Designing a model to detect diabetes using machine learning: A Survey

[1]Ms. Komal. Patil, [2]Dr. S. D. Sawarkar, [3]Mrs. Swati Narwane
[1]Research Assistant, [2]Principal, [3]Assistant Professor
Datta Meghe College Of Engineering

Abstract - Many of the interesting and important applications of machine learning are seen in a medical organization. The notion of machine learning has swiftly become very appealing to healthcare industries. The predictions and analysis made by the research community for medical dataset support the people by taking proper care and precautions by preventing diseases. Through a set of medical datasets, different methods are used extensively in developing the decision support systems for disease prediction. This paper explains various aspects of machine learning, the types of algorithm which can help in decision making and prediction. We also discuss various applications of machine learning in the field of medicine focusing on the prediction of diabetes through machine learning. Diabetes is one of the most increasing diseases in the world and it requires continuous monitoring. To check this we explore various machine learning algorithm which will help in early prediction of this disease.

Keywords - Diabetes, health care, decision tree, machine learning, applications, classification, approach, algorithm.

I. INTRODUCTION

Multiple opportunities for healthcare are created because machine learning models have potential for advanced predictive analytics. There are already existing models in machine learning which can predict the chronic illness like heart disorder, infections and intestinal diseases. There are also few upcoming models of machines learning to predict non-communicable diseases, which is adding more and more benefit to the field of healthcare. Researchers are working on machine learning models that will offer very early prediction of specific disease in a patient which will produce effective methods for the prevention of the diseases. This will also reduce the hospitalization of patients. This transformation will be very much beneficial to the healthcare organisations. [1]

The most explored area is the healthcare system which uses modern computing techniques is in healthcare research. As mentioned above the researchers in the related fields are already working with the healthcare organisation to come up with more technology ready systems. Diabetes is a disease which reduces the body's capability to produce insulin. In other words the body can not retaliate to the hormone insulin production. This results in anomalous metabolism of carbohydrates and increased blood glucose levels. Early detection of diabetes becomes very important because of the reason mentioned above. Many people in the world are getting affected by diabetes and this number is increasing day by day. This disease can damage many vital organs hence the early detection will help the medical organisation in treatment of it. As the number of diabetic patients is more there is a excessive important medical information which has to be maintained. With the support of increasing technology the researchers have to build a structure that store, maintain and examine these diabetic information and further see feasible dangers. [4]

The blood glucose levels become too high in the body when there is diabetes. Glucose is created in the body after eating food. The hormone insulin produced in the body helps balance the glucose levels and regulate the blood sugar levels, deficiency of insulin causes Diabetes. Type 1 diabetes is a scenario where the body does not produce insulin at all to balance the sugar levels in blood. Type 2 is a diabetes type where the body produces insulin but does not utilize this hormone completely to balance the blood sugar levels. The Type 2 diabetes is most common one. There is something called as prediabetes, this is a situation where the person can have high glucose level but not that high that he/she can be said to have diabetes. But the people who have prediabetes are prone to get type 2 diabetes. This disease can cause serious damage to many vital organs in the body like kidneys, heart, nerves and eyes. If a woman gets this disease during pregnancy then it is known as gestational diabetes. By managing our weight, meal plan and exercise we can control diabetes. One should always keep a check on its blood sugar levels.

II. Types of Approaches in machine learning

To make predictions or decisions, machine learning creates a mathematical model using a sample data which can also be called as training data. These algorithms do not explicitly perform a task or programmed to make predictions. Machine learning provides systems the capability to learn and improve from experience automatically without any explicit program execution because it is an application of artificial intelligence. Machine learning pays attention on the development of the computer programs which accesses data and uses it for learning.

2.1. Supervised learning algorithms and Semi-supervised algorithms

Supervised learning algorithms uses a mathematical model of a set of data which have both the inputs and the desired outputs. This set of data is called as the training data, which have of examples. These examples usually have one or more

inputs and the desired output, this is also mentioned as a supervisory signal. In **semi-supervised learning algorithms**, the desired output is missed by few of the training examples In the mathematical model, we represent the training example by an array and in matrix form is the training data. To predict the output related to new inputs supervised algorithms pursue a function through iterative optimisation. [20]
To flawlessly know the output for inputs which are not included in training data, an optimal function which allows the algorithm is used. This improves correctness of the output for a particular tasks. This algorithms involve regression and classification. These classifications are used only when we want the output to restricted value whereas the regression algorithm is used for numerical output which is in the range.[10]

2.2. Unsupervised learning algorithms
Unsupervised learning algorithms uses a dataset which has only input values, these inputs are in form of structure for example it could be in a cluster of data points or a group of data.The unsupervised algorithm recognizes similarities in the data and behave based on the existence and non-existence of such simlarities in the dataset.
The cluster is set of observations of the dataset, these cluster are analysed in such a manner that the similar observations which are pre-defined data or criteria of a data form one cluster and remaining data form a different cluster. These are then evaluated and metric to find the outputs[15]
2.3. Reinforcement Learning is a type of **Machine Learning**, and thereby also a branch of **Artificial Intelligence**. It allows **machines** and software agents to automatically determine the ideal behaviour within a specific context, in order to maximize its performance.[20]

