Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Type 2 Diabetes Prediction using K-Nearest Neighbor Algorithm

Dr. S Suriya[1], J Joanish Muthu[2]

[1]Associate Professor, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamil Nadu, India.

[2]PG Scholar, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamil Nadu, India.

Email: [1]suriyas84@gmail.com, [1]ss.cse@psgtech.ac.in, [2]22mz32@psgtech.ac.in

Abstract

Type 2 diabetes is a persistent disorder that affects millions of individuals globally. It is characterised by the excessive levels of glucose within the blood due to insulin resistance or the incapability to supply insulin. Early detection and prediction of type 2 diabetes can improve patient outcomes. K-Nearest Neighbor (KNN) is used in the present model to predict type 2 diabetes. The KNN set of rules is a simple but powerful machine learning set of rules used for categorization and regression. It's far a non-parametric approach that makes predictions based totally on the nearest k-neighbours in a dataset. KNN is widely used in healthcare and scientific studies to expect and classify sicknesses primarily based on the affected person's data. The intention of this work is to predict the threat of growing type 2 diabetes using the KNN set of rules. Data has been collected from electronic medical records of patients diagnosed with type 2 diabetes and healthy individuals. The dataset consists of various patient attributes, such as age, gender, body mass index, blood pressure, cholesterol levels, and glucose levels. Information has also been collected about lifestyle habits, such as physical activity, smoking status, and alcohol consumption. Data have been pre-processed by removing missing values and outliers, and normalization of the data has been done to ensure that all features have the same scale. Splitting the dataset into training and test sets, with training sets using 80% of the data and test sets using 20% of the data is performed. KNN algorithm have been used to classify the patients into two groups: those at high risk of developing type 2 diabetes and those at low risk. The model's performance has been assessed using a variety of metrics, including accuracy, precision, recall, and F1-score.

DOI: https://doi.org/10.36548/jtcsst.2023.2.007
Received: 02.05.2023, received in revised form: 02.06.2023, accepted: 17.06.2023, published: 27.06.2023

Keywords: Prediction, F1, Recall, K Fold, Confusion Matrix

1. Introduction

A metabolic condition known as Type 2 Diabetes mellitus (T2D) is characterised by high blood sugar levels. T2D is a serious disease that affects the way the body processes blood sugar, leading to high blood sugar levels over time. Early detection and prognosis of T2D is important to prevent or delay the onset of the disease and its complications. Machine learning algorithms have shown promising results in predicting T2D. One such algorithm is the K-Nearest Neighbour (KNN) algorithm. A non-parametric approach for classification and regression is the KNN algorithm.

In the context of T2D prediction, KNN is used as a classification system to predict whether a person is at risk for T2D based on clinical and lifestyle data. The KNN algorithm works by finding the nearest neighbours of the new data and assigning the new data points to the class most of their nearest neighbours. The k value is a hyperparameter that can be tuned to find optimal results. KNN is a lazy algorithm, meaning it doesn't require any training or learning steps. Instead, it stores all the training data and uses it to make predictions about new content.

To use the KNN algorithm to predict T2D, clinical and lifestyle data from a sample population are required. Information should include major factors that pave way for diabetes. The data should also include the objective variable, whether the person has been diagnosed with T2D.

After collecting the data, it is necessary to prepare it for the KNN algorithm. This includes processing missing values, measuring data, and coding categorical variables. The pre-processed dataset will be split into training and test sets. The KNN algorithm is trained through the training process, and its performance is assessed by testing.

To prepare the KNN calculation, the preparing set is stacked into memory, and the KNN calculation is initialized with an esteem of k. For each information point within the test set, the KNN calculation calculates the separate between the test point and each point within the preparing set. Once the separations are calculated, the KNN calculation chooses the k closest neighbours to the test point and relegates the test point to the lesson that's most common among its k closest neighbours.

