Original Article

Diabetes Mellitius Detection and Self Management based on Machine Learning

M. Ranjit Reddy[1], P. Lakshmi Sagar[2], Nazma Sultana Shaik[3]

[1]Professor, Department of Computer Science & Engineering, Srinivasa Ramanujan Institute of Technology, Anantapur.

E-mail: midderanjit@gmail.com

[2]Assistant Professor, Dept. of Computer Science and Systems Engineering, Sree Vidyanikethan Engineering College, Tirupati, Andhrapradesh, India.
E-mail: lakshmisagar.p@vidyanikethan.edu

[3]Assistant Professor, Vignan's Foundation for Science Technology and Research Deemed to be University. E-mail: yaminireddybode@gmail.com

Abstract

Diabetes Mellitus is considered to be a state evoked by unmonitored polygenic disorder which will cause various organs collapse in sufferers. An investigation of the identification, examination and autonomous methods of Diabetes Mellitus from six completely various sides viz. datasets of Diabetes Mellitus, preprocessing procedures, attribute extraction, machine learning based analysis, classifying and prediction of Diabetes Mellitus, and evaluating the results. Machine Learning Associate in Nursing computer science is advancing, which permits the first prediction and diagnosing the Diabetes Mellitus over an automatic method that is superior than a nonautomatic detection. There are various reports which are revealed on automated Diabetes Mellitus prediction, identification, examination and autonomous procedure through machine learning and artificial intelligence procedures and also three current analysis problems within the department of Diabetes Mellitus prediction are recorded. In this it provides the Diabetes Mellitus prediction procedures demonstrate importance to the research community utilized within a range of automated Diabetes Mellitus prediction and self supervision.

Keywords: Machine Learning, Diabetes Detection.

DOI: 10.47750/pnr.2022.13.04.138

INTRODUCTION

Diabetes is normally referred to as DM. It is a type of metabolic sicknesses wherein sufferers be afflicted by blood glucose issues because of atypical manufacturing and launch of insulin. As consistent with WHO document on 14th November 2016, i.e. on World Diabetes Day, 422 million adults are residing with diabetes, and 1.6 million individuals who misplaced their lifestyles because of DM. In 2016, 1.6 million deaths have been at once because of diabetes. So, it's miles one of the critically want to be taken into consideration type of persistent ailment across the world. DM can motive harm to extraordinary frame elements viz., nerves, eyes, heart, to call a few. Every 12 months hundreds of thousands of human beings were given stricken by this lifestyles threatening ailment in each civilized and uncivilized elements of the world. CDCP (Center for Disease Control and Prevention) projected that in 2001 to 2009 there may be 23% boom in Type II diabetes in US.

Many countries, Organization, and extraordinary fitness sectors also are involved approximately this persistent ailment for accomplishing manage and prevention with a view to mitigate it in early stages, in order that man or woman lifestyles may be saved. Different variations of DM are there viz. Type I, Type II, Juvenile and Gestational. Type I is insulin dependent, Type II is insulin independent, Gestational can take place at some stage in being pregnant and Juvenile diabetes after delivery of a baby. According to Canadian Diabetes Association (CDA) in coming 10 years this is at some stage in 2010 to 2020, there might be a predictable increase from 2.5 to 3.7 million for human beings tormented by persistent sicknesses. So, via way of means of searching at those information diabetes and different persistent sicknesses evaluation performs a crucial position in saving sufferers lifestyles.

The observe is primarily based totally on good sized capabilities to expand a gadget learning-primarily based totally prediction set of rules and find out the first-class classifier to get the first-class outcomes whilst as compared to scientific outcomes. Using predictive evaluation, the recommended approach specializes in deciding on the capabilities that resource with inside the early detection of Diabetes Mellitus. The Objective of the observe is to assess the overall performance of Support Vector Machine that offers the better accuracy.

RELATED WORK

The purpose of this study is to see which algorithm produces the greatest results in terms of recognising an existing disease and forecasting the likelihood of developing one in the future, based on the patient's diagnostic parameters.

