DIABETES PREDICTION USING RANDOM FOREST ALGORITHM

[1]Prof. DR.S.ILANGOVAN, [2]Ms.NAISHVINI A, [3]Ms. SRINITHI J P, [4]Ms. NIVEDITA T P

[1]Professor, [2,3,4]Students
Department of Information Technology
K.L.N.College of Engineering, Sivagangai, Tamil Nadu, India

Abstract: Diabetes is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques.The prediction model is introduced with different combinations of features and several known classification techniques. The system is developed based on classification algorithms includes Random Forest, and Logistic Regression algorithms have been used. Machine learning is a huge field which learns from past experiences and gives proper predictions.

***Index Terms*: Diabetes Prediction, Machine Learning, Missing values and outliers, Cloud data warehouse**

INTRODUCTION:

It is difficult to identify diabetes because of several contributory risk factors such as high blood pressure, high cholesterol and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of diabetes among humans. The severity of the disease is classified based on various methods like Random Forest and Logistic Regression. The nature of diabetes is complex and hence, the disease must be handled carefully. We have seen ML algorithms are used in predicting the accuracy of events related to diabetes. With the development of living standards, diabetes is increasingly common in people's daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels.

MOTIVATION OF THE PROJECT:

Rapid urbanization and lifestyle changes are major causes for the increase in its rapid growth. We wish to reduce the mortality rate of diabetic people by improving patient care, while also reducing the astronomical yearly cost of diabetic patient care. The motivation of the project is to predict the early stage of diabetic disease by using machine learning techniques and provides the better result. It will very helpful to lead to improved treatment.

PROBLEM STATEMENT:

Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation. Between 2000 and 2018, there was a 5% increase in premature mortality from diabetes. In 2020, diabetes was the ninth leading cause of death with an estimated 1.5 million deaths directly caused by diabetes. It is evident from the literature that the incidence of diabetes mellitus is increasing and that although there is evidence that the complications of diabetes can be prevented, there are still patients who lack the required knowledge and skills to manage and control their condition.

LITERATURE SURVEY:

1: Diabetes in developing countries

Author : A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. RezaAlbarrán, and K. L. Ramaiya

FINDINGS:

There has been a rapid escalation of type 2 diabetes (T2D) in developing countries, with varied prevalence according to rural vs urban habitat and degree of urbanization. Some ethnic groups (eg, South Asians, other Asians, and Africans), develop diabetes a decade earlier and at a lower body mass index than Whites, have prominent abdominal obesity, and accelerated the conversion from prediabetes to diabetes. The burden of complications, both macro- and microvascular, is substantial, but also varies according to populations. The syndemics of diabetes with HIV or tuberculosis are prevalent in many developing countries and predispose to each other. Screening for diabetes in large populations living in diverse habitats may not be cost-effective, but targeted high-risk screening may have a place. The cost of diagnostic tests and scarcity of health manpower pose substantial hurdles in the diagnosis and monitoring of patients. The quality of care is largely poor; hence, a substantial number of patients do not achieve treatment goals. This is further amplified by a delay in seeking treatment, "fatalistic attitudes", high cost and non-availability of drugs and insulins. To counter these numerous challenges, a renewed political commitment and mandate for health promotion and disease prevention are urgently needed. Several low-cost innovative approaches have been trialed with encouraging outcomes, including training and deployment of non-medical allied health professionals and the use of mobile phones and telemedicine to deliver simple health messages for the prevention and management of T2D.

2: Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset

Author : R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri

FINDINGS:

Diabetes Mellitus is a dreadful disease characterized by increased levels of glucose in the blood, termed as the condition of hyperglycemia. As this disease is prominent among the tropical countries like India, an intense research is being carried out to deliver a machine learning model that could learn from previous patient records in order to deliver smart diagnosis. This research work aims to improve the accuracy of existing diagnostic methods for the prediction of Type 2 Diabetes with machine learning algorithms. The proposed algorithm selects the essential features from the Pima Indians Diabetes Dataset with Goldberg's Genetic algorithm in the pre-processing stage and a Multi Objective Evolutionary Fuzzy Classifier is applied on the dataset. This algorithm works on the principle of maximum classifier rate and minimum rules. As a result of feature selection with GA the number of features is reduced to 4 from 8 and the classifier rate is improved to 83.0435 % with NSGA II in training rate of 70% and 30% testing.

3: IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045

Author: N. H. Choac, J. E. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. W. Ohlrogge, and B. Malandaa

FINDINGS:

Uncertainties persist about the magnitude of associations of diabetes mellitus and fasting glucose concentration with risk of coronary heart disease and major stroke subtypes. We aimed to quantify these associations for a wide range of circumstances.We undertook a meta-analysis of individual records of diabetes, fasting blood glucose concentration, and other risk factors in people without initial vascular disease from studies in the Emerging Risk Factors Collaboration. We combined within-study regressions that were adjusted for age, sex, smoking, systolic blood pressure, and body-mass index to calculate hazard ratios (HRs) for vascular disease.

