Diabetes Diagnosis using Electronic Health Records through Machine Learning

Rahul Sharma[1], Gursharan Singh[2], C. Lakshay Singhal[3]

[1,2,3] *Dr. Akhilesh Das Gupta Institute of Professional Studies*

***Abstract*—Electronic Health Records (EHRs) have become a critical part of modern healthcare systems. They offer numerous benefits such as improved patient care, better clinical decision- making, and increased efficiency. However, the vast amounts of data generated by EHRs can be overwhelming for healthcare professionals, making it difficult to extract meaningful insights. This is where machine learning (ML) techniques can be applied to help make sense of the data.**

The applications of these EHRs could easily be interpreted/found in numerous cases and could also be seen as a life-changing tool while trying to diagnose the specifics of a disease. Through the course of this project, EHRs of patients and their vitals are taken as the factors/bases for diagnosing Diabetics that affect the way your body regulates blood sugar or glucose. Metabolic disorder that occurs in diabetic patients is considered as an indirect cause of many diseases may it be Cardiovascular diseases, Kidney diseases, Eye problems, Neuropathy, Infections or Alzheimer's.

This project aims to develop an ML-based system for analysing EHR data to improve patient outcomes. The system will be trained using supervised and deep learning techniques using ML algorithms like KNN, SVM and RF to perform tasks such as predicting patient outcomes, identifying potential health risks, and thus finding out the important insights from the data. The system will be developed using

Python programming language and various ML libraries such as Pandas, NumPy and Scikit- learn.

The project will also involve data pre- processing, data cleaning, and data normalization to ensure the data is suitable for ML algorithms. The system will be evaluated using real-world EHR data and compared with existing methods used by healthcare professionals. The project's success will be measured based on the system's accuracy, efficiency, and ability to improve patient outcomes.

Overall, this project aims to demonstrate the potential of ML in healthcare and how it can be used to improve patient care and outcomes. The results of this project could have significant implications for healthcare providers and patients.

.

***Index Terms*—Healthcare, Electronic Health Records (EHRs), Machine Learning, Deep Learning, K- nearest neighbour(KNN), Support Vector Machine(SVM) Random Forest(RF), Disease Prediction, Diagnosis, Diabetic, Prediction.**

I. INTRODUCTION

Diabetes is a prevalent chronic disease affecting millions of people worldwide. Early and accurate diagnosis of diabetes is crucial for effective management and prevention of complications. Machine learning techniques can play a significant role in developing automated systems that aid in diabetic diagnosis.

The goal of this machine learning project is to develop a model that can accurately classify individuals as either diabetic or non-diabetic based on a set of relevant features. By leveraging a dataset consisting of clinical and demographic information, the model will learn patterns and relationships to make predictions about an individual's diabetic status. The dataset used in this project may include features such as age, gender, body mass index (BMI), blood pressure, glucose levels, insulin levels, and other relevant medical measurements. These features will serve as inputs to the machine learning algorithms, and the corresponding target variable (diabetic or non-diabetic) will be the output to be predicted.

The project will involve several stages, including data preprocessing, feature selection or engineering, model training, and evaluation. Data preprocessing steps may include handling missing values, normalizing or scaling features, and dealing with any outliers or inconsistencies in the dataset. Feature selection or engineering techniques may be employed to identify the most relevant features that contribute to the classification task.

Different machine learning algorithms, such as random forests, support vector machines and K-nearest neighbour, will be explored and evaluated for their performance in diabetic diagnosis. Model training will involve training the algorithms on a portion of the dataset and optimizing their parameters to achieve the best possible performance. The trained models will then be evaluated using appropriate metrics such as accuracy, specificity and sensitivity among others.

The ultimate aim of this machine learning project is to develop an accurate and reliable model that can assist healthcare professionals in making informed decisions about diabetic diagnosis. Such a model has the potential to augment the diagnostic process, provide early detection, and contribute to more effective management and prevention strategies for diabetes.

