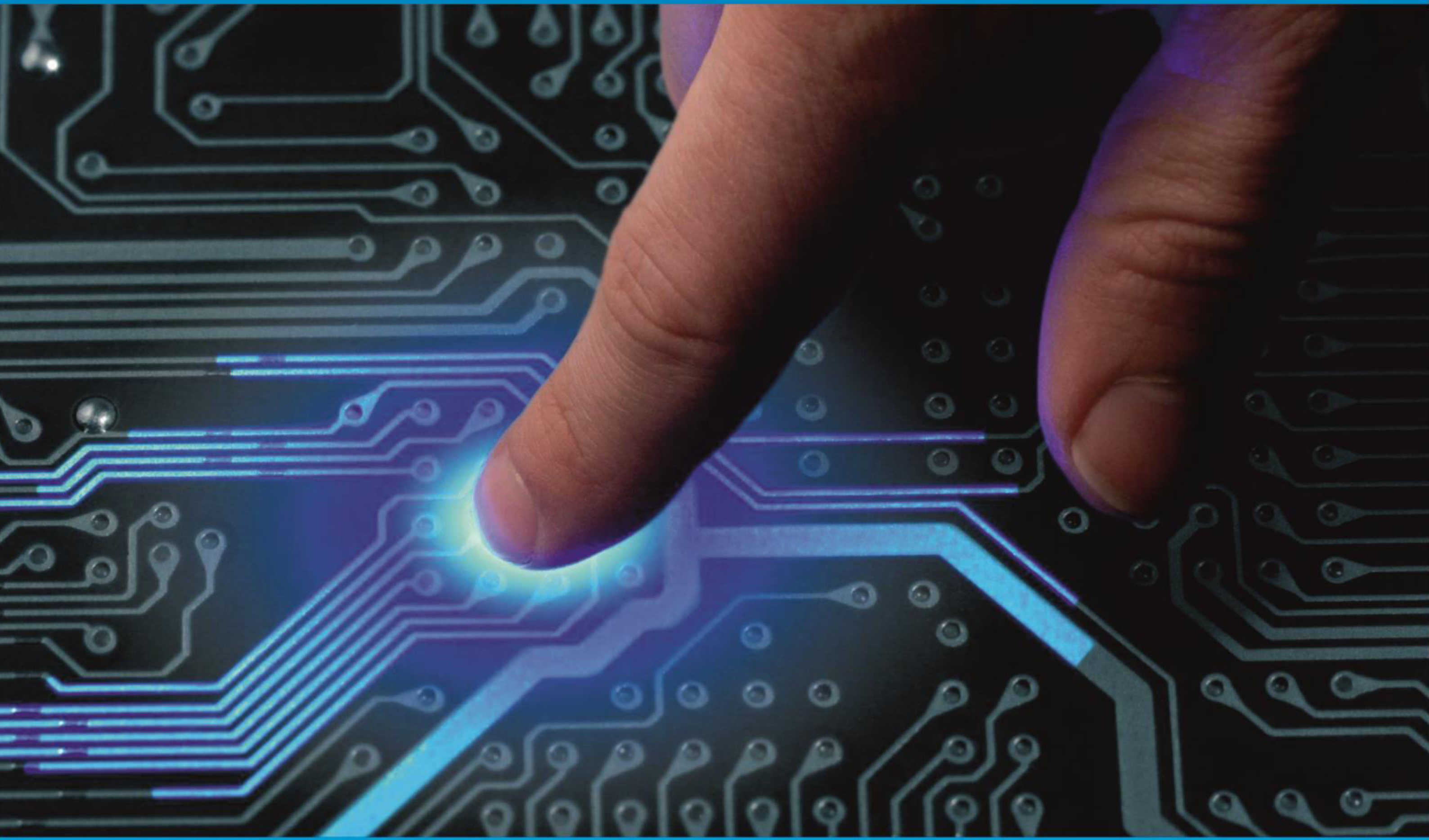
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 3, March 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903193 |

Prediction of Diabetes Mellitus Using Machine Learning Algorithms

Mr.V.R.Vimal[1], H.Suriya Babu[2], T.TamilArasan[3], P.Theepak Prakash[4]

Asst. Professor, Department of Computer Science and Engineering, Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Tamilnadu, India[1]

UG Student, Department of Computer Science and Engineering, Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Tamilnadu, India [2, 3,4,]

ABSTRACT: Diabetes is a metabolic disease affecting a multitude of people worldwide. An accurate system for diabetes prediction is proposed in this project. This project will propose a high precision diagnostic system by using K-Means Clustering algorithm. The main goal is to improve the diagnosis system by removing duplicate, uncertain and inconsistent data in shorter computational time. It will increase the classification performance as well as reduce the consumption time.

