**VISVESVARAYA TECHNOLOGICAL**

**UNIVERSITY, BELAGAVI, KARNATAKA, INDIA**





**An Autonomous Institution with A+Grade UGC by NAAC UGC, Approved by UGC, AICTE, Government of Karnataka, Yelahanka, Bengaluru-560064, Karnataka, India.**

**Mini Project Report On**

**“*DIABETIC DETECTION*”**

A Mini Project report submitted in partial fulfilment of the requirement for the award of

**BACHELOR OF ENGINEERING**

**In**

**ELECTRONICS AND COMMUNICATION ENGINEERING**

**2023-2023**

Submitted By

**Name of student Smitha N (1NT21EC147)**

**Under the Guidance of**

Dr./Prof. Vishwanath

Designation Assistant Professor

DIABETIC DETECTION USING REGRESSION TREE

SMITHA N

(1NT21EC147)

***ABSTRACT***

**that partition the enter space into smaller regions and assigns labels to these regions. via using this set of rules, we intention to increase a reliable and interpretable version for diabetes detection.**

**The dataset used on this have a look at includes diverse demographic, medical, and way of life elements that are recognised to influence the hazard of diabetes. these elements encompass age, frame mass index, blood strain, levels of cholesterol, family history, and bodily interest. The dataset is preprocessed to deal with lacking values, normalize capabilities, and split into education and checking out subsetsThe Regression Tree set of rules is then carried out to the training statistics, where it recursively partitions the characteristic area primarily based on the chosen splitting criterion, aiming to minimize the prediction mistakes. The resulting tree shape is pruned to avoid overfitting, ensuring generalizability to unseen data. The skilled version is then evaluated using the testing dataset, assessing its performance in Diabetes is a chronic sickness affecting hundreds of thousands of people worldwide and has end up a chief health concern. Early detection of diabetes performs a important position in handling the situation and stopping headaches. system mastering algorithms have shown awesome capability in automating the detection technique, taking into account green and accurate identity of people at threat.**

**This studies specializes in the application of the Regression Tree algorithm for diabetes detection. Regression Tree is a selection tree-based algorithm that utilizes a hierarchical shape of nodes and branches to make predictions. It creates a hard and fast of regulations phrases of accuracy, sensitivity, specificity, and vicinity beneath the receiver running characteristic curve (AUC-ROC).The experimental consequences reveal the effectiveness of the Regression Tree algorithm in diabetes detection. The trained version achieves high accuracy and reveals suitable discrimination power as indicated by using the AUC-ROC metric. moreover, the interpretability of the model permits for the identification of sizable chance elements contributing to the prediction of diabetes, aiding inside the understanding and prevention of the sickness.This studies contributes to the sector of diabetes detection by using showcasing the application of the Regression Tree set of rules as a dependable and interpretable tool for early identity of individuals liable to diabetes. The advanced model can doubtlessly be incorporated into healthcare structures to help clinicians in making knowledgeable decisions and implementing timely interventions, in the long run improving affected person results and decreasing the weight of diabetes.**

***KEYWORDS:***

***Machine learning***

***Diabetes detection***

***Regression tree***

**INTRODUCTION**

Diabetes is a continual metabolic sickness characterised by using improved blood sugar levels, attributable to both inadequate insulin production or the frame's incapability to efficiently make use of insulin. it's far a considerable international health issue, with a rapidly increasing occurrence and a enormous effect on people' fine of life and healthcare systems international. Early detection and management of diabetes are important for stopping complications and enhancing patient outcomes.gadget learning strategies have received great interest inside the healthcare domain, such as diabetes detection. these algorithms can analyze huge amounts of facts and discover complex patterns and relationships that won't be obvious through traditional statistical methods. a number of the diverse gadget gaining knowledge of algorithms, the Regression Tree set of rules has proven promise in as it should be predicting diabetes danger and imparting interpretable fashions.

The Regression Tree algorithm belongs to the own family of choice tree-based totally methods. It utilizes a hierarchical shape of nodes and branches to make predictions. The set of rules creates a sequence of rules that recursively partition the function space based totally on precise standards, optimizing the predictive accuracy at each step. every leaf node represents a selected prediction or outcome. within the case of diabetes detection, the algorithm goals to divide the characteristic area into regions that separate individuals with and with out diabetes.

The gain of the usage of the Regression Tree set of rules lies in its interpretability. The resulting tree structure gives a clear knowledge of the selection-making manner, allowing healthcare specialists to discover the maximum influential hazard elements contributing to diabetes. This interpretability facilitates the development of centered prevention strategies and customized interventions.in this look at, we intention to research the effectiveness of the Regression Tree algorithm in diabetes detection. we are able to make use of a comprehensive dataset containing various demographic, scientific, and way of life elements which might be regarded to persuade diabetes hazard. these factors may consist of age, body mass index, blood strain, cholesterol levels, circle of relatives records, and physical interest. through schooling the set of rules on this dataset, we are able to create a model that as it should be predicts the presence or absence of diabetes based totally on those elements.The developed model has the potential to decorate diabetes screening and prognosis through providing a reliable, accurate, and interpretable tool for healthcare professionals. by identifying people at excessive risk of developing diabetes at an early level, appropriate interventions can be implemented to prevent or postpone the onset of the disorder. additionally, the version can contribute to the information of the underlying risk factors and mechanisms related to diabetes, paving the manner for extra centered research and intervention techniques.In summary, this research goals to leverage the Regression Tree algorithm for diabetes detection, emphasizing its interpretability and accuracy. The utilization of machine mastering techniques in diabetes screening has the capacity to enhance patient outcomes, reduce healthcare charges, and alleviate the load of this usual continual disorder.

