A Study on Variable Selections and Prediction for Diabetes Pregnancy Dataset Using Data Mining With Machine Learning Approaches

G. K. Arun¹, Dr. P. Rajesh²

¹Assistant Professor, Department of Computer Science, Arigner Anna Government Arts College, Villupuram – 605 602. (Deputed from Dept. of Computer and Information Science, Annamalai University, Annamalainagar-608 002) Tamil Nadu, India.

²Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram – 608 102, (Deputed from Dept. of Computer and Information Science, Annamalai University, Annamalainagar-608 002) Tamil Nadu, India.

Email: arunnura2370@gmail.com¹, rajeshdatamining@gmail.com²

Abstract

Data mining aims to convert raw data into valuable insights that inform decision-making, predictions, and business-related research. Machine learning, a branch of artificial intelligence (AI), creates algorithms and models that empower computers to learn from data and make data-driven predictions or decisions. This paper considers diabetes-related parameters, namely gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose, for applying machine learning techniques to find diabetes and suitable variables for future predictions in diabetes using Gaussian Process, Linear Regression, Multilayer Perceptron, Simple Linear Regression, SMOreg, decision stump, M5P, random forest, random tree, and REP tree. Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

Keywords: Data Mining, Machine Learning, Decision Tree, Diabetes, And Test Statistics.

1. Introduction

Despite our increasing knowledge about how to prevent and treat diabetes, the global diabetes epidemic continues to escalate, particularly in low- and middle-income countries, leading to unnecessary suffering and death. Currently, more than 420 million individuals are affected by diabetes, which has quadrupled since 1980 and is projected to surpass half a billion by the end of this decade.

WHO declared that surging prevalence of diabetes can be predominantly attributed to the rising rates of obesity and physical inactivity? The percentage of overweight and obese children and adolescents aged 5-19 has surged from a mere 4% in 1975 to an alarming 18% in 2016. Alarming statistics also reveal a 70% global increase in diabetes-related deaths between 2000 and 2019. This condition now stands as a leading contributor to the escalating male mortality rate among the top 10 causes of death, marking an 80% increase since 2000. Furthermore, diabetes remains a significant factor in conditions such as blindness, kidney failure, heart attacks, strokes, and lower limb amputations. Shockingly, one out of every two adults with type 2 diabetes is unaware of their condition, highlighting a widespread lack of diagnosis and care. Across the globe, healthcare systems are falling short in their efforts to adequately address the needs of individuals living with diabetes.

Data mining encompasses diverse methods like clustering, classification, regression analysis, association rule mining, anomaly detection, text mining, and time series analysis. Its significance spans various fields such as business, marketing, finance, healthcare, and scientific research. Data mining provides valuable insights by analyzing structured and unstructured data.
Hybrid data mining and case-based reasoning user modeling system, which is used to monitor and predict the blood sugar level in diabetics. The practical objective for this project is to reduce the cost of direct blood sugar self-monitoring by minimizing the number of times that a diabetic needs to measure his or her sugar levels every day. From the technological point of view, the main aim is using the support vector machine as the classifier and implementing a case-based reasoning cycle as the retrieval cycle in order to indirectly determine and predict blood sugar level in diabetics and finally implement this software into a mobile device with wireless sensor networks and link it to a server which houses the relevant knowledgebase [1].

Diabetes affects between 2% to 4% of the global population and its avoidance and effective treatment are undoubtedly crucial public issues in the 21st century. Although human decision making is often optimal, it is poor when there are huge amounts of data to be classified. Medical data mining has been a great potential for exploring hidden patterns in the data sets of medical domain. Data mining algorithms can be trained in clinical data to predict the disease. Classification is the generally used technique in medical data mining. Presents results comparison of five supervised data mining algorithms using five performance criteria. The performance is evaluated by the five algorithms C4.5, SVM, k-NN, PNN, and BLR. Comparison of performance of data mining algorithms based on computing time, precision value, the data evaluated using 10 fold Cross Validation error rate, bootstrap validation and accuracy. A typical confusion matrix is furthermore displayed for quick check. The study describes algorithmic discussion of the dataset for the disease acquired from UCI and ICMR-INDIAB, on line repository of large datasets. Tanagra tool is used to achieve the best results. Tanagra is data mining matching set [2].

