HarvardX – Data Science Capstone: Prediction of Diabetes at Early Stage Capstone Project Report

Marisa Ivonne Zamora Carrillo

Contents

Overview

Nowadays health is a very important matter, we do not know when a global pandemic is going to occur and staying healthy, not having any disease or condition is crucial because the ones that do not have any of these are less likely to have complications. This is why I chose this dataset, to create a model that predicts the likelihood of having diabetes at early stage.

This dataset was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh.

Data Analysis

The data set used in this project is available in this website https://archive.ics.uci.edu/ml/machine-learning-databases/00529/diabetes_data_upload.csv

The 520-row dataset is divided into two datasets. The training and the validation dataset which is 10 percent of the data.

After select only the unique values it appears to be 269 duplicate values, since there is no patient ID , just the attributes, and the description of the dataset said that there were 520 we are going to assume that there are no duplicate values, just patients with the same characteristics.

This is a preview of the dataset

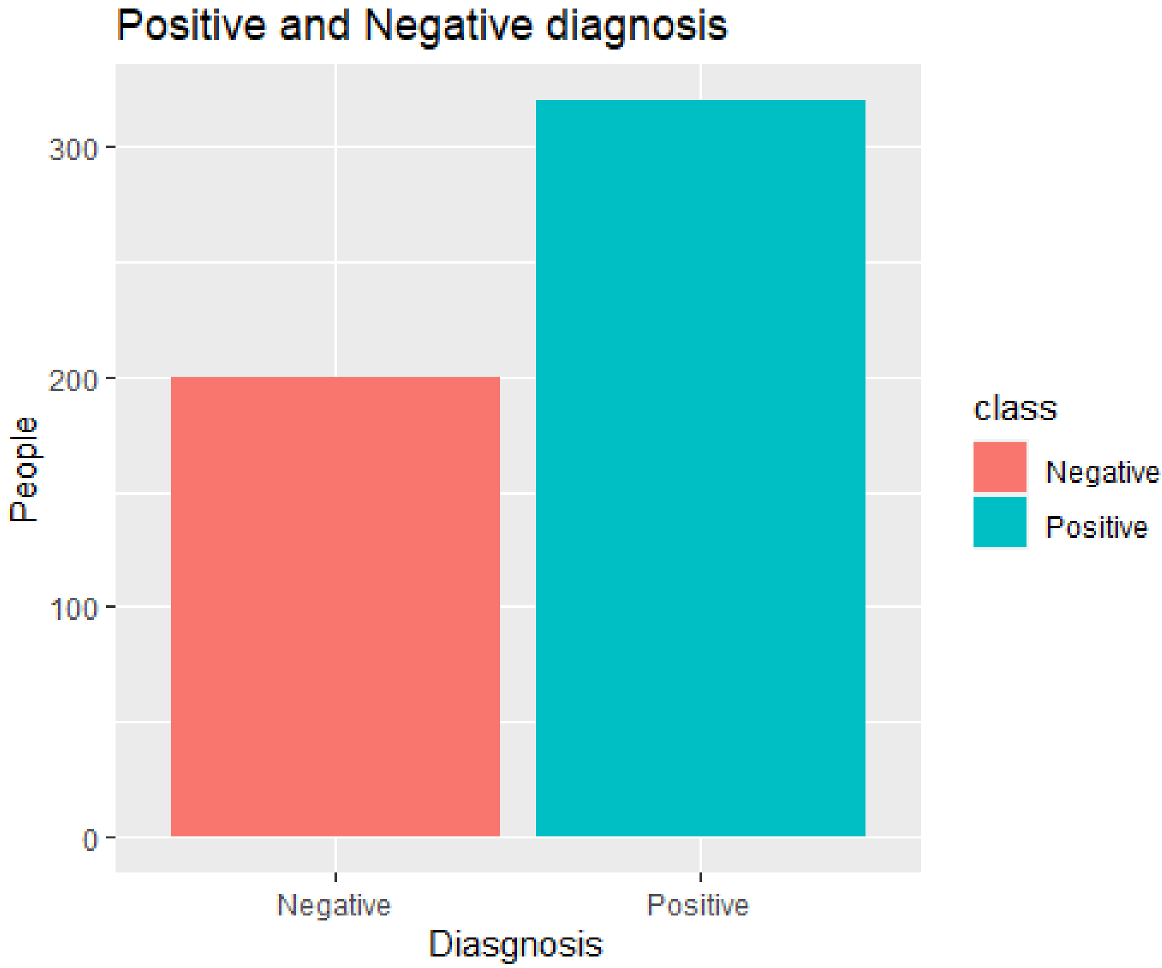
Age	Gender	Polyuria	Polydipsia	sudden.weight.loss	weakness	Polyphagia	Genital.thrush
40	Male	No	Yes	No	Yes	No	No
58	Male	No	No	No	Yes	No	No
41	Male	Yes	No	No	Yes	Yes	No
45	Male	No	No	Yes	Yes	Yes	Yes
60	Male	Yes	Yes	Yes	Yes	Yes	No
55	Male	Yes	Yes	No	Yes	Yes	No

visual.blurring	Itching	Irritability	delayed.healing	partial.paresis	muscle.stiffness	Alopecia	Obesity	class
No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
Yes	No	No	No	Yes	No	Yes	No	Positive
No	Yes	No	Yes	No	Yes	Yes	No	Positive
No	Yes	No	Yes	No	No	No	No	Positive
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
Yes	Yes	No	Yes	No	Yes	Yes	Yes	Positive

All the variables, except the age are character type values, because each of them describes a condition or symptom and is just marked as “Yes” of “No”, the class value is negative or positive which means if the patient has diabetes or no.

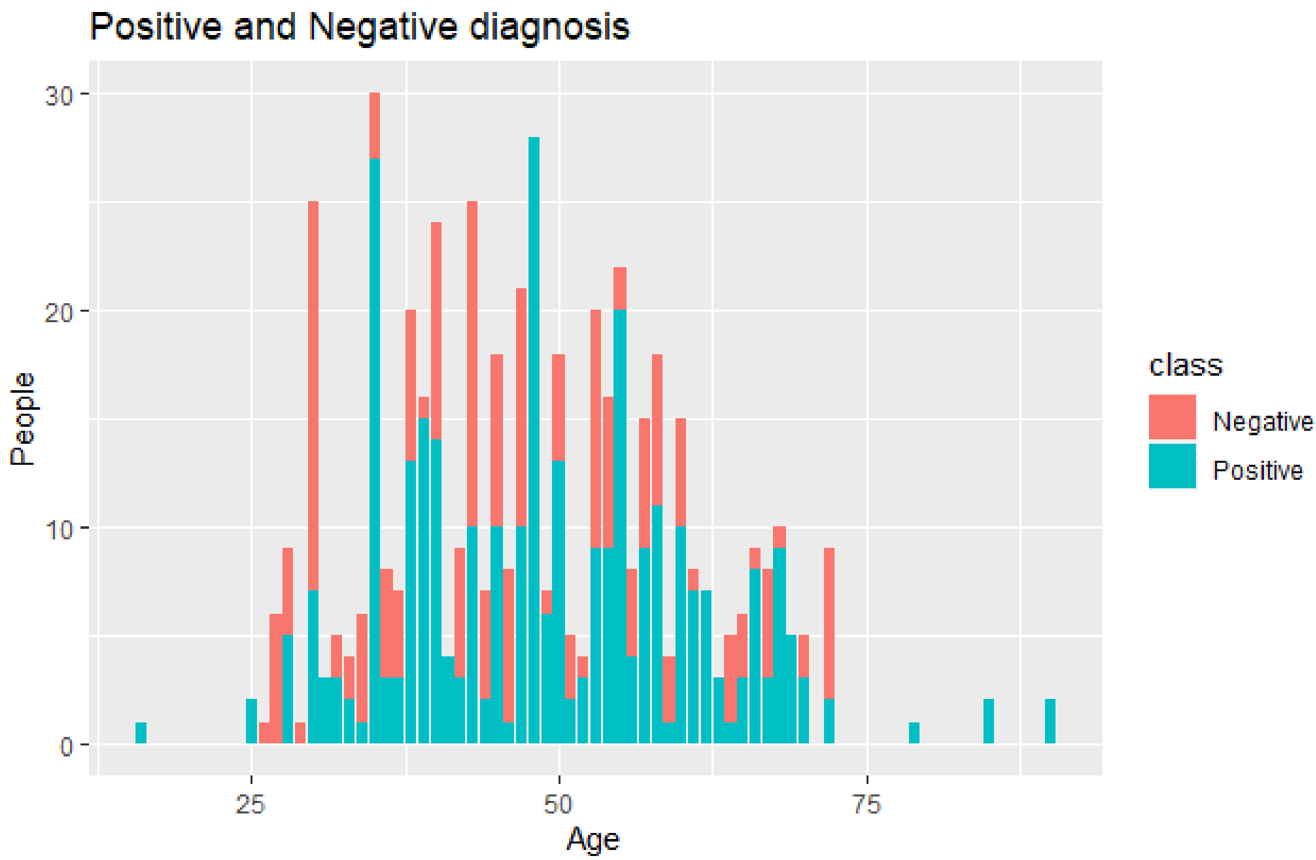
Exploratory Analysis

People with diabetes



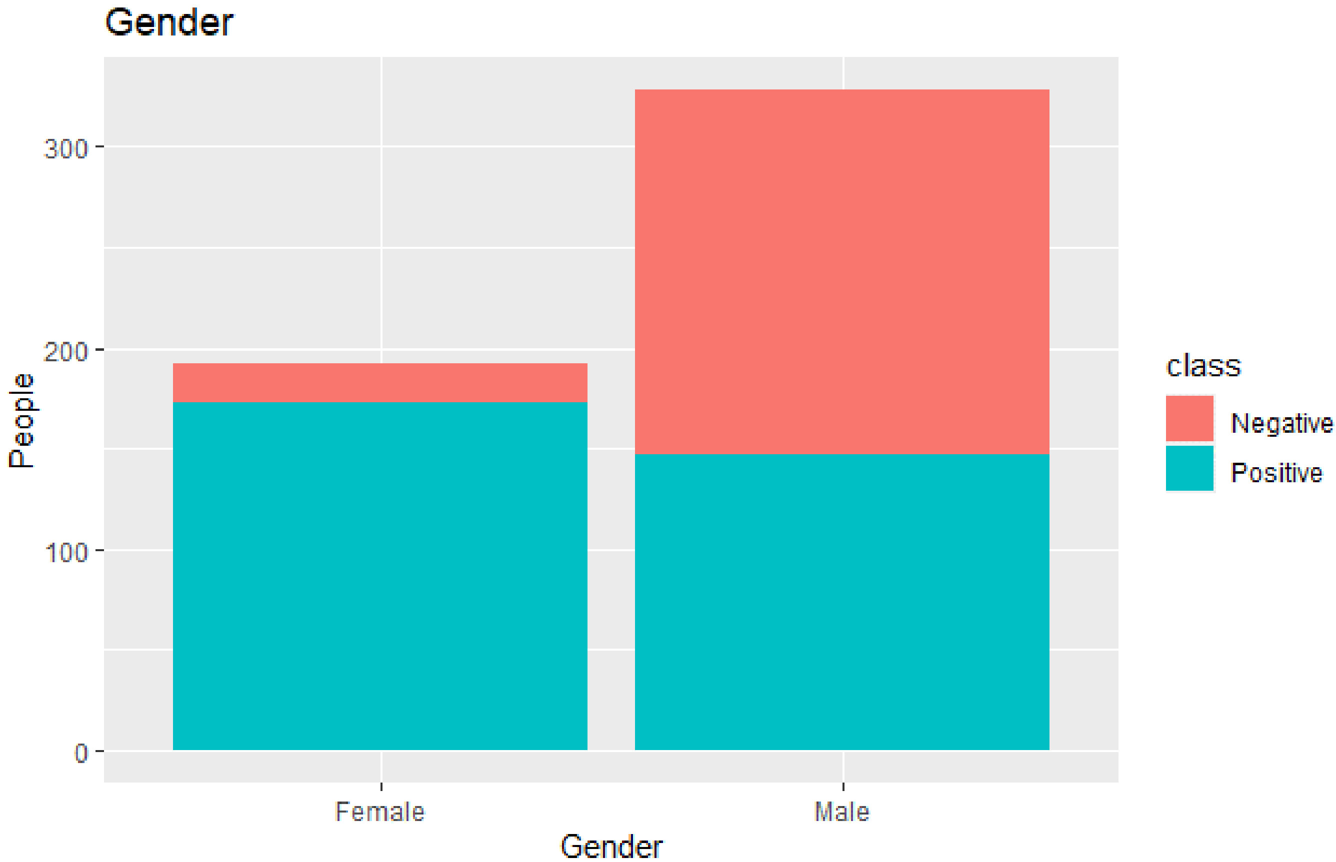
This graph shows that there are clearly more patients that participate in the questionaries that are positive on diabetes, but let’s get more insights

People grouped by age



It appears to be that most of the people are between 25 – 75 years old. There seems to be no pattern in the older the more positive cases.

People grouped by gender



This is very interesting plot because even tho there are more males in the dataset, more than the 50% of the are negative. Unlike the females wich only a very small percentage is negative.

Predictive Model building and evaluation

First we are going to create a data frame to keep all the results in there.

We are going to try tree different models, logistic regression, naive bayes and forest tree, compare their accuracy, sensitivity and specificity to choose the best one

Logistic Regression

After training the model

```
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were cost = 2, loss = L1 and epsilon = 0.001.
```

We predict the class using this model in the test dataset and generating this confusion matrix

```
          Reference
Prediction Negative Positive
  Negative       20        3
  Positive        0       29

               Accuracy : 0.9423
                 95% CI : (0.8405, 0.9879)
    No Information Rate : 0.6154
    P-Value [Acc > NIR] : 6.455e-08

                  Kappa : 0.8815

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 1.0000
            Specificity : 0.9062
         Pos Pred Value : 0.8696
         Neg Pred Value : 1.0000
             Prevalence : 0.3846
         Detection Rate : 0.3846
   Detection Prevalence : 0.4423
      Balanced Accuracy : 0.9531

       'Positive' Class : Negative
```

Achieving an accuracy of 0.9423, sensitivity of 1 and a specificity of 0.9062. Let's recall that the sensitivity tells us *the ability of an algorithm to predict a positive outcome when the actual outcome is positive* so here we have a perfect sensitivity but with the specificity we achieve a good value but we are going to see if another model is capable of improving *the ability of he algorithm to predict a negative when the outcome is negative*

Naive Bayes Model

After training, and predicting the class in the test dataset this is the confusion matrix

```
          Reference
Prediction Negative Positive
  Negative       20        5
  Positive        0       27

               Accuracy : 0.9038
                 95% CI : (0.7897, 0.968)
    No Information Rate : 0.6154
    P-Value [Acc > NIR] : 3.203e-06

                  Kappa : 0.806

 Mcnemar's Test P-Value : 0.07364

            Sensitivity : 1.0000
            Specificity : 0.8438
         Pos Pred Value : 0.8000
         Neg Pred Value : 1.0000
             Prevalence : 0.3846
         Detection Rate : 0.3846
   Detection Prevalence : 0.4808
      Balanced Accuracy : 0.9219

       'Positive' Class : Negative
```

Comparing this to our previous model, all the values decrease except sensitivity. So far the best model is Logistic Regression

Random Forest

```
            Reference
Prediction Negative Positive
  Negative       20        0
  Positive        0       32

               Accuracy : 1
                 95% CI : (0.9315, 1)
    No Information Rate : 0.6154
    P-Value [Acc > NIR] : 1.085e-11

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.3846
         Detection Rate : 0.3846
   Detection Prevalence : 0.3846
      Balanced Accuracy : 1.0000

       'Positive' Class : Negative
```

Seeing the results it seems that the random forest model is perfect for this dataset. This is the best model!

Results

This is the final table and visualization with the accuracy, sensitivity and specificity results of each model

model	accuracy	sensitivity	specificity
Logistic Regression	0.9230769	1	0.87500
Naive Bayes	0.9038462	1	0.84375
Random Forest	1.0000000	1	1.00000

The best model is the Random Forest, Logistic regression the second best and Naïve Bayes the worst but, not with bad results.

Conclusion

This project was interesting since the dataset selection. Once I choose this one I have to analyze the data, there was not so much of wrangling to do because the data was somehow clean and I say somehow because I think is important to register the ID of every patient so we can be certain that there are no duplicate patients in the dataset, but only people with the same attributes and characteristics.

The results of the training models are clear and I think this dataset and the predictions are very valuable, specially today with everything that is going on but the next steps should be gathered more data like this to create a larger dataset, adding a patients ID and of course replicate the principles of the dataset and the project with other diseases that can be detected and prevented at an early stage.