EdX: Harvard University: Data Science Capstone: custom project

Ajeet Parmar

August 4th, 2020

Contents

1 Introduction

In 2015, it was estimated that 400 million people, or 5% of the world's population, live with diabetes. If unregistered, one can suffer adverse health effects from a lack of insulin which can lead to death. The objective of this project is to generate a machine learning algorithm which will be able to predict cases of diabetes to an accuracy of over 90%.

This will be accomplished using a dataset from the University of California, Irvine consisting of 520 instances of patients with data collected via direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. The attributes in the dataset are as follows:

- Age
- Gender
- Polyuria
- Polydipsia
- sudden weight loss
- weakness
- Polyphagia
- Genital thrush
- visual blurring
- Itching
- Irritability
- delayed healing
- partial paresis
- muscle stiffness
- Alopecia
- Obesity
- class (whether the patient is diabetic or not)

The key steps involved are as follows:

- splitting the data into two groups: training data and testing data.
- training multiple algorithms using the training data and predicting using the testing data.
- obtaining the accuracy by comparing the output to the testing data's class.

- Aggregating the training algorithms into one ensemble.
- Finding the overall prediction for each case using the ensemble.
- Finding the final accuracy by comparing the ensemble's results to the testing data.

2 Methods/Analysis

2.1 Data cleaning

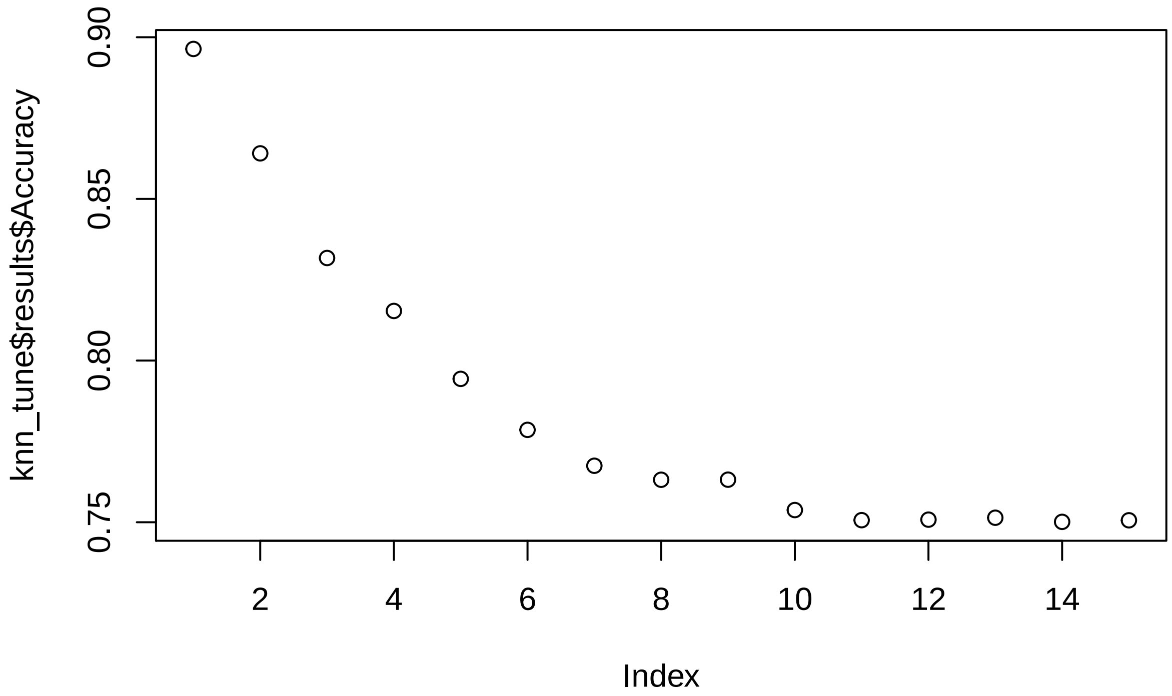
As the data was present in a .csv file, I simply used read_csv() to import the data. Then, I manually made the attributes factors to clean the data. All the data was complete so I did not need to check for NA's. Then, I split the data into the training data and testing data with 60% of the data going to the training data and 40% going to the testing data to avoid overfitting.

2.2 Modeling approaches

I tuned two models: knn and random forest to yield optimal results. The results are below:

2.2.1 K nearest neighbors

With knn, I had to tune my data so I could find out which k value to use. My results for tuning are below:



As seen, when K is 1, I get the highest accuracy. Therefore, k is 1 within my knn training.

2.2.2 Random Forest

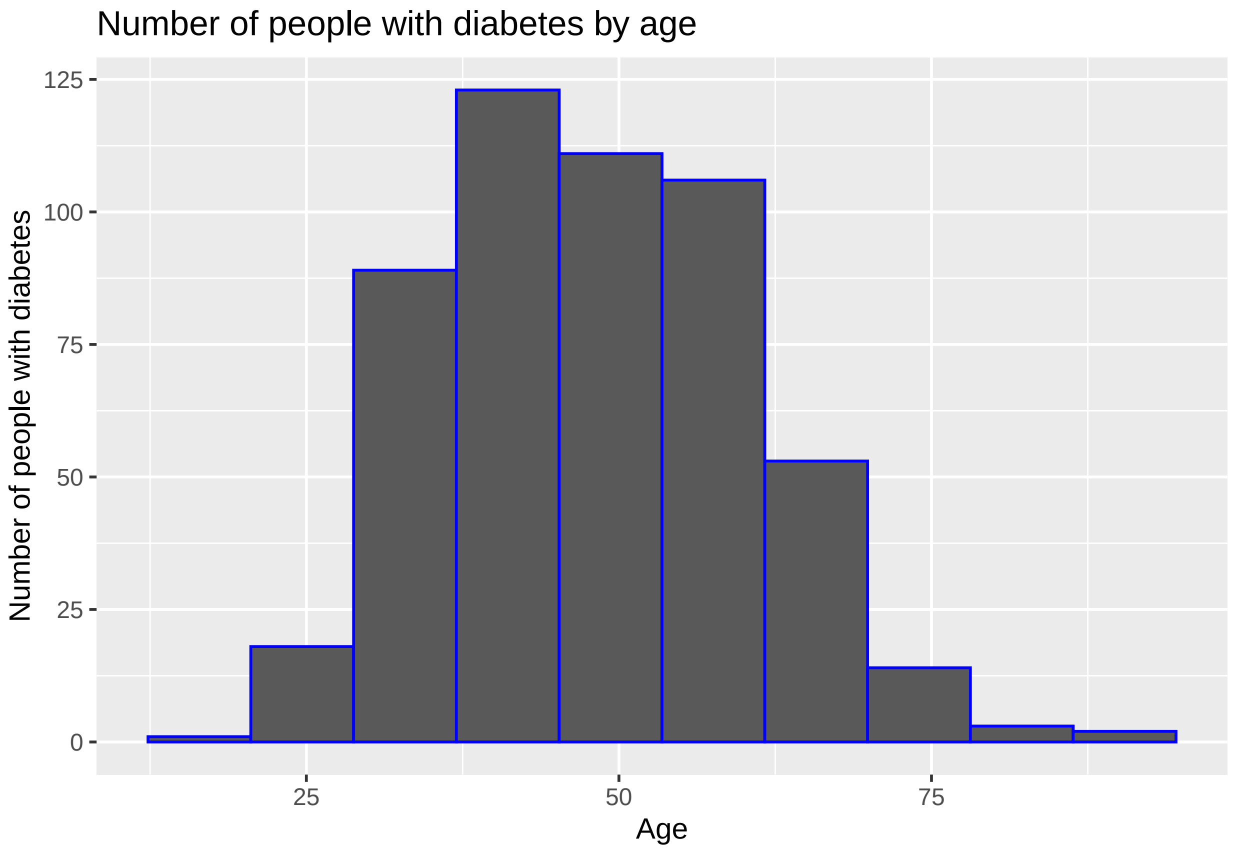
Similarly, I would have to optimize my random forest algorithm. In this, I needed to optimize the “mtry” parameter. Below are the results:

```
##    mtry  Accuracy     Kappa AccuracySD    KappaSD
## 1     1 0.9321493 0.8612653 0.01998277 0.04030402
## 2     2 0.9597015 0.9167771 0.02413118 0.04974522
## 3     3 0.9640155 0.9253308 0.01785501 0.03720360
## 4     4 0.9616882 0.9206871 0.01832602 0.03796320
## 5     5 0.9592843 0.9158954 0.01916703 0.03952567
## 6     6 0.9589951 0.9153468 0.01994290 0.04114611
## 7     7 0.9562163 0.9097266 0.02004743 0.04101212
## 8     8 0.9552308 0.9077679 0.02151277 0.04405839
## 9     9 0.9555145 0.9084152 0.02083027 0.04255363
## 10   10 0.9531079 0.9035173 0.01752697 0.03571838
## 11   11 0.9527282 0.9027955 0.01666439 0.03405012
## 12   12 0.9516988 0.9007050 0.01835021 0.03743760
## 13   13 0.9516909 0.9006858 0.01817777 0.03708414
## 14   14 0.9513612 0.8999871 0.01862730 0.03801525
## 15   15 0.9492926 0.8957435 0.01802152 0.03700511
## 16   16 0.9485531 0.8942425 0.01753547 0.03562920
## 17   17 0.9474740 0.8919757 0.01846258 0.03778049
## 18   18 0.9468472 0.8906200 0.01959684 0.04017915
## 19   19 0.9478887 0.8928544 0.01855028 0.03798913
## 20   20 0.9464802 0.8899600 0.01895437 0.03885143
## 21   21 0.9458543 0.8886131 0.01993260 0.04097450
## 22   22 0.9482200 0.8934806 0.01830845 0.03749073
## 23   23 0.9467952 0.8906242 0.01670230 0.03412990
## 24   24 0.9492882 0.8957188 0.01879852 0.03837812
## 25   25 0.9462279 0.8894842 0.01843518 0.03782187
```

The accuracy is at a maximum when mtry = 4. Thus, this is what I use in my parameter.

