Diabetes Prediction

Manav Mandal and Atharva Gurav

Abstract

Amid the global diabetes epidemic that affects over 420 million people and gives rise to critical health issues, early detection has emerged as a crucial intervention. This study harnesses the power of advanced machine learning techniques to predict diabetes risk among a population of US citizens, drawing exclusively from data within the United States. With a substantial dataset comprising more than 250,000 instances, the primary objective is to categorize individuals into one of three groups: Healthy, Diabetic, or Pre-Diabetic. Lifestyle modifications, including dietary adjustments and lifestyle changes, offer a promising avenue for reducing diabetes risk, particularly within high-risk demographics. The project seeks to uncover essential lifestyle factors that contribute to the onset of diabetes. Building upon prior research that compared prediction models and utilizing diverse data mining techniques, the study's outcomes will furnish valuable insights into diabetes risk predictors, facilitating early intervention and preventive strategies.

Keywords

Diabetes Prediction, Machine Learning, Data Mining, Health Indicators, BRFSS 2015

1 Introduction

Diabetes has become a global epidemic where over 420 million people suÐer from diabetes worldwide. It can cause other health related concerns such as cardiovascular disease, blindness and kidney failure if not managed on time. Early detection of diabetes has become crucial for timely treatment and prevention. Advances in machine learning has opened up new avenues for predictive modelling that helps us identify high risk individuals based on their day to day lifestyle. Machine learning models such as logistic regression, random forest and even deep learning networks such as CNN analyze huge amounts of data that help uncover latent relationships between patterns associated with diabetes.

We leverage machine learning approaches to detect diabetes using the Centers for Disease Control and Prevention database. This dataset contains information about the lifestyle, physical activity and other demographics. Our goal is to use machine learning approaches to classify an individual into one of three classes: Healthy, Diabeteic or Pre-Diabeteic. Answers to questions like Which lifestyle modifications can be made such as diet changes, prohibiting smoking and alcohol consumption to significantly reduce the progression even among high-risk groups. This project provides insights into the most important lifestyle predictors for diabetes onset.

2 Previous Work

A Survey on Diabetes Risk Prediction using Machine Learning Approaches (2023)[1]:
This article explored a range of supervised and unsupervised machine learning techniques for early detection of Diabetes mellitus(DM) a chronic condition. Further work proposed creation of amore precise and general model for early detection on DM.

Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type (2022)[2]:
The research provided an interesting approach for diabetes prediction using personalized lifestyle habits. The data used was from NHANES data from 17,833 respondents and employed AIC forward propagation to filter out 18 lifestyle variables for model training. They used a number of models out which CATBoost model achieved an accuracy of 82.1% and an AUC of 0.83.

Data-Driven Machine-Learning Methods for Diabetes Risk Prediction (2021)[3]:
This paper explored an important aspect of not only combining lifestyle data but also genetic and medical history for an even better performance from our model. Data exploration through risk factor analysis could help to identify associations between the features and diabetes. Performance analysis showed that data pre-processing is a major step in the design of efficient and accurate models for diabetes occurrence.

Predicting the Onset of Diabetes with Machine Learning Methods(2021)[4]:
This study investigates the use of machine learning algorithms for predicting the onset of diabetes in individuals with pre-diabetes. The research demonstrates the potential of machine learning in

identifying individuals who are likely to progress to diabetes, allowing for early intervention and prevention strategies. Prediction of diabetes based on personal lifestyle indicators (2015): This paper proposes a framework for predicting diabetes risk based on various lifestyle factors, including physical activity, dietary habits, and sleep patterns. The research demonstrates the feasibility of using lifestyle information for early diabetes detection.

Related Projects:
The PREDICT study (2020-2025)[5]:
This ongoing longitudinal study aims to develop a personalized risk prediction model for diabetes using wearable technology and machine learning.
The project collects data on various lifestyle factors and health metrics from participants over a New-year period to develop accurate and personalized risk assessments. The Diabetes Risk Engine (2019): This project developed a mobile application that uses machine learning algorithms to predict an individual's risk of developing diabetes based on their lifestyle habits and demographic information. The app aims to provide users with personalized risk assessments and recommendations for reducing their diabetes risk.

The SMART Diabetes Prevention Pro- gram (2018-2023)[6]:
This project investigates the effectiveness of a digital health intervention program that combines lifestyle coaching with personalized feedback based on data collected from wearable devices. The program aims to empower individuals at high risk of developing diabetes to adopt healthy lifestyle changes and prevent the disease.

3 Methods

We leveraged the Centers for Disease Control and Prevention database, rich with lifestyle, physical activity, and demographic data, to classify individuals into one of three categories: Healthy, Diabetic, or Pre-Diabetic. Our methodological approach is multifaceted, encompassing data preprocessing, exploratory data analysis, predictive analysis, and a comprehensive examination of feature importance.

SMOTE, which stands for Synthetic Minority Over-sampling Technique, is a method used in data science and machine learning to address the problem of class imbalance in datasets. In many real-world problems, some classes are under-represented compared to others, leading to biased models that do not perform well on the minority class. SMOTE is one of the techniques used to overcome this issue. We used SMOTE to remove imbalance from our dataset.

