Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07049-z

S.I.: AI-BASED E-DIAGNOSIS

Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms

Victor Chang[1] · Jozeene Bailey[2] · Qianwen Ariel Xu[2] · Zhili Sun[3]

Received: 16 August 2021 / Accepted: 31 January 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
This paper proposes an e-diagnosis system based on machine learning (ML) algorithms to be implemented on the Internet of Medical Things (IoMT) environment, particularly for diagnosing diabetes mellitus (type 2 diabetes). However, the ML applications tend to be mistrusted because of their inability to show the internal decision-making process, resulting in slow uptake by end-users within certain healthcare sectors. This research delineates the use of three interpretable supervised ML models: Naïve Bayes classifier, random forest classifier, and J48 decision tree models to be trained and tested using the Pima Indians diabetes dataset in R programming language. The performance of each algorithm is analyzed to determine the one with the best accuracy, precision, sensitivity, and specificity. An assessment of the decision process is also made to improve the model. It can be concluded that a Naïve Bayes model works well with a more fine-tuned selection of features for binary classification, while random forest works better with more features.

Keywords Diabetes mellitus · The Internet of Medical Things (IoMT) · Machine learning · Interpretable artificial intelligence

1 Introduction

Diabetes mellitus, or simply diabetes, is a leading non-communicable disease (NCD) globally, almost doubling in cases since 1980 [1]. It is a chronic illness that develops either when the pancreas are not able to generate sufficient insulin or when the body does not utilize the insulin produced effectively [1]. There is no cure for this disease. Diabetes is thought to result from a combination of genetic and environmental factors. Several risk factors that are attributed to diabetes include ethnicity, family history of diabetes, age, excess weight, unhealthy diet, physical inactivity, and smoking. In addition to this, the absence of early detection of diabetes has been known to contribute to the development of other chronic diseases such as kidney disease. Furthermore, additional pre-existing non-communicable diseases present a high risk for the patient, as they easily contract and are susceptible to infectious diseases such as COVID-19 [2].

Predicting the probability of an individual's risk and susceptibility to a chronic illness like diabetes is an important task. Diagnosing chronic illness at an early stage saves on medical costs and reduces the risk of more complicated health problems. Even in emergencies where a patient may be unconscious or unintelligible, it is pertinent that deductions can be made accurately from immediately measurable medical indicators to help clinicians make better decisions for patient treatment in high-risk situations.

✉ Victor Chang
victorchang.research@gmail.com

Jozeene Bailey
jozeenebailey@gmail.com

Qianwen Ariel Xu
qianwen.ariel.xu@gmail.com

Zhili Sun
z.sun@surrey.ac.uk

[1] Department of Operations and Information Management, Aston Business School, Aston University, Birmingham, UK

[2] Cybersecurity, Information Systems and AI Research Group, School of Computing and Digital Technologies, Teesside University, Middlesbrough, UK

[3] Institute for Communication Systems (ICS), 5G and 6G Innovation Centre (5G&6GIC), University of Surrey, Guildford, Surrey, UK

Published online: 24 March 2022

The majority of existing NCD cases remain undiagnosed, with patients suffering few symptoms during the initial phases of the disease, which causes a huge challenge in ensuring early detection and diagnosis. One advantage of providing treatments to patients in the early stage of their experience with non-communicable diseases is that they can avoid expensive treatments later in life as the disease gets worse. This is made more problematic with a lack of medical practitioners in underserved regions such as rural and remote villages. In such cases, the combination of the Internet of Medical Things (IoMT) and machine learning models can be made available to assist healthcare professionals in the early detection and diagnosis of NCDs by providing predictive tools for more efficient and timely decision-making.

However, it should be noted that machine learning solutions tend to be mistrusted by some people because of what may be referred to as a 'black-box' effect: an inability to show its internal decision-making process. This lack of explainability in machine learning models causes skepticism by consumers and results in slow uptake by end-users within the healthcare sector. The ability to explain both the reasonings behind and the process it takes to get a machine learning prediction is crucial to building trust, particularly in the healthcare field, where mistakes could be fatal.

This paper seeks to develop an e-diagnosis system for detecting and classifying diabetes as an IoMT application. Through the use of machine learning algorithms [Naïve Bayes, random forest, and decision tree (J48)], the system will be able to predict whether a person is at risk for diabetes based on several risk factors, provide doctors with a preliminary diagnosis, and feedback the doctor's guidance on diet, exercise, and blood glucose testing to patients.

These classification models were evaluated by the use of various methods, including accuracy, precision, sensitivity, F-measure, area under receiver operating characteristics (AUROC) curve to identify the best performing classifier. Several significant features that can be used to predict the severity of diabetes were extracted from the top classification model.

The Pima Indian Diabetes dataset is employed for this experiment. Pima Indians are a Native American group that lives in Mexico and Arizona, USA [3]. This group was deemed to have a high incidence rate of diabetes mellitus. Thus, research around them was thought to be significant to and representative of global health [4]. The Pima Indian Diabetes dataset consisting of Pima Indian females 21 years and older is a popular benchmark dataset [5]. This group is also significant to members of underrepresented minority or indigenous groups.

The features of the dataset comprise measures that do not require extensive testing. In emergency situations and patient self-care, which have become more popular, this function is essential.

The methodology is as follows: prepare the dataset, followed by data pre-processing such as dealing with missing values and categorical values, imputation, and standardization. Feature selection will be performed by using a variety of tools. Lastly, the classifiers' performance before and after feature selection will be evaluated further.

The organization of this paper is outlined as follows: Sect. 2 presents literature review, Sect. 3 provides details on data cleaning, exploration and feature selections, and Sect. 4 presents the methodology for analysis and evaluations of the dataset. Finally, Sect. 5 concludes the paper with discussions of future research.

2 Literature review

2.1 Internet of Medical Things (IoMT) and artificial intelligence algorithms

Internet of Medical Things (IoMT) is the application of the Internet of Things (IoT) in the medical field. Utilizing networking technologies, the IoMT aims to connect medical equipment and its applications with healthcare IT systems [6]. This innovative development has changed the medical field with its novel-designed remote healthcare system in terms of social benefits, perception, and reliable detection of illness. Benefiting from the constant computing of the IoT, it becomes easier to accomplish clinical goals such as patient data, medical orders, medical instruments, and remedies [7]. The development of the IoMT has brought about tremendous changes in promoting disease management, enhancing disease diagnostic and treatment techniques, as well as lowering healthcare costs and mistakes. This transformation has had a significant influence on the healthcare quality for both frontline healthcare professionals and patients. The IoMT is a thriving force for the researcher, the medical professional, the patient, and the insurer, enabling numerous use cases, for example, telemedical support, data insights, drug management, operation enhancement, patient tracking, etc. [8]. In particular, the IoMT offers various services to medical professionals, including delivering feedback to medical staff, equipment data and settings based on the needs of the patient and the specialist. IoMT gives rapid and easy access to various reports that help surgeons in operating rooms during surgeries [9].

The value of the IoMT is growing as a result of the symbiotic rise of artificial intelligence (AI). However, data production is one of the most significant challenges resulting from the development that a number of academics have confronted [10]. Because the amount of data acquired

is quite massive, it is necessary to use machine learning technology, which is good at processing and analyzing the data and extracting valuable information from the massive data and then visualizing them [11].

For chronic diseases like diabetes, AI, including machine learning and deep learning, plays an extremely important and effective role in supporting doctors' decision-making and monitoring and managing patients [12, 13]. Specifically, the combination of IoMT and AI can bring two benefits to the diagnosis and treatment of chronic diseases. On the one hand, an e-diagnosis system based on AI can efficiently analyze and classify the data obtained in IoMT to make a preliminary diagnosis of patients and provide support for the doctor to make the final diagnosis and specify the treatment plan. On the other hand, this e-diagnosis system makes it possible to realize remote supervision and management of patients with chronic illness. For example, the root of the diabetes management problem lies in the self-management of patients. The key to solving this problem is to tell patients how to monitor blood sugar, arrange diet, exercise, and rationally use drugs. The diabetes management system based on IoT technology provides the possibility to solve this problem. In remote areas lacking medical experts and professional medical equipment, mobile devices can provide data to the e-diagnosis system to use the services provided by IoMT to detect and classify diseases [9].

In this paper, an e-diagnosis system for detecting and classifying diabetes as an IoMT application is proposed, as shown in Fig. 1. Employing ML algorithms, this system aims to predict the diagnosis of diabetes based on patient data, provide doctors with a preliminary diagnosis, and return feedback on the doctor's guidance on diet, exercise, and blood glucose testing to patients. In addition, as shown on the left side of the figure, the IoMT enables the medical systems, applications and devices to connect with each other. Therefore, a patient's profile can be assessed by a doctor remotely through the Internet and shared by doctors from different medical institutions, no matter the community hospitals or large hospitals. In this way, the amount of paper medical records can be reduced to a large extent, and the patient does not need to go to the same hospital or even go to the hospital for follow-up visits in person.

2.2 Intelligent methods of diabetes prediction

By clarifying common problems, the emerging techniques in data science can bring benefits to other fields of science, including medicine. Numerous research has employed various machine learning or AI methods for diabetes prediction, such as artificial neural network (ANN), support vector machine, gradient boosting decision tree, and Naive Bayes.

