ISSN: 2093-5374 / E-ISSN: 2093-5382

Machine Learning for Early Diabetes Detection and Diagnosis

Sofiene Mansouri^{1*}, Souhaila Boulares^{2}, and Souhir Chabchoub^{3}

^{1*}Associate Professor, Department of Biomedical Technology, College of Applied Medical Sciences in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia; University of Tunis El Manar, Higher Institute of Medical Technologies of Tunis, Laboratory of Biophysics and Medical Technologies, Tunis, Tunisia. s.mansouri@psau.edu.sa, https://orcid.org/0000-0002-6191-3095

^{2}University of Tunis El Manar, Higher Institute of Medical Technologies of Tunis, Laboratory of Biophysics and Medical Technologies, Tunis, Tunisia, Souhaila.boulares@istmt.utm.tn, https://orcid.org/0009-0008-3870-1642

^{3}Assistant Professor, University of Tunis El Manar, Higher Institute of Medical Technologies of Tunis, Laboratory of Biophysics and Medical Technologies, Tunis, Tunisia, chabchoub_souhir@yahoo.fr, https://orcid.org/0000-0002-0683-7931

Received: September 27, 2023; Revised: November 28, 2023; Accepted: January 25, 2024; Published: March 30, 2024

Abstract

In this work, a machine learning (ML)-based e-diagnostic system is suggested specifically for the detection of gestational diabetes mellitus (GDM). Reviewing recent GDM data and outlining the intimate connection between GDM and prediabetic conditions, as well as the potential for future declines in insulin resistance and the emergence of overt Type 2 diabetes, were our goals. The present study explores the application of the K-nearest neighbors (KNN) algorithm to project diabetes diagnosis on the widely-used Pima Indians Diabetes database. The KNN algorithm, a non-parametric, instance-based learning method, was employed to classify individuals as either diabetic or non-diabetic, our objectives were to evaluate the algorithm's ability to make accurate predictions and explore factors influencing its performance. The study commenced with data preprocessing, including handling missing values, feature scaling, and data splitting into training and testing sets. The KNN classifier was trained and tested using these best-fit parameters. The results of this study revealed a model with an accuracy of approximately 0.76 in predicting diabetes diagnosis. This study looked at the various machine-learning approaches for diabetes patient classification, including recall, accuracy, precision, and F1-score. The study discusses the significance of hyperparameter tuning, data preprocessing, and imbalanced data handling in achieving optimal KNN model performance. Lastly, this study shows how the KNN algorithm may be used to project diabetes using the Pima Indians Diabetes Database. The findings suggest that KNN can serve as a viable tool in the early detection of diabetes, paving the way for more extensive applications in healthcare and predictive modelling.

Keywords: Machine Learning (ML), Gestational Diabetes Mellitus (GDM), Pima Indians Diabetes Dataset, K-nearest Neighbors (KNN).

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 15, number: 1 (March), pp. 216-230. DOI: *10.58346/JOWUA.2024.I1.015*

*Corresponding author: Associate Professor, Department of Biomedical Technology, College of Applied Medical Sciences in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia; University of Tunis El Manar, Higher Institute of Medical Technologies of Tunis, Laboratory of Biophysics and Medical Technologies, Tunis, Tunisia.

1 Introduction

One of the most prevalent metabolic diseases worldwide is diabetes, and during the past few decades, the disease has become more widespread in adults. (Shaw et al., 2010: Whiting et al., 2011). Diabetes mellitus, a collection of metabolic illnesses caused by deficiencies in either the metabolism or release of insulin, or both, is characterized by persistently high blood sugar levels. Protein, fat, and carbohydrate metabolism are all asymmetrical considering insulin is a hormone essential to anabolism (Amiruzzaman et al., 2022).

Low levels of insulin required to induce an adequate response or insulin resistance in target tissues mainly the liver, fat cells, and skeletal muscles at the level of receptors for insulin, the signal transduction system, effector enzymes, or genes are the causes of metabolic diseases (Nizam et al., 2023).

Gestational diabetes mellitus (GDM) is one of the most common medical conditions linked to pregnancy. If ignored, GDM can have a serious detrimental effect on the health of both the mother and the unborn child (Buchanan et al., 2012; Crowther et al., 2005). According to the most recent estimates provided by the International Diabetes Federation (IDF), 14% (95% confidence interval: 13.97–14.04%) of pregnancies worldwide or roughly 20 million babies annually are affected by GDM (Wang et al., 2022).

Mothers diagnosed with gestational diabetes mellitus (GDM) are susceptible to the development of gestational hypertension, pre-eclampsia, and caesarean delivery (Kondracki et al., 2022). Furthermore, complications include a history of cardiovascular disease, obesity, and impaired glucose metabolism all of which are made more likely by gestational diabetes mellitus (GDM) which may lead to type 2 diabetes (T2DM) both for the mother and the child (Lee et al., 2018; Mcintyre et al., 2019; Lenoir-Wijnkoop et al., 2015). In addition to being a serious financial burden, the rising prevalence of GDM warrants more focus and knowledge (Xu et al., 2017). Diabetes mellitus (T2DM) is a multifaceted, diverse set of metabolic disorders marked by increased blood glucose levels because of decreased insulin secretion or action (Das & Elbein, 2006). Regardless of blood glucose levels, pancreatic β-cells are physiologically obligated to continuously manufacture insulin.

