Chronic Kidney Disease Prediction using K-Means Algorithm

D.Charan[1] Mr.G.Ananthanath[2]
[1]Student [2]Assistant Professor
[1,2]Department of Computer Applications
[1,2]KMM institute of PG Studies, Tirupati, India

Abstract— The procedure of finding hidden and unidentified patterns and trends in big datasets, extracting information from them and building predictive models is defined as data mining. In other word, it's the process of the collection and exploration of data sets and building models by huge data stores to expose previously unknown outlines. Healthcare management is one of the areas which is using machine learning techniques broadly for different obejectives. Chronic kindney disease is agrowing disease in recent years and many researches are being done to predict its progression and classify the datasets based on related features. In this paper, we focus on applying different machine learning classification algorithm to a dataset with 400 observations and 24 attributes for diagnosis tree, linear regression, super vector machine and neural network.

Key words: Chronic Kidney Disease, K Means Algorithm, Logistic Regression, Support Vector Machine

I. INTRODUCTION

The procedure of finding hidden and unidentified patterns and trends in big datasets, extracting information from them and building predictive models is defined as data mining. In another word, it's the process of collection and exploration of data sets and building models by huge data stores to expose previously unknown outlines. Due to complexity and vagueness of data engendered by healthcare transactions, it is impossible to analyze them with traditional tools. In order to make the decision-making process easier and more trustable, data mining techniques are provided to transmute these data into useful information and make it feasible to get useful results and patterns and trends out of these huge amounts of data. Data mining has been widely used in many areas. One of these areas which is using it even more and more as an essential tool is healthcare management. All agents in a healthcare industry can significantly benefit Data mining applications. Data mining is not new. It has been used intensively and extensively by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling. A heart disease prediction is done using three data mining techniques namely Neural Network, Decision Tree and Naive Bayes. Their results disclose that a neural network with 15 features have surpassed two other techniques and accordingly is selected as the predictive model.

II. RELATIVE STUDY

A. An Analysis of Heart Disease Prediction using Different Data Mining Techniques.

The healthcare industry collects large amounts of Healthcare data, but unfortunately not all the data are mined which is required for discovering hidden patterns and effective decision making. We propose efficient genetic algorithm with the back propagation technique approach for heart disease prediction. This paper has analyzed prediction systems for Heart disease using more number of input attributes. The System uses medical terms such as Gender, blood pressure, cholesterol like13 attributes to predict the likelihood of patient getting a Heart disease.

B. Intelligent Heart Disease Prediction System using Data Mining Techniques.

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not ";mined"; to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer complex "; what if"; queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. IHDPS is Web-based, user-friendly, scalable, reliable and expandable. It is implemented on the .NET platform.

C. Intelligent Heart Disease Prediction System using Data Mining Techniques.

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not ";mined"; to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer complex ";what if"; queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. IHDPS is Web-based, user-friendly, scalable, reliable and expandable. It is implemented on the .NET platform.

III. PROPOSED ALGORITHM

The essential point of that is to interrupt down the use of records digging in medicinal space for forecast of incessant kidney illness. In the medicinal offerings territory countless kidney illness may be extraordinarily very lots expected utilizing information mining structures

A. Algorthims

1) K-means Algorithm

The number one point is one of the minimum complex unsupervised picking up statistics of figuring that good deal with the outstanding batching problem. The method seeks after a trustworthy and primary system to status quo a given enlightening accumulation via a particular amount settled apriority. The essential concept is to speak to alright centers, one for every organization. These centers need to be set shrewdly in light of various zone motives various effects. As such, the better preference is to position them however a horrendous parcel as could be foreseen a protracted way from each other. The accompanying degree is to take each guide having a vicinity in the direction of a given instructive amassing and accomplice it to the nearest attention. At the point while no aspect is pending, the underlying develop is executed and an early accumulating age is executed. Presently we want to re-discover all right new centroids as boycott acknowledgment of the clusters coming kind of due to the beyond increment. After we've got those k new centroids, every other coupling must be finished amongst similar instructional association centers and the nearest new cognizance. A circle has been made. Due to this circle we may additionally likewise take a look at that they all right facilities change their vicinity all round asked until the element that no greater noteworthy changes are finished or on the forestall of the day facilities don't stream any extra. At giant final, this figuring goes for proscribing goal limits recognize as squared botch work given with the manual of:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

Where

'||xi - vj||' is the Euclidean separation among xi and vj.
'ci' is the amount of records focuses in its organization.
'C' is the amount of organization focuses.

B. Algorithmic steps for k-means clustering

Let X = x1,x2,x3,… … ..,xn be the arrangement of information focuses and V = v1,v2,… … .,vc be the association of focuses.

1) Randomly pick 'c' group focuses.
2) Calculate the separation among every datum factor and bunch focuses.
3) Assign the information factor to the bunch awareness whose break free the organization focus is least of all of the organization focuses.
4) Recalculate the new bunch awareness making use of:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

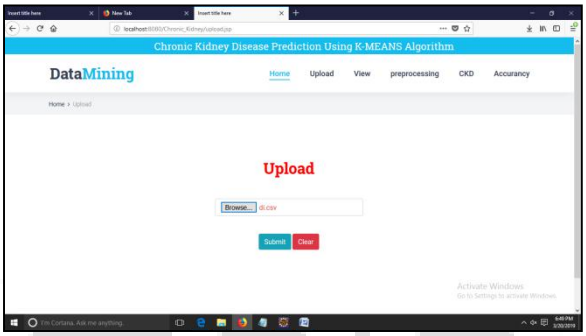
Where, 'ci' speaks to the amount of information focuses in ith group.

