EVOLUTION IN ELECTRICAL AND ELECTRONIC ENGINEERING
e-ISSN: 2756-8458

EEEE

Vol. 5 No. 1 (2024) 1-9
https://publisher.uthm.edu.my/periodicals/index.php/eeee

Development of Diabetes Diagnosis Tool Using Machine Learning

M.Aiman Haziq[1], N.S. Suriani[1]*

[1] *Internet of Things Focus Group, Faculty of Electrical and Electronic Engineering*
Universiti Tun Hussein Onn Malaysia, Batu Pahat, 86400, Johor, MALAYSIA

*Corresponding Author: nsuraya@uthm.edu.my

DOI: https://doi.org/10.30880/eeee.2024.05.01.001

Article Info

Received: 03 July 2023
Accepted: 14 February 2024
Available online: 30 April 2024

Keywords

Diabetes, Obesity rates, Logistic Regression, Support VectorMachine, K-Nearest Neighbors

Abstract

This work endeavors to create an anticipatory system employing a diverse array of machine learning methodologies, encompassing Logistic Regression, Support Vector Machine, and K-Nearest Neighbor algorithms. The efficacy of each technique is meticulously evaluated, with a keen focus on identifying the most precise model for the early prediction of diabetes. The overarching goal of this initiative is to enhance the early detection capabilities and overall awareness of diabetes within the context of Malaysia. By identifying and implementing the most accurate predictive model, the work aspires to contribute significantly to mitigating the pervasive impact of this prevalent non-communicable disease. Through the advancement of predictive analytics, the work holds the potential to alleviate the burden imposed by diabetes on individuals and healthcare systems, ultimately fostering a healthier population and promoting proactive health management in the Malaysian context.

1. Introduction

Diabetes is one of the worst diseases there is. Obesity, a high blood glucose level, and other factors can cause diabetes. It alters the function of the hormone insulin, which causes crabs to have anabnormal metabolism and raises blood sugar levels. It changes how the hormone insulin works, whichmakes crabs' metabolisms erratic and raises their blood sugar levels. Diabetes develops when the bodydoes not produce enough insulin. According to the World Health Organization, 422 million people worldwide, primarily in low- and middle-income countries, have diabetes. And this number could riseto 490 billion by the year 2030[1].

In Malaysia, a nationwide survey study states that in Malaysia, 3.6 million persons (18 and older) haddiabetes in 2019; 49% of these cases—3.7 million—were undiagnosed [2]. With a prevalence of 31.3%, diabetes is anticipated to impact 7 million Malaysian individuals 18 and older by 2025, creating a serious risk to the public's health [3]. According to papers that have been published, the prevalence of diabetes in Malaysia varies from 7.3% to 23.8% [4]. Numerous factors, including population growth, population ageing, urbanization, growing obesity and physical inactivity rates, and population growth all contribute to the upward trend [5]. This study was driven by the rising prevalence of diabetes and its consequences in Malaysia to systematically identify, characterize, and quantify the prevalence of diabetes and prediabetes.

2. Material and Methods

This section will describe the development process for a tool to diagnose diabetes using machine learning classification tools. It starts with a broad overview of the entire system. The dataset in the first subsection and followed by the flowchart of the system.

2.1 Diabetes Datasets

This work utilizes data from the Pima Indian Diabetes dataset and the Kaggle dataset from Kaggle.com, which were gathered from the female Pima Indian population. As in Table 1, the Pima Indian Diabetesdataset includes 8 factors or attributes, including demographic and medical information about patients. Whilst, Table 2 summarizes the Kaggle dataset that includes 100,000 patients, with 58552 female and 41430 male patients, and 9 factorsor attributes, including demographic and medical information. These data sets can be used to create machine learning models that predict the likelihood of diabetes in patients based on their demographicsand medical histories.

