SCIREA Journal of Information Science and Systems Science
ISSN: 2995-3936
http://www.scirea.org/journal/ISSS
January 9, 2024
Volume 8, Issue 1, February 2024
https://doi.org/10.54647/isss120327

Hybrid decision tree-based machine learning models for diabetes prediction

Efijemue Oghenekome Paul

Department of Computer Science, Austin Peay State University, Clarksville USA

Abstract

Due to the ever-increasing incidence of diabetes, effective screening strategies are needed for early diagnosis and intervention. This study proposes a novel approach that harnesses the power of artificial intelligence (AI) to predict diabetes risk. Using machine learning techniques and a database with demographic, clinical and lifestyle variables, the proposed model achieves the best accuracy in predicting the probability of developing diabetes. The prediction model uses advanced feature selection and cross-validation techniques to improve reliability and generalizability. Integrating AI into diabetes prediction paves the way for earlier healthcare, enabling personalized intervention and ultimately reducing the burden of diabetes on individuals and healthcare systems.

KEYWORDS: Diabetes prediction, Machine learning, Incidence of diabetes, Integrating AI into diabetes

Chapter one: Introduction

1.1 Introduction

Health professionals have many websites. Such a database may contain structured, semi-structured or unstructured data. Big data analytics involves analyzing large amounts of data and searching for hidden features to extract information from a given set of data. Considering the current scenario, diabetes mellitus (DM) has become a very serious disease in developing countries. According to the statistics of 2017, around 425 million people are suffering from diabetes every year, 2-5 million patients die from diabetes. This is projected to rise to 629 million by 2045 (Gauri et al 2017). Diabetes mellitus (DM) is classified as type-1, known as Insulin-Dependent Diabetes Mellitus (IDDM). The reason behind this type of DM is that the human body cannot produce enough insulin so the patient has to be injected with insulin. Also known as type-2 non-insulin-dependent diabetes mellitus (NIDDM). This type of diabetes occurs when the body's cells do not use insulin properly. Gestational Diabetes Type-3 This type of diabetes is caused by undiagnosed high blood sugar levels in pregnant women. DM has been associated with chronic problems. Also, a person with diabetes can have health problems. The machine learning approach includes various machine learning techniques, data mining algorithms, and algorithms that use current and past data to analyze and predict future events. By applying predictive analytics to healthcare data, important decisions and predictions can be made. Predictive analyzes can be done using machine learning and regression method. The goals of predictive analytics are to improve diagnosis, improve patient care, improve clinical outcomes, and optimize resource management (Gauri et al. 2017). Machine learning is considered to be the most important characteristic of artificial intelligence that supports computing systems that can draw from past experience without having to program in any case. Machine learning is considered necessary in today’s environment to eliminate human efforts by supporting automation with minimal errors. The current method of diagnosing diabetes uses laboratory tests such as fasting blood glucose and oral glucose tolerance. However, this method is time consuming. This paper describes how a prediction model using machine learning techniques and data mining techniques was developed to predict diabetes.

1.2 Problem statement

The prevalence of diabetes is a major health problem worldwide. Early diagnosis of diabetes in humans is important for early intervention and effective control of the disease. Existing risk assessment tools often rely on a limited number of variables and may not have the precision required for an accurate assessment. The goal of this research is to address these limitations by developing a predictive model that harnesses the power of artificial intelligence (AI). The model includes several demographics, clinical, and lifestyle variables to improve the accuracy of diabetes risk assessment. Thus, this study contributes to the development of preventive healthcare strategies by enabling healthcare providers to identify high-risk individuals early and implement appropriate prevention strategies, thereby reducing the severity of diabetes, leading to a reduction in population and health services.

1.3 Background of the study

Diabetes is considered one of the leading causes of death worldwide. If diabetes is not treated or detected early, it can lead to many complications. The aim of this study was to develop a model that could better predict the likelihood of patient satisfaction with diabetes. Classification methods are widely used in medicine to divide data into different groups according to certain criteria that do not allow individual classification. Thus, four machine learning classification algorithms, namely supervised learning algorithms (Random Forest, SVM and Naive Base, DT Decision Tree) and unsupervised learning algorithms (k-means), provide a way to detect diabetes at an early stage. The tests are carried out in two databases: one from a hospital in Frankfurt, Germany, and the other downloaded from the database. PIMA Indian Diabetes (PIDD) from the UCI Institute of Technology. Data from a clinic in Frankfurt, Germany showed that the Random Forest algorithm performed with the highest accuracy of 97.6%, and results from Indian PIMA datasets showed that the SVM algorithm outperformed the resolution of 83.1% with the highest accuracy of 97.6%. Compared to other algorithms. These results were confirmed by splitting the data into two groups: training and testing, described below. The training set is used to refine the model. An experimental setup is used to test the model and verify its accuracy.

1.4 Research objectives

i. To analyze the diabetes dataset
ii. To evaluate performance of the model
iii. To evaluate the model's sensitivity to changes in input features

1.5 Research questions

i. What is the visualization of the diabetes dataset?

ii. What will be the performance of the predictive model?

iii. How robust is the model to feature changes?

1.6 Significance to the field

Over the past decade, the number of people with diabetes has increased significantly. Modern lifestyle is the main cause of diabetes. There are three types of possible errors in current medical diagnosis.

i. False negative, in which the patient already has diabetes, but the test results show that they do not have diabetes.

ii. False-positive. With this method, the patient is not actually diabetic, but the test results show that they are diabetic.

iii. The third type is unclassified, in which the system does not recognize a specific disease.

This is due to insufficient knowledge of previous cases and the patient in question may be considered an unclassifiable patient. But in fact, the patient must indicate whether he belongs to the diabetic or non-diabetic category. Such diagnostic errors can lead to unnecessary treatment or no treatment at all. To avoid or reduce this effect, it is necessary to develop a system using machine learning and data analysis techniques that improve the accuracy of results and reduce human effort.

1.7 Limitations of the study

The model built in this research is limited to predicting medical conditions of diabetes only and cannot be used to predict other medical conditions.

Chapter two: Literature review

2.1 Introduction

Related Fields Analysis provides results based on a range of health data that used various methods and techniques to perform analysis and make predictions. Different researchers have developed and implemented various forecasting models using variations of data mining techniques, machine learning techniques, or combinations of these techniques. Saravana et al. (2015) developed a framework using Hadoop and map reduction techniques to analyze

diabetes data. This table identifies the type of diabetes and its associated risk factors. The system is based on Hadoop and is cost-effective for any healthcare organization. Aishwarya (2015) used a classification method to explore hidden aspects of diabetes data. This example uses Naive Bayes and decision trees. The performance of the two systems is compared and thus the effectiveness of the two systems is determined. Rajesh et al. (2012) used a classification method. The C4.5 decision tree structure was used to find the latent dimensions in the database and classify them correctly. Hammer et al. (2008) used an artificial neural network (ANN) with fuzzy logic to predict diabetes. Patil et al (2010) proposed a prediction model coupled with a simple K-means clustering algorithm followed by a classification algorithm applied to the results obtained by the clustering algorithm. C4.5 A decision tree algorithm is used to construct classes. Mani et al. (2015) presented a model using a random forest classifier to represent diabetic behavior. Nawaz et al. (2011) used C4.5 decision tree method, neural networks, K-means clustering algorithm and visualization to predict diabetes. The process of selecting a machine learning algorithm involves matching the type of data to be learned with the existing structure. The classifications of machine learning algorithms are described below. There are many machine learning techniques that fall into three categories: supervised learning, unsupervised learning, and semi-supervised learning.

2.2 Supervised learning

According to Michael et al (2017), supervised machine learning (SML) is the search for algorithms that can generate general hypotheses that predict future events based on external data. Classification of controlled tasks is one of the most frequently performed tasks in intelligent systems. According to Taiwo et al (2010), some effective machine learning algorithms for supervised classification are: linear classifiers, logistic regression, naive Bayes classifier, perceptron, support vector machine; Quadratic classifiers, K-means clustering, boosting, decision tree, random forest (RF); Neural networks, Bayesian correlations and much more.

2.2.1 Decision trees

Decision trees (DT) are trees that classify events by ordering them by attributes. Each decision tree node represents an attribute of the model to be classified, and each branch represents a possible value of the node. Models are trained from root nodes and sorted by their attribute values. Michael et al. (2017). Decision tree learning, used in data mining and machine learning, uses the decision tree as a predictive model that learns information about a factor and maps it to determine the target value of the factor in that representation. The most

descriptive names for these tree models are partition trees or regression trees. Backward pruning methods are commonly used to classify decision trees, which evaluate the pruning performance of decision trees using different types of evidence. Each node is extracted and assigned to locally distributed regular training samples.

