UE BC Early stage diabetes risk prediction

Kehan Liu

1. contexts

With the development of the economy, people's living conditions have undergone significant changes. This has led to alterations in dietary habits, with more high-sugar and high-fat foods being incorporated into daily consumption. Such lifestyle shifts have resulted in a plethora of health issues, including diabetes. Diabetes has become an increasingly prevalent health concern in today's society, particularly in developed countries. As urbanization and modernization accelerate, people are increasingly inclined towards convenient high-sugar, high-fat foods and sedentary lifestyles lacking in physical activity. These factors collectively contribute to the rising incidence of diabetes.

However, there are some early signs within the body that may help individuals realize more quickly that they may be developing this disease. For instance, some studies suggest that fluctuations in blood sugar levels and frequent thirst may be among the early signs of diabetes. Additionally, persistent feelings of fatigue, dramatic changes in weight, and blurred vision are symptoms that could also hint at potential diabetes risk. Therefore, by paying attention to and promptly detecting these bodily signals, individuals can identify potential health issues earlier and take appropriate preventive measures to reduce the risk of developing the disease.

2. problematique

Determine which signs in a person may lead to having diabetes. Calculate the possibility(NO, LOW, MIDDLE, HIGH, VERY HIGH) that the user has diabetes based on the corresponding signs in the user's body. It also tells the user which signs can lead to diabetes and gives advice on how to solve the corresponding signs.

3. Target audience:

Health-conscious and preventive population(Bangladesh)

4. Data selection and analysis

1. Data selected for the plan

1. **Obesity**: Obesity is one of the major risk factors for developing type 2 diabetes. Excess fat can lead to insulin resistance, increasing the risk of diabetes.
2. **Sex**: Gender may influence the risk of diabetes differently between males and females, although specific effects may vary.
3. **Age**: Age is a significant risk factor for diabetes. As age increases, so does the risk of developing diabetes.
4. **High blood pressure**: High blood pressure is closely associated with diabetes. Individuals with diabetes are more likely to develop high blood pressure, which also increases the risk of diabetes.
5. **High blood sugar**: High blood sugar is a core feature of diabetes. Sustained high blood sugar levels are indicative of type 2 diabetes.
6. **Frequent thirst**: Frequent thirst is an early symptom of diabetes, associated with fluctuations in blood sugar levels.
7. **Persistent feelings of fatigue**: Persistent fatigue may be an early symptom of diabetes, related to unstable blood sugar levels.
8. **Dramatic changes in weight**: Dramatic weight changes may be associated with diabetes, particularly in type 2 diabetes.
9. **Blurred vision**: Blurred vision may be an early symptom of diabetes, related to eye problems caused by high blood sugar.
10. **class of diabetes：** 1.Positive, 2.Negative.

2. dataset selection

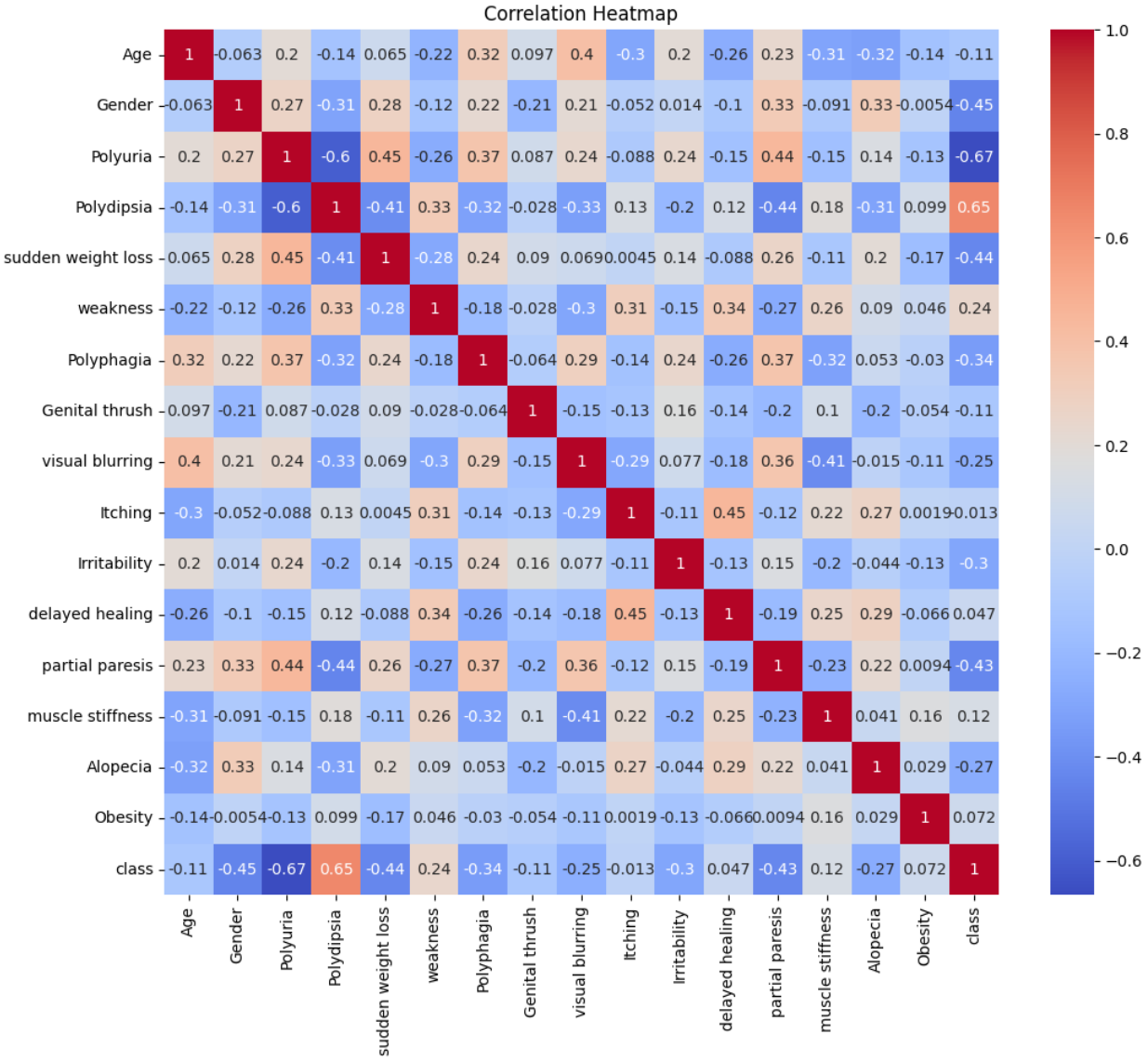
dataset：
https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset
info：This dataset contains the sign and symptpom data of newly diabetic or would be diabetic patient. And this has been collected using direct questionnaires from the patients of Sylhet Diabetes
Hospital in Sylhet, Bangladesh and approved by a doctor.

