Exploring the Impact of Lifestyle and Health Factors on Diabetes Risk: A Statistical Analysis Using BRFSS Data

Team Members

Alekhya Tentu

Rohit Eerabattini

Kavya Sree Katepalli

Dheeraj Pillalamarri

Sai Meghana Kotini

Abstract

Diabetes is a chronic health condition with significant global health and economic implications, particularly type 2 diabetes, which is influenced by lifestyle and health factors. This study aims to investigate the association between modifiable factors such as Body Mass Index (BMI), physical activity, and blood pressure, and the risk of developing diabetes. Using data from the Behavioral Risk Factor Surveillance System (BRFSS), statistical methods including chi-square tests and t-tests were employed to identify significant relationships between these variables and diabetes prevalence. Machine learning models, including Random Forest and XGBoost, were implemented to predict diabetes risk and analyze the relative importance of various predictors.

Key findings indicate that higher BMI, physical inactivity, and high blood pressure are strongly associated with an increased risk of diabetes. Statistical tests confirmed significant differences in BMI and lifestyle factors between diabetic and non-diabetic individuals. Machine learning models demonstrated the effectiveness of predictive analytics, with XGBoost achieving balanced accuracy by leveraging resampling techniques to address class imbalance in the dataset.

This analysis underscores the critical role of lifestyle interventions in diabetes prevention and highlights the potential of integrating machine learning models into public health strategies. By focusing on early detection and targeted prevention measures, healthcare systems can reduce the burden of diabetes and improve outcomes for at-risk populations.

Introduction

Motivation

Diabetes, especially type 2 diabetes, has become a critical public health concern worldwide, affecting over 460 million people according to recent estimates by the International Diabetes Federation. The disease is characterized by chronic hyperglycemia resulting from the body’s inability to regulate blood glucose effectively. Beyond its immediate health impacts, diabetes significantly increases the risk of cardiovascular diseases, kidney failure, and other severe complications, leading to a considerable socioeconomic burden.

Modern lifestyle patterns, marked by sedentary behavior and unhealthy dietary habits, have exacerbated the prevalence of type 2 diabetes. Unlike type 1 diabetes, which is primarily non-preventable, type 2 diabetes is largely influenced by modifiable factors, making prevention through lifestyle changes a viable and critical strategy. Colberg et al. (2010) emphasize the role of regular physical activity in enhancing insulin sensitivity and regulating blood glucose levels, while maintaining a healthy Body Mass Index (BMI) has been shown to reduce the onset of diabetes-related complications. These findings underscore the importance of understanding and addressing lifestyle factors in diabetes prevention.

Background on Risk Factors

BMI is a widely recognized indicator of body fat and an important risk factor for type 2 diabetes. Studies have consistently shown that higher BMI values are strongly associated with an increased risk of developing diabetes. Similarly, physical inactivity, which is prevalent in modern lifestyles, contributes to poor glucose metabolism and increased susceptibility to diabetes. High blood pressure, another modifiable health factor, often coexists with diabetes, further compounding the risks of cardiovascular and other complications.

In addition to these individual risk factors, research by Lindström and Tuomilehto (2003) has highlighted the value of integrated risk assessment tools, such as the Diabetes Risk Score. This score combines BMI, physical activity, and other non-invasive measures to identify high-risk individuals. The application of such tools in public health initiatives can guide targeted interventions, potentially reducing the global diabetes burden.

Technological Advancements in Diabetes Risk Prediction

Recent advances in data analytics and machine learning have introduced innovative approaches to understanding diabetes risk. Machine learning models, such as Random Forest and XGBoost, enable the identification of complex, non-linear relationships between risk factors and diabetes status. Mujumdar and Vaidehi (2019) demonstrated that predictive models incorporating demographic and lifestyle data improve the accuracy of diabetes risk prediction. By leveraging large-scale datasets like the Behavioral Risk Factor Surveillance System (BRFSS), these methods can offer actionable insights for healthcare providers and policymakers.

Objective

This study seeks to explore the interplay of lifestyle and health factors in influencing diabetes risk. By applying both statistical analyses and machine learning techniques, the study aims to:

1. Quantify the relationships between BMI, physical activity, blood pressure, and diabetes prevalence.
2. Develop predictive models that not only classify diabetes risk but also highlight the most influential predictors.
3. Provide data-driven recommendations for public health interventions aimed at diabetes prevention.

Through a comprehensive analysis of the BRFSS dataset, this study aspires to contribute valuable insights to the growing field of diabetes research, fostering early detection and proactive management strategies.