3.1 Exploratory data analysis

Our EDA was pivotal in uncovering underlying patterns within the demographic and lifestyle - related data. Among the visualizations, two graphs stood out for their potential implications on diabetes risk: the "Age Distribution" and "Physical Activity Distribution".

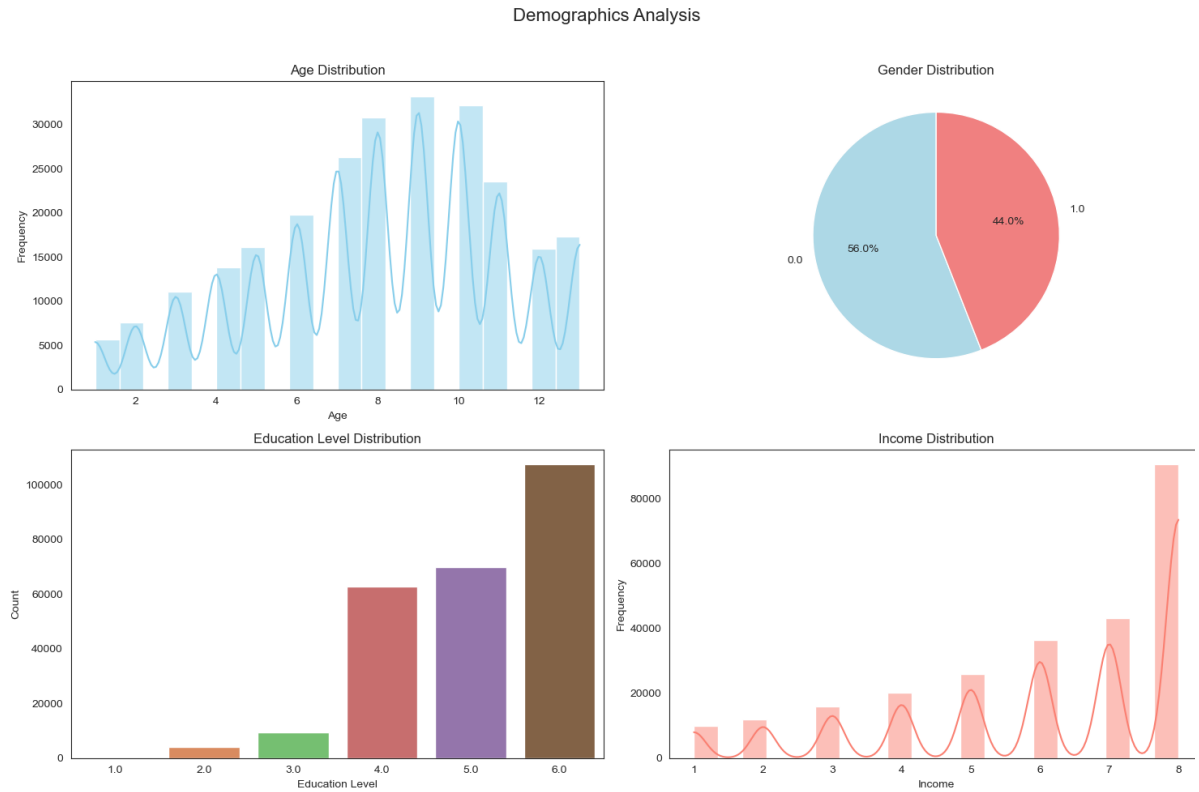

Figure 1

Figure 1 is a histogram depicting age distribution, complemented by a Kernel Density Estimate (KDE) indicating clusters of prevalence among certain age groups. We explore the age groups in 3-level age category. For example 1 = 18-24 9 = 60-64 13 = 80 or older. This insight is crucial, as age is a non-modifiable risk factor for diabetes, and the presence of age-related peaks may correlate with different intervention needs across the lifespan.

