Diabetes Prediction Using Machine Learning

JSPM'S Jayawantrao Sawant College of Engineering, Hadapsar, pune.
Department of Information Technology

Prof.V.V.Kalunge
vishvaskalumge@jspmjscoe.edu.in

Kalpesh Sonawane
Sonawanekalpesh489@gmail.com

Rohan Bhonsle
rohanbhonsle45@gmail.com

Saurav More
sauravmore06@gmail.com

Nikita Bhosle
bhoslenikita17@gmail.com

Abstract: Chronic diseases like diabetes have the potential to wreck the world's health care system. According to the International Diabetes Federation, there are 382 million diabetics worldwide. By 2035, this will increase to 592 million. Diabetes is a disease characterised by high blood glucose levels. The signs of this raised blood sugar level include increased thirst, appetite, and frequency of urinating. Diabetes is a significant contributing factor to heart failure, stroke, kidney failure, amputations, blindness, and kidney failure. When we eat, our bodies turn the food we consume into sugars like glucose. Then, we anticipate insulin to be released from our pancreas. Insulin functions as a key to unlock our cells, allowing glucose to enter and be used as fuel by us. However, in diabetes, this mechanism does not work. The most common types of diabetes are type 1 and type 2, but there are others, including gestational diabetes, which develops during pregnancy. Machine learning is a new area in data science that investigates how machines learn from experience. The purpose of this work is to develop a system that, by combining the findings of different machine learning algorithms, can more correctly identify early diabetes in a patient. Some of the approaches used include Logistic Regression, Random Forest, Support Vector Machine, and the Nave Bayes Algorithm. The accuracy of each algorithm is computed alongside.

INTRODUCTION

Diabetes is the disease that is spreading the fastest, even among young people. Understanding what happens in the body when there is no diabetes is critical to understanding diabetes and how it develops. Sugar (glucose) is derived from our diets, particularly those heavy in carbohydrates. Carbohydrates are essential for everyone, including diabetics, because they are the body's major source of energy. Diabetes is the fastest spreading disease, especially among young people. awareness diabetes and how it develops requires an awareness of what happens in the body when there is no diabetes. Our diets, particularly those high in carbohydrates, provide sugar (glucose). Carbohydrates are necessary for everyone, including diabetics, because they are the body's primary energy source. The rest of the glucose is taken up by the cells of our body. does not produce enough insulin (insulin deficiency) or if the body cannot use the insulin produced (insulin resistance), glucose accumulates in the bloodstream hyperglycemia and diabetes occurs. Diabetes Mellitus means high levels of sugar (glucose) in the bloodstream and in the urine. Types of diabetes Type 1 diabetes means that the immune system is weakened and the cells do not produce enough insulin. There is no conclusive research to show the root causes of type 1 diabetes, but currently there are no preventive measures.

Type 2 diabetes is characterized by either insufficient production of insulin by cells or poor use of insulin in the body. 90% of diabetics have this type of diabetes, so it is the most common type. Both genetic and lifestyle factors contribute to its occurrence. Pregnant women who acquire high blood sugar quickly get gestational diabetes. It will recur in two-thirds of the instances during consecutive pregnancies. After a pregnancy in which gestational diabetes was present, there is a high likelihood that type 1 or type 2 diabetes may develop.

A. *Symptoms of Diabetes:*

- Frequent Urination
- Increased thirst
- Tired/Sleepiness
- Weight loss
- Blurred vision
- Mood swings
- Confusion and difficulty concentrating
- Frequent infections

