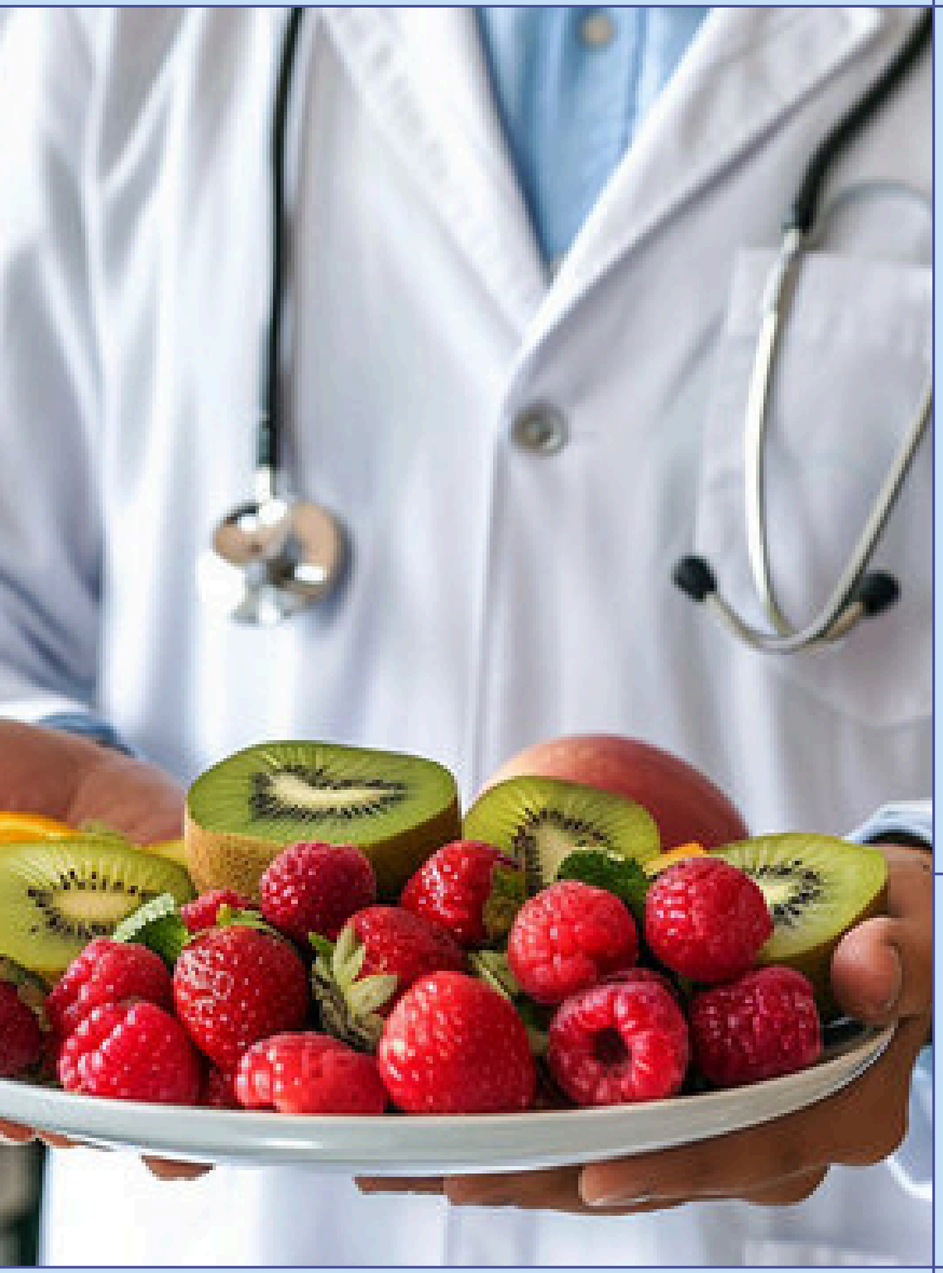


October 30, 2024

DIABETES & LIFESTYLE

Evaluating risks that can lead to diabetes

Jessica McCarty,
Cassandra Murray,
Armando Zamora,
Sarath Arja,
Alex King

Executive Summary

Our research aimed to determine if and how lifestyle factors impact the female population, focusing on identifying the most influential factors. This study builds on insights from our previous project

Source:
https://www.kaggle.com/datasets/prosperchuks/health-dataset.

Objective

We aim to uncover patterns and relationships between lifestyle choices and diabetes prevalence, providing insights that could potentially inform preventive health measures.

Approach

By employing machine learning models, including Random Forest classifiers, we identified key predictors of diabetes risk and evaluated the model’s accuracy in classifying individuals based on these lifestyle factors.

The findings from this analysis contribute to a better understanding of how daily habits and family history may influence diabetes risk, with implications for personalized healthcare and targeted interventions.

Comparison

MODEL	TRAINING SCORE	TESTING SCORE	ACCURACY	CONFUSION MATRIX	BALANCED ACCURACY SCORE
RandomForestClassifier	0.787	0.756	0.756	[[3667 936] [1405 35891]]	0.758
GradientBoostingClassifier	0.768	0.761	0.761	[[3596 1007] [1287 3707]]	0.762
AdaBoostClassifier	0.761	0.758	0.758	[[3508 1095] [1228 37661]]	0.758
DecisionTreeClassifier	0.955	0.686	0.686	[[2962 1641] [1374 36201]]	0.684
KNeighborsClassifier	0.759	0.756	0.756	[[3460 1143] [1194 38001]]	0.758

