

**Classification of Diabetes using ML techniques**

**Worked by:**

**Hiba Jamal Foqha 201811207**

**Shooq Nedal Balawneh 201812634**

**Aseel Ra’fat Qtait 201811932**

**Prof. Mohammad Awad**

**Computer System Department**

**Date 14/6/2022.**

1. **Introduction:**

Diabetes (DM) is quite possibly the most well-known disease where patients’ body capacity to create and react to insulin hormone is impeded, Which may bring about expanded glucose levels in the blood (Lonappan et al.,2007).

Over time, having too much glucose in your blood can cause health problems. Although diabetes has no cure, you can take steps to manage your diabetes and stay healthy.

Diabetes increases the risk of heart disease by about four times in women but only about two times in men. Women are also at higher risk of other diabetes-related complications such as blindness, kidney disease, and depression. (2)

The purpose of this research is the creation of a model to classify and to provide predictive analysis on the diagnosis of diabetes, which allow organizations that provide services to cover health access information on the diagnosis of diabetes mellitus quickly, through using Machine-learning techniques.

Machine learning is a subset of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly program. (3) Machine learning can be supervised, unsupervised, semi-supervised, or reinforcement based.



1. **Related works:**

The purpose of this research is the creation of a model to provide predictive analysis on the diagnosis of diabetes, which allows organizations that provide services to cover health access information on the diagnosis of diabetes, through using machine learning techniques.

In previous years, many researchers have used different machine learning techniques to predict and diagnose diabetes.

M.S. Barale and D.T. Shirke (17) used a 5 attributes dataset in diabetes predictions. Missing Values Imputation and Outlier Detection and Treatment are preprocessing technique he applied to clean dataset. Then, Artificial neural network (ANN), logistic regression (LR), support vector machine (SVM), k-mean and decision tree (DT) algorithms are applied to dataset. K-mean methodology reached the highest accuracy by up to 99.33%. M.S. Haneen Qteat (18) used a 9 attributes dataset in diabetes predictions. Missing Values Imputation and Treatment are preprocessing technique he applied to clean dataset. Then, Artificial neural network (ANN), logistic regression (LR), support vector machine (SVM), k-mean and decision tree (DT) algorithms are applied to dataset. PSO- MLPNNs model reached the highest accuracy by up to 98.73%. The PSO-MLPNNs model has outperformed each of the MLPNNs, SVM, K-NN, DT, DA, NB, and RF algorithms in classifying T1DM and T2DM.

Jozeene Bailey, Qianwen Ariel Xu, Zhili Sun. (19) the paper proposes an e-diagnosis system based on machine learning (ML) algorithms to be implemented on the Internet of Medical Things (IoMT) environment. used dataset with Several risk factors that are attributed to diabetes include ethnicity, family history of diabetes, age, excess weight, unhealthy diet, physical inactivity, and smoking.it used Feature selection. The data subsets were manually split into 538 and 230 samples, respectively (70/30 split). We can conclude that a Na ̈ıve Bayes model works well with a more fine-tuned selection of features with 97.13% on the 3-factor data subset.

1. **Dataset:**

This dataset is originally from the National Institute of Diabetes and Digestive and kidney Diseases.

**Data understanding:**

Using summary(data) command in Matlabfor describing the data.

* 1. **Content**

Several constraints placed on the selection of these instances from a larger database. (4) In particular, all patients here are females at least 21 years old of Pima Indian heritage. Dataset consists of eight Medical Variables (Independent) and one Outcome Variable (Dependent).