B. *Diabetes Causes:*

The main cause of diabetes is genetics. It is caused by at least two faulty genes on chromosome 6, a chromosome that affects how the body responds to various antigens. A viral infection can also affect the development of type 1 and type 2 diabetes. Studies show that viruses such as hepatitis B, CMV, mumps, rubella and coxsackie virus increase the risk of developing diabetes. Pancreatitis, pancreatic cancer and trauma can all damage beta cells or reduce their ability to produce insulin, leading to diabetes. When the damaged pancreas is removed, diabetes occurs due to the loss of beta cells. If there is a family

history of diabetes, a woman is more likely to develop gestational diabetes

Literature Review

Aiswarya E . [1] aims to find solutions for diabetes detection by exploring and exploring patterns in data through classification analysis using decision tree and Naive Bayes algorithms. The research is expected to provide a faster and more effective method to detect the disease, which will help patients recover in time. Using the PIMA dataset and cross-validation, the study found that the J48 algorithm yields 74.8 percent accuracy, while Naive Bayes yields 79.5 percent accuracy using a 70:30 distribution.

Lee et. [2] focuses on applying a decision tree algorithm called CART to a diabetes data set after adding a sampling filter to the data. The author emphasizes the problem of class imbalance and the need to address this problem before implementing any algorithm to achieve better accuracy. Categorical imbalance mostly occurs in data with dichotomous values, which means that a categorical variable has two possible outcomes and can be easily handled if detected in the data processing stage and helps to improve the accuracy of the prediction model.

Gupta S. [3] aims to find and calculate the accuracy, sensitivity and specificity rates of many classification methods and also tries to compare and analyze the results of several classification methods in WEKA, the study compares the performance of the same classifiers when implemented on some other tools which includes Rapidminer and Matlabusing the same parameters (i.e. accuracy, sensitivity and specificity). They used JRIP, Jgraft and BayesNet algorithms. The result shows that Jgraft shows the highest accuracy of 81.3%, sensitivity of 59.7% and specificity of 81.4%. It was also concluded that WEKA performs better than Matlab and Rapidminer.

METHODOLOGY

Dataset Description

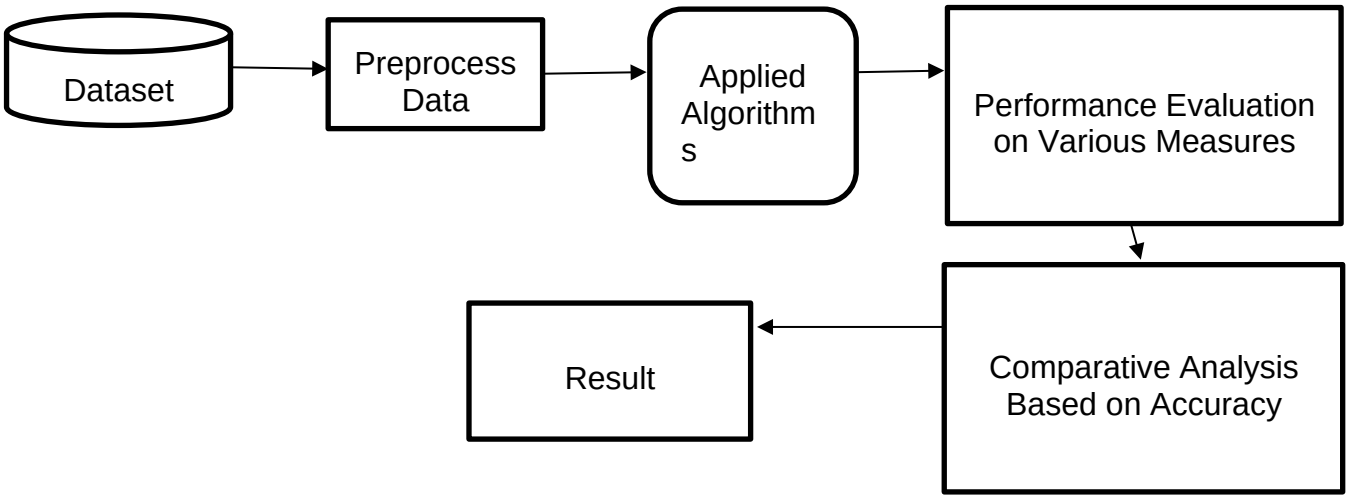
Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

* *The diabetes data set consists of 2000 data points, with 9 features each*
* "Outcome" is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.

