Journal of Optimization in Industrial Engineering,
Vol.17, Issue 1, Winter & Spring 2024, 1-16
DOI: 10.22094/QJIE.2024.951302

Research Article

Diabetes Detection via Machine Learning Using Four Implemented Spanning Tree Algorithms

Yas Ghiasi [a], Mehdi Seifbarghy [a], Davar Pishva [b,*]

[a] Department of Industrial Engineering, Faculty of Engineering, Alzahra University, Tehran, Iran

[b] Faculty of Sustainability and Tourism, Ritsumeikan Asia Pacific University (APU),Beppu, Oita, Japan

Received: 19 June 2023; Revised: 18 May 2024; Accepted: 21 May 2024

Abstract

This paper considers an accurate and efficient diabetes detection scheme via machine learning. It uses the science of data mining and pattern matching in its diabetes diagnosis process. It implements and evaluates 4 machine learning classification algorithms, namely Decision tree, Random Forest, XGBoost and LGBM. Then selects and introduces the one that performs the best towards its objective using multi-criteria decision-making methods. Its results reveal that Random Forest algorithm outperformed other algorithms with higher accuracy. It also examines the details of features that have a greater effect on diabetes detection. Considering that diabetes is one of the most deadly, disabling, and costly diseases observed today, its alarmingly increasing rates, and difficulty of its diagnosis because of many vague signs and symptoms, utilization of such approach can help doctors increase accuracy of their diagnosis and treatment schemes. Hence, this paper uses the science of data mining as a tool to gather and analyze existing data on diabetes and help doctors with its diagnosis and treatment process. The main contribution of this paper can therefore be its applied nature to an essential field and accuracy of its pattern recognition via several analytical approaches.

***Keywords*:** Diabetes; Data mining; Machine Learning; Multi-criteria decision-making; MCDM, Tree-based algorithms

1. Introduction

Considering that the focus of this paper is on data mining and artificial intelligence (AI) but applies it to diabetes detection in medical field, and because the two fields are quite different, and the fact that a typical reader may not be expert in both fields, it provides some relevant introductory information for easier understanding of its intended wide audience. It is important to highlight that diabetes is one of the most deadly, disabling, and costly diseases observed today, and its rates increase alarmingly. Diabetes is a metabolic disease in which people suffer from a lack or reduction in the ability to use insulin in their body (Dekamin, et al., 2021). According to the research of Abedian et al. (2019), the latest statistics published by the World Health Organization (WHO) and the International Diabetes Federation (IDF) showed that one out of every four people over the age of 60 suffer from this disease. Based on the investigation of Mosharrafzadeh et al. (2021) this disease is the fourth cause of death in the world. Mercaldo et al. (2017) mentioned in their research that in 2013 about 382 million people had diabetes and it is predicted that this number will rise to 595 million people by 2035. Another investigation (Standl et al., 2019) also showed that the number of people suffering from diabetes is increasing and estimates that by 2045 about 48% of the world's population will somehow suffer from this disease. Tireh et al. (2019) mentioned that various associated complications of this disease, both in terms of treatment costs and the disability it creates, have turned it into one of the most important healthcare issues. They also pointed out that the WHO has declared this disease as a hidden epidemic and since 1993, it has called all the countries of the world to deal with this epidemic. Dekamini et al. (2021) have found that factors such as obesity, stress, high cholesterol and fat levels, improper nutrition and sedentary lifestyle can lead to diabetes. Mosharrafzadeh et al. (2021) considered lack of timely or weak diagnosis of this disease as one of its major problems. They also mentioned that diabetes has destructive effects on people's health, and if diagnosed too late, can have irreparable damage to vision, kidneys, heart, arteries, etc.

Therefore, timely diagnosis of diseases is very important in medical science, and we can prevent disease, which ultimately leads to the improvement of society's health.

Mosharrafzadeh et al. (2021) also explained biological effects of diabetes in some detail. Accordingly, in a normal state, food in the stomach turns into glucose or blood sugar. Sugar then enters the bloodstream from the stomach, and the pancreas discharges the hormone insulin. This hormone causes sugar to enter the body's cells from the bloodstream, as a result, the blood sugar remains at a normal and balanced level. But in diabetes, there is either not enough insulin in the body, or the existing insulin is not able to perform its tasks properly. Consequently, blood sugar cannot enter the cells of the body effectively due to resistance and causes the blood sugar level to rise.

Rajeswari et al. (2019) classified diabetes into 3 general types of; Type 1 diabetes, Type 2 diabetes, and Gestational diabetes and the following three paragraphs provide some details on each of them.

* Corresponding Author Email Address: dpishva@apu.ac.jp

Accordingly, in Type 1 diabetes, the body is not able to produce insulin. This disease can exist in both adults and children. People with this type of diabetes take insulin every day. Jaiswal et al. (2021) cited this disease as an autoimmune disease in which the body's β-cells that are responsible for storing and releasing insulin are destroyed. Thus, leading to insulin deficiency.

Rajeswari et al. (2019) study also showed that in Type 2 diabetes, the body is not able to produce or use insulin. This type of diabetes is more common in middle-aged and older people. Azizi et al. (2015) also mentioned that Type 2 diabetes is the most common type of diabetes. They also believed that the presence of genetic and environmental factors can play a role in its occurrence and spread. An earlier study by Vinay et al. (2005) mentioned that Type 2 diabetes is also referred to as non-insulin-dependent diabetes since it usually occurs due to the resistance of the body's cells to insulin.

Rajeswari et al. (2019) study also stated that Gestational diabetes usually occurs in women and during their pregnancy stage. It causes an increase in blood sugar, which can both affect pregnancy and the baby's health.

Azizi et al. (2015) study also cited pre-diabetes, which many people get before developing diabetes. Accordingly, in many cases, lack of healthy nutrition and exercise initially causes prediabetes and then gets transformed into diabetes.

Diagnosing diabetes is often difficult because many of the signs and symptoms are vague and can only be diagnosed by experts. Therefore, countries that do not have enough health professionals for their population, such as developing countries like Bangladesh and India, are facing the problem of providing appropriate diagnostic methods for the wide range of their patient population. In addition, disease diagnosis often requires medical tests that low-income people find expensive and cannot afford. Since humans are also prone to error, it is not surprising that a mistake occurs in the diagnosis of a disease (over-diagnosis), which causes problems such as unnecessary treatment and adversely affects people's health and economy. According to the National Academics of Science, Engineering, and Medicine in 2015, most people will face at least one misdiagnosis of a disease during their lifetime (Ahsan et al., 2022).

Various factors may affect the misdiagnosis of diseases, the main of which are as follows (Ahsan et al. 2022, Tireh et al. 2019):

- Lack of proper symptoms, which are often unnoticeable.
- Common symptoms with other diseases.
- Omitted mistakenly from consideration.
- The condition of rare disease.

Considering the numerous difficulty factors that exist in diagnosing diabetes, techniques that can improve its detection accuracy and speed are needed. Decision support systems that have helped people in solving complex problems and decision making under uncertainty, have recently attracted the attention of many researchers towards their applications for diagnosing diseases. Data mining techniques can be considered as one of the essential methods that can be used for this purpose. It can discover patterns that may not be discovered by human intelligence under normal conditions (Tireh et al., 2019). According to the research of Mosharrafzadeh et al. (2021) data mining is relatively cheap and cost-effective. Ahsan et al. (2022) also pointed out that this method saves time. In 2019, Tireh et al. (2019) also proposed the use of data mining in the medical field for prevention or diagnosis of disease types as well as for the selection of their treatment methods. Mosharrafzadeh et al. (2021) also considered the use of machine learning methods and data mining science for diabetes diagnosis a necessity, due to existence of large amount of relevant data and the need to review and analyze them in the process. The severe social impact of the disease is also considered one of the main priorities in medical science research, which inevitably produces huge amounts of data.

