# 

# 

# **Exploring Predictive Models for Diabetes Risk Assessment Using Data Mining Techniques**

## **Zeta**

| **Names** | **UIS Email ID** |
| --- | --- |
| **Bhavana Yadav Krishnappa** | **bkris2@uis.edu** |
| **Shreya Vishwas** | **shre2@uis.edu** |
| **Sohan Shylaja Mohan** | **sshyl2@uis.edu** |

1. **Abstract and Highlights**

The growing global incidence of diabetes demands precise predictive modeling and focused intervention techniques. This study meets this necessity by taking a holistic strategy that includes data analysis, model creation, performance evaluation, and actionable suggestions.

Diabetes prevalence is caused by a variety of demographic and lifestyle factors. Early detection and action are critical for mitigating its impact. The purpose for this study stems from the urgent need for prediction models capable of detecting persons at risk of developing diabetes, allowing for prompt treatments and preventive actions.

Data exploration and preparation included a thorough evaluation of the Behavioral Risk Factor Surveillance System (BRFSS) dataset to ensure data integrity and find underlying trends and relationships. Clustering approaches helped to identify major risk factors related with diabetes development, establishing the framework for predictive modeling.

Model development used three data mining techniques—regression, neural network, and decision tree—to predict diabetes based on demographic and health-related information. The best-performing models were chosen based on rigorous evaluation criteria, and they underwent intensive training, validation, and testing to ensure robustness and generalizability.

Performance evaluation indicators, such as misclassification rates and average square error, offered information about the models' predicted accuracy and consistency. The chosen model, Decision Tree B3D6, outperformed the baseline misclassification rates and produced excellent accuracy in scoring on fresh data.

The study's findings led to practical recommendations, including targeted interventions, lifestyle modification programs, health education initiatives, improved access to healthcare services, cardiovascular health monitoring, age-specific interventions, socioeconomic support programs, multidisciplinary approaches, continuous monitoring, evaluation, and community engagement.

Finally, this study provides important insights into diabetes prediction and risk mitigation techniques, laying the groundwork for evidence-based treatments to combat the expanding diabetes epidemic and enhance public health outcomes.

1. **Problem Description**

Diabetes is a chronic metabolic condition defined by either insufficient insulin synthesis or poor insulin utilization, which leads to high blood glucose levels. Diabetes prevalence has increased dramatically worldwide, from 108 million in 1980 to 422 million in 2014(World Health Organization: WHO & World Health Organization: WHO, 2023). This increase is especially noticeable in low- and middle-income nations, emphasizing the critical need for comprehensive worldwide intervention measures (“Economic Costs of Diabetes in the U.S. in 2007,” 2008).

The Behavioral Risk Factor Surveillance System (BRFSS) dataset is a useful repository of information, covering demographic statistics and health-related activities of individuals from diverse locations(*Diabetes Health Indicators Dataset*, 2021). Leveraging this dataset opens up the possibility of using data-driven approaches to better understand and forecast diabetes risks.

The purpose of this study is to anticipate an individual's probability of having diabetes based on these complex parameters by leveraging machine learning techniques and predictive modeling. The research uses data analytics to examine the dataset in order to uncover patterns, correlations, and risk factors that traditional statistical methods may miss. The use of machine learning techniques is consistent with current research procedures, allowing for a nuanced examination of the intricate interplay of variables.

Understanding the predictive power of demographic and health-related indicators in diabetes not only helps with early detection but also informs focused prevention efforts. The project's emphasis on the BRFSS dataset(*Diabetes Health Indicators Dataset*, 2021) puts it at the vanguard of data-driven research, with the potential to make significant contributions to the fields of public health and diabetes management.

The importance of treating diabetes as a major public health concern is highlighted by a thorough literature review. The number of diabetics has grown over time, with a considerable increase in low- and middle-income countries(“Economic Costs of Diabetes in the U.S. in 2007,” 2008). Diabetes is associated with serious complications such as blindness, kidney failure, heart attacks, strokes, and lower limb amputation (Xie et al., 2019b). Furthermore, diabetes imposes a significant cost burden on healthcare systems, highlighting the critical need for effective preventative measures.

Research has shown that early detection and intervention can considerably reduce the impact of diabetes, making it critical to identify and understand the risk factors linked with its development (Herman et al., 2015; Simmons et al., 2017). Predictive modeling and exploratory analysis can provide useful information for preventative treatments and public health initiatives.

## ​​**Predictive Analysis**: Predict the likelihood of an individual having diabetes based on their demographic and health-related factors?

## **Exploratory Analysis**: What are the key risk factors associated with diabetes development among individuals surveyed in the BRFSS dataset?

**3. Data Exploration, Preparation, and Visualization**

Dataset Description:

Dataset link: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download>

The BRFSS dataset (*Diabetes Health Indicators Dataset*, 2021) examines the complex link between lifestyle choices, demographic characteristics, and the occurrence of diabetes using responses from over 400,000 Americans. The Target variable being “Diabetes\_012” showcasing three classes:

| Target variable class 0 | diabetes or only pregnancy |
| --- | --- |
| Target variable class 1 | prediabetes |
| Target variable class 2 | diabetes |

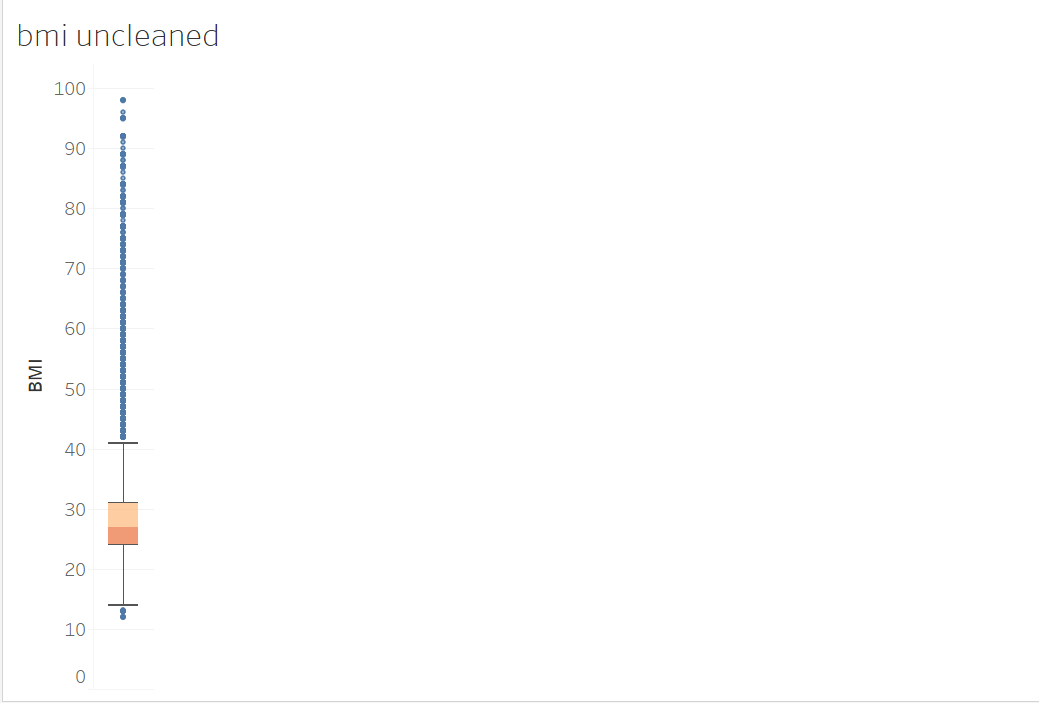
| **Input/Predictor Variable(s) Name** | **Variable Data Type** | **Description** |
| --- | --- | --- |
| HighBP | Binary | Indicates presence of high blood pressure (0 for no, 1 for yes). |
| DiffWalk | Binary | Indicates if there is any serious difficulty walking or climbing stairs? 0 = no 1 = yes |
| Education | Ordinal | Education level scale 1-6 indicates 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate) |
| HeartDiseaseorAttack | Binary | Indicates if coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes |
| HighChol | Binary | indicates the presence of high cholesterol (0 for no, 1 for yes). |
| BMI | Interval | Body Mass Index |
| Smoker | Binary | Indicates smoking status (0 for non-smoker, 1 for smoker) |
| PhysActivity | ordinal | indicates level of physical activity (0 for low, 1 for moderate, 2 for high) |
| Fruits | numeric | Consume Fruit 1 or more times per day 0 = no 1 = yes |
| Veggies | Binary | Consume Vegetables 1 or more times per day 0 = no 1 = yes |
| Age | Interval | represents age of the individual  13-level age category  1 = 18-24 9 = 60-64 13 = 80 or older |
| GenHlth | ordinal | Represents if individuals self rated general health on scale of 1 to 5  1 = excellent 2= very good 3 = good 4 = fair 5 = poor |
| HvyAlcoholConsumption | binary | Indicates if the individual is heavy drinker or not  Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes |
| Stroke | binary | Indicates of the individual ever had stroke.  0 = no 1 = yes |
| Sex | binary | Indicates indivdual’s sex  0 = female 1 = male |
| AnyHealthcare | Binary | Indicates if any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes |
| CholCheck | Binary | Indicates that 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years |
| Income | Ordinal | Income scale 1-8 indicates that 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more |
| MentHlth | Ordinal | Indicates about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days |
| NoDocbcCost | Binary | Indicates if there time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes |
| PhysHlth | Ordinal | Indicates your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days |

Utilizing the comprehensive features, we aim to analyze the data, answer the questions and understand the risk factors of diabetes.