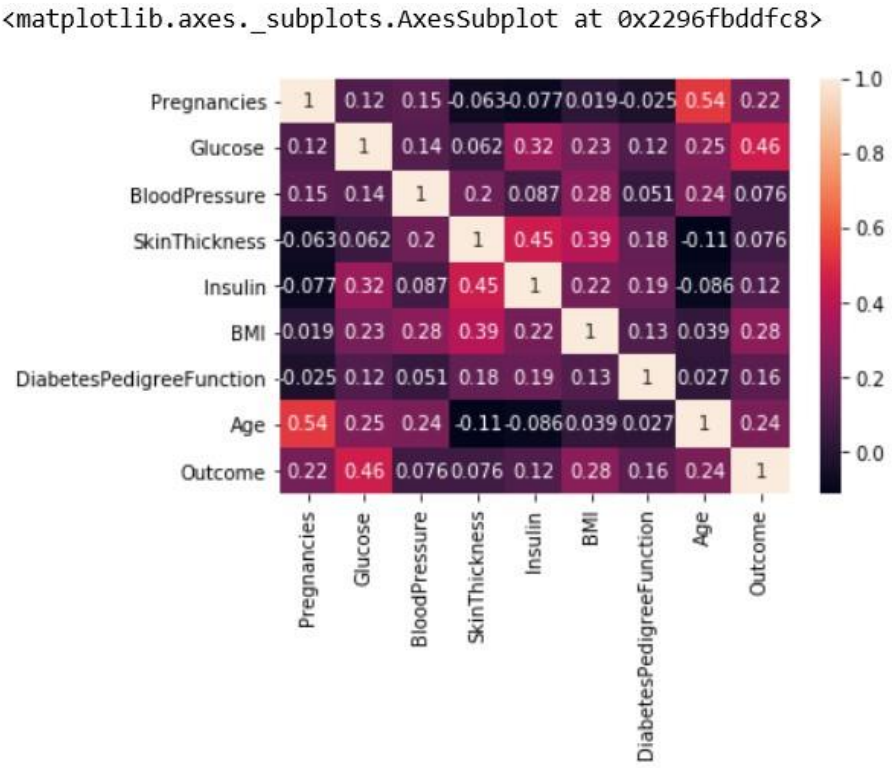
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

There is no null values in dataset.