The latest development in machine learning (ML) has improved the computer system's capacity to identify and label images, and project diseases, which leads to making better decisions by utilizing data analysis. Machine learning applications are designed to train a computer system to perform better than a single human (Shichkina et al., 2020). Testing data is used for evaluation, and the supervised learning technique is used to train the model (Maniruzzaman et al., 2020). According to the data acquired from the efforts of the various authors, improved methods are built to build a hybrid early diabetes illness prediction system (Figure 1). The recently created program makes accurate predictions about the changes by utilizing diabetes data that is available to the general population. Consequently, by lowering the quantity of false positives, the system can identify the diabetes condition. This research paper aims to develop predictive models using machine learning algorithms for diabetes risk assessment based on the Pima Indian Diabetes Dataset. The database provides valuable features related to diabetes, and the study uses the techniques of machine learning to build models that accurately predict diabetes risk. The research evaluates various algorithms and features to enhance predictive accuracy and, consequently, contribute to early diabetes detection applications in healthcare and predictive modelling (Juma et al., 2023).

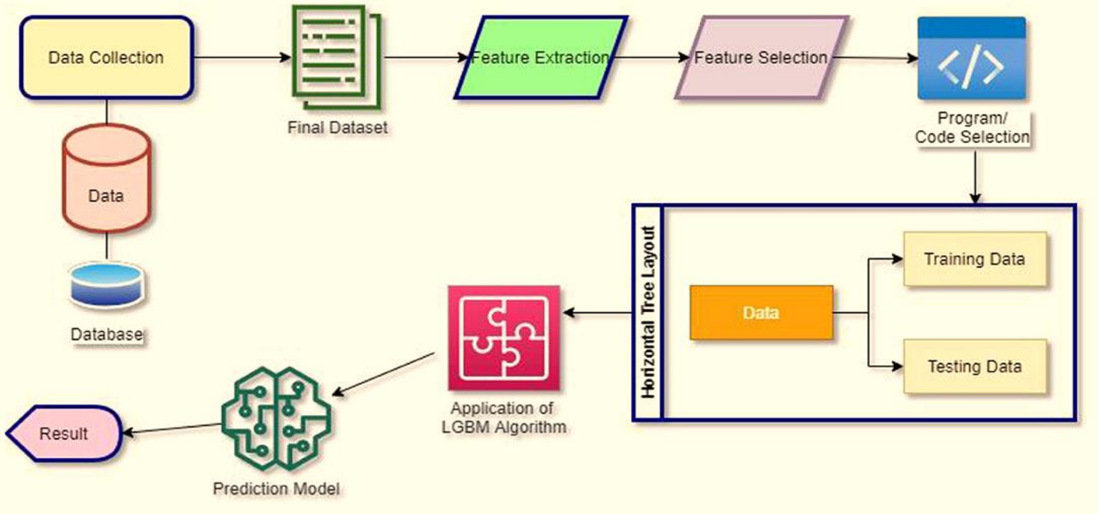


Figure 1: Type 2 diabetes mellitus disease prognosis using machine learning

When it comes to data science and machine learning, among the key methods for creating a prediction model is categorization. Different apparatus Studying strategies can help assess the facts from several angles and condense them into useful data. KNN stands for K-Nearest Neighbour (Smith & Everhart 1988) and is among the most well-known and basic devices. It helps in discovering how to classify things to anticipate illness using medical data. Only the KNN-based model includes the training data set stored. To anticipate an algorithm is used to find the closest information points in the training dataset "closest" (Ozyilmaz 2023).

2 Review of Literature

The reviews in this section of the literature concentrate on research that uses several machine learning algorithms to diagnose diabetes in its early stages. Lu et al., 2022, demonstrated the use of machine learning using an approach for Type 2 diabetes illness prediction. They created a bipartite graph using patient data from health insurance companies and projecting it onto the individual's network. The 2012 study by Kelarav et al. concentrated on using ensemble and DT classifiers to forecast cardiovascular nerve damage in diabetics. Ensemble classifiers allude to hybrid machine-learning models that use features from multiple models to enhance their efficacy. DT-based ensemble models, such as the J48, AD Tree, NB Tree, Random Tree, REP Tree, and Simple Cart DTs, were incorporated into their investigation. They also tried to find out the possibility of optimizing several ensemble approaches in this manner. Their best results, achieved with an accuracy of 94.84%, were from using AdaBoost with DT models and ensemble bagging.

They trained eight machine learning algorithms for prediction by employing particular features and attributes. The studies' AUC, which varied from 0.79 to 0.91, showed positive effects during the time of the investigation. According to the study's findings, combining network analysis and machine learning (ML) techniques can help predict sickness risk in the diabetes area more precisely. Diabetes is analysed by Sneha & Gangil, 2019 to predict conditions early on. The study suggests utilizing machine learning methods to create a prediction algorithm. The suggested approach focuses on applying predictive analysis to identify characteristics that help in the early detection of diabetes. Abdollahi & Moghaddam, 2022 employed an ensemble training methodology based on genetic algorithms to precisely identify and predict the consequences of diabetes mellitus. Actual Indian diabetes data from a University of California portal is used in this study together with experimental data.

To predict diabetes early on, Sisodia & Sisodia, 2018 employ classification algorithms. Pima Indians Diabetes Dataset trials are used to assess three ML algorithms using different approaches. Results indicated that an NB classification algorithm could interpret data with an accuracy of 76.30%. Further research can be carried out to improve the automation of diabetes analysis by employing additional machine-learning algorithms. Rajesh & Sangeetha, 2012 employ a range of ML classifiers, including NB, KNN, SVM, and DT techniques like ID3 and C4.5, to forecast the diagnosis of diabetes. They used the Pima Indians dataset to carry out their investigation. To achieve 100% accuracy, they also employed an RF algorithm. However, they concluded that the results should not be acknowledged because their RF model suffered from data overfitting. Rather, by applying the C4.5DT method, they obtain an accuracy of 90.62%. Since the method is widely used and effective in medical applications, they concluded that this model was the best for diabetes detection systems. Krishnamoorthi et al., 2022 developed a framework for using machine learning techniques to predict illnesses in healthcare.

