International Journal of Cyber Behavior, Psychology and Learning
Volume 13 • Issue 1

Integrating Machine Learning for Accurate Prediction of Early Diabetes:
A Novel Approach

Kailash Chandra Bandhu, Medi-Caps University, India*
https://orcid.org/0000-0002-4337-4198

Ratnesh Litoriya, Medi-Caps University, India

Aditi Rathore, Medi-Caps University, India

Alefiya Safdari, Medi-Caps University, India

Aditi Watt, Medi-Caps University, India

Swati Vaidya, Medi-Caps University, India

Mubeen Ahmed Khan, Medi-Caps University, India
https://orcid.org/0000-0002-7394-9052

ABSTRACT

In the current world, where diabetes is day by day becoming a very common and fatal disease, it's important that proper measures be taken in order to deal with it. As per the studies, early prediction of diabetes can lead to improved treatment to avoid further complications of the disease, and in order to do so efficiently, machine learning techniques are a great deal. In this study, various factors are taken into consideration, like blood pressure, pregnancy, glucose level, age, insulin, skin thickness, and diabetes pedigree function, which together can be useful to predict whether a person has a risk of developing diabetes or not and help society with the early diagnosis of diabetes. This model is trained using three main classification algorithms, namely support vector, random forest, and decision tree classifiers. The prediction results of each of the classifiers are summarized in this study, and the decision tree gives 78.89% accuracy.

KEYWORDS

Decision Tree Classifier, Diabetes Prediction, Disease, Health, Machine Learning, Random Forest Classifier, Support Vector Machine

1. INTRODUCTION

In many countries like India Diabetes has become a fatal disease. Millions of people lose their lives each year (WHO 2022). According to the stats it is concluded that almost Six hundred and twenty-five million people may be affected by the disease by 2045 and this could be concluded by using the large amount of existing data present with the Hospitals by making the use of techniques such as different Machine Learning Algorithms, Data mining techniques and statistical methods.

DOI: 10.4018/IJCBPL.333157

*Corresponding Author

Advanced analysis is an influential technique that utilises machine learning algorithms, and statistical methods to analyse already present data with us to predict future happenings. In healthcare, advanced analytics can help make critical decisions and forecasts. Machine Learning and Regression techniques can be used in predictive analytics. The aim is to diagnose diseases with the highest possible accuracy, enhance patient's care, increase and improve resources, and enhance clinical outcomes. Machine learning is a crucial component of advanced analytics, as it enables computer systems to learn from old experiences without any need of programming (Baranwal, Bagwe, and M 2020). It can also support automation with minimal errors and this paper largely focuses upon building a model that is effective in predicting the chances that a person may develop diabetes.

The paper's sequential arrangement looks like this: Section 2 discusses the summary of the scientific literature and prior studies, and Section 3 provides a detailed flow of materials and methodology used in this research. Section 4 provides details about the design and implementation parts, followed by Section 5, which reports the results obtained after the successful execution of the model. Section 6 is dedicated to the discussions and implications of the proposed work. The last section presents the conclusions of the proposed research and also sheds some light on future enhancements to this work.

2. LITERATURE REVIEW

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from either insufficient insulin production or inadequate utilization of insulin by the body. Early detection and prediction of diabetes are crucial for effective management and prevention of complications associated with the disease. In recent years, machine learning (ML) has emerged as a powerful tool in the field of healthcare, offering the potential to improve the accuracy of early diabetes prediction (Ahmed et al. 2021; Birjais et al. 2019; Laila et al. 2022; Malviya, Dave, and Kailash Chandra Bandhu 2023; Mustary and Singamsetty 2022; Pandey, Litoriya, and Pandey 2020; Parimala, Kayalvizhi, and Nithiya 2023). This literature review aims to provide an overview of the current state of research on integrating machine learning techniques for accurate prediction of early diabetes, focusing on recent developments and novel approaches in the field.

The motive is to link Machine Learning approaches with medical data to enable the algorithm to increase its efficiency of diabetes diagnosis (Dutta and Bandyopadhyay 2021)(Srivastava, Kumar, and Singh 2020).

Advanced analytics is gaining a strong reputation in the rapidly developing field of big data. There is a huge amount of information available about diseases, their symptoms, and effects on good physical condition. However, this collection of information is not always properly inspected to forecast or investigate a condition. The main motive is to provide an elaborative explanation of predictive approaches, including different types of predictive models, methods for developing them, and their applications in diabetes treatment and other areas of healthcare (Jayanthi, Babu, and Rao 2017).

Tests were conducted on 1,580+ aged siblings of newly born babies with increase in risk of Type 1 diabetes due to genetics. These tests involved analysing 5 HLA DQB1 alleles and 4 autoantibodies that are linked with diabetes. The DQB1 genotypes were then categorised into three groups based on their risk level: those that increased risk, those that somewhat increased risk, and those that lowered risk (Kukko et al. 2003).

Examined the ability of lone and combined diabetes in order to understand which group of population is most at risk of acquiring T1DM. Taking advantage of Sardinia's geographic uniqueness and high prevalence of T1DM, serum samples from 8448 healthy Sardinian school children were taken upon enrolment, and ICA, IA-2A and GADA were subsequently assessed. All SSC participants were tracked for almost 10 years (Velluzzi et al. 2016).

The motive of this research was to look into the consequence of autoantibodies from blood due to the danger of developing type 1 diabetes in the younger population. The study tracked 35,800+

children who were born between 2000 and 2004 and enrolled in the Prediction of diabetes in the Skane project. At birth the samples of blood were collected and examined for HLA genetic constitution, insulinoma associated protein 2, and glutamate decarboxylase 65 (GAD65). Multi component Cox proportional hazard systems are used to evaluate the independent associations with diabetes risk after controlling the affecting factors (Lundgren et al. 2015).

The reports of the World Health Organization shows that this problem of diabetes disease is currently the top global cause of fatality and has a high ranking in Korea. Unhealthy lifestyle choices such as inactivity, poor diet, being overweight, and smoking increase the risk of developing diabetes. The early stages of diabetes may not have noticeable symptoms, and many people may remain undiagnosed for years. However, over time, diabetes can cause complications such as kidney failure, cardiovascular disease, blindness, and limb amputation. Machine Learning approaches are increasingly used for predicting the disease in healthcare, and in this study, Deep neural network with a novel feature dependent on reconstruction error was developed to predict the disease. The RE-based feature was calculated using an intense autoencoder model and was then taken into practice by the DNN classifier, along with existing risk factors, to learn how to predict diabetes. The working of various ML models was evaluated with the help of data from Korea (Amarbayasgalan et al. 2021).

A study evaluated the effectiveness of a risk function in predicting almost the 8 years risk of diabetes in more than 3,200 individuals under the age of 20, using data from the SAHS. The accuracy of the SAHS function in predicting diabetes risk was compared to that of models made solely from the glucose and lipid study of Tehran. The assessment involved calibration, discrimination, and efficiency of fit comparisons (Bozorgmanesh et al. 2010).

Techniques of data mining are widely used in bioinformatics to analyse medical information. Pima Indians Diabetes has been taken into use to develop decision tree-based algorithms for prediction of diabetes disease, using the RapidMiner tool. We performed data pre-processing, which involved selecting and identifying attributes, removal of heretic values, normalising data, using discrete numerical data, visually analysing the data, identifying hidden relationships, and building a diabetes prediction model (Han, Rodriguez, and Beheshti 2009)(Sharma and Litoriya 2012).

In this article, author describe the findings from an examination of the genetic information from siblings who are currently healthy and children having (DMT1) Diabetes Mellitus of Type 1. Using given data, a decision support system is created to categorise and belatedly forecast this sickness in infants who have a genetic predisposition to DMT1. The system may advise adding any person for treatment of pre diabetes (Deja 2009).

Demonstrated that the improved model of prediction of diabetes from the Risk of Atherosclerosis in Study of communication could be applied for a population in the Middle East. In the Lipid of Tehran and Glucose Study, 3,721 participants were C20-year-olds without diabetes at baseline. Testing was done on all models for discrimination and calibration (Bozorgmanesh, Hadaegh, and Azizi 2013).

Diabetes and obesity are linked to more than only heart attacks, strokes, kidney failure, reduced lifespans, and untimely deaths. The three trillion questions of the twenty-first century are represented by them. Furthermore, the overall economic costs of diabetes and obesity in the world surpass USD 3 trillion, and they may increase by 50% or more by 2030. Other disease categories are incomparable. Budgets for healthcare, which are already constrained, may become overburdened globally. This introduction's and the book's goals are not to give harsh advice or make ominous threats. Instead, the objectives are to present and discuss workable, rational, and modern methods for preventing, diagnosing, and managing diabetes and obesity (Faintuch and Faintuch 2020).

Early research discovered a bidirectional relationship between diabetes and covid 19 illness. Here, we will analyse how COVID-19 will affect the prevalence and treatment of diabetes in the future. There is convincing proof that COVID-19 may contribute to newly established hyperglycemia. Hence, it appears that COVID-19 will alter current estimates of the prevalence of diabetes. The clinical management of diabetes is also faced with a host of difficulties. New-onset hyperglycemia is linked to a worse prognosis in COVID-19 patients. The management of comorbid conditions like diabetes mellitus must be very stringent (Hasanzad, Larijani, and Aghaei Meybodi 2022).

When diabetes is present for a long time, microvascular problems increase the risk of falling and are linked to diabetic individuals with greater rates of fracture. Diabetes impairs bone formation in a variety of ways, including insulin insufficiency, hyperglycemia, disruption of the Age axis, loss of factors influencing growth of insulin, and changes to the osteocalcin and Wnt signalling routes. As a result, maintaining bone health while managing diabetes requires sufficient glycemic control (Conti, Wolosinska, and Pugliese 2013).

Obesity and diabetes are pandemic levels over the world and contribute to early mortality. Body mass index of less than 30 kg per meter square indicates obesity, which is defined as an abnormal or excessive buildup of fat. Obesity raises the risk of musculoskeletal illnesses, some types of cancer, respiratory, psychiatric, and metabolic disorders. Obesity has increased significantly throughout the course of the 1980s. This increase in energy dense food consumption may be attributed to changing transportation patterns, rising urbanisation, and an overall decline in daily physical activity (Blüher and Stumvoll 2018).