2.3 Data exploration and data visualization

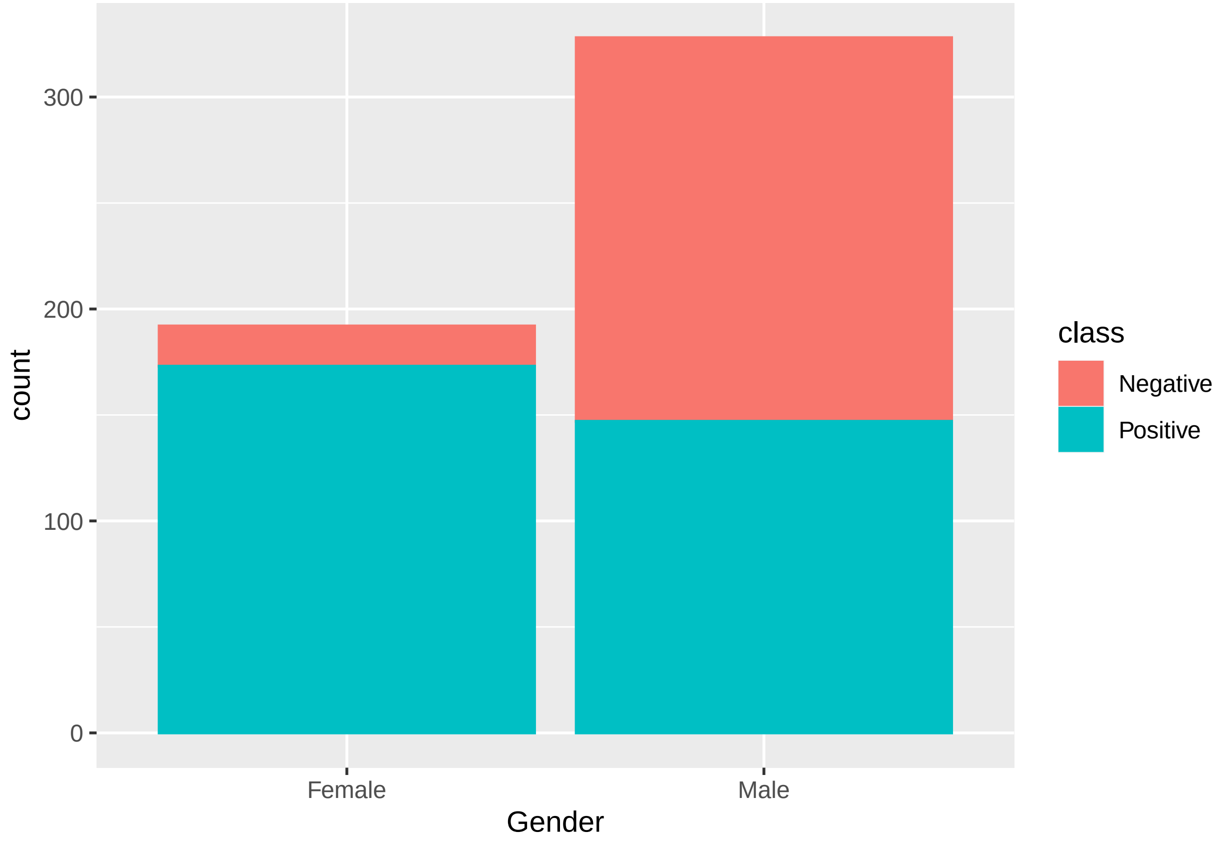
With my quantitative data (age), I was able to graph the number of people with diabetes by age. However, this did not yield any correlation which I could use:



As seen, there is not much correlation. Additionally, there are many ages where there is only one person with diabetes. There is a higher number of people at the middle with diabetes. However, this is attributed to the sampling since there are less samples near the lower end or the higher end of ages. This can be improved by expanding the dataframe to more people worlwide so the dataset can avoid being biased to one region and a certain age bracket. As a result of this, this algorithm can be tailored more to the global population rather than a specific age bracket in Bangladesh.

2.3.1 Cases by gender

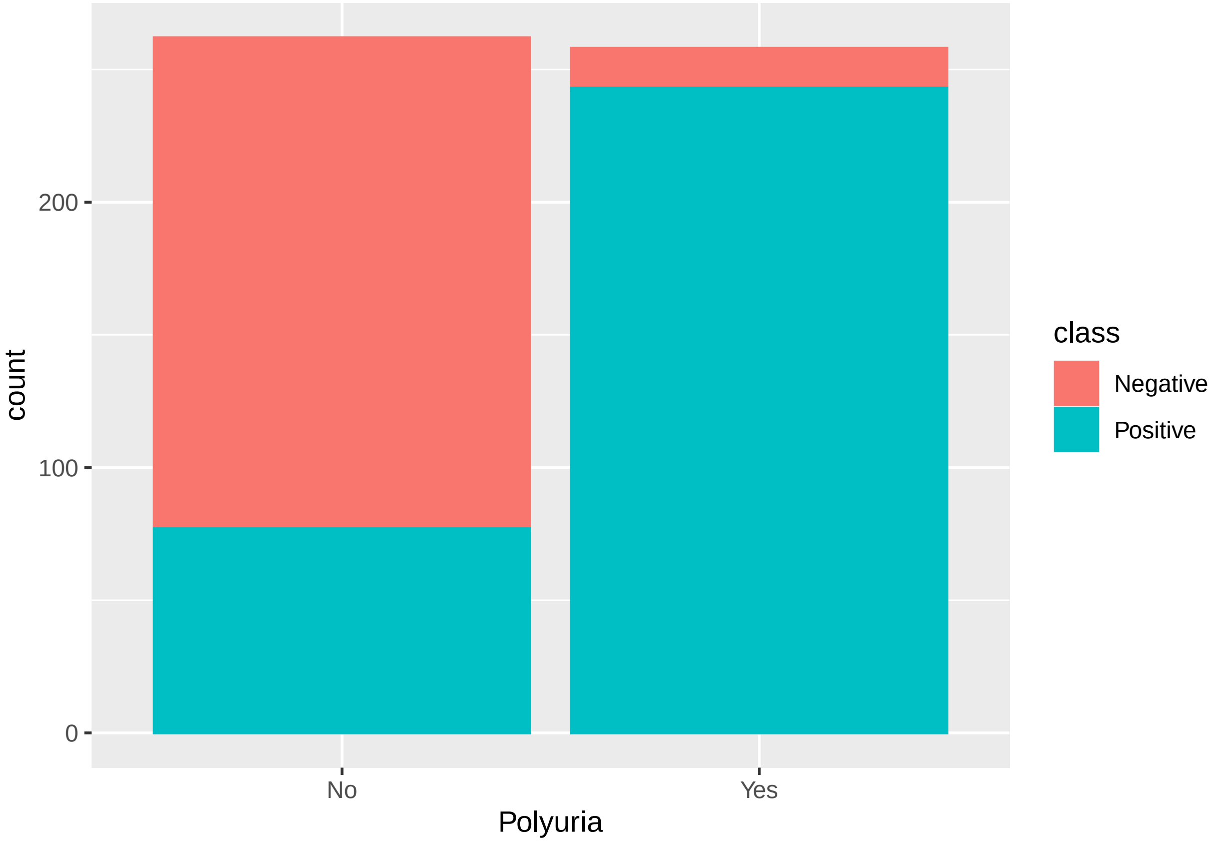
I also tried graphing the number of cases by gender. This did reveal something which could be beneficial:



As seen, females have a higher proportion of cases than males. This could be an insight later seen in our data. It should be noted that this is only unique to our dataset! In real life, diabetes is more common in males than in females due to testosterone levels in males.