Recently a lot of research has been done on a variety of diabetics detection using machine learning techniques for diabetics disease detection. 5 research articles were published in science direct and 446 articles found in Google scholar in a span of 5 years. In a recent study analysis of diabetes mellitus for early prediction using optimal features selection as they predicted an accuracy of 77.7% using Support Vector Machine algorithm. The optimal features from the diabetics suffering person are detected with an accuracy of 77.7%. The researcher used different algorithms like SVM, KNN for detection of diabetics and got an accuracy of 60%. The author proposed five cross validation for analyzing the models and they used principal component analysis (PCA) minimum redundancy and maximum relevance (mRMR) for deducting dimensionalities accuracy was achieved by 54%. The researchers took the data mining techniques to detect the diabetics and got an accuracy of 75.7%. The present literature demonstrates that identifying diabetes disease is difficult due to a lack of accuracy. To increase the accuracy of the support vector machine and K-Nearest neighbor algorithms for detecting diabetes by removing outliers from the dataset, which are the main cause of the model's poor findings. The goal is to increase diabetic identification accuracy by utilizing a Support vector machine algorithm instead of a K-Nearest Neighbor approach.

BACKGROUND

A. MACHINE LEARNING

Machine learning is being used to predict an outcome using past data. Machine learning (ML) is an AI technology that enables machines to understand without being explicitly programmed. Machine learning encompasses both the production of data-adaptive computer programmes, and the principles of algorithms, such as the development of something like a simple machine learning model in python. Sophisticated algorithm is employed as in learning and forecasting phase. It provides training information to an algorithms, it then utilized the trained data to forecast on new test data.

B. NLP

Machines can read and comprehend human language due to natural language processing (NLP). Natural-language interfaces design and direct effective learning of knowledge directly from sentient sources, such as newswire texts, it might possibly be doable with a sophisticated natural language processing system. Information extraction, text categorization, knowledge discovery, and translation software are some simple uses of natural language processing. To generate syntactic representations of text, many contemporary techniques use word co-occurrence frequencies. "Keyword spotting" search algorithms are popular and scalable, but they're also stupid; a search query for "dog" might only return pages that contain the literal word "dog" and ignore papers that contain the term "poodle." Lexical affinity techniques measure the emotion of a document by looking for words like "accident".

THE PROPOSED SCHEMES

The proposed plans and the system model will be presented within this part.

A. Analyse Exploratory Data

supervised classification using ml algorithms methods would be utilised the analyze a dataset and identify trends that will aid in categorizing comments, allowing apps to develop decent feature judgments going forward.

B. Data Manipulation

In this first import the file, verify to ensure cleanliness, then reduce and cleanse the catalog of inspection Within the report's section.

C. Data Collection

There are two components in the data collection to be able to classify information provided: testing and training. In most cases, data is split into train and test sets using a 70:30 ratio. The train set is then loaded with the SMLT created Data Model, and validation set forecast is performed based on the test result precision. The comments are examined, demonstrating the utility of machine learning to separate the comments.

MODULE DESCRIPTION

The basic functionalities of the employed modules are described in this section. The following is a list of modules.

- Process of Data Validation
- Exploration of data analysis and visualisation
- Comparing Algorithms to Predictions in the Form of the Highest Accuracy
- The best accuracy result is obtained by combining RNN and LSTM.
- Sentiment prediction output using input sentence.

a) Data Validation Process

Validation approaches for machine learning are used to determine the failure the ML algorithm's rate, which is as follows feasible to the specific sample's error rate, replicate the value, resulting in the data form description to get whether there is a missing value that's a float variable or an integer variable. To offer a neutral reference point for tuning model hyper parameters, sample data is employed estimate on the system to suit on the actual training set. The verification sample is employed to test a model, although it isn’t on a regular basis, it's used. Engineers working on machine learning used this data to control the modeling extraordinary parameters.

The processing of data can understand it, and it manages the data correctness, structure, analyze information take a long time. During the data recognition process, it is beneficial to understand the data and its qualities; the knowledge can assist you in deciding which algorithm to use to build your design. for ex, to show a dataset's data type format. The Pandas library is used for a variety of data cleaning jobs, with an emphasis on the most common data cleansing tasks, such as the ability to clean data more quickly, and missing values.

b) Exploration of data analysis and visualisation

Data visualization is a critical capability in applied analytics and machine learning. Objective data interpretations and estimations are the focus of statistics. Data visualization is a set of approaches for providing a qualitative analysis of the information. While exploring and getting to know a dataset, this might be helpful for finding new technologies, fake outliers, results, and other tidbits. Data visualizations can be utilized to express and exhibit crucial relationships in graphs and maps which are more intuitive and relevant to companies than signs of association or relevance based on restricted domain data.