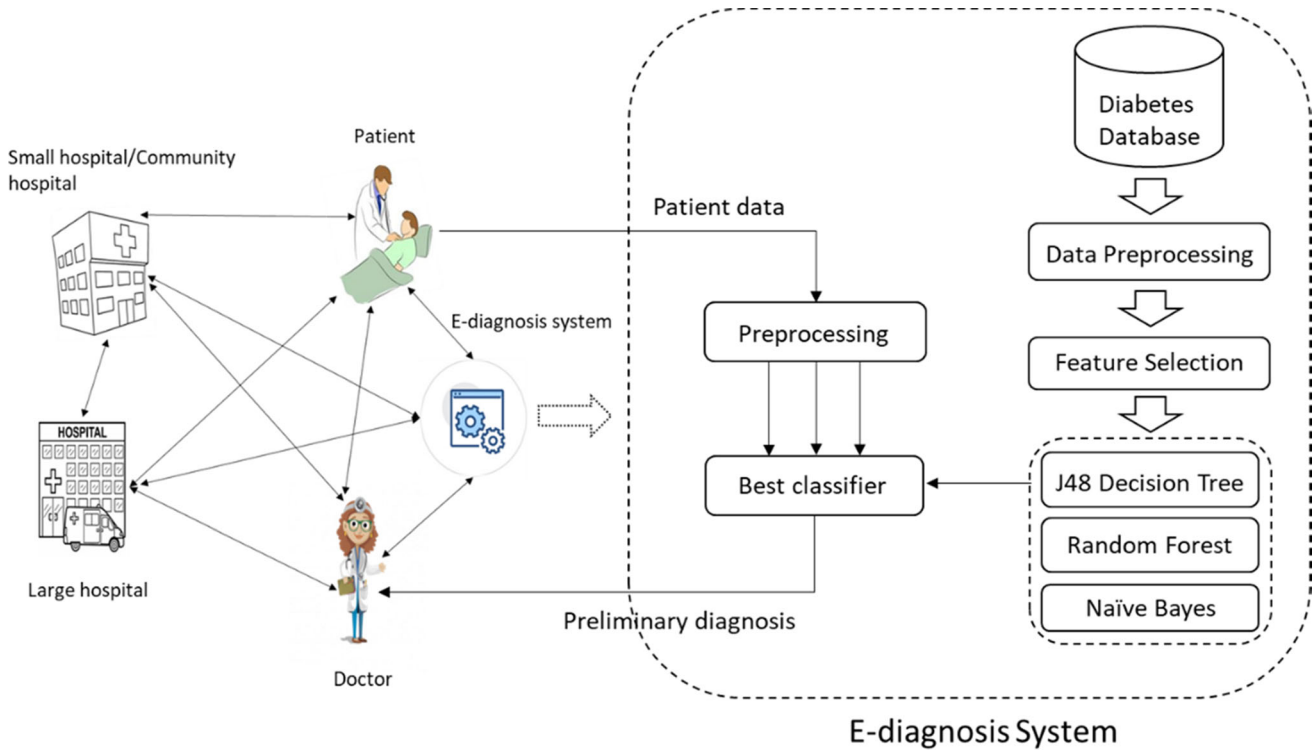


Fig. 1 An E-diagnosis system enabled in IoMT

In the study of Komi et al. [14], they use five various data mining techniques [ANN, elaboration likelihood model (ELM), Gaussian mixture model (GMM), support vector machine (SVM), and logistic regression] to explore the early prediction of diabetes. Their research results show that ANN performs best among the five techniques. Similar to Komi et al., Ramanujam et al. [15] and Kumar et al. [16] also contribute to the early prediction of diabetes, but with different approaches. The early diagnosis of diabetes and proper treatment will affect costs and mortality in the later stage. Early diagnosis and testing expenditures are significantly crucial. Therefore, people in rural areas are unlikely to afford early diagnosis and miss timely treatments, resulting in higher mortality [17]. In order to help the rural Indian people, Ramanujam et al. [15] develop a multilingual decision support system that integrates the predictive models and clinical decision support system. The design feature of the system is that users can not only evaluate diabetes with the help of nursing assistants but also evaluate diabetes by themselves. Kumar et al. [16] compare the performance of technique CatBoost with other ML techniques, including K-nearest neighbor, logistic regression, stochastic gradient descent, Gaussian Naive Bayes, and multilayer perceptron, in the early prediction of diabetes. In their research, CatBoost has the highest accuracy.

In addition, AI algorithms are also employed to analyze and classify iris images to diagnose diabetes. Samant and Agarwal [18, 19] study the diagnosis of diabetes through the changes in pigmentation in certain areas of the iris by using several ML algorithms. They use pre-image processing methods to obtain iris and crop out certain areas. Then, they use texture textural, statistical and wavelet features to observe the variances in the tissue pigmentation. Finally, five classifiers are employed to classify whether the patient has diabetes. Their results show that random forest outperforms other classifiers.

Although AI and machine learning pervade the fields of healthcare and non-communicable chronic diseases, due to the lack of explanation of these complex algorithms or models, their actual medical application rate is very low. Based on the existing literature, this paper chooses three classifier models, Naïve Bayes, random forest classifier, and J48 decision tree, to classify the Pima Indians Diabetes dataset in the R programming language. However, unlike the predecessors, the purpose of this study is to employ interpretable ML models to make our model clear and understandable to end-users regarding how we judge which features are important and how the choice of features affects the model's prediction results.

2.3 The selected machine learning algorithms

2.3.1 J48 decision tree

A decision tree (DT) is a supervised ML algorithm widely utilized in dealing with classification and regression issues. A leaf node in a decision tree represents the classification outcomes, and an internal node represents the judgment of attributes. Quinlan [20] calls the algorithm employed to establish the decision tree ID3, which uses a top-down learning method. The following steps describe the process of the DT: the first step is selecting the most appropriate attribute for the root node; secondly, the instances are divided into a number of subsets. For each subset, its instances are supposed to have identical attribute values; finally, every subset is repeated recursively until all instances have identical classes [21]. Figure 2 shows a part of a diagnosis decision tree, which can be interpreted easily. For instance, according to the tree, if a patient does not have inter-systolic noise, but has pre-cordial pain, then he or she has a prolapse.

The decision tree algorithm has been employed in many scientific regions, including the medical area. For example, Rochmawati et al. [22] use the DT algorithm to classify the COVID-19 symptom. They conclude that compared with Hoeffding tree, DT has a better performance but is more complicated. Other diseases can also be intelligently diagnosed by DT, for instance, Lupus disease [23] and coronary artery disease [24].

2.3.2 Random forest

Random forest (RF) is an extension of a decision tree and is composed of numerous single decision trees, each of which produces a category of prediction results. The category with the most votes in the forest contributes to the random forest classifier's final prediction result. For example, as shown in Fig. 3, among nine single decision trees in the forest, the prediction results of six trees are 1, and those of the remaining three trees are 0. Therefore, the prediction result of the RF is 1. The key to the good performance of this classifier is that the trees in the forest are relatively unrelated to each other, ensuring that the decision they make as a whole is better than the decisions made by each of them individually. [25].

Random forest uses a simple and powerful basic concept, called the wisdom of the crowd. The low correlation between trees is crucial to the success of the model. Under this premise, even if the prediction results of several trees are not correct, as long as the prediction results of most other trees are correct, then as a group, these trees can finally get the correct prediction results. In other words, the