Model Choice

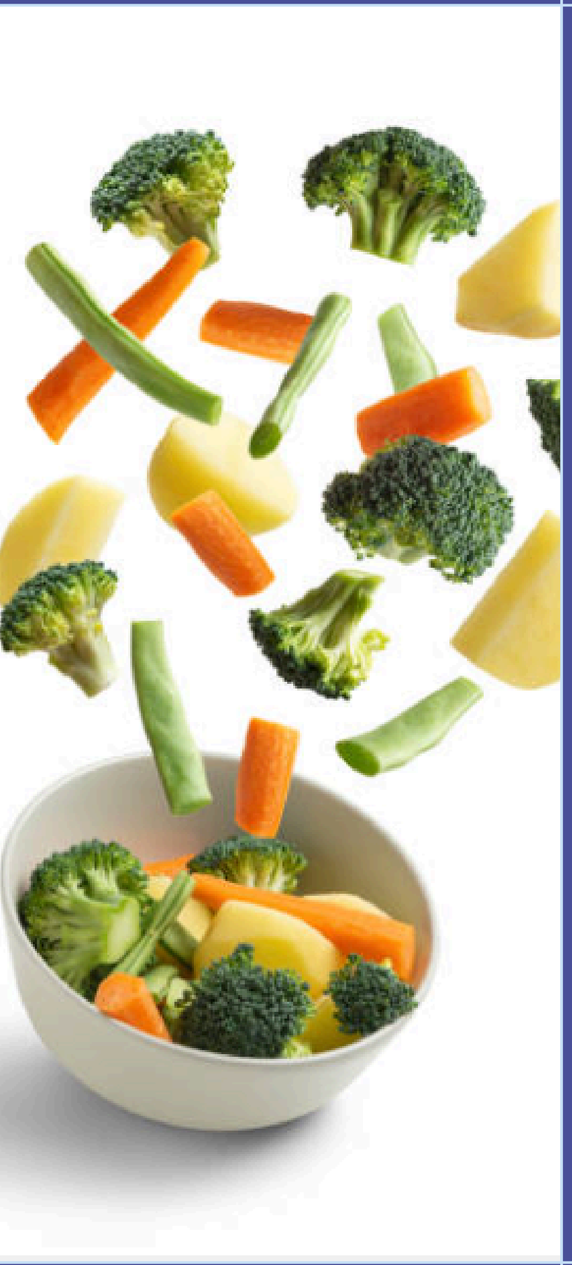
Evaluated various machine learning models to determine the best fit for our data.

Top Performers: Random Forest and Gradient Boosting emerged as the top two models, demonstrating the highest accuracy and reliability for our analysis.

Data Collection

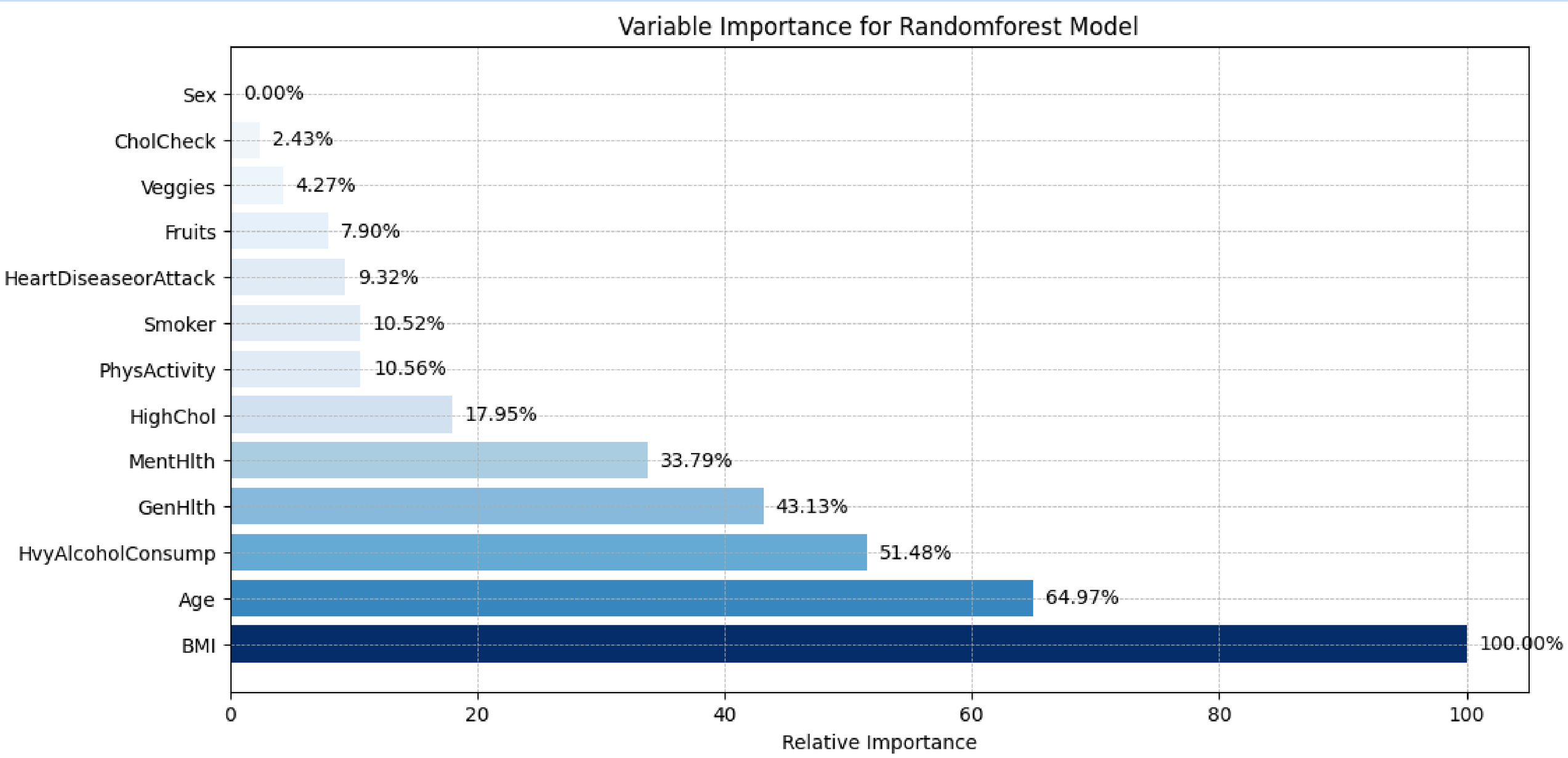
Feature ranking with recursive feature elimination