From the past few years, data mining got a lot of attention for extracting information from large datasets to find patterns and to establish relationships to solve problems. Well known data mining algorithms include classification, association, Naïve Bayes, clustering and decision tree. In medical science field, these algorithms help to predict a disease at early stage for future diagnosis. Diabetes mellitus is the most growing disease that needs to be predicted at its early stage as it is lifelong disease and there is no cure for it. This research is intended to provide comparison for different data mining algorithms on PID dataset for early prediction of diabetes [3].

2. Literature Review
Recently, several research teams conducted detailed research on the data mining platform to determine the precision of each other. Data mining can be used by parametric modeling from the health data, including diabetic patient data sets, to synthesize expertise in the field. Methods: This study proposes a new model for forecasting type 2 diabetes mellitus (T2DM) based on data mining strategies. The combined Particle Swarm Optimization (PSO) and Fuzzy Clustering Means (FCM) (PSO-FCM) are used to evaluate a set of medical data relating to a diabetes diagnosis challenge [4].

Diabetes mellitus is a disease that is caused due to increased blood sugar levels because of imbalance in insulin processing by the body. It can easily be diagnosed by hospitals and has major consequences if left untreated. By using efficient and reliable data mining techniques to identify trends and predict the onset of diabetes in people will help in preventing the disease early and for treatment. Data mining is the process where we take useful information from relevant datasets by applying algorithms and frameworks. This paper does a survey on the different kinds of predictions using machine learning techniques done on diabetes patients [5].

Data mining is a valuable tool for the practice of examining large pre-existing databases to generate previously unknown helpful information. The input for the weather data set denotes specific days as a row, attributes denote weather conditions on the given day, and the class indicates whether the conditions are conducive to playing golf. Attributes include Outlook, Temperature, Humidity, Windy, and Boolean Play Golf class variables. All the data are considered for training purpose, and it is used in the seven-classification algorithm likes J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP) and Random Forest (RF) are used to measure the accuracy. Out of seven classification algorithms, the Random tree algorithm outperforms other algorithms by yielding an accuracy of 85.714% [6].

This study addresses for applying data-mining techniques in diabetes research which gives a rational insight to model predicate patterns that can forecast incidence of Diabetes Mellitus disease (DMD) in human race. Clinical Patient records and Pathological test reports inherently represent data sets which
may be applied to data mining for diabetes research. Hidden knowledge rules may be extracted to new hypothesis for improving standards and quality in the field of health care for diabetes patients. Primary Data mining methods such as Rule classification and Decision trees are used [7].

In digitized world, data is growing exponentially and Big Data Analytics is an emerging trend and a dominant research field. Data mining techniques play an energetic role in the application of Big Data in healthcare sector. Data mining algorithms give an exposure to analyse, detect and predict the presence of disease and help doctors in decision-making by early detection and right management. The main objective of data mining techniques in healthcare systems is to design an automated tool which diagnoses the medical data and intimates the patients and doctors about the intensity of the disease and the type of treatment to be best practiced based on the symptoms, patient record and treatment history. This paper emphasises on diabetes medical data where classification and clustering algorithms are implemented and the efficiency of the same is examined [8].

Diabetic retinopathy the most common diabetic eye disease, is caused by complications that occurs when blood vessels in the retina weakens or distracted. It results in loss of vision if early detection is not done. Several data mining technique serves different purposes depending on the modeling objective. The outcome of the various data mining classification techniques was compared using rapid miner tool. We have used Naive bayes and Support Vector Machine to predict the early detection of eye disease diabetic retinopathy and found that Naive bayes method to be 83.37% accurate. The performance was also measured by sensitivity and specificity. The above methodology has also shown that our data mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class which we are trying to predict [9].

Data mining is discovering hiding information that efficiently utilizes the prediction by stochastic sensing concept. This paper proposes an efficient assessment of groundwater level, rainfall, population, food grains, and enterprises dataset by adopting stochastic modeling and data mining approaches. Firstly, the novel data assimilation analysis is proposed to predict the groundwater level effectively. Experimental results are done, and the various expected groundwater level estimations indicate the sternness of the approach [10] and [11].

