University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 7, Issue 2
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Evaluating Diabetes Detection Methods: A Multilinear Regression Approach vs. Other Machine Learning Classifiers

Hasnain Hyder[1], Khawaja Haider Ali[1], Dr. Abdul Aziz[1], Lubina Iram[1]

[1]Department of Electrical Engineering, Sukkur IBA University, Sukkur, Pakistan
engr.hasnainhyder@gmail.com, haiderali@iba-suk.edu.pk, aziz.memon@iba-suk.edu.pk and Lubinairam.mec17@iba-suk.edu.pk

Abstract: Machine learning has become an important tool in many fields, including healthcare. In this research paper, we aim to implement diabetes dataset in multi-linear regression and compare its performance with different classifiers of machine learning. The novelty of this research lies in the evaluation of the diabetes dataset using multilinear regression and subsequent comparison of its performance against several other classifiers, including Decision Trees (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Support Vector Machines (SVM). There is not much research on using Multi-linear regression to find diabetes, so it is important to check how well it works with diabetes data. Introducing Multi-linear regression to the analysis and measuring its success against other recognized machine learning classifiers will shed light on its suitability for diabetes detection. Our results show that multi-linear regression achieved an accuracy of 80.5%, However, other classifiers such as random forest, and logistic regression outperformed linear regression, achieving accuracy scores of 81.4% and 81.25%, respectively. Furthermore, we observed that decision tree, KNN, and SVM, which are often used for classification tasks, did not perform well on this dataset, achieving an accuracy of only 78.7%, 80.5%, and 79.6% respectively. This suggests that the model's performance can be greatly impacted by the classifier selection. Our findings suggest that linear regression can be used for predicting diabetes, other classifiers such as random forest, and logistic regression are more effective for this dataset. To choose the best classifier for a given job, it is crucial to assess and contrast the performance of several classifiers.

Keywords: Multi-Linear Regression; Diabetes Dataset; Logistic Regression (LR); Support Vector Machine (SVM); K-Nearest Neighbour (KNN); Decision Tree (DT);

I. INTRODUCTION

Diabetes is a widespread chronic disease affecting millions of individuals globally, and its early diagnosis can significantly improve patient outcomes [1-2]. Therefore, developing accurate prediction models for diabetes is crucial for effective treatment and management of the disease [3-5]. Diabetes comes in two primary forms: Type 1, typically diagnosed during childhood, often involves immune-related mechanisms. On the other hand, Type 2 diabetes tends to develop later in life, particularly as individuals age, and is often associated with pancreatic diseases [6].

In 2014, 8.5% of persons over the age of 18 had blood glucose levels affected by diabetes, a chronic illness. Almost half of the 1.5 million deaths that were directly caused by it in 2019 occurred in those under the age of 70. Diabetes also contributed to 460,000 fatalities from kidney illness, and high blood glucose levels were associated with 20% of deaths from heart and blood valve issues. [7].

Diabetes-related mortality rates increased by 3% after adjusting for age from 2000 to 2019. The increase was especially high in lower-middle-income countries, where diabetes caused 13% more deaths. [8].

In this research paper, we aim to implement the diabetes dataset on multi-linear regression and compare the performance of diverse machine learning classifiers. To accomplish this, we employ the diabetes dataset sourced from Kaggle, a widely recognized platform for data science competitions. By analyzing extensive data, our study aims to determine the effectiveness of various predictive models. Ultimately, our findings will empower healthcare professionals to make informed decisions when diagnosing and treating diabetes, leading to improved care for patients.

Our goal is to use multilinear regression to forecast diabetes and contrast it with other common classifiers such as LR, DT, SVM, and RF. We evaluate the performance of each classifier using metrics such as accuracy, precision, recall, and F1 score. We seek to identify the best classifier for diabetes forecasting and comprehend the factors that affect the performance of different classifiers.

The paper is divided into following different sections. In Section 1, we survey the previous work on machine learning for diabetes forecasting. In Section 3, provides the detailed analysis of previous reseach, In Section 3, we explain the data source, the data preparation, and the classifier training

methods used in this study. In Section 4, we report our findings and compare the accuracy of different classifiers. In Section 5, we discuss the significance of our results and summarize the paper.

II. LITERATURE REVIEW

This section provide a detailed analysis of prior research focused on diabetes detection. It comprises of a detailed survey of the scholarly work undertaken in this domain, offering an extensive summary of the methodologies, findings, and conclusions that constitute the current academic landscape regarding diabetes identification.

Smith et al. [9], conduct a research aiming to predict diabetes using multiple linear regression models while placing emphasis on feature selection techniques. The authors explore various methods to identify pertinent features from the diabetes dataset, with the goal of enhancing the predictive performance of the multiple linear regression model. According to the study's findings, their constructed model successfully predicts diabetes with an accuracy rate of 78%.

Similarly, Deepti et al. [10], developed a model to estimate the likelihood of diabetes in patients using machine learning classifiers. They employed the Pima Indians Diabetes Database and evaluated three classifiers (DT, SVM, and Naive Bayes). They discovered that Naive Bayes achieved the highest accuracy of 76.30%. They also assessed the performance of the classifiers using Precision, F-Measure, and Recall. They validated their results using ROC curves.

