International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Evaluating the Performance of Machine Learning Algorithms for Diagnosing Diabetes in Individuals

Idemudia Christian Uwa[1], Nehikhare Efehi[2]

[1]School of computer and communication technology, Lanzhou University of Technology, Lanzhou 730050, China

[2]School of life sciences, Lanzhou University of Technology, Lanzhou 730050, China

Abstract: ***Application of machine learning algorithms for the diagnosis of diabetes have become a trending research area, as effort to improve current techniques and methods used byhealth care institutions to determine the occurrence of diabetes in individuals is now given more attention than before. This study attempts to evaluate the performance of five (5) machine learning models on diabetic dataset using Python to predict the incidence of diabetes. Pima Indian diabetes dataset from UCI machine learning repository was used for the study. To ensure quality evaluation of the algorithms, a second dataset provided by Dr. John Schorling of the department of Medicine, University of Virginia was used for double evaluation. Result shows that Naive Bayes algorithm performs better when used for the prediction of diabetes.***

Keywords: Data Mining; Diabetes; Feature Selection; Naive Bayesian classifier; Machine Learning

1. Introduction

Technological improvements in the field of science and health care has given rise to the application of computer aided systems and applications in handling medical and health issues. Medical practitioners which include doctors, laboratory specialists, nurses are faced with hundreds of situations where they have to make a decision about the health condition of their patients based on the patient's symptoms and signs.

With the increasing need for efficient health care and the need for timely decisions, it is obvious that the traditional method of sieving knowledge from records can no longer be sustained as this result to delay and errors in medical decisions. The amount of time required for laboratory diagnostic results to be available must be optimized for better health care and for timely decision making. Data mining as a branch of computer science has evolved to assist medical personnel in performing their functions more effectively. With the availability of large amount of patient information in health organizations, decision making as regards patient's condition can be more optimized and made faster through data mining knowledge discovery techniques. Many computational tools and algorithms have been recently developed to increase the experiences and the abilities of physicians for taking decisions about different diseases [1].Diabetes mellitus is a chronic disease caused by inherited and/or acquired deficiency in production of insulin by the pancreas, or by the ineffectiveness of the insulin produced. Such a deficiency results in increased concentrations of glucose in the blood, which in turn damage many of the body systems, in particular the blood vessels and nerves [2]. According to the WHO media center, "The number of people with diabetes has risen from 108million in 1980 to 422 million in 2014. Also, in 2012 about 2.2 million deaths were attributable to high blood glucose while an estimated 1.6million deaths were reported to be directly caused by diabetes in 2015. with another 2.2 million deaths attributable to high blood glucose in 2012. Diabetes is deadly if there is no prompt diagnosis and offers a patient less lifeline of active living as it usually result to health complications such as heart attack, stroke, blindness, blindness, kidney failure, heart attack, stroke and lower limb amputation [2]. Although early diagnosis has been identified as a major step in the fight against the dangers of diabetes, medical personnel are often in situations where analysis of a patient test result may pose a challenge. Thus, the availability of an enhanced data mining algorithm for early diagnosis will definitely help to advert the dangers of late diagnosis and improve the management of diabetic cases.

In this study, an evaluation of the performance of machine learning classifiers in predicting diabetes disease in individuals is analyzed using experimental and statistical procedures. Classification is a form of data analysis that extracts models describing important data classes. Such models called classifiers predict categorical (discrete, unordered) class labels [3]. Classification splitsa dataset into mutually exclusive groups called a class based on suitable attributes [4]. Some of the numerous applications of classification include fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis. Machine learning models can show the result of a patients test with a pre-test probability (of the population), to predict or determine the chance of finding a particular disease. Therefore, the aim of this study is to evaluate the performance of machine learning classifiers in predicting and diagnosing diabetes using historical patient data. The completed study will provide a clear understanding of the data mining process for medical diagnosis, and also a confidence level on the application of machine learning models on diabetic patient data to determine their status.