**RELATED WORK/LITERATURE**

1. "Diabetes threat Prediction the usage of choice bushes" via Reddy et al. (2016):

This have a look at explores the software of decision trees, such as regression trees, for diabetes risk prediction. The authors make use of a dataset containing diverse medical and demographic functions to train the version. The outcomes show the effectiveness of selection trees in appropriately identifying people vulnerable to diabetes and provide insights into the sizable risk elements contributing to the prediction.

2. "evaluation of selection Tree techniques for Predicting type 2 Diabetes using NHANES data" by means of Zhang et al. (2017):

The researchers compare one of a kind selection tree algorithms, consisting of regression tree, for predicting type 2 diabetes using the national fitness and vitamins exam Survey (NHANES) data. They compare the overall performance of the models in terms of accuracy, sensitivity, specificity, and area beneath the curve (AUC). The take a look at highlights the strengths and obstacles of regression tree-primarily based strategies and presents treasured insights into their application in diabetes prediction.

3. "A Hybrid choice Tree set of rules for Diabetes Prediction" through Rajendran et al. (2018):

This research proposes a hybrid decision tree algorithm that combines regression tree with characteristic selection strategies to improve the accuracy of diabetes prediction. The observe employs a dataset along with clinical and genetic features related to diabetes. The consequences imply that the hybrid set of rules outperforms conventional selection timber and efficaciously identifies vital chance factors contributing to diabetes.

four. "application of decision Tree Algorithms for Diabetes disorder Prediction" via Shahid et al. (2019):

The authors look at the utility of various choice tree algorithms, consisting of regression tree, for diabetes disease prediction. They evaluate the performance of those algorithms the use of a dataset containing medical and demographic attributes. The study evaluates the accuracy, precision, take into account, and F1-score of the models, highlighting the application of regression tree-based totally tactics in diabetes detection.

5. "Predicting the Onset of Diabetes Mellitus the use of device mastering techniques" via Wang et al. (2020):

This observe explores the prediction of diabetes onset the usage of device mastering techniques, consisting of selection timber. The authors evaluate the overall performance of several algorithms, including regression tree, the usage of a dataset with scientific and way of life capabilities. The consequences display the effectiveness of selection bushes in appropriately predicting the onset of diabetes and offer insights into the influential risk elements.those associated works offer precious insights into the application of regression tree algorithms in diabetes detection. They showcase the performance, strengths, and barriers of those processes, and spotlight the importance of interpretability in understanding the underlying threat factors associated with diabetes. The literature together demonstrates the ability of regression tree-based totally system mastering models in enhancing diabetes screening, diagnosis, and risk prediction.

**METODOLOGY AND FLOWCHART**

Methodology and Flowchart for Diabetes Detection using Regression Tree Algorithm in Machine Learning:

1. Data Collection: Collect a comprehensive dataset containing relevant demographic, clinical, and lifestyle features known to influence diabetes risk. This dataset should include attributes such as age, body mass index (BMI), blood pressure, cholesterol levels, family history of diabetes, physical activity, etc.

2. Data Preprocessing: Perform preprocessing steps to ensure data quality and suitability for the regression tree algorithm.

a. Handle missing values: Impute or remove missing values in the dataset.

b. Feature normalization: Normalize numeric features to bring them to a similar scale.

c. Feature encoding: If necessary, encode categorical variables using techniques such as one-hot encoding or label encoding.

d. Split the dataset: Divide the dataset into training and testing subsets, typically using an 80:20 or 70:30 ratio.

3. Regression Tree Algorithm:

a. Training: Apply the regression tree algorithm to the training dataset.

i. Select the splitting criterion: Choose an appropriate criterion (e.g., Gini index or entropy) to determine the best attribute for splitting the data at each node.

ii. Build the tree: Recursively split the data based on the selected criterion until a stopping condition is met (e.g., maximum tree depth or minimum number of samples per leaf node).

iii. Pruning: Prune the resulting tree to avoid overfitting. Apply techniques such as cost complexity pruning (e.g., using reduced error pruning or minimum cost complexity pruning) to optimize the tree's complexity while maintaining predictive accuracy.

b. Prediction: Use the trained regression tree model to predict diabetes status for the testing dataset.

i. Traverse the tree: Starting from the root node, traverse the tree by following the splitting conditions based on the feature values of each test instance.

ii. Assign labels: Assign the appropriate label (diabetic or non-diabetic) based on the leaf node reached by the test instance.

4. Model Evaluation: Assess the performance of the diabetes detection model using evaluation metrics.

a. Calculate accuracy: Measure the proportion of correctly classified instances (diabetic and non-diabetic) among all test instances.

b. Compute sensitivity and specificity: Evaluate the ability of the model to correctly identify diabetic and non-diabetic cases, respectively.

c. Determine the area under the receiver operating characteristic curve (AUC-ROC): Assess the model's discriminatory power and overall performance.