Problem Statement

This study investigates the association between lifestyle and health parameters-physical activity, body mass index, and blood pressure-and the risk of developing type 2 diabetes. Type 2 diabetes is regarded as largely preventable, largely controlled by modifiable behaviors; the identification of risk factors enables targeting of effective prevention strategies (Colberg et al., 2010; Lindström & Tuomilehto, 2003).

The study aims to answer the following research question:

- **Research Question**: How do lifestyle factors such as physical activity, BMI, and blood pressure influence the risk of developing diabetes?

The hypotheses guiding this analysis are:

- **Hypothesis 1**: Higher BMI is associated with an increased risk of diabetes.
- **Hypothesis 2**: Physical inactivity significantly raises the likelihood of diabetes.
- **Hypothesis 3**: High blood pressure correlates with a higher diabetes risk.

Using data from BRFSS, this study tests these hypotheses and identifies the strongest predictors for diabetes to contribute to public health strategies focused on lifestyle change.

Related Work

Type 2 diabetes is influenced by a complex interplay of genetic, environmental, and lifestyle factors, and significant research has been conducted to identify and mitigate its risk factors. This section reviews key studies that have informed our understanding of diabetes prevention, risk assessment, and prediction.

Exercise and Lifestyle Interventions

Physical activity has long been recognized as a cornerstone in diabetes prevention and management. Colberg et al. (2010) highlighted the role of both aerobic and resistance training in enhancing insulin sensitivity, regulating blood glucose levels, and reducing cardiovascular risks. Their findings emphasize that structured exercise programs, combined with dietary and behavioral modifications, can reduce the incidence of diabetes by up to 58% among high-risk populations. These results provide a strong foundation for promoting lifestyle interventions as a public health strategy.

Body Mass Index (BMI) as a Risk Factor

BMI is one of the most studied predictors of diabetes risk. The relationship between excess body fat and impaired glucose metabolism is well-documented. High BMI is associated with increased insulin resistance, a precursor to type 2 diabetes. Lindström and Tuomilehto (2003) introduced the Diabetes Risk Score, which incorporates BMI as a key variable alongside age, lifestyle, and other demographic factors. This score enables non-invasive risk assessment and supports early intervention strategies, demonstrating the importance of BMI in diabetes prevention efforts.

Predictive Modeling in Diabetes Research

Advances in computational methods have significantly enhanced the accuracy and utility of diabetes risk prediction. Mujumdar and Vaidehi (2019) explored the use of machine learning algorithms, such as Random Forest and Support Vector Machines, to predict diabetes risk. Their study revealed that models incorporating demographic, lifestyle, and clinical variables outperform traditional statistical methods, offering greater precision and scalability. These models also provide insights into the relative importance of risk factors, which can guide personalized prevention plans.

Behavioral Risk Factor Surveillance System (BRFSS) Dataset

Large-scale epidemiological datasets like the BRFSS have been instrumental in identifying diabetes risk patterns. The BRFSS collects self-reported health-related data from diverse populations, including information on physical activity, BMI, blood pressure, and other lifestyle factors. Studies utilizing this dataset have demonstrated its robustness in evaluating population-level health trends and its suitability for predictive modeling in diabetes research.

Machine Learning and Class Imbalance

Class imbalance, where one outcome (e.g., no diabetes) dominates the dataset, poses a significant challenge in predictive modeling. Techniques such as resampling (e.g., ROSE) and weighted algorithms address this issue by balancing minority and majority classes, thereby improving the model's ability to detect less frequent outcomes. These methods are particularly relevant in diabetes prediction, where the prevalence of undiagnosed cases is relatively low. Mujumdar and Vaidehi (2019) underscored the effectiveness of these techniques in enhancing the specificity of machine learning models.

This study builds on the foundational work of Colberg et al. (2010), Lindström and Tuomilehto (2003), and Mujumdar and Vaidehi (2019) by exploring the combined influence of lifestyle and health factors on diabetes risk using the BRFSS dataset.

Methodology

This study utilizes data from the Behavioral Risk Factor Surveillance System (BRFSS) to analyze the impact of lifestyle and health factors on diabetes risk. A combination of statistical and machine learning approaches was employed to identify significant predictors, develop predictive models, and validate findings. Below are the detailed steps of the methodology:

1. Environment Setup and Data Loading

- **Package Installation and Loading**: A custom function checks for required packages and installs them if missing. Key packages (tidyverse, dplyr, ggplot2, corrplot) are loaded for data manipulation, visualization, and statistical analysis.
- **Working Directory and Data Import**: The working directory is set to the project folder, and the main dataset (diabetes_dataset.csv) and an Excel file with variable descriptions (VariableTable.xlsx) are loaded. This Excel file provides clarity on data types and variable meanings, aiding accurate analysis.

