Jenna Chan, Nidhi Shinde, Tammy Si

Team Cache Me Outside

Jun 3, 2025

**Predicting Diabetes Diagnosis from Health Indicators**

**Introduction**

This project seeks to identify trends among diabetes patients across the spectrum subject to varying socioeconomic factors and with differing underlying health conditions.

**Dataset Information**

The dataset was sourced from UC Irvine’s Machine Learning Repository and collected by the CDC. This data was collected for the purpose of better understanding the relationship between diabetes diagnoses and various lifestyle factors of people living in the US.

The target variable is diagnosis of diabetes, coded as healthy (0), prediabetes (1), and diabetes (2).

Overall, this dataset includes:

* **253,680 instances**
* **21 features**
* **Health indicators**, including but not limited to:
  + Presence of high blood pressure (0/1)
  + Presence of high cholesterol (0/1)
  + BMI
  + At least one stroke experienced (0/1)
  + Mental health score
  + Physical health score
  + Presence of heart disease / attack (0/1)
* **Lifestyle factors**, including but not limited to:
  + Smoker (0/1)
  + Heavy alcohol consumption (0/1)
  + Physical activity level
  + Consumes fruit at least once a day (0/1)
  + Consumes vegetables at least once a day (0/1)
* **Demographics**, including but not limited to:
  + Sex (0 = F, 1 = M)
  + Age
  + Educational Level
  + Income Level

**Preliminary Data Analysis**

The target variable was diabetes diagnosis.

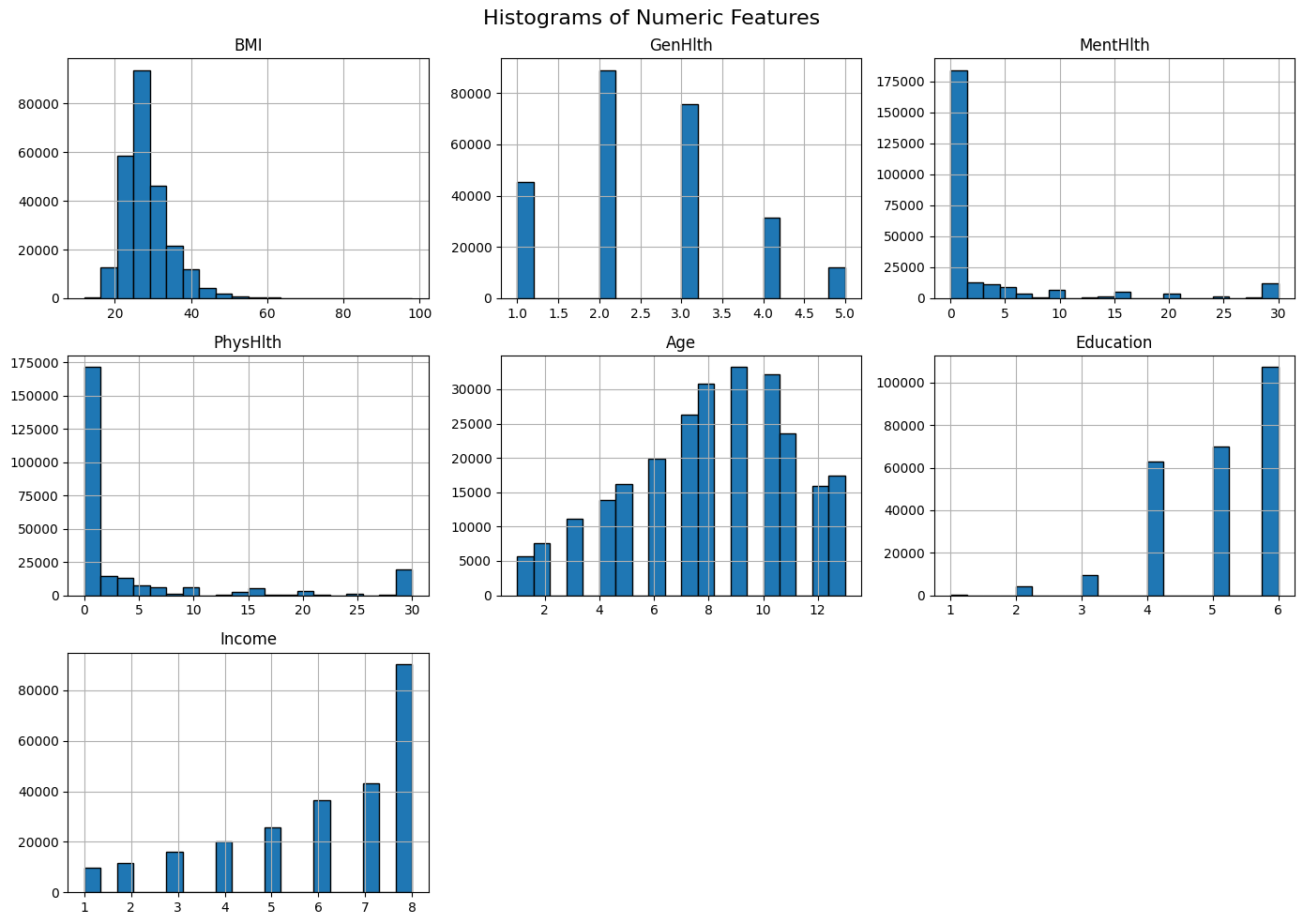
* 0 = Healthy
* 1 = Prediabetes
* 2 = Diabetes

