Journal of Industrial Engineering & Management
Vol. 6 No. 2
http://www.jiemar.org
e-ISSN : 2722-8878

Analysis to predict diabetes Using Data Mining

Abdulhalim Hamid [1], Yunifa Miftachul Arif [2] , M.Faisal[3]

Magister informatika, UIN Maulana Malik Ibrahim Malang, Jl. Gajayana No.50, Dinoyo, Kec. Lowokwaru, Kota Malang, Jawa Timur 65144
[1] *abdulhalimhamed3@gmail.com*
[2] yunif4@ti.uin-malang.ac.id
[3] mfaisal@it.uin-malang.ac.id

Abstract *Data mining is crucial for extracting patterns and valuable insights from extensive datasets, utilizing artificial intelligence and advanced data analysis techniques across various domains. Diabetes, a metabolic disorder characterized by elevated blood glucose levels, poses significant health risks, including cardiovascular and renal complications if untreated. Data mining plays a pivotal role in exploring and predicting diabetes by identifying high-risk populations, thereby enabling early intervention strategies such as lifestyle modifications and timely treatment initiation.*

Analyzing comprehensive datasets encompassing diabetes-related factors such as weight, blood pressure, blood glucose levels, and genetic predispositions data mining constructs predictive models to assess risks and implement targeted interventions. In a comprehensive study involving 768 cases (268 positive and 500 negative) Logistic Regression achieved 70% accuracy, with a recall of 57% and an F1 score of 0.63 , Naive Bayes (GaussianNB) achieved 68% accuracy, with a recall rate of 54% and an F1 score of 0.61, Decision Tree Classifier achieved 66% accuracy, with a recall rate of 62% and an F1 score of 0.64 , Random Forest achieved 70% accuracy, with a recall rate of 59% and an F1 score of 0.64 , XGBClassifier achieved 66% accuracy, with a recall rate of 58% and an F1 score of 0.62.

The analysis underscores a trade-off between precision and recall, particularly in classifying high-risk diabetes cases. High precision reduces false positives but may lower recall, potentially missing true positive cases. Conversely, emphasizing recall may increase false positives. Achieving a balance between these metrics is critical for effective diabetes prediction and tailored healthcare strategies This abstract encapsulates the pivotal role of data mining in diabetes research, emphasizing its impact on predictive modeling and healthcare decision making.

Keywords (Size 10 & Bold) — *Data mining, Analysis, predict diabetes, diabetes, Analysis predicts diabetes.*

I. INTRODUCTION

prediction of diabetes is a crucial aspect of proactive healthcare and disease management. Many studies and techniques have focused on predicting and preventing diabetes, as well as its associated complications. The research problem lies in data quality. This factor indicates that the data used in the study may be insufficient or Inaccurate, which could negatively affect the accuracy and validity of the study results. Data analysis. This factor indicates that the data analysis process may be complex and requires experience in data mining and statistical techniques to ensure accurate conclusions. Selecting important variables. This factor indicates difficulty. Identifying the main factors that affect the prediction of diabetes correctly, which can affect the accuracy of prediction models. Interpreting the results. This factor indicates the challenges that the researcher may face in interpreting the results and understanding the relationships between variables, especially if there are multiple influences. Controlling the variable factors. This factor indicates that there are other unstudied factors that may affect the results of the study, making it difficult to achieve high accuracy in predictions. Reliance on prediction models. This factor indicates that the results of prediction models may not be accurate enough to rely on in making medical decisions properly Crucial.

The research deals with the field of data analysis. Sections include interpretive data analysis, which is the process of interpreting and analyzing the collected data to understand the meanings and trends inherent in it. Identifying the relationships between variables. This section includes studying the relationships and effects between different variables and how they affect each other. Identifying the influencing factors. This section focuses on identifying Factors that may affect the phenomena or phenomena studied, evaluating graphical models. This section includes evaluating statistical or mathematical models that are used to represent and analyze data. Evaluating the performance of the models. This section includes evaluating the efficiency and accuracy of the models used in analyzing the data. Clarifying and discussing the results. This section includes explaining and interpreting the results. Extracted from the analysis and discussion around it, comparing the results with previous literature: This section includes comparing the results extracted from the research with previous studies and research in the same field, explaining the importance of the results. This section focuses on explaining the importance and impact of the results on the academic field or society in general.

Data analysis can effectively pinpoint diabetes risk factors and predict the probability of developing the condition. By scrutinizing variables such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, and the influence of diabetes-related genetic markers, valuable insights and probabilistic models can be extracted from comprehensive datasets. These insights reveal critical trends, relationships, and patterns, empowering informed strategic decisions in various domains. This process involves data preparation, model creation, evaluation, and application of results in practical settings. In the healthcare sector, data mining offers numerous advantages including enhancing treatment targeting and patient outcomes, improving the management and prevention of chronic diseases, discovering new trends in public health, enhancing the efficiency of healthcare services, reducing costs, optimizing resource utilization, and enhancing decision-making processes within healthcare institutions.

II. Literature Review

Data mining techniques can enhance patient care and improve healthcare delivery by enabling predictive analytics. By analyzing large datasets, patterns can be identified to predict outcomes, such as a patient's risk of readmission or identifying at-risk populations. This proactive approach allows healthcare providers to intervene early, design treatment plans, and allocate resources effectively. Implementing data mining strategies improves patient care and the efficiency of health operations. In the healthcare sector, these techniques are crucial for improving the quality of care and making accurate, evidence-based decisions. By analyzing health data, significant patterns and vital information can be discovered and used for decision-making. Data mining plays a vital role in identifying and detecting diseases, aiding researchers in developing effective health policies, and creating recommendation systems and patient health profiles. Additionally, these technologies contribute to better disease management, predicting health developments, reducing unnecessary spending, and enabling informed health decisions. In the healthcare sector, they play a crucial role in improving disease management and making precise health decisions. By analyzing healthcare data, important patterns and vital information can be identified for decision-making processes. These technologies help identify and detect diseases, aid researchers in developing effective healthcare policies, and create patient health profiles.

III. History of Data Mining

Data mining has developed into a distinct specialty within the fields of artificial intelligence and knowledge engineering. The origin of this discipline dates back to the late 1980s when these terms began to attract attention within the research community. Initially, there was some confusion in the definition of data mining, but it can be broadly described as a set of mechanisms and techniques implemented in software to uncover information hidden in data. Data extraction is not limited to SQL queries; Rather, it involves extracting valuable insights from data.

One important aspect of data mining is its application within a broader process called knowledge discovery in databases, where it focuses on discovering hidden information. This subprocess includes data preparation, analysis, and visualization of results. Over time, the scope of data mining has expanded to include broader areas of non-traditional data such as images, documents, videos, graph and network data.

These technologies saw increased demand during the 1990s as a result of technological advances in computer processing power and data storage capabilities. This development has led to the emergence of dedicated conferences and increased interest in massive data mining. Data mining techniques have their foundations in the fields of machine learning and statistics, with a strong focus on pattern extraction, clustering, and classification. Pattern mining in this field involves identifying recurring patterns in data, such as customer purchasing patterns or time trends. Clustering aims to group data into categories based on similarity, while classification creates classifiers to classify unseen data into pre-defined groups. These techniques have found applications beyond traditional tabular data mining, including text mining, image mining, and graphs, where they play a critical role in various fields by providing a wide range of techniques for extracting valuable insights from different types of data. This field is expected to witness further development, especially in dealing with ever-increasing and diverse datasets, as well as enhancing post-processing tasks such as visualization and annotation generation.

IV. Reviewing past literature

The study by Cut Fiarni et al (2019) focuses on analyzing and predicting diabetes complications in Indonesia using data mining algorithms. Techniques such as C4.5, Naive Bayes, and k-means clustering were employed to analyze the dataset and predict the three main diabetes-related complications. The research aims to develop a predictive model and identify key factors linked to these complications. Findings revealed that significant risk factors include high blood pressure for retinopathy, disease duration over four years for nephropathy, and body mass index for neuropathy. The study also demonstrated the effectiveness of the analysis techniques in generating predictive rules for diabetes complications.

In a study conducted by Ahed J. Al-Khatib, published in 2020, the primary objective was to predict type 2 diabetes using neural network analysis and assess the significance of various risk factors associated with the disease. The findings demonstrated that the model achieved prediction accuracies of 78.3% for training data and 76.9% for testing data. Glucose levels, BMI, diabetes pedigree function, number of pregnancies, age, blood pressure, insulin levels, and skin thickness emerged as the most influential predictors of diabetes .

Study by Muhannad Muhammad Al-Saleh (2022), the text deals with a study that aims to create a machine learning model that uses the Diabetes and the term This approach aims to understand why decisions are made by AI models and explain them in a way that can be understood by humans, through the use of techniques such as transparency, visualization and analysis.