* 2.1.1 Attributes :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attributes from PIMA Indian dataset** | **Description of attributes** | **Min Val** | **Max Val** | **Median Val** | **Mod Val** |
| Pregnancies | Number of times a participant is pregnant | 0 | 17 | 3 | 1 |
| Glucose | Plasma glucose concentration a 2 h in an oral glucose tolerance test | 0 | 199 | 117 | 100 |
| Diastolic Blood Pressure | It consists of Diastolic blood pressure (when blood exerts into arteries between heart) (mm Hg) | 0 | 122 | 72 | 70 |
| Skin Thickness | Triceps skinfold thickness (mm). It concluded by the collagen content | 0 | 99 | 23 | 0 |
| Serum Insulin | 2-Hour serum insulin (mu U/ml) | 0 | 846 | 30.5 | 0 |
| BMI | Body mass index (weight in kg/(height in m)^2) | 0 | 67.1 | 32 | 32 |
| Diabetes pedigree Function | An appealing attributed used in diabetes prognosis | 0.078 | 2.42 | 0.3725 | 0.254 |
| Age | Age of participants | 21 | 81 | 29 | 22 |

* 2.1.2 Outcome:

|  |  |  |
| --- | --- | --- |
| outcome | number | Class name |
| Healthy | 500 | 0 |
| Diabetic | 268 | 1 |
| sum | 768 | - |

1. **Data prepossessing:**

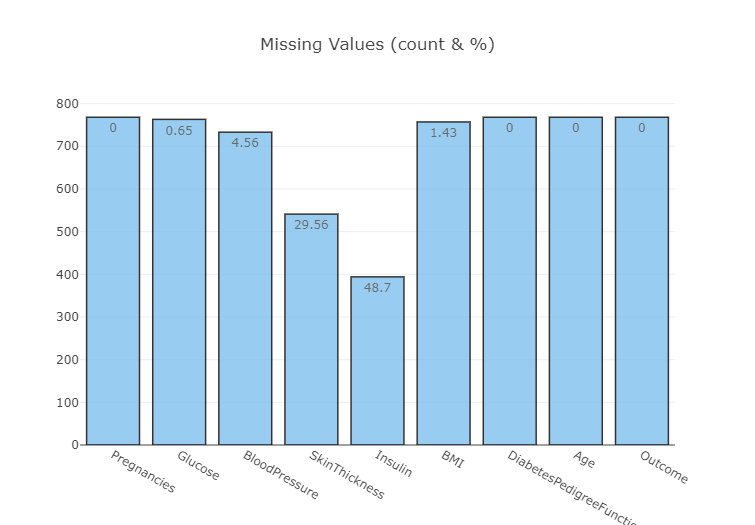
Data preprocessing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. Data preprocessing transforms the data into a format that is more easily in data mining, machine learning and other data science tasks. The techniques generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate results. (5)

* 1. **Data preprocessing Technique**

* + 1. **Data cleaning**

Handling data errors and issues to create complete and accurate data sets. For example \*faulty data is removed or fixed. \*handle missing values. \*inconsistent entries harmonized. \*removing for outlier. (6)

* Handling missing values



It is not medically possible for some data record to have 0 value such as Blood Pressure or Glucose levels. So it consider as missing value

Percentage of Missing values:

* Insulin = 48.7%
* Skin Thickness = 29.56%
* Blood Pressure = 4.56%
* BMI = 1.43%
* Glucose = 0.65%

We replace insulin and skin thickness with the median value of its class.

* **Insulin** - Skin Thickness

|  |  |
| --- | --- |
| Outcome | insulin median |
| 0 | 102.5 |
| 1 | 169.5 |

|  |  |
| --- | --- |
| outcome | Skin Thickness median |
| 0 | 107.0 |
| 1 | 140.0 |

We replace blood pressure and BMI with the mean of five values before and five after each missing value.

We replace glucose missing values by average of all record value.