This study's objective is to evaluate how type 1 diabetes patients respond to the u blood glucose monitoring and treatment system based on the internet. Diabetic patients who had been receiving pump infusion therapy for less than three months were eventually randomised for either use of the Care Link with or without contact initiated by the diabetes team for four months (intervention group; n = 36). The same staff made monthly adjustments to both groups' treatments during the first four months. HbA1c levels and results on the questionnaires measuring diabetes treatment satisfaction and quality of life were the primary end measures. Patients who provided information less than three times per four months were considered noncompliant (Shalitin et al. 2014).

Authors developed and analysed 5 alternative predictive models to categorise the patients into groups with or without diabetes using the R programme. And for this, supervised type of machine learning techniques was used, including the multifactor dimensionality reduction, k-nearest neighbour and others (Kaur and Kumari 2022).

According to a 2018 WHO report, diabetes, one of the chronic diseases that can be fatal, already affects 422 million people worldwide. From its diagnosis to its treatment stages, this is quite expensive for individuals, the government, and groups. Its expense is caused, among other things, by the disease's need for long-term therapy. This disease will continue impacting more people due to its lengthy asymptotic period, which makes it difficult to detect early (Oladimeji, Oladimeji, and Oladimeji 2021).

The African and Caribbean (AfC) communities in the UK have a higher incidence of type 2 diabetes (T2D) than other populations. Structured education is essential for managing T2D, and it emphasises adopting a healthy diet and engaging in physical activity. However, cultural barriers may impede community participation in these programs. To encourage healthy behaviour change, it is important to understand how social norms affect lifestyle behaviour. This study is aimed at using the BCW to create a self-management program for T2D tailored to the needs of UK AfC communities (Moore et al. 2019).

According to the World Health Organization, chronic diseases are expected to cause a significantly higher number of death cases than other causes combined by the end of 2025, and they may account for 60% of the world's disease problems. The chronic diseases, including chronic kidney disease, are more likely to occur as people age, and are therefore regarded as serious conditions that put adults at higher risk. To address this issue, researchers aim to develop a unique feature selection strategy, along with the combination of machine learning algorithm, In prediction of chronic disease at an early stage (Hegde and Mundada 2021).

The field of disease diagnosis is advancing rapidly with the help of machine learning Methods. This research aims on developing an ensemble learning model to predict diabetes at an early stage. The study uses a group of different machine learning models to improve the overall accuracy. The dataset used in the study is the NHANES 2013-14, which includes 54 feature variables and 10,172 samples for the diabetes segment. The feature variables are based on the NHANES recommended set of questions for diabetes diagnosis. The ensemble model uses a voting technique by merging non weighted probabilities of predictions of multiple machine learning models. The use of the algorithm is confirmed using real user input data. The ensemble model enhances the overall performance with an AUC of 75% (Husain and Khan 2018).

To predict the generality of diabetes, a Bayesian network could be used to know the cause-and-effect relationship between risk variables and the interdependencies between direct and indirect hazards. This study is aimed to compare and examine the performances of Bayesian models with type-2 diabetes, including non-hierarchical (BNNH), non-hierarchical and reduced variables, expert judgement based, and hierarchical learning structures. To compare the effectiveness of these classification algorithms, ROC curves, AUC, percentage error, and F1 score were used. The Thai National Health Examination Survey IV dataset was used to evaluate the performance of the models, and the findings indicated that BNHE had the highest AUC values of 0.76 and 0.77 after training and testing datasets, respectively, making it the best for diabetes prevalence (Leerojanaprapa and Sirikasemsuk 2019).

Diabetes is a fatal disease that raises sugar levels in blood and can lead to several complications if not diagnosed and treated properly. The task of going to a clinic and seeing a doctor can be difficult. With the increase of machine learning techniques, this work has become easy. The field of health informatics, which focuses on the technology of the presentation, generation, and application of clinical information in healthcare, has experienced significant growth in recent years (Kaul and Kumar 2020).

A Decision Support system is developed to categorise and forecast illness in kids with genetic DMT1 sensitivity based on genetic information. The system can recommend pre diabetes therapy for an individual. During the development of the system, classification issues were encountered, and rough set theory techniques were utilized to improve the classification accuracy (Deja 2011).

In this study, the effectiveness of 3 Machine Learning classification algorithms: Decision-Tree, SVM, and Naive-Bayes, is evaluated for early identification of diabetes (Nagaraj and Deepalakshmi 2021). Pima Indians Diabetes Database from the UCI Machine Learning repository is used for experimentation, and various metrics that is Recall, Precision, Accuracy, and F-measure are used to assess the working and efficiency of each algorithm. The accuracy is evaluated based on correctly and incorrectly classified examples. The results indicate that Naive Bayes performs better than the other algorithms, with a maximum accuracy of 76.30%. The findings are validated systematically using Receiver Operating Characteristic (ROC) curves (Sisodia and Sisodia 2018).

Accurately predicting diabetes is a crucial aspect of health prognosis. However, the problem of overfitting hinders the accuracy of diabetes prognosis. To overcome this, the study proposes a prediction system for diabetes using a dropout method. A deep neural network is used, which includes entirely connected layers and after that dropout layers. The proposed network outperforms other techniques for the Pima Indians Diabetes Data Set in terms of prediction scores (Ashiquzzaman et al. 2018).

The quantitative measurement of morphological changes in medical imaging is crucial, particularly for analysing the effects of ageing and disease on organ morphology. This study shows a novel method for diabetes prediction with the help of an abdominal shape based deep neural network. The method works directly on raw clouds, removing the need for mesh processing or shape alignment. The network is trained completely, allowing it to learn an optimal representation without relying on manually created shape descriptors. To evaluate the proposed method, we extend the Brain Print form descriptor to Abdomen Print and compare the results (Gutiérrez-Becker et al. 2018).

Blockchain technology can also be utilized as a promising and secure technology in various applications, including healthcare(Bandhu et al. 2022, 2023; Soner, Litoriya, and Pandey 2021, 2022a, 2022b). A Blockchain-enabled diabetes disease detection framework that provides an earlier detection of this disease by using various machine learning classification algorithms and maintains the EHRs of the patients in a secure manner (Chen et al. 2021).

Diabetes disease is harmful and expensive caused by increased blood sugar or low insulin levels. Late detection of diabetes can have negative health effects and is a growing concern for government and health experts. The paper uses experimental data from the University of California website and real data on Indian diabetes. The proposed methodology can provide more intelligent health strategies for predicting disease outcomes in daily life and hospitals, which can prevent disease progression and its complications. Current advancements in information and communication technology (ICT) include machine learning, data mining, and the Internet of Things (Abdollahi and Nouri-Moghaddam 2022).

It is being increasingly used in various industries, from self-driving cars to healthcare. In the medical sector, large amounts of patient data are generated and processed in multiple ways. With machine learning, a prediction system has been developed that can identify multiple diseases simultaneously. Unlike most current systems that can only predict one disease at a time with low accuracy, this system aims to predict diabetes disease providing remarkable efficiency and possibly more in near future. By entering several disease related parameters, the system can output whether the user has the disease or not. This system has the potential to benefit many people by allowing for the monitoring of their condition and taking the necessary actions to lengthen their life (Singh et al. 2022).

Using Machine Learning approaches to discover diabetes mellitus has proven to be a promising approach to predict and detect the disease. The project's motive is to explore and understand the utilisation of various Machine Learning approaches such as 'Decision Trees', 'Logistic Regression', and 'Support Vector Machines' in diabetes prediction. The findings of the project suggest that Machine Learning Algorithms can nicely forecast the occurrence of diabetes depending upon different features such as age, BMI, and glucose levels. The Bayesian model, in particular, has demonstrated high accuracy in detecting diabetes and its associated risk factors.

However, the project also encountered some problems, including the need for a larger and more diverse dataset for better model performance, as well as the challenges of interpreting the models' predictions and features importance. Future research could address these issues by incorporating more data sources and using advanced visualisation techniques.

Thus, the use of machine learning models, including Bayesian models, can improve diabetes diagnosis and early intervention. As technology advances, we can expect more sophisticated algorithms and techniques to enable healthcare providers to make more informed decisions and ultimately improve patient outcomes.

3. PROPOSED METHODOLOGY

We employed a dataset comprising clinical records, including patient demographics, medical history, and laboratory test results. Feature selection involved statistical analysis and domain expertise to identify relevant variables. We utilized Python's scikit-learn library for data preprocessing, feature engineering, and model development. The dataset was split into training (80%) and testing (20%) subsets. We employed a set of supervised machine learning approach, support vector machines, decision tree and random forests, to predict early diabetes onset. Model performance was evaluated using standard metrics, including accuracy, precision, recall, and F1-score, in a 5-fold cross-validation framework to ensure robustness and generalizability.

The proposed methodology constitutes of data preprocessing, splitting of data set, building and training of different machine learning model and their comparison which is depicted by figure 1 and elaborated in subsequent sub sections.

