Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

Using Machine Learning Techniques to Predict Type 2 Diabetes

Alzain Meftah Alzain
Information technology department
Faculty of education
Misurata university
Misurata, Libya
A.alzain@edu.misuratau.edu.ly

Hanan Meftah Al-Futaisi
Computer Science Department
Faculty of information Technology
Misurata University
Misurata, Libya
h.alftesi.pg@itmisuratau.edu.ly

Abstract

Diabetic patients face multiple risks of potential complications; therefore, early diagnosis is so important to avoid these consequences. In healthcare scientific research, the literature revealed that machine learning techniques are widely used to diagnose many diseases, including diabetes. This paper aims to predict diabetes using some machine-learning techniques. A dataset from Kaggle was used in the study, which is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. Random Forest, Logistic Regression, and Support Vector Machine techniques were used for classification. The results revealed that the highest obtained accuracy value was with the 0.831 algorithm Logistic Regression and Random Forest.

keywords— predict diabetes, machine learning, classification, data pre-processing

استخدام تقنيات التعلم الآلي للتنبؤ بمرض السكري من النوع الثاني

الزين مفتاح الزين
قسم تقنية المعلومات
كلية التربية-جامعة مصراتة

حنان مفتاح الفطيسي
قسم علوم الحاسوب
كلية تقنية المعلومات-جامعة مصراتة

الملخص:

يواجه مرضى السكري مخاطر متعددة من المضاعفات المحتملة؛ لذلك فإن التشخيص المبكر مهم جدًا لتجنب هذه العواقب. كشفت أدبيات البحث العلمي في مجال الرعاية الصحية أن تقنيات التعلم الآلي

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

تستخدم على نطاق واسع لتشخيص العديد من الأمراض، بما في ذلك مرض السكري. وتهدف هذه الورقة إلى التنبؤ بمرض السكري باستخدام تقنيات التعلم الآلي، حيث تم استخدام مجموعة بيانات من **Kaggle** في هذه الدراسة، والتي هي في الأصل من المعهد الوطني للسكري، وأمراض الجهاز الهضمي والكلى. كما تم استخدام تقنيات الغابة العشوائية والانحدار اللوجستي وآلة الدعم الموجه للتصنيف.

أظهرت النتائج أن أعلى قيمة دقة تم الحصول عليها كانت مع خوارزمية الانحدار اللوجستي والغابة *العشوائية* *0.831*.

الكلمات المفتاحية – التنبؤ بمرض السكري، والتعلم الآلي، والتصنيف، والمعالجة المسبقة للبيانات.

I.Introduction

In the last few years, the amount of available data has increased dramatically, posing a challenge in extracting valuable information and important patterns hidden within it. Moreover, the healthcare field is one of the hot research topics, which has a great amount of available data and requires more intensive research. The literature explains that machine learning techniques are widely used to reach this aim. Diabetes is a global health problem that affects millions of people around the world, and diagnosing and treating diabetes requires a precise understanding of the factors affecting the development of this disease and determining the appropriate treatment for each case. Diabetes is a disease with high levels of blood sugar, which occurs because the pancreas does not produce enough amount of insulin, or the body cannot effectively use it. According to the International Diabetes Federation statistics, one out of every 11 adults has diabetes, and every 6 seconds 1 person dies due to diabetes-related complications (Aguirre, F., Brown, A., Cho, N. H., Dahlquist, G., Dodd, S., Dunning, T., ... & Whiting, D. 2013).

The purpose of the research is to employ some machine-learning techniques (Random Forest, Logistic Regression, and Support Vector Machine) to predict diabetes in its early stage, which might contribute to the prevention of the dangerous complications related to this disease.

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

II.RELATED WORKS

in 2018, Quan Zou and others conducted research titled "Predicting Diabetes Mellitus with Machine Learning Techniques". The research aimed to predict diabetes with the highest accuracy. They used a dataset from Luzhou Hospital in China. This dataset contains 14 features of 86994 people. The authors also used a decision tree and random forest algorithms for prediction, and the results showed that random forest had the highest accuracy (0.8084) (Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. 2018).

In 2021, Ram and others conducted a study titled “Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches”. The study aimed to predict type 2 diabetes in Pima Indian women utilizing a logistic regression and decision tree, the authors argue that their model can be used to make a reasonable prediction of diabetes (type 2), and they reached a prediction accuracy of 78.26% (Joshi, R. D., & Dhakal, C. K. 2021).

