**DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES**

**A report on**

**PROJECT – PHASE I**

#### Submitted by

**Abhay Kumar (2019105194)**

**Sungjemkaba** (**2019105221**)

***in partial fulfillment of the requirements for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

****

**DEPARTMENT OF** **COMPUTER SCIENCE AND ENGINEERING**

**NATIONAL INSTITUTE OF TECHNOLOGY NAGALAND**

**DIMAPUR 797 103**

**DEC 2022**

**DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES**

**A report on**

**PROJECT – PHASE I**

***Submitted by***

**Abhay kumar (2019105194)**

**Sungjemkaba** (**2019105221**)

***in partial fulfillment of the requirements for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

Under the supervision of

**Dr. Arambam Neelima**

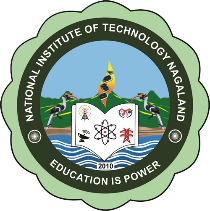
****

**DEPARTMENT OF** **COMPUTER SCIENCE AND ENGINEERING**

**NATIONAL INSTITUTE OF TECHNOLOGY NAGALAND**

**DIMAPUR 797 103**

**DEC 2022**

**राष्ट्रीय प्रौद्योगिकी संस्थान नागालैंड**

**NATIONAL INSTITUTE OF TECHNOLOGY NAGALAND**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**Chumukedima, Dimapur – 797 103, Nagaland**

**BONAFIDE CERTIFICATE**

Certified that this Project titled “**DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES**” is the bonafide work of ABHAY KUMAR (2019105194) AND SUNGJEMKABA (2019105221) who carried out the work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other student.

**Dr. Arambam Neelima**

**Project Coordinator**

Asst. Professor

Department of Computer Science and Engineering

NIT Nagaland

Dimapur - 797103

**Dr. Arambam Neelima**

**Project Supervisor**

Asst. Professor

Department of Computer Science and Engineering

NIT Nagaland

Dimapur - 797103

**Dr. Lithungo K Murry**

**Head of Department**

Asst. Professor

Department of Computer Science and Engineering

NIT Nagaland

Dimapur - 797103

**External Examiner**

Professor Department of CST

IIEST Shibpur Howrah, India

Howrah - 711103

**ABSTRACT**

The project work reported is focused on the development…..

Line Spacing is 1.5

**ACKNOWLEDGEMENT**

We wish to place on record our deep sense of gratitude to   
our honorific Guide **Dr. Arambam Neelima** , Assistant professor, National Institute of Technology Nagaland for her supervision, valuable guidance and moral support leading to the successful completion of the work. Without her continuous encouragement and involvement, this project would not have been a reality.

We would like to extent my sincere thanks to **Dr. Lithungo Murry ,** Assistant professor, National Institue of Technology Nagaland, for their valuable suggestions. We wish to thank **Dr. Lithungo Murry**, Department of computer science and engineering, NIT Nagaland for continuous support. We would also like to thank all our friends who have developed us to gain a sense of dutifulness, perfection and sincerity in the effort (if Desired).

We wish to dedicate this work to **Mr. PRITHVI RAJ**, for they are the pillars of support giving us confidence in whatever we do. We would like to thank **Ankit kumar** who has motivated us to work harder and do our best. Last but not least, We would like to owe our sincere and incessant gratitude to the almighty God for the immense blessing on us.