Table 1. Literature summary

S. No.	Title	Work Done	Techniques Used	Findings	Future Scope
1	"Multiple Disease Prediction System (Singh et al. 2022)"	Created a system that predicts more than one disease with high accuracy.	KNN Algorithm, Random forest, XGBoost.	Random forest: 88%, XGBoost: 89%	Scope of improvement in accuracy of prediction in order to decrease the mortality rate
2	"Diabetes Prediction Method Based on Machine Learning (Dutta and Bandyopadhyay 2021)"	They used supervised machine learning algorithms in the training to build a model by using the data of more than 519 diabetic patients aged 16 to 90.	Naive Bias, SVM, LightGBM	Accuracy Naive Bias: 93.27%, SVM- 96.54% LightGBM:88.46%	-
3	"Survey on clinical prediction models for diabetes prediction (Jayanthi et al. 2017)"	Implements advanced models from basic to the level that describes various types of predictive models and steps to develop the same.	OLAP, Predictive analytics, visualisation tools	Accuracy of the model in which the tool WEKA was used on PIMA dataset was 98.9247%, Accuracy of the system based on Hybrid Twin Support vector machine is 98.924%	It has huge scope in Homeland security, prevention of crime, management of infrastructure, Cyber security, Health care etc.
4	"Discovery of Decision Tree Based Diabetes Prediction Model (Han et al. 2009)"	chose the Rapid_ I's Rapid Miner as a tool to discover diabetes prediction model build by decision tree algorithm from a Pima Indians Diabetes Data Set	RapidMiner software	-	Understanding relations between characteristics of patients and likelihood of developing diabetes disease
5	"Study of Bayesian Networks for Diabetes Prediction (Leerojanaprapa and Sirikasemsuk 2019)"	To understand the interdependencies between indirect and direct risks hierarchically	ROC curves, AUC	Maximum AUC values of 0.767 and 0.776	In near future Bayesian Network could be used to forecast the commonness of disease from the cause-and-effect relationship within the factors producing risk.
6	"Artificial Intelligence based Learning Technique for Diabetes Prediction (Kaul and Kumar 2020)"	The main objective is to enhance the result prediction of prediction models	'Genetic Algorithm', 'Decision Tree', 'Random Forest', 'Logistic Regression', 'SVM' and 'Naive Bayes'.	-	In coming years, advanced classifiers like evolutionary algorithms for prediction of diabetes could be applied along with ML Algorithms.
7	"Diabetes and COVID 19 a bitter nightmare (Hasanzad et al. 2022)"	Effects of COVID 19 in the time ahead of diabetes in terms of generality and treatment	-	Estimation shows that the count of people suffering with diabetes to reach more than 570 million by 2030, and more than 700 million by 2045	Effective interventions should be implemented by policy makers.
8	"Diabetes and bone fragility- a dangerous liaison (Conti et al. 2013)"	Puts light on the fact that people suffering from diabetes show an increased danger of osteoporotic fractures	-	There is a hazardous linkage between diabetes and bone fragility with hyperglycemia or anti hyperglycemic agents influencing health of the bone	Research can be done to gain more clarity upon the correlation between diabetes and Bone fracture

continued on following page

Table 1. Continued

S. No.	Title	Work Done	Techniques Used	Findings	Future Scope
9	"Predictive approach modelling and analytics for diabetes using a machine-learning approach (Kaur and Kumari 2022)"	The model using supervised-learning Algorithms in *R* programming for Pima dataset of India to understand patterns for discovery process in diabetes	Linear kernel, Support Vector Machine, Radial Basis Kernel, kNN, and ANN Algorithms	Accuracy: Linear Kernel SVM:89.00% Radial Basis Kernel SVM:84.00% KNN: 88.00% ANN: 86.00%	Some more algorithms could be implemented to get better accuracy and results
10	"Hybrid stacked ensemble combined along with genetic algorithms for diabetes prediction (Abdollahi and Nouri-Moghaddam 2022)"	Identification and prediction of results of diabetes prediction model. Data on Indian diabetics is done in the website of University of California	Internet of Things, Machine learning, and Data mining'.	98.80%, and 99.00% accuracy	intention is to expand the research related to diagnosis of disease such as breast-cancer Metastasis, Lung-cancer, Covid 19 by data mining tools and implemented algorithms

3.1 Algorithm: Diabetes Prediction

3.1.1 Step 1: Data Collection

Collection of type of data containing features that will be helpful in predicting the chances of a person likely to develop diacetate data set used in this study contains features such as Blood Pressure, Pregnancy, Glucose level, Age, Insulin, Skin thickness, Diabetes pedigree function. Total of 768 such records have been taken into account.

3.1.2 Step 2: Exploratory Data Analysis

Investigating the dataset will help you grasp it more clearly and that is done using 2 main approaches i.e., data visualisation and descriptive statistics i.e., mean and standard deviation.

3.1.3 Step 2: Data Preprocessing

This model deals with erratic data which produces more precise and reliable outcomes. This model of diabetes prediction system data is pre-processed by using feature selection and by standardising the data using scikit-learn.

3.1.4 Step 3: Division of Dataset

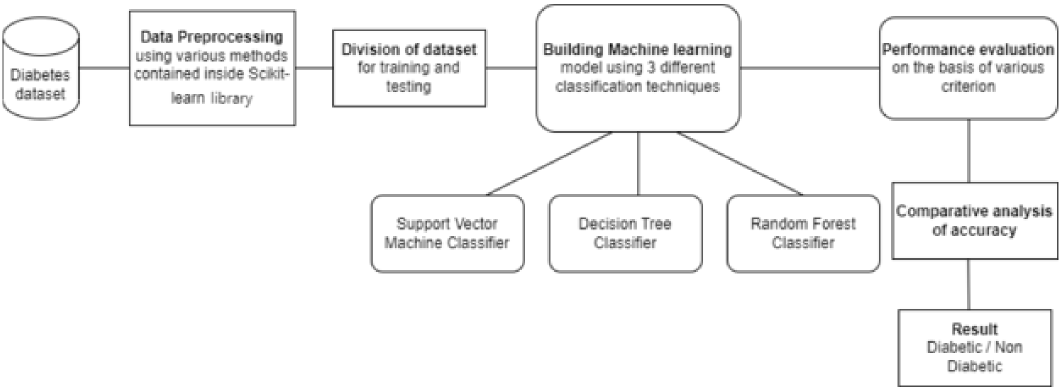
We divide the dataset into 80% of train data and 20% of test data. On training the data we will use Algorithms like Decision Tree, Support vector and Random Forest classifier. After training, we will be using test data to test the accuracy of our model.

3.1.5 Step 4: Model Building Using Support-Vector-Classifier, Decision-Tree, and Random-Forest

1. Training with Support-Vector-Machine Classifier
   a. Loading the Pandas library and the dataset by using Pandas
   b. Defining features and targets.
   c. Have a look at features and targets.
   d. Splitting dataset into training and testing using the sklearn library before building the SVM model.
   e. Importing the SVC function from the Sklearn library SVM module. Create the SVM model using SVC function.

   f. Evaluate the SVM model.
   g. Evaluate the model
2. Training with Decision Tree Classifier
   a. Import pandas' library and a Decision Tree Classifier using Scikit-learn
   b. Loading the Pima Indian Diabetes dataset using panda's read CSV function.
   c. Select the features by dividing the columns in target variable and feature variables).
   d. For understanding the model's performance, splitting the given dataset into the train and test set.
   e. Build a Decision Tree model using Scikit-learn.
   f. Evaluate the efficiency and accuracy of our model against the test set and visualise the decision tree chart.
   g. Optimise the performance of the decision tree by either attribute selection measure or split strategy.
   h. Visualise the decision tree again.
   i. Evaluate the model.
3. Training with Random Forest Classifier
   a. Import Pandas and NumPy into the python environment. Also import seaborn and matplotlib for visualisation.
   b. Load the dataset into pandas' data frame.
   c. Understanding the existence of missing values in any dataset is crucial prior to beginning data analysis and drawing any conclusions. Use the df.info() function to accomplish this; it will provide the names of the columns together with the quantity of non-null values for each column.
   d. Replace the zero values with NaN and after that impute them by their mean value.
   e. A heatmap is used to show the relationship between each column. The output shows that there is stronger correlation when the hues are lighter.
   f. Separate the dataset into features and target variables and then visualise.
   g. Scale the feature variables of our dataset using sklearn's StandardScaler() function. This function standardised the features doing elimination of the mean and doing scaling to unit variance.
   h. Import Random Forest classifier using scikit-learn.
   i. Plot decision boundary for each two possible features.
   j. Evaluate the model.

Figure 1. Proposed methodology


Diabetes dataset
Data Preprocessing using various methods contained inside Scikit-learn library
Division of dataset for training and testing
Building Machine learning model using 3 different classification techniques
Performance evaluation on the basis of various criterion
Support Vector Machine Classifier
Decision Tree Classifier
Random Forest Classifier
Comparative analysis of accuracy
Result Diabetic / Non Diabetic

3.2 Analysis and Accuracy Evaluation

Here, we assess the outcomes utilising various assessment criteria, including classification accuracy, confusion matrix, and f1 score.

1. **Classification Accuracy:** It is expressed as the proportion of right prediction to all input samples.

$$Accuracy = \frac{Number\ of\ Predictions\ that\ are\ Correct}{Total\ number\ of\ Predictions\ that\ are\ Wrong} \tag{1}$$

Following is another way to understand the formulae.

$$Accuracy = \frac{True\ Negative\ values + True\ Positive\ Values}{True\ Positive\ Values + False\ Positive\ Values + True\ Negative\ Values + False\ Negative\ Values} \tag{2}$$

Formal Accuracy formula for the binary Classification case

2. **Confusion Matrix:** A matrix used for determining the complete accomplishment of the model.

Table 2 depicted the general terminology which will help us to determine the metrics we are looking:

- True Positive: Both the actual value as well as value of prediction is Positive.
- True Negative: Both the actual value as well as value of prediction is Negative.
- False Positive: Negative Actual value and positive predictive value.
- False Negative: Positive Actual value and negative predictive value.

3. **Precision:** The proportion of total correctly classified positive classes to the total positive prediction classes. Precision needs to be high (ideally 1).
4. **Recall:** The proportion of total correctly classified positive classes to the total positive classes. Recall needs to be high (ideally 1).
5. **F Measure/F1 Score:** It is a value between 0 & 1 and represents the numerical average mean of Recall as well as Precision. It is kind of maintaining a balance between the recall & precision for the classifier. F- score need to be high (ideally 1).
6. **Dataset Description:** This dataset has 9 attributes and 768 records.

Source of Data: “ https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database ” (UCI Machine Learning 2016).

Name of Dataset: Pima Indians Diabetes Database

The different feature of dataset represented using different unit of measurement which is given below as per the table 3.