2.1.2 Multi-layer perceptron

This is a class of network weights obtained by solving a quadratic programming problem with linear constraints instead of solving an unconstrained, nonconvex minimization problem as in conventional neural network training. Other popular algorithms are based on the concept of the perceptron. The perceptron algorithm is used to learn from a set of training samples by running the algorithm repeatedly through the training set until it finds the best prediction vector in the learning set. These prediction rules are used to obtain prediction scores in Michael et al. (2017).

2.2.3 Support vector machines (SVM).

This is a recent supervised machine learning approach developed by Michael et al (2017). Support vector machine (SVM) models are closely related to classical multi-layer perceptron neural networks. SVMs revolve around the concept of a "stack" on each side of a hyperplane separating two sets of data. Increasing the margin, and therefore the maximum distance, between a particular hyperplane and the samples at its two ends reduces the upper limit of the normalization error adopted in the 19th century.

2.2.4 Random forest classifiers

According to Siddharth et al. (2020), RF classifier is an ensemble method that trains multiple decision trees in parallel with bootstrapping followed by aggregation, collectively called bagging. Bootstrapping showed that multiple individual decision trees were trained side by side on different subsets of training data using different available subsets. Bootstrapping ensures that each decision tree in the random forest is unique, reducing the overall variance of the RF classification. In the final decision, an RF class combines the decisions of individual trees; therefore, the RF classes show good generalization. The RF classifier usually outperforms most other classification methods in terms of accuracy without any problems. Like the DT class, the RF class does not require feature scaling. Unlike the DT classifier, the RF classifier is more robust in selecting noisy training samples in the training data set. The RF classifier is more difficult to define, but the hyperparameter is easier to construct compared to the DT classifier.

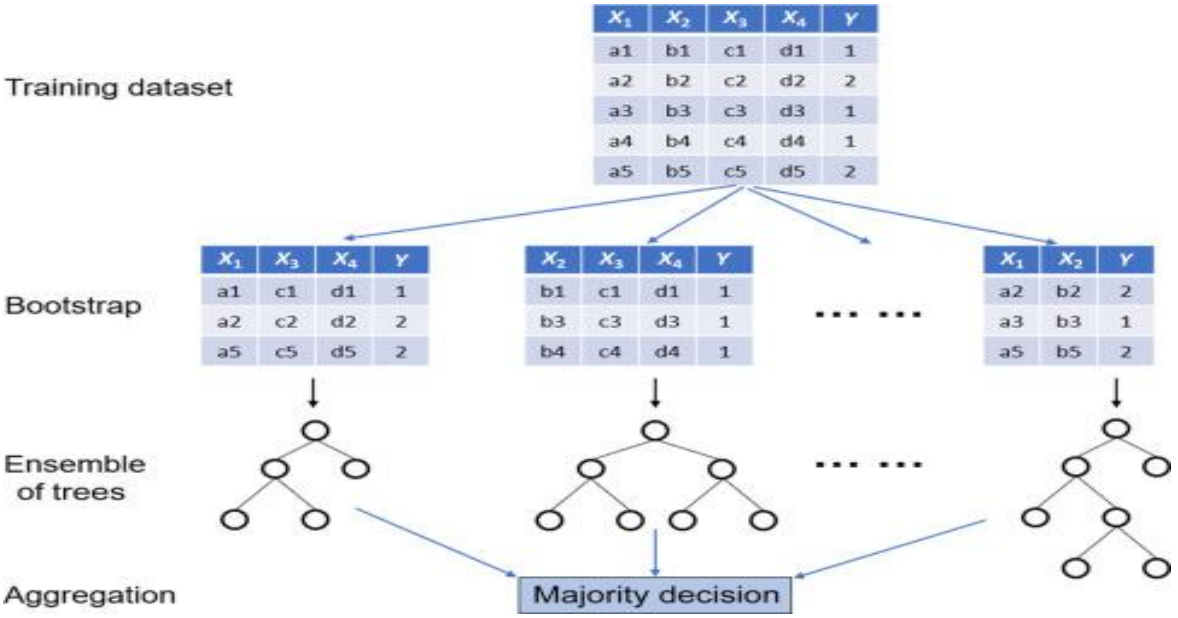

Figure 2.1 Random Forest classifier

2.2.5 Logistic regression

It is a multinomial logistic regression model and classification function that uses learning to generate statistics. Logistic regression generally shows where the boundary between classes is and also shows that the probability of a class depends on the distance to the boundary. As the amount of data increases, it quickly changes to extreme values (0 and 1). This probabilistic information makes logistic regression more than just classification. It contains clear, detailed and sensitive instructions; But these strong predictions might not be true. Logistic regression is an estimation method similar to ordinary least squares (OLS) regression. However, in logistic regression, the prediction leads to biased results by Newsom et al. (2015) and the author. Logistic regression is one of the most used tools in applied statistics and discrete data analysis. Logistic regression is simple. According to Priyanka et al. (2021), logistic regression has proven to be one of the effective algorithms for building prediction models. The accuracy of the model depends not only on the chosen algorithm, but also on other factors. Data preprocessing is one such aspect. Removing redundant and null values is essential to improve performance. Normalization of values also plays an important role when characteristics vary on a large scale. In Priyanka et al. (2021) paper, they used logistic regression algorithm in two different ways which gave them impressive results. Their logistic regression was based on featured selection and univariate feature selection. These two methods gave them impressive performance report as shown in figure 2.1 and figure 2.2.

Classification report for dataset 1 after Feature Selection.

	Precision	Recall	F1 score	Support
0	0.81	0.82	0.81	151
1	0.65	0.62	0.64	80
Accuracy			0.75	231
Macro Avg	0.73	0.72	0.73	231
Weighted Avg	0.75	0.75	0.75	231

Figure 2.2 classification report for feature selection

Classification report for dataset 2 after univariate feature selection.

	Precision	Recall	F1 score	Support
0	0.89	0.98	0.93	93
1	0.87	0.54	0.67	24
Accuracy			0.89	117
Macro Avg	0.88	0.76	0.80	117
Weighted Avg	0.89	0.89	0.88	117

Figure 2.3 classification report for univariate feature selection.

2.3 Semi-supervised learning

Semi-supervised learning (SSL) is a form of machine learning (ML). This falls somewhere between controlled and uncontrolled studies. The main goal of SSL is to overcome the problems of supervised and unsupervised learning. Supervised learning requires more training data to classify the test data, which is expensive and time-consuming. On the other hand, unsupervised learning does not require labelled data, its groups data into similar data points using a clustering method or maximum likelihood method. The main disadvantage of this method is that it does not accurately incorporate unknown information. To overcome these problems, the research team proposed SSL which can learn from small training data and unlabeled (or) testing data. SSL creates instances, some of which are identified as training data and the rest of which are treated as test data.

2.3.1 Semi-supervised clustering

According to Vishwanath (2018), a semi-managed team is simple teamwork. Generally, when clustering, we use unlabeled data models for grouping. But in semi-supervised clustering, we use labelled and unlabeled data as well as secondary data in such a way that we have pairwise constraints (which must be correlated and uncorrelated) that affect the properties of the cluster data. The semi-supervised single link layer (SSL) approach solves the hierarchical layer state problem of Reddy et al. (2016). SSL solves the problem of noise bridging, i.e., distance

between layers, by considering a predefined distance matrix with certain restrictions. The authors tested the results on synthetic and real data sets. The self-learning method is a widely used method in SSL. In this approach, the algorithm separates some labelled training data and then unlabeled data and then adds these prediction models to the training set. The process is repeated until the test suite is empty. Some algorithms attempt to "unlearn" or move beyond an unlabeled point when the estimated features of the data are below a threshold. Self-processing has been used in applications such as natural language processing (NLP) projects. Support vector machines are a standard class that uses only labelled data, while TSVMs use unlabeled data. TSVM is an extension of SVM designed to assign labels to normally distributed unlabeled data samples and increase the variance of the original labelled data samples for unlabeled data. TSVMs are very popular and are used in many applications such as image retrieval, bioinformatics, and name recognition. The probabilistic framework was proposed by Basu et al. (2004) for a semi-controlled group. The authors minimize an objective function derived from the background energy of hidden Markov fields (HMRF). Its structure, presented in a multi-information repository, illustrates the benefits of semi-supervised learning.

