Bayesian Data Analysis: Pima Indians Diabetes

UNIVERSITY AT ALBANY
State University of New York

Nidhi Vadnere and Jahnavi Bonagiri
University at Albany, SUNY

ABSTRACT

Introduction and Purpose

The Pima Indians are a group of North American Indians who traditionally lived in present day 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 and apply 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.

METHODS

Our objective is to use the diagnostic measurements from the dataset to predict if a patient has diabetes.

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

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

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

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

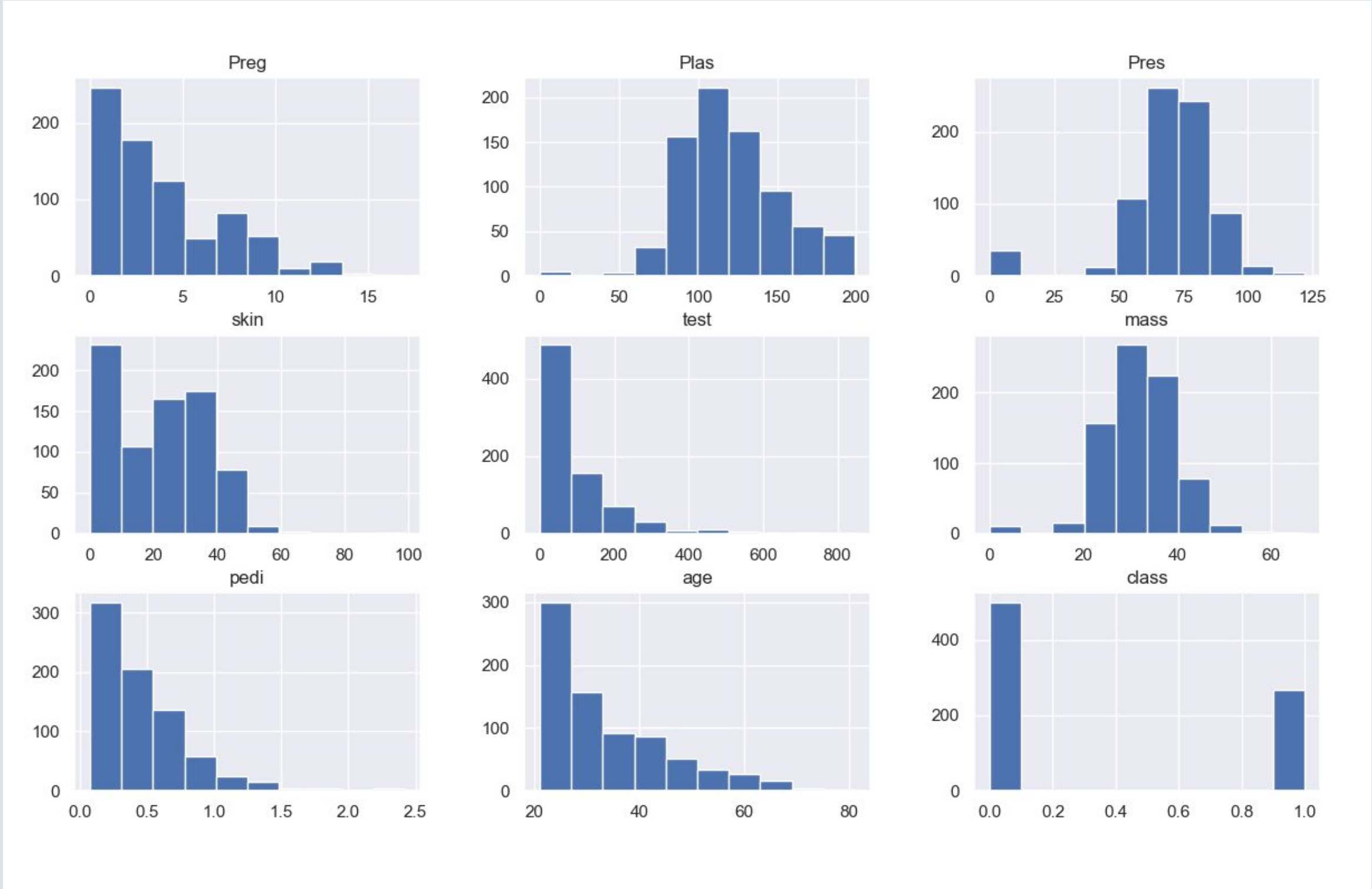
$$\hat{y} = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