Table 2. Confusion Matrix

	Actually, Positive Values (1)	Actually, Negative Values (0)
Predicted Positive Values (1)	True Positive Values	False Positives Values
Predicted Negative Values (1)	False Negative Values	True Negatives Values

Table 3. Table showing attributes and its type contained in dataset

Attribute	Type
Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome	Num

Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
BloodPressure: Diastolic blood pressure (mm Hg)
SkinThickness: Triceps skin fold thickness (mm)
Insulin: 2 Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction: Diabetes pedigree function
Age: Years

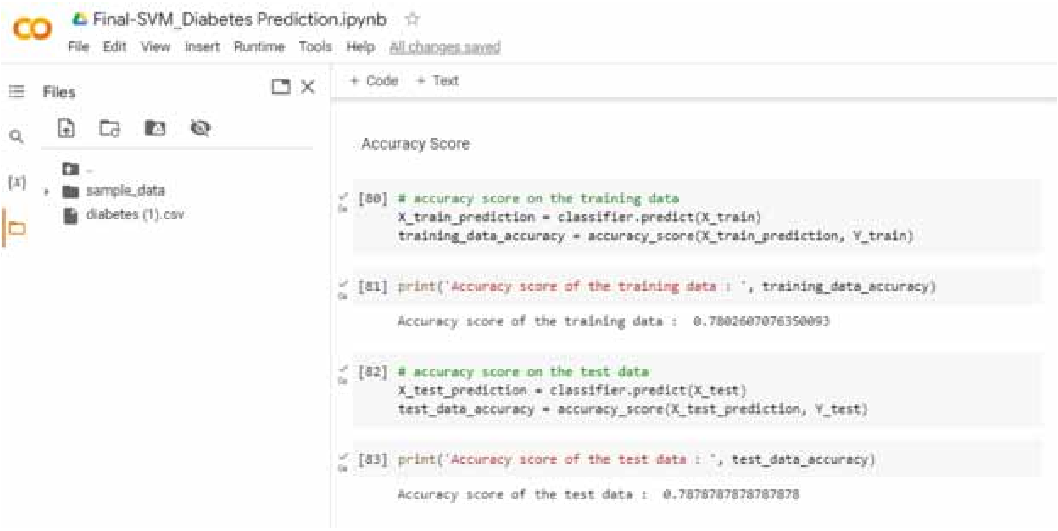
4. DESIGN AND IMPLEMENTATION

This work design and implemented the three machine learning models for diabetes prediction which is represented by Figures 2, 3, 4, 5, 6,7, 8, and 9.

4.1 Support Vector Machine

1. Model Training: After importing SVM from sklearn library model is trained and the trained model is used as the name "classifier".
2. Accuracy: Accuracy of testing and training set is calculated as shown in figure 2.
3. Manual Input and Output: Figure 3 shows the result generated for manual input values by this model. Image contains one example test case and more such test records are shown in table 4.

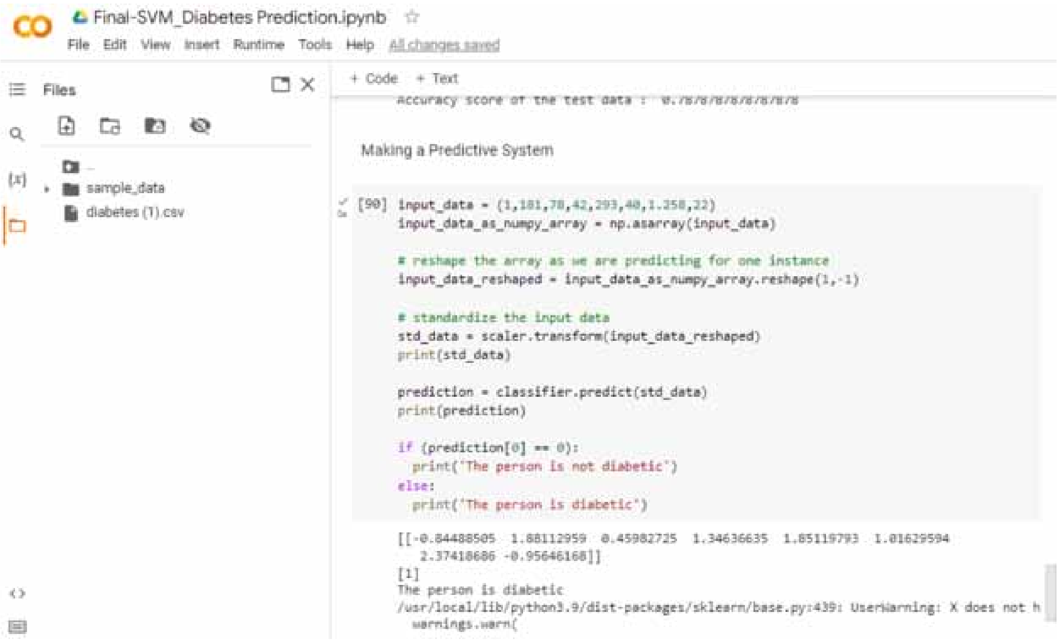
Figure 2. SVM model implementation


Final-SVM_Diabetes Prediction.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Files
sample_data
diabetes (1).csv
+ Code + Text

Accuracy Score

```
[80] # accuracy score on the training data
     X_train_prediction = classifier.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[81] print('Accuracy score of the training data : ', training_data_accuracy)

     Accuracy score of the training data :  0.7802607076350093

[82] # accuracy score on the test data
     X_test_prediction = classifier.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[83] print('Accuracy score of the test data : ', test_data_accuracy)

     Accuracy score of the test data :  0.7878787878787878
```

Figure 3. SVM model manual testing


Final-SVM_Diabetes Prediction.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Files
sample_data
diabetes (1).csv
+ Code + Text

```
Accuracy score of the test data :  0.7878787878787878
```

Making a Predictive System

```
[90] input_data = (1,181,78,42,293,40,1.258,22)
     input_data_as_numpy_array = np.asarray(input_data)

     # reshape the array as we are predicting for one instance
     input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

     # standardize the input data
     std_data = scaler.transform(input_data_reshaped)
     print(std_data)

     prediction = classifier.predict(std_data)
     print(prediction)

     if (prediction[0] == 0):
       print('The person is not diabetic')
     else:
       print('The person is diabetic')
```

```
[[-0.84488505  1.88112959  0.45982725  1.34636635  1.85119793  1.01629594
   2.37418686 -0.95646168]]
[1]
The person is diabetic
/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not h
  warnings.warn(
```

Table 4. Manual testing table for support vector classifier

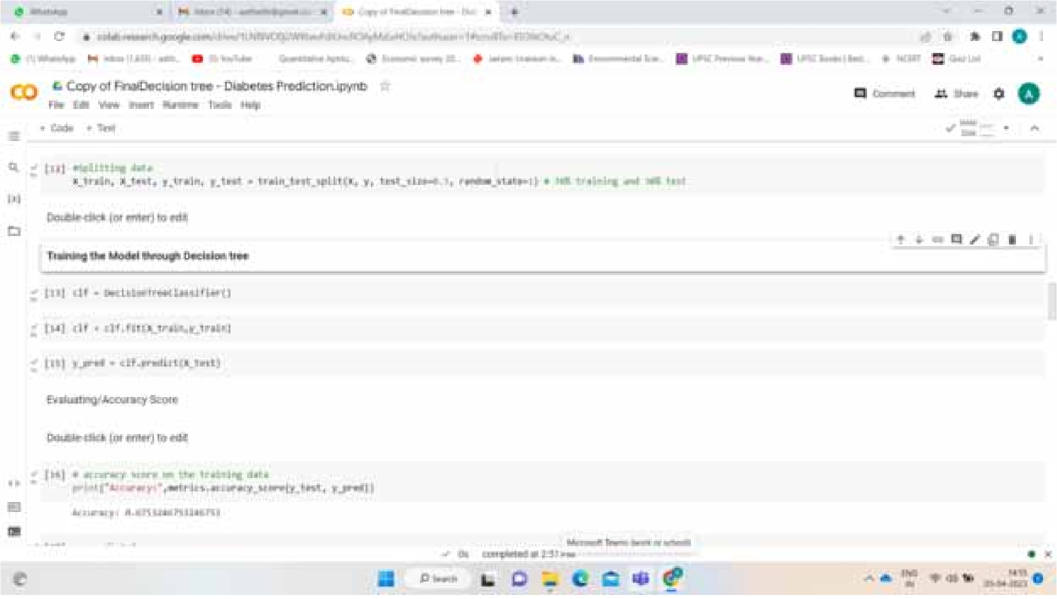
Manual Input Given by User (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome)	Output Generated by Our Model for Given Input	Actual Output of the Given Input	Result
7,100,0,0,0,30,0.484,32	Diabetic	Diabetic	True
1,103,30,38,83,43.3,0.183,33	Not diabetic	Non diabetic	True
2,90,68,42,0,38.2,0.503,27	Not diabetic	Diabetic	False
1,79,75,30,0,32,0.396,22	Not diabetic	Non diabetic	True
3,170,64,37,225,34.5,0.356,30	Diabetic	Diabetic	True
4,141,74,0,0,27.6,0.244,40	Not diabetic	Non diabetic	True
5,114,74,0,0,24.9,0.744,57	Not diabetic	Non diabetic	True
8,197,74,0,0,25.9,1.191,39	Diabetic	Diabetic	True
1,181,78,42,293,40,1.258,22	Diabetic	Diabetic	True

Manual Testing Accuracy is 90% as 9 out of 10 manual input records gives correct values.

4.2 Decision Tree

1. **Training Model:** For training the data we are using DecisionTreeClassifier() with splitting the data as 70% training and 30% testing.
2. **Accuracy:** Accuracy of testing and training set is calculated as shown in Figure 4 and Figure 5.
3. **Manual Input and Output:** Figure 6 shows the result generated for manual input values by this model. Image contains one example test case and more such test records are shown in Table 5.