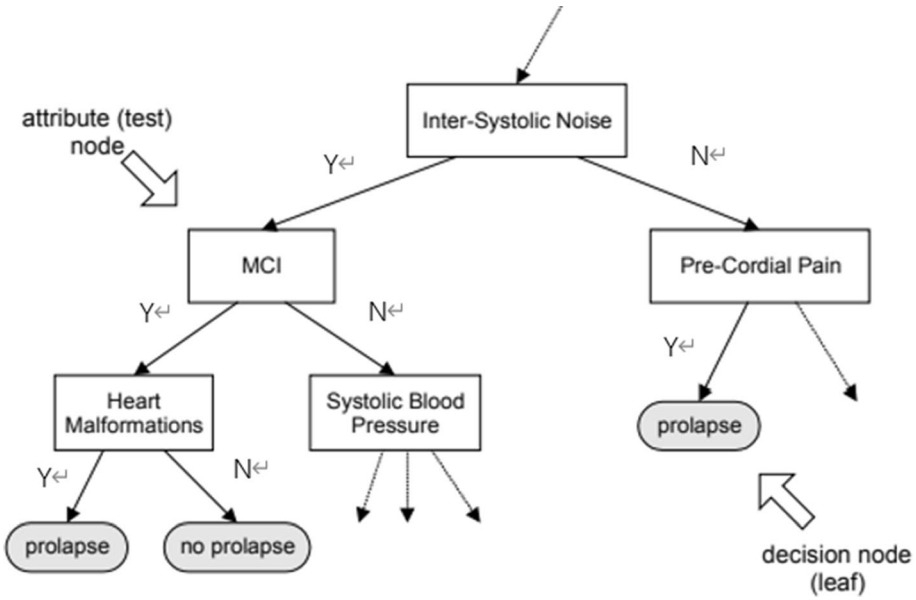


Fig. 2 Decision tree example

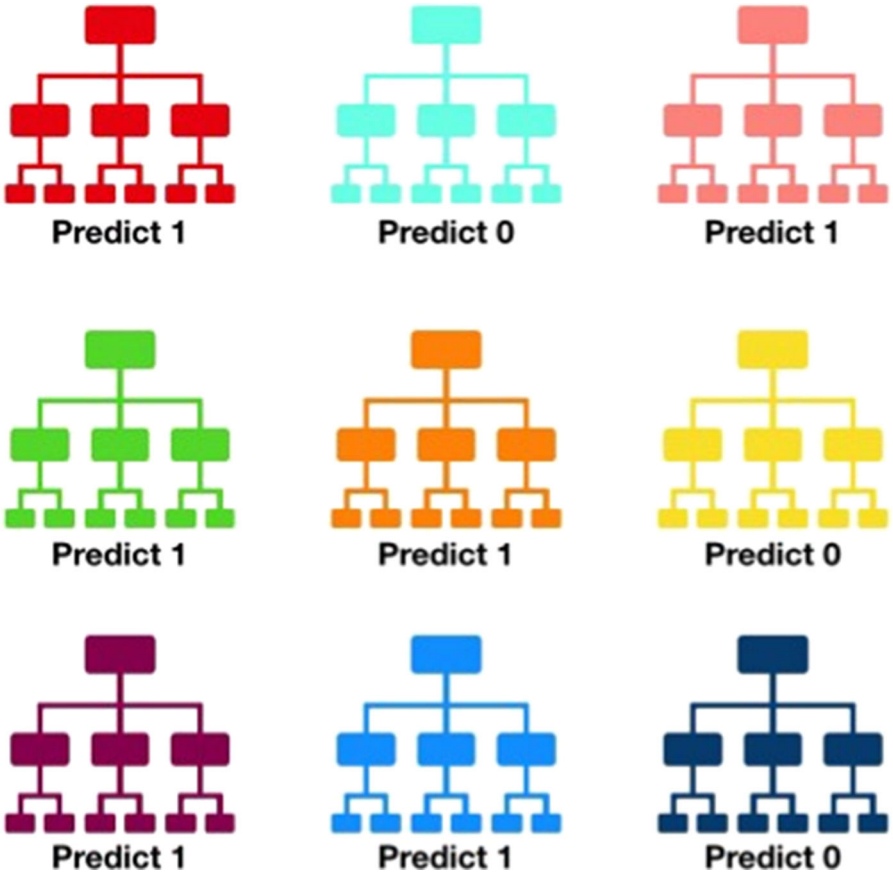


Fig. 3 Visualization of a random forest model making a prediction

random forest model performs well because abundant relatively unrelated models that operate as a whole perform better than any single constituent model.

Surface-enhanced Raman scattering (SERS) technology is very useful for analyzing biological samples. Nevertheless, it is difficult to obtain the required information from the collected data in the absence of labeled molecules. Therefore, Seifert [26] combines the random forest method with SERS data to solve this problem. The outcomes indicate that this approach is able to enhance the

performance of SERS technology. Apart from biology, RF can also be used in the areas of agriculture [27] and medical science [18, 19].

2.3.3 Naïve Bayes

The Bayesian classifier is a statistical classifier, and it is operated according to the Bayes theorem, classifying data into predetermined categories using conditional probability. Conditional probability can be understood as the probability that an event will take place if other events have already taken place. A Bayesian rule is an approach used to estimate the possibility of an attribute given a data set as input. The term "naive" of the algorithm's name refers to that it assumes that each attribute value is independent.

Naive Bayes (NB) is regarded as a descriptive as well as a predictive algorithm. The probabilities are descriptive and then employed to predict the categories of the untrained data. This method has several merits, as follows. First of all, it is easy to use. Secondly, the amount of training data NB needs for classification is not necessarily large. In addition, although the NB classifier is naively designed and its assumption seems to be too simple, it performs well in a number of complicated real-world situations [28].

Pandiangan et al. [29] consider that in his applied AI research, a student's study time and duration is an essential index to evaluate the quality of the university. They then employ the NB classification algorithm and DT algorithm to predict the student's study period, evaluate academic performance and identify correlations for improving the quality of the university. In the field of education, Daniati [30] develops a decision support system for students to select suitable programs using DBSCAN and Naive Bayes. Different from them, Akbar et al. [31] integrate the Internet of Things with the NB algorithm to develop an intelligent laundry mobile application.

3 Data and methodology

3.1 Dataset exploration and pre-processing

Although there are now larger, more complex diabetes datasets, the Pima Indian Diabetes dataset has remained a benchmark for diabetes classification research. Given the presence of a binary outcome variable, the dataset naturally lends itself to supervised learning and, in particular, logistic regression. However, various ML algorithms have been employed to produce classification models based on this dataset for not being limited to a singular type of model.

In this research, our focus is to analyze the Pima Indian Dataset with advanced algorithms to work with IoMT effectively. The dataset was downloaded from Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database) and is available via a CC0: Public Domain License and is properly anonymized and does not contain any identifiable features of the patient subjects. As seen in Table 1, it records eight causal characteristics and the corresponding classification. The dataset has 9 columns and 768 rows (500 non-diabetics and 268 diabetics). The binary classification outcome variable takes (0 or 1) values, where 0 indicates a negative test for diabetes, and 1 implies a positive test. Table 1 shows the dataset features (columns) and descriptions.

The dataset has no null values and no missing values. However, according to domain knowledge [32], there are inconsistent values for the attributes: glucose concentration (Gluc), blood pressure (BP), skin fold thickness (Skin), insulin and BMI, whereby zero values are not within the normal range and are therefore inaccurate (Table 2).

The scatterplot matrix is helpful to identify the pair-wise relationships of the features preliminarily. If the points are scattered, it means that there is no obvious relationship, while if the points are roughly arranged in a straight line, it means that they are linearly related. While referring to the scatterplot matrix in Fig. 4, the most closely correlated/proportional features include [pregnancy and age], [skin thickness and BMI], and [glucose and insulin] because their scatterplot figures all show a positive correlation.

As shown in Fig. 5, there are outliers in DPF, age, insulin, glucose, BMI, and blood pressure features, which might be due to other underlying factors. It would be best to standardize the data to avoid the ill effects of the outliers. The dataset is not a very large one, so it would be better to avoid removing rows unnecessarily.

There seems to be a demonstrable difference in the performance and efficiency of prediction classification models depending on the pre-processing methodology. Therefore, in the first round of experiments, minimal pre-processing was done. The second time around, however, feature selection algorithms were applied.

Since there were no missing or null values, only one data pre-processing technique was applied in the first round. This was to impute the median value on the features that had invalid zero values.

Tree algorithms such as decision trees, Naïve Bayes models, and random forests are not highly sensitive to non-normalized data, so no scaling was done as a way to keep the testing similar for the three machine learning models.

With only eight features, it may seem counter-intuitive to reduce the features further, but it can reduce some of the noise in classification and pinpoint subtle groupings synthesized by combining existing classes. In rounds two and

Table 1 Overview of Pima Indian diabetes dataset

Feature	Description	Data type	Range
Preg	Number of times pregnant	Numeric	[0, 17]
Gluc	Plasma glucose concentration at 2 Hours in an oral glucose tolerance test (GTIT)	Numeric	[0, 199]
BP	Diastolic Blood Pressure (mm Hg)	Numeric	[0, 122]
Skin	Triceps skin fold thickness (mm)	Numeric	[0, 99]
Insulin	2-Hour Serum insulin (μh/ml)	Numeric	[0, 846]
BMI	Body mass index [weight in kg/(Height in m)]	Numeric	[0, 67.1]
DPF	Diabetes pedigree function	Numeric	[0.078, 2.42]
Age	Age (years)	Numeric	[21, 81]
Outcome	Binary value indicating non-diabetic /diabetic	Factor	[0,1]

Table 2 Statistical summary of Pima Indians diabetes dataset

Features	Preg	Gluc	BP	Skin	Insulin	BMI	DPF	Age
Min.	0.000	0.0	0.00	0.00	0.0	0.00	0.0780	21.00
1st Qu.	1.000	99.0	62.00	0.00	0.0	27.30	0.2437	24.00
Median	3.000	117.0	72.00	23.00	30.5	32.00	0.3725	29.00
Mean	3.845	120.9	69.11	20.54	79.8	31.99	0.4719	33.24
3rd Qu.	6.000	140.2	80.00	32.00	127.2	36.60	0.6262	41.00
Max	17.000	199.0	122.00	99.00	846.0	67.10	2.4200	81.00

three of experiments, the feature selection methodologies that were employed included: PCA, *k*-means clustering, and importance ranking.

3.2 Methods

As it relates to the classification and prediction of diabetes and other non-communicable diseases, ML and DL models have become an important research area for many years. Numerous tools and models have been put forward to help solve the diagnosis prediction problem, including convolutional neural networks (CNN), artificial neural networks (ANN), and combined or hybrid machine learning models.

Based on recent research, the top algorithms (exclusive of combined models and neural networks) used to train and forecast the Pima Indian Diabetes dataset included: the J48 decision tree with 94.44% accuracy [32], as well as random forest (94% accuracy) and Naïve Bayes (91% accuracy) models [5]. All the proposed ML algorithms chosen to be used in this research paper are classification models.

All models that have previously been applied to the dataset have used additional machine learning techniques to pre-process or engineer the dataset, including bootstrapping, resampling and *k*-folds, as well as the information gain method [5]. Mercaldo et al. [33] made use of feature selection algorithms, such as GreedyStepwise and BestFirst, to determine the discriminatory clauses. Iyer et al. [34] and Zia and Khan [32] also used a percentage split of 70:30 for the training and testing sets. Other than the 70:30 percentage split, none of these techniques were used for simplicity for this project. Instead, the feature selection methodologies mentioned in the previous section were used.