Furthurmore, Priyanka et al. [11], created a logistic regression diabetes prediction model and investigated methods to improve its accuracy and performance. They made use of two datasets, PIMA Indians Diabetes and Vanderbilt, and used feature selection and ensemble methods. They achieved the highest accuracy of 78% for Dataset 1 and 93% for Dataset 2, using ensemble techniques. The research highlighted the importance of data preprocessing, feature selection, and ensemble methods in improving model accuracy and speed. Logistic regression was recognized as an effective algorithm for developing prediction models for diabetes analysis.

Sihao Wang et al. [12], explores the application of the LASSO (Least Absolute Shrinkage and Selection Operator) regression model in predicting diabetes incidence. The study outlines the methodology of data collection and preprocessing, feature selection using LASSO regression, and the construction and training of the LASSO regression model. It emphasizes the model’s ability to deal with multicollinearity and overfitting by shrinking certain coefficients to zero, thus simplifying the model and focusing on significant predictors.

Boon Feng Wee et al. [13], presents the application of machine learning (ML) and deep learning (DL) techniques for identifying diabetes. The study reviews a range of ML models including SVM, decision tree, random forest, k-nearest neighbor, and logistic regression. In addition, DL methods such as Convolutional Neural Network (CNN), Deep Belief Network (DBN), and Deep Neural Network (DNN) are utilized to enhance the accuracy and efficiency of diabetes identification. Additionally, the paper discusses the use of oversampling techniques and feature selection to enhance the performance of diabetes detection systems. It concludes by highlighting the potential future direction of employing advanced feature selection techniques to further improve the reliability and accuracy of diabetes detection models.

Adel Al-Zebari et al. [14], focuses on evaluating various machine learning methods for detecting diabetes. The study uses the Pima Indian Diabetes Dataset (PIDD) and employs multiple classifiers including decision tree, logistic regression, discriminant analysis, support vector machines, k-nearest neighbors, and ensemble learners. All methods were implemented using MATLAB Classification Learner Tool. The study concludes that Logistic Regression was the most effective model that acieves the highest accuracy that is 77.9%. The author suggested that the future work may includes applying deep learning techniques and advanced feature selection methods to enhance classification accuracy.

Turki Alghamdi et al. [15], prsents a study in which the author used data mining and machine learning techniques to predict diabetes and its complications effectively. This study highlights the application of various computational intelligence techniques like decision trees, logistic regression, support vector machines, neural networks, and particularly the XGBoost classifier, which demonstrated a notable accuracy rate of 89%. This research underscores the potential of computational intelligence in transforming healthcare approaches towards diabetes, emphasizing early diagnosis and personalized treatment plans to mitigate the disease's impact.

Md Shahin Ali el al. [16], discusses a machine learning approach to improve diabetes detection through optimal parameter selection and feature engineering. The study introduces a fine-tuned Random Forest algorithm with best parameters (RFWBP), which incorporates feature engineering techniques to enhance early diabetes detection. Several data processing and mining techniques were applied to enrich the primary dataset used for training multiple machine learning models, including AdaBoost, SVM, logistic regression, and more. The proposed RFWBP achieved impressive accuracy rates of 95.83% with 5-fold cross-validation and 90.68% without, outperforming traditional machine learning methods. The study underscores the significance of early and accurate diabetes detection to manage and mitigate the adverse impacts of the disease. Through rigorous testing and comparison with conventional methods, this research demonstrates the potential of machine learning in improving diagnostic processes for chronic conditions like diabetes.

III. AFTER REVIEWING THE LITERATURE RESEACH IT IS CLEAR THAT THERE IS NOT MUCH RESEARCH ON USING MULTI-LINEAR REGRESSION TO FIND DIABETES, SO IT IS IMPORTANT TO CHECK HOW WELL IT WORKS WITH DIABETES DATA. INTRODUCING MULTI-LINEAR REGRESSION TO THE ANALYSIS AND MEASURING ITS SUCCESS AGAINST OTHER RECOGNIZED MACHINE LEARNING CLASSIFIERS WILL SHED LIGHT ON ITS SUITABILITY FOR DIABETES DETECTION. THIS IS A KEY STEP BECAUSE IT FILLS A GAP IN THE RESEARCH AND HELPS US UNDERSTAND THE PROS AND CONS OF MULTI-LINEAR REGRESSION COMPARED TO OTHER METHODS. THIS DETAILED STUDY WILL HELP US LEARN MORE ABOUT THE BEST WAYS TO DETECT DIABETES. METHODS

The study outlines its methodology in a step-by-step process as illustrated in Fig. 1. This structured approach not only improves the clarity and understandability of the analysis but also each step is designed to add to the overall soundness and reliability of the research findings.

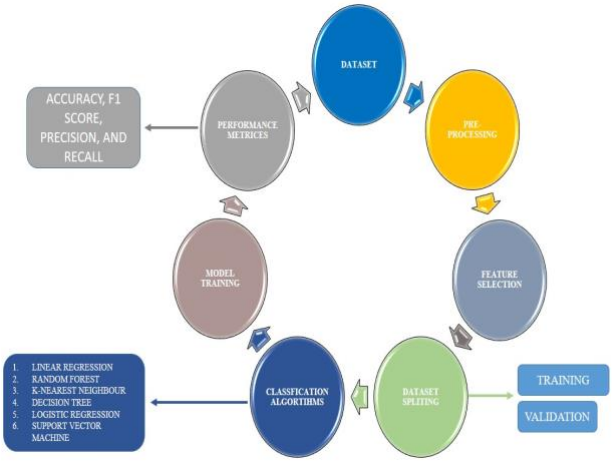

Figure 1 Proposed design flow for diabetes detection using various classifiers.