Figure 4. Decision tree model implementation



Copy of FinalDecision tree - Diabetes Prediction.ipynb

```
[12] #splitting data
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) # 70% training and 30% test
```

Double-click (or enter) to edit

Training the Model through Decision tree

```
[13] clf = DecisionTreeClassifier()
```

```
[14] clf = clf.fit(X_train,y_train)
```

```
[15] y_pred = clf.predict(X_test)
```

Evaluating/Accuracy Score

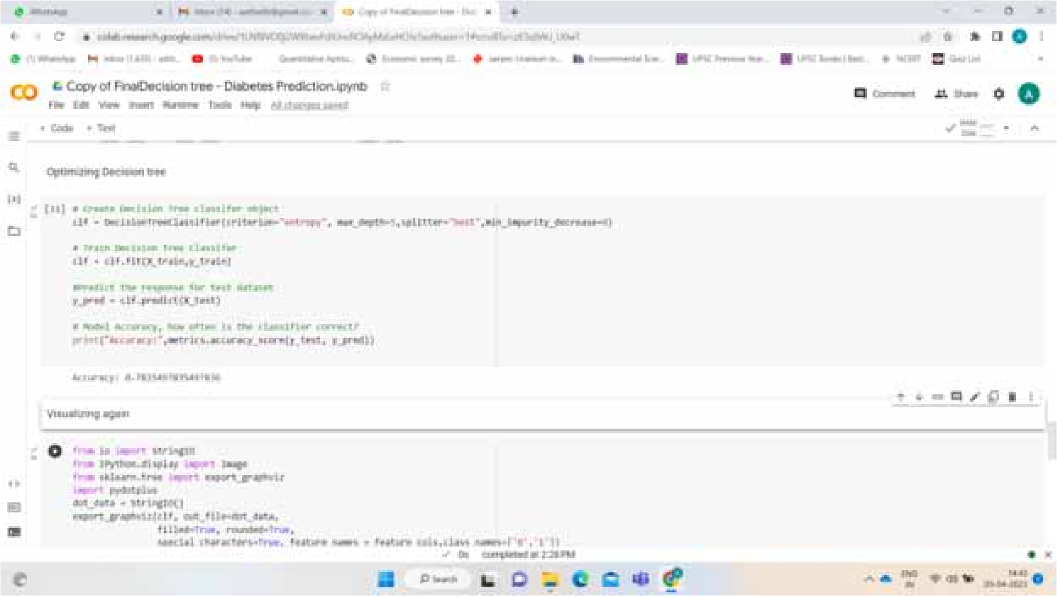
Double-click (or enter) to edit

```
[16] # accuracy score on the training data
     print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

     Accuracy: 0.6753246753246753
```

Microsoft Teams (work or school)

✓ 0s completed at 2:51 PM

Figure 5. Decision Tree model accuracy implementation



Copy of FinalDecision tree - Diabetes Prediction.ipynb

All changes saved

Optimizing Decision tree

```
[31] # Create Decision Tree classifer object
     clf = DecisionTreeClassifier(criterion="entropy", max_depth=3,splitter="best",min_impurity_decrease=0)

     # Train Decision Tree Classifer
     clf = clf.fit(X_train,y_train)

     #Predict the response for test dataset
     y_pred = clf.predict(X_test)

     # Model Accuracy, how often is the classifier correct?
     print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

     Accuracy: 0.7835497835497836
```

Visualizing again

```
from io import StringIO
from IPython.display import Image
from sklearn.tree import export_graphviz
import pydotplus
dot_data = StringIO()
export_graphviz(clf, out_file=dot_data,
                filled=True, rounded=True,
                special_characters=True, feature_names = feature_cols,class_names=['0','1'])
```

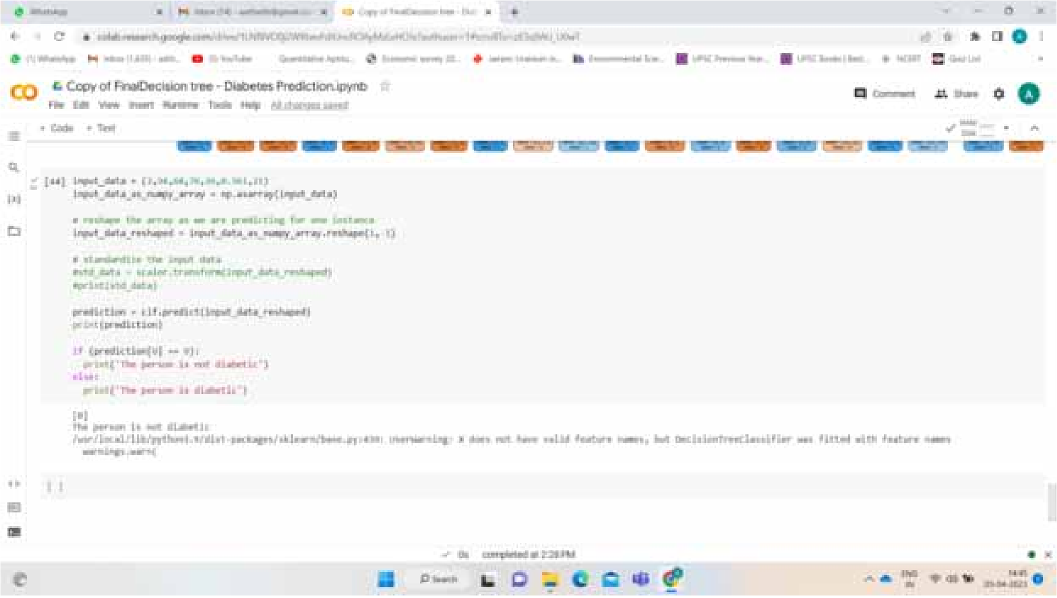
✓ 0s completed at 2:28 PM

4.3 Random Forest

1. **Training Model:** Imported Random Forest classifier from sklearn and the training data labels are fit into the function to train the model shown in Figure 7.
2. **Accuracy:** Figure 8 shows the Accuracy of the model.

Figure 9 and Table 6 shows the manual input and result generated.

Figure 6. Decision tree model manual testing



Copy of FinalDecision tree - Diabetes Prediction.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Comment
Share

```
[44] input_data = (2,94,68,76,26,0.561,21)
     input_data_as_numpy_array = np.asarray(input_data)

     # reshape the array as we are predicting for one instance
     input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

     # standardize the input data
     #std_data = scaler.transform(input_data_reshaped)
     #print(std_data)

     prediction = clf.predict(input_data_reshaped)
     print(prediction)

     if (prediction[0] == 0):
       print('The person is not diabetic')
     else:
       print('The person is diabetic')

     [0]
     The person is not diabetic
     /usr/local/lib/python3.x/dist-packages/sklearn/base.py:430: UserWarning: X does not have valid feature names, but DecisionTreeClassifier was fitted with feature names
       warnings.warn(
```

Table 5. Manual testing table for decision tree algorithm

Manual Input Given by User (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome)	Output Generated by Our Model for Given Input	Actual Output of the Given Input	Result
7,100,0,0,30,0.484,32	Diabetic	Diabetic	True
1,103,30,83,43.3,0.183,33	Non diabetic	Non diabetic	True
2,90,68,0,38.2,0.503,27	Non diabetic	Diabetic	False
1,79,75,0,32,0.396,22	Non diabetic	Non diabetic	True
3,170,64,225,34.5,0.356,30	Non diabetic	Diabetic	False
4,141,74,0,27.6,0.244,40	Non diabetic	Non diabetic	True
5,114,74,0,24.9,0.744,57	Non diabetic	Non diabetic	True
8,197,74,0,25.9,1.191,39	Non diabetic	Diabetic	False
2,94,68,76,26,0.561,21	Non diabetic	Non diabetic	True
1,181,78,42,293,40,1.258,22	Diabetic	Diabetic	True

Manual Testing Accuracy is 70% as 7 out of 10 manual input records gives correct values.

5. RESULTS AND OBSERVATIONS

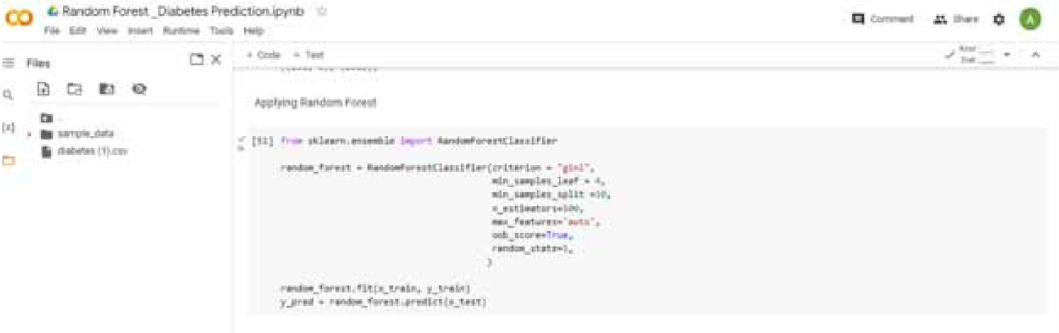
After performing hyperparameter tuning on all the 3 algorithms the following observation is derived:

5.1 Support Vector Machine

Hyperparameters used for support vector machine are as follows:

1. When using SVM, a penalty is assigned for each data point that is classified incorrectly. The value of the penalty, denoted as C, affects how the SVM algorithm determines the decision boundary.