Based on the network b-colouring approach, Vijayalakshmi & Thilagavathi, 2012 created a clustering algorithm for diabetes prediction. They put their strategy into practice, carried out tests, and contrasted it with KNN classification and K-means clustering. The outcomes demonstrated that, in terms of accuracy and purity, graph colouring-based clustering performs better than the alternative clustering techniques. The suggested method ensures the inter-cluster disparity in partitioning and provides an accurate representation of clusters through dominating items, which can be used to assess a cluster's quality. To predict diabetes, In 2016, Barale and Shirke combined the use of logistic regression classifiers with the K-means clustering technique and an artificial neural network (ANN). The updated artificial immune recognition system (AIRS), which was used by Chikh et al. in 2012, achieves the highest accuracy in this work. To diagnose diabetes, they employed the fuzzy K-nearest neighbor technique is 98%. The K-means clustering technique is employed to uncover latent patterns within the dataset. The diabetes dataset was classified by Christobel & Sivaprakasam 2013 using class-wise K-nearest neighbor (CkNN). Data preprocessing is completed in the first step, and missing values are replaced with the mean value. Using a modified KNN, classification is finished in the second stage of diabetes. In this paper, the highest accuracy of 78.16% was attained.

k-NN, or k-Nearest Neighbor

Despite being a simple method, k-Nearest Neighbor produces excellent results. It is a non-parametric, instance-based approach for lazy learning. This method can be applied to both regression and classification problems. When classifying, k-NN is used to determine which class a newly discovered unlabelled object belongs to. This is accomplished by choosing an odd number, "k," to represent the number of neighbours to be taken into account. The distance between the data points nearest to the objects is then determined using approaches such as Minkowski, Hamming, Manhattan, and Euclidian distances. The "k" closest neighbours are selected after the distance computation, and their votes are used to establish the new object's class.

Step by Step Diabetes Classification-KNN-detailed

Dataset

The Pima Indians Diabetes Dataset is a publicly accessible dataset that was used in this study (Chang et al., 2022). In total, 768 female patients' data are included in this dataset, which has eight independent parameters: age, diabetes pedigree function, skin thickness (ST), insulin, blood pressure (BP), glucose, and pregnancies. There is also one dependent component, outcome. Table 1 displays the dataset's first ten records.

The goal of this research is to use sophisticated algorithms to examine the Pima Indian Dataset to efficiently use IoMT. The dataset, which can be accessed under a CC0: Public

Domain License was obtained via Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database

Dataset Description:

Number of Instances: 768

Number of Attributes: 8

Parameters

1. Pregnancies: No. of times pregnant.
2. Glucose: The plasma glucose concentration was more than two hours during an oral glucose tolerance test.
3. Blood Pressure: Diastolic heart rate (mm Hg).
4. Insulin: Two-hour serum insulin (mu U/ml).
5. Skin Thickness: Triceps skinfold thickness (mm).
6. BMI: The body mass index.
7. Diabetes Pedigree Function (DPF): is a tool that uses family history to score a person's risk of developing diabetes.
8. Age: The age of the person (in years).
9. Outcome/Result: Class variable.
   0 means nondiabetic
   1 means diabetic.

Table 1: Shows the first ten records in the dataset

Sr. No.	Pregnancies	Glucose	Blood Pressure	Insulin	Skin Thickness	BMI	DPF	Age	Outcome
1	6	148	72	0	35	33.6	0.627	50	1
2	1	85	66	0	29	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	94	23	28.1	0.167	21	0
5	0	137	40	168	35	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	88	32	31	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	543	45	30.5	0.158	53	1
10	8	125	96	0	0	0	0.232	54	1

The dataset contains no null or missing values. Domain knowledge, however, suggests that the values of the following attributes are inconsistent: skin fold thickness (Skin), blood pressure (BP), insulin, and glucose concentration (Gluc) and BMI. Zero values, on the other hand, are not within the normal range and are therefore erroneous. The goal is to predict the value of the ninth variable (diabetes = yes (1) diabetes = no (0)) using the first eight variables.

It is discovered that there are numerous missing values in this dataset, despite the owners' claims that there are none. To achieve that, we must first preprocess the data before utilizing it. When data processing techniques are used before mining, both the overall quality of the patterns mined and the amount of time needed for actual mining can be significantly increased. As excellent data are required to support excellent conclusions, preprocessing data is a crucial step in the knowledge discovery process.

Exploratory Data Analysis (EDA) and Statistical Analysis

The extraction of as much information as possible from the provided data is known as EDA. Here, we look at the attributes that are given to see what they represent (numerical values, categories, etc.). Table 2 displays the quality of the data, identifying and fixing any null, missing, or filler value. Look for any trends, correlations, or unusuality in the data. The ultimate goal is to identify useful qualities that will yield the best prediction outcomes.

