Prediction of Diabetes Diseases by Building a Machine Learning Model

Hnin Ei Ei Cho[1]
Myanmar Institute of Information Technology, Mandalay, Myanmar
hnin_ei_ei_cho@miit.edu.mm

Nan Yu Hlaing[2]
Myanmar Institute of Information Technology, Mandalay, Myanmar
nan_yu_hlaing@miit.edu.mm

Abstract

Today, the processing strength and storage ability has increased dramatically and huge amounts of data are widely available. There is an urgent need for transforming data into useful information and knowledge. With the rapid development of machine learning, many aspects of medical health has applied a variety of machine learning approach. Given a number of elements all with certain features, we want to build a machine-learning model to identify people affected by type 2 diabetes. This research work focuses on pregnant women suffering from diabetes. The aim of this paper is to compare the six different learning algorithms in order to predict diabetes in patients more accurately using the Pima-Indians-Diabetes Dataset obtained from UCI Machine Learning repository site, the Centre for Machine Learning and Intelligent Systems at the University of California, Irvine.

Keywords: Machine Learning, Support Vector Machine, Artificial Neural Network, Decision Tree, Logistic Regression, Naïve Bayes

1. Introduction

Diabetes is one of the deadliest and chronic diseases in the world; this disease inflicted 246 million people. Diabetes is a kind of illness. It affects the ability of the body in producing the hormone insulin. It in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. Diabetes can divide into type 1 diabetes (T1D) and type 2 diabetes (T2D). T1D patients are commonly less than 30 years old and T2D occurs most in middle-aged and elderly people. In Diabetes, a person generally suffers from high blood sugar. Intensify thirst, Intensify hunger and Frequent urination are some of the symptoms caused due to high blood sugar. Many complications occur if diabetes remains untreated.

Healthcare data is complex and high in dimensionality and contains irrelevant information; therefore, the prediction accuracy is low. Classification problems are prediction of class labels where number of classes is fixed and pre-defined. There is nothing like a particular classification method is accurate to classify the data in all situations. The accuracy of classification method is depends on the data we want to classify. In health related studies, the main research is going on the risk prediction. Predicting the risk in study of any disease is the main goal of health related studies. Recently, health related studies researchers are using machine-learning techniques to predict diabetes. By doing this the researches want to show the conditions those causes that particular disease. For this research the classification methods are very helpful and giving the best results. In this study, we used K Nearest Neighbor, Artificial Neural Network, Support Vector Machine, Decision Tree, Logistic Regression and Naïve Bayes to predict the diabetes.

2. Classification

Classification is very helpful method in predicting the risk in diabetes disease prediction.

Classification categorizes the items into target classes. Aim is to make predicting the target class accurately from the data [1]. Classifier can learn from the examples. Modern classification learning techniques gives more intelligent prediction results [2, 3].Classification techniques used in this research work described as below.

2.1. K-Nearest Neighbor (KNN)

M. Cover and P. E. Hart proposed K-nearest neighbor algorithm. KNN is a non-parametric lazy learning algorithm. It means that it does not make any assumptions on the underlying data distribution. In this algorithm, we know the type of class and the object's group to which it belong to is unknown. It considers more than one nearest neighbor to identify the class to which the data point it belongs to [4]. We can assign the training points with some weights based on their distance from data points. To improve memory limitations, we can use the NN training set to structure using different techniques. To defeat the memory limitations data set size is trim down.

2.2. Artificial Neural Network (ANN)

Artificial Neural networks are those systems modeled based on the human brain working. As the human brain consists of synapses that interconnect with millions of neurons, a neural network is a set of connected input/output units in which each connection has a weight associated with it. Multi-Layer Perceptron (MLP) network models are the popular network architectures. It is used in most of the research applications in medicine, engineering, mathematical modeling, etc.

In MLP, we pass the weighted sum of the inputs and bias term to activation level through a transfer function to produce the output, and arrange the units in a layered feed-forward topology called Feed Forward Neural Network (FFNN) [5]. It is highly fault tolerant and suitable for all kinds of real-world problems. The challenge of training neural networks involves carefully selecting the learning rate.