Lindong Zhang and Min Liu (2022). The study aimed to analyze the risks of developing diabetes and treatment patterns for diabetic patients using data mining techniques. The main objective was to develop a diabetes risk prediction model and explore medication patterns for diabetic patients. The study concluded that the joint prediction model of K-means and logistic regression showed promising results in predicting diabetes risk and analyzing medication use patterns.

In a study conducted by K. Saravananathan and T. Velmurogan (2021), the research focused on analyzing clustering algorithms specifically k-Means and k-Medoids using diabetes data for disease prediction. The study compared the algorithms based on their execution count and runtime performance. Ultimately, the research concluded that the k-Means algorithm exhibited superior accuracy in predicting diabetes compared to the k-Medoids algorithm.

V. Previous study methods.

This section explains how previous research conducted by researchers in the same field was carried out, including a description of the processes used and the techniques that were applied, as well as the results reached in those previous studies.

Table 1. Previous Study Methods

Author Name	The method of work	Result accuracy
Study by Cut Fiarni et al(2019) .	Naive Bayes, C4.5 and k-means data mining techniques. The study identified the main influencing factors for each of the three major complications of diabetes, namely retinopathy, nephropathy, and neuropathy	overall accuracy 68%
Al-Khatib et al(2020) .	Using neural network analysis (SPSS) to	Prediction rate of 78.3% for training and 76.9%

	predict type 2 diabetes using a dataset from Kaggle.	for testing.
Study by Muhannad Muhammad Al-Saleh ,(2022)	An XGBoost classifier with Bayesian optimization was used to tune the hyperparameters. The SHAP technique was applied for global and local interpretability, revealing that polydipsia and polyuria were the largest contributors to the model predictions.	A high-performance score of 0.99, indicating 99% accuracy in distinguishing between diabetics and non-diabetics. Confusion matrix metrics included a precision of 97%, a precision of 0.981, a recall of 0.968, and an F1 score of 0.974.
Lindong Zhang and Min Liu.(2022)	Use data mining techniques to analyze diabetes risks and medication patterns for diabetics	The accuracy of the prediction model is 90.7%. The model also demonstrated an accuracy of 91.6%, recall of 96.4%, a Matthews correlation coefficient (MCC) of 75.2, and a receiver operating characteristic (ROC) area of 95.7. The kappa statistical value was determined to be 75.2.
Saravananathan and T. Velmurogan.(2021).	The study used k-Means and k-Medoids clustering algorithms to analyze datasets for diabetes and disease prediction.	The k-Means algorithm achieved an accuracy of 87%, while the k-Medoids algorithm achieved an accuracy of 80%.

The difference between previous studies and the currently proposed study.

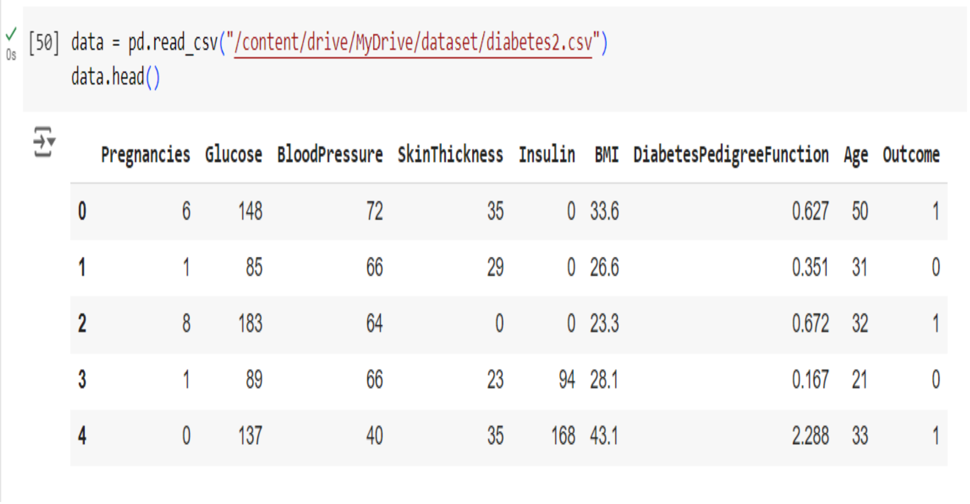
This research aims to diagnose diabetes in patients through exploratory data collection and analysis, then data modeling and evaluation. According to the study by Cut Fiarni and colleagues, the research focuses on analyzing and predicting diabetes complications in Indonesia using data mining algorithms. As for the study of Ahed J. Al-Khatib, the main goal was to predict type 2 diabetes using neural network analysis. Muhannad Muhammad Al-Saleh's study focuses on developing a machine learning model utilizing XAI (Explainable Artificial Intelligence) to provide interpretable and dependable predictions of diabetes. In contrast, K. Saravananathan and T. Velmurogan's research analyzes k-Means and k-Medoids clustering algorithms with diabetes data, aiming to predict disease trends and identify treatment patterns.

VI. METHODS

The methods section typically outlines the procedures and techniques used in a research study to collect and analyze data. It provides a detailed description of the research design, participants, materials, data collection methods, and data analysis techniques employed in the study. This section is crucial for understanding how the research was conducted and evaluating the validity and reliability of the study's findings Data mining in healthcare involves the application of various data mining techniques to extract valuable insights from healthcare data. The process includes disciplines like machine learning, artificial intelligence, probability, and statistics. Healthcare data mining employs diverse models such as predictive and descriptive models, supporting tasks like classification, association rules, clustering, and anomaly detection. Methods utilized include statistical analysis, discriminant analysis, decision trees, swarm intelligence, k-nearest neighbor, logistic regression, Bayesian classifiers, and support vector machines.

a. Data source.

The National Institute of Diabetes and Digestive and Kidney Diseases compiled data from the Pima Indian Diabetes Database, which consists of samples from a larger database with specific constraints. The dataset exclusively comprises female patients of Pima Indian descent who are aged 21 years or older. The data table used for predicting diabetes is depicted in Figure (1).



```
[50] data = pd.read_csv("/content/drive/MyDrive/dataset/diabetes2.csv")
     data.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

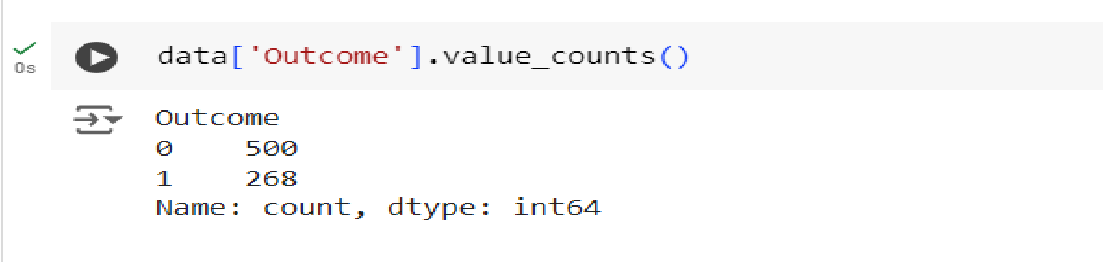
Figure 1 . Data used in the research**.**

The details of the data used are integer numbers and numbers with a decimal point as shown in Table (2).

Table 2. variables and attributes of the dataset.

Variable	Data type
Pregnancies	int64
Glucose	int64
BloodPressure	int64
SkinThickness	int64
Insulin	int64
BMI	int64
DiabetesPedigreeFunction	float64
Age	float64
Outcome	int64

The percentage of data used for infected and uninfected patients is as in Figure (2)



```
data['Outcome'].value_counts()