After the KNN calculation has been prepared, its execution is assessed on the test set. Usually done by comparing the anticipated values with the genuine values of the test set. The execution of the KNN calculation can be assessed utilizing measurements such as exactness, exactness, review, and F1 score.

In conclusion, the KNN calculation may be a promising machine learning calculation for T2D forecast. It can be utilized to anticipate whether an individual is at hazard of creating T2D or not based on their restorative and way of life information. In any case, it is vital to note that the execution of the KNN calculation is exceedingly subordinate on the quality of the dataset and the choice of hyperparameters such as k and the remove metric. In this manner, cautious pre-processing of the information and hyperparameter tuning are essential to get the most excellent results.

2. Literature Review

Research [1], employed decision tree algorithms to examine the risk variables for "Type 2 Diabetes Mellitus" (T2DM). The data were collected from 450 T2DM patients and 450 healthy individuals and utilized a decision tree algorithm to develop a model for predicting the likelihood of developing T2DM. The findings of the study indicate that age, Body Mass Index (BMI), family history of diabetes, and systolic blood pressure are the most significant risk factors for T2DM. These variables were used to construct a decision tree model, which demonstrated an impressive accuracy of 92.9% in predicting the occurrence of T2DM. Based on their results, the authors highlighted the practical value of decision tree algorithms in identifying T2DM risk factors and creating prediction models. The future research could explore alternative data mining techniques and incorporate additional risk factors to enhance the prediction accuracy.

Research [2] examined the application of machine learning techniques and medical data for predicting diabetes. The authors handled missing values, scale features, and encode categorical variables as part of the pre-processing of the data. The feature selection methods were employed to identify key predictors for diabetes, including age, BMI, glucose level, and blood pressure. The study's results show that the Random Forests algorithm performs best, with accuracy ratings of 78.88%, sensitivity ratings of 67.50%, specificity ratings of 84.76%, and an AUC-ROC of 0.82. Additionally, the authors develop a diagnostic tool based on the Random

Forests algorithm, which accurately predicts diabetes with a 78.8% accuracy rate. These results have significant implications for early diagnosis and prevention of diabetes, ultimately leading to improved health outcomes for patients.

The study **[3]** focused on the construction of a prediction model for T2DM using data mining techniques. The research involved collecting data from 168 T2DM patients and 168 healthy individuals. Through a combination of feature selection techniques and classification algorithms, the authors successfully developed the prediction model. The study highlights age, BMI, Fasting Plasma Glucose (FPG), and triglycerides as the most significant risk factors for T2DM. These variables were used to create a prediction model that achieved an accuracy of 81.55% in forecasting T2DM. The authors emphasized the potential of the developed prediction model as a valuable tool for early detection and prevention of T2DM. However, the importance of further validation and refinement of the model before its practical application in clinical settings was acknowledged. Overall, this research underscores the potential.

Study **[4]** focused on developing a prediction model for T2D using machine learning classification techniques. The dataset comprised information from 768 patients, and the author employed various machine learning. A comparison with previously published studies demonstrated that the prediction model developed in this study outperformed others in terms of accuracy for predicting T2D. The author concludes that machine learning algorithms have significant potential in predicting T2D and could be integrated into clinical practice to assist in early diagnosis. However, it is important to note that further validation and refinement of the model are necessary before its implementation in real-world clinical settings. Overall, this study highlights the promising role of machine learning methods in healthcare and disease prediction.

In **[5]**, age, BMI, and haemoglobin A1c (HbA1c) were identified as the most significant risk factors for T2D. These variables are used to create risk prediction models employing different machine learning algorithms. The accuracy rates of these models range from 77.6% to 89.2% in predicting T2D. The authors highlight the high effectiveness of machine learning algorithms in developing risk prediction models for T2D. These models have the potential to contribute to early detection and prevention of the disease. Furthermore, the authors suggest that future research could explore alternative machine learning algorithms and incorporate additional risk factors to further improve the accuracy of the prediction models.

