Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 10
October 2023

An Ensemble Classifier by Combining Natural Language Processing (NLP) AND Machine Learning Models to Detect the Diabetes Detection

[1]Dr. B. V. V. Padmavathi, [2]Jasti Pushpalatha

[1]Asst.Prof of English, Velagapudi Ramakrishna Siddhartha Engineering College, Chalasani Nagar, Kanuru, Vijayawada.

[2]Assistant professor, Dhanekula Institute Of Engineering And Technology , Ganguru,Andhra Pradesh.

Abstract: Diabetes is a chronic disease that affects millions of people worldwide and poses significant health challenges. Early detection and management of diabetes can lead to improved outcomes and better quality of life for patients. In recent years, there has been growing interest in utilizing Natural Language Processing (NLP) and Machine Learning (ML) algorithms to aid in diabetes detection and diagnosis. This research focuses on developing a robust and accurate system for diabetes detection using NLP and ML techniques. The proposed approach involves the analysis of textual data, such as electronic health records, patient notes, and medical literature, to extract relevant information related to diabetes risk factors, symptoms, and medical history. The NLP component of the system employs advanced techniques, including text preprocessing, named entity recognition, and sentiment analysis, to extract meaningful features from unstructured text data. These features are then used to build a comprehensive feature set for ML model training. Various ML algorithms, such as Support Vector Machines, Random Forest, and Gradient Boosting, are employed to create predictive models based on the extracted features. The models are trained on labeled datasets containing information about diabetic and non-diabetic individuals. To evaluate the system's performance, extensive experiments are conducted using cross-validation techniques and performance metrics like accuracy, precision, recall, and F1-score. The results demonstrate the effectiveness and efficiency of the proposed approach in accurately detecting diabetes from textual data. The developed system shows promising results in early diabetes detection, which can help healthcare providers in timely interventions and personalized treatment plans for patients at risk. By leveraging the power of NLP and ML, this research contributes to the ongoing efforts in improving diabetes management and ultimately reducing the burden of diabetes on individuals and healthcare systems.

Keywords: Diabetes, Natural Language Processing, Machine Learning, Text Analysis, Disease Detection, Healthcare, Medical Records.

Introduction

All around there are various constant diseases that are vast in advanced and agricultural countries. One of such ailment is diabetes. Diabetes is a metabolic issue that causes glucose by making a critical proportion of insulin in the human body or by creating a little proportion of insulin. Diabetes is maybe the deadliest disorder on earth. It isn't simply an illness yet, likewise a creator of various kinds of infections like a coronary disappointment, visual lack, kidney sicknesses and nerve hurt, etc. Hence, the ID of such persistent metabolic disease at a starting period could help experts all over the planet in hindering loss of human existence. As of now, with the climb of AI, computer based intelligence, and brain frameworks, and their application in different areas we might have the choice to track down a solution for this issue. ML techniques and brain frameworks assist researchers with finding new real factors from existing prosperity related educational lists, which might help in disease management and discovery. The ongoing work is finished using the Pima Indians Diabetes Data set. The place of this casing work is to make a ML model, which can expect with accuracy the probability or the chances of a patient being diabetic. The customary distinctive cycle for the area of diabetes is that the patient necessities to visit a suggestive concentration. One of the central questions of bio-informatics assessment is to accomplish exact results from the data. Human mix-ups or different lab tests can trap the method of distinguishing proof of the illness. This model can predict regardless of whether the

patient has diabetes, supporting experts to guarantee that the patient needing clinical thought can get it on time and furthermore assist with expecting the deficiency of living souls. DNA settles on brain networks the clear decision. Brain networks use neurons to send information across different layers, with every hub dealing with an alternate weighted boundary to assist with foreseeing diabetes.

In the event that diabetics patient is untreated for quite a while, it might prompt increment glucose. Presently a days, Medical care enterprises creating huge volume of information. AI calculations and measurements are utilized to foresee the sickness with the assistance of current and past information. AI procedures assists the specialists with anticipating beginning phase for diabetics. Diabetics patient clinical record and various kinds of calculations are added in dataset for trial examination. we utilize different ML models to foresee whether a patient has diabetes in view of demonstrative estimations. Execution and precision of the applied calculations is talked about and looked at.