The explanatory variables were:

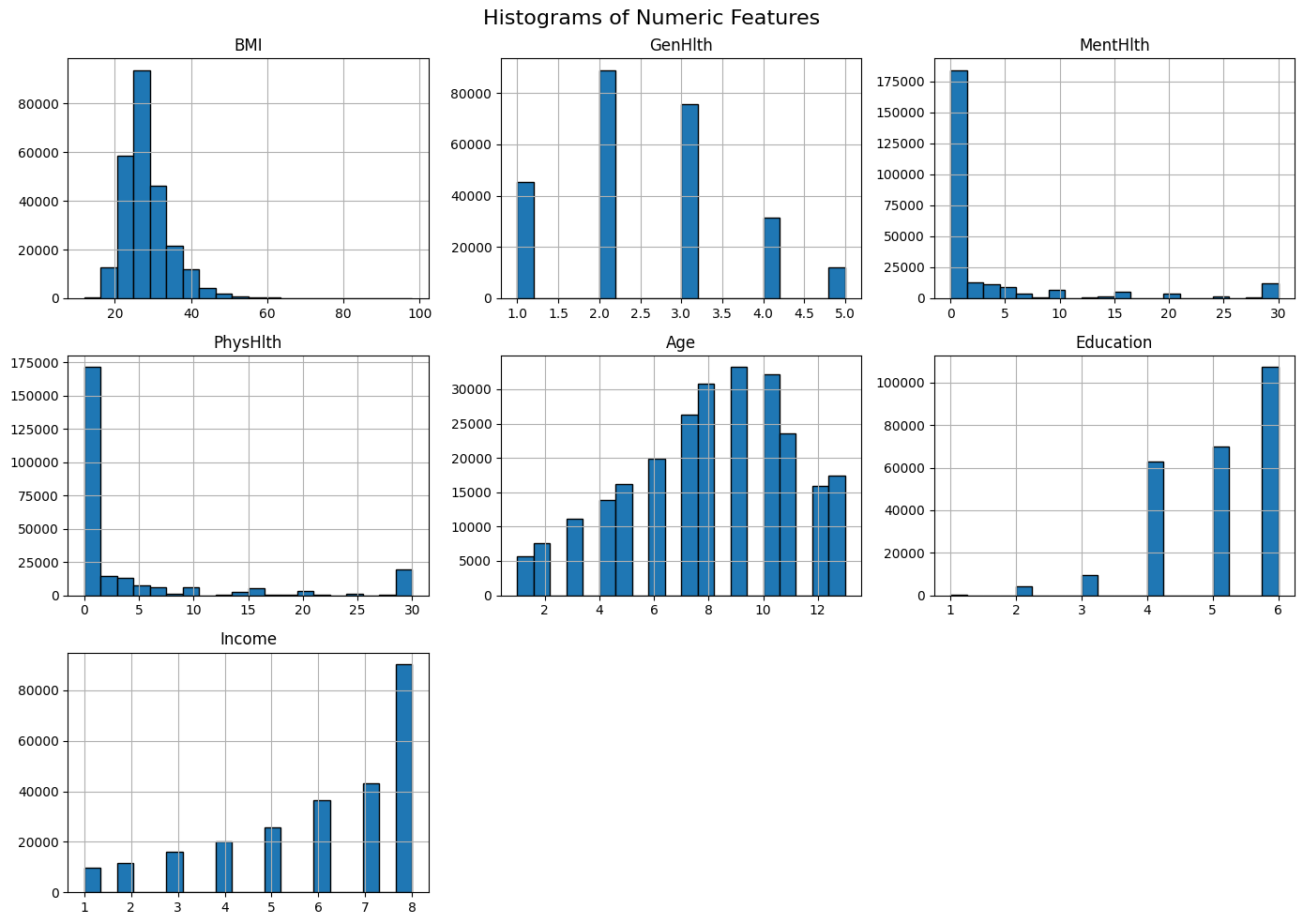
* Demographics
  + Sex
  + Age
  + Education level
  + Income
* Lifestyle Factors
  + Smoking status
  + Alcohol consumption
  + Physical activity level
  + Fruit intake
  + Vegetable intake
  + Access to healthcare
* Health Conditions
  + High blood pressure (HighBP)
  + High Cholesterol (HighChol)
  + Body Mass Index (BMI)
  + History of stroke
  + Mental Health (days of poor mental health per month)
  + History of heart disease or heart attack