Range Index: Entries= 0 to 767

Data columns (9)

Table 2: Null, missing or filler values

Column	D type	Non-Null Count
0 Pregnancies	int_64	768
1 Glucose	int_64	768
2 B.P.	int_64	768
3 Insulin	int_64	768
4 Skin Thickness	int_64	768
5 BMI	int_64	768
6 DPF	int_64	768
7 Age	int_64	768
8 Outcome	int_64	768

D types: float_64(2), int_64(7)

The Data Frame describe () method: Excludes NaN values, and produces descriptive statistics that highlight the distribution's shape, central tendency, and dispersion. Numerous things about a dataset are revealed by this method. The describe () technique only works with numerical values, which is a crucial distinction. No categorical values can be used with it. Put another way, the describe () method will display the summary for the other columns and ignore any category values in the column unless option include="all" is provided.

Let's now examine the statistics produced by the describe () function:

- count indicates how many non-empty rows there are in a feature.
- mean provides the feature's mean value.
- The Standard Deviation Value of the feature is provided by std.
- min provides the feature's lowest value.
- For every attribute, the corresponding percentile and quartile are 25%, 50%, and 75%.
- max indicates the feature's maximum value.

Table 3: Statistical measures of the data

	Pregnancies	**Glucose**	**Blood Pressure**	**Insulin**	**Skin Thickness**	**BMI**	**Diabetes Pedigree Function**	**Age**	**Outcome**
Count	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000	768.000000	768.000000	768.000000
Mean	3.845052	120.894531	69.105469	79.799479	20.536458	31.992578	0.471876	33.240885	0.348958
Std	3.369578	31.972618	19.355807	115.244002	15.952218	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	30.500000	23.000000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	127.250000	32.000000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	846.000000	99.000000	67.100000	2.420000	81.000000	1.000000

Table 3 displays the data's statistical measurements and the diabetes dataset attributes are represented as a histogram in Figure 2. Visual representations of data points clustered into intervals that the viewer chooses are called histograms. A series of statistics is summarized into an understandable visual by the histogram, which looks like a bar chart. The numbers are clustered into distinct groups called "bins."

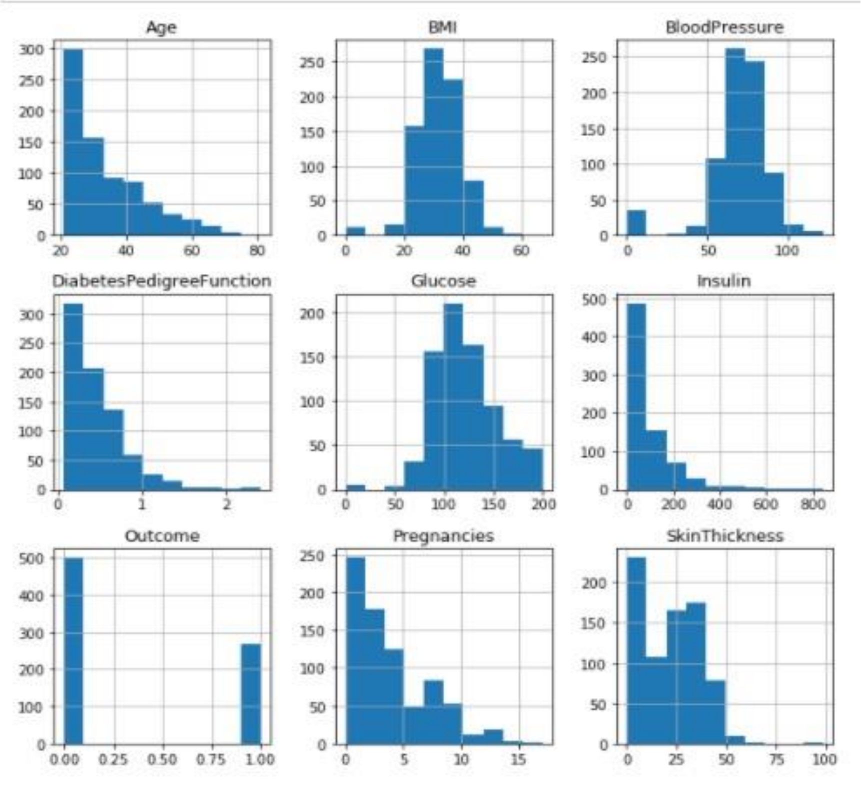
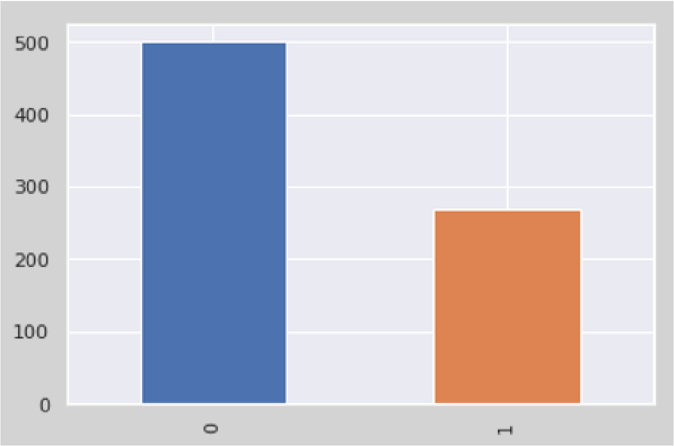


Figure 2: Histogram of each attribute

Check for Missing Values: It is imperative to verify if the Pima Indians Diabetes Dataset contains any missing values before utilizing the K-nearest Neighbors (KNN) technique. One of the most important steps in data preprocessing is handling missing values to guarantee the precision and dependability of your machine-learning model (Table 4).