Therefore, the science of data mining and analysis of existing data can serve as a tool to improve the prevention of the disease as well as help doctors to increase the accuracy of their diagnosis and treatment methods. Fig. 1 shows a summary of the services that can be provided to the field of health science via data mining (Moghaddassi et al., 2012).

Machine learning is one of the subsets of data mining science. The computer can act like a human being, learn more by using its own experiences and the additional information that it receives from the clients (Faraz et al., 2022).

Machine learning algorithms learn through a set of data called "Training Dataset" and create the required models. When new data is introduced to the machine learning algorithm, the system can perform the prediction process based on the created model. Bansal et al. (2022) in their research indicated that machine learning has recently been used for text detection, hate speech detection, recommender system, face detection, and others. Ahsan et al. (2022) also classified machine learning algorithms into three general categories of: supervised learning, unsupervised learning, and semi-supervised learning. There is also a fourth category, which is called reinforcement learning, but this study uses a supervised learning approach, details of which are explained below.

Adoption of supervised learning methods becomes useful when the value of its input variables is known to us. Finding error prediction models in insurance claims of a health institution is an example of a supervised learning strategy. In this strategy, models and features are known to us and are used with the purpose of data prediction and information discovery (Moghaddassi et al., 2012).

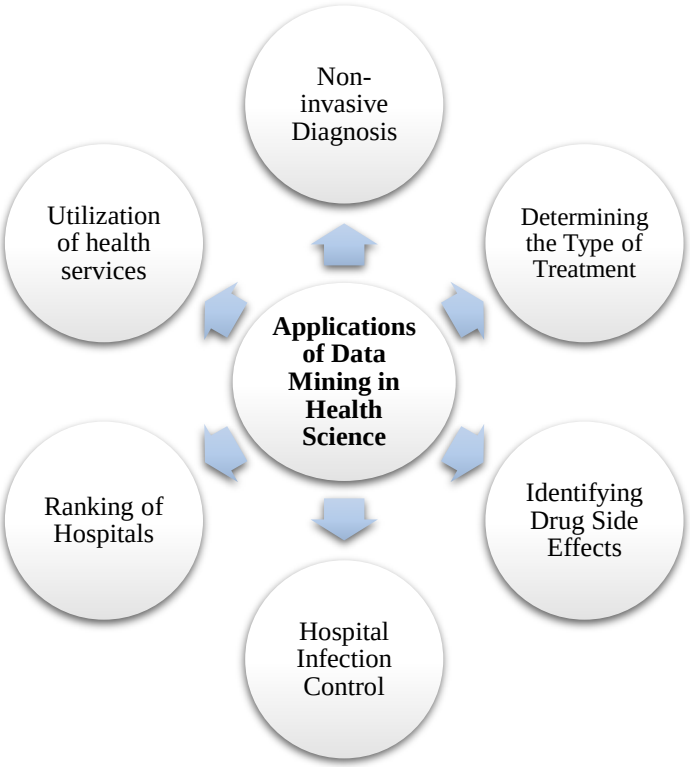

Fig. 1. Applications of data mining in the field of health science (Moghaddassi et al., 2012)

Supervised learning can be associated with students learning from a teacher. Data acts like a teacher and performs the task of teaching the machine. Once the machine is trained, it can make the necessary predictions and decisions about new data input to the system. A technical example is classification and regression.

According to the investigation of Dekamini et al. (2021), one of the most common applications of data mining in the medical field is the diagnosis of diseases. To diagnose diabetes, physiological and blood variables of some diabetic patients and healthy people are given to machine learning classification algorithms. These algorithms provide models for patients and classify them into two categories of "diabetic patient" and "healthy person". The created models can be used to classify new clients and people who are suspected of diabetes and predict their health status.

In this research, we use a binary approach to determine whether a person has diabetes or not. It is based on the supervised learning approach since we are facing a classification problem in which the characteristic of our target is clear and has a label. Considering that the information about patients has already been extracted, by implementing classification algorithms, we distinguish sick people from healthy ones. We then check the performance of our algorithms by evaluation indicators, and finally, with the help of 6 multi-criteria decision-making methods, we choose the result of the best performing classification algorithm.

In this research, we therefore seek to diagnose and predict diabetes with the help of data mining and implemented models. In its process, we implement and evaluate 4 machine learning algorithms, then select and introduce the one that performs the best towards our goal. We also examine the details of features that have a greater impact on diabetes detection.

2. Literature Review

Data mining has recently been widely used in the medical field since it can help doctors in diagnosing diseases and reducing errors. This section reviews some previous studies that have been done during the past five years on the application of machine learning techniques in diagnosing diabetes. Its focus is on the modeling and analysis of datasets that are taken from well-known sources such as Pima Indian Diabetes Dataset (PIDD).

Febrian et al. (2023) used K-Nearest Neighbor (KNN) algorithm and the Naïve Bayes (NB) algorithm to predict diabetes based on different features in the dataset using supervised machine learning. According to their results, it was concluded that the NB algorithm with 76.07% accuracy, 71.37% recall (i.e., sensitivity) and 73.37% precision outperformed KNN algorithm in predicting diabetes using PIDD.

Yakut (2023) divided the PIDD into 2/3 and 1/3, for training and test, respectively; then, both datasets were fed into random forest (RF) classifier, extra tree classifier and Gaussian process classifier machine learning. It was concluded that RF Classifier had the highest prediction accuracy. The addressed accuracy was 81.71%, recall was 84.83%, precision was 88.79%, F-score was 86.76% and area under the receiver operating characteristic (ROC) curve was 88.03%.

Perdana et al. (2023) used KNN algorithm to analyze important features in PIDD and to classify people as

diabetic or non-diabetic. They found out that KNN has the highest accuracy of 83.12% by setting k equal to 22. Moreover, they ranked important features, using KNN in the following order: Glucose, Age, Insulin, Blood Pressure, Body mass index (BMI), Pregnancy, Skin thickness, and Diabetes Pedigree function.

Tasin et al. (2022) studied PIDD and a private dataset from 203 individuals aged between 18 and 77 from a local textile factory in Bangladesh. A merged dataset was used in this work. They used synthetic minority oversampling technique, known as SMOTE and adaptive synthetic sampling, known as ADASYN for preprocessing to handle the class imbalance issue. Class imbalance problem is defined as not having the same number of diabetic patients as non-diabetic patients; in other words, a dataset is said to have imbalanced target feature when there exists a cost of misclassification error for balancing the data in the preprocessing stage. Various machine learning classification methods are applied for this purpose, namely: decision tree (DT), RF, support vector machine (SVM), logistic regression (LR), KNN, and various ensemble techniques such as bagging, adaptive boosting (AdaBoost: AB), extreme gradient boosting (XGBoost: XGB) and voting to find which algorithm has better prediction results. The best-performed prediction model was XGBoost classifier, with 81% accuracy, 0.81 F1 score and AUC of 0.84.