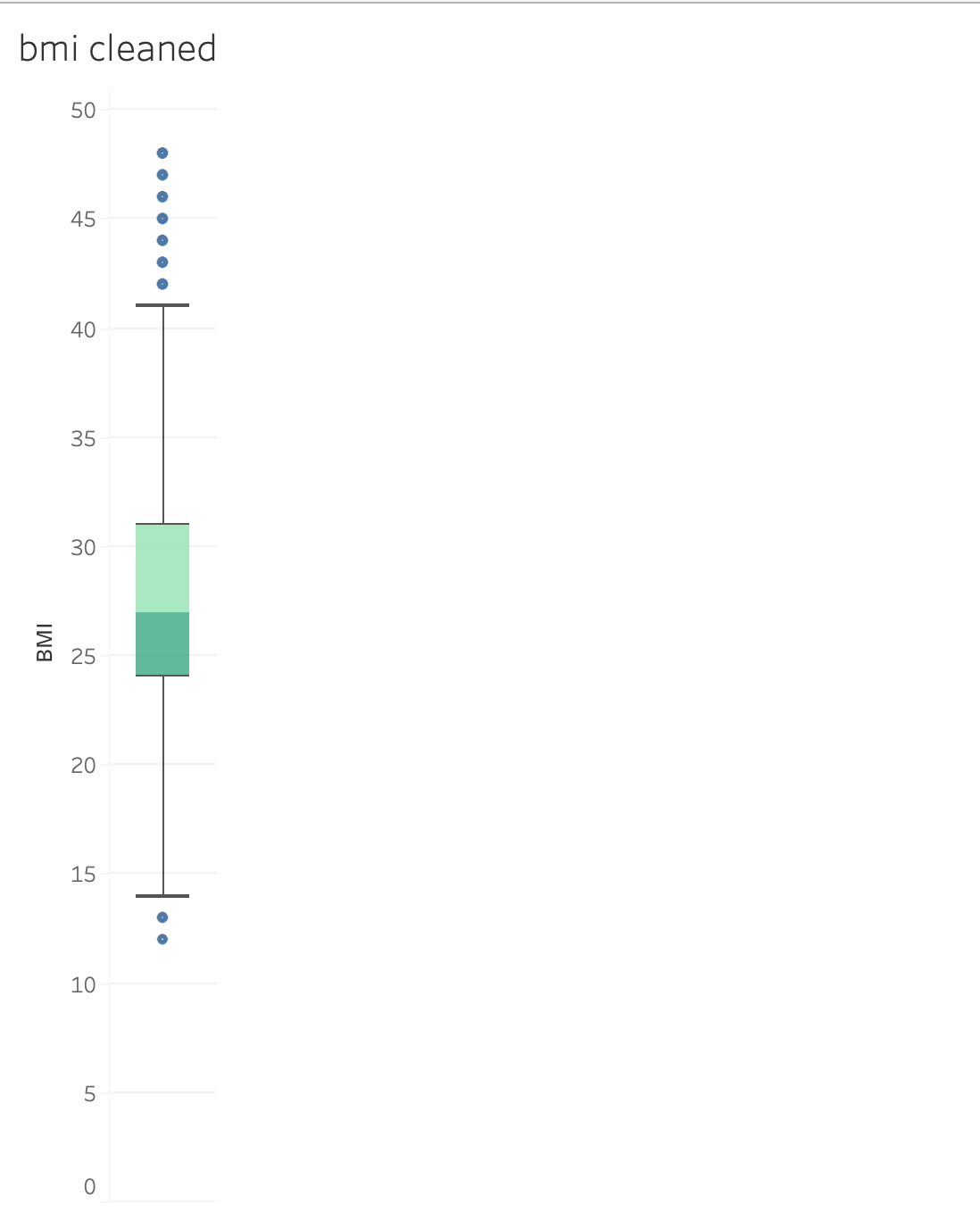
**Data Cleaning and Preparation:**

The majority of the data had binary values, and there were no missing values across the dataset. The only data cleaning we had to do was deal with outliers in the "BMI" feature. To find extreme values, we used Excel to calculate the Z-score and then remove them. This focused strategy made sure that there were no missing values in the dataset and that outliers in the "BMI" feature were handled correctly, laying the groundwork for further analysis.

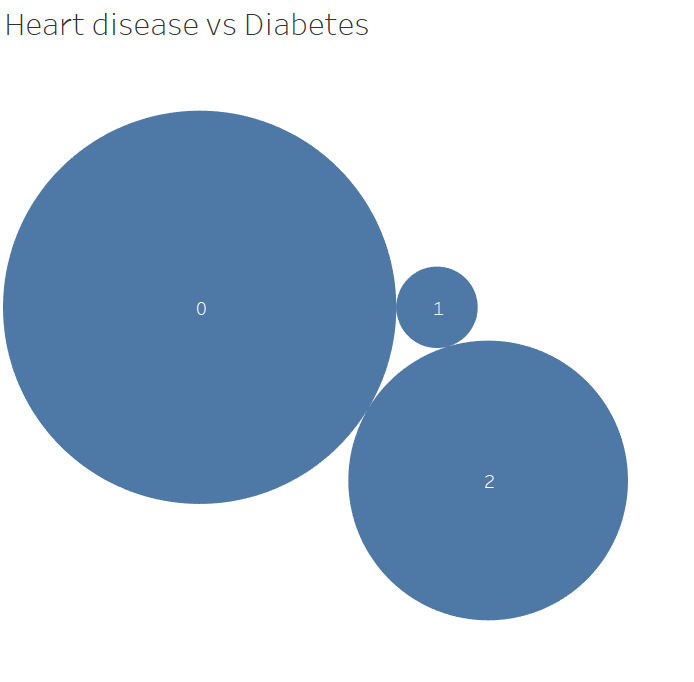
**Data Visualization to explore the data sets :**



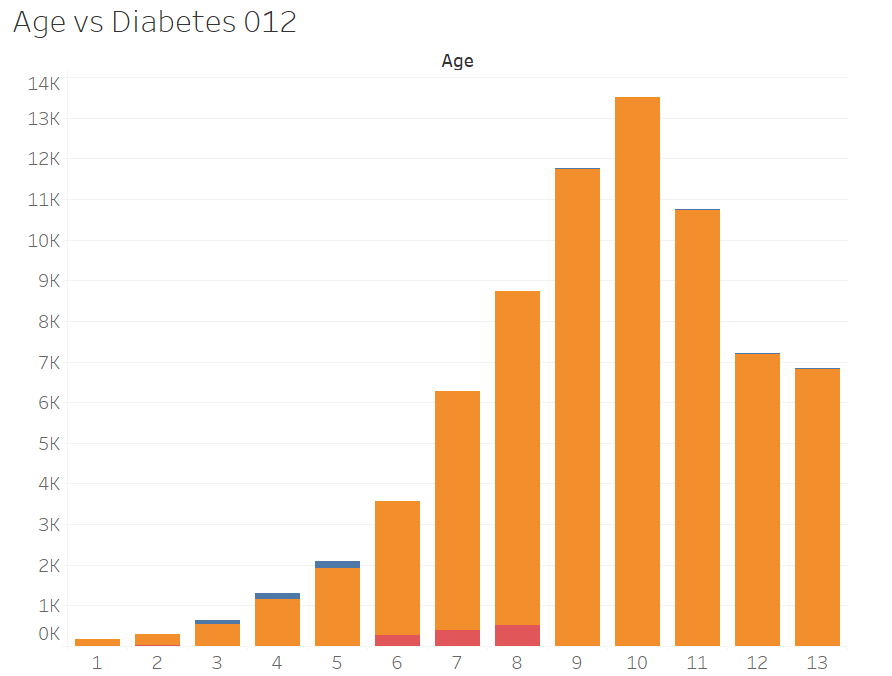
Box plots were used for this visualization because they are effective at representing numerical data distributions and finding outliers. In this scenario, the box plot is used to show the distribution of Body Mass Index (BMI) data. The box plot helps us comprehend the distribution of BMI values by providing a quick assessment of the central tendency, variability, and presence of outliers within the BMI dataset. Notably, outliers, particularly those more than 40 and less than 15, are easily identified and can be subject to additional inspection or data cleansing operations. This visualization helps to identify extreme values and potential data discrepancies, which contributes to the assessment of dataset quality and integrity. Finally, insights gained from the box plot help to guide further analytical stages and ensure the credibility of data-driven studies.