Research **[6]** used data mining, machine learning and the deep learning in diabetes research, highlighting their significance in handling large volumes of data and uncovering hidden patterns that may elude traditional statistical approaches. The work examined numerous machine learning. Concrete examples of their use in diabetes research are provided. The authors also acknowledge challenges associated with these techniques, such as overfitting, feature selection, and model validation. The authors advocate for further research to fully exploit the capabilities of these methods and develop innovative approaches to effectively address the unique challenges posed by diabetes.

In **[7]**, the application of machine learning techniques for anticipating the onset of diabetes mellitus was investigated by the author. The goal of the study is to determine which machine learning model is the best accurate at predicting the development of diabetes mellitus. The study utilizes a dataset of 45,712 individuals who participated in a health screening program in China. The dataset includes demographic, clinical, and laboratory data. The results indicate that the artificial neural network model achieves the highest accuracy (85.46%) and AUC-ROC (0.854) in predicting diabetes mellitus. Conversely, the logistic regression model demonstrates the lowest accuracy (73.52%) and AUC-ROC (0.729). Furthermore, the study identifies age, body mass index, fasting plasma glucose, and triglycerides as the most significant factors in predicting diabetes mellitus. In conclusion, the study highlights the potential of machine learning techniques in accurately predicting the onset of diabetes mellitus. It underscores the importance of considering multiple factors when predicting disease outcomes.

In **[8]**, the authors discussed the global increase in diabetes prevalence and the importance of early detection for effective disease management. The potential of machine learning algorithms to uncover patterns within extensive datasets, enabling the prediction of diabetes before symptoms become apparent was highlighted. The study covers many machine learning methods and their use in predicting diabetes, including logistic regression, decision trees, random forests, support vector machines, and artificial neural networks. The requirement for high-quality data, fairness, avoiding biases, and the models' interpretability and explainability are a few of the difficulties the authors discuss when trying to deploy these algorithms in a healthcare setting. Additionally, the authors present a case study that uses logistic regression to predict diabetes based on demographic and clinical data. The study demonstrates a high level of accuracy in diabetes prediction, achieving an impressive AUC value of 0.84. Overall, the work emphasized the potential of machine learning algorithms in

predicting diabetes and highlights the need for further research to develop accurate and reliable models suitable for healthcare implementation. The authors propose that predictive models created using machine learning algorithms can assist healthcare professionals in identifying individuals at risk of developing diabetes, facilitating early intervention and disease prevention.

Research **[9]** presented a study centred on the application of machine learning techniques for analysing and predicting diabetes. The study begins by emphasizing the global prevalence of diabetes and the importance of early detection and treatment. Various machine learning approaches for data analysis, including decision trees, logistic regression, support vector machines, and k-nearest neighbors were studied. The results show that the decision tree algorithm demonstrated the highest performance with an accuracy of 77.21%, while the k-nearest neighbour algorithm achieved the highest sensitivity at 75.34%. Moreover, the authors compared the findings of the system with previous studies and found that the models outperformed traditional statistical models. In conclusion, the study indicates that machine learning techniques can effectively analyse and predict diabetes, potentially providing more accurate results than traditional statistical models. The authors emphasized the need for additional research to improve the performance of these models and to enable their use in real-world situations.

In **[10]**, a study centred on using machine learning techniques for classifying and predicting diabetes wqas conducted. The study begins by discussing the global prevalence of diabetes and emphasizing the importance of early detection and treatment. Several machine learning algorithms for diabetes classification and prediction was studied. Additionally, the authors compared the findings of the system with previous studies, demonstrating that the suggested models outperformed traditional statistical models. In conclusion, the study suggests that machine learning techniques can effectively classify and predict diabetes, potentially yielding more accurate results than traditional statistical models. The authors underscore the importance of further research to enhance the performance of these models and their application in real-world scenarios.

