International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

STATISTICAL DATA ANALYSIS USING MACHINE LEARNING CLASSIFIERS TO DEVELOP A PREDICTION MODEL FOR DIABETES PATIENTS

Shweta Kamble[1], Abhishek Devkhile[2], Pooja Phule[3] , Prof. Dnyaneshwar Kudande[4]

[1]Dept. of Computer Engineering, SRTTC-FOE, Pune, Maharashtra, India
[2]Dept. of Computer Engineering, SRTTC-FOE, Pune, Maharashtra, India
[3]Dept. of Computer Engineering, SRTTC-FOE, Pune, Maharashtra, India
[4]Dept. of Computer Engineering, SRTTC-FOE, Pune, Maharashtra, India

---

Abstract - Diabetes is a chronic condition that has the potential to wreak havoc on the global health-care system. According to the International Diabetes Federation, 382 million people worldwide suffer with diabetes. By 2035, this number will have risen to 592 million. Diabetes mellitus, or just diabetes, is a disease caused by a rise in blood glucose levels. Diagnosing diabetes can be done using a variety of traditional approaches based on physical and chemical tests. However, because of the complicated interdependence of different elements and the fact that diabetes affects human organs such as the kidney, eye, heart, nerves, and foot, early diabetes prediction is a difficult assignment for medical practitioners. Other scientific domains could benefit from data science methodologies. One of these responsibilities is to assist in the prediction of medical data. Machine learning is a new discipline of data science that studies how machines learn from their past experiences. The goal of this project is to create a system that can predict diabetes in a patient earlier and with more accuracy by merging the findings of various machine learning approaches. The goal of this research is to predict diabetes using supervised machine learning algorithms such as SVM, Logistic regression.

Key Words: Diabetes Prediction, Machine Learning, Deep Learning, Statistical Analysis, Data Visualization, Prediction Model

1. INTRODUCTION

Sugar and fat are abundant in most people's everyday diets. Diabetes risk has increased globally as a result of these variables. As a result, many people visit health centers to have blood tests done. However, many of them may not have even the tiniest chance of developing diabetes. These tests consume a significant amount of time and money from health-care organizations and individuals each year. There are studies underway to develop novel methods for diagnosing diabetes more quickly and cheaply, avoiding the need for blood testing in those who have a low risk of developing diabetes.

In medical prediction, machine learning techniques are commonly used. The learning algorithms employ previously recorded datasets of patient information to create a model, which is then used with data from an unknown patient to predict whether or not the patient has the targeted condition.

The goal of this research is to create a more accurate diabetes prediction classifier. As a result, we employ machine and deep learning algorithms to increase prediction system performance in terms of time, cost, and accuracy.

2. LITERATURE SURVEY

1. Prediction of Diabetes Using a Combination of Machine Learning Classifiers

Diabetes, often known as chronic sickness, is a collection of metabolic illnesses caused by a persistently high blood sugar level. If exact early prediction is achievable, the risk factor and severity of diabetes can be considerably decreased. In this literature, we are proposing a robust framework for diabetes prediction where the outlier rejection, filling the missing values, data standardization, feature selection, K-fold cross-validation, and different Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) were employed.

2. Classification of Diabetes using Deep Learning

Deep Learning (DL) is a research area that has different kinds of activation functions and their efficiency is flourished significantly in recent years and has shown remarkable reported by comparative analysis. Potential for artificial intelligence in the field of medical applications. The rest of paper is illustrated in respective manner: We have implemented the DL algorithm for the diabetes basics and background of deep learning techniques is declassification. This paper applied the Multi-Layer Feed Forward Neural Networks (MLFNN) for the diabetes classification.

3. Deep Belief Neural Network Model for Prediction of Diabetes Mellitus

Diabetes Mellitus is metabolic chronic disease in which blood glucose levels are too high. In India nearly 8.7% of population suffers from diabetes in age range from 20 to 70. Unidentified and untreated diabetes leads to so many health difficulties such as damage of heart, kidneys, eyes, nerves and blood vessels. There are already several methods exists to support clinical decision making but still need improvements to solve the issues and challenges.

4. A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques

With the continuing increase in the number of the deadly diseases that threaten both human health and life, medical Decision Support Systems (DSS) continue to prove their effectiveness in providing physicians and other healthcare professionals with support in clinical decision making. Among these dangerous diseases, diabetes continues to be one of the leading one that has caused several deaths in the world. It is characterized by an increase in blood sugar levels which can have severe effects on other human organs.