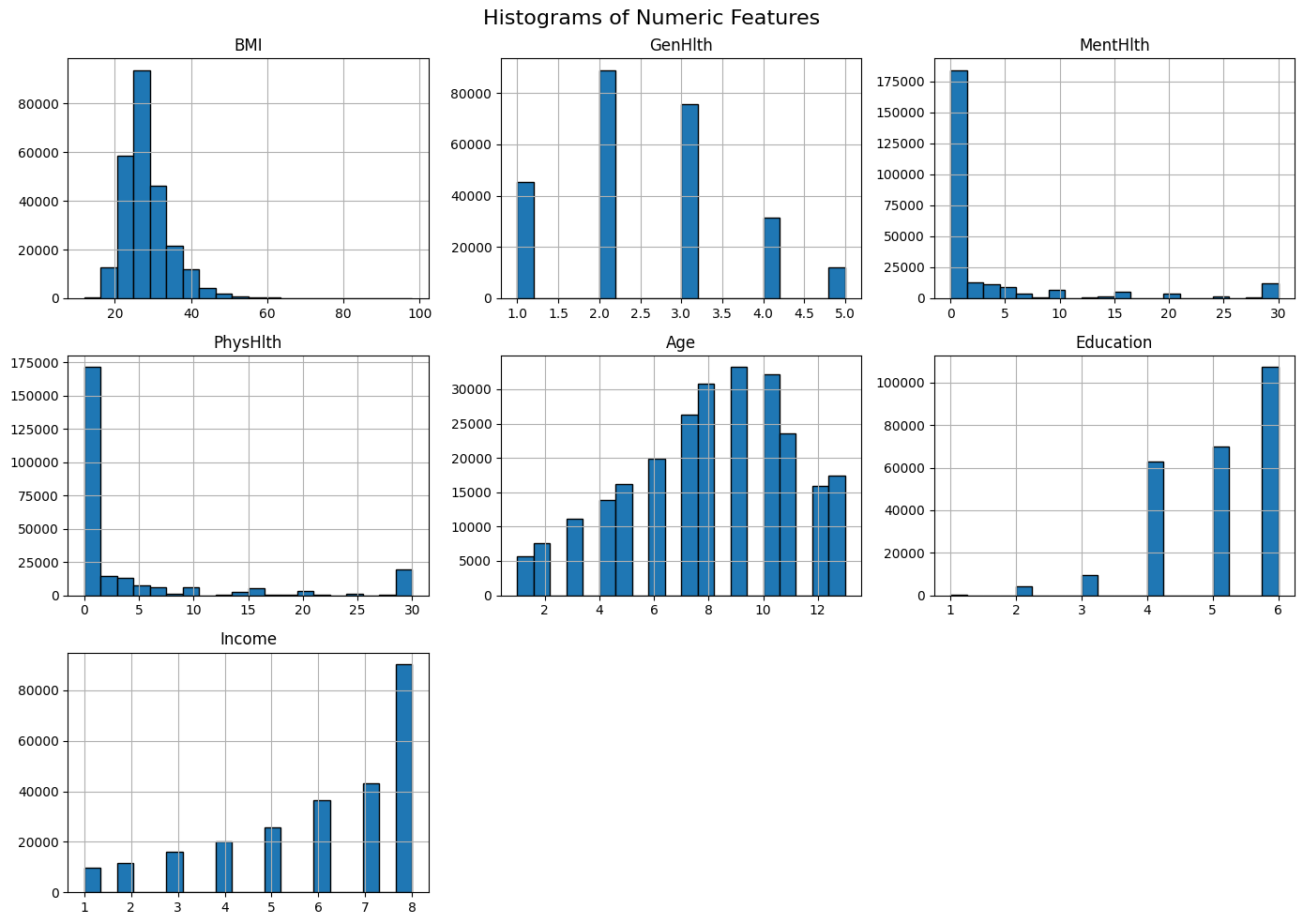
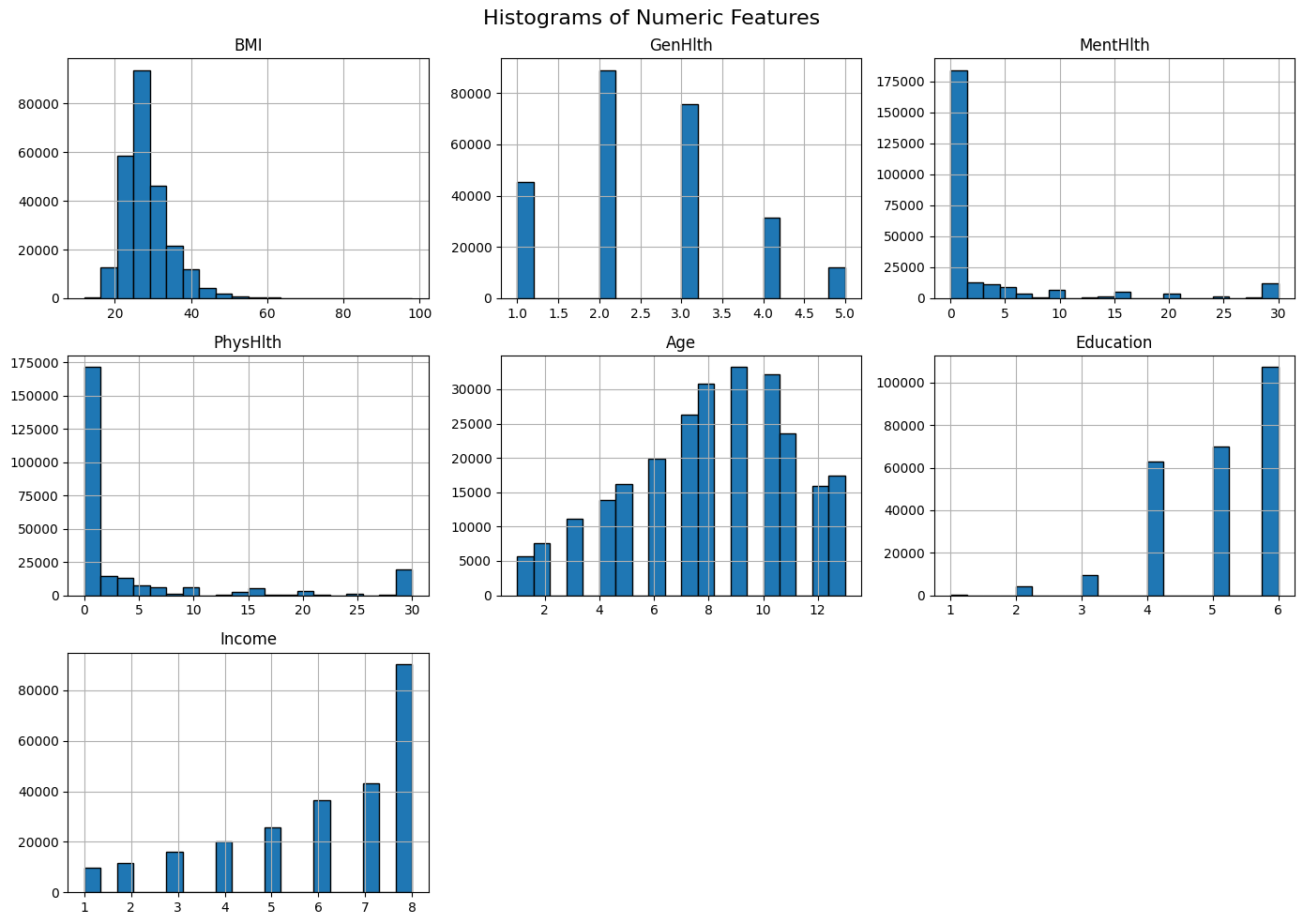