The input for the chronic disease data denotes a specific location as a row; attributes denote topics, questions, data values, low confidence limit, and high confidence limit. All the data are considered for training and testing using five classification algorithms. In this paper, the authors present the various analysis and accuracy of five different decision tree algorithms; the M5P decision tree approach is the best algorithm to build the model compared with other decision tree approaches [12].

3. Materials and Methods
A data mining decision tree is a widely used machine learning technique for classification and regression tasks. It visually depicts a sequence of decisions and their possible outcomes in a tree-like structure. Each internal node represents a decision based on a specific feature, and each branch corresponds to the potential result of that decision. The tree's leaf nodes represent the final decision or the predicted outcome. The "CART" (Classification and Regression Trees) algorithm is the most used algorithm for building decision trees [13].

Linear Regression
Linear regression is a statistical technique employed to comprehend and forecast the connection between two variables by discovering the optimal straight line that most effectively aligns with the data points. It aids in ascertaining how alterations in one variable correspond to changes in another, proving valuable for predictions and trend recognition. The core idea of linear regression is to find the best-fitting straight line (also called the "regression line") through a scatterplot of data points. This line represents a linear equation of the form:

\[ y = mx + b \]

Where \( y \) is the dependent variable, \( x \) is the independent variable, \( m \) is the slope of the line, representing how much, \( y \) changes for a unit change in \( x \) and \( b \) is the \( y \)-intercept, indicating the value of \( y \) when \( x \) is 0.

REP Tree
REP (Repeated Incremental Pruning to Produce Error Reduction) Tree is a machine learning algorithm for classification and regression tasks. A decision tree-based algorithm constructs a decision tree using
a combination of incremental pruning and error-reduction techniques. The key steps involved in building a REP Tree are as follows:

Step 1. Recursive Binary Splitting

Step 2. Pruning

Step 3. Repeated Pruning and Error Reduction

Step 4. Model Evaluation

**Correlation coefficient**

The correlation coefficient, often denoted by the symbol "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is commonly used to assess the degree to which changes in one variable are associated with changes in another. The correlation coefficient takes values between -1 and 1:

A correlation coefficient of +1 indicates a perfect positive linear relationship where the two variables increase together. A correlation coefficient of -1 indicates a perfect negative linear relationship, where one variable increases as the other decreases. A correlation coefficient close to 0 indicates a weak or no linear relationship between the variables.

\[
r = \frac{\Sigma ((X - \bar{x}) (Y - \bar{y}))}{\sqrt{\Sigma (X - \bar{x})^2 * \Sigma (Y - \bar{y})^2}}
\]... (1)

**Mean Absolute Error**

Mean Absolute Error (MAE) is a metric used to measure the average absolute difference between predicted and actual (true) values in a regression problem. It is commonly used to assess the accuracy of a regression model's predictions [14]. The formula to calculate Mean Absolute Error (MAE) is as follows:

\[
MAE = \frac{\Sigma \mid (Actual \ Value - Predicted \ Value) \mid}{n}
\] ... (2)

Where Σ represents the summation symbol, which sums up the values for all data points, | | denotes the absolute value, ensuring the differences are positive. In this formula, Actual Value Refers to the true value of the target variable (ground truth) for a specific data point. Predicted Value: Refers to the value predicted by the regression model for the same data point and n represents the total number of data points in the dataset.

**Root Mean Squared Error (RMSE)**

Root Mean Squared Error (RMSE) is a commonly used metric to assess the accuracy of a regression model's predictions. It measures the average magnitude of the errors between the predicted and actual (true) values, considering both the direction and magnitude of the errors. The formula to calculate Root Mean Squared Error (RMSE) is as follows [15]:

\[
RMSE = \sqrt{\frac{\Sigma (Actual \ Value - Predicted \ Value)^2}{n}}
\] ... (3)

Where Σ represents the summation symbol, which sums up the values for all data points, (Actual Value - Predicted Value)² denotes the squared difference between the actual and predicted values for each data point and n is the total number of data points in the dataset.

**Relative Absolute Error (RAE)**

Relative Absolute Error (RAE), also known as Mean Absolute Percentage Error (MAPE), is a metric used to evaluate the accuracy of predictions in regression tasks. It measures the average percentage difference between the absolute and actual (valid) values, providing a relative measure of the prediction errors [16]. The formula to calculate Relative Absolute Error (RAE) is as follows:

\[
RAE = \frac{\Sigma \mid (Actual \ Value - Predicted \ Value) \mid}{\Sigma \mid Actual \ Value \mid} \times (100 / n)
\] ... (4)

Where Σ represents the summation symbol, which sums up the values for all data points, | | denotes the absolute value, ensuring the differences are positive, n is the total number of data points in the dataset.