Table 4: Check for missing values

	Pregnancy	Glucose	Blood Pressure	Insulin	Skin Thickness	BMI	DPF	Age	Outcome/Result
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
763	0	0	0	0	0	0	0	0	0
764	0	0	0	0	0	0	0	0	0
765	0	0	0	0	0	0	0	0	0
766	0	0	0	0	0	0	0	0	0
767	0	0	0	0	0	0	0	0	0



```
0   500
1   268
```

Figure 3: Outcome

The data is skewed in favour of data points with an outcome value of 0, which suggests that diabetes was not present, as the graph above shows. About twice as many people do not have diabetes as do those who do. It does not contain any identifiable information on the patient participants and is suitably anonymized. As Figure 3, shows, it enumerates eight causal attributes together with the appropriate classification. The dataset consists of 768 rows with 9 columns (consisting of 268 diabetics and 500 non-diabetics). A positive result from the diabetes test is represented by a value of 1, and a negative result by a value of 0, which is the outcome variable for the binary classification. The above graph shows how the data is biased toward data points with result values of 0, which signal the absence of diabetes. The proportion of the population without diabetes is about double that of those who have the condition.

Techniques for Test Train Split and Cross-Validation

Split Train Test: to test the data using unknown data points as opposed to the same points that were used to train the model. This greatly improves the model's ability to record performance.

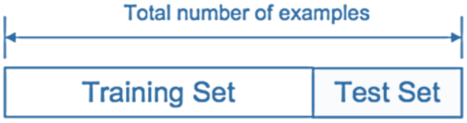


Figure 4: Training and Test Set

Cross Validation: It is common for a certain type of data point to be used only in the testing or training component of a model that has been split into training and testing halves (Figure 4). The model would perform below par as a result. Cross-validation techniques can therefore successfully avoid problems with over-fitting and under-fitting.

Stratify: The Stratify parameter separates the data to ensure a match between the percentage of values created in the sample and the proportion of values supplied to the Stratify parameter. If variable y is a binary categorical variable with values of 0 and 1, then stratify=y ensures that 25% of your random split will contain 0s and 75% of 1s. (Figure 5). For split, cross-validation, testing, and training, use this URL to access the Python reference page: 80b61beca4b6.

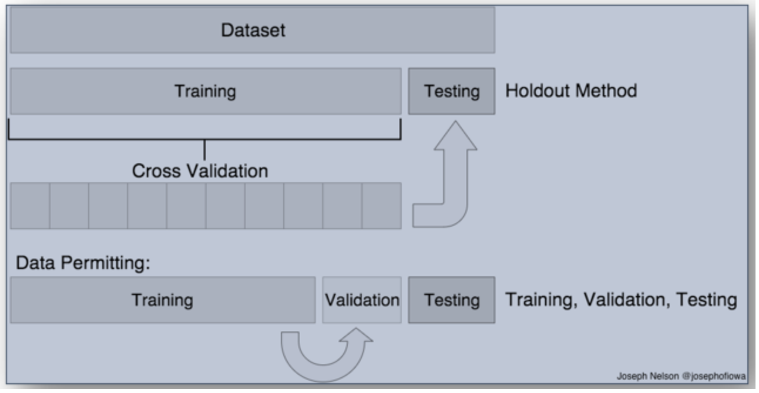


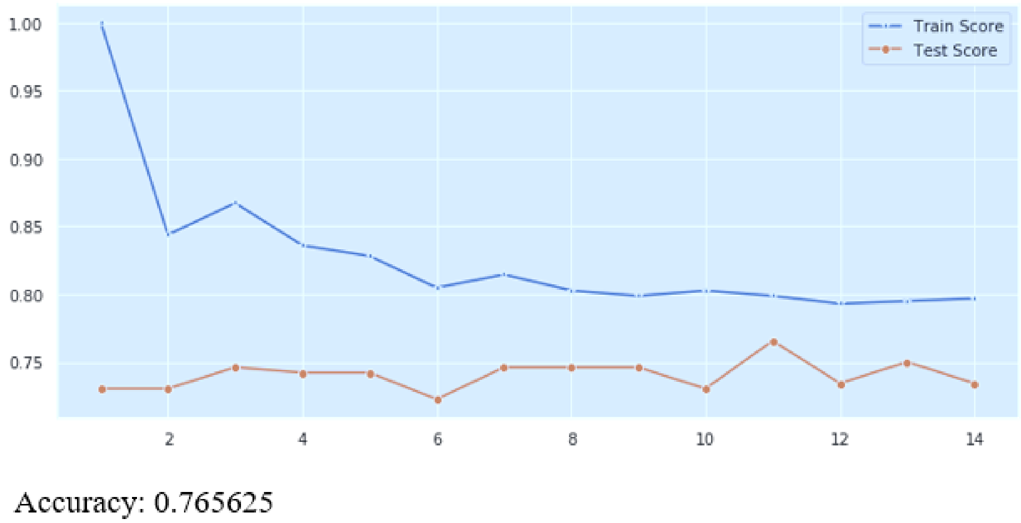
Figure 5: Cross Validation

Splitting Input and Target Variable

Train/Test Split

The train-test split approach is used to measure the performance of machine learning algorithms used for non-training data prediction. It is quick and easy to evaluate the effectiveness of machine learning algorithms for a specific predictive modeling scenario, and the process produces results. The approach is easy to understand and apply, but it should not be used in certain scenarios (e.g., tiny datasets). In other cases, further preparation is needed (e.g., imbalanced datasets and classification using the procedure). (Figure 6, 7).

```
linkcode
plt.figure (fig size=(12,5))
p=sns.line plot (range(1,15), train_scores, market='*', label=’Train Score’)
p=sns.lineplot (range(1,15), test_scores, market=’o’,label=’Test Score’)
```



```
Accuracy: 0.765625
```