Variable Information

Age 1.20-65
Sex 1. Male, 2.Female
Polyuria 1.Yes, 2.No.
Polydipsia 1.Yes, 2.No.
sudden weight loss 1.Yes, 2.No.

weakness 1.Yes, 2.No.
Polyphagia 1.Yes, 2.No.
Genital thrush 1.Yes, 2.No.
visual blurring 1.Yes, 2.No.
Itching 1.Yes, 2.No.
Irritability 1.Yes, 2.No.
delayed healing 1.Yes, 2.No.
partial paresis 1.Yes, 2.No.
muscle stiness 1.Yes, 2.No.
Alopecia 1.Yes, 2.No.
Obesity 1.Yes, 2.No.
Class 1.Positive, 2.Negative.

3. Correlation



The numbers shown in the figure are Pearson's correlation coefficients, which measure the strength and direction of the linear correlation between two variables.

1. When the correlation coefficient is -1, it indicates a perfect negative correlation. This means that as one variable increases, the other decreases.
2. When the correlation coefficient is 0, it means that there is no linear correlation. This means that there is no linear correlation between the two variables.
3. When the correlation coefficient is 1, it indicates a perfect positive correlation. This means that when one variable increases, the other variable also increases.

We can see that the coefficient of the target variable (class) is -0.45 with Gender, -0.67 with Polyuria, -0.44 with sudden weight loss, and -0.43 with partial paresis.We can assume that these variables have a strong negative correlation with our target variable relationship.

We can see that the coefficient of the target variable (class) is -0.34 with Polyphagia, -0.25 with visual blurring, -0.3 with Irritability, and -0.27 with Alopecia.We can assume that there is some negative correlation between these variables and our target variable Relationship.

We can see that the coefficient of the target variable (class) with Polyuria is 0.65. we can assume that this variable has a strong positive correlation with our target variable.

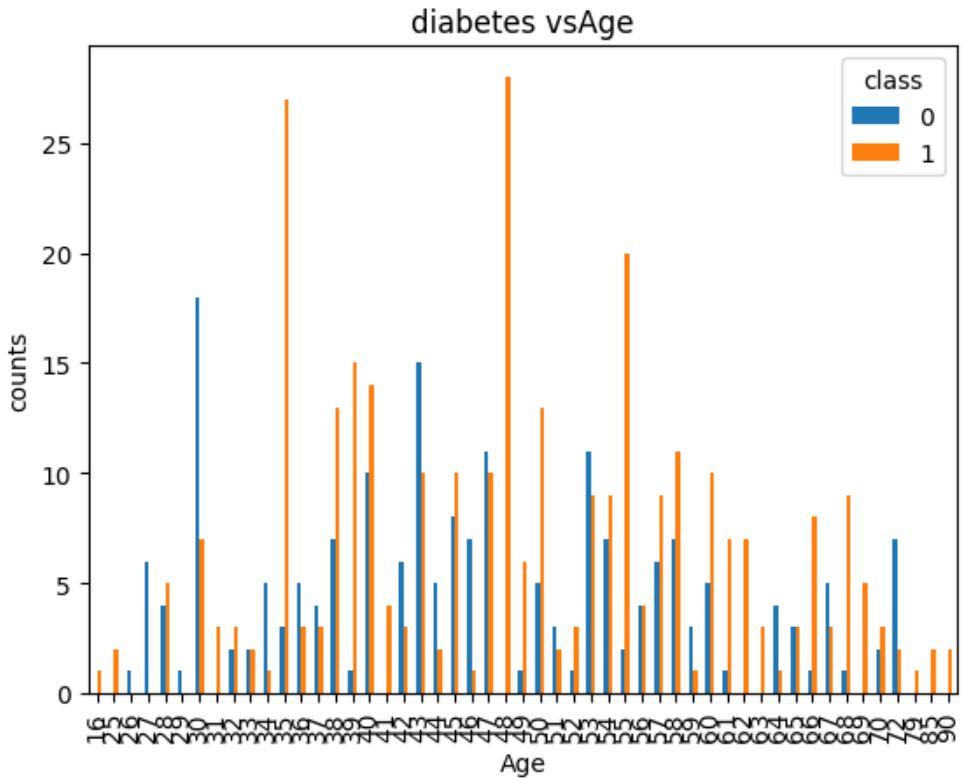
We can see that the coefficient of the target variable (class) with weakness is 0.24. we can assume that there is some positive correlation between this variable and our target variable.

conclude

So we can find that only Gender,Polyuria, Polydipsia,sudden weight loss, weakness, Polyphagia, visual blurring, Irritability, partial paresis, Alopecia have relation with class(target variable)

4. static analysis

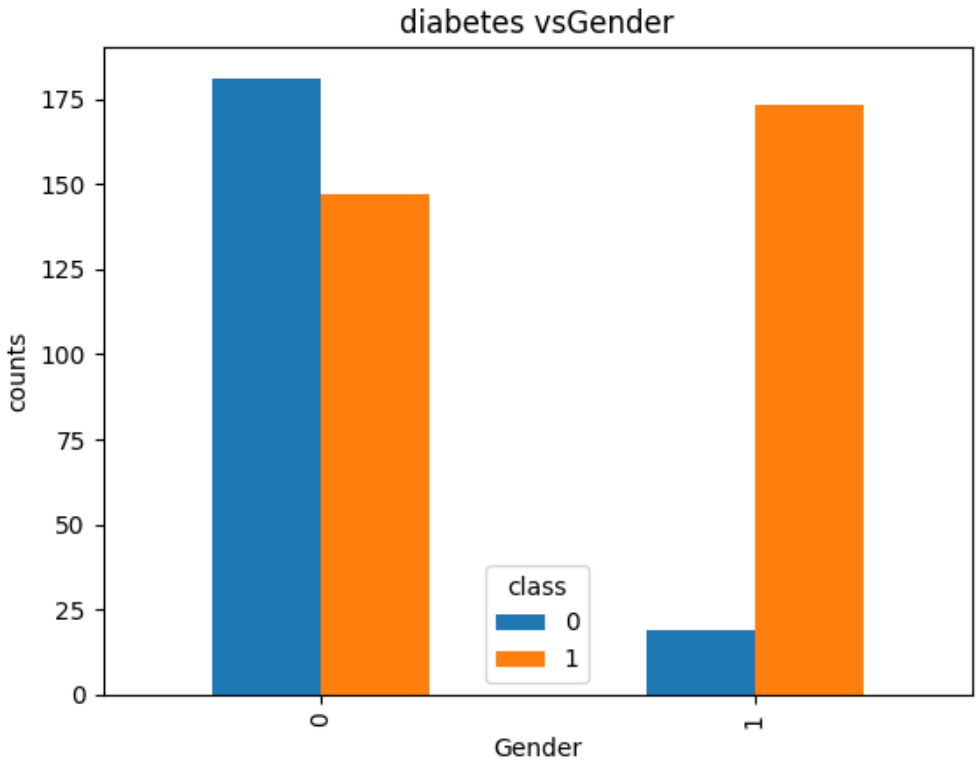
1. **diabetes vs age**



Although our experience tells us that older people are more likely to get diabetes. But we don't see a clear trend in the graphs.