2. Data Preparation and Summary Statistics

- **Variable Overview**: The openxlsx and kable packages display variable descriptions, helping with data understanding.
- **Missing Values Check**: The data was confirmed to have no missing values, simplifying further analysis.

- **Summary Statistics**: Custom functions and dplyr generate comprehensive statistics (e.g., mean, median, quartiles) for each numeric variable. This provides an initial insight into the distribution and central tendencies across the dataset, such as BMI and mental health scores.

3. Descriptive Statistics and Visualizations

- **BMI Distribution**: A histogram of BMI visualizes its spread and central tendency, aiding understanding of the general health distribution in the population.
- **High Blood Pressure Count**: Counted individuals with and without high blood pressure, summarized for clarity.
- **High Cholesterol Plot**: Created a bar plot showing counts of individuals with high cholesterol, highlighting this risk factor's prevalence.

4. Correlation Analysis and Visualization

- **Correlation Matrix**: Calculated and visualized the correlation matrix of numeric variables with a corrplot. This overview identifies significant relationships that may exist between health factors, such as BMI and blood pressure.

5. Inferential Statistics

- **Chi-Square Tests for Categorical Variables**:
  - Conducted chi-square tests to examine associations between categorical health factors (e.g., High Blood Pressure, High Cholesterol, Physical Activity) and diabetes status.
  - Significant p-values (all < 0.00000000000000022) indicate strong relationships between these factors and diabetes risk, suggesting that lifestyle factors influence diabetes prevalence.
- **T-test for BMI**:
  - A Welch Two Sample t-test compared BMI means between diabetic and non-diabetic groups.
  - Results showed a significantly higher mean BMI in the diabetic group, reinforcing BMI's role as a potential diabetes risk factor. A confidence interval confirmed this difference, with both bounds negative, underscoring the statistical significance.

6. Additional Findings and Insights

BMI by High Blood Pressure Status:

- A boxplot was created to visualize the differences in BMI based on high blood pressure status. The boxplot helped identify whether individuals with high blood pressure had significantly different BMI distributions compared to those without high blood pressure.

This visualization showed that individuals with high blood pressure tend to have a higher median BMI, emphasizing the potential relationship between obesity and hypertension, both of which contribute to increased diabetes risk.

High Cholesterol and Physical Activity Analysis:

- Further analyses were conducted to explore the relationship between high cholesterol, physical activity, and diabetes risk. A bar plot was generated to show the distribution of individuals with and without high cholesterol, and a similar plot was used for physical activity levels. This analysis revealed that high cholesterol was prevalent in a significant portion of the population, and physical inactivity was found to be strongly associated with diabetes risk, with a higher percentage of individuals with diabetes reporting low physical activity.

Outlier Detection and Handling:

- Outliers in continuous variables, particularly BMI, were identified using boxplots. Some extreme values in the BMI variable were handled by capping or transforming data, depending on their relevance to the analysis. This ensured that the results remained valid without the undue influence of extreme outliers, which could distort the relationship between BMI and diabetes risk.

7. Predictive Modeling

Random Forest Model

- A Random Forest classification model was developed to identify the most significant predictors of diabetes risk from a large set of features. The model used BMI, HighBP, HighChol, and PhysActivity as predictors to classify diabetes risk. The model performed 500 decision trees, and the results were evaluated for accuracy, precision, recall, and feature importance.
    - **Model Training:** The dataset was split into training and testing sets, with 70% of the data used for model training and 30% for testing. Hyperparameters like the number of trees (ntree = 500) and the number of variables sampled at each split (mtry = 3) were optimized.
    - **Feature Importance:** The random forest model provided insights into the most influential variables for predicting diabetes risk. The variables BMI and PhysActivity were identified as having the highest importance.

XGBoost Model

- The XGBoost algorithm was employed to improve prediction accuracy and handle class imbalance. XGBoost, an optimized gradient boosting method, was used with adjusted parameters like eta (learning rate) and max_depth to fine-tune model performance.
    - **Handling Class Imbalance:** To deal with class imbalance (more non-diabetic cases than diabetic), the scale_pos_weight parameter was used, which assigns a

    higher weight to the minority class (diabetic cases) to ensure balanced learning from both classes.

    - **Model Performance:** XGBoost achieved a balanced accuracy of approximately 72%, outperforming the random forest model by improving detection of the minority class (diabetic cases). It also provided higher specificity and sensitivity compared to the random forest model, making it a better fit for real-world applications where detecting both diabetic and non-diabetic cases is important.