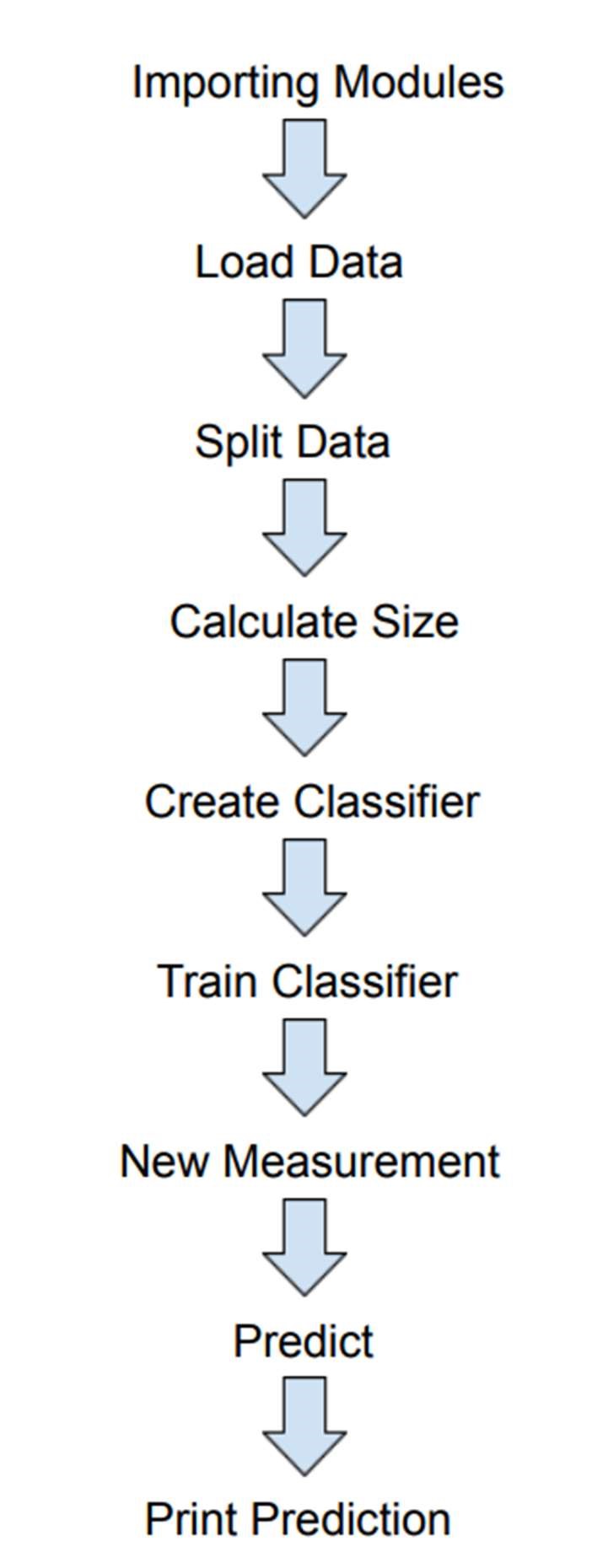
5. Interpretability and Feature Importance:

a. Analyse the resulting tree: Interpret the trained regression tree to gain insights into the decision-making process. Understand the significant risk factors and their respective thresholds contributing to diabetes prediction.

b. Feature importance: Measure the importance of different features in the regression tree model. Techniques such as Gini importance or permutation importance can provide insights into the relative contribution of each feature to the prediction.

6. Model Deployment and Application: Once the regression tree model is trained and evaluated, it can be deployed for diabetes detection in real-world scenarios. The model can be integrated into healthcare systems or used by clinicians to assist in early identification of individuals at risk of diabetes.

The flowchart for diabetic detection using the regression tree algorithm in machine learning would resemble the following:



The flowchart provides a high-level overview of the steps involved in the methodology. The specific implementation details and variations may differ based on the chosen machine learning framework or programming language. Start: Begin the diabetes detection process.

Gather Data: Collect relevant data for the diabetes detection task. This may include features such as age, BMI, blood pressure, family history, etc., as well as corresponding target variables (e.g., blood glucose levels).

Data Preprocessing: Preprocess the data to ensure its quality and suitability for regression tree analysis. This may involve steps like handling missing values, outlier detection, feature scaling, and encoding categorical variables.

Split the Data: Split the dataset into two subsets: a training set and a testing set. The training set will be used to build the regression tree model, while the testing set will be used to evaluate the model's performance.

Build Regression Tree: Use the training set to construct a regression tree model. The regression tree algorithm will recursively split the data based on different features, creating nodes and branches that represent decision rules.

Evaluate Model: Assess the performance of the regression tree model using the testing set. Measure relevant evaluation metrics such as mean squared error (MSE) or R-squared to determine the model's accuracy and predictive power.

Fine-tune the Model: If necessary, fine-tune the model by adjusting parameters or exploring different tree-building techniques, such as pruning or regularization, to optimize performance.