It is important to note that while machine learning can be a powerful tool in diabetic diagnosis, the model's predictions should always be considered in conjunction with clinical expertise and further medical evaluations. Machine learning models should not replace healthcare professionals but rather serve as a supportive tool in the diagnostic process.

II. LITERATURE REVIEW

Diabetes is a collective term for a group of symptoms related to hyperglycaemia. It is a chronic metabolic disorder in which patients may have problems with insufficient insulin secretion, insulin resistance, or both. The main clinical symptoms of diabetes are polyuria, thirst, hunger, fatigue, blurred vision, weight loss, and difficulty in wound healing. The American Diabetes Association classifies diabetes into two types.[1]4 Type I diabetes, formerly known as insulin-dependent diabetes mellitus (IDDM), often occurs in childhood, mainly because islet cells are damaged by immune responses. It may be due to the patient’s heredity, living environment, or a viral infection that triggers an autoimmune response that damages the beta cells in the pancreas so that the patient’s body cannot produce enough insulin. Type II diabetes, formerly known as noninsulin-dependent diabetes mellitus (NIDDM), is commonly seen in adults and generally occurs when people are about 40 years old. Patients with NIDDM usually have insufficient insulin secretion and insulin resistance concurrently. The cause of Type II diabetes is multifactorial. It is generally considered to be related to heredity, obesity, and lack of exercise. Other special types of diabetes include diabetes caused by genetic defects, diabetes caused by pancreatic exocrine destruction, and diabetes caused by drugs or chemicals.

Applying machine learning and data mining methods in DM research is a key approach to utilizing large volumes of available diabetes- related data for extracting knowledge. The severe social impact of the specific disease renders DM one of the main priorities in medical science research, which inevitably generates huge amounts of data. Undoubtedly, therefore, machine learning and data mining approaches in DM are of great concern when it comes to diagnosis, management and other related clinical administration aspects. Hence, in the framework of this study, efforts were made to review the current literature on machine learning and data mining approaches in diabetes research.

The algorithms used in data mining, machine learning, or any field of artificial intelligence perform predictive modelling, that is, the use of data and statistics to predict future outcomes based on historical data. The most common symptoms of diabetes include abnormal metabolism, hyperglycaemia, and associated risk for specific complications affecting the eyes, kidneys, and nervous system, which are major parts of the body. Such symptoms are used to gather data, and then the modelling is performed based on age and gender categories. One such algorithm is ordering points to identify cluster structure (OPTICS), which is a set of ordering points to identify clustering structures. OPTICS is an advanced version of density-based spatial clustering of applications (DBSCAN) with noise, and it eliminates all negative aspects of DBSCAN. The data clustering method used in this algorithm is a balanced iterative reducing and clustering algorithm using hierarchies (BIRCH) that selects the most suitable data for further analysis. Thus, the naïve Bayes (NB) data mining technique is used, and BIRCH and OPTICS are used for clustering similar types of data and used for the identification of the correct algorithm for better accuracy. Apache Spark is among the fastest-growing platforms for health analysis. It operates more rapidly than the Hadoop platform, making it more easily usable and applicable to clinical practice.

The research about using machine learning techniques to identify diabetes can be traced back to 1988 when J. W. Smith and his cooperators published a paper about using a so- called 'ADAP' algorithm to identify diabetes [4]. They use the Indian Pima data set of diabetes onset of women as their training and testing data, and the accuracy of their algorithm is about 76%. Kayaer's team used the GRNN technique to identify diabetes. They discussed how to build the network and had a similar result as Gail A. Carpenter and his group made, which used a very complicated ARTMAP-IC network [6]. The technique Kayaer used was much simplified compared to Gail's, but it was still a complex one in regard to the scale of the data set. From all those researches we can see that they all explored diabetes identification through one particular method, and modified and improved it to its best or approximate best.

