# Application of classification algorithms in predicting diabetes

Pavan Movva
PG Student, School of Computing and IT
REVA University, BENGALURU.

Dr. Kiran Kumari Patil
Professor and Director UIIC at REVA University,
BENGALURU.

*Abstract*— **Machine learning is one kind of Artificial Intelligence (AI) that enables computers to learn without the need of programming. Classification problems is the task of classifying examples into one of a discrete set of possible categories. Classification algorithms helps in classifying the instances into a given set of categories. Decision tree, Naive Bayes classifier, K nearest Neighbors classifier, support vector machines are a few examples of classification algorithms. Classification algorithms are very helpful for classifying various medical data. As part of this paper, different classification algorithms will be used to classify diabetes patients. PIMA Indian diabetes data set is used for this purpose. Performance of all the algorithms will be analyzed and the best algorithm will be chosen as the outcome of this paper.**

**Keywords—Classification; PIMA Indian diabetes dataset; Type 2 diabetes.**

## I. INTRODUCTION

Diabetes mellitus (DM) is commonly known as diabetes. In this disease, there will be high levels of blood sugar over a prolonged period. Frequent urination, increased thirst and increased hunger are the symptoms of diabetes. If left untreated, diabetes can result into various complications. There are two reasons for diabetes:

a) Pancreas not producing the required insulin.

b) Body cells not responding properly to the produced insulin.

There are three types of diabetes:

a) Type 1 diabetes: This results when pancreas are not producing the required insulin. This occurs at a very young age of below 20 years. This is also referred to as juvenile diabetes.

b) Type 2 diabetes: This begins with a condition where cells fail to respond to insulin. The most common cause is lack of exercise and immoderate body weight. This is also referred as adult-onset diabetes. Type 2 diabetes is more prevalent.

c) Gestational diabetes: This occurs when pregnant women develop high levels of blood sugar.

Per 2015 worldwide statistics, 422 million people are estimated to have diabetes. Among the patients, 90% are suffering from Type 2 diabetes. This constitutes 8.3% of the adult population.

A person is diagnosed as diabetic if the blood sugar levels are above normal i.e.,

a) Fasting glucose level is greater than 6.9 mmol/l

b) Plasma glucose is greater than 11 mmol/l in case of glucose tolerance test.

This paper studies the onset of diabetes in pregnant women. Logistic regression, K Nearest neighbors, Naïve Bayes, CART and Support Vector Machines algorithms are used to predict the onset of diabetes. Results from all the algorithms are compared and presented.

## II. DATA SET DESCRIPTION

We will apply machine learning classification algorithms to data set which has females who are under a high risk of the onset of diabetes. The data set for this paper is the Pima Indian Population in Phoenix, Arizona. This data set has been taken from the UCI Machine Learning repository website. This data set was originally donated by Vincent Sigillito from the Applied Physics Laboratory at the Johns Hopkins University. This is one of the most popular datasets for the testing of classification algorithms. The minimum age of females is 21 years. There are eight features and one class in this dataset. All features are numerical values. There are in total 768 instances in the Pima Indians diabetes dataset out of which 268 samples are for diabetes positive and 500 samples are for diabetes negative.

Following are the attributes present in this dataset.

1. Number of times a female is pregnant
2. Plasma glucose concentration a Two hour in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. Two-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age of the female (years)
9. Class variable (0 or 1)

Class variable takes the value of 0 or 1 where 0 means tested negative for diabetes and 1 means tested positive for diabetes. The Pima Indian diabetes date set is widely used for testing various binary classification algorithms.

## III. LITERATURE REVIEW

Considerable research has been performed for diabetes diagnosis. Several classification algorithms were used.

"Examining Classification Techniques in Data Mining for PIMA Indian Diabetes Dataset" is a research paper presented by S.Janani, Research Scholar, D. Ramya Chitra, Assistant Professor, Department of Computer Science and Engineering, Bharathiar University, Coimbatore.

As part of this paper, the performance of five classification algorithms are assessed to predict the onset of diabetes. J48, FT, Decision Table, Multilayer Perceptron and Naive Bayes are the selected classification algorithms. PIMA Indian Diabetes dataset from UCI machine learning repository is used for this binary classification problem. WEKA tool is used to study the performance of these algorithms. Various performance factors like execution time, error rate and

classification accuracy are taken into consideration to assess the performance of these selected classification algorithms.
As part of the experimental results, it is observed that the Multilayer Perceptron is performing better when compared to the other four classification algorithms. [1]

"Classification Of Diabetes Disease Using Support Vector Machine", is a research paper presented V. Anuja Kumari, PG Student, R. Chitra, Associate Professor, Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, Kumaracoil, Kanyakumari District, India.
As part of this paper, Support Vector Machine(SVM) with radial basis function kernel is used for the binary classification of the diabetes patient records. PIMA Indian Diabetes dataset from UCI machine learning repository is used for this binary classification problem. MATLAB R2010a is used to conduct the experiments. MS Excel is used for the purpose of storing the datasets. Various performance factors such as accuracy, sensitivity and specificity are considered for assessing the performance of the SVM classifier. As part of the results, it is observed that 78% accuracy, 80% sensitivity and 76.5% specificity is achieved using the SVM classifier.
From the experimental results, it can be concluded that SVM classifier can be used as a good option for diabetes data set classification [2].