Table 1 *Attributes of PIMA Dataset*

List of Attributes	Datatype
Number of time pregnant	Numeric
Glucose	Numeric
Blood Pressure	Numeric
Skinfold thickness	Numeric
Insulin	Numeric
Body mass index (BMI)	Numeric
Diabetes predigree function	Numeric
Age (years)	Numeric
Outcome class	Nominal

Table 2 *Attributes of Kaggle Dataset*

List of Attributes	Datatype
Gender	Numeric
Age	Numeric
Hypertension	Numeric
Heart Rate	Numeric
Smoking history	Nominal
Body mass index (BMI)	Numeric
HbAic level	Numeric
Blood pressure	Numeric
Outcome class	Nominal

2.2 Programming Environment

The Python programming language and the Anaconda Navigator application, both of which serve as useful tools for Machine Learning algorithms, are the two different environments used in this work. Both environments have well-researched, efficient Machine Learning packages, and they are the most frequently used environments worldwide for data mining and predictive analysis techniques. Because Anaconda Navigator is a multipurpose platform, it was chosen to finish this work. Python ispreferred because its syntax is simple to understand. The Anaconda Navigator Jupiter Notebook was employed in this work to write code and create classification models.

These are several packages in python, as described earlier, which contain algorithms. Some of those libraries, packages, and modules are used in this work are described as follows [6]:

- **NumPy**: This mathematic operation to calculate mean values is performed using the main libraryfor scientific calculations, NumPy.
- **Pandas:** is primarily a data analysis and manipulation library. It previously read data from a.CSVfiles.
- **Scikit-learn**: also known as Sklearn, is a Python library for machine learning. It includes variousclustering, regression, and classification algorithms.
- **Matplotlib**: This library for plotting uses Matplotlib. This library includes a module calledmatplotlib. pylot that offers a plotting approach in MATLAB

2.3 Diabetes Prediction Model

In illustrating the workflow within this section, a machine learning model is presented. A series of actions must be undertaken to establish a robust model for decision-making, enabling the anticipation of outcomes for patients with diabetes. Constructing the final model involves seven core phases, each accompanied by corresponding sub-steps. As shown in Fig. 1, the first step was Data Collection, which requiresfinding and data gathering related to the problem. We used a diabetes dataset "Pima Indian Diabetes Dataset" [7] and "Kaggle Dataset" [8]. In the second step, the data was pre-processed since it can includemany gaps, missing values, and outliers. In the third step, the dataset was analyzed in terms of description and separated into three sets which is logistic regression, support vector machine and k- nearest neighbor as per the model requirements. In the fourth step, different parametric performance. In thefifth step, the implementation was done in coding environments. In the sixth step, the results were interpreted and analyzed for different algorithms in different algorithm in different programming environments in the seventh step, and the eighth and final step, the model was ready to compare the conclusion for both the environments.

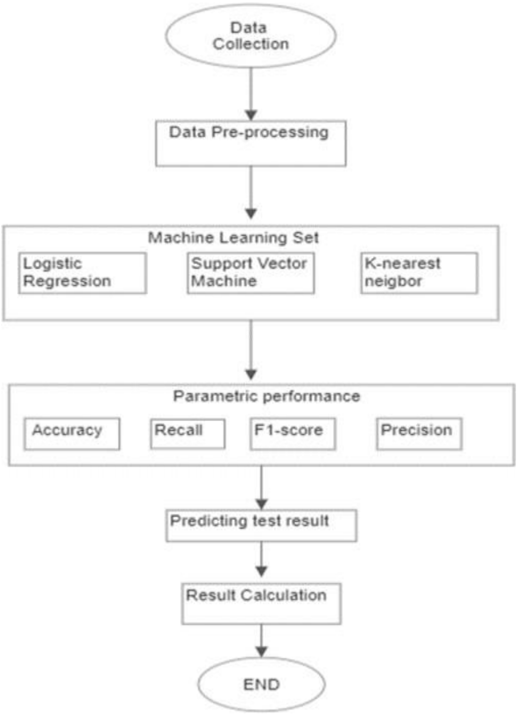


Fig. 1 *Flowchart Diabetes Prediction Model*

2.4 Classification Algorithm

The work deals with K-nearest neighbor, Support vector machines, and Logistic regression. The Python programming language is used in Jupiter Notebook's Anaconda navigator library to implement K-nearest neighbor, Logistic Regression, and Support Vector Machine. PyQt5 is also used to implement a graphical user interface. A supervised learning algorithm called logistic regression uses input variables (x) and a target variable to train the model (y). Contrary to linear regression, the output or target variable in logistic regression is a categorical variable, making logistic regression a binary classification algorithm that assigns a datapoint to one of the data classes [9]. The basic formula for logistic regression is:

$$\text{Log}(p(X)/(1-p(X))) = \beta\sigma+\beta 1X \qquad (1)$$

By using its underlying logistic function to estimate probabilities, logistic regression determines the relationship between the independent variables, the input, and the dependent variable, the output. L2 penalty is used for regularization. The logistic function, also known as the sigmoid function, which is derived from the resultant probabilities, then converts them to binary values of 0 or 1. Any real numbercan be transformed into a value between 0 and 1, excluding the limits themselves, using the sigmoid function [10]. A threshold classifier then converts the outcome to a binary value. The input features should be independent of one another, meaning that one variable should have little to no co-linearity with the other variables. This is one of the main premises of logistic regression. Consequently, PCA isperformed on the data in advance, to convert the correlated variables to a set of uncorrelated variables.

A supervised learning model called Support Vector Machines, or SVM, analyses data used forclassification, regression analysis, and outlier identification [11]. The associated learning algorithms are also presented. Despite being a binary linear non-probabilistic classifier, the method can be changed to performnon-linear and probabilistic classification, making it a versatile one. To categorize and clearly separate the examples, an SVM model represents them as points in space that are mapped. New instances are then projected into the same space and mapped into the appropriate category, depending on which side of the gap they fall on. The main advantage of SVM is that it performs well in multidimensional environments. Since only a portion of it the fact that dense datasets often yield the best results. In this work, SVM was implemented using the Scikit-Learn SVC class.

KNN is another supervised machine learning algorithm. KNN aids in the resolution of both the classification and regression issues. KNN is a sluggish prediction method. KNN assumesthat related things are located close to one another [12]. Similar data points are frequently located closeto one another. Based on a similarity metric, KNN aids in grouping new work. The KNN algorithm collects all the records and categorizes them based on how similar they are. Uses a tree-like structure tocalculate the distance between two points. The algorithm locates the nearest data points in the trainingdata set, or the data points nearest neighbors, to make a prediction for a new data point. Here, K is standfor the number of close neighbors, which is always a positive integer. Neighbor's value is selected from a list of classes.

2.5 Performance Metrics

The proportion of accurate predictions among all the model's predictions is called accuracy. Despitebeing widely used, it is not a very reliable indicator of performance, particularly when the dataset is unbalanced as it was in this instance. The accuracy equation is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

The F1-score is used to measure how accurate a test is. The harmonic mean between recall andprecision is the F1-score. The F1-score range is [0, 1]. It reveals the classifier's robustness and precision.

$$F1 - score = \frac{TP}{TP + 1/2(FP + FN)} \tag{3}$$

Precision is the number of correct positive results divided by the number of all relevant samples. In mathematical form it is given as:

$$Precision = \frac{TP}{(TP + FP)} \tag{4}$$

Recall is the number of correct positive results divided by the number of all the relevant samples. In mathematical form it is given as:

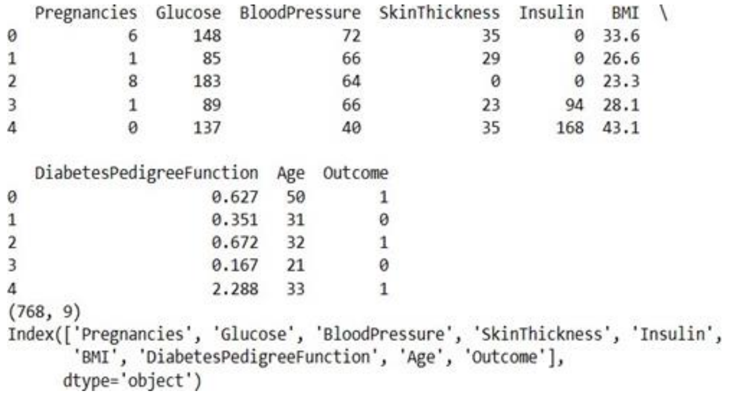
$$Recall = \frac{TP}{(TP + FN)} \tag{5}$$

3. Results and Discussion

This section explores classification experiments using scikit-learn Python and supervised learning algorithms K-nearest neighbor, Support vector machine, and Logistic Regression on "PIMA Diabetic" and "Kaggle" patients' datasets. The experiments evaluate the effectiveness of these algorithms based on specific assessment metrics and identify the best classifier or model for these tests.