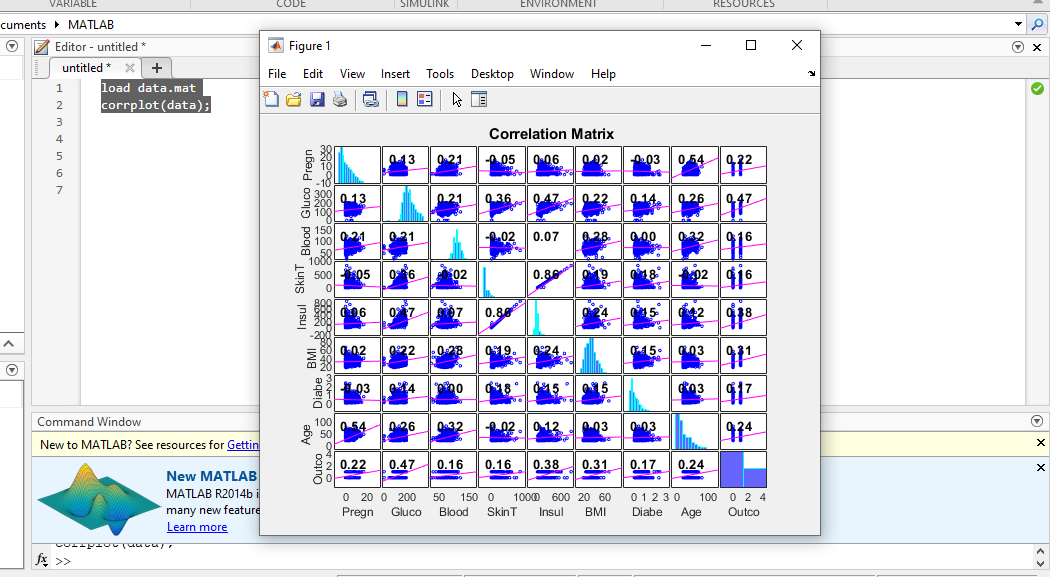
We keep the values of pregnancies, age, diabetes Pedigree Function as they are, “they don’t contain any missing values”.

* + 1. **Feature selection**

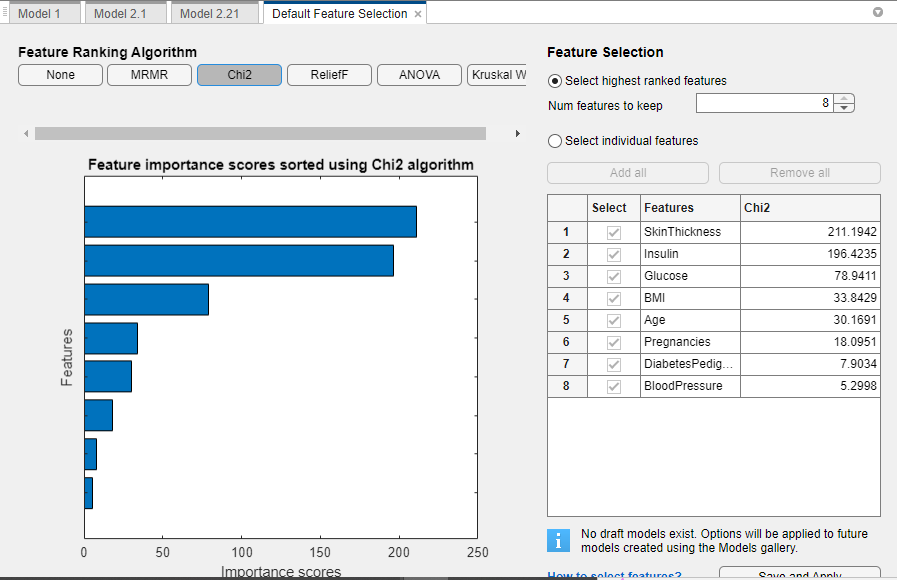
In this process, the reduction of attributes done so that the dimensionality of the data set can be reduced. Combining and merging the attributes of the data without losing its original characteristics. This also helps in the reduction of storage space and computation time reduced. (7)

**Pearson's Correlation Coefficient:** Helps you find out the relationship between two quantities. It gives you the measure of the strength of association between two variables. The value of Pearson's Correlation Coefficient can be between -1 to +1. 1 means that they are highly correlated and 0 means no correlation. (8)

We use correlation matrix to displays the correlation coefficients ofall the possible pairs of attribute in a dataset.