**ABHAY KUMAR**

**2019105194**

**&**

**SUNGJEMKABA**

**2019105221**

**TABLE OF CONTENTS**

**CHAPTER NO. TITLE PAGE NO.**

**ABSTRACT**  iii

**LIST OF TABLES** xi

**LIST OF FIGURES** xii

**LIST OF ABBREVIATIONS** xv

**1** **INTRODUCTION**

1.1 GENERAL 13

1.2 STATE OF THE ART 14

1.3 OBJECTIVES 16

1.4 ORGANIZATION OF THE REPORT 16

**2** **Data Pre-Processing**

2.1 INTRODUCTION 15

2.2 NEED OF DATA PRE-PROCESSING 17

2.2.1 RESCALE DATA 18

2.2.2 BINARIZE DATA(MAKE BINARY) 18

2.2.3 STANDARDIZE DATA 18

2.3 PROCESSING TECHNIQUES USED IN MODEL 19

2.3.1 REMOVING NULL VALUE 19

2.3.2 NORMALISING DATA 19

2.4 SPLITTING DATASET 19

2.4.1 HOE TO DIVIDE DATA 20

**CHAPTER NO. TITLE PAGE NO.**

**3** **PROPSED METHOD**

3.1 INTRODUCTION 22

3.2 PROSED MODEL 23

3.3 PROPOSED ALGORITHM 23  
 3.3.1 LOGISTIC REGRESSION23 23

3.3.2 DECISION TREE 24

3.3.3 SUPPORT VECTOR MACHINE 26

**4** **CHAPTER TITLE**

4.1 INTRODUCTION 27

4.2 DATA SET 27

4.2.1 DATA SET DESCRIPTION 27

4.2.1 GRAPHICAL REPRESENTATION OF DATASET 28

4.2.3 REPRESENTATION IN THE FORM OF DATA FRAME 28

4.3 FILLING IN MISSIING VALUE 28

4.3.1 ATTRIBUTE CONSISTING OF VALUE AS 0 OR NULL VALUE 29

4.4 CORRELATION 29

4.4.1 CORRELATION COEFFICIENT 30

4.5 RESULT FROM DIFFERENT ALGORITHM 31

4.5.1 SUPPORT VECTOR MACHINE 32

4.5.2 DECISION TREE 32

4.5.3 LOGISTIC REGRESSION 33

4.6 COMPARING DIFFERENT ALGORITHM 33

4.7 COMPARING THE F SCORE OF DIFFERENT MODEL34

4.8 CONFUSION MATRIX 35

**CHAPTER NO. TITLE PAGE NO.**

**5 CONCLUSION**

5.1 INTRODUCTION 36

5.2 HIGHLIGHTS OF THE WORK DONE 36

5.3 FUTURE WORK 37

**REFERENCES** 38

**LIST OF TABLES**

**TABLE NO. TITLE PAGE NO.**

4.1 REPRESENTING THE DATASET ATTRIBUTE 27

4.2 DATA FRAME OF DIABETES DATASET 28

4.3 CORRELATION TABLE 30

**LIST OF FIGURES**

**FIGURE NO. TITLE PAGE NO.**

2.1 DATA PREPROCESSING STEPS 17

3.1 PROPOSED MODEL 23

3.2 GRAPH REPRESENTING THE SIGMOID FUNCTION 24

3.3 REPRESENTATION OF DECISION TREE 25

3.4 DIAGRAM REPRESENTING THE SVM CLASSIFIER 26

4.1 REPRESENTING THE DIABETES DATASET 28

4.2 NULL VALUE ATTRIBUTE 29

4.3 CORRELATION DATASET HEATMAP 31

4.4 SUPPORT VECTOR ACCURACY 32

4.5 DECISION TREE ACCURACY 32

4.6 LOGISTIC REGRESSION ACCURACY 33

4.7 ALL MODEL ACCURACY COMPARISION 34

4.8 CONFUSION MATRIX 35

5.1 DESCRIPTION OF THE ACCURACY 37

**LIST OF ABBREVIATIONS**

|  |  |  |
| --- | --- | --- |
| SVM | - | Support vector machine |
| RF | - | Random forest |
| KNN | - | K nearest neighbour |
| i.e | - | That is |
| LR | - | Logistic Regression |
| DF | - | Data frame |
| nan | - | Null |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  | - |  |
|  |  |  |

**CHAPTER -1**

**INTRODUCTION**

**1.1 GENERAL**

In underdeveloped nations, diabetes is one of the main causes of death. The government and private citizens are funding research projects in an effort to find a cure for the serious illness. Due to a shortage of insulin, which impacts blood sugar metabolism, diabetes is a condition in which blood sugar levels keep rising. Patients with diabetes are unable to efficiently convert the carbs they consume into the glucose sugar that provides energy for daily activities.

The most prevalent type of diabetes is type 1, which develops when the body does not create enough insulin. Low insulin production is a common occurrence in the diabetic population due to immune system attack and loss of pancreatic function. According to the research, both children and adults can get this type of diabetes. To ensure their continuing existence on this world, they must ingest enough insulin. The most important risk factors for type 1 diabetes are pancreatic disease, pancreatic infection, and family history. The next stage of diabetes, type 2, develops when the body's insulin is improperly utilised.

Different ML techniques can be used with various data structures. This study looks at predictive modelling in the healthcare industry. Healthcare data sets are analysed using ML algorithms. The focus of this experiment is gestational diabetes in the research. On the Pima Indian Diabetes Database (PIDD) data set, KNN, SVM, logistic regression, and random forest ML approaches are used to examine the prediction of diabetes. In order to obtain precision, this test is performed using a number of factors, including blood pressure, BMI, and glucose .