2.4 Unsupervised learning

Unsupervised learning studies how systems can learn to express certain inputs in a way that reflects the overall mathematical structure of the inputs. Unlike monitoring or validation studies, there are no clear objectives or ecological expectations associated with the input data; Instead, unsupervised learning exposes existing biases about which parts of the input system should be preserved in the output. Unsupervised learning is important because it can occur more frequently in the brain than unsupervised learning. For example, each eye contains approximately 106 cameras whose functions constantly change with the visible world and provide general information about recognizing objects in the world, exposure, lighting conditions, etc. Adult growth and plasticity are important for animal vision. Indeed, the structural and anatomical properties of neocortex synapses strongly influence the organization of nascent sensory neurons. However, the study contained virtually no information on incidence. This makes uncontrolled mechanisms important and allows them to be used as mathematical models of synaptic adaptation. The only objects that unsupervised learning methods need to deal with are sequences of observed inputs xi, which are generally considered to be independent samples starting from an unknown probability distribution PI[x] and in some cases emergent or a priori significant. The basic assumption is that inputs, such

as scene images, have their own peripheral causal effects, such as objects located in certain regions being illuminated by a certain light. Since we naturally have to work with these highly dependent factors, the best indicator of a process is its expression. Two different approaches to unsupervised learning have been proposed. Density estimation methods explicitly create mathematical models (BAYESIAN NETWORKS) that describe how objects form the input data. Feature extraction methods attempt to extract statistical patterns (or sometimes anomalies) directly from input data.

Chapter three: Methods

3.1 Introduction

This chapter provides an overview of compound development strategies for diabetes detection using machine learning-based decision trees. This study used machine learning techniques such as unsupervised and supervised machine learning and training methods. The researcher relied on the Pima data model (dataset) developed by National Institute of Diabetes and Digestive and Kidney Diseases for the analysis and results.

3.2 Methodology

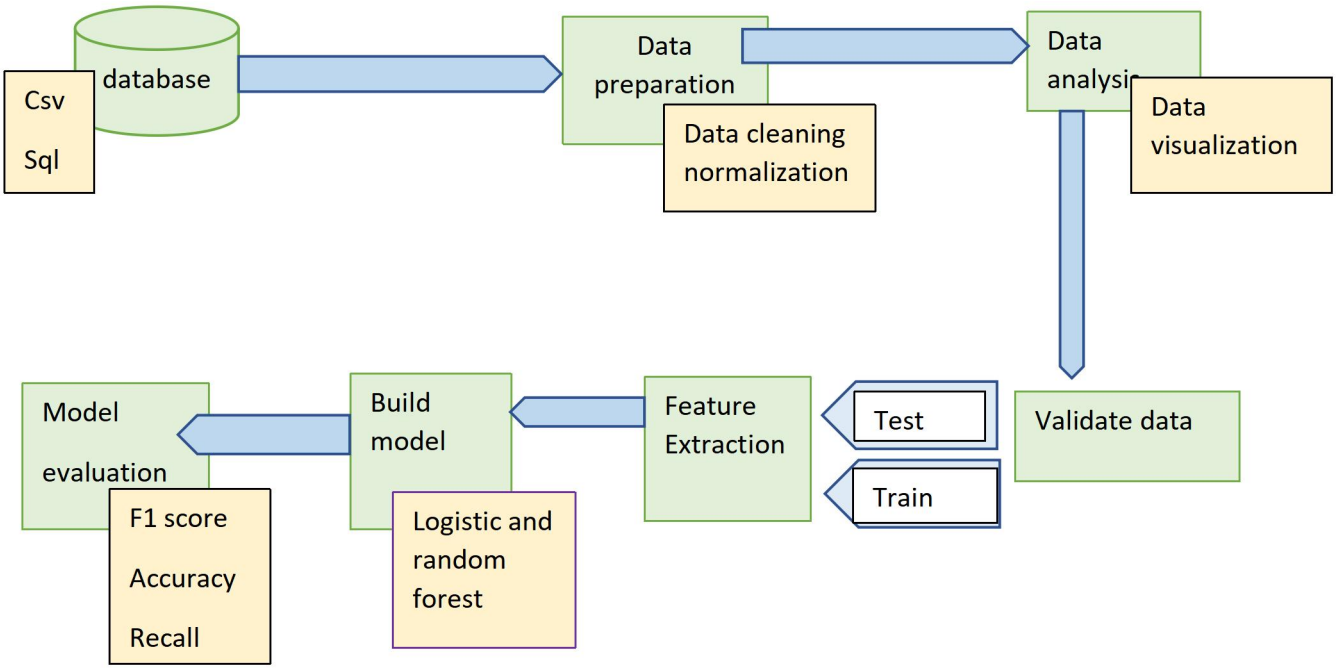

Figure 3.1 Architectural framework for hybrid model

3.2.1 Database (Pima Dataset)

Data on 768 women near Phoenix, Arizona, United States form the PIMA India Diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. The test result was diabetes: 258 positive and 500 negatives. Thus, the following variables and parameters are available:

• Pregnancy (frequency),

• Oral glucose tolerance test - OGTT (glycemia two hours after ingestion of 75 g of anhydrous glucose per mg/dL).

• High blood pressure (blood pressure mmHg),

• Body weight (triceps skinfold thickness, mm),

• Insulin (2-hour plasma insulin in internal units/ml),

• BMI (body mass index kg/m2),

• years and years),

• Diabetes genetic function (“a function that indicates the likelihood that your ancestors had a history of the disease”);

The Pima are descendants of people who lived for centuries in the Sonoran Desert and Sierra Madre region. Around 300 BC e. At that time, they settled in the Gila River Valley of Mexico, but in 1853 this territory was incorporated into the United States. They founded the Pima Reservation in Arizona in 1959 and converted their wilderness space in the 19th century to force water to flow from it. This includes livelihoods. agriculture in 1900, the number of white settlers increased and they diverted water. This affected the Pima's diet and lifestyle. The Pima Indians practice sustainable agriculture using manual labor with little labor and food. As a result, their diet is high in fat and their lifestyle is largely sedentary. This led to diabetes among the Pima people of Arizona, who have received attention because they have the highest prevalence and prevalence of type 2 diabetes (T2D) among a geographically defined population (SCHULZ et al., 2015). The National Institute of Diabetes and Digestive and Kidney Diseases studied the Pima population over a two-year period beginning in 1965. Since epidemiologic evidence suggests that T2D is caused by genetic and environmental interactions, all data on Pima Indians with diabetes included qualitative information. The onset of diabetes and future complications must be linked. Chandra-Selvi et al (2016) defined adipose tissue as "a loose connective tissue composed of adipose cells with virtually unlimited

growth potential at every stage of life, producing excess adipose tissue that is affected by many diseases – the development of insulin resistance. In diabetic patients, as the disease progresses, the thickness of the skin, study showed a gradual decrease. However, Indian PIMA diabetes data does not provide information on diabetes duration. The researchers analyzed the results and concluded that free fat is produced in adipose tissue in diabetes. in diabetic patients, leading to a decrease in skin mass as the disease progresses (Chandra-Selv. et al., 2016), but there are still few prospective reports examining the relationship between skin lesions and the development of T2DM. Triceps skinfold thickness in millimeters, median, standard error of mean, race, ethnicity, and selected age percentiles: United States, 2007-2010. Obesity, as measured by body mass index (BMI), is closely associated with diabetes, and its impact on the development of T2DM has been widely discussed in large cohort studies (Van Gaal et al, 2015). Indeed, many people with T2D are overweight or obese. Despite the link between obesity and T2D, not all obese people develop diabetes and not all diabetics are obese. In healthy people with diabetes, the genetic component of T2D may be stronger than in people who are overweight or obese (Wilding, 2014). Consequently, there is no recommendation for the use of anthropometric measures other than BMI and waist circumference as risk factors for T2DM. Unfortunately, waist circumference is not included in diabetes data in India. BMI provides a simple and accurate way to measure nutritional status. It can be calculated by dividing a person's weight (in kilograms) by the square of their height (in meters).

3.2.2 Data preparation

Most technical methods require some organization of data; Therefore, websites usually require some preparation to provide useful information. Some databases contain missing, incorrect, or programmatically complex values. Without data, the algorithm cannot be used. If the information provided is incorrect, the statistics will be incorrect or misleading. Some databases are relatively clean but require sorting (e.g. compiling or converting), and many databases simply don't have a useful workflow (e.g. incorrectly defined identifier values), the performance must therefore be improved. Proper data preparation results in cleaner, more organized data, leading to more sensitive and accurate model results.