**Numerical Illustrations**

The corresponding dataset was collected from the open source Kaggle data repository [17]. The diabetes pregnancy dataset include 9 parameters which have different categories of data like gender, age,
hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, diabetes [18]. A detailed description of the parameters is mentioned in the following Table 1.

<table>
<thead>
<tr>
<th>gender</th>
<th>age</th>
<th>hypertension</th>
<th>heart_disease</th>
<th>smoking_history</th>
<th>bmi</th>
<th>HbA1c_level</th>
<th>blood_glucose_level</th>
<th>diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>80</td>
<td>0</td>
<td>1</td>
<td>Never</td>
<td>25.19</td>
<td>6.6</td>
<td>140</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>No Info</td>
<td>27.32</td>
<td>6.6</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>Never</td>
<td>27.32</td>
<td>5.7</td>
<td>158</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>36</td>
<td>0</td>
<td>0</td>
<td>Current</td>
<td>23.45</td>
<td>5</td>
<td>155</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>76</td>
<td>1</td>
<td>1</td>
<td>Current</td>
<td>20.14</td>
<td>4.8</td>
<td>155</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>Never</td>
<td>27.32</td>
<td>6.6</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>Never</td>
<td>19.31</td>
<td>6.5</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>79</td>
<td>0</td>
<td>0</td>
<td>No Info</td>
<td>23.86</td>
<td>5.7</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>Never</td>
<td>33.64</td>
<td>4.8</td>
<td>145</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>Never</td>
<td>27.32</td>
<td>5</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>53</td>
<td>0</td>
<td>0</td>
<td>Never</td>
<td>27.32</td>
<td>6.1</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>Former</td>
<td>54.7</td>
<td>6</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>Former</td>
<td>36.05</td>
<td>5</td>
<td>130</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>Never</td>
<td>25.69</td>
<td>5.8</td>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>Female</td>
<td>76</td>
<td>0</td>
<td>0</td>
<td>No Info</td>
<td>27.32</td>
<td>5</td>
<td>160</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>No Info</td>
<td>27.32</td>
<td>6.6</td>
<td>126</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Machine Learning Models with Linear Regression

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Correlation coefficient</th>
<th>MAE</th>
<th>RMSE</th>
<th>RAE</th>
<th>RRSE</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.7061</td>
<td>16.2325</td>
<td>19.4204</td>
<td>85.1001</td>
<td>86.2482</td>
<td>0.6500</td>
</tr>
<tr>
<td>hypertension</td>
<td>0.2979</td>
<td>0.1367</td>
<td>0.2512</td>
<td>98.6989</td>
<td>95.4593</td>
<td>0.5600</td>
</tr>
<tr>
<td>heart_disease</td>
<td>0.2832</td>
<td>0.0804</td>
<td>0.1866</td>
<td>106.1361</td>
<td>95.9063</td>
<td>1.2400</td>
</tr>
<tr>
<td>bmi</td>
<td>0.3882</td>
<td>4.4376</td>
<td>6.1164</td>
<td>100.4019</td>
<td>92.1577</td>
<td>0.4400</td>
</tr>
<tr>
<td>HbA1c_level</td>
<td>0.4006</td>
<td>0.8639</td>
<td>0.9810</td>
<td>97.4624</td>
<td>91.6254</td>
<td>0.5500</td>
</tr>
<tr>
<td>blood_glucose_level</td>
<td>0.4193</td>
<td>30.6934</td>
<td>36.9564</td>
<td>99.4240</td>
<td>90.7831</td>
<td>0.5600</td>
</tr>
<tr>
<td>diabetes</td>
<td>0.8906</td>
<td>0.1538</td>
<td>0.2251</td>
<td>98.8652</td>
<td>80.6968</td>
<td>11.4400</td>
</tr>
</tbody>
</table>

![Correlation Coefficient](image_url)

Fig. 1. R2 Score for Machine Learning Approaches
Fig. 2. Machine Learning Models with MAE and RMSE