A. Data Collection

First, The dataset is taken from Kaggle website that comprises a total of 768 instances. Among these instances, 268 are labeled as positive, indicating the presence of diabetes, while 500 instances are labeled as negative, indicating the absence of diabetes. The dataset consists of 9 variables in total, with 8 input variables and 1 output variable. The table I displays the details of the first five instances, including their features and target values.

B. Data Pre-Processing

In the initial phase of our methodology, we address the need for data refinement, normalization, and preparation for subsequent analysis. This crucial step involves several processes, including the removal of missing values, the handling of outliers, and the conversion of categorical variables into numerical values. Firslty, we check the missing values in our dataset weather any missing values. The Fig. 2 shows the bar chart of missing values.

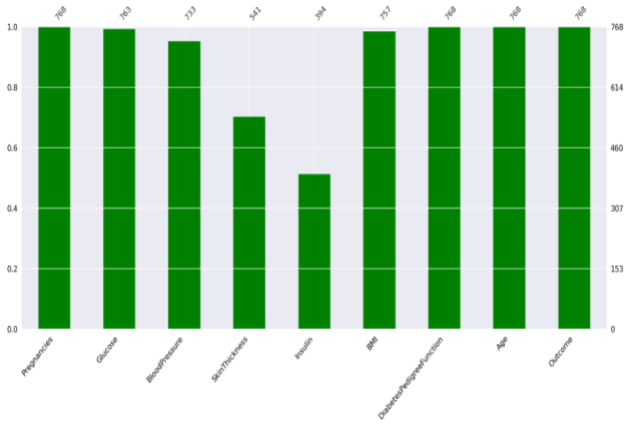

Figure 2 Details of a missing values in a diabeties dataset.

The Fig. 2 shows bar chart presents the count of missing values in a diabeties dataset. In the bar chart x-axis shows the variables included that are Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and Outcome. The vertical scale ranging from 0 to 768 shows the number of missing values. In bar chart the variables such as Pregnancies, Glucose, BloodPressure, BMI, DiabetesPedigreeFunction, Age, and Outcome—reach up to the chart maximum height, indicating they have almost no missing data. However, the bar for SkinThickness shows a moderate amount of missing data, roughly half of the maximum count, while Insulin has a noticeably lower count of missing values but more than the others. This visualization highlights which variables are most impacted by missing data and may require strategies for data imputation or removal to ensure robust data analysis. After analyzing the missing values, we filled the missing values. The Fig. 3 show the result after filling missing values in our dataset. This shows that now we do not have missing value in our dataset.

```
Pregnancies                 False
Glucose                     False
BloodPressure               False
SkinThickness               False
Insulin                     False
BMI                         False
DiabetesPedigreeFunction    False
Age                         False
Outcome                     False
dtype: bool
```

Figure 3 Details of missing values in our dataset

After conducting an assessment for missing values within our dataset, we proceeded to investigate the presence of outliers. Upon inspection, it was evident that outliers were present within the dataset. To enhance the dataset's robustness and ensure the reliability of our analysis, we employed outlier removal techniques. By eliminating these outliers, we aimed to refine the dataset's integrity and

enhance the accuracy of our subsequent analyses and model training processes.

C. Feature Selection

Following the data pre-processing stage, the focus shifts to feature selection. In this step, a meticulous approach is employed to identify and choose relevant features from the preprocessed data. In order to improve the model accuracy and interpretability, this selection process is essential. For feature selection firstly calculate the weight of the feature that highlight how much which feature is most important. The Fig. 4 shows a bar chart illustrating the relative importance of various features in a diabetes dataset. The chart is designed to show which factors are most predictive of diabetes outcomes based on the data analyzed. At the top of the importance scale is glucose, with a value close to 0.175, indicating that glucose levels are a critical predictor of diabetes. Following closely is insulin, suggesting its significant role in diabetes management and as a predictor of the condition. Body Mass Index (BMI) also features prominently, underscoring the connection between body weight and diabetes risk.

Age is another factor considered, with a moderate importance, reflecting its role in increasing diabetes risk as it advances. The Diabetes Pedigree Function, which gauges genetic predispositions to diabetes, also shows a notable level of importance, though less than the aforementioned factors. Skin thickness and blood pressure appear further down the scale, suggesting they have a lesser, yet still measurable, impact on diabetes prediction. The least important feature, according to the chart, is the number of pregnancies, which holds some relevance, particularly in the context of gestational diabetes, but is less critical compared to other factors.

After evaluating the feature importance, it was determined that glucose, insulin, BMI, age, diabetes pedigree function, and skin thickness are pivotal parameters. Consequently, these six features were selected for model training, ensuring a comprehensive consideration of factors crucial to diabetes prediction.