Mohammed Al-Nussairi et al. (2022) proposed a new classification algorithm for artificial neural networks (ANN) based on an enhanced version of the grey wolf optimizer (GWO) algorithm. The results of enhanced grey wolf optimizer (EGWO) were compared with other nature-inspired algorithms-trained ANN, such as, genetic algorithm (GA), deferral evolution (DE), evolutionary strategies (ES), particle swarm optimization (PSO), bat algorithm (BA) and GWO. EGWO gave better results for classification accuracy than the other algorithms.

Faraz et al. (2022) studied SVM algorithm. The reason behind applying this algorithm was due to having labels in PIDD dataset. In the data preparation stage, the data were standardized so that all data were set between zero and one. They concluded that SVM had around 79% accuracy.

Chang et al. (2022) selected three supervised machine learning algorithms to predict diabetes. These three models were: NB classifier, RF classifier and J48 DT model. They trained the selected algorithms utilizing PIDD. Then, the performances of the algorithms were analyzed to determine which one had the higher accuracy, sensitivity, precision, and specificity. The results showed that on the full PIDD (i.e., without eliminating unimportant features), RF classifier gave better results than those of NB and J48 DT with 79.57% accuracy, 89.40% precision, 75.00% specificity, 85.17% f-score, and 86.24% AUC. However, NB worked well in fine-tuned selection of features. In addition, it was mentioned that the most important features in diagnosing diabetes and its occurrence were glucose, BMI, age, insulin, and skin thickness.

Roy et al. (2021) used three different approaches for handling missing data, including median value imputation, K-nearest neighbor imputation, and iterative imputation. They applied LR, SVM, KNN, classification and regression tree (CART), Gradient Boosting (GB), ANN, RF and light gradient-boosting machine (LGBM) algorithms. Results showed that GB classifier had the best performance among others; it had accuracy of 91.06%; the second order algorithm was LGBM classifier, with accuracy of 90.69%. Ultimately, all algorithms showed acceptable performance with more than 60% across all evaluation metrics.

Khanam et al. (2021) studied seven algorithms including DT, RF, NB, LR, KNN, AdaBoost and SVM. All mentioned algorithms provided accuracy of higher than 70 percent. LR and SVM had approximately 77–78 percent accuracy for both K-fold and splitting. LR (K-fold) and LR (Splitting) had accuracy of 76.82 percent and 78.85 percent while SVM (K-fold) and SVM (Splitting) had accuracy of 76.82 percent and 77.71 percent. They also designed and used ANN structure for diabetic prediction utilizing PIDD with accuracy of 88.6 percent.

Sivaranjani et al. (2021) applied SVM and RF machine learning algorithms for diabetes prediction. After preprocessing step, they selected features which had higher impacts on diabetes prediction. Finally, RF with accuracy of 83 percent showed higher performance than that of SVM with accuracy of 81.4 percent.

Ghosh et al. (2021) used four different algorithms including GB, AdaBoost, RF and SVM for diabetes prediction. They evaluated the algorithms once considering all features and then with selected features using the minimal redundancy maximal relevance (MRMR) approach. Selected features were pregnancies, glucose, BMI, Diabetes Pedigree Function (DPF) and Age. In the end, seven types of performance evaluation metrics were used, and results revealed that RF had the highest performance among others with accuracy of 99.35 percent.

Naz et al. (2020) mentioned that a lot of research were done by using Deep Learning (DL) in anomaly detection. For diabetes, they applied Decision Tree, NB, ANN, and DL and evaluated the four classifiers by six performance evaluation metricsTheir proposed DL model showed the highest performance on the PIDD.

Maniruzzaman et al. (2020) worked on the diabetes datasets of 2009-2012, derived from the National health and nutrition examination survey (NHANES) which studies the health and nutritional states of people, including children in the United States. By eliminating missing values and unusual observations, the final dataset included about 6561 records for 6561people among whom there were 657 diabetic and 5904 non-diabetic people. Moreover, it had 14 features including age, gender, race, education, material status, occupation, weight, height, BMI, systolic blood pressure (BP), diastolic BP, direct cholesterol, total cholesterol, and Physical activity. By using LR for feature selection, they found out that 7 factors out of 14 were more important compared to others. The important features or risk factors were age, education, BMI, systolic BP, diastolic BP, direct cholesterol, and total cholesterol. They also predicted diabetes using DT, RF, NB and AdaBoost algorithms.

Benbelkacem et al. (2019) used RF algorithm to detect diabetes. Initially, they run several RF algorithms with different number of trees to find the optimum size of the forest. They set the size of RFs to 40 trees. Then, they compared its performance with other algorithms including: C4.5, Reduced Error Pruning Tree (REPTree), Simple CART, Best First Tree (BFTree) and SVM. In this study they calculated the error rate to evaluate the performance of models. The error rate for SVM was 0.23 while for RF was 0.21; other algorithms had higher error rates. Therefore, the RF had the lowest error rate and was selected as the best.

Mujumdar et al. (2019) studied a dataset similar to PIDD. The features of this dataset were the number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, age and outcome which were the same with those of PIDD; but this dataset also included the feature of the job type (Office-work/Field-work/Machine-work). This dataset contained 800 records and 10 attributes. In this research, they studied the performance of ML algorithms on both datasets. The implemented algorithms were RF Classifier, DT Classifier, Extra Tree Classifier, AdaBoost algorithm, Perceptron, Linear Discriminant Analysis (LDA) algorithm, SVC, LR, KNN, Gaussian NB, Bagging algorithm and Gradient Boost (GB) Classifier. Results showed that LR achieved the highest accuracy equal to 96 percent for the new diabetes dataset.

Sisodia et al. (2018) used three ML algorithms namely DT, SVM and NB to diagnose diabetes at an early stage. After evaluation based on various measures, they concluded that NB was the best with accuracy of 76.30 percent. They confirmed their results using receiver operating characteristic (ROC) curves.

A Summary of the discussed research is illustrated in Table 1. In the "Classification" column, all the applied algorithms are presented while in the "The best selected" column, the best algorithm is given according to its performance considering accuracy measures.

Table 1
Summary of the literature review on the performance of different algorithms in diabetes prediction

No	Authors	Year	Applied Classification Methods	The Best Performing Algorithms
1	Febrian et al.	2023	KNN, NB	NB with accuracy of 76.07%
2	Yakut	2023	RF, Extra Tree, Gaussian Process	RF with accuracy of 81.71%
3	Perdana et al.	2023	KNN	Accuracy of 83.12%
4	Tasin, et al.	2022	DT, SVM, RF, LR, KNN, XGB, Bagging, AB, Voting	XGBoost with accuracy of 81%
5	Mohammed Al-Nussairi, et al.	2022	EGWO, GA, DE, ES, PSO, BA, GWO	EGWO with accuracy of 76.3%
6	Faraz, et al.	2022	SVM	Accuracy of 79%
7	Chang et al.	2022	NB, RF, J48, DT	RF with accuracy of 79.57%
8	Roy, et al.	2021	LR, SVM, RF, LGBM, KNN, GB, CART, ANN	LGBM with accuracy of 84.2%
9	Khanam et al.	2021	DT, RF, NB, LR, KNN, AB, SVM	SVM and LR with accuracy of 77-78 %
10	Sivaranjani et al.	2021	SVM, RF	RF with accuracy of 83%
11	Ghosh et al.	2021	GB, SVM, AB, RF	RF with accuracy of 99.35%
12	Naz et al.	2020	DL, ANN, DT, NB	DL with accuracy of 98.07%
13	Maniruzzaman, et al.	2020	NB, DT, AB, RF	Combination of LR and RF with accuracy of 94.25%
14	Benbelkacem et al.	2019	C4.5, REP Tree, Simple Cart, BFTree, SVM, RF	RF with accuracy of 79%
15	Mujumdar et al.	2019	LR, GB, LDA, AB, Extra tree, Gaussian NB, Bagging, RF, DT, Perceptron, SVC, KNN	LR with accuracy of 96%
16	Sisodia et al.	2018	DT, SVM, NB	NB with 76.30% of accuracy was the best
17	This Study	2023	Random Forest, XGB, Decision tree, LGBM	Random Forest with 75.6% of accuracy was the best

3. Random Forest, XGB, Decision tree, LGBM

This section highlights the importance of the problem; explains our analytical approach and provides additional info about the source and nature of the data that was used in the process.

