Diabetes Prediction Using Machine Learning

Sethupathi M[1] and Privietha P[2]

[1]*MCA Student, Department of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India.*
[2]*Assistant Professor, Department of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India.*
[1]*sethupathioffl@gmail.com* , [2]*priviethaprabhakar@gmail.com*

Abstract. The project entitled "DIABETES PREDICTION USING MACHINE LEARNING" is based on Data analytics. Diabetes is a long-standing disease with the likely to cause a global health care crisis. As per International Diabetes Federation (IDF), 382000 thousand people are living with diabetes over the entire world. After a decade, this will be doubled as 592000 thousand. Diabetes is caused due to the acceleration of blood glucose. The higher levels of sugar produce the sign of frequent urination, high levels of thirst and hunger. Diabetes is one of the main causes of myopia, kidney failure, amputations, heart failure and stroke. When we eat food, our body turns it into sucrose, or glucose. At that point, our pancreas is about to release insulin. Insulin serves as a manager to open our cells and to allow the glucose to enter and permit us to use the glucose for energy. However, with diabetes, this procedure will not be completed. Type 1 and type 2 diabetes are the most common forms of the disorder, but there are also other types, such as gestational diabetes, which occurs during gestation, as well as other forms. Machine learning is a booming scientific technology in data science dealing with the ways in which machines understands from experience. The goal of this project is to build a system, which can do early prediction of diabetes for a patient with a higher level of accuracy by combining the results of various machine-learning aspects. The algorithms like K-means algorithm, Random Forest, Logistic Regression, Support Vector Machine and Decision tree are used in this project. The correctness of the model based on algorithms is calculated and the unique model is taken for prediction of diabetes.

Keywords: Diabetes Prediction, Machine Learning, SVM, Logistic Regression.

1. Introduction

Diabetes is a chronic illness that is caused due to higher amount of blood sugar in it. Human body needs energy, and glucose and it is one of the major sources of energy to build the muscles and tissues of the body. In general, unhealthy lifestyle and lack of workouts for body are the main causes of type 2 diabetes. That means presence of excessive amount of sugar in the blood results in diabetes. Sometimes, the pancreas is not able to convert the food into insulin.

Since the sugar remains unabsorbed, it causes diabetes. Diabetes will affect kidneys, nervous system, eyes, blood vessels, and so on. Diabetes is of three types. The First type is juvenile diabetes, which occurs mostly in infants and children and kills the cells which produce insulin in the pancreas. Second is type 2 diabetes, which generally happens after the age of 40 due to the lack of exercise and unhealthy lifestyle. Diabetes is a type of disease that cannot be reversed but can be controlled with the help of medicines, regular walk and healthy lifestyle, exercise, and by maintaining a proper diet. Type 2 diabetes is also known as insulin-independent diabetes since patients are not required to be injected with insulin after a gap of regular intervals, but in the case of type 1 diabetes, insulin will be injected at a regular interval of time to the patient, so this is known as insulin-dependent diabetes. The third type of diabetes is gestations,

which occurs during pregnancy time due to the change of hormones, and this generally disappears after the baby born.

There is another type, that is known as pre-diabetes, in which the intake levels of sugar is on the borderline, and this condition can be cured with the help of regular workouts and by maintaining a healthy lifestyle. In this study, we have tried machine learning algorithms to predict diabetes. Machine learning is known as a branch of artificial intelligence (AI) in which the machine predicts the result based on certain data and previous outcomes or datasets. Machine learning is of two types. First type is supervised learning, in which data act as a tutor and the model is built around the dataset. Second type is unsupervised learning, in which data will be trained by itself by finding certain patterns in the dataset and labelling them. In modern years, many authors have published and produced their work on diabetes prediction using machine learning algorithms. In this study different prediction methods are used using machine learning and presented a comparative study of few methods.

2. Related Work

Machine learning is one the branches of AI. At a high level, Machine Learning is about to teach a computer how to make accurate and exact predictions when it is fed with data [1][6][7].

For instance, such a system could detect whether an orange or an apple is in a picture. It can spot people those crosses the road in front of a self-driven tramp. ML can also differentiate important emails from spam. It can even recognize speech to provide captions on YouTube. [6][8]

Deep learning comprises many artificial intelligence (AI) applications and services that improves the elimination of human efforts, performs analytical and physical tasks without any human intervention. [6][9] Deep learning technology lies with the everyday products and services (such as voice-enabled TV remotes, digital assistants, and credit card fraud detection) as well as emerging technologies (such as self-driving vehicles). [7]

3. Methodology

The study deals with the advancements in machine learning applied to Logistic regression [3][5][10]. First of all, the data has been collected from online file with extension.CSV. The data has been split into 32% testing and 68% training data. Then the logistic regression model will be applied to the obtained data. Results will be obtained and generated based on the accuracy classification. Figure 1 represents the flow diagram of the methodoloy to prepare the model.