3.1 Data Attributes for PIMA and Kaggle Dataset

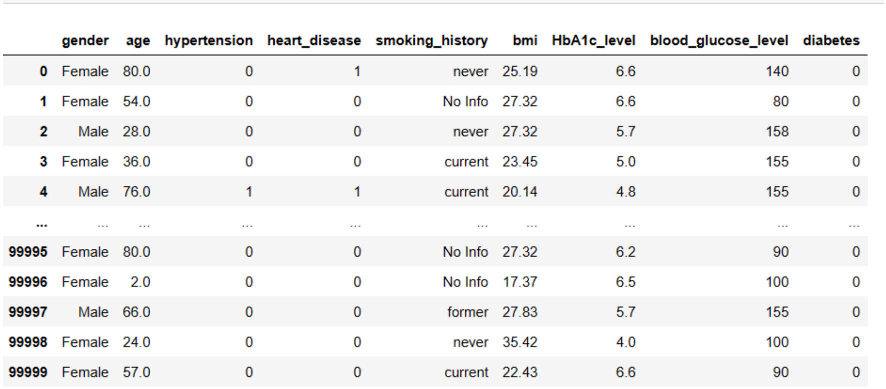
The PIMA dataset contains 769 instances with 9 attributes, each with a description in Fig. 2. The coding in CSV format outputs various values for these attributes, including pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome.The output also indicates whether the patient has diabetes or not. For Fig. 3 the Kaggle dataset 100,000 instances, the coding shows the output of 9 attributes, including gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, blood glucose level, and diabetes. The outcome indicates whether the patient has diabetes or not.



```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1
(768, 9)
Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
      dtype='object')
```

Fig.2 *Data attributes information patient from the PIMA Dataset*



	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...	...
99995	Female	80.0	0	0	No Info	27.32	6.2	90	0
99996	Female	2.0	0	0	No Info	17.37	6.5	100	0
99997	Male	66.0	0	0	former	27.83	5.7	155	0
99998	Female	24.0	0	0	never	35.42	4.0	100	0
99999	Female	57.0	0	0	current	22.43	6.6	90	0

Fig.3 *Data attributes information patient from the Kaggle Dataset*

3.2 Bar Plot for PIMA and Kaggle Dataset

The bar plot graph in Fig. 4 compares the number of people classified as diabetes and non-diabetic. The non-diabetic group is represented by a bar with a value of 500, indicating those who do not have diabetes, while the diabetic group is represented by a bar with a value of 250. The significantvariance in bar heights highlights the difference between non-diabetic and diabetic people in the datasetor population under investigation. The bar plot in Fig. 4 provides an easy-to-understand visual representation of the imbalance by displaying the relative proportions of these two categories. This toolis useful for determining the distribution and prevalence of diabetes in the dataset or population under study. The bar plot provides a clear visual representation of the relative proportions, making it easy to understand how the two categories are imbalanced. By using the bar plot, category counts can be easilycompared, making it easier to understand the prevalence of diabetes in the dataset or population understudy.

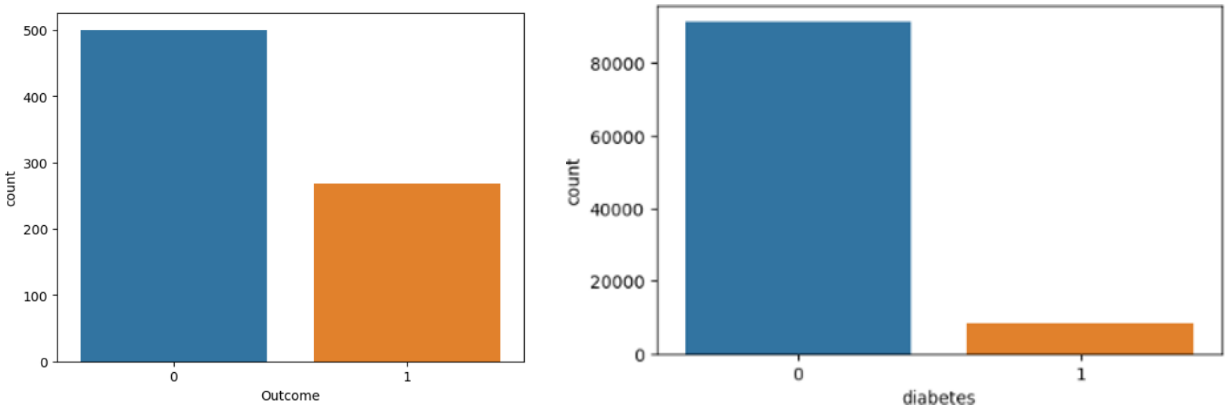


Fig. 4 *Bar Plot tested positive and negative for PIMA Dataset (Left) and Kaggle Dataset (Right)*

