IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

www.ijitce.com

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

Sugar Sense: Machine and Deep Learning for predicting Diabetes risks

T. Anil Karuna Kumar[1], P. Hima Kumari [2], A. Devi Manasa[3], D. Mounika [4], P. Supriya[5]

[1] Associate Professor, Dept. of Computer Science & Engineering, Vijaya Institute of Technology for Women, Enikepadu, Vijayawada-521108

[2,3,4,5] Students, Dept. of Computer Science & Engineering, Vijaya Institute of Technology for Women, Enikepadu, Vijayawada-521108

Email id: anilkarunakumar@gmail.com[1] , phimakumari799@gmail.com [2], atmurimanasa@gmail.com[3], dmounika5160@gmail.com[4] supriyapilli5134@gmail.com[5]

Abstract:

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading causes of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning and deep learning techniques. The algorithms like K-Nearest Neighbors (KNN), Decision tree, multi-layer perceptron (MLP) are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes

Keywords: Diabetes risks, Sugar Sense, Machine and Deep Learning

1.Introduction

Machine Learning is a system of computer algorithms that can learn from example through self- improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to predict an output which can be used to make actionable insights. The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers. A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation. Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

2.Literature Review

Yashoda and M. Kannan **[1]** the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient's data set is established by gathering data from hospital warehouse which contains two hundred and forty-nine instances with seven attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. N. Niyati Gupta, A. Rawal, and V. Narasimhan **[2]** This aims to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyze the results of several classification methods in WEKA, the study compares the performance of same classifiers when implemented on some other tools which includes RapidMiner and Matla busing the same parameters (i.e. accuracy, sensitivity and specificity). Salian and G. Hari Sekaran [3] Big Data analytics have been applied to evaluate the risk of readmission for diabetes patients. Predictive modeling has been employed by applying decision tree classification method. It has been observed that chance of readmission in diabetic patient is successfully predicted using the above analysis. Many analysis methods can be explored to improve the accuracy of the existing system P. Lee [4] In recent years, focus on applying a decision tree algorithm named as CART on the diabetes dataset after applying the resample filter over the data. The author emphasis on the class imbalance problem and the need to handle this problem before applying any algorithm to achieve better accuracy rates. K. Sharmila and S. Manickam [5] This paper discusses about to analyze the data in predicting the diabetes from medical record of the patients. The study states that approximately 40 million Indians suffer from diabetes till now. his. This study is analyzing the diabetes from huge medical records by using decision trees with statistical implication using R tool.

EXISTING SYSTEM

The existing system for predicting diabetes using machine learning involves leveraging various algorithms and data sets to develop predictive models. Researchers and data scientists gather datasets containing information such as medical history, lifestyle factors, and biomarkers from individuals with and without diabetes. These datasets are then used to train machine learning algorithms, such as logistic regression, decision trees, support vector machines, or neural networks. Through a process called model training, the algorithms learn to identify patterns and relationships within the data that are indicative of diabetes risk. Once trained, the models can be used to predict the likelihood of an individual developing diabetes based on their input data. These predictive models hold promise for early detection, risk stratification, and personalized interventions to prevent or manage diabetes effectively. Ongoing research aims to enhance the accuracy and scalability of these models while addressing challenges such as data quality, interpretability, and implementation in better accuracy.

PROPOSED SYSTEM

The proposed system for predicting diabetes integrates a range of machine learning algorithms and deep learning methods including K-Nearest Neighbors (KNN) and Decision Trees, alongside deep learning methods such as Neural Networks. By leveraging these diverse techniques, the system aims to enhance accuracy and reliability in diabetes prediction. Additionally, the integration of Standard Scalar normalization ensures optimal preprocessing of the data, contributing to improved model performance. This multifaceted approach allows for a comprehensive analysis of various features and patterns within the data, ultimately leading to more precise predictions of diabetes risk. Through the fusion of traditional machine learning and cutting-edge deep learning methods, the proposed system endeavors to advance the field of diabetes prediction, offering potentially groundbreaking insights for early detection and effective management strategies.