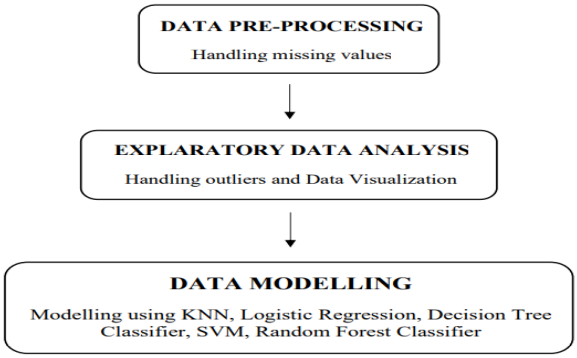

Figure 1: Methodology.

a) Collection of data:

Most relevant input data has been obtained, data can be obtained from any sources. In this research study the datasets are collected from various websites. [5][9]

b) Lung cancer dataset as input:

In this stage we will provide the dataset as an input to the proposed system such that it gives the results. The dataset is divided into two, Training and testing datasets. [3][5]

c) Feature selection:

The main aim of feature selection policy is to identify the inputs or features which are correlated with output values where the values are dependent and they lie upon a specific input which is collected by applying some useful test.

d) Splitting dataset:

In this stage, the Diabetes dataset has 768 test cases where the dataset is divided into train and test data. Many combinations of train and test data have been done.

e) perform lr technique on training datasets:

Logistic Regression is a popular mathematical modeling procedure used in the analysis of epidemiologic datasets, especially in the field of machine learning. Logistic Regression is a method mainly used here for classification of datasets.

4. Dataset

The data used in this study was taken from one of the online repositories. This is the data collected from the websites of the Diabetes prediction.

The used dataset comprises of 8 attributes, 768 rows and 9 columns of data. The attributes include “Pregnancy”, “Blood Pressure”, “Glucose”, “Skin Thickness”, “Insulin”, “BMI”, “Age”, and “Diabetes Pedigree Function”. Furthermore, pre-processing data is processed only by checking the missing values and it turns out if there is no missing value in the dataset.

There are various aspects to deal with irrelevant classification problems for imbalanced data, such as resampling the training data and developing other versions of the existing machine learning algorithms specific use. The finest way to improve the accuracy for an imbalanced dataset in implementing classification techniques using the linear and logistic model is optimizing the threshold parameter[5][11]. The threshold parameter is known as a tolerant probability of a single data classified between two classes. Data preparation lies on specifying independent variables and dependent variables. The next stage is to split the dataset into two, data test and data train.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 2: Representation of Dataset.

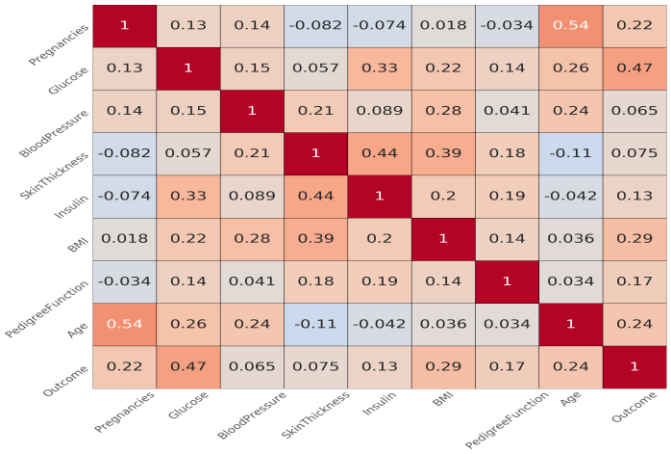

Figure 3: Pictorial Representation of Dataset.

Figure 2 and 3 represents the dataset with background variables as symptoms and the mean, standard deviation, minimum value and the maximum value range in the dataset is sorted based on the background variables for future process.

5. Library and Package Details

a) NumPy:

Numpy is a library of Python used to work with arrays. It also has functions for working in domain of linear fourier transform, algebra, and matrices. It is an open-source project and we can use it for free of cost.

In Python NumPy arrays are faster and more compact than lists. An array consumes lower amount of memory and is more convenient to use. NumPy array uses very less memory to store data and it allows a mechanism of specifying the data types.

b) Pandas:

Pandas is a powerful, fast, flexible and easy to use open-source data analysis and manipulation tool, and it is built on top of the Python programming language.