Electronic health records were used for many studies related to diabetes. A method that enables risk assessment from electronic health records (EHR) on a large population, they also added administrative claims and pharmacy records. Another study proposed a model that predicts the severity as a ratio interpreted as the impact of diabetes on different organs of the human body, the algorithm estimated the severity on different parts of the body like the heart and kidney. A rapid model for glucose identification and prediction based on the idea of model migration. Despite the data being collected from different sources and places, some attributes were used in many studies such as age, gender, body mass index (BMI), glucose, blood pressure, time of diagnosis, and smoking.

III. RESEARCH METHODOLOGY

The research methodology for this project will follow a quantitative approach, using statistical analysis and ML techniques to analyse EHR data. The data will be collected from real-world sources, and the results will be interpreted and reported objectively. These stages include:

1. Literature review: This stage will involve a comprehensive review of existing literature/work on EHRs, ML, and their applications in healthcare. The aim of this stage is to gain a better understanding of the current state of research and identify any gaps that the project can address.
2. Data collection and pre-processing: This stage will involve collecting real-world EHR data and preparing it for analysis. This will include data cleaning, data normalization, and feature selection to ensure that the data is suitable for ML algorithms.
3. ML algorithm selection and development: This stage will involve selecting Machine learning techniques such as supervised learning, unsupervised learning, and deep learning, and developing the ML models. The models will be trained on the pre- processed EHR data to perform tasks such as predicting patient outcomes, identifying potential health risks, and recommending personalized treatment plans, for which a number of algorithms may it be KNN, RF or SVM are to be tested/examined to find the best fit on the basis of accuracy and expected results following the train and test dataset.

Model evaluation: This stage will involve evaluating the effectiveness of the ML models using appropriate metrics such as accuracy, specificity and sensitivity score. The models will be compared with existing methods used by healthcare professionals to determine their effectiveness in improving patient outcomes.

Interpretation and reporting of results: This stage will involve interpreting the results of the ML models and reporting them in a clear and concise manner. The results will be presented using appropriate visualizations such as graphs and charts, and the implications of the results will be discussed.

IV. DATA COLLECTION AND PRE-PROCESSING

Data collection and preprocessing are crucial steps in developing a machine-learning model for diabetic diagnosis. Here's an overview of the process:

1. Data Collection:

Identify relevant data sources: This can include electronic health records, clinical databases, research studies, or publicly available datasets related to diabetes.

Ensure data compliance: Adhere to data privacy and security regulations, obtain necessary permissions, and ensure compliance with ethical guidelines.

Gather diverse data: Collect a variety of data types, such as demographic information, medical history, laboratory test results, and other relevant clinical measurements.

2. Data Preprocessing:

Data preprocessing is an iterative process that may involve multiple iterations to refine the dataset for optimal model performance. It is important to document and maintain a clear record of the preprocessing steps performed to ensure reproducibility and transparency

Handling missing data: Identify and handle missing values in the dataset. This can involve techniques such as imputation (replacing missing values with estimated values based on existing data) or excluding incomplete records if the missing data is extensive.

Dealing with outliers: Detect and handle outliers, which are data points that significantly deviate from the norm. Outliers can be treated by either removing them if they are due to measurement errors or applying appropriate transformations to mitigate their impact.

Feature selection/engineering: Identify the most relevant features for diabetic diagnosis. This can involve statistical analysis, domain expertise, or feature selection algorithms to choose the subset of features that contribute the most to the predictive task. Additionally, feature engineering techniques may be applied to create new informative features from existing ones.

Data normalization/scaling: Normalize or scale numerical features to a common scale, ensuring that they have similar ranges and distributions. Common techniques include min-max scaling or standardization (mean = 0, standard deviation = 1).

Handling categorical variables: Convert categorical variables into numerical representations suitable for machine learning algorithms. This can involve one-hot encoding, label encoding, or ordinal encoding, depending on the nature of the categorical variables and their relationship to the target variable.

