APPLIED MACHINE LEARNING BUAN 6341.001
PROJECT REPORT – GROUP 12

Lifestyle Habits and Medical Conditions Influencing Diabetes

Group Members:
Raghuraman Shankar - RXS210139
Ashwin Lakshminarasimhan - AXL210094
Sivasankaran Rajasekaran - SXR210114
Dinesh Raj Eswaran - DXE210014
Sofia Rajan - SXR220034
Premi Jawahar Vasagam – PXJ220007

INDEX

INTRODUCTION

Diabetes stands as one of the most pervasive chronic health conditions in the United States, affecting millions of Americans annually and imposing a substantial economic burden on society. The disease, characterized by the inability to regulate blood glucose levels effectively, significantly impacts the quality of life and life expectancy of individuals.

There are different types of diabetes, but type II diabetes is the most common form, and its prevalence varies by age, education, income, location, race, and other social determinants of health. Diabetes also places a massive burden on the economy, with diagnosed diabetes costs of roughly $327 billion dollars and total costs with undiagnosed diabetes and prediabetes approaching $400 billion dollars annually.

While there is no cure for diabetes, proactive strategies, including weight management, healthy eating, physical activity, and medical interventions, can alleviate its impact. Early detection is crucial, enabling timely lifestyle adjustments and more effective treatment, highlighting the importance of predictive models for diabetes risk.

MOTIVATION

The scale of the diabetes problem necessitates innovative approaches to tackle its prevalence and associated risks. According to the Centers for Disease Control and Prevention (CDC), millions of Americans are affected, with alarming numbers unaware of their risk. We want to understand the relationship between lifestyle habits and health condition with diabetes in the U.S.

By analyzing responses from over 400,000 individuals in the 2015 dataset, we seek to identify patterns, risk factors, and potential indicators that can aid in early diagnosis and intervention. The diverse datasets, including imbalanced and balanced versions, provide a nuanced understanding of the challenges posed by diabetes, offering valuable insights for public health officials, researchers, and policymakers. Ultimately, our project strives to contribute to the development of effective strategies for diabetes prevention, early intervention, and improved public health outcomes.

DATA DESCRIPTION

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. A csv of the dataset for 2015 was used, with a clean dataset of 253,680 survey responses The target variable Diabetes_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables. 6 out of the 22 variables have numerical data, 16 columns have variables that are binary in nature.

Column Description:

Diabetes_binary : you have diabetes (0,1)

HighBP : Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional (0,1)

HighChol : Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high? (0,1)

CholCheck : Cholesterol check within past five years (0,1)

BMI : Body Mass Index (BMI)

Smoker : Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] (0,1)

Stroke : (Ever told) you had a stroke. (0,1)

HeartDiseaseorAttack : Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) (0,1)

PhysActivity : Adults who reported doing physical activity or exercise during the past 30 days other than their regular job (0,1)

Fruits : Consume Fruit 1 or more times per day (0,1)

Veggies : Consume Vegetables 1 or more times per day (0,1)

HvyAlcoholConsump : Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)(0,1)

AnyHealthcare : Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? (0,1)

NoDocbcCost : Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? (0,1)

GenHlth : Would you say that in general your health is: rate (1 ~ 5)

MentHlth : 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? (0 ~ 30)

PhysHlth : Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0 ~ 30)

