Diabetes Impacted Cardiovascular Disease Prediction using Machine Learning

C.Akash Mahadevan, S. Kanishka, Saisurya. S, V. Arun

Abstract: Utilizing big data growth in biological and health communities, an accurate analogy of medical data can benefit the detection of diabetes impacting cardiovascular diseases. Using k-Means clustering (kMC) algorithm for structured data of heart disease patients, we narrow down to cardiovascular diseases impacted by diabetes. To our knowledge, none of the previous work focused on predicting heart diseases specifically for diabetes patients. Contrasted to multiple other prediction algorithms, the accuracy of predicting in our proposed algorithm is faster than that of other prediction systems for cardiovascular diseases.

Keywords: Cardiovascular diseases, Diabetes, Prediction.

I. INTRODUCTION

Diabetes is a major chronic disease which occurs due to increase in blood glucose or sugar levels. It is a disease that is drastically increasing during the recent times. Blood glucose is the body's source energy which is acquired from the food we eat. Insulin, a hormone secreted by pancreas helps the glucose from food to be used for energy. Diabetes has no cure but steps can be taken to manage and stay healthy. The World Health Organization (WHO) estimated that the number of diabetic patients has increased from 108,000,000 in the year 1980 to 422,000,000 in the year 2014. It was estimated that around 1.6 million deaths were caused by diabetes and was the seventh leading cause of death in 2016.Type 1 and type 2 diabetes have the highest percentages of occurrence. Type 2 diabetes is the most prevalent diabetes. Cardiovascular diseases are diseases that damage the heart or its vessels. Hypertension or commonly known as high blood pressure is one of the main causes of CVD's. It is estimated that around 17.9 million people died due to CVD's, making up 31% of all global deaths in 2016. People with diabetes are more likely to die from cardiovascular diseases. Early detection and management using appropriate counselling and medicines can prevent the high cardiovascular risk.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author

C.Akash Mahadevan, B. Tech CSE, SRM Institute of Science and Technology, Chennai, India. Email: akash200400@gmail.com

S.Kanishka, B. Tech CSE, SRM Institute of Science and Technology, Chennai, India. Email: skanishka885@gmail.com

Saisurya.S, B. Tech CSE, SRM Institute of Science and Technology, Chennai, India. Email: saisuryas0112@gmail.com

V.Arun, CSE, SRM Institute of Science and Technology, Chennai, India. Email: arunpro3284@gmail.com

II. SYSTEM ARCHITECTURE

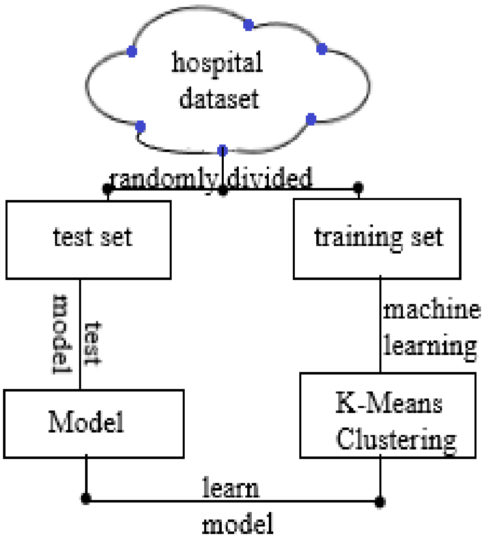

Fig 1: Proposed system architecture

The hospital data used comprises of real-life data which is stored in the data center. The dataset is fed as an input which is divided as two datasets, namely training and test set.

The training set is written manually where the model should follow the same definitions given in the training set. The test set is where we apply our model and test it to get the outcome. The training set is processed using an algorithm known as K-Means Clustering algorithm. The ideology is to define k centers, one center for each cluster. Then, take each point in a given dataset and group it to the nearest center. After we obtain the k-centroids, a binding is done among the same dataset points and the nearest new center.

This algorithm targets to minimize a function known as objective function which is thereby known as squared error function which is given by:

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

Fig 2: Formula for kMC

C is the total number of clusters present in the system

||xi-vj|| denotes the Euclidean distance between the two points xi and vj respectively.

Ci denotes the ith cluster of the system.