2. Literature Survey

2.1 Data Mining and Disease Diagnosis (Prediction)

With the high level of success recorded in the application of data mining techniques in areas such as business, retail sales, banking and marketing, experts in both the medical and information technology sector have sort for the application of data mining in the health sector to help provide for better health care services. Experts believe that the health care environment is still 'information rich' but 'knowledge poor' [6]. Health care organizations have hundreds of thousands of data within its various data repositories (which include both

the electronic and traditional methods of storing patient records) but there is still less analysis of such data to uncover hidden and interesting patterns in the data. This fate is most common in the third world (developing) nations. With the incidence of sudden deaths increasing daily, preventing sudden death and unnecessary loss of life begins with accurately identifying risk factors and early diagnosis which will aid medical officials to provide efficient prevention and management methods and techniques.

Medical diagnosis is regarded as an important yet complicated part of medical services that needs to be executed accurately and efficiently. The automation of this stage (medical diagnosis) would be extremely advantageous. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is shortage of resource persons at certain places. Therefore, an automated medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. Appropriate computer-based information and/or decision support system can aid in achieving clinical tests at a reduced cost [6]. For example, Aiswarya Iyer et al noted in their study that automatic diagnosis of diabetes is an important real-world medical problem. Detection of diabetes in its early stages is the key for treatment [7]. As earlier stated due to the massive application of electronic health system in health organizations, if automated medical diagnosis systems are designed using data mining techniques, it will help doctors and health laboratory staffs diagnose disease in record time.

2.2 Classification

Classification is one of the most frequently studied data mining techniques by data mining and machine learning researchers. Classification consist of predicting a certain outcome based on a given input [8]. Classification technique is basically a two (2) step process that normally consist of (a) construction of the classification model (simply referred to as the learning phase and (b) application of the constructed model (simply referred to as the classification phase where given a set of data the model is used to predict class labels.)

Classification technique assigns items in a group or collection to target categories or classes. The end result is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify students based on grade as excellent, average, or poor.

A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case [15]. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating. In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown [15]. In finding out the effectiveness of our classification problem, the predicted values are compared to known target values in a test data set. The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared. Typically, the build data and test data come from the same historical data set. A percentage of the records is used to build the model; the remaining records are used to test the model. Test metrics are used to assess how accurately the model predicts the known values. If the model performs well and meets the business requirements, it can then be applied to new data to predict the future [14].

The accuracy of a classifier on a given set is the percentage of test set tuples that are correctly classified by the classifier. Various algorithms exist for carrying out classification example include, Decision tree, K-means Neighbor(KNN), Support Vector Machine (SVM), Naïve Bayes, etc.

3. Methodology

The experiment on data was carried out using Python programming language.

3.1 Description of Dataset

The Pima Indians dataset used for this study was obtained from the UCI machine learning repository. It is listed under the “Pima-Indians-diabetic” data name [12]. The dataset was originally owned by the National Institute of diabetes and Digestive and Kidney Diseases. The data set is made up of 768 number of instances with nine (8) attributes (features) for each instance of the dataset. The attributes in the dataset are listed as follows, and table 3.1 shows a sample of the dataset.

- Number of Times Pregnant (preg)
- Plasma Glucose Concentration (plas)
- Diastolic Blood Pressure (mm Hg) (bld pres)
- Triceps Skin fold thickness (mm) (skin fold)
- 2-hour serum insulin (insulin)
- Body mass index (Kg/m^2) (bmi)
- Diabetes Pedigree Function (pedi)
- Age(years) (age)
- Class Variable (class)

Table 3.1: Samples of pima-indians dataset

preg	plas	bld pres	skin fold	insulin	bmi	pedi	age	class
6	148	72	35	0	33.6	0.627	50	yes
1	85	66	29	0	26.6	0.351	31	no
8	183	64	0	0	23.3	0.672	32	yes

The second dataset used in this work are courtesy of Dr John Schorling, Department of Medicine, University of Virginia School of Medicine. The data consist of 15 variables on 403 instances from 1046 subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in

central Virginia for African Americans. The 403 subjects were the ones who were actually screened for diabetes. Glycosylated hemoglobin > 7.0 is usually taken as a positive diagnosis of diabetes [16].

3.2 Preprocessing Technique

Data quality is a top issue when dealing with prediction algorithms. We applied different preprocessing method for our two dataset in order to create variation in the data that we feed to the different algorithm that will be used.

3.2.1 Missing Values

Missing values in a dataset are generally regarded as noise and can impart the quality of classifiers. In order to get optimal result, we handled missing values in our data set by replacing missing values with mean value. After performing the above step we trained and tested our different classifier with the data and recorded the result.