3.3 Distribution for PIMA and Kaggle Dataset

In Fig. 5, the age distribution graph shows varying density values along the y-axis, indicatingthe concentration of individuals in specific age groups. The pregnancies distribution graph displays concentrations of pregnancies within different categories. The glucose distribution graph provides insights into the concentration of individuals at different blood glucose levels. The skin thickness distribution graph illustrates concentrations within various skin thickness categories. The Diabetespedigree function distribution graph depicts concentrations at different levels of the diabetes. The insulin distribution graph reveals concentrations at different insulin levels. Furthermore, the BMI distribution graph indicates concentrations within various BMI categories.These graphs offer valuable insights into the concentration of individuals across different factors.

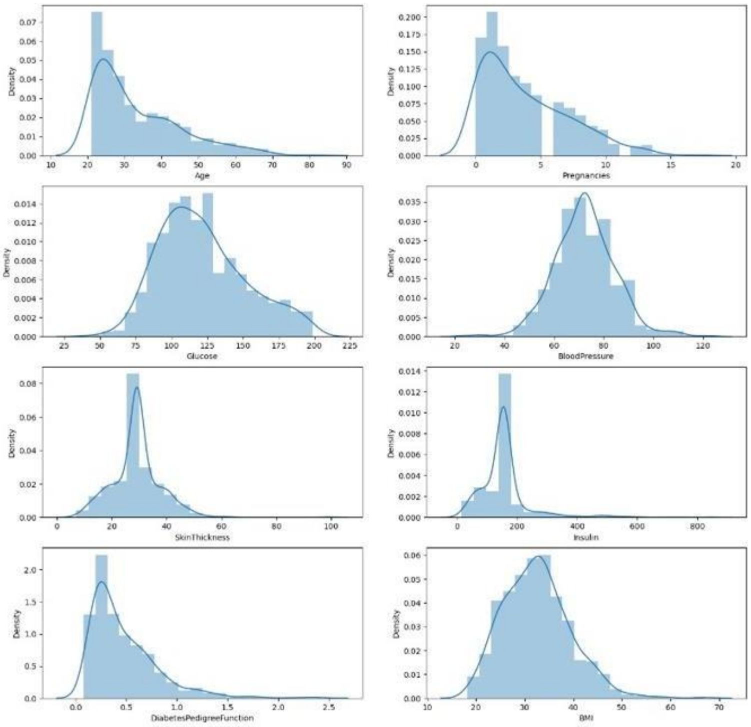


Fig. 5 *Distribution graph for PIMA Dataset*

Fig. 6 shows the age distribution graph revealing a concentration of individuals in specificage groups, with a notable increase in density at age 40 and the highest density at age 80. The BMI distribution graph shows a moderate concentration at a healthy weight range (BMI 20) and a significantconcentration at overweight or obese levels (BMI 30), with a decrease in density at severely obese levels (BMI 40). The HbA1c distribution graph indicates a higher concentration at HbA1c 3, consistent densityat HbA1c 4 and 5, a significant increase at HbA1c 6, and a decrease at HbA1c 7, 8, and 9. Lastly, the blood glucose distribution graph displays a higher concentration at blood glucose 100, with decreasingconcentrations at higher blood glucose levels.

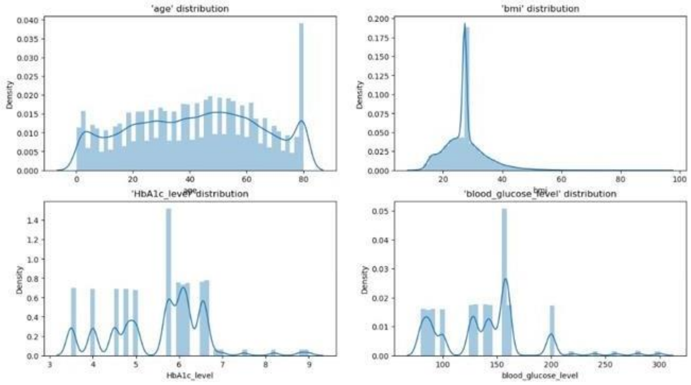


Fig. 6 *Distribution graph for Kaggle Dataset*

3.4 Heatmap for PIMA and Kaggle Dataset

The heatmap in Fig. 7 and Fig. 8 is a graphical representation of individual values of a matrix represented by colors, using a logistic regression model. The heatmap displays nine attributes, including pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes predigree formation, age, and outcome. The heatmap uses a warm to cold color spectrum, with warm areas havinghigh values and cold areas having low values. The frequency of pregnancies is a significant determinantof diabetes in women.