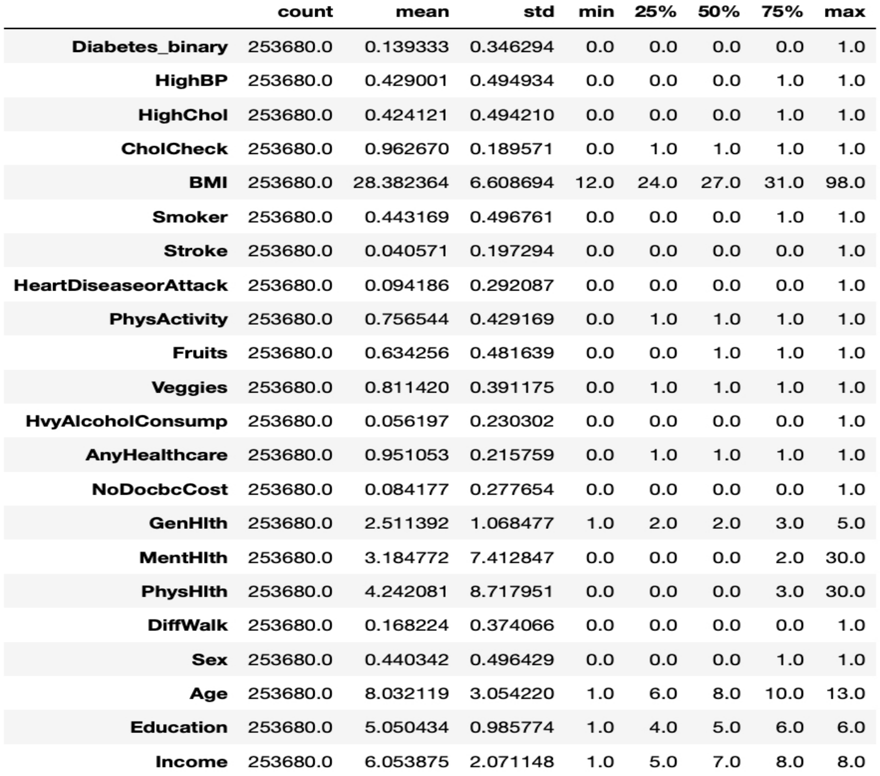
DiffWalk : Do you have serious difficulty walking or climbing stairs? (0,1)

Sex : Indicate sex of respondent (0,1) (Female or Male)

Age : Fourteen-level age category (1 ~ 14)

Education : What is the highest grade or year of school you completed? (1 ~ 6)

Income : Is your annual household income from all sources: (If respondent refuses at any income level, code "Refused.") (1 ~ 8)



	count	mean	std	min	25%	50%	75%	max
Diabetes_binary	253680.0	0.139333	0.346294	0.0	0.0	0.0	0.0	1.0
HighBP	253680.0	0.429001	0.494934	0.0	0.0	0.0	1.0	1.0
HighChol	253680.0	0.424121	0.494210	0.0	0.0	0.0	1.0	1.0
CholCheck	253680.0	0.962670	0.189571	0.0	1.0	1.0	1.0	1.0
BMI	253680.0	28.382364	6.608694	12.0	24.0	27.0	31.0	98.0
Smoker	253680.0	0.443169	0.496761	0.0	0.0	0.0	1.0	1.0
Stroke	253680.0	0.040571	0.197294	0.0	0.0	0.0	0.0	1.0
HeartDiseaseorAttack	253680.0	0.094186	0.292087	0.0	0.0	0.0	0.0	1.0
PhysActivity	253680.0	0.756544	0.429169	0.0	1.0	1.0	1.0	1.0
Fruits	253680.0	0.634256	0.481639	0.0	0.0	1.0	1.0	1.0
Veggies	253680.0	0.811420	0.391175	0.0	1.0	1.0	1.0	1.0
HvyAlcoholConsump	253680.0	0.056197	0.230302	0.0	0.0	0.0	0.0	1.0
AnyHealthcare	253680.0	0.951053	0.215759	0.0	1.0	1.0	1.0	1.0
NoDocbcCost	253680.0	0.084177	0.277654	0.0	0.0	0.0	0.0	1.0
GenHlth	253680.0	2.511392	1.068477	1.0	2.0	2.0	3.0	5.0
MentHlth	253680.0	3.184772	7.412847	0.0	0.0	0.0	2.0	30.0
PhysHlth	253680.0	4.242081	8.717951	0.0	0.0	0.0	3.0	30.0
DiffWalk	253680.0	0.168224	0.374066	0.0	0.0	0.0	0.0	1.0
Sex	253680.0	0.440342	0.496429	0.0	0.0	0.0	1.0	1.0
Age	253680.0	8.032119	3.054220	1.0	6.0	8.0	10.0	13.0
Education	253680.0	5.050434	0.985774	1.0	4.0	5.0	6.0	6.0
Income	253680.0	6.053875	2.071148	1.0	5.0	7.0	8.0	8.0

DATA PREPROCESSING

1. We checked for null values in our dataset and found none
2. We also identified the number of unique values to get more understanding of the dataset
3. We then checked for outliers for 7 columns: ‘BMI’, ‘GenHlth’, ‘MentHlth’, ‘PhysHlth’, ‘Age’, ‘Education’, and ‘Income’. We did not identify any outliers
4. We identified 24,206 duplicated rows in our dataset which we then removed. The change was verified, and the new shape of the dataset showed 229,474 rows and 22 columns