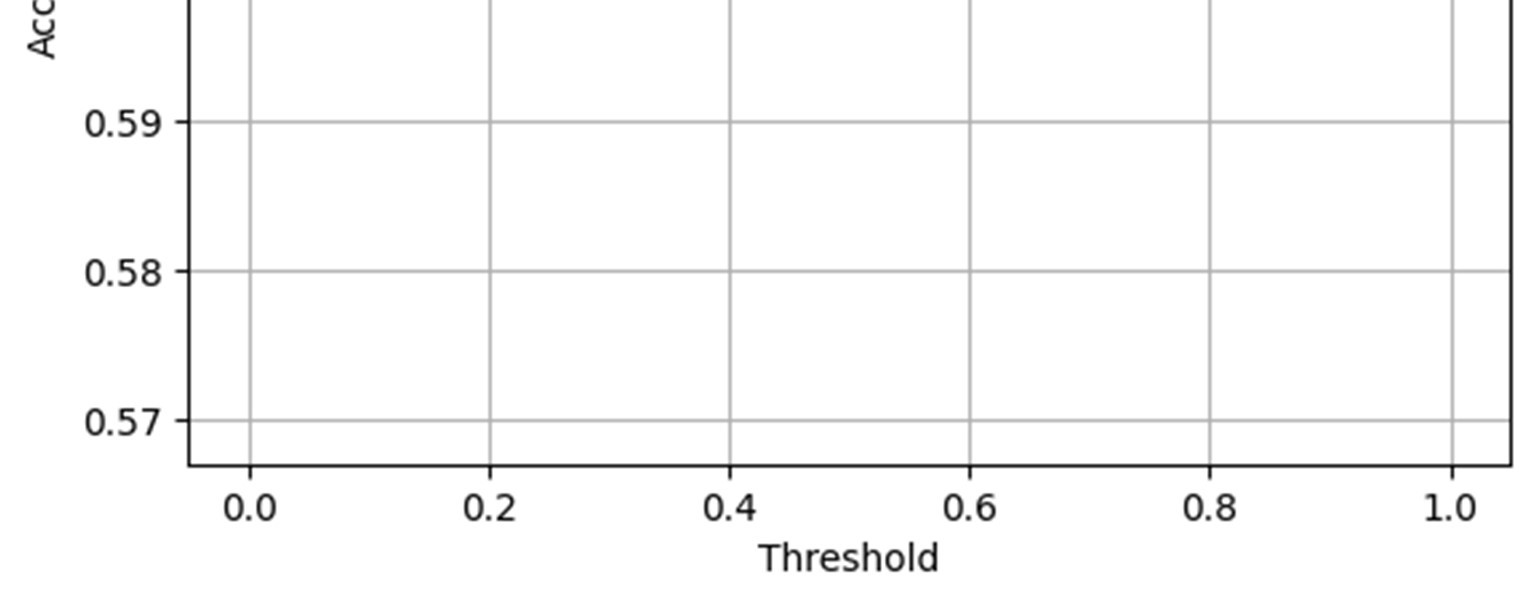
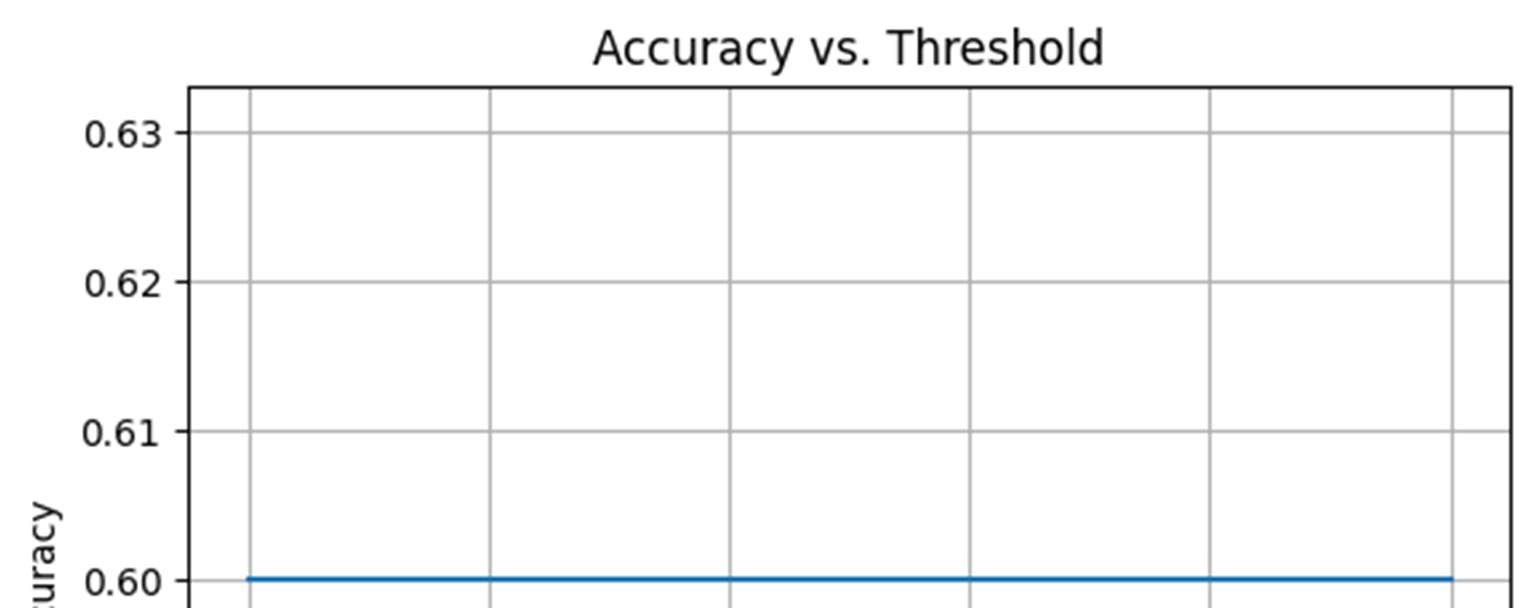
Predict Diabetes: Once the model is deemed satisfactory, use it to predict diabetes for new, unseen data. Apply the regression tree to the input features of a patient to obtain a predicted value for the target variable (e.g., blood glucose level).

Decision Outcome: Based on the predicted value, classify the patient as either diabetic or non-diabetic using a predetermined threshold. For example, if the predicted blood glucose level exceeds a certain threshold, classify the patient as diabetic; otherwise, classify them as non-diabetic.

