Predicting Diabetes Risk from Lifestyle and Demographics

By Hanna Pan

Summary of Research Questions and Results

1. **What lifestyle factors (physical activity, smoking, drinking, etc.) have the biggest influence in developing diabetes?**

   Smoking and physical activity showed that they had the strongest influence on diabetes diagnoses. Smoking increased the odds of someone developing the disease by ~41% and physical activity accumulated over 50% of the feature importance, when used to classify diabetic individuals.

2. **Do socioeconomic factors correlate to an increased risk of developing diabetes?**

   Yes, there are correlations between socioeconomic factors and risks of developing diabetes. Those with healthcare are significantly more likely to be diagnosed due to medical access, while income showed an inversely proportional relationship with diabetes risk.

3. **Are there any demographic factors that are associated with the likelihood of being diagnosed with diabetes?**

   Yes, demographic factors have some effects on diabetes diagnoses. Age and BMI in particular, showed a positive correlation to developing diabetes.

Motivation

This dataset and overall topic intrigues me because diabetes is a highly prevalent disease globally, and especially in the US. There are a variety of metrics used to classify diabetic, pre-diabetic, and healthy individuals in this dataset, making it a potentially conclusive study to find patterns across medicine and daily life. Considering how many cases go undiagnosed for years, it is important to develop an understanding of the signs for early prevention, and to enhance the quality of life in general. Studying how different behaviors may influence the likelihood of diabetes will hopefully improve health literacy for medical professionals, researchers, and most importantly, the average human-being.

Dataset

I will be using a dataset collected by the CDC as part of a Behavioral Risk Factor Surveillance System (BRFSS) survey. It's called CDC Diabetes Health Indicators, hosted by the UCI Machine Learning Repository: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators. This dataset can also be found on Kaggle for download:

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. It contains 253,680 instances and 22 features. These features range from lifestyle survey questions to medical test results, all leading towards conclusions to classify individuals as health, pre-diabetic, or diabetic.

Method

Overall:

- Import necessary imports (like pandas, seaborn, matplotlib.pyplot)
- Create dataframe by reading the csv
- Check to make sure that there are no missing or null values (in this case, there are none)
- For each attribute/factor being used in each research question, find each value count or 7-number summary (i.e summarize the variable)

1. **RQ1**

   Relevant features:

   - Diabetes_012 (our target variable), Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump

   Steps:

   - Data cleaning:
     - In this particular dataset, all variables are already binary (0 meaning no and 1 meaning yes) and have no missing values.
   - Perform data analysis:
     - Plot a multi-bar barplot (using `plt.bar()`), where the x axis represents each relevant lifestyle factor, split into their respective yes and no answers. The y axis represents the percentage of individuals diagnosed with diabetes. This barplot will indicate the prevalence of diabetes for individuals who answered yes or no to each lifestyle factor.
   - Challenge tasks:
     - Import necessary imports (sklearn models, linear models, random forest, and metrics, and numpy)
     - Modeling:
       - Fit and train a Logistic Regression model to the dataset. Check the coefficients (using `.coef_`) of the logistic regression model to help determine general/overall effects of each factor. Determine odds ratios (using `np.exp` of the coefficient) to help understand which factor is more likely to influence diabetes.
       - Fit and train a Random Forest model to the dataset. Find feature importances (using `.feature_importances_`) of each factor to determine which is more likely to influence or contribute to a diabetes prognosis.
     - Validity:

        - Perform chi-squared tests (using `.crosstab()` and `.chi2_contingency()`) on each lifestyle feature. Using a p value of 0.05 as the marker (95% confidence), determine whether or not each factor is statistically significant and if the null hypothesis should be rejected or not.