2. **diabetes VS gender**

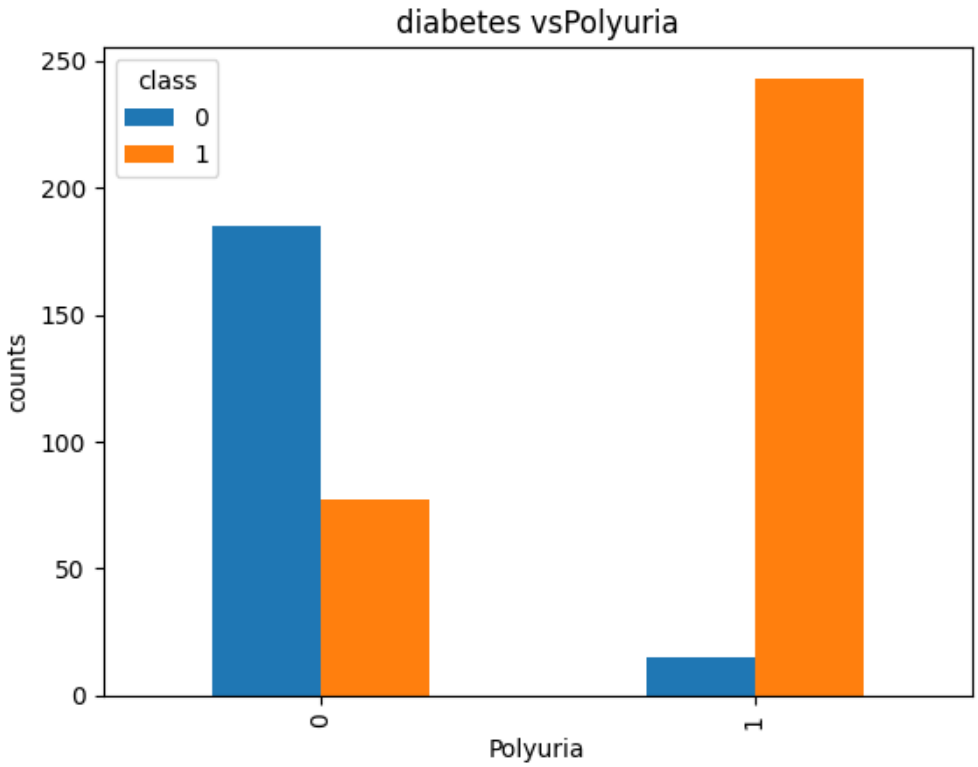
   male(0),Female(1)



We can clearly see that the percentage of women with diabetes is higher than that of men.This proves that gender does have an effect on developing diabetes

Causes:
Physiological differences: Women's physiology is different from men's, for example, changes in hormone levels after menopause may increase a woman's risk of developing diabetes, especially after estrogen levels decline.
Pregnancy and postpartum: During pregnancy, women are at increased risk of developing gestational diabetes. After giving birth, some women may maintain higher body weight and blood sugar levels, increasing the risk of diabetes.

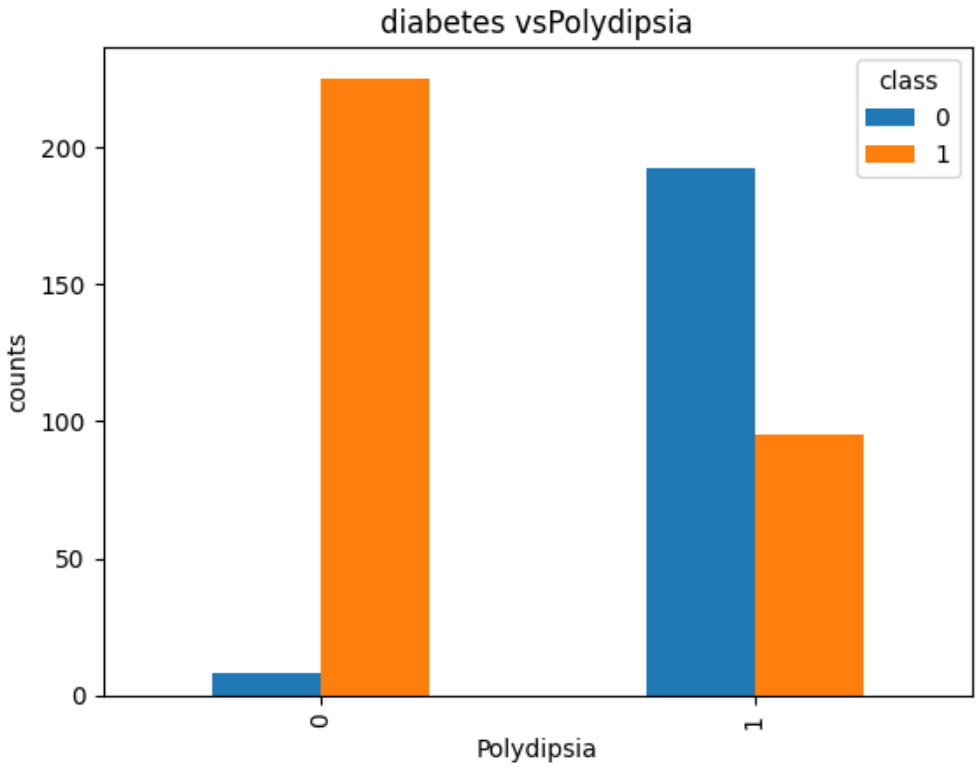
3. **diabetes VS polyuria**



We can see that those with polyuria are much more likely to have diabetes than those without.And most of the people who have this sign are sick.

Cause:
Polyuria is a hallmark symptom of diabetes. High blood sugar levels cause the kidneys to filter the blood to get rid of excess glucose, which can increase the amount of urine produced and lead to frequent urination.