We also use feature ranking (selection) algorithms epically chi2 to do data reduction.



**We remove 3 attributes (blood pressure, Diabetes Pedigree Function and Pregnancies) and still have the same accuracy (=90.5%).**

1. **Machine learning techniques** 
   1. **Decision trees**

Decision are most**popular methods for classification.**

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand that asks a question, and based on the answer (Yes/No), it further split the tree into subtrees. (9)

* 1. **SVMs:**

Stands for Support Vector Machine. Constructing a hyperplane where the distance between two classes of data points is at its maximum, this hyperplane minimizes the error bound. Separating the classes of data points. That finds the perfect boundary between the possible results. (10)

* 1. **KNN:**

K-Nearest Neighbor algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. (11)

* 1. **Ensemble:**

Linear combine the predictions from two or more models. Specially applied for supervised learning tasks including classification and regression, where the outputs of the trained base learner are real-valued probability estimates of class label given the input data. Although there are an almost unlimited number of ways in which this can achieved, such as the strategies they are Bagging and Boosting. (12)

* 1. **Artificial neural network**

Parallel processing abilities, ability to generalize and infer unseen relationships means ANNs can predict the output of unseen data, works with numerical information, which must translated into numerical values before they can presented to the ANN.

The ANN can make decisions based on the observations, has three or more layers that are interconnected. The first layer consists of input neurons. Those neurons send data on to the deeper layers, which in turn will send the final output data to the last output layer.

All the inner layers are hidden. The units in the neural layer try to learn about the information gathered by weighing it according to the ANN’s internal system. These guidelines allow units to generate a transformed result, which is provided as an output to the next layer. (20)

1. **Performance metrics**

To measure the performance of a model, we need several elements:

* **Confusion matrix** : also known as the error matrix, allows visualization of the performance of an algorithm :
  + True positive (TP): Diabetic correctly identified as diabetic.
  + True negative (TN): Healthy correctly identified as healthy.
  + False positive (FP): Healthy incorrectly identified as diabetic.
  + False negative (FN): Diabetic incorrectly identified as healthy.
* **Metrics**:
  + Accuracy : 𝑇𝑃+𝑇𝑁 /𝑇𝑃+𝐹𝑃+𝑇𝑁+𝐹𝑁 × 100 % (13)
  + 𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 = 𝑇𝑃 /𝑇𝑃+𝐹𝑃 × 100 % (14)
  + 𝑆𝑒𝑛𝑠𝑖𝑡𝑖𝑣𝑖𝑡𝑦 = 𝑇𝑃/ 𝑇𝑃+𝐹𝑁 × 100 % (15)
  + 𝑆𝑝𝑒𝑐𝑖𝑓𝑖𝑐𝑖𝑡𝑦 = 𝑇𝑁/ 𝑇𝑁+𝐹𝑃 × 100 % (16)
* **Roc Curve**: The ROC curve created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | Performance Matrices | | | |
|  | Accuracy (%) | Sensitivity (%) | Specificity (%) | Roc (%) |
| **ANN** | 87.9% | 90.4% | 82.44% | 91% |
| **Decision trees** | 90.7% | 92.8% | 86.12% | 92% |
| **SVM** | 88.5% | 92.17% | 81.28% | 91% |
| **Ensemble** | 91% | 92.07% | 88.44% | 96% |
| **KNN** | 89.1% | 91.48% | 84.06% | 93% |

The most powerful machine learning algorithms has been applied to the Pima Indians Diabetes Dataset. It can be noticed from table that the Ensemble model was accurate in classifying the diabetes by up to 91%.