In **[11]**, the authors begin by addressing the prevalence of diabetes and emphasizing the importance of early detection and treatment. Several machine learning techniques were used for the study in forecasting diabetes, such as decision trees, Naive Bayes, and random forests. The study utilized the Pima Indians Diabetes dataset, which consisted of data from 768 patients belonging to the Pima Indian tribe. The data underwent pre-processing, and feature selection techniques were applied to identify the most relevant features for prediction. Subsequently, the

authors implemented an ensemble approach, which combines the predictions of multiple models, to enhance the accuracy of their predictions. The results demonstrated that the ensemble approach achieved an accuracy of 78.74%, surpassing the performance of individual models used in the study. In conclusion, the study suggests that machine learning algorithms, particularly the ensemble approach, can effectively be utilized for the analysis and prediction of diabetes. The authors also highlight the importance of further research aimed at improving the accuracy of these models and their practical application in real-world scenarios.

In **[12]**, the UCI Machine Learning Repository provided the diabetes dataset for the study, which included data from 768 patients with the disease. To find the most pertinent aspects for prediction, the data underwent pre-processing, and feature selection approaches were used. The results revealed that the ANN algorithm achieved the highest accuracy of 77.60%, followed by the decision tree algorithm with an accuracy of 73.70%. Additionally, the authors compared the findings of the work with those of previous studies and found that the suggested models outperformed certain traditional statistical models. The authors emphasized the importance of further research aimed at improving the performance of these models and expanding their application to different datasets and real-world scenarios.

The study **[13]** demonstrated the effectiveness of machine learning algorithms, particularly the ensemble approach, in analysing and predicting diabetes. The authors emphasized the importance of further research to improve the accuracy of these models and their practical implementation.

In **[14]**, the authors begin by performing data pre-processing, which involves addressing missing values and standardizing the features. Then a feature selection technique was applied to identify the most relevant attributes for diabetes prediction, including age, BMI, glucose level, and blood pressure. The results demonstrate that the Support Vector Machines algorithm outperforms the others, achieving an accuracy of 78.2%, sensitivity of 61.3%, specificity of 84.2%, precision of 68.9%, and an F1-score of 64.8%. This research has potential implications for timely diagnosis and treatment of diabetes, ultimately leading to improved health outcomes for patients.

In **[15]**, by addressing missing values, encoding categorical variables, and normalising numerical features, the authors begin by pre-processing the data. Then, feature selection approaches were applied to determine which characteristics are most crucial for diabetes prediction. The authors find that the Random Forests algorithm achieves the highest

performance, with an accuracy of 85.3%, sensitivity of 77.2%, specificity of 89.4%, and AUC-ROC of 0.93. Additionally, a diagnostic tool based on the Random Forests algorithm was developed, which accurately diagnoses diabetes mellitus with an accuracy of 87.8%.

3. Proposed Methodology

In this study, KNN has been used as it is easy to implement and has highest accuracy when compared to other popular machine learning algorithms like SVM, Decision Tree, etc. Even though Random Forest achieves better accuracy, it is difficult to tune the parameters to achieve that accuracy level; so in this study, detection of diabetes with KNN algorithm is performed.

KNN is essentially a classification calculation that accumulates all information that is readily available, and categorises a point of unused information based on similitude. This means that as new information appears, it may be efficiently categorised into a suitable category by using K-NN calculations.

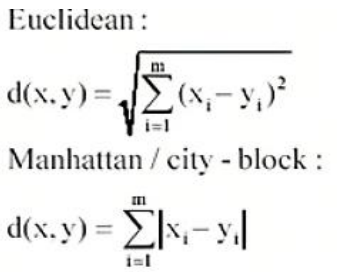

Figure 1. Formula for Euclidean and Manhattan Distance

$$eucledian\ Distance = \sqrt{\sum (x_i - y_i)^2}$$

$$Manhattan\ Distance = |x_2 - x_1| + |y_2 - y_1|$$

3.1 Steps in the KNN Algorithm

Step-1: Select K neighbours

Step-2: Calculate the Euclidean distance between K neighbours

Step-3: Take K neighbours according to Euclidean distance calculations

Step-4: Calculate the data points of each type of neighbour.