3.1 Problem statement

As mentioned above, in 2013, around 382 million people had diabetes and it was estimated to get increased to 595 million by 2035. As such it is important to take note of this disease, its increasing trend and perform timely diagnosis to prevent its numerous irreparable damages. Hence, researchers in the field of health science are exploring numerous approaches for its timely diagnosis and prevention of the associated damages. The technical difficulty of its diagnosis and lack of skilled doctors in the field, particularly in developing countries, has made the traditional approach of its diagnosis quite difficult. This has led researchers to apply the science of data mining to partially automate its diagnosis and help doctors in the field. In the research conducted so far, various machine learning algorithms have been investigated for its diagnosis and the differentiation of sick people from healthy ones.
The focus of this research is on tree-based algorithms. We evaluate the performance of these algorithms by 4 indicators of accuracy, sensitivity, precision, and F-score. We then apply multi-criteria decision-making methods to choose the best performing algorithm. This is mainly because this approach has several evaluation indicators and enables us to choose the best option more securely. This way, we can select the best performing algorithm more accurately and confidently propose it for the diagnosis of this disease.
Even though numerous studies have shown that machine learning approach has a great potential for diabetic diagnosis, it has not yet been widely applied in practice. This is mainly because it deals with human life and the fact that the science of machine learning is quite different from that of medical science. Hence, the focus of this research is on several convincing evaluation indicators for medical experts and appropriate managerial guidance on its practical usage. Nonetheless, the validity of its created model is limited to women's patients from India. In other words, universal data modeling is beyond the scope of this research, but its summary results and conclusion section show how it can eventually be achieved.

3.2 Data source and characteristics

The dataset used in this research was extracted from the Kaggle website. It was acquired for the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). All patients are women over 21 years old of Pima Indian heritage, and their data contains 768 records together with nine characteristics of diabetic patients. The characteristics include glucose, number of pregnancies, blood pressure, insulin, body mass index (BMI), age, history of diabetes, skin thickness, and finally a binary number, indicating whether the patient has diabetes or not Chang et al. (2022). Table 2 shows a summary of the characteristics.

Table 2
Description of features Chang et al. (2022)

Feature	Description	Data type
Pregnancies	Number of times pregnant	Numeric
Glucose	Plasma glucose concentration at 2 Hours in an oral glucose tolerance test (GTIT)	Numeric
Blood Pressure	Diastolic Blood Pressure (mm Hg)	Numeric
Skin Thickness	Triceps skin fold thickness (mm)	Numeric
Insulin	2-Hour Serum insulin (μh/ml)	Numeric
BMI	Body mass index [weight in kg / (Height in m)]	Numeric
DPF	Diabetes pedigree function	Numeric
Age	Age (years)	Numeric
Outcome	Binary value indicating non-diabetic /diabetic	Factor

Tables 3, 4 and 5 provide details of the characteristics that were selected and used from the data for analysis of diabetes.

Table 3
Summary info on data characteristics

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
Mean	3.84	121.70	72.08	24.96	128.64	32.38	0.47	33.24	0.35
Std	3.37	30.44	12.31	11.34	97.47	6.92	0.33	11.76	0.48
Min	0	44	24	6	7	18.2	0.08	21	0
Max	17	199	122	99	846	67.1	2.42	81	1

Table 4
Mean characteristics of diabetes

Outcome	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
0	3.30	110.73	70.75	23.86	115.54	30.78	0.43	31.19
1	4.86	142.14	74.54	27.03	153.10	35.35	0.55	37.07

Table 5
Median characteristics of diabetes

Outcome	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
0	2	107.5	70	23	99.5	30.1	0.34	27
1	4	140	74	27	138.5	34.25	0.45	36

4. Data Mining Steps

In this research, prior to creating machine learning models, we go through a data preprocessing stage. According to Khanam et al. (2020) data preprocessing can help us transform data to a format that could lead to a better machine learning model and provide higher accuracy.

We then select and adjust the desired hyper-parameters for the machine learning algorithms. After this stage, we implement the algorithms and check their performance by evaluating them with the indicators. A schematic diagram of the process is shown in Fig. 2.

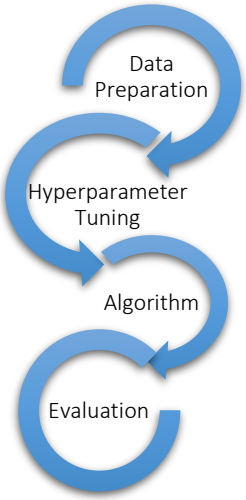

Fig. 2. Implementation process of data mining

4.1 Data preparation

Methods that are used for data preprocessing depend on the type and nature of dataset as well as the intended modeling approach. Common approaches include checking for outliers, missing data, duplicate data, normalization, etc. Since in this research we use tree-based models, normalization or standardization of the dataset was not needed. Considering that all our data were either integer or floating-point numbers, we also did not need to perform data series adjustments. As such we checked for outliers, missing and duplicate data. Our finding showed no outliers or duplicate data in the dataset, but spotted missing data were appropriately replaced in the operation process. Finally, as part of its most important step, balancing of the target characteristics was checked, and the unbalanced nature of the dataset was properly adjusted.

The following three subsections provide some details about our employed techniques and after going through all these processes, our dataset became ready for modeling.

4.1.1 Outlier data detection

To detect outlier data, we create a box plot for all characteristics. In this process, we need to determine the first and third quartiles. We considered the first quartile as 25% and the third quartile as 75%. Then, by applying the following formulas (1 and 2), we obtained the Upper Bound (UB) and Lower Bound (LB) for each of the characteristics. Table 6 shows our associated obtained results, and any data outside this range is considered incorrect.

$$LB = Q_1 - 1.5 * (Q_3 - Q_1) \quad (1)$$
$$UB = Q_3 + 1.5 * (Q_3 - Q_1) \quad (2)$$

Table 6
Upper and lower bounds of each characteristic

Characteristics	LB	UB
Pregnancies	-6.5	13.5
Glucose	39.0	201.0
Blood Pressure	40.0	104.0
Skin Thickness	-5.5	54.5
Insulin	-92.0	324.0
BMI	13.6	50.4
Diabetes Pedigree Function	-0.33	1.2
Age	-1.5	66.5

A graphical view of the obtained values for each characteristic, and their associated incorrect values are shown in Fig. 3. But these data are not necessarily outliers since an outlier data must have incorrect values for at least 3 characteristics. Therefore, it should be checked which data is jointly located in 3 or more incorrect characteristics values.

