Predicting Diabetes Risk: Identifying Risk Factors

University of Texas at Arlington

ASDS 6303 Data Mining with Info Visual

Introduction

Research Question:

How do demographic, behavioral, and health-related variables contribute to the likelihood of developing diabetes or prediabetes, and which factors have the strongest predictive power?

Relevance and Importance

Diabetes is a major global public health challenge, with significant implications for healthcare systems, patient quality of life, and economic burdens. 136 million adults are currently living with diabetes or prediabetes in the United States alone, according to the CDC in 2024. Additionally, diabetes costs the United States economy $413 billion annually, which includes medical costs and productivity. Even more so, 1 in 5 diabetics and 8 in 10 prediabetics are unaware of their condition. Understanding the predictive factors associated with diabetes can help address several critical issues plaguing the United States:

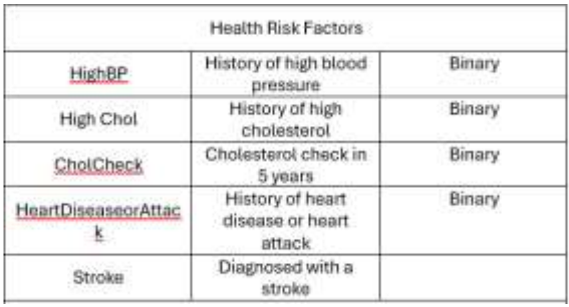
1. **Preventive interventions:** By being able to identify high-risk patients, we can help guide early screening programs and lifestyle interventions. This would enable health care providers to reduce diabetes onset and potentially dangerous complications.
2. **Public Health Strategies**: Insights into demographic information such as Age, or BM, can help inform policy decisions targeting specific populations. Furthermore, this would enable resource allocation to underserved groups such as locations with food deserts.
3. **Health Equity**: Addressing disparities in risk factors such as income and education can provide more equitable healthcare outcomes by tailoring strategies to vulnerable populations. Additionally, it would help for a push in a more comprehensive health education.
4. **Data-Driven Decision Making**: By using machine learning models and traditional statistical models to enhance accuracy of predictions, we can provide actionable insights and functional visualizations for healthcare practitioners and policymakers.

By exploring the relationship between these variables, this study not only informs diabetes prevention and treatment efforts but also contributes to broader conversations about chronic disease management and health equity.

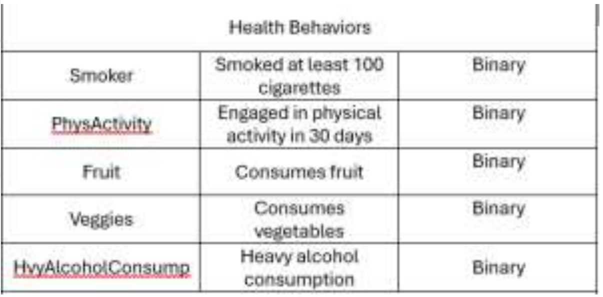
Data Overview

Data Source:

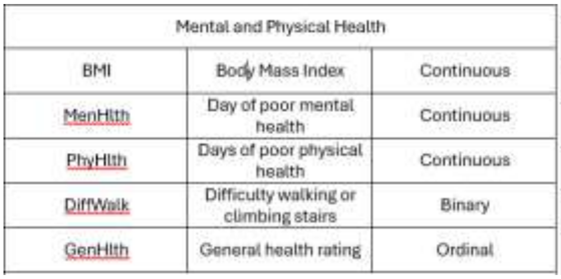
The dataset comes from the publicly available CDC Behavioral Risk Factor Surveillance System (BRFSS) from 2015. This data is comprised of survey data, where it is a mix of some continuous variables, but primarily categorical variables. To further expand on these variables, they comprise of demographic, behavioral, and health-related factors, specifically linked to diabetes and prediabetes. For a more in-depth description, refer to the table below:



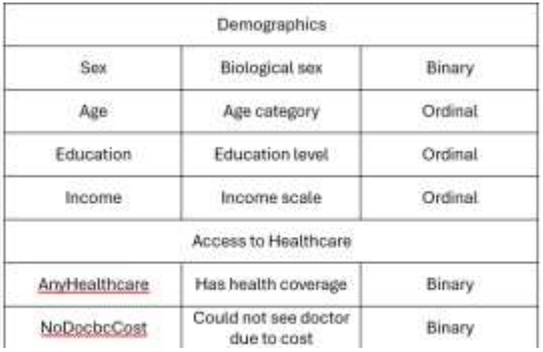
Health Risk Factors		
HighBP	History of high blood pressure	Binary
High Chol	History of high cholesterol	Binary
CholCheck	Cholesterol check in 5 years	Binary
HeartDiseaseorAttack	History of heart disease or heart attack	Binary
Stroke	Diagnosed with a stroke	



Health Behaviors		
Smoker	Smoked at least 100 cigarettes	Binary
PhysActivity	Engaged in physical activity in 30 days	Binary
Fruit	Consumes fruit	Binary
Veggies	Consumes vegetables	Binary
HvyAlcoholConsump	Heavy alcohol consumption	Binary



Mental and Physical Health		
BMI	Body Mass Index	Continuous
MenHlth	Day of poor mental health	Continuous
PhyHlth	Days of poor physical health	Continuous
DiffWalk	Difficulty walking or climbing stairs	Binary
GenHlth	General health rating	Ordinal



Demographics		
Sex	Biological sex	Binary
Age	Age category	Ordinal
Education	Education level	Ordinal
Income	Income scale	Ordinal
Access to Healthcare		
AnyHealthcare	Has health coverage	Binary
NoDocbcCost	Could not see doctor due to cost	Binary

Target Variable		
Diabetes_binary	Presence of diabetes	Binary

Data Preparation and Exploration

Preprocessing steps:

For our data preprocessing steps, we took a stepwise approach to make sure we covered each area before our exploratory data analysis:

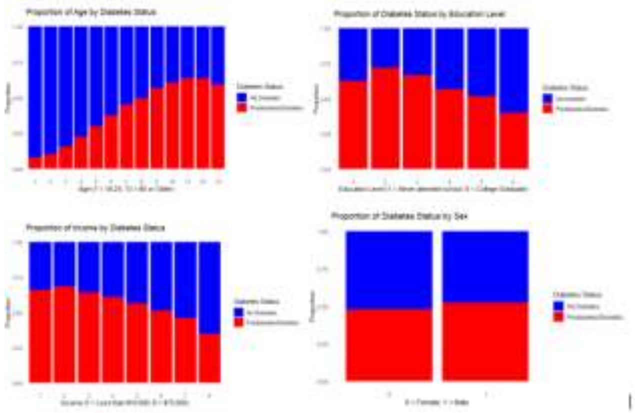
1. Cleaning:
    a. No duplicate entries were found
    b. No missing values were found
2. Outlier Detection and Removal:
    a. Outliers were found in BMI, but we decided to keep them because of clinical relevance. Outliers might reveal true diabetic cases in the United States.
3. Transformations:
    a. Categorical variables were transformed as factors, while ordinal variables were handled as factors with appropriate levels.
4. Balancing Classes:

a. The binary classes were both balanced, so no balancing techniques were necessary.

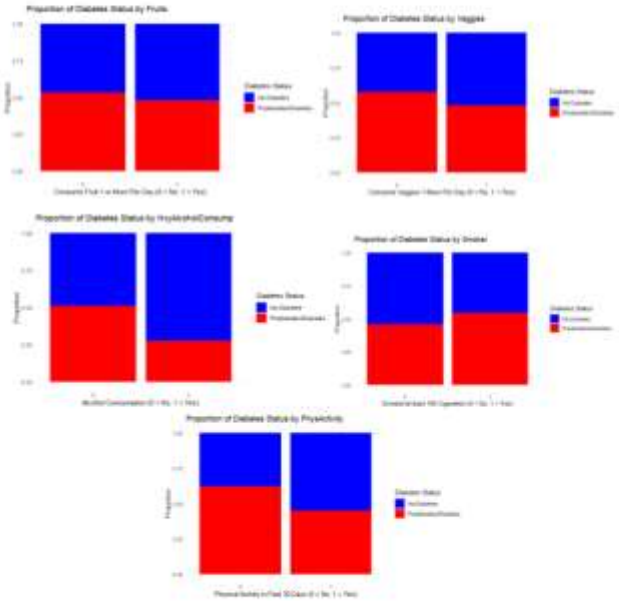
<u>Exploratory Data Analysis:</u>

Continuing with EDA, we aimed to investigate relationships between key variables. To keep our observations concise, we grouped some variables together under new categories purely to visualize the relationships with the presence of diabetes:

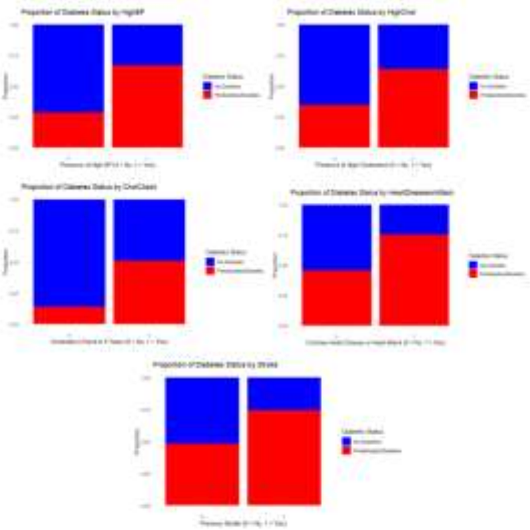
1. *Demographics:* Diabetes prevalence seems to increase with age and decreases with higher education and income levels. Males seem to slightly more affected than females.



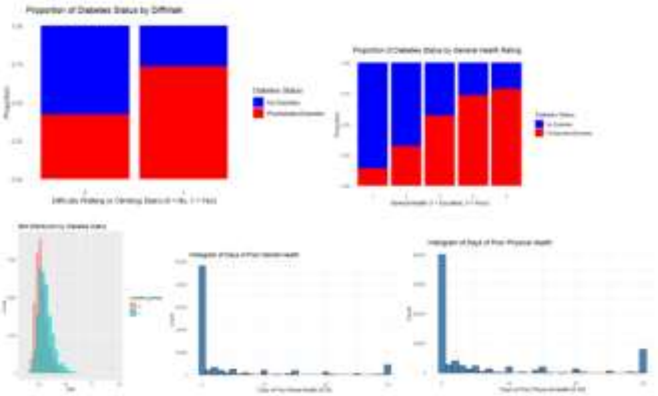
2. *Health Behaviors: Healthy* behaviors such as physical activity, and daily vegetable and fruit consumption are associated with lower diabetes prevalence. On the other hand, smoking seems to increase diabetes risk.



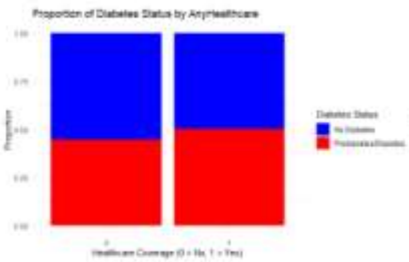
3. *Health Behaviors*: High blood pressure, high cholesterol, and cardiovascular diseases show a strong association with diabetes risk. Additionally, the presence of stroke also seems to have increased diabetes prevalence.

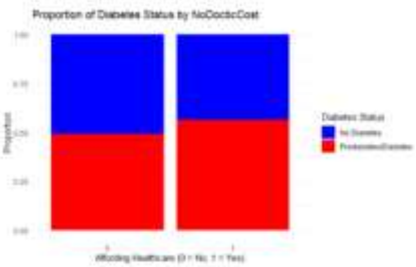


4. *Mental and Physical Health:* Indicators of physical and mental health, including mobility challenges, higher BMI and self-reported poorer general health, seem to be associated with diabetes prevalence. Although we have only the histograms of mental health and physical health, research would suggest that they both may also contribute to diabetes risk.



5. *Access to Health Care:* From the graphs alone, healthcare coverage does not seem to directly influence diabetes prevalence. However, affordability does seem to play a role.





Model Building and Performance

For our chosen models, we decided on three different models:

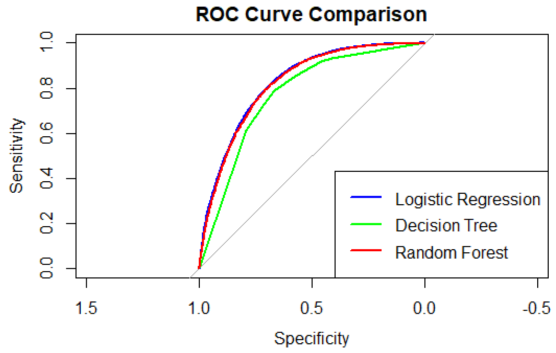
1. Logistic Regression: Chosen for its interpretability, allowing us to quantify the effect of predictors like BMI and age on diabetes risk. It was also our baseline model.
2. Decision Trees: Used to provide an intuitive visualization of risk factors and their interactions.
3. Random Forest: Chosen to handle non-linear relationships and rank predictors through feature importance metrics. It does this by combining multiple decision trees to improve predictive accuracy and reduce overfitting.