Common evaluation methods for model performance included accuracy, precision, sensitivity, specificity, *F*-measure (*F*-score), and mean square error (MSE), as well as comparing performance on pre-processed versus non-pre-processed data [5].

Explainable AI (XAI) is the concept within artificial intelligence whereby decisions made by a machine learning model can be understood by its users [35]. The complementary concept of interpretability refers to the ability to observe cause and effect within such a model. Model interpretability can be intrinsic, and such as with decision trees. However, interpretability can also be introduced to a model (post hoc) by applying functions on pre-trained models to generate explanations. Concerning non-communicable diseases, not much work has been done in examining explainable machine learning models [35].

It is important to note that interpretable models may not always be explainable to an extent where the human mind fully comprehends the steps taking place to arrive at a decision made by a machine learning model.

An example of a post hoc interpreting method is Shapley Additive exPlanations (SHAP), which is a framework (based on the idea in game theory of 'Shapley values', a

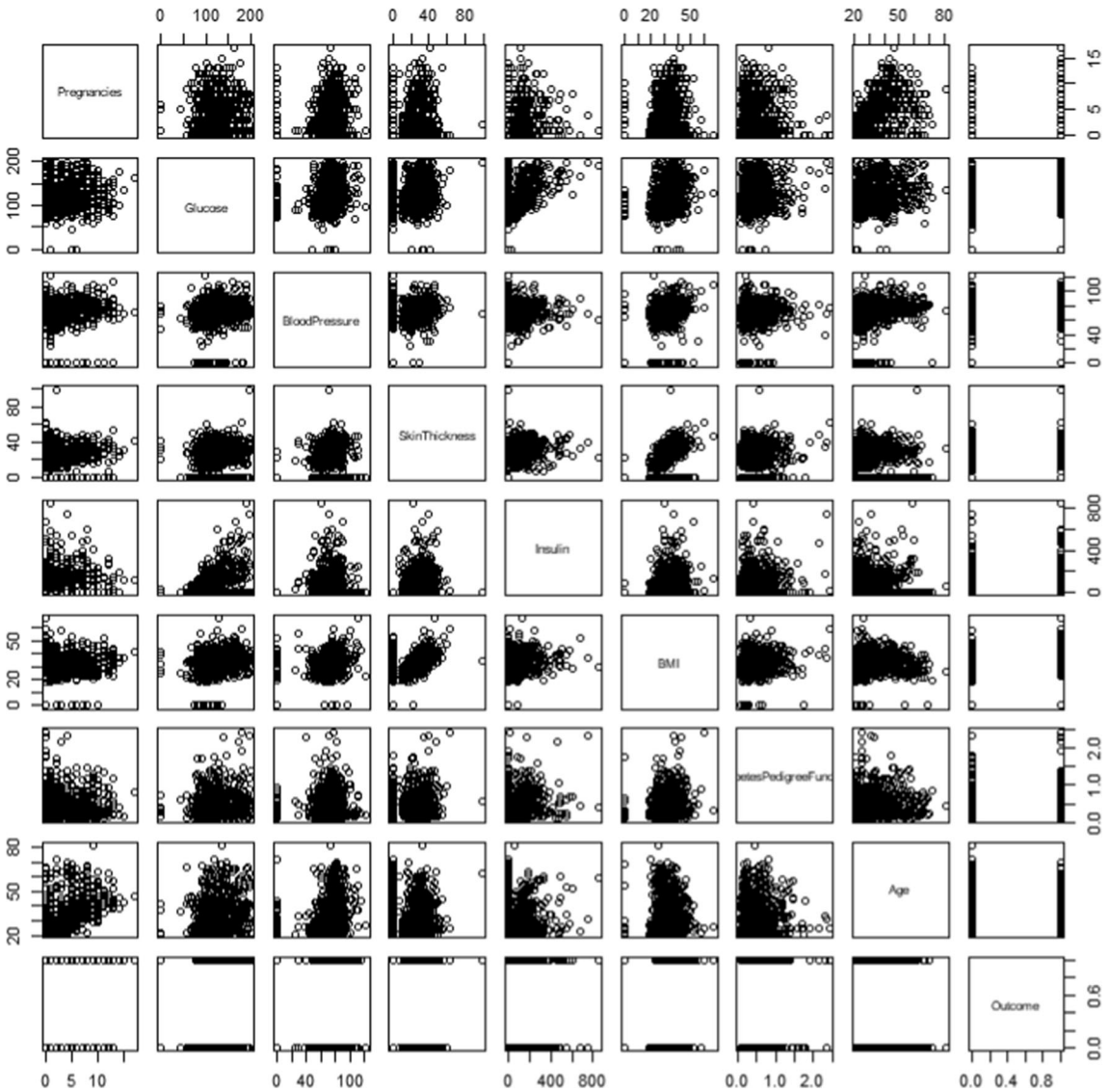


Fig. 4 Scatterplot matrix of features

description of a player's contribution to the result of a coalitional game) that builds on an additive feature attribution approach, to generate exclusive interpretation models that can offer interpretations on a classification model's decisions in the form of specific feature contributions [36]. Benefits of the SHAP framework include a capability to meet local accuracy as well as consistency.

The SHAP summary plot shows feature importance ranked in descending order denoted by the *y*-axis as well as the effect on how the feature value is associated with the prediction as denoted by the *x*-axis, which in turn can be used to interpret the correlation between a feature and the outcome [35].

We can use post hoc interpretation to identify whether the model we have trained has accurately captured the details of the real-world decision-making process from the dataset. In addition, it can be used to highlight biases and errors in the machine learning models [35]. In this paper, the goal is to use the interpretable AI method to make our model clear and understandable to end-users from two aspects. The first one is to judge which features are important, and details are presented in Sect. 4.1. In Sect. 4.2, we will discuss the other aspect of how the choice of features affects the models' prediction results.

4 Experiment and results

4.1 Feature selection

In order to make our model clear and explainable, this part shows the end-users how we judge which features are

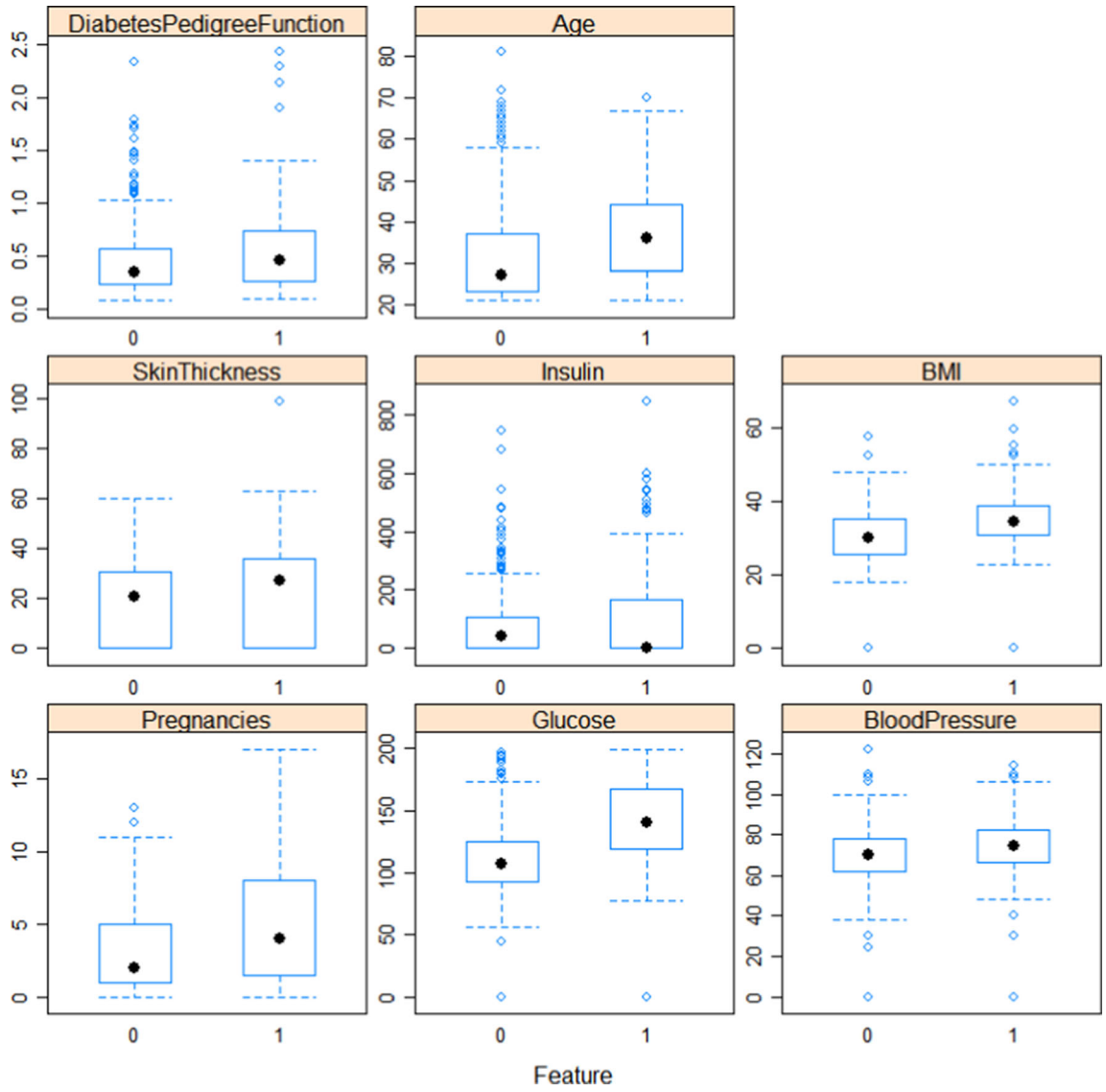


Fig. 5 Box and whisker plots showing feature distribution for each outcome class

important. For feature selection purposes within the experiments with the machine learning models, this research performs *k*-means clustering, principal component analysis (PCA), and importance ranking on the dataset.