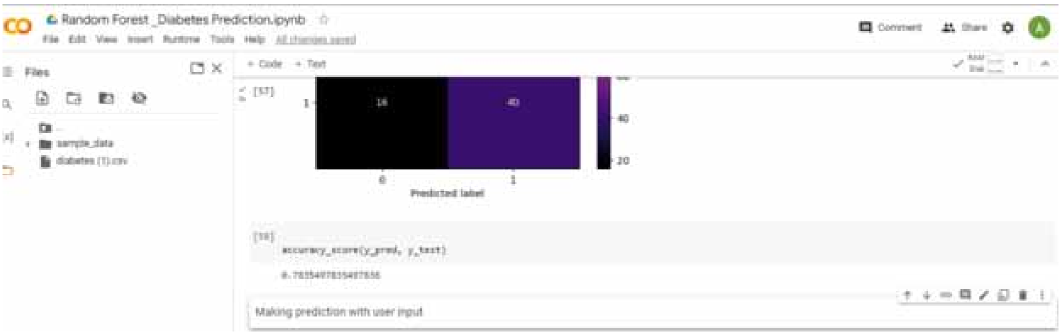
Figure 7. Random forest model implementation


Random Forest _Diabetes Prediction.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
Files
sample_data
diabetes (1).csv
+ Code
+ Text

Applying Random Forest

```
[51] from sklearn.ensemble import RandomForestClassifier

     random_forest = RandomForestClassifier(criterion = "gini",
                                            min_samples_leaf = 4,
                                            min_samples_split =10,
                                            n_estimators=100,
                                            max_features='auto',
                                            oob_score=True,
                                            random_state=1,
                                           )

     random_forest.fit(x_train, y_train)
     y_pred = random_forest.predict(x_test)
```

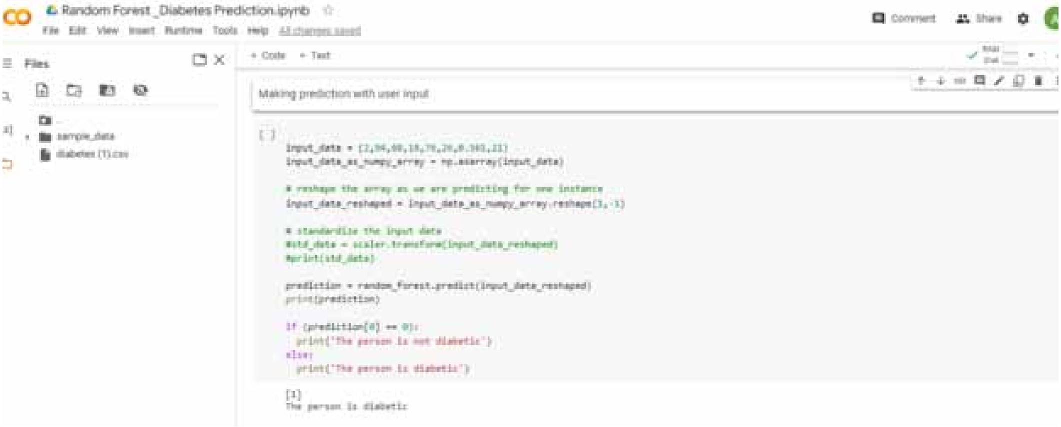
Figure 8. Random forest model accuracy implementation


Random Forest _Diabetes Prediction.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
diabetes (1).csv
+ Code
+ Text
[57]
1
16
40
40
20
0
1
Predicted label

```
[58]
     accuracy_score(y_pred, y_test)

     0.7835497835497836
```

Making prediction with user input

Figure 9. Random forest model manual testing


Random Forest _Diabetes Prediction.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
sample_data
diabetes (1).csv
+ Code
+ Text

Making prediction with user input

```
[ ]
     input_data = (2,94,60,18,70,26,0.381,21)
     input_data_as_numpy_array = np.asarray(input_data)

     # reshape the array as we are predicting for one instance
     input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

     # standardize the input data
     #std_data = scaler.transform(input_data_reshaped)
     #print(std_data)

     prediction = random_forest.predict(input_data_reshaped)
     print(prediction)

     if (prediction[0] == 0):
       print('The person is not diabetic')
     else:
       print('The person is diabetic')

     [1]
     The person is diabetic
```

If C is set to a precise value, the penalty for incorrect data points is also low, which means that the SVM algorithm may choose a decision boundary with a higher margin at the cost of more incorrect classifications. Conversely, if C is set to a high value, the penalty for incorrect data points is higher, which leads the SVM algorithm to try to reduce the number of incorrect classifications

Table 6. Manual testing table for random forest

Manual Input Given by User (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome)	Output Generated by Our Model for Given Input	Actual Output of the Given Input	Result
7,100,0,0,0,30,0.484,32	Diabetic	Diabetic	True
1,103,30,38,83,43.3,0.183,33	Diabetic	Non diabetic	False
2,90,68,42,0,38.2,0.503,27	Diabetic	Diabetic	True
1,79,75,30,0,32,0.396,22	Diabetic	Non diabetic	False
3,170,64,37,225,34.5,0.356,30	Diabetic	Diabetic	True
4,141,74,0,0,27.6,0.244,40	Diabetic	Diabetic	True
5,114,74,0,0,24.9,0.744,57	Diabetic	Non diabetic	False
8,197,74,0,0,25.9,1.191,39	Diabetic	Diabetic	True
2,94,68,18,76,26,0.561,21	Diabetic	Non diabetic	False
1,181,78,42,293,40,1.258,22	Diabetic	Diabetic	True

Manual Testing Accuracy is 60% as 6 out of 10 manual input records gives correct values.

by choosing a decision boundary with a lesser margin. The penalty for each misclassified data point is not the same and depends on the distance from the decision boundary.

2. Kernel: The kernel is a mathematical function used in support vector machines to transform the training dataset into a higher dimensional space where it becomes separable linearly. In Python's implementation of the support vector classifier, the default kernel function is the Radial Basis Function, commonly denoted as "rbf."
3. Gamma: The coefficient kernel is a very important parameter in support-vector-machines that affects the model's performance. It is used in the RBF, polynomial, and sigmoid kernel functions and is defined as the inverse of the support vector influence radius. The value of the can be set to "scale", "auto", or a floating-point number. Table 7, 8, 9 represented the performance and hyper-tuning of SVM model.

Table 7 the Kernel is taken as Variable and rest of the hyperparameters as constant and test train split ratio is 20% 80% respectively.

Table 8 the C is taken as Variable and rest of the hyperparameters as constant and the test train split ratio is 20% 80% respectively. Kernel set to Poly.

Table 9 the C is taken as variable and rest of the hyperparameters as constant and test train split is 30% 70% respectively. Kernel set to Linear.

Table 7. Support vector machine performance with Kernel

Kernel	C	Gamma	Training Set Accuracy	Test Set Accuracy
Linear	0.1	Scale	78.8	77.27
Poly	0.1	Scale	75	70.12
rbf	0.1	Scale	78.6	71.4
Sigmoid	0.1	Scale	76.3	74.6

Table 8. Support vector machine performance with C value

Kernel	C	Gamma	Training Set Accuracy	Test Set Accuracy
Poly	0.1	Scale	75	70
Poly	1.0	Scale	80	70.7
Poly	10.0	Scale	85	66
Poly	100.0	Scale	88.5	64
Poly	1000.0	Scale	89.5	62

Table 9. Support vector machine performance with C value and constant Kernel and Gamma

Kernel	C	Gamma	Training Set Accuracy	Test Set Accuracy
Linear	0.1	Scale	77.6	77.9
Linear	1.0	Scale	78.21	77.4
Linear	10.0	Scale	78.21	77.4
Linear	100.0	Scale	77.8	77.4
Linear	1000.0	Scale	78.02	78.7

5.1.1 Observations

a. In most of the cases linear kernel gives highest accuracy of test set.
b. When the test train split in 30% and 70% respectively our model gives the highest accuracy of the test set that is 78.7 and training set accuracy is 78.02. Here the Kernel is set to linear, Gamma to Scale and value of C parameter is taken as 1000.0.
c. As shown in (Abdollahi and Nouri-Moghaddam 2022) the accuracy of SVM with the same dataset is 65.10 whereas our model gives the highest accuracy of 78.02 of Support vector classifier.

5.2 Decision Tree

Working of a Decision Tree:_We must comprehend the basic splitting parameters that our model employs to create those conditions, such as Gini Index, Entropy, Max Depth, etc. in order to comprehend how it divides our training data and develops into a decision tree.

Gini Score/ Gini Index:_Every machine learning model has a cost function, also known as a loss function, whose goal is to reduce the cost, which is the ambiguous gap between the anticipated value and real value. Probabilities of the anticipated class are employed in classification issues. Decision trees employ the Gini Index as the cost/loss function to decide which feature to use for splitting the data and where to split the column.

Entropy:_Entropy gauges a system's chaos or randomness. We might describe randomness in terms of data as the element of uncertainty in the information we are processing. Entropy increases as unpredictability increases. Making conclusions from that information more difficult.

Max Depth: Max depth of any tree will be shown using this parameter. In the absence of any specific function, the given tree grows till the last leaf node acquires only a single value. Hence, by reducing, we can prevent overfitting by not letting the tree learn all training sample data. Table 10, 11, 12 represented the performance and hyper-tuning of the Decision Tree model.

Table 10. Decision tree performance with gini index and entropy

Max Depth	Splitter	Accuracy When Criterion = Gini	Accuracy When Criterion = Entropy
1	Random	63.32	63.42
2	Random	74.55	75.63
3	Random	71.37	75.72
4	Random	71.24	74.31
5	Random	68.38	77.65
6	Random	74.82	68.23
7	Random	72.29	70.23

Table 11. Decision tree performance with gini index and entropy

Max Depth	Splitter	Accuracy When Criterion = Gini	Accuracy When Criterion = Entropy
1	Best	75.13	76.38
2	Best	75.34	77.12
3	Best	75.47	77.34
4	Best	76.38	78.45
5	Best	76.43	78.79
6	Best	74.22	76.12
7	Best	72.09	77.32

Table 12. Decision tree performance with gini index and entropy

Max Depth	Splitter	Accuracy When Criterion = Gini	Accuracy When Criterion = Entropy
1	Entropy	63.32	76.44
2	Entropy	74.65	77.43
3	Entropy	76.34	77.34
4	Entropy	73.78	78.12
5	Entropy	75.88	78.89
6	Entropy	73.12	77.54
7	Entropy	71.09	77.67

Table 10 the Splitter set to “RANDOM” and accuracy find out for criteria “GINI” & “ENTROPY” with increasing “MAX DEPTH”.

Table 11 the Splitter is set to “BEST” and accuracy find out for criteria “GINI” & “ENTROPY” with increasing “MAX DEPTH”.

Table 12 the Criteria is set to “ENTROPY” and accuracy find out for splitter “RANDOM’ &” BEST” with increasing “MAX DEPTH”.

5.2.1 Observations

1. In most of the cases 'best' splitter gives highest accuracy of test set
2. As shown in article (Abdollahi and Nouri-Moghaddam 2022), the accuracy of Decision Tree with the same dataset is 73.82 whereas our model gives the highest accuracy of 78.89 of Decision tree classifier.