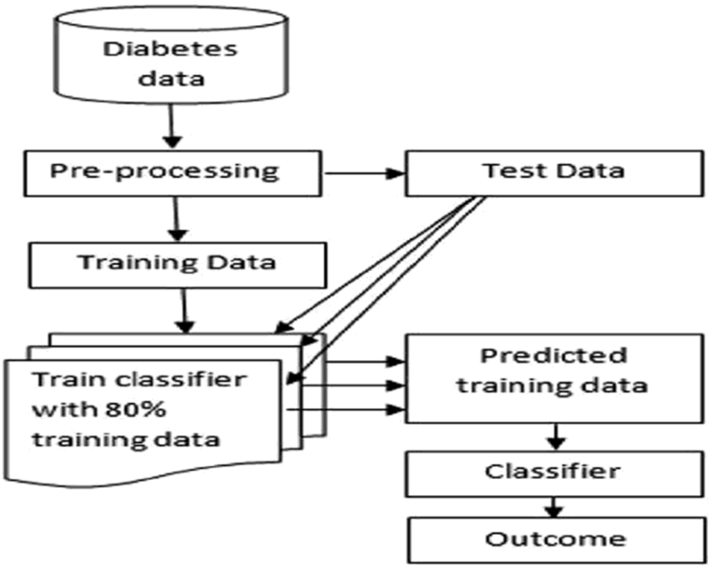


Fig: System Architecture

INPUT DESIGN

The input design is the link between the information system and the user. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. Input Design considered the following things:

Input - Expanding the Feature Set:

- Demographic Data: Include zip code or neighborhood information to account for socioeconomic factors that may influence diabetes risk.
- Medical History: Consider incorporating details like history of hypertension, history of cardiovascular disease, and history of mental health conditions (e.g., depression) that can linked to diabetes development.
- Lifestyle Factors: Deepen the diet data by including details on specific food groups (e.g., high intake of processed foods, sugary drinks), sleep patterns, and stress levels.
- Biomarkers: Explore incorporating additional biomarkers like triglycerides, C-reactive protein (CRP) which can indicate inflammation, and liver function tests.
- Genetic Data: If ethically obtained and anonymized, genetic information can provide various insights into an individual's predisposition to diabetes.

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user.

Classification (Enhanced):

- Multi-class Classification: Instead of just high-risk and low-risk, categorize individuals into groups like normal, prediabetic, and high-risk for diabetes.
- Time-based Risk: Indicate the timeframe associated with the risk prediction (e.g., high risk of developing diabetes within 2 years).

Regression (Enhanced):

- Confidence Intervals: Along with the risk score, provide a confidence interval to indicate the range of possible risk based on model uncertainty.
- Actionable Insights: Translate the risk score into actionable recommendations for the user. For example, a score above 0.5 might suggest consulting a doctor for further evaluation and personalized risk management strategies.

IMPLEMENTATION

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, thebetter our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc. Link:https://www.kaggle.com/datasets/saurabh00007/diabetes.csv.

Data Ingestion:

Data ingestion plays a curial role in acquiring, integrating, and preparing diverse datasets essential for predicting diabetes. Through a meticulous data collection process, information is retrieved from these sources using appropriate methods and protocols. Throughout the data ingestion phase, strict quality assurance measures are applied to validate data accuracy, completeness, and reliability. Finally, the ingested data is securely stored in scalable data repositories, ensuring accessibility and compliance with regulatory requirements.

Data preprocessing and Manipulation:

In data preprocessing and manipulation for the Sugar Sense project, raw data from diverse sources such as electronic health records, continuous glucose monitoring devices, and mobile applications undergo rigorous cleaning, normalization, and feature engineering processes. This ensures data quality, consistency, and relevance for subsequent analysis and modeling, ultimately enabling accurate predictions and personalized insights into diabetes management within a robust computational framework.

Overall, data preprocessing and manipulation are iterative processes that require careful consideration of data quality, consistency, and suitability for analysis. By effectively cleaning, transforming, and preparing the data, analysts and data scientists can derive meaningful insights and build reliable models for decision- making and predictive analytics.