Fig. 3. Machine Learning Models with RAE (%) and RRSE (%)

Fig. 4. Machine Learning Models and its Time Taken to Build the Model (Seconds)
### Table 3: Machine Learning Models with REP Tree

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Correlation coefficient</th>
<th>MAE</th>
<th>RMSE</th>
<th>RAE</th>
<th>RRSE</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.6155</td>
<td>14.1062</td>
<td>17.7570</td>
<td>73.9530</td>
<td>78.8609</td>
<td>2.0300</td>
</tr>
<tr>
<td>hypertension</td>
<td>0.2860</td>
<td>0.1244</td>
<td>0.2529</td>
<td>89.8174</td>
<td>96.0877</td>
<td>1.3000</td>
</tr>
<tr>
<td>heart_disease</td>
<td>0.2776</td>
<td>0.0681</td>
<td>0.1878</td>
<td>89.9223</td>
<td>96.5061</td>
<td>1.2400</td>
</tr>
<tr>
<td>bmi</td>
<td>0.4776</td>
<td>4.1701</td>
<td>5.8595</td>
<td>94.3502</td>
<td>88.2876</td>
<td>1.8400</td>
</tr>
<tr>
<td>HbA1c_level</td>
<td>0.3855</td>
<td>0.8661</td>
<td>0.9885</td>
<td>97.7138</td>
<td>92.3247</td>
<td>2.8500</td>
</tr>
<tr>
<td>blood_glucose_level</td>
<td>0.4046</td>
<td>30.8231</td>
<td>37.2464</td>
<td>99.8442</td>
<td>91.4956</td>
<td>2.9100</td>
</tr>
<tr>
<td>diabetes</td>
<td>0.8296</td>
<td>0.0460</td>
<td>0.1558</td>
<td>29.5894</td>
<td>55.8757</td>
<td>1.4700</td>
</tr>
</tbody>
</table>

**Fig. 5.** R2 Score for Machine Learning Approaches

**Fig. 6.** Machine Learning Models with MAE and RMSE
4. Results and Discussion

Table 1 explains 8 parameters with one class, which includes different categories of data like gender, age, hypertension, heart disease, smoking history, BMI, HbA1c_level, and blood_glucose_level. Based on the dataset, it is evident that different machine learning decision tree approaches are used to find the hidden patterns and which is the best or influencing parameter to decide future predictions. Related results and numerical illustrations are shown between Table 1 to Table 3 and Figure 1 to Figure 8. They are based on Equation 1, Table 2, and Figure 1, which is used to find the R² score or correlation coefficient by comparing 9 parameters. Numerical illustrations suggest that there may be a significant difference from one parameter to another. In this case, using the linear regression modeling approach, the age and class label diabetes return a strong positive correlation. These two parameters are essential for deciding whether the patient is affected by diabetes and predicting the future outcome. A similar analysis conducted used REPtree, which is used to find the diabetes predictions. In this case, the age and diabetes parameters only return positive correlations. Similar results are shown in Table 3 and Figure 8.

Further data analysis revealed a gradual improvement in test scores over time. The MAE is used to find model errors using Equations 2. The linear regression and REP tree return a maximum except for using age and blood_glucose_level. Similar results and discussion are shown in table 2, figure 3, table 3, and figure 6.
The RMSE (root mean square error) measures the difference between predicted and actual values using Equation 3. In this case, the linear regression and REP tree return a minimum error hypertension, heart disease, smoking history, BMI, and HbA1c_level. RAE and RRSE also produce the same error. The related numerical illustration is shown in Table 2, table 3, figure 3 and figure 7.

Time taken is one of the significant tasks in machine-learning approaches. Based on Table 2 and Figure 4, for using a linear regression model, all the parameters take less time to build the model except class label diabetes. Another ML approach REPtree, was used to analyze the dataset, in this case, HbA1c_level and blood_glucose_level taking more time to build the model. The results and discussions are included in Table 3 and Figure 8.

5. Conclusion
Further research
The findings presented in this study contribute to our understanding that age and diabetes parameters return robust positive correlations. Future studies can build upon these, finding the suitable variable for future prediction with increased diabetes-related parameters and also increasing accuracy level using different machine learning and decision tree approaches.

Reference