Leah Maciel

Project 1

Link to dataset: <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>

Before performing the following tasks, I removed duplicate rows from the dataframe and confirmed there were no missing values.

# Task 2- 5 types of summaries

## View information about the columns and metadata

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

From this we can see a summary of all the columns, what they represent, the data type, and if there are missing values. All the features are binary or integers, and there are no missing values.

## Summary statistics

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

This step is valuable for understanding the general distribution of each feature in the dataframe. With this step I can see a summary of each feature and use this to inform which features I may want to investigate further. For example, BMI appears to have a large outlier. This may mean that I want to use other measures to understand BMI, like median or a trimmed mean.

## Visualize distribution of diabetes diagnosis

A blue and red graph

Description automatically generated

Since I am focusing on diabetes and trying to find relationships between the other features and whether or not a patient has diabetes, I first wanted to understand how many of the patients in the dataset have diabetes, and what the distribution of positive to negative diagnosis is. Out of the 253,680 patients, 35,346 have diabetes. I also created a bar chart to visualize how there are many more people in the dataset without diabetes than there are with.

### Calculate number of outliers for each integer feature

A screenshot of a computer

Description automatically generated

I was also interested in investigating how many outliers are in each feature. I chose to focus only on the integer features, since it isn’t logical to have outliers in binary features because there’s only 2 options (0 or 1). To do this I calculated the IQR and counted how many values are below q1 – 1.5\*IQR or above q3 +1.5\*IQR. I found that mental health and physical health have the most outliers. Since these features have a range of values from 1-30 (days of poor physical or mental health) and these are subjective questions, it makes sense there can be a large range of answers, resulting in outliers.

## Count number of unique values for each feature

A screenshot of a computer screen

Description automatically generated

While the metadata describes the possible values each feature could have, I wanted to investigate the actual number of unique values present for each feature. Most of the features are categorical and binary, explaining the 2 different values. Notably, BMI has 84 unique values, so I may expect to see some variation in that feature.

### Visualize distribution of each feature

A screenshot of a graph

Description automatically generated

Along with looking specifically at the diabetes diagnosis, I wanted to understand the distribution of each feature. I did this by creating a histogram of each feature. From visual observations we can see that most of the features do not seem to be evenly or normally distributed. This is especially for the binary features because many of them appear to have a dominant/more common value (for example CholCheck and Stroke).

## Task 3- 5 types of exploratory analysis

## Correlation heatmap

A screenshot of a graph

Description automatically generated

Correlation is a valuable tool for identifying relationships between features. The closer to 1 the correlation of two features is, the more likely they are to be associated. I created and visualized the correlation between all features. From this we can see there does not appear to be a strong association between any of the features, however there are some correlations that might be valuable to investigate further such as physical health with general health or with difficulty walking. Based on what these features represent, it makes sense that there would be an association between them (i.e. someone with poor physical health is likely to have more difficulty walking). There is also a negative correlation between general health and income which could be interesting to investigate further and see if income negatively impacts a person’s health and contributes to diabetes.

### Cosine similarity of top 3 features with greatest correlation values

A close up of black text

Description automatically generated

Based on the results of the correlation performed in the step above, I wanted to investigate how similar the top three most correlated features were. These features are physical health, general health, and difficulty walking. Since I know these features have correlation values > 0.45, I also expected them to be similar. For all three pairs, the cosine similarity was greater than 0.5, showing that there is some similarity between the features. Logically it also makes sense that someone struggling with one of these features would also struggle with the other two.

### Box plot of features against Diabetes\_binary

A screenshot of a graph

Description automatically generated

I generated boxplots of diabetes diagnosis versus each feature to summarize and understand the distribution of the data for each feature, see where there are outliers, and see if this changes based on diagnosis. HighBP, PhysActivity, and DiffWalk all have notably different boxplots based on Diabetes\_binary. For example, almost all patients (except for outliers) with a positive diagnosis also have high blood pressure, compared to those with a negative diagnosis that have a range of high blood pressure distribution.