KEYWORDS: K-Means Clustering, diabetes mellitus, Elbow method, diagnosis.

I. INTRODUCTION

The aim of the project is of detecting the Diabetes at earlier stage by using various machine learning algorithm and to improve efficiency rate.Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood.Person generally suffers from high sugar level in blood which can have severe effects on other human organs.Insulin is an essential hormone produced by the pancreas that allows the cells to absorb glucose (blood sugar) from food supplies in order to provide them the necessary energy. Some of the symptoms are Intensify thirst, Intensify hunger and frequent urination.In medicine, doctors and current research confirm that if this disease is discovered at an early stage, the chances of recovery will be greater.But the identifying process is tedious, visiting to a diagnostic center and consulting doctor, i.e. these tests take a lot of time and waste budget of health care systems and people every year.But the rise in machine learning approaches gives solution to this problem.The learning algorithms use recorded datasets of former patient's information to prepare a model and then use this model with information of an unseen patient to predict if the patient has the desired disease or not.

II. LITERATURE SURVEY

Developments in machine learning (ML) offer an opportunity for improved care of individuals at risk of DFUs, future research should address direct comparison of ML applications with current standards of care, health economic analyses and large scale data collection.There is currently no evidence to contently suggest that ML methods in DFU diagnosis are ready for implementation and use in healthcare settings. SVM is the standard supervised learning algorithm. It does the complex data transformations and separates the data based on the outputs In KNN algorithm using set of rules the complete dataset is sorted. The data is divided into classes. The Dataset i.e. the statistics set has many impartial variables along with Glucose, Blood pressure, skinThickness, BMI etc.Diabetes prediction has been accomplished using the model from the Pima Indian Diabetes dataset. The quality of the dataset was improved by the proposed pre-processing scheme. Performance of model is decided with the help of confusion matrix. Support vector machine (SVM)algorithm is used in complex data classification the SVM based on a feature selection algorithm for differential space fusion (DSF-FS) is proposed. Machine learning enables machines to improve performance automatically as experience data accumulates.Original data should be preprocessed,The two key points in data preprocessing are (i) Filling in the missing value in data.(ii)normalize the data. Feature extraction and statistical modelling on PIDD is presented in this research work. The PIDD (Pima Indian Diabetes Data) features are extracted to a new space using PCA(Principal component analysis).These newly projected features are then modelled using Linear Regression Model.The results obtained in this

| DOI: 10.15680/IJIRCCE.2021.0903193 |

study have achieved high accuracy rate (82.1%) for predicting diabetes when compared with other existing methods.In this study, systematic efforts area unit created in coming up with a system that finally ends up among the prediction of illness like genetic defect.Throughout this work Random Forest algorithms area unit studied and evaluated on varied measures. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using machine learning technique which provides advance support for predicting the accuracy rate of diabetes. A comparative analysis of machine learning and deep learning-based algorithms for prediction of diabetes. It showed that RF was more effective for classification of the diabetes in all rounds of experiments which produced overall accuracy for diabetic prediction to be 83.67%. The prediction accuracy for SVM reached 65.38% while DL method produced 76.81% on dataset. In this study, there are several comparison between algorithms with different preprocessing techniques and identify algorithms best performance in which preprocessing technique. It also shows that Neural Network was given best accuracy(80.4%) than any other methods.Comparing the execution time of several methods in it's the best accuracy and found Naive Bayeswas taken less execution time than any other methodsThis paper focuses on recent developments in machine learning which have made significant impacts in the detection and diagnosis of diabetes.Each network uses algorithms to compare the system output to the desired output value, and uses the calculated error to direct the training.In the comparison of application of Artificial Neural Networkand Bayesian Network for classification of diabetes and CVD, different values for the network accuracy have been achieved.Algorithms which are used in machine learning have various power in both classification and predicting.Support vector machine (SVM),NaïveNet (NN, and Decision Stump (DS) classification algorithm are combined the prediction of them in to one.In future multiple data set can be used for prediction. In this study only limited base classifier used it is possible to use another base classifier like KNN.

