PREDICTION OF DIABETES MELLITUS THROUGH MACHINE LEARNING

THOTA SITHA RAMANJANEYULU[1], JAI SAI VENKATA ROHIT GOWRA[2], RAKESH KOMMINENI[3], TETALLA BHAVANA REDDY[4]

[1][2][3][4]*Dept. of Computer Science and Engineering, Koneru Lashmaiah Educational Foundation, Guntur*

---***---

Abstract - *Diabetes Mellitus (or) Diabetes, it is a chronic disease and an increasing health problem. Diabetes is a disease that has the potential to cause a worldwide health crisis. The main cause of the type 2 diabetes is due to the absence of insulin. When pancreas is unable to produce insulin, the body develops type 2 diabetes. There are also other contributing environmental factors for the cause of diabetes such as inactive and overweight. Due to the rapid growth potentiality of the disease, it is our responsibility to decrease the range of it. In the current project, we propose to involve experts from various fields to generate data and perform interdisciplinary studies to obtain knowledge about Insulin Resistance. The aim of this project is to develop a model that performs the prediction of diabetes with better accuracy by combining the results of four supervised machine learning techniques including: Decision tree, Random forest, ANN, Logistic regression. Dataset we used in this project is "Diabetes" which is taken from Kaggle.*

***Key Words*: Diabetes Mellitus, Pancreas, Supervised, Machine Learning.**

1. INTRODUCTION

As said before, Diabetes is an infection that can possibly cause an overall wellbeing emergency. The development of this sickness is so high so that individuals are being influenced by it even without the information on having it. Anyway, how it i.e., Diabetes happens? Prior to getting into the explanation, we need to know a few terms, for example, Blood glucose, Insulin, and so on

The principle wellspring of the energy that comes from the food we eat is Blood glucose. A significant degree of Blood glucose additionally called glucose is the justification the introduction of Diabetes. The glucose which we get from the food ought to get into our cells for energy. It is finished by Pancreas. The pancreas delivers a chemical named Insulin which helps the glucose to enter our cells.

Insulin Opposition, which is an antecedent of Diabetes Mellitus results from an interaction of genetics and ecological elements, starting from before birth i.e., in the antenatal period, and proceeding later til' the very end. Ladies and their babies (kids) have been appeared to have insulin obstruction, which is an antecedent for the later advancement of type 2 diabetes mellitus and related complexities. An assortment of conventional strategies, in view of physical and substance tests, are being used for diagnosing diabetes. Nonetheless, early forecast of diabetes might be a bitch allenging task for clinical experts because of its reliance on different elements as diabetes can influence the heart, kidneys, eyes, and nerves, and so on Utilizing AI strategies one can examine gigantic datasets and can discover covered up data to create results fittingly.

In the current venture, we propose to involve experts from various fields to generated at and perform interdisciplinary studies to get information about Insulin Opposition, which is far reaching in the Indian Populace. The point of this undertaking is to foster a model that plays out the expectation of diabetes with better precision by consolidating the consequences of different managed AI strategies without leaving the essential reason. The AI calculations which are being utilized in this task were very recognizable tous.

2. LITERATURE REVIEW

The goal of this project is to develop a machine learning model which gets us the best results while predicting the diabetes. There is a lot of research work going on and there are so many models that were proposed earlier for the prediction of diabetes.

Here in a paper, "Diabetes Prediction Using Machine Learning Techniques", it is shown that the SVM is the best model to predict the diabetes. Whereas they compared the results in between the three Logistic regression, ANN and SVM. The advantage of this project is about the dataset that they worked on. The dataset which they took consists of huge amount of data. As all of us know that more amount of data helps us in building the model more accurately i.e., while training the model which gets us the best results. Same with the disadvantage, the dataset. As we said earlier, insulin resistance is the major cause of diabetes. There is no attribute related to insulin in the dataset which they took. So, we cannot rely on the model which they built in prediction of diabetes.