3.2.3 Data visualization

Data visualization is the process of presenting information using simple techniques such as graphs and charts. These visuals convey complex data relationships and data-driven insights in an easy-to-understand way. It is important to note that data visualization can be used for a

variety of purposes and is not limited to data structures. Managers also use it to define organizational policies and procedures, and researchers and data scientists use it to analyze and interpret patterns and models.

3.2.4 Data validation

In machine learning, we try to learn patterns present in datasets and generalize those lessons. This puts data at the forefront of our machine learning applications, and data quality is critical to the success of our machine learning projects. Each step in our machine learning pipeline determines whether the task can proceed to the next step or whether the entire process needs to be stopped and restarted (e.g., with additional training data). Data validation is a very important area of testing because it detects changes in data entering the machine learning pipeline before it reaches the tedious pre-processing and training stages.

Validating our data is essential if we aim to truly improve our machine learning models. Essentially, when we talk about verification, we're talking about three different ways of looking at your data:

• Check if the data is incorrect.

• Make sure the data structure has not been modified.

• Ensure that predictions from our new datasets continue to match those from our previous training datasets.

The data validation stage of our engine performs this check and highlights any errors. If a failure is detected, we can stop operations and resolve the data problem manually, for example by creating a new database.

The data validation stage is also called the data processing stage, which is the next step in our pipeline. Data validation evaluates the quality of your data and determines if a feature has a high percentage of missing values or if the features are highly correlated. This is useful information for deciding which conditions to include in corrective actions and which condition to fulfill.

3.2.5 Splitting data into test and train

In machine learning, splitting our sample data into training and testing sets is probably one of the first steps before preprocessing them. Creating different models for training and testing helps us evaluate the performance of the model. Many common problems arise when training models. This phenomenon occurs when a model performs well with the data, we train it on,

but fails to reach new data points that we haven't actually seen. This could be due to noise in the data or because the model learned to predict specific inputs instead of estimating parameters to make better predictions. In general, the more complex a model is, the more likely it is to overfit. On the other hand, underfitting occurs when the model performs poorly even on training data. Inconsistencies often arise because the model is not suited to a particular task. Overall, this means that the model is not complex enough to predictably model useful thresholds.

3.2.6 Feature extraction

Feature extraction refers to the process of transforming raw data into useful digital assets while preserving the original database information. This produces better results than directly applying machine learning to raw data. Feature extraction can be done manually or automatically:

• Manual characterization involves identifying and defining the characteristics of the problem and developing strategies to eliminate these characteristics. In many cases, a better understanding of context or genetics can help you make informed decisions about profitable activities. Over decades of research, engineers and scientists have developed methods to extract features from images, symbols and text. An example of a simple object is the center of a window on a label.

• Automatic extraction uses specialized algorithms or deep networks to automatically extract features from signals or images without human intervention. This approach is particularly useful if you want to quickly get started developing machine learning algorithms from raw data. Wave propagation is an example of automatic feature extraction.

With the advent of deep learning, most feature extraction has been replaced by the underlying layers of deep networks, especially for image data. In signal and time series processing, feature reduction remains a fundamental problem that requires specialized knowledge to build effective predictive models.

3.2.7 Build model

Modeling is at the heart of data analysis and is used to extract information and insights from data to inform business decisions and strategies. During this phase of the project, the data team must create databases for training, testing, and product development. These databases allow data scientists to develop and train an analytical approach while storing some of the data for testing models. The goal of data analytics modeling is not only to achieve high

accuracy on training data, but also to be comprehensive and effective for new and emerging data. Therefore, the focus is on developing models that can capture patterns and relationships in the data rather than memorizing training data.

3.2.8 Model evaluation

Model evaluation involves using various benchmarks to understand the performance of a technical model, as well as its strengths and weaknesses. Model calibration is important for evaluating model performance at the beginning of a study and also plays a role in model validation. There are many analytics metrics you can use to understand how your models perform with new data. Accuracy measures how often the classifier makes correct predictions, as it is the ratio of the number of correct predictions to all predictions. Accuracy measures the proportion of truly good predictions. Precision is a good strategy when testing metrics if you want to have more confidence in your predictions. For example, if you are developing a chart to estimate whether you will reduce your credit limit for a certain account, you need to be very confident in your estimate, otherwise it will make customers unhappy. The confusion matrix (or confusion table) shows a detailed breakdown of correct and incorrect classifications for each category. Using a confusion matrix is useful when you want to understand differences between studies, especially when two classes may have different classification costs or when you have a lot of test data for one class. For example, the consequences of a false positive or false negative result for a cancer diagnosis are very different.

Chapter four: Result

4.0 Introduction

This chapter illustrates the applicability of the proposed anomaly detection methods to detect botnets in industrial IoT networks in an IoT environment and their potential to answer the three research questions discussed. This section describes the experimental / testing process and the results / implications.

4.1 Glucose tolerance test

The glucose tolerance test (GTT), also known as the oral glucose tolerance test (OGTT), is a method that helps diagnose diabetes or insulin resistance. This test is used to determine if the body is having difficulty absorbing glucose/carbohydrates. The patient is asked to drink a

sugary drink and blood sugar is measured before and after drinking the sugary drink. To take the test, you first take a blood test to measure your blood sugar before the test. The next step is to drink a very sweet and sugary drink. Subsequent blood samples were routinely collected 30, 60 minutes, or 2 hours later in a single study. The test can take up to 3 hours.

People without diabetes

• Fasting value (before the test): less than 6 mmol/L

• Within 2 hours: less than 7.8 mmol/L

People with resistant diabetes mellitus (IGT).

• Fasting value (pre-test): 6.0 to 7.0 mmol/L

• At 2 hours: 7.9 to 11.0 mmol/L

The amount of sugar in the body

• Fasting value (before the test): greater than 7.0 mmol/L

• Within 2 hours: more than 11.0 mmol/L

If you have hypoglycemia, you may be advised to make lifestyle changes. In some cases, medications that lower blood sugar may be prescribed.

If you have diabetes, you may be prescribed glucose medications to help your body regulate blood sugar levels. We conclude that insulin sensitivity and secretion can be predicted with reasonable accuracy from populations matching values obtained during OGTT. The derived equations should be applied to clinical situations where the use of forceps or minimal sampling is not practical.

4.2 Visualization of pima diabetes datasets

Data visualization is the representation of data using simple visual aids such as graphs, charts, reports, and tables. These visualizations intuitively communicate complex data relationships and data-driven meanings. This section views the pima database to derive information.

4.2.1 Nutritional status of the pima dataset

According to Somayyeh et al (2015), the average age of the subjects was 56.7 ± 9.9 years and the average duration of diabetes was 6.5 ± 5.0 years. Subjects' mean hemoglobin A1c was 7.6% ± 1.4%, with only 20.2% achieving the target value of <6.5%, with no significant difference between genders. Mean body weight was 26.9 ± 4.7 kg/m2 and 86.5% were

overweight or obese. The Jupyter code below adds the parent attribute of each row in the database.

```
for i in range(len(pima)):
    if pima['BMI'][i] == 0.0:
        Nutritional_status[i]="NA"

    elif pima['BMI'][i] < 18.5:
        Nutritional_status[i]="Underweight"

    elif pima['BMI'][i] < 25:
        Nutritional_status[i]="Normal"

    elif pima['BMI'][i] >= 25 and pima['BMI'][i] < 30:
        Nutritional_status[i]="Overweight"

    elif pima['BMI'][i] >= 30:
        Nutritional_status[i]="Obese"

    else:
        Nutritional_status[i]= pima['BMI'][i]
```

The pima dataset was analyzed to ascertain the nutritional status of each patient and the figure below shows the nutritional status.

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

Figure 4.1 nutritional status of pima dataset

Figure 4.2 shows the aggregated results of the nutritional content of the pima dataset.

```
Obese          472
Overweight     179
Normal         102
NA              11
Underweight      4
Name: Nutritional Status, dtype: int64
```

Figure 4.2 aggregated nutritional status of the dataset

4.2.2 Glucose level in the pima database

• Normal blood sugar is less than 140 mg/dL (7.8 mmol/L).

• A blood sugar level between 140 and 199 mg/dL (7.8 and 11 mmol/L) is considered hypoglycemia or prediabetes. If the patient has prediabetes, you are at risk of eventually developing type 2 diabetes. Patients can have a heart attack even without diabetes.

• A blood sugar level of 200 mg/dL (11.1 mmol/L) or higher may indicate diabetes.