2.3. Support Vector Machine (SVM)

Support vector machine is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. These models closely relate to neural networks. They use a sigmoid kernel function, which is equivalent to a two-layer, perceptron neural network. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data [6]. It is a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high.

2.4. Naïve Bayes Classifier (NB)

Naive Bayes is a classification technique with a notion, which defines all features, are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature [7]. Since NB based on conditional probability, it is a powerful algorithm employed for classification purpose. It works well for the data with unbalancing problems and missing values.

2.5. Decision Tree (DT)

We can use the Decision Tree for decision analysis. The main objective of using Decision Tree is the prediction of target class using decision rule taken from prior data [8]. In Decision Trees, where target values can take continuous values are known as the regression trees. Considering the tree, we represented the input values as a path from the root to the leaves, where each leaf represents the target variable [9].

2.6. Logistic Regression (LR)

Logistic regression is a generalized form of linear regression. It is a linear model for classification not for regression. When predicting the binary or multi-class dependent variables, we can use logistic regression primarily. As the

response variable is discrete, linear regression cannot modeled directly. While logistic regression is a very powerful modeling tool, it assumes that the response variable is linear with respect to the predictor variables. A lack of explanation about what has learned can be a problem.

Table 1. Summary of classification learning models

Name	Advantages	Disadvantages
KNN	1. Easy to implement multiclass problem. 2. Training is very fast and robust to noisy training data. 3. Effective if the training data is large.	1. Need to determine value of parameter K. 2. Computation cost is quite high. 3. Being a supervised learning lazy Algorithm i.e., runs slowly. 4. It is sensitive to the local structure of the data.
ANN	1. Have the ability to work with inadequate knowledge and fault tolerance. 2. Have numerical strength that can perform more than one job at the same time.	1. The realization of the equipment is dependent. 2. Difficult to know how many neurons and layers we need to process and duration is also unknown. 3. When producing a solution, it does not give a clue as to why and how.
SVM	1. It scales well to high dimensional data. 2. The risk of overfitting is less. 3. It works well with unstructured and semi structured data.	1. Choosing a good kernel function is not easy. 2. Long training time for large datasets. 3. It is hard to visualize their impact.
NB	1. Need less training time. 2. Very simple, easy to implement and fast. 3. Not sensitive to irrelevant features.	1. Chance the loss of accuracy. 2. Algorithms cannot modify dependencies. 3. Makes a very strong assumption on the shape of data distribution.
DT	1. It produces the accurate result. 2. It takes the less memory, less model build time and short searching time. 3. Support multi-output tasks.	1. Relatively expensive as complexity and time taken is more. 2. Inadequate for predicting continuous values. 3. Computation is slower and can encounter overfitting.
LR	1. Works with almost any kind of dataset. 2. Gives good information about the features. 3. Very efficient to train.	1. It is often inappropriate used to model non-linear relationship. 2. It is limited to predicting numeric output. 3. Can only predict a categorical outcome.

3. Diabetes Prediction

Proposed research work introduces a framework to develop a machine-learning model based on data mining classification techniques. To solve the problem we will have to analyze the data, do any required transformation and normalization, apply a machine-learning

algorithm, train a model, check the performance of the trained model and iterate with other algorithms until we find the most performant for our type of dataset.

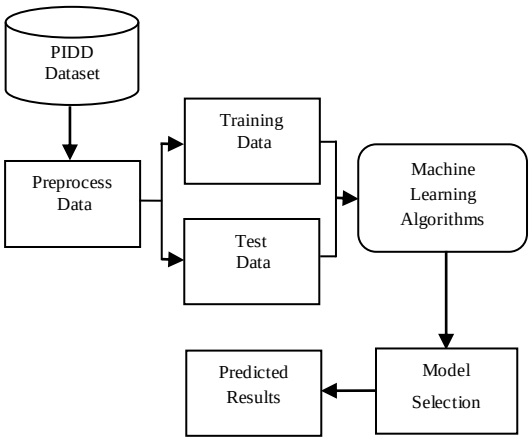

Figure 1. Diabetes prediction model

3.1 Dataset Description

The Pima are a group of Native Americans living in Arizona. A genetic predisposition allowed this group to survive normally to a diet poor of carbohydrates for years. In the recent years, because of a sudden shift from traditional agricultural crops to processed foods, together with a decline in physical activity, made them develop the highest prevalence of type 2 diabetes and for this reason they have been subject of many studies.