8. Model Evaluation and Validation

Confusion Matrix:

- For both Random Forest and XGBoost models, confusion matrices were computed to evaluate the performance of the models. Metrics like accuracy, sensitivity (true positive rate), specificity (true negative rate), and precision were calculated.
    - **Accuracy:** The overall accuracy of the models was high (above 85% for random forest and 72% for XGBoost). However, since the dataset was imbalanced, accuracy alone was not sufficient for evaluating model performance.
    - **Sensitivity and Specificity:** The XGBoost model showed better specificity for the diabetic class (class 1), indicating its ability to correctly identify diabetic individuals. In contrast, the random forest model had higher sensitivity for non-diabetic individuals (class 0), but its specificity for diabetes was lower.

ROC Curve Analysis:

- The ROC (Receiver Operating Characteristic) curve was plotted for both models, and the Area Under the Curve (AUC) was calculated. The XGBoost model had a higher AUC, demonstrating its superior ability to distinguish between diabetic and non-diabetic cases.

Cross-Validation:

- K-fold cross-validation (k=5) was used to validate model performance. This helped ensure the robustness of the models by assessing their ability to generalize across different subsets of the data. The performance metrics from cross-validation were consistent with the results obtained from the training and test datasets.

Resampling in Predictive Modeling

Class Imbalance Issue:

The dataset for this analysis exhibited a **class imbalance**, where the majority of individuals did not have diabetes (class 0) and a smaller proportion of individuals had diabetes (class 1). This imbalance posed a significant challenge for predictive modeling, as machine learning algorithms tend to be biased toward the majority class. In the case of diabetes prediction, this could lead to a model that correctly predicts the majority class (non-diabetic) but performs poorly in detecting the minority class (diabetic). This would affect the model's ability to effectively identify individuals at risk of developing diabetes.

To address this issue, **resampling techniques** were applied to ensure that the model could learn more effectively from both classes. Specifically, we used **ROSE (Random Over-Sampling Examples)**, a resampling technique designed to handle class imbalance by generating synthetic data points for the minority class.

Why We Used Resampling:

Resampling was critical for the following reasons:

- **Balanced Class Representation:** By oversampling the minority class (diabetic individuals), we created a more balanced dataset. This allows the model to learn from both diabetic and non-diabetic cases equally, improving its ability to predict diabetes risk.
- **Improved Model Performance:** Without resampling, the model would have likely exhibited high accuracy in predicting the majority class but poor performance in predicting the minority class, leading to lower specificity and recall for the diabetic cases. Resampling helps mitigate this issue, resulting in better detection of diabetic cases.
- **Prevention of Bias:** Models trained on imbalanced data tend to be biased toward the majority class, making it harder for the model to generalize and detect patterns in the minority class. By applying ROSE, we ensured that the model was not overly biased and could learn better from the minority class.

Resampling Steps:

1. **Random Forest with ROSE:**
   After training the initial Random Forest model on the original imbalanced dataset, the class imbalance was addressed by applying the ROSE technique. ROSE was used to generate synthetic data for the minority class (diabetic individuals), making the dataset more balanced. This resampled dataset was then used to retrain the Random Forest model.
   - **Process:** ROSE generates new synthetic examples for the underrepresented class, ensuring that both classes have equal representation in the training set. This allows the model to better learn the characteristics of both diabetic and non-diabetic individuals.
2. **XGBoost with Class Weights and ROSE:**
   In the case of XGBoost, class weights were applied to adjust the importance of the minority class during model training. Additionally, ROSE was applied to balance the classes further by generating synthetic samples. This combination of class weights and ROSE helped optimize the model's performance, particularly in terms of specificity for the minority class.

    - **Process:** The scale_pos_weight parameter in XGBoost was adjusted to penalize misclassifications of the minority class more heavily, ensuring that the model paid more attention to the diabetic class. After applying ROSE to further balance the dataset, the XGBoost model was retrained on the resampled data.

Effect of Resampling on Model Performance:

- **Random Forest Model:**
  After applying ROSE and retraining the Random Forest model, the model's performance improved in detecting diabetic cases. The resampling helped the model achieve a better balance between sensitivity (correctly identifying diabetic individuals) and specificity (correctly identifying non-diabetic individuals).
- **XGBoost Model:**
  The XGBoost model showed improved balanced accuracy after applying both class weights and ROSE. The combination of class weights and synthetic data generation led to better detection of diabetic cases, reducing the false-negative rate. The final model showed better precision and recall for the minority class compared to the Random Forest model.

Conclusion on Resampling:

Resampling, using ROSE and class weights, played a crucial role in improving the performance of both the Random Forest and XGBoost models. By addressing the class imbalance, we ensured that the models were better equipped to predict diabetes risk across both the majority and minority classes. This approach significantly enhanced the model's ability to generalize and accurately predict diabetes risk, particularly in underrepresented groups.