EXPLORATORY DATA ANALYSIS

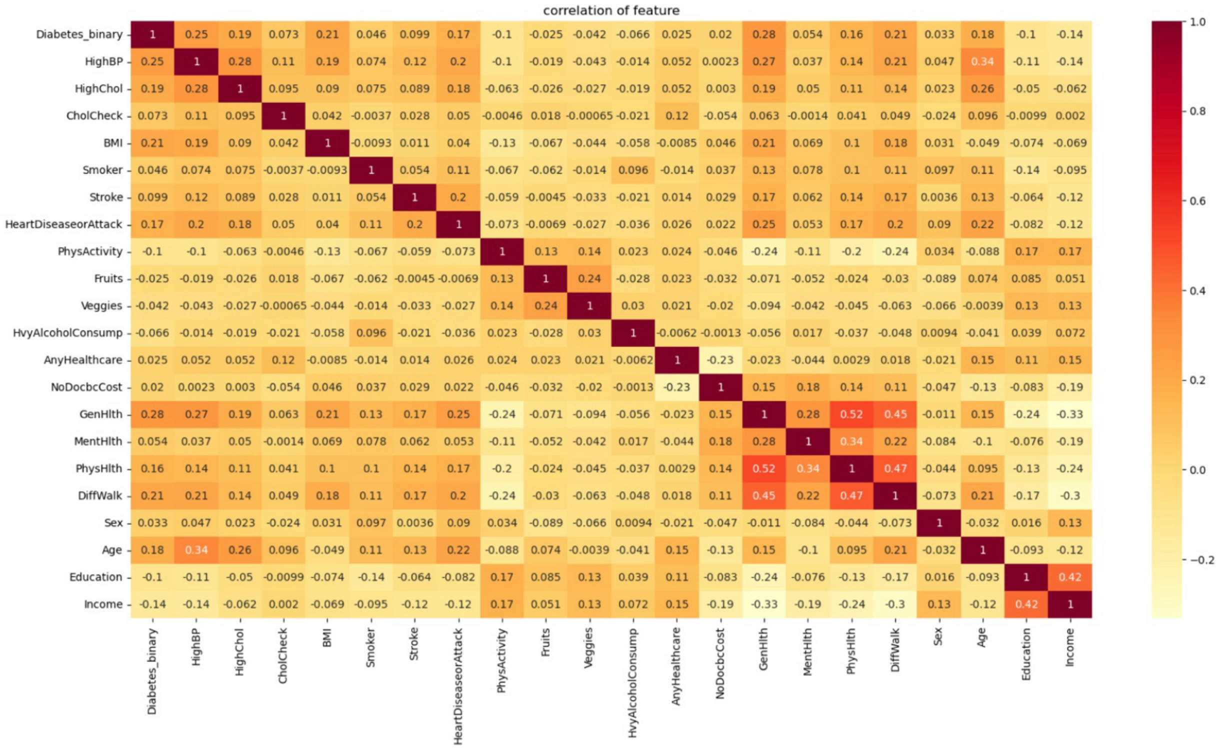


Figure 1

After checking for correlation using a heatmap, the following was observed:

- (GenHlth ,PhysHlth), (PhysHlth, DiffWalk), (GenHlth ,DiffWalk) are highly correlated with each other => positive relation
- (GenHlth, Income) , (DiffWalk , Income) are highly correlated with each other => Negative relation