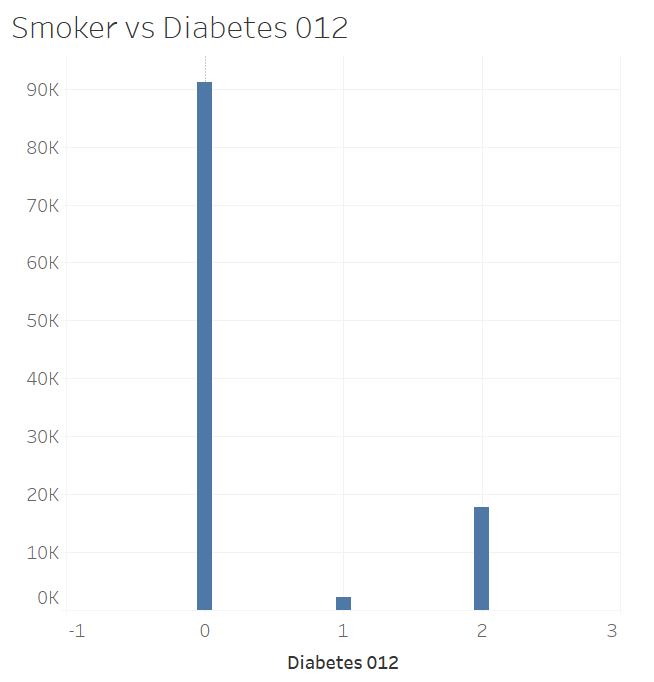
Box plots were chosen as the visualization tool for this investigation because they are effective at illustrating numerical data distributions and finding outliers. Specifically, the box plot is used to show the distribution of Body Mass Index (BMI) data after cleaning. This image provides a concise summary of the improved BMI dataset's central tendency, variability, and outlier presence. Notably, after the purification procedure, the number of outliers above approximately 40 and below 15 has decreased significantly, indicating that extreme and useless data points have been removed. This decrease improves the data's quality and relevance for analysis. Overall, the box plot clearly illustrates changes in BMI data distribution after cleaning, stressing the necessity of data refinement for better analytic results and data integrity.



Packed bubble charts are used to visually represent hierarchical data and compare proportions. They efficiently represent two categorical variables—Diabetes\_012 (three classes) and Heart Disease or Attack (two classes), with bubble size representing observation frequency. The picture demonstrates the link between diabetes and heart disease or attack. Larger bubbles represent more frequent occurrences, indicating that the majority of people without heart disease or an attack do not have diabetes (Sullivan et al., 2005). Prediabetes, on the other hand, has fewer people with heart disease or an attack, and the smallest bubble reflects those who have both. This image depicts the distribution of people across various combinations of diabetes and heart health status, which helps with pattern recognition and analysis.

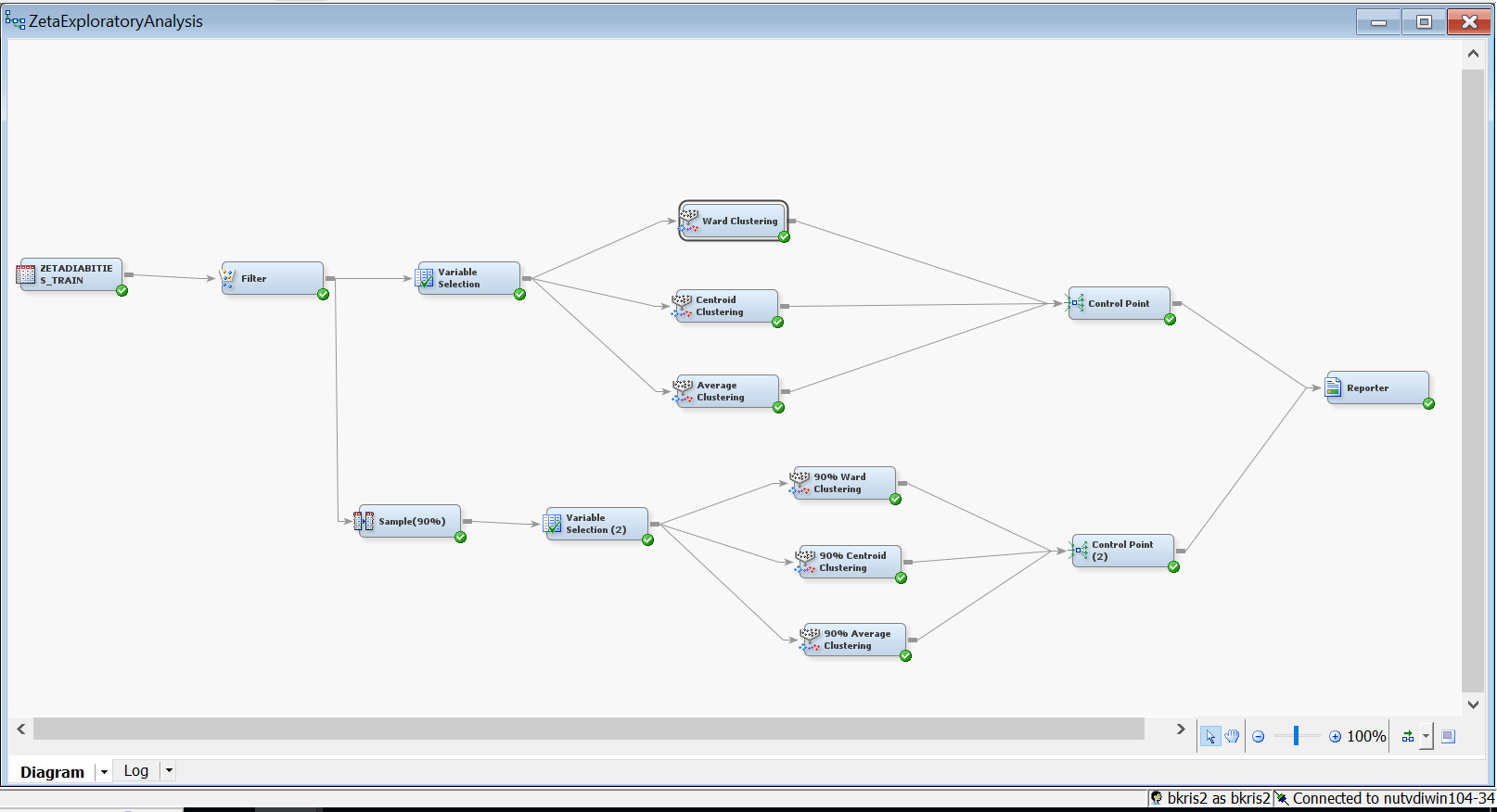


This visualization uses bar charts to efficiently display categorical data and allow for comparisons between age groups and diabetes prevalence. Each bar represents an age group, with the height reflecting the proportion of people in that age range who have diabetes. Stacked bars improve the comparison by visually representing different categories of diabetes prevalence within each age group. The graphical analysis shows a progressive increase in diabetes prevalence with age, peaking at 65-69, and then gradually declining in older age groups (Bellary et al., 2021). Diabetes is more common in those aged 65 to 69, compared to other age groups. Overall, this visualization allows for the detection of trends and patterns in diabetes prevalence across various age groups, providing valuable insights into the relationship between age and diabetes occurrence.