2. **RQ2**

Relevant features:

- Diabetes_012 (our target variable), Income, Education, AnyHealthcare

Steps:

- Data cleaning:
    - In this particular dataset, all variables are already binary (0 meaning no and 1 meaning yes) and have no missing values.
- Perform data analysis:
    - Plot a barplot (`sns.barplot()`), where the x axis are the buckets of income ranges, and the y axis is the proportion of individuals who are diabetic. This plot will describe the distribution of people who are diabetic in each income group.
    - Plot a barplot (`sns.barplot()`), where the x axis are the buckets of education ranges, and the y axis is the proportion of individuals who are diabetic. This plot will describe the distribution of people who are diabetic in each education level group.
    - Plot a barplot (`sns.barplot()`), where the x axis are whether or not someone has healthcare, and the y axis is the proportion of individuals who are diabetic. This plot will describe the distribution of people who are diabetic depending on if they have healthcare.
- Challenge tasks:
    - Uses the same imports as RQ1
    - Modeling:
        - Fit and train a Logistic Regression model to the dataset. Check the coefficients (using `.coef_`) of the logistic regression model to help determine general/overall effects of each factor. Determine odds ratios (using `np.exp` of the coefficient) to help understand which socioeconomic factor is more likely to influence diabetes.
        - Fit and train a Random Forest model to the dataset. Find feature importances (using `.feature_importances_`) of each socioeconomic factor to determine which is more likely to influence or contribute to a diabetes prognosis.
    - Validity:
        - Import spearmanr
        - Perform chi-square tests (using `.crosstab()` and `.chi2_contingency()`) on AnyHealthcare because it's

categorical. Using a p value of 0.05 as the marker (95% confidence), determine whether or not this factor is statistically significant and if the null hypothesis should be rejected or not.
        - Perform correlation test on Income and Education. Use `spearmanr()` because it's better for analyzing qualitative and quantitative data. Determine correlations between these factors and having diabetes to see if there is any relevance.

3. **RQ3**

Relevant features:

- Diabetes_012 (our target variable), Age, Sex, BMI

Steps:

- Data cleaning:
    - In this particular dataset, all variables are already binary (0 meaning no and 1 meaning yes, and 0 meaning female and 1 meaning male) and have no missing values.
- Perform data analysis:
    - Plot a barplot (`sns.barplot()`), where the x axis represents age groups, and the y axis represents the proportion of individuals with diabetes.
    - Plot a barplot (`sns.barplot()`), where the x axis represents sex (female or male in this dataset), and the y axis represents the proportion of individuals with diabetes.
    - Plot a boxplot (`sns.boxplot()`), where the x axis represents each group of classified diabetes groups (0=healthy, 1=pre-diabetic, 2=diabetic), and the y axis represents the BMIs of individuals in each group.
- Challenge tasks:
    - Modeling:
        - Fit and train a Logistic Regression model to the dataset. Check the coefficients (using `.coef_`) of the logistic regression model to help determine general/overall effects of each factor. Determine odds ratios (using `np.exp` of the coefficient) to help understand which demographic factor is more likely to influence diabetes.
        - Fit and train a Random Forest model to the dataset. Find feature importances (using `.feature_importances_`) of each demographic factor to determine which is more likely to influence or contribute to a diabetes prognosis.
    - Validity:
        - Import scipy.stats
        - Perform chi-squared tests (using `.crosstab()` and `.chi2_contingency()`) on Age and Sex. Using p value of 0.05, determine statistical significance and whether or not to reject null hypothesis.

- Perform t-test using `ttest_ind` on BMI and Sex to find any significance in their influence on diabetes. Use a p value of 0.05 as well.

EDA

The dataset contains 253,680 rows and 22 columns, with no missing values. Each row represents one survey response from an individual, and each column represents one feature variable asked in the survey. It's important to also note that most variables are binary or ordinal, which led to minimal cleaning required for the dataset's preparation. The dataset is also imbalanced—our target variable (Diabetes_012) has a breakdown of ~84% healthy, ~14% diabetic, and ~2% pre-diabetic.

Some initial statistics and revelations from the plots revealed notable patterns:

1. **Lifestyle:** Smoking and physical inactivity showed higher proportions of individuals with diabetes.
2. **Socioeconomic:** Lower income and education levels illustrated higher diabetes prevalence
3. **Demographics:** Older age groups and those with high BMIs showed substantially greater diabetes rates.