4. **diabetes VS polydipsia**

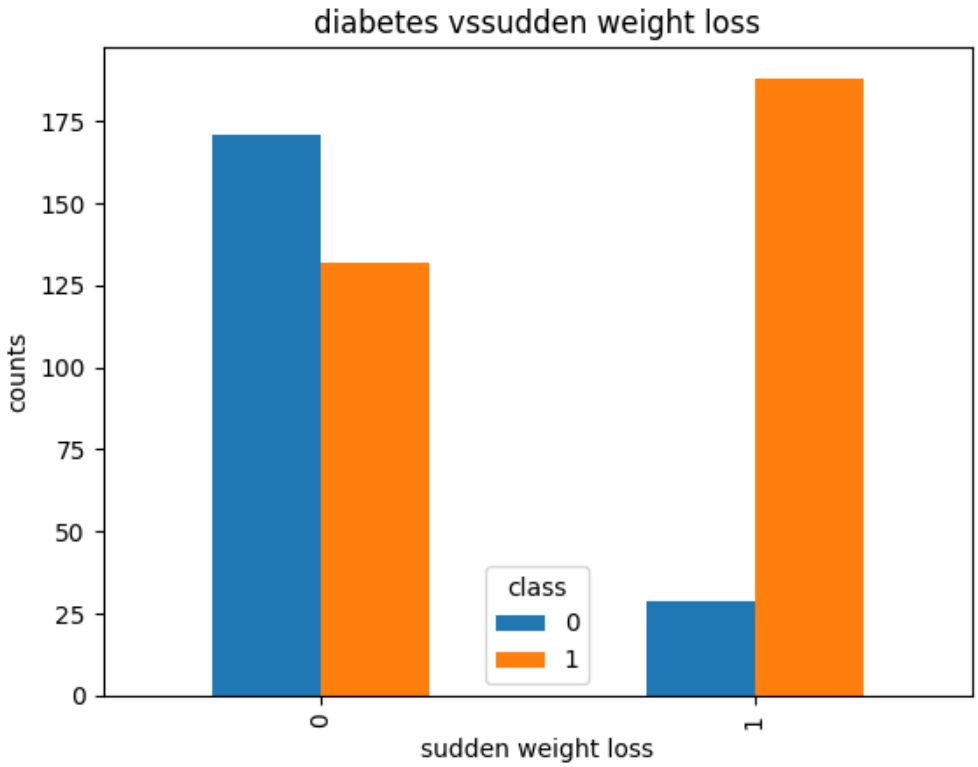


We can see that those without poldipsia are much more likely to have diabetes than those with.

Cause:

Polyhydramnios is due to the stimulation of the thirst center by high blood glucose, causing the patient to feel thirsty and need to drink a lot of water. Therefore, polydipsia is often one of the manifestations of untreated or poorly controlled diabetes.

5. **diabetes VS sudden weight loss**

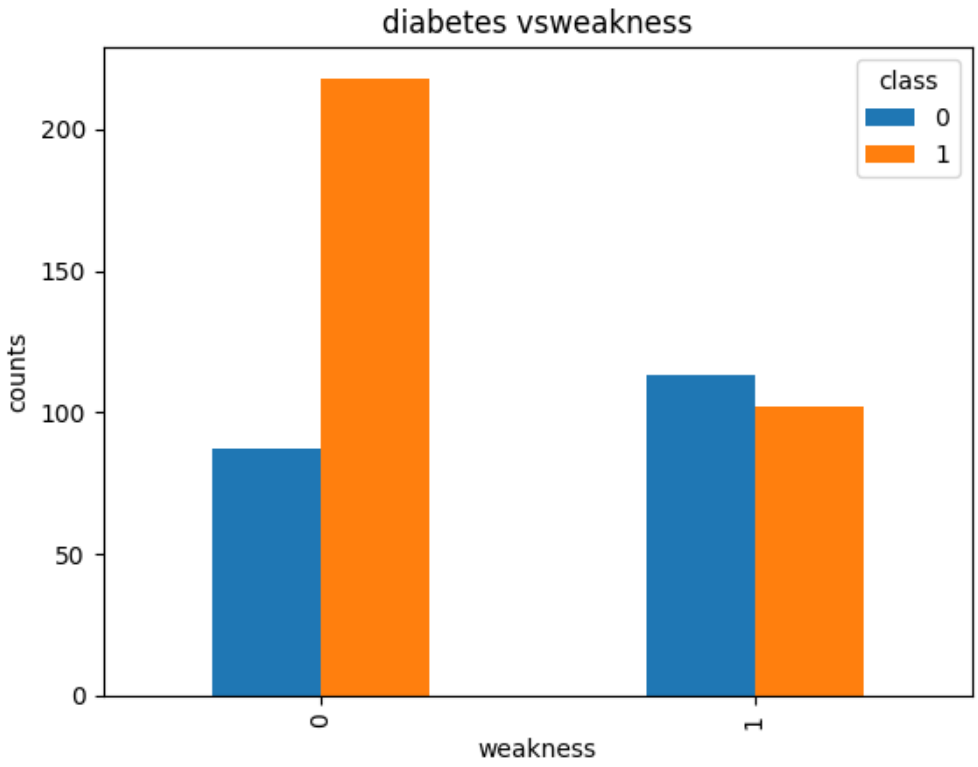


We can clearly see that the rate of disease is higher in those who have this sign than those who don't.And most of the people who have this sign are sick.

Cause:

For individuals with diabetes, the body cannot effectively utilize glucose in the blood to obtain energy. This leads to the breakdown of fats and muscles to provide energy, resulting in weight loss.

6. **diabetes VS weakness**

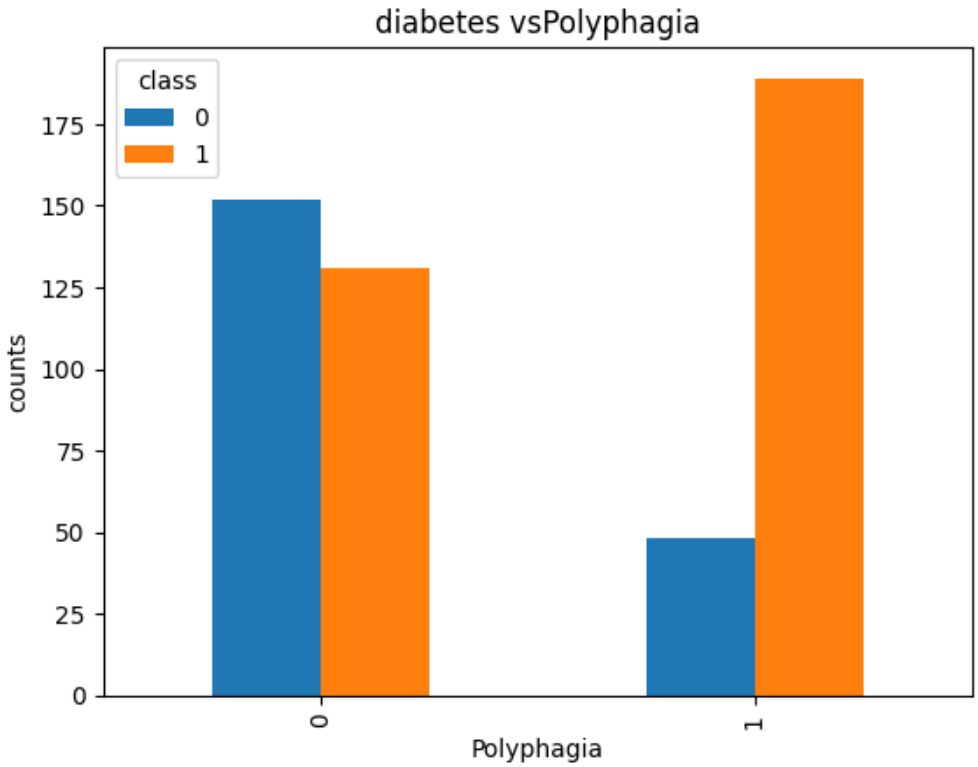


We can see that the rate of disease is higher in those who don't have this sign than those who have. But this difference is not as pronounced as the previous variables. It also means that the relationship with the target variable is not as strong as the previous variable

Cause:

High blood sugar is a common characteristic of diabetes. When blood sugar levels are too high, body cells cannot obtain enough energy, which can lead to a feeling of weakness.