3. PROPOSED SYSTEM

In many real-world problems, classification is a critical decision-making tool. The primary goal of this project is to increase classification accuracy by classifying data as diabetes or non-diabetic. The greater the number of samples used in a classification task, the lower the classification accuracy. In many circumstances, the algorithm's performance is excellent in terms of speed, but the accuracy of data classification is poor. Our model's primary goal is to attain great accuracy. If we employ a large portion of the data set for training and only a small portion for testing, we can improve classification accuracy. This study looked into a variety of classification approaches for diabetic and non-diabetic data. As a result, approaches like Support Vector Machine, Logistic Regression, and Artificial Neural Network are shown to be the best appropriate for creating the Diabetes prediction system.

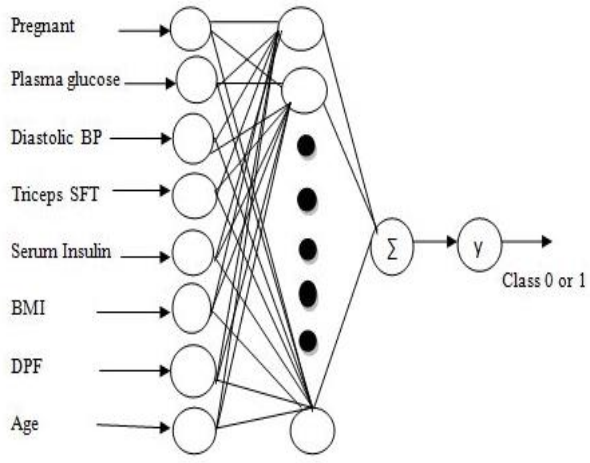


Fig -1: System Architecture

4. METHODOLOGIES

[1] Machine Learning: Machine learning is a branch of artificial intelligence (AI) that allows computers to learn and improve on their own without having to be explicitly programmed. Machine learning is concerned with the creation of computer programmes that can access data and learn on their own. The learning process starts with observations or data, such as examples, direct experience, or instruction, so that we can seek for patterns in data and make better decisions in the future based on the examples we provide. The fundamental goal is for computers to learn on their own, without the need for human involvement, and to change their behaviour accordingly.

Machine learning algorithms are often categorized as supervised or unsupervised.

a. Supervised machine learning algorithms can use labelled examples to apply what they've learned in the past to fresh data and predict future events. The learning algorithm creates an inferred function to generate predictions about the output values based on the examination of a known training dataset. After enough training, the system can provide targets for any new input. The learning algorithm can also compare its output to the correct, intended output and detect faults, allowing the model to be modified as needed.

b. Unsupervised machine learning techniques, on the other hand, are utilised when the data being trained is neither classed nor labelled. Unsupervised learning investigates how computers might infer a function from unlabeled data to describe a hidden structure. The system does not determine the correct output, but it examines the data and can infer hidden structures from unlabeled data using datasets.

c. Reinforcement machine learning algorithms are a type of learning algorithm that interacts with its surroundings by generating actions and detecting failures or rewards. The most important elements of reinforcement learning are trial and error search and delayed reward. This technology enables machines and software agents to automatically select the best behaviour in a given situation in order to improve their efficiency. For the agent to learn which action is better, simple reward feedback is required; this is known as the reinforcement signal.

[2] Classification: Classification is a supervised learning strategy in machine learning and statistics in which a computer programme learns from the data input supplied to it and then applies that learning to classify fresh observations. This data set could be bi-class (for example, determining whether the person is male or female or whether the mail is spam or non-spam) or it could be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. Here we have the types of classification algorithms in Machine Learning: Linear Classifiers: Logistic Regression, Naive Bayes Classifier, Nearest Neighbour, Support Vector Machines, Decision Trees, Boosted Trees, Random Forest, Neural Networks.

[3] Logistic Regression: Under the Supervised Learning approach, one of the most prominent Machine Learning algorithms is logistic regression. It's a method for predicting a categorical dependent variable from a set of independent variables. A categorical dependent variable's output is predicted using logistic regression. As a result, the result must be a discrete or categorical value. It might be Yes or No, 0 or 1, true or false, and so on.