Outcome
0    500
1    268
Name: count, dtype: int64
```

Figure (2) The number of positive and negative states

A total of (768) data were collected, including (268) positive cases and (500) negative cases, as shown in the table (3)

Table 3. percentages of values in the dataset.

*	Outcome	Total	Percent (%)
Positive	268	768	35%
Negative	500	768	65%

b. Algorithms Used in The Research.

Algorithms are sequential procedures or formulas used to solve problems or accomplish tasks, particularly in data mining and machine learning. They analyze data, derive insights from it, and make forecasts or decisions. In the realm of intrusion detection systems, algorithms such as k-means clustering scrutinize data patterns and group them based on similarities, playing a vital role in uncovering anomalies or malicious behaviors. In data mining, algorithms consist of a series of instructions or guidelines that guide a computer program in addressing a specific issue or executing a particular task. They are crafted to scrutinize data, identify patterns, and make forecasts or categorizations based on the provided data, and the following algorithms were used in the current proposed study.

1. Random Forest (RF).

The Random Forest algorithm, widely recognized for its effectiveness in handling classification and regression tasks, combines the outputs of multiple decision trees. During training, it generates a large number of decision trees, and the final output is determined either by the majority class (for classification) or the average prediction (for regression) of the individual trees. This ensemble method enhances accuracy and helps mitigate overfitting, making it a popular choice in various machine learning applications. In Python, the Random Forest algorithm can be implemented for both classification and regression tasks.

2. Gaussian Naive Bayes (GaussianNB).

is a type of Naive Bayes method used in machine learning for classification tasks, especially when dealing with continuous or scalar features. It is based on Bayes' theorem and assumes that the presence of a particular feature in a category is unrelated to the presence of any other feature. In the case of Gaussian Naive Bayes, it specifically deals with continuous features and assumes that data features follow a Gaussian distribution. This makes it suitable for dealing with numerical data and is commonly used in various classification functions within machine learning algorithms. Gaussian Naive Bayes is implemented in Python where it can be used to build Classification models and training on them a simple probabilistic classifier based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. In the Gaussian Naive Bayes variant, it is assumed that the likelihood of the features is Gaussian (normal distribution). This classifier is commonly used for classification tasks, especially when dealing with continuous data.

3. Classification Report Evaluation (RE).

A classification report is a key tool for assessing machine learning model performance. It provides a detailed summary of the model's performance by presenting important metrics such as precision, recall, and F1 scores. These metrics are crucial for understanding the model's effectiveness in correctly identifying different classes within the dataset. Precision measures the accuracy of positive predictions, while recall indicates the proportion of actual positives correctly identified by the model. The F1 score, a harmonic mean of precision and recall, balances the two metrics. Overall, the classification report is invaluable for evaluating the classification model's ability to handle various categories and make accurate predictions.

4. Decision tree or classification tree (DT).

is a learning technique used in statistics, data mining, and machine learning. This method is used to classify or predict outcomes using input variables. Decision trees are a form of predictive modeling that help to make decisions or find different solutions for specific outcomes. This method is widely used in tasks that aim to classify or predict outcomes using input data.

5. Linear regression (LR)

Linear regression is a statistical model designed to estimate the linear relationship between a dependent variable (response) and one or more independent variables (predictors). The goal of linear regression is to determine the best-fitting linear equation that represents the relationship between these variables. The model assumes that this relationship is linear, meaning that any change in the independent variables corresponds to a proportional change in the dependent variable.

VII. RESULTS AND DISCUSSION.

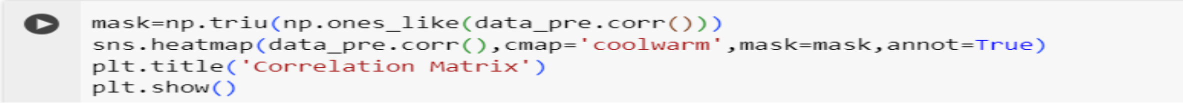
In this section, the study's results are presented and interpreted. The methods of data analysis are detailed using tables, figures, or statistical analysis. Following the presentation of results, the discussion section analyzes and interprets these findings, compares them to existing literature, and explains their significance.

A. Exploratory Data Analysis.

Exploratory Data Analysis (EDA) involves analyzing datasets to summarize their key characteristics, typically employing visual methods. Its primary aim is to uncover insights and patterns inherent in the data.

formal modeling or hypothesis testing. It involves looking at the data from different angles, summarizing their main characteristics, and detecting patterns, anomalies, and relationships between ariables. Exploratory Data Analysis (EDA) is a crucial initial step in the data analysis process, as it aids in identifying patterns, testing assumptions, and guiding subsequent analyses.

Figure (1) shows a strong relationship between BMI and skin thickness. The term "body mass index" (BMI) refers to a scale used to determine the percentage of body fat based on a person's weight and height. The figure shows that there is a strong relationship between BMI and skin thickness, indicating that the higher the BMI, the thicker the skin. BMI is commonly used to evaluate whether a person is underweight, normal weight, overweight, or obese. It also indicates that there is a relationship between glucose level and body mass index, Figure (3) also indicates that there is no “inverse relationship” between the score and other independent variables. An inverse relationship usually means that as one variable increases, the other decreases, and vice versa.



```
mask=np.triu(np.ones_like(data_pre.corr()))
sns.heatmap(data_pre.corr(),cmap='coolwarm',mask=mask,annot=True)
plt.title('Correlation Matrix')
plt.show()
```

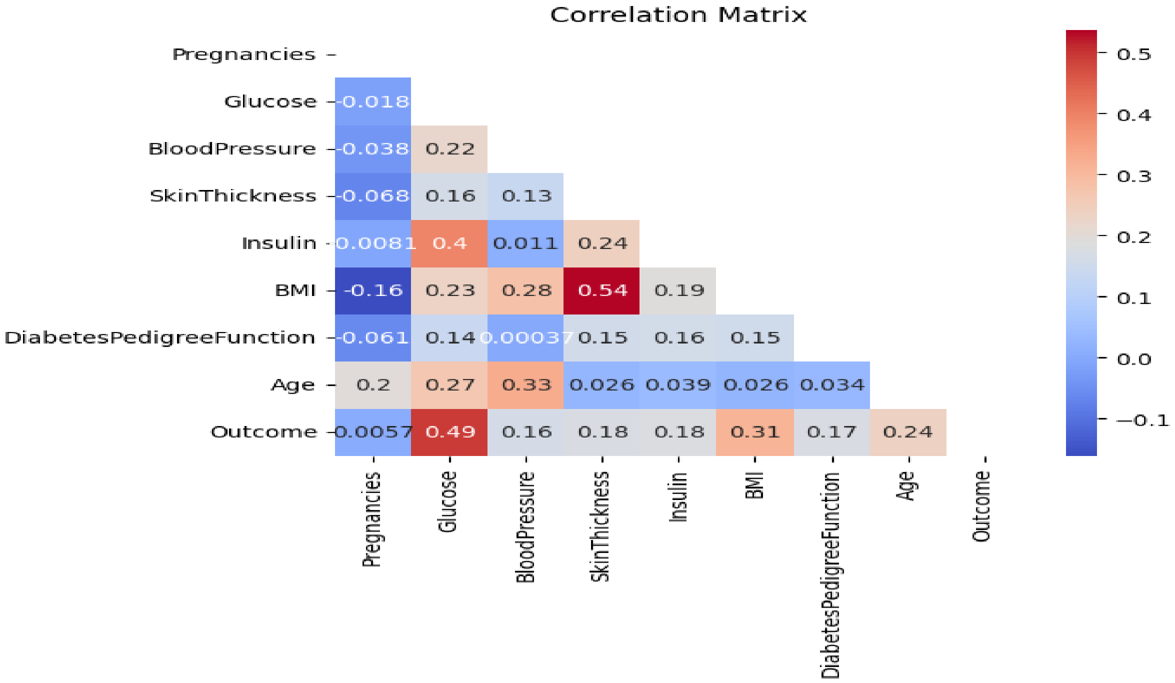
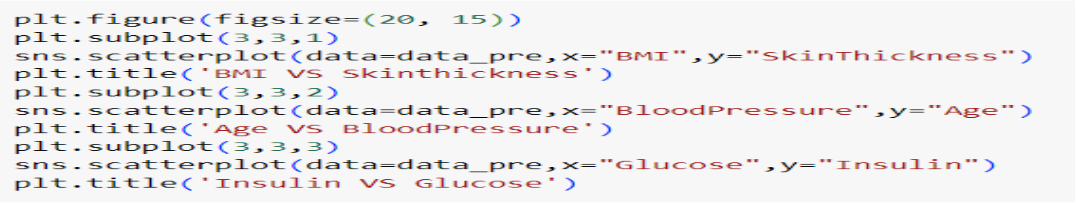


Figure 3. Exploratory Data Analysis: the relationship between two variables.

In addition, the report notes that there is a "small association" between age and blood pressure, as well as between insulin and glucose level. Simple correlation indicates that there is a direct relationship between variables.