2.3.2 Cases by polyuria

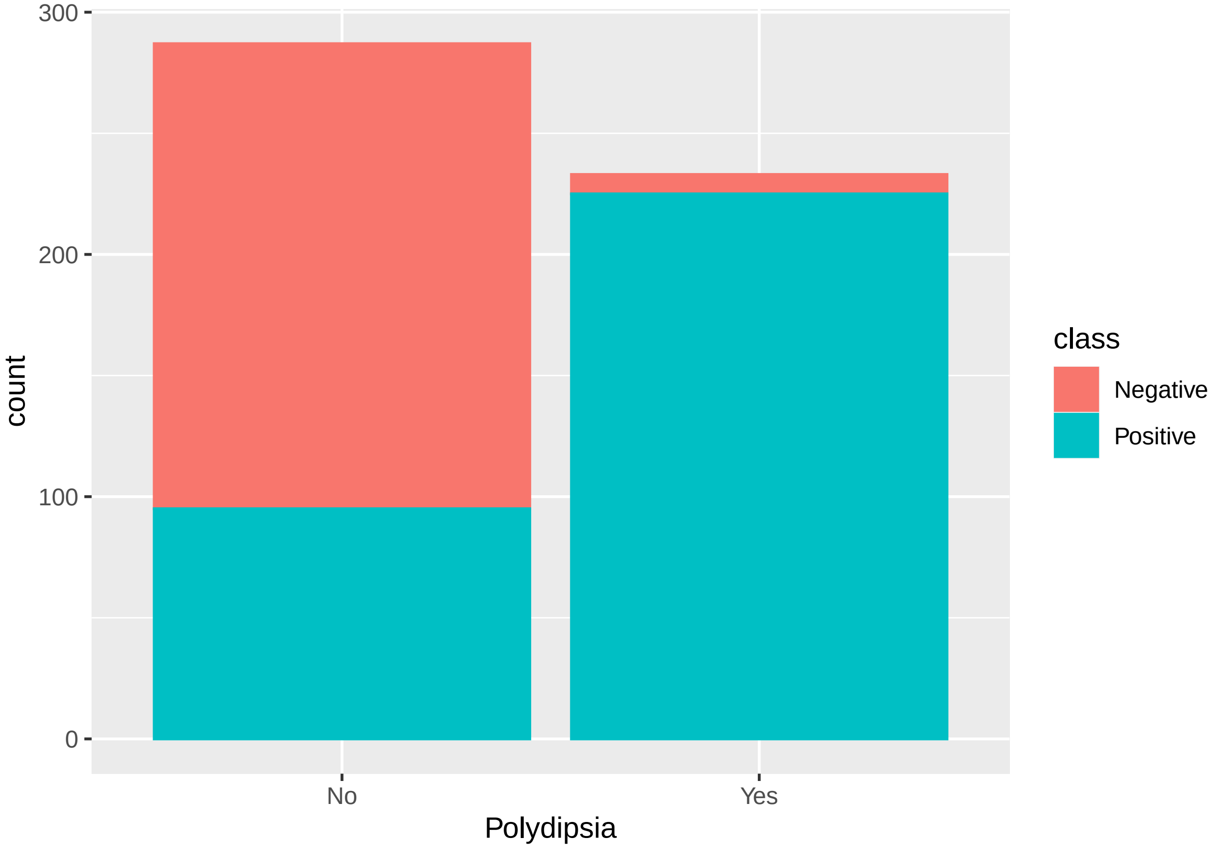
Below is number of cases by polyuria:



Polyuria does play a substantial role as well. It seems that if one suffers from polyuria, they're more likely to have diabetes.

2.3.3 Cases by polydipsia

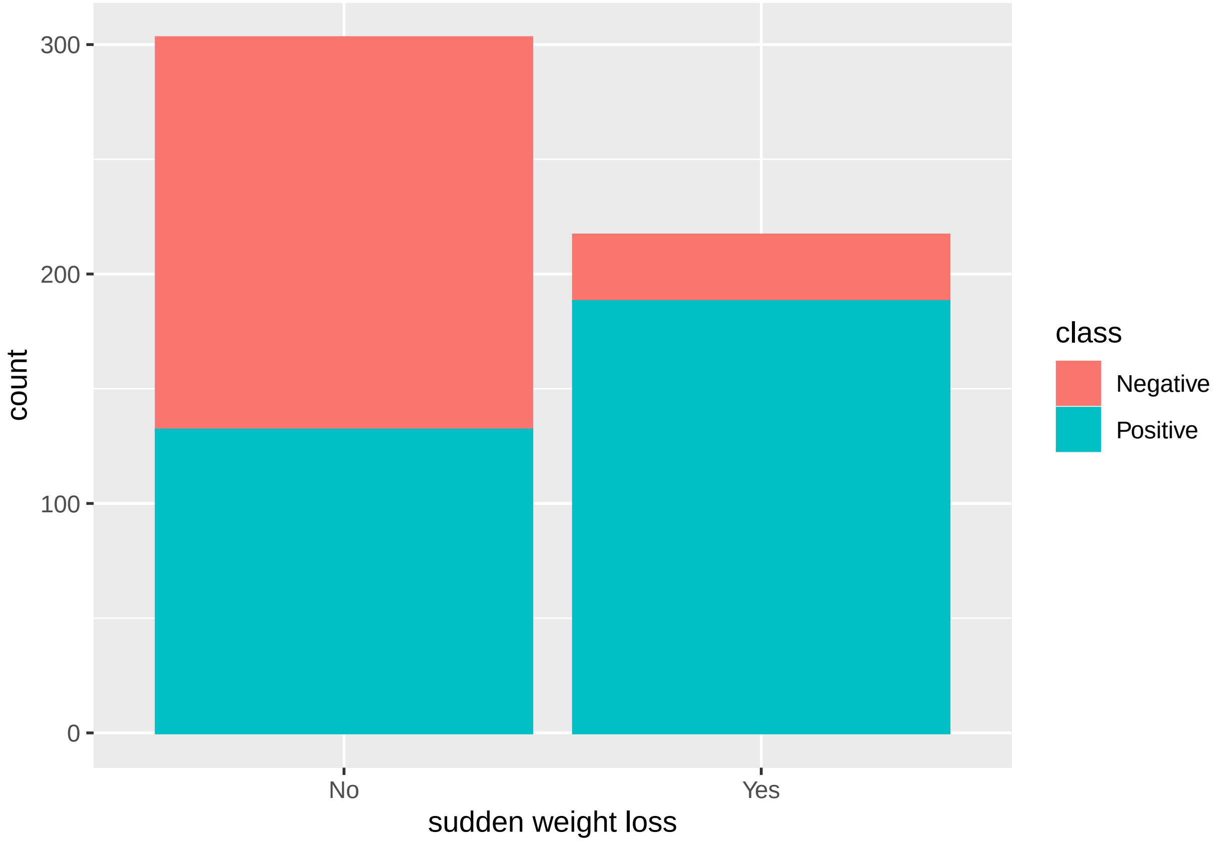
Below is a comparison between polydipsia states and diabetes state.



This also reveals that if one has polydipsia, they are more likely to be a positive case of diabetes.

2.3.4 Cases by sudden weight loss

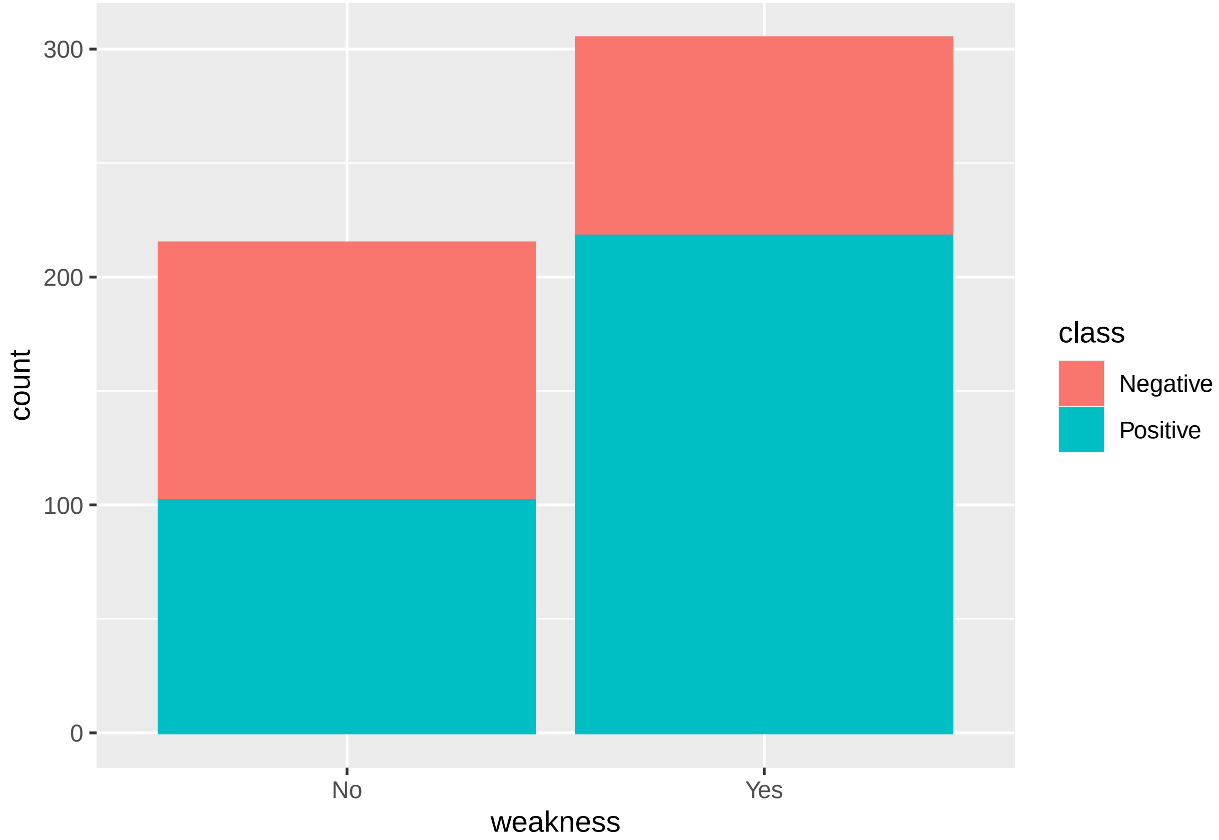
Below is a comparison between sudden weight loss and diabetes state.



This reveals a correlation between weight loss and diabetes where if one experiences sudden weight loss, they are more likely to be a case of diabetes.

2.3.5 Cases by weakness

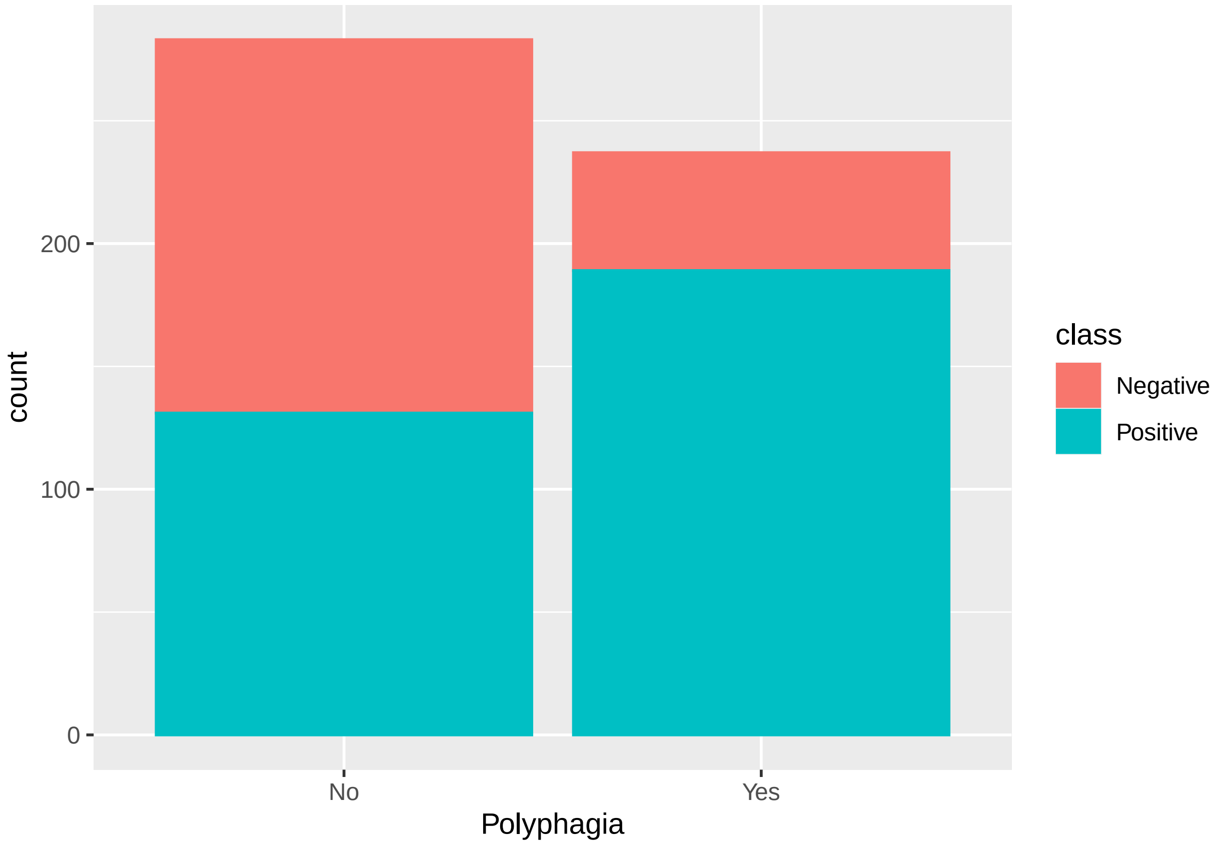
Below is a comparison between weakness and if one has diabetes.



This reveals a slightly weak correlation between weakness and having diabetes. This implies that weakness may not be as valuable of an attribute as other attributes.

2.3.6 Cases by polyphagia

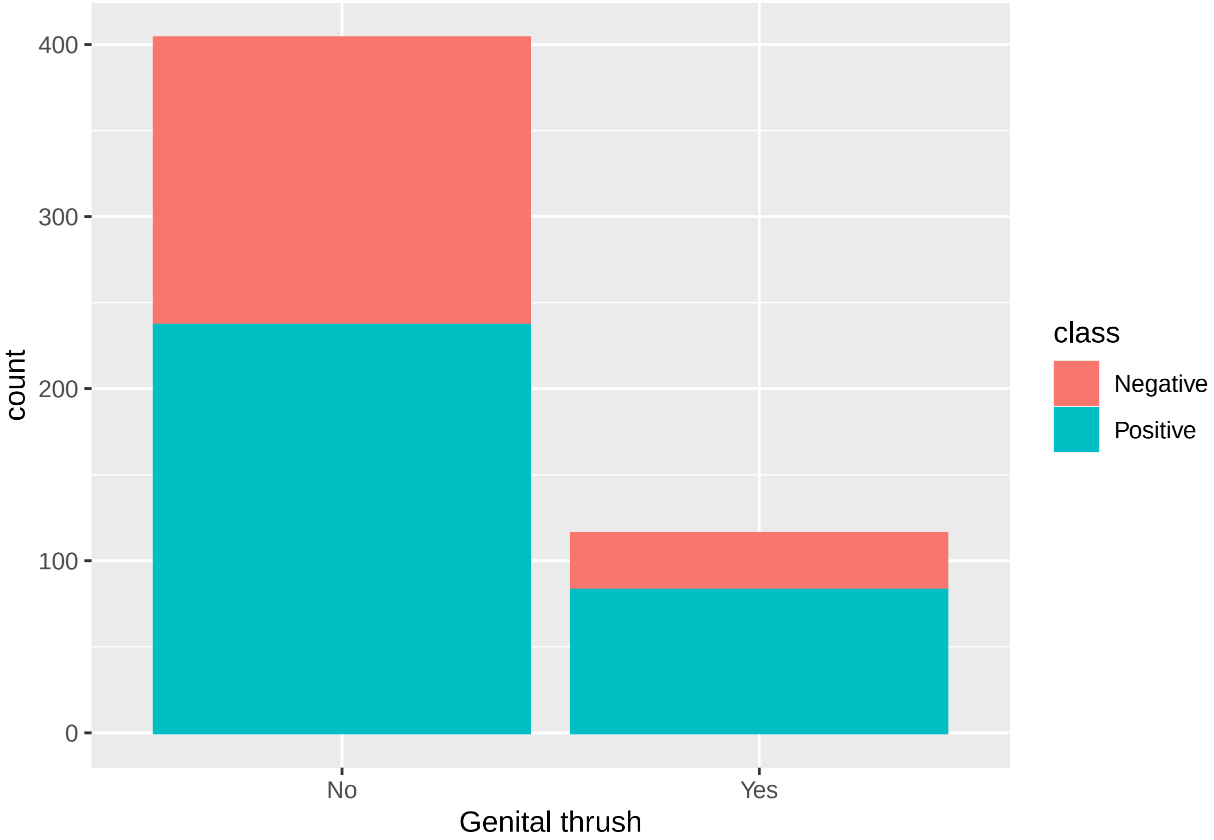
Below is a comparison between polyphagia and if one has diabetes.



There is a relatively strong correlation between having diabetes and polyphagia. However, it is equally likely to have diabetes whether or not one has polyphagia. Therefore, this is a somewhat valuable attribute.