The EDA was important to my analysis in a multitude of ways:

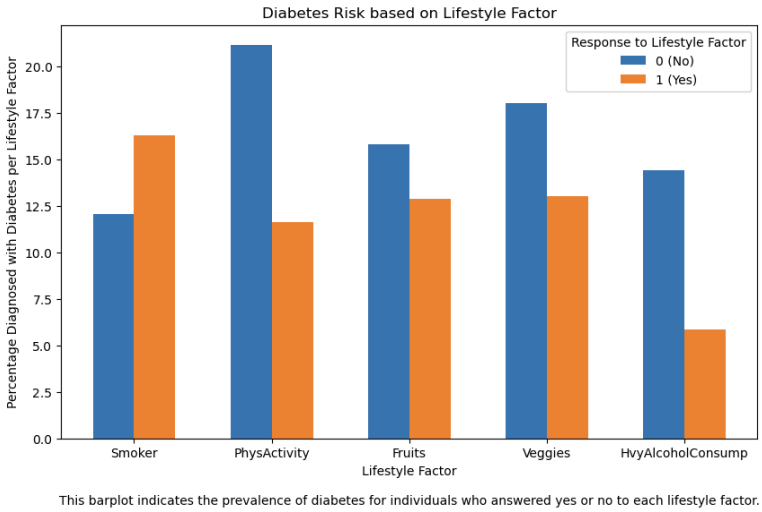
1. Minimal preprocessing made it easier to focus on the research questions at hand.
2. Understanding variable types and data collection strategies guided challenge goals—model and statistical tests choices.
    a. Doing so allowed for certain assumptions to be made, like independency, categorical, ordinal, or continuous variables.
3. Pinpointed relevant features for each research question.

This is not an exhaustive list by any means, but these early insights helped shape my analysis and guide any changes, and allowed me to focus on modeling and validity of promising predictors.

Results

RQ1: Lifestyle Factors and Diabetes Risk

Findings:



This barplot indicates the prevalence of diabetes for individuals who answered yes or no to each lifestyle factor.

This barplot shows the percentage of individuals diagnosed with diabetes based on their responses to certain lifestyle factors (e.g., smoking, physical activity, diet, alcohol use, etc.). Each factor is binary-coded (0 = No, 1 = Yes). This multi-barplot reveals that the lack of physical activity, poor diet, and smoking are linked to higher diabetes rates, while heavy alcohol use shows a lower association.

Machine Learning Results:

- **Logistic Regression model:** Smoking had the largest positive coefficient of 0.342 and an odds ratio of 1.41, indicating that ~41% of smokers are at higher odds to develop diabetes compared to non-smokers, if controlling for other factors.
- **Random Forest model:** Physical activity was identified as the most influential lifestyle factor in predicting diabetes risk, accounting for over 50% of the feature importance. This means that the model found splitting on physical activity to be most useful.

While these models predicted different features, it's important to note that "influence" is defined differently for each model. Based on interpretable data, the Logistic Regression model infers that smoking is a more likely cause of diabetes compared to other factors. As for the Random Forest model, physical activity metrics are a better predictor for whether or not someone might develop the disease. In either case, these outcomes do reinforce what we see visually in the barplot.

Statistical Validation Results:

- Null hypothesis: There is no association between lifestyle factors and the likelihood of developing diabetes.
- All variables relevant to this research question were categorical and the chi-square tests for all factors revealed a p-value of less than 0.05 for all.

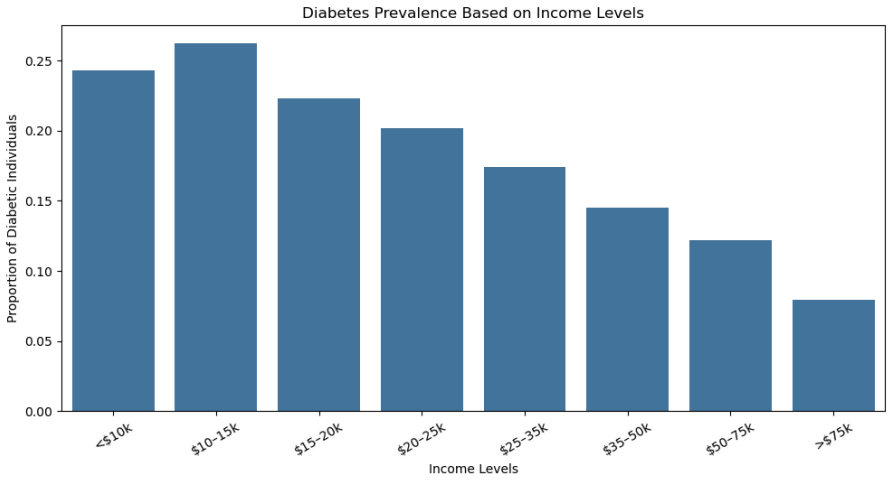
I rejected the null hypothesis. This means that there is some statistical significance between each lifestyle factor and its influence on developing diabetes. The relationships between our daily lives and risks of diabetes are not due to random chance, but rather a reflection of consistent patterns in lifestyle that influence health outcomes over time.