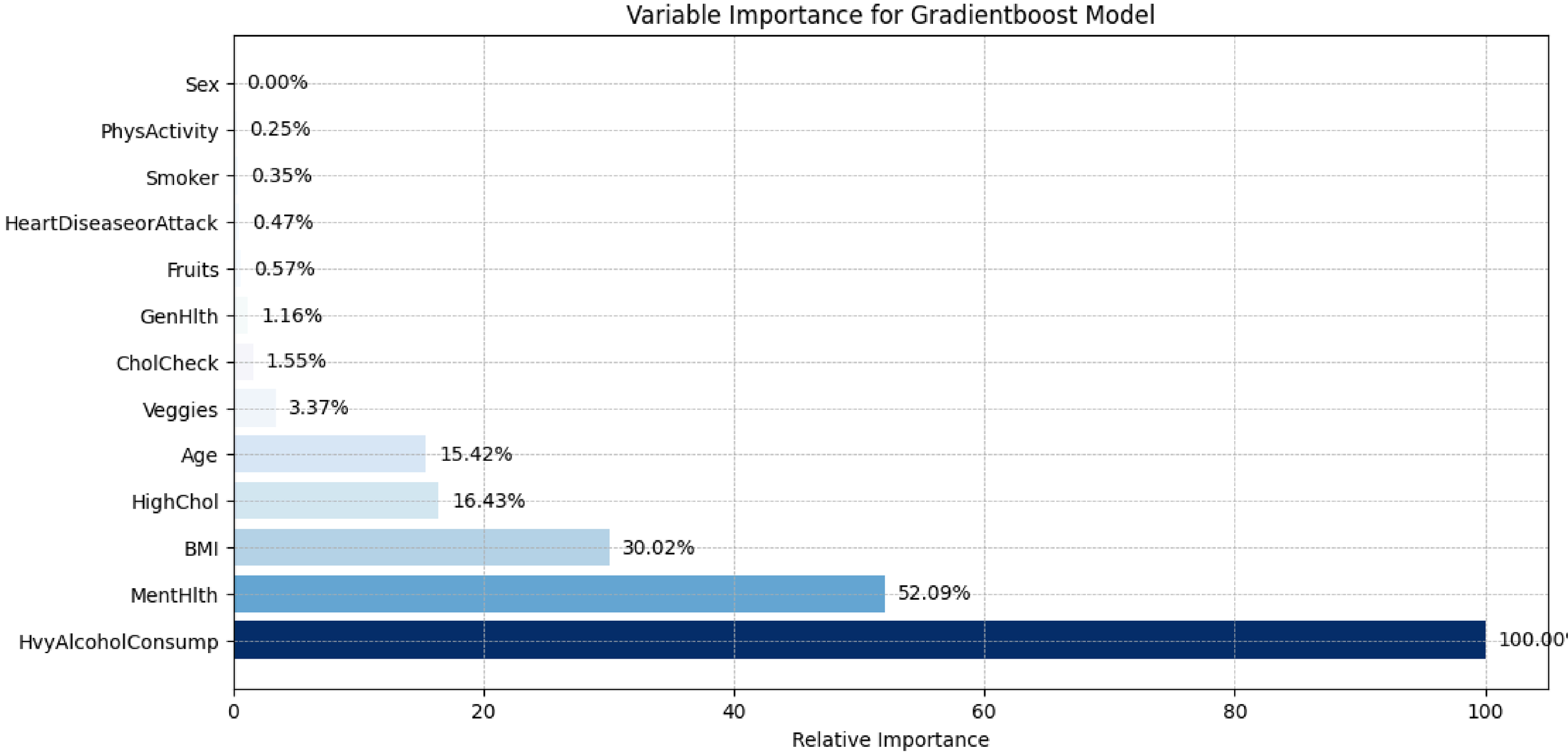
As part of our data optimization efforts, we implemented Recursive Feature Elimination to identify the top five ranked features based on their weight assignments. We then used these selected features to evaluate the performance of our models.

