9808:675X
HIGHLY CITED JOURNAL
ACCEPTANCE RATION BELOW: 8%

International Machine learning journal and Computer Engineering

**Diabetes Risk Assessment using Machine Learning: A Comparative Study of Classification Algorithms**

**Dr. Emily Johnson**

**Affiliation: Department of Computer Science, Stanford University**

**Abstract**

Diabetes is a serious health condition with high blood glucose/sugar levels. Diabetes is a chronicle disease that can cause worldwide health care crisis but we can take some steps to manage these crisis. IN Diabetes Blood sugar/glucose is the main source of energy that is drawn from the food we eat in our day to day life.

Insulin is a hormone that is produced by the pancreases in our body which helps the glucose gets into the cells which can be used for energy to perform day to day activities. When body doesn't make enough or any insulin then glucose stays in the blood which might lead to various health problems like heart attack, strokes etc.

There are many types of diabetes like type1, type2, gestational and monogenic diabetes where type1 and type2 are the most common ones.type1 is mostly diagnosed in young adults and children and type2 is mostly diagnosed in middle-age or older group of people.

Machine learning is a scientific field here machine learn from the human experiences the aim of the project is to build a system which can predict whether the patient is diabetic or not with a high accuracy by combining result of various machine learning technique with the algorithm used like KNN, logistic regression, random forest etc.

The accuracy of the model using each algorithm is calculated then the one with more percentage of accuracy is taken as the model for predicting diabetes.

**1. Introduction**

Diabetes is growing faster among the young adults and children because of the hectic schedule due to which they have less amount of time to perform physical activity. In order to understand diabetes first need to understand what all happen in the body and how blood glucose/sugar is produced.

The food that we eat is the main source from where we gain glucose. Then the glucose produced travels around the body in the blood streams out of which some amount of sugar/glucose is taken to our brain

which help the brain to think clearly without any confusion and function properly. The remaining glucose is further taken to the cells of our body and also to the liver. Liver is the

Body's glucose (or fuel) reservoir, and helps to keep your circulating blood sugar levels steady and constant. The liver both stores and manufactures glucose in the form of stored energy which can be used by the body as and when required depending upon the body's need. The need to store or release glucose is by the hormones insulin and glucagon. Insulin is required for using the stored energy and is produced by pancreas. Some time when the pancreases are not able to produce enough insulin in the body then the amount of sugar/glucose increases in the blood stream and gives rise to diabetes.

Diabetes mellitus is another name given to diabetes which occurs when blood glucose/sugar is too high in urine or bloodstream.

**1.1 Types of diabetes**

- **Type 1**

In this type of diabetes the body produces very less amount or no insulin. The immune system attacks and destroys the cells in your pancreas that makes insulin so the immune system is compromised.it can be caused by an autoimmune reaction. Approximately 5-10% of the people that are predicted diabetic fall under this category. Symptoms are developed qquickly. usually Children and young adults are the ones that are mostly diagnosed with type 1 diabetes. There are currently No preventative measures are known.

- **Type 2**

It is the most common type of diabetes. In this type of diabetes the body produces a low quality of insulin or doesn't use insulin well. Type 2 diabetes can developed at any age. Usually the middle-age and older people are diagnosed with type 2 diabetes. 95% of the people have type 2 diabetes but can be prevented or delayed with the healthy lifestyle changes **(eat healthy, sleep on time ,being active etc.)**

- **Gestational diabetes**

It is most common in some women when they are pregnant. This type of diabetes is not permanent in most of the cases goes away after the baby is born if to third of the cases it can reappear in the net pregnancies. The person had gestational diabetes, and then there are greater chances of developing type 2 diabetic later in future.

- **Monogenic diabetes**

It's is rare condition that result from changes in a single gene. The monogenic diabetes can be inherited from a parent who is predicted with diabetes.2-5% group of the people are affected with this type of diabetes.

**1.2 Symptoms of diabetes**

The symptoms vary from person to person depending upon how high there blood sugar is. But the person with type 2 or pre-diabetic may not have symptoms.

- Losing weight
- Urinating often
- Constantly tired and weak
- Blurry vision
- Slow healing of wound
- Feeling thirsty than usual
- Mood swing

**1.3 Cause of diabetes**

The most common causes that boost the diabetes factor are the genetic, lifestyle and environment. Eating an unhealthy/junk food, being overweight/ obese and not exercising enough can be the reason for developing diabetes, particularly Type 2 diabetes. Type 1 diabetes is caused by an autoimmune response.