One year after in (2022), Kırğıl conducted research titled " Predicting Diabetes Using Machine Learning Techniques" which aims to predict diabetes with high accuracy. in this research Pima Indian Diabetes dataset was used, as well as some classification algorithms which are respectively Decision Tree, Naïve Bayes, Support Vector Machine, Logistic Regression, Multilayer Perceptron, K Nearest Neighbor, Logistic Model Tree, and Random Forest. The results of this study concluded that the Random Forest algorithm scored the highest accuracy value (80.869) (Kirgil, E. N. H., Erkal, B., & Ayyildiz, T. E. 2022).

another study conducted in 2023 by Chun Zhu et al, titled “Predicting the onset of diabetes using machine learning methods”, the study aimed to predict diabetes with high accuracy and used a dataset for diabetic patients from the Taipei Municipal Medical Center. Where 15000 women aged between 20 and 80 participated. Moreover, the study investigated the impact of 8 different features including the number of pregnancies, body mass index, plasma glucose level, blood pressure, insulin level, sebum thickness, age, and diabetes pedigree function. The results showed

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

that the boosted decision tree algorithm obtained the highest accuracy (0.991). (Lugner, M., Rawshani, A., Helleryd, E., & Eliasson, B. 2024).

recently (2024), Moa Lugner and others published a paper titled "Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data", the study investigated the most predictive feature for the development of diabetes (type 2). In this research, 450,000 patients aged 40–69 participated, and the XGboost classification model was used. The results revealed that HbA1c is the first indicator, followed by body mass index, waist circumference, blood glucose, family history of diabetes, gamma-glutamyl transferase, waist-to-hip-ratio-, high-density lipoprotein cholesterol, age, and urine (Shah, K., Patel, H., Sanghvi, D., & Shah, M. 2020).

III.METHODOLOGY

A. Machine Learning Techniques

1. Logistic Regression (LR)

Logistic Regression is frequently used for linear classification this technique investigates the relationship between the dependent and independent variables in the dataset, it uses the binary variable (yes/no classification) to measure the result (Aroef, C., Rivan, Y., & Rustam, Z. 2020).

2. *Support Vector Machine (SVM)*

Support Vector Machine (SVM): the mechanism of this algorithm depends on determining a margin that separates the two different categories in the data. for this reason, it is used for classification and prediction (Saxena, R. 2021).

3. Random Forest (RF)

Random Forest: is a popular machine learning algorithm used for regression and classification depending on a set of decision trees, where several

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

independent decision trees are generated and their predictions are combined to perform the final classification (Shah, K., Patel, H., Sanghvi, D., & Shah, M. 2020).

4. *Gradient Boosting* (GB)

Gradient Boosting: It is also used for classification and prediction, and gradually builds a machine-learning model by overcoming and modifying the weaknesses of previous models (Fafalios, S., Charonyktakis, P., & Tsamardinos, I. 2020).

B. Dataset

This study used an open-access dataset "Pima Indians Diabetes", which contains 8 features and 768 samples, all patients are female and above 20 years old. Table 1 explains dataset features, and Figure 1 shows the proposed framework of this study.

Table 1. Dataset Features

Features
Pregnancies
Glucose
Blood Pressure
BMI (Body Mass Index)
Skin Thickness
Diabetes Pedigree Function
Age
Insulin
Outcome

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

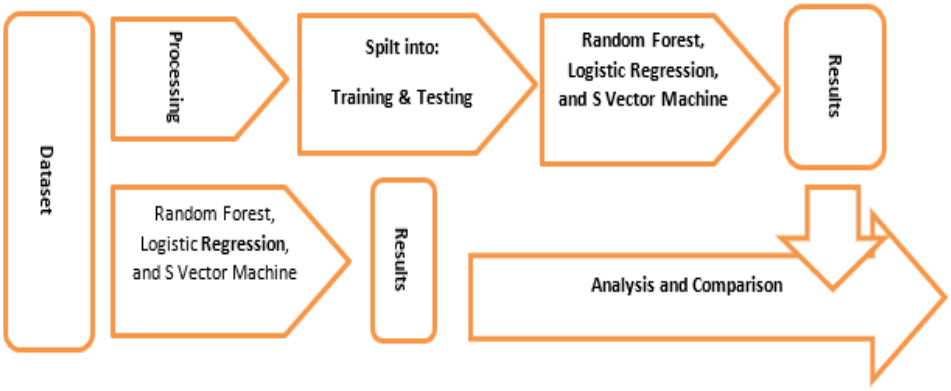

Fig 1. Proposed Framework

C. *Data Pre-processing*

To get valuable results the dataset has to be prepared to be processed later by machine learning techniques. That is why, many issues have been considered including missing values as well as the range of data, its mean, and standard deviation. In this study, the missing values were filled using the average value (Acock, A. C. 2005), the normalization technique (equation 1) was used to cast the data into the specific range, and standardization (equation 2) to make a data set with mean=0, and sd =1. (Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. 2014).

