# HEALTHCARE PREDICTIVE ANALYTICS

**M. Shanthipriya[1], Dr. G. T. Prabavathi[2]**

*[1]M.Phil Research Scholar, Dept. Of Computer Science, Gobi Arts & Science College, Gobichettipalayam, TamilNadu, India*
*[2]Associate Professor, Dept. Of Computer Science, Gobi Arts & Science College, Gobichettipalayam,TamilNadu, India*

---

**Abstract -** *Healthcare analytics is a systematic use of information and similar clinical and business insights development through applied analytical disciplines. Healthcare analytics finds insights from unstructured complex and noisy healthcare records of patients for making better health care decision. Predictive analytics is used for future prediction and forecasting based on an existing patterns and trends, and also used to identify the current trends, calculate the probability of pattern occurrences and reduce uncertainties. In this research, a study has been made on various researches on predictive analytics that deals with prediction of diseases such as Asthma, Cancer, Kidney, Heart and Diabetes. Synthetic diabetes data is taken for analyzing and used for prediction. The clustering algorithms K-Means, Hierarchical and Density-Based algorithms are used to evaluate the accuracy of the prediction.*

***Key Words*: Big data, Healthcare Analytics, Predictive Analytics, Disease Prediction, Clustering.**

## 1. INTRODUCTION

Big data plays an important role in the field of healthcare. Big data have characteristics like volume, velocity, variety, value and veracity.Big data analytics is the process of examining huge and different data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions (Priyanka et al., 2014). Healthcare analytics as the "systematic use of data and relatedclinical and business insights developed through applied analytical disciplines such as statistical, contextual, quantitative, predictive, and cognitive spectrums to drive fact-based decision making for planning, management, measurement and learning". The major objective of healthcare analytics is to "help people to make and execute rational decisions, defined as being data driven, transparent, verifiable and robust" (Raghupathi et al., 2013).

Predictive Analytics is the use of existing data to estimate on consumer behavior and trends. Predictive analytics can also define process that uses machine learning to analyze data and make predictions. Predictive analytics supports medical applications to achieve a high level of useful complete care and preventive care, as predictive systems' results allow treatments and actions to be taken when all the risks are recognized in early stages, which aids for reducing costs (Conley et al 2008).

The research work describes the prediction of diseases using various clustering techniques. Chapter 2 provides detailed literature review on various chronic diseases like asthma, cancer, heart diseases and diabetes. Chapter 3 provides the methodology and detailed description about data sets. The results and discussions of this research work presented in Chapter 4. Chapter 5 presents the conclusion and future enhancements of this work.

## 2. REVIEW OF LITERATURE

ArchananaBakare et al. [1] introduced a new novel method introduced which collect results and prediction data form many social networking sites like twitter, Facebook, Google search. They discussed a generalized method which predicts the fore-coming strokes of various diseases depending on data gathered from social networking sites. The proposed model can predict the number of diseases stroke visits based on near-real-time environment and social media data with approximately 70% precision. A multirank algorithm is used for prediction of asthma. The results showed that 80% accuracy can be achieved in prediction diseases.

RostomMennour et al. [16]used Machine learning algorithms on top of MapReduce and Mahout in order to pre-filter the huge set of ligands to effectively do virtual screening for the breast cancer protein receptor. The authors explored five Machine Learning algorithms to do the classification of ligands into dockable and non-dockable ones. After that, we have selected three algorithms Multilayer Perceptron, Naive Bayes and Random forests are proved that these algorithms produce good results. The obtained results have shown that we can reach about 80%.

Bojana et al. [4]described the practical application of data mining methods for estimation of survival rate and disease relapse for breast cancer patients. The model based on ANN algorithm with 20 neurons in hidden layer achieved the best accuracy (0.836). However, sensitivity of this model is 0.595 which represent a low value to account it as a reliable classifier. Decision Tree algorithm has achieved the best performance in classification task. These results have created an opportunity to further investigate this field in order to implement techniques which can additionally improve results and to implement and evaluate the methodologies for estimating reliability of individual predictions.