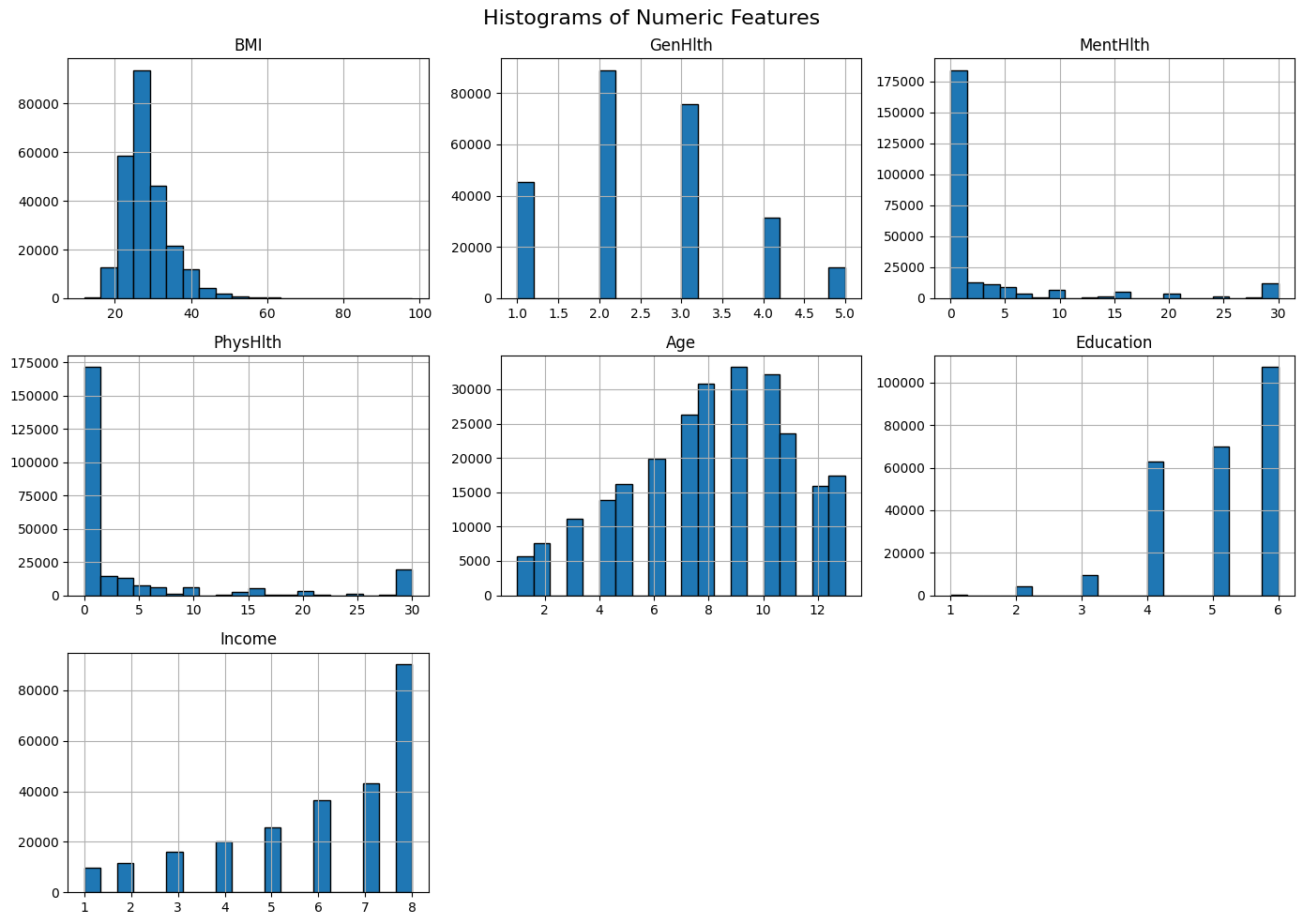
Several notable trends were observed in the dataset:

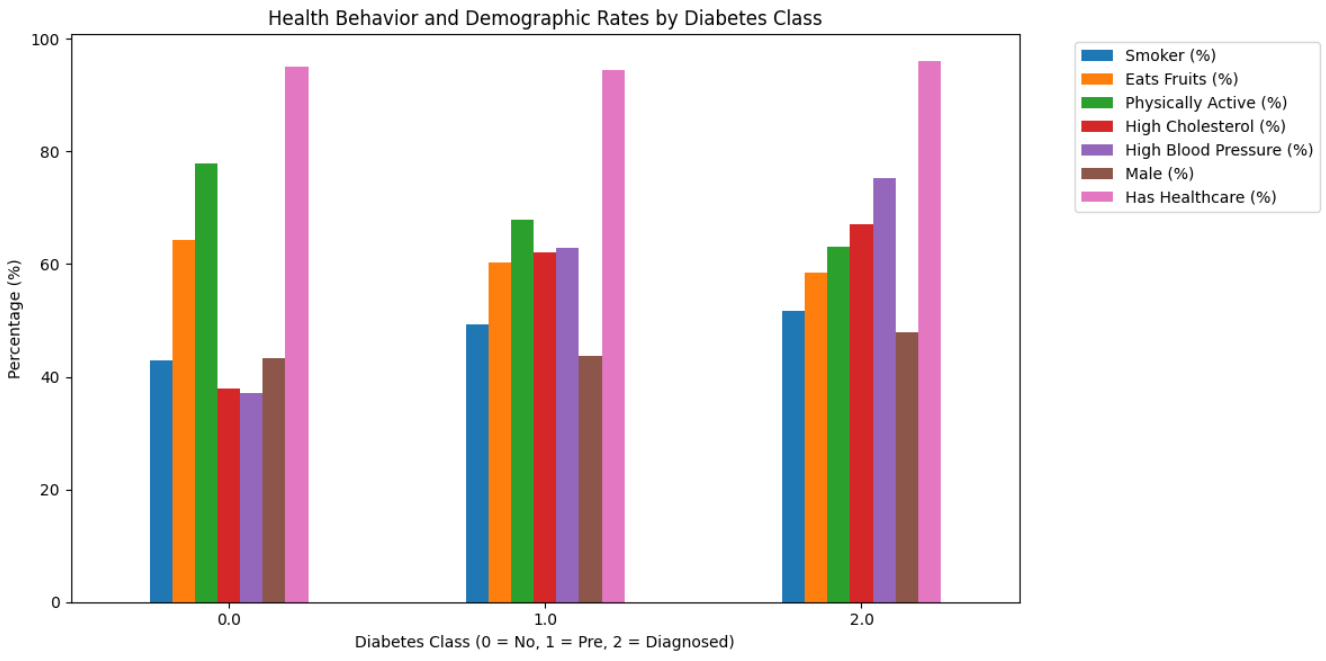
* Body Mass Index (BMI):
  + The average BMI among individuals in the dataset is approximately 28, which classifies as overweight according to standard health guidelines.