3.2.2 Data Feature Relevance Selection

Feature selection is a vital technique that allows for the selection of only important data feature from an entire data set and then passed as input to the classification algorithm [15]. This ensures that only features in your data that contribute the most to the prediction variable or output in which you are in interested are considered. Feature selection helps in overcoming the problem of over fitting, improves accuracy level and reduces the training time required for our model. Since this study is on evaluating the performance of machine learning algorithms for predicting the incidence of diabetes, we decided to experiment with selecting features that serves best for the classification task. To perform feature relevance analysis, we applied the Univariate selection with chi squared (chi2) statistical test for non-negative features to select the top 5 features that meet our relevance analysis. From the pima-indian data set we obtained the following attributes as the top 5 features:
(1)plas (2) bld pres (3) age (4) pedi (5) bmi

The 7 most relevant features from our second data set include: cholesterol (2) bldg (3) age (4) weight (5) gender (6) hip (7) diastolic

3.3Target Variable

The class variable is nominal (categorical) and determines if a person had diabetes or not. The value 0 means diabetes is absent and the value 1 means it is present. The distribution of the class variable for the pima-Indian dataset is a follows:

- Patients determined to have diabetes (1's). Of the 768 encounters 264 were diagnosed of diabetes.
- Patients determined not to have diabetes (0's). 488 of the 768 encounters did not have diabetes. We want to predict when an individual would be diagnosed of diabetes based on the given attributes.

Figure 3.3 below shows the distribution of the target (class) variable for pima-Indian dataset while figure 3.4 represent the distribution for the target class for Dr Scholing dataset.

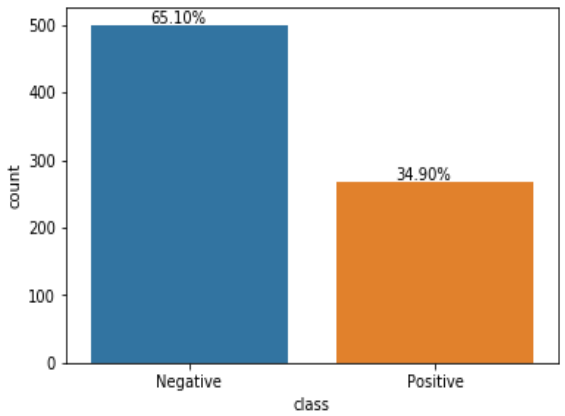

Figure 3.3

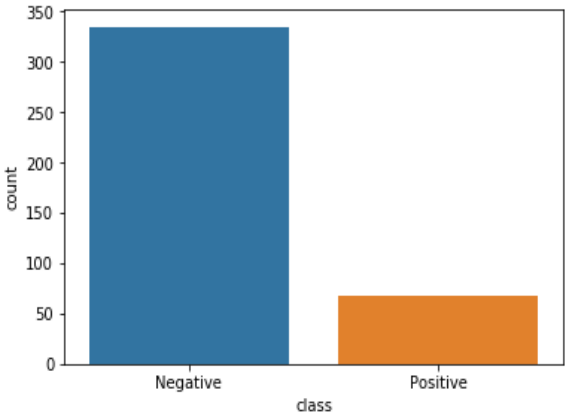

Figure 3.4

3.4 Classification Models

To evaluate the prediction of diabetes using machine learning technique we decided to use five (5) different classifiers which are:
(a) Logistic Regression
(b) Naïve Bayes
(c) Support Vector Machine
(d) XGBoost and
(e) K-Nearest Neighbor algorithms.

The performance and results are presented in section 5 below.

4. Results and Discussion

As proposed, Python programming language was used. Pima Indian diabetes and Dr. John Schorling data sets were used in evaluating the effectiveness of five (5) datamining algorithm for predicting the onset of diabetes. The process of feature selection and data cleaning presented a unique case for testing the performance of the various classifiers. We had to split our dataset into training and test data for the classification task using the 70/30 split ratio for training and test data respectively.