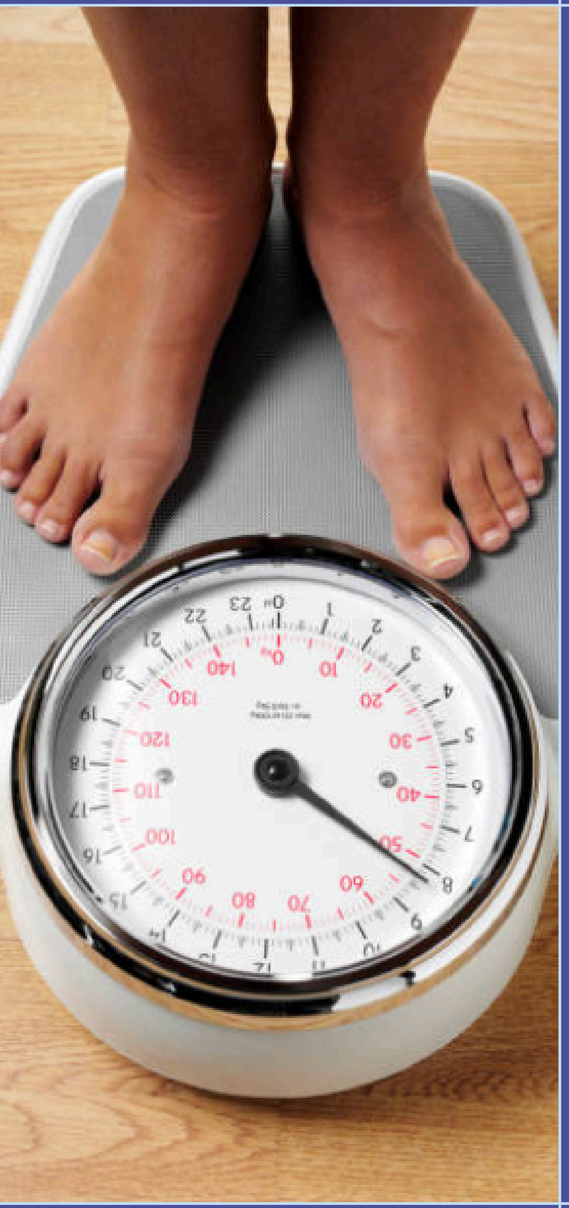


Data Collection

- Feature Evaluation: During preprocessing, we assessed the impact of removing certain health-related features, specifically those related to stroke and heart attack.
- Testing Outcomes: Excluding these features was tested but did not significantly enhance the model's performance.
- Decision: Based on testing results, we chose to retain these features in the dataset to ensure the model’s robustness and accuracy.
- Conclusion: This approach allowed for a more comprehensive analysis, supporting reliable and actionable insights.


Variable Importance for Randomforest Model
Sex
CholCheck
Veggies
Fruits
HeartDiseaseorAttack
Smoker
PhysActivity
HighChol
MentHlth
GenHlth
HvyAlcoholConsump
Age
BMI
0.00%
2.43%
4.27%
7.90%
9.32%
10.52%
10.56%
17.95%
33.79%
43.13%
51.48%
64.97%
100.00%
0
20
40
60
80
100
Relative Importance


Variable Importance for Gradientboost Model
Sex
PhysActivity
Smoker
HeartDiseaseorAttack
Fruits
GenHlth
CholCheck
Veggies
Age
HighChol
BMI
MentHlth
HvyAlcoholConsump
0.00%
0.25%
0.35%
0.47%
0.57%
1.16%
1.55%
3.37%
15.42%
16.43%
30.02%
52.09%
100.00
0
20
40
60
80
100
Relative Importance



Key Lifestyle Factors Identified

- BMI (Body Mass Index)
- Heavy Alcohol Consumption
- Age

Conclusion: These models effectively highlighted the primary lifestyle factors impacting the female population, guiding our focus for actionable insights.

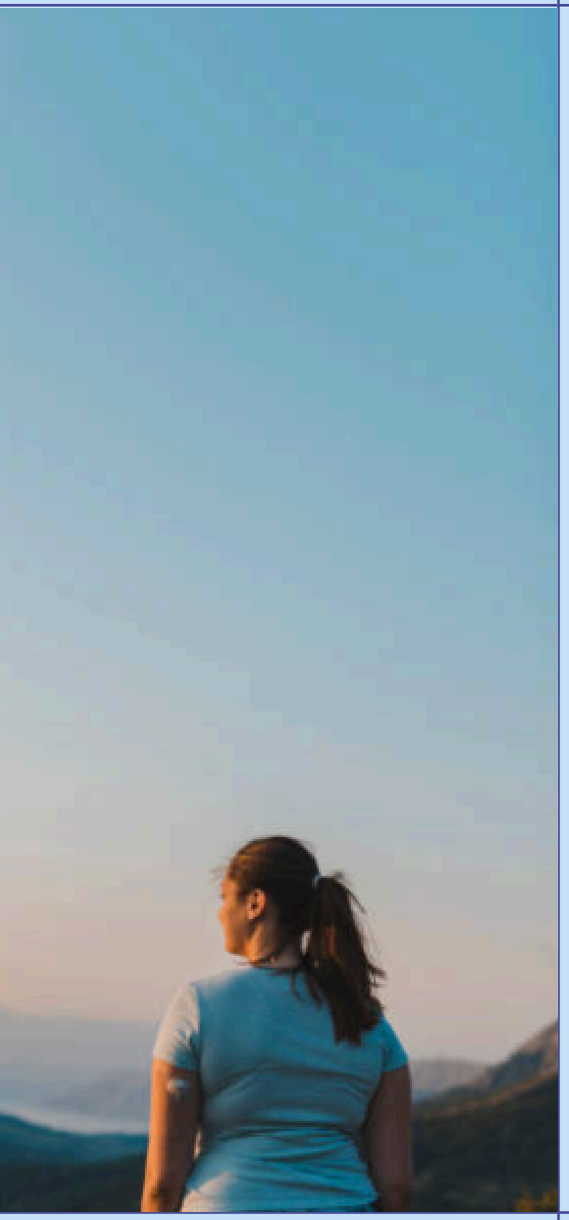
Conclusions

- Key Findings: All lifestyle factors influence the likelihood of developing diabetes, but some factors had a lesser impact than initially hypothesized.

- Unexpected Results: Factors like healthy eating, smoking, cholesterol, and physical activity showed a smaller impact individually than expected.

- Interconnected Effects: These lower-impact factors may still contribute indirectly by influencing major factors such as BMI, underscoring the complex interplay between lifestyle choices and health outcomes.

Let’s run the model



Future Research Directions

- Comparative Analysis: We aim to explore how lifestyle factors impact males versus females, prompted by our finding that male-only data produced a lower accuracy score than female-only data.

- Expanded Insights: By analyzing gender-specific effects, we aim to understand potential differences in lifestyle factor influences across genders.

- Objective: This comparative investigation could provide a more holistic view, enabling tailored health recommendations for both males and females.

- Expected Impact: Insights could inform gender-specific wellness programs and public health strategies.