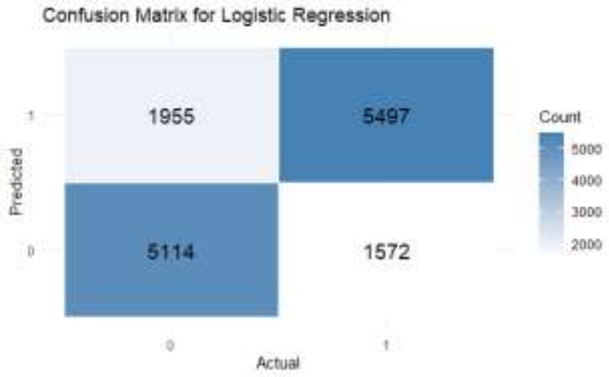
In terms of model evaluation, the Random Forest and Logistic Regression model performed similarly, with Decision Tree struggling with sensitivity. Based on all our models, we decided to take all of them into consideration, as they each offer insights into diabetes risk factors.

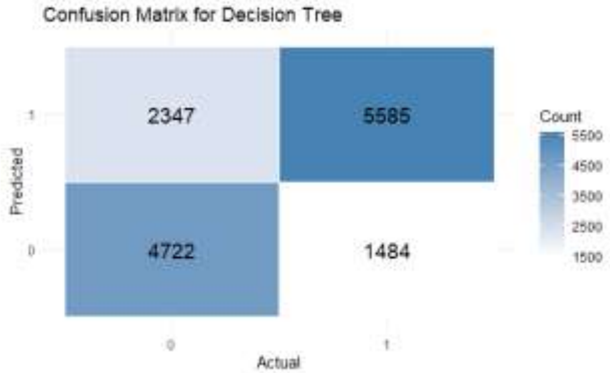
Models	Accuracy	Sensitivity	Specificity	F1-Score
Logistic Regression	**0.7505**	**0.7234**	**0.7776**	**0.7436**
Decision Tree	**0.7290**	**0.6680**	**0.7901**	**0.7114**
Random Forest	**0.7495**	**0.7145**	**0.7846**	**0.7405**

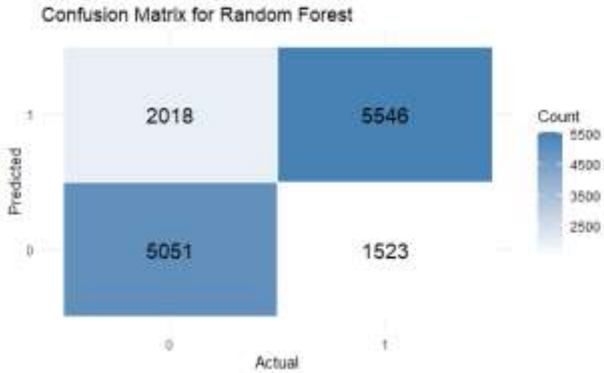
<u>ROC Curve and Confusion Matrices</u>

In our ROC Curve comparison, our Logistic Regression Model and Random Forest performed similarly, showcasing strong discriminatory power. This is again shown in our confusion matrices, where actual vs predicted had a good balance, with the models leaning more towards false positives vs false negatives.