While there is another paper, "A Machine Learning Based Approach for the identification of Insulin Resistance with Non-Invasive Parameters using Homa-IR", it is shown that the SVM, Logistic regression is the best model for prediction. Whereas they compared the results in between several machine learning models. The advantage of this model is that they proposed a non-invasive technique to identify and monitor insulin resistance. While there is a con, since we all know that the predicted values are somewhat inaccurate when they compared with the actual, there might be a chance of getting some errors in outputs which results us to less accuracy. The values i.e., the insulin resistance values which they go tare the combinations of the eighteen parameters such as gender, waist size, height etc.

These are two different papers which worked under diabetic prediction. While there are so many other papers which are under prediction of diabetes and even the accuracy of methods which were proposed in those models are high in number, but the major disadvantage of those papers is because of the dataset which they relied on. As said earlier, Insulin Resistance is the reason for the patients who are being affected from diabetes. Resistance of insulin, which we get from insulin (released by pancreas) is the reason for the birth of diabetes. In most of the projects, it is missing in the dataset which they considered for the prediction. They went for the dataset which has regular factors such as Glucose, BMI, weight, etc.

In our project, the dataset which we are using consists of an attribute named Insulin resistance (IR) which plays a major role in the prediction of diabetes. And, from the resources, it is observed that the birth of diabetes in a body is taking place from the antenatal stage itself. Most of the woman and their infants have been shown to have insulin resistance. So, we minimized the dataset which is taken from Kaggle to a dataset which consists of only woman data.

3. THEORITICAL ANALYSIS

The title of the project itself says that prediction needs to be done using machine learning techniques. So, what is machine learning?

3.1 Machine Learning:

Name itself tells that the machine is trying to learn something. A machine cannot learn on its own. So, we help them in learning and asks them back the things which we want. On a professional note, Machine Learning is a part of Artificial Intelligence that makes the computer systems with the ability to learn and make predictions (on its own) on the data which we fed. While coming to the programming language, there is no single best language for machine learning. Everyone uses their own choice of language that suits best for their problem statement. But the topmost languages which are frequently being used are Python, R, Java and JavaScript, Julia, LISP. Every language has their own pros and cons. In our project we used Python.

3.2 Why Python?

One of the main reasons to use python as programming language for machine leaning is its extensive collection of libraries and packages. These libraries and packages setups a base-code for the classifications in machine learning and there is no need in developing the code from scratch for the programmers. This also helps us in reduction of development time. And the major reason for the use of python is its flexibility. It is easily understandable for any English-speaking person. Another reason for choosing python is because of its portable and extensible nature.

Machine learning has so many algorithms which are categorized as: Supervised learning, Unsupervised learning, Semi-supervised learning.

3.3 Supervised Learning/Predictive Models:

Supervised learning algorithms are used to develop predictive models. A predictive model predicts missing values using other values present in the dataset. Supervised learning algorithm has a set of input data and a set of output and buildsa model to make realistic predictions. These include Decision Tree, Bayesian Method, Artificial Neural Network, Instance based learning.

3.4 Unsupervised Learning / Descriptive Models:

Descriptive models are developed using unsupervised learning methods. In this model we have a known set of inputs, but output is unknown. Unsupervised learning is mostly used on transactional data. This method consists of clustering algorithms such as k-Means clustering and k-Medians clustering.

3.5 Semi-supervised Learning:

Semi Supervised learning method uses both data i.e., labeled, and unlabeled data on training dataset. These include Classification, Regression techniques Logistic Regression, Linear Regression.

The dataset which we took is a labeled dataset. So, we used supervised learning models. Those are ANN, Decision tree, Random forest, Logistic regression.

3.6 Algorithms:

3.6.1 Decision Tree:

A general, predictive modeling technique that has applications covering a variety of different areas is Decision Tree Analysis. Decision trees are usually built by an algorithmic approach that recognizes ways to divide a data set based on various conditions. It is one of the most used methods for supervised learning and is practical. A non-parametric supervised learning approach used for classification problems including regression is Decision Trees. In general, the rules for decisions are in the form of if-then-else sentences. The deeper the tree, the more complicated the laws are, and the model is more suitable.

