A Gaussian Naive-Bayes Model Analysis on the correlation between Blood Pressure, getting Fever, having Diabetes and Vomiting in predicting whether an individual has an underlying disease or not

By: Jec S. Castaños

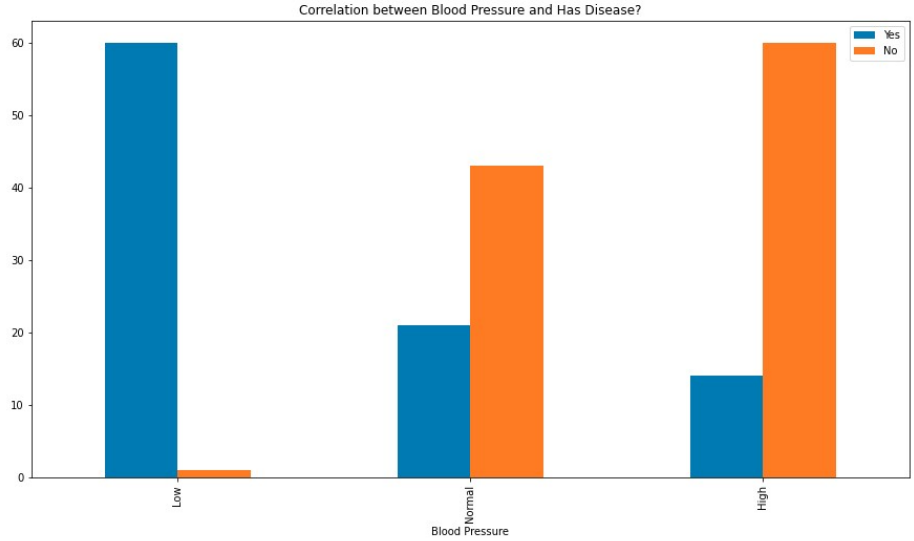
Introduction

Early diagnosis of the underlying disease is very essential in the medical field as early prevention in some diseases has been found helpful to the patients. Early detection is part of the initial examination that is done by physicians so getting a grasp on the situation is proven vital in providing the proper health care treatment.

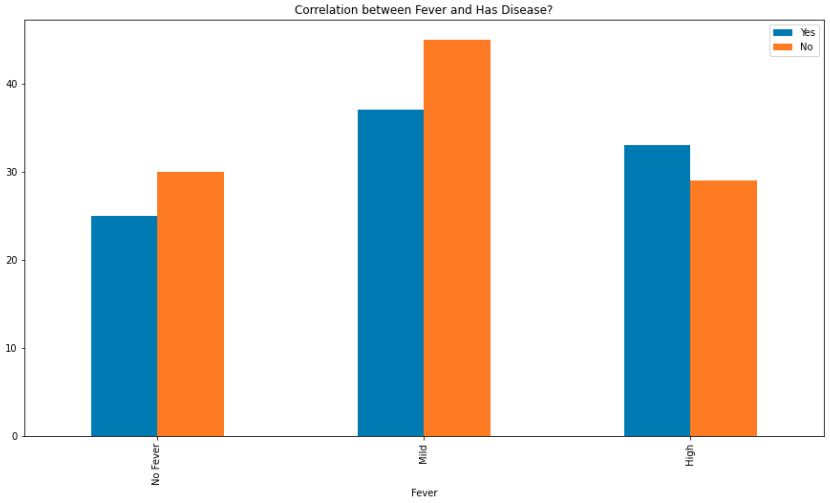
Since the early days of medical treatments, a face-to-face physical examination has been the initial step for treatment. But with the rise of technology and artificial intelligence, there are a lot of studies to target the early detection of diseases based on patient data. With this kind of technology, we can create a system to diagnose and analyze a patient's health from the comfort of their home with them providing the essential data.

Correlation Between Variables

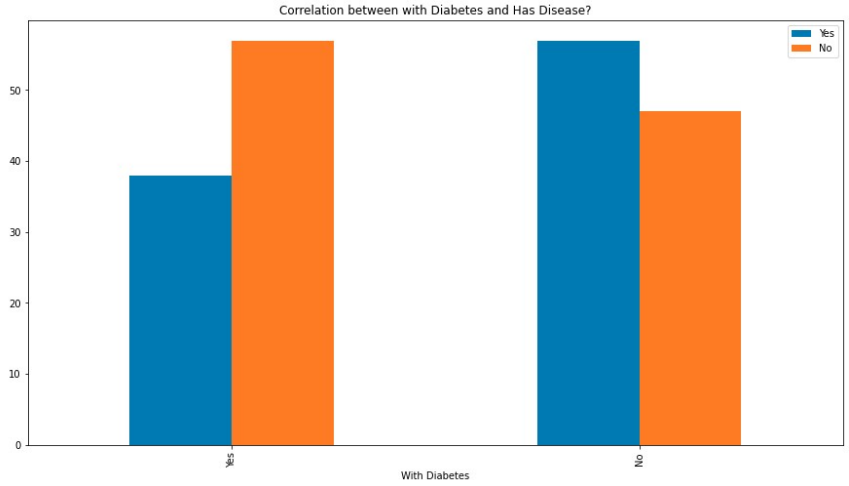
For us to know have an idea of what data we are dealing with and later help us build our Naive-Bayes Model let's visualize the correlation between our x data to our y data; where x is either blood pressure, fever, diabetes, and vomiting, and where y has a disease.

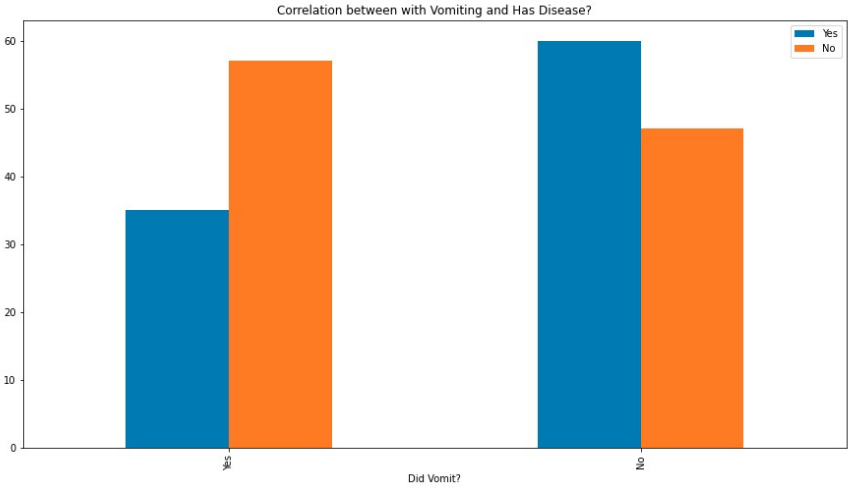


As we can see above the we can identify that a patient with Low Blood Pressure has the highest chance of having an underlying disease, on the other hand the patient with High Blood Pressure has seen to be having the lowest chance of having a disease but there's still a chance that the patient could have a disease.



With correlation above we can say that it is hard to determine whether the patient has a disease just by having a fever. This means that fever should be paired to other variables in order for us to determine the correct assessment to our patient.





For the last two graphs, the correlation between having diabetes or vomiting, respectively, and having a disease is hard to determine given there is a possibility that having diabetes and/or did vomit will not constitute to having a disease.

While the above correlations give us a rough idea of what output we can expect given the data. We can't determine that x data will result in y data. With this assessment, it's better for us to use all of the variables that we have and create a model that would help assess correctly.

Gaussian Naive-Bayes

Naive Bayes Classifier is a technique for constructing models that assign class labels to problem instances. Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood. And Gaussian Naive-Bayes is a type of Naive Bayes that follows Gaussian normal distribution and supports continuous data that is computed using the formula below.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

For this analysis we will be using a dataset with 200 rows that describes if patient has a diseases given the other variables discussed above.

	Blood Pressure	Fever	Diabetes	Vomit	Has Disease?
0	high	high	yes	no	no
1	high	high	yes	yes	no
2	low	mild	no	yes	yes
3	high	no fever	yes	no	no
4	low	no fever	no	no	yes
...	...	...	...	...	...
194	high	high	no	yes	yes
195	normal	mild	yes	yes	no
196	high	high	yes	no	no
197	high	high	yes	yes	no
198	normal	high	yes	yes	yes

With the use of sckit-learn, a free machine learning library for Python, we can build the model and assess its accuracy.

```
# Build the Model
model = GaussianNB()
model.fit(Features_train,Targets_train)

# Estimate Error
accuracy = model.score(Features_test,Targets_test)
print("The Gaussian Naive Bayes Model has ", str(round(accuracy * 100, 2)), "% in accuracy prediction")
```

```
The Gaussian Naive Bayes Model has  82.05 % in accuracy prediction
```

As shown above, we can see that our model has an 82.05% accuracy in predicting whether a patient has a disease or not.