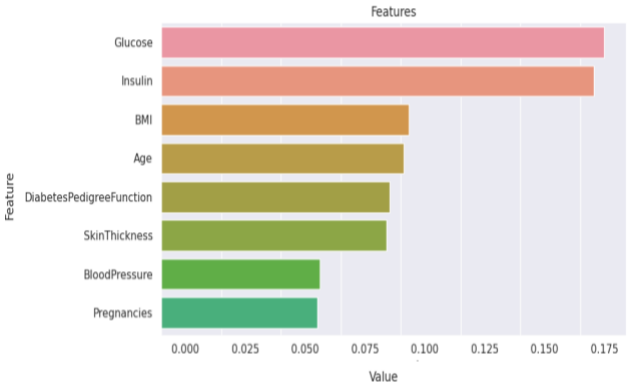

Figure 4 Relative importance of various features in diabetes prediction

In order to further find the relationship between the features, we create a correlation heatmap for various factors associated with diabetes, using a color scale from green to red to indicate the strength and direction of correlations between variables. The colors reflect the correlation strength with green indicating positive and red indicating negative correlations. The values range from 1, which signifies a perfect positive correlation, to -1, indicating a perfect negative correlation, with values around 0 showing little to no correlation.

The heat map as shown in Fig.5 highlights several relationships such as a strong positive correlation of 0.54 between age and the number of pregnancies, suggesting that older women tend to have had more pregnancies. There is also a notable positive correlation of 0.55 between skin thickness and BMI, indicating that higher BMI may be associated with greater skin thickness. Glucose levels demonstrate moderate positive correlations with insulin and BMI, with coefficients of 0.36 and 0.23 respectively, suggesting that higher glucose levels might be linked to higher insulin levels and a higher body mass index.

This visualization is crucial for understanding the interrelationships among different physiological factors in the context of diabetes, which can aid in medical research and the development of treatment strategies.

Table I: DETAILS OF THE FIRST FIVE INSTANCES OF DIABETES DATASET

Pregnancies	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**	**Outcome**
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

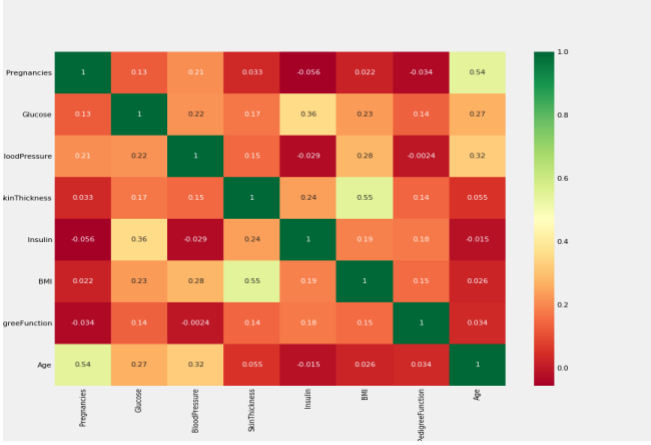

Figure 5 Correlation heatmap of various factors associated with diabetes

D. Classifications Algorithms

Upon the completion of dataset splitting, the subsequent phase involves the strategic selection of models for analysis. In this time, a diverse classifiers will be employed, encompassing methodologies such as LR, RF, KNN, DT, LR, SVM, among others. The primary objective is to train these models comprehensively using the diabetes disease dataset and subsequently evaluate their performance.

E. Dataset Splitting

The diabetes dataset was divided into two parts: a testing set and a training set. The training set has 428 instances and helps the model learn from the data. The testing set has 108 instances and tests the model's performance and generalization. This way of splitting the data is essential to measure how well the model applies its learned knowledge to new data.

$$Y=b0+b1X1 +e \quad (1)$$

In regression, the intercept (b0) is the value of the outcome variable (Y) when the predictor variable (X) is zero. The slope (b1) indicates how the outcome variable varies with the predictor variable. The error term (e) is the discrepancy between the observed and estimated values of the outcome variable, which represents the variation in the data that the regression equation fails to account for.The regression is classified into two parts as follows:

i. Linear Regression:

Linear regression is a statistical method used to model the relationship between two or more variables. It assumes a linear relationship between the dependent variable (the one you want to predict) and one or more independent variables (the ones used to make the prediction).

ii. Multi-Linear Regression:

The utilization of various classifiers serves a dual purpose: it allows for a thorough examination of how well each model handles the intricacies of the dataset, and it facilitates the assessment of their efficacy in predicting diabetes. By employing a range of classifiers, our goal is to learn more about the advantages and disadvantages of each approach, providing a comprehensive understanding of their applicability in the context of diabetes detection. This diversified approach ensures a robust evaluation and comparison of the models, paving the way for informed decisions on the most suitable model for our specific dataset. The following model will be used in order to gauge the performance of different models.

- Regression: Regression is a method used to nderstand how independent variables relate to a dependent variable. It is commonly employed in machine learning to forecast continuous values [17]. The objective of regression is to determine the values of the dependent variable, also called the outcome, based on a set of independent variables, also known as features [17]. For instance, regression can be used to predict weather condition or the likelihood of a disease occurrence [18]. The general mathematical equation for regression is given in Eq. (1).

  Multilinear regression is a technique to forecast a variable that is influenced by two or more other variables. In contrast to simple linear regression, which has one predictor variable and one outcome variable, multilinear regression has more than one predictor variable along with the outcome variable. The formula for multilinear regression is given in Eq. (2).

$$Y = b0 + b1X1 + b2X2 + ... + bnXn + e \quad (2)$$

  where, Y is the variable that depends on other variables, X1, X2, ..., Xn are the variables that affect Y, b0, b1, b2, ..., bn are the values that show how much each variable changes Y, and e is the difference between the actual and predicted values of Y.