2.3.7 Cases by genital thrush

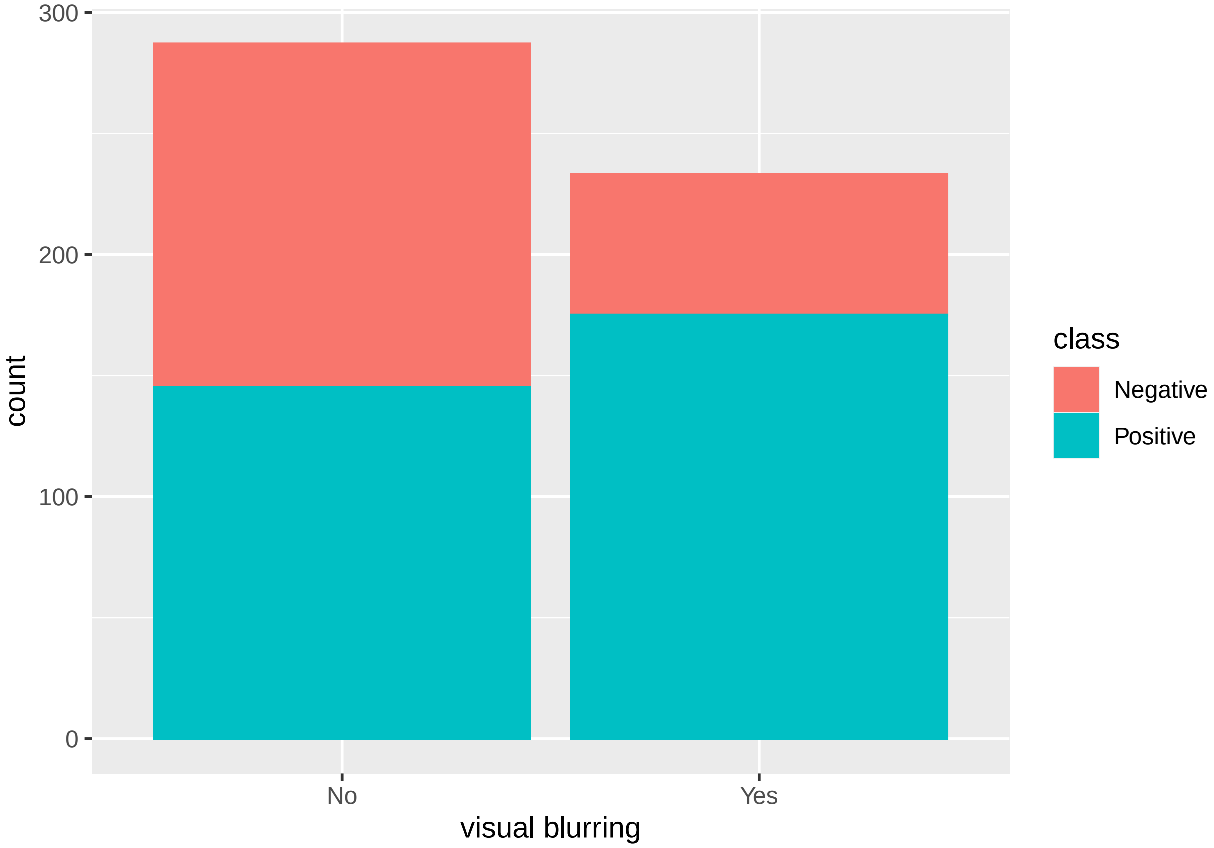
Below is a comparison between genital thrush and if one has diabetes.



Just like with polyphagia, there is a relatively strong correlation between having diabetes and genital thrush although there is a higher proportion of diabetic patients without genital thrush than non diabetic patients. This also may not be valuable as an attribute.

2.3.8 Cases by visual blurring

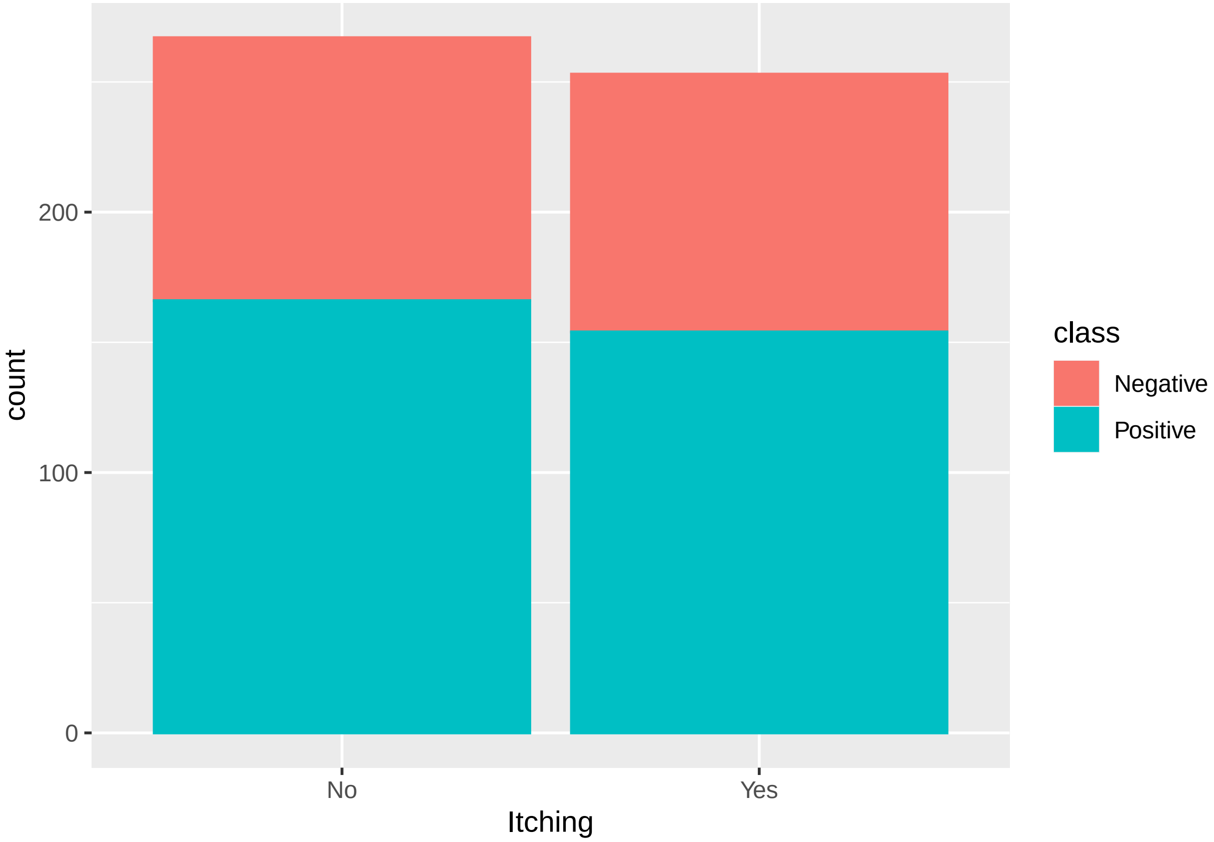
Below is a comparison between visual blurring and if one has diabetes.



Usually, diabetic patients have visual blurring. However, there is an equal proportion between diabetis patients and non diabetic patients relative to visual blurring.

2.3.9 Cases by itching

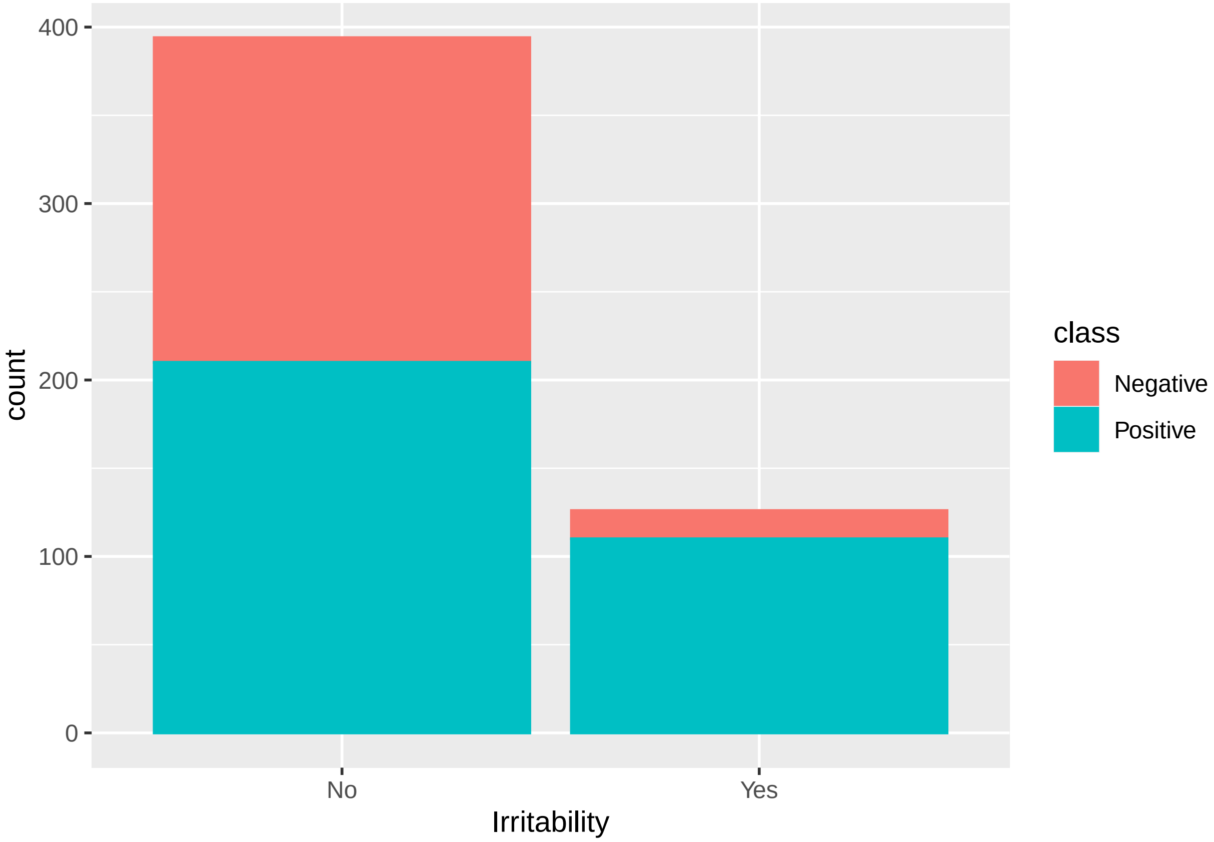
Below is a comparison between itching and if one has diabetes.