Questions

Credits

Lead coders

Jessica McCarty
Cassandra Murray
Armando Zamora
Sarath Arja

Code validation

Cassandra Murray
Armando Zamora
Sarath Arja
Alex King

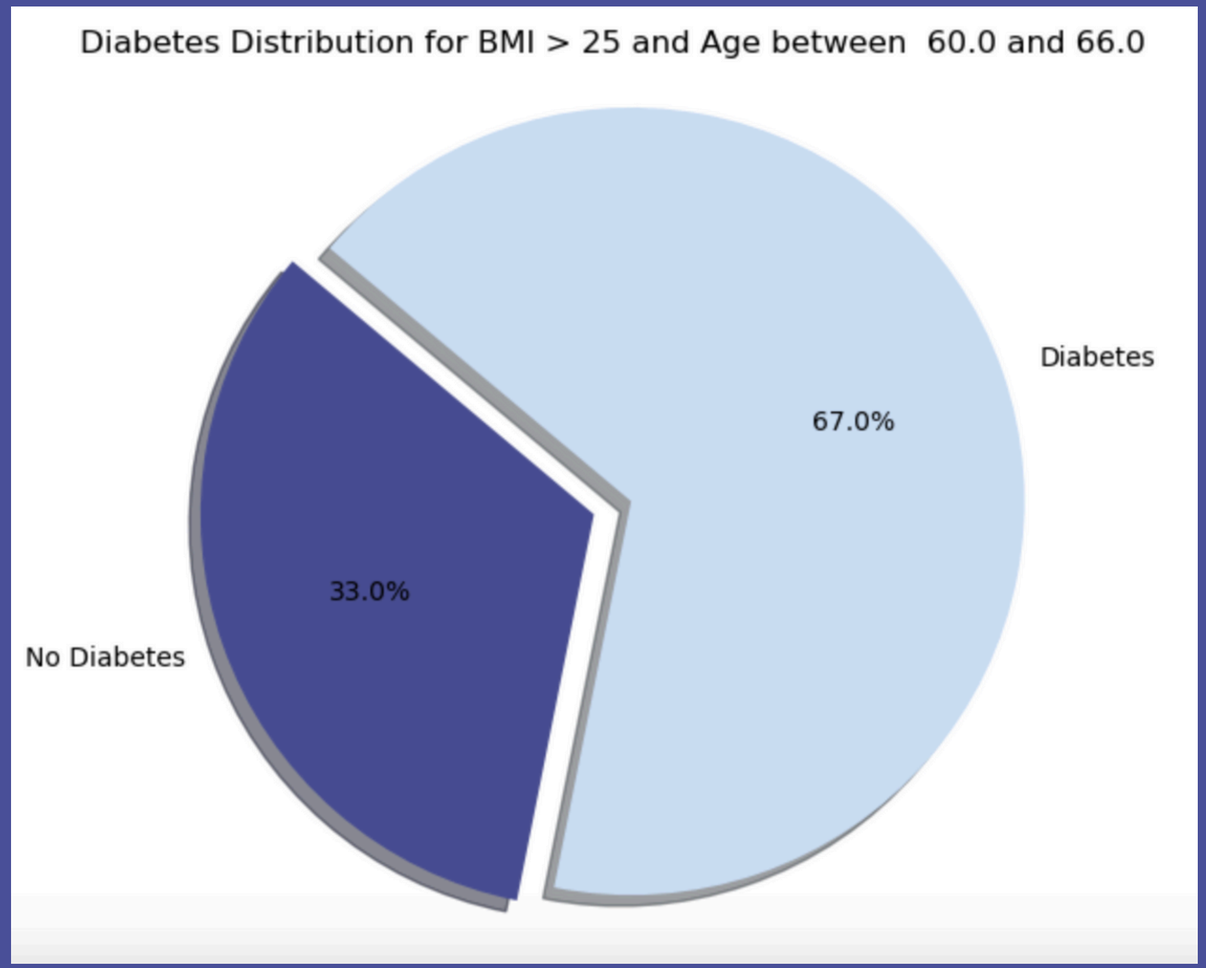
Readme and Repository