The most recent advancement in ML has improved the computer system's ability to recognise and categorise images, forecast diseases, and enhance decision-making through data analysis. The goal of machine learning applications is to teach the computer to perform tasks more effectively than a human. The model is trained using a supervised learning technique, and it is then evaluated using test data.

**1.2 STATE OF THE ART**

In the world of medicine, ML algorithms are well-known for their ability to predict disease. In an effort to produce the best and most accurate findings possible, many researchers have used ML approaches to predict diabetes .

Multiple classifiers, including SVM, K-Nearest Neighbors (KNN), and Random Forest, were utilised by Shubham Joshi , Ali Rizwan & Basant Tiwari [1]. A dataset from the UCI repository was used for the categorization . Based on the accuracy, sensitivity, and specificity values, the classifiers' outputs were compared. Using 5-fold cross validation, the classification was carried out in two scenarios: with and without pre processing of the dataset. The pre-processing procedure that was used on the dataset was not described by the authors; they merely said that the noise had been eliminated. According to their findings, KNN (k = 1) and Random Forest classifiers had the best accuracy rates of 100%, while the decision tree classifier had the highest accuracy rate of 73.82% without pre-processing.

Furthermore, Neha prerna tigga & shruti garg [2] detected diabetes using Decision Tree, SVM, and Naive Bayes classifiers. Finding the classifier with the maximum accuracy was the goal. For this investigation, the Pima Indian dataset was utilised. Cross-validation on 10 folds is used to partition the dataset. The preprocessing of the data wasn't discussed by the authors. Accuracy, precision, recall, and the F-measure were used as metrics to assess performance. The Naive Bayes model had the highest level of accuracy (76.30%).

V. jackins & S. Vimal , M Kallipan [3] sought to apply the SVM to predict diabetes in addition to the other research. The Diabetes 130-US and Pima Indians datasets were integrated to create a single dataset. Given that many other studies only used one dataset, the goal of this study was to confirm the accuracy of the findings. 64,419 of the 102,538 samples in the dataset were positive, while 38,115 were negative. The dataset consists of 49 attributes and 102,538 samples. The characteristics employed in this study were not discussed by the authors. The dataset is pre-processed by converting the non-numerical values to numerical values, replacing missing values and out-of-range data with zero, and normalising the data between 0 and 1. Before the SVM model was applied, various feature selection techniques were employed. While the Wrapper and Ranker methods (from the Weka Tool) picked nine and twenty characteristics, respectively, the Fselect script from the LIBSVM package only selected four attributes. The authors employed a 10-fold cross validation technique for the validation process. The diabetes prediction may be more accurate, at 72%, when using a combined dataset.

Six distinct classifiers were utilised by Ionnis kavakiotis, Olga Tsave [4]. Multilayer Perceptron, RandomForest are the classifiers. For this study, the Pima Indian dataset was also used. Although the authors didn't mention a pretreatment step, they used the GreedyStepwise and BestFirst algorithms to identify the discriminating characteristics that improve classification performance. Body mass index, plasma glucose concentration, diabetic pedigree function, and age are the four criteria that have been chosen. The dataset is validated via a 10 fold cross. Based on the importance of the precision, recall, and F-Measure, the classifiers were compared. The Hoeffding Tree algorithm produced a precision value of 0.757, recall value of 0.762, and F-measure value of 0.759.

**1.3 OBJECTIVES**

Objective are as follows:

* Removing the noise.
* Balancing the dataset.
* Finding running time of different algorithm.
* Predicting the diabetes.

**1.4 ORGANIZATION OF THE REPORT**

The organization of the thesis is as follows. Chapter 1 gives an insight of our objective and Literature survey. In main content, Chapter 2 deals with the data cleaning and normalisation part. Chapter 3 is the description of the models used in the diabetes prediction like svm, decision tree and logistic regression . Chapter 4 consists of the work done in creating the diabetes prediction model, its results and outputs. Chapter 5 sums it all giving an overview of what is done, what is to be done and how the accuracy can be improved by considering time bounded limitation.

**CHAPTER 2**

**DATA PRE-PROCESSING**

**2.1 INTRODUCTION**

The adjustments made to our data prior to feeding it to the algorithm are referred to as pre-processing. Data pre-processing is a method for transforming unclean data into clean data sets. In other words, anytime data is acquired from various sources, it is done so in a raw manner that makes analysis impossible.