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

Fig 3: Formula for kMC-2

Recalculation of the data points is necessary and to be done using the formula mentioned above.
Here, ci denotes the number of data points in the ith cluster.
An example of the result is given below,

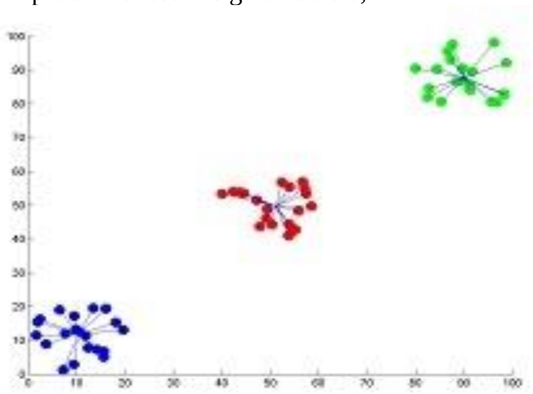

Fig 4: Example of a K-means Graph

III. EXISTING SYSTEM

This system utilizes the structured data and text data to multi-dimensionally meld the organized information and unstructured information to anticipate whether the patient is at high-danger of cardiovascular malady affected by diabetes. For unstructured, they select the highlights consequently utilizing CNN algorithm. The system proposed is one of the latest CNN-based multi-model disease risk prediction (CNN-MDRP) algorithm.
The drawbacks of the existing system are:

1. Accuracy of prediction of cardiovascular diseases is low.
2. In this system the set of data is considerably smaller, for diseases with detailed conditions, the characteristics are chosen through experience. However, these pre-chosen characteristics may not satisfy or affect the changes in the disease and its influencing factors.
3. This system has lower accuracy and is more time consuming.

IV. PROPOSED SYSTEM

For organized information, we use K-Means Clustering algorithm. It is frequently chosen in light that it is anything but difficult to utilize, simple to prepare, and simple to decipher the outcomes. It is regularly utilized in scout applications when you are attempting to discover non-distinguished items.
The advantages of the proposed system are:

1. Prediction of Diabetes impacted cardiovascular diseases.
2. Light weight algorithm and more efficient in deduction of disease.
3. The algorithm requires lesser datasets than the existing systems algorithm which makes the algorithm to learn faster and takes lesser time to increase experience with the data set.

V. CHALLENGES AND LIMITATIONS

The algorithm that is proposed has multiple advantages but also has a couple of challenges and limitations. Some of the limitations are:

1. The algorithm requires precise specification of the clusters to be able to learn properly.
2. If there is an existance of 2 or more overlapping data, the algorithm will not be able to distinguish between the different clusters.
3. The Euclidean distance that is measured will weigh in underlying factors that are not necessary and may influence the result.
4. It cannot manage or handle noisy or outlying data and it fails for all non-linear data sets.

VI. RESULT AND DISCUSSION

The proposed system is to implement a program to predict cardiovascular diseases impacted by diabetes using the K-Means Cluster algorithm.
The algorithm used here helps to predict the cardiovascular diseases based on the input given by the user. The algorithm compares the input data and the different means of the algorithm and categorizes the input to the specified cluster. It then narrows down the result from the input given and states the final output.

VII. CONCLUSION

The concluded aim is to provide the user with an accurate result of the prediction of cardiovascular diseases impacted by diabetes.

REFERENCES

1. “The’big data’revolution in healthcare: Accelerating value and innovation,” P.Groves, B. Kayyali, D. Knott, and S. V. Kuiken..
2. “Big data: A survey,” M. Chen, S. Mao, and Y. Liu.
3. “Mining electronic health records: towards better research applications and clinical care,” P.B.Jensen, L. J. Jensen, and S.Brunak..
4. “A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics," D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi.
5. “Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System,” M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, C.Youn.
6. “Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring”, M. Chen, Y. Ma, J. Song, C. Lai, B. Hu.
7. “Emotion Communication System,” M. Chen, P.Zhou, G.Fortino.
8. “Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems,” M.Qiu and E.H.M. Sha.
9. “Enabling real-time information service on telehealth system over cloud-based big data platform,” J.Wang, M.Qiu, and B.Guo.
10. “Big data in health care: using analytics to identify and manage high-risk and high-cost patients,” D.W.Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar.

AUTHORS PROFILE

C.Akash Mahadevan is a prefinal year student pursuing his B. Tech degree in Computer Science Engineering in SRM Institute of Science and Technology, Ramapuram, Chennai. He is Game Designing enthusiast and has industrial exposure in the field of Database Management.

S.Kanishka is a prefinal year student pursuing her B. Tech degree in Computer Science Engineering in SRM Institute of Science and Technology, Ramapuram, Chennai. She has an industrial exposure in Python Programming.

Saisurya.S is a prefinal year student pursuing his B. Tech degree in Computer Science Engineering in SRM Institute of Science and Technology, Ramapuram, Chennai. He has an industrial exposure in the field of Database Management.

V.Arun is Assistant Professor in the Department of Computer Science and Engineering in SRM Institute of Science and Technology, Ramapuram, Chennai, He has published several papers in past and is currently working on many research oriented projects.