- Random Forest: This approach is a versatile and straightforward algorithm commonly employed for both classification and regression purposes [19] [20]. It utilizes multiple individual decision trees working collectively as a unified model. Each tree categorizes instances into specific classes, and the predicted class is determined by the one with the highest number of votes [19].
- K-Nearest Neighbour (KNN): It is a basic supervised machine learning technique that operates by locating the k nearest points from the training set to a specific input point. [21-22]. The predicted class of the new input data point is determined by

considering the majority vote of the classes among its k nearest neighbors [21].

- Decision Tree: A decision tree is a versatile supervised learning technique that creates forecasts using a structure resembling a tree [23]. It may create a hierarchical decision process by examining the input features and addressing problems with regression and classification [23].
- Logistic Regression: One statistical method for binary classification issues is logistic regression. It uses independent variables to compute the probability of a result or a class's membership. It differs from linear regression, which predicts continuous outcomes, by modeling the connection between independent variables and a binary outcome [24-25].
- Support Vector Machine (SVM): This supervised machine learning method is mostly applied to categorization issues. It looks for a border in a high-dimensional space that best separates various classes. This boundary is formed by choosing important instances called support vectors. These support vectors are essential for defining the decision boundary [26-27].

F. Model Training

Once the classification algorithm is chosen, we will proceed to train the model and assess its performance on the diabetes dataset. In order to compare the performance of linear regression with other classifiers, we will iteratively train different models. This iterative process allows us to evaluate how well the multi-linear regression model performs in contrast to other classifiers when applied to the diabetes disease dataset. The following hyperparamerters are used while training the models:

- Multi-linear Regression: In multi-linear regression, we use default parameters that are automatically used when you create an instance of the LinearRegression class without explicitly specifying any parameters. These default parameters are the settings that the model relies on unless you specify otherwise. Random Forest: In Random Forest, we utilize the parameter n_estimators to define the number of trees in the forest. By setting it to 100, an ensemble comprising 100 decision trees is created. Additionally, we employ max_depth to establish the maximum depth of each decision tree, which is set to 5, thereby capping the depth at 5 levels. To ensure consistent results across different runs, we utilize random_state, fixing it at 42 to set the random seed for reproducibility.
- Logistic Regression: In logistic regression, we used the default parameters to train the model. We use as default penalty parameter that is L2 that shows the type of regularization applied. Max_iter specifies the maximum number of iterations for the solver to converge that is with a default of 100.
- K-Nearest Neighbour: In KNN, we use n_neighbors parameter that defines the number of neighbors to consider when making predictions. we set this parameter to 5, meaning the classifier will consider the labels of the five nearest neighbors. We use metric parameter that is the distance metric used for calculating the distance between points. Here, minkowski is employed, which is a generalization of other distance metrics like Euclidean distance and Manhattan distance. We also used p parameter that is used when the 'minkowski' metric is selected. It defines the power parameter for the Minkowski metric. When p is set to 2, it corresponds to using the Euclidean distance.
- Decision Tree: In this we use max_depth parameter that determines the maximum depth of the decision tree. we set this parameter to 5, meaning the tree will grow to a maximum depth of 5 levels.
- Support Vector Machine: In this we use kernel parameter that specifies the type of kernel used for the SVM. We use linear kernel that indicates a linear decision boundary. We use random_state parameter that sets the seed used by the random number generator. By fixing it to 0, it ensures reproducibility of results across different runs.

G. Performance Metrices

There are various evaluation measures that may be used to assess each classifier's performance. The following evaluation metrices are used in this research:

i. Accuracy: Accuracy is the most straightforward metric, measuring the proportion of correctly classified instances among all instances. The accuracy can be calculated using the Eq. (3).

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ predictions} \quad (3)$$

ii. Precision: Precision measures the proportion of true positive predictions among all positive predictions made by the classifier. The precision can be calculated using Eq. (4).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

iii. Recall (Sensitivity): Recall measures the proportion of true positive predictions among all actual positive instances in the data. The recall can be calculated using Eq. (5).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

iv. F1-score: F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. The F1 score can be calculated using Eq. (6)

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

IV. RESULTS AND DISCUSSION

The results of using different classifiers on a diabetic dataset are shown in this section. We will now discuss the findings in detail, highlighting the performance of each classifier on the given dataset.

A. Multi-Linear Regression (MLR)

Our main objective was to utilize a diabetes dataset and apply multi-linear regression to build a predictive model. We then proceeded to compare its performance against other classification algorithms. We trained a multi-linear regression model and tested it on a different dataset that was not used for training. The model had an accuracy of 80.5%. It also had an F1 score of 0.76, a precision of 0.78, and a recall of 0.73. The confusion matrix of the model is shown in Fig. 6.

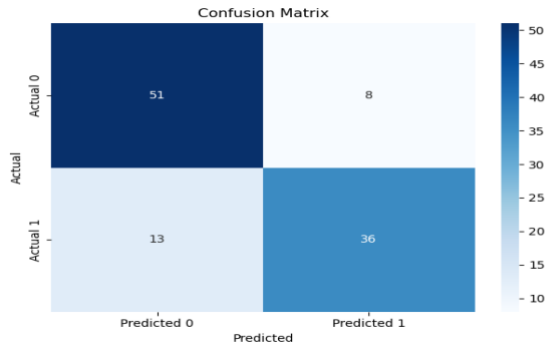

Figure 3 Representation of the confusion matrix illustrating the performance evaluation of a multi-linear regression model applied to the diabetes dataset.