Splitting the dataset: Divide the dataset into training, validation, and test sets. The training set is used for model training, the validation set helps tune hyperparameters and assess model performance, and the test set is reserved for the final evaluation of the trained model.

3. Addressing class imbalance (if applicable):

In the case of imbalanced classes (e.g., significantly more non-diabetic cases than diabetic cases), techniques like oversampling the minority class (diabetic cases) or under sampling the majority class (non-diabetic cases) can be applied to balance the class distribution. Alternatively, algorithms like SMOTE (Synthetic Minority Over-sampling Technique) can generate synthetic samples to balance the classes.

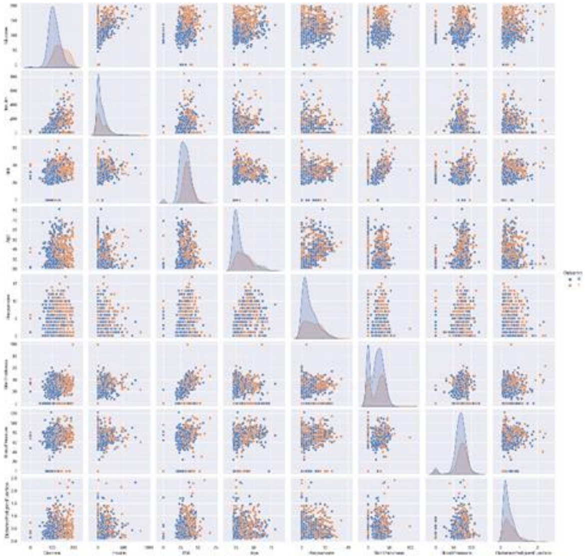
4. Data validation and quality checks:

Verify the accuracy, consistency, and integrity of the data.

Check for potential data entry errors, anomalies, or inconsistencies and rectify them as appropriate.

Ensure the dataset is representative, unbiased, and suitable for the model's intended purpose.

5. Also, finding correlations between features in a dataset is important because it provides insights into the relationships and dependencies among variables. This knowledge helps in feature selection, identifying redundant or highly influential features, improving model performance, and gaining a deeper understanding of the data and underlying patterns.



V. DATASET DEFINITION

A dataset for diabetic diagnosis typically consists of various features or variables related to an individual's demographic information, medical history, and clinical measurements. Here's a general description of the types of data that may be included:

1. Demographic Information:

Age: The age of the individual.

Gender: The gender of the individual (male or female).

2. Medical History:

Family History: Information about any family history of diabetes or related conditions.

Personal Medical History: Details of past medical conditions, including diabetes, hypertension, cardiovascular disease, etc.

Medication: Any medications currently being taken by the individual.

3. Clinical Measurements:

Body Mass Index (BMI): A measure of body fat based on height and weight.

Blood Pressure: Measurements of systolic and diastolic blood pressure.

Glucose Levels: Measurements of fasting or random blood glucose levels.

Insulin Levels: Measurements of insulin levels in the blood.

Cholesterol Levels: Measurements of total cholesterol, LDL (low-density lipoprotein), HDL (high-density lipoprotein), and triglyceride levels.

HbA1c: Measurement of glycated haemoglobin, which indicates average blood glucose levels over a few months.

4. Target Variable:

Diabetic Diagnosis: A binary variable indicating the presence (1) or absence (0) of diabetes.

It's important to note that the specific dataset may vary depending on the source or context in which it was collected. The dataset may also include additional features or measurements relevant to diabetic diagnosis. Furthermore, privacy and ethical considerations should be taken into account when handling and using medical data.

VI. MACHINE LEARNING ALGORITHM SELECTION

When selecting and developing machine learning algorithms for diabetic diagnosis, it's important to consider factors such as the nature of the data, the interpretability of the model, the performance metrics, and the computational requirements. Here's an overview of the process:

A. Selecting Machine Learning Algorithms:

Random Forests (RF): Random forests combine multiple decision trees, providing improved accuracy and robustness. They can handle high-dimensional datasets and are less prone to overfitting.