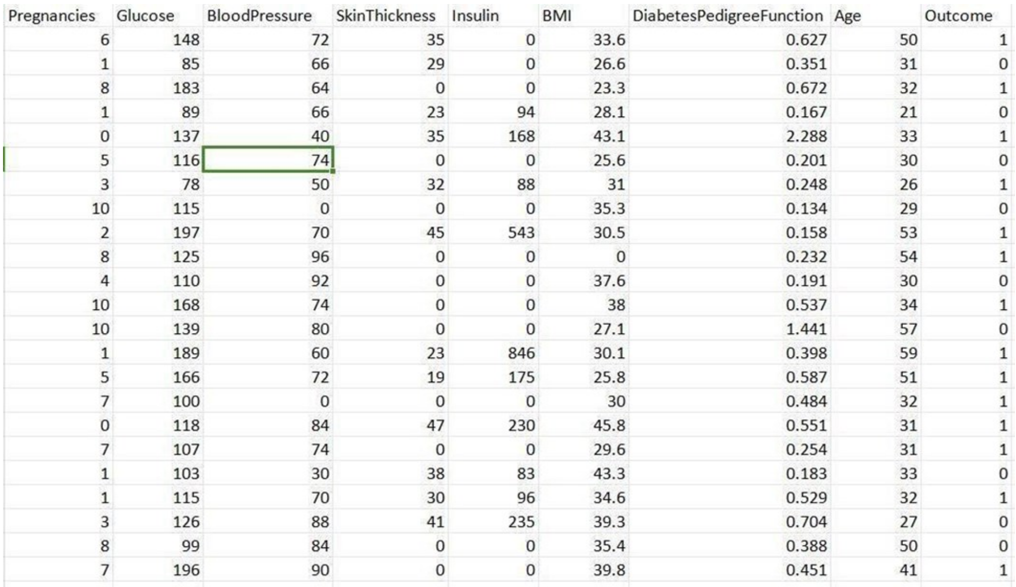
Data Analysis and Visualization:

Data analysis is the process of organizing, cleaning, and interpreting data sets to extract insights, while data visualization is the process of transforming findings into visual representations. Data visualization can help develop trends and conclusions by organizing information into charts, graphs, and other visual representations. Some common techniques used for data visualization include bar chart, graph, or other visual format that helps inform analysis and interpretation.

Here various algorithms and methods are utilized for analysis in the Sugar Sense project. These include both traditional statistical methods and machine learning techniques. Here are some commonly used algorithms and methods for analysis:

Statistical Methods:

- Descriptive Statistics: Mean, median, standard deviation, percentiles, and other descriptive statistics are used to summarize the distribution of blood glucose levels and other variables.
- Correlation Analysis: Pearson correlation, Spearman correlation, or Kendall's tau correlation coefficients are calculated to assess the relationship between blood glucose levels and other factors. Regression Analysis: Linear regression, logistic regression, or other regression models are employed to predict blood glucose levels based on predictor variables.



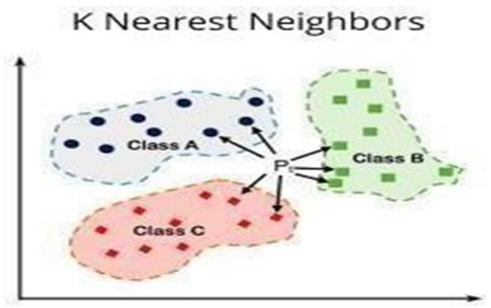
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1

Figure: Sample data set for diabetes prediction

ALGORITHMS

K-NEAREST NEIGHBOUR(KNN) ALGORITHM

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category byusing K- NN algorithm.K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.



Steps involved in KNN

Step-1: Select the number K of the neighbors.

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

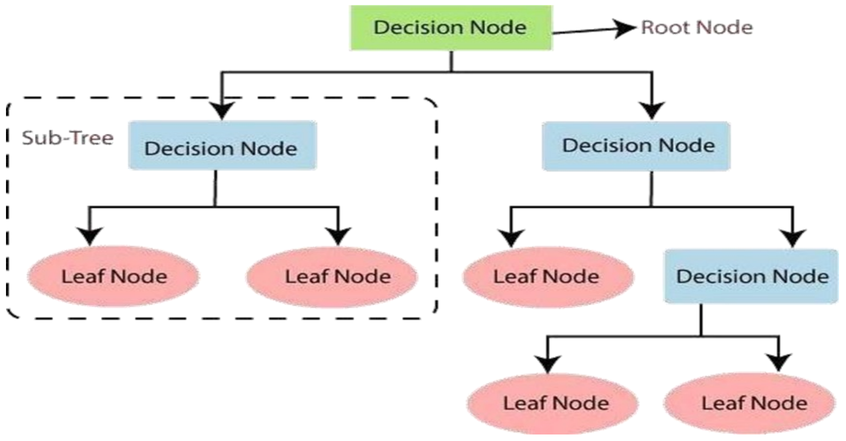
Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