Starting with PCA, the feature groupings were observed within the plot in Fig. 6, where arrows that are close together represent closely related features. It can be seen that the following are closely related:

- Pregnancy and age
- Glucose and blood pressure (BP)
- BMI, DPF, insulin level, and skin thickness

Following this, *k*-means clustering was used where various values of k were used and observed. According to Fig. 7, we can see that $k = 2$ has the best separation of boundaries or groupings, but $k = 3$ also has decent clustering as well and could be useful.

Lastly, importance ranking was used to identify the features that had the highest mathematical importance to the outcome.

Through the use of the above-mentioned methodologies, as shown in Fig. 8, we can see that the features Glucose, BMI, age, insulin and skin rank very highly in helping to classify the data. In contrast, the DPF (Diabetes Pedigree Function), Blood Pressure, and (number of) Pregnancies rank very low. By using these methodologies, the dataset was scaled back to two versions. In the second round of experiments, three (3) factors (glucose, BMI, and age) were used as the features for classification by the ML algorithms. In contrast, five (5) factors were chosen (glucose, BMI, age, insulin, and skin thickness) in the final round.

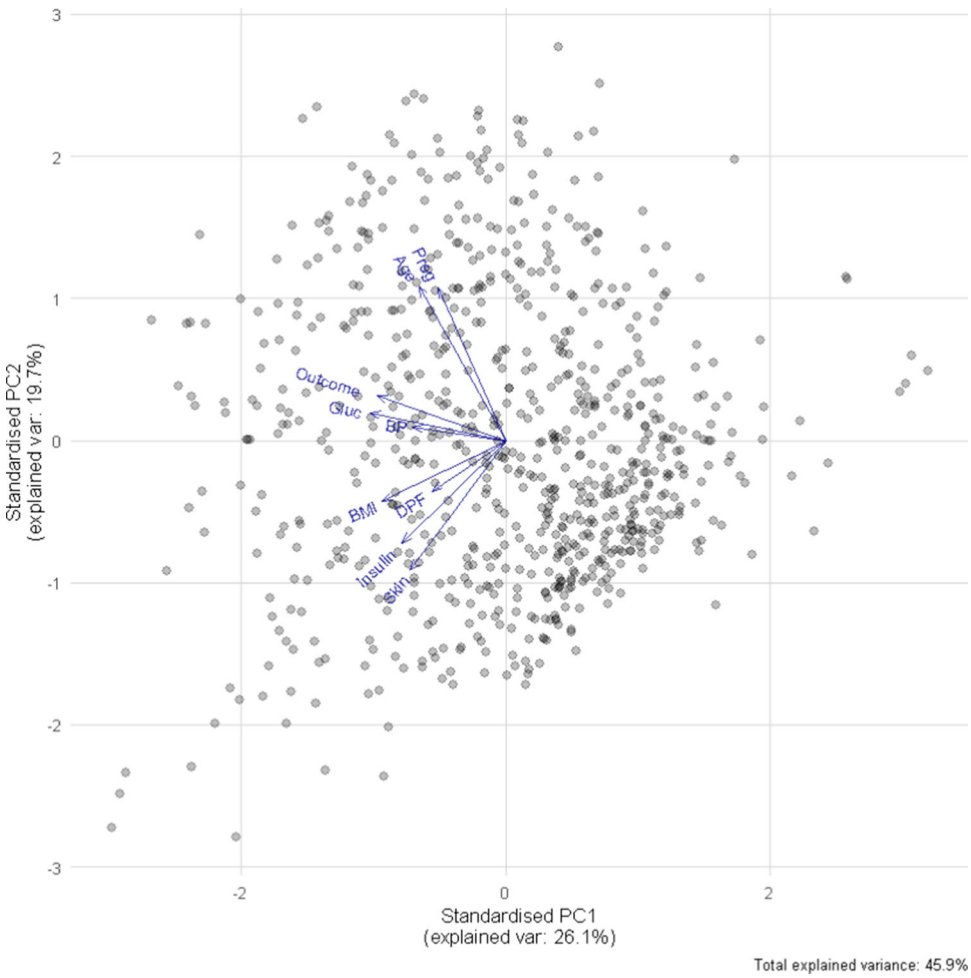


Fig. 6 Principal component analysis

4.2 Results of machine learning algorithms

In this paper, the three ML algorithms that were used to analyze the Pima Indian Diabetes dataset are J48 decision tree, random forest, and Naïve-Bayes. The same training and testing sets were used for all three as a sort of control environment. The data subsets were manually split into 538 and 230 samples, respectively (70/30 split).

Six metrics were used to evaluate the results, including the accuracy, precision, sensitivity, specificity, *F*-score, and area under the curve (AUC). These variables are computed through the confusion matrix, a matrix showing the values of the actual outcome classes and the predicted outcome classes on the testing set (see Table 3 and formulas below).

The accuracy refers to the percentage of all samples that have been predicted correctly. It is the ratio of the sum of true positives and true negatives to the total number of predictions made.

Precision refers to the percentage of all samples that have been correctly predicted as true among all those which were predicted as true, even if they were false.

The sensitivity refers to the percentage of all samples that have been correctly predicted as true among all those which were predicted true as well as those predicted false but were true.

The specificity refers to the percentage of all samples that have been correctly predicted as false among all those which were false even if predicted incorrectly.

The standard *F*-score (*F*1-score) is an indicator of a binary classification model's accuracy, calculated by the weighted average of the precision and sensitivity. To be specific, it is calculated by dividing the product of the precision and sensitivity by the sum of the precision and sensitivity and multiplying the result by two.

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

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

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

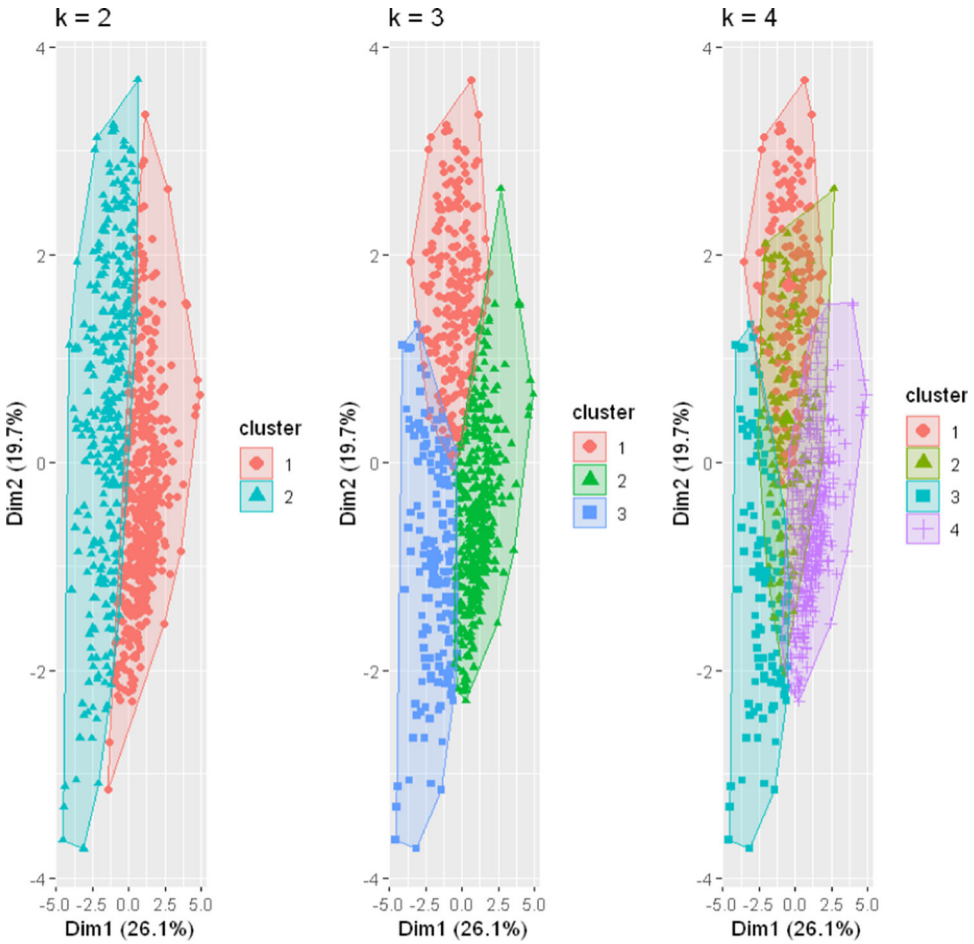
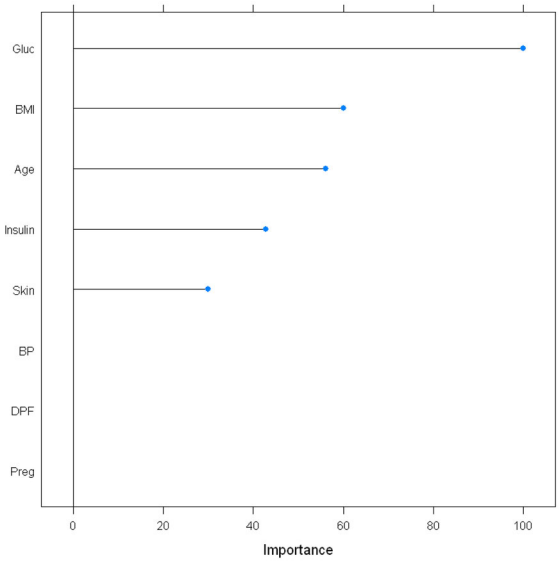