7. **diabetes VS polyphagia**

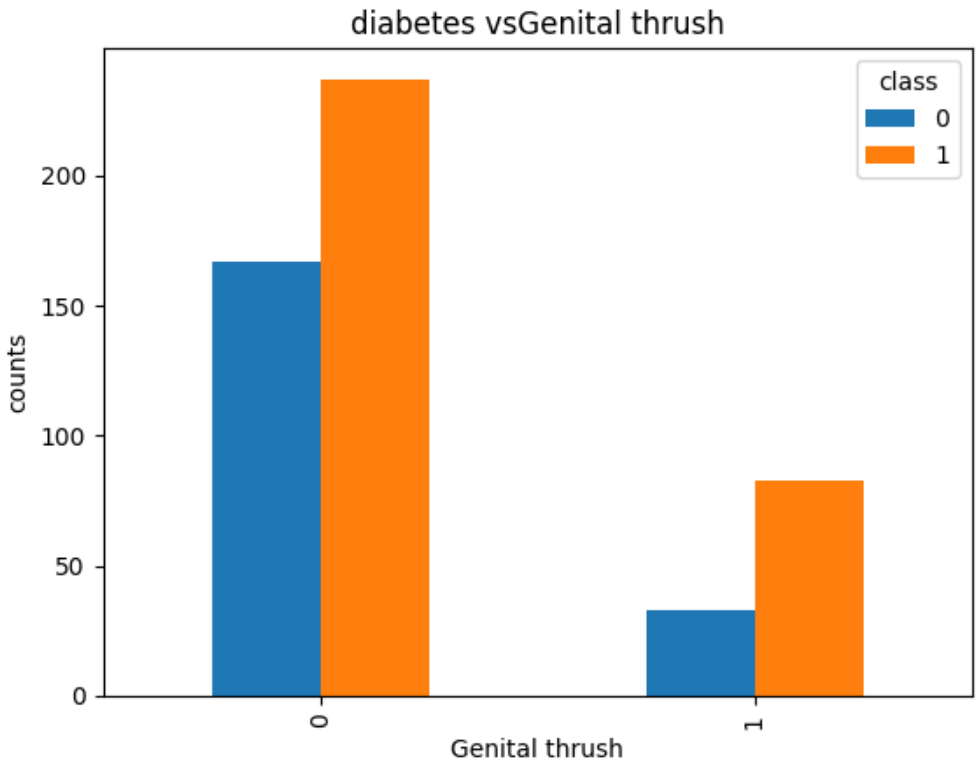


We can clearly see that the rate of disease is higher in those who have this sign than those who don't.And most of the people who have this sign are sick.

Cause:

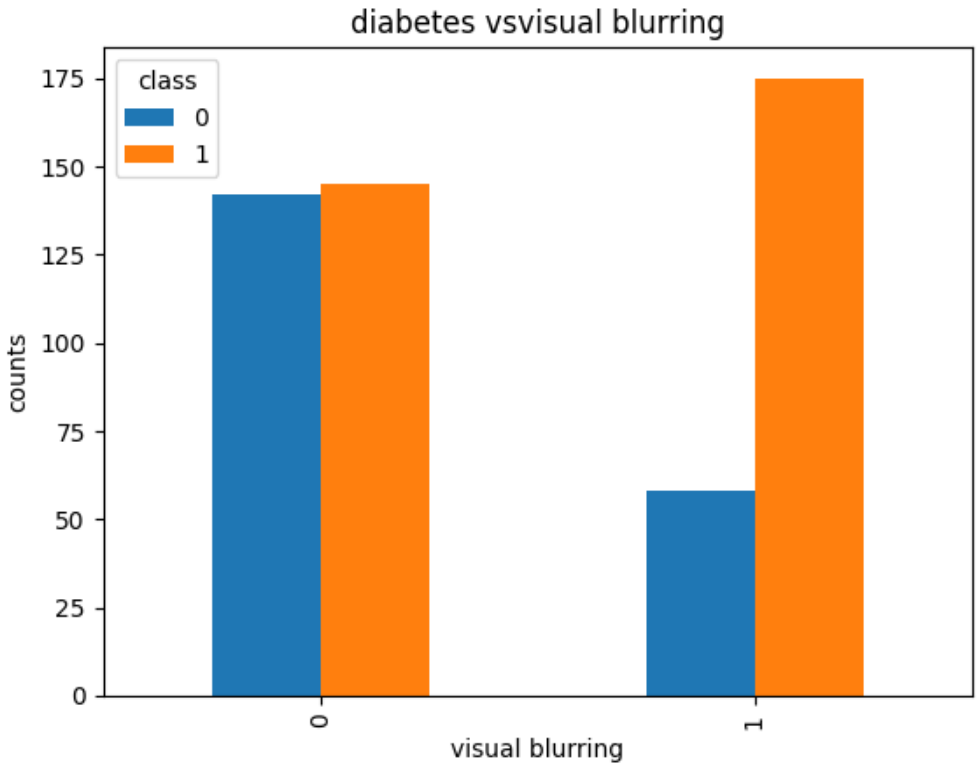
Polyphagia, or increased appetite, is a common symptom of diabetes and is associated with poor blood sugar control and abnormal insulin levels. Because body cells cannot effectively utilize glucose to obtain energy, this leads to feelings of hunger and increased appetite in patients, who attempt to compensate for energy loss through eating.

8. **diabetes VS genital thrush**



The rate of illness is about the same whether or not there is a sign of it.We can assume that it is independent of our target variable