Literature Work

Sun and Zhang [1] studied several important learning and representational techniques, such as false mental associations, decision trees, unpredictable forests, and the SVM. Qawqzeh et al. [4] implemented a feedback-based strategy to collect data on diabetes. Data preparation considers 460 patients and test data consolidates 130 patients. The accuracy of the query obtained by the developers was 92% using the reverse key.A major disadvantage of this model was that it was undifferentiated and could not support other models for estimating diabetes in this way. Tafa et al. [5] divides the data set into half an array set and half a test set. A model has been proposed that uses a combination of innocent Bayesian vector machine estimates and sponsorship for the diabetes hypothesis. The data set consisted of three distinct regions and the proposed model was based on this data set. Eight credits were available in the dataset, involving 402 patients, including 80 patients with type 2 diabetes. The simple Bayesian machine and Sponsor vector set achieved an accuracy of 97.6%, which is significantly better than calculations made with the dataset alone or nothing else were performed, Guiltless Bayes achieves an accuracy of 94.52 and Support Vector Machine achieves 95.52%.The developers did not refer to pre-processing steps to filter out unwanted features from the data set. Karanet al. [6] presented another method to build confidence in diabetes by organizing a distributed, end-to-end, three-tiered system of specific clinical benefits using false mental association representations (ANS). At the most basic level, sensors and wearables are used to study key markers in the human body. The third tier integrates energy zones to consolidate regional servers that provide clients with wellness maps and information index maps. Using a fake brain network is used to study disorders at both a concurrent and future level. Mental network dummy calculations require the client and server to model them on their basis. Depending on the possible infection, this technology accelerates both client-side and server-side evaluations and exchange structures. Sisodia and Sisodia [7] applied Guileless Bayesian methods, DT, and computer assist vectors to the Pima Indian diabetes dataset, and the best diabetes prediction accuracy was obtained using a simple Bayesian classifier. Sisodia used a ten-part peer review methodology, dividing the data set into ten identical parts: 9 segments were used for preparation and an additional part for testing. The expected assessment limits for diabetes were precision, accuracy, control, and displaced range. A study of different AI calculations was presented by Hussain et al., [8] which considered Unpredictable Forest, Unsuspecting Bayes and Brain Network for accuracy. To test these calculations of human intelligence, the developers used Matthews' association coefficient. In [9] worked on the real-time dataset, applying flawless bayes, uneven woods and key backslides and varying these three routines and the clothing and pattern hits with a accuracy of 79%. In [10] used meaningful understanding, in other words mental association, which is a multi-level and feed-forward association. The developers made an estimate of the Pima Indian Diabetes dataset and the dataset was decoupled to the point where 500 features were used for planning and 268 for testing. The

data set was normalized to obtain a numerical fit before any pre-processing tasks were performed.

Proposed Methodologies

Dataset collection

It incorporates information assortment and understanding the information to concentrate on the secret examples and patterns which assists with anticipating and assessing the outcomes. Dataset carries 1405 lines i.e., absolute number of information and 10 segments i.e., complete number of elements.

Data Pre-processing:

This period of model handles conflicting information to come by additional exact and exact outcomes like in this dataset Id is conflicting so we dropped the component. This dataset doesn't contain missing qualities. In this way, we ascribed missing qualities for few chose credits like Glucose level, Pulse, Skin Thickness, BMI and Age on the grounds that the ocean credits can't have values zero. Then, at that point, information was scaled utilizing Standard Scaler. Since there were fewer highlights and significant for expectation so no component choice was finished.

Missing value identification:

Utilizing the Panda library and SK-learn, we got the missing qualities in the datasets, We supplanted the missing worth with the comparing Mean worth.

