Prediction of Type - I and Type –II Diabetes: A Hybrid Approach using Fuzzy Logic and Machine Learning Algorithms

Geeta Pattun *
*Corresponding Author, Department of Computer Science and Information Technology, School of Technology, Maulana Azad National Urdu University, Hyderabad, Telangana, India-500032. E-mail: geeta781@manuu.edu.in

Khaleda Afroaz
Department of Computer Science and Information Technology, School of Technology, Maulana Azad National Urdu University, Hyderabad, Telangana, India-500032. E-mail: khaleda.afroaz@manuu.edu.in

Ahmad Talha Siddiqui
Department of Computer Science and Information Technology, School of Technology, Maulana Azad National Urdu University, Hyderabad, Telangana, India-500032. E-mail: ahmadtalha2007@gmail.com

Shaheena Ghazala
Department of Computer Science and Information Technology, School of Technology, Maulana Azad National Urdu University, Hyderabad, Telangana, India-500032. E-mail: shahinaghazala@gmail.com

Abstract

Diseases like diabetes are chronic and require long-term management. Inadequate production of insulin results in high blood sugar levels. Such diseases lead to serious health issues such as heart ailments, blood vessel complaints, eye ailments, kidney function disorders, and nerve ailments. Hence, accurate assessment and management of risk factors are crucial for the onset of diabetes. Our proposed approach combines fuzzy logic & machine learning algorithms for diabetes risk prediction. Three machine learning models were trained to classify patients into two categories of diabetes (Type-I and Type-II) based on their clinical dataset collected from Katihar Medical College & Hospital and Suvadhan Lab. The polynomial regression algorithm achieved a score of 0.947, while the support vector regression algorithm with the rbf kernel achieved a score of 0.954, with a linear kernel achieved a score of 0.73. Our proposed
Prediction of Type - I and Type –II Diabetes...

approach performed well with respect to the conventional approaches with improved accuracy by identifying the patients at diabetes risk. In future work, we further analyze the relationship between other ignored factors which contribute to diabetes risk.

**Keywords:** Diabetes, Blood Sugar, Machine Learning Algorithm, Fuzzy Logic, Disease Management, Risk Factors, Insulin Resistance, Polynomial Regression, Support Vector Regression.

Journal of Information Technology Management, 2023, Vol. 15, Special Issue, pp. 35- 56
Received: July 06, 2023
Published by University of Tehran, Faculty of Management
Received in revised form: August 24, 2023
doi: https://doi.org/10.22059/jitm.2023.95244
Accepted: November 09, 2023
Published online: December 24, 2023

© Authors

**Introduction**

Nowadays, diabetes is a chronic disease that hinders the body's ability to regulate blood sugar level (Geman et al., 2017) (Bressan et al., 2020). Insulin, a hormone produced by the pancreas, plays a crucial role in breaking down the sugars we consume through our daily diet, converting them into glucose to fuel our body for daily activities (Howsalya Devi et al., 2020). However, when the body fails to produce adequate insulin, it can lead to severe complications such as heart disease, blood vessel disorders, eye damage, kidney dysfunction, and nerve disorders (Thakkar et al., 2021).

The two categories of diabetes are type –I and type –II. Type - I diabetes is a perennial autoimmune disease found in infants and adults (Devi & Uma, 2019; Howsalya Devi et al., 2020). This situation occurs when insulin production by pancreatic beta cells is impaired, leading to insulin deficiency and spikes in blood sugar levels. Symptoms to watch out for are excessive hunger, frequent thirst, frequent urination, weight loss, fatigue, allergic reactions, and blurred vision. Premature treatment can lead to serious complications and even death. Insulin therapy is a common treatment given for regulating blood sugar levels of type 1 diabetes.

Type 2 diabetes is a long-term disease that primarily affects the population between the ages of 30 and under 100 (Yahyaoui, Jamil & Rasheed, 2019; Howsalya Devi et al., 2020; Rajeswari et al., 2018; Vijiyakumar et al., 2019). It occurs when the body does not produce enough insulin, causing blood sugar levels to rise. Common symptoms to look out for are excessive hunger, extreme thirst, weakness, distorted vision, itching, excessive urination, slow recovery, and the development of hemorrhoids. These symptoms in any circumstances left unchecked can lead to serious complications to name few heart ailments, kidney functionality...
disorders, myopia, and nerve ailments. Several lifestyle factors also contribute to type 2 diabetes.