If the patient has type 2 diabetes, the doctor may repeat the test later or use another blood test to confirm the diagnosis. The accuracy of a glucose tolerance test can be affected by several factors, including illness, activity levels, and certain medications. The code below helped in getting the glucose level of each patient in the dataset.

```
for i in range(len(pima)):
    if pima['Glucose'][i] == 0.0:
        OGTT_Interpretation [i]="NA"

    elif pima['Glucose'][i] <= 140:
        OGTT_Interpretation [i]="Normal"

    elif pima['Glucose'][i] > 140 & pima['Glucose'][i] <= 198:
        OGTT_Interpretation [i]="Impaired Glucose Tolerance"

    elif pima['Glucose'][i] > 198:
        OGTT_Interpretation[i]="Diabetic Level"

    else:
        OGTT_Interpretation [i]= pima['Glucose'][i]
```

```
Normal                          571
Impaired Glucose Tolerance      192
NA                                5
Name: Glucose Result, dtype: int64
```

Figure 4.3 OGTT level of pima dataset

4.2.4 Skin thickness

Andrea et al. (2020) found a strong association between subscapular skinfold thickness and developing T2DM and HT. Skinfold assessment can be a laboratory-free strategy to identify high-risk HT and T2DM cases. Skin thickness of the pima dataset was also determined.

Figure 4.4 below shows the skin thickness percentile of the pima dataset.

```
 0 NA                227
19 >P95th            145
11 P50th - P75th      87
9 P25th - P50th       79
15 P85th - P90th      50
13 P75th - P85th      46
17 P90th - P95th      23
8 P25th               20
7 P15th - P25th       20
3 P5th - P10th        18
12 P75th              17
1 <P5th               15
5 P10th - P15th       14
4 P10th                6
20 >P95th              1
Name: Percentile skin thickness, dtype: int64
```

Figure 4.4 skin thickness percentiles

4.2.5 Attribute distribution

Figure 4.5 illustrates the histogram charts for the interfered attributes with respect to the two-class labels. Thus, the algorithm that is capable of handling the interfered attributes correctly will make the most accurate decisions. Accordingly, in the next subsections, we will justify the performance by explaining how each technique resolved the interfered attributes.

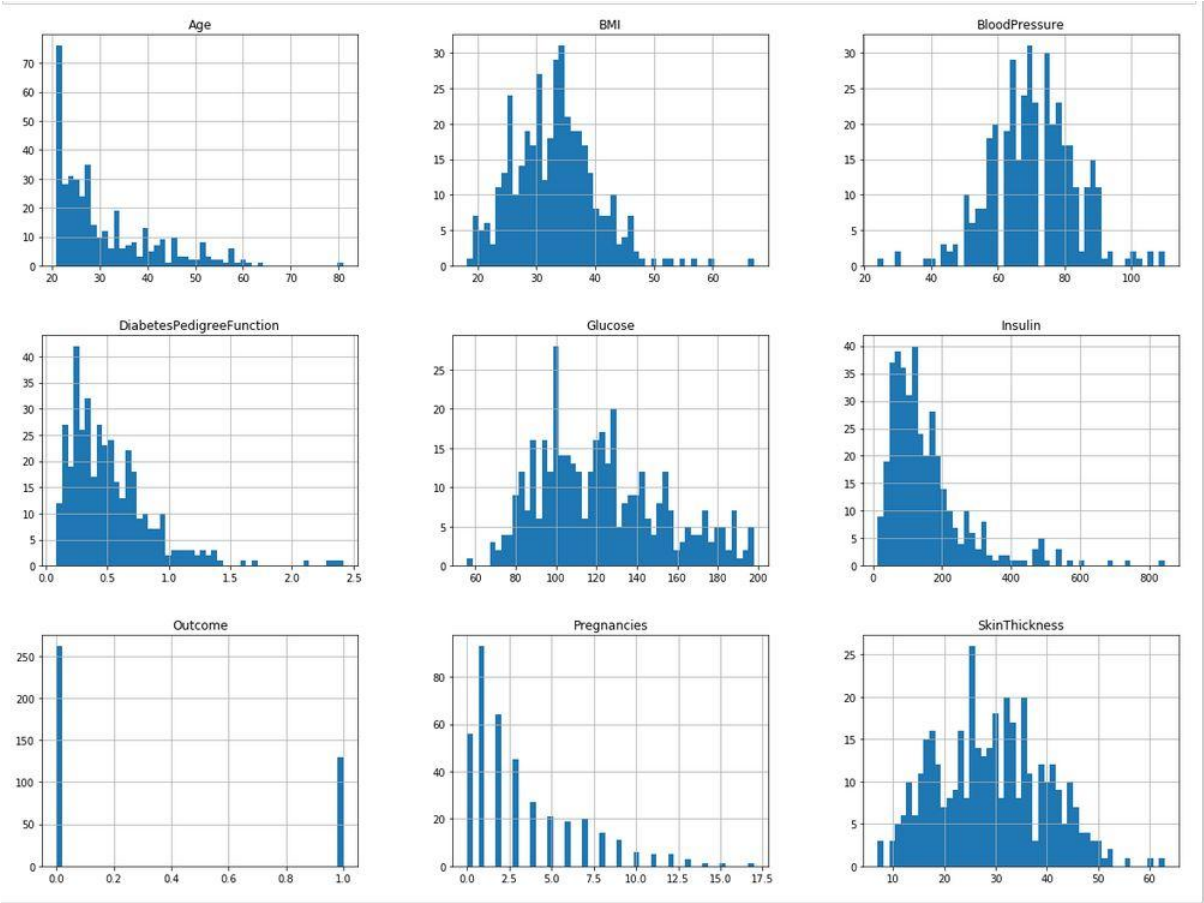

Figure 4.5 histogram of pima dataset variables

4.2.6 Density plots for all attributes

As such density plots work better at determining the distribution shape (the concept of the shape of a probability distribution arises in questions of finding an appropriate distribution to use to model the statistical properties of a population, given a sample from that population) because they're not affected by the number of bins. The most well-known form of density plot is kernel density estimation, where a continuous curve is drawn at every individual data point. Then all of these curves are sum up to make a single smooth density estimation.