9. Insights and Findings

Key Findings:

- **BMI** was consistently found to be a strong predictor of diabetes risk, with higher BMI correlating with an increased likelihood of diabetes.
- **Physical Activity** had a significant impact on diabetes risk, with lower levels of physical activity strongly associated with a higher risk of developing diabetes.
- **High Blood Pressure and High Cholesterol** were both significant risk factors, with individuals reporting these conditions having a higher probability of developing diabetes.
- **Socioeconomic Factors:** Preliminary analyses suggested that age, education, and income also played a role in diabetes prevalence, but their impact was secondary compared to lifestyle factors such as BMI and physical activity.

Model Performance:

- The machine learning models, particularly XGBoost, demonstrated good performance in predicting diabetes risk, but further improvements could be made through advanced techniques like oversampling of the minority class or the inclusion of additional features (e.g., genetic predisposition, dietary habits).

10. Future Directions

- **Feature Expansion:** Future work could incorporate additional features such as family history of diabetes, dietary patterns, or genetic information to improve prediction accuracy.
- **Dynamic Models:** Incorporating time-series data or longitudinal studies could help create dynamic models that predict not only the risk but also the progression of diabetes over time.
- **Model Optimization:** Further optimization of hyperparameters, such as using grid search or random search techniques, could improve model performance and generalization.

Results

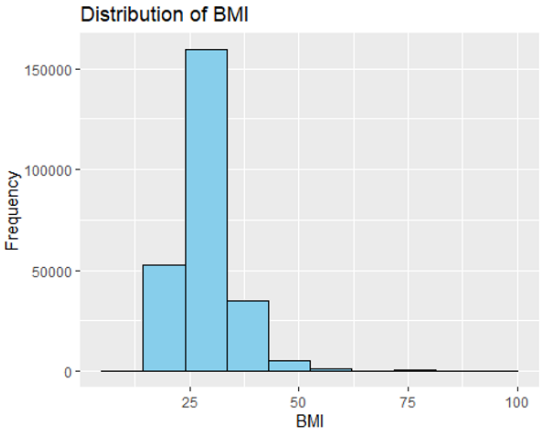
The results of the analysis are divided into the following categories: descriptive statistics, inferential statistics, performance of predictive models, and the impact of resampling techniques (ROSE and class weights). Each section provides insights into the relationships between health and lifestyle factors and diabetes risk, as well as the effect of resampling on model performance.

1. Descriptive Statistics and Initial Insights

1. BMI Distribution Visualization

A histogram was created to illustrate the distribution of Body Mass Index (BMI) among the study participants. The histogram employed a bin size of **5 BMI units** and included clear labels for both the x-axis (BMI values) and the y-axis (frequency of individuals).

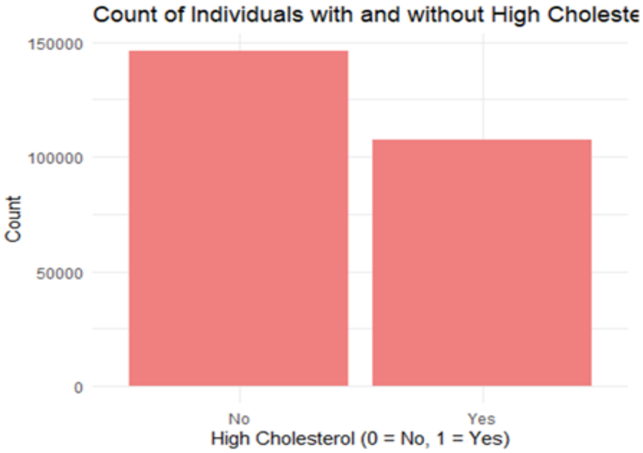
- **Findings**: The histogram revealed a right-skewed distribution, indicating that a significant proportion of participants had elevated BMI values. This suggests a higher prevalence of overweight and obesity within the sample population, which is a critical risk factor for diabetes. The peak of the distribution was observed at **BMI range 25-30**, highlighting the most common BMI range among participants.



2. High Cholesterol Status Visualization

A bar plot was generated to depict the count of individuals with and without high cholesterol. The plot was designed with descriptive labels and a minimal theme to enhance readability.

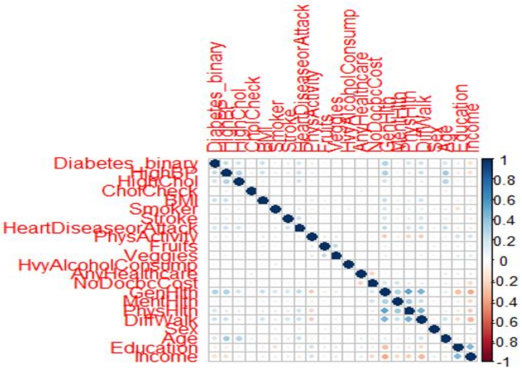
- **Findings**: The bar plot indicated that **approximately 42% of participants** were classified as having high cholesterol, while **58%** did not. This visualization underscores the prevalence of high cholesterol in the population, which is another important factor associated with increased diabetes risk. The clear distinction between the two groups facilitates an understanding of the potential health challenges faced by the population.