Cassandra Murray

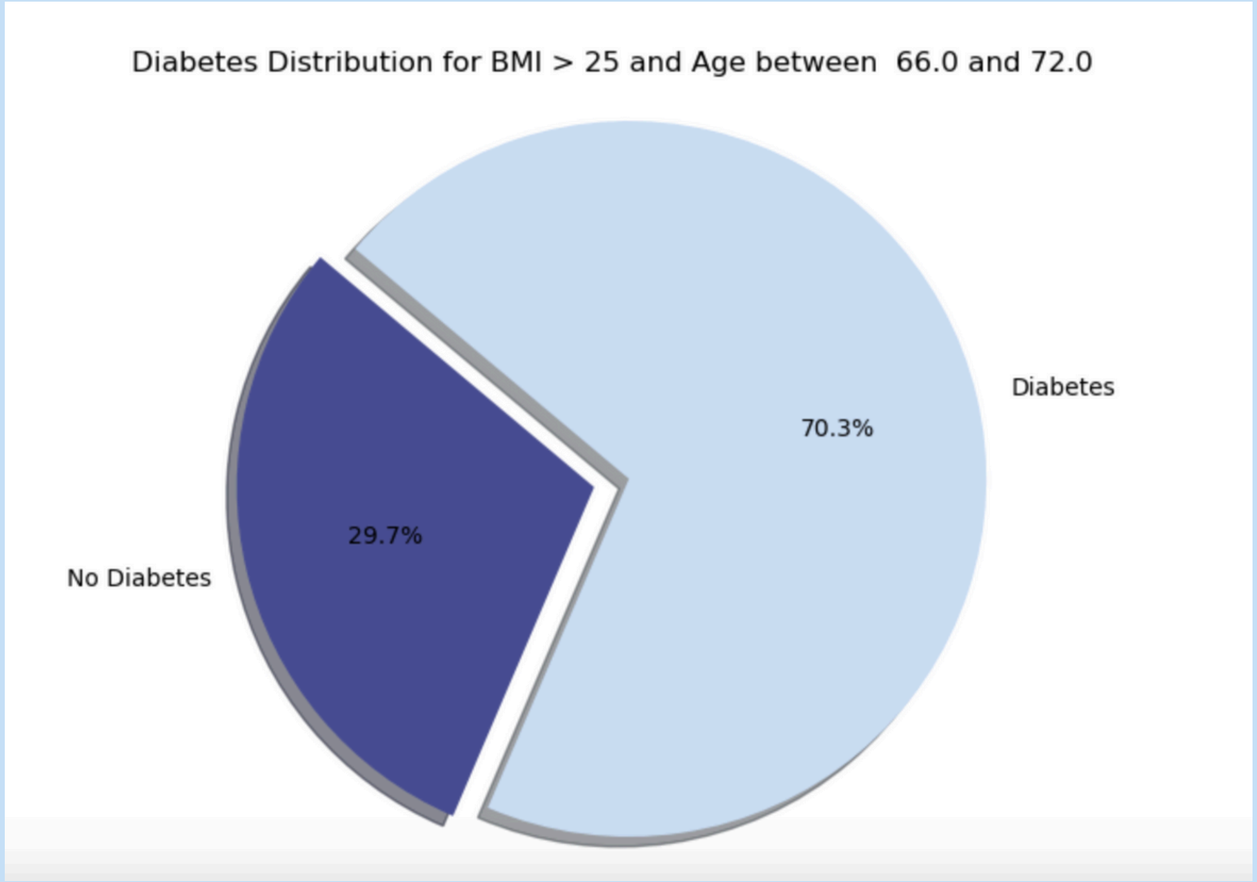
Presentation

Alex King

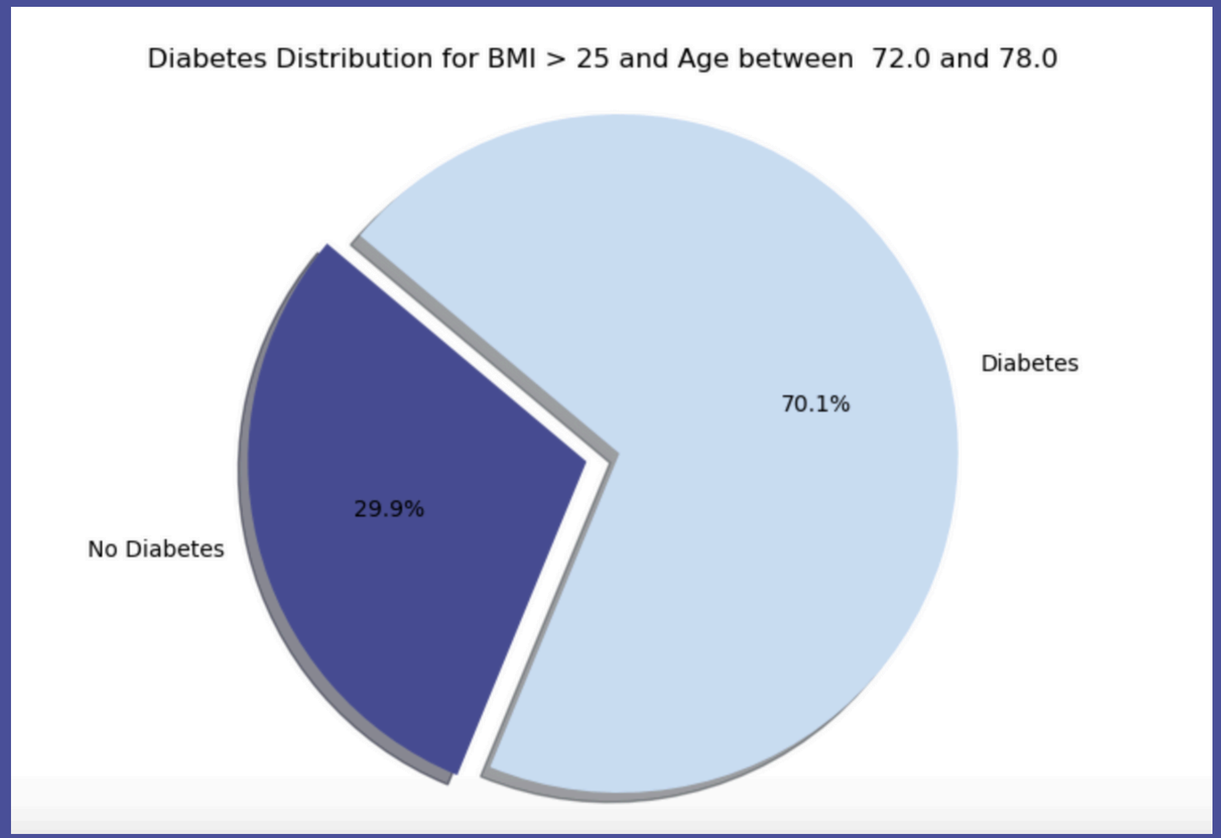
Appendix


Diabetes Distribution for BMI > 25 and Age between 60.0 and 66.0
Diabetes
67.0%
33.0%
No Diabetes

