Dylan Cozloff
AST03
CS 422 1001

# Diabetes Prediction NBC Model

**Installs:**
pip install pandas
pip install scikit-learn
pip install matplotlib
pip install seaborn

**Dataset:** https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

I used a diabetes prediction dataset (100,000 entries) which is a collection of medical and demographic data from patients, with their diabetes status. All features: age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. But, I only manually extracted four continuous features (age, BMI, HbA1c level, and blood glucose level) from this dataset. The binary target was whether the patient was positive or negative for diabetes. I loaded the dataset using the pandas library, in which I use the function read_csv().

**Training/Test Data Splits:** I used the sklearn.model_selection function train_test_split() to use a random 20% of the dataset for testing, and the remaining for training.

**Model Training:** I used the sklearn.naive_bayes functions GaussianNB(), alongside with fit() to fit the training data using a gaussian naive based classifier model.

**Model Predictions:** I used the sklearn.naive_bayes prediction() function to perform classification on both the training and test vectors. I also used predict_proba() to get the probability estimates for these vectors as well.

**Model Accuracy:** I used the sklearn.metrics function accuracy_score() to pass the training & test data with it's classification predictions to get a more precise model accuracy as the training and test accuracy were very close: **Training Accuracy** = 0.95685, **Test Accuracy** = 0.95585. As you can see the model generalized very well, accuracy dropped of by only 0.1%.

**NEXT PAGE**

Dylan Cozloff
AST03
CS 422 1001

## Classification Report:

**Test Data:**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.99 | 0.98 | 18292 |
| 1 | 0.85 | 0.59 | 0.69 | 1708 |
| accuracy | | | 0.96 | 20000 |
| macro avg | 0.91 | 0.79 | 0.84 | 20000 |
| weighted avg | 0.95 | 0.96 | 0.95 | 20000 |

Sensitivity (Test Data): 59%
Specificity (Test Data): 99%
Log Loss (Test Data): 0.13328660919175253

**Training Data:**

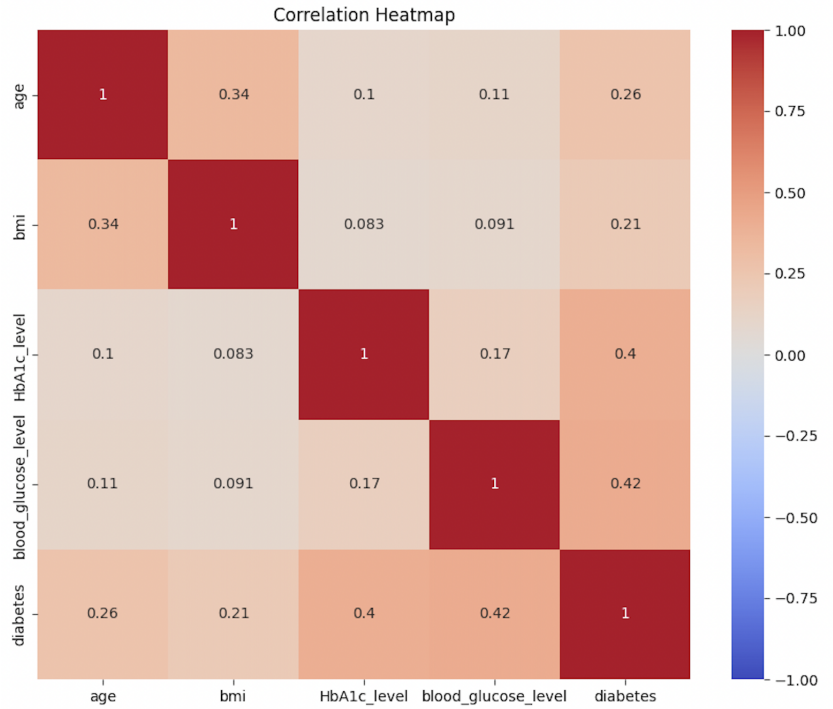
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.99 | 0.98 | 73208 |
| 1 | 0.85 | 0.60 | 0.70 | 6792 |
| accuracy | | | 0.96 | 80000 |
| macro avg | 0.91 | 0.79 | 0.84 | 80000 |
| weighted avg | 0.95 | 0.96 | 0.95 | 80000 |

Sensitivity (Training Data): 60%
Specificity (Training Data): 99%
Log Loss (Training Data): 0.12902673860634378

**NEXT PAGE**

Dylan Cozloff
AST03
CS 422 1001

**Correlation Heat Map:**




I used the corr() function to evaluate the dataset, and matplotlib.pyplot / seaborn to visualize the correlation matrix. As you can see there is definitely a correlation between diabetes and HbA1c_levels, blood glucose levels, age, and bmi.