**FIG (2.1) DATA PREPROCESSING**

**2.2 NEED OF DATA PRE-PROCESSING**

* The format of the data in machine learning projects must be correct in order to get better results from the applied model. For example, the Random Forest algorithm does not tolerate null values, hence null values must be handled from the original raw data set in order to execute the Random Forest algorithm. Some specific Machine Learning models require data in a specific format.
* Another consideration is that the data set should be organised so that many Machine Learning and Deep Learning algorithms can run simultaneously and the best one is selected.

**2.2.1 RESCALE DATA**

* Many machine learning methods can benefit from rescaling the attributes to have the same scale when our data contains attributes with different scales.
* This is helpful for optimization techniques like gradient descent, which are at the heart of machine learning methods.
* It is especially helpful for algorithms like regression and neural networks that weight inputs, as well as for algorithms like K-Nearest Neighbors that require distance measurements.
* Using the MinMaxScaler class from scikit-learn, we can rescale your data.

**2.2.2 BINARIZE DATA (MAKE BINARY)**

* Our data can be transformed using a binary threshold. The threshold is recorded as 1, and all values that are equal to or below it are marked as 0.
* This is referred to as binarizing or thresholding your data. When you wish to give clear values to probabilities, it can be handy. It is also helpful when adding new features that convey important information during feature engineering.
* Using scikit-Binarizer learn's class and new binary attributes, we can build new binary attributes in Python.

**2.2.3 STANDARDIZE DATA**

* In order to convert attributes with a Gaussian distribution and varying means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1, standardisation is a useful technique.
* Scikit-StandardScaler Learn's class allows us to standardise data.

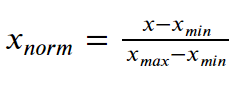
**2.3 PROCESSING TECHNIQUES USED IN MODEL**

**2.3.1 REMOVING NULL VALUE**

* By finding the null value in the data frame and replacing those value with the mean value of that particular attribute by using the function SimpleImputer from the scikit learn module of tensorflow.
* The scikit-learn class SimpleImputer is useful for handling the missing data in the dataset for the prediction model. It inserts a designated placeholder in lieu of the NaN values.
* The SimpleImputer() method, which accepts the following inputs, is used to implement it.

**2.3.2 NORMALISING DATA**

* When the various characteristics (variables) are on a smaller scale, machine learning algorithms typically perform better or converge more quickly. Therefore, normalising the data before training machine learning models on it is standard practise.
* Additionally, normalisation reduces the sensitivity of the training process to the magnitude of the characteristics. As a result, following training, the coefficients improve.
* Feature scaling is the process of rescaling features to make them more training-friendly.
* The Normalization formula is provided below:



**Equation (2.1) normalising the dataset**

<https://journaldev.nyc3.digitaloceanspaces.com/2020/10/Normalization.png>

**2.4 SPLITTING DATASET**

The core of every ML problem is data. ML models are like bodies without a soul if they are not fed with the right data. But gathering data is no longer a significant issue in the 'big data' world of today. Every day, we consciously or unconsciously produce enormous datasets. Having an abundance of data available, however, does not make the issue go away. We need to feed massive amounts of high-quality data into ML models in order for them to produce accurate results.

The quality data is useless unless it is used effectively, even though extracting meaning from raw data is an art in and of itself and necessitates solid feature engineering skills and domain expertise (in special circumstances). How to divide the data for training and testing is the main issue that ML/DL practitioners encounter. Even though it initially appears to be a straightforward issue, only by delving deeply into it can its complexity be determined. Inaccurate training and testing sets may have unanticipated consequences on the model's output. It could result in the data being overfitted or underfitted, and our model could end up producing biassed results.

**2.4.1 HOW TO DIVIDE DATA**

Ideally, the data should be split into three sets: a train set, a test set, and a holdout cross-validation or development (dev) set. Let's first briefly discuss the meaning of these sets and the kinds of data they ought to contain.

* Train Set: The data that would be fed into the model would be in the train set. Simply put, our model would gain knowledge from this data. For instance, a regression model could uncover gradients to lower the cost function by using the instances in this data. These gradients will then be applied to lower costs and improve data prediction.
* Dev Set: The trained model is validated using the development set. As the foundation for our model evaluation, this setting is the most crucial. If there is a significant disparity between the error on the training set and the error on the development set, the model is over-fitting and has a high variance.
* Exam Set: The data used to test the trained and approved model are in the test set. It reveals the effectiveness of our entire model and the likelihood that it will forecast an illogical event. Numerous assessment criteria (such as precision, recall, accuracy, etc.) can be utilised to assess the effectiveness of our approach.