Fuzzy logic is a mathematical framework that provides a solution for processing data that is inaccurate, uncertain, or ambiguous. It is particularly beneficial in dealing with real-world issues where knowledge is often incomplete or imprecise. By utilizing Fuzzy logic, it becomes easier to develop complex relationships between variables and rules, simplifying the decision-making process and making it easier to comprehend (Bressan et al., 2020; Niswati et al., 2018; Rajeswari et al., 2018; Swain et al., 2013; Thakkar et al., 2021). Fuzzy logic has widespread applications, ranging from control theory to artificial intelligence. Its usefulness lies in its ability to incorporate fuzzy risk factors, such as "slightly obese" or "moderately high" blood sugar levels, into predictive models that cannot be represented by traditional binomial taxonomies (Bressan et al., 2020). This integration of ambiguous risk factors into predictive models can significantly enhance model accuracy and robustness, which is particularly important when predicting the risk of diabetes.

A potent tool for predicting diabetes has emerged: machine learning. These algorithms are able to recognize patterns and relationships that might escape the notice of human experts by training models on vast datasets of patient data. In order to produce precise predictions of a patient's risk for developing diabetes, machine learning models can consider a wide range of variables, such as age, weight, family history, and blood glucose levels. The outcomes of patients may be improved, and healthcare costs may be decreased, if healthcare professionals can intervene earlier with preventative measures. Machine learning algorithms can also continuously increase their predictive accuracy (Lukmanto et al., 2019; Sarwar et al., 2018; Thakkar et al., 2021) as more data is gathered over time, which makes them a valuable tool in the battle against this crippling illness.

Our research introduces a novel approach to diagnosing diabetes. We used fuzzy logic systems to determine boundary values for predicting abnormalities, enabling us to identify individuals who need additional care. By in-depth studying of risk factors associated with type 1 and type 2 diabetes across different age groups, we can contribute valuable insights that inform public health initiatives. Our innovative hybrid approach, combining fuzzy logic with machine learning, enhances the precision of diabetes risk prediction models. This cutting-edge method (Khan et al., 2022; Kumar et al., 2021) effectively handles uncertain data and improves early identification and intervention strategies. Our research has significant implications for healthcare practices and decision support systems, ultimately benefiting individuals living with diabetes.

**Literature Review**

We have tabulated the previous work done in the context of the prediction of diabetes types. Through Table 1 we attempted to project fuzzy logic adopted for diabetes prediction, in
particular, it describes the methods implemented by researcher enthusiasts & the datasets utilized for training and evaluating the system. In Figure 1 we projected the accuracy of the models developed through a bar plot, in Figure 2 we projected the dataset adopted in the previous study through the count plot, and through the graph we could make out that much of the models developed were based on the Pima Indian diabetes dataset.

