ISSN 2319 – 7595
Volume 8, No.2, March - April 2019
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs13822019.pdf
https://doi.org/10.30534/ijiscs/2019/13822019

Machine Learning Approach For Diabetes Prediction

Jensia Thomas[1], Anumol Joseph[2], Irene Johnson[3], Jeena Thomas[4]
[1]St.Joseph's College of Engineering,India, jensia.thomas7@gmail.com
[2]St.Joseph'sCollege of Engineering,India, anu.joseph0511@gmail.com
[3]St.Joseph's College of Engineering,India, irene.johnson310@gmail.com
[4] St.Joseph's College of Engineering,India, jeena.thomas@sjcetpalai.ac.in

ABSTRACT

Diabetes the silent killer which kills part by part of our life Diabetes can strike anyone, from any walk of life. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases etc. It is found that in last decades the cases of people living with diabetes jumped almost 350 million worldwide. Early detection of diabetes can reduce the health risk in patients. With rise of new technology like machine learning we have a solution to this issue, a system for predicting the chance of diabetes. Machine learning techniques increase medical diagnosis accuracy and reduce medical cost. This paper aims to predict diabetes via supervised machine learning algorithms such as decision tree. In our work we have used the Pima Indians Diabetes Dataset from UCI Machine Learning Repository.

Key words: *diabetes,decision tree, machine learning.*

1. INTRODUCTION

Diabetes is a disease caused by decreased production of insulin or by decreased ability to use insulin.Diabetes[8] is one of the most endocrine disorders that affect 425 million people worldwide.It affect both male and female with different age group.The early identification of diabetes[9] is only remedy to escape from the threat of this deadliest disease. Eating an unhealthy diet being overweight or obese and not exercising enough may play a role in developing diabetes. The early symptoms of untreated diabetes are frequent urination,increased thirst,feeling tired, hungry and vision problems and they have easy chance of developing infections in bladder,skin and in vaginal area.

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science encompasses the fields of data mining and big data. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data

Machine Learning (ML) is a division of algorithm that allows software applications to become more definite in predicting outcomes without being explicitly programmed. The basic aim of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. Machine learning algorithms are classified into supervised and unsupervised.

1.1 Supervised learning

Supervised algorithms need a data scientist with machine learning skills to give both input and desired output.Data scientists determine which variables, or features, the model should analyze and use to develop predictions. Once training is finished, the algorithm will apply what was learned to new data.

Supervised machine learning systems provide the learning algorithms with known measures to support future judgments. chatbots, facial recognition programs, expert systems and robots are among the systems that may use either supervised or unsupervised learning. Supervised learning systems are mostly associated with recovery-based AI but they may also be capable of using a generative learning model.

Supervised learning problems can be further grouped into regression and classification problems.
Classification: A classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease".
Regression: A regression problem is when the output variable is a real value, such as "dollars" or "weight".

1.2 Unsupervised learning

Unsupervised algorithms do not need to be trained with expected outcome data. Instead, they use an iterative method called deep learning to analyze data and arrive at conclusions. Unsupervised learning algorithms are also called neural networks.They are used for more complicated processing tasks than supervised learning systems, including image recognition, speech-to-text, pattern recognition and automatic target recognition. These neural networks work by linking through millions of examples of training data and identifying often subtle correlations between many variables. Once trained, the algorithm can use its store of associations to interpret new data. These algorithms have only become beneficial in the age of big data, as they demand massive amounts of training data.

The work focus on predicting the chance for diabetes using various machine learning techniques such as decision tree,logistic regression and naive Bayes.Performance of these algorithms are compared and better one is chosen for prediction.The remaining part of the paper is organized as follows.Section II includes related works of diabetes prediction. Section III is about proposed system,it shows the working of our system from data collection, classification and to final results .Section IV is about the conclusion and future works it shows about further improvement we can apply on this system.

2. RELATEDWORKS

The section analyses various research works that are related to the proposed work.VaishaliAggarwal[4] offered a performance analysis based on the competitive learning algorithms on Gaussian data for spontaneous cluster selection and also studied and evaluated the performance of these algorithms and randomized results have been analysed on 2-D Gaussian data with the learning rate parameter kept simple for all algorithms. Algorithms used in their work include clustering algorithm and frequency sensitive algorithm. Supervised learning machine algorithms are used for classification of the Gaussian data. K.Srinivas [7] established applications of datamining techniquesin healthcare and prediction of heart attacks. This research used medical profiles such as age, sex, blood pressure and blood sugar and predicted the possibility of patients getting a heart and kidney problems .

Song et al. [1] defined and explained different classification algorithms using different parameters such as Glucose, BloodPressure, Skin Thickness and insulin. The studies were not included pregnancy parameter to predict diabetes disease (DD). In this work, the researchers were using only small sample amount data for prediction of Diabetes. The algorithms were used by this paper were five different algorithms GMM, ANN, SVM and EM. Finally. The researchers conclude that ANN (Artificial Neural Network) was providing High precision for prediction of Diabetes.