Interpretations and Implications:
The results from the preliminary data analysis and challenge tasks illustrate an established link between activity, diet, and smoking, and disease. Heavy alcohol consumption showing low prevalence rates is rather surprising, but this could very well be a product of response or self-reporting bias. Additionally, those diagnosed with diabetes may have adopted healthier habits in an effort to battle the disease, thus making the relationship appear inverse.

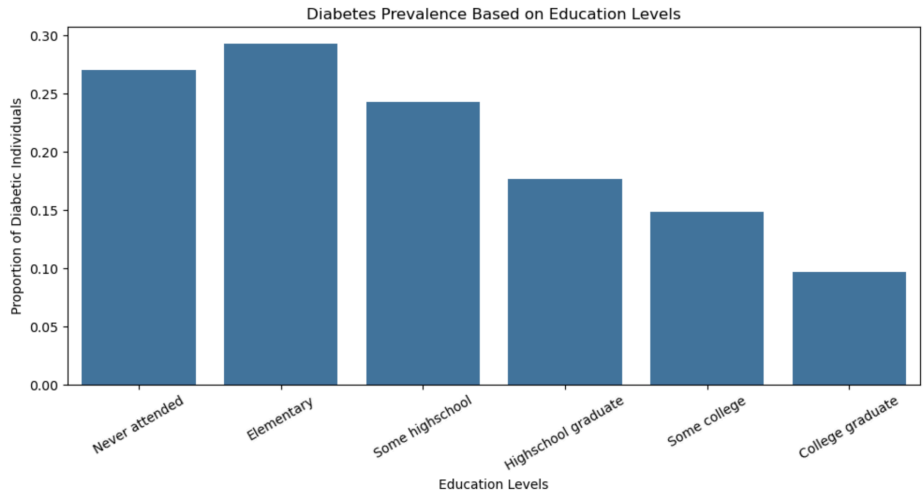
It's also widely known that “correlation doesn’t equal causation,” but these consistent patterns across multiple variables suggest that lifestyle choices play a substantial role in shaping long-term health outcomes. It’s worth spreading public health messages and interventions that target tobacco use and physical inactivity. Such lifestyle factors are formidable forces in other diseases and aspects of life, so it’s important to spread awareness of the potential health implications.

RQ2: Socioeconomic Factors and Diabetes Risk

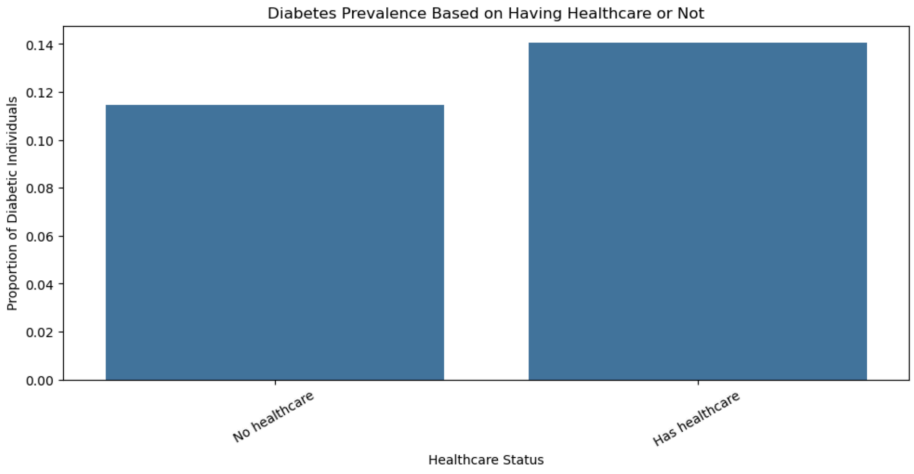
Findings:



This plot describes the distribution of people who are diabetic in each income group.