Fig. 7 *k*-Means clustering with $k = 2, 3, 4$



Gluc	100.00
BMI	59.98
Age	56.04
Insulin	42.80
Skin	29.91
DPF	0.00
BP	0.00
Preg	0.00

Fig. 8 Results of rpart_importance function in *R*

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})}$$

$$F - \text{Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Table 3 Confusion matrix template

	Actual positive	Actual negative
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative

Table 4 J48 decision tree confusion matrix

	Actual positive	Actual negative
Predicted positive	107	44
Predicted negative	14	65

Table 5 J48 decision tree confusion matrix with feature selection (3-factor)

	Actual positive	Actual negative
Predicted positive	106	45
Predicted negative	12	67

Table 6 J48 decision tree confusion matrix with feature selection (5-factor)

	Actual positive	Actual negative
Predicted positive	107	44
Predicted negative	12	67

where TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

4.2.1 J48 decision tree

The J48 decision tree is an implementation of algorithm ID3 (Iterative Dichotomiser 3) decision tree, developed by the WEKA (Java-based ML software) team and included in R in the package RWeka. Each attribute of the dataset is used to split the data into smaller modules used to classify/make a prediction. Tables 4, 5, and 6 show the confusion matrixes of the J48 decision tree on the 3-factor subset, 5-factor subset, and full dataset; further analysis on the evaluation metrics based on these matrixes will be provided in Sect. 5.

Table 7 Random forest confusion matrix

	Actual positive	Actual negative
Predicted positive	136	15
Predicted negative	28	51

Table 8 Random forest confusion matrix with feature selection (3-factor)

	Actual positive	Actual negative
Predicted positive	123	28
Predicted negative	31	48

Table 9 Random forest confusion matrix with feature selection (5-factor)

	Actual positive	Actual negative
Predicted positive	121	30
Predicted negative	30	49

4.2.2 Random forest

A random forest model refers to a tree ensemble that works similarly to a decision tree but, instead of splitting at a single attribute, forms random groups of attributes to make classifications. As a result, more processing is done, improving the accuracy of this model over a single tree model. Tables 7, 8, and 9 show the confusion matrixes of random forest on the 3-factor subset, 5-factor subset, and full dataset.

4.2.3 Naïve Bayes

Naïve Bayes is regarded as a simplistic model based on the "naive" assumption that all predictor variables are independent of each other, therefore, will not impact another. It uses Bayes probability theory on all attributes to determine the likelihood of the class outcome and make a prediction. Tables 10, 11, and 12 display the confusion matrixes of

Table 10 Naive Bayes confusion matrix

	Actual positive	Actual negative
Predicted positive	131	29
Predicted negative	20	50

Table 11 Naive Bayes confusion matrix with feature selection (3-factor)

	Actual positive	Actual negative
Predicted positive	133	30
Predicted negative	18	49

Table 12 Naive Bayes confusion matrix with feature selection (5-factor)

	Actual positive	Actual negative
Predicted positive	130	30
Predicted negative	21	49

Naive Bayes on the 3-factor subset, 5-factor subset, and full dataset, respectively, and further analysis on the evaluation metrics based on these matrixes will be provided in Sect. 5.

4.2.4 AUC-ROC curves

The receiver operating characteristics (ROC) curve and the resulting area under the curve (AUC) provide a vital performance measurement for classification models and represent the degree of separability of classes.

AUC can be seen as the likelihood that the represented model ranks a random positive example higher than a

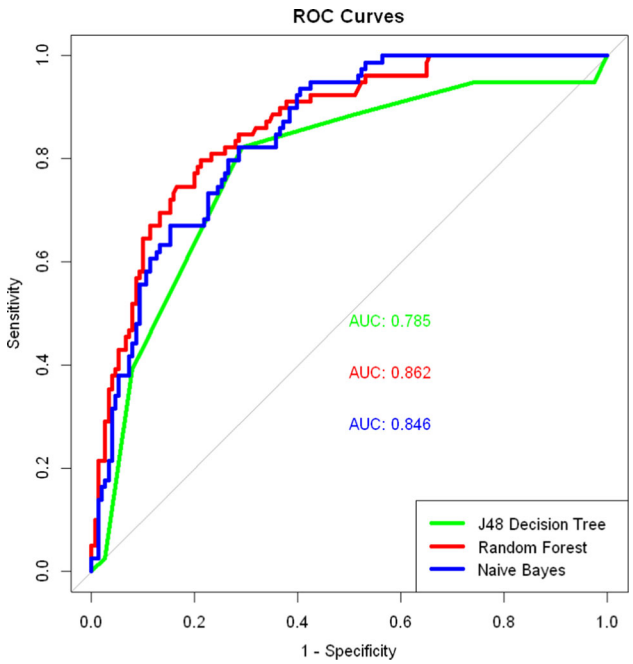


Fig. 9 ROC curves for all models on imputed data

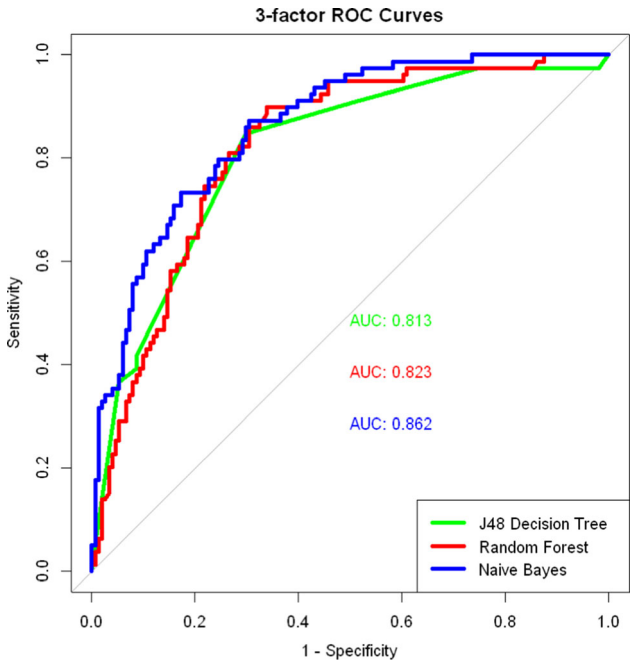


Fig. 10 ROC curves for all models using PIMA dataset with 3 features

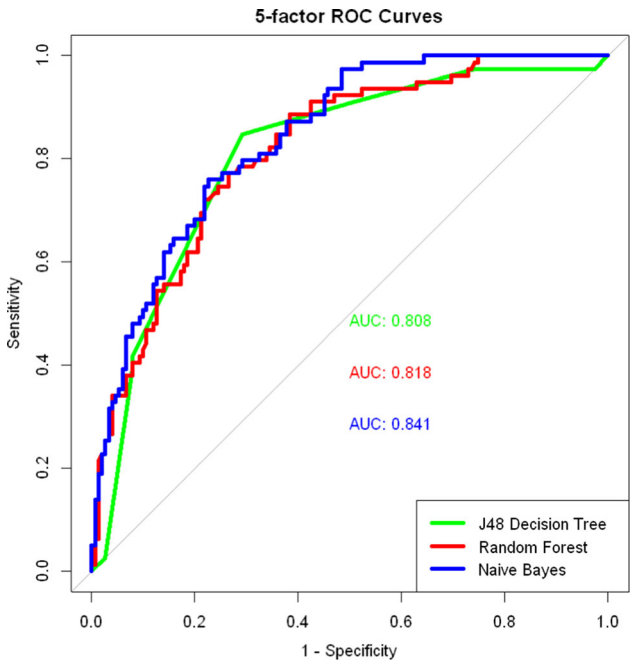


Fig. 11 ROC curves for all models using PIMA dataset with 5 features

Table 13 Results of all models using the only imputation

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	*F*-score (%)	AUC (%)
J48 decision tree	74.78	70.86	**88.43**	59.63	78.68	78.55
Random forest	**79.57**	**89.40**	81.33	**75.00**	**85.17**	**86.24**
Naïve Bayes	78.67	81.88	86.75	63.29	84.24	84.63

Bold values represent the highest values in each set of measurement among three methods

Table 14 Results of all models using feature selection (3-factor)

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	*F*-score (%)	AUC (%)
J48 decision tree	75.22	70.20	**89.83**	59.82	78.81	81.28
Random forest	75.22	**82.12**	80.52	**64.47**	81.31	82.27
Naïve Bayes	**79.13**	81.60	88.08	62.03	**84.71**	**86.15**

Bold values represent the highest values in each set of measurement among three methods

Table 15 Results of all models using feature selection (5-factor)