Figure 6: Train/Test Split for KNN model

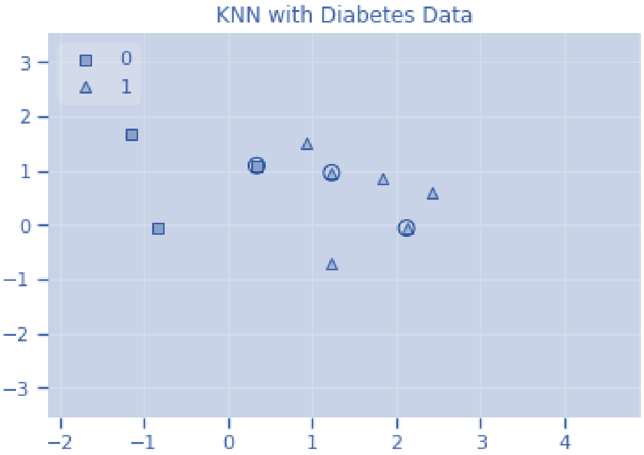


Figure 7: KNN with Diabetes Data

Confusion Matrix

1. The dataset is split into features (X) and the target variable (y), where ‘Outcome’ represents the binary classification target (1 for diabetic, 0 for non-diabetic).
2. Train_test_split divides the data into two sets: a training set (X_train, y_train) and a testing set (X_test, y_test).
3. A K neighbours Classifier is created with a specified value for n_neighbours. Based on the tuning of your hyperparameters, we ought to select a suitable value.
4. Using knn. fit (X_train, y_train), the KNN classifier is trained on the training set of data.
5. On the test data, predictions are generated using knn.predict(X_test).
6. The confusion matrix is computed using confusion_matri x (y_test, y_pred).
7. Finally, the confusion matrix is printed to the console, providing information about the classifier’s performance on the Pima Indians Diabetes dataset (Figure 8).

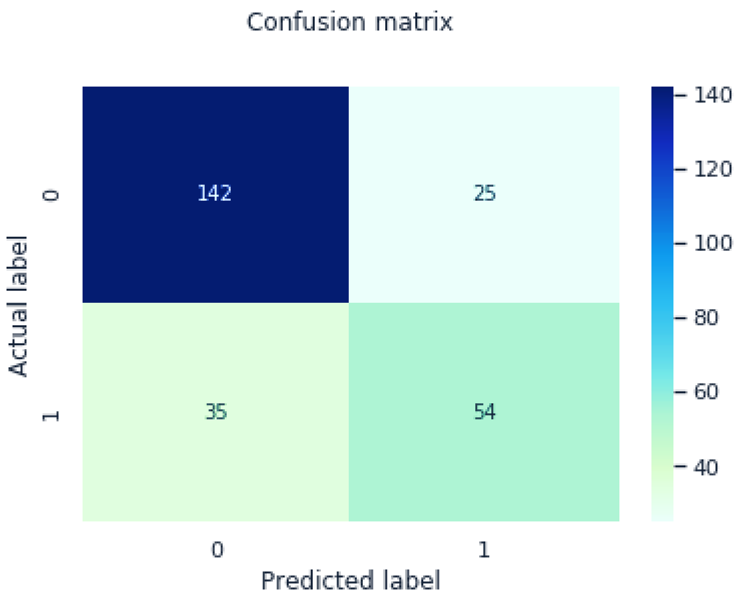


Figure 8: Confusion Matrix

Classification Report

A report on classification that includes F1-score, recall, and precision.

True Positives, or TP, stands for Precision Score

False Positives, or FP

Precision: Positive forecast accuracy.

$$\frac{TP}{TP+FP} \text{ equals precision}$$

Recall Rating: FN, or False Negatives

Recall: The fraction of accurately identified positives (also known as true positive rate or sensitivity).

$$\frac{TP}{TP+FN} \text{ is the recall}$$

F1 Score

F1 Score: A useful statistic for comparing two classifiers, also known as F-Score or F-Measure.

Recall and precision are factors in the F1 Score.

The harmonic mean of recall and precision is found to make it.

F1 is equal to 2 x (recall x precision)/ (recall + precision).

Precision

The fraction of accurately predicted positive observations among all expected positive observations is measured by a statistic called precision. The question of how many of the passengers who were supposed to have survived did so is addressed by this measure. A low false positive rate is connected with high accuracy. Our precision is 0.80, which is quite good.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall (Sensitivity): Recall is defined as the proportion of correctly anticipated positive observations to all observations made during the actual class. How many of the passengers who survived were we able to identify? A recall of more than 0.8 is regarded as good.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1 score: The F1 Score is comprised of weighted averages for Precision and Recall. Thus, this score takes into account both false positives and false negatives. F1 is usually more advantageous than accuracy, even though it is not as intuitively clear-cut as accuracy. This is especially true when there is an uneven class distribution. If the expense of false positives and false negatives is the same, accuracy will be at its best. If the cost of false positives and false negatives differs significantly, it is recommended to include both Precision and Recall.

(Recall + Precision) / 2(Recall Precision) equals the F1 score.