B. Random Forest (RF)

Following the implementation of the diabetes dataset, we trained a random forest classifier as the next algorithm. The random forest model demonstrated a higher accuracy of approximately 81.4% on the unseen test data compared to the multi-linear regression. Moreover, the random forest classifier achieved an improved F1 score of 0.795, precision of 0.795, and recall of 0.795.This means that the random forest algorithm was more accurate and better at making predictions compared to the multi-linear regression model. The Fig. 7 demonstrates the confusion matrix of random forest.

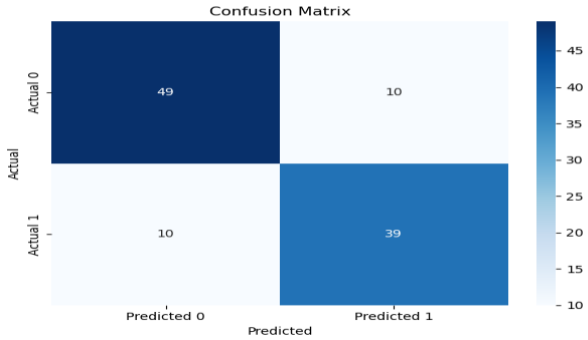

Figure 4 Representation of the confusion matrix illustrating the performance evaluation of a random forest model applied to the diabetes dataset.

C. Logistic Regression (LR)

Continuing with our classifier evaluation, we used logistic regression on the diabetes dataset. On the test dataset, the LR model outperformed the multi-linear regression, scoring 81.4% accuracy. Furthermore, the LR obtained an F1 score of 0.79, precision of 0.80, and recall of 0.85. These results demonstrate that the logistic regression algorithm outperformed the multi-linear regression model in all aspects. The logistic regression confusion matrix is shown in Fig. 8.

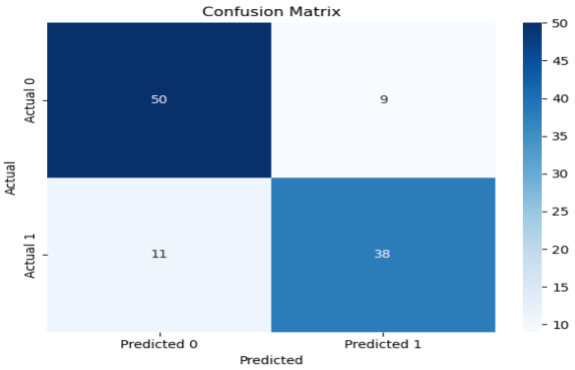

Figure 5 Representation of the confusion matrix illustrating the performance evaluation of a logistic regression model applied to the diabetes dataset.

D. K-Nearest Neighbour (KNN)

Another classifier we utilized for training the diabetes dataset was K-Nearest Neighbors (KNN). Interestingly, the KNN classifier exhibited similar performance to the multi-linear regression model. It achieved a test accuracy of 80.5%, along with an F1 score of 0.79, precision of 0.75, and recall of 0.86. These results indicate that the KNN classifier performed at a comparable level to the multi-linear regression model. The confusion matrix of a k-nearest neighbor is displayed in Fig. 9.

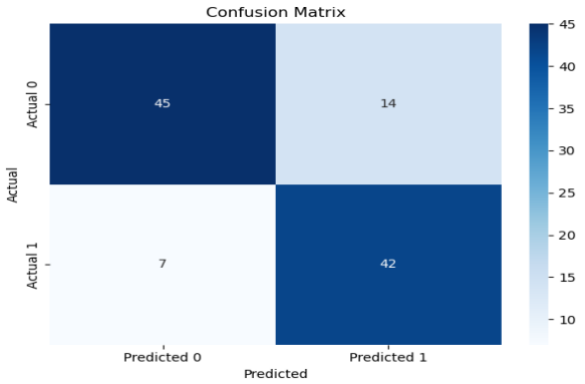

Figure 6 Representation of the confusion matrix illustrating the performance evaluation of a K-Nearest Neighbour model applied to the diabetes dataset.

E. Decision Tree (DT)

We applied the decision tree to the diabetes dataset as well. However, this model showed poor performance compared to the multi-linear regression. The decision tree classifier achieved 78.7% accuracy on the test data, with an F1 score of 0.76, precision of 0.78, and recall of 0.73. These metrics indicate that the decision tree model performed less well than multi-linear regression in appropriately identifying occurrences. Figure 10 depicts the confusion matrix of a decision tree.

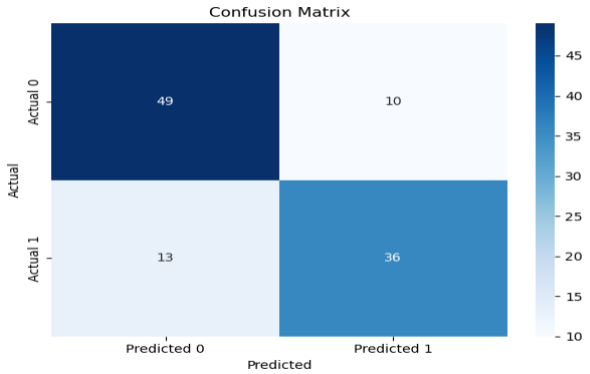

Figure 7 Representation of the confusion matrix illustrating the performance evaluation of a decision tree model applied to the diabetes dataset.