### Table 1. Methods implemented in previous work for diabetes prediction

<table>
<thead>
<tr>
<th>S.No</th>
<th>Author</th>
<th>Method</th>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Vijiyakumar et al., 2019)</td>
<td>Random Forest classifier</td>
<td>pima Indian diabetes dataset (UCI)</td>
<td>90%</td>
</tr>
<tr>
<td>2</td>
<td>(A. Yahyaoui et al., 2019)</td>
<td>CNN, SVM, Random Forest</td>
<td>pima Indian diabetes dataset (UCI)</td>
<td>83.67%</td>
</tr>
<tr>
<td>3</td>
<td>(Sarwar et al., 2018)</td>
<td>SVM, KNN, LR, Decision Tree, NB</td>
<td>Pima Indian diabetes dataset (UCI)</td>
<td>77%</td>
</tr>
<tr>
<td>4</td>
<td>(Bressan et al., 2020)</td>
<td>Neuro Fuzzy, Decision Tree</td>
<td>Moreira Glycemic index data</td>
<td>70%</td>
</tr>
<tr>
<td>5</td>
<td>(Geman et al., 2017)</td>
<td>FNFIS Neuro Fuzzy System</td>
<td>Pima Indian diabetes dataset (UCI)</td>
<td>84%</td>
</tr>
<tr>
<td>6</td>
<td>(Howsalya Devi et al., 2020)</td>
<td>farthest First &amp; Smo classifier</td>
<td>Pima Indian diabetes dataset (UCI)</td>
<td>99.40%</td>
</tr>
<tr>
<td>7</td>
<td>(Khalil &amp; Al-Jumaily, 2017)</td>
<td>SVM, PNN, K-means, Fuzzy e-Mean, probabilistic NN</td>
<td>Black Lion general specialized hospital Ethiopia</td>
<td>96%</td>
</tr>
<tr>
<td>8</td>
<td>(Lukmanto et al., 2019)</td>
<td>Fuzzy Support Vector Machine</td>
<td>Pima Indian diabetes dataset (UCI)</td>
<td>89%</td>
</tr>
<tr>
<td>9</td>
<td>(Niswati et al., 2018)</td>
<td>Fuzzy System Mamdani Method</td>
<td>6 puskesnas east jakarta sample data</td>
<td>96%</td>
</tr>
<tr>
<td>10</td>
<td>(Raj et al., 2019)</td>
<td>Naïve Bayes, SVM</td>
<td>sample from multiple hospitals</td>
<td>82%</td>
</tr>
<tr>
<td>11</td>
<td>(Rajeswari et al., 2018)</td>
<td>Fuzzy Association classifier</td>
<td>pima Indian diabetes dataset (UCI)</td>
<td>84%</td>
</tr>
<tr>
<td>12</td>
<td>(Swain et al., 2013)</td>
<td>NN, Hybrid Adaptive Neuro Fuzzy System</td>
<td>Local inhabitants of Bhubaneshwar, India</td>
<td>66%</td>
</tr>
<tr>
<td>13</td>
<td>(Thakkar et al., 2021)</td>
<td>Fuzzy System</td>
<td>Pima Indian diabetes dataset (UCI)</td>
<td>96%</td>
</tr>
<tr>
<td>14</td>
<td>(Undre et al., 2015)</td>
<td>Fuzzy Neural Network, cbr approach</td>
<td>Pima Indian diabetes dataset (UCI)</td>
<td>90%</td>
</tr>
<tr>
<td>15</td>
<td>(Verma &amp; Mishra, 2017)</td>
<td>Naïve Bayes, SVM, RED Tree, J48</td>
<td>Pima Indian diabetes dataset (UCI)</td>
<td>76.80%</td>
</tr>
</tbody>
</table>
Figure 1. Accuracy obtained in diabetes analysis

Figure 2. Diabetes dataset used lated work
**Methodology**

We developed a novel approach for diagnosing diabetes by combining fuzzy logic and machine learning. Our hybrid approach employs a fuzzy system set by experts based on clinical data (Undre et al., 2015). Using this method, we can accurately classify individuals with type I or type II diabetes by extracting pertinent features from diabetic datasets. One of the key strengths of our study lies in the utilization of rule-based feature extraction techniques to enhance the diagnostic precision of the model. By employing these methods, we could effectively extract relevant features from the data, leading to improved accuracy and reliability in the diagnostic process. This model contributes to its overall strength and effectiveness in diagnosing the target condition of the individuals with such a chronic condition (Lukmanto et al., 2019).

![Figure 3. Hybrid model based on machine learning & Fuzzy based system](image)

A mathematical framework called fuzzy logic can be used to reason and make conclusions based on input that is ambiguous or uncertain. It can be used in decision-making systems, engineering, medicine, finance, and economics, among other areas (Bressan et al., 2020; Rajeswari et al., 2018). To convert precise values into degrees of membership using fuzzy logic, membership functions, and linguistic concepts are used (Thakkar et al., 2021). The IF-THEN statements are used to create fuzzy rules and logical operators like AND, OR, and NOT are used in the inference process to create a single fuzzy output result. In our research, we emphasize the critical step of defuzzification, where the fuzzy output value is transformed into a crisp value for decision-making. With the defuzzification process we can ensure clarity and facilitates informed decision-making based on the output of the fuzzy logic system (Niswati et al., 2018; Swain et al., 2013).
Figure 4. Fuzzy Expert system

Machine learning algorithms

Several algorithms have been applied to train the dataset for diabetes patient prediction in a hybrid environment that integrates fuzzy logic-based systems with machine-learning approaches (Howsalya Devi et al., 2020; Swain et al., 2013).

