International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 11 | Nov 2019 www.irjet.net p-ISSN: 2395-0072

Comparison of Techniques for Diabetes Detection in females using Machine Learning

Aditya Dhall[1], Sarthak Kar[2], A. Nagaraja Rao[3]

[1,2]UG Scholar, Vellore Institute of Technology, University, Vellore, Tamil Nadu
[3]Associate Professor, Department of Computational Intelligence, Vellore Institute of Technology, University, Vellore, Tamil Nadu, India

---

Abstract - *Diabetes is a cause of immense concern these days. Everybody adjusts their diets in such a way so as to avoid diabetes completely. This paper discusses the different factors present in the human body and how they indicate the presence of diabetes in women. It discusses the different machine learning techniques that have been employed in this project to predict whether the patient has diabetes or not. This paper also highlights the advantages of each of the algorithms used and how they are better suited for this problem.*

1. INTRODUCTION

Due to the immense advances that have taken place in health sciences and biotechnology, there has been a sudden availability in data that indicate the different factors that affect the likelihood of a person to get diabetes. Diabetes is an extremely common, and one of the most widespread conditions harming the bodily functions of many humans and also results in a lot of deaths. Due to such widespread effects of this disease, many specialists have researched this data extensively and as a result, a lot of data has been derived. In this day and age, the presence of Machine Learning in the field is more than it ever has been. In this project, we will be using certain concepts of Machine Learning on a given dataset that takes certain factors that affect the likelihood of diabetes into account and predict whether or not the *female,* who's details have been entered, has diabetes or not.

2. LITERATURE REVIEW

Diabetes is a disease that has spread all around this world and affected countless people. It occurs when the blood glucose or the blood sugar level is higher than the advised amount. Blood glucose comes from the food that you consume and is your main source of energy. A hormone, made by the pancreas in your body called insulin is responsible for helping the glucose in the food to get into your cells, which will later be used as energy by your body. At times, one's body may fail to create sufficient or any insulin or isn't able to use the insulin well. When this happens, glucose ends up staying in the blood cells and ends up accumulating. This condition is called diabetes. The presence of excess glucose in your body can cause some adverse health conditions over time, such as heart disease, strokes, kidney diseases, eye problems, nerve damage and many more.

Over time, after immense research that has been conducted across the world, some basic factors have emerged as some major indicators of whether or not a patient is likely to have diabetes. The number of pregnancies helps in the detection of diabetes. According to research conducted by scientists, it was noted that women who had at least four pregnancies were more likely to have diabetes when compared to women with one to three pregnancies. On the contrary, it was also noted that women with one to three pregnancies had a higher likelihood to have diabetes than women with no pregnancies at all. This observation seems to have baffled scientists who are unable to come up with a proper explanation for this behavior. Another contributing factor is the glucose level. If the glucose level present in ones blood stream is above a certain level indicates that she is highly likely to have diabetes. On the other hand, it is also possible that a person's glucose level is slightly below the given level considered harmful, but they still are diagnosed with diabetes. According to research, people with higher blood pleasure are 50% more likely to have diabetes. In a study conducted by scientists, it was shown that skin thickening was found in 22% of the patients with diabetes and 4% of the patients without diabetes. Low insulin levels in a person indicate an increase in the likelihood of a person to have diabetes. Studies suggest that the chances of having diabetes increase with the increase in body mass index (BMI) of the person. Over the years, scientists have made the following observations. Among people between the ages of 20-40 years, an estimated 3.7% of the people were diagnosed with diabetes. Similarly, among people between the ages of 45-64 and at least 65 have an estimated 13.7% and 26.9% respectively, of confirmed cases of diabetes. Another factor we have included is the *Diabetes Pedigree Function*. It is a function that scores the likelihood of a person to have diabetes based on the family history provided.