Step 5: Assign the new data point to the class with the largest number of neighbours.

Step 6: The model is ready

3.2 Dataset Description:

The research data was obtained from Pima Indian heritage dataset which is available online at Kaggle website Pima Indians Diabetes Database | Kaggle. The attributes present in the dataset are,

1. Pregnancies
2. Glucose
3. Blood Pressure
4. Skin Thickness
5. Insulin
6. BMI
7. Diabetes Pedigree Function
8. Age
9. Outcome

The description of the dataset is as follows.

- Pregnancy means how many times a woman has gone through
- Glucose represents the “2-hour plasma glucose concentration in the oral glucose tolerance test”.
- Blood Pressure signifies the Diastolic blood (mm /Hg)
- Skin Thickness indicates triceps skinfold thickness in millimetres
- Affront level implies the “2-hour serum affront in mm U/ml”

- The BMI indicates the Body mass list (weight in kg/ (tallness in m) ^2)
- The diabetes family work could be a degree of hereditary impact that appears the innate hazard of diabetes mellitus based on the diabetes history of relatives
- The age notices the age of the individual in a long time
- The result may be a classification variable and has the course variable of 1

4. Experimental Results and Discussion

A. Python

Python was the platform used and NumPy, pandas, scikit learn and mat plot were the libraries used in data pre-processing, model fitting, making predictions and plotting the results. The results achieved for various diabetes dataset are shown below.

Dataset 1: Diabetes Health Indicators Dataset | Kaggle

Table 1. Diabetic Dataset 1

Diabetes_	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	1	1	1	40	1	0	0	0	0	1	0	1	0	5	18	15	1	0	9	4	3
0	0	0	0	25	1	0	0	1	0	0	0	0	1	3	0	0	0	0	7	6	1
0	1	1	1	28	0	0	0	0	1	0	0	1	1	5	30	30	1	0	9	4	8
0	1	0	1	27	0	0	0	1	1	1	0	1	0	2	0	0	0	0	11	3	6
0	1	1	1	24	0	0	0	1	1	1	0	1	0	2	3	0	0	0	11	5	4
0	1	1	1	25	1	0	0	1	1	1	0	1	0	2	0	2	0	1	10	6	8
0	1	0	1	30	1	0	0	0	0	0	0	1	0	3	0	14	0	0	9	6	7
0	1	1	1	25	1	0	0	1	0	1	0	1	0	3	0	0	1	0	11	4	4
1	1	1	1	30	1	0	1	0	1	1	0	1	0	5	30	30	1	0	9	5	1
0	0	0	1	24	0	0	0	0	1	0	0	1	0	2	0	0	0	1	8	4	3
1	0	0	1	25	1	0	0	1	1	1	0	1	0	3	0	0	0	1	13	6	8
0	1	1	1	34	1	0	0	0	1	1	0	1	0	3	0	30	1	0	10	5	1
0	0	0	1	26	1	0	0	0	0	1	0	1	0	3	0	15	0	0	7	5	7
1	1	1	1	28	0	0	0	0	0	1	0	1	0	4	0	0	1	0	11	4	6
0	0	1	1	33	1	1	0	1	0	1	0	1	1	4	30	28	0	0	4	6	2
0	1	0	1	33	0	0	0	1	0	0	0	1	0	2	5	0	0	0	6	6	8
0	1	1	1	21	0	0	0	1	1	1	0	1	0	3	0	0	0	0	10	4	3
1	0	0	1	23	1	0	0	1	0	0	0	1	0	2	0	0	0	1	7	5	6
0	0	0	0	23	0	0	0	0	1	0	0	1	0	2	15	0	0	0	2	6	7
0	0	1	1	28	0	0	0	0	0	0	1	1	0	2	10	0	0	1	4	6	8
0	1	1	1	22	0	1	1	0	1	0	0	1	0	3	30	0	1	0	12	4	4
0	1	1	1	38	1	0	0	0	1	1	0	1	0	5	15	30	1	0	13	2	3
0	0	0	1	28	1	0	0	0	0	1	0	1	0	3	0	7	0	1	5	5	5
1	1	0	1	27	0	0	0	1	1	1	0	1	0	1	0	0	0	0	13	5	4
0	1	1	1	28	1	0	0	0	0	1	0	1	0	3	6	0	1	0	9	4	6
0	0	0	1	32	0	0	0	1	1	1	0	1	0	2	0	0	0	0	5	6	8
1	1	1	1	37	1	1	1	0	0	1	0	1	0	5	0	0	1	1	10	6	5