$$\text{Xnorm} = \frac{x - min}{Xmax - Xmin} \quad (1)$$

$$\text{Xstand} = \frac{x - mean\ (x)}{Deviation\ (x)} Standard \quad (2)$$

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

IV. RESULTS

A. performance metrics

Accuracy value calculated using equation (3)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Where:

TP: TRUE POSITIVE RATE

TP: TRUE POSITIVE RATE

TN: True Negative Rate

FP: False Positive Rate

FN: False Negative Rate

B. Experimental Results

To decide whether or not patients had diabetes, four machine-learning classifiers were utilized which are Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting. The dataset was split into 70 % and 30% for model training and testing. And then, the classifiers were run without any pre-processing of the dataset first. The results shown in Table (2), and Figures (2,3,4,5,6).

TABLE II. SUMMARY OF PERFORMANCE METRICS BEFORE PREPROCESSING

	Accuracy (%)	Recall (%)	Score (%) f1
Logistic Regression	.0741	0.555	0.581
SVM	0.720	0.536	0.536
Random Forest	0.766	0.558	0.571
Gradient Boosting	.0694	0.534	0.547

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

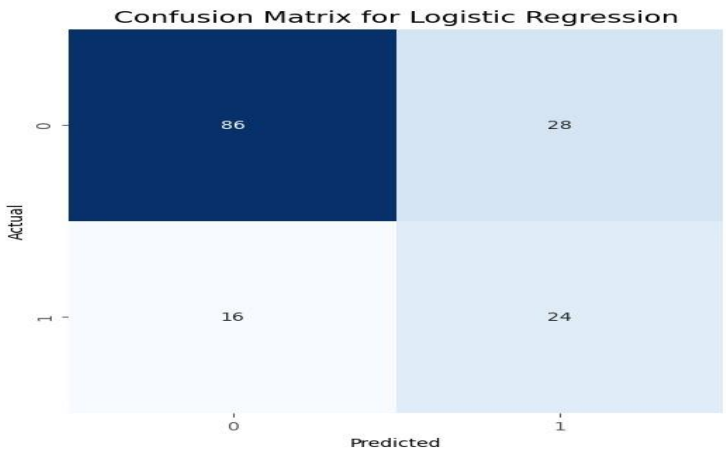

Fig. 2. Confusion matrix for logistic Regression

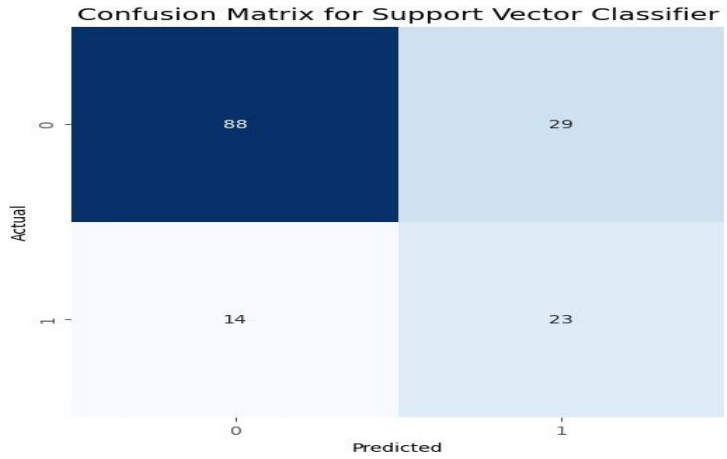

Fig. 3. Confusion matrix for SVM

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

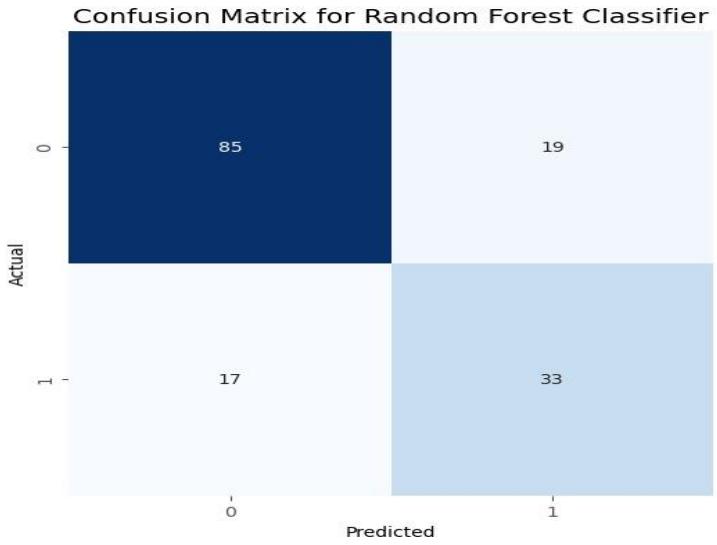

Fig. 4. Confusion matrix for Random Forest

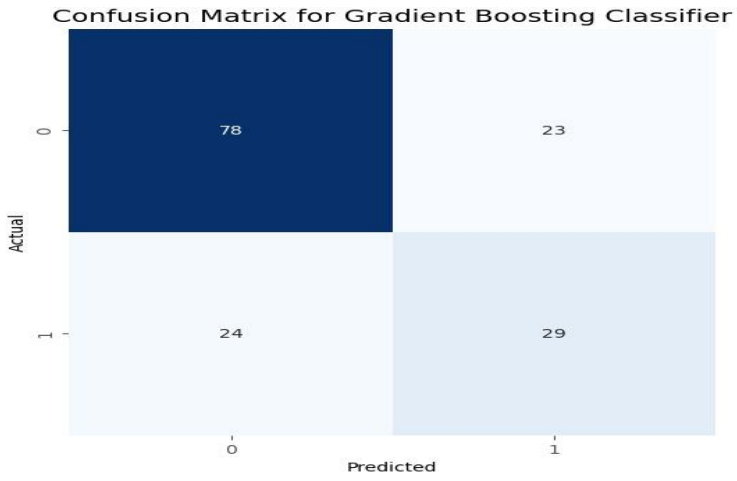

Fig. 5. Confusion matrix for Gradient Boosting

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

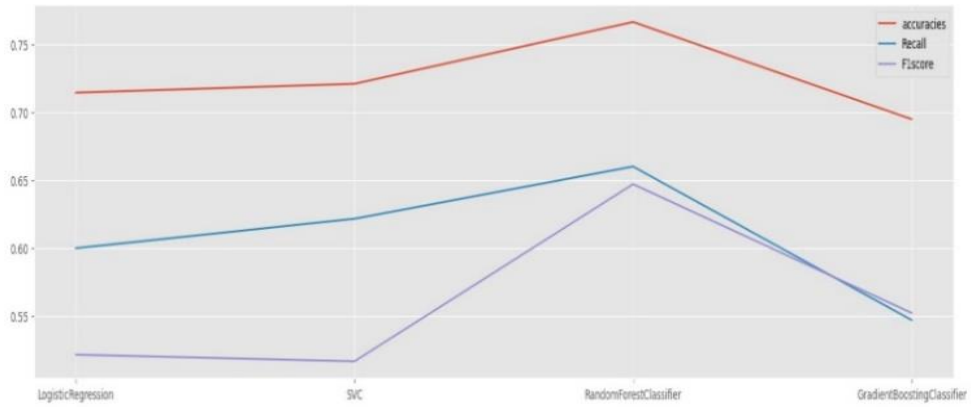

Fig. 6. Comparing Algorithms Accuracy Before Data Prepressing

Afterward, pre-processing methods (Normalization, standardization, and filling in missing data) were made on the dataset, and then the classifiers were run again. The obtained results are shown in Table (3), and Figures (7,8,9,10,11).