**CHAPTER 3**

**PROPOSED METHOD**

**3.1** **INTRODUCTION**

For data analysis, the suggested method has been utilised with the Anaconda tool (AEN 4.1 Version). The package versions for predictive analysis and data management are managed via the anaconda package management system . those with diabetes disease—have been used as inputs. These datasets were chosen because they share characteristics with each other in real life and because there is some correlation between them. These data are loaded and examined to determine whether there are any missing values. Any missing values are replaced with null values if they are discovered. After that, it was determined if any of the data's columns were related to any other column in the data on their own. One of the columns is eliminated if any correlation between the two is discovered. If data has either a true or false value, they are replaced with 1 and 0, accordingly. The original data has been divided into training data, which contains 70% original data, and test data, which contains 30% original data.

To determine how many instances in the original, practise, and test data are true and false The decision tree method has been trained on our different class datasets, and it has been used to calculate the accuracy of the classes individually using confusion matrix . [4.8] Fig. displays the proposed method's block diagram. 1. The accuracy calculation performance metrics are displayed separately for each kind of data in the performance report. In a similar manner, it trained the random forest algorithm with our three separate class data. The algorithm uses a confusion matrix to calculate the results accuracy and presents the three classifications separately. The accuracy calculation performance metrics are displayed separately for each kind of data in the performance report. Every epoch has one or more batches, which change internal model parameters. The epoch can be used to reduce dataset error rates.

For each class, a few example test results have been collected independently. We can determine whether or not the sample data are associated with the disease by applying them to each trained model of it. One can see that the model trained with Random forest provides the most accurate classification results when comparing the results of both models for each type of data. The trained data are compared individually with the suggested algorithms in order to determine the efficacy of the procedure, and the performance of the test data is also examined. The suggested approach can be used to test real-time disease data for classification and to determine whether a patient has a specific disease or not.

**3.2 PROPOSED MODEL**

Dataset

Pre-processing

1.Min-Max scaler

2. Replacing missing value with mean.

Classification algorithm such as SVM, Decision tree and Logistic Regression have been used

Result

**Fig(3.1) Propsed model**

**3.3 PROPOSED ALGORITHM**

**3.3.1 LOGISTIC REGRESSION**

Essentially, supervised categorization is what logistic regression does. For a specific collection of features (or inputs), X, the target variable (or output), y, can only take discrete values in a classification issue.

Contrary to popular assumption, a regression model is a logistic regression. In order to determine the likelihood that a specific data entry falls under the category designated by the number "1," the programme creates a regression model. Logistic regression models the data using the sigmoid function, just like linear regression assumes that the data follows a linear distribution.



**Fig(3.2) graph representing the sigmoid function**

https://media.geeksforgeeks.org/wp-content/uploads/20190522162153/sigmoid-function-300x138.png

**3.3.2 DECISION TREE**

The most effective and well-liked technique for categorization and prediction is the decision tree. A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label.

Glucose

low

High

Normal

Age

Yes

BMI

Old

Young

Normal

High

Yes

No

Yes

No

**Fig(3.3) representation of decision tree**

Decision Tree Representation: Decision trees classify instances by ordering them from the root of the tree to a leaf node, which then indicates the instance's categorization. As seen in the above diagram, to classify an instance, one tests the attribute given by the root node of the tree before continuing down the branch of the tree that corresponds to the attribute's value. The subtree rooted at the new node is then subjected to the same procedure once more.

According to whether a specific morning is ideal for diabetes prediction, the decision tree in the above diagram assigns a classification to each leaf and returns that classification (in this case Yes or No).

**3.3.3 SUPPORT VECTOR MACHINE**

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues.

The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary.

SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method. Consider the diagram below, where a decision boundary or hyperplane is used to categorise two distinct categories:

|  |
| --- |
| Support Vector Machine Algorithm |

https://static.javatpoint.com/tutorial/machine-learning/images/support-vector-machine-algorithm.png **fig (3.4) diagram representing the svm classifier.**

**CHAPTER 4**

**RESULT AND DISCUSSION**

**4.1 INTRODUCTION**

In this chapter we will discuss the working procedure of the model and at last we will see the accuracy gain from the use of different algorithm .