### t test for BMI



I chose to perform a t test for BMI since it’s a numeric feature, comparing the average means for the positive diabetes diagnosis and negative diagnosis groups. Although BMI is not normally distributed, the dataset is large enough that we can use the central limit theorem and assume normality. The p value is < 0.05, meaning there is a statistically significant difference in mean BMI values for the two groups. This can lead to further analysis and questions, such as does having a greater BMI contribute to developing diabetes.

### Chi-squared test for categorical variables

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Similar to the t test I performed with BMI, for the categorical features I wanted to see if there is a significant association between the feature and the diagnosis. For all the features the p value is < 0.05, meaning there is a statistically significant association. While this means that there is an association with diagnosis, the results could also be impacted by the size of the dataset and the differences in distribution of diabetes diagnosis. Because of this I should perform more analysis to see which features would be the best predictors of diabetes outcome.

## Task 4- Exploratory analysis to answer questions

### Question 1- How do lifestyle choices (smoker, physical activity, fruits, veggies and alcohol consumption) impact diabetes outcome?

To understand how the lifestyle features smoker, physical activity, consuming fruits and vegetables, and heavy alcohol consumption impact a person's diabetes diagnosis I first investigated the proportion of people in each group with and without diabetes. Since all these features are binary, I could split them into 4 groups. For example: is a smoker and doesn’t have diabetes, is a smoker and has diabetes, is not a smoker and doesn’t have diabetes, and is not a smoker and does not have diabetes. I calculated these proportions for each feature, and plotted bar charts of this information.

A screenshot of a computer screen

Description automatically generated

A close-up of a number

Description automatically generated

A screenshot of a graph

Description automatically generated

For most of the features, the proportion of people with a specific value for the feature seems to be fairly even between having vs not having diabetes (ie for fruits the proportion of people with and without diabetes for 1 (eats fruits daily) is not very different, 58% vs 61%). However, there are some notable gaps, largely seen in physical activity. The non-diabetes group appears to have more physical activity than the diabetes group.

I next wanted to look at potential interaction effects between the features, and see if combining some of them would effects diabetes outcome. I created the interactions Smoker\*PhysActivity, Smoker\*HvyAlcoholConsump, PhysActivity\* HvyAlcoholConsump, and Fruits\*Veggies. I then ran logistic regression with these and the original features.

A screenshot of a document

Description automatically generated

I can use the coefficients to understand how each feature or interaction affects the likelihood of predicting a patient having diabetes. It’s important to note that all of them have p values < 0.05 except for Smoker\*HvyAlcoholConsump, meaning that interaction is statistically significant for all except Smoker\*HvyAlcoholConsump. A negative coefficient means that a positive value in the feature is associated with a decreased likelihood of diabetes. This includes physical activity, veggies, heavy alcohol consumption, physical activity\*heavy alcohol consumption, and fruits\*veggies. A positive coefficient means that having a positive value in the feature is associated with an increased likelihood of diabetes. This includes smoker, fruits, smoker\*physical activity, and smoker\*heavy alcohol consumption (not statistically significant).

This reveals that overall making healthy lifestyle choices appears to decrease the odds of having diabetes. Physical activity and eating fruits and vegetables all contribute to lowered odds, while smoking and its interactions with physical activity increase odds. It’s interesting that heavy alcohol consumption appears to have a positive impact on lowering the likelihood of diabetes. This is a feature that could be investigated further.

Lastly, I generated a heatmap of correlation to visualize the association between the features. Not surprisingly, the interactions I created as most strongly correlated with the features they include. Other associations of note are physical activity with fruits and veggies, and fruits with veggies.

A screenshot of a graph

Description automatically generated

Overall, there appears to be a relationship between having a healthy lifestyle (having physical activity and eating healthy) with not having diabetes. Also having a poor lifestyle (smoking, no physical activity, poor diet) being related to developing diabetes. While I can’t say that these variables cause diabetes, there does appear to be a relationship.