"Analysis of a Population of Diabetic Patients Databases with Classifiers", is a research paper presented by Murat Koklu from Selcuk University Technical Education Faculty, Department of Electronics and Computer Education, Konya, Turkey, and Yavuz Unal, Education Faculty from Amasya University, Computer Education and Instructional Technology Department, Amasya, Turkey.
As part of this paper, J48, Multilayer Perceptron and Naive Bayes classifier algorithms were used to analyse the onset of diabetes. PIMA Indian Diabetes dataset from UCI machine learning repository is used for this binary classification problem. WEKA tool is used for the analysis. It was observed that Naive Bayes classifier was the best classifier among the selected binary classifiers. It achieved an accuracy of 76.302%. [3]

"DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES", is a research paper presented by Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, Department of Computer Science, BITS Pilani Dubai, United Arab Emirates.
Decision Tree and Naïve Bayes classifier techniques are used for diagnosis of diabetes. There are several decision tree algorithms available. In this paper, J48 algorithm is selected for predicting diabetes. PIMA Indian Diabetes dataset from UCI machine learning repository is used for this binary classification problem. WEKA tool is used for the analysis. [4]

## IV. METHODOLOGY

### A. LOGISTIC REGRESSION

Logistic regression is a technique present in statistics. Logistic regression belongs to linear machine learning algorithm class. It is the best method used for solving binary classification problems. Logistic function is used as part of this algorithm. Hence the name logistic regression. Sigmoid function is the other name for logistic function. Its input is a real-valued number and the output is a value ranging between 0 to 1. Basically, this function maps a real number to a value between 0 to 1.

Below is an example logistic regression equation:
y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

### B. K NEAREST NEIGHBOURS

In the KNN model, entire training data set is stored. KNN model uses the complete training data set for prediction. As part of prediction, entire training data set is searched for K neighbours. To find out the K neighbours, distance metric will be used. There are several distance metrics available like Euclidean distance, Hamming distance, Manhattan distance, Mahalanobis and Minkowski distance. For real-valued attributes, Euclidean distance measure can be used. The complexity of KNN increases with the volume of training dataset. KNN can be applied for both classification and regression problems. KNN is a straightforward algorithm. It is easy to understand and simple to implement.

### C. NAIVE BAYES

Naive bayes is a very powerful algorithm used for prediction. It is simple to understand. This algorithm can be used for both binary and multi-class classification problems. It belongs to supervised learning methods. Probability is used for the classification purpose. For naive bayes, two types of probabilities need to be obtained.

a) class probabilities : The probability of each class is calculated from the given training data set.
In a binary classification problem, there will be two classes.
E.g. : Let's says class A and B are the 2 classed in a binary classification problem.
class probability(A) = number of instances (class A)/ (number of instances (class A) + number of instances (class B))

b) conditional probabilities : the conditional probability is the probability of each input value given each class value.

### D. CLASSIFICATION & REGRESSION TREES

Decision trees are a pivotal type of algorithms for prediction in machine learning. There are several classical decision tree algorithms that are around for decades. Several modern variations are also available like random forest.ID3, C4.5, CART, J48 are some of the examples of decision tree algorithms. We will use CART algorithm which means Classification and Regression

Trees. Leo Breiman introduced the term CART. This term refers to the decision tree algorithms that can be applied for classification or regression type of prediction problems. CART provides base for several other algorithms like random forest, boosted decision trees and bagged decision trees.

Decision trees are extremely popular because the model is very easy to understand. Model representation in CART. CART uses binary tree for model representation. Input variable(x) is represented by root node. Leaf nodes represents the output variable(y). Decision tree can also be represented as a set of rules. Making Predictions When a new input is given, the tree will be traversed.

E. SVM

Support Vector Machine is one of the most popular machine learning algorithms. SVM belongs to supervised learning method. SVM requires a dataset which is labelled. It is an algorithm with high performance. It requires little tuning.

How SVM works:

the input variables in the data set forms a n-dimensional space. A hyperplane splits these input variables. Classifications are made based on this hyperplane. If the new input point is above the hyper plane then it belongs to class 0. If the new input point is below the hyper plane then it belongs to class 1. If the new input point is close to hyper plane then it is difficult to classify. Margin is the distance between the line and the closest input data points. Maximal-margin hyper plane is the best line that separates the 2 classes and has the largest margin. the input data points which are closest to the line which separate the classes are known as support vectors. These support vectors play a pivotal role in placing the line. SVM algorithms are implemented using kernel.

## RESULT AND CONCLUSION

Python programming language is used to analyze the performance of the selected classification algorithms. Scikit-learn machine learning library is used to implement these algorithms. Table shows the classification results.

| ALGORITHM | ACCURACY |
|---|---|
| Logistic Regression | 76.95 % |
| KNN | 72.65% |
| Naïve Bayes | 75.51% |
| CART | 69.12% |
| SVM | 65.10% |

**Table 1**

From the experimental results, the selected techniques achieved the prediction accuracy of 76.95%, 72.65%, 75.51%, 69.12% and 65.10% for Logistic Regression, KNN, Naïve Bayes, CART and SVM respectively. It can be observed that Logistic Regression has the best performance among the others on Pima Indian diabetes dataset.

Acknowledgment

I express sincere thanks to Dr. Kiran Kumari Patil and Prof. Shilpa Chaudhari for their great support and guidance.

## REFERENCES

[1] S.Janani , D. RamyaChitra," Examining Classification Techniques in Data Mining for PIMA Indian Diabetes Dataset" , IJSRD - International Journal for Scientific Research & Development| Vol. 4, Issue 09, 2016.

[2] V. Anuja Kumari, R.Chitra,"Classification Of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications 2013.

[3] Murat Koklu and Yavuz Unal,"Analysis of a Population of Diabetic Patients Databases with Classifiers", World Academy of Science, Engineering and Technology
International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering Vol:7, No:8, 2013.

[4] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes using classification mining techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.