Linear Regression is a simple algorithm that models the linear relationship between input and output variables, which can be used to predict blood glucose levels based on factors such as age, BMI, and blood pressure. Each input value or column receives one scale factor from the linear equation, known as a coefficient, and is denoted by the capital Greek letter Beta (B). This second coefficient, often known as the intercept or bias coefficient, offers the line an extra degree of freedom, allowing it to move up and down on a two-dimensional plot. As an illustration, the model form for a simple regression problem (one x and one y) might be as follows:

$$y = B_0 + B_1 \times X$$

B0 stands for the Intercept. The independent variable is denoted by the symbol B1, which represents the coefficient X. y stands for the dependent variable or outcome.

Support Vector Regression (SVR) is one such technique that converts input data into a high-dimensional feature space and locates a hyperplane that maximizes the separation between the data points. For predicting continuous variables like blood glucose levels, this is helpful (Yahyaoui, Jamil & Rasheed, 2019; Khalil & Al-Jumaily, 2017). The SVR seeks to fit the best line within a threshold value (Distance across the hyperplane and boundary line) a, in contrast to other regression models that aim to minimize the deviation between the real and predicted value. As a result, we can say that the SVR model tries to meet

Condition $$-a < y - wx + b < a$$

To forecast the value, it use the locations along this boundary. Because of its adaptability, SVR discovers a suitable line (or hyperplane in greater dimensions) to fit the
data and lets us select how much variation in the model is acceptable. (Howsalya Devi et al., 2020; Lukmanto et al., 2019)

Polynomial Regression, this is a variant of linear regression that uses a polynomial function to describe non-linear correlations between input and output variables. This variation can help anticipate intricate relationships between patient parameters and blood glucose levels. However, it is vulnerable to overfitting and might not work well when used with fresh data.

**Data Collection**

This research is an experiment on primary diabetes disease on the original dataset. To conduct the study we selected participants from clinical records of individuals diagnosed with type 1 or type 2 diabetes, who fall into different age groups.

We used data from 836 diabetic patients of different ages who were treated at Katihar Medical College & Hospital and Suvadhan Laboratory. We only included patients with type 1 or type 2 diabetes. We collected the data using a convenient method.

The data analysis graph describes a dataset of diabetic patients who received treatment at Katihar Medical College & Hospital and Suvadhan Laboratory. The dataset includes 836 rows and 16 columns, representing various attributes of the patients such as patient id, patient name, gender, age, blood pressure, fasting plasma, two hours plasma, clinical history, Hemoglobin, BMI, family history, blood sugar, total cholesterol, Triglycerides, insulin and target column diabetic type. The patients in the dataset have different ages, and they were diagnosed with either type 1 or type 2 diabetes.

**Table 2. Describes the variables of the diabetes dataset**

<table>
<thead>
<tr>
<th>S.no</th>
<th>Attribute</th>
<th>Units/measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Patient Id</td>
<td>Required</td>
</tr>
<tr>
<td>2</td>
<td>Patient Name</td>
<td>Required</td>
</tr>
<tr>
<td>3</td>
<td>Gender(F/M)</td>
<td>Category (male, female)</td>
</tr>
<tr>
<td>4</td>
<td>Age</td>
<td>Years(1-100)</td>
</tr>
<tr>
<td>5</td>
<td>Blood Pressure</td>
<td>(1-130) mmHg</td>
</tr>
<tr>
<td>6</td>
<td>Fasting Plasma Glucose</td>
<td>126 mg/dl</td>
</tr>
<tr>
<td>7</td>
<td>Two Hours After Eating</td>
<td>Mg/dl</td>
</tr>
<tr>
<td>8</td>
<td>Clinical History</td>
<td>If any</td>
</tr>
<tr>
<td>9</td>
<td>Hemoglobin A1c</td>
<td>HbA1c(4.0-14.0)</td>
</tr>
<tr>
<td>10</td>
<td>Body Mass Index</td>
<td>Kg/height (0-80)</td>
</tr>
<tr>
<td>11</td>
<td>Family history</td>
<td>If any</td>
</tr>
<tr>
<td>12</td>
<td>Blood Sugar</td>
<td>(mg/dl)</td>
</tr>
<tr>
<td>13</td>
<td>Total Cholesterol</td>
<td>(mg/dl)</td>
</tr>
<tr>
<td>14</td>
<td>Triglycerides</td>
<td>(mg/dl)</td>
</tr>
<tr>
<td>15</td>
<td>Insulin</td>
<td>µIU/mL</td>
</tr>
<tr>
<td>16</td>
<td>Diabetes Status</td>
<td>Type –I, Type –II</td>
</tr>
</tbody>
</table>
Data Preprocessing

Family clinical history has the highest percentage of missing results when compared to fasting plasma glucose, two hours after eating, triglycerides, and HbA1c. The locations of missing values in columns were located using a heatmap visualization of the correlation matrix, and a distribution plot was generated to illustrate the distribution of missing data.