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	*F*-score (%)	AUC (%)
J48 decision tree	75.65	70.86	**89.92**	60.36	79.26	80.84
Random forest	73.91	80.79	79.74	**62.34**	80.26	81.77
Naïve Bayes	**77.83**	**81.25**	86.09	62.03	**83.60**	**84.10**

Bold values represent the highest values in each set of measurement among three methods

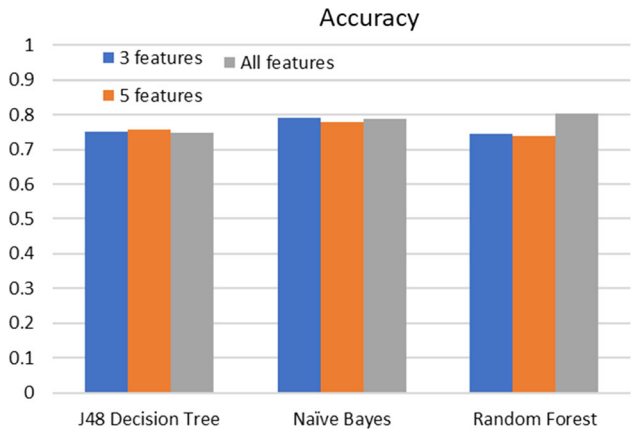


Fig. 12 Graph comparing accuracy across models and datasets

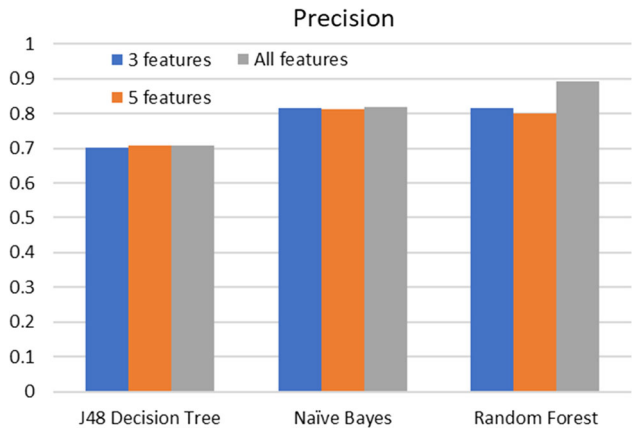


Fig. 13 Graph comparing precision across models and datasets

random negative example. AUC measures the entire two-dimensional area underneath the entire ROC curve from (0.0) to (1.1) with a maximum value of 1. The ROC curves for all models using a 3-factor subset, 5-factor subset, and full dataset are shown in Figs. 9, 10, and 11.

4.3 Final results

Tables 13, 14, and 15 combine the data of the above tables and compare the performance of the classifiers on each dataset, while Figs. 12, 13, 14, 15, 16, and 17 compare the performance of the algorithms across datasets from the perspective of each evaluation metric.

5 Discussion and conclusion

This paper presented classification models suitable for electronic diagnostic systems to be implemented in the IoMT environment. The models were trained using three machine learning algorithms and evaluated to predict

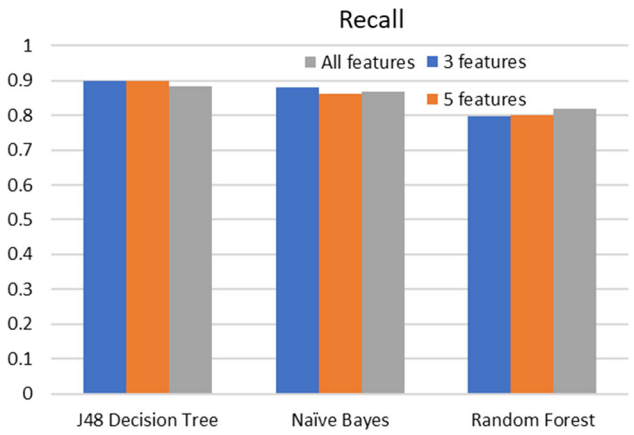


Fig. 14 Graph comparing sensitivity across models and datasets

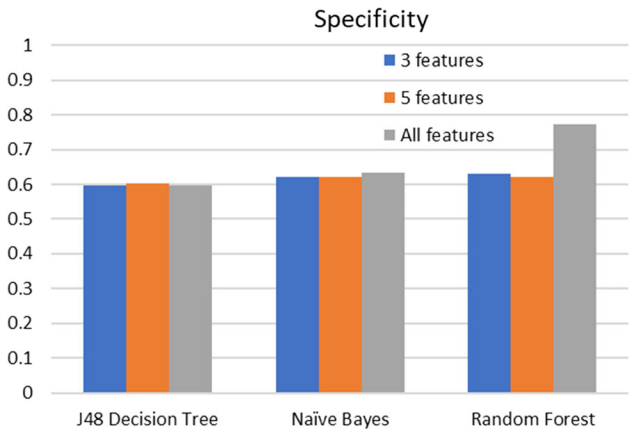


Fig. 15 Graph comparing specificity across models and datasets

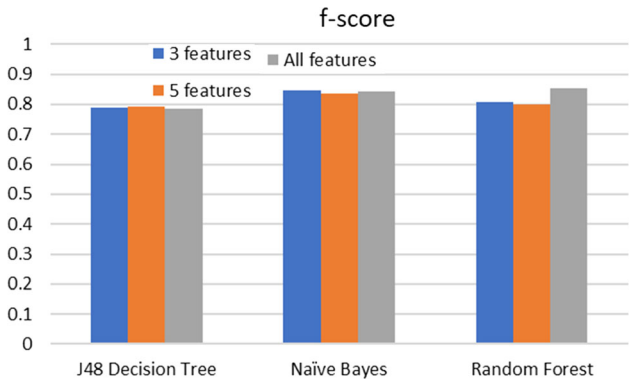


Fig. 16 Graph comparing F-score across models and datasets

whether a subject's diabetes mellitus diagnosis is positive according to eight given attributes.

The experimental results in Sect. 4.3 show that, on the full Pima Indian Diabetes dataset, the random forest classifier outperformed both the Naïve Bayes and J48 decision tree with accuracy metric (79.57%), precision (89.40%), specificity (75.00%), *f*-score (85.17%), and AUC (86.24%), while the J48 had the best sensitivity (88.43%) of the three.

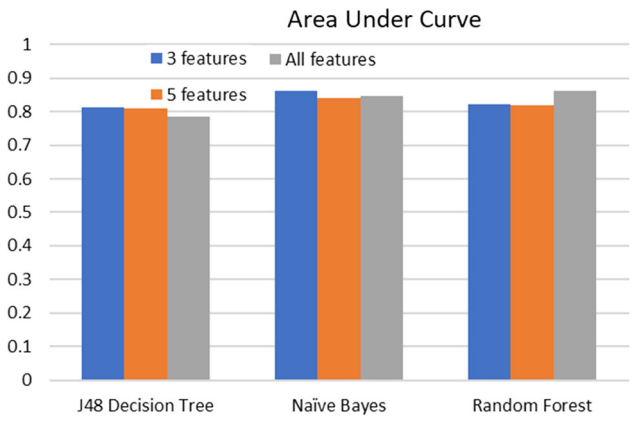


Fig. 17 Graph comparing AUC across models and datasets

The vast difference between sensitivity and specificity is likely due to the imbalance of samples of class 0 and class 1.

However, the results for the 3-factor and 5-factor data subsets that used feature selection show that the Naïve Bayes classification model outperformed both the random forest and the J48 decision tree models for accuracy. The Naïve Bayes model on the 3-factor data subset performed just as well as the random forest model on the full dataset with an accuracy of 79.13% compared to 79.57%, which was the highest accuracy in this experiment. We can conclude that a Naïve Bayes model works well with a more fine-tuned selection of features for binary classification but falls short with numerous correlated features, while random forest works better with more features.

The J48 decision tree model consistently performed with a sensitivity rate range of 88.43% (full dataset) to 89.92% (5-factor data subset), showing that it is good at predicting the presence of diabetes no matter how many features it has to work with.

Although the models in this experiment are close to 80% accuracy, our research outputs are in line with similar work by Iyer et al. [34] and can be improved. A positive sign was that there was no overfitting in our approach. In other words, results are more genuine and close to reality. The top five most important features as indicated by the feature selection methodologies (glucose, BMI, age, insulin, and skin fold thickness) are in line with existing guidelines, which indicate that age and weight (indicated through BMI and skin fold thickness) play a huge role in the diagnosis and occurrence of diabetes mellitus.

Based on this experiment, an e-diagnosis system for detecting and classifying diabetes as an IoMT application is proposed. The data from IoMT and the advanced ML algorithms enable the e-diagnosis system to predict the diagnosis of diabetes based on patient data, provide doctors with a preliminary diagnosis, and return feedback on the doctor's guidance on diet, exercise, and blood glucose

testing to patients. This system also contributes to the remote monitoring and management of patients with chronic diseases. Employing the IoMT can make data collection and analysis easy. In the IoMT, different medical systems, applications, devices, patients and doctors are able to connect with each other, leading to the accessibility of massive data to the e-diagnosis system. With these medical data, the system can be improved constantly. Moreover, we can develop more algorithms to provide better accuracy if we cannot control the quality of datasets provided by medical researchers.

Our future work will include developing innovative methods and applying them to other types of medical analysis. For example, the accuracy may be enhanced by using suitable pre-processing techniques for data management and analysis. New automation and automated processes with IoMT can be developed to improve diabetes mellitus prediction and other non-communicable diseases.