DECISION TREE ALGORITHM

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.



Steps involved in Decision Tree:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

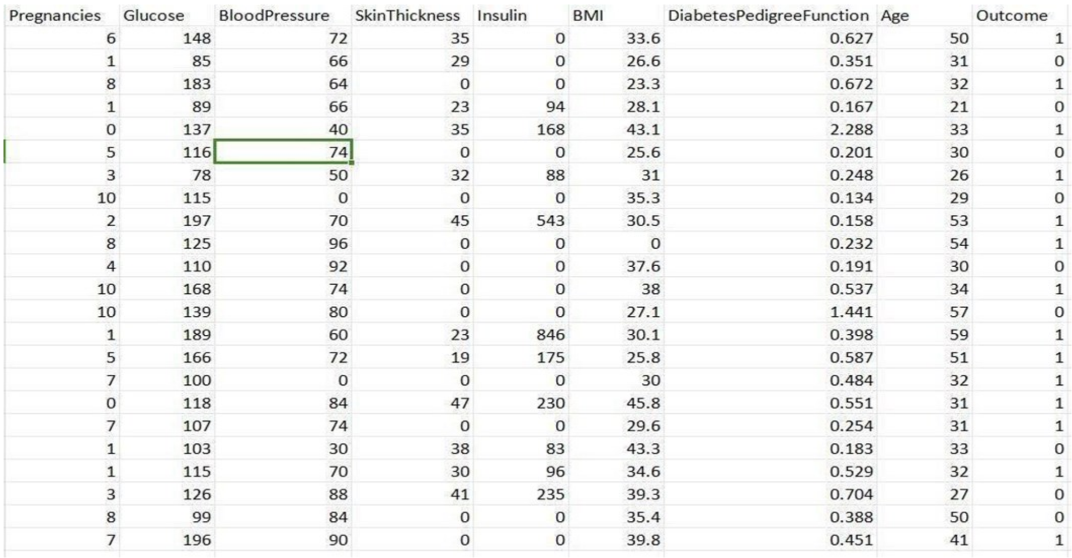
Step-2: Find the best attribute in the dataset using Attribute Selectio Measure.

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

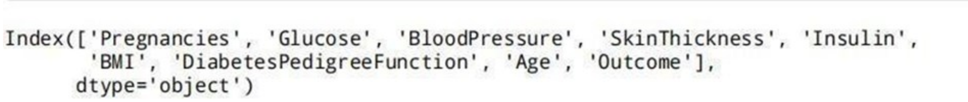
Step-5: Recursively make new decision trees using the subsets of the datasetcreated instep -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

This is the dataset which we downloaded from Kaggle



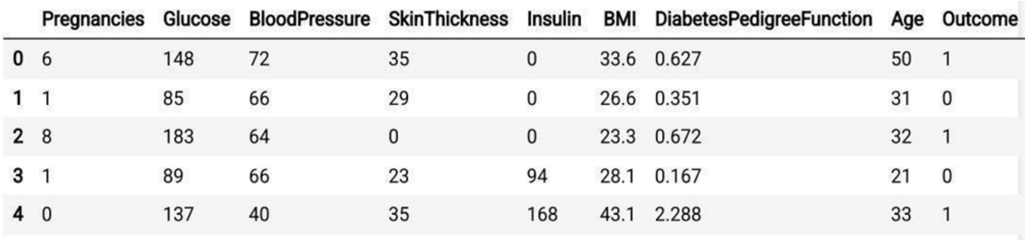
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1

These are the 9 columns in our dataset, here outcome is the dependent variable and remaining all are independent variables.