### Question 2- How does socio economic factors effect on diabetes (anyhealthcare, nodocbccost, education, income)?

I also wanted to investigate how socioeconomic factors such as having access to healthcare, being able to afford a doctor, education, and income impact diabetes.

Similar to the above question, I first investigated the proportion of people in each group with and without diabetes. I split the proportions up by feature value. AnyHealthcare, and NoDocbcCost are binary, so the potential values are only 0 or 1. For Education the values range from 1-6 and Income ranges from 1-8. These ranges represent different education and income levels. I then calculated the proportion for each category that has diabetes and does not have diabetes. I then created bar charts with these results.

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

The results for AnyHealthcare and NoDocbcCost appear to be very similar for diabetes and non-diabetes. However, there are some notable differences in education and income. The highest level of education for both education and income have large differences between diabetes and non-diabetes. Non-diabetic people appear to disproportionately have higher education and income compared to diabetic people. For education this difference is 29% vs 40%, and for income its 20% vs 33%.

I wanted to further investigate income and see if stratifying the remaining data by income group would reveal any insights. To do this I first divided the income column by label and looked at the distribution.

A screenshot of a computer code

Description automatically generated

The distribution follows what we would expect after seeing the bar chart, with each bin being larger than the one below.

I then repeated a similar process to look at the proportion of each feature, but this time instead of calculating proportions based on Diabetes\_binary, I based it on the income bins. I first looked at the AnyHealthcare feature.

A graph of a number of people

Description automatically generated

We can see that overall people in the dataset have healthcare. However, there is a notable decline in people without healthcare (0) as income increases.

I next looked at NoDocbcCost which looks at if someone in the past year couldn’t see a doctor because of the cost.

A graph of income bars

Description automatically generated with medium confidence

Again, most of the people in the dataset were able to access doctors, but the amount of people not being able to access doctors decreases as income increases.

Lastly, I looked at Education.

A graph of income in different colors

Description automatically generated with medium confidence

There are 6 levels of education, 1 being no school and 6 being 4 years of college. As income increases so does the percentage of people with higher levels of education.

From all of these results, we can see that income impacts access to healthcare, affording doctors, and education. We also saw that all 4 of these features appear to impact diabetes diagnosis, with education and income being the more important ones. While the results are logical and don’t reveal any surprising information (it makes sense that higher income people are able to access healthcare and education more easily than lower income people) it is still an important observation to make when thinking about preventing diabetes and who needs to most access to resources.

### Question 3- Which features are most important in predicting diabetes?

Lastly, I wanted to understand which features will be most important in predicting diabetes. While most of my work so far focused on EDA and understanding the dataset, I also wanted to look ahead to potential machine learning applications and know which features I might want to use in a project predicting a person’s Diabetes\_binary value.

To do this I used a forest of trees from scikit-learn to evaluate the importance of features.

A graph with blue and white bars

Description automatically generated

From the above plot, we can see that BMI, Age, Income, PhysHlth, and Education are the top 5 most important features. During previous analysis, I already identified Income, PhysHlth, and Education are seemingly important features that are contributing to diabetes diagnosis. Knowing these features can help inform future classification and prediction models.

Since I haven’t investigated BMI or age yet, I decided to follow a similar approach I did in the previous 2 questions to better understand these features.

A screenshot of a white paper with numbers and text

Description automatically generated A group of blue and orange bars

Description automatically generated

I’ve already investigated Income, PhysHlth, and Education so I won’t focus on that analysis here, but for BMI and Age there are some interesting trends going on. At lower BMIs there are many more people without diabetes, and at the largest BMI there is more people with diabetes. At younger ages, there are many more people without diabetes, and as age increases so does the percentage of people with diabetes.

I also looked at the impact of potential interactions: BMI\*Age, BMI\*Income, PhysActivity\*BMI, Age\*Education. I ran logistic regression with all these features and looked at the impact.