Dimensions: 1873 x 22

Results:

Figure 2 shows the k value vs accuracy graph for dataset 1. It can be observed that the k value attains a peak value at k=40.

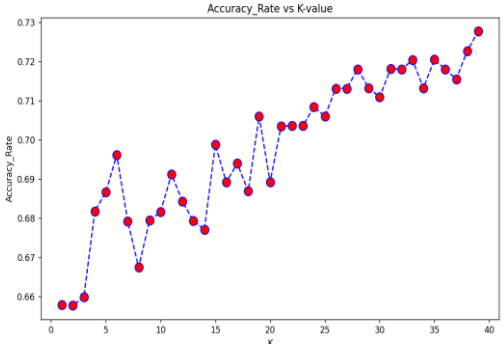

Figure 2. K value vs Accuracy for Dataset 1

Figure 3 shows the confusion matrix for dataset 1. It can be observed that the dataset has got 897 true positives and 12 true negatives for the k value of 35.

```
For K=5

[[819   0  92]
 [ 15   0   3]
 [165   1  41]]
For K=35

[[897   0  14]
 [ 18   0   0]
 [195   0  12]]

Process finished with exit code 0
```

Figure 3. Confusion Matrix for Dataset 1

Dataset 2: [Diabetes dataset | Kaggle]

Table 2. Diabetes Dataset 2

	A	B	C	D	E	F	G	H	I
1	pregnanci	glucose	bloodpres	skinthickn	insulin	bmi	diabetespe	age	outcome
2	6	148	72	35	0	33.6	0.627	50	TRUE
3	1	85	66	29	0	26.6	0.351	31	FALSE
4	8	183	64	0	0	23.3	0.672	32	TRUE
5	1	89	66	23	94	28.1	0.167	21	FALSE
6	0	137	40	35	168	43.1	2.288	33	TRUE
7	5	116	74	0	0	25.6	0.201	30	FALSE
8	3	78	50	32	88	31	0.248	26	TRUE
9	10	115	0	0	0	35.3	0.134	29	FALSE
10	2	197	70	45	543	30.5	0.158	53	TRUE
11	8	125	96	0	0	0	0.232	54	TRUE
12	4	110	92	0	0	37.6	0.191	30	FALSE
13	10	168	74	0	0	38	0.537	34	TRUE
14	10	139	80	0	0	27.1	1.441	57	FALSE
15	1	189	60	23	846	30.1	0.398	59	TRUE
16	5	166	72	19	175	25.8	0.587	51	TRUE
17	7	100	0	0	0	30	0.484	32	TRUE
18	0	118	84	47	230	45.8	0.551	31	TRUE
19	7	107	74	0	0	29.6	0.254	31	TRUE
20	1	103	30	38	83	43.3	0.183	33	FALSE
21	1	115	70	30	96	34.6	0.529	32	TRUE
22	3	126	88	41	235	39.3	0.704	27	FALSE
23	8	99	84	0	0	35.4	0.388	50	FALSE
24	7	196	90	0	0	39.8	0.451	41	TRUE
25	9	119	80	35	0	29	0.263	29	TRUE
26	11	143	94	33	146	36.6	0.254	51	TRUE
27	10	125	70	26	115	31.1	0.205	41	TRUE
28	7	147	76	0	0	39.4	0.257	43	TRUE

Dimensions: 769 x 9

Results:

Figure 4 shows the k value vs accuracy graph for dataset 2. It can be observed that the k value attains a peak value at k=21.