9. **diabetes VS visual blurring**

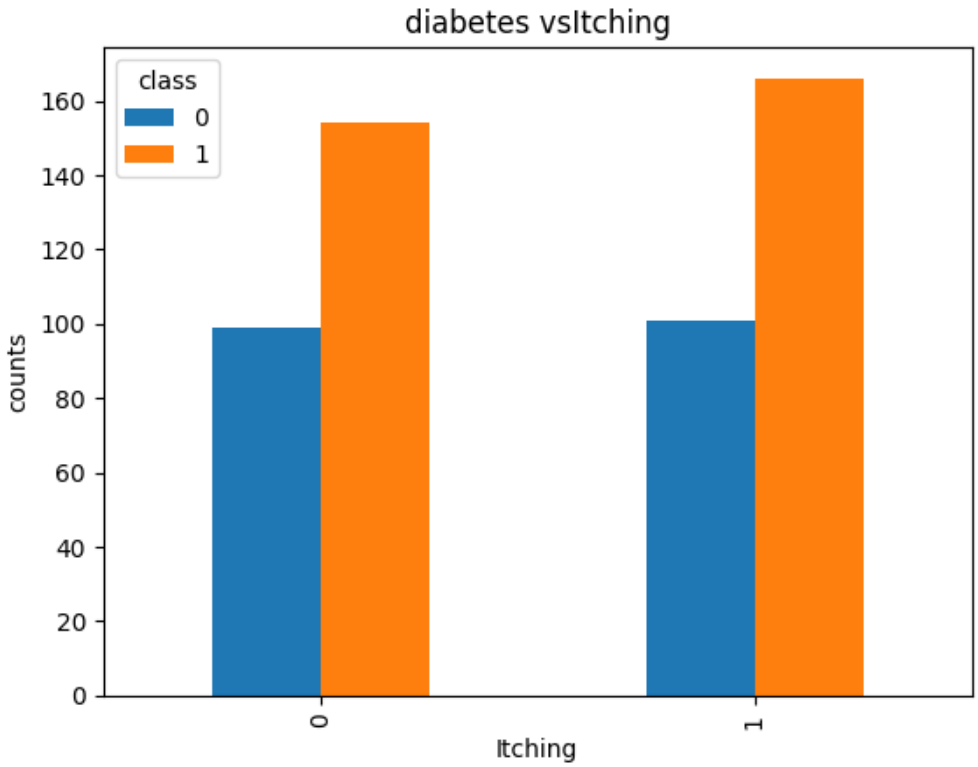


We can clearly see that the rate of disease is higher in those who have this sign than those who don't.And most of the people who have this sign are sick.

Cause:

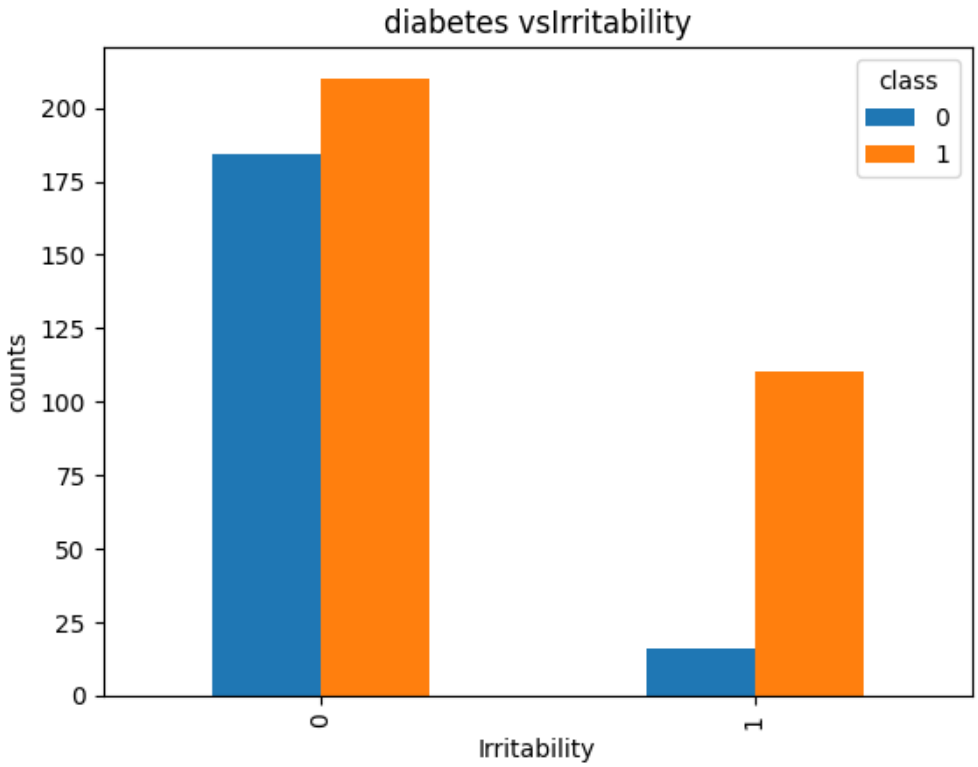
Fluctuations in high blood sugar levels can lead to refractive errors in the eyes, such as nearsightedness or farsightedness, resulting in blurred vision.

10. **diabetes VS itching**



The rate of illness is about the same whether or not there is a sign of it.We can assume that it is independent of our target variable

11. **diabetes VS irritability**

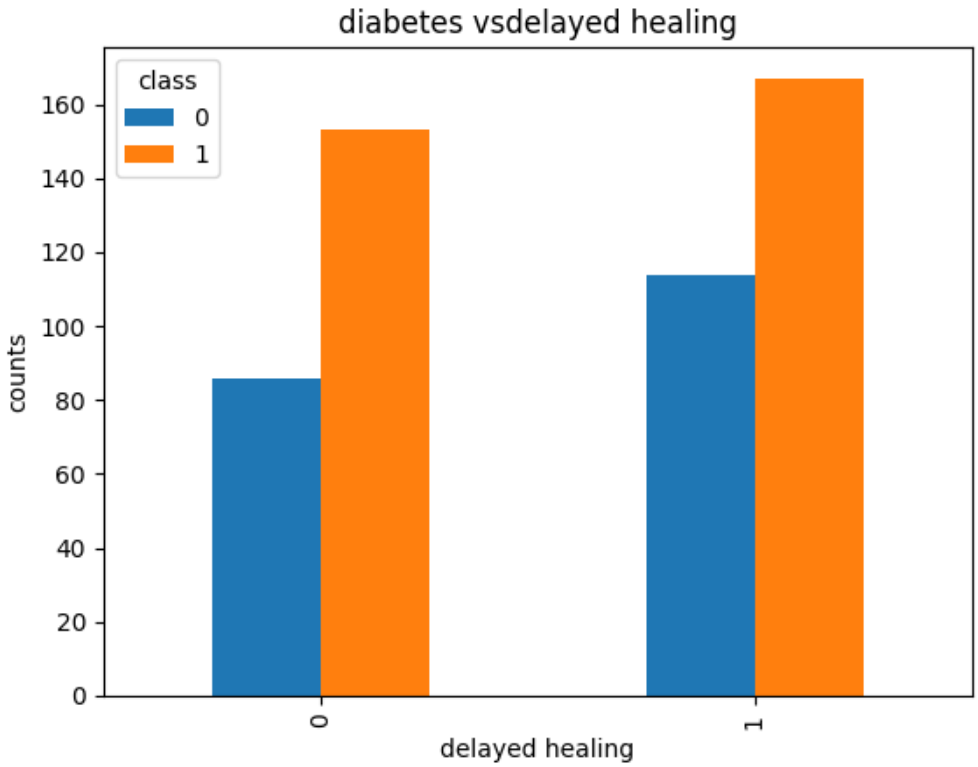


We can clearly see that the rate of disease is higher in those who have this sign than those who don't.And most of the people who have this sign are sick.