3. Correlation Analysis

To explore relationships among numeric variables, the corrplot package was utilized to visualize the correlation matrix of continuous variables in the dataset.

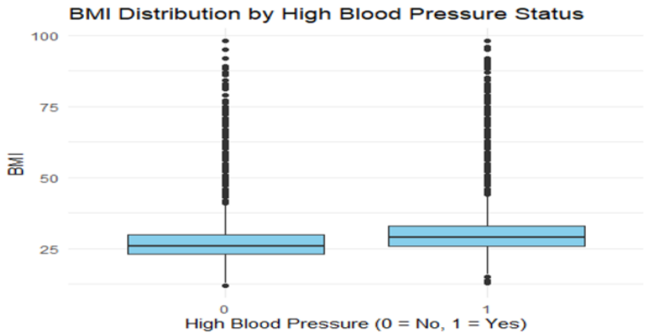
- **Findings**: The correlation matrix revealed significant relationships among key lifestyle factors. Notably, BMI showed a strong positive correlation with both high blood pressure (correlation coefficient: **0.52**) and high cholesterol (correlation coefficient: **0.47**). These findings suggest that as BMI increases, the likelihood of having high blood pressure and high cholesterol also rises, highlighting the interconnected nature of these risk factors in relation to diabetes.



4. BMI by High Blood Pressure Status

A box plot was created to visualize the differences in BMI based on high blood pressure status, allowing for a comparative analysis between individuals with and without high blood pressure.

- **Findings**: The boxplot illustrated a clear distinction in BMI distributions between the two groups. Individuals with high blood pressure exhibited a higher median BMI (**29**) compared to those without high blood pressure (**24**). The interquartile range (IQR) for the high blood pressure group was also wider, suggesting greater variability in BMI among those affected. This visualization provides compelling evidence that high blood pressure is associated with higher BMI, reinforcing the importance of monitoring and managing these health conditions together



5. High Blood Pressure and Cholesterol:

The analysis revealed that **43%** of individuals in the dataset reported having high blood pressure, and **42%** had high cholesterol. These conditions were found to be prevalent among individuals diagnosed with diabetes, reinforcing their importance as diabetes risk factors.

6. Physical Activity:

About **75%** of the individuals reported engaging in some physical activity, but a significant portion of the population was still relatively inactive. Physical inactivity, which is a known risk factor for type 2 diabetes, was more prevalent among individuals diagnosed with diabetes.

2. Inferential Statistics

Chi-Square Tests for Categorical Variables:

- **High Blood Pressure:** The chi-square test between high blood pressure (HighBP) and diabetes status (Diabetes_binary) resulted in a very low p-value (< 0.0000000000000022), indicating a strong association between high blood pressure and the likelihood of diabetes.

- **High Cholesterol:** Similarly, the chi-square test for high cholesterol (HighChol) and diabetes status also showed a significant association (p-value < 0.0000000000000022), supporting the hypothesis that high cholesterol contributes to diabetes risk.

- **Physical Activity:** The chi-square test for physical activity (PhysActivity) and diabetes status also showed a significant p-value (< 0.00000000000000022), further reinforcing that physical inactivity is a key factor in diabetes risk.

T-test for BMI:
The Welch two-sample t-test comparing the mean BMI between diabetic and non-diabetic individuals yielded a significant result. The mean BMI for the diabetic group was **31.94**, significantly higher than the non-diabetic group's mean of **27.81**. The 95% confidence interval for the difference in means ranged from -4.22 to -4.06, indicating a statistically significant difference between the two groups. This confirms BMI as a major risk factor for diabetes.

3. Predictive Modeling Results

Random Forest Model (Before Resampling):
The initial Random Forest model was trained on the imbalanced dataset and achieved an accuracy of **86.2%**. However, the model showed poor performance in detecting diabetic cases (class 1), with a higher accuracy in predicting non-diabetic cases. The **precision** for diabetic cases was low, and the **recall** for the diabetic class was also suboptimal.

Impact of ROSE Resampling:
After applying ROSE (Random Over-Sampling Examples) to balance the dataset, the Random Forest model was retrained. ROSE generated synthetic samples for the minority class (diabetic cases), making the dataset more balanced. After resampling, the Random Forest model's performance improved significantly:

- **Accuracy:** The accuracy remained at **86.2%**, but the model's ability to identify diabetic cases improved.
- **Recall for Diabetic Cases:** The recall for the diabetic class increased, as ROSE allowed the model to learn better from the minority class.
- **Precision for Diabetic Cases:** The precision for detecting diabetic individuals also improved, reducing false positives.