Xue- HuiMeng et al. [3] in this study the researchers used different data mining techniques to predict the diabeticdiseases using real world data sets by collecting material by spread questioner .in this study SPSS and weka tools were used for data analysis and prediction respectively .in this study the researchers equate three techniques ANN, Logistic regression, and j48 .finally it was concluded as j48 [6] machine learning technique provide efficient and better accuracy.

Bashir,S.Qamar[5]proposed individual disease risk prediction based on medical history. This paper also predicts each patient's greatest disease risks based on their own medical history data.Dataset are used for medical coding and concerted assessment and recommendation engine (CARE) information technique. From this literature, it is observed that the machine learning algorithms place a important role in knowledge detection form the databases especially in medical diagnosis with the medical data.

Sajidaet al.[2]by using CPCSSN(Canadian primary care sentinel surveillance Network) dataset and three machine learning methods to predict the diabetes Disses (DD) in early stage to safe human life at from early death .on this study Bagging ,Adaboost,and decision tree(J48) were used to predict the diabetes and the researcher was compare the result of those methods and concluded that Adaboost method was provide active and better accuracy than the other methods inweka data mining tools.

Pradhan et al in [4] used Genetic programming (GP) for prediction of diabetes. Results achieved using Genetic Programming gives optimal accuracy as compared to other implemented techniques.

3.PROPOSED SYSTEM

Machine learning classification algorithms namely Decision Tree are used in this prediction. From them the best algorithm is selected and used for earlyprediction of diabetes. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository.

3.1 DECISION TREE

Decision tree is one of the supervised machine learning algorithms. It is useful in making decision by producing tree like model of decision where the data is continuously split according to a certain parameter. The tree contain two units namely decision nodes and leaves. Decision nodes are where the data is split and the leaves are the decisions or the final outcomes. From the training data, the algorithm generates decision trees to solve classification and regression problem.

3.2 DECISION TREE ALGORITHM

• Select the best attribute using attribute selection measure like gini index
• Make that attribute as the decision node and breaks the dataset into smaller subsets
• Start tree building by repeating this process recursively for each child until one of the condition match
• All the tuples belong to the same attribute value
• There are no more remaining attribute

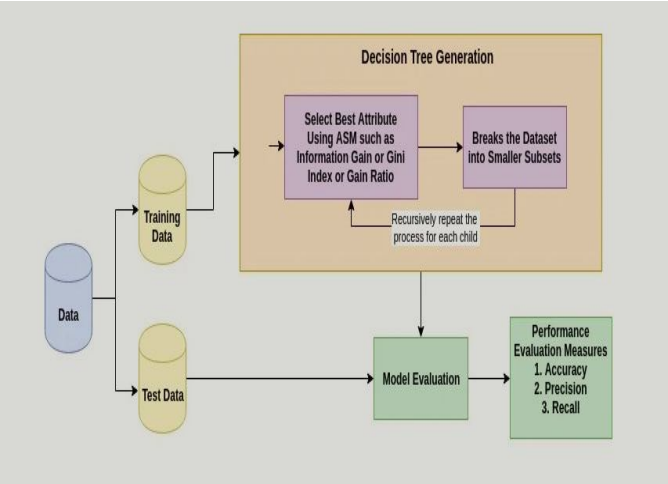

Figure 1: Decision Tree Generation

This figure shows proposed model for decision tree. Firstly the data is divided into two types training data and testing data. Training data is used to train the model and testing data is used to test the trained model.The trained data is passed into decision tree generator.In decision tree generator, from the training data

the best attribute is selected using gini index. Model is evaluated by test data in terms of accuracy,precision and recall.

Accuracy = TP+TN/TP+FP+FN+TN
(1)

Precision = TP/TP+FP
(2)

Recall = TP/TP+FN
(3)

3.3 DATASET

Dataset is taken from the National Institute of Diabetes and Digestive and Kidney Diseases.
This dataset contain data of 768 people.
Attributes:
Plasma glucose concentration a 2 hours in an oral glucose tolerance
Diastolic blood pressure (mm Hg)
Hour serum insulin (mu U/ml)
Body mass index (weight in kg/(height in m)^2)
Age (years)
Class variable (0 or 1)
0-it indicates person has diabetes
1-It indicates no diabetes for person

Glucose	BloodPres	Insulin	BMI	Age	Outcome
148	72	0	33.6	50	1
85	66	0	26.6	31	0
183	64	0	23.3	32	1
89	66	94	28.1	21	0
137	40	168	43.1	33	1
116	74	0	25.6	30	0
78	50	88	31	26	1
115	0	0	35.3	29	0
197	70	543	30.5	53	1

Figure 2: Pima Indian Diabetes Data Set

3.4 BUILDING DECISION TREE MODEL

```
#splitting the data into train and test
X_train,X_test,Y_train,Y_test=train_test_split(x,y,test_size=0.05,random_state=0)
```

3.5 PREDICTION

```
#predicting

y_pred=clf.predict(X_test)
```

Accuracy can be computed by comparing actual test set values and predicted values.

```
#testing the accuracy

accuracy=accuracy_score(Y_test,y_pred)
print(str(accuracy*100)+"% accuracy")

#testing the precision
average_precision = average_precision_score(Y_test,y_pred)

print('Average precision-recall score: {0:0.2f}'.format(average_precision))
```

4. EXPERIMENTAL RESULTS

Decision tree structure of Pima dataset is shown in Figure 3. The root node is glucose which can show the glucose has the max gini index.