**4.2 DATA SET**

**4.2.1 DATA SET DESCRIPTION**

The Pima Indian Diabetes Dataset, a UCI repository, is where the information came from. The collection contains information about 768 patients.

|  |  |
| --- | --- |
| S.NO | ATTRIBUTE |
| 1 | **PREGNANCY** |
| 2 | **GLUCOSE** |
| 3 | **BLOOD PRESSURE** |
| 4 | **SKIN THICKNESS** |
| 5 | **INSULIN** |
| 6 | **BMI** |
| 7 | **DIABETES PEDIGREE FUNCTION** |

**(4.1) representing the dataset attribute**

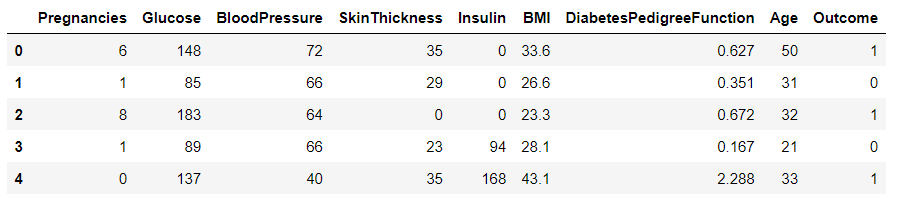
**4.2.2 GRAPHICAL REPRESENTATION OF DATASET**

Here the graphical representation of the dataset has been shown this consist of 500 negative dataset and 268 value of the positive value sample.

**(4.1) representing the diabetes dataset**

**4.2.3 REPRESENTATION IN THE FORM OF DATA FRAME**

Here the dataframe representation of the above dataset is being displayed using the pandas function.



**(4.2) data frame of diabetes dataset**

**4.3 FILLING IN THE MISSING VALUE**

Any data that is missing a value indicates that the event did not occur, the data were unavailable, or the data were not applicable. Here, null values have been used to replace any missing or unavailable information.

**4.3.1 ATTRIBUTE CONSISTING OF VALUE AS 0 OR NULL VALUE**

Below diagram shows all the attribute rows which consist of the value ‘0’ as well as value null .

**(4.2) null value attribute**

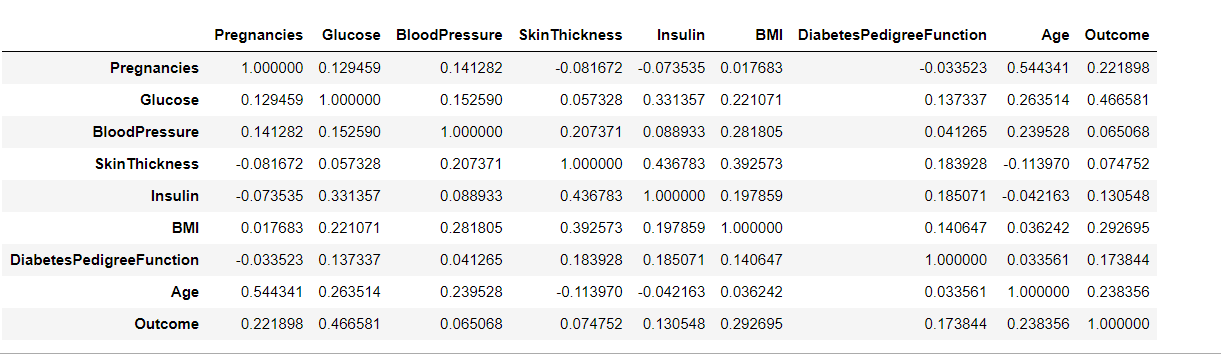
The above graph represents the attribute value which are either 0 or null so as shown in the graph we can see that the total no of rows in the dataset are 768 and that of column are 8 out of that insulin having the 374 column which is either 0 or null value ,pregnancies having 35 rows which consist of 0 or null value glucose attribute consist of 5 ‘0’ or null value and the blood pressure column consist of the 35 rows ,skinthickness consist of the 227 rows having the value as 0 or null , BMI (body mass index) this column consist of the 11 0 or null value ,Diabetes pedigree function and age having the 0 rows which consist of the ‘0’ or null value.