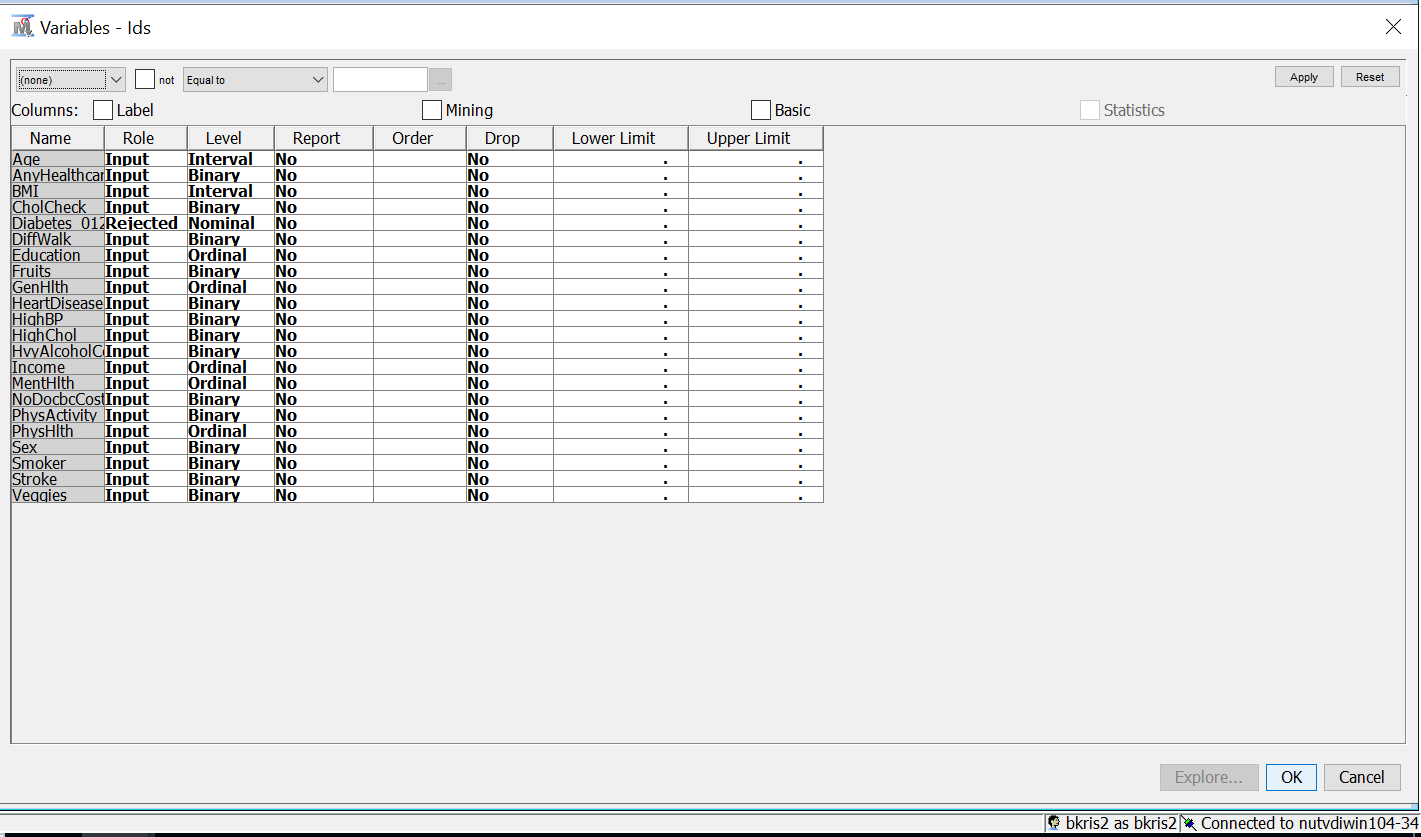


Bar charts are a useful tool for studying the association between smoking status and diabetes prevalence because they are good at showing categorical data and comparing proportions across different categories. Because each bar in this graphic represents a different category of smokers, it is easy to compare the prevalence of diabetes within each group. The data clearly shows that the majority of smokers are not diabetics, with pre-diabetic smokers and diabetic smokers following closely behind. Remarkably, compared to other groups, pre-diabetic smokers seem to have a lower prevalence of diabetes, suggesting a possible link between smoking and a lower risk of developing diabetes. The intermediate prevalence of diabetes among smokers with diabetes, however, suggests that the association between smoking and diabetes is not direct (Bellary et al., 2021).Overall, the bar chart does a good job of conveying these results, providing insightful information about how diabetes is distributed throughout various smoking categories and encouraging more research into the association between smoking and diabetes.

**4. Exploratory Analysis**



In order to identify major risk factors related with diabetes development in the BRFSS dataset (*Diabetes Health Indicators Dataset*, 2021), we used three different clustering techniques: Ward, Average, and Centroid clustering. Using unsupervised data mining approaches, where no predetermined goal variable exists, we identified the "Diabetes\_012" variable as "Rejected," recognizing its importance in our exploration. Similarly, all additional variables were allocated the job of "Input." This method enabled us to objectively investigate trends in the dataset without being swayed by predetermined outcomes. To test the stability of our found clusters, we applied the same clustering procedures on 90% of the sample data, ensuring that our findings were robust.By using this thorough methodology, we hoped to uncover intrinsic structures and correlations between variables, providing significant insights into the potential risk factors linked with diabetes development in the surveyed population.(Shmueli et al, 2018)



| Centroid Clustering | 90% Centroid Clustering |
| --- | --- |
| Number of Clusters : 2 | Number of Clusters : 4 |
| Result : | Result : |

The number of clusters before sampling in centroid clustering is 2 and after sampling is 4 which is within (N+/-2). The values of the mean statistics before and after sampling are similar so the centroid clusters are stable.

| Average Clustering | 90% Average Clustering |
| --- | --- |
| Number of Clusters : 4 | Number of Clusters : 5 |
| Result : | Result : |

The number of clusters before sampling in average clustering is 4 and after sampling is 5 which is within (N+/-2). The values of the mean statistics before and after sampling are similar so the centroid clusters are stable.

| Ward Clustering | 90% Ward Clustering |
| --- | --- |
| Number of Clusters : 3 | Number of clusters : 3 |
| Result : | Result : |

Both the above ward clusters are stable because the number of clusters before sampling and after sampling are same (3) which is within (N+/- 2 clusters) also the mean statistics data is similar before and after sampling. All the above clusters are stable but ward clustering is the only cluster with the same number of clusters before and after, so I am going to provide the answers using ward clustering.

**Cluster 1 analysis:**

Cluster 1 has distinguishing characteristics, depicting a population that, on average, faces physical health challenges, as evidenced by a higher mean value for variables such as DiffWalk (indicating difficulty walking) at 0.9646 and a significant prevalence of smokers with a mean value of 0.7601. Furthermore, this cluster had lower mean values for favorable health markers such as fruit consumption (0.0354) and physical activity (0.3020), indicating probable lifestyle-related health issues. The average age in this cluster is approximately 29.77, indicating a somewhat youthful demographic. Surprisingly, there is a significant prevalence of heavy alcohol intake (mean value: 0.5807) and a higher percentage having a history of stroke. Overall, this cluster represents a population with possible health risks and lifestyle characteristics that could lead to inferior health outcomes(Sullivan et al., 2005).