	Precision	Recall	F1-score	Support
1	0.68	0.61	0.64	89
0	0.80	0.85	0.83	167
avg (micro)	0.77	0.77	0.77	256
avg (macro)	0.74	0.73	0.73	256
avg (Weighted)	0.76	0.77	0.76	256

Figure 9: Results for the KNN Model

3 Result and Discussion

By identifying the k patients in the Pima dataset who are the most comparable to the patients being classed, KNN can be used to categorize patients as either diabetic or non-diabetic. To enhance the KNN algorithm's performance, one can adjust the value of k. A prediction that is more stable but potentially less accurate will be produced by a higher value of k. Using the Pima dataset, KNN has been demonstrated to be a successful machine-learning technique for early diabetes prediction. Research has demonstrated that KNN can use the Pima dataset to detect diabetes early with up to 76% accuracy. By determining which features are most crucial for diabetes prediction, feature selection can help KNN perform better Finding the most pertinent features in a dataset is the process of feature selection. To do this, a variety of methods can be applied, such as correlation analysis and feature importance rating. Once the most important features have been identified, a KNN classifier can be trained with them. The final classifier will most likely be more accurate than a classifier that is trained on every feature in the dataset. In conclusion, even if the accuracy of 0.76 is a good beginning, there are several ways to enhance the KNN model's performance using the Pima Indians Diabetes dataset. More investigation, testing, and taking into account various assessment criteria will contribute to the development of a more reliable and accurate diabetes prediction model.

The initial model was created to use the KNN algorithm to predict a person's diabetes condition. The K Nearest Neighbors Classifier function from the Python programming library was used in this experiment, and K = 8 neighbors were taken into consideration at point 2 when the Euclidean distance method was used to compute the number of neighbors. The experiment yielded an accuracy of 76%, with precision of 0.80, recall of 0.85 f, F1 Score of 0.83, and support of 167 when 70% of the data were used for training and the remaining 30% for testing (Figure 9).

4 Conclusion

Determining whether or not the KNN classification algorithm is suitable for prediction was the main objective of this study. We can see this by examining the performance analysis, on which we scored 76%. To find this accuracy, we use the Scikit-Learn module in Python. It is appropriate to utilize this accuracy for forecasting. The experiment's findings can be utilized in the medical field to predict and make prompt judgments about treating diabetes and saving lives. The general population and hospital administration would both greatly benefit from using this pattern to identify individuals with diabetes. We can find the results easily. Finally, utilizing the Pima dataset, KNN is a straightforward and powerful machine-learning technique that may be applied to the early detection of diabetes. By utilizing feature selection to determine the most crucial features for diabetes prediction, the performance of KNN can be enhanced. This task was accomplished in multiple steps. This method used the K-Nearest Neighbours (KNN) classification algorithm. This machine-learning technique allows us to assess the accuracy of the KNN algorithm's diabetes prediction. Furthermore, our 76% accuracy score suggests that using it for prediction is preferred. The study concludes that we can attain high-performance accuracy by employing this KNN method. There exist other variations of the KNN method that may result in an accuracy score that deviates from the current one. Regression and classification problems can be solved with the KNN approach. For each new data point, the KNN algorithm predicts its value using a method known as "feature similarity". Put another way, the training set's points are compared to the new point to determine its worth. Outlier reduction and management of missing values are necessary feature engineering techniques since the KNN dataset is noisy.

References

[1] Amiruzzaman, M., Islam, M.R., Islam, M.R., & Nor, R.M. (2022). Analysis of COVID-19: An infectious disease spread. *Journal of Internet Services and Information Security (JISIS), 12*(3), 1-15.

[2] Barale, M.S., & Shirke, D.T. (2016). Cascaded modeling for PIMA Indian diabetes data. *International Journal of Computer Applications*, *139*(11), 1-4.

[3] Buchanan, T.A., Xiang, A.H., & Page, K.A. (2012). Gestational diabetes mellitus: risks and management during and after pregnancy. *Nature Reviews Endocrinology*, *8*(11), 639–649.

[4] Chikh, M.A., Saidi, M., & Settouti, N. (2012). Diagnosis of diabetes diseases using an artificial immune recognition system (AIRS2) with fuzzy k-nearest neighbor. *Journal of medical systems*, *36*(5), 2721–2729.

[5] Christobel, A., & Sivaprakasam P. (2013). A New Class wise k Nearest Neighbor (CKNN) method for the classification of diabetes dataset. *International Journal of Engineering and Advanced Technology*, *2*(3), 396–200.

[6] Crowther, C.A., Hiller, J.E., Moss, J.R., McPhee, A.J., Jeffries, W.S., & Robinson, J.S. (2005). Effect of treatment of gestational diabetes mellitus on pregnancy outcomes. *New England journal of medicine*, *352*(24), 2477-2486.

[7] Das, S.K., & Elbein, S.C. (2006). The genetic basis of type 2 diabetes. *Cellscience*, *2*(4), 100–131.

[8] Juma, J., Mdodo, R.M., & Gichoya, D. (2023). Multiplier Design using Machine Learning Alogorithms for Energy Efficiency. *Journal of VLSI Circuits and Systems*, *5*(1), 28-34.

[9] Kelarev, A.V., Stranieri, A., Yearwood, J.L., & Jelinek, H.F. (2012). Empirical study of decision trees and ensemble classifiers for monitoring of diabetes patients in pervasive healthcare. *In 15th International Conference on Network-Based Information Systems*, 441-446.

[10] Kondracki, A.J., Valente, M.J., Ibrahimou, B., & Bursac, Z. (2022). Risk of large for gestational age births at early, full and late term in relation to pre-pregnancy body mass index: mediation by gestational diabetes status. *Paediatric and Perinatal Epidemiology*, *36*(4), 566-576.