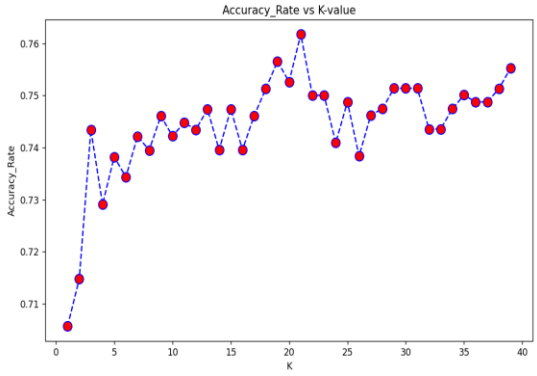

Figure 4. K value vs Accuracy for Dataset 2

Figure 5 shows the confusion matrix for dataset 2. It can be observed that the dataset has got 259 true positives and 92 true negatives for the k value of 21.

```
For K=5

[[247  61]
 [ 51 102]]
For K=21

[[259  49]
 [ 61  92]]
```

Figure 5. Confusion Matrix for Dataset 2

Dataset 3: Diabetes Dataset | Kaggle

Table 3. Diabetes Dataset 3

	A	B	C	D	E	F	G	H	I
1	Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesPe	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1
25	9	119	80	35	0	29	0.263	29	1
26	11	143	94	33	146	36.6	0.254	51	1
27	10	125	70	26	115	31.1	0.205	41	1
28	7	147	76	0	0	39.4	0.257	43	1

Dimensions: 692 x 9

The Table 3 dataset is similar to the second dataset (Table 2) with the difference that this dataset contains less rows than the previous dataset. The dimensionality of the dataset is reduced with the help of Principal Component Analysis (PCA).

Results:

Figure 6 shows the k value vs accuracy graph for dataset 3. It can be observed that the k value attains a peak value at k=5.

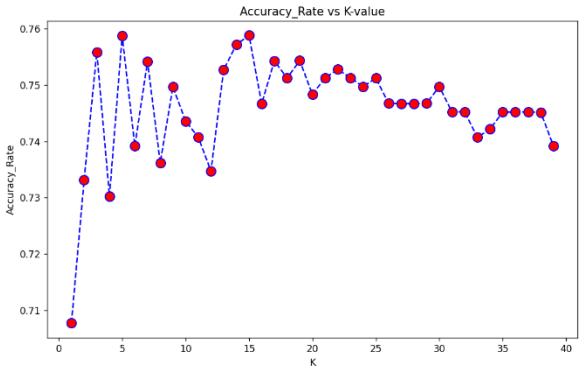

Figure 6. K Value vs Accuracy for Dataset 3

Figure 7 shows the confusion matrix for dataset 3. It can be observed that the dataset has got 228 true positives and 72 true negatives for the k value of 5.

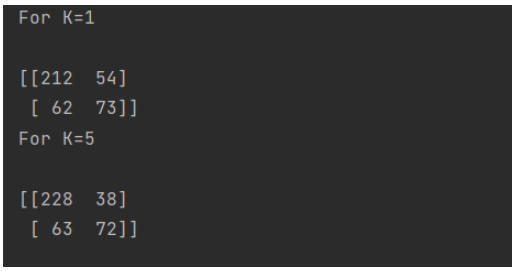

Figure 7. Confusion Matrix for Dataset 3

B. Discussion

In this study, the dataset is pre-processed to remove null values and scaled using Standard Scalar method to achieve better performance. K Fold has also been used with a fold value of 10 to avoid overfitting of data

In Figure 2 it is observed that, since the attributes and dataset is huge, the KNN accuracy gets affected and it achieves a highest accuracy score of 70% at k value of 40.