III. THE PROPOSED SYSTEM

Diabetes prediction is very critical, we need algorithms with high accuracy. Hence in the proposed system, we will apply combination of PCA and K-Means algorithm. KMeans algorithm is an iterative algorithm that tries to partitions the dataset into non-overlapping subgroups**.**We will train the machine learning algorithms with training dataset of Diabetes information collected from public domain. Later we will test the accuracy of testing dataset with multiple algorithms. The modulus we are going to implement in our project as follows at first (i)Loading and Preprocessing Dataset (ii)Build and Train Dataset (iii)Testing Dataset (iv)Display Evaluation Results. We collect dataset of Diabetes Patient from public domain. We use python pandas data frame to load csv as data frame. As dataset might contain null values or missing values, we perform pre-processing such as (i)Replace missing values (ii)Normalization of values. Build and Train Dataset. We build multiple models such as SVM, ANN and K-Means. These models will be trained with pre-processed dataset.Testing Datasetwe create confusion matrix and test dataset with user input data. User input data will be loaded with separate datasetEvaluation ResultsTesting accuracy of various models will be displayed in chart for comparison. And best model to predict diabetes will be identified.

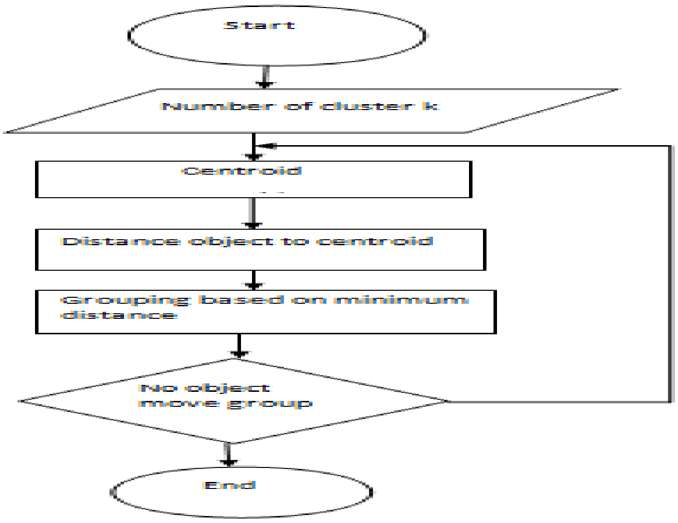

Start
Number of cluster k
Centroid
Distance object to centroid
Grouping based on minimum distance
No object move group
End

Fig. 1 Flow diagram of Proposed System

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903193 |

IV. WORKING OF PROPOSED SYSTEM

First we collect the diabetes datasetfrom a public domain then we load the dataset as csv. As dataset might contain null values or missing values, we perform pre-processing techniques. We build multiple models with the preprocessed dataset. In this project we focused on kmeans clustering algorithm to improve the efficiency in the result dataset. We create a confusion matrix and test the user input data. At last test accuracy of the various model is displayed and the best model will be identified. K-means clustering is a type of unsupervisedlearning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are: The centroids of the K clusters, which can be used to label new data. Labels for the training data (each data point is assigned to a single cluster). At first we have to select the number of clusters to be formed. Then K number of clusters has to be assigned. For each of the selected set of K clusters we have to select far and definite centroids.Now we select each item of the available set and examine its range to all the centroids of the K clusters. Therefore on the basis of evaluated range the item is joined to the cluster whose centroid is closer to the item. Therefore for every evaluation or a set of evaluations the centroids of the cluster should be recomputed. Finally, this is a repetitive method and gradually revised.

4.1.Advantages:

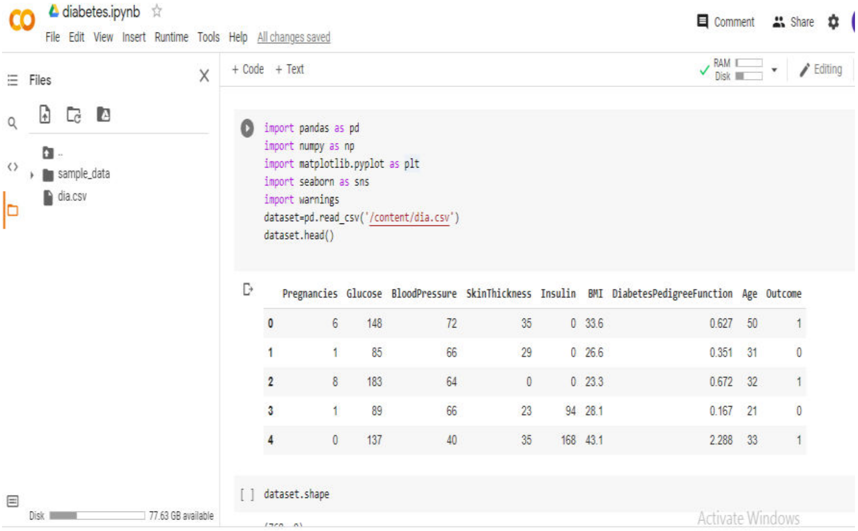
These machine learning algorithm provides various advantages when compared to other algorithms. Some of the advantages are listed below.