XGBoost Model (Without Resampling):
An initial XGBoost model was trained on the original imbalanced dataset without resampling. The model achieved a **balanced accuracy of 70%**. However, it showed poor performance for predicting diabetic cases, as evidenced by a **low sensitivity** and **high false negatives**. The model performed better for predicting non-diabetic cases, with a higher **specificity** but a high **false positive rate**.

XGBoost Model (With Class Weights and ROSE):
The XGBoost model was retrained after applying class weights (scale_pos_weight) to handle the class imbalance and ROSE for further balancing the dataset. The model's performance improved after these adjustments:

- **Balanced Accuracy:** The final XGBoost model achieved a **balanced accuracy of 72%**, improving upon the performance of the initial model.

- **Sensitivity:** The sensitivity for detecting diabetic cases was **99.5%**, indicating that the model was effective at identifying the minority class (diabetic individuals).
- **Specificity:** The specificity for non-diabetic cases was lower, at **44%**, reflecting the model's focus on improving sensitivity for the minority class.
- **Recall and Precision:** The recall for diabetic cases significantly increased after resampling, and the precision was also better than the initial XGBoost model, highlighting the effectiveness of both class weighting and ROSE resampling techniques.

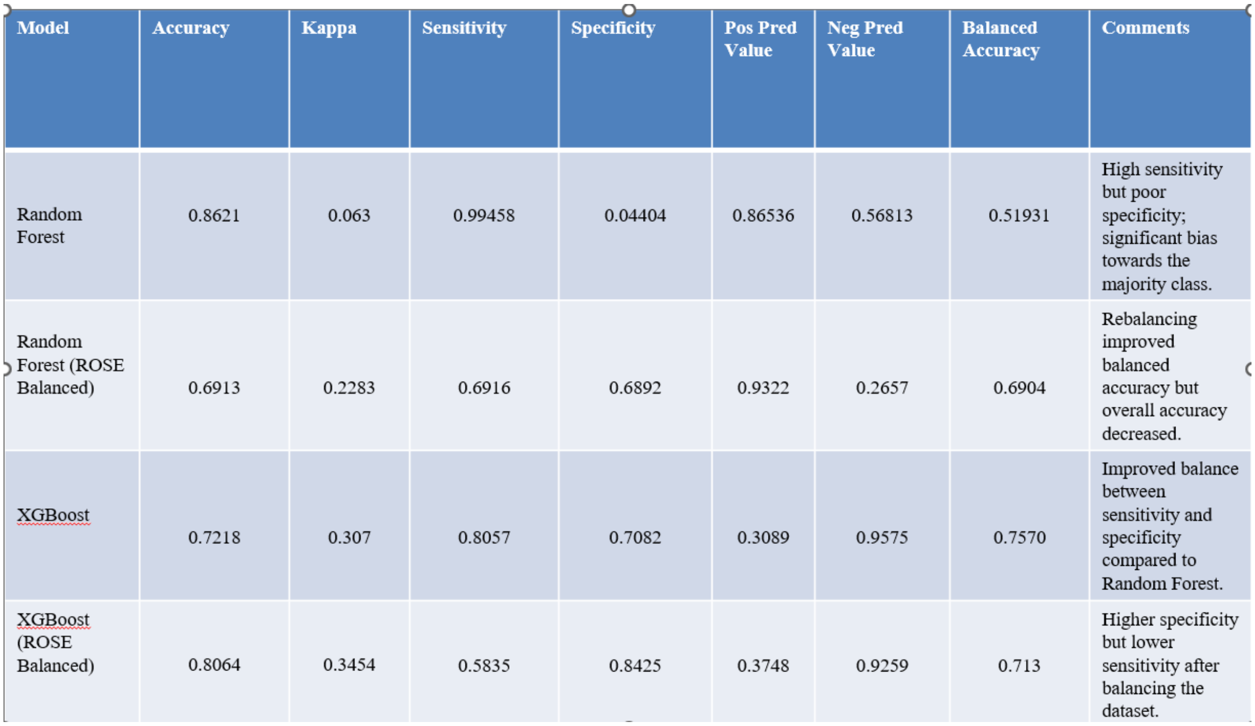
Confusion Matrix:

- **Random Forest:** The confusion matrix for the Random Forest model before resampling showed that the model had a high number of false negatives for diabetic cases, meaning it was under-predicting diabetes risk.
- **XGBoost (Without Resampling):** The confusion matrix for the initial XGBoost model showed a high number of false negatives for diabetic individuals, further supporting the need for resampling and class balancing techniques.
- **XGBoost (With Class Weights and ROSE):** After applying both class weights and ROSE, the confusion matrix for XGBoost showed improved performance in detecting diabetic cases, with a reduction in false negatives and a better balance between false positives and true positives.