**2. Methodology**

The objective of this paper is to identify the model to predict diabetes with best accuracy. Various classifiers are used in order to predict whether the person is diabetic or not. Five methods are used for the prediction. The accuracy of the model using each algorithm is calculated then the one with more percentage of accuracy is taken as the model for predicting diabetes.

**Dataset description:**

The data is gathered from https://www.kaggle.com/johndasilva/diabetes.

The dataset have many attributes of 2000 patients.

| SNO | Attributes |
|---|---|
| 1 | Pregnancies |
| 2 | Glucose |
| 3 | BloodPressure |
| 4 | SkinThickness |
| 5 | Insulin |

| | |
|---|---|
| 6 | BMI |
| 7 | DiabetesPedigreeFunction |
| 8 | Age |
| 9 | Outcome |

Table 1: Dataset Description

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 138 | 62 | 35 | 0 | 33.6 | 0.127 | 47 | 1 |
| 1 | 0 | 84 | 82 | 31 | 125 | 38.2 | 0.233 | 23 | 0 |
| 2 | 0 | 145 | 0 | 0 | 0 | 44.2 | 0.630 | 31 | 1 |
| 3 | 0 | 135 | 68 | 42 | 250 | 42.3 | 0.365 | 24 | 1 |
| 4 | 1 | 139 | 62 | 41 | 480 | 40.7 | 0.536 | 21 | 0 |

Table 2: Dataset record

- Dataset consist data of 2000 patient with 9 features.
- “Outcome” attribute is a class variable that shows the outcome 0 and 1 which means positive or negative for diabetes.
- The model developed to predict diabetes is slightly imbalanced having around 200 patients with 0 means no diabetes and 500 labeled 1 means diabetic

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

Fig 1: represents that there is no feature it zero value

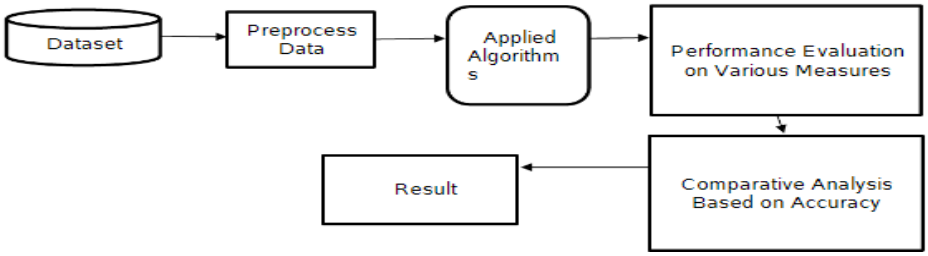



Fig 2: proposed model diagram flow

**Data processing**

Mostly the dataset used healthcare contains missing values and impurities that can affect the effectiveness of the data. In order to improve the quality and effectiveness of data obtain after mining, data processing is performed. So that machine learning techniques can be used on the dataset effectively.

- Missing values removal

This removes all the zero values as they have zero worth which is not possible therefore the value is eliminated

- Splitting of data

Once the cleaning of the data is done data is normalized in training and testing models. The aim of normalization is to bring all the attributes under same scale.

**Apply Machine learning Algorithms**

No the data is ready and we can apply machine learning technique on it. Different techniques are used to predict diabetes. The main objective of applying machine learning techniques is to analyze the performance of these methods and estimate accuracy.

Correlation Matrix:

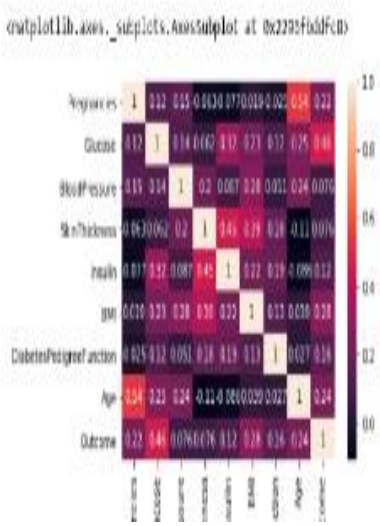


Fig3: Graph showing correlation

As we can see there is no feature that has high correlation with the outcome feature. Some features have negative correlation and rest have positive correlation