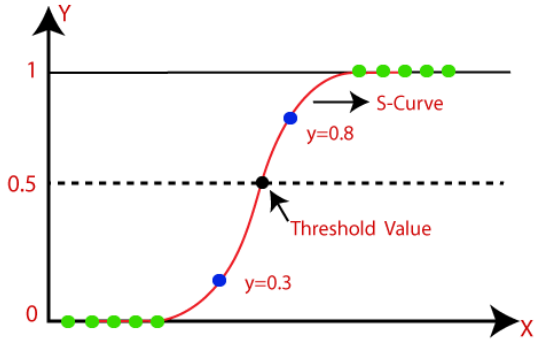


Fig -2: Logistic Regression

[4] Naïve Bayes: It's a classification method based on Bayes' Theorem with the assumption of predictor independence. A Naive Bayes classifier, in simple terms, posits that the existence of one feature in a class is unrelated to the presence of any other feature. For example, if a fruit is red, round, and roughly 3 inches in diameter, it is termed an apple. Even if these features are reliant on one another or on the presence of other features, they all add to the probability that this fruit is an apple, which is why it is called 'Naive'. The Naive Bayes model is simple to construct and is especially good for huge data sets. Naive Bayes is known to outperform even the most advanced classification systems due to its simplicity. Bayes theorem provides a way of calculating posterior probability P (c|x) from P(c), P(x) and P (x|c). Look at the equation below:

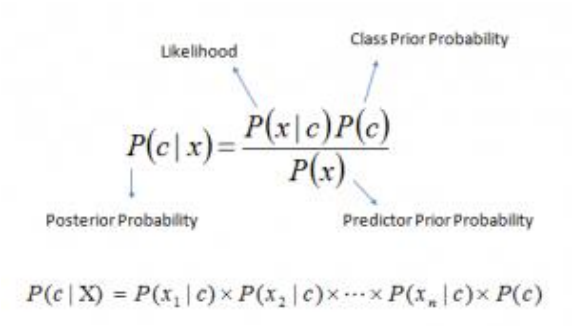


Fig -3: Naïve Bayes Formula

Above,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

[5] Support Vector Machine: SVM (Support Vector Machine) is a supervised machine learning technique that may be used to solve both classification and regression problems. It is, however, mostly employed to solve classification problems. The value of each feature is the value of a particular coordinate in this technique, which plots each data item as a point in n-dimensional space (where n is the number of features you have). Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).

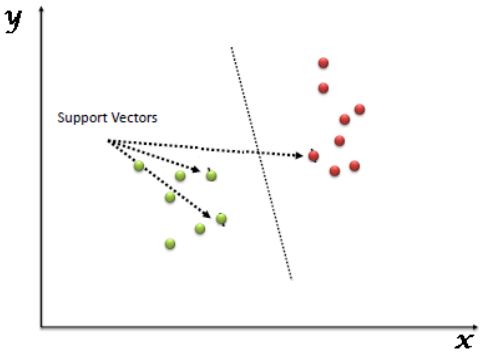
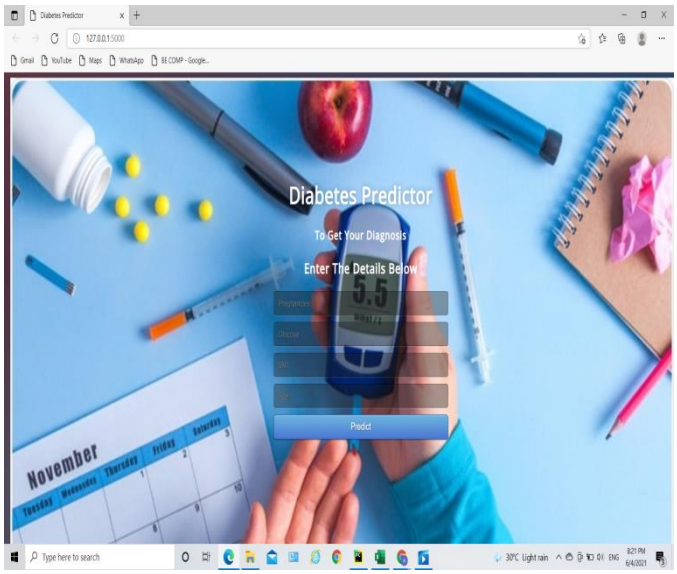


Fig -4: Support Vector Machine

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

5. RESULT AND DISCUSSION

In the proposed system, the novel approach is presented for diabetes classification using power of machine learning techniques. As diabetes is a chronic disease with the potential to cause a worldwide health care crisis, it's very important to address this disease with proper measures. We have implemented web based system for diabetes classification using Python along with Flask framework. We have created simple user interface so that anyone can use it and make benefit out of this. User need to add basic details and our system will predict the result in just few seconds with simple popup message.