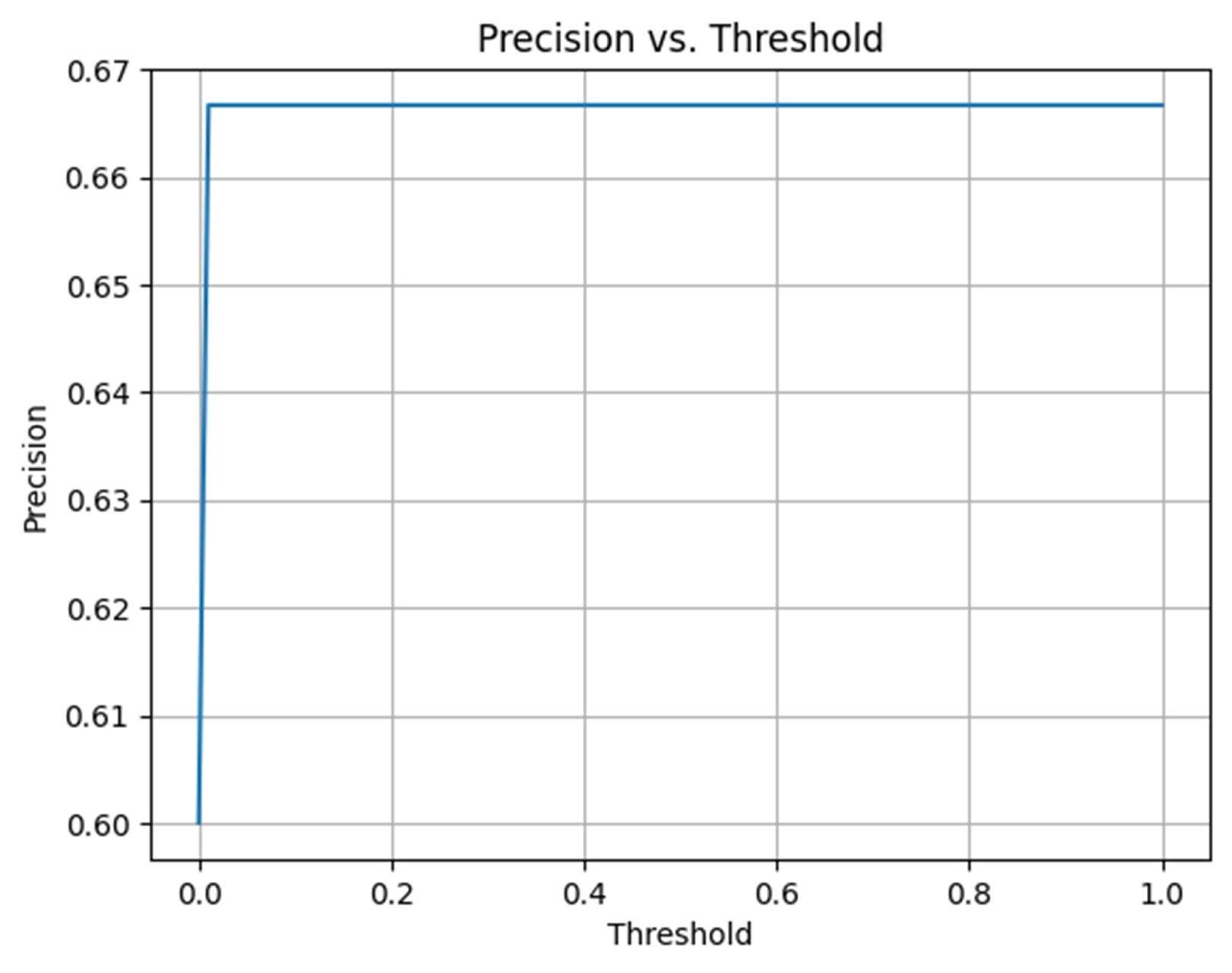
End: Conclude the diabetes detection process.