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

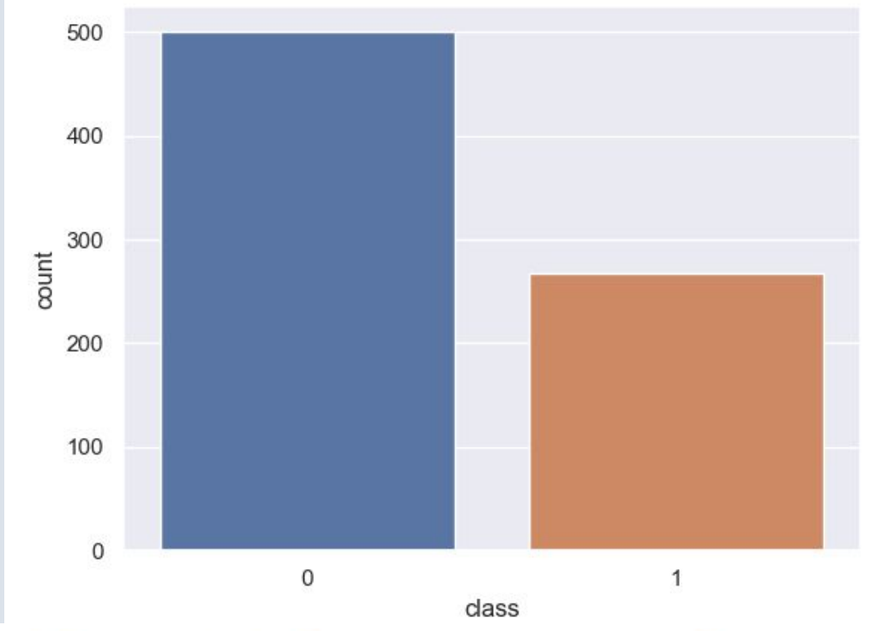
DATASET

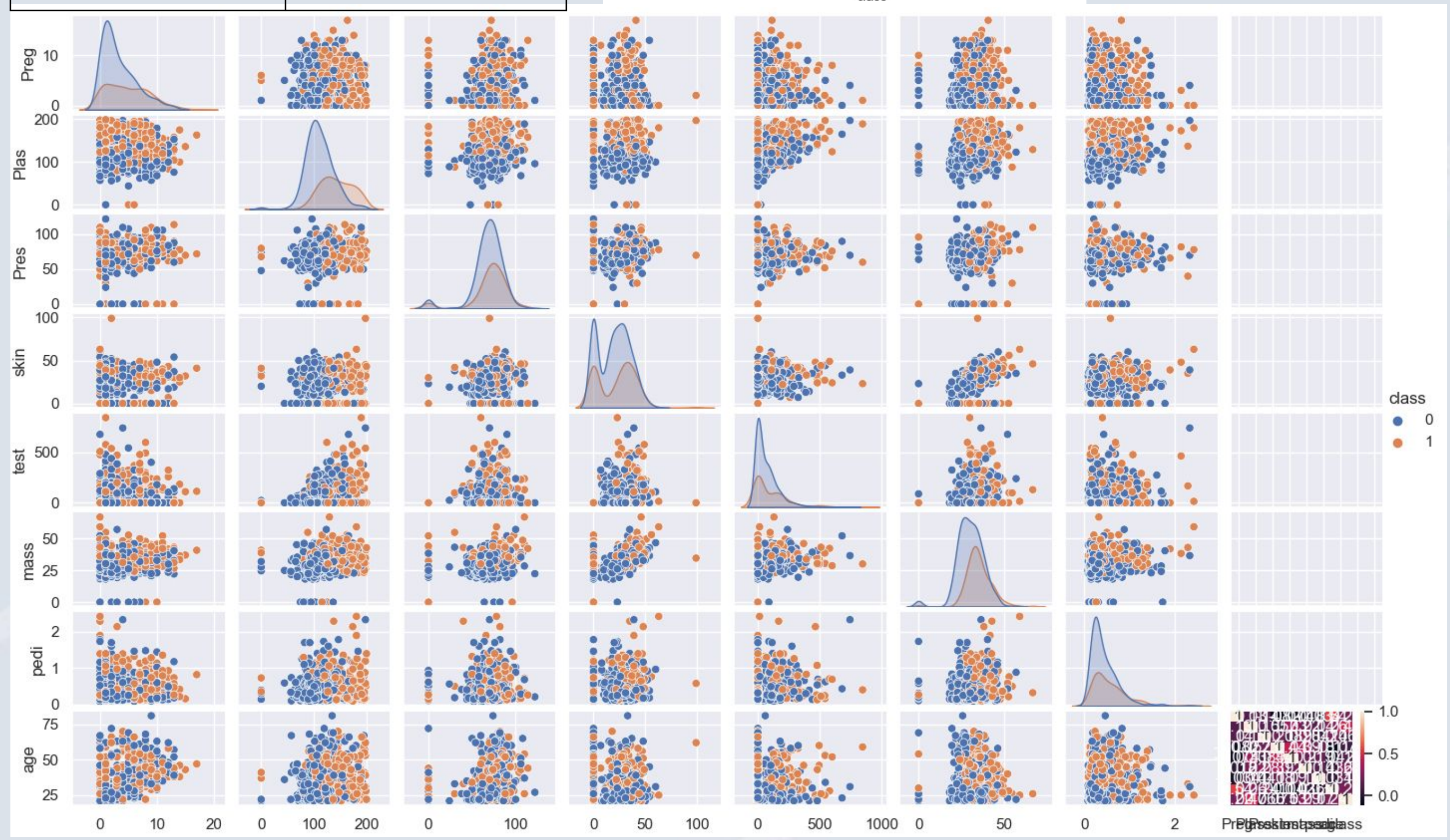
The Pima Indian Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information on 768 women from a population near Phoenix, Arizona, USA.

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)