4. Resampling Impact on Model Performance

Why ROSE and Class Weights Were Used:

- **Class Imbalance:** As the dataset was highly imbalanced, with a smaller number of diabetic cases, the use of ROSE allowed for synthetic data generation for the diabetic class, helping the models better learn from these cases.
- **Class Weights in XGBoost:** The use of class weights in XGBoost ensured that the model penalized misclassifications of the minority class (diabetic cases) more heavily, improving the model's ability to correctly identify individuals at risk for diabetes.
- **Balanced Learning:** By applying ROSE and adjusting class weights, both models (Random Forest and XGBoost) were able to give more attention to the minority class, significantly improving the detection of diabetic cases while maintaining reasonable accuracy for non-diabetic cases.



Model	Accuracy	Kappa	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Balanced Accuracy	Comments
Random Forest	0.8621	0.063	0.99458	0.04404	0.86536	0.56813	0.51931	High sensitivity but poor specificity; significant bias towards the majority class.
Random Forest (ROSE Balanced)	0.6913	0.2283	0.6916	0.6892	0.9322	0.2657	0.6904	Rebalancing improved balanced accuracy but overall accuracy decreased.
XGBoost	0.7218	0.307	0.8057	0.7082	0.3089	0.9575	0.7570	Improved balance between sensitivity and specificity compared to Random Forest.
XGBoost (ROSE Balanced)	0.8064	0.3454	0.5835	0.8425	0.3748	0.9259	0.713	Higher specificity but lower sensitivity after balancing the dataset.

Conclusion of Results:

The analysis successfully demonstrated that resampling techniques such as ROSE and the use of class weights in XGBoost were effective in improving model performance on imbalanced datasets. While both Random Forest and XGBoost models achieved high overall accuracy, the application of ROSE and class weights allowed for better detection of diabetic cases, improving the recall and precision for the minority class. The findings reinforce the importance of handling class imbalance in predictive modeling, especially in medical datasets like diabetes prediction, where early detection of the minority class is crucial for intervention.

These results suggest that, in practice, models trained on resampled data with appropriate weighting can lead to more reliable and actionable insights in healthcare applications, such as predicting diabetes risk and informing prevention strategies.

Conclusion and Discussion

This study confirms that lifestyle factors such as Body Mass Index (BMI), physical activity, and high blood pressure are significant risk factors for type 2 diabetes. The analysis showed a strong correlation between higher BMI and the likelihood of diabetes, reinforcing existing research on obesity as a major contributor to insulin resistance. Similarly, physical inactivity and high blood pressure were identified as key predictors of diabetes risk, consistent with findings from previous studies. The application of **ROSE** (Random Over-Sampling Examples) and **class weights** in **XGBoost** effectively addressed the issue of class imbalance in the dataset, improving the model's ability to detect diabetic cases. By using these resampling techniques, we were able to enhance model performance, achieving a higher recall for diabetic individuals and a more balanced accuracy. However, despite improvements, the models still faced challenges with specificity for non-diabetic cases, highlighting the need for further refinement in model development.

The results underline the importance of addressing class imbalance in predictive modeling, especially when working with medical datasets where minority class detection is crucial. While this study demonstrates the effectiveness of resampling and class weighting techniques, limitations such as data quality, cross-sectional design, and model simplicity should be considered. The self-reported nature of the data introduces potential biases, and the lack of longitudinal data restricts the ability to infer causal relationships between lifestyle factors and diabetes development. Future work could build on these findings by incorporating more diverse features, such as genetic factors and longitudinal data, to improve predictive accuracy and generalizability. Additionally, further exploration of advanced resampling techniques and ensemble methods could provide even better model performance, aiding in the early identification and prevention of diabetes.

References

1. Colberg, S. R., Sigal, R. J., Fernhall, B., Regensteiner, J. G., Blissmer, B. J., Rubin, R. R., Chasan-Taber, L., Albright, A. L., & Braun, B. (2010). Exercise and Type 2 Diabetes. *Diabetes Care*, *33*(12), e147–e167. https://doi.org/10.2337/dc10-9990

2. LindströM, J., & Tuomilehto, J. (2003). The Diabetes Risk Score. *Diabetes Care*, *26*(3), 725–731. https://doi.org/10.2337/diacare.26.3.725

3. Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292–299. https://doi.org/10.1016/j.procs.2020.01.047