Support Vector Machines (SVM): SVMs aim to find the best hyperplane that separates the two classes. They work well with high- dimensional data and have effective mechanisms for handling nonlinear relationships.

K Nearest Neighbour(KNN): The K-Nearest Neighbours (KNN) algorithm is a non- parametric classification algorithm that predicts the class of a data point by finding the k closest labelled data points in the training dataset and assigning the majority class label among those neighbours to the new data point.

B. Model Development:

Split the dataset into training, validation, and test sets.

Train the selected algorithm(s) on the training set and tune hyperparameters using cross- validation and grid search techniques.

Evaluate the model's performance on the validation set and iteratively refine the model.

VII. SIMULATION

Model evaluation in machine learning refers to the process of assessing the performance and effectiveness of a trained model. It involves various techniques and metrics to measure how well the model generalizes to unseen data and performs on the task it was trained for. Here are some key aspects of model evaluation:

Accuracy: The proportion of correctly classified instances.

Precision: The proportion of true positive predictions out of the total predicted positives. It measures the model's ability to avoid false positives.

Sensitivity, also known as true positive rate or recall, is a performance metric measuring the proportion of true positive predictions out of all positive instances in a binary classification problem, indicating how well a model identifies positive cases.

Specificity: The specificity of an ML model measures the proportion of correctly classified negative instances. It is calculated as the ratio of true negatives to the sum of true negatives and false positives. Higher specificity indicates better identification of negative cases.

Cross-validation is a technique used to assess the model's performance by repeatedly splitting the data into different train-test splits. It helps to estimate how well the model will perform on unseen data and mitigate the impact of the specific train-test split.

Over-fitting occurs when a model performs well on the training data but fails to generalize to new data.Under-fitting, on the other hand, occurs when a model is too simplistic and fails to capture the underlying patterns. Model evaluation helps identify and address these issues.

VIII. RESULTS

In this section, we show the performance of machine learning classification techniques for diabetes classification. For this, we analyse various popular classification techniques that include k-nearest neighbour (KNN), support vector machine (SVM) and random forest (RF). Moreover, three classifiers have been also adapted and compared their performance based on accuracy, sensitivity and specificity. The next section represents the related work.

A. Comparison of Performance of the models based on the Accuracy of Algorithms

Figure 1 demonstrates the comparison of the performance of 3 machine learning algorithms based on accuracy. KNN has the most accurate result and after that two algorithms, random forest and support vector machine, with an accuracy of 72.44 and 75.02 respectively.

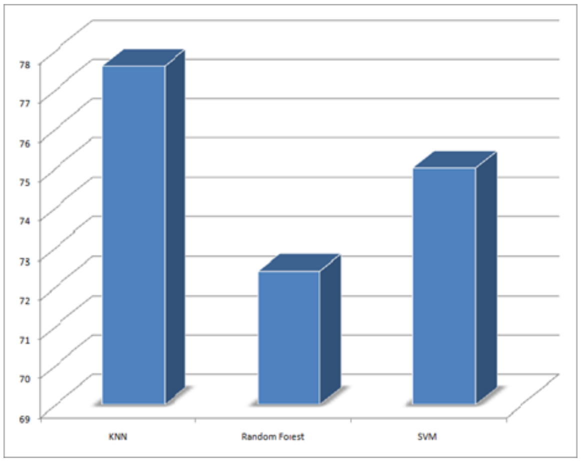


Figure 1 Comparison of Performance of the models based on the Accuracy of Algorithms

When making predictions using KNN, the algorithm identifies the k data points in the training set that are closest (most similar) to the new data point being classified or predicted. The class or value assigned to the new data point is determined by a majority vote or average of the classes or values of its k nearest neighbours.