Cause:

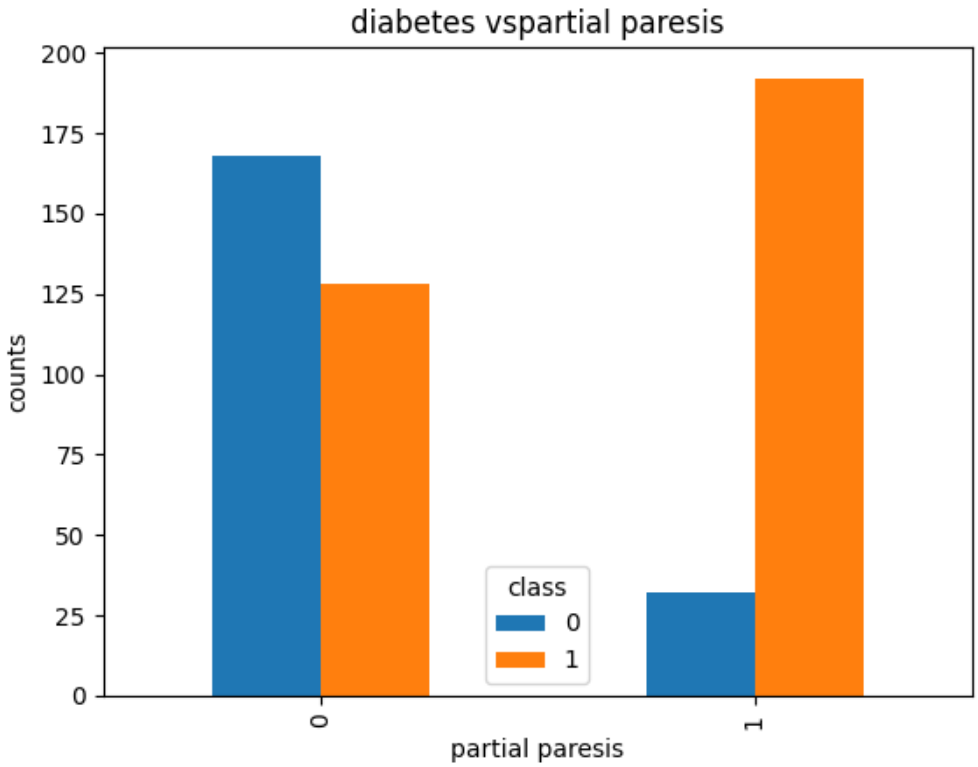
Fluctuations in blood sugar levels may lead to emotional instability, including irritability. Diabetes may be accompanied by other physical discomforts such as pain, fatigue, etc., which can also affect the patient's emotional state, increasing the likelihood of irritability.

12. **diabetes VS delayed healing**



The rate of illness is about the same whether or not there is a sign of it.We can assume that it is independent of our target variable

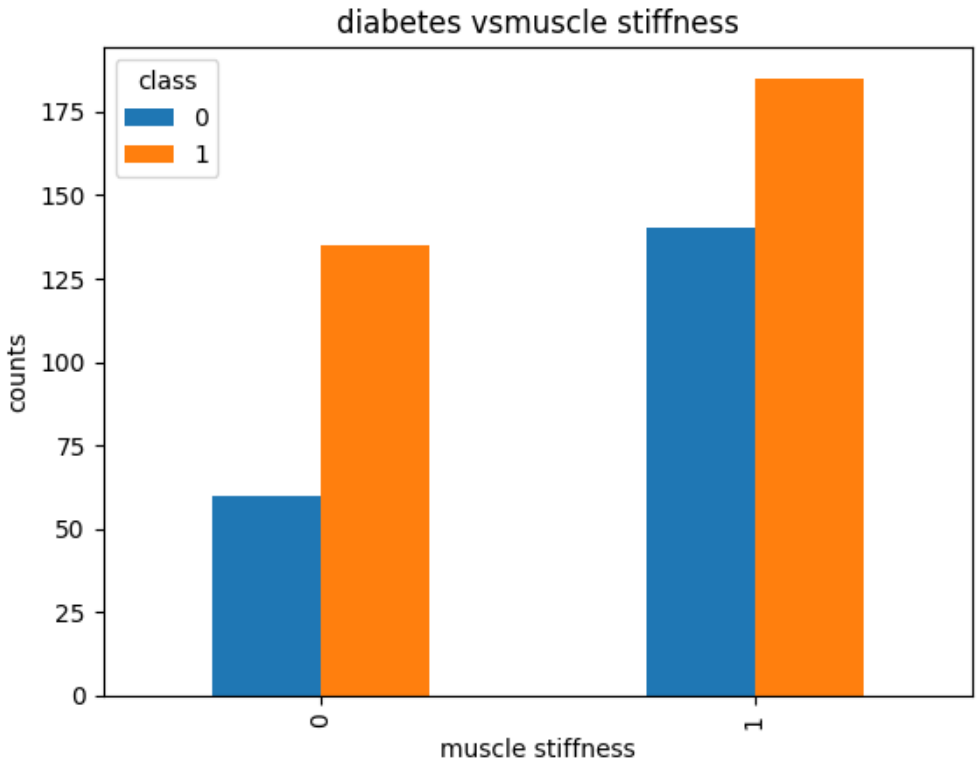
13. **diabetes VS partial paresis**



Cause:

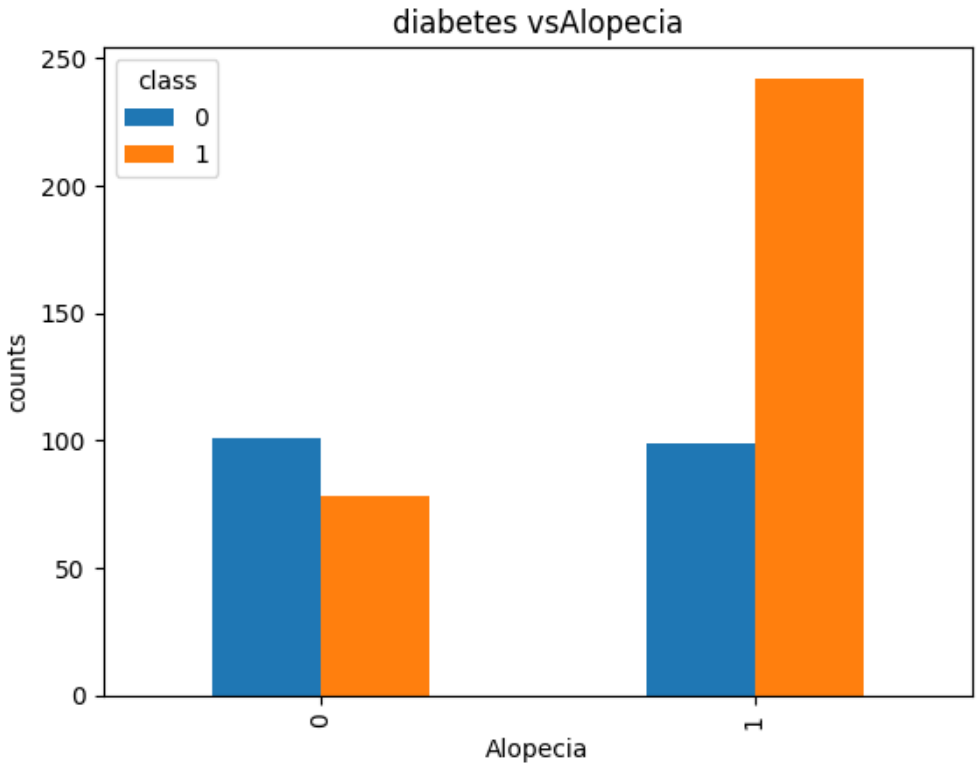
Diabetic neuropathy can cause weakness or partial paralysis in various parts of the body, typically starting in the feet and legs and gradually progressing upward.