3. METHODOLGY

Approach is based on research conducted by Md. Aminul Islam and Nusrat Jahan in their paper [4] 'Prediction *of Onset Diabetes using Machine Learning Techniques'.*

The dataset consists of the following columns

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test

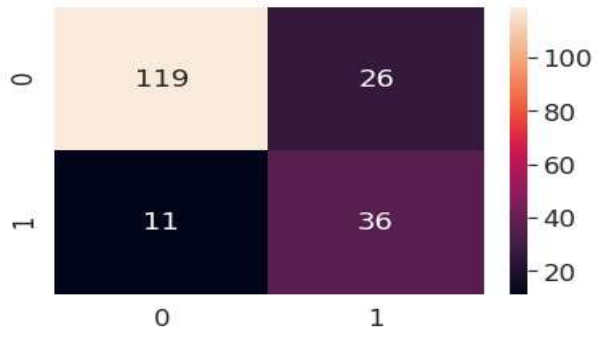
- Blood Pressure: Diastolic blood pressure (mm Hg)
- Skin Thickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- Diabetes Pedigree Function: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. In logistic regression we measure the relationship between the categorical dependent variable or one or more independent variable by estimating the probabilities using a logistic or sigmoid function. It is mostly used to predict an outcome of binary values.

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

In case of the above a and b are ppregnancies and glucose, pregnancies and blood pressure, pregnancies and skin thickness, pregnancies and insulin, glucose and blood pressure, glucose and skin thickness blood pressure and skin thickness, insulin and BMI, age and outcome, so on and so forth

[5] We chose logistic regression model because, the predicted values will be more than one and less than zero. We assume that in linear regression that the variance of Y is constant across X which is not possible. In linear regression model K=PQ where K is constant. As we move to more extreme values, the variance will increase and approaching zero, it will decrease.



We also use decision tree in these kinds analysis. It divides the decision into smaller and smaller subset in order to take correct decisions. Decision tree assembles regression or characterization models as a tree structure. It breaks a dataset into littler and littler subsets while simultaneously a related decision tree is steadily created. The conclusive outcome is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has at least two branches (e.g., Sunny, Overcast and Rainy), each speaking to values for the characteristic tried. Leaf node (e.g., Hours Played) speaks to a decision on the numerical objective. The highest decision node in a tree which compares to the best indicator called root node. Decision trees can deal with both downright and numerical information. Then in this decision we try to predict how much changes we can make it will break down the problem into many sub-problems like the whole model will be breaking down into many factors like how BMI and no of pregnancies will affect in diabetes.

[2] There are two ways to build a decision tree .i.e. CART which uses gini index classification and metric and ID3 which uses entropy function and information as metrics. In this case, we use ID3 algorithm.

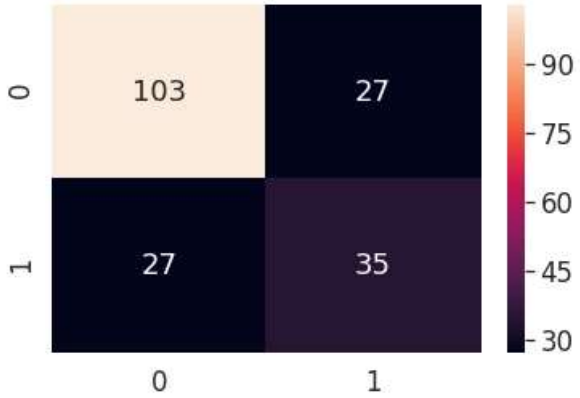
It works in the following fashion: it will first predict which of the parameters is going to distinguish the most. It does this by determining which parameter we can distinguish more data in the group. Then we will again calculate the second highest parameter using the highest parameter as constant. Then so on and so forth until all the parameters are done. We need to compute the entropy for each parameter form dataset. First, we need to calculate the entropy for all categorical values then take the average information provided for the current attribute. After that we need to calculate gain for the current attribute. We use the decision tree because of the fact that it mimics the human level of thinking, so it is simple to understand the data and make some and good interpretations. It helps us to see the logic of the data interpret.

$$Entropy = -\sum p(X)\log p(X)$$

Here, x is the ratio of number of outcomes to the total no of outcomes.

$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

This is the difference between the total entropy and the entropy which satisfies outcome x.