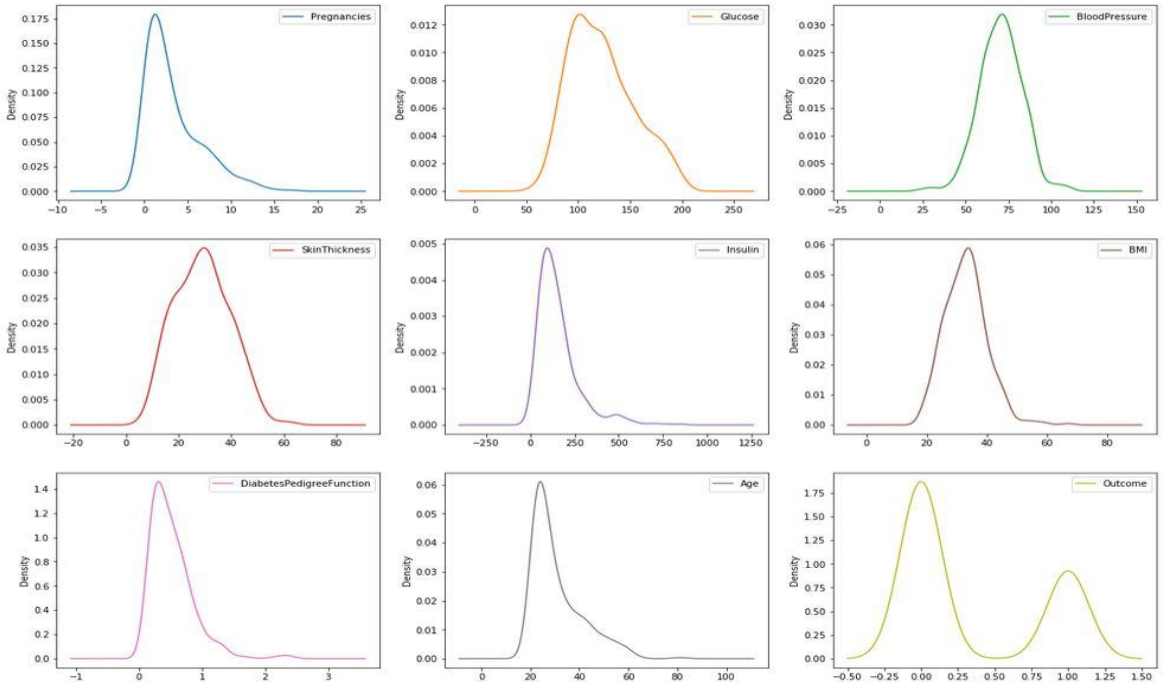

Figure 4.6 density plot of pima dataset variables

4.2.7 Correlation matrix on feature attributes

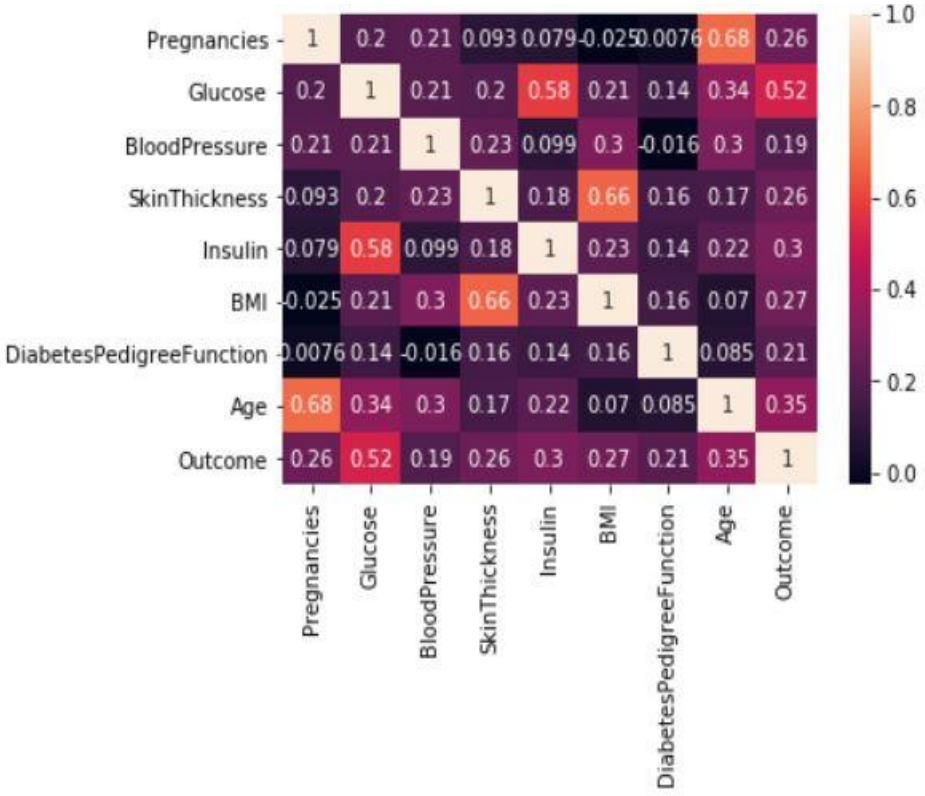

Figure 4.7 heat map of pima dataset feature variables

There is no strong correlation between the features. The 'strongest' ones are the following:

- Age x pregnancies (0.68) - Older women tend to have higher number of pregnancies
- Glucose x insulin (0.58)
- Glucose x outcome (0.52) - Women that have higher level of glucose tend to have higher level of insulin and have DM
- Skin fold thickness x BMI (0.66) - Women with higher skin fold thickness value have higher BMI (and probably are overweight/obese)

Negative correlation:

- BMI and Pregnancies (-0.025)
- Blood Pressure and Diabetes Pedigree Function (-0.016)

4.2.8 Scattered plot

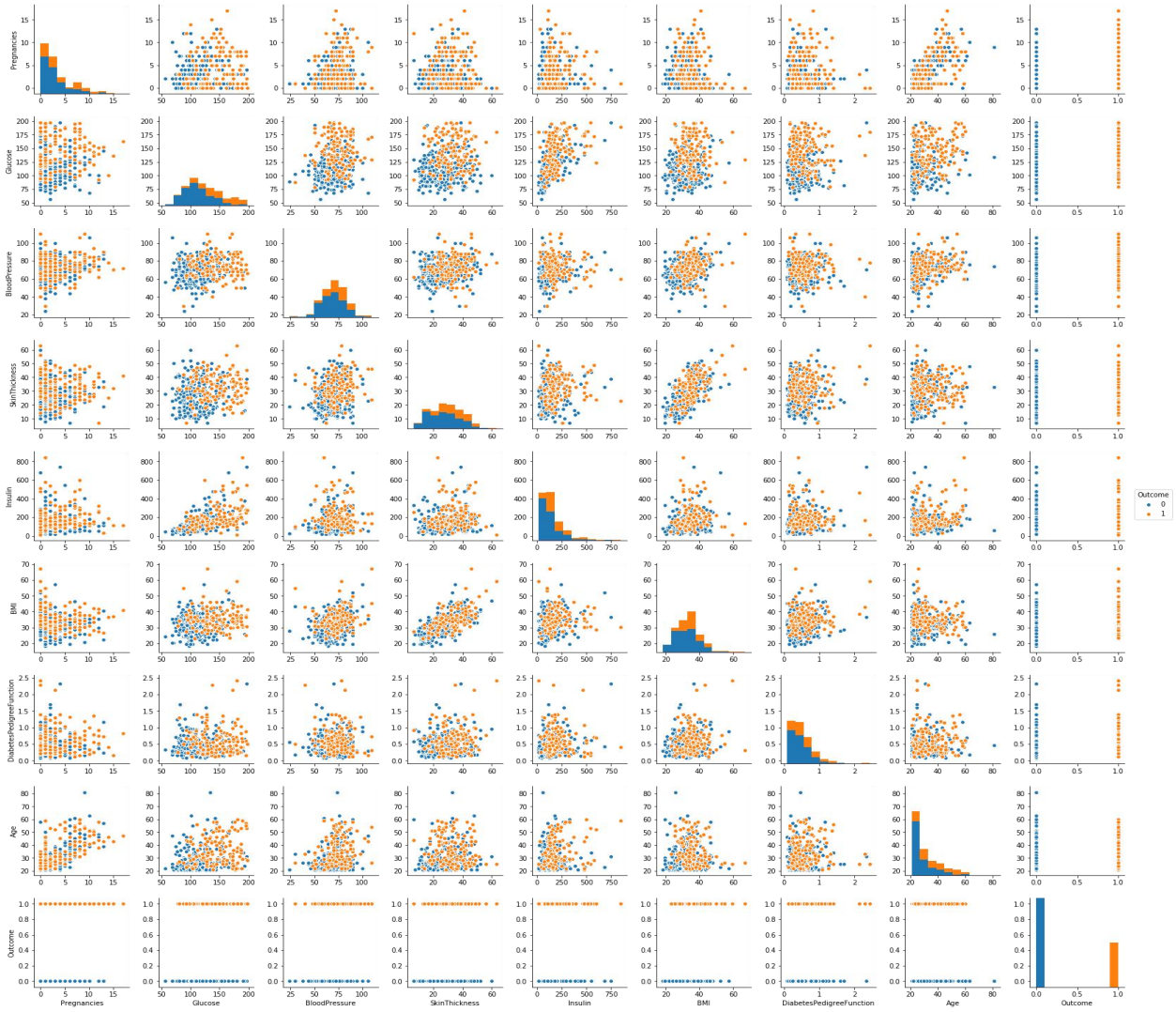

Figure 4.8 scattered plot of pima variables

One pattern of special interest is a *linear* pattern, where the data has a general look of a line going uphill or downhill. Looking at the preceding figure, you can see that a positive linear relationship does appear between the temperature and the number of cricket chirps. That is, as the temperature increases, the number of cricket chirps increases as well. Note that the scatterplot only suggests a linear relationship between the two sets of values. It does *not* suggest that an increase in the temperature *causes* the number of cricket chirps to increase

4.3 Model building using logic regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Variables used for building the model are "Pregnancies", "Glucose", "Blood Pressure".

```
Optimization terminated successfully.
         Current function value: 0.563677
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  392
Model:                          Logit   Df Residuals:                      384
Method:                           MLE   Df Model:                            7
Date:                Mon, 30 Oct 2023   Pseudo R-squ.:                  0.1128
Time:                        03:45:21   Log-Likelihood:                -220.96
converged:                       True   LL-Null:                       -249.05
Covariance Type:            nonrobust   LLR p-value:                 8.717e-10
==============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Pregnancies                    0.1299      0.049      2.655      0.008       0.034       0.226
Glucose                        0.0174      0.005      3.765      0.000       0.008       0.026
BloodPressure                 -0.0484      0.009     -5.123      0.000      -0.067      -0.030
SkinThickness                  0.0284      0.015      1.898      0.058      -0.001       0.058
Insulin                        0.0019      0.001      1.598      0.110      -0.000       0.004
BMI                           -0.0365      0.022     -1.669      0.095      -0.079       0.006
DiabetesPedigreeFunction       0.4636      0.344      1.347      0.178      -0.211       1.138
Age                            0.0005      0.016      0.031      0.976      -0.031       0.032
==============================================================================================
```

Figure 4.9 logit regression results

This model can explain 56% of the variation in dependent variable. Skin thickness, BMI, Diabetes Pedigree Function, Age, and Insulin are attributes to be eliminated in the next model. as variables that influence the outcome should have a p-value less than 0.05.

```
Optimization terminated successfully.
         Current function value: 0.574607
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  392
Model:                          Logit   Df Residuals:                      389
Method:                           MLE   Df Model:                            2
Date:                Mon, 30 Oct 2023   Pseudo R-squ.:                0.09558
Time:                        03:50:06   Log-Likelihood:                -225.25
converged:                       True   LL-Null:                       -249.05
Covariance Type:            nonrobust   LLR p-value:                 4.597e-11
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Pregnancies    0.1405      0.037      3.826      0.000       0.069       0.212
Glucose        0.0210      0.004      5.709      0.000       0.014       0.028
BloodPressure -0.0525      0.007     -7.449      0.000      -0.066      -0.039
==============================================================================
```

Figure 4.10 feature selection

This model can explain 57% of the variation in dependent variable. All the variables show a p-value < 0.005.