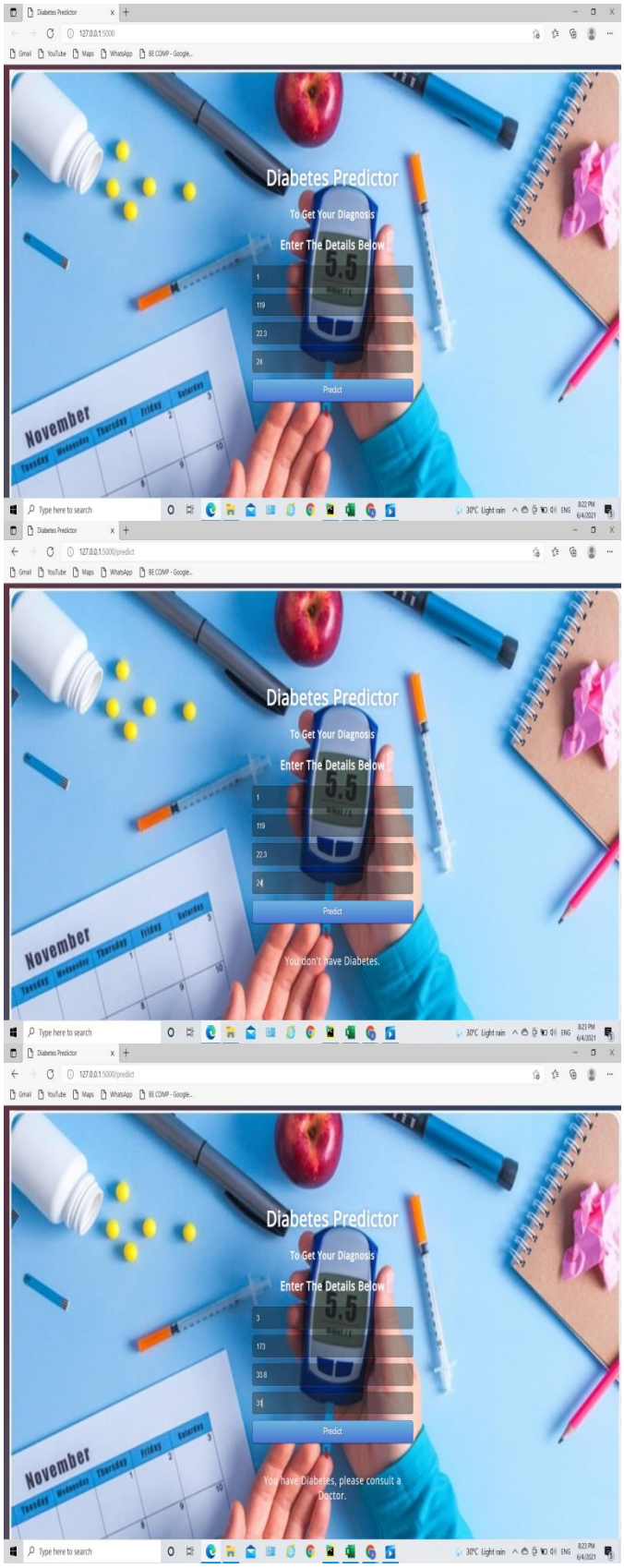


Fig -4: Results

Comparative results of existing and proposed system is as follow,

Table -1: Comparative Results

Parameters	**Existing System**	**Proposed System**
Diabetes classification using web system	No	Yes
Accuracy	70-80	90+
Time efficient	No	Yes
Dataset Support	No	Yes
Cost	More	Less

With reference to Table-1 it is clear that we overcome various problems in existing system and our approach works efficiently.

6. CONCLUSIONS

Diabetes is the most common disease in the world. In our study, we compare several algorithms with different pre-processing techniques and identify algorithms best performance in which pre-processing technique. We found many of machine learning will give us best accuracy than any other methods. In future we will apply more advanced tricks in Neural Network, such as more hidden layers, algorithm optimization would be more accurate in this case.

ACKNOWLEDGEMENT

I wish to express my profound thanks to all who helped us directly or indirectly in making this paper. Finally I wish to thank to all our friends and well-wishers who supported us in completing this paper successfully. I am especially grateful to our guide Prof. Dnyaneshwar V. Kudande for his time to time, very much needed, and valuable guidance. Without the full support and cheerful encouragement of my guide, the paper would not have been completed on time.

REFERENCES

1. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers
2. Classification of Diabetes using Deep Learning
3. Deep Belief Neural Network Model for Prediction of Diabetes Mellitus
4. A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques
5. 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.
6. P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karurangaa, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova,

J. E. Shaw, D. Bright, R. Williams, and IDF Diabetes Atlas Committee, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation," Diabetes Research and Clinical Practice, vol. 157, pp. 107843, Nov. 2019.

7. M. Maniruzzaman, M. J. Rahman, M. A. M. Hasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, "Accurate diabetes risk stratification using machine learning: role of missing value and outliers," Journal of Medical Systems, vol. 42, no. 5, pp. 92, May 2018.
8. 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.
9. 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.