- These machine learning algorithms will provide high accuracy compared to optimal ANN algorithm.
- With a large number of variables, K-Means computationally will be faster than hierarchical clustering.
- The algorithm itself generalizes to clusters of different shapes and sizes, such as elliptical clusters.

V. EXPERIMENTAL RESULT

LOADING DATASET

We collect dataset of Diabetes Patient from public domain. We use python pandas data frame to load csv as data frame. As dataset might contain null values or missing values. Load the Pandas libraries with alias as pd. Read the data from file filename.csv. Control delimiters, rows, column names with read_csv.
Preview the first 5 lines of the loaded data.



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
dataset=pd.read_csv('/content/dia.csv')
dataset.head()
```

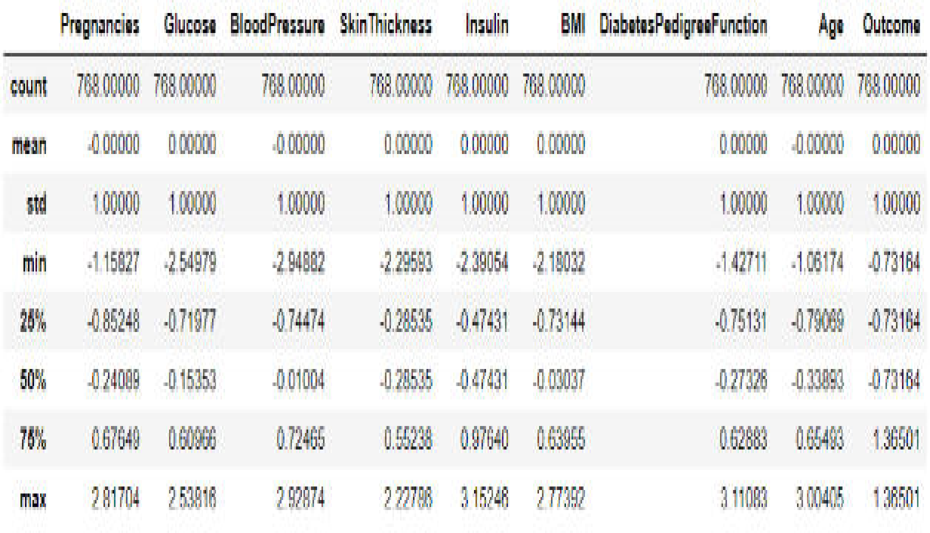
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
[ ] dataset.shape
```