Acknowledgements This research is partly supported by VC Research (VCR 0000159) for Prof Chang.

Declarations

Conflict of interest There are no any conflicts of interest from authors.

References

1. World Health Organisation (2016) Global Report on Diabetes. https://www.who.int/publications-detail/global-report-on-diabetes.Accessed: 24 Apr 2020
2. World Health Organization (2013) Global action plan for the prevention and control of NCDs 2013–2020. https://apps.who.int/iris/bitstream/handle/10665/94384/9789241506236_eng.pdf;jsessionid=5A344A4B152EABE6C021C9A7EEE444C8?sequence=1.Accessed 24 Feb 2021
3. Schulz LO, Bennett PH, Ravussin E, Kidd JR, Kidd KK, Esparza J, Valencia ME (2006) Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the US. Diabetes Care 29(8):1866–1871
4. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the annual symposium on computer application in medical care, pp 261–265
5. Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T (2019) Current techniques for diabetes prediction: review and case study. Appl Sci 9(21):4604
6. Pratap Singh R, Javaid M, Haleem A, Vaishya R, Ali S (2020) Internet of Medical Things (IoMT) for orthopaedic in COVID-19 pandemic: roles, challenges, and applications. J Clin Orthop Trauma 11(4):713–717. https://doi.org/10.1016/j.jcot.2020.05.011
7. Pustokhina IV, Pustokhin DA, Gupta D, Khanna A, Shankar K, Nguyen GN (2020) An effective training scheme for deep neural network in edge computing enabled Internet of Medical Things (IoMT) systems. IEEE Access 8:107112–107123. https://doi.org/10.1109/ACCESS.2020.3000322
8. Alsubaei F, Abuhussein A, Shandilya V, Shiva S (2019) IoMT-SAF: Internet of Medical Things security assessment framework. Internet Things 8:100123. https://doi.org/10.1016/j.iot.2019.100123
9. Khan SU, Islam N, Jan Z, Din IU, Khan A, Faheem Y (2019) An e-Health care services framework for the detection and classification of breast cancer in breast cytology images as an IoMT application. Futur Gener Comput Syst 98:286–296. https://doi.org/10.1016/j.future.2019.01.033
10. Divya K, Sirohi A, Pande S, Malik R (2021) An IoMT assisted heart disease diagnostic system using machine learning techniques. In: Hassanien AE, Khamparia A, Gupta D, Shankar K, Slowik A (eds) Cognitive Internet of Medical Things for smart healthcare, vol 311. Springer, New York, pp 145–161. https://doi.org/10.1007/978-3-030-55833-8_9
11. Kumar PM, Devi Gandhi U (2018) A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. Comput Electr Eng 65:222–235. https://doi.org/10.1016/j.compeleceng.2017.09.001
12. Kaur H, Kumari V (2020) Predictive modelling and analytics for diabetes using a machine learning approach. In: Applied computing and informatics, ahead-of-print (ahead-of-print). https://doi.org/10.1016/j.aci.2018.12.004
13. Adeniyi EA, Ogundokun RO, Awotunde JB (2021) IoMT-based wearable body sensors network healthcare monitoring system. In: Marques G, Bhoi AK, de Albuquerque VHC, Hareesha KS (eds) IoT in healthcare and ambient assisted living, vol 933. Springer, Singapore, pp 103–121. https://doi.org/10.1007/978-981-15-9897-5_6
14. Komi M, Li J, Zhai Y, Zhang X (2017). Application of data mining methods in diabetes prediction. In: 2017 2nd international conference on image, vision and computing (ICIVC), Chengdu, China, pp 1006–1010. https://doi.org/10.1109/ICIVC.2017.7984706
15. Ramanujam E, Chandrakumar T, Thivyadharsine KT, Varsha, D (2020) A multilingual decision support system for early detection of diabetes using machine learning approach: case study for rural Indian people. In: 2020 fifth international conference on research in computational intelligence and communication networks (ICRCICN). Bangalore, India, pp 17–21. https://doi.org/10.1109/ICRCICN50933.2020.9296187
16. Kumar PS, Kumari AK, Mohapatra S, Naik B, Nayak J, Mishra M (2021) CatBoost ensemble approach for diabetes risk prediction at early stages. In: 2021 1st Odisha international conference on electrical power engineering, communication and computing technology (ODICON). Bhubaneswar, India, pp 1–6. https://doi.org/10.1109/ODICON50556.2021.9428943
17. Maan V, Vijaywargiya J, Srivastava M (2020) Diabetes prognostication—an aptness of machine learning. In: 2020 international conference on emerging trends in communication, control and computing (ICONC3). Lakshmangarh, Sikar, India, pp 1–5. https://doi.org/10.1109/ICONC345789.2020.9117465
18. Samant P, Agarwal R (2018) Machine learning techniques for medical diagnosis of diabetes using iris images. Comput Methods Programs Biomed 157:121–128. https://doi.org/10.1016/j.cmpb.2018.01.004
19. Samant P, Agarwal R (2018) Comparative analysis of classification based algorithms for diabetes diagnosis using iris images. J Med Eng Technol 42:35–42. https://doi.org/10.1080/03091902.2017.1412521
20. Quinlan JR (1996) Learning decision tree classifiers. ACM Comput Surv (CSUR) 28(1):71–72
21. Saxena R (2017) How decision tree algorithm works. https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/. Accessed Apr 40.

22. Rochmawati N, Hidayati HB, Yamasari Y, Yustanti W, Rakhmawati L, Tjahyaningtijas HPA, Anistyasari Y (2020) Covid symptom severity using decision tree. In: 2020 third international conference on vocational education and electrical engineering (ICVEE). Surabaya, Indonesia, pp 1–5. https://doi.org/10.1109/ICVEE50212.2020.9243246
23. Gomathi S, Narayani V (2015) Monitoring of Lupus disease using decision tree induction classification algorithm. In: 2015 international conference on advanced computing and communication systems. Coimbatore, India, pp 1–6. https://doi.org/10.1109/ICACCS.2015.7324054
24. Abdar M, Nasarian E, Zhou X, Bargshady G, Wijayaningrum VN, Hussain S (2019) Performance improvement of decision trees for diagnosis of coronary artery disease using multi filtering approach. In: 2019 IEEE 4th international conference on computer and communication systems (ICCCS). Singapore, pp 26–30. https://doi.org/10.1109/CCOMS.2019.8821633
25. Yiu T (2019) Understanding random forest—how the algorithm works and why it is so effective. https://towardsdatascience.com/understanding-random-forest-58381e0602d2. Accessed 22 May
26. Seifert S (2020) Application of random forest based approaches to surface-enhanced Raman scattering data. Sci Rep 10:5436. https://doi.org/10.1038/s41598-020-62338-8
27. You J, van der Klein SAS, Lou E, Zuidhof MJ (2020) Application of random forest classification to predict daily oviposition events in broiler breeders fed by precision feeding system. Comput Electron Agric 175:105526. https://doi.org/10.1016/j.compag.2020.105526
28. Burdi F, Setianingrum AH, Hakiem N (2016) Application of the Naive Bayes method to a decision support system to provide discounts (case study: PT. Bina Usaha Teknik). In: 2016 6th international conference on information and communication technology for The Muslim World (ICT4M). Jakarta, pp 281–285. https://doi.org/10.1109/ICT4M.2016.064
29. Pandiangan N, Buono MLC, Loppies SHD (2020) Implementation of decision tree and Naïve Bayes classification method for predicting study period. J Phys Conf Ser 1569:022022. https://doi.org/10.1088/1742-6596/1569/2/022022
30. Daniati E (2019) Decision support systems to determining programme for students using DBSCAN and Naive Bayes: case study: engineering faculty of Universitas Nusantara PGRI Kediri. In: 2019 international conference of artificial intelligence and information technology (ICAIIT). Yogyakarta, Indonesia, pp 238–243. https://doi.org/10.1109/ICAIIT.2019.8834474
31. Akbar R, Nasution SM, Prasasti AL (2020) Implementation of Naive Bayes algorithm on IoT-based smart laundry mobile application system. In: 2020 international conference on information technology systems and innovation (ICITSI). Bandung - Padang, Indonesia, pp 8–13. https://doi.org/10.1109/ICITSI50517.2020.9264938
32. Zia UA, Khan N (2017) Predicting diabetes in medical datasets using machine learning techniques. Int J Sci Eng Res 5(2):257–267
33. Mercaldo F, Nardone V, Santone A (2017) Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Comput Sci 112:2519–2528
34. Iyer A, Jeyalatha S, Sumbaly R (2015) Diagnosis of diabetes using classification mining techniques. Int J Data Min Knowl Managt Process (IJDKP) 5(1):1–14
35. Cheng D, Ting C, Ho C, Ho C (2020) Performance evaluation of explainable machine learning on non-communicable diseases. Solid State Technol 63:2780–2793
36. Athanasiou M, Sfrintzeri K, Zarkogianni K, Thanopoulou A, Nikita K (2020) An explainable XGBoost-based approach towards assessing the risk of cardiovascular disease in patients with type 2 diabetes mellitus. In: 2020 IEEE 20th international conference on bioinformatics and bioengineering (BIBE), Cincinnati, OH, USA, 2020, pp 859–864. https://doi.org/10.1109/BIBE50027.2020.00146

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.