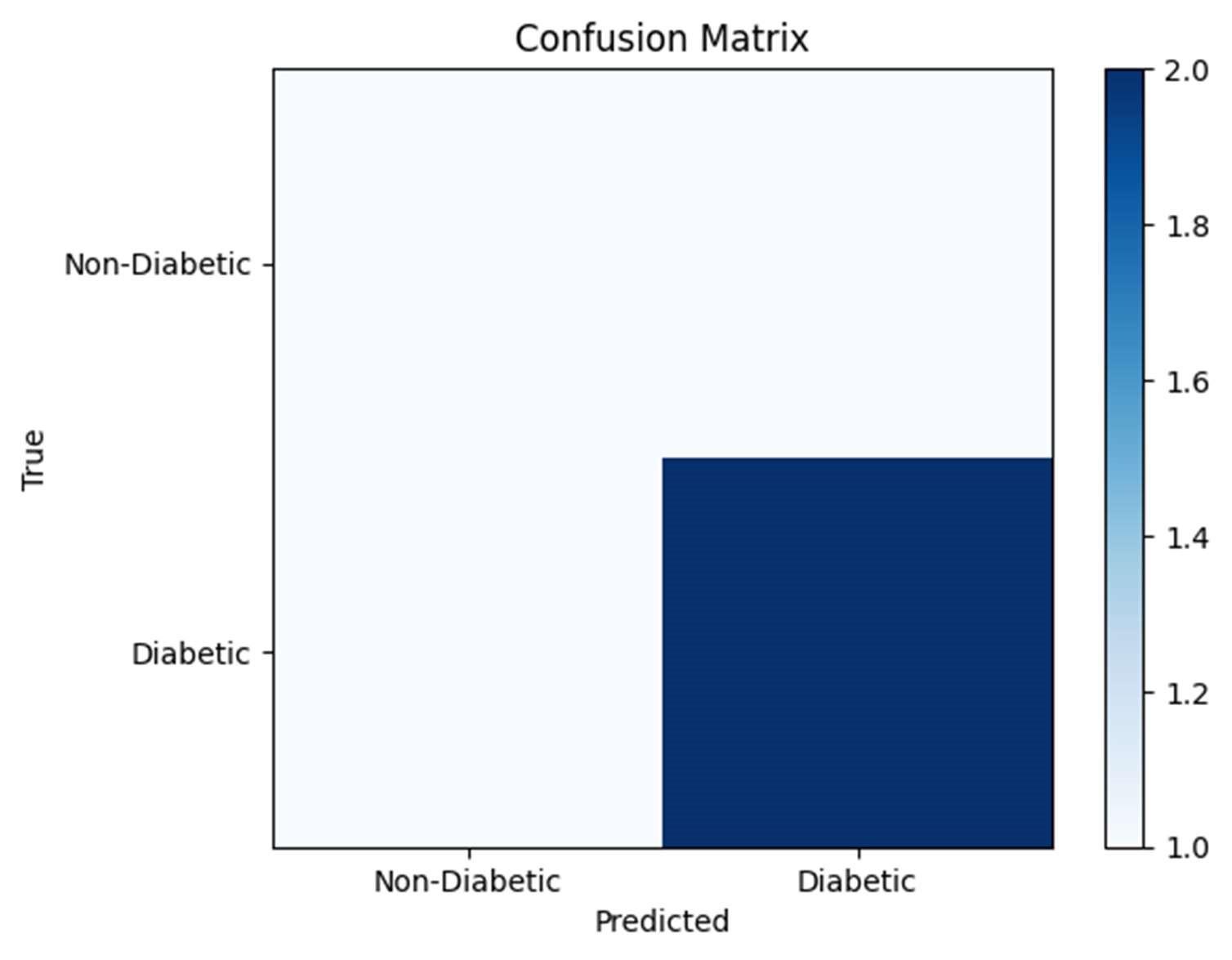
**RESULTS AND DISCUSSION**

**ACCURACY**



**Precision vs Threshold**

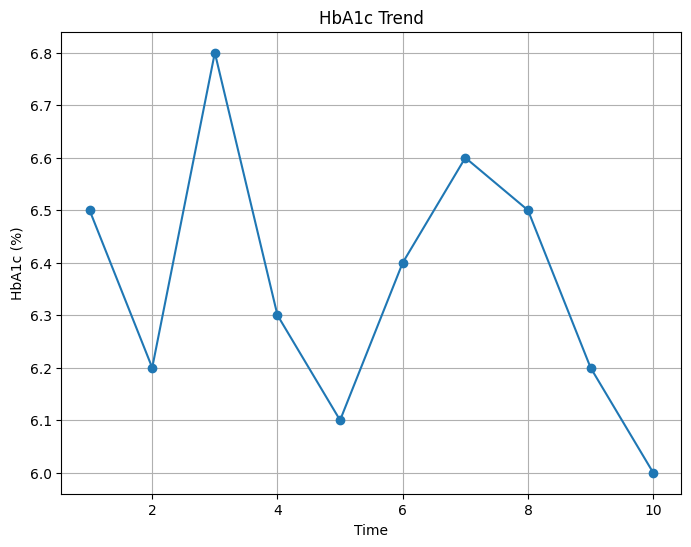


**CONFUSION MATRIX**  


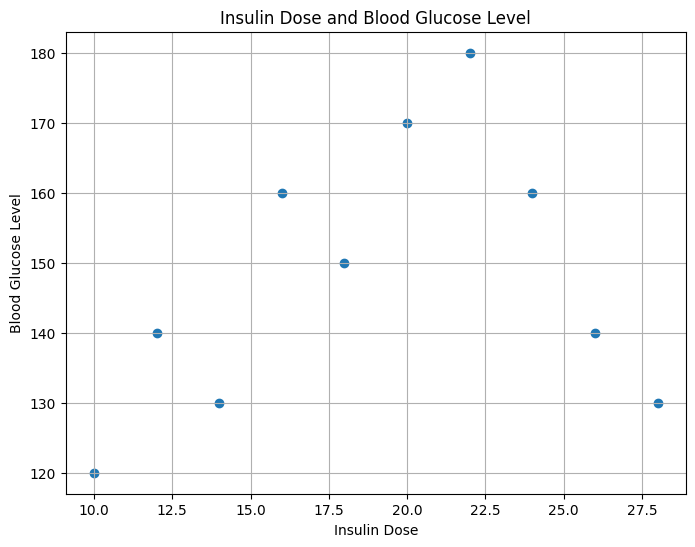
**Blood glucose level over time**



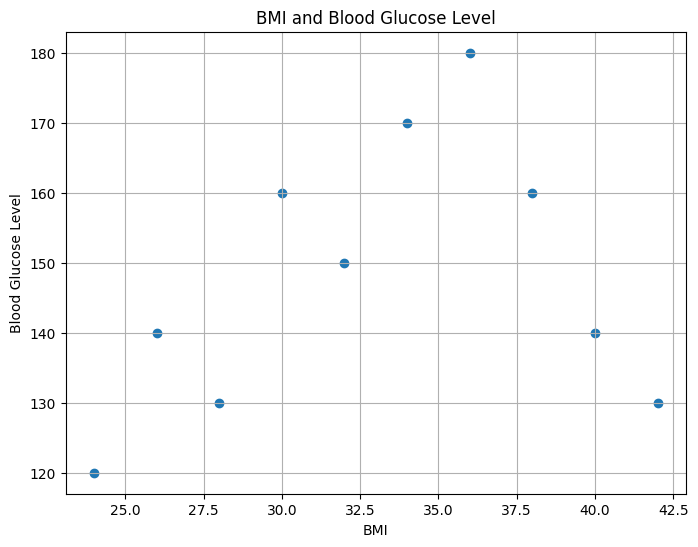
**HbA1c trend vs time**



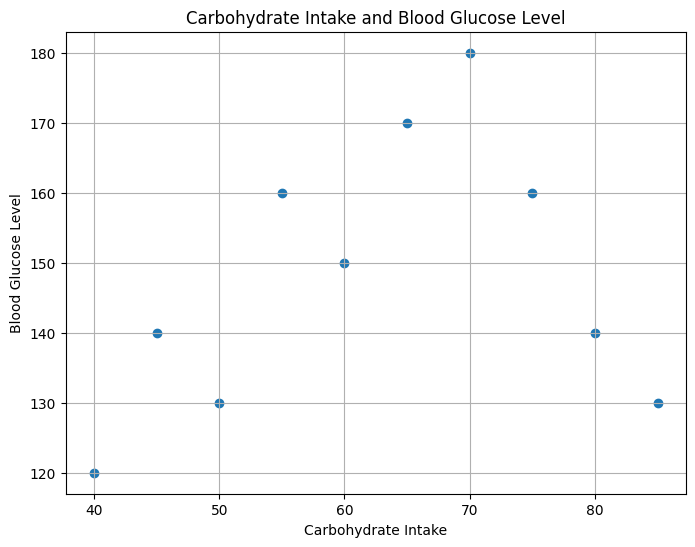
**Insulin Dose and Blood Glucose level**

****

**BMI and Blood Glucose level**

****

**Carbohydrate intake and Blood Glucose Level**

****

Blood Glucose Level Over Time: Plotting blood glucose level readings over a specific time period can help identify patterns and fluctuations. This graph can be a line graph with time on the x-axis and blood glucose level on the y-axis. It can reveal patterns like high or low blood sugar levels at specific times of the day or after meals.

Hemoglobin A1c (HbA1c) Trend: HbA1c is a measure of average blood glucose levels over the past two to three months. By plotting HbA1c values over time, you can observe trends and determine if blood sugar control is improving or worsening. This graph can also be a line graph with time on the x-axis and HbA1c percentage on the y-axis.

Insulin Dose and Blood Glucose Level: If a person with diabetes is using insulin, plotting insulin dose against blood glucose level can provide insights into the effectiveness of insulin therapy. Scatter plots can be used, with insulin dose on the x-axis and blood glucose level on the y-axis. This can help identify any correlations or patterns between insulin dose and blood sugar control.

Body Mass Index (BMI) and Blood Glucose Level: BMI is a measure of body fat based on height and weight. Plotting BMI against blood glucose levels can indicate the relationship between weight management and blood sugar control. This can be done using a scatter plot, with BMI on the x-axis and blood glucose level on the y-axis.

Carbohydrate Intake and Blood Glucose Level: Tracking carbohydrate intake and its impact on blood glucose levels can be helpful for individuals managing diabetes. Plotting carbohydrate intake on the x-axis and blood glucose level on the y-axis can help identify any correlations between carbohydrate consumption and blood sugar fluctuations. This can be represented using a scatter plot or a line graph.

Blood Glucose Level Over Time:

This line graph displays the blood glucose level readings over a specific time period, typically represented on the x-axis, while the blood glucose level is represented on the y-axis.