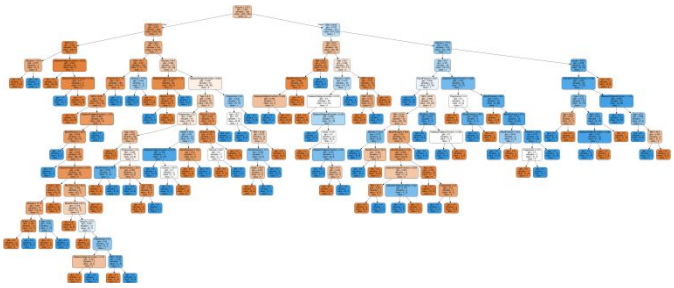

Figure 3: Decision Tree

In the decision tree chart[7], each internal node has a decision rule that splits the data. Gini referred as Gini ratio, which measures the impurity of the node. A node is pure when all of its records belong to the same class, such nodes known as the leaf node.

$$\text{Gini index} = 1-\sum_j p_j^2 \quad (4)$$

Here, the resultant tree is unpruned. This unpruned tree is unexplainable and not easy to understand**.**

4.1 OPTIMIZING DECISION TREE PERFORMANCE

Criterion: optional (default="gini") or Choose attribute selection measure: This parameter allows us to use the different-different attribute selection measure. Supported criteria are "gin" for the Gini index and "entropy" for the information gain.

Splitter: string, optional (default="best") or Split Strategy: This parameter allows us to choose the split strategy. Supported strategies are "best" to choose the best split and "random" to choose the best random split.

Max depth: int or None, optional (default=None) or Maximum Depth of a Tree: The maximum depth of the tree. If None, then nodes are expanded until all the leaves contain less than

min_samples_split samples. The higher value of maximum depth causes over fitting, and a lower value causes under fitting.

```
#training the classifier
clf=tree.DecisionTreeClassifier(criterion='gini',min_samples_split=30,splitter="best")
clf=clf.fit(X_train,Y_train)
```

4.2 VISUALISATION

```
#visualising the training set results
precision, recall, _ = precision_recall_curve(Y_test, y_pred)
plt.step(recall, precision, color='b', alpha=0.2,
         where='post')

plt.fill_between(recall, precision, step='post', alpha=0.2,color='b')
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.ylim([0.0, 1.05])
plt.xlim([0.0, 1.0])
plt.title('2-class Precision-Recall curve: AP={0:0.2f}'.format(
          average_precision))
plt.show()
```

5. CONCLUSION

One of the important real-world medical problems is the detection of diabetes at its early stage. . Experiments are performed on Pima Indians Diabetes Database. The prediction analysis is the technique in which user predicts the future on the basis of current situations. In this work, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, decision tree algorithm is studied and evaluated based on accuracy. Experimental results showed the adequacy of the designed system with an achieved accuracy of 87 %. Decision trees tend to over fit quickly at the bottom and also have poor prediction accuracy for responses with low sample sizes. The work can be extended and improved by using logistic regression on our prediction system. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

REFERENCES

1. Komi, Messan, Jun Li, YongxinZhai, and Xianguo Zhang, **Application of data mining methods in diabetes prediction**, *IEEE Trans. On Data Mining*,Vol.3 pp.570-578,july 1996

2.Perveen, S., Shahbaz, M., Guergachi, A., &Keshavjee, K. (2016),**Performance analysis of data mining classification techniques to predict diabetes**.*Procedia Computer Science, 82, 115-121.*
https://doi.org/10.1016/j.procs.2016.04.016

3.X.H Huang, Y. X.Rao(2013),**Comparison of three data mining models for predicting diabetes or prediabetes by risk factors.** The Kaohsiung journal of medical sciences, 29(2), 93-99.
https://doi.org/10.1016/j.kjms.2012.08.016

4.Pavate, Aruna, and NazneenAnsari,Risk**Prediction of Disease complications in Type 2 Diabetes patients using soft computingTechniques.**

5.Bashir, S., Qamar, U., Khan, F. H., &Javed, M. Y. (2014, December).**An Efficient Rule-Based Classification of Diabetes Using ID3, C4.5,& CART Ensembles**. In Frontiers of Information Technology (FIT), *2014 12th International Conference on (pp. 226-231).IEEE.*
https://doi.org/10.1109/FIT.2014.50

6.Pradeep, K. R., & Naveen, N. C. (2016, December).**Predictive analysis of diabetes using J48 algorithm of classification techniques.** In Contemporary Computing and Informatics (IC3I), *2016 2nd International Conference on (pp. 347-352). IEEE.*
https://doi.org/10.1109/IC3I.2016.7917987

7.AsmaA.Aljarullah, **Decision tree discovery for the diagnosis type 2 diabetes**,*IEEE, International conference on innovation in information technology*, pp 303-307, 2011
https://doi.org/10.1109/INNOVATIONS.2011.5893838