The choice of the k value affects the performance of the KNN algorithm. Smaller values of k can result in more flexible and potentially noisy predictions, while larger values of k can provide more stable but potentially biased predictions. Selecting the optimal k value is often determined through experimentation and cross-validation techniques to find the best balance between bias and variance in the model. In the above experimentation, the value of k comes around 13.08-13.34 when different percentages of the dataset are proportionated into train and test datasets. (as in Figure 2)

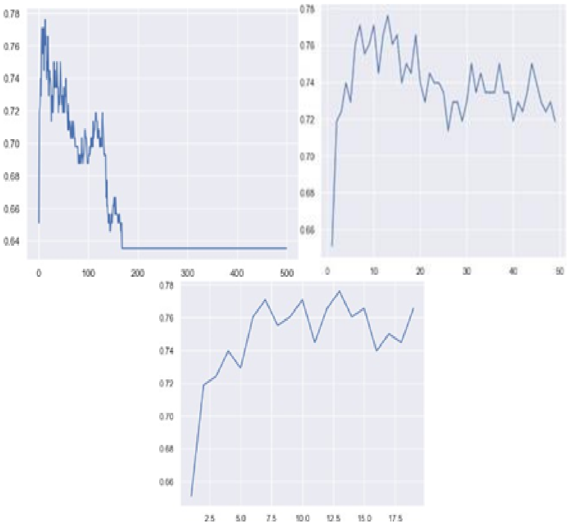


Figure 2

A. Comparison of Performance of the models based on the Specificity of Algorithms

A comparison between learning algorithms based on specificity is depicted in Figure 3. KNN has the most specificity of 86.06.

Following this, RF and SVM have 84.08 and 74.31 as their respective specificity.

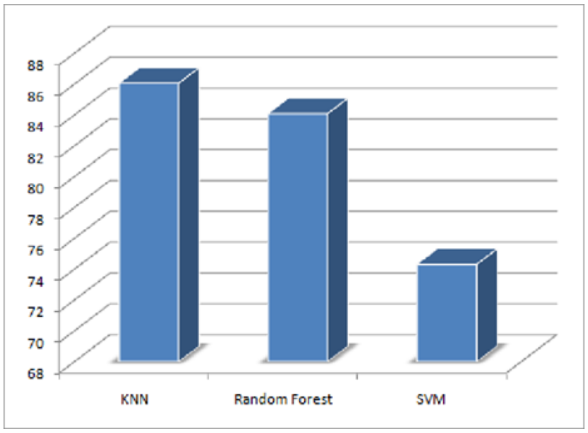


Figure 3 Comparison of Performance of the models

based on the Specificity of Algorithms

B. Comparison of Performance of the models based on the Sensitivity of Algorithms

Figure 4 demonstrates the comparison of the performance of 3 machine learning algorithms based on their sensitivity. SVM has the most sensitivity of 78.21. After that two algorithms, random forest and KNN show a sensitivity of 60.80 and 62.85 respectively.

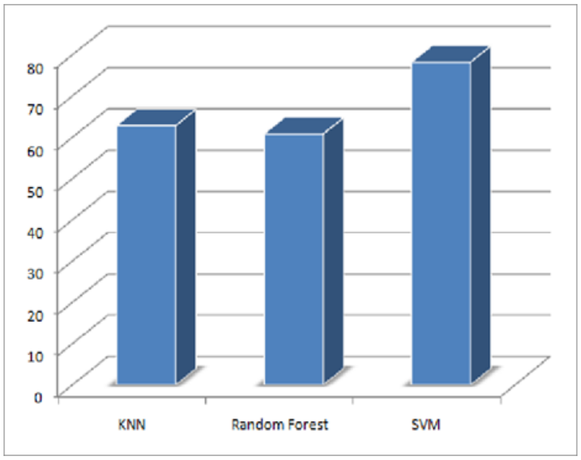


Figure 4 Comparison of Performance of the models based on the Sensitivity of Algorithms

IX. CONCLUSION

In conclusion, this research paper explored the application of machine learning techniques for diabetes prediction. We analysed several popular classification algorithms, including random forest (RF), support vector machine (SVM), and K nearest neighbour (KNN).