4. COMPARATIVE ANALYSIS

There have been many other studies that have been conducted on detecting diabetes using various machine learning techniques. All these data scientists have employed a wide range and combinations of existing machine learning techniques and have applied them to existing, or fresh datasets. All these datasets consist of a different set of parameters or attributes that have been recorded. Due to the availability of such convenient tools, studies can be conducted more easily and accurately than ever before. We have used the Prima Indians dataset which consists of the above discussed data for women at least 21 years of age.

[3]Another similar study was conducted by Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang. They attempted to develop a method to detect diabetes from *hospital physical examination data* in Luzho, China. They had two different datasets available to them. They made complete use of one of these datasets with 164431 instances present in the training set and chose 13700 samples from the other dataset as a test set. There are a total of 14 attributes that have been recorded in these datasets. They are pulse rate, age, breathe, right systolic pressure (RSP), right diastolic pressure (RDP), left systolic pressure (LSP), left diastolic pressure (LDP), height, physique index, weight, waistline, fasting glucose, high density lipoprotein (HDL) and low density lipoprotein (LDP). On cleaning the data, it was discovered that only 151598 and 69082 data was viable for the training and sample test respectively. Similar to us, they also made use of the decision tree algorithm which makes use of the tree structure to classify instances based on the features provided. Apart from this, they also make use of Random Forest algorithm and Neural Networks Algorithm. They have also performed a similar approach to the Prima Indians dataset.

Random Forest is an algorithm that performs classification by utilizing many decision trees. It has the capability to perform both classification and regression simultaneously. Random Forest algorithm repeated comes up in the field of applied machine learning in biomedical research. While a new object is being predicted by Random Forest algorithm, given certain attributes, all the decision trees present in the Random Forest will perform their own classification and cast a vote. The output is decided by taking the one with the largest taxonomy. *Neural Networks* is a model based on mathematics. It is designed in such a way that it imitates the working of an animal's neural networks. This algorithm takes advantage of the complexity of the system. It is able to process the information by altering how each internal node is related to each other. As discussed before, since they noticed data imbalance, they had to take the output 5 different times and then took their average as the result.

Table 1: Comparing Accuracy of Correct Predictions made by the different approaches on different datasets.

Category	Dataset	Classifier Name	Accuracy
The discussed approached	Prima Indians	Logistic Regression	80.72
		Decision Tree	75.0
[3]		Random Forest	76.04
		Neural Network	76.7
	Luzuhou	Random Forest	80.84

From the above table, it is evident that Logistic Regression on Prima India dataset and Random Forest on Luzuhou dataset have approximately the same accuracy. But, since Prima Indians dataset consists of much lesser attributes, training this data is much more cost efficient than training Luzuhou. As a result, it is safe to say that Logistic Regression on Prima Indians is the best available approach to predict diabetes among the approaches discussed.

5. CONCLUSION

Diabetes is a widely spread condition that has affected millions of lives and shows no signs of stopping. Researchers have predicted that at the current rate, by the year 2040, there will be approximately 640 million reported cases of diabetes all across the world. Another way to put this in perspective is that 1 in every 10 adults will be suffering from diabetes. Such widespread effects of this disease make this topic an important field of research. Due to the sudden enhancement in machine learning, it is becoming much easier to reach a solution. By conducting such researches, symptoms that indicate the early stages of diabetes may help medical researchers in finding a cure for this disease.

6. REFERENCES

[1] Kavakiotis, Ioannis, et al. "Machine learning and data mining methods in diabetes research Computational and structural biotechnology journal 15 (2017): 104-116

[2] Hanna, Wedad, et al. "Pathologic features of diabetic thick skin." Journal of the American Academy of Dermatology 16.3 (1987): 546-553.

[3] Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018).

[4] Islam, Md Aminul, and Nusrat Jahan. "Prediction of Onset Diabetes using Machine Learning Techniques." International Journal of Computer Applications 975: 8887.

[5]http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html

[6] https://medium.com/deep-math-machine-learning-ai