Itching does not seem to be a valuable attribute since there are relatively equal proportions of diabetic patients who are itching and not itching. Since Bangladesh is a tropical, hot climate, itching can be due to other factors such as mosquitos or natural skin irritability.

2.3.10 Cases by irritability

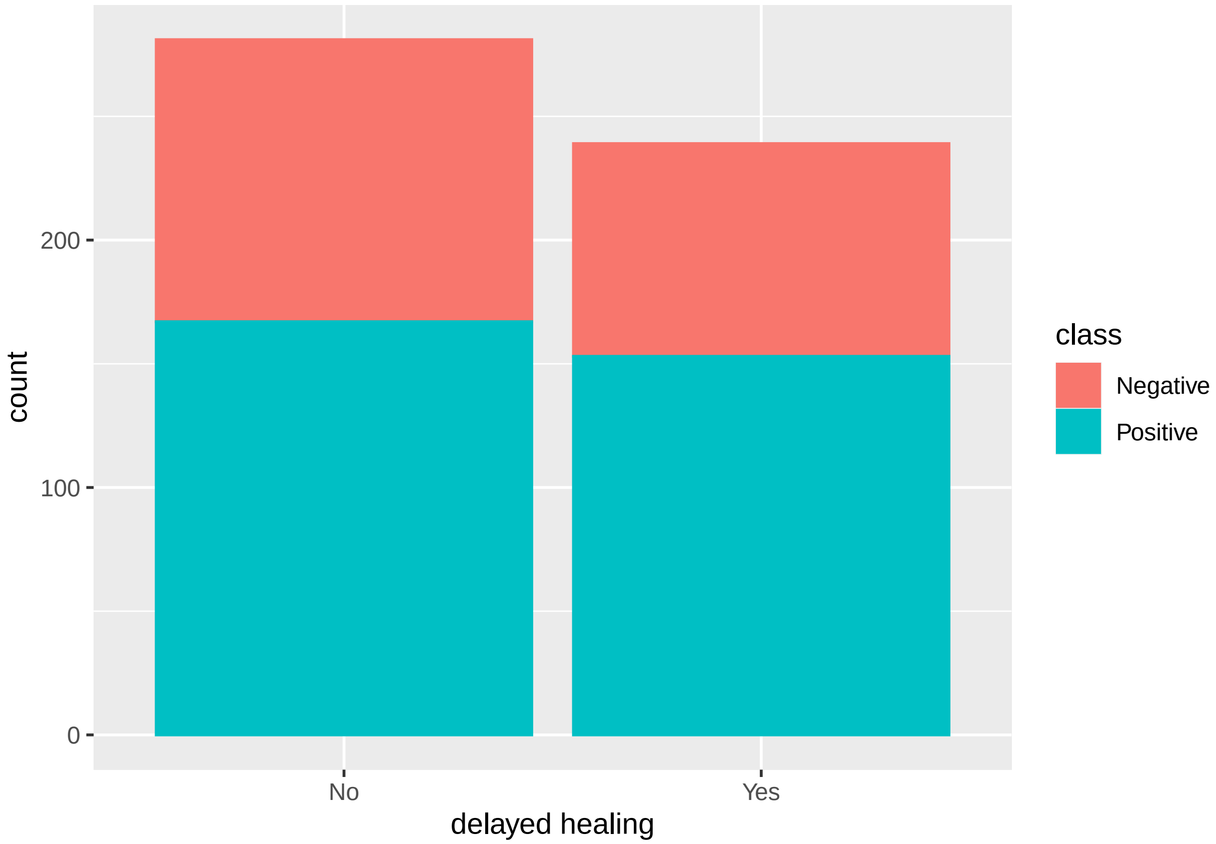
Below is a comparison between irritability and if one has diabetes.



Irritability is definitely a factor of diabetic patients. Judging by this, it seems to be a good attribute to determine if one is diabetic.

2.3.11 Cases by delayed healing

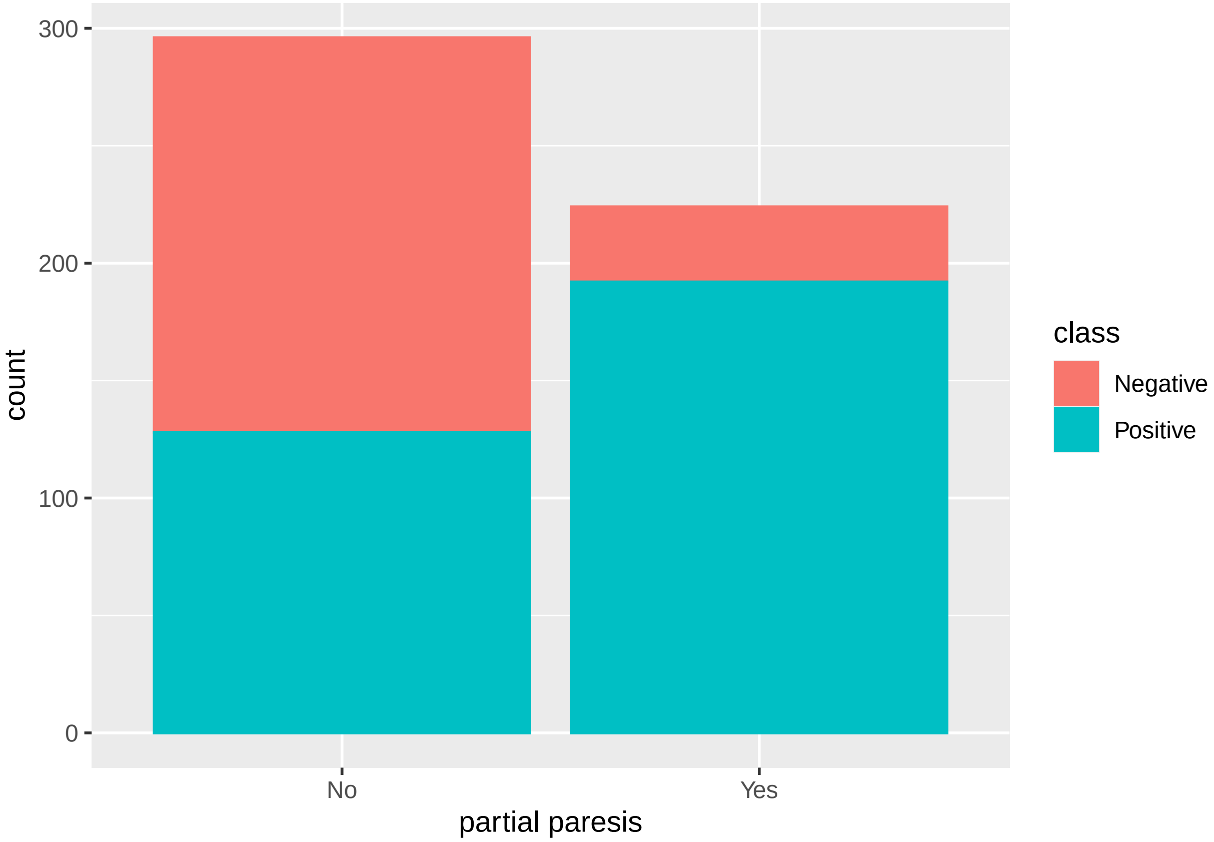
Below is a comparison between delayed healing and if one has diabetes.



Delayed healing does have, to an extent, equal proportions of diabetic patients across the options. Although diabetes does compromise the immune system and result in delayed healing, reasons for delayed healing in non diabetic patients could be immune system disorders or compromising diseases such as HIV or general vitamin deficiencies. This is not a strong attribute we can use.

2.3.12 Cases by partial paresis

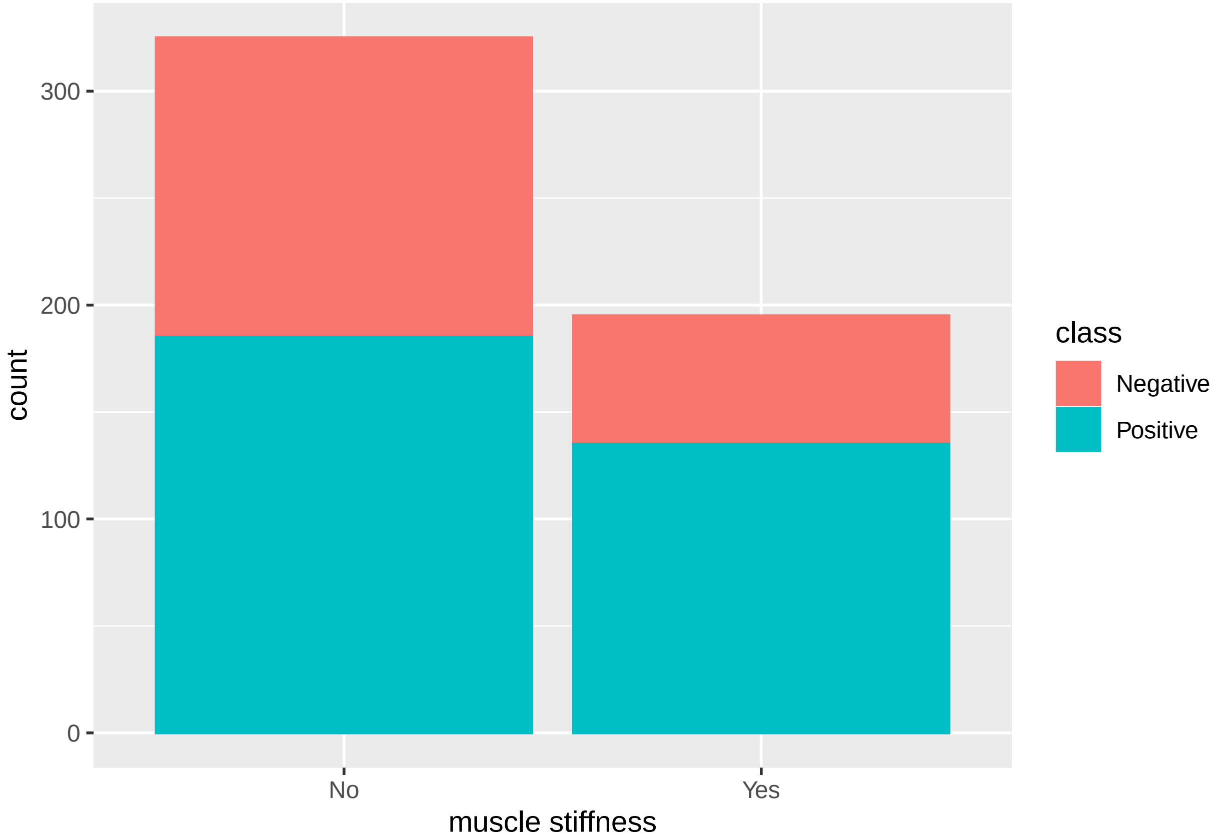
Below is a comparison between partial paresis and if one has diabetes.



As a high portion of patients with partial paresis have diabetes, partial paresis could be a valuable indicator for diabetes. However, its strength is limited by the proportion of diabetic patients who don't have partial paresis.

2.3.13 Cases by muscle stiffness

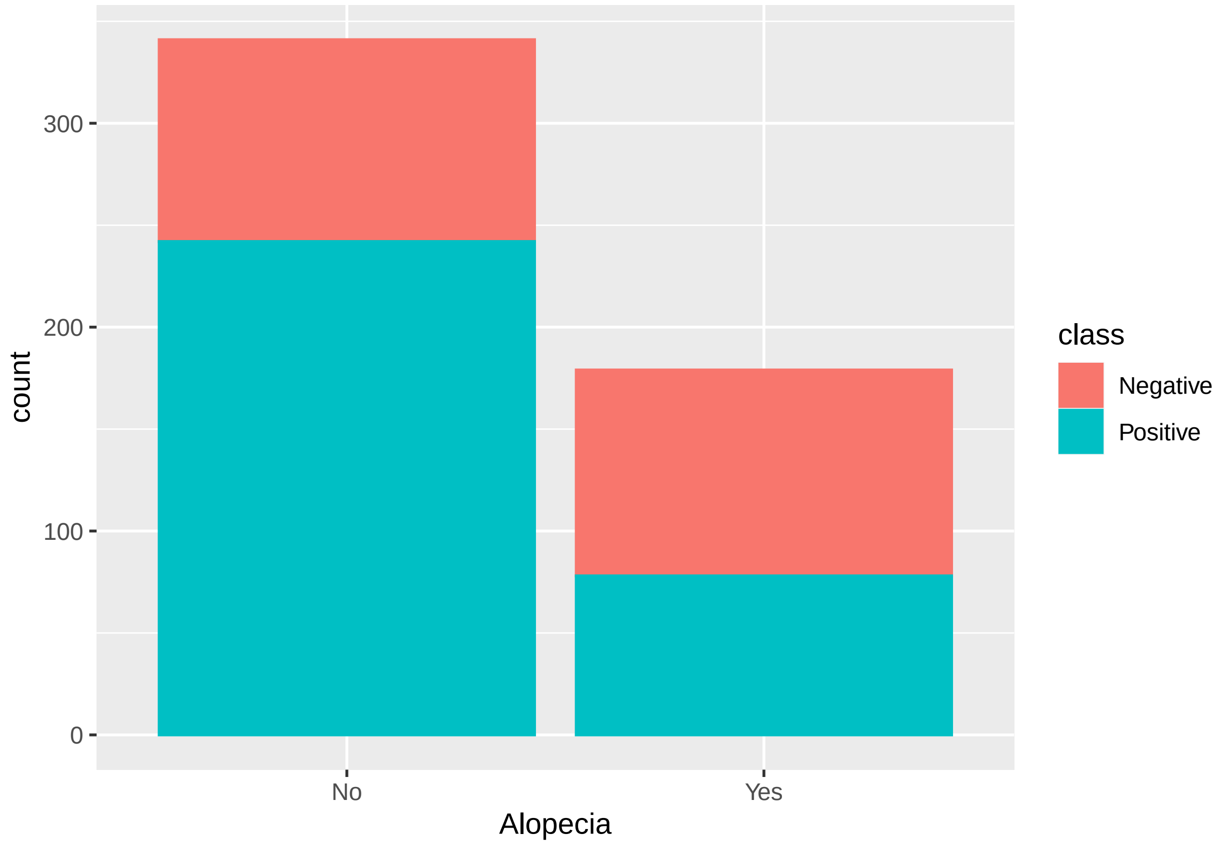
Below is a comparison between muscle stiffness and if one has diabetes.



Muscle stiffness does have a higher proportion of diabetic patients than without muscle stiffness. However, the number of diabetic patients between the two is similar. Therefore, this is not as strong of an attribute as the others.

2.3.14 Cases by Alopecia

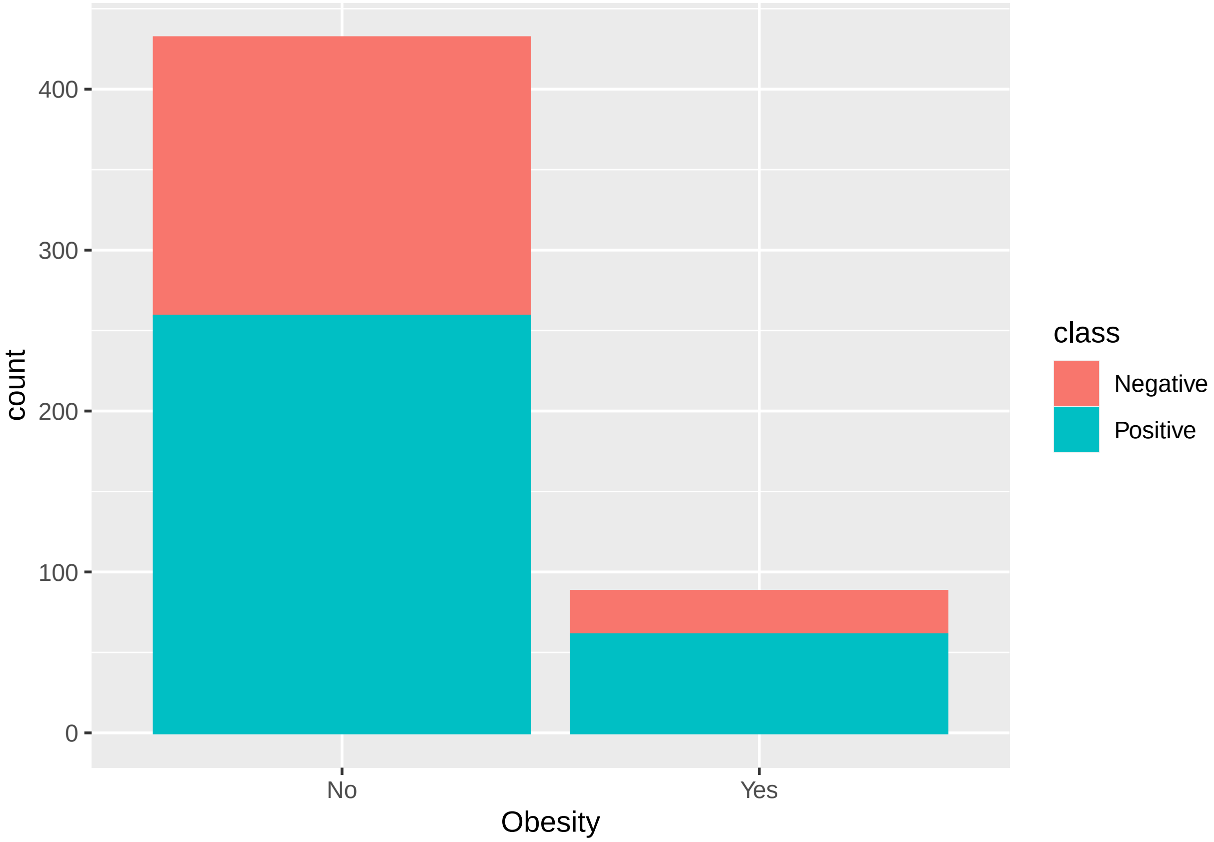
Below is a comparison between alopecia and if one has diabetes.



There is a higher proportion of diabetic patients without alopecia than with alopecia. Conversely, not having alopecia can be a symptom of diabetes according to this data. Therefore, we can use this as a decently strong attribute.

2.3.15 Cases by obesity

Below is a comparison between obesity and if one has diabetes.



As a high number of patients who are obese are diabetic as well, we can use this attribute. However, since non-obese patients can also be diabetic, the use of this attribute is slightly limited.

We can classify the attributes in the following manner:

Strong attributes:
- gender
- polyuria
- polydypsia
- sudden weight loss
- irritability
- alopesia

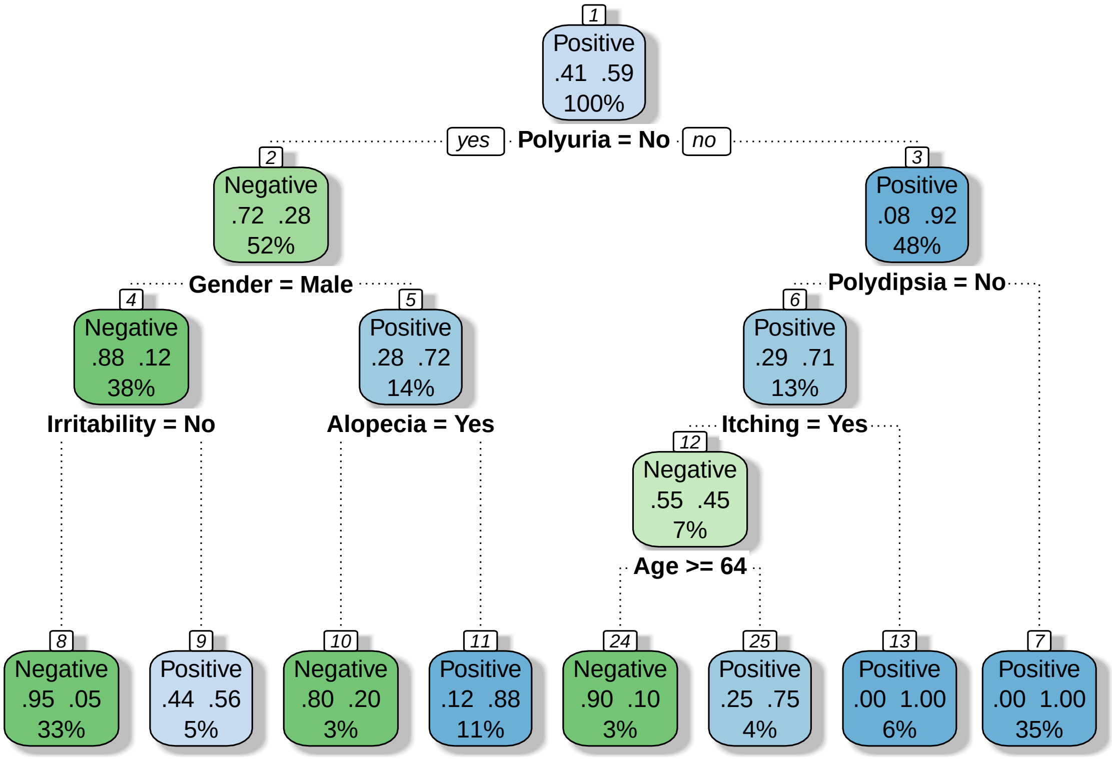
medium strength attributes:
- polyphagia
- partial paresis
- muscle stiffness
- obesity

weak attributes:
- weakness
- genital thrush
- visual blurring
- itching
- delayed healing

We can expect to see the strong attributes (perhaps utilized with the medium attributes) within our algorithm. We should not expect to see the weak attributes based of this hypothetical data analysis.

2.4 Insights

Lastly, I created a regression tree using the random forest algorithm:



decision tree for the data

From the information, we can observe that the main attributes are Polyuria, age, gender, alopecia, irritability, itching and polydipsia.(As seen beforehand with our graph of gender vs cases, we can see that gender does play a substantial role). As an insight, this is beneficial to us since we can use this tree to classify individual cases as either positive or negative. However, this is a product of machine learning and is therefore imperfect. We must also account for bias. As this data is from one region of the world (Bangladesh), we cannot use it to safely estimate whether someone from another nation is diabetic, albeit this tree is an excellent foundation. Compared with previous data analysis, we did see some attributes classified as “strong” show up such as gender, polyuria, polydipsia, alopecia, and irritability. In fact, all of our strong attributes showed up except for sudden weight loss. However, it is suprosing that itching showed up as an attribute on our tree since it was classified as a weak attribute. Perhaps it was combined with another string attribute, such as age, within the decision tree generation.

Relative to the dataset, a limitation is that the data mainly consists of factors. These are mostly boolean values such as Yes/No or Male/Female. Although it is possible to create a Least Squares estimate with this data by coercing the values into 0s and 1s, the final values will likely end up becoming decimals, which does not make sense in the real world (one person cannot be 0.876 positive for diabetes, for example) and would be more of a hindrance than a help. Instead, I can use modelling algorithms since my dataset is not large and my computer can support such computational operations.

I used several modelling algorithms within my algorithm. I used logistic regression, LDA, QDA, Naive Bayes, k nearest neighbors (knn), and random forest. Notably, random forest, logistic regression, and QDA gave me the highest individual accuracies out of any other method while knn and LDA gave me the lowest. A possible method to increase my accuracy could be to neglect the models which give me low accuracy. However, this leads to the fallacy of incomplete evidence (colloquially known as “cherry-picking”) which is a violation of scientific research and leads to pseudoscience. Therefore, I will not neglect insufficient models.

3 Results

3.1 Model results

Below is a table of general accuracies according to each method with the ensemble included:

```
##         GLM       LDA       QDA Naive_Bayes       knn Random_Forest  Ensemble
## 1 0.9417476 0.8980583 0.9757282   0.9320388 0.8737864     0.9854369 0.9563107
```

As seen, the overall accuracy of this ensemble is around 0.95. We have achieved our goal of achieving an accuracy over 90%. We could neglect the less accurate models. However, as aforementioned, this would be bias on our behalf and would violate scientific principles. Therefore, the less accurate models should be kept. With the inclusion of more quantitative data, other methods can be used such as a least squares estimate which could be a beneficial inclusion to the ensemble. Other algorithms which were not included in the Data Science course could be used to expand upon the ensemble such as nnet and earth.

3.2 Model performance

The model performance, while relatively high, does not achieve an extremely high level of accuracy. This can be attributed to multiple factors, either in the dataset or due to the algorithms. A limitation of the algorithms can be, for instance, that the Knn has an extremely low k value. This can lead to overtraining on a particular dataset which would not work as well on another section of the data.

A limitation of the dataset is that it has only a certain segment of a population and includes middle ages males and females without including many older and younger patients. Using these other patients for the algorithm can help produce less skewed results. Another limitation of the dataset is that is somehow contradicts previously observed medical phenomenae. As mentioned in the hypothetical exploratory data analysis, males are more likely than females to have diabetes due to testosterone imbalances. Somehow, more female patients have diabetes than male populations. This could be due to sampling bias or a biased population sample. However, the dataset could be expanded to account for a more accurate population sample.

4 Conclusion

4.1 Summary

In summary, this project explored the dataset on diabetic patients from the University of California, Irvine from patients from Sylhet Diabetes Hospital in Sylhet, Bangladesh, performing exploratory data analysis on each of the attributes to see if they indicate diabetes, used machine learning algorithms such as logistic regression, LDA, QDA, Naive Bayes, k nearest neighbors (knn), and random forest, tuned knn and random forest according to the k-value and mtry respectively, then concluded with an ensemble of the algorithms with the accuracy of the algorithms compared.

4.2 Potential impact

Since many people suffer from diabetes worldwide, a efficient testing system would be largely beneficial to many to easily test for diabetes. However, the accuracy should be improved and the algorithms should be tested on a more representative dataset before this algorithm is used en masse throughout the world.

4.3 Limitations

Limitations of this project are present both in the algorithms and the dataset. Some of the algorithms, such as knn, are guilty of overfitting and need to be rectified properly. The dataset is local to a single population in Bangladesh with a specific age bracket. To make the accuracy better, the data could be expanded to patients worldwide and could include an even distribution of ages and sex.

4.4 Future work

Future work could include multiple other algorithms such as linear regression and least squares estimates. However, for a least squares estimate to work, more quantitative data needs to be used to measure a tangible distance from the mean. Thus, the dataset could be expanded to include useful quantitative data such as daily or yearly sugar consumption, testosterone imbalance (for men) or testosterone surplus (for women). The main expansion of the dataset needs to include a more representative sample of ages so that the data is more uniform. Similarly, the data could include patients worldwide to create an algorithm which can be used worldwide. This does tie in to an excellent machine learning quote I heard: “data is to a machine learning algorithm as grapes are to wine. To make a good wine, you need good grapes. Similarly, to make a good machine learning algorithm, you need good data.” So expanding the data to get better data will inevitably improve our algorithm substantially.