

Prediction of Diabetes Mellitus Using Machine Learning Algorithm

S.V.K.R.Rajeswari¹, VijayakumarPonnusamy^{2*}

¹²SRM Institute of Technology , Tamil Nadu, India.

* vijayofsr@gmail.com

ABSTRACT

Diabetes Mellitus collectively known for Type 1 ,Type 2 and Gestational Diabetes is a condition that impairs the ability to provide blood sugar in the body. From the statistics of International Diabetes Federation in 2020, 463 Million people in between ages 20-79 which represents the 9.3% of whole world's population are diabetic .This statistics sums up the need of lifestyle changes and seriousness that one should have towards each self. Machine learning is the current trend in predicting and diagnosing diseases. It is vital for predictive analysis by using data and machine learning algorithms to recognize the future outcomes based on the data available or historical data. Our paper has extracted available diabetic data from Institution and Hospital and has provided a solution to predict diabetes providing a powerful insight. The models used in Machine Learning to predict diabetes are the Linear Regression, Support Vector Machine. Other algorithms require more computational time and Deep Learning algorithms requires a larger dataset .Hence in this paper, we have considered using classical algorithms.

Keywords

Diabetes Prediction, Machine Learning, Linear Regression, Support Vector Machine

Introduction

Diabetes Mellitus

Diabetes Mellitus is collectively known for all types of diabetes that occur in human body. It is an Endocrinological disorder. Diabetes Mellitus happens due to the defect in insulin action or secretion. It is broadly categorized into two types known as Hyperglycemia that happens due to high levels of glucose in the body and Hyperglycemia is due to less amount of glucose in the human body.

Under fasting blood glucose test, If the glucose level is in between 70-100 mg/dl, it is known as Normal .If the glucose level is in between 100-125 mg/dl, it is known as Pre-diabetes. It is considered as diabetes if the blood glucose level is more than 126mg/dl. Diabetes Mellitus is divided further into Type1,Type 2,Type 3/Type 1.5,MODY(Maturity Onset Diabetes of Young), Maternally inherited diabetes with deafness due to mitochondrial DNA mutation, Pancreatic defect(chronic pancreatitis, hemochromatosis, cystic fibrosis),Endocrinopathies associated with diabetes(cushing syndrome, growth hormone adenoma and hyperthyroidism),Infection associated diabetes etc.

Prediction of Diabetes Mellitus by implementing Machine Learning

Predicting diabetes is very important for any type of diabetes mellitus as it will provide an insight of the patient .By comparing the history of diabetes patient and predicting the possible outcome in the future will help doctors, care takers to have a better understanding of the disease on the patient and provide better and quality treatment. There are many approaches towards the prediction of diabetes mellitus by many researchers . Machine Learning is the work of algorithms in Artificial Intelligence. In [1],it is stated that Machine learning is learning can be classified into large data very quickly .Working with larger amount of data is feasible in machine learning approaches[1].The algorithms work based on the data. If a challenge has to be solved, Machine

Learning helps in developing algorithms and trains the machine on how to execute the instructions following a flow[2]. There are three different Machine Learning algorithms depending on the "signal" and "feedback" available to the system[1]. Supervised learning, Unsupervised learning and Reinforcement learning are the three methods.

High accuracy can be achieved using classification algorithms[3]. Machine learning approaches work on structured and unstructured data. The hospital data is a complete dataset of structured and unstructured data thus ML play a major role in classification and forming models of the dataset.

Contribution

This paper provides an insight on classical algorithms by implementing LR and SVM. Literature review on LR and SVM is also explored. To analyze and predict a smaller dataset, the traditional algorithms are a great fix. In order to have auto detection and prediction of diabetes, ML algorithm is applied for low complexity.

Literature Review

Prediction of any disease has its own importance and many researchers have always been very keenly working on the prediction algorithms. Here in Table 1 ,we discuss the different use cases and their outcomes of each research for predicting the diabetes using Machine Learning approaches[2]. From the literature review[3],[5],[7] comparison of various algorithms is presented. Linear regression, SVM[has produced more accuracy than other algorithms[14-15]. In [4] ,prediction of diabetes is worked on SVM with 5-fold validation. Integration of K-Means and PCA with Linear regression is discussed in [6].

Table 1. Literature Review.

References	Proposed work And Use Case	Outcome
R. Sehlyet.al[3]	This work proposes the comparative analysis of classifications models for pima diabetes.	Linear SVM gives 0.7721, Linear Regression produced 0.7695 providing highest accuracy. Receive K-NN produces 0.762,RBF SVM 0.612, ,Sigmoid SVM 0.6510,LDA 0.7734,CART is 0.6952,NB 0.7551.
R. Deo et.al[4]	This work proposes the performance assessment of ML based models for diabetes prediction. Prediction model is based on algorithms Linear SVM and bagged trees.	The Linear SVM model gives 91% of highest accuracy with a 5-fold cross validation.
P. Sonar et.al[5]	This work proposes the prediction of diabetes using different ML approaches.	Support vector classifier produced 82% accuracy, Decision tree classifier proposed 74%,Gaussian

		Naive Bias produces 80% whereas Artificial Neural Network produces 82%.
Zhuet.al[6]	This work proposes the improved logistic regression model for predicting diabetes by integrating K-means and PCA.	LR produced highest accuracy more than 98% than other algorithms when combined with K-means and PCA.
Jobedaet.al[7]	This work proposes the comparison of ML algorithms for diabetes prediction.	LR and SVM produced 77%-78% accuracy than DT,KNN,RF,NB,AB.
Alshamalan et.al[8]	This work proposes the prediction of gene function in Type 2 diabetes.	Fisher score and LR was implemented where LR produced highest accuracy of 90.23%.
Syed et.al[9]	This work proposes an approach for data classification in diabetes disease prediction.	89.67% of highest accuracy is achieved with SVM algorithm.
Miao et.al[10]	This work proposes an algorithm for future prediction of diabetes .	Accuracy of 96.5% is achieved using SVM.
Al-Zebari et.al [11]	This work proposes the comparison of different ML techniques in detecting diabetes.	Among the algorithms used such as DT,LR,DA,SVM,k-NN,best accuracy is produced by LR with 77.9.

Chakour et.al [12]	This work proposes early detection of diagnosis in diabetes.	Among the different algorithms used, i.e. ANN,SVM and LR,LR has produced accuracy of 0.97 and max sensitivity of 0.93.
--------------------	--	--

This paper	This work proposes prediction of diabetes.	Our work has achieved coefficient of 38.23786125 with mean squared error of 2548.07 and determination of coefficient is 0.47 with LR and testing accuracy of 0.82,training accuracy of 0.75.Precision of 0.63,recall of 0.82,F1 of 0.71 and AUC of 0.77.
------------	--	--

From the literature review [8],[9],[11],[12],among many algorithms, LR has produced max accuracy in detecting diabetes. In [10], ANN,SVM,LR are implemented for early detection and diagnosis of diabetes. Linear Regression produced ore accuracy(98%) when combined with K-Means and PCA.

Motivation for using LR and SVM :The main purpose of executing LR and SVM is that on the considered dataset, traditional algorithms can be implemented. LR and SVM provide low computational complexity than other algorithms .Other algorithms require more computational time and complexity is more in Deep Learning algorithms requiring a larger dataset .

By summarizing the above literature review ,in this paper diabetes prediction will be performed using Linear Regression and SVM. The comparison results will be depicted in the further sections that would give an insight of the classical algorithms.

System Architecture

Architectureof diabetes prediction system is presented in figure 1.The algorithms used to predict diabetes are the Linear Regression and SVM.

Linear Regression: LR is a supervised learning algorithm that fits a linear equation in a dataset by modeling two or more variables.

SVM:SVM is a supervised learning algorithm that is implemented for classification and regression analysis.SVM classifies data points in multidimensional space into parallel lines known as hyperplane.

LR and SVM has been implemented as both perform great when dataset is less .In this paper, less dataset is considered.

Working model of Diabetes Prediction system architecture:

Features are extracted for the prediction of diabetes from the original dataset. In our paper, already extracted features are considered.. The features dataset is divided into Training Data,

Testing Data and Validation Data. The Training Data helps in learning algorithms by understanding data and its relationship with all the data. A model is created by the training data in machine learning .Hence the model data depends on the training data. The testing dataset is extracted from where the training data was taken. Once the data is trained in the trained dataset, to function the training dataset, testing set can be executed on it. Features are extracted from the training data and the target database is formed. Linear Regression and SVM are applied at the Classifier .Model is predicted for the testing data. Both classifier results are then compared for accuracy.

The statistical model helps in depicting the relationship between two variables. Two variables are represented based on predictor and response variables. Depending on the strength of the correlation of the variables, the quality of the linear regression depends[6].The process is explained in a much detailed method:

Dataset: The dataset considered for predicting diabetes are extracted from NC State University. The dataset contains 442 instances with 10 attributes. The feature set includes age, sex, bmi ,bp ,tc, ldl, hdl, tch, ltg, glu. Diabetes dataset is loaded and one feature set is used.

Training data and Test data :As discussed in preceding section, the training data helps to train a model to perform a specific action[2] .To train a model, 10 features

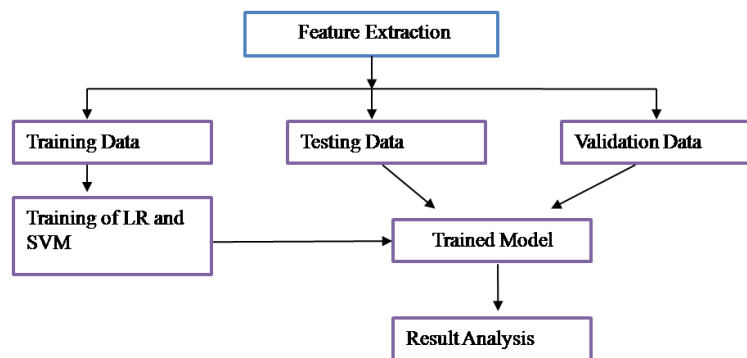


Figure 1. Diabetes Prediction system architecture.

are extracted. Among the 442 instances and 10 features,70% is used for training and 15% for testing and 15% for validation dataset. The data and target are splitted into training and testing datasets. Support Vector Machine (SVM) assists prediction and classification which falls under supervised learning.SVM first uses the nonlinear mapping function to map training data into higher dimensional space. To separate the data, linear regression is performed in the higher dimensional space[5].

Feature Extraction: Input information is transformed to the outcome of the features. The input class designs, square measures are considered as the characteristic of input design[5].

Feature Scaling: To avoid under fitting , over fitting curve, feature scaling is done. When a dataset has two different features BMI and Glucose, with different ranges, uniformity is not met

.To meet the uniformity, maximum range of the data is pulled out. Uniform range is always chosen to meet more accuracy.

Result and Discussion

This research paper has extracted an already available dataset from NC State University for Linear Regression and SVM. The feature set of dataset include age ,sex ,bmi ,bp,tc, ,ldl, hdl, tch, ltg, glu, ID, estimated salary, purchased etc. Further this paper clearly represented two algorithms i.e, Linear Regression, Support Vector Machine with obtained results.

In this experimental study, two machine learning algorithms Linear Regression and SVM were used. The dataset used was extracted from NC State University. The data was divided into training data(70%) and test data(15%).Figure 2. depicts the Linear Regression plot. The coefficient obtained was 38.23786125 with mean squared error of 2548.07.The determination of coefficient is found out to be $R_2 = 0.47$.The results are depicted in table .2

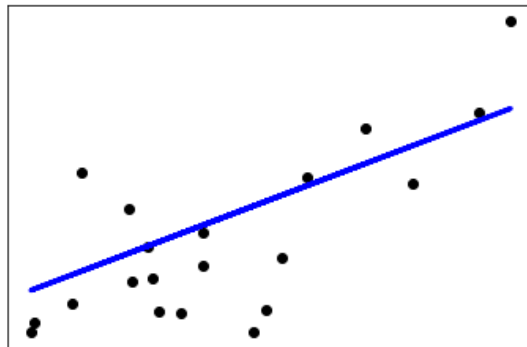


Figure 2. Linear Regression Plot

Table 2.Accuracy of Linear Regression

Linear Regression	Coefficient obtained	Mean squared error	Determination Coefficient
	38.23786125	2548.07	$R_2 = 0.47$

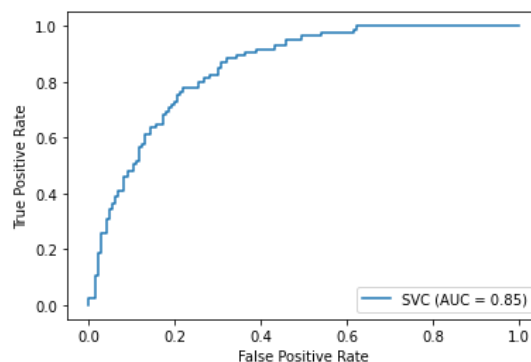


Figure 3.SVM Plot

Table 3. Accuracy of SVM

SVM	Precision	Recall	F1-Score	AUC
	0.63	0.823	0.71	0.77

Figure 3 depicts the SVM plot. The testing accuracy is found out to be 0.8290960451977402 with Training accuracy: 0.7575757575757576. Precision of 0.6306306306306306, Recall of 0.8235294117647058 with F1-Score 0.7142857142857143, ROC AUC Score 0.7713537469782433 is obtained. The accuracy results are depicted in table 3.