**4.4 CORRELATION**

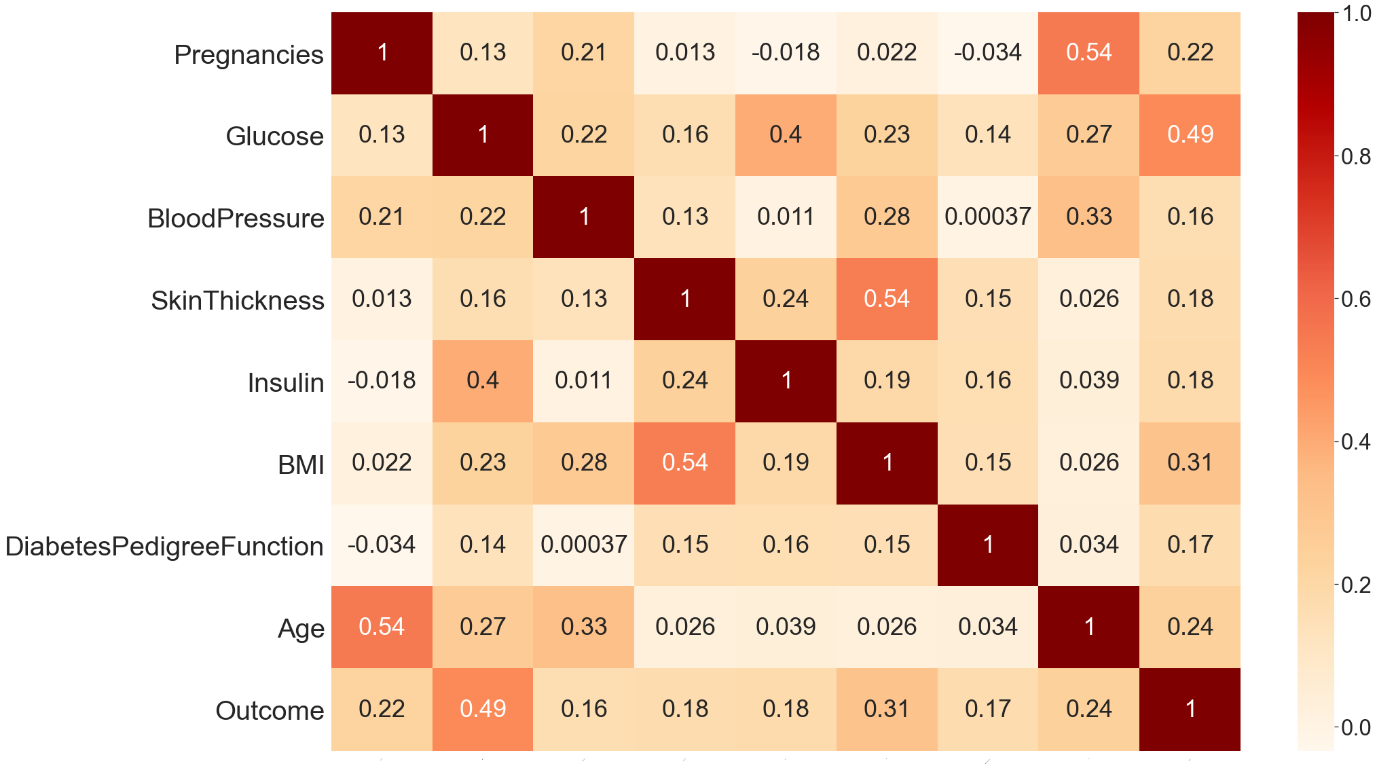
A statistical measure called correlation shows how much two or more variables fluctuate in connection to one another. When two variables rise or decrease simultaneously, there is a positive correlation; when there is a negative correlation, one variable increases as the other falls.

**4.4.1 CORRELATION COEFFICIENT**

In statistics, correlation coefficients are used to assess how closely two variables are related. It is a metric used in statistics to express the linear relationship between dependent and independent variables. It is symbolised by the letter "r" in lower case. Here, the correlation between each dataset's columns is determined to determine how closely related they are (Table 4.3). The outcomes provide the value of each column's correlation with every other column in the dataset. One of the columns in a dataset is eliminated to prevent value duplication if two columns in the dataset have identical associated values.



**(4.3) corelation diagram**



**(4.3) coorelation dataset heatmap**

As stated in the above discussion the value of corrlation function is as close to 1 two attribute are highly corelated so from the diagram ploted above we can say that the corelation between BloodPressure and output is lowest i.e(0.16) and that of blood pressure and glucose is maximum i.e (0.49) except that of outcome and outcome i.e (1) because both the outcome and outcome value are same so they are highly corelated so from here we can say that the outcome is more dependent on the value of glucose concentration rather than that of Blood Pressure.

**4.5 RESULT FROM THE DIFFEERENT ALGORITHM**

Now from here onwards we will see the accuracy achieved by different model and their f1 score as well as the precision and the recall of every model used in this classification.