In Figure 4, using the 2nd dataset which has the dimensions of 769 x 9, it is observed that k value gets a peak at 21 and then decreases slowly and even the true positive and true negative is greater at k=21 which can be seen in Fig. 5.

In Figure 6 which is using the PIMA Indian dataset which has 692 x 9 this dataset is similar to the previous dataset with the exception of less data and it is shown that the k value reaches a max at a k value of 5 itself and the true positive and true negative are shown in fig. 7 for the dataset.

5. Conclusion and Future Works

It is observed that KNN gets a better accuracy with less dataset and in case of huge dataset, the KNN value needs to be increased to get a higher accuracy. It is also observed that the KNN performs well after the data has been scaled and preprocessed. K Fold has been done to prevent the overfitting of data. The future works that can be performed is achieving a better accuracy by providing a better dataset and using better classification techniques.

References

[1] Habibi, S., Ahmadi, M. and Alizadeh, S., 2015. Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Global journal of health science, Volume no 7 Issue no 5, page no 304.

[2] Singh, D.A.A.G., Leavline, E.J. and Baig, B.S., 2017. Diabetes prediction using medical data. Journal of Computational Intelligence in Bioinformatics, Volume no 10 Issue no 1, page no 1-8.Wu, H., Yang, S., Huang, Z., He, J. and Wang, X., 2018. "Type 2 diabetes mellitus prediction model based on data mining". Informatics in Medicine Unlocked, 10, pp.100-107.

[3] Wu, H., Yang, S., Huang, Z., He, J. and Wang, X., 2018. "Type 2 diabetes mellitus prediction model based on data mining". Informatics in Medicine Unlocked, Volume no 10, page no 100-107.

[4] Neha Prerna Tigga, Shruti Garg, 2019." Prediction of Type 2 Diabetes using Machine Learning Classification Methods". International Conference on Computational Intelligence and Data Science (ICCIDS 2019), Volume no 167, page no.706-716.

[5] Zidian Xie, Olga Nikolayeva, MS, Jiebo Luo and Dongmei Li, 2019." Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques". National Centre for Biotechnology information, Volume no 50, Page no 100-105

[6] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I., 2017. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, Volume no 15, Page no 110.

[7] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H., 2018. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, Volume no 9, page no 115.

[8] Sarwar, M.A., Kamal, N., Hamid, W. and Shah, M.A., 2018, September. Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) IEEE, Page no 1-6

[9] Saru, S. and Subashree, S., 2019. Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering, Volume no 5, Issue no 4

[10] Maniruzzaman, M., Rahman, M.J., Ahammed, B. and Abedin, M.M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, Volume no 8, Page no 1-10

[11] Alehegn, M., Joshi, R. and Alehegn, M., 2017. Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology, Volume no 4 , Issue no 10, page no 426-436.

[12] Alanazi, A.S. and Mezher, M.A., 2020, September. Using machine learning algorithms for prediction of diabetes mellitus. In 2020 International Conference on Computing and Information Technology (ICCIT-1441) IEEE, page no 1-3

[13] Panda, M., Mishra, D.P., Patro, S.M. and Salkuti, S.R., 2022. Prediction of diabetes disease using machine learning algorithms. IAES International Journal of Artificial Intelligence, Volume no 11, Issue no 1. Page no 20-40

[14] Shafi, S. and Ansari, G.A., 2021, May. Early prediction of diabetes disease & classification of algorithms using machine learning approach. In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021). Volume no 30, Issue no 4, Page no 50-58

[15] Olisah, C.C., Smith, L. and Smith, M., 2022. Diabetes mellitus prediction and diagnosis from a data pre-processing and machine learning perspective. Computer Methods and Programs in Biomedicine, Volume no 220, page no 10-20