![Figure 5. Distribution plot for missing values in dataset](image)

We employed one-hot encoding and label encoding approaches to handle the missing data and transform the family clinical history categorical variable into a numerical form for analysis. Each categorical variable was handled similarly and we used the results as features to enhance our predictions.
We were able to determine the important features of a continuous variable in the diabetic dataset through histograms. We discovered, in particular, that fasting plasma glucose is right-skewed with a few outliers, insulin data is bimodal with some outliers, triglycerides are right-skewed and contain outliers, blood sugar is symmetrically distributed but have more outliers than other variables, and the data for glucose two hours after eating has multiple peaks and a few outliers that need further investigation. Figure 6 & Figure 7 describes the distribution of data using the histogram.

Figure 6. Histogram for fasting plasma glucose, insulin, total cholesterol, triglycerides
Figure 7. histogram for blood sugar and two hours after eating

It is important to note that outliers can significantly affect statistical analyses, so it's crucial to handle them appropriately.
Figure 8. Correlation matrix describing the variables relation with respect to type 1 diabetes patients

RQ1: What are the factors that contributes to the risk of type 1 diabetes among the varied age groups?

Age, insulin, and clinical history were revealed to be the risk factors for Type-1 diabetes by utilizing a heat map of the data. We can build prevention and treatment strategies using this information to better understand the causes of Type-1 diabetes.
Figure 9. Line plot for Age Vs Diabetes Type

Figure 10. Insulin Vs Diabetes Type

Figure 11. Clinical History Vs Diabetes Types
A deeper analysis was conducted on the three variables - age, insulin, and clinical history against the target variable. The results showed that Type-1 diabetes is more common in lower age groups and the amount of insulin is negligible in Type-1 patients. The plot also revealed that individuals who have an autoimmune disorder are more likely to develop Type-1 diabetes. This information can help healthcare professionals better diagnose and treat Type-1 diabetes, as well as potentially develop preventative measures for individuals at risk (Howsalya Devi et al., 2020).

RQ2: What are the factors that contribute to the risk of type 2 diabetes among the varied age groups?

The analysis showed that age, insulin, and clinical history are important factors for identifying type-II diabetes patients. Glucose and triglyceride levels after eating are also important. Type-II diabetes is more common with increasing age and can lead to hypertension due to unstable blood glucose levels (Bressan et al., 2020).

![Correlation matrix for diabetes dataset variables with type-II diabetes patients](Figure 12)
Figure 13. Describing the number of patients identified as diabetes patients for various age groups

The count plot projects the distribution of glucose levels among the type – II diabetes sufferers, which covers the complete graph, while the distribution of glucose levels for patients with type I diabetes has only a little overlap covering the smaller section comparatively, which suggests that glucose levels are a good predictor for distinguishing between the two classes. Since type II diabetes diseases patients are characterized by insulin resistance and high blood glucose levels, patients with type II diabetes patients often have higher glucose levels than those with type-I diabetes individuals.

Figure 14. Describes the glucoses fluctuations after two hours among diabetes patients
According to our study, high amounts of total cholesterol and triglycerides accumulate in Type-2 diabetic individuals who have a clinical history of obesity. Furthermore, the findings show that normal to moderate insulin levels are often present in Type-2 diabetic patients.
RQ 3: What are the boundary values for prediction of outliers in risk factors for diabetes using a fuzzy logic system?