**4.5.1 SUPPORT VECTOR MACHINE**

Using the support vector machine the accuracy gain on the training and testing data are as follows:

**(4.4) support vector accuracy**

**4.5.2 DECISION TREE**

Using the decision tree classifier accuracy gain on the training and testing dataset is:

**(4.5) decision tree accuracy**

**4.5.3 LOGISTIC REGRESSION**

Using the Logistic Regression classifier the accuracy of the model is being represented with the help of the bar chart shown below:

**(4.6) logistic regression accuracy**

**4.6 COMPARING DIFFERENT ALGORITHM**

From the above three graph , on analysis its being find that the decision tree works well on the PIMA diabetes dataset and gives an accuracy of 85% on testing the data as well as 100% on training data. Where as support vector machine gives an accuracy of 76% on training data where as 74% on testing data and Logistic regression model gives an accuracy of 74% on training data and 76% on testing data.

**(4.7) all model accuracy comaprision**

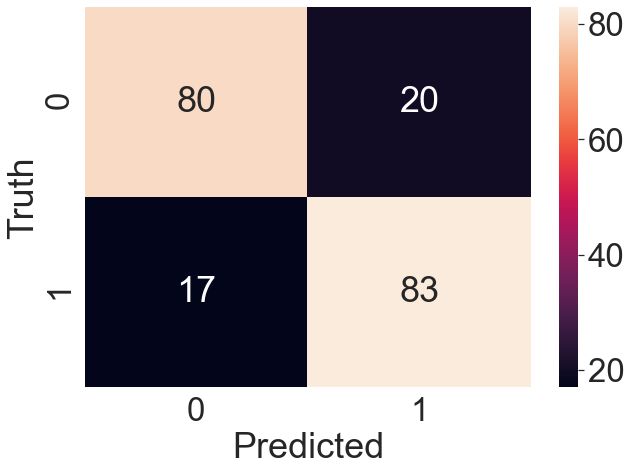
**4.7 COMPARING THE F SCORE OF THE DIFFERENT MODEL**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **F1 score (class 0/class 1)** | **Precision (class 0 / class 1)** | **Recall (class 0 /class1)** |
| **SVM** | **77/73** | **72/79** | **82/68** |
| **Decision Tree** | **85/85** | **86/84** | **83/87** |
| **Logistic Regression** | **77/75** | **74/78** | **80/72** |

**(4.4) precission, recall , f-score**

From the table above its clear that the decision tree has given the best f1 score as well as the best accuracy but decision tree takes more time to run a model.

**4.8 CONFUSION MATRIX**



**(4.8) confusion matrix**

**CHAPTER 5**

**CONCLUSION**

**5.1 INTRODUCTION**

Machine learning method is thought to be helpful in identifying the diabetes disease. The benefits of early diagnosis for patients include a medical checkup. This report used a small number of ML classification models for predicting people with have diabetes on the basis of the accuracy, been discussed. a manifestation of The classification issue's accuracy has been located.

**5.2 HIGHLIGHTS OF THE WORK DONE**

The aim of this study is to discover a model for the diagnosis of diabetes on the basis of PIMA diabetes dataset. The dataset is chosen from the online repositories . The techniques of pre-processing the data such as removing the null value as the missing value with the mean of that particular attribute and using normalization techniques to normalize the data in some particular range between (0-1) which helps the model to train in very efficient ways.

Upon doing the data preprocessing , three different model has been used which are as follows:

1. SVM
2. Decision Tree
3. Logistic Regression.

The accuracy achieved from the above model has been represented in the form of the graph;

**(5.1) description of accuracy**

Decision tree performs well out of these three model in terms of accuracy.

**5.3 FUTURE WORK**

Although the proposed method in this report have addressed the critical issues by predicting the diabetes. There is much more to do with this model like increasing their efficiency as well as doing comparative analysis with the existing model. Use the concept of neural network to increase the efficiency as well as the running time of the algorithm.

**REFERENCES**

[1] Shubham Joshi, Ali Rizwan , Basant Tiwari; A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques; Journal of Healthcare engineering; vol. 2022, page (1-10), 2022.

[2] Neha prerna tigga , shruti garg; Prediction of Type 2 Diabetes using Machine Learning Classification Methods; Jounal of procedia computer science; vol. 167 ,page (706-716),2020.

[3] V. Jackins, S. Vimal, M. Kaliappan ,M.Y. Lee ;AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes; Journal of Supercomputing; vol. 77,page (5198–5219), 2021.

[4] Ionnis kavakiotis, Olga Tsave,Athanasios Salifoglou; Machine Learning and Data Mining Methods in Diabetes Research; computational and structural biotechnology journal;vol. 15 , page (104-116),2020.

[5] Aishwarya Majumdar,Dr. V. vaidehi; Diabetes prediction using machine learning algorithm; International conference on recent trends in advanced computing; vol.165,page (192-199),2019.