14. **diabetes VS muscle stiffness**



The rate of illness is about the same whether or not there is a sign of it.We can assume that it is independent of our target variable

15. **diabetes VS alopecia**

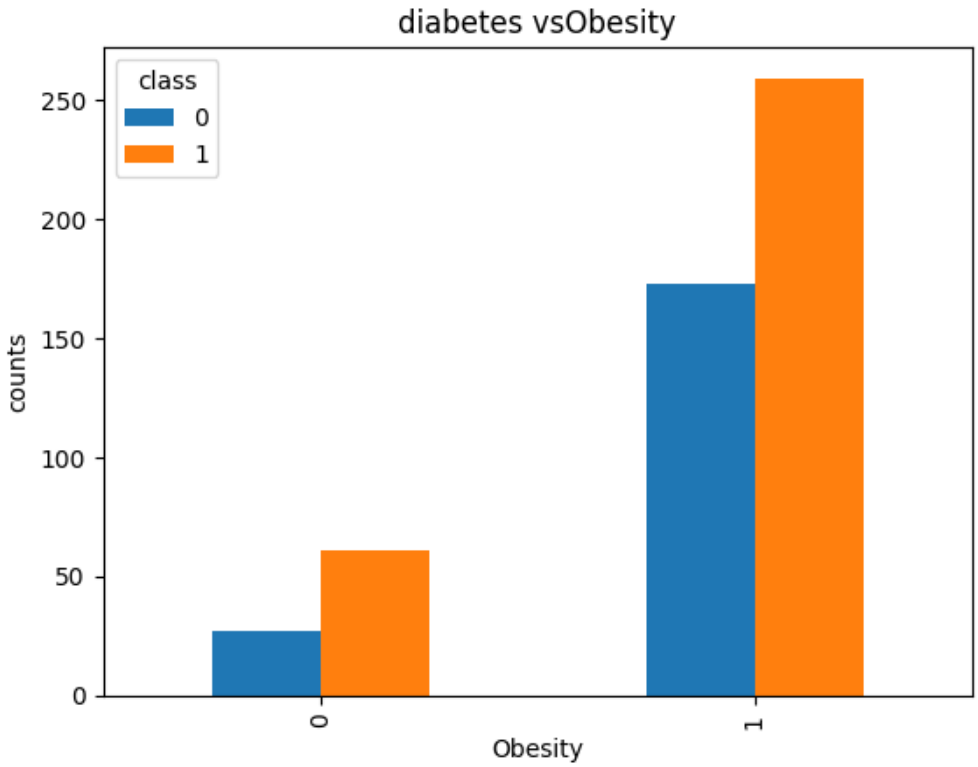


We can clearly see that the rate of disease is higher in those who have this sign than those who don't.And most of the people who have this sign are sick.

Cause:

Poor Circulation: Diabetes can affect blood circulation, including blood flow to the scalp. Reduced blood flow to hair follicles may lead to weakening of the hair and eventual hair loss.

16. **diabetes VS obesity**



The rate of illness is about the same whether or not there is a sign of it.We can assume that it is independent of our target variable

5. Conclude

Based on the heatmap, we can learn the relationship between the elements in the dataset and thus get those variables that are related to our target variable. We then verify that the relationships from the heatmap are correct through static analysis (analyzing the percentage distribution of the target variable under the effect of the variable). And to identify the reasons that led to this relationship. So I think these variables are sufficient to predict the risk of disease

Finally, i decide to choose Gender,Polyuria, Polydipsia,sudden weight loss, weakness, Polyphagia, visual blurring, Irritability, partial paresis, Alopecia and class(target variable) as selected variables.

5. Algorithm and model

The model I chose is a logistic regression model

reason：

1. Logistic regression is a simple and effective model with fast computation speed.
2. The output of the logistic regression model is a probability value, which can be intuitively understood as the probability of belonging to a certain class. Additionally, the magnitude and direction of the feature coefficients can explain the extent and direction of the impact of features on the target variable.

Model Evaluation

Training set accuracy: 0.9086538461538461

Test Set Accuracy: 0.8942307692307693

This means that the model performs relatively well on both the training and test data, and there is no overfitting or underfitting. The model can generalize well to unseen data and has a certain predictive capability.

Add function

I've added the ability to export the factors that increase the possibility of the disease and provide suggestions for them.

6. Final result

By entering your medical condition, we will first provide you with the possibility of having diabetes(no, low , middle, high , high). Then, we will tell you what input variables are the manifestations of diabetes that affect you, and how to solve them.

For example：

> Your recent physical condition is: male, Polyuria✓, Polydipsia×, sudden weight loss×, weakness✓, Polyphagia✓, visual blurring✓, Irritability×, partial paresis×, Alopecia×
>
> you will get the following:

Predicting the probability of disease: very high

Feature Name: Polydipsia
Recommendation: Polydipsia is usually caused by dehydration due to polyuria. It is important to replenish sufficient water to maintain body hydration. However, if polydipsia persists and occurs with other symptoms, medical attention should be sought as soon as possible.

Feature Name: Polyuria
Recommendation: This may be caused by high blood sugar levels, as the kidneys attempt to excrete excess glucose through urine. It is recommended to control blood sugar levels, avoid

consuming too many sugary drinks and foods, and drink plenty of water to maintain body hydration.

Feature Name: Polyphagia
Recommendation: Polyphagia occurs because the body cannot effectively utilize glucose, leading to increased hunger. It is recommended to choose low-GI (glycemic index) foods, such as whole grains, vegetables, and healthy proteins, to control blood sugar fluctuations and reduce the tendency to overeat.

7. system evaluation

I plan to make a questionnaire targeting doctors (who have experience in treating diabetes). This questionnaire will contain 50 scenarios and the doctor will need to determine the probability of developing the disease in each scenario (NO,LOW,MIDDLE,HIGH,VERY HIGH). And give the corresponding solution for the 10 influencing factors. In this way we can compare the results obtained from the questionnaire with our predictions. If the percentage of correctness reaches 90%, we can consider the prediction of the system as true and accurate.And optimize the recommendations of our system again according to the doctor's recommendations.

8. Reference

dataset:https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset

Related articles：https://dergipark.org.tr/en/download/article-file/2050893