Data cannot be understood until it is presented graphically, like those in graphs and charts. for both applied statistics and related machine learning the capacity to visualize data samples and other objects quickly is an asset significant talent. It will show you how to better understand your data by using the many sorts of diagrams accessible when displaying data in Python.

Raw data is converted to clean data after preprocessing. To put it another way, data is gathered from variety of sources and processed in raw format, making interpretation difficult. The data must be appropriately organized in order to produce better results from the implemented Machine Learning model. Any machine learning a model is a collection of data in a specific format of set procedure that does not allow null values. As a result, in order to execute the random forest algorithm, all missing values from the initial input data collection are processed.

FALSE POSITIVE (FP): A defaulter is someone who has failed to pay a debt. If the intended class is yes, but the actual class is no. The actual class, for example, suggests that the passenger will survive, but the true class implies that an individual did not.

FALSE NEGATIVE (FN): A defaulter is more than likely a person who pays. When the class is really held higher than the projected class. If the passengers true class value indicates that they survived, yet the predicted class value predicts that this passenger would perish.

TRUE POSITIVE (TP): A defaulter is someone who just doesn't pay their debts. Both successfully forecast positive value, displayed the real class value as well as the coming class value. For example, if both the actual and predicted class values suggest that the passenger made it, then this passenger is alive.

TRUE NEGATIVE (TN): A defaulter is more than likely a payer. It's both precisely calculated negative number's, showing the real and projected classes have the same value. For instance, if the traveler did not survive in both the real and expected classes.

c) Comparing Algorithm with prediction in the form of best accuracy result

The following five algorithms are compared:

- Logistic Regression
- Decision Tree
- Random Forest
- K-Nearest neighbour

The K-fold training and testing procedure is employed to judge all algorithms, this assures to that of identical divides into training dataset/example is made and also checks that every algorithm had been accessed with similar way. Separate the testing and learning set as well. By evaluating accuracy, it's feasible to predict the conclusion.

A linear model with potential predictors is frequently utilized in the logistic regression process to estimate a value. The anticipated value ranges from negative and positive infinity. By comparing the best accuracy, the logistic regression model produces a higher accuracy prediction result.

TPR= TruePositive / (TruePositive + FalseNegative)

FPR= FalsePositive / (FalsePositive + TrueNegative)

Accuracy Calculation

Accuracy = (TruePositive + TrueNegative) / (TruePositive + TrueNegative + FalsePositive + FalseNegative)

The vast straight forward and efficient statistic is precision, that is defined as the proportion of accurately outcome expectations to all occurrences. If the model's accuracy is high, one might believe it is the best. Only when the records are symmetrical as well as the rates of false positive and negative is mostly now equal is the accuracy metric relevant.

precision = TruePositive / (TruePositive + FalsePositive)

Recall = TruePositive / (TruePositive + FalseNegative)

Fmeasure = 2*TP / (2*TP + FP + FN)

F1score = 2*(Recall * Precision) / (Recall + Precision)

ALGORITHMS AND TECHNIQUES

Categorization as an ensemble learning strategy in which the computing algorithm studies from input data and after it apply what it has learnt to categorise new finds.

The dataset can have two or more classes. Recognition of handwriting and speech, biometric authentication, and other categorization issues are some examples. In Supervised algorithms educated from labelled results. After knowing the information, the algorithm evaluates whether a mark must be given to the newly information premised on the patterns and connecting the pattern onto the unlabeled new information.

Logistic Regression

It just the method of reviewing the dataset into the which a single or multiple distinct factors have impact on the outcome. The outcome is evaluated using a variable that is either true or false. The goal of the logistic regression is the to select the top model for explaining the association between such a number of autonomous factors and an interesting dualistic traits of interests[11].

Decision tree Algorithm

It just the well-known and highly effective method. A supervised learning method, the decision tree algorithm is classified as such. It can be used to create output variables that are both continuous and categorical. Hypotheses for decision trees We presume that the entire training data is the base at first. characteristics are expected to just be categorical in regard to knowledge gain, however continuous characteristics are presumptions. Based on attribute values, entries are repeatedly allocated. It dismantles a large dataset into smaller subgroups over time while also building a decision tree.