```
plt.figure(figsize=(20, 15))
plt.subplot(3,3,1)
sns.scatterplot(data=data_pre,x="BMI",y="SkinThickness")
plt.title('BMI VS Skinthickness')
plt.subplot(3,3,2)
sns.scatterplot(data=data_pre,x="BloodPressure",y="Age")
plt.title('Age VS BloodPressure')
plt.subplot(3,3,3)
sns.scatterplot(data=data_pre,x="Glucose",y="Insulin")
plt.title('Insulin VS Glucose')
```

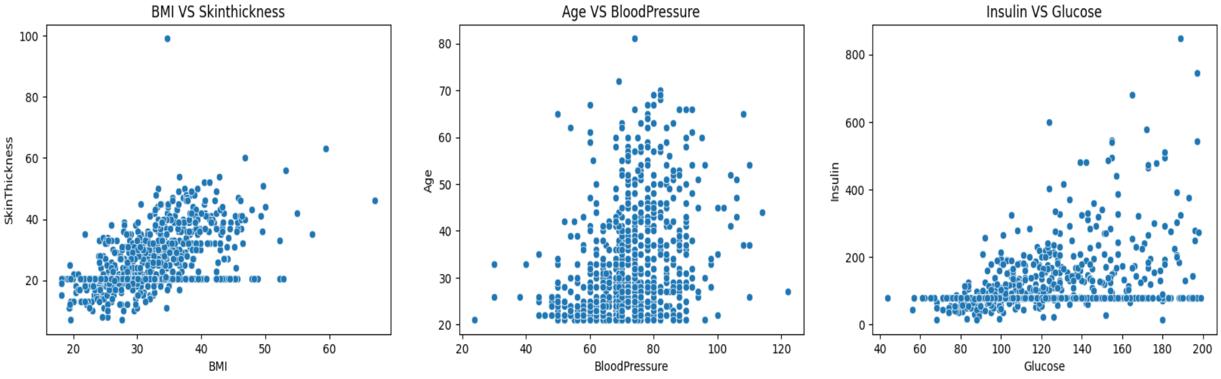
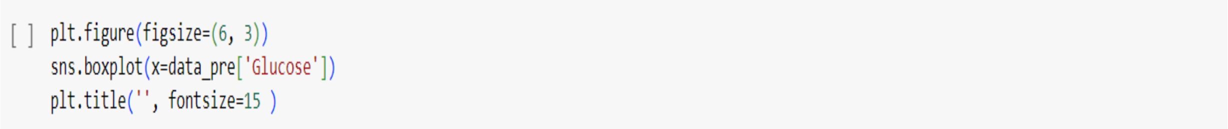


Figure 4. calculate point biserial correlation (BMI VS Skin Thickness, Age VS Blood Pressure, Insulin VS Glucose)

Figure 4 indicates that there is a strong relationship between BMI and skin thickness, as it shows that people with a higher BMI often have less skin thickness. However, BMI can have a stronger relationship with the outcome related to health and fitness. Skin thickness can therefore be dropped as an independent indicator



```
[ ]  plt.figure(figsize=(6, 3))
     sns.boxplot(x=data_pre['Glucose'])
     plt.title('', fontsize=15 )
```

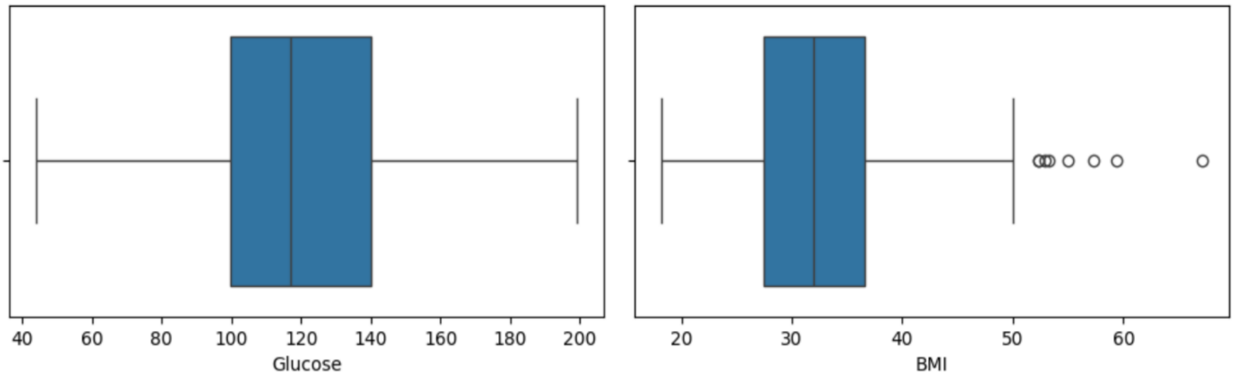


Figure 5. outliers Identification (Glucose , BMI).

Glucose provides energy to the body once it is absorbed into the bloodstream. On the other hand, Body Mass Index (BMI) measures whether an individual is underweight or overweight relative to their height. It is calculated by dividing a person's weight in kilograms by their height in meters squared.

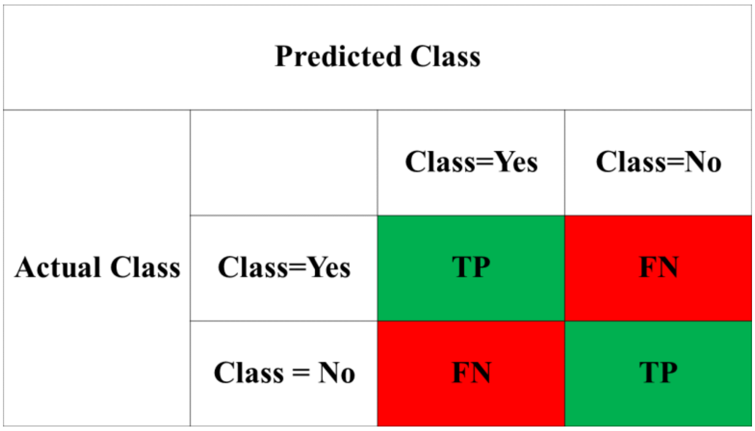
Extreme values are those that are far from the statistical mean of the sample, whether above the average (upper extreme) or below the average (lower extreme). It is necessary to identify outliers in glucose levels and BMI, as determining them based on the two most influential variables means we will identify values that are far from the average and significantly impact the results.

From the figure (5) In statistics, an outlier is a data point that differs significantly from other observations in a dataset. Outliers can have a disproportionate influence on the results of statistical analysis, especially in regression analysis where they can heavily impact the estimated coefficients and predictions.

'Glucose' and 'BMI' are two variables that are often considered influential in health and medical data analysis. In the context of a health-related dataset, identifying outliers in these variables is important because they can significantly affect the interpretation of the data and the conclusions drawn from it. Therefore, it is essential to check for outliers in 'Glucose' and 'BMI' to ensure that the analysis accurately represents the underlying patterns in the data and the process of removing outliers from the database used is Outliers are values that deviate from the statistical average of the data and are considered abnormal or unexpected the process of removing outliers aims to improve the accuracy and validity of statistical analyzes and results

B. Data Modelling and evaluation.

Model evaluation is the process of using different metrics to understand the performance of a machine learning model, as well as its strengths and weaknesses this includes evaluating whether the model is working as intended and understanding its predictive capabilities. In the context of data science, model evaluation often includes techniques such as cross-validation to evaluate a model's ability to generalize to new data. Data modeling and model evaluation are important because they enable efficient representation and evaluation of data and models.



		Predicted Class	
		Class=Yes	Class=No
Actual Class	Class=Yes	TP	FN
	Class = No	FN	TP

Figure 6. Confusion Matrix.

In Figure (6), the confusion matrix displays both actual and predicted values. True Positive occurs when both predicted and actual labels are positive, whereas False Positive happens when the model predicts a positive label incorrectly. These definitions similarly apply to True Negative and False Negative. The evaluation metrics we will explore are grounded in these definitions. First, we will discuss precision, which assesses the false positive rate. A high precision score signifies a low rate of false positives, determined by the proportion of true positive predictions to the overall positive predictions generated by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (1)$$

Recall, also known as sensitivity in Classification models, measures the proportion of correctly predicted positive values relative to all actual positive values. A score exceeding 0.5 is generally deemed satisfactory, indicating the model's effectiveness in identifying positive results.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (2)$$

The F1-score is the harmonic mean of precision and recall in classification models. It holds greater significance than accuracy as it accounts for both false positives and false negatives. While accuracy is appropriate for balanced class distributions, the F1-score is particularly valuable for assessing performance in scenarios with uneven class distributions.

$$\text{F 1} - \text{score} = \frac{2 \times (\text{Precision} \times \mathit{Recall})}{\text{Precision} \times \mathit{Recall}} \qquad (3)$$

Accuracy measures the proportion of correct predictions relative to the total number of observations.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (4)$$

True Positive (TP): Indicates cases that were correctly predicted as positive. For example, in a diagnostic test for a particular disease, True Positive is when the patient is correctly predicted to have the disease.

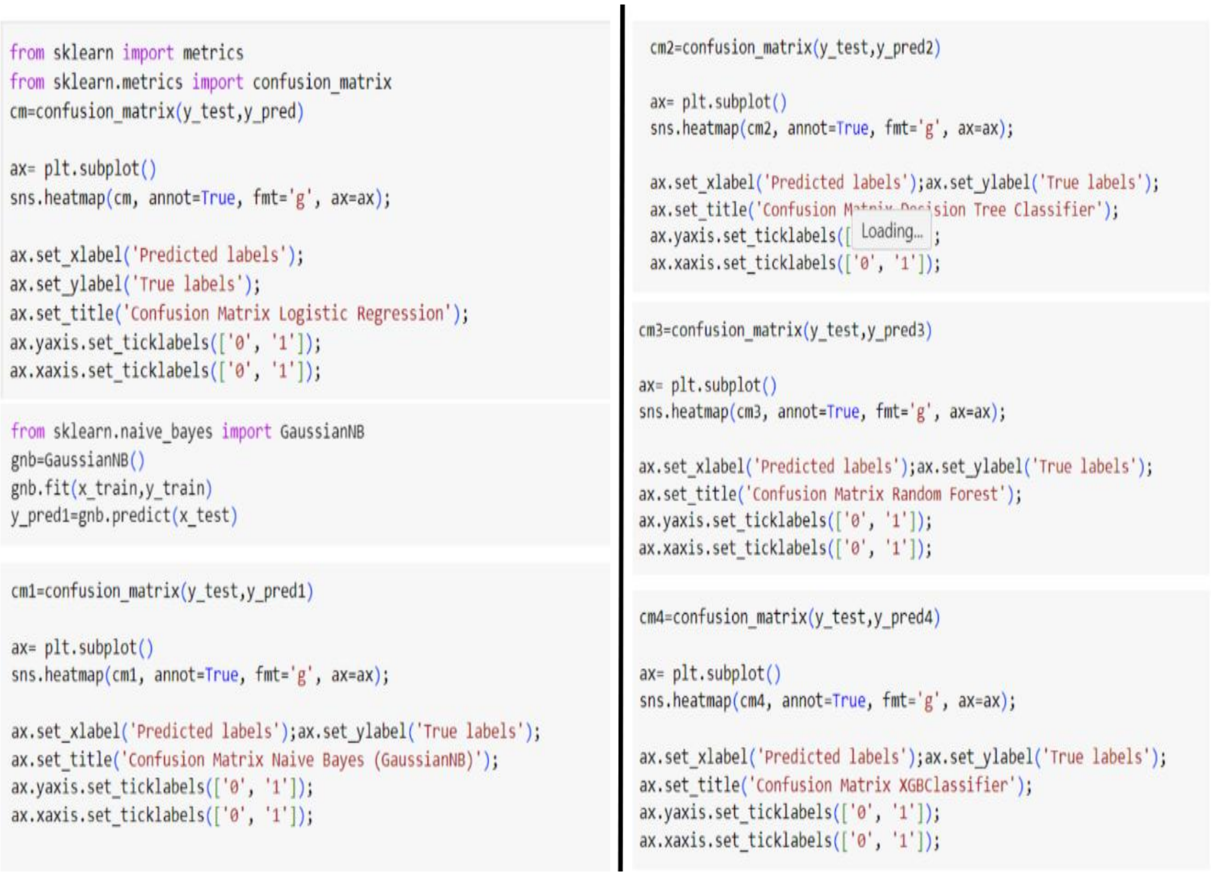
True Negative (TN): Indicates cases that were correctly predicted as negative. For example, in the same diagnostic test, True Negative occurs when a patient is correctly predicted not to have the disease.

False Positive (FP): Occurs when a condition is incorrectly predicted to be positive even though it is in fact negative. For example, if there is an incorrect prediction that someone will have a disease when in fact they are not.

False Negative (FN): Occurs when a condition is incorrectly predicted to be negative even though it is in fact positive. For example, if there is an incorrect prediction that someone will not have a disease when in fact they do

Binary classification in the study, five binary classification models were used, namely Logistic Regression

Classification, Decision Tree Classifier, Random Forest, XGBClassifier, Naive Bayes, Prediction in this section to compare the ratings.