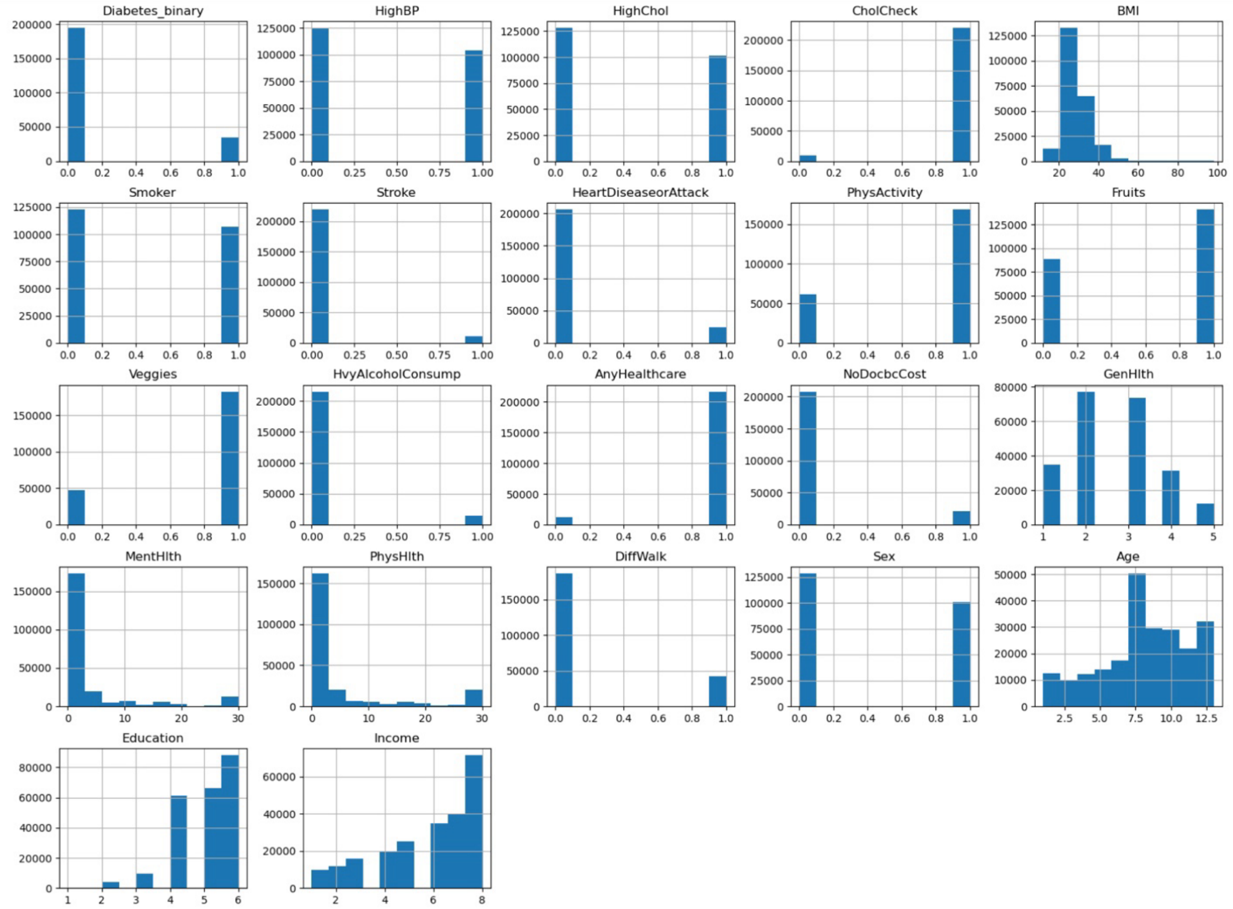


Figure 2

We further plotted different variables to study their ranges and occurrences:

- The binary bar chart indicates that a total of 35,347 individuals were identified as having diabetes
- According to the Heavy Alcohol Consumption bar chart, it can be observed that 14,256 individuals are reported to be consumers of alcohol
- The Physical Activity bar chart suggests that a total of 191,920 individuals are engaged in physical activity