III. Classification Algorithms
Classification is the procedure followed to predict the class of the given dataset. These classes are defined or also known as targets/ labels or categories. Classification is a type of predictive modelling which maps the function from input to desired output. Let us take an example of spam detection in email to understand further the concept of classification. In spam detection in email the service providers can be pointed as a classification problem. This will be a binary type of classification as only two classes namely, as spam and not spam are defined. A classifier deploys some data to learn how given input variables relate to the class. In this example, spam and non-spam emails have to be given as the training data. When the classifier is trained correctly, it can be used to detect a spam email. In classification the targets are also given along with the input data, hence this comes under supervised learning category.[19]

3.1. Classification algorithms
Depending on the application and nature of the dataset used we can use any classification algorithms mentioned below. As there are different applications, we can not differentiate which of the algorithms are superior or not. Each of classifiers have its own way of working and classification. Let us discuss each of them in details.[5]

3.1.1. Naive Bayes Classifier
This classifier can also be called as Generative Learning Model. The classification here is based on Baye's Theorem, it assumes independent predictors. In simple words this classifier will assume that the existence of specific feature in a class is not related to the existence of any other feature. If there is dependency among the features of each other or on the presence of other features, all of these will be considered as an independent contribution to the probability of the output. This classification algorithm is very much useful to large datasets and is very easy to use. [14]

3.1.2. Logistic Regression
Logic regression is used for Predictive Learning Model. To determine output in this classifier, we use a statistical method to analyse the dataset. These data set can have one or more than one independent values. The output is calculated with a data in which there could be two outputs. The aim of this classification algorithm is to find the relationship between the dichotomous category and predictor variables.[6][14]
3.1.3. Decision Trees
This classification algorithm builds the regression models. These models are builded in form of structure which is similar to tree - a tree like structure is created by this classifier. It keeps on dividing the data set into subsets and smaller subsets which develops an associated tree, incrementally. The decision tree is finally created which has decision nodes and leaf nodes. In this tree the leaf node will have details about the classification or the decision taken for classification whereas the decision will have branches. The highest decision node which will be at the top of the tree will correspond to the root node. This will be the best predictor. [3][14]

3.1.4. Random Forest
This classification algorithm are similar to ensemble learning method of classification. The regression and other tasks, work by building a group of decision trees at training data level and during the output of the class, which could be the mode of classification or prediction regression for individual trees. This classifier accuracy for decision trees practice of overfitting the training data set.[8][14]

3.1.5. Neural Network:
As the name suggests this classifier has units known as neurons, which are arranged in layers that convert the input vector to relevant output. Each single neuron takes an input, this is most often a non-linear input, this is given to a function which is

them passed to next layer to get the output. The input given to the first layer will act as an output for the next layer and so on, thus this classification algorithm follows a feed-forward method. But in this method there is no feedback to the previous layer, so weighting are also given to the signals passing through the neurons and the layers, these signal then are turned into a training phase this eventually then become a network to handle any particular problem.[2][14]

3.1.6. Nearest Neighbor

As the name suggests the nearest neighbour algorithm is based on the nearest neighbour and this classification algorithm is supervised. It is also called as k-nearest neighbour classification algorithm. A cluster of labeled points are used to understand how the other points should be labelled. For labelling a new point it checks the already labelled points which could be closest to the point to be labelled, i.e closest to the neighbour. In this way depending on the votes of the neighbour the new point is labelled the same label which most of neighbours have. In in algorithm 'k' is the number of neighbours which are checked.[5][14]

3.1.7 Support vector machine (SVM)

This is also one of the classification algorithm which is supervised and is easy to use. It can used for both classification and regression applications, but it is more famous to be used in classification applications. In this algorithm each point which is a data item is plotted in a dimensional space, this space is also known as n dimensional plane, where the 'n' represents the number of features of the data. The classification is done based on the differentiation in the classes, these classes are data set points present in different planes.