3.6.2 Random Forest:

Random Forest is an algorithm for supervised machine learning. The forest that it constructs is an arrangement of decision trees that are typically educated in the bagging method. The fundamental principle of the bagging method is that a mixture of learning models increases the cumulative result. The random forest takes the prediction from each tree and it is based on the majority votes of predictions instead of depending on single decision tree, and it predicts the final model. The larger proportion of trees in the forest leads to higher precision and avoids the overfitting problem.

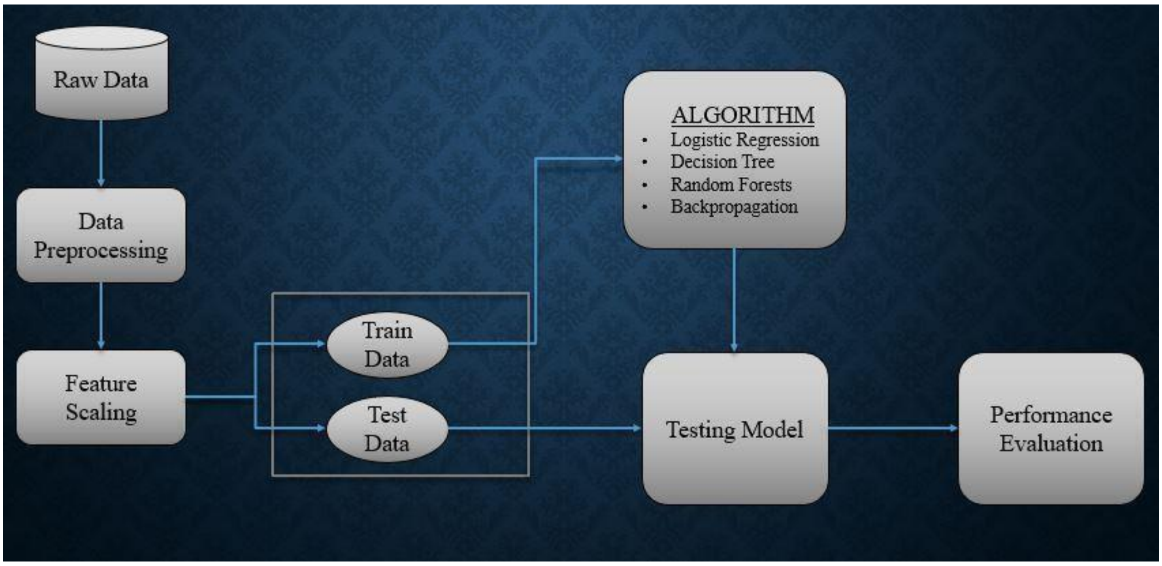
3.6.3 Logistic Regression:

A supervised learning classification algorithm used to predict the likelihood of a target variable is logistic regression. The existence of the target or dependent variable is dualistic, implying that only two possible forms a repossible. In plain English, with data coded as either 1 (stands for yes) or 0 (stands for no), the dependent variable is binary in nature. In general, Logistic regression means binary., it has two unordered target values for a dependent variable. It also can be divided into Multinomial Logistic Regression having three or more unordered target values for a dependent variable and Ordinal Logistic Regression having three or more ordered target values for a dependent variable.

3.6.4 Back propagation:

Before getting into Back propagation, we need to know about Artificial Neural Networks (ANN). A neural network is a group of connected I/O units where their computer programs have a weight associated with each connection. This allows you to create predictive models from large datasets. The human nervous system builds upon this model. It allows you to understand machine voice, pictures, human learning etc. Now, what is Back propagation? It is the technique of fine-tuning a neural net's weights based on the error rate obtained in the previous period (i.e., iteration). Back propagation is the core of Neural Net Training is back-propagation. By increasing its generalization, careful tuning of the weights helps you to reduce errorrates and make the model reliable.