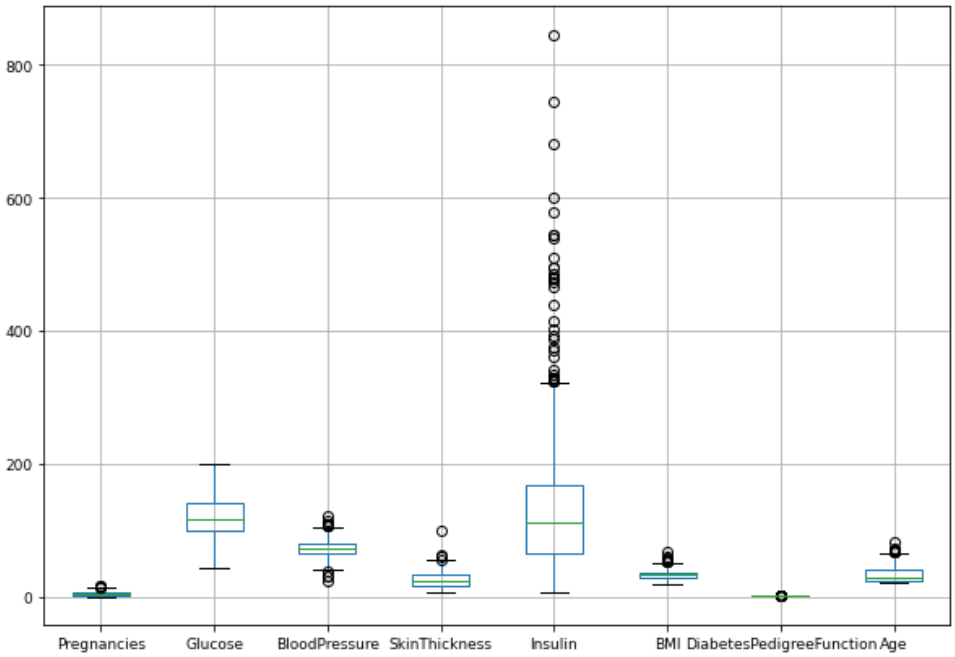

Fig. 3. Detected incorrect values for the 8 characteristics

According to Fig. 3, people who may be identified as having invalid data in several characteristics can only belong to the characteristics of blood pressure, skin thickness, BMI, and age. A list of numerically identified such people under each of these four characteristics is shown in Table 7. The numbers in each column, commonly called as record, identify people who have invalid values for the characteristics specified in the column.

By examining the data of Table 7, we can check whether a particular person is jointly located in 3 or more incorrect characteristics values. For example, patient number 120 has incorrect information for skin thickness and BMI but is not an outlier record since an outlier requires at least 3 incorrect characteristics. The same is true for patient number 177 who has incorrect information for blood pressure and BMI. The fact that none of the data are common in the 3 characteristics, indicates that the dataset does not have outliers.

Table 7
Record of correct values of each characteristic

Blood Pressure	Skin Thickness	BMI	Age
43	57	120	123
84	120	125	363
106	445	177	453
177	579	193	459
362	-	247	489
549	-	303	537
658	-	445	666
662	-	673	674
672	-	-	684
691	-	-	-

4.1.2 Missing data replacement

A dataset is considered to have missing data when one or more of its observations have either unrecorded or missing values in the dataset characteristics. The existence of such values and their extent are directly related to the accuracy of its associated model. Selecting the right method for its management is based on the nature of its characteristics. In medical science, since each patient is completely different from another, it is usually recommended to replace missing data with a new patient's dataset, something which is costly and time-consuming. In this research, since it was impossible to have access to new patients' information of Pima Indian heritage, we adapted an appropriate estimation and replacement method.

Table 8 shows an overview of our missing dataset, and its succeeding paragraph provides some details on our estimation and replacement method.

Table 8
Number and percentage of the missing data

	Insulin	Skin Thickness	Blood Pressure	Glucose	BMI	Diabetes Pedigree Function	Age	Pregnancies
Number	374	227	35	5	11	0	0	0
Percentage	48.70	29.56	4.56	0.65	1.43	0	0	0

To handle missing values in our data, we first identify the highest and lowest values for each feature. We then divide the range (the difference between the highest and lowest values) of each feature into a specific number of smaller, equally sized intervals, which are set based on the range of the feature. Next, we select the interval that contains the maximum number of people and generate random numbers within their range and replace the missing values with them. During this process, we save the new dataset that does not have corresponding missing value for our future references and continue the random number data generation until all missing data are replaced. Table 9 shows the selected intervals and their associated number of people.

Table 9
Selected ranges and values for each attribute

	Insulin	Skin Thickness	Blood Pressure	Glucose	BMI	Diabetes Pedigree Function	Age	Pregnancies
Range	(0-200)	(0-25)	(51-80)	(100-149)	(21-35)	(0.10001-0.8)	(0-35)	(0-4)
Largest Number	683	427	539	428	494	653	498	492

4.1.3 Balancing of the target characteristics

In the last stage, we need to assure that the characteristics of our goal are well-balanced, and it is done after dividing the data into training and test datasets. The objective is to ensure that the number of patients with diabetes is the same as those without diabetes in the training dataset that is used for modeling. If such issues are not balanced, the generated model could lose its efficiency and get overwhelmed by data of the group that has bigger dataset and yield to inaccurate prediction. In our entire dataset, the characteristics data of the target were not balanced. As shown in Fig. 4, out of a total 768 data, 268 of them had a value of 1 and 500 a value of zero.

One way to handle this problem is to randomly split the data into training and test dataset and identify their number of ones and zeros. Then examine the training dataset, identify the group that has smaller size, and supplement its pertinent shortcomings from the test dataset. In its implementation process, our total 768 data was randomly split into training and test dataset in a ratio of 70 to 30. It is of course natural to assign a bigger portion of the data to the training set to increase the reliability of its model. Hence, initially 537 data were assigned to the training dataset and 231 data to test dataset. After supplementation of the training dataset, we can then achieve a balanced dataset for training.

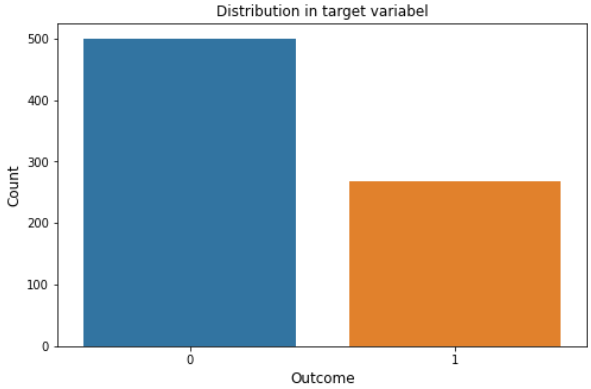

Fig. 4. Number of zeros and ones in the target characteristics

As mentioned, we check the number of zeros and ones in the training data. Out of 537 data, we have 350 zeroes and 187 ones, which is unbalanced. Considering that our total number of ones in the dataset is only 268 (Fig. 4), it is

impossible to increase the number of ones in the training data from 187 to 350. In order to make the modeling more accurate, our strategy is to split the training dataset into two subsets, each of which contains the same 187 ones, but grouped with two different 187 zeros and use the average of their generated result. Therefore, first we increase the number of zeros to 2 times the number of ones in the training data so that we can have 2 separate subsets in the above-mentioned manner and run the model on both. Hence, we take 24 zeros from test data and add them to our training dataset so that the number of zeros becomes 374, which is 2 times 187 (the number of ones). Now we can divide them into 2 subsets, use them for modeling and present their averaged answer as the result.
At this stage, all the data preparation steps are considered complete, and the data set is ready for modeling. In the next step, we introduce and adjust the hyper parameters of each model. Finally, the algorithms are implemented and evaluated.