4.3.1 Model for dataset with missing data (model 1)

Initial Logic regression for a comparison to further LR models, this LR model is completed on the diabetes dataset that contained the missing data and also had pregnancies as a discrete variable.

Confusion Matrix report

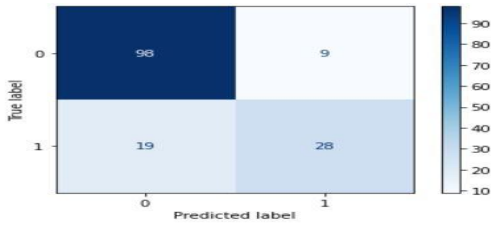

Figure 4.11 confusion matrix for model1

Model Performance

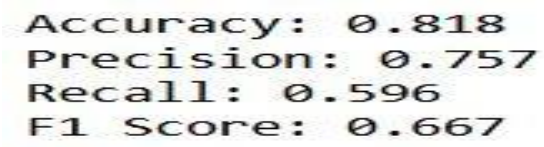

Figure 4.12 model1 performance report

ROC Curve

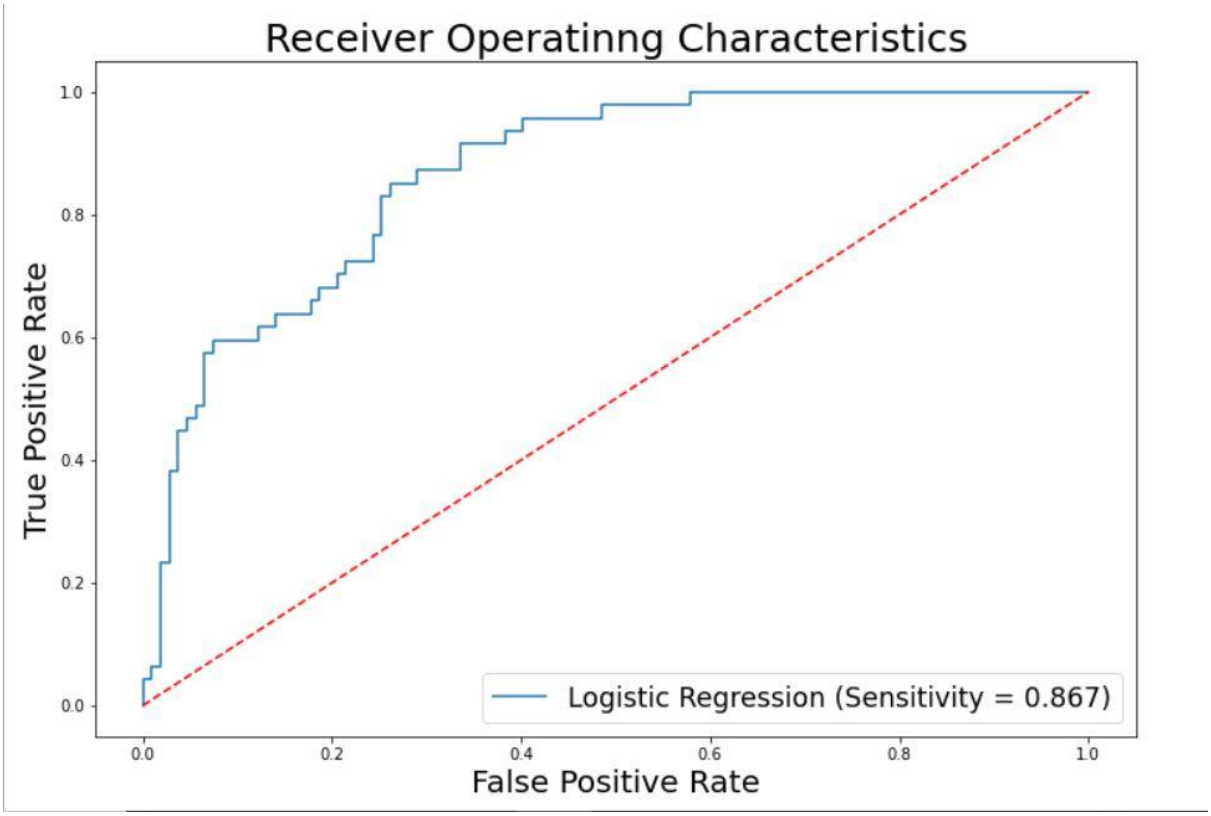

Figure 4.13 Roc curve for model1

Classification report

	precision	recall	f1-score	support
0	0.84	0.92	0.88	107
1	0.76	0.60	0.67	47
accuracy			0.82	154
macro avg	0.80	0.76	0.77	154
weighted avg	0.81	0.82	0.81	154

Figure 4.14 classification report for model 1

4.3.2 Logistic regression: hyperparameter optimization (model 2)

Hyperparameter optimization for this final logistic regression model the four significant variables will be used, but some hyperparameters will be optimized to try and improve the performance of the model in terms of diabetes prediction

Confusion Matrix

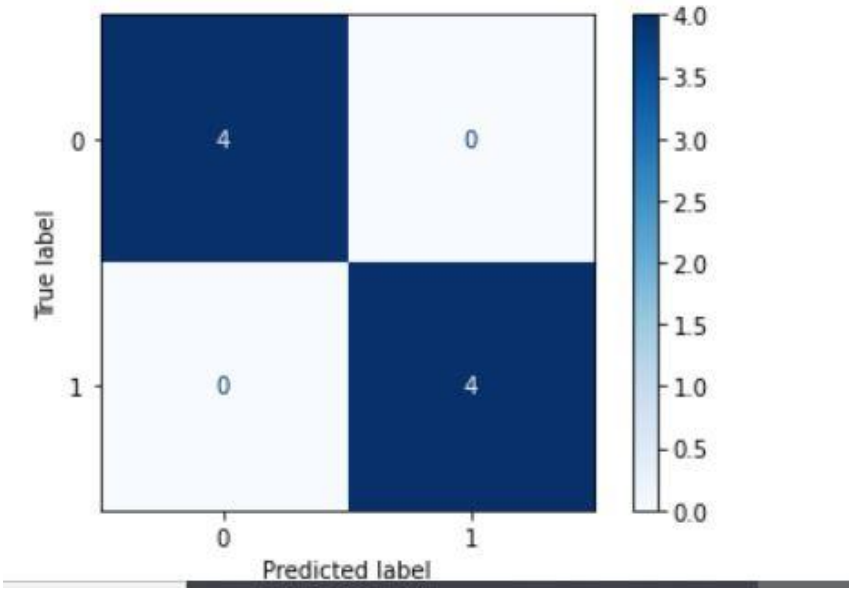

Figure 4.15 confusion matrix for model2

Model Performance

```
Accuracy: 1.000
Precision: 1.000
Recall: 1.000
F1 Score: 1.000
```

Figure 4.16 model performance for model2

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4
1	1.00	1.00	1.00	4
accuracy			1.00	8
macro avg	1.00	1.00	1.00	8
weighted avg	1.00	1.00	1.00	8

Figure 4.17 classification report for model 2

4.3.3 Hybrid tree-based classifier model

Model makes use of Random Classifier and Logistic classifier.