TABLE III. SUMMARY OF PERFORMANCE METRICS AFTER PREPROCESSING

	Accuracy (%)	Recall (%)	Score (%) f1
Logistic Regression	0.831	0.68	0.723
SVM	0.772	0.50	0.588
Random Forest	0.831	0.70	0.729
Gradient Boosting	0.798	0.72	0.699

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

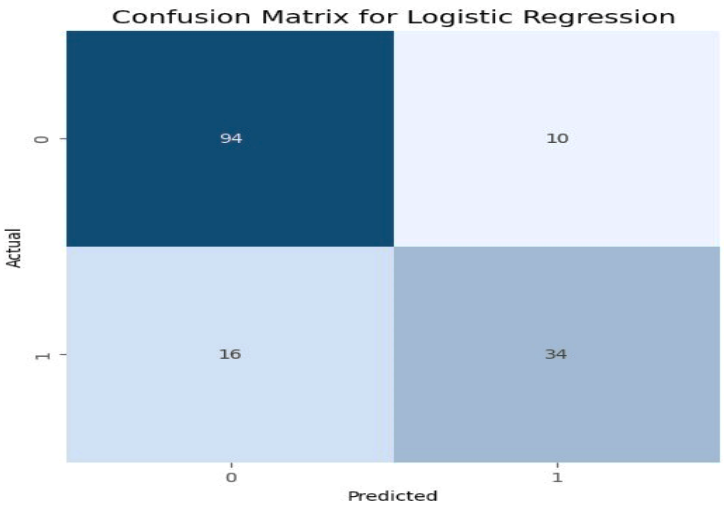

Fig. 7. Confusion matrix for logistic Regression

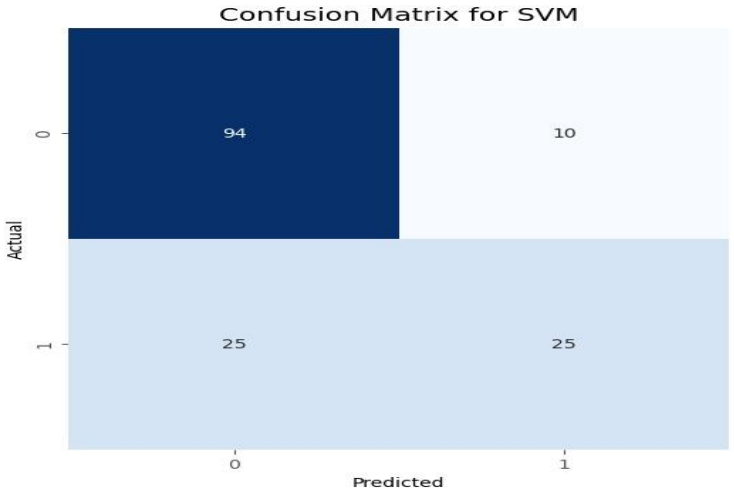

Fig. 8. Confusion matrix for SVM

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

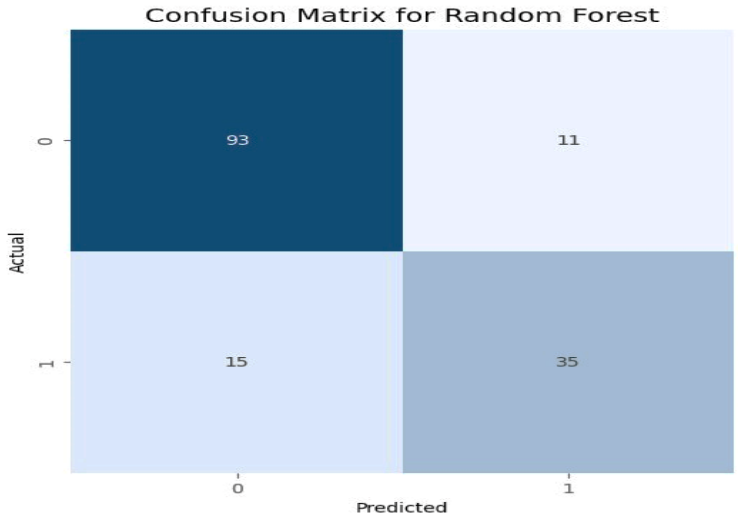

Fig. 9. Confusion matrix for Random Forest

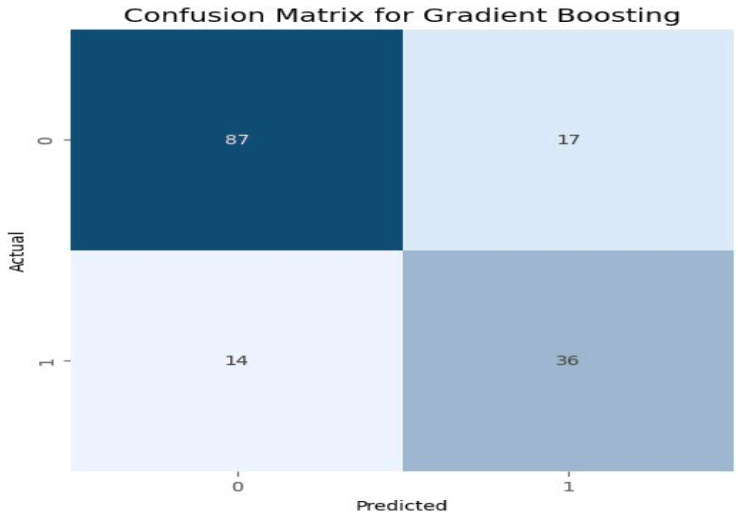

Fig. 10. Confusion matrix for Gradient Boosting

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

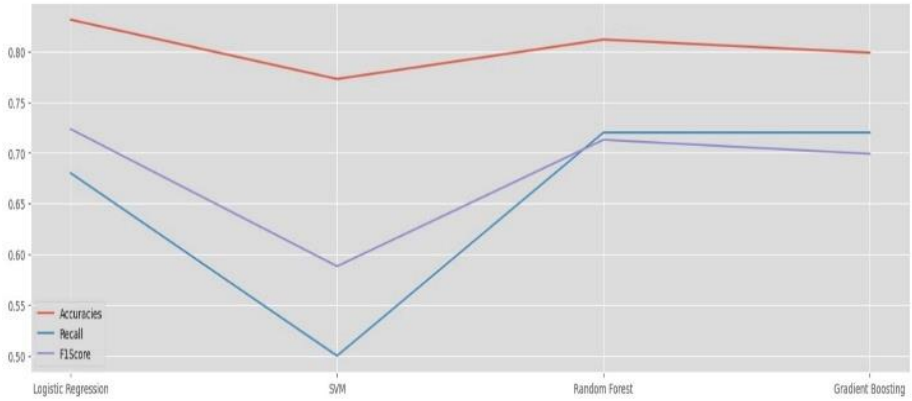

Fig. 11. Comparing Algorithms Accuracy After Data Prepressing

By comparing the values of obtained results before and after pre-processing, it is seen that there is an improvement. After conducting data pre-processing. The Logistic Regression and Random Forest classifiers achieved the best results with (0.831) followed by, Gradient Boosting and SVM with 0.798, 0.772.

The results of this study are consistent with the results of many related studies including Quan Zou (Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. 2018). and Kırğıl (Kirgil, E. N. H., Erkal, B., & Ayyildiz, T. E. 2022). In these studies, the Random Forest algorithm scored the highest accuracy value (about 80%). This prediction accuracy was higher than the value obtained by Ram in 2021 when he scored 78% prediction accuracy value (Joshi, R. D., & Dhakal, C. K. 2021). In contrast, in 2023 Chun Zhu obtained a higher accuracy value when he used the boosted decision tree algorithm (Lugner, M., Rawshani, A., Helleryd, E., & Eliasson, B. 2024). Another result of this study was that the prediction accuracy value was increased after conducting data pre-processing.