The age value of 1 indicates a positive association between age and diabetes prediction, suggesting that as people age, their chances of developing diabetes may also rise. Age is a known risk factor for diabetes, and this correlation aligns with the understanding that the risk tends to rise with advancing age. The BMI value of 0.34

represents a measure of body fat based on an individual's heightand weight. Higher BMI values are associated with a higher risk of developing diabetes due to increasedinsulin resistance and unstable glucose metabolism.

The HbA1c level value of 0.4 demonstrates a positive correlation between HbA1c levels and diabetes prediction. A higher risk of developing diabetes is associated with poor long-term glucose management, as shown by higher HbA1c values. HbA1c is widely recognized as a useful marker for evaluating and tracking long-term blood glucose management, and its clinical importance is consistentwith its correlation with diabetes prediction.

Blood glucose levels with a value of 0.42 show a positive correlation with the diagnosis of diabetes, supporting the idea that higher blood glucose levels are associated with an increased risk of developing the condition, as elevated blood glucose levels are a key component of diabetes.

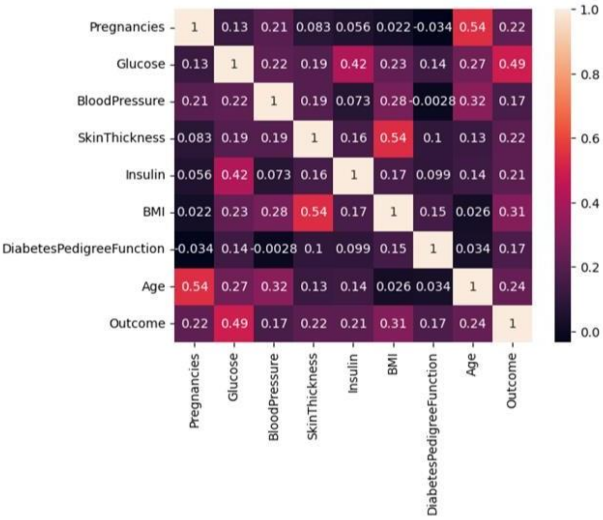


Fig.7 *Heat map correlation for PIMA Dataset*

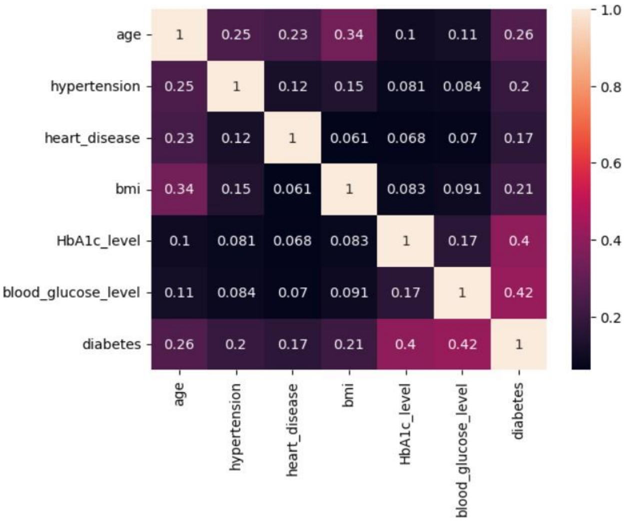


Fig. 8 *Heat map correlation for Kaggle Dataset*

3.5 Comparison of Performance Matric of PIMA and Kaggle Dataset

The following figures show the performance metrics for accuracy, precision, recall and F1-score for all three classification methods. Overall, Fig. 9 (left) shows the highest accuracy for diabetes detection is using linear regression method and Support Vector Machine for PIMA Dataset. While for Kaggle dataset in Fig. 9 (right), all performance metrics achieve almost 100% for linear regression method. Both Support Vector Machine and k-NN methods achieve more than 90% accuracy and precision.