Jen Chih-Hung et al. [5] also investigated chronic diseases such as diabetes, disease of the liver, hypertension cardiovascular disease, and renal disease. The data that was used in the study was collected from a Taiwan medical center. The analysis revealed that the screening of important factors and classification of materials were all effective at a higher rate (corrected rate of 80%). From the analysis results, using the early-warning system can obtain the best classification rate for evaluations in chronic illnesses and the K-NN method is suitably applied to the classifications of clinical diseases as well. The classification result of the K-NN is used to construct the early-warning criteria for evaluating different chronic illnesses.

BasmaBoukenze et al. [3] presented an overview on the evolution of big data in the healthcare system and using Decision Tree (C4.5) algorithm for predicting chronic kidney diseases. The authors applied C4.5 to make classification and prediction on a database to extract knowledge and classify patients into two categories: chronic kidney disease (ckd) and not chronic kidney disease (notckd). The authors proved that the C4.5 classifier is powerful, according to the number of correctly classified instances (396) and just 4 instances misclassified. This is mentioned with a low error rate (0.37). The results displayed that this classifier proves its performance in predicting with best results in terms of accuracy and minimum execution time.

Mounika et al. [11] proposed a predictive analysis of diabetic treatment using classification algorithms such as Naive bayes, OneR and ZeroR. The dataset contains the 8 parameters such as age, diet, smoke, type, drug, insulin, obesity and level; these parameters are used to predicting the blood glucose level of patients. They analysed the error rates and performances of different classification algorithms and the results showed that Naive Bayes is faster than ZeroR and OneR for analysing performance and error rates.

ArunCullotta [2] proposed a method based on analysis of messages posted on the social media website such as twitter to determine if a homogeneous connection can be uncovered Analyzing user messages in social media can measure different population features, including public health measures. The author suggested that supplying the classifier with more sophisticated linguistic features will improve its accuracy. Perhaps more importantly, given the informal syntax and number of spelling mistakes in Twitter messages, it is likely that more sophisticated pre-processing could improve the quality of analysis.

Nithya et al. [12]used diabetic dataset for comparing the performance of three clustering algorithms: K-Means Clustering, Hierarchical Clustering and Density based Clustering algorithms. The comparison of those clustering algorithms based on the performance of execution time and the number of clustered instances. The diabetic dataset is collected from UCI repository. From the experimental results it is inferred that by using the training set parameter for the hierarchical clustering algorithm number of clustered instances in the group and the execution time is higher than the density based clustering and k means clustering algorithms. From the experimental results it is inferred that the K-Means algorithm performs better as compared to the hierarchical clustering and density based clustering. Various healthcare predictive analytics research review has been done by Prabavathi et al.[13].

## 3. METHODOLOGY

Data collection plays a vital role for obtaining accurate results for any research work being carried. The accurate results depend highly on the nature of the collected data. The data is normally noisy and variant. The dataset used for this research work has been taken from two sources: The diabetes dataset from synthetic data set created by collecting data from students within particular age groups in a specific geographical region.

For this research work synthetic data set has been created by collecting the data from a specific group of college students. Data is collected from college women students in rural areas. The data set contains 1560 instances and 16 attributes. The attributes used for data analyzes are: Age, Family History, Overweight, Blurred Vision, Intake of More Sweets, Excessive Thirst, Feeling of Urination Frequently, Heal Slowly, UTI Symptoms, Have PCOs Problem, Unexplained Weight Loss, Feet and Hands Tingle, Taken Diabetes Check-up, Bumpiness Sensation, Exercise Regularly and Regular Walking.

The dataset needs to be pre-processed for removing the noisy data and convert it into meaningful data. The pre-processing is the first step for analyzing the results. Since a synthetic data as heterogeneous in nature and due to inconsistencies, pre-processing is a challenging task. The dataset is first loaded into in the file system and the table is created with attributes present in data collection. After pre-processing the data is ready for predictive analysis.

### Filtering

The dataset was collected only from girl students who belong to age group 17-21, 22-25 and 25-27. Assigning appropriate values to identify the importance of each attribute in the diabetes dataset is more important to analyze the dataset. The attributes are assigned 1 if the response is Yes and 0 if the response is No. All the attributes are of Yes/No type except exercise attribute. The exercises regularly attribute has 3 values: 4, 5 and 6. The value 4 denotes the students do exercise regularly, 5 means do exercise weekly and 6 for those who do not do exercise.

