IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907062 |

Diabetes Prediction Using Machine Learning Techniques like Logistic Regression and K-Means Algorithms

M.Sai Vineela[1], P.Krishna Sri[2], M.Lakshmi Keerthana[3], M.Jyostna[4], Mr.A.Sudharsan Reddy, M.Tech[5]

UG Students, Dept. of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur district, Andhra Pradesh, India[1,2,3,4]

Associate Professor, Dept. of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur district, Andhra Pradesh, India[5]

ABSTRACT: Diabetes causes a large number of deaths every year and a large number of people living with the disease do not realize their health condition early enough. Early prediction of diabetes is an important issue in Health Care Services (HCS). In this study, a model for early diagnosis and prediction of diabetes using the Pima Indians Diabetes dataset is proposed. Various techniques and algorithms are designed for application in extracting knowledge and information in the diagnosis and treatment of disease from medical databases.

This proposed model comprises PCA (Principal Component Analysis), K-means and Logistic Regression algorithm. To enhance the K-means clustering algorithm, PCA will be used to reduce the dataset to a lower dimension. Logistic regression algorithm is used to classify data items into categories. The model is useful for automatically predicting diabetes using patient electronic health records data.

KEYWORDS: Diabetes, Prediction, PCA, K-means, Logistic Regression.

I. INTRODUCTION

Diabetes is one of the deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. Early prediction of diabetes is an important issue in Health Care Services (HCS).Various techniques and algorithms are designed for application in extracting knowledge and information in the diagnosis and treatment of disease from medical databases.

II. LITERATURE SURVEY

Iyer, Aiswarya & Jeyalatha, S & Sumbaly, Ronak proposed Diagnosis of Diabetes Using Classification Mining Techniques.This paper aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms. The research hopes to propose a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907062 |

Yılmaz, Nihat & Inan, Onur & Uzer, Mustafa proposed A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Diabetes Diseases.This article presents a new data preparation method based on clustering algorithms for diagnosis of heart and diabetes diseases. In this method, a new modified K-means Algorithm is used for clustering based data preparation systems for the elimination of noisy and inconsistent data and Support Vector Machines is used for classification.

K. Rajesh and V. Sangeetha used classification techniques. They used the C4.5 decision tree algorithm to find hidden patterns from the dataset for classifying efficiently B.M. Patil, R.C. Joshi and Durga Toshniwal (2010) proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of a classification algorithm to the result obtained from the clustering algorithm.

III. DATA SET

The data is gathered from the Pima Indian Diabetes Dataset which is present in source kaggle.Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.
The dataset has 768 entries and has the following attributes:

1. Pregnancies - Number of times pregnant
2. Glucose - Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Blood Pressure - Diastolic blood pressure (mm Hg)
4. Skin Thickness - Triceps skin fold thickness (mm)
5. Insulin - 2-Hour serum insulin (mu U/ml)
6. BMI - Body mass index (weight in kg/(height in m)^2)
7. Diabetes Pedigree Function - Diabetes pedigree function
8. Age - Age (years)
9. Outcome - 0 or 1

IV. IMPLEMENTATION

4.1 Data Preprocessing:

In this step we retrieve data from the database and convert the raw data into an efficient format. We transform the numeric attribute into a nominal attribute of value 0 and 1.Here we also handle missing values for a few selected attributes like Glucose level, Blood Pressure, Insulin, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	1	148.000000	72.000000	35.000000	79.799479	33.600000	0.627	50	1
1	1	85.000000	66.000000	29.000000	79.799479	26.600000	0.351	31	0
2	1	183.000000	64.000000	20.536458	79.799479	23.300000	0.672	32	1
3	1	89.000000	66.000000	23.000000	94.000000	28.100000	0.167	21	0
4	0	137.000000	40.000000	35.000000	168.000000	43.100000	2.288	33	1
5	1	116.000000	74.000000	20.536458	79.799479	25.600000	0.201	30	0
6	1	78.000000	50.000000	32.000000	88.000000	31.000000	0.248	26	1
7	1	115.000000	69.105469	20.536458	79.799479	35.300000	0.134	29	0
8	1	197.000000	70.000000	45.000000	543.000000	30.500000	0.158	53	1

Figure 1: Dataset after data pre processing

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907062 |

4.2 Data Standardization:

In this process the data is converted into a common format in order to analyze it. StandardScaler is imported.

StandardScaler: standardizes a feature by subtracting the mean and then scaling to unit variance.

4.3 Principle Component Analysis:

PCA algorithm is generally used to find the most important features of the data set. Here, we used the PCA algorithm for the most important features/records.

- Find the mean of data
- Scale the data with use mean of data
- Calculate data's covariance matrix
- Get diagonals of covariance matrix and find variances of data
- Sort variances of data from the max to min
- Finally delete records which is first n minimum variance

By applying PCA the dataset is reduced into 3 principal components.

4.4 K-means:

The k-means algorithm searches for a predetermined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like.

The "cluster center" is the arithmetic mean of all the points belonging to the cluster. Each point is closer to its own cluster center than to other cluster centers. Those two assumptions are the basis of the k-means model.

By using K-means clustering the data set is partitioned into two predefined clusters. Those clusters are 0 and 1.