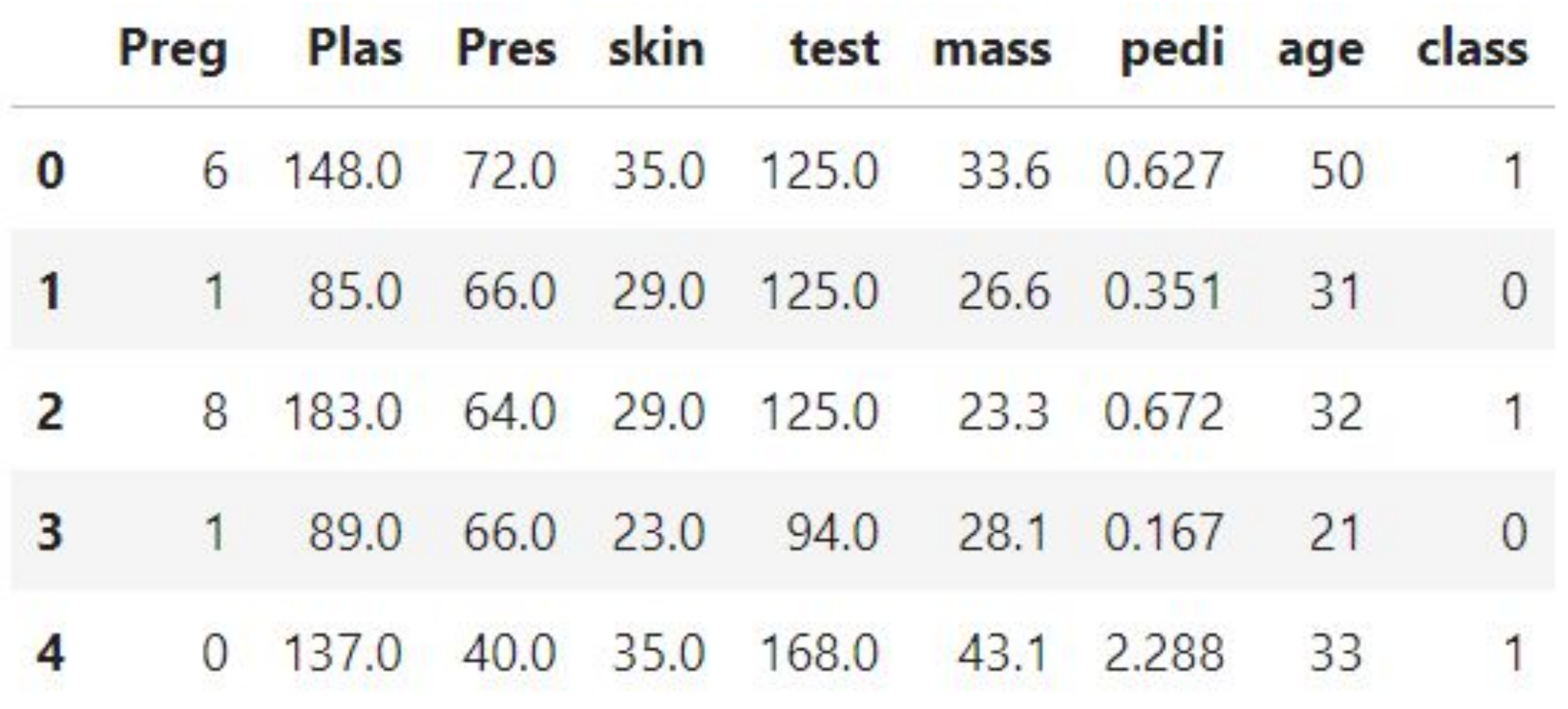
Test Negative for Diabetes	Test for Positive Diabetes
500	268





MODEL BUILDING

As we realized that the data may have many missing values, we converted the 0's into NaN. Then using the mean result to get more precise results. we created training and testing variables. Once we completed this, we split the data into a 70:30 percentage for training and testing respectively. After scaling the data, we were able to see the amount of oversampled data and model score.