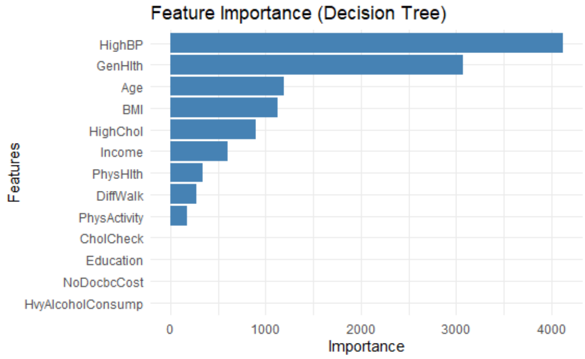


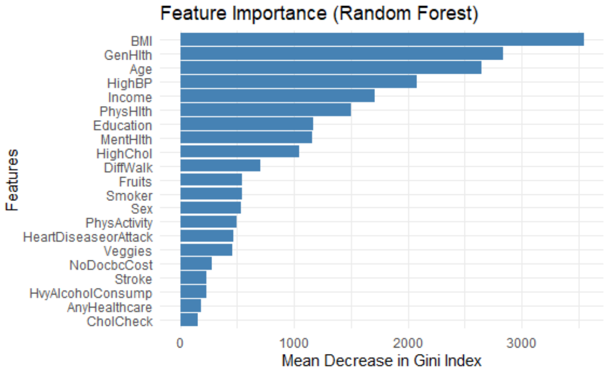


Results and Key Insights

Based on our models, diabetes had a few key insights:

1. Diabetes is heavily influenced by a mix of health factors and demographic elements such as BMI, High Blood Pressure, Poor Self Rated Health, Income, Education, Age and Gender.
2. Health behavior such as physical activity and diet showed modest impact
3. An unexpected inverse correlation as observed between heavy alcohol consumption and diabetes risk, which warrants further exploration.





Conclusion

This study explored the predictors of diabetes risk using a dataset from the CDC that specifically contained demographic health, and behavioral factors. Key findings revealed that health indicators such as BMI, High Blood Pressure, and Poor Self-Rated Health, as well as demographic factors like Age and Education, are significant contributors to diabetes risk. Behavior factors like physical activity and diet were found to have some modest impact. Interestingly, an inverse correlation was found between heavy alcohol consumption and diabetes risk, which is contrary to previous research, highlighting a further need to investigate. In terms of models, we decided on using all, but using the logistic regression as the forefront model. Logistic regression models are known to be easier to interpret, as well as providing tangible numbers that can be used to quantify risk and odds. However, using the other models to create visuals such as decision trees or feature importance can still be valuable as the relationship between variables is often non-linear.

Discussion and Limitations

While the study produced meaningful insights, there were still a few limitations that come from the nature of the dataset:

1. Self-Reported Data: Variables such as physical activity and alcohol consumption are self-reported, introducing potential biases or inaccuracies.
2. Limited context: The dataset lacks information regarding genetic predispositions and regional or environmental factors, which are crucial for a holistic understanding of diabetes risk.
3. Inverse correlation with Alcohol: The negative relationship between alcohol use and diabetes risk warrants caution, as it may result from confounding factors.

Recommendations:

Despite these limitations, based on our results we recommend the following:

1. Incorporate genetic markers or geographic information and explore the causal relationship behind observed trends.
2. Target high-risk groups and promote early screening and lifestyle intervention.
3. Emphasize the importance of health equity for low income and lower-education groups that are at risk of developing diabetes.
4. Encourage health fairs in locations with high risk once geographic information is incorporated but push for education.
5. Suggest a class that is required in public schools as an increase in education showed a decreased risk in developing diabetes.

References

American Diabetes Association; Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care* 1 May 2018; 41 (5): 917–928. https://doi.org/10.2337/dci18-0007

Centers for Disease Control and Prevention. (2015). *Behavioral Risk Factor Surveillance System (BRFSS)*. Retrieved from https://www.cdc.gov/brfss/

Centers for Disease Control and Prevention (CDC) (2010). Incidence of end-stage renal disease attributed to diabetes among persons with diagnosed diabetes --- United States and Puerto Rico, 1996-2007. *MMWR. Morbidity and mortality weekly report*, *59*(42), 1361–1366.

Centers for Disease Control and Prevention. (2024). *Data & research portal: Diabetes*. Retrieved from https://www.cdc.gov/diabetes/php/data-research/index.html

Huxley, R., Barzi, F., & Woodward, M. (2006). Excess risk of fatal coronary heart disease associated with diabetes in men and women: meta-analysis of 37 prospective cohort studies. *BMJ (Clinical research ed.)*, *332*(7533), 73–78. https://doi.org/10.1136/bmj.38678.389583.7C

Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., Nathan, D. M., & Diabetes Prevention Program Research Group (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *The New England journal of medicine*, *346*(6), 393–403. https://doi.org/10.1056/NEJMoa012512

University of California, Irvine. (n.d.). *CDC diabetes health indicators dataset*. Retrieved December 8, 2024, from https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators