they are applied independently from other models, thus promoting parallelism. The principle advantage of this execution is that it integrates the ordinance in it and all you need is to select good guidelines for the foundation of algorithms. Standardizing the models leads to discarding (or, at least, improvement) for the unstable models which can be derived from biased data.

CONCLUSION

The amount of data gathered regarding the different disease and their respective symptoms in health-care system are available in numerous amount. With the help of the data and using data mining algorithms we can find interesting patterns in the data-set and thus can derive new relations between various symptoms. The principle difficulty is to build a precise and efficient model using data mining algorithms in health-care. Each algorithm produces result according to its merits, attributes and its accuracy. These different algorithms produce non-identical results based on the parameters used by each algorithm. To further improvise the results generated by the different algorithms and to select the algorithm having highest accuracy among them we use the ensemble learning method- Bagging. With the help of bagging parallel computing can be done on all the algorithms such that they all are considered as independent to one another. This gives us output as to which algorithm generates highest accuracy of prediction for a particular disease and thus reducing the chances of wrong prediction. With the help of our proposed model we can predict diseases beforehand depending upon its symptoms of the patients and thus help to reduce the cost of treatment and various test. It also provides knowledge regarding the particular disease and its precautions to ensure safety of the patient. We can gain awareness among people living in rural areas, with lack of proper infrastructure, equipment and facilities.

REFRENCES

[1] Ajinkya Kunjir, Harshal Sawant, Nuzhat F. Shaikh, -Data Mining and Visualization for Prediction of Multiple Diseases in Healthcare 978-1-5090-6399-4/17/$31.00 2017 IEEE

[2] Alok Nagargoje, Anand Bohara, Vijay Kakade, Mrunal Kale, Prof. P.S.Hanwate, -A survey on prediction of diseases through data mining,International Engineering Journal For Research & Development, Volume 5, Issue 1, E-ISSN NO:-2349-0721, 2020

[3] K.Manimekalai, ―Prediction of Heart Diseases using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 2, ISSN(Online):2320- 9801, ISSN (Print):2320- 9798, February 2016.

[4] Theresa Princy R, “Human Heart Disease Prediction System using Data Mining Techniques”, International Conference on Circuit, Power and Computing Technologies [ICCPCT], IEEE (2016)

[5] Dr. S. Vijayarani, Mr.S.Dhayanand, “Kidney Disease Prediction using SVM and ANN algorithms”, International Journal of Computing and Business Research (IJCBR), Volume 6, Issue 2, ISSN (Online):2229-6166, March 2015.

[6] Vivekanandan T et al., “Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease”,

www.elsevier.com/locate/compbiomed, https://doi.org/10.1016/j.compbiomed, Pages: 125-136 (2017)

[7] Polat K., Sahan S., Gunes S., "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing", ScienceDirect, 2007.

[8] Parvathi I, Siddharth Rautaray, ―Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain, International Journal of Computer Science and Information Technologies, Vol. 5 (1), 838-846, ISSN: 0975- 9646, 2014

[9] Vahid Rafe, Roghayeh Hashemi Farhoud, ―A Survey on Data Mining Approaches in Medicine, International Research Journal of Applied and Basic Sciences, Vol 4 (1), ISSN 2251-838X, 2013.

[10] T. Revathi, S. Jeevitha, ―Comparative Study on Heart Disease Prediction System Using Data Mining Techniques, Volume 4 Issue 7, ISSN (Online): 2319-7064, July 2015.

[11] Devendra Ratnaparkhi, Tushar Mahajan, Vishal Jadhav, ―Heart Disease Prediction System Using Data Mining Technique, International Research Journal of Engineering and Technology (IRJET), Volume: 02

[12] M.A.Nishara Banu , B Gomathy, -DISEASE PREDICTING SYSTEM USING DATA MINING TECHNIQUESInternational Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Volume 1, Issue 5 (Nov-Dec 2013), PP. 41-45

[13] K. Aparna, Dr. N. Chandra Sekhar Reddy, I. Surya Prabha, Dr. K. Venkata Srinivas, -Disease Prediction in Data Mining TechniquesIJCST Vol. 5, Issue 2, April - June 2014

[14] K.Gomathi, Dr. D. Shanmuga Priyaa, -Multi Disease Prediction using Data Mining Techniques Article can be accessed online at http://www.publishingindia.com

[15] Shwetambari Kharabe, C. Nalini," Robust ROI Localization Based Finger Vein Authentication Using Adaptive Thresholding Extraction with Deep Learning Technique", Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 07-Special Issue, 2018.

[16] Shwetambari Kharabe, C. Nalini," Using Adaptive Thresholding Extraction - Robust ROI Localization Based Finger Vein Authentication", Journal of Advanced Research in Dynamical & Control Systems, Vol. 10, 13-Special Issue, 2018.

[17] Shwetambari Kharabe, C. Nalini," Evaluation of Finger vein Identification Process", International Journal of Engineering and Advanced Technology (IJEAT), International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6S, August 2019.