Pandas is flexible and powerful quantitative analysis tool. Pandas has been grown into one of the most popular Python libraries. It has a vast amount of active community and contributors.

c) Seaborn:

Seaborn is one of the Python data visualization libraries based on matplotlib. It provides a high-level interface for drawing informative and attractive statistical graphics.

Seaborn is built on top of matplotlib and it integrates closely with panda's data structures. Seaborn helps us explore and understand our data.

d) Matplotlib:

Matplotlib is a fantastic visualization library in Python for two-dimensional plots of arrays. Matplotlib is a multi-platform data visualization library in Python, built on NumPy arrays and programmed to work with the vast amount of SciPy stack. John Hunter invented it in the year 2002.

One of the greatest aspects of visualization is that it allows us visual access to large amounts of data in easily digestible visuals. Matplotlib consists of many plots like bar, line, histogram, scatter etc.

6. Training and Testing

When trained over original data, Logistic Regression proved to be the best among other classifiers in the term of recall value, with accuracy percentage of 78.01%.

After performing the Major Component Analysis, when data is trained our models, let found there is no improvement in KNN results; also, Random Forest too did not yield any better results. However, the Logistic Regression remains the better result with the accuracy level of 78.01%.

7. Result

In this model, the test and evaluations were performed using a variety of machine learning algorithms, including Support Vector Machine, Linear Regression, Logistic Regression. With the help of these algorithms the testing and training were performed and by using the accuracy level of those testing and training datasets the Logistic Regression has been applies to obtain the great accuracy level with the percentage of 78.01%. Figure 4 shows the pictorial representation of outcome flowchart.

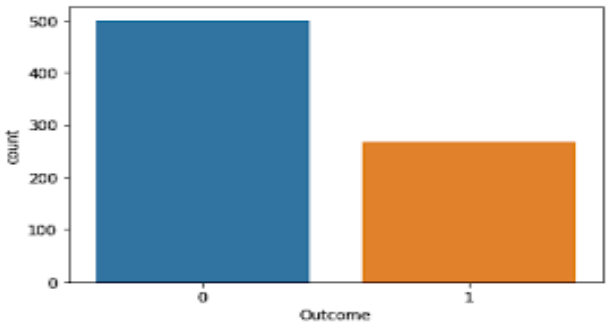

Figure 4: Pictorial Representation of Outcome Flowchart.

The above graph shows that the data is biased towards data points having outcome values as 0 where it means that diabetes was not present actually and the value 1 indicates the presence of diabetics. The result of non-diabetics is almost twice the number of diabetic patients.

8. Future Work

In future, this research work can be expanded with Unsupervised dataset. The missing values can be deleted or replaced with quality or valid data. The parameters needed to improve the accuracy can be chosen to maintain the train and test time. Instead of prediction from dataset and numerical features we can use deep neural network and image recognition. We can create a leverage of machine learning model by creating a licensed online portal for Diabetes prediction and can provide the software to hospitals.

References

1. Kandhasamy, J. P. & Balamurali, S., 2015. Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, Volume 47, p. 45–51.
2. Lekha, S. & Suchetha, M., 2018. Real-time non-invasive detection and classification of diabetes using modified convolution neural Network. IEEE Journal of Biomedical Health Information, Volume 22, p. 1630–1636.
3. G. A. Pethunachiyar, "Classification of diabetes patients using kernel-based support vector machines," in Proceeding of the 2020 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–4, IEEE, Coimbatore, India, January 2020.
4. T. Anand, R. Pal, and S. K. Dubey, "Cluster analysis for diabetic retinopathy prediction using data mining techniques," International Journal of Business Information Systems, vol. 31, no. 3, pp. 372–390, 2019.
5. N. Yuvaraj and K. R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster," Cluster Computing, vol. 22, no. 1, pp. 1–9, 2019.
6. Zou et al., 2018 Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang Predicting diabetes mellitus with machine learning techniques Frontiers in Genetics, 9 (2018), p. 515
7. Tigga and Garg, 2020 Prediction of type 2 diabetes using machine learning classification methods Procedia Computer Science, 167 (2020), pp. 706-716
8. Kopitar et al., 2020 L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, G. Stiglic Early detection of type 2 diabetes mellitus using machine learning-based prediction models
9. Scientific Reports, 10 (2020), pp. 1-12
10. J. Han, J. C. Rodriguez, and M. Behesti, "Discovering Decision Tree-Based Diabetes Prediction Model," in Proceedings of the International Conference on Advanced Software Engineering and its Applications, pp. 99–109, Springer, Jeju Island, Korea, December 2018.
11. Y. K. Qawqzeh, A. S. Bajahzar, M. Jemmali, M. M. Otoom, and A. Thaljaoui, "Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modelling", BioMed Research International, vol. 2020, Article ID 3764653, 2020.