### Conditions

***Positive Criteria***

1. Diabetes Family = 0 AND Do Regular Exercise = 4 AND Go Regular Walking = 1

2. Take More Sweets = 0 AND Heal Slowly = 0 AND Overweight = 0 AND Go Regular Walking = 0

3. Diabetes Family = 1 AND Do Regular Exercise = 4 AND Go Regular Walking = 1 AND Feel Excessive Thirst = 0 AND Overweight = 0

4. Overweight = 0 AND Weight Loss = 0 AND Have PCO Problem = 0 AND Go Regular Walking = 1

5. Diabetes Family = 0 AND Blurred Vision = 0 AND Experienced in Urinary Tract Infection = 0 AND Do Regular Exercise = 1

***Negative Criteria***

1. Diabetes Family = 1 AND Do Regular Exercise = 6 AND Go Regular Walking = 0

2. Diabetes Family = 1 AND Take More Sweets = 1 AND Heal Slowly = 1

3. Diabetes Family = 0 AND Take More Sweets = 1 AND Feel Excessive Thirst = 1 AND Do Regular Exercise = 1

4. Overweight = 1 AND Weight Loss = 0 AND Have PCO Problem = 1 AND Go Regular Walking = 0 AND Experienced in Urinary Tract Infection = 1

5. Diabetes Family =1 AND Blurred Vision = 1 AND Do Regular Exercise = 5 AND Go Regular Walking = 0 AND Feel Urinate Frequently = 1

If the above conditions are satisfied in filtering, that particular instance has a probable chance for occurrences of diabetes.

## 4. RESULTS AND DISCUSSIONS

### SYNTHETIC DATA RESULT

The comparative analysis of diabetes dataset is used to predict the better algorithm. The Accuracyof the compared clustering techniques is depicted in Table 4.1

**Table 4.1 Comparing Accuracy using Clustering Techniques**

| Algorithms | No of clusters | Cluster A | Cluster B | Cluster A (%) | Cluster B (%) |
|---|---|---|---|---|---|
| K-Means | 2 | 956 | 604 | 62 | 38 |
| Hierarchical | 2 | 1034 | 526 | 67 | 33 |
| Density Based | 2 | 1198 | 362 | 76 | 24 |

From the above tables, the experimental measures are evaluated by using accuracy and execution time factors. The result shows k-means algorithm is better than hierarchical and density based clustering, based on the number of cluster accuracy and the minimum execution time. In this research work, the predictive measures are calculated using the factors clustering accuracy and clustering execution time. The datasets are compared using three clustering technique to predict more suitable algorithm. The accuracy and execution time measures for the compared algorithm are represented in Figure 4.1 and 4.2

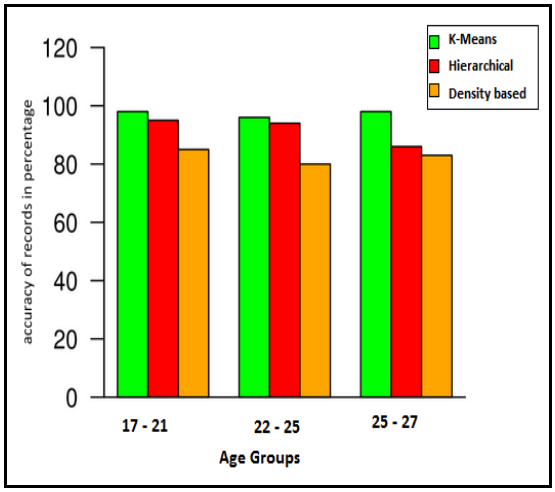


**Fig-4.1 Prediction Accuracy**

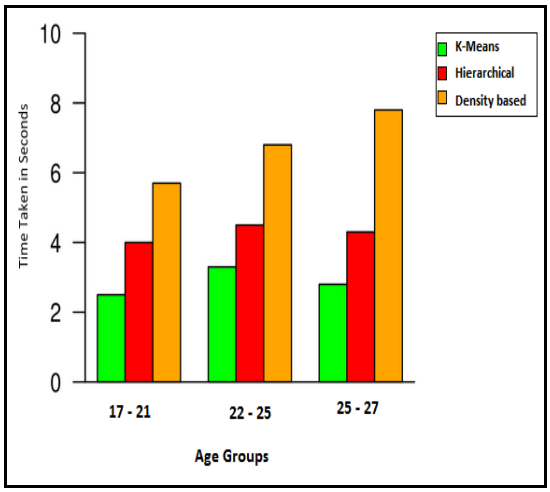


**Fig-4.2 Processing Time**

Figure 4.1 gives the accuracy of the disease occurrence prediction for age groups (17-21, 22-25 and 25-27) using *K*-Means clustering, Hierarchical clustering, Density based clustering. The execution time measures for the compared algorithm are shown in Figure 4.2. From the above chart, K-means algorithm takes minimum time to process the synthetic dataset. Hierarchical and density based clustering time taken is higher than the k-means clustering. K-means algorithm gives better accuracy in minimum time with more number of records.