4.1Performance Analysis

After designing and implementing our classification model, it is important to analyze (measure) how effective our model has performed. The performance of the various machine learning algorithm are shown in table 5.1 and 5.2 below:

Table 5.1: Performance table for Pima-Indian dataset

Model	Precision	Recall	F1-Support
Logistic Regression	0.66	0.81	0.72
XGBoost	0.71	0.79	0.73
KNN	0.58	0.74	0.71
Naive Bayes	0.76	0.83	0.78
SVM	0.77	0.81	0.79

Table 5.2: Performance table for Dr. Schorling dataset

Model	Precision	Recall	F1-Support
Logistic Regression	0.75	0.75	0.75
XGBoost	0.72	0.71	0.71
KNN	0.64	0.69	0.68
Naive Bayes	0.79	0.79	0.76
SVM	0.62	0.65	0.64

From the above tables, we can observe the performance of each algorithm when used for predicting the incidence of diabetes. For the pima-indian dataset, Naive Bayes model performed the best with an accuracy of 0.79% followed by logistic regression with 0.75% recall value. On the other hand, SVM had the least performance of 0.65%. When we used our second dataset (dr. Schorling dataset) result show that Naive Bayes algorithm remained the top performing model with an accuracy of 0.83% followed by logistic regression and svm models performing at an equal rate of 0.81% accuracy while KNN was the least having a value of 0.74% accuracy.

From our result, we can clearly note that Naive Bayes model showed consistency across two different dataset and as such is highly recommended for use in health institutions which seek to deploy the use of machine learning for diabetes prediction. Another model with good consistency is logistic regression.

5. Conclusion/ Future Scope

The application of data mining technique to medical diagnosis is important in the world of medicine. The sudden deaths, heart attack, strokes, blindness, etc. associated with diabetes can be avoided through early diagnosis and treatment. This paper shows the data mining process involved in medical diagnosis and how different models perform when used for predicting the incidence of diabetes. Naïve Bayesian classifier from our experimental result is the most effective model among the five evaluated models.

We noticed also how model performance varied for the different dataset and wondered what could be the cause. Future study can focus on gathering new dataset that will present new insight and knowledge to enhance the prediction of diabetes using data mining technique. Also fine tuning techniques can be used to improve the performance of the models while means of handling imbalance class data can also be explored.

Availability of data and materials
http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/diabetes.html

Competing interests: The authors have no competing interest to declare.

References

[1] Alaa Elsayad, Mahmoud Fakhr, "Diagnosis of Cardiovascular Diseases with Bayesian Classifiers", Journal of Computer Sciences 11 (2), pp. 274 – 282.
[2] WHO,"Diabetes Factsheet," WHO Media Center, updated Nov. 2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs312/en/. [Accessed: Dec. 8, 2017].
[3] J. Han, M. Kamber, J. Pei, Data Mining Concepts and Techniques 3rd ed, Morgan Kaufmann Publishers, USA, 2012.
[4] B. Tamilvanan, V. Bhaskaran, "An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques", IOSR Journal of Computer Engineering (IOSR-JCE), p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 39-44.
[5] M. Langarizadeh, F. Moghbeli, "Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review", ACTA INFORM MED. 2016 OCT; 24(5): 364-369.
[6] J. Soni, U. Ansari, D. Sharma, S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011
[7] A. Iyer, S. Jeyalatha, R. Sumbaly, "Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
[8] R. Sanakal, S. T. Jayakumari, "Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine", International Journal of Computer Trends and Technology (IJCTT) – volume 11 number 2 – May 2014.
[9] American Diabetes Association, "Classification and diagnosis of diabetes", Sec. 2. In Standards of Medical Care in Diabetesd2015. Diabetes Care 2015;38(Suppl. 1):S8–S16
[10] R.. R.Patil, "Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, May 2014.
[11] P.Ramachandran, N.Girija, T.Bhuvaneswari, "Early Detection and Prevention of Cancer using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, July 2014 .
[12] UCI Machine Learning Repository [Online]. Available: https://archive.ics.uci.edu/ml/datasets.html. [Accessed: Dec. 4, 2017].
[13] B. Strack, J. DeShazo, C. Gennings, et al, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records", Hindawi Publishing Corporation BioMed Research International Volume 2014, Article ID 781670.
[14] Oracle Help Center, "Data Mining Concepts". [Online]. Available: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/toc.htm [Accessed: Jan. 10, 2018].
[15] J. Brownlee, "Feature Selection For Machine Learning in Python", [Online]. Available https://machinelearningmastery.com/feature-selection-machine-learning-python/ [Accessed: Jan 15, 2018]