4.5 Logistic Regression:

In simple, linear regression, predict scores on one variable from the scores on a second variable.The variable that is predicted is called the criterion variable and is referred to as Y. The variable base for predictions is called the predictor variable and is referred to as X.When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

	precision	recall	f1-score	support
0	0.98	1.00	0.99	371
1	1.00	0.96	0.98	179
accuracy			0.99	550
macro avg	0.99	0.98	0.99	550
weighted avg	0.99	0.99	0.99	550

Figure 2: Diabetes prediction after applying logistic regression+PCA+K-Means algorithm

V. RESULTS AND DISCUSSIONS

The proposed work is implemented in Anaconda 3 with libraries Scikit-learn, pandas, matplotlib and other mandatory libraries. The PIMA Indian diabetes dataset is considered from UCI repository uci.edu. Machine learning algorithms such as Logistic regression,SVM,Naive bayes are used. We used these machine learning algorithms and identified diabetes. Also we applied Principal component Analysis (PCA) for feature reduction and K-means clustering for clustering. The result shows PCA and k-

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907062 |

means on machine learning algorithms achieved higher accuracy than machine learning algorithms.
The below table shows accuracy metrics of algorithms before and after PCA and K-means.

Algorithm	Before PCA+ k-means Accuracy (%)	After PCA+ k-means Accuracy (%)
Logistic Regression	77	99
SVM	82	99
Naive Bayes	75	94
K-NN	85	98

Table: Experimental Results of proposed system

VI. CONCLUSIONS

Medical data needs to be processed to find out the pattern and extraction of data for analysis purposes, data mining and machine learning were used. In different sectors of medicine, these techniques were found useful including medical image processing like brain tumour, cancer disease detection, diabetes, liver disease and heart disease, Parkinson disease identification, early detection of leukaemia etc. In this work, we considered diabetes, as it was found to be a common and major disease amongst Indians. We considered PIMA Indian diabetes database, and evaluated in machine learning algorithm SVM, Naive Bayes, Logistic regression and KNN algorithm. Also we evaluated the Novel model of feature reduction and clustering using Principal Component Analysis (PCA) and K-means algorithm. This model consists of PCA for dimensionality reduction, k-means for clustering, and logistic regression for classification. The proposed model gives better accuracy when compared with other models. Experimental results show that PCA enhanced the k-means clustering algorithm and logistic regression classifier accuracy versus the result of other published studies, with a k-means output of 25 more correctly classified data, and a logistic regression accuracy of 1.98% higher.

REFERENCES

[1] Sajida Perveena, Muhammad Shahbaza, Aziz Guergachib, Karim Keshavjeec, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes" Procedia Computer Science 82 (2016) 115 – 121.
[2] Deepti Sisodia, Dilip Singh Sisodia, "Prediction of Diabetes using Classification Algorithms", International Conference on Computational Intelligence and Data Science (ICCIDS 2018)
[3] Santi Wulan Purnami, Abdullah Embong, Jasni Mohd Zain and S.P. Rahayu, " A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis", Journal of Computer Science 5 (12): 1003-1008, 2009
[4] Faezeh Ensan, Mohammad Hossien Yaghmaee, Ebrahim Bagheri, "FACT: A new Fuzzy Adaptive Clustering Technique", The 11th IEEE Symposium on Computers and Communications, Sardinia, 26-29 June 2006
[5] Aishwarya, Gayathri, Jaisankar, "A Method for Classification Using Machine Learning Technique for Diabetes", International Journal of Engineering and Technology 2013.
[6] Nilesh Jagdish Vispute, Dinesh Kumar Sahu, Anil Rajput,"ASurvey on naive Bayes Algorithm for Diabetes Data Set Problems", International journal for research in Applied Science & Engineering Technology (IJRASET),Volume 3 issue XII,December 2015
[7] Haldurai Lingaraj, Rajmohan Devadass, Vidya Gopi, Kaliraj Palanisamy," Prediction of Diabetes Mellitus using Data Mining Techniques": A Review, Journal of Bioinformatics & Cheminformatics, February 19,2015.
[8] Isha Vashi, Prof. Shailendra Mishra,"A Comparative Study of Classification Algorithms for Disease Prediction in Health Care", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907062 |

[9] VrushaliBalpande, Rakhi Wajgi," Review on Prediction of Diabetes using Data Mining Technique", International Journal of Research and Scientific Innovation (IJRSI) |Volume IV, Issue IA, January 2017 | ISSN 2321–2705

[10] Haritha, R., Babu, D. S., & Sammulal, P. (2018). A Hybrid Approach for Prediction of Type-1 and Type-2 Diabetes using Firefly and Cuckoo SearchAlgorithms. International Journal of Applied Engineering Research, 13(2), 896-907.

[11] Rashid, T. A., & Abdullah, S. M. (2018). A Hybrid of Artificial Bee Colony, Genetic Algorithm, and Neural Network for Diabetic Mellitus Diagnosing. Aro-THE Scientific JOURNAL of Koya University, 6(1), 55-64.

[12] Zhang, Y., Lin, Z., Kang, Y., Ning, R., & Meng, Y. A Feed- Forward Neural Network Model For The Accurate Prediction Of Diabetes Mellitus.

[13] Kadhm, M. S., Hindawi, I. W., Muhawi, D. E. (2018). An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach. International Journal of Applied Engineering Research, 13(6), 4038-4041.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details