```
from sklearn import metrics
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test,y_pred)

ax= plt.subplot()
sns.heatmap(cm, annot=True, fmt='g', ax=ax);

ax.set_xlabel('Predicted labels');
ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix Logistic Regression');
ax.yaxis.set_ticklabels(['0', '1']);
ax.xaxis.set_ticklabels(['0', '1']);
```

```
from sklearn.naive_bayes import GaussianNB
gnb=GaussianNB()
gnb.fit(x_train,y_train)
y_pred1=gnb.predict(x_test)
```

```
cm1=confusion_matrix(y_test,y_pred1)

ax= plt.subplot()
sns.heatmap(cm1, annot=True, fmt='g', ax=ax);

ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix Naive Bayes (GaussianNB)');
ax.yaxis.set_ticklabels(['0', '1']);
ax.xaxis.set_ticklabels(['0', '1']);
```

```
cm2=confusion_matrix(y_test,y_pred2)

ax= plt.subplot()
sns.heatmap(cm2, annot=True, fmt='g', ax=ax);

ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix Decision Tree Classifier');
ax.yaxis.set_ticklabels([ Loading... ;
ax.xaxis.set_ticklabels(['0', '1']);
```

```
cm3=confusion_matrix(y_test,y_pred3)

ax= plt.subplot()
sns.heatmap(cm3, annot=True, fmt='g', ax=ax);

ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix Random Forest');
ax.yaxis.set_ticklabels(['0', '1']);
ax.xaxis.set_ticklabels(['0', '1']);
```

```
cm4=confusion_matrix(y_test,y_pred4)

ax= plt.subplot()
sns.heatmap(cm4, annot=True, fmt='g', ax=ax);

ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix XGBClassifier');
ax.yaxis.set_ticklabels(['0', '1']);
ax.xaxis.set_ticklabels(['0', '1']);
```

Figure 7. Code Confusion Matrix (Logistic Regression, Naive Bayes, Decision Tree Classifier, Random Forest, XGBClassifier).

The figure (8) presents the performance of different classifiers in terms of True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP) rates.

For the Logistic Regression classifier, the True Positive (TP) rate was 57.01%, the True Negative (TN) rate was 19.7%, the False Negative (FN) rate was 8.3%, and the False Positive (FP) rate was 14.9%.

The Naive Bayes classifier had a TP rate of 56.5%, TN rate of 18.8%, FN rate of 8.7%, and FP rate of 15.7%.

The Decision Tree Classifier showed a TP rate of 54.3%, TN rate of 21.4%, FN rate of 10.9%, and FP rate of 13.15%.

In the case of the Random Forest classifier, the TP rate was 56.5%, TN rate was 20.6%, FN rate was 8.7%, and FP rate was 14%.

Finally, the XGBClassifier exhibited a TP rate of 54.8%, TN rate of 20.17%, FN rate of 10.5%, and FP rate of 14.5%.

The classification report evaluates the quality of the classification model by providing metrics such as precision, recall, F1 score, and support for each class as in Figure (21).

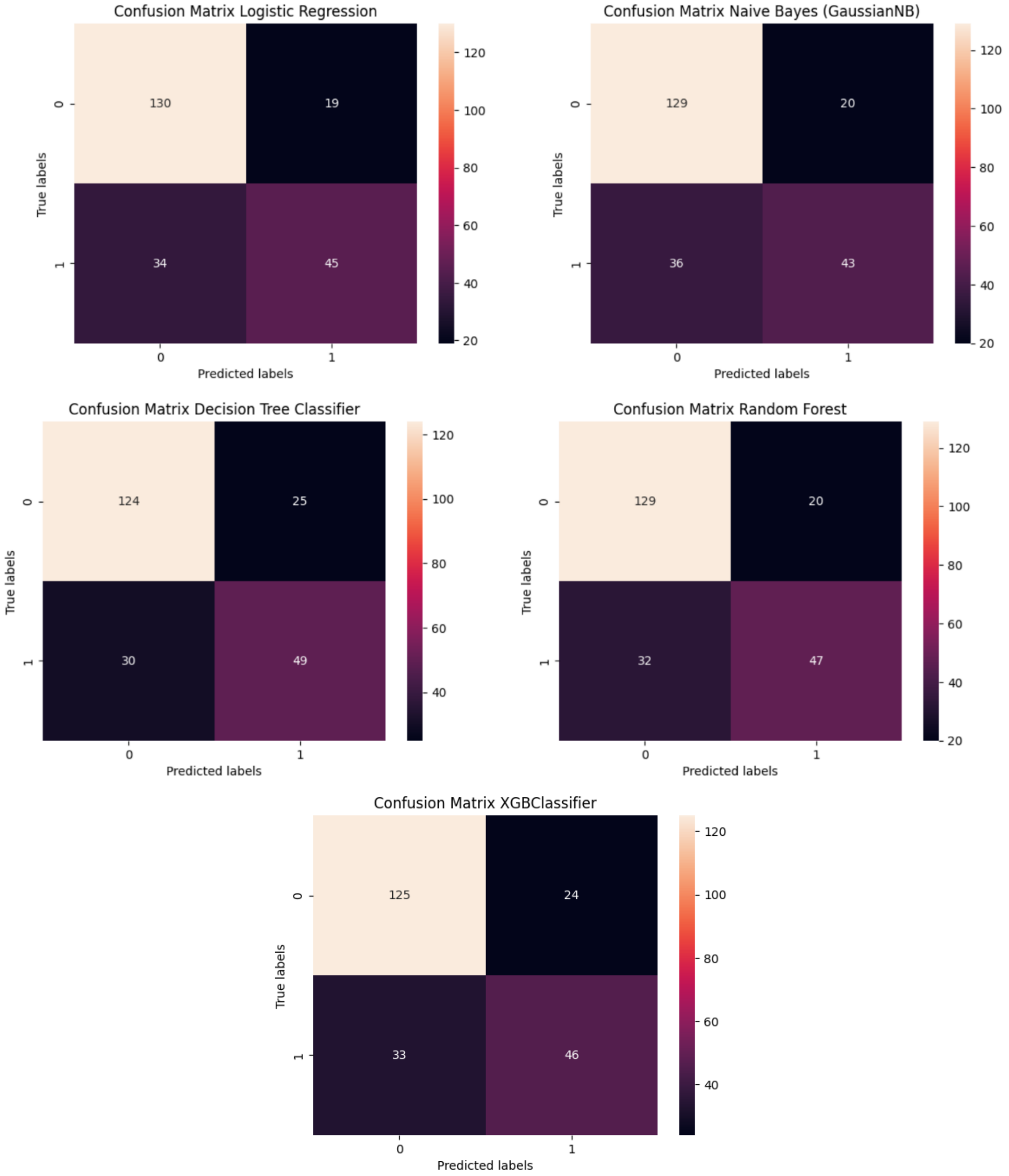
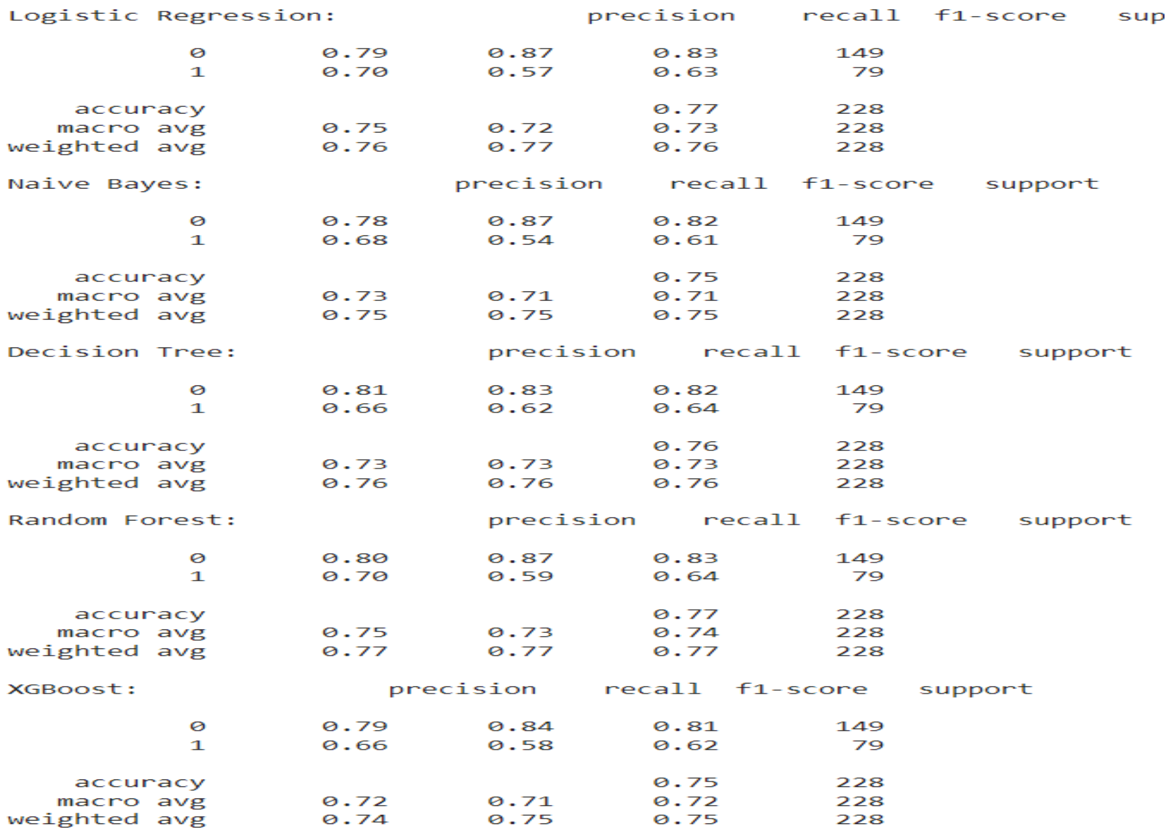


Figure 8. Confusion Matrix (Logistic Regression, Naive Bayes, Decision Tree Classifier, Random Forest, XGBClassifier).



```
Logistic Regression:                precision    recall  f1-score   sup

                 0       0.79      0.87      0.83       149
                 1       0.70      0.57      0.63        79

          accuracy                           0.77       228
         macro avg       0.75      0.72      0.73       228
      weighted avg       0.76      0.77      0.76       228

Naive Bayes:                precision    recall  f1-score   support

                 0       0.78      0.87      0.82       149
                 1       0.68      0.54      0.61        79

          accuracy                           0.75       228
         macro avg       0.73      0.71      0.71       228
      weighted avg       0.75      0.75      0.75       228

Decision Tree:                precision    recall  f1-score   support

                 0       0.81      0.83      0.82       149
                 1       0.66      0.62      0.64        79

          accuracy                           0.76       228
         macro avg       0.73      0.73      0.73       228
      weighted avg       0.76      0.76      0.76       228

Random Forest:                precision    recall  f1-score   support

                 0       0.80      0.87      0.83       149
                 1       0.70      0.59      0.64        79

          accuracy                           0.77       228
         macro avg       0.75      0.73      0.74       228
      weighted avg       0.77      0.77      0.77       228

XGBoost:                precision    recall  f1-score   support

                 0       0.79      0.84      0.81       149
                 1       0.66      0.58      0.62        79

          accuracy                           0.75       228
         macro avg       0.72      0.71      0.72       228
      weighted avg       0.74      0.75      0.75       228
```

Figure 9. Classification Report Evaluation.

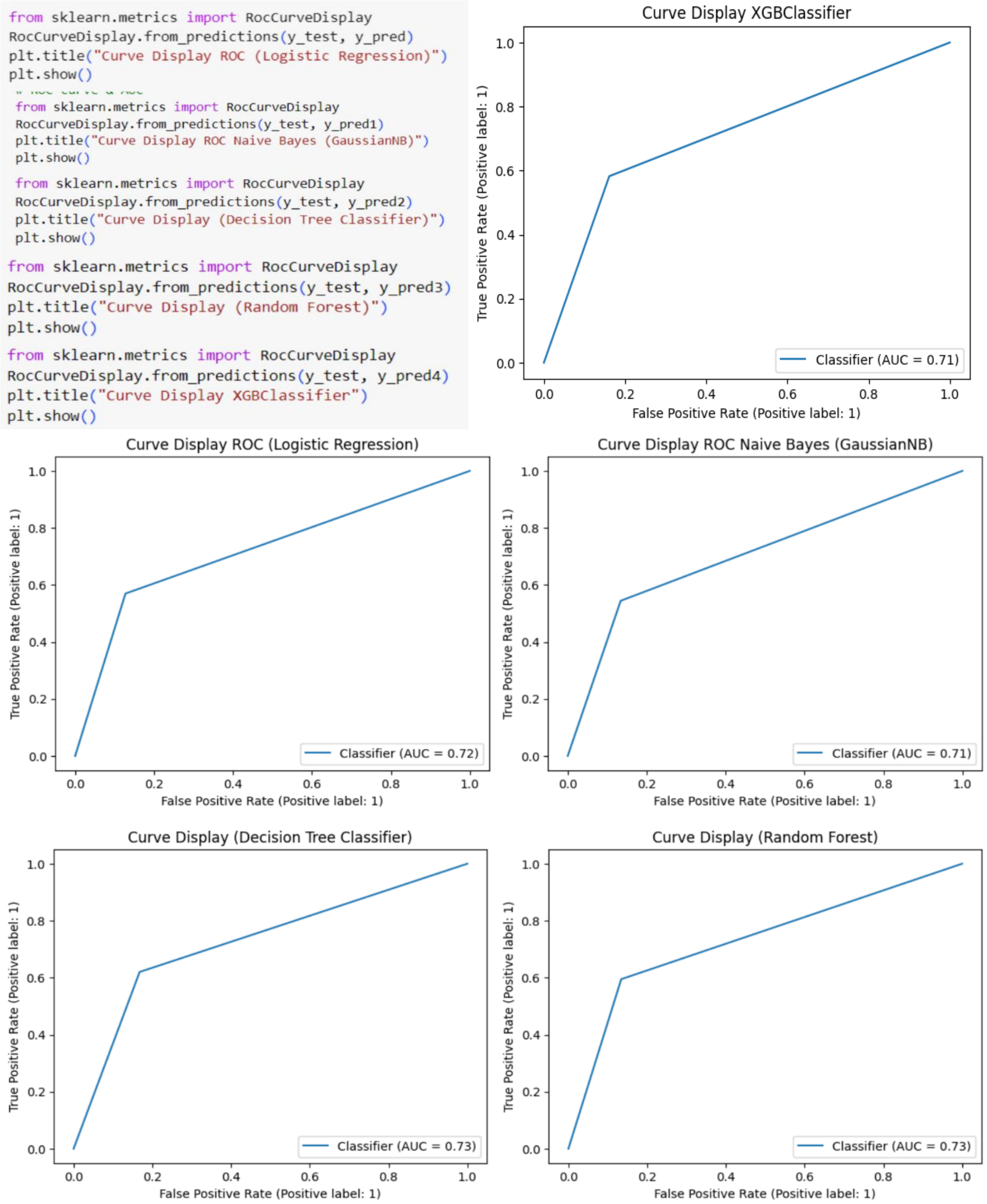
Accuracy measures the ratio of correctly predicted positive observations to the total predicted positives. Recall is the ratio of correctly predicted positive observations to all actual positive observations. The F1 score is the harmonic average of precision and recall, providing a balance between the two. Support refers to the number of actual occurrences of a class in the data set. These metrics provide insight into the performance of the classification model for each category, as shown in Table (3).

Table 3. Logistic Regression, Naive Bayes, Decision Tree Classifier, Random Forest, XGBClassifier produced for the Classification models.

		precision	recall	f1-score
logistic Regression	negative	0.79	0.87	0.83
	positive	0.70	0.57	0.63
Naive Bayes	negative	0.78	0.87	0.82
	positive	0.68	0.54	0.61
Decision Tree	negative	0.81	0.83	0.82
	positive	0.66	0.62	0.64
Random Forest	negative	0.80	0.87	0.83
	positive	0.70	0.59	0.64
XGBoost	negative	0.79	0.84	0.81
	positive	0.66	0.58	0.62

Open-space ROC tools for practicing classification models, especially in binary classification economics. The ROC lights clearly represent the trade-off between true motor and correct motor rate, while the forest AUC metric exists to summarize model performance across all classification thresholds [39][41].

In evaluating a classification model, the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) are important metrics. The ROC curve visually depicts the model's performance by plotting the true positive rate (sensitivity) against the false positive rate across various thresholds. A ROC curve that curves towards the upper left corner indicates that the model is effective, achieving high true positive rates while keeping false positive rates low.