F. Support Vector Machine (SVM)

Finally, we performed training on a Support Vector Machine (SVM) using diabetes data. The SVM model exhibited an accuracy of 79.6%, which is comparable to the accuracy achieved by the multilinear regression model. Moreover, the F1 score, precision, and recall of the SVM model were 0.76, 0.81, and 0.71, respectively. The SVM confusion matrix is shown in Fig. 11.

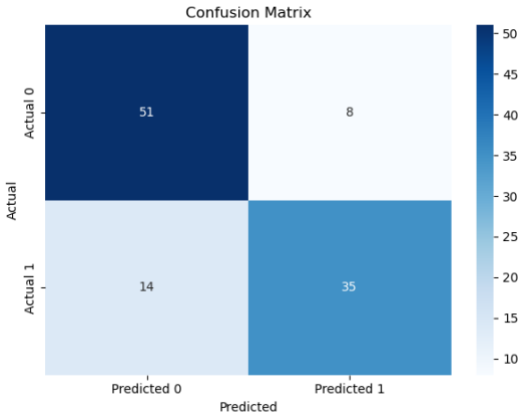

Figure 8 Representation of the confusion matrix illustrating the performance evaluation of a support vector machine model applied to the diabetes dataset.

G. Comparsion Table

A comparison of several classifiers and multilinear regression based on performance criteria, such as accuracy, F1 score, precision, and recall, is shown in table II. This comparison allows for an assessment of the classifiers' effectiveness in terms of their ability to accurately predict outcomes.

Table II: COMPARISON TABLE OF MODEL PERFORMANCE IN DIABETES DETECTION.

Classifier	Accuracy	F1 Score	Precision	Recall
MLR	80.5	0.76	0.78	0.73
RF	81.4	0.795	0.795	0.795
LR	81.4	0.79	0.80	0.85
KNN	80.5	79.9	0.75	0.86
DT	78.7	0.76	0.78	0.73
SVM	79.6	0.76	0.81	0.71

H. Enhancing Diabetes Prediction through Classifier Comparison and Analysis

i. Random Forest and Logistic Regression: Both RF and LR outperformed Multi-Linear Regression (MLR) in terms of accuracy, F1 score, precision, and recall. This could be attributed to their ability to capture non-linear relationships between features and the target variable. RF, being an ensemble method, combines multiple decision trees, which can handle complex interactions between features better than MLR. LR, on the other hand, is inherently suited for binary classification tasks like diabetes prediction and can learn complex decision boundaries.

ii. K-Nearest Neighbors: KNN exhibited similar performance to MLR, suggesting that its simple approach might not be as effective in capturing the underlying patterns in the dataset compared to more sophisticated algorithms like RF and LR. KNN relies heavily on the choice of the number of neighbors (k)

and the distance metric, which might not be optimal for this particular dataset.

iii. Decision Tree and Support Vector Machine: DT performed less well than MLR, indicating that it might have overfit the training data or failed to capture the underlying patterns effectively. SVM, while achieving comparable accuracy to MLR, had lower recall, suggesting that it might have misclassified some positive instances as negative. This could be due to the choice of hyperparameters or the nature of the dataset.

iv. Model Complexity and Generalization: The performance differences among classifiers highlight the importance of selecting appropriate models that balance complexity and generalization. More complex models like RF and LR might perform better on the test data but could be prone to overfitting if not regularized properly.

V. LIMITATIONS AND FUTURE WORK

This section outlines the limitations encountered during the course of this research. The study encountered several constraints, including:

- Imbalanced Class Distribution: The dataset exhibited an imbalanced distribution among the classes, potentially affecting the classifiers performance and result generalization.

Biases in Data Collection: Biases such as sampling bias or selection bias were identified in the data collection process. These biases may have ramifications for the generalizability of the study findings, as they could skew the representation of certain population segments.

- Dataset Size: Additionally, the size of the dataset presents a notable limitation, as it is relatively small. The limited number of instances may restrict the robustness of the models developed and could contribute to poorer performance in terms of predictive accuracy and generalization.

Deep Neural Network (DNN) can be used as a game-changer for improving diabetes prediction models. They are known for their precision in various classification and prediction tasks, DNNs could significantly boost the accuracy and strength of these models.

By analyzing diabetes data with DNNs, we can expect to see major strides in how accurately we can predict the condition. DNNs' ability to learn complex patterns through multiple layers means they can find hidden relationships in the data that simpler models might miss.

DNN are also incredibly flexible and can handle big, diverse datasets, which is often a challenge in medical research. They can automatically pick out important features from the data, saving time and possibly revealing new insights. They are complex and need a lot of computing power, which can make them hard to work with and understand. So, there is a push to make DNN more transparent and efficient, especially if they are going to be used in healthcare.

VI. CONCLUSION

The aim of this research paper was to apply multilinear regression to the diabetes dataset and compare its performance with various machine learning classifiers. The novelty of this research stems from the assessment of the diabetes dataset using multilinear regression and the subsequent comparison of its performance with several other classifiers, such as decision trees, random forests, K-nearest neighbors, logistic regression, and SVM . We discovered that linear regression obtained an accuracy of 80.5%, while other classifiers such as random forest and logistic regression surpassed linear regression, with accuracy scores of 81.4% and 81.25% respectively. Conversely, the decision tree, k-nearest neighbour, and SVM were less effective, obtaining accuracies of 78.7%, 80.5%, and 79.6% respectively.