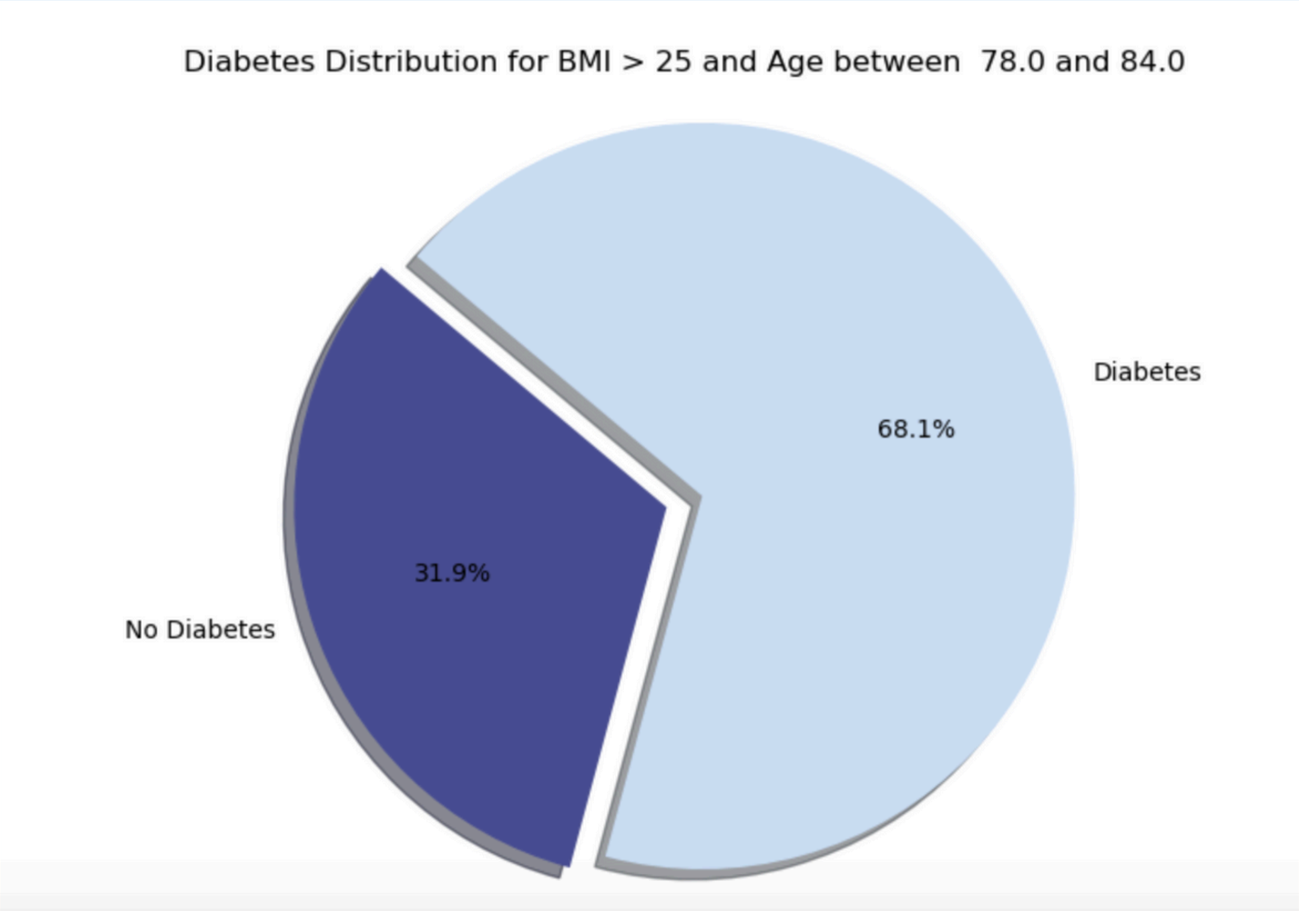
Appendix


Diabetes Distribution for BMI > 25 and Age between 66.0 and 72.0
Diabetes
70.3%
29.7%
No Diabetes