4.2 Hyper-parameter tuning

One of the most important ways to control and improve the machine learning process is to adjust the hyper-parameters of the algorithm. Scikit-Learn Python library has considered hyper-parameters by default for each model, but in general such values are not optimal for all problems (including ours). Although setting of the best hyper-parameters is usually an impossible task, with some trial and error, we can obtain optimal values. To achieve this, we need to test many combinations of hyper-parameters and evaluate the performance of each model. In its process, we use the Grid Search method to set the hyper-parameters. In the Grid Search method, all combinations are evaluated. This method allows us to set multiple states for each hyper-parameter of every model, then it runs all the models and shows us the best state among the set hyper-parameters. We have applied the Grid Search method for adjusting the hyper-parameters of our 3 implemented algorithms of Random Forest, LGBM and XGBoost. For the decision tree algorithm, however, we have taken another approach, details of which are explained in the next subsection.

4.3 Classification of proposed models and hyper-parameters

Considering the advantages of tree-based algorithms, decision tree, random forest, XGBoost and LGBM algorithms were selected. One of these advantages according to Chang et al. (2022) is that it does not need normalization. Jijo et al. (2021) also labeled decision trees as one of the powerful methods that are commonly used in various fields such as machine learning, image processing, and pattern recognition. Decision tree structures are powerful for analyzing big data in a short time. It also has the capacity to select effective features in the algorithm.

4.3.1 Decision tree model

The criterion hyper-parameter is very important in the decision tree model, indicating that each leaf in the decision tree has reached the end and can no longer continue. For this, we have two indices of Gini and Entropy. The leaf that sets these indexes to zero is the final leaf. These two hyper-parameters indicate the degree of impurity of each leaf. The more this index reaches its minimum value, the better. This means that the leaves that reach the end have zero impurity. The formula for obtaining these indicators are as follows:

$$Gini(t) = 1 - \sum_j [p(j|t)]^2 \quad (3)$$

$$Entropy(t) = -\sum_j p(j|t)\, log_2 p(j|t) \quad (4)$$

Between Gini and Entropy, we should choose the one for our model that is more suitable for our problem. To determine which index and with what maximum number of levels is more accurate for our decision tree, we should test with both on various maximum levels. Thus, we run the model once with the Gini index with a maximum level from 1 to 20 and once with the Entropy index with a maximum level from 1 to 20, with both subsets that we have, and measure their accuracy levels. Results of the investigations are graphically shown in Fig. 5 and Fig. 6, and their summary is listed in Table 10.

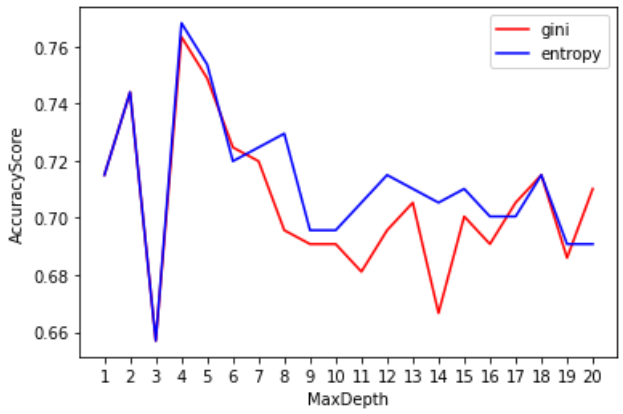

Fig. 5: Subset 2

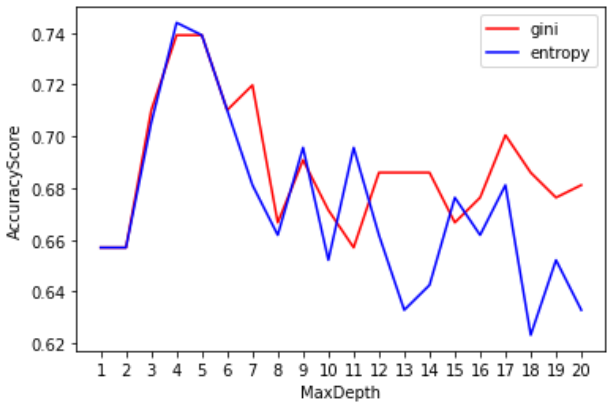

Fig. 6: Subset 1

4.3.2 Random forest model

The algorithms used in random forest model is the same as that of the decision tree, but this model runs several decision trees under random conditions. Its final result is based on the result that a greater number of implemented trees would have produced the same result.
In this model, the maximum level of ending the model has been checked in the range of 1 to 20, with the two indices of Gini and Entropy. We also checked the number of trees 4 times with 50, 100, 150 and 200 trees. The obtained best values for the 2 subsets are listed in Table 10.

4.3.3 Extreme Gradient Booting (EGB) tree model

In this model, the maximum level of ending the model has been checked in the range of 3 to 10. The learning rate index is in the range of 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30. We also checked Min child weight with numbers 1, 3, 5, 7. The obtained best values for the 2 subsets are listed in Table 10.

4.3.4 LGBM model

This algorithm is a gradient boosting framework that uses tree-based learning algorithms. It is designed for distributed and efficient use. This algorithm is quite fast and the term Light in its title refers to this point.
In the LightGBM model, the maximum termination level of the model has been checked in the range of 0 to 26, and the number of leaves has been calculated in the range of 7, 14, 21, 28, 31, and 50. Also, the learning rate was checked in the range of 0.1, 0.03, 0.003, and finally, the number of trees was considered between 50, 100, 200 and 500 trees. Table 10 shows the obtained best values for the 2 subsets, its adjusted and implemented values in the model.

Table 10
Obtained hyper-parameters of the Models via grid search algorithm

Model	Algorithms	Hyper-parameters	Description	Value
1	Decision Tree	Max_depth	Maximum depth of the tree	Subset1: 4 Subset2: 4
		Criterion	Function to measure the quality of a split	Subset1: "Entropy" Subset2: "Entropy"
2	Random Forest	Max_depth	Maximum depth of the tree	Subset1: 18 Subset2: 17
		Criterion	Function to measure the quality of a split	Subset1: "Entropy" Subset2: "Entropy"
		n_estimator	Numbers of trees in the forest	Subset1: 50 Subset2: 50
3	XGB	Learning_Rate	The shrinkage done at every step	Subset1: 0.2 Subset2: 0.3
		Max_depth	Maximum depth of the tree	Subset1: 9 Subset2: 8
		Min_child_weight	Minimum sum of instance weight (hessian) needed in a child	Subset1: 5 Subset2: 3
4	LGBM	Learning_Rate	The shrinkage done at every step	Subset1: 0.1 Subset2: 0.03
		Max_depth	Maximum depth of the tree	Subset1: 50 Subset2: 200
		n_estimator	Numbers of trees in the forest	Subset1: -1 Subset2: 9
		num_leaves	the maximum number of leaves per tree	Subset1: 14 Subset2: 14

4.4 Evaluation of Machine Learning Models' Performances

After implementation of the algorithms, to check which algorithm has performed better, several indicators have been used, which are labeled based on the confusion matrix (Table 11). These indicators include accuracy, precision, F-score (also called F1Score) and recall that we obtained for each model. We obtain the confusion matrix for each

algorithm and its associated two subsets (Table 12). Using the obtained matrix and formulas (5) to (8), we calculate the values of the evaluation indicators that are presented in Table 13.