**Cluster 2 analysis:**

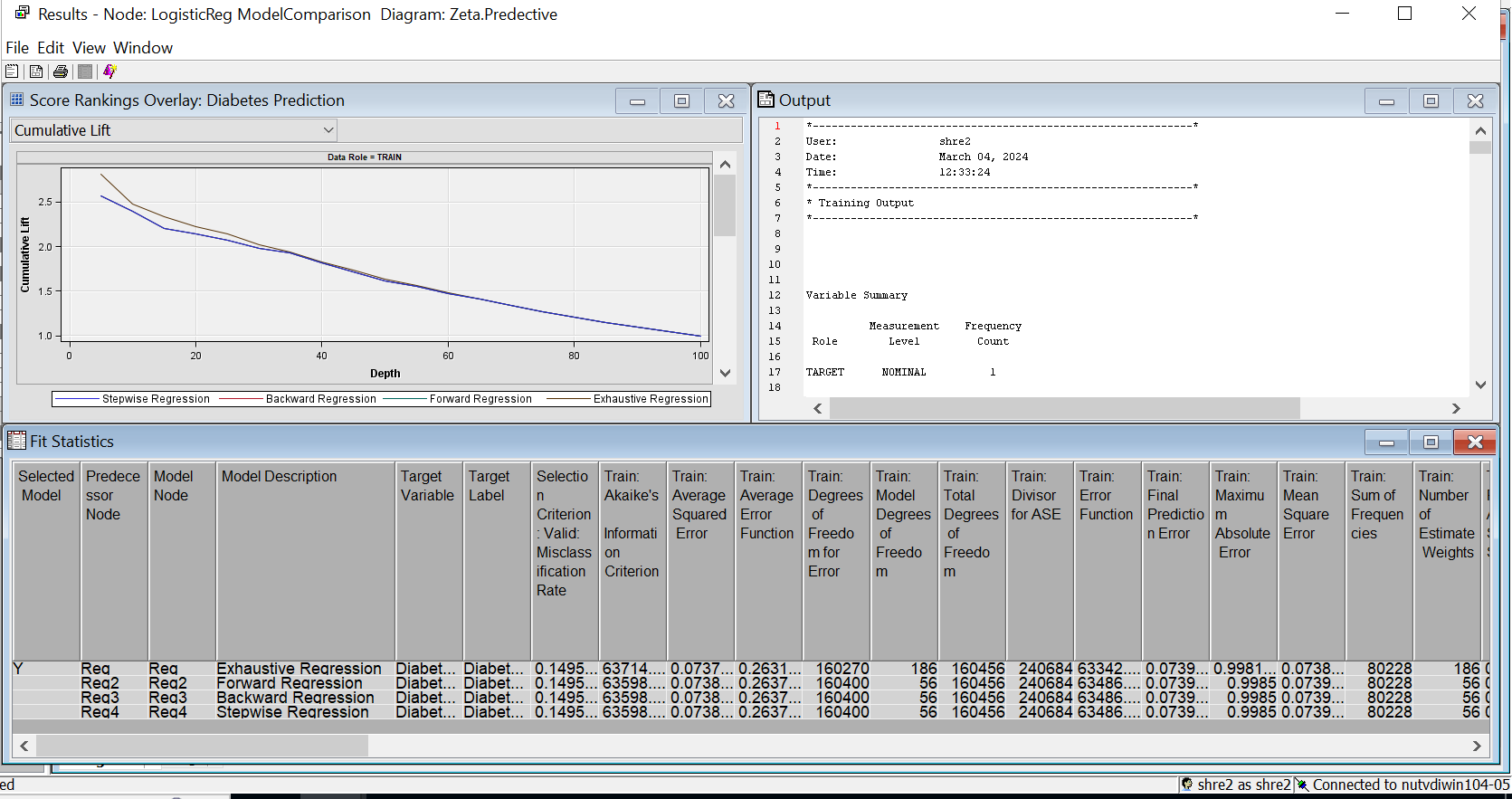
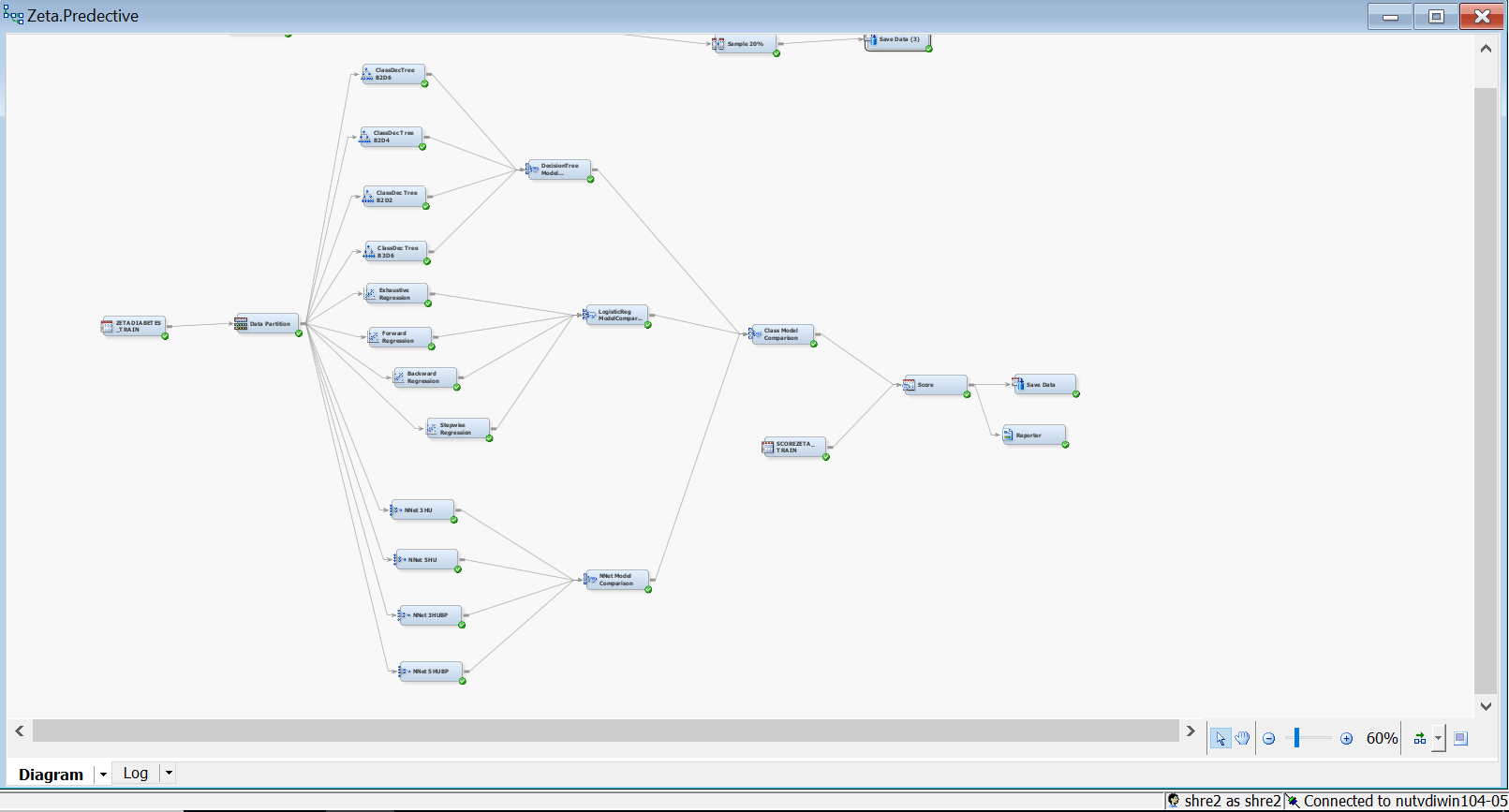
This population has modest health traits and educational backgrounds(Cruz-Cobo & Cano, 2020). The average education level is 3.452, which corresponds to people who have completed grades 9 through 11. The average BMI is 27.58, which places it in the overweight category, and gender representation is balanced. Notably, this cluster contains a moderate prevalence of smokers (mean: 0.2633) and people with elevated cholesterol (mean: 0.5181). The mean age of this cluster is not explicitly stated, but the overall pattern indicates a heterogeneous group with a diversity of health indicators, underlining the need for specific interventions tailored to the cluster's unique health requirements.

**Cluster 3 analysis:**

Cluster 3 appears as a cluster with overall positive health characteristics. The mean values for physical health variables such as DiffWalk (0.0059) and PhysActivity (0.1386) are particularly positive, indicating a population with less reported walking difficulties and higher levels of physical activity. Educational attainment is relatively higher, with a mean value of 6.15, indicating college education. This cluster has the highest average age (59.17), indicating an elderly demography. Furthermore, people in this cluster had a better health profile, with lower rates of smoking (mean: 0.0758), excessive alcohol intake, and a higher likelihood of seeking medical attention when necessary. Overall, Cluster 3 indicates a more health-conscious and educated population, which is expected to lead to improved overall health outcomes.

| **Variable Importance Before Sampling** | **Variable Importance After Sampling** |
| --- | --- |
|  |  |