These results indicate that the selection of classifier has a considerable influence on the predictive performance of the model. Although linear regression can be utilized for predicting diabetes, other classifiers such as random forest and logistic regression demonstrate greater efficiency for this specific dataset. To select the best classifier for a given job, it is therefore crucial to assess and contrast the performance of several classifiers.

REFERENCES

[1] G Chowdhury, M. M., Ayon, R. S., & Hossain, M. S. (2024). An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. Healthcare Analytics, 5, 100297.

[2] Alam, A., Dhoundiyal, S., Ahmad, N., & Rao, G. K. (2024). Unveiling diabetes: Categories, genetics, diagnostics, treatments, and future horizons. Current Diabetes Reviews, 20(4), 10-22.

[3] Franks, P. W., Cefalu, W. T., Dennis, J., Florez, J. C., Mathieu, C., Morton, R. W., ... & Stehouwer, C. D. (2023). Precision medicine for cardiometabolic disease: a framework for clinical translation. The Lancet Diabetes & Endocrinology, 11(11), 822-835.

[4] Syed Muhammad Nabeel Mustafa, Hassan Zaki, Syeda Sundus Zehra, & Muhammad Shoaib. (2022). Significance and Challenges of Big Data in Healthcare: A Review. University of Sindh Journal of Information and Communication Technology , 6(1), 25-30. Retrieved from https://sujo.usindh.edu.pk/index.php/USJICT/article/view/6265

[5] Abdul Hafeez Muhammad, & Amna Faisal. (2022). Integration of Artificial Intelligence and Human Computer Interaction in Healthcare. University of Sindh Journal of Information and Communication Technology , 6(3), 101-107. Retrieved from https://sujo.usindh.edu.pk/index.php/USJICT/article/view/6279

[6] ElSayed, N. A., Aleppo, G., Aroda, V. R., Bannuru, R. R., Brown, F. M., Bruemmer, D., ... & American Diabetes Association. (2023). 2. Classification and diagnosis of diabetes: standards of care in diabetes—2023. Diabetes care, 46(Supplement_1), S19-S40.

[7] Mumtaz, M. T., Khan, M. M. F., Uzair, M., Khan, M. A., Salman, A., & Khan, H. F. T. (2023). The Silent Killer: Investigating the Influence of Stress on Cardiovascular Health of Diabetic Patients. Pakistan Journal of Medical & Health Sciences, 17(05), 397-397.

[8] Liu, Y., Wang, D., Huang, X., Liang, R., Tu, Z., You, X., ... & Chen, W. (2023). Temporal trend and global burden of type 2 diabetes attributable to non-optimal temperature, 1990–2019: an analysis for the Global Burden of Disease Study 2019. Environmental Science and Pollution Research, 1-10.

[9] Alhussan, A. A., Abdelhamid, A. A., Towfek, S. K., Ibrahim, A., Eid, M. M., Khafaga, D. S., & Saraya, M. S. (2023). Classification of Diabetes Using Feature Selection and Hybrid Al-Biruni Earth Radius and Dipper Throated Optimization. Diagnostics, 13(12), 2038.

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

[11] Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032.

[12] Wang, S., Chen, Y., Cui, Z., Lin, L., & Zong, Y. (2024). Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model. Journal of Theory and Practice of Engineering Science, 4(01), 58-64.

[13] Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., & Juwono, F. H. (2024). Diabetes detection based on machine learning and deep learning approaches. Multimedia Tools and Applications, 83(8), 24153-24185.

[14] Al-Zebari, A., & Sengur, A. (2019, November). Performance comparison of machine learning techniques on diabetes disease detection. In 2019 1st international informatics and software engineering conference (UBMYK) (pp. 1-4). IEEE.

[15] Alghamdi, T. (2023). Prediction of diabetes complications using computational intelligence techniques. Applied Sciences, 13(5), 3030.

[16] Ali, M. S., Islam, M. K., Das, A. A., Duranta, D. U. S., Haque, M., & Rahman, M. H. (2023). A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: Machine learning insights. BioMed Research International, 2023.

[17] Kumari, K., & Yadav, S. (2018). Linear regression analysis study. Journal of the practice of Cardiovascular Sciences, 4(1), 33-36.

[18] Leung, P., & Tran, L. T. (2000). Predicting shrimp disease occurrence: artificial neural networks vs. logistic regression. Aquaculture, 187(1-2), 35-49.

[19] Brown, S. H. (2009). Multiple linear regression analysis: a matrix approach with MATLAB. Alabama Journal of Mathematics, 34, 1-3.

[20] Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. PloS one, 11(6), e0156571.

[21] Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. Briefings in Bioinformatics, 24(2), bbad002.

[22] Cunningham, P., & Delany, S. J. (2021). k-Nearest neighbour classifiers-A Tutorial. ACM computing surveys (CSUR), 54(6), 1-25.

[23] Syriopoulos, P. K., Kalampalikis, N. G., Kotsiantis, S. B., & Vrahatis, M. N. (2023). k NN Classification: a review. Annals of Mathematics and Artificial Intelligence, 1-33.

[24] Shehadeh, A., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in Construction, 129, 103827.

[25] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), 3-14.

[26] Loh, W. Y. (2023). Logistic regression tree analysis. In Springer handbook of engineering statistics (pp. 593-604). London: Springer London.

[27] Shahi, T. B., & Pant, A. K. (2018, February). Nepali news classification using Naive Bayes, support vector machines and neural networks. In 2018 international conference on communication information and computing technology (iccict) (pp. 1-5). IEEE.