5.3 Random Forest

Hyperparameters used to optimise this algorithm:

- **Entropy:** Entropy is a metric that measures the degree of impurities present in a group of observations. In decision trees, entropy is used to determine how to divide the data. For a dataset with N classes, the entropy measures how much uncertainty there is in the classification of a randomly chosen observation.
- **Gini:** The Gini index, also known as Gini impurity, is another metric used in decision trees to assess the impurity of a group of observations. It calculates the probability that a randomly chosen observation from a group is misclassified based on the distribution of class labels in the group. When a group contains only one class, it is considered pure and has a Gini index of zero.
- **Log Loss:** log loss function is used to measure how well a predicted probability is aligned with the true value of the target variable. The log loss value increases as the predicted probability deviates from the actual value. Specifically, when the predicted probability is close to the true value, the log loss value is low, and when the predicted probability is far from the true value, the log loss value is higher. Therefore, the goal of a binary classification model is to minimise the log loss function to make accurate predictions.

Tables 13, 14, 15, and 16 represented the performance and hyper-tuning of the Random Forest model.

Table 13. Random forest criterion is taken as variable and rest of the hyperparameters as constant

Criterion	Min. Sample Leaf	Sample Split	n_estimator	Accuracy
Entropy	5	10	100	0.76
Gini	5	10	100	0.77
Log loss	5	10	100	0.76

Table 14. Min. sample leaf is taken as variable and rest of the hyperparameters as constant

Criterion	Min. Sample Leaf	Sample Split	n_estimator	Accuracy
Gini	1	10	100	0.77
Gini	2	10	100	0.77
Gini	3	10	100	0.77
Gini	4	10	100	0.7835
Gini	5	10	100	0.77
Gini	6	10	100	0.77

Table 15. Min. sample leaf is taken as variable and rest of the hyperparameters as constant

Criterion	Min. Sample Leaf	Sample Split	n_estimator	Accuracy
Gini	4	1	100	-
Gini	4	2	100	0.76
Gini	4	3	100	0.76
Gini	4	4	100	0.76
Gini	4	5	100	0.76
Gini	4	6	100	0.76
Gini	4	7	100	0.76
Gini	4	8	100	0.77
Gini	4	9	100	0.77
Gini	4	10	100	0.78

Table 16. n_estimator is taken as variable and rest of the hyperparameters as constant

Criterion	Min. Sample Leaf	Sample Split	n_estimator	Accuracy
Gini	4	10	10	0.74
Gini	4	10	20	0.75
Gini	4	10	30	0.75
Gini	4	10	40	0.77
Gini	4	10	50	0.77

5.3.1 Observations

1. 4 min. Sample leaf gives highest accuracy.
2. As shown in (Jayanthi et al. 2017) the Accuracy of Random Forest with the same dataset is 72.00 whereas our model gives the highest accuracy of 78.35 of the Support vector classifiers which is shown in Table 17.

6. DISCUSSIONS

The integration of machine learning techniques for the accurate prediction of early diabetes represents a significant advancement in the field of healthcare. This study aimed to develop a novel approach that harnesses the power of machine learning algorithms to enhance the early detection of diabetes, ultimately leading to improved patient outcomes and healthcare resource allocation. Our findings demonstrate the

Table 17. Comparison of accuracy

Classification Algorithm	Accuracy Obtained
Support Vector Machine Classifier	78.02
Random Forest Classification model	78.35
Decision Tree Classifier	78.89

potential of machine learning models in predicting early diabetes with high accuracy. The incorporation of diverse datasets, including clinical data, genetic markers, and lifestyle information, allowed us to create a comprehensive and holistic predictive model. Our model not only identified individuals at risk of developing diabetes but also provided insights into the contributing factors, enabling personalized prevention strategies. One notable contribution of our approach is its ability to adapt and learn from new data continuously. By implementing a dynamic learning framework, our models can incorporate emerging research findings and evolving patient data, ensuring its relevance and effectiveness over time. This adaptability is crucial in a healthcare context where the understanding of diabetes risk factors is constantly evolving. While our results are promising, there are limitations to consider. The quality and completeness of input data can impact the model's performance. Additionally, the implementation of this approach in real-world clinical settings may require addressing practical challenges, such as data privacy and model interpretability.

In conclusion, our study highlights the potential of machine learning in revolutionizing early diabetes prediction. As we continue to refine and expand this approach, it holds the promise of transforming diabetes care by enabling proactive interventions, reducing healthcare costs, and ultimately improving the quality of life for individuals at risk of diabetes. Further research and collaboration with healthcare practitioners are essential steps in translating these findings into clinical practice.

7. CONCLUSION

In this research, the performance of all 3 machine learning models, namely Support Vector Machine, Random Forest Classifier, and Decision Tree, was analysed on a given dataset. The models were then hyper-tuned to improve their accuracy. The outcomes showed that the Decision Tree model contains the maximum accuracy of 78.89%, while the Random Forest Classifier also performed well, achieving a best accuracy of 78.35% when using a minimum sample leaf. Additionally, the study compared the accuracies of different Machine-learning Algorithms on two datasets and found that the contained model had better accuracy and precision in predicting diabetes compared to existing datasets. Future research could expand on this work to investigate the probability of nondiabetic individuals developing diabetes in the next few years.

8. FUTURE SCOPE

The World Health Organization has projected that by 2030, over 350 million individuals worldwide will suffer from diabetes. Furthermore, with the current susceptibility of people's bodies to the disease in the post COVID era, early detection of diabetes is critical in raising awareness and allowing individuals to take precautionary measures to mitigate the risk of developing serious health complications. In the future, if the dataset is expanded and updated, there may be a possibility of improving the model's accuracy.

ETHICAL DECLARATION

In this study no experiments have been conducted on human respondents. No animal experiments are involved in the study.

CONFLICT OF INTEREST

The authors of this publication declare there are no competing interests.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Abdollahi, J., & Nouri-Moghaddam, B. (2022). Hybrid Stacked Ensemble Combined with Genetic Algorithms for Diabetes Prediction. *Iran Journal of Computer Science*, *5*(3), 205–220. doi:10.1007/s42044-022-00100-1

Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., & Paul, B. K. (2021). Machine Learning Based Diabetes Prediction and Development of Smart Web Application. *International Journal of Cognitive Computing in Engineering*, *2*, 229–241. doi:10.1016/j.ijcce.2021.12.001

Ashiquzzaman, A., & Tushar, A. K. (2018). Reduction of Overfitting in Diabetes Prediction Using Deep Learning Neural Network BT - IT Convergence and Security 2017. Springer Singapore.

Bandhu, Litoriya, Bag, Barwaniwala, & Garg. (2022). Blockchain and Smart Contract Enabled Smart and Secure Electronic Voting System. *Int. J. of Electronic Governance*.

Bandhu, K. C., Litoriya, R., Lowanshi, P., Jindal, M., Chouhan, L., & Jain, S. (2023). Making Drug Supply Chain Secure Traceable and Efficient: A Blockchain and Smart Contract Based Implementation. *Multimedia Tools and Applications*, *82*(15), 23541–23568. doi:10.1007/s11042-022-14238-4 PMID:36467435

Baranwal, Bagwe, & Vanitha. (2020). *Machine Learning in Python*. Academic Press.

Birjais, R., Mourya, A. K., Chauhan, R., & Kaur, H. (2019). Prediction and Diagnosis of Future Diabetes Risk: A Machine Learning Approach. *SN Applied Sciences*, *1*(9), 1112. doi:10.1007/s42452-019-1117-9

Blüher, M., & Stumvoll, M. (2018). *Diabetes and Obesity BT - Diabetes Complications, Comorbidities and Related Disorders*. Springer International Publishing.

Bozorgmanesh, M., Hadaegh, F., & Azizi, F. (2013). Transportability of the Updated Diabetes Prediction Model from Atherosclerosis Risk in Communities Study to a Middle Eastern Adult Population: Community-Based Cohort Study. *Acta Diabetologica*, *50*(2), 175–181. doi:10.1007/s00592-010-0241-1 PMID:21120544

Bozorgmanesh, M., Hadaegh, F., Zabetian, A., & Azizi, F. (2010). San Antonio Heart Study Diabetes Prediction Model Applicable to a Middle Eastern Population? Tehran Glucose and Lipid Study. *International Journal of Public Health*, *55*(4), 315–323. doi:10.1007/s00038-010-0130-y PMID:20217177

Chen, M., Malook, T., Rehman, A. U., Muhammad, Y., Alshehri, M. D., Akbar, A., Bilal, M., & Khan, M. A. (2021). Blockchain-Enabled Healthcare System for Detection of Diabetes. *Journal of Information Security and Applications*, *58*, 102771. doi:10.1016/j.jisa.2021.102771

Conti, F., Wolosinska, D. T., & Pugliese, G. (2013). Diabetes and Bone Fragility: A Dangerous Liaison. *Aging Clinical and Experimental Research*, *25*(1), 39–41. doi:10.1007/s40520-013-0084-z PMID:23907773

Deja, R. (2009). Applying Rough Set Theory to the System of Type 1 Diabetes Prediction. In Advances in Intelligent and Soft Computing (Vol. 64). Springer Berlin Heidelberg. doi:10.1007/978-3-642-05019-0_14

Deja, R. (2011). Accuracy Evaluation of the System of Type 1 Diabetes Prediction. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 6954). Springer Berlin Heidelberg. doi:10.1007/978-3-642-24425-4_42

Dutta, S., & Bandyopadhyay, S. K. (2021). *Diabetes Prediction Using Machine Learning Approaches BT - Advanced Prognostic Predictive Modelling in Healthcare Data Analytics*. Springer Singapore. doi:10.1007/978-981-16-0538-3_9

Faintuch, J., & Faintuch, S. (2020). Introduction to Obesity and Diabetes: The Windows of Opportunity BT - Obesity and Diabetes: Scientific Advances and Best Practice. Cham: Springer International Publishing.

Gutiérrez-Becker, B., Gatidis, S., Gutmann, D., Peters, A., Schlett, C., Bamberg, F., & Wachinger, C. (2018). *Deep Shape Analysis on Abdominal Organs for Diabetes Prediction BT - Shape in Medical Imaging*. Springer International Publishing.