Model performance for hybrid tree-based

```
Accuracy: 1.000
Precision: 1.000
Recall: 1.000
F1 Score: 1.000
```

Figure 4.18 model performance for hybrid tree-based

Classification report for hybrid

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4
1	1.00	1.00	1.00	4
accuracy			1.00	8
macro avg	1.00	1.00	1.00	8
weighted avg	1.00	1.00	1.00	8

Figure 4.19 classification report for hybrid tree-based model

4.4 Model comparison

Table 4.1 comparison of two logit models with Hybrid model

	Logistic with missing data	**Hyperparameter: Logit**	**Hybrid Model**
Accuracy	0.818	1.0	1.0
Precision	0.757	1.0	1.0
Recall	0.596	1.0	1.0
F1 Score	0.667	1.0	1.0

The models are doing well. The model which included all the missing data and pregnancy variables produced 8.1% and the same logit model with hyperparameter tuning produced a perfect score with 100%. The hybrid model also produced a 100% for detection of diabetes.

Table 4.2 hybrid results compared with that Priyanka et al. (2021) results.

	Logistic (feature selection)	**Univariate feature Selection: Logit**	**Hybrid Model**
Accuracy	0.75	0.89	1.0
Precision	0.81	0.89	1.0
Recall	0.82	0.98	1.0
F1 Score	0.81	0.93	1.0

The hybrid performance report is performing better than that of the results of Priyanka et al. (2021).

Chapter five: Discussion, conclusion and Recommendation

5.1 Discussion

The model performed well and several researchers demonstrated good performance on this classification problem, thus providing a complete answer to the research question posed. The main objective of the study is to predict diabetes by developing models that will improve the accuracy and efficiency of the results. After a comprehensive review of the literature, an initial understanding of current limitations and gaps was obtained. The difference is that there is no model build using two machine learning techniques for diagnosing diabetes. Existing models use one technique for prediction. The aim of the study was to develop the most appropriate classification model for diabetes prediction. In terms of performance, all classification models performed very well.

5.2 Recommendations and future work

The research achieved its objectives and future work will focus on how to integrate the three machine learning approaches when creating hybrid prediction models.

5.3 Conclusion

The present study addresses the research objective of the study and answers the research question of how a combination of machine learning and methods can be used to predict diabetes. Clinical values used included precision, recall, accuracy, and FI. The comparative analysis adequately answers the research question by applying logistic regression in two different scenarios and applying the research objectives to the analysis using a hybrid model using logistic regression and a random forest model. The representation of each class can be processed by oversampling and under sampling to estimate representative classes in datasets where subclasses of the dataset are not adequately represented by the classifier. A hybrid two-person machine learning model was developed, yielding an overall accuracy of 100.0%. As a result, optimal feature selection and data weighting in classification algorithms can improve model performance. The accuracy of the present study is higher than previous studies, despite using the same machine learning model.

References

[1] Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T. and Sain, M. (2018). *A classification approaches with different feature sets to predict the quality of different types of wine*

using machine learning techniques, 2018 20th International conference on advanced communication technology (ICACT), IEEE, pp. 139–143.

[2] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," *Diagnosis of Diabetes Using Classification Mining Techniques*", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[3] Andrea Mechelli and Sandra Vieira "*Machine learning*", Method and Application: https://doi.org/10.1016/C2017-0-03724-2, 2020

[4] Aurit, S., Kleffner, A. and Robinson, E. (2021). *Final project proposal: Statistical learning imbalanced classification and prediction of wine quality*, red 94: 92–6.

[5] Ayush Anand and Divya Shakti," *Prediction of Diabetes Based on Personal Lifestyle Indicators*", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[6] Barnaghi, P., Sheth, A. and Henson, C. (2013). *From data to actionable knowledge: Big data challenges in the web of things*, IEEE Intelligent Systems 28(6): 6–11.

[7] Basu, M. Bilenko, and R. J. Mooney, *"A probabilistic framework for semi-supervised clustering,"* in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 59–68.

[8] Bhardwaj, P., Tiwari, P., Olejar Jr, K., Parr, W. and Kulasiri, D. (2022). *A machine learning application in wine quality prediction, Machine Learning with Applications* 8: 100261.

[9] Brokamp, C., Jandarov, R., Hossain, M. and Ryan, P. (2018). *Predicting daily urban fine particulate matter concentrations using a random forest model*, Environmental science &technology 52(7): 4173–4179.

[10] Caissie, A. F., Riquier, L., De Revel, G. and Tempere, S. (2021). *Representational and sensory cues as drivers of individual differences in expert quality assessment of red wines*, Food Quality and Preference 87: 104032.

[11] Canizo, B. V., Escudero, L. B., Pellerano, R. G. and Wuilloud, R. G. (2019). *Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes,* Computers and Electronics in Agriculture 162: 514–522.

[12] Cardoso Schwindt, V., Coletto, M. M., D´ıaz, M. F. and Ponzoni, I. (2022). Could modelling and machine learning techniques be useful to predict wine aroma? Food and Bioprocess Technology pp. 1–19.

[13] Chandra-Sely, Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M,. "*Predicting At-risk University Students in a Virtual Learning Environment via a Machine Learning Algorithm*". Computers in Human Behavior (2016). https://doi.org/10.1016/j.chb.2018.06.032.

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16: 321–357.

[15] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K. et al. (2015). Xgboost: extreme gradient boosting, R package version 0.4-2 1(4): 1–4.

[16] Dahal, K., Dahal, J., Banjade, H. and Gaire, S. (2021). Prediction of wine quality using machine learning algorithms, Open Journal of Statistics 11(2): 278–289.

[17] Elder, J. (n.d). *Introduction to Machine Learning and Pattern Recognition*. Available at LASSONDE University EECS Department York website: http://www.eecs.yorku.ca/course_archive/2011-12/F/4404/

[18] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar*," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop"*, International Conference On I-SMAC,978-1-5090-3243-3,2017.

[19] Georgieva, P. and Rocha, E. (n.d.). *Machine learning in wine classification*.

[20] Gupta, M. and Vanmathi, C. (n.d.). *A study and analysis of machine learning techniques in predicting wine quality*, International Journal of Recent Technology and Engineering.

[21] Humar Kahramanli and Novruz Allahverdi*," Design of a Hybrid System for the Diabetes and Heart Disease"*, Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.

[22] Nawaz, R., Thompson, P., & Ananiadou, "*Identification of Manner in BioEvents*". In LREC (pp. 3505-3510). May, 2011

[23] Nithya and Dr. V. Ilango*," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques",* International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[24] Patil, R.C. Joshi and Durga Toshniwal*," Association Rule for Classification of Type-2 Diabetic Patients"*, ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing.

[25] Priyanka Rajendra, Shahram Latifi, "*Prediction of diabetes using logistic regression and ensemble techniques",* Department of Electrical and Computer Engineering, UNLV, Las Vegas, Nevada, United States, October 2021: https://doi.org/10.1016/j.cmpbup.2021.100032

[26] Rajesh and V. Sangeetha, *"Application of Data Mining Methods and Techniques for Diabetes Diagnosis"*, International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[27] Saravana kumar N M, Eswari T, Sampath P and Lavanya S*," Predictive Methodology for Diabetic Data Analysis in Big Data"*, 2nd International Symposium on Big Data and Cloud Computing,2015

[28] Siddhartha Bhattacharyya, Koyel Chakraborty , Surbhi Bhatia , Jan Platos, Rajib Bag, Aboul Ella Hassanien *"Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers-, A study to show how popularity is affecting accuracy in social media*", Applied Soft Computing, Volume 97, Part A, December 2020, 106754

[29] Somayeh Najafi-Ghobadi, Hosein Tarhsaz, Sharareh Parami, and Leili Tapak, "*Machine Learning-based Classifiers for the Prediction of Low Birth Weight*", Published online 2023 Jan 31. doi: 10.4258/hir.2023.29.1.54

[30] Taiwo, O. A. (2010). *Types of Machine Learning Algorithms, New Advances in Machine Learning*, Yagang Zhang (Ed.), ISBN: 978-953-307-034-6, InTech, University of Portsmouth United Kingdom. Pp 3 – 31.

[31] Ruiz-Alejos , and B. E. Reddy, *"Semi-supervised single-link clustering method,"* in Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on. IEEE, 2016, pp. 1–5. https://doi.org/10.1109/ICCIC.2016.7919689.

[32] Mani Butwall and Shraddha Kumar," *A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier*", International Journal of Computer Applications, Volume 120 - Number 8,2015.

[33] Clement Odooh , Regina Robert, Efijemue Oghenekome Paul " A Review Of Data Intelligence Applications Within HealthCare Sector In The United States" , International Journal on Soft Computing (IJSC),Vol.14,No.4,2023. DOI: 10.5121/ijsc.2023.14401