CSI 551: Bayesian Data Analysis

Bayesian Data Analysis

Pima Indians Diabetes

Instructor: Kevin Knuth

Team: Jahnavi Bonagiri and Nidhi Vadnere

November 30, 2022

University at Albany, SUNY

Abstract

The Pima Indians are a group of North American Indians who traditionally lived in Arizona, US. They are one of the most studied groups for the causes and effects of diabetes. This paper will focus on analyzing a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases and applying the concepts of Naive Bayes Classifier to test various models and see which model has the best predictive ability to help diagnose populations with diabetes early on.

Section 1: Introduction

Pima Indians are a group of North American Indians who traditionally lived along the Gila and Salt rivers in Arizona, U.S. According to various recognized medical journals, the Pima Indian population is one of the highest at-risk populations to obtain diabetes compared to any other population in the world. The Pima Indians are one of the most studied groups for the causes and effects of diabetes. Additionally, diabetes is currently one of the major causes of death, with the number of affected people reaching up to 629 million by 2045 (Naz, Huma, and Sachin). The purpose of this project is to understand and recognize the various factors that can cause diabetes so these populations can better remedy their daily habits.

Section II: Methods

The Pima Indian Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information on 768 women from a population near Phoenix, Arizona, USA. Our objective is to use the diagnostic measurements from the dataset to predict if a patient has diabetes. To approach this problem, we will be using the method of the Bayes Naive Classifier and Gaussian model to test various models and see which model has the best predictive ability to help diagnose populations with diabetes early on and prevent future complications.

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

$$\begin{aligned} p(C_k \mid x_1, \dots, x_n) &\propto p(C_k, x_1, \dots, x_n) \\ &\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k) \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k), \end{aligned}$$

$$p(C_k \mid x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

$$Z = p(\mathbf{x}) = \sum_k p(C_k)\, p(\mathbf{x} \mid C_k)$$

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

Section III: Dataset

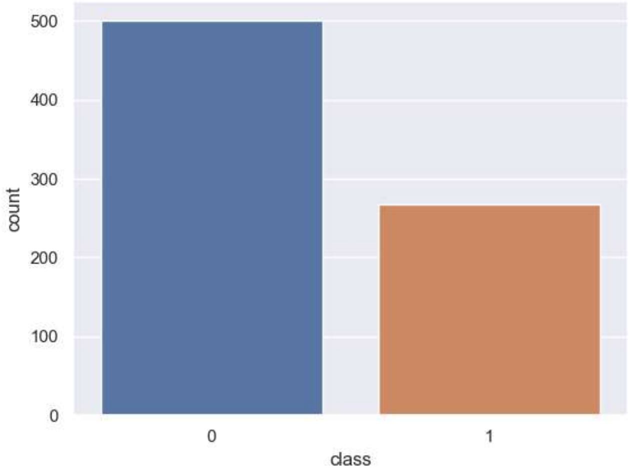
We got our dataset, *Pima Indians Diabetes*, from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset consists of 9 attributes as follows:

1. Pregnancies: Number of times pregnant
2. Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. BloodPressure: Diastolic blood pressure (mm Hg)
4. Skinthickness: Triceps skin fold thickness (mm)
5. insulin: 2-Hour serum insulin (mu U/ml)
6. bmi: Body mass index (weight in kg/(height in m)^2)
7. diabetespedigreefunction: Diabetes pedigree function
8. age: Age (years)
9. outcome: Class variable (0 or 1), 1 being the patient tested positive for diabetes (response variable)

There were 768 instances in this dataset. The constraints placed on selecting the data instances were that all patients were female, at least 21 years of age, and of Pima-Indian heritage. Our objective is to use the diagnostic measurements using 8 of the attributes above to predict if a patient has diabetes.

Section IV: Data Analysis

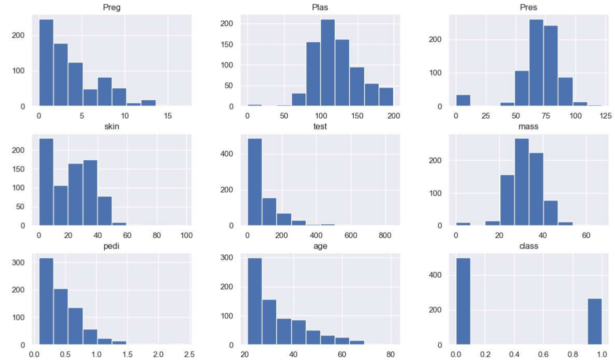
First, let us look at a general summary of the dataset



Here, we see the number of patients who were not diagnosed with diabetes (0) and the number of patients who were diagnosed with diabetes (1)

Test Negative for Diabetes	Test Positive for Diabetes
500	268

The attribute distribution is as follows:

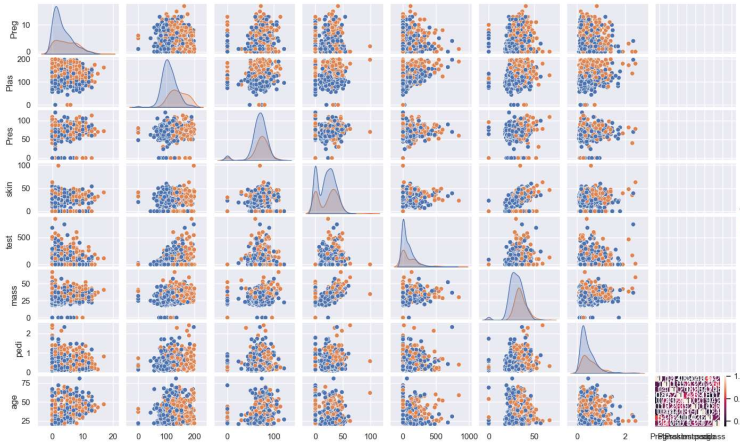


As we can see, blood pressure appears to have a bell curve, whereas age, insulin(test), pregnancies, diabetes pedigree function (pedi) are skewed to the right.

Analysis of the Histogram:

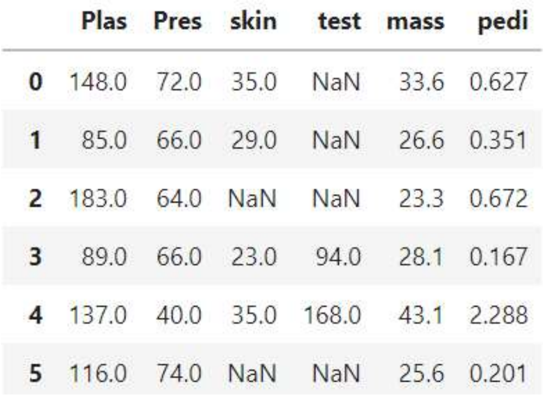
Bell shape curve: Blood Pressure
Right-Skewed: Age, Insulin, Pregnancies, Diabetes Pedigree Function

Short IQR: insulin, Diabetes Pedigree Function, Blood Pressure, and BMI

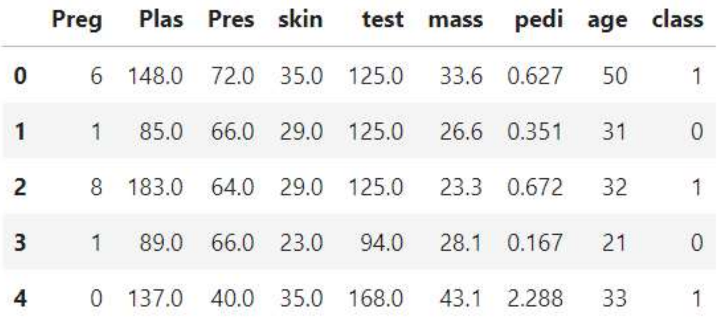


Section V: Model Building

To ensure that the data we were using is accurate, we first looked at the values within the dataset. We noticed that there were a lot of 0's in many of the rows, which could possibly be missing values. We replaced the 0's to NaN and made sure the new values were reflected in the dataset. Then, we replaced the NaN with the median values. We first went about the model building by creating a training and testing dataset. Once we completed this, we split the data into a 70:30 percentage and scaled the data.



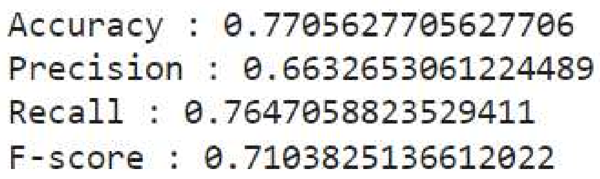
	Plas	Pres	skin	test	mass	pedi
0	148.0	72.0	35.0	NaN	33.6	0.627
1	85.0	66.0	29.0	NaN	26.6	0.351
2	183.0	64.0	NaN	NaN	23.3	0.672
3	89.0	66.0	23.0	94.0	28.1	0.167
4	137.0	40.0	35.0	168.0	43.1	2.288
5	116.0	74.0	NaN	NaN	25.6	0.201



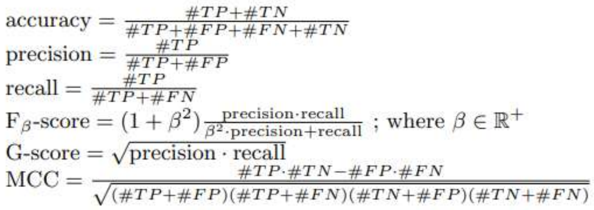
	Preg	Plas	Pres	skin	test	mass	pedi	age	class
0	6	148.0	72.0	35.0	125.0	33.6	0.627	50	1
1	1	85.0	66.0	29.0	125.0	26.6	0.351	31	0
2	8	183.0	64.0	29.0	125.0	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