[11] Krishnamoorthi, R., Joshi, S., Almarzouki, H.Z., Shukla, P.K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of Healthcare Engineering*, *2022*.

[12] Lee, K.W., Ching, S.M., Ramachandran, V., Yee, A., Hoo, F.K., Chia, Y.C., & Veettil, S.K. (2018). Prevalence and risk factors of gestational diabetes mellitus in Asia: a systematic review and meta-analysis. *BMC pregnancy and childbirth*, *18*, 1-20.

Lenoir-Wijnkoop, I., Van der Beek, E.M., Garssen, J., Nuijten, M.J., & Uauy, R.D. (2015). Health economic modeling to assess short-term costs of maternal overweight, gestational diabetes, and related macrosomia–a pilot evaluation. *Frontiers in pharmacology*, *6*, 103. https://doi.org/10.3389/fphar.2015.00103

[13] Lu, H., Uddin, S., Hajati, F., Moni, M.A., & Khushi, M. (2022). A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. *Applied Intelligence*, *52*(3), 2411-2422.

[14] Maniruzzaman, M., Rahman, M.J., Ahammed, B., & Abedin, M.M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, *8*, 1-14.

McIntyre, H.D., Catalano, P., Zhang, C., Desoye, G., Mathiesen, E.R., & Damm, P. (2019). Gestational diabetes mellitus. *Nature reviews Disease primers*, *5*, 47. https://doi.org/10.1038/s41572-019-0098-8.

[15] Nizam, M., Zaneta, S., & Basri, F. (2023). Machine Learning based Human eye disease interpretation. *International Journal of Communication and Computer Technologies (IJCCTS)*, *11*(2), 42-52.

[16] Ozyilmaz, A.T., & Bayram, E.I. (2023). Glucose-Sensitive Biosensor Design by Zinc Ferrite ($ZnFe_2O4$) Nanoparticle-Modified Poly (o-toluidine) Film. *Natural and Engineering Sciences*, *8*(3), 202-213.

[17] Rajesh, K., & Sangeetha, V. (2012). Application of data mining methods and techniques for diabetes diagnosis. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(3), 224-229.

[18] Shaw, J.E., Sicree, R.A., & Zimmet, P.Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes research and clinical practice*, *87*(1), 4-14.

[19] Shichkina, Y.A., Kataeva, G.V., Irishina, Y.A., & Stanevich, E.S. (2020). The use of mobile phones to monitor the status of patients with Parkinson's disease. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JOWUA)*, *11*(2), 55-73.

[20] Sisodia, D., & Sisodia, D.S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, *132*, 1578-1585.

[21] Smith, J.W. & Everhart, J.E. (1988). "Predict the onset of diabetes based on diagnostic measures," *Pima Indians Diabetes Database*.

[22] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big data*, *6*(1), 1-19.

[23] Vijayalakshmi, D., & Thilagavathi, K. (2012). An approach for prediction of diabetic disease by using b-colouring technique in clustering analysis. *International Journal of Applied Mathematical Research*, *1*(4), 520-530.

[24] Wang, H., Li, N., Chivese, T., Werfalli, M., Sun, H., Yuen, L., & Yang, X. (2022). IDF diabetes atlas: estimation of global and regional gestational diabetes mellitus prevalence for 2021 by International Association of Diabetes in Pregnancy Study Group's Criteria. *Diabetes research and clinical practice*, *183*, 109050. https://doi.org/10.1016/j.diabres.2021.109050.

[25] Whiting, D.R., Guariguata, L., Weil, C., & Shaw, J. (2011). IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes research and clinical practice*, *94*(3), 311-321.

[26] Xu, T., Dainelli, L., Yu, K., Ma, L., Zolezzi, I. S., Detzel, P., & Fang, H. (2017). The short-term health and economic burden of gestational diabetes mellitus in China: a modelling study. *BMJ open*, *7*(12), e018893. https://doi.org/10.1136/bmjopen-2017-018893

Authors Biography

Sofiene Mansouri is an Associate Professor at Department of Biomedical Technology, Prince Sattam bin Abdelaziz University, Saudi Arabia. He obtained his PhD from AL-Manar University, Tunis (Tunisia) in Medical Electronics in 2011. His current research interests include Biomedical Engineering and Devices, Instrumentation; Signal Processing, Bioimpedance and AI in medical diagnosis. He is the author of more than 40 scientific papers.

Souhaila Boulares was born in Kef Village, Tunisia in 1991 She obtained her diploma of applied science in biomedical engineering, having also two Master’s degrees: The first is Research Master’s degree in Biophysics, Medical Physics and Medical Imaging, and the second is Professional Master’s degree in Biomedical Engineering from Tunis AL Manar University, Tunisia. Since 2015, she is manager of the research laboratory in biophysics and medical technologies, at the Higher Institute of Medical Technologies of Tunis, Tunisia. She is currently enrolled in a doctoral thesis.

Souhir Chabchoub is an Assistant Professor in Biophysics at the Higher Institute of Medical Technologies of Tunis, TUNISIA. She received her M.S. degree in Biomedical Engineering and M.S. degree in Biophysics, Medical Physics and Medical imaging from Tunis-EL Manar University, TUNISIA, in 2011 and 2013, respectively. She received the PhD degree in Biophysics, Medical Physics and Medical imaging from Tunis-EL Manar University, in 2017. She is currently member of the research laboratory in Biophysics and medical technologies, at the Higher Institute of Medical Technologies of Tunis. Her research interests include Biophysics and Biomedical signal processing.