References

Aguirre, F., Brown, A., Cho, N. H., Dahlquist, G., Dodd, S., Dunning, T., ... & Whiting, D. (2013). IDF diabetes atlas.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, *9*, 515.

Scientific Journal of Faculty of Education, Misurata University-Libya, No25, Sep 2024
Published online in September
المجلة العلمية لكلية التربية، جامعة مصراتة، ليبيا -العدد الخامس والعشرون، سبتمبر 2024
Issn: 2710-4141
تاريخ الاستلام:2024/05/13 تاريخ النشر:2024/09/01

Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. *International journal of environmental research and public health*, *18*(14), 7346.

Kirgil, E. N. H., Erkal, B., & Ayyildiz, T. E. (2022). Predicting diabetes using machine learning techniques.

Lugner, M., Rawshani, A., Helleryd, E., & Eliasson, B. (2024). Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Scientific Reports*, *14*(1), 2102.

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, *5*(1), 12.

Aroef, C., Rivan, Y., & Rustam, Z. (2020). Comparing random forest and support vector machines for breast cancer classification. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *18*(2), 815-821.

Saxena, R. (2021). Role of K-nearest neighbour in detection of Diabetes Mellitus. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(10), 373-376.

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, *5*(1), 12.

Fafalios, S., Charonyktakis, P., & Tsamardinos, I. (2020). Gradient boosting trees. *Gnosis Data Analysis PC*, 1-3.

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and family*, *67*(4), 1012-1028.

Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, *1*(1), 1-6.