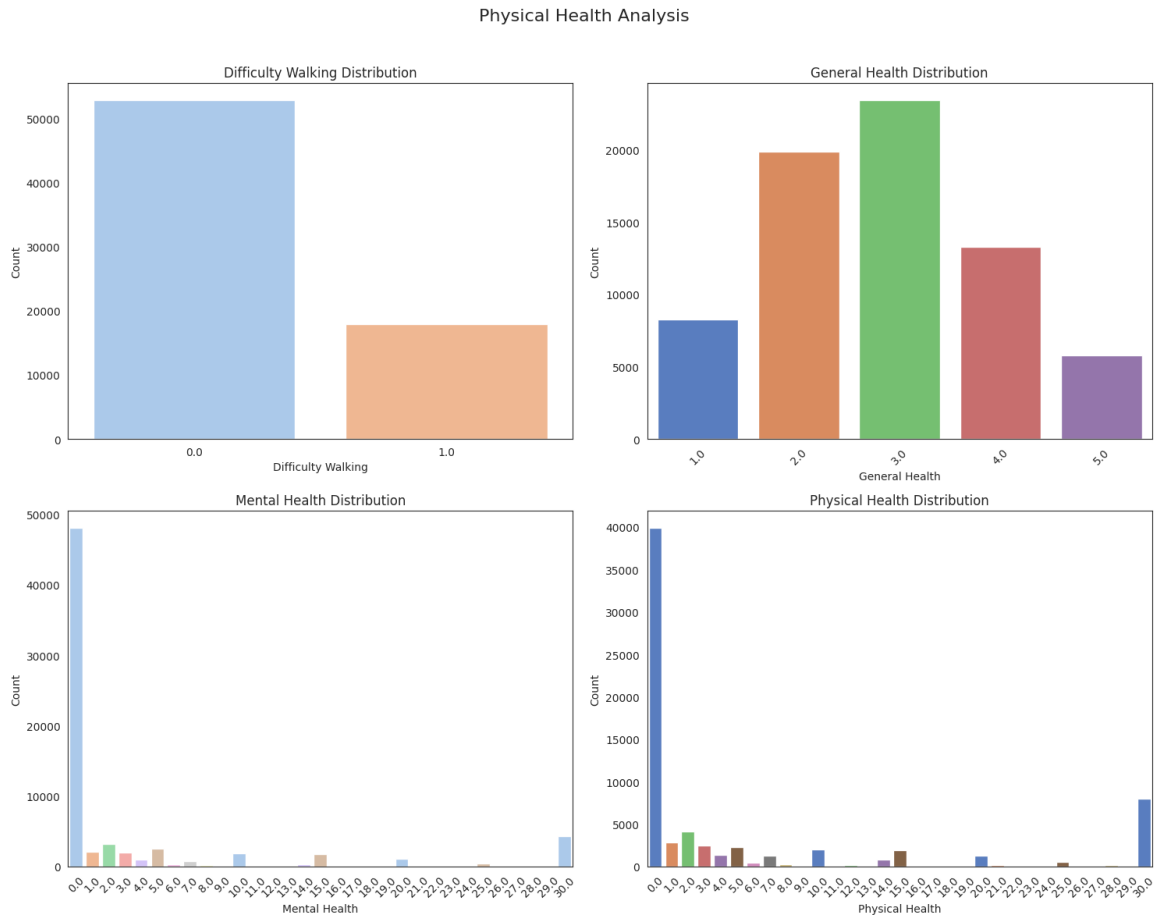

Figure 2

Figure 2 is a bar chart distinguished between two groups of physical activity levels. Given the established link between physical inactivity and increased diabetes risk, the graph showed a concerning number of individuals falling into the lower activity category. When it comes to mental health the participants of the survey were asked questions such as, Now thinking 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? While other graphs, such as those representing smoking habits, fruit and vegetable consumption, and alcohol intake, also provide valuable insights, the age and physical activity distributions were particularly telling.

We also explore the access to healthcare and services present to the participant to figure out whether the access to healthcare, medicines or doctor would some how detect the individuals chances of having diabetes. We saw that a majority portion of the population has access to healthcare and the results for the second graph were obtained by asking the question was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes.

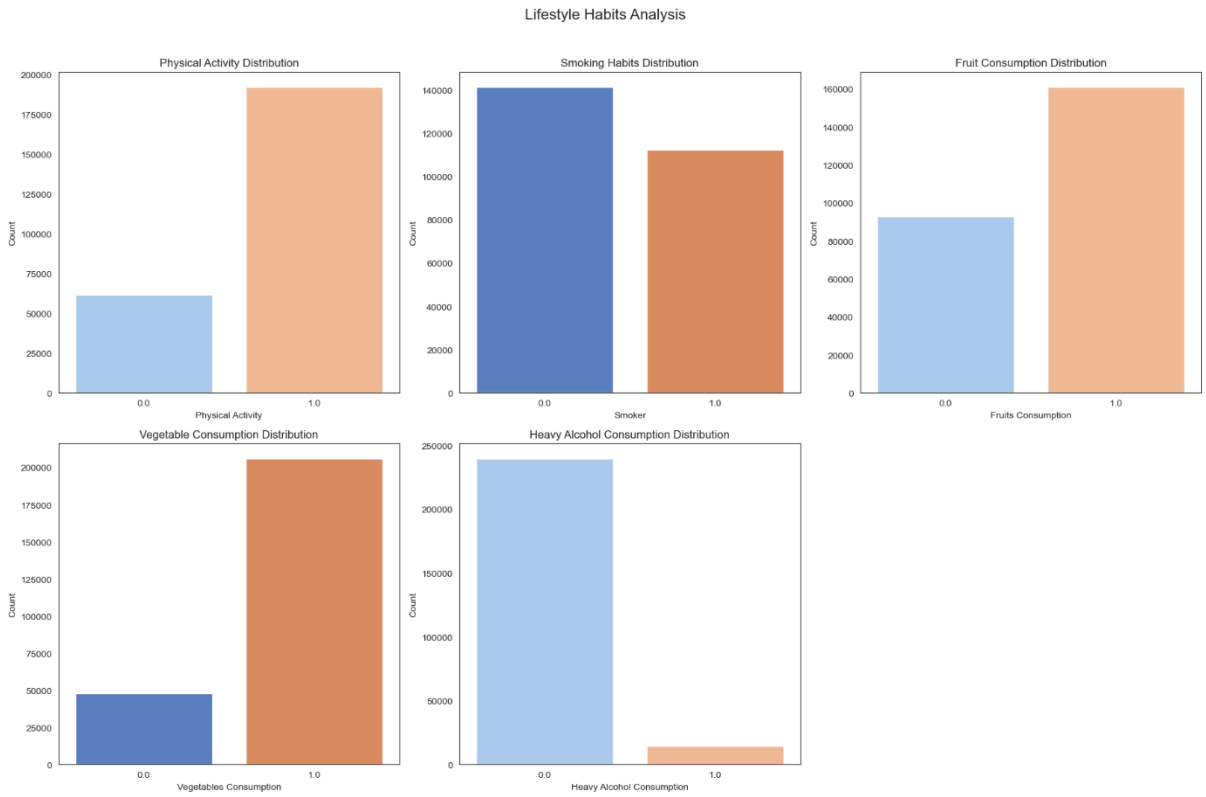

Figure 3

The image contains six bar charts representing the distribution of various lifestyle habits among a population. Each chart has two bars corresponding to two categories within each lifestyle habit:

Physical Activity Distribution: The data analysis reveals several key trends in lifestyle habits. There is a noticeable preference for higher levels of physical activity, as indicated by a larger count of individuals categorized under '1.0' compared to '0.0'. In terms of smoking habits, the data suggests a higher number of smokers than non-smokers or those with an alternative smoking status labeled '1.0'. Fruit and vegetable consumption patterns follow a similar trend, with more individuals regularly consuming fruits and vegetables, as represented by the taller bars under '1.0'. Contrasting these trends, the data on heavy alcohol consumption stands out distinctly. A significant portion of the surveyed population appears not to engage in heavy alcohol consumption, evidenced by a very tall left bar labeled '0.0' and an almost non-existent right bar for the '1.0' category. This implies either a very small number or complete absence of individuals in the higher alcohol consumption bracket.

Finally, we plot the correlation matrix heatmap, which is a graphical representation of the Pearson correlation coefficients between several variables.

Some key inferences from this heatmap include:

1. **High Blood Pressure (HighBP) and Age**: There is a relatively strong positive correlation (around 0.34), indicating that as age increases, the likelihood of having high blood pressure increases.

2. **General Health (GenHlth) and Physical Health (PhysHlth)**: There is a strong positive correlation (above 0.5), suggesting that individuals who report better general health also report better physical health.

3. **Difficult Walking (DiffWalk) and Age**: A moderate positive correlation (around 0.2), indicating that older individuals may have more difficulty walking.

4. **Smoking and all other variables**: The correlations are quite low, implying that smoking does not have a strong linear relationship with the other health-related variables in this dataset.

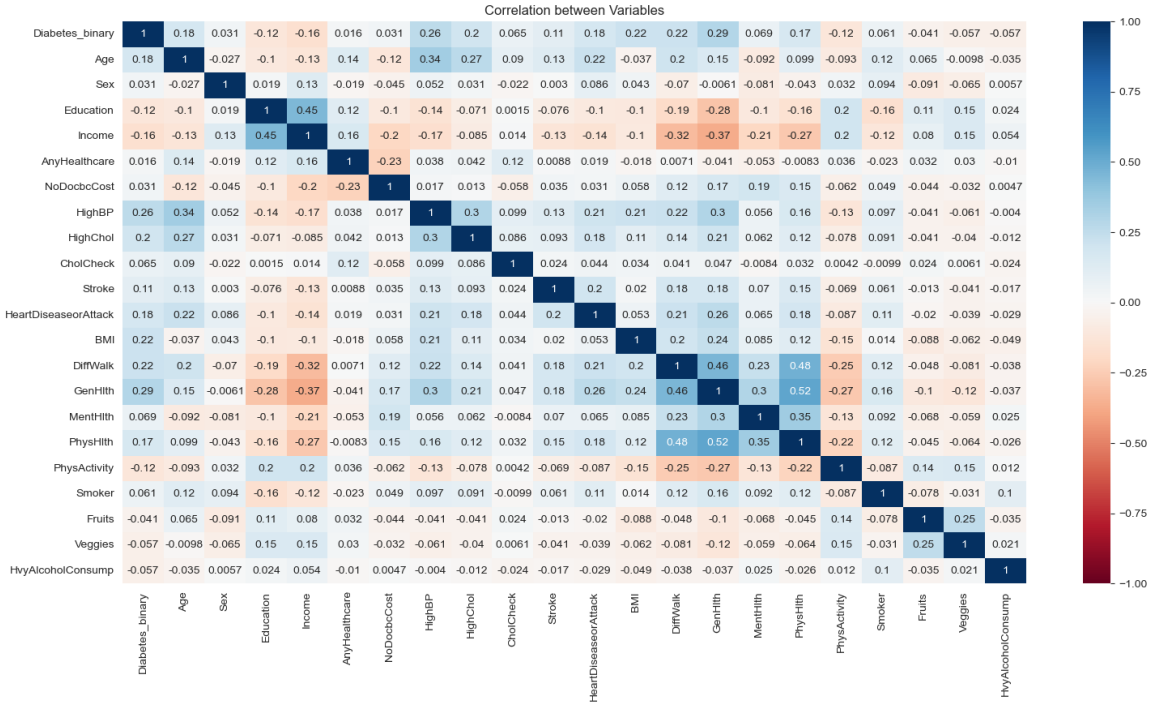

Figure 4

5. **Fruit and Vegetable Consumption**: Both have negative correlations with several health issues like diabetes, high blood pressure, and general health, which could suggest that higher consumption of fruits and vegetables is associated with better health outcomes.