## 5. CONCLUSION AND FUTURE WORK

In this research work, clustering algorithms are evaluated to predict which algorithm is more suitable for diabetic datasets. For predicting the health-care, synthetic datasets have been used in this work. Proper pre-processing and prediction filtering conditions are used to enhance the quality of data and reduce the size of dataset. For predicting the possibility of occurrence of diabetes, many conditions have been applied for finding better accuracy. The algorithms are compared and it has been identified that *K*-Means clustering Algorithm produces comparatively better prediction than Hierarchical clustering and Density Based clustering algorithms. In future, k-means algorithm can be modified for producing more efficient and accurate results. Further various machine learning algorithms for incrementing the accuracy of prediction can also be done.

## REFERENCES

[1] ArchanaBakare M, R. V. Argiddi : " Prediction of Disease using Big Data Analysis", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2016.

[2] ArunCulotta: "Towards Detecting Influenza Epidemics by Analyzing Twitter Messages", 1st Workshop in Social Media Analytics (SOMA '10), July 25, 2010.

[3] Basma Boukenze1, HajarMousannif and AbdelkrimHaqiq: "Predictive Analytics in Healthcare System using Data Mining Techniques", CCNET-2016, pp. 01–09, 2016.

[4] Bojana R. AndjelkovicCirkovic, Alesksandar M. Cvetkovic, Srdjan M. Ninkovic, and Nenad D. Filipovic : " Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients".

[5] Chih-Hung Jen, Chien-Chih Wang, Bernard C Jiang, Yan-Hua Chu and Ming-Shu Chen. Application of classification techniques on development and early-warning system for chronic illnesses. Expert Systems with Applications, 39 (10) : 8852-8858, 2012.

[6] Frost & Sullivan: Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations. http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technologycomplexities-ar.pdf.

[7] Harshit Kumar and Nishant Singh: "Review paper on Big Data in healthcare informatics", International Research Journal of Engineering and Technology (IRJET), Vol. 4, Issue. 2, February 2017.

[8] Linda A: "Seven ways Predictive analytics Can improve Healthcare", Elsevier, October 2016.

[9] Lidong Wang and Cheryl Ann Alexander: " Big Data in Medical Applications and Health Care ", Americal Medical Journal, Vol. 6(1), 2015.

[10] Madhura A. Chinchmalatpure, Dr.Mahendra P. Dhore: "Review of Big Data Challenges in Healthcare Application ", IOSR Journal of computer Engineering , e-ISSN: 2278-8727, e-ISSN : 2278-8727, PP. 06-09.

[11] Mounika M, S. D. Suganya, B. Vijayashanthi, S. Krishna Anand: "Predictive Analysis of Diabetic Treatment Using Classification Algorithm ", International Journal of Computer Science and information Technologies, Vol. 6(3),2015.

[12] Nithya R, P. Manikandan, D. Ramyachitra : " Analysis of Clustering Technique for the diabetes dataset using the training set parameter", International journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2015, ISSN : 2278-1021.

[13] Prabavathi G.T. and M. Shanthipriya: "Review of Healthcare Informatics", International Journal of Innovative Research in Computer and Communication Engineering, Volume 5(7), ISSN: 2320-9801July 2017.

[14] Priyanka K, NagarathnaKulennavar: "A Survey on Big Data Analytics in Health Care", International Journal of Computer Science and Information Technologies, Vol. 5(4), 2014, 5865-5868.

[15] RaghunathNambiar, Adhiraaj Sethi, Ruchie Bhardwaj, Rajesh Vargheese: "A Look at Challenges and Opportunities of Big Data Analytics in Healthcare", IEEE International Conference on Big Data, 2013.

[16] RostomMennour and Mohamed Batouche: "Drug Discovery for Breast Cancer Based on Big Data Analytics Techniques".