This plot describes the distribution of people who are diabetic in each education level group.



This plot describes the distribution of people who are diabetic depending on if they have healthcare or not.

These barplots illustrate a clear trend: lower income and education levels show a higher prevalence of diabetes. And interestingly, those who have healthcare seem to show a higher proportion in diabetes prevalence too.

Machine Learning Results:

- **Logistic Regression model:** Healthcare coverage had the highest positive correlation of 0.643, with an odds ratio 1.902. This means that healthcare is important in determining someone's diabetes diagnosis, but also shows that those with healthcare are ~90% more likely to be diagnosed. This metric might seem misleading at first, but it's important to clarify that these statistics aren't insinuating that healthcare *causes* disease, rather, having healthcare *increases* the likelihood of someone being diagnosed.
- **Random Forest model:** Income was identified as the most influential lifestyle factor in predicting diabetes risk, accounting for over 66% of the feature importance. This means that the model found predicting based on income to be most useful.

Again, it's important to note that "influence" is defined differently for each model. Based on interpretable data, the Logistic Regression model infers that healthcare is a more likely cause of high diabetes diagnosis prevalence. As for the Random Forest model, income metrics are a better predictor for whether or not someone might develop the disease. In either case, these outcomes do reinforce what we see visually in plots.

Statistical Validation Results:

- Null hypothesis: There is no association between socioeconomic factors and the prevalence of diabetes diagnoses.
- For AnyHealthcare, a categorical variable, the chi-square test revealed a p-value of less than 0.05.
- For Income and Education, ordinal variables, the Spearman correlation tests revealed both factors had a weak, but negative correlation with developing diabetes.

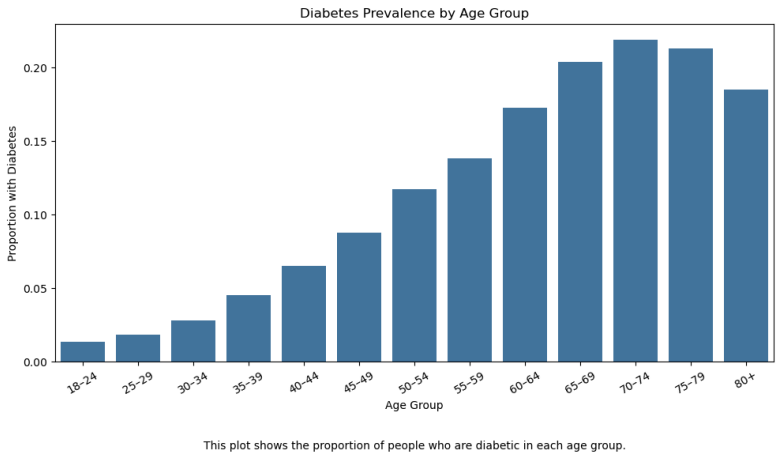
I rejected the null hypothesis. This means that there is some statistical significance (i.e not due to random chance) between having healthcare and diabetes diagnoses. Additionally, the correlation tests revealed that the more educated and wealthy an individual is, the less likely it is to develop diabetes.

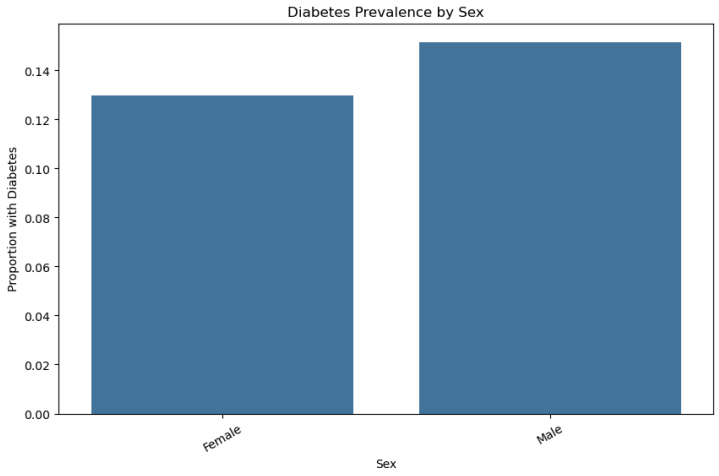
Interpretations and Implications:
These results show a connection between low socioeconomic status and diabetes. Research has shown that limited access to opportunities, food, and more, can lead to harsher lifestyles that might negatively impact an individual's health. In the case of healthcare, those without insurance may remain undiagnosed even if they are diabetic.