Section VI: Model Building and Naive Bayes Model

After scaling the data, we were able to see the amount of oversampled data and model score. We used the Gaussian function to find the best fit model on the training data we gathered. The accuracy was around 77% as shown in the results. We also measured the performance of Naive Bayes by classification. The accuracies and precisions are shown below. Since the attributes are not completely independent the accuracy of the Naive Bayes model is not that much improved



```
Accuracy : 0.7705627705627706
Precision : 0.6632653061224489
Recall : 0.7647058823529411
F-score : 0.7103825136612022
```



$$\text{accuracy} = \frac{\#TP+\#TN}{\#TP+\#FP+\#FN+\#TN}$$

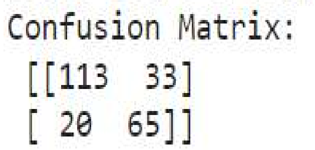
$$\text{precision} = \frac{\#TP}{\#TP+\#FP}$$

$$\text{recall} = \frac{\#TP}{\#TP+\#FN}$$

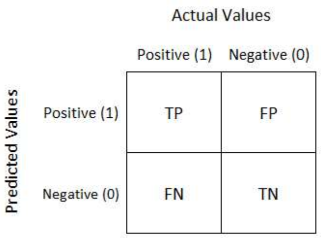
$$\text{F}_{\beta}\text{-score} = (1+\beta^2)\frac{\text{precision}\cdot\text{recall}}{\beta^2\cdot\text{precision}+\text{recall}} \text{ ; where } \beta \in \mathbb{R}^+$$

$$\text{G-score} = \sqrt{\text{precision}\cdot\text{recall}}$$

$$\text{MCC} = \frac{\#TP\cdot\#TN-\#FP\cdot\#FN}{\sqrt{(\#TP+\#FP)(\#TP+\#FN)(\#TN+\#FP)(\#TN+\#FN)}}$$



```
Confusion Matrix:
 [[113  33]
 [ 20  65]]
```



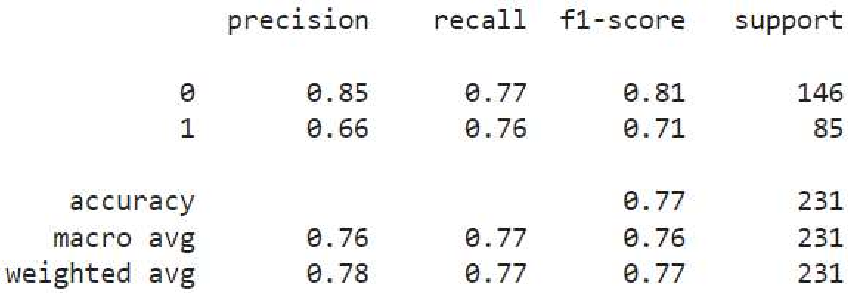
Actual Values

Predicted Values

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

The confusion matrix showed that the number of True Positives for diabetes was 113, false positives 33, false negatives 20, and true negatives 65. In terms of the accuracy, precision, recall, and f-score, the majority of the values ranged from 66% to 77% showing that the accuracy was a fairly good fit but not ideal.

Measure the performance of Naive Bayes by classification below:



```
              precision    recall  f1-score   support

           0       0.85      0.77      0.81       146
           1       0.66      0.76      0.71        85

    accuracy                           0.77       231
   macro avg       0.76      0.77      0.76       231
weighted avg       0.78      0.77      0.77       231
```

Section VII: Conclusion

This paper focused on finding prediction models for the risk of being diagnosed with diabetes. A large human population is diagnosed with diabetes, especially the Pima Indians group. If it remains untreated, it would cause a huge population of the world to be at severe risk leading to further complications.

Therefore, we have used the Naive Bayes classifier on the Pima dataset to experiment with the data. The experimental results show that, on the full Pima Indian Diabetes dataset, accuracy was (77.05%), precision (66.36%), and *f*-score (71.03%). Since the attributes are not completely independent, the accuracy of the Naive Bayes model is not the best.

We hope to continue working on this project and use models to perform the analysis. Due to the time constraint with this project, we were not able to use Bernoulli regression. Using more models could help improve the accuracy.

Section VIII: Works Cited

"..." ... - *YouTube*, 17 January 2019, https://diabetesjournals.org/diabetes/article/64/12/3993/34762/Dissecting-the-Etiology-of-Type-2-Diabetes-in-the. Accessed 2 December 2022.

"Deep learning approach for diabetes prediction using PIMA Indian dataset." *NCBI*, 14 April 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7270283/. Accessed 2 December 2022.

"Pima Indians Diabetes Database - dataset by data-society." *Data.world*, https://data.world/data-society/pima-indians-diabetes-database. Accessed 2 December 2022.

Rossi, Ryan A., and Nesreen K. Ahmed. "pima-indians-diabetes | Machine Learning Data." *Network Repository*, https://networkrepository.com/pima-indians-diabetes.php. Accessed 2 December 2022.