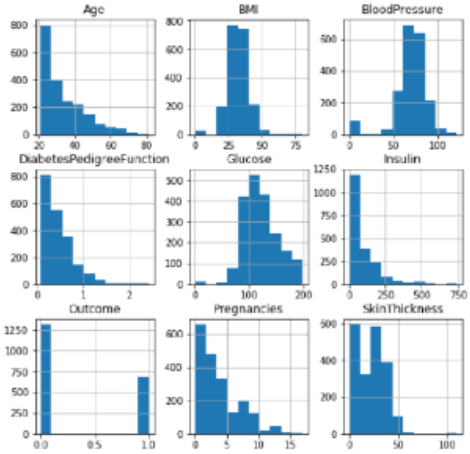



Fig 4: label outcome

In the above graph we can see that different label and features are distributed along different ranges. Each of the bar is actually a categorical variable which need to be handle before applying machine learning.

**2.1 Decision tree algorithm**

Decision tree is mostly preferred to solve classification problem but can be used for both classification and regression problem. That is why t is a supervised learning technique. The classifier (tree structured) create a decision tree based on that, it assigns the class value to each point.

The maximum number of features can vary while creating the model.

Training accuracy=100%

**Algorithm-**

Step 1: Construct a tree with nodes as input are the feature.

Step 2: Select feature from which you will predict the output from input feature.

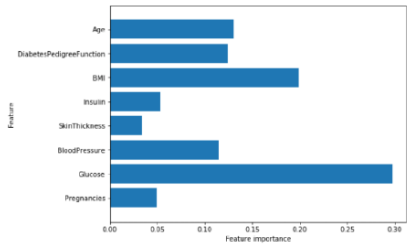
Step 3: The highest information gain is calculated for each attribute.

Step 4: Repeat step 2

Step 5: After that form a sub tree using the feature which is not used in above node.

**Feature importance**

How important a particular feature is in decision tree making is determined by the feature importance rate.

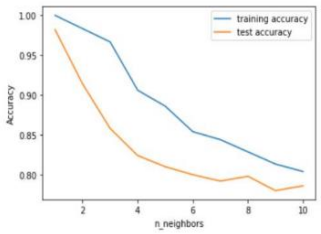



From the above bar graph we can see that glucose is the most important feature

**2.2 KNN**

Also known as k-Nearest Neighbor based on supervised learning and is one of the simplest machine learning algorithms. KNN can be used to solve classification and regression problem but is a lazy prediction as compared to other method. The algorithm builds the model consisting only stored training dataset.in order to predict the new data point the algorithm finds the closest data point “nearest neighbor”. The closest data point in the training data set its nearest neighbor. The k= number of nearby neighbors and is always a positive integer. The closeness is defined in terms of Euclidean distance

$$\text{Euclidean Distance between } A_1 \text{ and } B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$$




In the above graph here the the y-axis represent accuracy of training and testing and x-axis are the n neighbors against whom the y axis setting. But hen more number of neighbors are considered then training accuracy drops and the model becomes too complex.

Training accuracy=.81

Testing accuracy=.78

**Algorithm**

Step 1: Take a sample dataset.

Step 2: Take a test dataset.

Step 3: Find the Euclidean distance with the help of equation.

Step 4; Decide a random value of K. is the no. of nearest neighbors

Step 5: with minimum distance and Euclidean distance find out the nth column of each.

Step 6: Find output values.

**2.3 Random forest**

Random forest was developed by Leo Bremen. It's used for classification as well as for the regression tasks. Comparatively gives greater accuracy then the other models present.it can handle large datasets.it is type of ensemble learning method. Improves performance by reducing variance.

Algorithm-

Step 1: select the D features from the total features m where D<<M.

Step 2: from the R features, the node using the best split point.

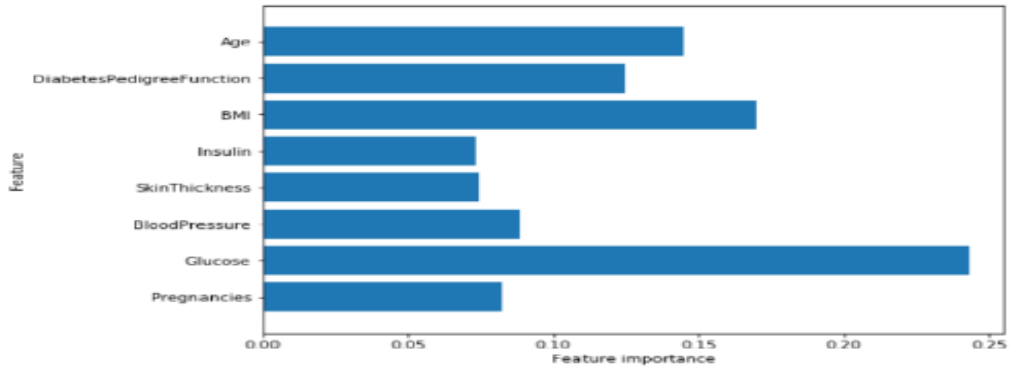
Split the node into sub nodes.