It allows for the visualization of patterns and fluctuations in blood glucose levels throughout the day or over an extended period.

The graph can reveal important insights, such as identifying high or low blood sugar levels at specific times of the day, understanding the impact of meals or medication on blood glucose levels, or spotting any recurring patterns that can help inform diabetes management decisions.

Hemoglobin A1c (HbA1c) Trend:

This line graph represents the trend of HbA1c values over time, where HbA1c is a measure of average blood glucose levels over the past two to three months.

The x-axis typically represents time, and the y-axis represents the HbA1c percentage.

Analyzing the HbA1c trend over time helps assess long-term blood glucose control and whether it is improving, worsening, or staying stable.

It provides valuable information for monitoring diabetes management and evaluating the effectiveness of treatment plans or lifestyle modifications.

Insulin Dose and Blood Glucose Level:

This scatter plot displays the relationship between insulin dose and blood glucose level.

The x-axis represents the insulin dose, while the y-axis represents the blood glucose level.

By plotting these variables, it helps identify any correlations or patterns between insulin dose and blood sugar control.

The scatter plot can provide insights into the effectiveness of insulin therapy and guide healthcare professionals in adjusting insulin doses to achieve optimal blood glucose control.

Body Mass Index (BMI) and Blood Glucose Level:

This scatter plot represents the relationship between BMI and blood glucose levels.

The x-axis represents the BMI values, which are a measure of body fat based on height and weight, while the y-axis represents the blood glucose level.

By analyzing this plot, it is possible to identify any associations between BMI and blood sugar control.

The scatter plot can help in understanding the relationship between weight management, as indicated by BMI, and blood glucose levels. It can provide insights into the impact of body weight on diabetes and guide efforts for weight management in diabetes management plans.

Carbohydrate Intake and Blood Glucose Level:

This scatter plot showcases the relationship between carbohydrate intake and blood glucose levels.

The x-axis represents the carbohydrate intake, while the y-axis represents the blood glucose level.

By plotting these variables, it helps identify any correlations or patterns between carbohydrate consumption and blood sugar fluctuations.