3.7 Methodology:



As we can see the first step in the methodology is Data Preprocessing. It is nothing but making a raw data into a data which is suitable to a machine learning model. Feature scaling is one of the preprocessing techniques where we need to standardize the independent variables of the dataset in a specific range. And next step is splitting of the data set into training and testing. Train data is used to train the model by feeding the values of the data. After training the model it is ready for predicting the output. Then test data is to be used for predicting the outputs and comparingit with the actual values. This helps us in finding out the accuracy of the developed model.

The dataset which we took for prediction is labeled. So, we went with the supervised learning algorithms such as Logistic Regression, Decision Tree, Back Propagation, Random Forest.

4. EXPERIMENTAL INVESTIGATION

4.1 About Dataset:

Thedataset used in this project is “Diabetes” dataset whichis taken from Kaggle. The dataset consists of the following attributes.

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Age

- Outcome

As saidearlier, this dataset consists of only woman data since most of the woman and their infantshave been shown to have insulin resistance.

The datasetconsists of nearly 1000 rowsand 9 columns. We all know about the parameters that are there in the dataset.
Pregnancies, Glucose, Blood Pressure, Skin Thickness, BMI, Age are not new to us. Let us know about the remaining.

Insulin: It is a hormone thathelps the glucose enterour body cells, like limiting the level of glucoseentering our body, where that can be used for energy or can be stored for future use.

Diabetes Pedigree Function: DPF isa function that'scores probability or chance of diabetes to a person ba sed on the family history considering age.

Outcome: Outcome has the valueseither 1 or 0. 1 indicates that the person is Diabeticand 0 indicates that the personis Non- Diabetic

4.2 About Libraries:

As weare developing the machine learning model in python programming language because of a huge collection of libraries, we all need to know some of the libraries which we used in our project.

pandas: pandas stand for Python Data Analysis Library. It is used for data manipulation and data analysis.

NumPy: NumPy stand for Numerical Python. It isa library which canbe used while working witharrays. Itcanalso be used in domain of Fourier transform, linear algebra, and matrices.

sklearn: Scikit-learn, also known as sklearn is a machine learning library for python. It holds a lot of well-organized tools for machine learning and statistics. Statistical models including classification, clustering, and regression. It is necessary to build any type of machine learning model.

These are the required libraries or the libraries which we used during this project.

5. RESULTS

Accuracy for all the four models:

- Decision Tree - 76.380519
- Random Forest - 77.9220779
- Logistic Regression - 79. 2207792
- Back propagation - 75%

Prediction for a new entry [1, 85, 66, 29, 0, 26.6, 0.351, 31]:

- Decision Tree - 0
- Random Forest - 0
- Logistic Regression - 0
- Back propagation - 0

6. DISCUSSION OF RESULTS

Prediction of Diabetes Mellitus through MachineLearning performsa bettersolutionwith Logistic Regression. Basedon the results, by comparing the accuracy, we can see that the Logistic regression tops when compared to the other three models i.e., Random Forest, Back propagation, Decision tree. So, Logistic Regression may give us the better results in prediction.

7. CONCLUSION AND RECOMMENDATIONS

In this project, A systematic study was made in Prediction of Diabetes Mellitus through Machine Learning. This may be a better solution when compared with other papers since the dataset which we used consists of the major attribute insulin and that too the tested values. It becomes a robust and trustworthy solution for small scale as well as large scale of data. Since the model is trained with less data the accuracy was less and the prediction may vary with the actual. So, if we consider large data, then the model can be trained well, and the prediction will be much more accurate than now.

8. REFERENCES

[1]. Komi, Zhai. 2017. Application of Data Mining Methods in Diabetes Prediction

[2]. Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus, Omar Kassem Khalil
Aissa Boudjella, 2016 Sixth International Conference on Developments in eSystems Engineering.

[3]. Alan Siper, Roger Farley and Craig Lombardo, “Machine Learning and Data Mining Methods in Diabetes Research”, Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 6th, 2005.

[4]. Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1 (2016): 727-730.

[5]. Berry, Michael, and Gordon Linoff. Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., 1997

[6]. Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[7]. Giri, Donna, et al. "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICAand discrete wavelet transform." Knowledge-Based Systems 37 (2013):274-282.