Step 3: Repeat step 1 to 2 until l number of nodes has been reached.

Step 4: Built forest by repeating from step 1 to 3 for a number of times to create n number of trees.

Now the classifier takes the decision tree t net level by creating forest of trees. Where each tree is generated by a random selection of features

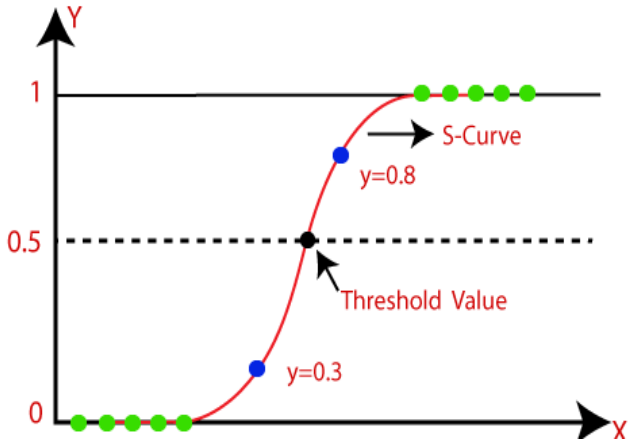
| | |
|---|---|
| Training accuracy | 1.00 |
| Testing accuracy | 0.98 |




Still glucose is the most important feature and followed by BMI (As second most important feature).

**2.4 Logistic regression**

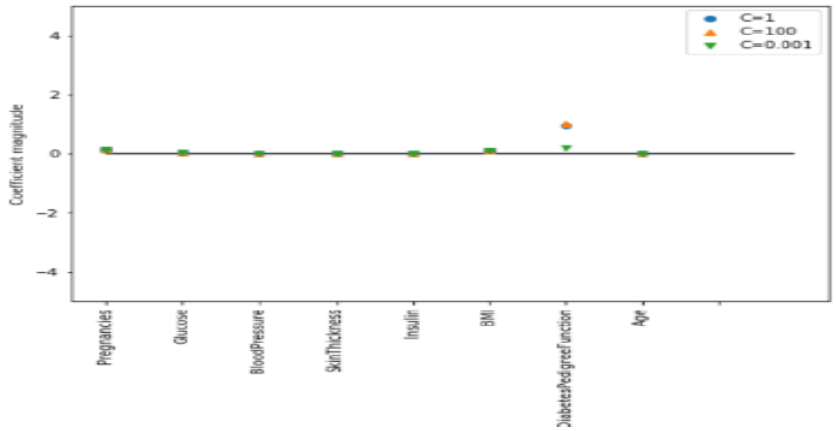
Popular machine learning algorithm that is why it's also a supervised learning algorithm it estimate the probability of response(binary) based on individual or group of predictors. The predicted output is of the categorical dependent variable therefore it must be categorical or discrete i.e.(yes or no, 0 or 1, true or false).the objective the algorithm is to best fit that describe relation between target and predictor variable. This model uses sigmoid function for prediction that whether person is diabetic or not as shown in fig below. Sigmoid function P = 1/1+e – (a+bx) Here P = probability, a and b = parameter of Model.



**Condition involved**

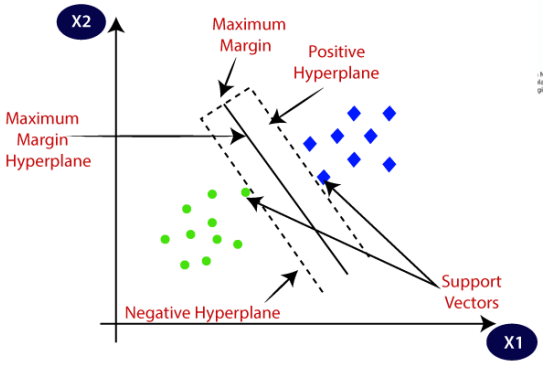
| conditions | Training accuracy | Testing accuracy |
|---|---|---|
| C=1 | 0.779 | 0.784 |
| C=.01 | 0.784 | 0.780 |
| C=100 | 0.778 | 0.792 |