This analysis underscores exactly the importance of increasing care access amongst individuals who need it. Whether that takes the form of introducing new policies to bridge the gap between insured and uninsured populations, or community-level initiatives, it's necessary to address the barriers between groups to promote healthy living.

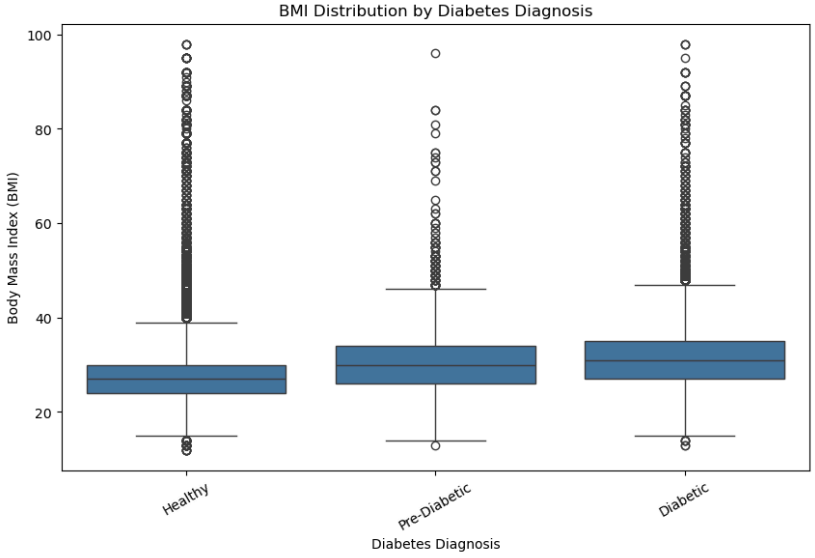
RQ3: Demographic Factors and Diabetes Risk

Findings:





This plot shows the proportion of people who are diabetic in each sex group.



This plot shows the BMI of people in each diabetes diagnosis group.

Based on these barplots, diabetes prevalence increases as age increases and males have slightly higher prevalence than females. Boxplots showed that average BMI was markedly higher among diabetic individuals.

Machine Learning Results:

- **Logistic Regression model:** All three demographic factors show positive and meaningful associations with diabetes risk. Age has the strongest effect (coefficient of 0.246 and odds ratio of 1.279), likely reflecting increased vulnerability to chronic conditions over time.
- **Random Forest model:** The model identified BMI as the most influential lifestyle factor in predicting diabetes risk, accounting for over 56% of the feature importance. This means that the model found splitting on BMI to be most useful. Additionally, age also showed to have ~43% importance as well.

The two models seem to agree overall, but it's still important to note that "influence" is defined differently for each model. Based on interpretable data, the Logistic Regression model infers that age is more likely to cause higher proportions of diabetes. As for the Random Forest model, BMI metrics are a better predictor for whether or not someone might develop the disease—though age seems to also be almost equally as useful.

Statistical Validation Results:

- For age, an ordinal variable, the Spearman correlation tests revealed a weak, but positive correlation with developing diabetes.
- For Sex and BMI, the null hypothesis used was that there is no difference between average BMI for those with or without diabetes and that there is no difference between diabetes rates between sex.
    - T-tests confirmed that the p-value for both variables is less than 0.05.

I rejected the null hypothesis. There is some difference between BMIs of individuals with and without diabetes, as well as a difference between diabetes rates for sex. Additionally, the correlation tests revealed that the older someone is or gets, the more likely it is to develop diabetes.

Interpretations and Implications:
These findings align with established medical understanding: age and obesity are primary risk factors for general health issues, including diabetes. The minimal difference between sex may be influenced by lifestyle, genetic predisposition, or other environmental factors.

As we age, our bodies' mechanisms gradually slow down and no longer function as they should, which highlights the importance of additional screening for older adults. Also, weight management is crucial to leading a healthier lifestyle, not just in preventing diabetes, but also other harmful diseases. And although pre-existing medical conditions, or other factors may indirectly lead to such issues, the message of healthy living remains the same: maintaining a healthy lifestyle impacts future health.