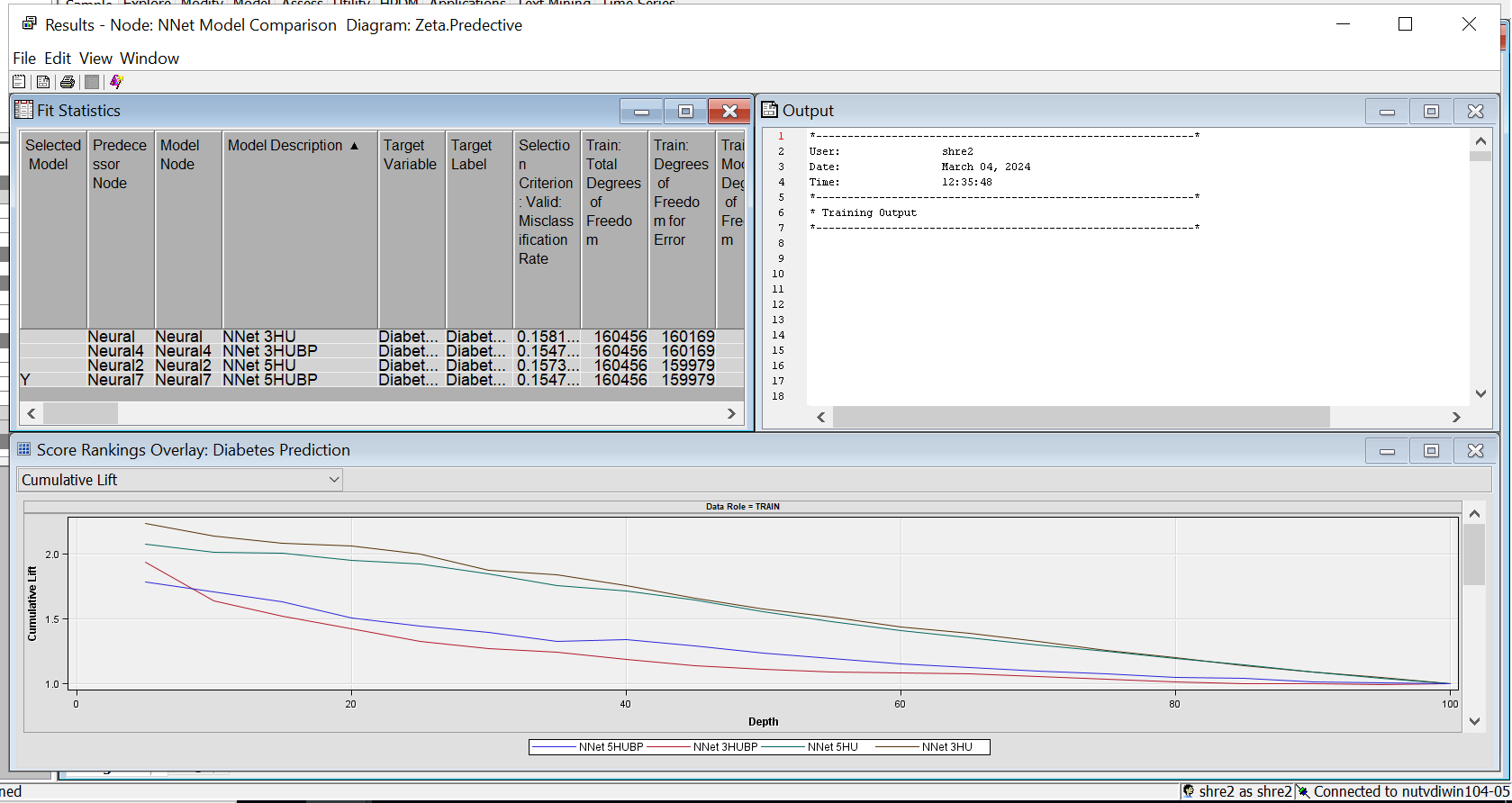
**5. Predictive Analysis (Decision Tree, Regression, Neural Network)**

* Descriptions of best models in each method:
* **Regression**
* 
* Model selection is critical in regression analysis to effectively predict outcomes. In this case, exhaustive regression was chosen as the best model based on the valid misclassification rate, which assesses the accuracy of the model's predictions (Shmueli et al, 2018). The misclassification rates for many regression algorithms are fairly similar, but exhaustive regression excels them all, with the lowest rate of 14.9512. Forward, backward, and stepwise regressions had a slightly higher misclassification rate of 14.9579.
* It's also crucial to evaluate the model's performance using additional metrics like average square error. In this scenario, the average square error for exhaustive regression falls from training, validation, and testing sets.
* - Training dataset: 0.07376.
* - Validation set: 0.07358.
* - Test set: 0.073382.
* This diminishing trend implies that the model is generalizing well to new data, as the error from the data on which it was trained lowers when applied to new data.

The exhaustive regression model identifies multiple relevant variables that have a considerable influence on the outcome being predicted.These variables and their significance are as follows:

* Age
* BMI (Body Mass Index)
* Cholesterol Check (CholCheck)
* Education
* General Health (GenHlth)
* Heart Disease or Heart Attack (HeartDiseaseorAttack)
* High Blood Pressure (HighBP)
* High Cholesterol (HighChol)
* Heavy Alcohol Consumption (HvyAlcoholConsump)
* Income
* Mental Health (MentHlth)
* No Doctor Because of Cost (NoDocbcCost)
* Sex
* Stroke

**Neural Network**

* 

In the neural network analysis, four distinct configurations were tried, each with a different number of hidden units and use of backpropagation. The purpose was to choose the most optimal model based on the misclassification rate, which is an important parameter for evaluating model correctness.

The following are the misclassification rates for each neural network configuration.

1. A neural network with three hidden units has a misclassification rate of 15.812.
2. Neural network with 5 hidden units: misclassification rate = 15.737.
3. Neural network with 3 hidden units and backpropagation: misclassification rate = 15.4714
4. Neural network with 5 hidden units and backpropagation: misclassification rate = 15.4714

With the least amount of misclassification among these configurations, the neural network with five hidden units and backpropagation turned out to be the best model. It is worth mentioning that the models with backpropagation outperformed the others, demonstrating that this training approach works well.

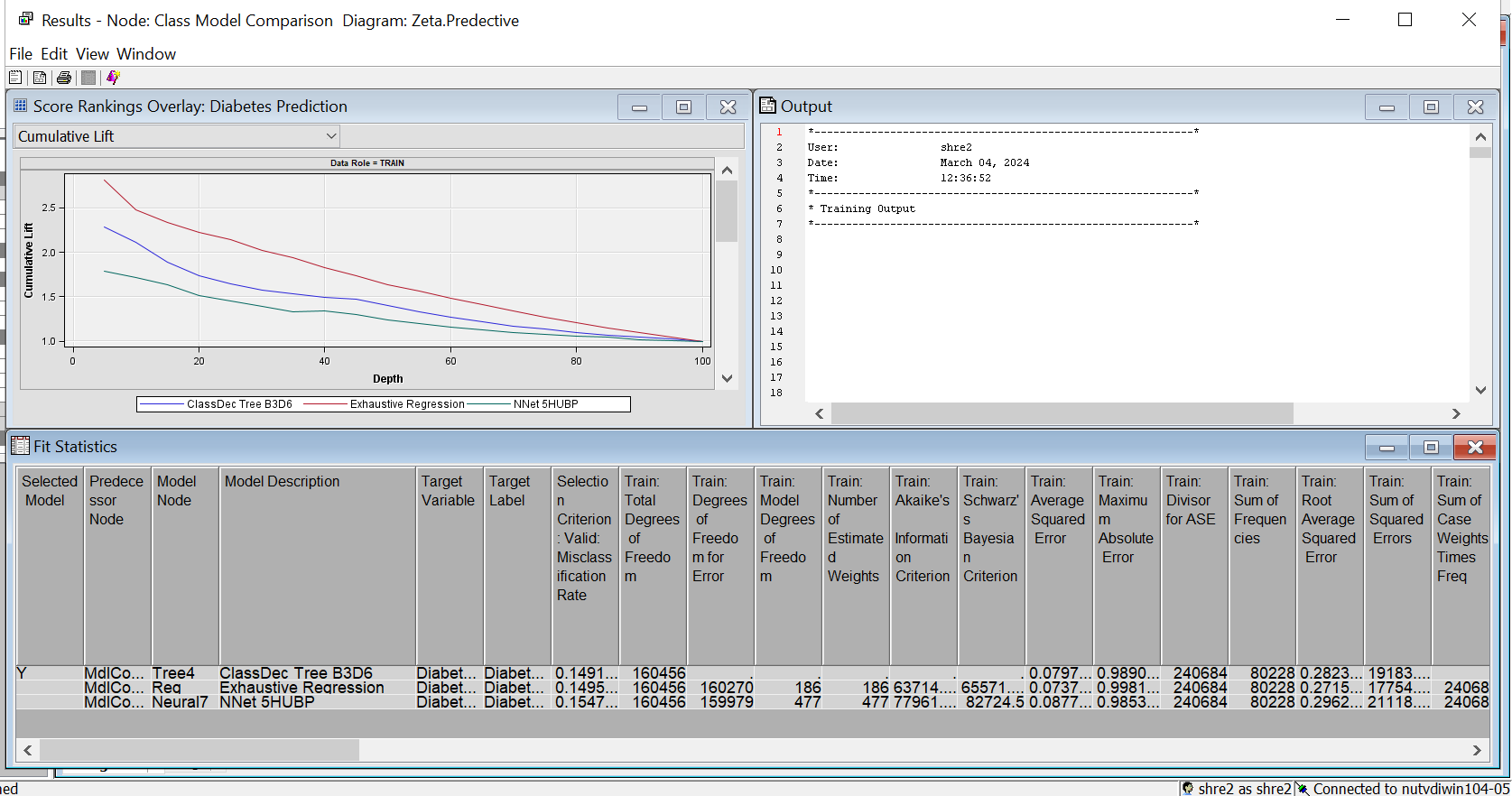
For the specified model (neural network with 5 hidden units with backpropagation), the average square error across multiple datasets was as follows:

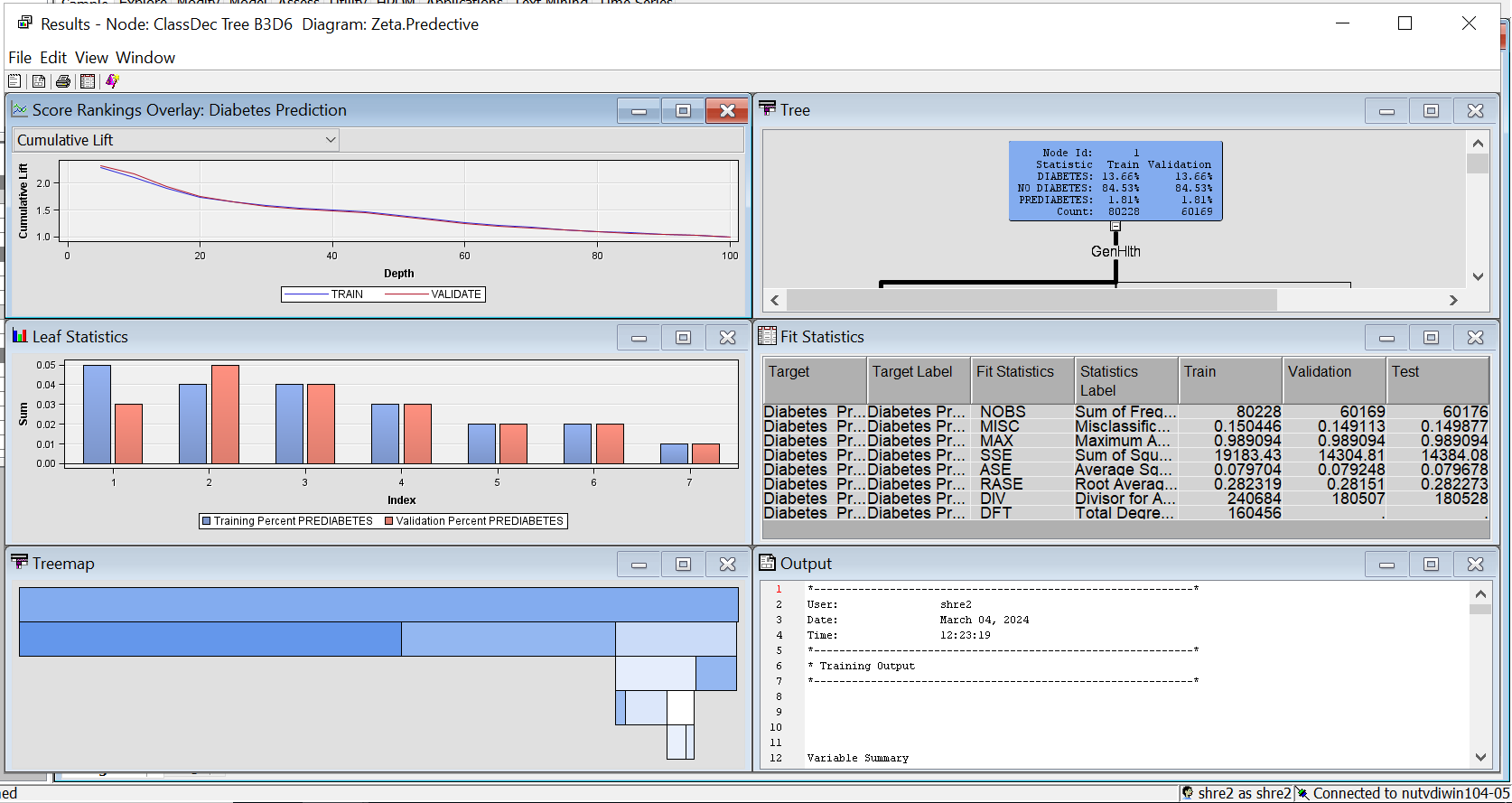
* Training set: 0.087742
* Validation set: 0.087737
* Test set: 0.087736.

These results show consistent performance across diverse datasets, implying that the model generalizes well and does not overfit to the training data.

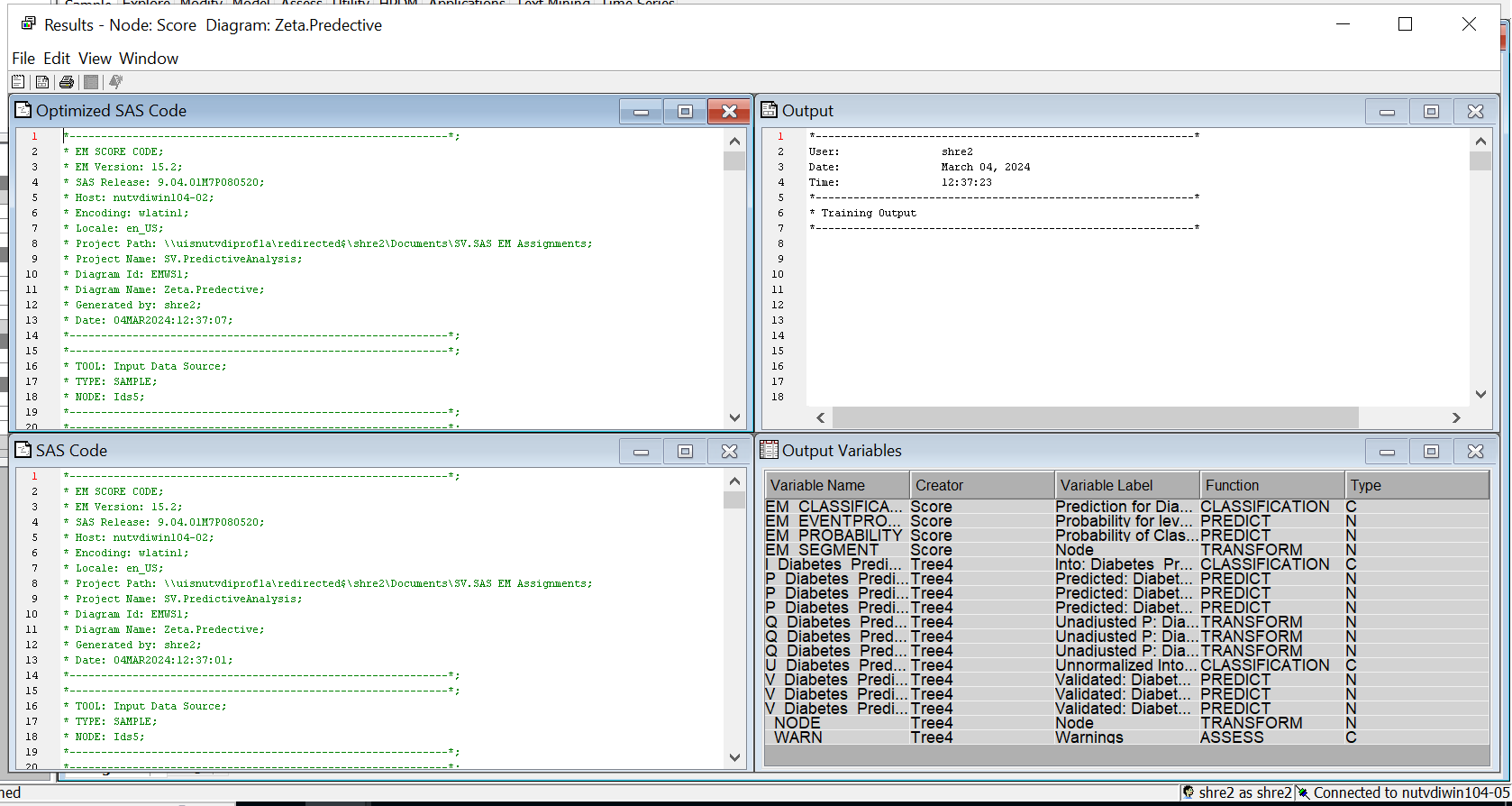
**Decision tree:**

Tree B3D6 was selected among the other decision trees and also it was selected as the best model among the neural network and regression model.





The process of selecting the best diabetes predictive model begins with a study of many decision tree models: B2D6, B2D4, B2D2, and B3D6. Among them, B3D6 stood out as the best performance, with a misclassification rate of 14.91%, beating out its competitors, who varied between 14.98% and 15.47%. Following its success, B3D6 was tested against neural networks and exhaustive regression models. Despite the competition, B3D6 retained its dominance, with a 15.4714% misclassification rate for neural networks and 14.9512% for exhaustive regression. The model's consistency across datasets was clear, with average squared error (ASE) values of 0.079704, 0.079248, and 0.079678 for the training, validation, and test datasets, respectively, showing low overfitting and strong generalization ability.The identification of significant variables such as "GenHlth," "HighBP," "BMI," and "HighChol" based on the tree diagram increased the model's predictive power by highlighting crucial elements in diabetes prediction. Using these variables promises to improve forecast accuracy and efficacy in real-world circumstances, providing useful insights for prevention and management efforts. With its superior performance, limited overfitting, and attention to essential predictors, the B3D6 decision tree model emerges as a dependable tool for accurate diabetes prediction, ready to make significant contributions to healthcare interventions and outcomes.