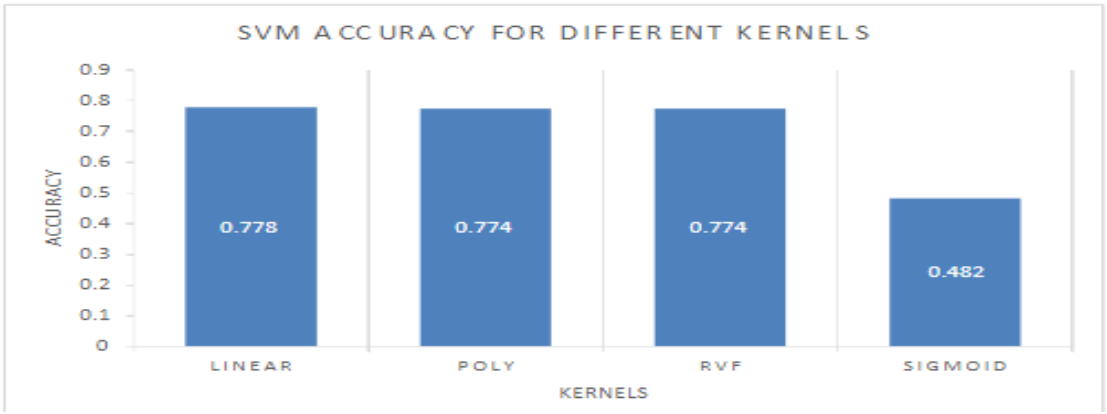
Therefore we choose c=1 because in c=0.01 both training and testing have the same accuracy and c=100 training test result less accuracy then the testing set.




**2.5 Support vector machine**

Another name given to the support vector machine is SVM. In support vector machine hyper plane are created that differentiate between the two classes as much as possible by adjusting the distance between the data points and hyper plane. It can create an individual or a set of hyper plane in the high dimensional space. Several kernels based on which the hyper plane is decided. The kernel used are linear, poly, rvf and sigmoid as shown below.




**Algorithm-**

- Select the hyper plane which divides the class bet- ter.
- To find the better hyper plane you have to calcu- late the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to
- Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

**3. Output**

Accuracy comparison table

| Algorithms | Training Accuracy | Testing Accuracy |
|---|---|---|
| k-Nearest Neighbors | 81% | 78% |
| Logistic Regression | 78% | 78% |
| Decision Tree | 98% | 99% |
| Random Forest | 94% | 97% |
| SVM | 76% | 77% |

While designing the project the aim of the project was to develop a model for the prediction of diabetes and perform analysis using machine learning method which has been achieved. As we know diabetes is the chronicle disease so it becomes important that the detection of diabetes is done at the early stage that is why systematic effort are made in designing the system and five machine learning algorithms are evaluated on various measure. The result computed after evaluating the model on various measures only decision tree provides 1% of inaccurate result (minimum as compared to other approaches that we evaluated). In future these system can be used for prediction of disease not only diabetes but disease as well to spread more awareness among the younger generation.

**4 References**

[1] Sahoo, Abhaya Kumar, Chittaranjan Pradhan, and Himansu Das. "Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making." In Nature inspired computing for data science, pp. 201-212. Springer, Cham, 2020.

[2] Prajapati, Gend Lal, and Rekha Saha. "REEDS: Relevance and enhanced entropy based Dempster Shafer approach for next word prediction using language model." Journal of Computational Science 35 (2019): 1-11.

[3] Ambulgekar, Sourabh, Sanket Malewadikar, Raju Garande, and Bharti Joshi. "Next Words Prediction Using Recurrent NeuralNetworks." In ITM Web of Conferences, vol. 40, p. 03034. EDP Sciences, 2021.

[4] Stremmel, Joel, and Arjun Singh. "Pretraining federated text models for next word prediction." In Future of Information and Communication Conference, pp. 477-488. Springer, Cham, 2021.

[5] Xiaoyun, Qu, Kang Xiaoning, Zhang Chao, Jiang Shuai, and Ma Xiuda. "Short-term prediction of wind power based on deep long short-term memory." In 2016 IEEE PES Asia- Pacific Power and Energy Engineering Conference (APPEEC), pp. 1148-1152. IEEE, 2016.

[6] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349, no. 6245 (2015): 255-260.