4: Global and regional diabetes prevalence estimates for 2019

Author : P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karurangaa, N. Unwin, S. Colagiuri, L. Guariguata

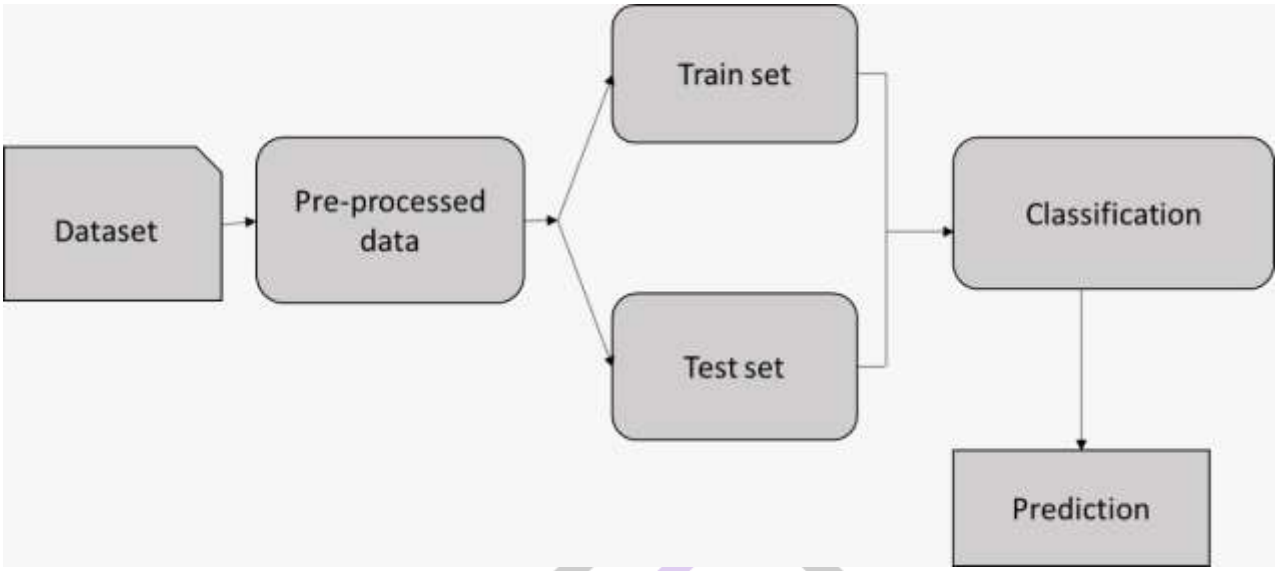
FINDINGS:

Diabetes is one of the most challenging health problems in the 21st century that brings a considerable economic burden on worldwide healthcare resources.Indeed, people with diabetes have a higher lifetime healthcare expenditure due to the long-term complications, which include micro and macrovascular complications. This study sought to estimate the frequency of diabetes complications, and to investigate the associated risk factors. Methodology: Data were obtained from the medical records of 2401 diabetic patients followed at the Reference Center of Diabetes and Chronic Diseases (RCD) in Oujda (Morocco) during the period 2006-2011. Results: Our sample of 2401 diabetic patients include 64.7% women. 32% of patients have one or more complications; retinopathy is the most frequent complication (16.8%), followed by nephropathy (12.4%), cardiovascular diseases (5.4%), neuropathy (3.6%) and diabetes foot (2%). Logistic regression in univariate followed by multivariate analysis has showed that age, duration of diabetes and high albuminuria are the major risk factors for the development of diabetic complications in both type 1 and type 2 diabetes. Conclusions: Nearly one third of diabetic patients were affected by at least one diabetic complication; retinopathy is the most common complication in these patients. Strengthening programs to improve diabetes management and to reduce the risk of these complications should be a high priority in order to control the cost of treatment.

PROPOSED SYSTEM :

The proposed model is introduced to overcome all the disadvantages that arises in the existing system. This system will increase the accuracy of the Supervised classification results by classifying the data based on the social network mental disorders and others using Decision tree classification algorithm.It enhances the performance of the overall classification results.Apply ensemble data mining techniques to the dataset to investigate if ensemble ML techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis.

SYSTEM ARCHITECTURE:



ADVANTAGES:

- High performance.
- Provide accurate prediction results.
- It avoid sparsity problems.

LIMITATIONS:

- Doesn't Efficient for handling large volume of data.
- Theoretical Limits
- Incorrect Classification Results.
- Less Prediction Accuracy.

SYSTEM REQUIREMENTS:

Software Requirements:

Operating System : Windows 7

Language : Python

IDE : Anaconda - Spyder

Hardware Requirements:

Hard Disk : 1000 GB

RAM : 4GB

CONCLUSION:

In this framework, we scramble the information present in the cloud twice by utilizing intermediary reencryption technique.This is recommended that utilizes AES in half breed with symmetric intermediary reencryption conspire.So information can be re-encoded by cloud servers. The information proprietor just needs to create a lot of re-encryption keys and AES figure content scrambling the new keys then send both to the cloud for reencryption.

REFERENCES:

[1] A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. RezaAlbarrán, and K. L. Ramaiya, "Diabetes in developing countries," Journal of Diabetes, vol. 11, no. 7, pp. 522-539, Mar. 2019.

[2] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in Proc. International Conference on Computing Networking and Informatics, Oct. 2017, pp. 1-5.

[3] Emerging Risk Factors Collaboration and other, “Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies,” The Lancet, vol. 375, no. 9733, pp. 2215-2222, Jul. 2010.

[4] N. H. Choac, J. E. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. W. Ohlrogge, and B. Malandaa, “IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,” Diabetes Research and Clinical Practice, vol. 138, pp. 271-281, Apr. 2018.