* Education Level:
  + The majority of patients have an education level of 4 or higher, indicating that most individuals are high school graduates or have pursued further education



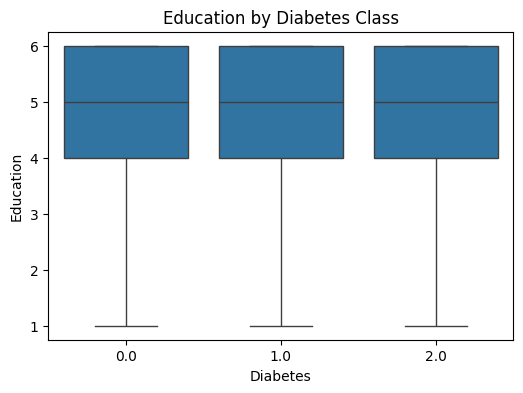
* Income Distribution
  + Income levels are left-skewed, meaning a larger portion of individuals fall into higher income brackets, particularly earning $75,000 or more per year.
* Mental Health:
  + Most individuals report experiencing fewer than 5 days per month of poor mental health, suggesting that mental health issues are relatively limited in frequency for the majority of the sample.
* Physical Health:
  + Similarly, most individuals report fewer than 5 days per month of poor physical health, indicating positive physical well-being across the dataset.
* Patient behaviors and other health conditions 