```
from sklearn.metrics import RocCurveDisplay
RocCurveDisplay.from_predictions(y_test, y_pred)
plt.title("Curve Display ROC (Logistic Regression)")
plt.show()
```

```
from sklearn.metrics import RocCurveDisplay
RocCurveDisplay.from_predictions(y_test, y_pred1)
plt.title("Curve Display ROC Naive Bayes (GaussianNB)")
plt.show()
```

```
from sklearn.metrics import RocCurveDisplay
RocCurveDisplay.from_predictions(y_test, y_pred2)
plt.title("Curve Display (Decision Tree Classifier)")
plt.show()
```

```
from sklearn.metrics import RocCurveDisplay
RocCurveDisplay.from_predictions(y_test, y_pred3)
plt.title("Curve Display (Random Forest)")
plt.show()
```

```
from sklearn.metrics import RocCurveDisplay
RocCurveDisplay.from_predictions(y_test, y_pred4)
plt.title("Curve Display XGBClassifier")
plt.show()
```

Curve Display XGBClassifier

Curve Display ROC (Logistic Regression)

Curve Display ROC Naive Bayes (GaussianNB)

Curve Display (Decision Tree Classifier)

Curve Display (Random Forest)

Figure .10 Curve Display (Logistic Regression, Naive Bayes, Decision Tree Classifier, Random Forest, XGBClassifier).

The database contains more negative results than positive ones, which can bias the classifier towards the negative class. The Area Under the Receiver Operating Characteristic (ROC) Curve, or AUC, measures the performance of a classification model. It evaluates how well the model can differentiate between two classes, providing insight into its effectiveness in distinguishing positive from negative outcomes. The AUC value ranges from 0 to 1, Performance improves as the AUC value approaches 1. ROC, which stands for Receiver Operating Characteristics, is a graph that illustrates the performance of a classification model as the threshold is varied. This graph shows the relationship between true positive rates and false positive rates. AUC and ROC are valuable for comparing the performance of different classification models. AUC is an indicator of overall model performance, while the ROC curve demonstrates how performance changes with different threshold settings. (4).

Table 4. ROC curve & AUC Logistic Regression, Naive Bayes, Decision Tree Classifier, Random Forest, XGBClassifier produced for the Classification models.

Algorithm	Classifier AUC
logistic Regression	0.72
Naive Bayes	0.71
Decision Tree	0.73
Random Forest	0.73
XGBoost	0.71

VIII. CONCLUSION.

Class 0 contains a significantly larger number of cases than Class 1. The text indicates that data-driven models have difficulty correctly predicting cases of the minority class (Class 1), resulting in the precision, recall, and F1 score for Class 1 being lower than Class 0.

It is explained that increasing precision (reducing false positives) may lead to a decrease in recall (increasing false negatives) and vice versa and the text notes the importance of choosing classifiers that achieve higher recall for class 1 even if this comes at the cost of lower precision.

(Class 0 and Class 1) refer to the two target classes in the data classification problem, where class 0 is the majority class and class 1 is the minority class. Accuracy is an assessment of the performance of a classification model, and is defined as the ratio of the number of correctly classified cases to the total number of cases. Recall is also known as with sensitivity, which is an assessment of the ability of the classification model to recognize all instances of the target category, the F1 score is a comprehensive measure of the performance of the classification model that takes into account both precision and recall. False positives and false negatives indicate instances of misclassification, where a false positive is the classification of a negative instance as positive. A false negative is the classification of a positive condition as a negative.

REFERENCES

[1] Cut Fiarni, Evasaria M. Sipayung, Siti Maemunah ,"Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm" Vol161,Pages 449-457, 2019, https://doi.org/10.1016/j.procs.2019.11.144.

[2] Ahed J. Alkhatib, Amer Mahmoud Sindiani , Eman Hussein Alshdaifat, "Prediction of Risk Factors Leading to Diabetes Using Neural Network Analysis" vol3, Issue 2,2020 , https://asclepiusopen.com/clinical-research-in-diabetes-and-endocrinology/volume-3-issue-2/4.pdf

[3] Mohanad M.Alsaleha , Kyung-Mo Yeonb , SohailAkhtara , Qazi Mohammad Sajid Jamala," XAI Implementation on Preliminary Data Analysis Phase: Explainable Output Application with Prediction of Diabetes Mellitus at Early Stage" Vol.13 No.02 (2022), 1070-1078 , https://doi.org/10.17762/turcomat.v13i2.12677.

[4] Lindong Zhang , Min Liu," Analysis of Diabetes Disease Risk Prediction and Diabetes Medication Pattern Based on Data Mining",Vol 2022, Article ID 2665339, p9, https://doi.org/10.1155/2022/2665339.

[5] K. Saravananathan, T. Velmurugan , "Quality Based Analysis of Clustering Algorithms using Diabetes Data for the Prediction of Disease", vol-8, Issue-11S2, 2019, 2278-3075, http://dx.doi.org/10.35940/ijitee.K1072.09811S219.

[6] Hong Guo1,ZhiChao Fan1,Yan Zeng," Novel Data Mining Analysis Method on Risk Prediction of Type 2 Diabetes",94:1183–1198,2020, https://doi.org/10.1007/s11265-021-01717-4.

[7] Joyce Jackson,"data mining a conceptual overview",vol 8 267-296,2002, https://doi.org/10.17705/1CAIS.00819.

[8] David Crockett, Ryan Johnson, and Brian Eliason ," What is Data Mining in Healthcare", vol 8 ,2002, 267-296, https://www.healthcatalyst.com/wp-content/uploads/2014/06/What-is-data-mining-in-healthcare.pdf.

[9] Ogundele I.O, Popoola O.L, Oyesola O.O, Orija K.T," A Review on Data Mining in Healthcare",vol 7, Issue 9, September 2018, ISSN: 2278 – 1323, https://www.researchgate.net/publication/370899263.

[10] FRANS COENEN," Data Mining: Past, Present and Future",vol 7,26(01):25-29,2018, https://www.researchgate.net/publication/220254364.

[11] Felipe Israel Marinho , Mario Henrique Akihiko da Costa Adaniya , "DATA MINING, MACHINE LEARNING, AND BUSINESS INTELLIGENCE - A CASE STUDY ON CRYPTOCURRENCIES",vol39, 2596-2809, 2023, http://periodicos.unifil.br/index.php/Revistateste/article/download/2891/2640/.

[12] Bernd Kirchhof," 170 years of data-mining: history and future", vol 262, pages 1013–1014, 2024, https://doi.org/10.1007/s00417-023-06359-9.

[13] Kuldeep Nagi , "From Bits and Bytes to Big Data-An Historical Overview ", (June 9, 2020), , https://ssrn.com/abstract=3622921 or http://dx.doi.org/10.2139/ssrn.3622921.

[14] Ravindra Maan , "The Evolution of Python Programming Language", 2040-0748 , Vol-9 Issue-02 July 2020 , https://ijgst.com/admin/uploadss/The%20Evolution%20of%20Python%20Programming%20Language.pdf.

[15] Neesha Jothia , Nur'Aini Abdul Rashidb , Wahidah Husainc , "Data Mining in Healthcare – A Review",vol 72, P 306-313, 2015, https://doi.org/10.1016/j.procs.2015.12.145.