Feature selection:

Pearson's relationship strategy is a famous technique to track down the most pertinent characteristics/highlights. The connection coefficient is determined in this strategy, which relates with the result and info credits. The coefficient esteem stays in the reach by somewhere in the range of −1 and 1. The worth above 0.5 and beneath −0.5 shows an outstanding connection, and the zero worth means no relationship

Scaling and Normalization:

We performed highlight scaling by normalizing the information from 0 to 1 territory, which supported the calculation's computation speed. scaling implies that you're changing your information so it fits inside a particular scale, similar to 0-100 or 0-1. You need to scale information while you're utilizing strategies in view of proportions of how farapart information focuses are, similar to SVM or KNN.With these calculations, a difference in "1" in any numeric element is given a similar significance.

Splitting of data:

The dataset is split into two folders such as training and testing.

Design and implementation of classification model:

In this exploration work, thorough examinations are finished by applying different ML characterization strategies like KNN, RF, NB, SVM, MLP.

Experimental Results

In this section the performance is measured by using the confusion matrix. Confusion matrix used to predict the model performance for the diabetes detection. Here, TP: True positive; FP: False positive; TN: True negative; FN: False negative.

The following performance metrics are used to calculate the presentation of various algorithms.

TP – Actual positive predicted positive.

TN – Actual positive predicted negative.

FP – Actual negative predicted positive.

FN – Actual negative and predicted negative.

Precision is TP/ total number of person have prediction result is yes.

Accuracy is the total number of correctly classified records.

Results and Discussions

ML algorithms are most widely used to predict diabetes in the earlier stages. The training is 30%, and testing is 80%. The MLP shows immense accuracy compared with the 89%. The comparative results are shown in table-1.

Machine Learning Algorithms	Accuracy
K-Nearest Neighbors	67%
SVM	78%
Naïve Bayes	75%
Random Forest	72%
Multilayer Perceptron	89%

Among all the used algorithms, MLP is gives more accuracy that is 89% which due to it filter the data through various layers which was acquired from Artificial Neural Network (ANN) .

Conclusion

The project predicts the onset of diabetes in a specific individual based on the relevant medical data collected. Once a person has entered all the required relevant medical data into the data set, this data is then passed to the trained model so that it can predict whether or not the person has diabetes. The model then makes a prediction with an accuracy of 89%, which is pretty good and reliable. The figure below shows the results of the downloaded dataset. This makes the more intuitive and easier to understand for beginners.

References

[1] Y. L. Sun and D. L. Zhang, “Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey,” Technical Gazette, vol. 26, pp. 872–880, 2019.

[2] S. Malik, S. Harous, and H. El-Sayed, “Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women,” in Proceedings of the International Symposium on Modelling and Implementation of Complex Systems, pp. 95–106, Springer, Algeria, October 2020.

[3] J. Han, J. C. Rodriguez, and M. Behesti, “Discovering Decision Tree-Based Diabetes Prediction Model,” in Proceedings of the International Conference on Advanced Software Engineering and its Applications, pp. 99–109, Springer, Jeju Island, Korea, December 2018.

[4] Y. K. Qawqzeh, A. S. Bajahzar, M. Jemmali, M. M. Otoom, and A. Thaljaoui, “Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling,” BioMed Research International, vol. 2020, Article ID 3764653, 2020.

[5] Z. Tafa, N. Pervetica, and B. Karahoda, “An Intelligent System for Diabetes Prediction,” in Proceedings of the 4th Mediterranean Conference on Embedded Computing (MECO), pp. 378–382, Budva, Montenegro, June 2015.

[6] O. Karan, C. Bayraktar, H. Karlık, and B. Karlik, “Diagnosing diabetes using neural networks on small mobile devices,” Expert Systems with Applications, vol. 39, no. 1, pp. 54–60, 2012.

[7] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.

[8] A. Hussain and S. Naaz, “Prediction of diabetes mellitus: comparative study of various machine learning models,” Advances in Intelligent Systems and Computing, vol. 1166, pp. 103–115, 2021.

[9] S. Kumari, D. Kumar, and M. Mittal, “An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier,” International Journal of Cognitive Computing in Engineering, vol. 2, pp. 40–46, 2021.

[10] E. O. Olaniyi and K. Adnan, “Onset diabetes diagnosis using artificial neural network,” International Journal of Scientific Engineering and Research, vol. 5, pp. 754–759, 2014.