K-Nearest Neighbour

This was the supervised computer learning approach this tracks every occurrences in multi-dimensional spaces that correspond to experimental data sets. Whenever unspecified separate data is collected, this examines the closest k preserved occurences and gives its most common in the estimation category, whereas legitimate data has to be returned the the k closest average neighbours. This uses the set of the designated areas as input and makes use of them to instruct on its own to identify further. It polls the specified spots closest to it and polls its neighbours to mark a new point[13].

Random Forest Algorithm

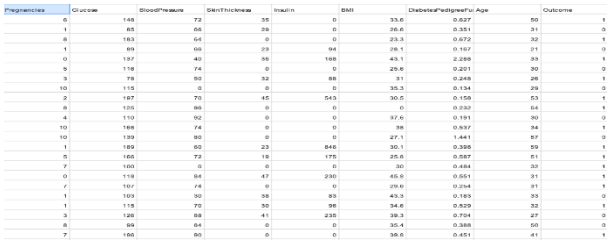
It is just regression approach or supervised classification that could develop a lot of decision trees and produce an outcome class based on the weighted average of all the trees. The algorithm is built on a system of ensembles training. Ensembles training is a method of creating a more efficient predictive model by integrating different versions of the same or comparable algorithms[13, 14].

Support vector machine

It's just a classification system that sorts data into categories by finding the best hyperplane among the key points. This classification was selected because it is extremely versatile with relation to the number of various functions of kernelling that can be applied, and that also has a high prediction score. When this was initially invented in the 1990s, they were a huge hit, and they're still a must have tool for such a high-performing algorithms with no adjustment today.

RESULTS

The findings show that the proposed classification method is accurate.



Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFu	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1