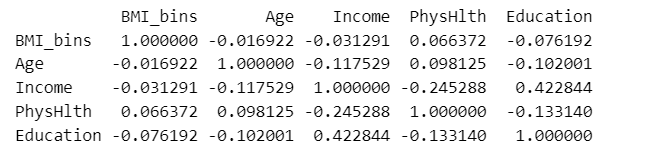
A screenshot of a logit graph

Description automatically generated

I can use the coefficients to understand how each feature or interaction affects the likelihood of predicting a patient having diabetes. It’s important to note that all of them have p values < 0.05 so they’re statistically significant. A negative coefficient means that a positive value in the feature is associated with a decreased likelihood of diabetes. This includes age, income, education, and PhysActivity\*BMI. A positive coefficient means that having a positive value in the feature is associated with an increased likelihood of diabetes. This includes BMI, PhysHlth, BMI\*Age, BMI\*Income, and Age\*Education.

The biggest takeaway from this is that a greater BMI is associated with having diabetes, while having an increased income and education (i.e. improved socioeconomic resources) is associated with a lower chance of diabetes.

Lastly, I also computed the cosine similarity of the top features to understand how similar they are.



We can see that many of the features are not similar to each other. The greatest similarity is between education and income, and the greatest dissimilarity is between physical health and income. This is interesting because although these features are the most important in the dataset in terms of diabetes, they are not very similar. This could be an advantage because it means I have a range of information informing predictions.

## Task 5- Explore biases

### Demographic bias

I first wanted to investigate potential bias among the demographic features (sex, age, income) in the dataset. Like I have in other problems, I calculated the proportion of diabetes for the different features and looked at the distribution. A screenshot of a data

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A graph of different colored lines

Description automatically generated

The distribution for sex is fairly even. There are slightly more female (0) than male (1) but overall does not seem concerning. There is a skew to the age distribution. With most of the people, especially the people with diabetes, being older. However, this is logical because people tend to develop type 2 diabetes later in life so we would expect to see more of those people being older. The most concerning aspect of potential bias is in income. We can clearly see a disproportionate difference in the ratio of diabetes to non-diabetes at low and high incomes. Lower income appears to be mostly diabetes, whereas high income is mostly non-diabetes. This could be due to bias in the dataset and how sampling was conducted. However, it could also be due to larger issues surrounding food and healthcare access. Lower income people may not be able to access healthy foods or support a healthy lifestyle, leading to increased risk of diabetes.

### Sample size bias

I next wanted to look into sample size bias and check the sample size for different groups and features. I already know that there are less people in the data set with diabetes, but I wanted to further investigate this issue.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A number with numbers on it

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

In general, the distribution of diabetes vs non-diabetes within a specific value for a feature does not seem to be a huge concern. There are some instances among the dataset, such as higher-level education and income which do have clear differences, but I’ve already explored those above. However, it is clear there is a big difference between different values in a feature and those proportions. For example, HvyAlcoholConsump, AnyHealthcare, and NoDocbcCost all have 1 value that is the dominant value in that feature for both diabetes and non-diabetes. This may skew my analysis and is important to keep in mind when drawing conclusions.

### Bias in Diabetes\_binary

Although it has already indirectly come up by looking into other features, I wanted to directly investigate bias in Diabetes\_binary and its distribution/proportion in the dataset. According to the CDC, approximately 11.6% of the US has diabetes, and 38% has prediabetes (<https://www.cdc.gov/diabetes/php/data-research/index.html>). To help avoid bias, the dataset should have similar percentages to the real-world prevalence of diabetes.

A blue circle with orange triangle in center

Description automatically generated

A close up of a text

Description automatically generated with medium confidence

The proportion of people in the dataset is 15%. While this is lower than the 38% of people that are prediabetic (included in the Diabetes\_binary ==1), it is still within the range of diabetic and prediabetic prevalence. While having more entries of diabetic people would be valuable insight, I don’t think there is a huge bias in the dataset against diabetic entries.