6. **Income and Health Variables**: There are generally negative correlations with various health issues, indicating that higher income might be associated with better health outcomes.

7. **Diabetes (Diabetes_binary)**: Shows positive correlations with variables like HighBP, HighChol (High Cholesterol), and BMI, which could suggest that these factors are associated with the prevalence of diabetes.

3.2 Predictive analysis

We evaluated a range of machine learning models, from the straightforward Logistic Regression, SVM and k-NN to complex ensemble methods such as Random Forest, Decision Tree, and Naïve Bayes. The SVM model achieved the highest validation accuracy of 0.8613, conversely, the Naive Bayes model achieved the lowest accuracy of 0.7713.

This ranking indicates that the SVM model performed best in terms of validation accuracy, while the Naive Bayes model in its second instance had the lowest accuracy. It's important to note that while validation accuracy is a quick way to compare models, the classification report provides more detailed insights into how each model performs across different classes.

Here's a breakdown of what each part means:

- **Validation Accuracy**: The SVM model has an accuracy of 86.13%. This means that when the model was given a set of data to predict (the validation set), it correctly predicted the outcome (whether class 0 or class 1) for 86.13% of the instances.
- **Class Precision**: Precision for class 0 is 0.86. This means that when the model predicted an instance to be class 0, it was correct 86% of the time. Precision for class 1 is 0.66. This means that when the model predicted an instance to be class 1, it was correct 66% of the time.
- **Class Recall**: Recall for class 0 is 1.00, which is the highest possible score. This means that the model identified 100% of the actual class 0 instances correctly. Recall for class 1 is 0.02. This is very low, indicating that the model identified only 2% of the actual class 1 instances correctly.
- **Class F1-Score**: The F1-score for class 0 is 0.93. The F1-score is a measure of a test's accuracy that considers both the precision and the recall to compute the score. A score of 0.93 is quite high and suggests a good balance between precision and recall for class 0. The F1-score for class 1 is 0.03. This low score indicates that the model is not doing well on class 1 in terms of both precision and recall.
- **Class Support**: The support for class 0 is 43,645. This is the number of actual occurrences of class 0 in the dataset. The support for class 1 is 7,091. This is the number of actual occurrences of class 1 in the dataset.

In summary, the SVM model performs very well on class 0 but performs poorly on class 1, with very low recall and F1-score. The high overall accuracy is largely driven by the model's performance on class 0, which is the majority class.

3.3 Feature importance

Complementing our Findings from the EDA, the feature importance graph illustrates the relative importance of various factors in predicting diabetes. BMI stands out as the most influential feature, which aligns with existing literature that acknowledges the strong association between body weight and diabetes risk. The graph strikingly highlights BMI, Age, and Physical Health as the top determinants in diabetes risk, validating the emphasis on age and activity in our EDA .

The image displays a horizontal bar chart illustrating the permutation feature importance for a Support Vector Classifier (SVC) model. The permutation feature importance is a technique for calculating the importance of features by observing how random re-shuffling (permutation) of each feature affects the model performance.

1. **BMI as a Key Predictor**: 'BMI' (Body Mass Index) appears to be the most significant predictor for the model, given its longest bar on the chart. This implies that changes in the BMI feature greatly affect the model's accuracy.

2. **General Health and Smoking**: 'GenHlth' (general health) and 'Smoker' status are the next most important features, suggesting they also have a considerable impact on the model's predictions.

3. **Socioeconomic Factors**: 'Income' and 'Education' show moderate importance, indicating that these socioeconomic factors contribute to the model's decision-making process to a certain degree.

4. **Health-Related Factors**: Other health-related factors, such as 'HighBP' (high blood pressure), 'DiffWalk' (difficulty walking), 'MentHlth' (mental health), and 'HighChol' (high cholesterol), are also influential, but less so compared to 'BMI', 'GenHlth', and 'Smoker'.