[16] Furqan Alama , Rashid Mehmoodb , Iyad Katiba , Aiiad Albeshri," Analysis of Eight Data Mining Algorithms for Smarter Internet of Things (IoT)", Volume 98, P 437-442, 2016, https://doi.org/10.1016/j.procs.2016.09.068.

[17] Steven J. Rigatti, MD, DBIM, DABFM," Random Forest", vol 47, : 31–39, https://doi.org/10.17849/insm-47-01-31-39.1.

[18] Solane Duquea , Dr.Mohd. Nizam bin Omar ," Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)", Vol 61, Pages 46-51, 2015, https://doi.org/10.1016/j.procs.2015.09.145.

[19] Nabila Farnaaz , M. A. Jabbar ," Random Forest Modeling for Network Intrusion Detection System",89 213 – 217 , 2016, https://doi.org/10.1016/j.procs.2016.06.047.

[20] Aritz Pe´rez *, Pedro Larran˜aga, In˜aki Inza," Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes",vol 43, p 1–25 , 2006 , https://doi.org/10.1016/j.ijar.2006.01.002.

[21] Nurul Rismayanti , Ahmad Naswin , Umar Zaky , Muhammad Zakariyah , Dwi Amalia Purnamasari ," Evaluating Thresholding-Based Segmentation and Humoment Feature Extraction in Acute Lymphoblastic Leukemia Classification using Gaussian Naive Bayes",Volume 1 Issue 2 ISSN 3025-4167, https://doi.org/10.56705/ijaimi.v1i2.99.

[22] Ivan Rodrigues, Alitta Parayil, Tarun Shetty, Imran Mirza," Use of Linear Discriminant Analysis (LDA), K Nearest Neighbours (KNN), Decision Tree (CART), Random Forest (RF), Gaussian Naive Bayes (NB), Support Vector Machines (SVM) to Predict Admission for Post Graduation Courses",7 Pages Posted: 26 Oct 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3683065.

[23] Sheikh Amir Fayaza , Majid Zamanb, Muheet Ahmed Buttc ," To Ameliorate Classification Accuracy using Ensemble Distributed Decision Tree (DDT) Vote Approach: An Empirical discourse of Geographical Data Mining", Volume 184, 2021, Pages 935-940, https://doi.org/10.1016/j.procs.2021.03.116.

[24] Zeljko Vujovic, "Classification Model Evaluation Metrics", Volume 12 Issue 6, 2021, https://dx.doi.org/10.14569/IJACSA.2021.0120670.

[25] Ching-Lung Fan,"Article Evaluation of Classification for Project Features with Machine Learning Algorithms", 2022, 14(2), 372; https://doi.org/10.3390/sym14020372.

[26] Karan Bhowmick , Vivek Sarvaiya," A COMPARATIVE STUDY OF THE DIFFERENT CLASSIFICATION ALGORITHMS ON FOOTBALL ANALYTICS ", Int. J. Adv. Res. 9(08), 392-407, http://dx.doi.org/10.21474/IJAR01/13280.

[27] D.Y. Lin,"Linear regression analysis of censored medical costs", Volume 1, Issue 1, March 2000, Pages 35–47, https://doi.org/10.1093/biostatistics/1.1.35.

[28] Gülden Kaya Uyanık , Neşe Güler ,"A Study on Multiple Linear Regression Analysis" , Volume 106, 10 December 2013, Pages 234-240, https://doi.org/10.1016/j.sbspro.2013.12.027.

[29] Peter C. Austina, Ewout W. Steyerbergd ,"The number of subjects per variable required in linear regression analyses" , VOLUME 68, ISSUE 6, P627-636, JUNE 2015, http://dx.doi.org/10.1016/j.jclinepi.2014.12.014.

[30] Kolawole Ogunsina , Ilias Bilionis b , Daniel DeLaurentis , "Exploratory data analysis for airline disruption management", Volume 6, 15 December 2021, 100102, https://doi.org/10.1016/j.mlwa.2021.100102.

[31] joan Stelmack, OD; Janet P. Szlyk, PhD; Thomas Stelmack, OD; Judith Babcock-Parziale, PhD; Paulette Demers-Turco, OD; R. Tracy Williams, OD; Robert W. Massof, PhD, "Use of Rasch person-item map in exploratory data analysis: A clinical perspective", Volume 41, Number 2, Pages 233–242,2004, http://dx.doi.org/10.1682/JRRD.2004.02.0233.

[32] Kunitoshi Iseki 1, Yoshiharu Ikemiya, Kozen Kinjo, Taku Inoue, Chiho Iseki, Shuichi Takishita,"Body mass index and the risk of development of end-stage renal disease in a screened cohor", VOLUME 65, ISSUE 5, P1870-1876, MAY 2004 , https://doi.org/10.1111/j.1523-1755.2004.00582.x.

[33] Massimo Cirillo, Pietro Anastasio , Natale G. De Santo, "Relationship of gender, age, and body mass index to errors in predicted kidney function", (2005) 20: 1791–1798, https://doi.org/10.1093/ndt/gfh962.

[34] Chandra L. Jackson, PhD, MS, Hsin-Chieh Yeh, PhD , Moyses Szklo, MD, DrPH, Frank B. Hu, MD, PhD , Nae-Yuh Wang, PhD , Rosemary Dray-Spira, MD, PhD, and Frederick L. Brancati, MD, MHS," Body-Mass Index and All-Cause Mortality in US Adults With and Without Diabetes ", 29(1):25–33,2013, DOI: 10.1007/s11606-013-2553-7.

[35] George A Bray, Kathleen A Jablonski, Wilfred Y Fujimoto, Elizabeth Barrett-Connor, Steven Haffner, Robert L Hanson, James O Hill, Van Hubbard, Andrea Kriska, Elizabeth Stamm, and F Xavier Pi-Sunyer , " Relation of central adiposity and body mass index to the development of diabetes in the Diabetes Prevention Program ", r 2008;87:1212– 8, https://doi.org/10.1093/ajcn/87.5.1212.

[36] Ari Karppinen , Jaakko Kukkonen , Jari Härkönen , Mari Kauhaniemi , Anu Kousa , Tarja Koskentalo,"A modelling system for predicting urban air pollution: Comparison of model predictions with the data of an urban measurement network in Helsinki", 34(22):3735-3743 , https://www.researchgate.net/publication/222829613_A_modelling_system_for_predicting_urban_air_pollution_Comparison_of_model_predictions_with_the_data_of_an_urban_measurement_network_in_Helsinki

[37] Daniel L. Moody , "Measuring the Quality of Data Models: An Empirical Evaluation of the Use of Quality Metrics in Practice", Proceedings of the 11th European Conference on Information Systems, ECIS 2003, Naples, Italy 16-21 June 2003, http://aisel.aisnet.org/ecis2003/78.

[38] A.L.Sayeth Saabith , MMM.Fareez , T.Vinothraj , " Python Current Trend Applications-An Overview" , Volume 6, Issue 10, October-2019, e-ISSN: 2348 - 4470, print-ISSN: 2348-6406 , https://www.scribd.com/document/544106143/IJAERDV06I1085481.

[39] Andre M. Carrington , Douglas G. Manuel, Paul W. Fieguth , Tim Ramsay , Venet Osmani , Bernhard Wernly, Carol Bennett, Steven Hawken , Olivia Magwood, Yusuf Sheikh, Matthew McInnes, and Andreas Holzinger , Senior Member, "Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation" , Volume: 45, Issue: 1, 01 January 2023 , https://doi.org/10.1109/TPAMI.2022.3145392.

[40] Weichao Xu; Shun Liu; Xu Sun; Siyang Liu; Yun Zhang ,"A Fast Algorithm for Unbiased Estimation of Variance of AUC Based on Dynamic Programming", vol 9553 – 9560 , 2016 , https://doi.org/10.1109/ACCESS.2016.2628102.

[41] Krzysztof Gajowniczek , Tomasz Ząbkowski , "ImbTreeAUC: An R package for building classification trees using the area under the ROC curve (AUC) on imbalanced datasets",Volume 15, July 2021, 100755, https://doi.org/10.1016/j.softx.2021.100755.

[42] Farrukh Aslam Khan, Khan Zeb, Mabrook Alrakhami, Abdelouahid Derhab , " Detection and Prediction of Diabetes Using Data Mining A Comprehensive Review",vol9 IEEE Access PP(99):1-1, https://ieeexplore.ieee.org/document/9354154.