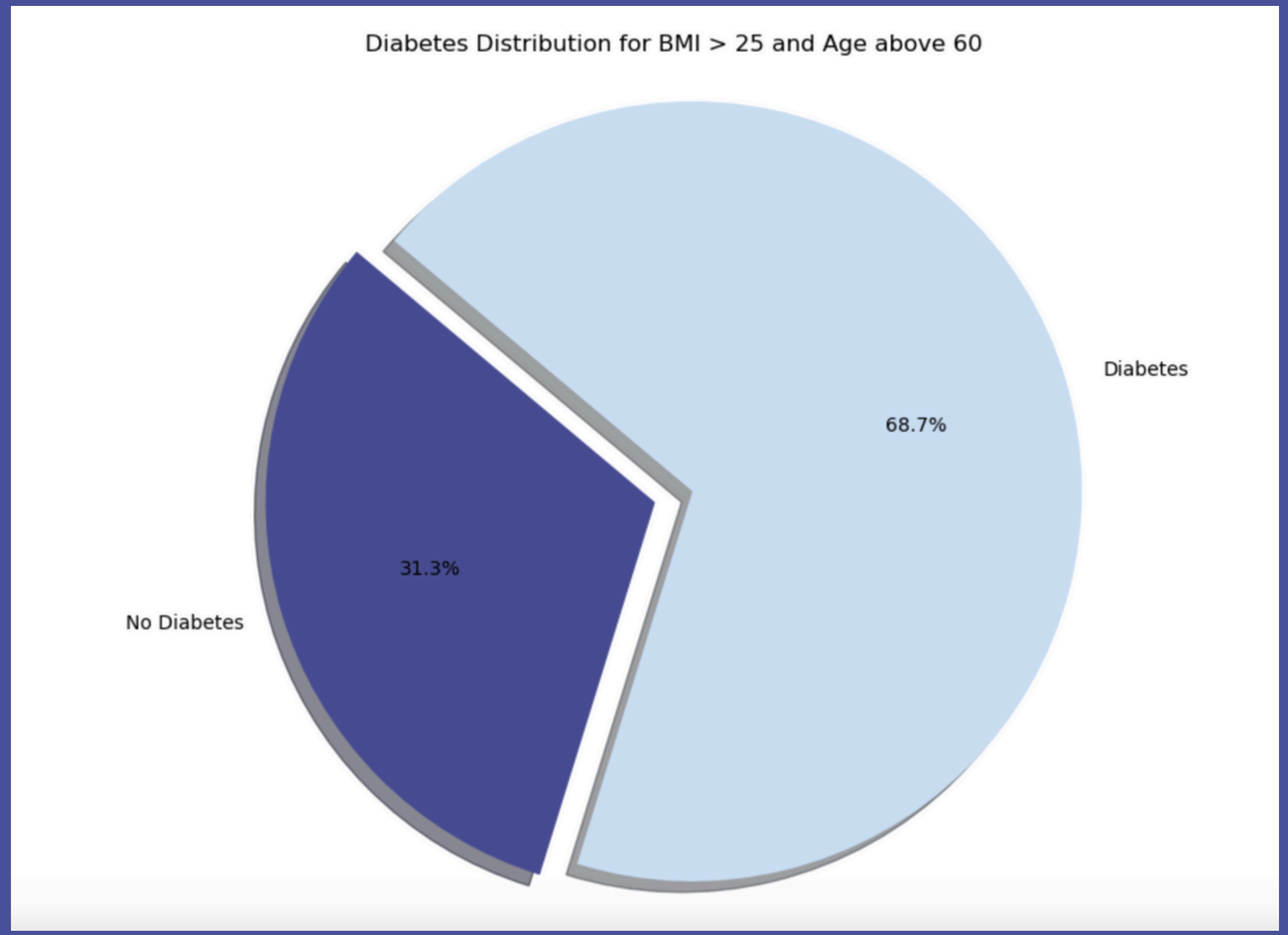
Appendix


Diabetes Distribution for BMI > 25 and Age between 72.0 and 78.0
Diabetes
70.1%
29.9%
No Diabetes

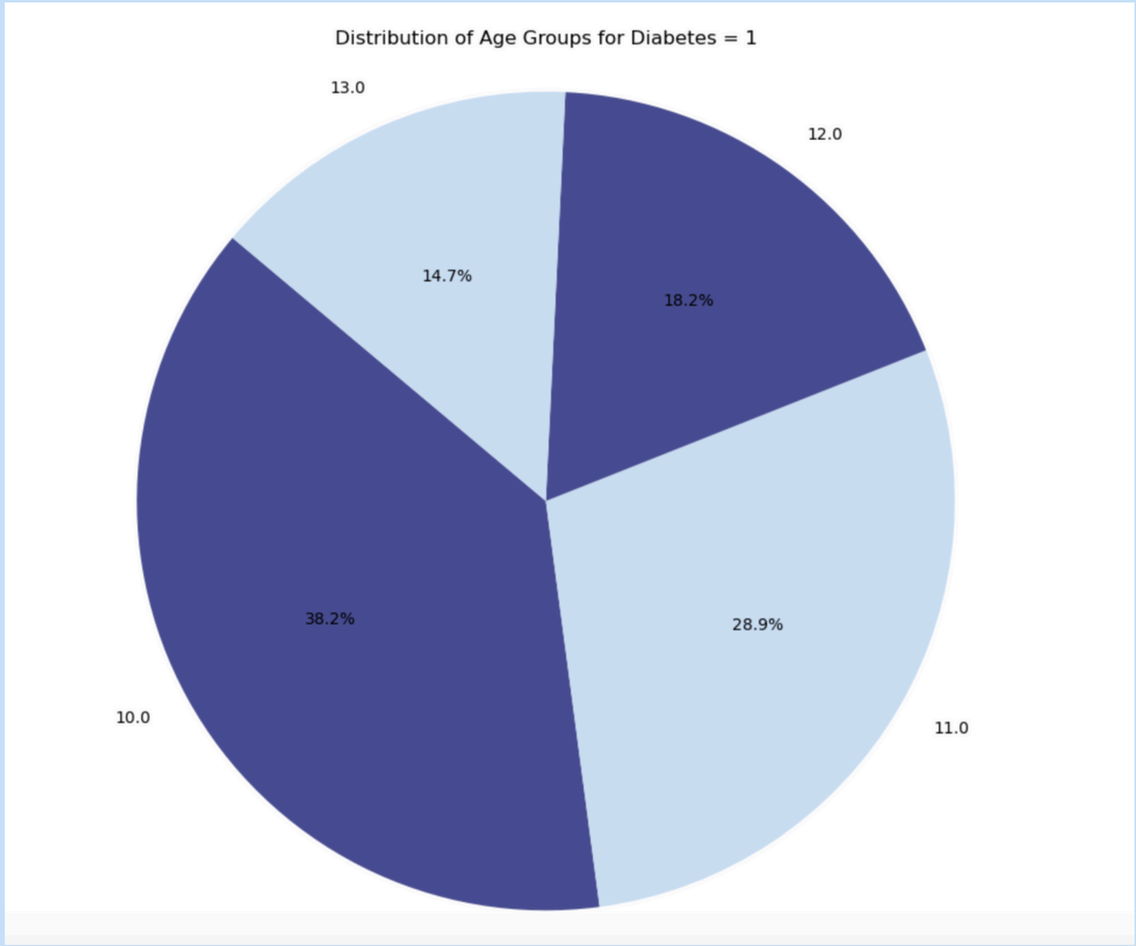
Appendix


Diabetes Distribution for BMI > 25 and Age between 78.0 and 84.0
Diabetes
68.1%
31.9%
No Diabetes

Appendix


Diabetes Distribution for BMI > 25 and Age above 60
Diabetes
68.7%
31.3%
No Diabetes

Appendix


Distribution of Age Groups for Diabetes = 1
13.0
12.0
14.7%
18.2%
38.2%
28.9%
10.0
11.0

Thank You