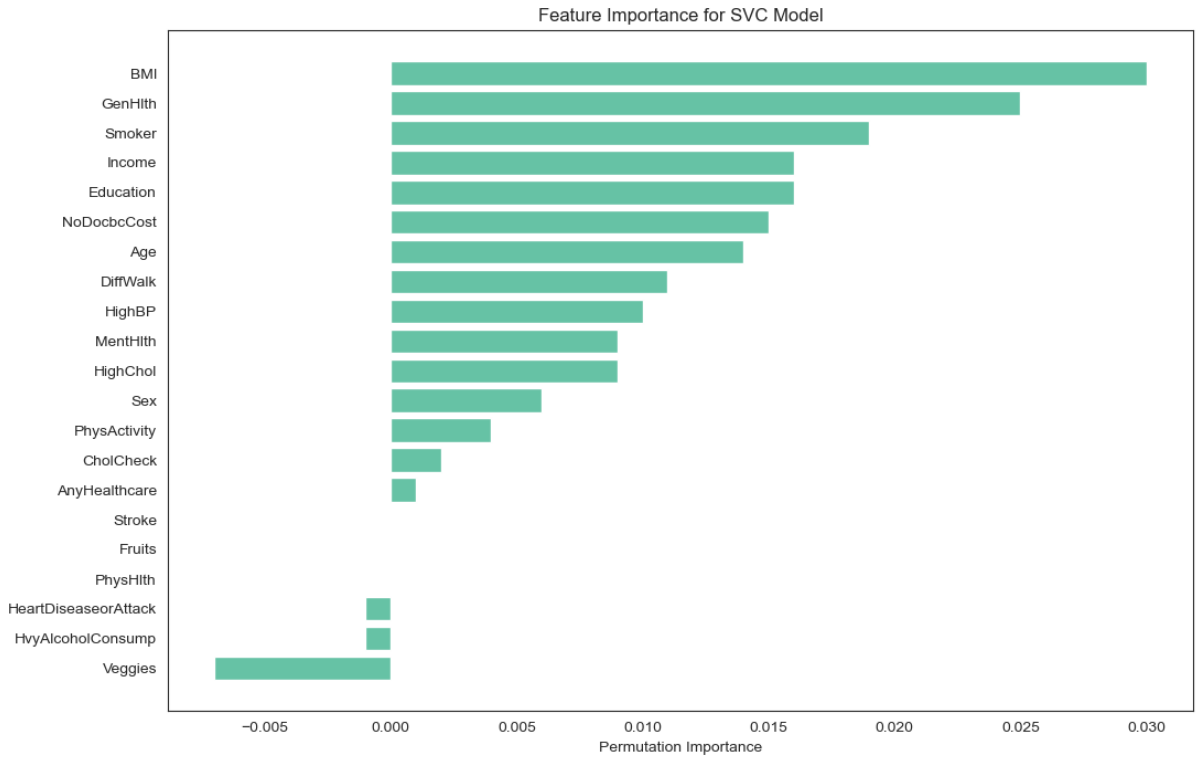

Figure 5

5. **Least Important Features**: Towards the bottom of the chart, features like 'Veggies' (vegetable consumption), 'HvyAlcoholConsump' (heavy alcohol consumption), and 'HeartDiseaseorAttack' have the smallest bars, suggesting they have the least influence on the model out of the features listed.

6. **Scale of Importance**: The scale is quite granular, with the highest importance being just above 0.030. This suggests that the model's performance is sensitive to small changes in the feature values, or that the model does not rely heavily on any single feature.

7. **Negative Importance**: None of the features show a negative importance, which would indicate a decrease in model performance when the feature is shuffled. All features contribute positively to some extent.

This chart is useful for identifying which features might be most valuable for predictive accuracy in the SVC model and could be prioritized for data collection or feature engineering efforts. However, it's important to note that while permutation importance can provide insight into the predictive power of each feature, it does not necessarily imply causation.

This type of chart is typically used to illustrate how different features (or independent variables) influence the prediction of a model, with the length and direction of the bars indicating the magnitude and direction of the feature's coefficient in the model.

Here are some inferences that can be drawn from this chart:

1. **Strong Positive Influence**: Features with bars extending to the right have a positive influence on the model's predictions. For instance, 'CholCheck' (cholesterol check) seems to have the strongest positive impact, suggesting that if an individual has had their cholesterol checked, it increases the likelihood of the outcome the model is predicting.

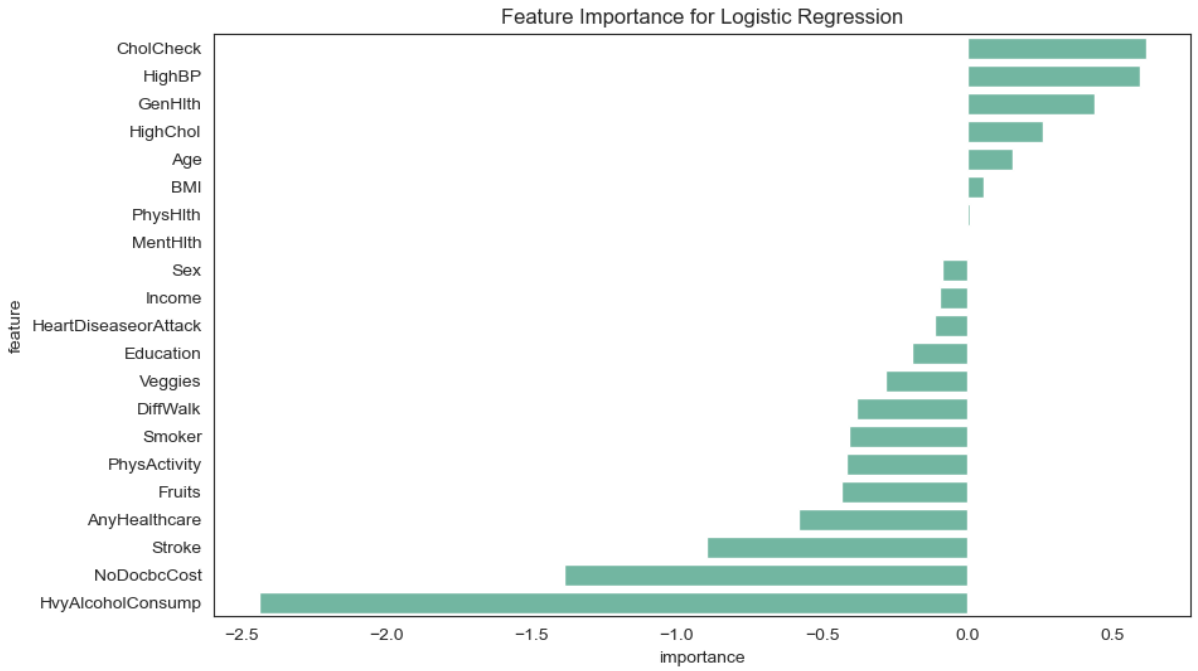

Figure 6

2. **Strong Negative Influence**: Conversely, features with bars extending to the left, such as 'HvyAlcoholConsump' (heavy alcohol consumption), negatively influence the prediction. This implies that higher values for heavy alcohol consumption decrease the likelihood of the outcome being predicted by the model.