Impact and Limitations

Implications
The findings of this analysis can certainly help those who are looking for interventions regarding prevention or delaying of onset diabetes. By identifying lifestyle factors that may be harmful, health organizations can rally support behind programs targeting behavior changes for better health. And by highlighting the stark contrast between resource access, policies can be enacted to close the health gap in underserved communities. Overall, the results of this analysis can help policymakers and healthcare providers identify populations for early screening.

Beneficiaries

- **Healthcare providers:** can develop better and more personalized screening protocols for patients.
- **Policy makers:** can enact policies to address accessibility gaps.
- **Researchers:** can use insights to study specific groups and causes.
- **Public health agencies and officials:** can use this information to organize informational and resourceful programs

Who Might Be Excluded or Harmed

- **Underserved communities:** may be undiagnosed or not accounted for in the survey, and thus these generalizations may not reflect their circumstances.
- **Stereotypes:** these findings can reinforce stereotypes in communities if misinterpreted; systemic and environmental factors may be overlooked when hyperfocusing on certain results.

Data Biases

- **Response bias/Self-reported data:** this was a self-reported survey, thus people may inaccurately report certain responses. This could be a conscious or subconscious decision, but with health being a sensitive topic, there are possibilities of inaccurate responses, either due to social pressure or what's deemed acceptable.
- **Sampling bias:** the survey may not have been given in all parts of the country, and therefore, may not capture all populations.
- **Imbalanced data of target variable:** there are far more individuals in this survey that are healthy, which might limit or affect a model's predictive power on diabetic cases.

Limitations of Analysis

- **Correlation ≠ causation:** although mentioned earlier, correlation still does not equate to causation, despite the strength of each association.
- **Genetic predispositions and other unmeasurable factors:** genetics plays a huge role in someone's health, and is unfortunately not controllable.
- **Static observational timeline:** these answers come from a one-time survey, which doesn't account for change over time. This data gives us a picture from a snapshot of time, which limits our ability to process actual causality.

Use of Conclusions

While these findings are informative, they are by ***no*** means a diagnostic tool. The identified associations can be used to reallocate resources and inform prevention strategies, but health has a gradual timeline and these diseases are constantly being studied, and thus, these conclusions should be used as a guiding framework for future research, policies, or practices.

Challenge Goals

1. **Machine learning:** Instead of training decision trees, I decided to train and evaluate a Logistic Regression and Random Forest model to better understand which features have the heaviest influence on diabetes. Using these models, I can work on answering a more realistic goal: *which* factor contributes to higher risk, *why* is someone at high risk or not and *what* to look for.
2. **Result validity:** I didn't change much of my plan for this challenge goal. I used chi-square tests, correlation tests, and two sample t-tests to validate my findings from initial relationships. These helped me reject or accept the null hypothesis.

Work Plan Evaluation

Upon revisiting my original plan, I would still say 90% of the proposed work plan I set up was pretty accurate. Cleaning the data took far less time because the data was already pre-cleaned and I just had to double check. As expected, the actual data analysis and challenge tasks took most of the time because I needed to do research and actually learn how to implement the tasks, as well as debug. As I'm writing finishing touches of the final report, it took less time than I expected as my previous EDA was very detailed and already contained much of the necessary and expected information.

Testing

I tested my code by using smaller data files and assert statements. Because my dataset has over 250,000 entries, it would be too time consuming and inefficient to test it. So I created a test_data.csv file that has the relevant variables I studied with some entries. I then used assert statements to test my methods. I chose not to test certain methods (e.g data_summary) because there are no returns and no print statements, thus nothing to test. For the ML methods, I tested intermediate steps because I can't control how my model splits all the data. For the validity tests, I also tested the internal logic as my method doesn't return anything (just prints). I didn't test any plots because I'm unsure of what the expected output should look like. But overall, my code is trustworthy because all the set up for each method is correct and there are no internal logical errors.

Collaboration

While completing this part of my report, I referenced these sources:

- https://www.geeksforgeeks.org/machine-learning/chi-square-test-in-Data-Science-and-Data-Analytics/
- https://www.geeksforgeeks.org/python/plotting-multiple-bar-charts-using-matplotlib-in-python/
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html
- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html
- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html#r3566833beaa2-2