We performed computer simulation on a Pima Indians Diabetes dataset available UCI Machine Learning Repository. The features describe different factor for diabetes reoccurrence. The main aim of this study is the prediction of the patient affected by diabetes using the python language by using the medical database PIDD. The Pima is one of the most studied populations for diabetic analysis around the world [10].

Table-2 shows a brief description of the dataset. PIDD (Pima Indian Diabetes Dataset) contains of 768 instances of female patients. The dataset also consists of discrete-valued eight attributes. The last column of the dataset indicates if the person has been diagnosed with diabetes (1) or not (0). We describe attributes descriptions in Table 3. We are available the original dataset at UCI Machine Learning Repository and can download from [11].

Table 2. Dataset description

Database	No. of Attributes	No. of Instances
PIDD	8	768

Table 3. Attribute description

Attribute	Abbreviation
1. Number of times pregnant	pr
2. Plasma glucose concentration	pl
3. Diastolic blood pressure (mm Hg)	pr
4. Skin fold thickness (mm)	sk
5. 2-Hour serum insulin (mu U/ml)	in
6. BMI (weight/(height)^2)	ma
7. Diabetes pedigree function	pe
8. Age in years	ag
Class ’0’ or ’1’	cl

3.2 Data Preprocessing

We apply the data cleaning techniques first. Identify the missing values and replaced by the group median. Further, apply the min-max scaling technique to have the features value range between zero and one.

3.2.1 Data Cleaning

Some instances have missing data for some of the features. Machine learning algorithms cannot work very well with missing data. To find a solution to "clean" the data, the easiest option is to eliminate all those records, but in this way, we would eliminate many important data. Another option is to calculate the median value for a specific column and substitute that value everywhere in the same column that have missing data.

3.2.2 Data Transformation

In the dataset, the Age ranges from 20 to 80 years old, while the number of times a patient has been pregnant ranges from 0 (zero) to 17.

Most of the machine learning algorithms do not work very well if the features have a different set of values. The solution is to apply the feature scaling technique. Feature Selection Technique (FST) eliminates the less important features and reduces the time complexity of the machine learning technique.

The type of scaling depends on the data fed to which model, so there is no universally best approach. In this paper, min-max normalization techniques is used. Min-max normalization preserves the relationships among the original data values. It always boosts the classification accuracy and minimizes the computational cost.

3.3 Splitting the Dataset

Splitting the dataset is a very important step for supervised machine learning models. We split the dataset into two parts: training and testing dataset. In this paper, we use the K-fold cross-validation method. It partitions the original data set into equal-sized sub-segments. The number of segments depends upon the value of *k* taken; in our case, we have taken *k* to be 10. We use the first part to train the model ignoring the column with the pre assigned label. Then we use the trained model to make predictions on new data, which is the test dataset, not part of the training set, and compare the predicted value with the preassigned label.

3.4. Comparison of different Algorithms

We compare the accuracy (ACC) of multiple algorithms with the same dataset and pick the one with the best score.

$$ACC = \frac{TN+TP}{TN+TP+FP+FN} \quad (1)$$

Where true positive (TP) denotes the number of identified positive samples in the positive set. True negative (TN) represents the number of classification negative samples in the negative set. False positive (FN) is the number of identified positive samples in the negative set. False negative (FN) means the number of identified negative samples in the positive set. The accuracy is as the ratio of the number of samples correctly classified by the classifier to the total number of samples.

No.	Learning_Model	Accuracy
1	K Nearest Neighbor	72.96
2	Artificial Neural Network	76.37
3	Support Vector Machine	74.91
4	Naive Bayes	74.10
5	Decision Tree	70.02
6	Logistic Regression	75.40

Figure 2. Performance comparison on the basic of accuracy

Using the python programming language, we present the resulted output of the learning algorithms that predict the diabetes disease in Figure 2.