3. **Health-Related Features**: Several health-related features like 'HighBP' (high blood pressure), 'GenHlth' (general health), 'HighChol' (high cholesterol), 'Age', and 'BMI' (Body Mass Index) are among the most influential, which suggests that these factors are critical in the model's decision-making process.

4. **Lifestyle Features**: Lifestyle choices such as 'PhysActivity' (physical activity), 'Fruits', and 'Veggies' (vegetable consumption) have less influence compared to health-related features but are still significant. Interestingly, 'Smoker' has a smaller negative influence, and 'Fruits' has a small positive influence.

5. **Socioeconomic Features**: Features like 'Income' and 'Education' also show up in the chart, indicating that socioeconomic factors have some predictive power in the model, albeit less than medical or lifestyle factors.

6. **Access to Care Features**: 'AnyHealthcare' (access to any healthcare) and 'NoDocbcCost' (not seeing a doctor because of cost) seem to have a small positive and negative influence, respectively, indicating that access to healthcare and financial barriers to healthcare play a role in the model's predictions.

The chart does not provide a clear indication of what outcome the Logistic Regression model is predicting, but the features and their importance suggest it could be related to predicting a health-related

event or condition. It is also important to note that the feature importance in a Logistic Regression model is based on the coefficients obtained from the model, which are dependent on the scale of the feature unless standardized, and the model's specific structure and regularizations applied.

4 Results

Our analytical models underwent rigorous validation processes to determine their accuracy in predicting diabetes. The SVM model demonstrated the highest validation accuracy at 86.13%, narrowly surpassing the Logistic Regression and Random Forest models, which both showed strong performances with accuracies of 86.01% and 85.98%, respectively.

The k-NN algorithm, known for its simplicity, performed admirably, securing an accuracy of 84.90%. However, it was the Naive Bayes model that, despite having the lowest overall accuracy of 77.13%, revealed a notable balance in identifying the positive cases, which could be particularly useful in early screening contexts where sensitivity is crucial.

The Decision Tree model, while being the most interpretable, yielded an accuracy of 79.42%, suggesting that the model might be too simplistic to capture the complex patterns present in the data or could be prone to overfitting.

5 Discussion

Based on the comprehensive analysis conducted in this study, we can draw the following conclusions regarding diabetes prediction using machine learning models:

Model Performance: The Support Vector Machine (SVM) model outperformed other machine learning algorithms with the highest validation accuracy of 86.13%, indicating its robustness in predicting diabetes risk. This high accuracy demonstrates the SVM's capability in handling the complex patterns associated with the onset of diabetes.

Feature Importance: Body Mass Index (BMI) emerged as the most influential predictor, underscoring the critical role of body weight in diabetes risk. Age and general health status also proved to be significant factors. These findings reinforce the importance of considering physiological and demographic factors in diabetes risk assessments.

Lifestyle and Health Indicators: Physical activity levels, dietary habits, and smoking status were identified as key lifestyle indicators that impact diabetes risk. The study suggests that interventions promoting physical activity and healthy eating, along with smoking cessation, may effectively reduce the risk of developing diabetes.

Socioeconomic Influence: Income and education showed a moderate level of importance, suggesting that socioeconomic status has a considerable impact on an individual's risk of diabetes. This highlights the need for targeted diabetes prevention efforts in lower-income and less educated populations.

Healthcare Access: Access to healthcare services was found to have a relationship with diabetes risk, with individuals lacking healthcare access being more vulnerable. This aspect points to the potential benefits of improving healthcare accessibility for effective diabetes prevention.

In summary, the study successfully leverages machine learning to identify individuals at high risk of diabetes and underscores the significance of various predictors, including BMI, age, general health, and lifestyle factors. The insights gained from this study could be instrumental in guiding public health strategies and individual lifestyle modifications to mitigate the growing burden of diabetes.

5.1 Tradeoffs and Consideration

The endeavor to leverage machine learning for diabetes prediction is fraught with challenges, particularly when handling vast datasets and contending with resource constraints. Our study

encountered two primary issues: managing large data volumes and overcoming limited computational resources.

1. Data Volume Management:
The substantial size of our dataset, comprising over a quarter-million records, posed significant challenges. Large datasets enhance the potential for uncovering nuanced patterns and relationships; however, they also demand considerable memory and processing power. To address this, we implemented data reduction techniques such as feature selection and dimensionality reduction, which helped to mitigate memory issues but also introduced the trade-off of potentially overlooking relevant predictors.

2. Computational Resource Limitations:
Our limited access to high-performance computing resources necessitated trade-offs between model complexity and computational feasibility. While more complex models, such as deep learning networks, might provide improved accuracy, they were computationally infeasible. Consequently, we focused on models that strike a balance between predictive power and resource efficiency, such as SVM and logistic regression.