5) Recalculate the separation among every datum point and new acquired group focuses.
6) If no records point was reassigned then prevent, typically rehash from step 3).

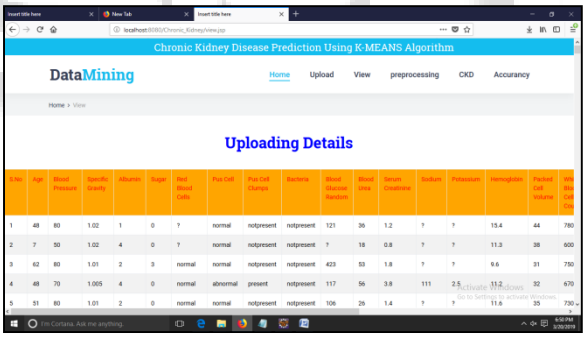
IV. RESULTS

A. Home:

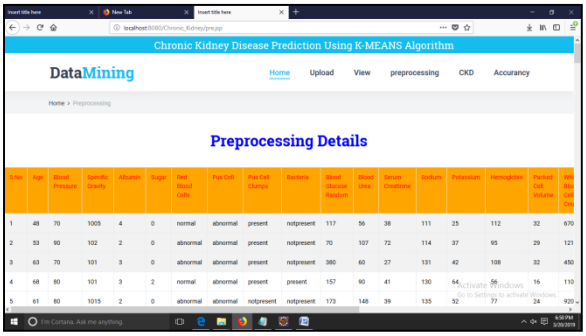
B. Upload:



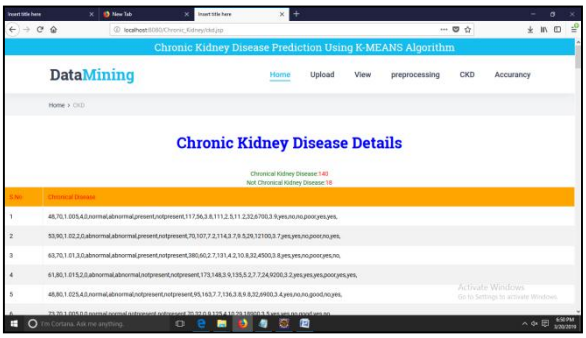
C. View:



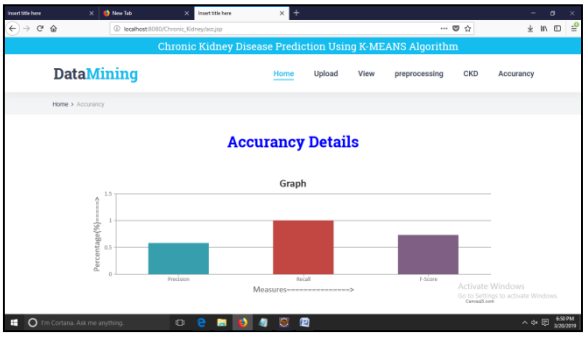
D. Preprocessing:



E. Chronic Kidney Disease:



F. Accuracy:



V. CONCLUSION

In this paper first the data set containing 400 samples and 24 features was selected from UCI data base and preprocessing was done to remove noisy and unreliable data. In order to do so, first missing value is filled via mean for nominal features and filled via mode for categorical features. Then, dataset has been normalized to have a unit scale for all data. The correlation matrix of features is obtained and it is observed that features are highly correlated to each other. Classification has been done in three stages. In the first stage, two L1-based feature selections is done for different values of controlling factors and accordingly different number of features are selected. Next performance of five different techniques, namely, DT, NN, LR, SVM and NB in classifying both original and normalized data based on their AUC is compared. In the third step, by using classification is done for all features of original and normalized data set and the performance of classifiers is compared by their sensitivity, specificity, accuracy and AUC. The aim is to analyze the results and see the importance of features on the classification results. Results show that, first except of NN which is sensitive to scale of data, performance of other classifiers are almost the same for original and normalized data set. Second, same results are obtainable using 8 or 9 features instead of 24 features which validate the results correlation matrix which shows high correlativity between the features.

REFERENCES

[1] Milley, A. (2000). Healthcare and data mining. Health Management Technology, 21(8), 44-47.
[2] Koh Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.
[3] Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., & Meyer, D. (2016). Fusing Data Mining, Machine Learning and Traditional Statistics to Detect Biomarkers Associated with Depression. PloS one, 11(2), e0148195.
[4] SA, S. (2013). Intelligent heart disease prediction system using data mining techniques. International Journal of Healthcare & Biomedical Research, 1, 94-101.
[5] Boukenze, B., Mousannif, H., & Haqiq, a. predictive analytics in healthcare system using data mining techniques. Computer Science & Information Technology, 1.
[6] Vijayarani, S., Dhayanand, M. S., & Phil, M. (2015). Kidney disease prediction using svm and ann algorithms‖. International Journal of Computing and Business Research (IJCBR) ISSN (Online), 2229-6166.
[7] Sharma, S., Sharma, V., & Sharma, A. (2016). Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis. arXiv preprint arXiv:1606.09581.
[8] UCI Machine Learning Repository: Chronic_Kidney_Disease Data Set. (n.d.). https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
[9] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36(4), 2431-2448.
[10] Mitchell, T. M. (1997). Machine Learning.
[11] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.
[12] Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.