1. **Conclusion:**

The accuracy achieved by functional classifiers Artificial Neural Network (ANN), Decision Tree (DT), Ensemble, KNN, and SVMs within the range of 88–91%. Among the fifth of them, Ensamble provides the best results for diabetes onset with an accuracy rate of 91% on the PIMA dataset. Hence, this proposed system provides an effective prognostic tool for healthcare officials. The results obtained can be used to Classify of Diabetes that can be helpful in early detection of the disease.

**References**

2-[https://www.cdc.gov/diabetes/library/features/diabetes-and-women.html](#_top)

3-<https://www.ibm.com/cloud/learn/machine-learning>

4-[Kaggle: Your Home for Data Science](https://www.kaggle.com/)

5-[https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation?\_gl=1\*vs6rjd\*\_ga\*MjEyMjYzMzE1NS4xNjU0MjU0OTgz\*\_ga\_TQKE4GS5P9\*MTY1NDI1NDk4Mi4xLjEuMTY1NDI1NTAzMy4w&\_ga=2.215485143.1900114441.1654254983-2122633155.1654254983](https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation?_gl=1*vs6rjd*_ga*MjEyMjYzMzE1NS4xNjU0MjU0OTgz*_ga_TQKE4GS5P9*MTY1NDI1NDk4Mi4xLjEuMTY1NDI1NTAzMy4w&_ga=2.215485143.1900114441.1654254983-2122633155.1654254983)

6-[https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation?\_gl=1\*vs6rjd\*\_ga\*MjEyMjYzMzE1NS4xNjU0MjU0OTgz\*\_ga\_TQKE4GS5P9\*MTY1NDI1NDk4Mi4xLjEuMTY1NDI1NTAzMy4w&\_ga=2.215485143.1900114441.1654254983-2122633155.1654254983](https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation?_gl=1*vs6rjd*_ga*MjEyMjYzMzE1NS4xNjU0MjU0OTgz*_ga_TQKE4GS5P9*MTY1NDI1NDk4Mi4xLjEuMTY1NDI1NTAzMy4w&_ga=2.215485143.1900114441.1654254983-2122633155.1654254983)

7 <https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/>

8- [Kaggle: Your Home for Data Science](https://www.kaggle.com/)

9-[Machine Learning Decision Tree Classification Algorithm - Javatpoint](https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm)

10- [Diabetes Prediction | Kaggle](https://www.kaggle.com/code/mvanshika/diabetes-prediction)

11- [K-Nearest Neighbor (KNN) Algorithm for Machine Learning - Javatpoint](https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning)

12- [Ensemble Models. A guide to learning ensemble techniques… | by Mohammed Alhamid | Towards Data Science](https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c)

13- M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, “Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project”, Plos One, Vol. 12, No. 7, 2017.

14- D. K. Choubey, S. Paul, and S. Kumar, “Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection”, In: Proc. of International Conf. on Communication and Computing System, London, pp. 451–455, 2017.

15-Kolukisa, H. Hacilar, G. Goy, M. Kus, B. BakirGungor, A. Aral, and V. C. Gungor, “Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease”, In: Proc. of International Conf. on Big Data (Big Data), Seattle, Westin Seattle, USA, pp. 2232-2238, 2018.

16-R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, “Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset”, In: Proc. of International Conf. on Computing Networking and Informatics, Nigiria, Lagos, pp. 1-5, 2017.

17- <https://www.researchgate.net/profile/Mahesh-Barale/publication/301335647_Cascaded_Modeling_for_PIMA_Indian_Diabetes_Data/links/59c0ba97a6fdcca8e5724d61/Cascaded-Modeling-for-PIMA-Indian-Diabetes-Data.pdf>

18-<https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm>

19-<https://link.springer.com/content/pdf/10.1007/s00521-022-07049-z.pdf>

20-<https://www.techopedia.com/definition/5967/artificial-neural-network-ann#:~:text=Artificial%20neural%20networks%20%28ANNs%29%20use%20learning%20algorithms%20that,very%20effective%20tool%20for%20non-linear%20statistical%20data%20modeling>.