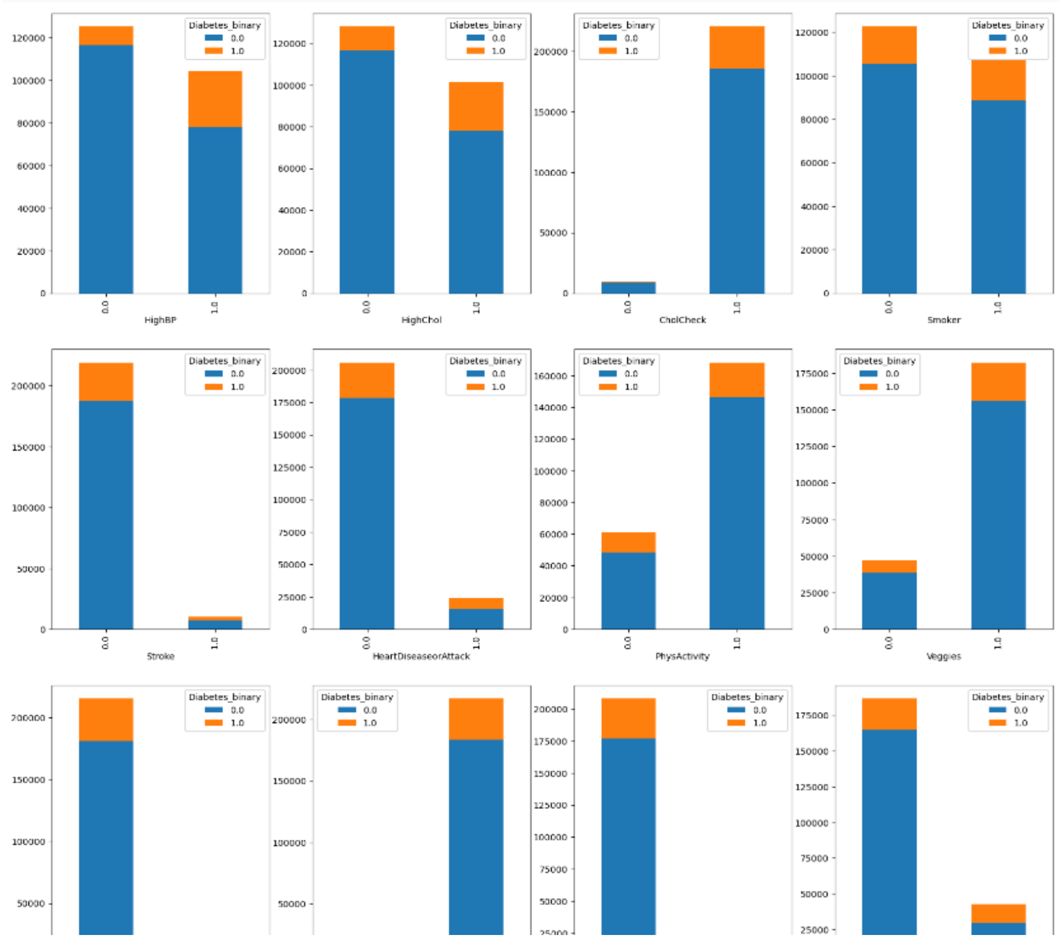


Figure 3

Since blood pressure is one of the main risk factors, we decided to explore it more:

- The stacked bar chart for high blood pressure (highbp) indicates that 8,743 individuals with diabetes also do not have high blood pressure
- The stacked bar chart for high blood pressure (highbp) indicates that 26605 individuals with diabetes also have high blood pressure
- The stacked bar chart for high blood pressure (highbp) indicates that 8226 individuals with no diabetes also have high blood pressure
- The stacked bar chart for high blood pressure (highbp) indicates that 136110 individuals with no diabetes also does not high blood pressure

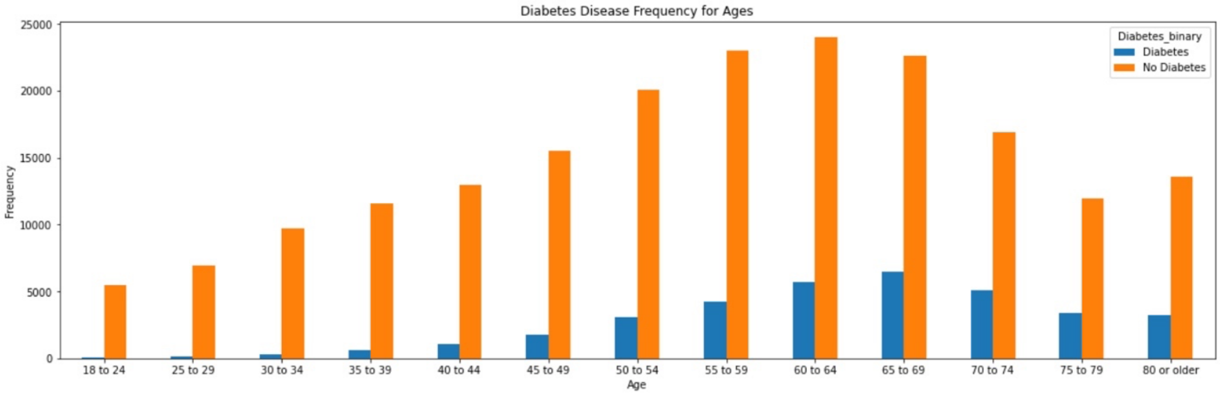


Figure 4

We also decided to explore the age groups are more affected by diabetes than the others

- The above plot suggests that individuals without diabetes typically fall within the age range of 18 to 44
- Conversely, those with diabetes are generally found in the age range of 45 or older, according to the same plot

MODELS USED AND PERFORMANCE EVALUATION

For our dataset, we trained 3 models with cross validation and hyperparameter tuning. We have listed the models used along with the accuracy metrics.