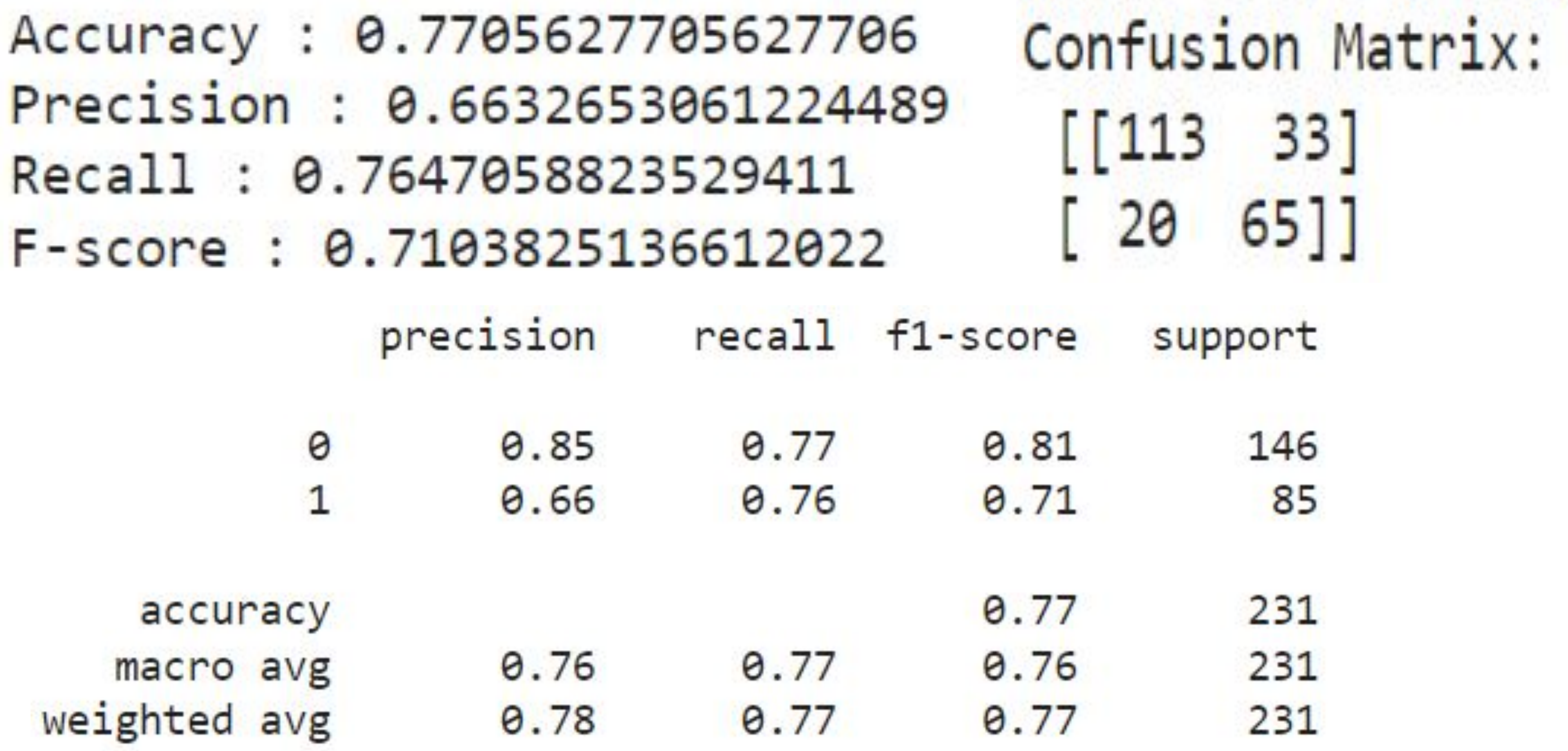


	Preg	Plas	Pres	skin	test	mass	pedi	age	class
0	6	148.0	72.0	35.0	125.0	33.6	0.627	50	1
1	1	85.0	66.0	29.0	125.0	26.6	0.351	31	0
2	8	183.0	64.0	29.0	125.0	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

RESULTS

Naive Bayes Model

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
```

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

```
              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
```

CONCLUSIONS

This paper focused on finding prediction models for the risk of being diagnosed with diabetes.

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 on. Due to the time constraint with this project, we were not able to use Logistic or Multinomial regression. Using more models could help improve the accuracy.

Works Cited:

https://data.world/data-society/pima-indians-diabetes-database

https://www.samyzaf.com/ML/pima/pima.html

Ewan R. Pearson; Dissecting the Etiology of Type 2 Diabetes in the Pima Indian Population. Diabetes 1 December 2015; 64 (12): 3993–3995. https://doi.org/10.2337/dbi15-0016

Naz, Huma, and Sachin Ahuja. "Deep learning approach for diabetes prediction using PIMA Indian dataset." Journal of diabetes and metabolic disorders vol. 19,1 391-403. 14 Apr. 2020, doi:10.1007/s40200-020-00520-5