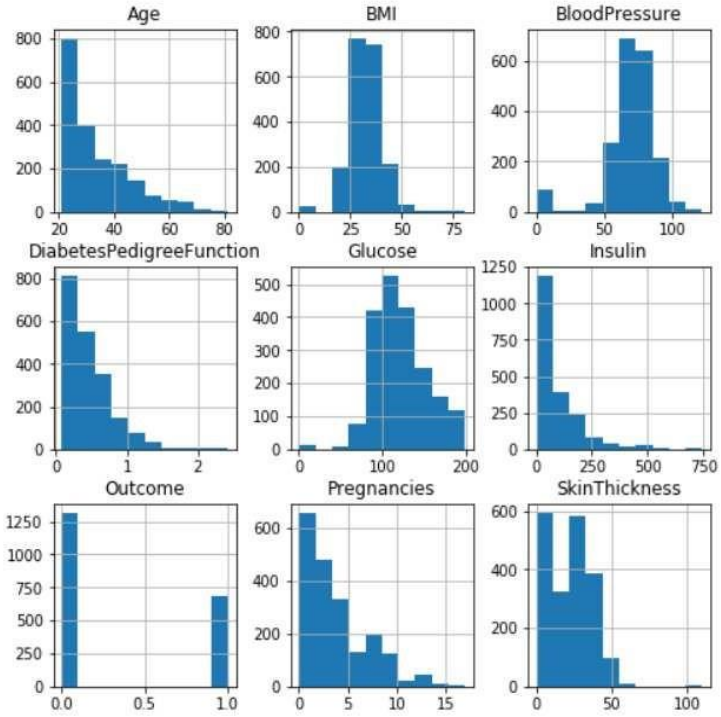


Proposed Model Diagram

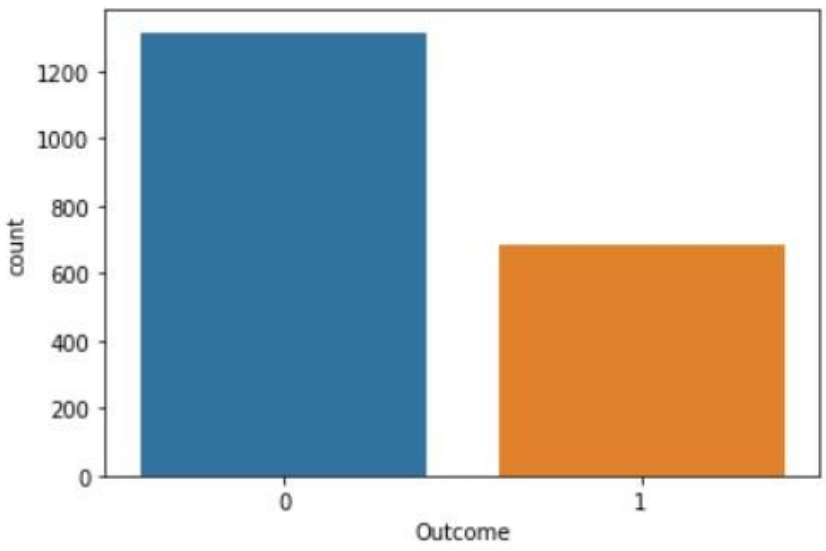
RESULT & DISCUSSION



It's easy to see that there's not a single feature that's very correlated with our results. Some of these characteristics have a negative correlation with the endpoint value, while some are positive. Histogram:



Let's see the plot. It shows how each feature and label is distributed over different regions, further confirming the need for scaling. Then when you see discrete bars, it basically means that each one is actually a categorical variable. Before applying machine learning, we need to process these categorical variables. The result label has two classes, 0 for no disease and 1 for disease.



The chart above shows that the data is skewed towards data points with a score of 0. This means that diabetes did not actually exist. The number of non-diabetics is almost double that of diabetics.

Algorithms:

Support Vector Machine (SVM)

SVM is a division of Supervised Learning Algorithms. The strategy used to perform regression, classification and outlier detection of data. SVM will be grouping the information dependent on the hyperplane.

The hyper plane is used to totally isolate the two classes in the best way and the most extreme edge hyper plane ought to be picked as a best separator.

The two types of SVM Classifiers that have been used are: Linear Classifier and Non-Linear Classifier.

Accuracy:78%

Random Forest

The outfit learning technique used for the classification and regression operates by constructing the multitude of decision trees at training time and outputting the class i.e mode of the classes or the regression of the individual trees. Irregular choice woods right for choice trees propensity which is used for over fitting on to their preparation set.

Accuracy:74%

Naive Bayes algorithm:

It is a classification method built on the Bayes Theorem with the assumption of predictor independence. A Naive Bayes classifier, to put it simply, believes that the presence of one feature in a class is unrelated to the presence of any other features. To determine which class a test point belongs to, it employs probability. A purely statistical model, naive Bayes. Because it is presumed that the features and attributes in the datasets are independent of one another, this approach is known as naive.