In our pursuit to understand the intricate relationship between lifestyle habits, medical conditions, and the prevalence of diabetes, we employed three distinct models, namely Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree. The focal point of our analysis was the binary variable representing the presence or absence of diabetes. Below, we outline the models used and their respective performance evaluations.

Logistic Regression:

Features of Interest:

Medical Conditions: HighBP, HighChol

Lifestyle Habits: Smoker, BMI

Performance Metrics:

Test Accuracy for Medical Conditions: 0.86

Test Accuracy for Lifestyle Habits: 0.85

Interpretation:

- The Logistic Regression model indicates that the presence of High Blood Pressure (HighBP) and High Cholesterol (HighChol) significantly contributes to predicting diabetes in the context of medical conditions
- For lifestyle habits, Smoking (Smoker) and Body Mass Index (BMI) emerge as key factors influencing the likelihood of diabetes in the predictive model
- The test accuracy of 0.86 for medical conditions and 0.85 for lifestyle habits suggests that the model performs well in capturing and predicting diabetes based on the selected features
- While effective, Logistic Regression assumes a linear relationship between features and the log-odds of diabetes. This may oversimplify complex interactions in our dataset

Decision Tree:

Model Specifications:

Max Depth: 4 (Less complex and less prone to overfitting)

Max Leaf Nodes: 10 (Limits the growth of terminal nodes)

Min Sample Split: 2 (Node will be split even if there are 2 records)

Key Features:

Features where maximum data split occurs are High Blood Pressure (HighBP) and General Health (Gen Health).

Model Generalization:

The decision tree model is designed to be generalized for unseen data, with a focus on simplicity and reduced risk of overfitting.

Interpretation:

- The Decision Tree model, with a maximum depth of 4, focuses on simplicity and avoids overfitting
- High Blood Pressure (HighBP) and General Health (Gen Health) are identified as pivotal features, indicating their significant impact on predicting diabetes
- The model's design for generalization ensures that it can effectively handle new, unseen data while maintaining its predictive accuracy
- The decision to limit the tree's depth to 4 mitigates overfitting, but there's a tradeoff between simplicity and capturing intricate patterns

K-Nearest Neighbors (KNN):

Optimal Parameters:
Number of Neighbors: 205 (Determined for optimal diabetes prediction).

Performance Metrics:
Prediction Accuracy: 86%

Interpretation:

- KNN, being a distance-based algorithm, determines that considering 205 neighbors optimally captures patterns for predicting diabetes
- The prediction accuracy of 86% suggests that the model effectively identifies individuals at risk of diabetes based on their proximity to other data points
- The KNN model is particularly useful for recognizing patterns in the dataset, providing a robust tool for predicting diabetes
- KNN is computationally intensive, requiring careful consideration of computational resources, especially with large datasets

Overall Assessment:

Our comprehensive analysis using logistic regression, decision tree, and KNN underscores the multifaceted nature of diabetes prediction. The models considered both medical conditions and lifestyle habits, shedding light on the interplay of factors influencing diabetes prevalence. The results provide valuable insights for public health officials, researchers, and policymakers, offering a foundation for the development of targeted strategies in diabetes prevention, early intervention, and improved public health outcomes.

CONCLUSION

From our analysis, we can see that all the models show a very similar test accuracy, but we prefer to rely on logistic regression or decision tree for the analysis. The processing time of decision tree and KNN was very long due to the size of the dataset containing over 250,000 data points and 21 variables whereas logistic regression was quickest in terms of processing.

Given the importance of understanding feature impacts in the context of public health, logistic regression remains a strong choice. If the aim is to capture more intricate patterns, particularly in non-linear relationships, decision tree could offer valuable insights. If computational resources permit, KNN's proximity-based approach may be beneficial, especially if the goal is to emphasize pattern recognition.

Further exploration, perhaps through ensemble methods or hyperparameter tuning, could enhance the predictive capabilities of the chosen model. All models exhibit high accuracy, indicating their capability in predicting diabetes based on our selected features. While logistic regression provides clear interpretability, decision tree and KNN sacrifice some interpretability for capturing more complex relationships. Logistic regression is computationally efficient, which may be advantageous for large datasets. However, KNN's computational intensity warrants consideration.