After this we will score the model on the new data to validate the accuracy of the best model on new data. Using this we identified the true positive and true negative and false positive and false negatives and using that we calculated the accuracy. The accuracy percentage for the scoring of our analysis was 98.18.

We have three classes in the target variable where 0 is no diabetes with frequency 8224, 1 is pre diabetes with frequency 176 and 2 is diabetes with frequency 1600. So the baseline misclassification is 1.76%.

**6. Conclusion and practical (actionable) recommendations**

Based on the analysis, here are practical and actionable recommendations for addressing diabetes risk using predictive models and exploratory analysis.

1. Implement targeted actions based on predictive model findings. For example, use the Decision Tree model (ClassDec Tree B3D6) to identify high-risk individuals and adapt interventions based on their demographic and health-related characteristics.

2. Develop lifestyle change programs for weight control, physical activity, and smoking cessation. The research identified target demographics with high BMI, low physical activity levels, and a greater smoking prevalence (Sullivan et al., 2005, Xie et al., 2019c).

3. Health Education Initiatives: - Improve health literacy, especially among those with lower education levels. Focus on promoting knowledge of diabetes risk factors, good lifestyle choices, and the value of frequent health exams(Cruz-Cobo & Cano, 2020).

4. Access to Healthcare Services: - Address barriers to healthcare access, especially for populations with limited access to healthcare services. Provide cheap healthcare choices and encourage regular check-ups, especially for people who have experienced difficulty getting healthcare due to cost.

5. Cardiovascular Health Monitoring: - Monitor cholesterol levels and check for heart disease or heart attack history. Integrate diabetes risk assessment into cardiovascular health initiatives to effectively detect and control comorbid risk factors.

6. Age-Specific Interventions: - Target interventions for specific age groups, taking into account the higher risk of diabetes with advanced age. Implement age-appropriate medications and screenings to successfully manage diabetes risk in older adults.

7. Socioeconomic Support Programs: - Implement socioeconomic support programs targeting populations with lower income levels. Address socioeconomic determinants of health, such as income inequality, to enhance healthcare access and encourage healthier lifestyle choices.

8. Multidisciplinary Approach: Encourage collaboration among healthcare professionals, community organizations, and public health agencies to prevent and manage diabetes effectively. Use the experience of various stakeholders to create comprehensive and successful interventions.

9. Continuous Monitoring and Evaluation: Establish procedures for monitoring and evaluating intervention initiatives to reduce diabetes risk. Review and adjust strategy on a regular basis, taking into account developing trends and participant input.

10. Community Engagement: - Promote diabetes prevention through outreach programs, support groups, and community-based initiatives. Empower people to take charge of their health and create circumstances that encourage healthy behaviors.

By applying these practical guidelines, healthcare organizations, governments, and community stakeholders can collaborate to effectively address diabetes risk and improve health outcomes in a variety of groups.

**Exploratory Analysis Question:**

## **What are the key risk factors associated with diabetes development among individuals surveyed in the BRFSS dataset?**

**Answer :** The variable significance analysis reveals characteristics and factors linked with diabetes development in the BRFSS dataset. The key factors contributing to the understanding of diabetes risk are:

1. Body Mass Index (BMI) is a significant factor in both pre- and post-sampling studies. Elevated BMI is well recognized as a risk factor for diabetes, highlighting the significance of weight management in diabetes prevention methods (Cruz-Cobo & Cano, 2020).

2. General Health (GenHealth) and Physical Health (PhysHealth): Diabetes risk assessment requires consideration of both general and physical health. Individuals who report poor overall health and are experiencing physical health issues are more likely to develop diabetes. These variables emphasize the general health status as a predictor of diabetes.

3. Difficult walking (DiffWalk): The existence of difficulties walking is strongly connected with diabetes risk, as evidenced by its importance. This shows a relationship between mobility challenges and underlying health factors that contribute to diabetes.

4. Education: Educational attainment is a key component, where greater education levels are noted. Education generally corresponds with health literacy, lifestyle choices, and access to healthcare, all of which influence diabetes risk (Cruz-Cobo & Cano, 2020).

5. Age: Advanced age is a well-established risk factor for diabetes, emphasizing the importance of age-specific therapies and screenings(Bellary et al., 2021).

6. Smoking and Alcohol Consumption: - Clusters with higher prevalence of smoking and excessive alcohol intake, such as Cluster 1, have been linked to an increased risk of diabetes. These lifestyle choices improve general health and may exacerbate diabetes development(Bellary et al., 2021).

7. Healthcare Access (AnyHealthcare and NoDocbcCost): - Important variables pertaining to healthcare access, such as AnyHealthcare and NoDocbcCost, are identified. Limited access to healthcare services may impede diabetes preventive and management efforts.

8. Heart illness History (HeartDiseaseorAttack) and Cholesterol Check (CholCheck): -Diabetic risk assessment is influenced by heart-related variables, such as cholesterol checks and a history of heart illness or attack. Cardiovascular health is inextricably linked to diabetes risk, and monitoring these factors is critical (Xie et al., 2019b).

9. Income: - Income is a significant factor in predicting diabetes risk, particularly after sampling. Socioeconomic factors, such as income levels, influence lifestyle choices and access to healthcare, which affects diabetes outcomes (“Economic Costs of Diabetes in the U.S. in 2007,” 2008).

Understanding the varying relevance offers a more nuanced perspective on the multifaceted nature of diabetes risk. These factors highlight the need for comprehensive and targeted public health initiatives that address lifestyle, healthcare access, and socioeconomic determinants in order to effectively reduce the risk of diabetes development in various groups.

## **Predict the likelihood of an individual having diabetes based on their demographic and health-related factors?**

* The Decision Tree model labelled as **"ClassDec Tree B3D6"** demonstrates superior performance in predicting the likelihood of an individual having diabetes based on demographic and health-related variables.

* Mean statistics consistently show higher diabetes prediction values associated with Decision Tree B3D6 across various datasets, including training, validation, and test sets.
* Decision Tree B3D6 exhibits lower variability in its predictions compared to the other two models, indicating greater reliability and consistency in capturing patterns within the data.
* The interpretability of Decision Tree models allows healthcare professionals to understand the decision-making process behind the predictions, facilitating informed clinical decisions and interventions.
* Therefore, based on the mean statistics provided, Decision Tree B3D6 stands out as the most reliable and effective model for predicting diabetes risk, leveraging demographic and health-related variables with superior accuracy and consistency compared to the other models.

**7. Team Member Contributions**

| **Phase** | **Contribution** |
| --- | --- |
| Finding Dataset | Shreya,Sohan |
| Dataset Cleaning and Preparation | Bhavana |
| Data Visualization | Sohan,Shreya |
| Exploratory Analysis | Bhavana,Sohan |
| Predictive Analysis | Shreya,Bhavana |
| Report | Bhavana,Shreya,Sohan |

**8. References**

[1] Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R. & Lichtendahl Jr, K. C. (2018) Data

Mining for Business Analytics: Concepts, Techniques, and Applications in RJohn Wiley

& Sons

[2]Xie, Z., Nikolayeva, О., Luo, J., & Li, D. (2019c). Building risk prediction models for Type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, *16*. https://doi.org/10.5888/pcd16.190109

[3]Cruz-Cobo, C., & Cano, M. (2020). Efficacy of diabetes education in adults with diabetes mellitus type 2 in primary care: a systematic review. *Journal of Nursing Scholarship*, *52*(2), 155–163. https://doi.org/10.1111/jnu.12539

[4] Bellary, S., Kyrou, I., Brown, J. E., & Bailey, C. J. (2021). Type 2 diabetes mellitus in older adults: clinical considerations and management. *Nature Reviews Endocrinology*, *17*(9), 534–548. https://doi.org/10.1038/s41574-021-00512-2

[5] Xie, Z., Nikolayeva, О., Luo, J., & Li, D. (2019b). Building risk prediction models for Type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, *16*. https://doi.org/10.5888/pcd16.190109

[6]World Health Organization: WHO & World Health Organization: WHO. (2023, April 5). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes

[7]Economic costs of diabetes in the U.S. in 2007. (2008). *Diabetes Care*, *31*(3), 596–615. https://doi.org/10.2337/dc08-9017

[8]https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download

[9]Obesity, inactivity, and the Prevalence of Diabetes and Diabetes-Related Cardiovascular Comorbidities in the U.S., 2000–2002. *Diabetes Care*, *28*(7), 1599–1603. https://doi.org/10.2337/diacare.28.7.1599