Fig. 2 Loading the dataset as a CSV

| DOI: 10.15680/IJIRCCE.2021.0903193 |

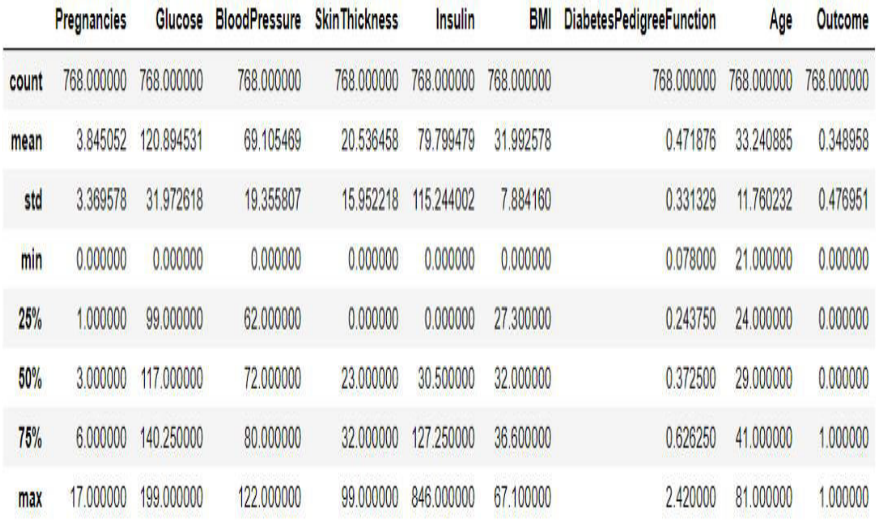
PREPROCESSING DATASET



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000	768.00000
mean	-0.00000	0.00000	-0.00000	0.00000	0.00000	0.00000	0.00000	-0.00000	0.00000
std	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
min	-1.15827	-2.54979	-2.94882	-2.29593	-2.39054	-2.18032	-1.42711	-1.06174	-0.73164
25%	-0.85248	-0.71977	-0.74474	-0.28535	-0.47431	-0.73144	-0.75131	-0.79069	-0.73164
50%	-0.24089	-0.15353	-0.01004	-0.28535	-0.47431	-0.03037	-0.27326	-0.33893	-0.73164
75%	0.67649	0.60966	0.72465	0.55238	0.97640	0.63955	0.62883	0.65493	1.36501
max	2.81704	2.53816	2.92874	2.22798	3.15248	2.77392	3.11083	3.00405	1.36501

Fig. 3 Description of Raw Data

As dataset might contain null values or missing values, we perform pre-processing such as (i)Replace missing values(ii)Normalization of values. When using machine learning algorithms we should always split our data into a training set and test set. (If the number of experiments we are running is large, then we can should be dividing our data into 3 parts, namely training set, development set and test set). In our case, we will also separate out some data for manual cross checking. The data set consists of record of 767 patients in total. To train our model we will be using 650 records. We will be using 100 records for testing, and the last 17 records to cross check our model.

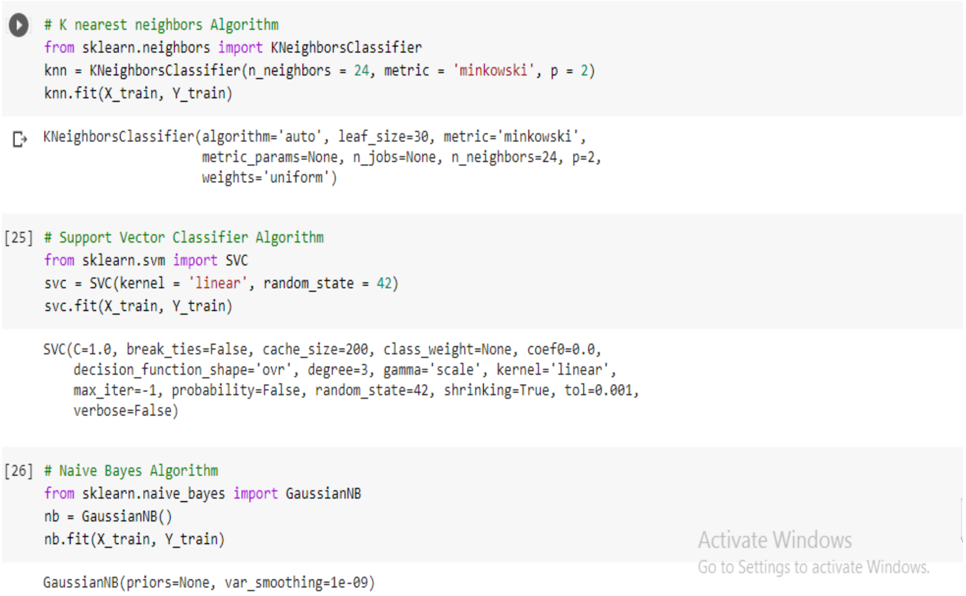


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Fig. 4 Description of Processed Data

BUILD AND TRAIN DATASET

In this module we build and train the dataset on multiple models such as SVM, ANN and K-Means. These models will be trained with pre-processed dataset.



```
# K nearest neighbors Algorithm
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors = 24, metric = 'minkowski', p = 2)
knn.fit(X_train, Y_train)
```

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=24, p=2,
                     weights='uniform')
```

```
[25] # Support Vector Classifier Algorithm
     from sklearn.svm import SVC
     svc = SVC(kernel = 'linear', random_state = 42)
     svc.fit(X_train, Y_train)
```

```
SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='linear',
    max_iter=-1, probability=False, random_state=42, shrinking=True, tol=0.001,
    verbose=False)
```