3. Model Complexity:
The complexity of a model often correlates with its predictive accuracy. However, more complex models can be less interpretable—a crucial factor in healthcare applications where understanding the rationale behind predictions is as important as the predictions themselves. Therefore, we had to consider simpler models that allowed for greater interpretability at the cost of some predictive accuracy.

In summary, while managing large datasets with limited resources, it is crucial to make informed decisions that consider the trade-offs between model accuracy, complexity, interpretability, and computational demands. The choices made in this regard can significantly impact the applicability and efficacy of machine learning in real-world settings such as diabetes risk prediction.

5.2 Future Work
Our study has laid the groundwork for several potential avenues of exploration that could further enhance the predictive modeling of diabetes. Future work should focus on addressing the limitations encountered in the current study and exploring innovative approaches to diabetes prediction. Here are the key areas for future research:

1. Advanced Computational Resources:
To overcome the limitations posed by large datasets, future studies could leverage more advanced computational resources such as cloud computing services, which offer scalable and elastic computing capabilities. Utilizing such resources could allow the deployment of more complex models like deep learning networks, which may uncover deeper insights from the data.

2. Incorporation of Additional Data Sources:
Expanding the dataset to include additional variables such as genetic markers, environmental factors, and more detailed dietary patterns could enhance the model's predictive power. Integrating electronic health records (EHRs) and data from wearable health devices could also provide more personalized risk assessments.

3. Interdisciplinary Collaboration:
Collaborating with healthcare professionals and data scientists from diverse fields can introduce new perspectives and expertise in the model development process. This interdisciplinary approach could lead to the creation of more holistic and patient-centric models.

4. Model Interpretability and Explainability:

Future research should also prioritize improving the interpretability of complex models. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be utilized to make the model's decision-making process more transparent to end-users, particularly healthcare professionals.

5. Personalized Diabetes Risk Predictions:
Tailoring models to provide individualized risk predictions based on personalized health profiles could significantly impact preventive healthcare strategies. This personalized approach would account for the unique health trajectories of individuals.

In conclusion, while this study has made significant strides in diabetes prediction, the path forward offers rich opportunities for innovation and improvement. Embracing these opportunities can lead to more accurate, actionable, and personalized tools for diabetes prevention and management.
The study, while comprehensive, has limitations. The dataset, though large, is limited to a specific population (US citizens), which might affect the generalizability of the models. Future studies could include more diverse datasets to enhance the models' applicability across different demographics. Additionally, integrating more complex models like deep learning could be explored to potentially improve prediction accuracy.

6 Author Contribution

Manav Mandal and Atharva Gurav collaboratively contributed to the conception and design of this study. Their contributions are detailed as follows:

Manav Mandal

Conceptualization and Design: Played a pivotal role in conceptualizing the study and formulating the research questions.
Data Preprocessing and Analysis: Led the data preprocessing efforts, ensuring the dataset's integrity and suitability for analysis.
Model Training and Evaluation: Took charge of training the machine learning models, meticulously tuning their parameters for optimal performance.
Interpretation of Results: Contributed significantly to interpreting the results, drawing meaningful conclusions about the models' performance and their implications for diabetes prediction.

Atharva Gurav

Literature Review and Background Research: Conducted an extensive review of existing literature, providing a solid theoretical foundation for the study.
Data Visualization and Reporting: Spear headed the creation of data visualizations, elucidating complex data patterns and model outcomes.
Statistical Analysis and Model Validation: Focused on the statistical analysis of the models, ensuring the validity and reliability of the findings.
Drafting the Report: Played a crucial role in drafting and revising the final report, ensuring clarity, coherence, and adherence to academic standards.

7 References

[1] Firdous, S., Wagai, G. A., & Sharma, K. (2022). A survey on diabetes risk prediction using machine learning approaches. Journal of Family Medicine and Primary Care, 11(11), 6929.

[2] Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., ... & Ren, Z. (2022). Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. International Journal of Environmental Research and Public Health, 19(22), 15027.

[3] Chou, C. Y., Hsu, D. Y., & Chou, C. H. (2023). Predicting the onset of diabetes with machine learning methods. Journal of Personalized Medicine, 13(3), 406.

[4] Dritsas E, Trigka M. Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. Sensors (Basel). 2022 Jul 15;22(14):5304. doi: 10.3390/s22145304. PMID: 35890983; PMCID: PMC9318204.

[5] Kuhn MJ, Chen N, Sahani DV, Reimer D, van Beek EJ, Heiken JP, So GJ. The PREDICT study: a randomized double-blind comparison of contrast-induced nephropathy after low- or isoosmolar contrast agent exposure. AJR Am J Roentgenol. 2008 Jul;191(1):151-7. doi: 10.2214/AJR.07.3370. PMID: 18562739.

[6] Alva ML, Chakkalakal RJ, Moin T, Galaviz KI. The Diabetes Prevention Gap And Opportunities To Increase Participation In Effective Interventions. Health Aff (Millwood). 2022 Jul;41(7):971-979. doi: 10.1377/hlthaff.2022.00259. Epub 2022 Jun 27. PMID: 35759735; PMCID: PMC10112939.