Table 11
The confusion matrix

		PREDICTED CLASS	
ACUAL CLASS	Not Purchased	Not Purchased (TP)	Purchased (FN)
	Purchased	(FP)	(TN)

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$precision = \frac{TP}{TP+FP} \quad (6)$$

$$recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1Score = \frac{2*precision*sensitivity}{precision+sensitivity} \quad (8)$$

Table 12
Values of confusion matrix and indices of different models

Model	**Algorithms**	**Subsets**	**Values**
1	Decision Tree	Subset1	$\begin{bmatrix}97 & 29\\19 & 62\end{bmatrix}$
		Subset2	$\begin{bmatrix}103 & 23\\30 & 51\end{bmatrix}$
2	Random Forrest	Subset1	$\begin{bmatrix}93 & 33\\16 & 65\end{bmatrix}$
		Subset2	$\begin{bmatrix}93 & 33\\19 & 62\end{bmatrix}$
3	XGB	Subset1	$\begin{bmatrix}88 & 38\\22 & 59\end{bmatrix}$
		Subset2	$\begin{bmatrix}90 & 36\\18 & 63\end{bmatrix}$
4	LGBM	Subset1	$\begin{bmatrix}91 & 35\\21 & 60\end{bmatrix}$
		Subset2	$\begin{bmatrix}95 & 31\\20 & 61\end{bmatrix}$

Table 13
The values of the evaluation indices of each model

	Accuracy	**Sensitivity (Recall)**	**Precision**	**F-score**
Random Forest	0.756	0.784	0.658	0.715
XGB	0.725	0.753	0.622	0.681
Decision tree	0.756	0.697	0.685	0.689
LGBM	0.741	0.747	0.647	0.693

5. The Evaluation

This section carries out the evaluation by considering the implemented algorithms as our options, and indicators that are used to evaluate them, as the evaluation criteria. Therefore, the numerical values of the evaluation indices that are shown in Table 13 form the same decision matrix.

5.1 Evaluation of the models

Performance of each model can be checked by using the numerical values of their evaluation indices that we have obtained (Table 13). Through multi-criteria decision-making methods, their performance can be compared, and the best performing algorithm can be selected.
Table 14 shows the weights of the criteria that we obtained via the Shannon Entropy method.

Table 14
Weight of the obtained criteria

	Accuracy	Sensitivity (Recall)	Precision	F-score
Wj	0.085	0.488	0.335	0.094

5.2 Decision making models

In this research, we have implemented and used six Multiple Attribute Decision Making (MADM) models, which are: Simple Additive Weighting (SAW), TOPSIS, WASPAS, TAXONOMY, VIKOR and MOORA. Table 15 shows their obtained ranking for each of the implemented spanning tree algorithms.

Table 15
Generated ranking of each MADM model for the four algorithms

	SAW	TOPSIS	WASPAS	TAXONOMY	VIKOR	MOORA	Average	Rank
Random Forest	1	1	1	1	1	1	1	**1**
XGB	4	3	4	4	3	4	3.67	**4**
Decision tree	3	4	3	3	4	2	3.17	**3**
LGBM	2	2	2	2	2	3	2.17	**2**

5.3 Proper decision making via spearman's rank correlation coefficient

To decide which method and ranking is appropriate, we first average the generated ranking of each model and sort them in ascending order of the average to reveal their rank. In Table 15, the red column shows the respective ranking of each model based on the average of the associated ranking. Now, using the average, we calculate the Spearman index for each of the decision-making techniques by means of the following formula.

$$\rho = 1 - \frac{6 \sum d_i^2}{n\,(n^2-1)} \quad (9)$$

Table 16 shows the result of Spearman correlation index for each of the utilized MADM models.

Table 16
Spearman correlation index

MADM models	Spearman correlation index
SAW	0.983
TOPSIS	0.883
WASPAS	0.983
TAXONOMY	0.983
VIKOR	0.883
MOORA	0.783

As can be observed from Table 16, SAW, WASPAS, and TAXONOMY models have the highest correlation, which means that their ranking is more reliable. Table 17 shows the final ranking of the implemented spanning tree algorithms based on the ranking of the three highly correlated MADM models.

Table 17
The final ranking

	SAW	WASPAS	TAXONOMY	Rank
Random Forest	1	1	1	**1**
XGB	4	4	4	**4**
Decision tree	3	3	3	**3**
LGBM	2	2	2	**2**

5.4 Summary of the evaluation results

According to the results of the decision-making techniques, the random forest model has performed the best with an accuracy of 75.6% and a sensitivity of 78.3%, and the second place belongs to the LGBM model, then the decision tree model has won the third place and The XGB model is ranked last. The values obtained by the random forest model for each index are listed in Table 18.

Table 18
Evaluation index value of random forest model

	Accuracy	Sensitivity (Recall)	Precision	F-score
Random Forest	0.756	0.784	0.658	0.715

Because this model is one of the tree-based models, it has the power to select the most effective characteristics in diagnosing diabetes and can provide a priority for it. Table 19 shows this priority separately for the 2 subsets.

Table 19
The priority of characteristics in the diagnosis of diabetes

Attributes	Subset1	Subset2
Pregnancies	0.066	0.062
Glucose	0.240	0.223
Blood Pressure	0.087	0.083
Skin Thickness	0.090	0.103
Insulin	0.114	0.104
BMI	0.186	0.162
Diabetes Pedigree Function	0.108	0.132
Age	0.110	0.130

As can be observed from Table 19, glucose has top priority in both subtests and can be considered as the most effective characteristic in the diagnosis of diabetes. The second priority belongs to BMI which is an index for body mass and according to Roy et al. (2021), BMI is directly related to diabetes. Therefore, it can be concluded that obesity increases the risk of type 2 diabetes.

5.5 Managerial issues on its practical usage

Employment of artificial intelligence (AI) in medical field, however, requires precautions and supervised learning approach. This is mainly because it is related to human life and the nature of human physiology is quite complex. Initially, region specific models should be created from physiological and blood variables of some diabetic patients and healthy people who have been diagnosed with the traditional approach. The model can then be used on other people after confirming its accuracy. Nonetheless, precision of AI specified diabetic patients, should still be double checked with the traditional approach in its early implementation stage. This approach can still significantly relieve regions and countries that do not have enough health professionals for their population since their focus will be on AI specified diabetic patients. Passage of time and reanalysis of misdiagnosis can further help to improve the accuracy of the model. This can eventually lead to universal modeling via region-specific data normalization and correlation factor analysis.

6. Conclusions and Future Work

This study clearly showed that with the help of data mining science and implementation of machine learning models on health datasets, valuable information can be extracted. Such an approach can help doctors to both improve and speed up their diagnosis. The finding identified the random forest model as the best performing model (Table 18). It also showed that glucose level is the most effective characteristic in the diagnosis of diabetes and confirmed that obesity (Body Mass Index) increases the risk of type 2 diabetes (Table 19).