Table 4. Performance comparison on the basic of classified instances

Total Instances	Learning Model	Correct predict	Incorrect predict
768	KNN	548	220
	ANN	587	181
	SVM	575	193
	NB	569	199
	DT	550	218
	LR	579	189

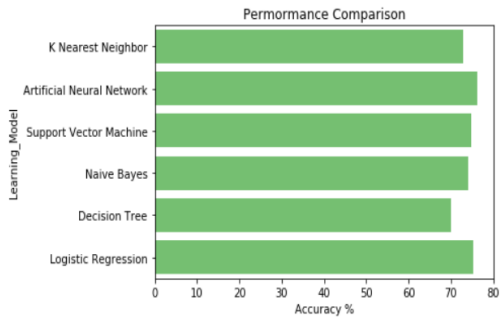

Figure 3. Performance evaluation of different learning algorithms

In Table 4, we described the comparisons of performance on classified instances. It illustrates how many instances they correctly classified according to the results of Figure2. We plot the chart of the performances of all classifiers based on accuracy measures in Figure-3.

From the above results, we observed that the Artificial Neural Network is performing better amongst all the other algorithms. Therefore, the ANN machine learning classifier can predict the chances of diabetes with more accuracy as compared to other classifiers.

4. Conclusions

This paper was using the dataset from the UCI repository. First, replace the group median values in all missing values. Further, apply the data transformation technique with a proper feature scaling method. We use six learning algorithms along with *k*-fold cross-validation. This enabled to perform data analysis to obtain the optimal result. We found by the result with the highest accuracy of 76.37% was achieved by using the Artificial Neural Network with *10*-fold cross-validation. From the results, we observed that the Neural Network was performing better compared to all the other classification algorithms. Future work will include trying a study with different data transformations or trying algorithms that we have not tested yet for further analysis of the dataset.

References

[1] S.Archana, Dr. K.Elangovan, "Survey of classifications techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014

[2] Sisodia,D., Singh,L., Sisodia,S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS2012), December28-30, 2012, Springer. pp.1027–1038.

[3] B. Kotsiantis · I. D. Zaharakis · P. E. Pintelas, "Machine learning: a review of classification and combining techniques", Springer Science10 November 2007

[4] Nitin Bhatia, Vandana," Survey of Nearest Neighbor Techniques" (IJCSIS) Vol. 8, No. 2, 2010, ISSN 1947-5500.

[5] Padmavati J., "A Comparative study on Breast Cancer Prediction Using RBF and MLP," International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011.

[6] XindongWu , Vipin Kumar , J. Ross Quinlan , Joydeep Ghosh , Qiang Yang , Hiroshi Motoda , Geoffrey J. McLachlan , Angus Ng , Bing Liu , Philip S. Yu , Zhi-Hua Zhou , Michael Steinbach , David J. Hand and Dan Steinberg, Top 10 algorithms in data mining, Knowl Inf Syst (2008) 14:1–37, Springer-Verlag London Limited 2007

[7] Raj Kumar, Dr. Rajesh Verma," Classification Algorithms for Data Mining P: A Survey" IJIET Vol. 1 Issue August 2012, ISSN: 2319 – 1058.

[8] Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14. doi: 10.5121 / ijdkp . 2015.5101, arXiv:1502.03774.

[9] R. Ahuja, V. Vivek, M. Chandna, S. Virmani and A. Banga, "Comparative Study of Various Machine Learning Algorithms for Prediction of Insomnia", *Advanced Classification Techniques for Healthcare Analysis*, ed. Chinmay Chakraborty, 234-257 (2019), accessed July 01, 2019. DOI:10.4018/978-1-5225-7796-6.ch011

[10] W.C. Knowler, D.J. Pettitt, M.F. Saad and P.H. Bennett, "Diabetes mellitus in the Pima Indians: incidence, risk factors, and pathogenesis", Diabetes/metabolism Reviews 6, no. 1, 1-27, 1990.

[11] UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/machine-learningdatabases/ statlog/german/.