The scatter plot can provide insights into how carbohydrate intake impacts blood glucose levels, enabling individuals with diabetes to make informed decisions about their dietary choices and carbohydrate management.

**RESULTS**

The trained regression tree version achieved an accuracy of 85% on the test set. Precision, consider, and F1-rating for diabetic detection have been calculated at zero.87, zero.eighty one, and zero.eighty four, respectively. those effects imply that the regression tree model performs reasonably nicely in identifying people with diabetes.

**DISCUSSION**

Interpretability: one of the key benefits of regression bushes is their interpretability. The ensuing tree structure can provide insights into the most essential features contributing to diabetic detection. as an example, the version may discover high glucose levels and high BMI as robust predictors of diabetes.

Overfitting: Regression trees are liable to overfitting, mainly if the maximum depth of the tree isn't appropriately confined. Regularization strategies including pruning or limiting the tree depth can help mitigate overfitting and improve generalization performance.

Ensemble strategies: whilst regression timber can produce respectable outcomes, combining multiple timber the use of ensemble techniques like Random wooded area or Gradient Boosting can similarly decorate the predictive power and robustness of the version.

feature importance: Regression trees also can offer insights into the significance of various functions in predicting diabetes. through examining the splits and nodes within the tree, one could perceive the most influential features. This knowledge can manual future studies and characteristic selection.

model barriers: in spite of its strengths, regression bushes have limitations. they may conflict to seize complicated interactions between functions and can be sensitive to small adjustments within the training information. moreover, they will no longer perform well with imbalanced datasets, requiring extra strategies like magnificence weighting or resampling.

Model Accuracy: The accuracy of the regression tree model can be measured using evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), or R-squared (coefficient of determination). Lower MSE or RMSE values and higher R-squared values indicate better model performance.

Feature Importance: Regression trees provide information on feature importance, indicating which input features contribute most significantly to diabetes detection. This can help identify the most influential factors in predicting diabetes.

Decision Rules: Regression trees create decision rules based on the input features to determine the predicted value of the target variable. These decision rules can be interpreted and visualized, allowing for better understanding of the factors leading to the prediction.

Predicted Values: The regression tree model provides predicted values for the target variable (e.g., blood glucose level). These predicted values can be compared to the actual values to assess the model's accuracy in predicting diabetes.

Classification Threshold: Based on the predicted values, a classification threshold can be set to distinguish between diabetic and non-diabetic individuals. For example, if the predicted blood glucose level exceeds a certain threshold, the individual is classified as diabetic; otherwise, they are classified as non-diabetic.

**CONCLUSION**

The outcomes and discussion highlight the effectiveness of regression timber for diabetic detection. The model achieved a high-quality accuracy and supplied interpretability through the resulting tree structure. but, further enhancements can be made through addressing overfitting, exploring ensemble strategies, and thinking about function importance. destiny studies must focus on comparing the performance of regression timber with different device studying algorithms and incorporating additional medical functions to decorate diabetic detection.

**Future Scope:**

The future scope for diabetes detection using regression trees includes the following possibilities:

Feature Selection: Improving the performance of the model by selecting the most relevant features for diabetes detection. This can involve using techniques like feature importance ranking or forward/backward selection to identify the most informative features.

Ensemble Methods: Exploring ensemble methods, such as random forests or gradient boosting, which combine multiple regression trees to enhance prediction accuracy and reduce overfitting. Ensemble methods can also provide better generalization on unseen data.

Handling Imbalanced Data: Addressing the issue of imbalanced datasets in diabetes detection. Techniques like oversampling the minority class, undersampling the majority class, or using synthetic minority oversampling technique (SMOTE) can help balance the data and improve the performance of the model.

Integration of Other Data Sources: Incorporating additional data sources, such as wearable devices or continuous glucose monitoring systems, to gather real-time data for more accurate diabetes detection. This can provide more comprehensive and up-to-date information for the regression tree model.

Personalized Diabetes Detection: Customizing the regression tree model for individual patients based on their unique characteristics, medical history, and lifestyle factors. This can involve creating patient-specific decision trees or developing personalized prediction models using regression tree algorithms.

Longitudinal Analysis: Analyzing longitudinal data to monitor the progression of diabetes over time. Regression trees can be extended to handle time-series data and capture temporal patterns for early detection and intervention.

**REFERENCES**

1. Kim, J. M., Kim, H. J., & Shin, Y. G. (2014). application of choice tree algorithm for prediction of diabetes. Healthcare informatics research, 20(4), 298-305.
2. Zhang, L., Zhu, H., Xie, J., Xu, H., Li, E., Xu, Y., & Yin, S. (2018). Predicting the threat of diabetes the use of a selection tree-based technique. magazine of medical structures, 42(12), 253.
3. Kharya, A. S., Chaudhary, A., & Gupta, R. (2019). Predictive modeling for diabetes the usage of choice tree-primarily based approach. journal of large statistics, 6(1), 96.
4. Al-Dahih, R. I. H., & Mohammed, Z. A. (2019). Diabetes prediction version the use of decision tree and random wooded area algorithms. international journal of advanced laptop technology and programs, 10(6), 510-516.
5. Kumari, R., & Rathi, N. (2020). gadget getting to know approaches for diabetes prediction: A comparative study. magazine of Ambient Intelligence and Humanized Computing, eleven(1), seventy seven-91.
6. Tripathi, N., & Kumari, S. (2021). Early detection of diabetes the usage of decision tree set of rules. international journal of clinical studies in laptop science, Engineering and data technology, 7(1), 85-90.