Considering the importance and complex nature of the problem, numerous strategies must be explored in the adoption of the machine learning approach. However, based on the variation of the generated result of each model, appropriateness of the adopted methods must be ensured by adjusting pertinent hyper-parameters and preventing the overfitting of machine learning. As such, this study evidently showed that by implementing several algorithms and evaluating them with evaluation indicators, the best algorithm can be obtained using decision-making methods. Its evaluation process initially started with average ranking of each implemented spanning tree algorithm, then calculated their associated Spearman index, and carried out a final evaluation based on the performance of the models that had the highest correlation index.

Considering the relatively new and attractive nature of the field, this study concentrated on the best performing model and identification of the most effective characteristics in the diagnosis of diabetes. In this process, however, this study used only women patients' data of Pima Indian heritage. Due to importance of the field, its worldwide application and the complex nature of human physiology, further studies along this field can extend this study from various perspectives, a few of which are as follows:

- Apply the same strategy to data that are taken from patients of other countries and regions and do a comparative analysis.
- Apply the same strategy to data that are taken from patients who live in different climate zone (e.g., tropical, temperate, and frigid zones).
- Apply the same strategy to data that are taken from patients of the same region, who have different lifestyles (e.g., degree of fatty food and sweets intakes, alcohol consumptions, sports activities, etc.).

- Examine a large-scale instance of the problem by integrating numerous types of data that are used in various region-specific studies and carry out region-specific data normalization and correction factor approach in the process for its universal application.

References

Abedian, I., Ayoobi, A., Ghaffary, H., Zabbah, I. (2019). Diagnosis of diabetes by using a data mining method based on native data. *Journal of Torbat Heydariyeh University of Medical Sciences*, Volume 7, No.1: 1-14

Ahsan, M. M., Luna, S. A., Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare.* 10(3), 541. https://doi.org/10.3390/healthcare10030541.

Azizi, F., Hadaegh, F. (2015). The upward trend of diabetes and pre-diabetes in Iran. *Iranian Journal of Endocrinology and Metabolism*. 17 (1) :1-3

Bansal, M., Goyal, A., Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal.* Volume 3, 100071.

Benbelkacem, S., Atmani B. (2019). Random Forests for Diabetes Diagnosis. *International Conference on Computer and Information Sciences (ICCIS).* Sakaka, Saudi Arabia, pp. 1-4, doi: 10.1109/ICCISci.2019.8716405.

Chang, V., Bailey, J., Xu, Q. A., Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Nature Public Health Emergency Collection*. https://doi.org/10.1007/s00521-022-07049-z

Dekamini, F., Ehsanifar, M. (2021). Comparison of the Efficiency of Data Mining Algorithms in Predicting the Diagnosis of Diabetes. *Iranian Journal of Diabetes and Metabolism*. Vol. 21, No 4.

Faraz, S., Singh ,P. (2022). Diabetes Prediction using Machine Learning. *Journal of Applied Science and Education.* Vol. 02, Iss. 02, S. No. 003, pp. 1-12.

Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *7th International Conference on Computer Science and Computational Intelligence. Procedia Computer Science*, Vol 216, Pages 21-30.

Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., Jonkman, M. (2021). A Comparative Study of Different Machine Learning Tools in Detecting Diabetes. *Procedia Computer Science. 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems.* Volume 192, Pages 467-477. https://doi.org/10.1016/j.procs.2021.08.048.

Jaiswal, V., Negi, A., Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. *Primary Care Diabetes.* Volume 15, Issue 3, Pages 435-443.

Jijo, B. T., Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*. Vol. 02, No. 01, pp. 20)

Kaggle Data Science, Pima Indians Diabetes Database, San Francisco, USA, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (Last access: 2023.5.30).

Khanam, J. J., Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*. Volume 7, Issue 4, Pages 432-439.

Maniruzzaman, M., Rahman M. J., Ahammed B., Abedin M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems 8, 7.* https://doi.org/10.1007/s13755-019-0095-z.

Mercaldo, F., Nardone, V., Santone, A. (2017). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. *Procedia Computer Science*. Volume 112, Pages 2519-2528

Moghaddassi, H., Hoseini, A., Asadi, F., Jahanbakhsh, M. (2012). Application of Data Mining in Health. *Health Information Management*; 9(2): 304.

Mohammed Al-Nussairi, M., Eljinini, M. A. H. (2022). A Hybrid Approach for Enhancing the Classification Accuracy for Diabetes Disease. *Journal of Information Technology Research.* Volume 15, Issue 1. DOI: 10.4018/JITR.298024.

Mosharrafzadeh, S., Ravaei, B., Koozegar, E. (2021). Diagnosis of Diabetes Using a Random Forest Algorithm. *Iranian Journal of Diabetes and Metabolism*; Vol. 21, No 2.

Mujumdar, A., Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *International Conference on Recent Trends in Advanced Computing (ICRTAC)*. Procedia Computer Science 165 (2019) 292–299.

Naz, H., Ahuja S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders.* 19(1):391-403 https://doi.org/10.1007/s40200-020-00520-5.

Perdana, A., Hermawan, A., Avianto, D. (2023). Analyze Important Features of PIMA Indian DatabaseFor Diabetes Prediction Using KNN. *Journal SISFOKOM (Sistem Informasi dan Komputer)*, Volume 12, Nomor 01, PP 70-75.

Rajeswari, M., Prabhu, P. (2019). A Review of Diabetic Prediction Using Machine Learning Techniques. *International Journal of Engineering and Techniques.* Volume 5 Issue 4.

Roy, K., Ahmad, M. et al. (2021). An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values. *Complexity.* Volume 2021, Article ID 9953314, 21 pages.

Sisodia, D., Sisodia D. S. (2018). Prediction of Diabetes using Classification Algorithms. *International Conference on Computational Intelligence and Data*

Science. Procedia Computer Science vol. 132, 1578–1585.

Sivaranjani, S., Ananya, S., Aravinth, J., Karthika, R. (2021). Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction. *7th International Conference on Advanced Computing and Communication Systems (ICACCS).* Coimbatore, India, pp. 141-146, doi: 10.1109/ICACCS51430.2021.9441935.

Standl, E., Khunti, K., Hansen, T. B., Schnell, O. (2019). The global epidemics of diabetes in the 21st century: Current situation and perspectives. *European Journal of Preventive Cardiology*. Vol. 26(2S) 7–14

Tasin, I., Nabil, T., Islam, S. (2022). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters published by John Wiley. The Institution of Engineering and Technology.* DOI: 10.1049/htl2.12039

Tireh, H., Taghi Shakeri, M., Rasoulinezhad, S., Esmaily, H., Yousefi, R. (2019). Comparison of the efficiency of data mining methods in predicting type 2 diabetes. *Tehran University Medical Journal*, Vol. 77, No. 5: 301-307

Vinay, K., Nelson, F., Abul, A. K., Ramzi, C. S., Stanley, R. L. (2005). *Robbins and Cotran Pathologic Basis of Disease* (7th ed.). Philadelphia, Pa.: Saunders. pp. 1194–1195. ISBN 0-7216-0187-1.

Yakut, Ö. (2023). Diabetes Prediction Using Colab Notebook Based Machine Learning Methods. *International Journal of Computational and Experimental Science and Engineering* (IJCESEN). Vol. 9-No.1, pp. 36-41.