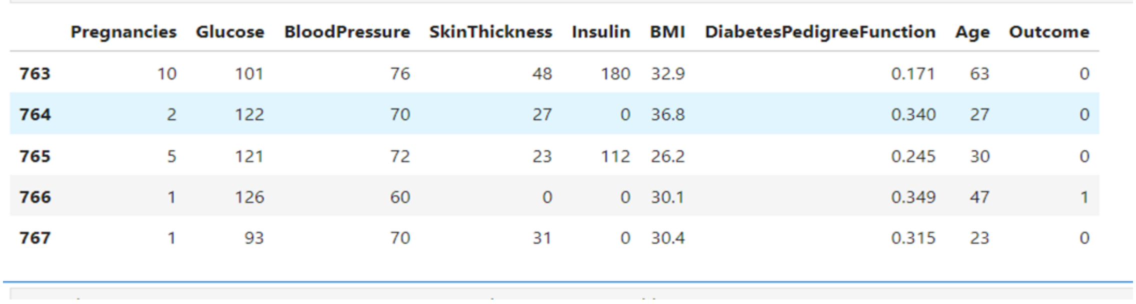
```
Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
      dtype='object')
```

By using head function First 5 rows were displayed.



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

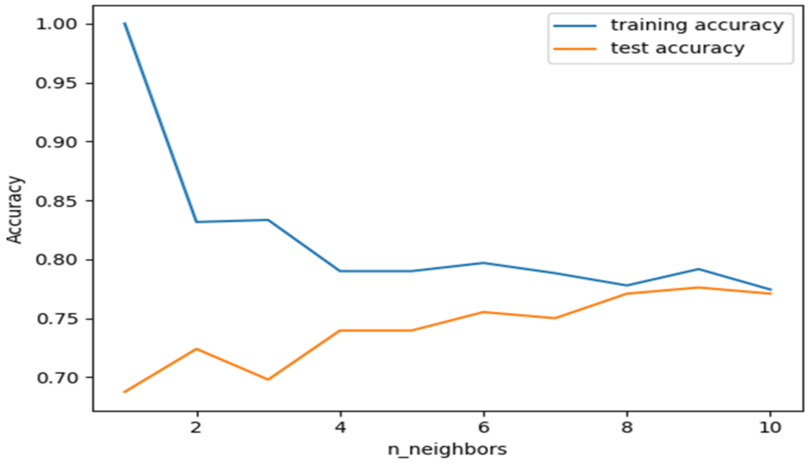
By using tale function bottom 5 rows were displayed

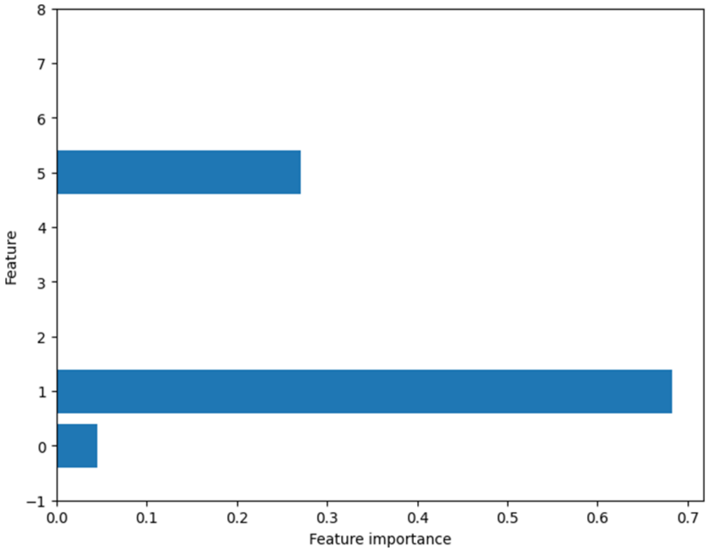


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

By using shape function, we displayed number of rows and number of columns were displayed in our dataset.