Han, J., Rodriguez, J. C., & Beheshti, M. (2009). Discovering Decision Tree Based Diabetes Prediction Model. In Communications in Computer and Information Science. Springer Berlin Heidelberg. doi:10.1007/978-3-642-10242-4_9

Hasanzad, M., Larijani, B., & Hamid, R. A. M. (2022). Diabetes and COVID-19: A Bitter Nightmare. *Journal of Diabetes and Metabolic Disorders*, *21*(1), 1191–1193. doi:10.1007/s40200-022-00994-5 PMID:35284345

Hegde, S., & Mundada, M. R. (2021). Early Prediction of Chronic Disease Using an Efficient Machine Learning Algorithm through Adaptive Probabilistic Divergence Based Feature Selection Approach. *International Journal of Pervasive Computing and Communications*, *17*(1), 20–36. doi:10.1108/IJPCC-04-2020-0018

Jayanthi, N., Vijaya Babu, B., & Sambasiva Rao, N. (2017). Survey on Clinical Prediction Models for Diabetes Prediction. *Journal of Big Data*, *4*(1), 26. doi:10.1186/s40537-017-0082-7

Kaul, S., & Kumar, Y. (2020). Artificial Intelligence-Based Learning Techniques for Diabetes Prediction: Challenges and Systematic Review. *SN Computer Science*, *1*(6), 322. doi:10.1007/s42979-020-00337-2

Kaur, H., & Kumari, V. (2022). Predictive Modelling and Analytics for Diabetes Using a Machine Learning Approach. *Applied Computing and Informatics*, *18*(1/2), 90–100. doi:10.1016/j.aci.2018.12.004

Kukko, M., Kimpimäki, T., Kupila, A., Korhonen, S., Kulmala, P., Savola, K., Simell, T., Keskinen, P., Ilonen, J., Simell, O., & Knip, M. (2003). Signs of Beta-Cell Autoimmunity and HLA-Defined Diabetes Susceptibility in the Finnish Population: The Sib Cohort from the Type 1 Diabetes Prediction and Prevention Study. *Diabetologia*, *46*(1), 65–70. doi:10.1007/s00125-002-0976-5 PMID:12637984

Laila, Mahboob, Khan, Khan, & Taekeun. (2022). An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study. *Sensors, 22*(14), 5247. .10.3390/s22145247

Leerojanaprapa, K., & Sirikasemsuk, K. (2019). *Comparison of Bayesian Networks for Diabetes Prediction BT - Advances in Computer Communication and Computational Sciences*. Springer Singapore.

Lundgren, M., Lynch, K., Larsson, C., & Larsson, H. E.Diabetes Prediction in Skåne study group. (2015). Cord Blood Insulinoma-Associated Protein 2 Autoantibodies Are Associated with Increased Risk of Type 1 Diabetes in the Population-Based Diabetes Prediction in Skåne Study. *Diabetologia*, *58*(1), 75–78. doi:10.1007/s00125-014-3394-6 PMID:25273346

Malviya, S., Dave, S., Bandhu, R. L. K. C., & Litoriya, R. (2023). A Cryptographic Security Mechanism for Dynamic Groups for Public Cloud Environments. *Journal of Automation, Mobile Robotics and Intelligent Systems*, *16*(2), 46–54. doi:10.14313/JAMRIS/2-2022/15

Moore, A. P., Rivas, C. A., Stanton-Fay, S., Harding, S., & Goff, L. M. (2019). Designing the Healthy Eating and Active Lifestyles for Diabetes (HEAL-D) Self-Management and Support Programme for UK African and Caribbean Communities: A Culturally Tailored, Complex Intervention under-Pinned by Behaviour Change Theory. *BMC Public Health*, *19*(1), 1146. doi:10.1186/s12889-019-7411-z PMID:31429735

Mustary & Singamsetty. (2022). *Prediction and Recommendation System for Diabetes Using Machine Learning Models.* Academic Press.

Nagaraj, P., & Deepalakshmi, P. (2021). Diabetes Prediction Using Enhanced SVM and Deep Neural Network Learning Techniques. *International Journal of Healthcare Information Systems and Informatics*, *16*(4), 1–20. doi:10.4018/IJHISI.20211001.oa25

Oladimeji, Oladimeji, & Oladimeji. (2021). Classification Models for Likelihood Prediction of Diabetes at Early Stage Using Feature Selection. *Applied Computing and Informatics*. .10.1108/ACI-01-2021-0022

Pandey, M., Litoriya, R., & Pandey, P. (2020). Applicability of Machine Learning Methods on Mobile App Effort Estimation: Validation and Performance Evaluation. *International Journal of Software Engineering and Knowledge Engineering*, *30*(1), 23–41. doi:10.1142/S0218194020500023

Parimala, G., Kayalvizhi, R., & Nithiya, S. (2023). Diabetes Prediction Using Machine Learning. In *2023 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE.

Shalitin, S., Ben-Ari, T., Yackobovitch-Gavan, M., Tenenbaum, A., Lebenthal, Y., de Vries, L., & Phillip, M. (2014). Using the Internet-Based Upload Blood Glucose Monitoring and Therapy Management System in Patients with Type 1 Diabetes. *Acta Diabetologica*, *51*(2), 247–256. doi:10.1007/s00592-013-0510-x PMID:23982170

Sharma, N., & Litoriya, R. (2012). Incorporating Data Mining Techniques on Software Cost Estimation : Validation and Improvement. *International Journal of Emerging Technology and Advanced Engineering*, *2*(3), 301–309.

Singh, Shah, Nagpure, & Ashish. (2022). Multiple Disease Prediction System. *International Research Journal of Engineering and Technology*, *9*(3), 5.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes Using Classification Algorithms. *Procedia Computer Science*, *132*, 1578–1585. doi:10.1016/j.procs.2018.05.122

Soner, S., Litoriya, R., & Pandey, P. (2021). Exploring Blockchain and Smart Contract Technology for Reliable and Secure Land Registration and Record Management. *Wireless Personal Communications*, *121*(1), 2495–2509. doi:10.1007/s11277-021-08833-1

Soner, S., Litoriya, R., & Pandey, P. (2022a). Combining Blockchain and Machine Learning in Healthcare and Health Informatics: An Exploratory Study. In Blockchain Applications for Healthcare Informatics. Elsevier.

Soner, S., Litoriya, R., & Pandey, P. (2022b). Integrating Blockchain Technology with IoT and ML to Avoid Road Accidents Caused by Drunk Driving. *Wireless Personal Communications*, *125*(4), 3001–3018. doi:10.1007/s11277-022-09695-x

Srivastava, A. K., Kumar, Y., & Singh, P. K. (2020). A Rule-Based Monitoring System for Accurate Prediction of Diabetes. *International Journal of E-Health and Medical Communications*, *11*(3), 32–53. doi:10.4018/IJEHMC.2020070103

UCI Machine Learning. (2016). *Pima Indians Diabetes Database*. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Velluzzi, F., Secci, G., Sepe, V., Klersy, C., Shattock, M., Foxon, R., Songini, M., Mariotti, S., Locatelli, M., Bottazzo, G. F., & Loviselli, A.The Sardinian Autoimmunity Study Group. (2016). Prediction of Type 1 Diabetes in Sardinian Schoolchildren Using Islet Cell Autoantibodies: 10-Year Follow-up of the Sardinian Schoolchildren Type 1 Diabetes Prediction Study. *Acta Diabetologica*, *53*(1), 73–79. doi:10.1007/s00592-015-0751-y PMID:25896008

WHO. (2022). *Diabetes.* World Health Organization.

Kailash Chandra Bandhu was born in India in 1982. He received the B.E. (Computer Science and Engineering), M.Tech. (Computer Science and Engineering), and Ph.D. (Computer Science and Engineering) degrees from different reputed Universities of India in 2005, 2010, and 2017, respectively. He has been with the Department of Computer Science and Engineering, Medi-Caps University Indore, India, where he is currently a professor. His Research interest includes Machine Learning, Big Data Analysis, Wireless Network and Blockchain Technology Dr. Bandhu supervised various B.Tech. Projects on Machine Learning, Blockchain Technology and Internet of Things. He also supervised M.Tech. Research projects.

Ratnesh Litoriya was born in India in 1983. He received the B.Tech (Information Technology), ME (Computer Engineering), and PhD (Computer Engineering) degrees from different reputed Universities of India in 2004, 2007, and 2015, respectively. He has been with the Department of Computer Science and Engineering, Medi-Caps University Indore, India, where he is currently a professor. His research interests cover software engineering, machine learning, Fuzzy intelligence, elderly care, Blockchain technology, and their application areas. Dr. Litoriya is a Microsoft certified professional in dot net technology and the recipient of International Award for Professor with Huge Potential in Engineering conferred by World Federation of Science & Technology. He has published various research papers in international journals of repute. He has also published an Indian patent for intelligent and adaptive control for micro hydro plant. He has been on the Editorial Board of several International journals.

Aditi Rathore has completed her B.Tech. (Computer Science and Engineering) in 2022 from Medi-Caps University, Indore, India. Her research interests are Artificial Intelligence, Machine Learning.

Alefiya Safdari has completed her B.Tech. (Computer Science and Engineering) in 2022 from Medi-Caps University, Indore, India. Her research interests are Artificial Intelligence, Machine Learning.

Aditi Watt has completed her B.Tech. (Computer Science and Engineering) in 2022 from Medi-Caps University, Indore, India. Her research interests are Artificial Intelligence, Machine Learning.

Swati Vaidya has completed her M. Tech and Bachelor of Engineering from RTMNU Nagpur University, Maharashtra, India and presently she is working as an Assistant Professor in the Department of Computer Science and Engineering at Medi-Caps University, Indore, Madhya Pradesh, India. She is having total 7 years of teaching experience. Mrs. Vaidya has published papers in various international conferences and journals. Her research interests include digital image Processing, Wireless Sensor Networks and Machine Learning.

Mubeen Ahmed Khan is received B.E. and M.Tech. in 2005 and 2012, Mr. Khan is working as Assistant Professor, Department of Computer Science and Engineering, Medi-Caps University, Indore India. His research area includes wireless networks, computer networking, and sensor networks and machine learning.