Accuracy:77%

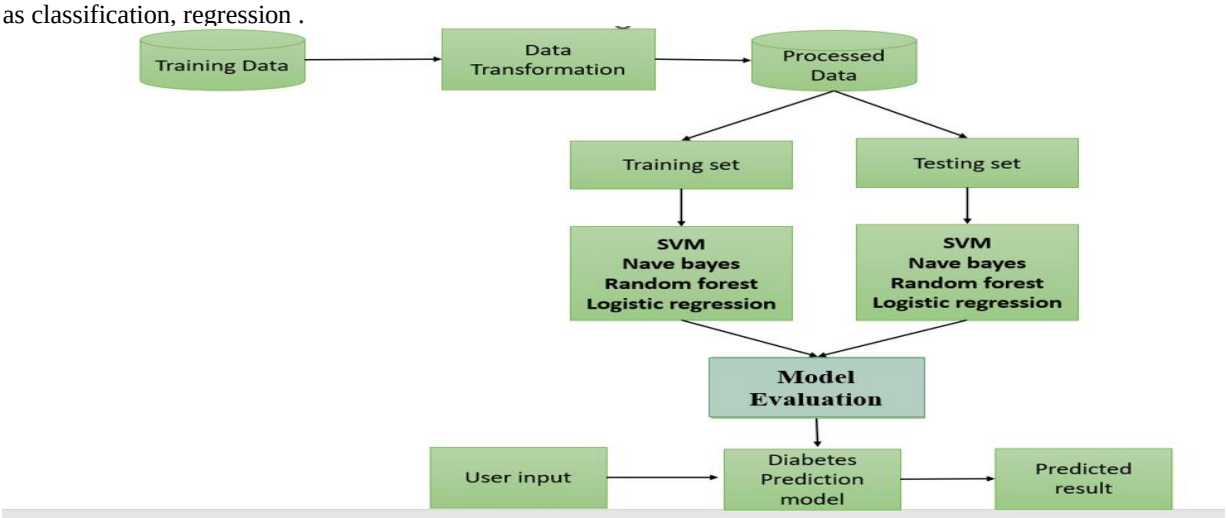
Logistic regression:

It is a classification technique based on predictor independence and the Bayes Theorem. Simply expressed, a Naive Bayes classifier thinks that the existence of one feature in a class has no bearing on the existence of any other characteristics. It uses probability to determine which class a test point belongs to. a naïve Bayes model that is solely statistical. This strategy is referred to as naive since it assumes that the features and attributes in the datasets are independent of one another

Accuracy:75%

PROPOSED DIAGRAM

To predict the diabetic patient, we offer a classification model with improved accuracy. We used several MLtechniques such as classification, regression .



CONCLUSION AND FUTURE WORK

One of the primary medical difficulties that exists in real life is the early detection of diabetes. In this study, concerted efforts are made to develop a system that can anticipate diabetes. This research examines and evaluates five machine learning classification methods utilising a range of measures. Experiments are carried out using the PIMA Indian Diabetes Database. In our project, the outcome is either Yes or No. We employ a time prediction module if the outcome is classed as No. Time Prediction - In this section, we forecast the "time" of onset of the diabetic condition. We examine the diabetes prediction result and examine the accuracy of the diabetes prediction, the time required to compute the accuracy of the diabetes prediction, properly classified and erroneously classified patients The categorization of diabetes prognosis results. We employed naive Bayes, SVM, logistic regression, and the Random forest Algorithm to predict diabetes, with the outcome categorised as Yes or No, and the same KNN Algorithm is used for the time prediction module. More diseases may be predicted or detected in the future using the developed system and machine learning classification algorithms.

ACKNOWLEDGEMENT

This assignment was completed under the supervision of Mr. V. Kalunge (Professor and Head, Department of Information Technology & Engineering, Jaywantrao Sawant College of Engineering, Pune). We would want to express our gratitude to our mentors for inspiring us to complete this task and write this post. I would not have made any progress on this thesis without their active leadership, aid, participation, and encouragement. We are really grateful for their important advise and assistance in completing this report.

REFERENCES

• Azra Ramezankhani, Omid Pournik, Jamal Shahrabi, Fereidoun Azizi and Farzad Hadaegh, ”An Application of Association Rule Mining to Extract Risk Pattern for Type 2 Diabetes Using Tehran Lipid and Glucose Study Database”, Int J Endocrinol Metab, April 2015.

• Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

• Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584- 1589). IEEE.

• Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451– 455

• M. F. Faruque, Asaduzzaman and I. H. Sarker, ”Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus”, 2019 International Conference on Electrical Computer and Communication Engineering (ECCE), pp. 1-4, 2019.