This is performing k-nearest neighbors classification on a diabetes dataset. It first splits the dataset into training and testing sets, ensuring that the proportion of outcomes in each set remains similar. Then, it iterates through different numbers of neighbors (1 to 10) to train the KNeighbors Classifier model and evaluates its performance on both the training and testing sets. Finally, it plots the training and testing accuracies against the number of neighbors to visualize the model's performance. The aim is to find the optimal number of neighbors that maximizes accuracy without overfitting or underfitting





This segment first imports and trains a Multi-Layer Perceptron (MLP) classifier from sklearn. neural_ network module, initializing it with default parameters and evaluating its accuracy on both training and test sets. Then, it standardizes the input features using Standard Scaler to ensure they have mean zero and unit variance. The scaled data is then used to train another MLP classifier, and its performance is evaluated. Subsequently, the code initializes another MLP classifier with a higher maximum number of iterations (max_iter=1000) and regularization strength (alpha=1). It's trained on the scaled data, and its accuracy on both training and test sets is assessed.

FUTURE WORK

Integration of Wearable Technology:

- Investigate the incorporation of wearable devices like continuous glucose monitors (CGMs) or fitness trackers to gather real-time physiological data. This could enhance the accuracy of predictions by providing a continuous stream of relevant information.

Longitudinal Data Analysis:

- Extend the analysis beyond cross-sectional data to longitudinal studies. By tracking individuals over time, it's possible to identify trends and patterns in diabetes risk development, leading to more personalized risk assessments.

Feature Engineering and Selection:

- Explore advanced feature engineering techniques to extract more meaningful features from the data. Additionally, employ feature selection algorithms to identify the most relevant predictors of diabetes risk, reducing dimensionality and potentially improving model performance.

Conclusion

The SUGARSENSE model, employing machine learning and deep learning techniques, shows great promise in predicting diabetes risks. Our study demonstrates the efficacy of various algorithms, including decision trees, support vector machines, and neural networks, in accurately assessing the likelihood of diabetes onset. Moreover, the integration of deep learning methodologies, particularly deep neural networks, enhances predictive performance by extracting intricate patterns from raw data, thereby improving the robustness of risk prediction. The interpretability of the developed models provides valuable insights into the complex interplay of risk factors influencing diabetes susceptibility. Healthcare professionals can utilize these insights to make informed decisions regarding patient care and intervention strategies. The actionable nature of the information derived from the SUGARSENSE framework enables proactive measures, including lifestyle modifications and targeted interventions, aimed at mitigating diabetes risks and improving patient outcomes. While our findings are promising, further research is needed to enhance the generalizability and scalability of the model across diverse populations and healthcare settings. Additionally, the incorporation of additional data sources, such as genetic information and environmental factors, could enrich the predictive power of the model and enable more comprehensive risk assessment. Continued refinement and validation of predictive analytics frameworks like SUGARSENSE have the potential to revolutionize diabetes care by enabling early detection, personalized interventions, and ultimately, reducing the burden of diabetes on health systems.

References:

1. Al-Majeed, Salah H., et al. "Deep learning techniques for diabetes disease prediction: A comprehensive review." IEEE Access 9 (2021): 55796-55813.
2. Bansal, Aditi, and Ankush Mittal. "A review on application of machine learning techniques in diabetes prediction. "International Journal of Advanced Research in Computer Science 8.5 (2017): 146-149.
3. Chatterjee, Sayak, et al. "A survey on diabetes prediction using machine learning techniques." Procedia Computer Science 167 (2020): 1395-1405.

4. Patel, Tejas B., et al. "Deep learning approach for predicting diabetes risk using support vector machine." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 8.8 (2019): 412-417.
5. Rajalakshmi, P., and Dr R. Sivakumar. "A survey on machine learning approaches for diabetes prediction." International Journal of Computer Applications 138.1 (2016): 26-33.
6. Singh, Ashutosh, et al. "Diabetes prediction using machine learning algorithms." International Journal of Computer Applications 139.18 (2016): 22-29.
7. Tripathy, Bikash C., and Ram Bilas Pachori. "Machine learning techniques for diabetes detection and prediction." In 2020 IEEE International Conference on Power, Electrical, and Electronics & Communication Engineering (PEECE), pp. 1-5. IEEE, 2020.
8. Zeng, Jianqiang, et al. "Machine learning models for diabetic risk prediction and their performance in real-world settings: A systematic review and meta-analysis." BMC Medicine 18.1 (2020): 1-15.