Conclusion

The dataset from NC State University is taken. Using the features, two algorithms i.e., Linear Regression and SVM have been derived. The determination of coefficient is found to be $R_2 = 0.47$ in Linear Regression. SVM works well with structured and semi-structured data like text, images, trees. The outcome of SVM using the same hospital dataset for testing accuracy is 0.82 and training accuracy is 0.75.

REFERENCES

- [1] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.
- [2] Augustine UhunomaOsarogiagbon, Faisal Khan, Ramachandran Venkatesan, Paul Gillard,
- [3] Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations, Process Safety and Environmental Protection, Volume 147, 2021, Pages 367-384, ISSN 0957-5820, <https://doi.org/10.1016/j.psep.2020.09.038>.
- [4] R. Sehly and M. Mezher, "Comparative Analysis of Classification Models for Pima Dataset," 2020 International Conference on Computing and Information Technology (ICCIT-1441), Tabuk, Saudi Arabia, 2020, pp. 1-5, doi: 10.1109/ICCIT-144147971.2020.9213821.
- [5] R. Deo and S. Panigrahi, "Performance Assessment of Machine Learning Based Models for Diabetes Prediction," 2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT), Bethesda, MD, USA, 2019, pp. 147-150, doi: 10.1109/HI-POCT45284.2019.8962811.
- [6] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.
- [7] Zhu, Changsheng, et al. "Improved Logistic Regression Model for Diabetes Prediction by Integrating PCA and K-Means Techniques." *Informatics in Medicine Unlocked*, vol. 17, 2019, p. 100179, 10.1016/j.imu.2019.100179. Accessed 1 Feb. 2021.

- [8] Jobeda Jamal Khanam, Simon Y. Foo, "A comparison of machine learning algorithms for diabetes prediction", *ICTExpress*, 2021, ISSN 24059595, <https://doi.org/10.1016/j.icte.2021.02.004>.
- [9] H. Alshamlan, H. B. Taleb and A. Al Sahow, "A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression," 2020 11th International Conference on Information and Communication Systems (ICICS), 2020, pp. 1-4, doi: 10.1109/ICICS49469.2020.239549.
- [10] R. Syed, R. K. Gupta and N. Pathik, "An Advance Tree Adaptive Data Classification for the Diabetes Disease Prediction," *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)*, 2018, pp. 1793-1798, doi: 10.1109/ICRIEECE44171.2018.9009180.
- [11] L. Miao, X. Guo, H. T. Abbas, K. A. Qaraqe and Q. H. Abbasi, "Using Machine Learning to Predict the Future Development of Disease," 2020 International Conference on UK-China Emerging Technologies (UCET), 2020, pp. 1-4, doi: 10.1109/UCET51115.2020.9205373.
- [12] A. Al-Zebari and A. Sengur, "Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection," 2019 1st International Informatics and Software Engineering Conference (UBMYK), 2019, pp. 1-4, doi: 10.1109/UBMYK48245.2019.8965542.
- [13] I. Chakour, Y. El Mourabit, C. Daoui and M. Baslam, "Multi-Agent System Based on Machine Learning for Early Diagnosis of Diabetes," 2020 IEEE 6th International Conference on Optimization and Applications (ICOA), 2020, pp. 1-6, doi: 10.1109/ICOA49421.2020.9094511.
- [14] Vijayakumar, Ponnusamy, and S. Malarvihi. "Green spectrum sharing: Genetic algorithm based SDR implementation." *Wireless Personal Communications* 94, no. 4 (2017): 2303-2324.
- [15] Ponnusamy, Vijayakumar, and S. Malarvihi. "Hardware Impairment Detection and Preshwhitening on MIMO Precoder for Spectrum Sharing." *Wireless Personal Communications* 96, no. 1 (2017).