```
[26] # Naive Bayes Algorithm
     from sklearn.naive_bayes import GaussianNB
     nb = GaussianNB()
     nb.fit(X_train, Y_train)
```

```
GaussianNB(priors=None, var_smoothing=1e-09)
```

Fig. 5 Build and Train Dataset

V. FIND CLUSTER USING ELBOW METHOD

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The **Elbow Method** is one of the most popular methods to determine this optimal value of k.We now demonstrate the given method using the K-Means clustering technique using the **Sklearn** library of python.

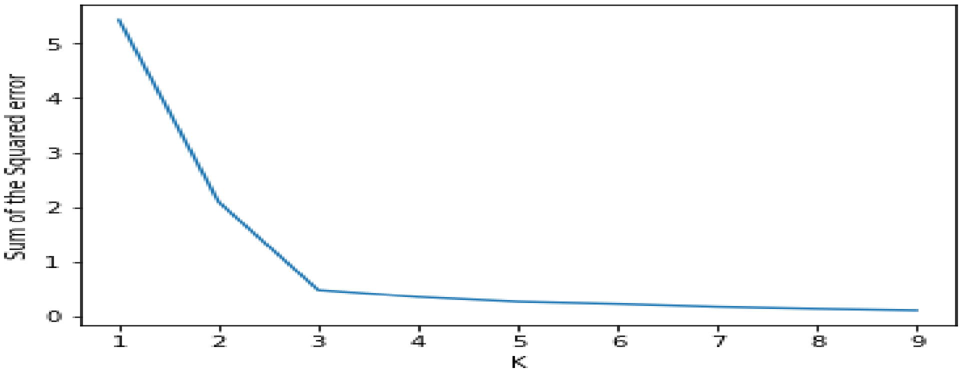


Fig. 6 Find cluster using Elbow method

VI. CONCLUSION

In this project we have investigatedthe early prediction of diabetes by taking into account several risk factors related to this disease using machine learning techniques.We have done our investigation using popular machine learning algorithm namely Support vector machine(SVM), Naïve Bayes(NB), Kmeans Clustering, Logistic Regression(LR), Random Forest (RF) on adult population data to predict diabetes. The technique which accomplishes the highest performance in terms of accuracy is considered to be the best choice.Based on that we have concluded that Kmeans Clustering achieved better accuracy to predict diabetes mellitus utilizing a given medical dataset.

REFERENCES

[1] Dong Wen*, Member, IEEE, Peng Li, Yanhong Zhou* , Yanbo Sun, Jia, "Feature Classification Method of Resting-state EEG Signals from Amnestic Mild Cognitive Impairment with Type 2 Diabetes Mellitus based on Multi-view Convolution Neural Network.DOI 10.1109/TNSRE.2020.3004462, IEEE Transactions on Neural Systems and Rehabilitation Engineering

[2] Aiswarya I., S. Jeyalatha and Ronak S., "Diagnosis Of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, ,No. 1, pp. 1-14, 2015.

[3] G. Krishnaveni*, T. Sudha," A Novel Technique To Predict Diabetic Disease Using Data Mining Classification Techniques" in International Conference on Innovative Applications in Engineering and Information Technology (ICIAEIT2017), vol. 3, Issue 1, pp. 5-11, 2017.

[4]Harleen and Dr. Pankaj B.,"A Prediction Technique in Data Mining for Diabetes Mellitus," Journal of Management Sciences and Technology, vol. 4, Issue 1, pp. 1-12, 2016.

[5] K.Rajesh and V.Sangeetha,"Application of Data Mining Methods and Techniques for Diabetes Diagnosis," in proceedings of International journal of Engineering and Innovative Technology, vol.2, Issue 3, pp. 43-46, 2012.

[6] Lowongtrakool C., Hiransakolwong N., "Noise filtering in unsupervised clustering using computation intelligence," International Journal of Math, vol. 6, no. 59, pp. 2911–2920, 2012.

[7] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019.

[8] Ravi S. and Smt T., "Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine," International Journal of Computer Trends and Technology (IJCTT), vol. 11, No. 2, pp. 94-98, 2014.

[8] V. Anuja and R.Chitra., "Classification Of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications (IJERA), vol.3,Issue 2, pp. 1797-1801, 2013.

[9] Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 2019, pp. 1-4, doi:10.1109/UBMYK48245.2019.896555

[10]Yilmaz N., Inan O., Uzer M.S., " A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases," J Med Syst, vol. 38, no. 5 2014

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor:
7.488

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details