The clinical data of the diabetic dataset revealed the presence of 15 outliers. These outliers were observed in variables such as age, insulin, triglycerides, glucose, and blood pressure. We employed the center of gravity defuzzification method to detect uncertain variables in the diabetic dataset. Further analysis showed that some patients had clinical values that were lower than the range set by member functions (Swain et al., 2013), while others had values that exceeded the range. For example, the defuzzification plot of Triglycerides shows more overlap than the plots of other variables like blood pressure and glucose, it may indicate that the Triglycerides variable has more ambiguity or uncertainty associated with it than the other variables. This could potentially lead to ambiguity in the decision-making process of the fuzzy system. It is important to consider the overall performance of the fuzzy logic system; Therefore, it is necessary to explore alternative techniques to identify outliers in the any other diabetes dataset for better predictions.

```
# Defining Constants
# Age (0-100)
age_low = 35
age_med = (30, 50)
age_high = 45
# Blood Pressure (0 - 130)
BP_low = 105
BP_med = (95, 135)
BP_high = 125
# Fasting Plasma Glucose (126mg/dl or Higher)
Glucose_low = 110
Glucose_med = (100, 350)
Glucose_high = 340
# BMI (height, weight) Level(0 - 30)
BMI_low = 55
BMI_med = (50, 80)
BMI_high = 75
# Insulin
Insulin_low = 25
Insulin_med = (15, 330)
Insulin_high = 320
# Total Cholesterol
TC_low = 210
TC_med = (200, 300)
TC_high = 320
# Triglycerides
TG_low = 130
TG_med = (120, 220)
TG_high = 210
```

Figure 18. linguistic variable for fuzzy system
RQ4: How can a hybrid approach incorporating fuzzy logic and machine learning algorithms enhance the accuracy of diabetes risk prediction models?

The diabetic dataset describes a binary classification problem with two classes, type-I and type-II diabetic patients, with the dataset being class-imbalanced. To address this class imbalance, the under-sampling technique was adopted, but this approach suffered from the loss of data.

Feature scaling

Standardization and min-max scaling were used in the feature scaling on the diabetes datasets numerical variables. Some rows were removed to enhance the quality of the data for the training and testing procedure.
Table 3. Descriptive analysis of the diabetes dataset

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Std</th>
<th>Min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender(F/M)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (0-100)</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>Blood Pressure (0-140)</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>Fasting Plasma Glucose (126mg/dl or Higher)</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>Two Hours after Eating (100-500)</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>Clinical History (Prior Medical Condition)</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>HbA1C (4.0-14.0)</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>BMI(height, weight Level0-00)</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>Blood Sugar</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>Total Cholesterol</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
<tr>
<td>Insulin</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
<td>65.2</td>
</tr>
</tbody>
</table>

Feature extraction

Regression and correlation coefficient analyses identified the age, fasting plasma glucose, clinical history, total cholesterol, and triglycerides as influential features. As per literature, the rules in the expert system are used to extract features from the training data, which are then used as input to the machine learning algorithm. Even the defuzzification output values can also be used as a feature or input variable in your machine learning algorithm. For example, you can use the defuzzification output value as a predictor variable in a logistic regression, decision tree, or support vector machine algorithm to predict the likelihood of a patient being type I or type II diabetic. The crisp output labels can be used as the ground truth or target variable in a supervised machine learning algorithm. Before training the machine learning algorithms on the KMC diabetic dataset collected from the Suvadhan laboratory, preliminary activities such as handling missing values, transforming data, scaling features, and feature extraction were done from fuzzy rules. Three different machine learning models were trained to classify patients into two categories of disease (Type-I and Type-II) based on their clinical data.

Performance Evaluation

The three algorithms used for training were support vector regression (SVR), linear regression, and polynomial regression. These algorithms were given fuzzy rules as features to tune the training process with respect to the dataset. The linear regression model achieved a score of 0.73, with a mean square error (MSE) of 0.033 and a root mean square error (RMSE) of 0.1833. The polynomial regression algorithm achieved a score of 0.947, while the support
vector regression algorithm with 'rbf' kernel achieved a score of 0.954, and the one with a 'linear' kernel achieved a score of 0.73 (Lukmanto et al., 2019).

![Score Comparison](image)

**Figure 21. Describing the scores obtained by the machine learning algorithm**

**Conclusion**

Our Machine learning models have performed well and very well utilized the ground truth of the fuzzy system. However, it's important to note that glucose levels can vary widely among individuals, and other factors such as age, lifestyle, and genetics can also contribute to the risk of developing diabetes. Therefore, additional analysis may be necessary to fully understand the relationship between glucose levels and diabetes risk. In future work we may involves investigating into new risk factors, enhancing forecasting techniques, creating more precise and user-friendly diagnostic tools, and assessing the success of remedies. We can work towards reducing the effects of diabetes takes on individuals as well as the community at large by expanding our understanding and putting evidence-based practices into action.

**Conflict of interest**

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.
References


Bibliographic information of this paper for citing:


Copyright © 2023, Geeta Pattun, Khaleeda Afroaz, Ahmad Talha Siddiqui and Shaheena Ghazala