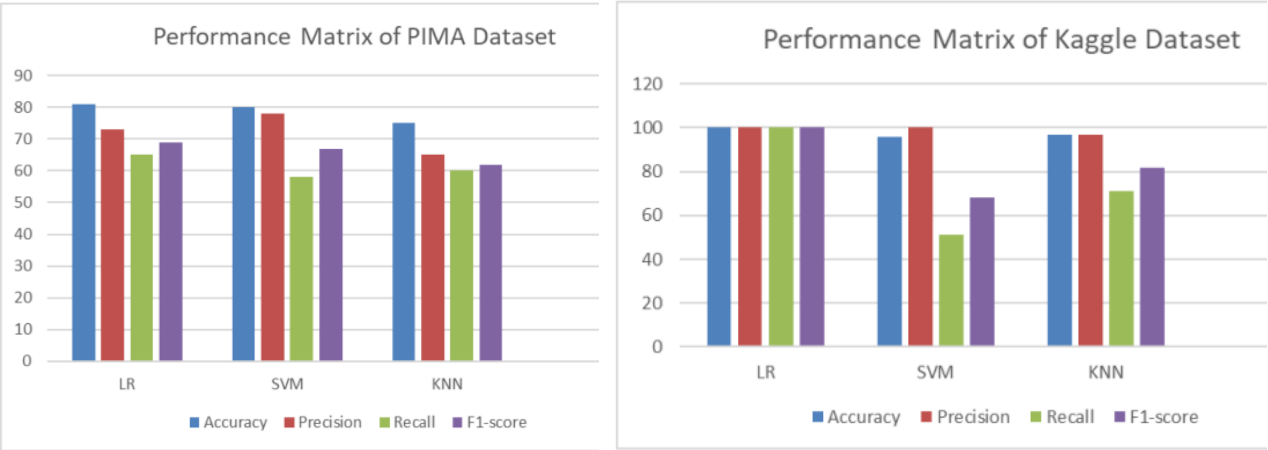
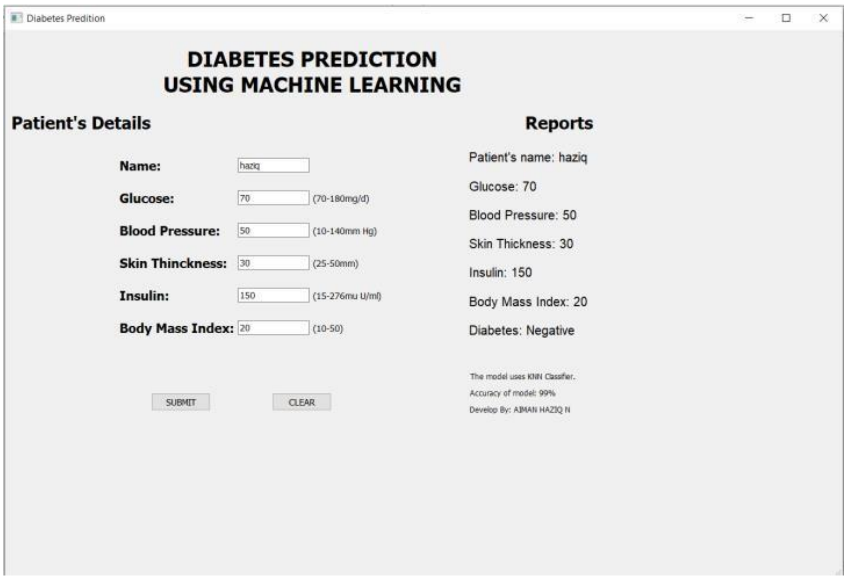


Fig. 9 *Performance of Machine Learning Algorithms for PIMA (left) and Kaggle (right) Dataset*

4. Computer Aided Interface for Diabetes Prediction

Fig. 10 shows the final step in the work involves creating a user interface for the model, which inputs hidden data for the model to read. The model predicts the onset of diabetes based on medical details collected. In Fig. 10 shows when the user enters the necessary information, thetrained model uses this information to predict whether the user has diabetes. The model achieves a 60% accuracy test rate, making it a reliable and accurate tool for assessing a person's risk of developing diabetes. The accuracy of overall diabetes prediction is 0.66 and for result for diabetes prediction system is normal.