3.1.8. XGBoost

Recently, the researches have come across an algorithm "XGBoost" and its usage is very useful for machine learning classification. It is very much fast and its performance is better as it is an execution of a boosted decision tree. This classification model is used to improve the performance of the model and also to improve the speed. [21]

IV. Machine learning in medical diagnosis of diabetes

The doctors uses the methods like pattern recognition in an image, segmentation filtering, and data processing with the help of machine learning classification algorithms mentioned in the above section to analyse the disease or predict the accuracy of the existence of the disease. By defining the region of interest one can analyse the data and come up with predictions. Many diseases like cancer diabetes are predicted by used machine learning technology. We shall broadly talk about diabetes. For example, for detection of breast cancer a X-ray of a female breast is examined to detect this disease. A computer aided diagnosis by the doctors is used to screen the mammography. This methodology is used for the early prediction/ detection of the cancer in females. These system classify the tumor as a benign or malignant. The researches are still working on this topic. [6][17]

Diabetes is a disease which occurs due to ability of a human body to produce insulin which increase the glucose levels in the body. The variation in glucose levels is cause of diabetes. Insulin balances the blood glucose level in the body, deficiency of which cause diabetes. For the prediction of diabetes machine learning is used, these have many steps like image pre-processing/data preprocessing followed by a feature extraction and then classification. We can use any of the mentioned machine learning classifiers to predict this disease. In the above section we have learning about many classification algorithms, we can either use any one of these to predict the disease or we can explore the techniques to use the hybrid methodology to improve the accuracy over using a single one. Currently, the researches have used the a single classification algorithm and have come up to accuracy of 70 to 80% for detection of the diabetes disease. [7][9]

We have already learned about all the machine learning classification algorithms and approaches used to predict the disease. After doing this survey we would be proposing to use more than one classification algorithm along with any of the learning approaches which will improve the prediction accuracy of the disease by more than 80%.

It is good to use the combination of more than 2 classifiers to get the desired accuracy. We shall be using Decision tree along with other classifiers, we shall design a model to evaluate the training data We shall evaluate each of the classifier and either use XGBoost along with Decision tree/ RF/ SVM / Naive Bayes or we can use Decision Tree / RF along with the Naive Bayes.by using the combination mention in this section we shall improve the accuracy by more than 80%. [3][7]

V. Conclusion

At an early period of diagnosing the diabetic disease, machine learning techniques can help the physicians to diagnose and cure diabetic diseases.We shall conclude that the increase in classification accuracy helps to improvise the machine learning models and yields better results. The performance analysis is analyzed in terms of accuracy rate among all the classification methods such as decision tree, logistic regression, k-nearest neighbors, naive Bayes, and SVM. After learning about the approaches and the machine learning classification, we shall explore the use of using the different classification algorithms in machine learning. We have also see than the accuracy of the current system is less than 80% hence we after the survey we recommend to use combination of either XGBoost along with Decision tree/ RF/ SVM / Naive Bayes or we can use Decision Tree / RF along with the Naive Bayes.by using the combination mention in this section we shall improve the accuracy by more than 80%. In future we can explore a combination of other classifiers and then evaluate the results.

VI. References

[1] [Harleen Kaur, Vinita Kumari], Predictive modelling and analytics for diabetes using a machine learning approach, 2018

[2] [Jake A. Carter, Christina S. Long, Beth P. Smith, Thomas L. Smith, George L. Donati], Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes, 2018
[3] [Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras,Ioannis Vlahavas, Ioanna Chouvarda], Machine Learning and Data Mining Methods in Diabetes Research, 2017
[4] [S M Hasan Mahmud, Md Altab Hossin, Md. Razu Ahmed, Sheak Rashed Haider Noori, Md Nazirul Islam Sarkar], Machine Learning Based Unified Framework for Diabetes Prediction, 2018.
[5] [Ratna Patil, Sharavari Tamane], A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes, 2018
[6] [Arianna Dagliati, Simone Marini, Lucia Sacchi, Giulia Cogni, Marsida Teliti,Valentina Tibollo, Pasquale De Cata, Luca Chiovato, and Riccardo Bellazzi], Machine Learning Methods to Predict Diabetes Complications, 2017
[7] [Rabindra Kumar Barik, R. Priyadarshini, Harishchandra Dubey, Vinay Kumar and S. Yadav], Leveraging Machine Learning in Mist Computing Telemonitoring System for Diabetes Prediction, 2017
[8] [Ambika Choudhury and Deepak Gupta], A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques, 2019
[9] [Piyush Samant , Ravinder Agarwal], Machine learning techniques for medical diagnosis of diabetes using iris images, 2017
[10] [Irene Dankwa-Mullan,Marc Rivo,Marisol Sepulveda, Yoonyoung Park, Jane Snowdon, and Kyu Rhee], Transforming Diabetes Care Through Artificial Intelligence: The Future Is Here, 2018
[11] [Rabindra Kumar Barik, R. Priyadarshini, Harishchandra Dubey ,Vinay Kumar and S.Yadav], Leveraging Machine Learning in Mist Computing Telemonitoring System for Diabetes Prediction, 2017
[12] [Andrew L. Beam, Isaac S. Kohane], Big Data and Machine Learning in Health Care, 2017
[13] [Ratna Patil, Sharavari Tamane], A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes, 2018
[14] https://www.greycampus.com/opencampus/machine-learning/different-types-of-classification
[15] https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4
[16] https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623
[17] https://en.wikipedia.org/wiki/Machinelearning
[18] https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/
[19] https://www.sciencedirect.com/science/article/pii/S2405959518304624
[20] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232260/
[21] https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/