Through the use of feature selection, we identified high-risk factors associated with diabetes, enabling us to focus on the most relevant features for classification. We then compared the performance of three classifiers based on various evaluation metrics, including accuracy, sensitivity, and specificity.

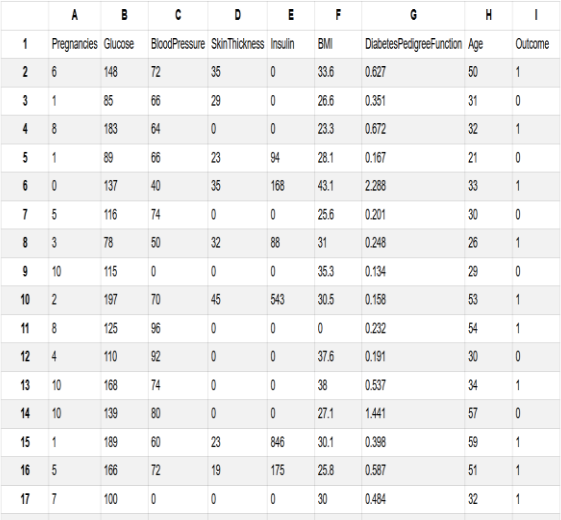
The results demonstrated the potential of machine learning in accurately predicting diabetes. The model which achieved notable performance, throughout the three classifiers namely accuracy, sensitivity and specificity are best suited for diabetic diagnosis. And from the above observations, it is evident that the model which consistently outperforms the other model in terms of having the highest accuracy, highest sensitivity and second highest specificity is K-nearest neighbour (KNN).

This research contributes to the field by showcasing the effectiveness of machine learning techniques for diabetes prediction. The findings suggest that such approaches can assist healthcare professionals in identifying individuals at risk of developing diabetes, enabling early intervention and tailored preventive measures.

Overall, this study highlights the potential of machine learning as a valuable tool for diabetes prediction, supporting efforts to improve early diagnosis, personalized treatment, and ultimately, the management of this prevalent and chronic condition.

X. DATASET USED

https://www.dropbox.com/scl/fi/0uiujtei423te1q4kvrny/diabetes.csv?rlkey=20xvytca6xbio4vsowi2hdj8e&e= 1&dl=0



	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1

REFERENCE

[1] Predicting the Onset of Diabetes with Machine Learning Methods by Chun-Yang Chou 1, *ORCID, Ding-Yang Hsu 2ORCID and Chun-Hung Chou 3

[2] Machine Learning and Data Mining Methods in Diabetes Research by Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda.

[3] A comprehensive review of machine learning techniques on diabetes detection By Toshita Sharma and Manan Shah

[4] A comprehensive exploration to the machine

learning techniques for diabetes identification By Wei, Sidong; Zhao, Xuejiao; Miao, Chunyan

[5] A Comparative Analysis of Machine Learning Algorithms to Build a Predictive Model for Detecting Diabetes Complications by Ali A. Abaker

[6] "Machine Learning Approaches for Diabetes Prediction from Healthcare Data: A Review" by S. Kavakiotis et al.

[7] "Predicting the Onset of Diabetes Mellitus Using Machine Learning" by G. N. Deepak and S. Sumathi (2017)

[8] "Diagnosis of Type 2 Diabetes Mellitus Using Machine Learning Algorithms" by H. S. Alshawi et al. (2018)

[9] "DeepDiabetes: Deep Learning-Based Automated Diagnosis of Diabetes Using EHR Data" by Z. Yu et al.

[10] "Predicting Type 2 Diabetes Mellitus Using Machine Learning Techniques" by A. Al-Dubaee (2019)