DIABETES PREDICTION
USING MACHINE LEARNING

Patient's Details

Name: haziq

Glucose: 70 (70-180mg/d)

Blood Pressure: 50 (10-140mm Hg)

Skin Thinckness: 30 (25-50mm)

Insulin: 150 (15-276mu U/ml)

Body Mass Index: 20 (10-50)

Reports

Patient's name: haziq

Glucose: 70

Blood Pressure: 50

Skin Thickness: 30

Insulin: 150

Body Mass Index: 20

Diabetes: Negative

The model uses KNN Classfier.

Accuracy of model: 99%

Develop By: AIMAN HAZIQ N

Fig.10 *Result for Diabetes Prediction*

5. Conclusion

The study analyzed the performance of various machine learning algorithms on the Pima and Kaggle datasets to predict the presence of diabetes. Logistic Regression achieved an accuracy of 81%,with precision, recall, and F1-score values of 73%, 65%, and 69%, respectively. SVM performed slightly lower with an accuracy of 80%, precision of 78%, recall of 58%, and F1-score of 67%. K- nearest neighbors achieved 75% accuracy, with precision, recall, and F1-score values of 65%, 60%, and62%, respectively.

On the Kaggle dataset, Logistic Regression demonstrated excellent performance, achieving 100% accuracy, precision, recall, and F1-score. SVM achieved 96% accuracy, but a lower recall of 51% andan F1-score of 68%. K-nearest neighbors achieved 97% accuracy, with precision, recall, and F1-score values of 97%, 71%, and 82%, respectively.

The work aimed to develop an accurate prediction model and enhance it by creating a GUI simulation for a user-friendly experience. Various machine learning algorithms, including Logistic Regression, SVM, and K-nearest neighbors, were employed to achieve the objective. The models weretrained and tested using relevant datasets, with accuracy as the primary evaluation metric.

In conclusion, this work successfully achieved its objectives of diabetes prediction using machine learning techniques, developing a graphical user interface simulation, and evaluating the prediction model basedon accuracy. The work contributes to the field of healthcare by providing a reliable and user- friendly tool for early diabetes detection, potentially leading to improved patient care and management.

Acknowledgement

The authors would like to thank the Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia for their support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The author attests to having sole responsibility for the following: planning and designing the study, data collection, analysis and interpretation of the outcomes, and paper writing.

References

[1] Soni, M, and Varma, "Diabetes prediction using Machine Learning Technique" International Journal of Engineering Research and Technology, October 4, 2020. [Online]. Available: https://www.ijert.org/diabetes-prediction-using-machine-learning- techniques. [AccessedOctober 4, 2020]

[2] The Star News. Available: https://www.ijert.org/diabetes-prediction-using-machine-learning- techniques.

[3] Institute for Public Health (IPH), National Institutes of Health, Ministry of Health Malaysia 2020 "Communicable Disease: Risk Factors and other Health Problem" National Health and Morbidity Survey (NHMS) 2019, Vol.I

[4] Harris, Oii YB, Lee JS and Matanium P, "Non-communicable disease among low income adultsin rural coastal communities in Eastern Sabah, Malaysia" Malaysia BMC Public Health, 19(S4). https://doi.org/10.1186/s12889-019-6854-6

[5] Samsudin S, Abdullah N, Applanaidu SD. International Journal of Public Health Research. 2016 Jan 1; 6(2):741–9.

[6] Nr.(2019). Numpy,pandas and scikit learn explained. Medium,from https://medium.com/personal-project/numpy-pandas-and-scikit-learn-explained- e7336baecedc

[7] Khare,A.D.(2022). Diabetes dataset. Kaggle,from https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset Retrieved January26,2023

[8] Mustafa,M.(2023).Diabetespredictiondataset.Kaggle. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[9] Sarwar, A., Ali, M., Manhas, J., & Sharma, V. (2018). "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model". International Journal of Information Technology, https://doi.org/10.1007/s41870-018-0270-5 12(2), 419–428.

[10] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic RegressionJohn Wiley & Sons (Vol. 398).

[11] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). "Support VectorMachines. IEEE Intelligent Systems and their applications", 13(4), 18-28.

[12] Mitushi Soni, Dr. Sunita Varma. "Diabetes Prediction using Machine Learning Techniques". INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT). 2020. Vol 09, Issue 09.