Figure 1: Diabetes dataset

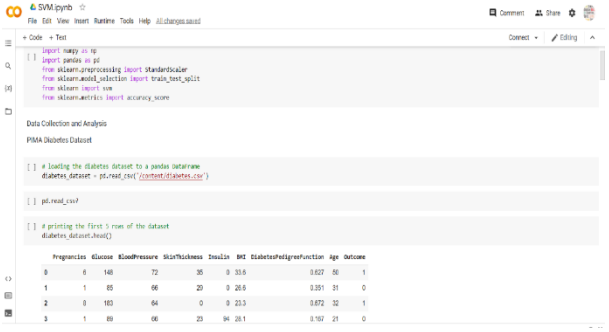


Figure 2: data collection

Figure 3: Data Standardisation

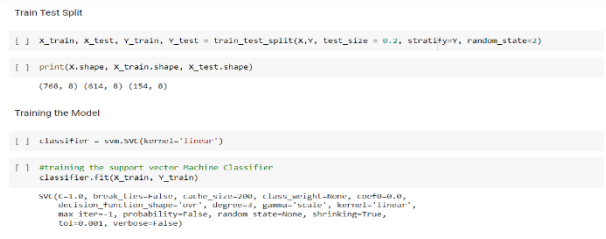


Figure 4: Train-Test split

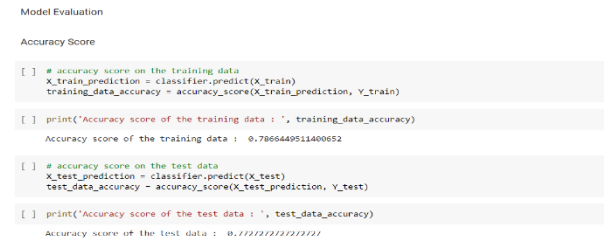


Figure 5: Model Evaluation

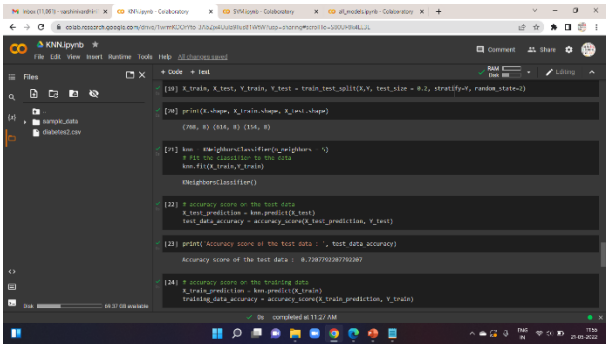


Figure 6: KNN accuracy

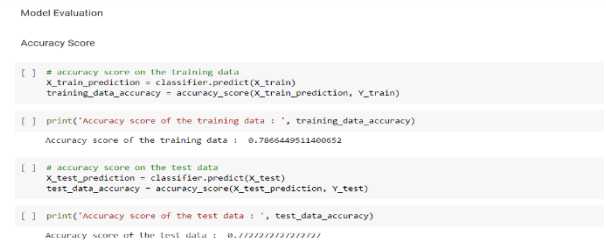


Figure 7: Naive bayes accuracy

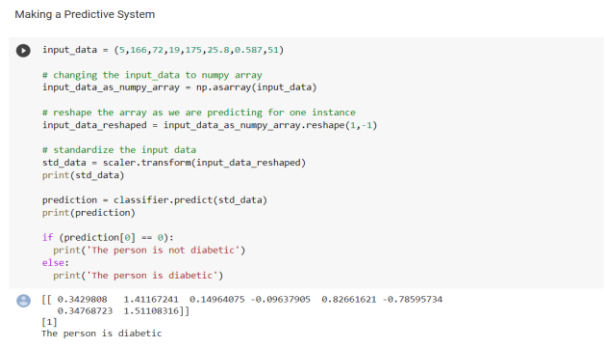


Figure 8: Predictive Analysis

Conclusion and Future Work

The analytical procedure starts from data cleaning and preparation, missing value, exploratory analysis and finally model creation and testing. The best accuracy on public testing set is higher overall accuracy score will be find out.

This tool can assist you in determining the Prediction of Google play store review classification prediction.

Future work of this app review prediction using machine learning is to review classification prediction to connect with cloud and to optimise the work for Artificial Intelligence implementation.

References

Mujumdar, Aishwarya, and V. Vaidehi. 2019. “Diabetes Prediction Using Machine Learning Algorithms.” Procedia Computer Science. https://doi.org/10.1016/j.procs.2020.01.047.

Pasha, Syed Matheen. 2020. “Diabetes and Heart Disease Prediction Using Machine Learning Algorithms.” International Journal of Emerging Trends in Engineering Research. https://doi.org/10.30534/ijeter/2020/60872020.

Agarwal B, Balas VE, Jain LC “Deep Learning Techniques for Biomedical and Health Informatics”, Academic Press, Jan 14, 2020, pages 367.

Machine learning and artificial intelligence based Diabetes Mellitus detection and self management: A systematic review. Journal of King Saud University - Computer and Information Sciences. 2020 [cited 30 Sep 2021]. doi:10.1016/j.jksuci.2020.06.013

Agarwal B, Balas VE, Jain LC, Poonia RC, Sharma M. Deep Learning Techniques for Biomedical and Health Informatics. Academic Press; 2020.

Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T. Current Techniques for Diabetes Prediction: Review and Case Study. NATO Adv Sci Inst Ser E Appl Sci., 2019; 9: 4604.

Kriještorac M, Halilović A, Kevric J. The Impact of Predictor Variables for Detection of Diabetes Mellitus Type-2 for Pima Indians. Advanced Technologies, Systems, and Applications IV -Proceedings of the International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (IAT 2019). 2020. pp. 388–405. doi:10.1007/978-3-030-24986- 1_31.

Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction.

Lakshmi Haritha.M & K. Ramani. Impact of Deep Learning on Localizing and Recognizing Handwritten Text in Lecture Videos. International journal of Advanced Computer Science and Applications; 12(4), 2021.

Sakthivel M, Sivanantham S, Kamalraj R & Krishnamoorthy V (2022). An Analysis of Machine Learning Depend on Q-MIND for Defencing the Distributed Denial of Service Attack on Software Defined Network. International Journal of Early Childhood Special Education. 2022; 14(05): 3769 – 3776.

Silpa C, RamPrakash Reddy Arava, K.K. Baseer. Agri Farm: Crop And FInternational Journal of Early Childhood Special Education ertilizer Recommendation Systemfor High Yield Farming Using Machine Learning Algorithms; 14(5): 2022.

P. Dhanalakshmi et al. (2022)Application of Machine Learning in Multi-Directional Model to Follow Solar Energy Using Photo Sensor Matrix. International journal of Photoenergy; 9:1-9.

K. Pujitha & R. Vidya (2017).Semantic based search engine for real images and web url’s using hypergraph distance measure algorithm. International journal of pure and applied mathematics; 115(6): 1311-8080.

Siva Kumar Depuru & Dr.K. Madhavi (2019). Autoencoder Integrated Deep Neural Network for effective analysis of malware in distributed Internet of Things (IoT) Devices. The International journal of analytical and experimental modal analysis; 9(11): 226-232.