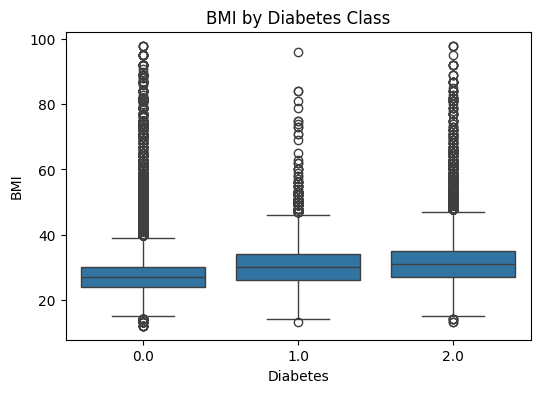
* Many characteristics seemed to be consistent across the different types of patients: diabetic, prediabetic, and non-diabetic. The percentage of patients with access to healthcare stayed fairly consistent across all 3 classes of patients in the dataset. Non-diabetic patients tended to be more physically active than diabetic and pre-diabetic patients. Additionally, prediabetic and diabetic patients exhibited higher rates of high cholesterol and high blood pressure.

Distributions of health and demographic features

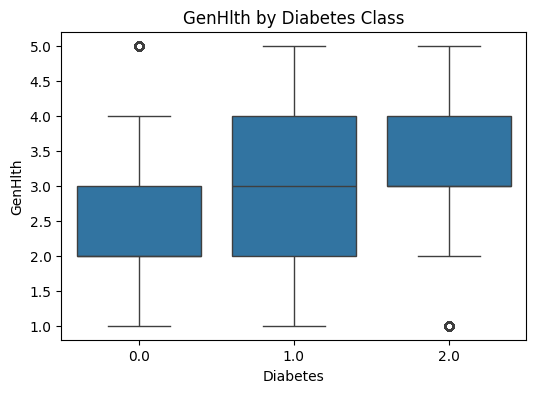
* Education by diabetes class
  + The median level of education is consistent across the 3 classes. Of note, the distribution also appears to be uniform across the 3 classes.



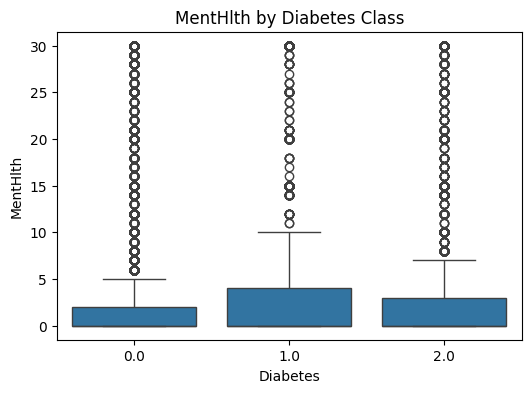
* BMI by diabetes class
  + Median BMI appears to increase slightly for patients with diabetes.



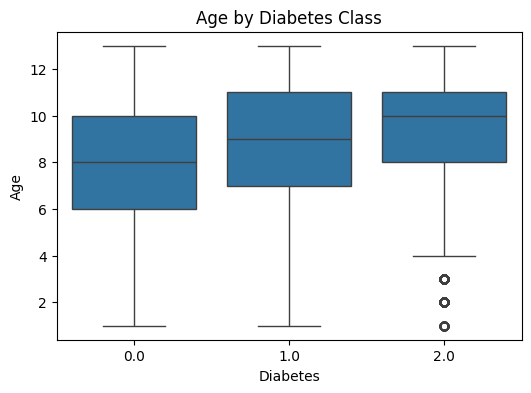
* General health by diabetes class
  + Median health rate appears to worsen from class 0 to class 2, with higher values indicating worsening health



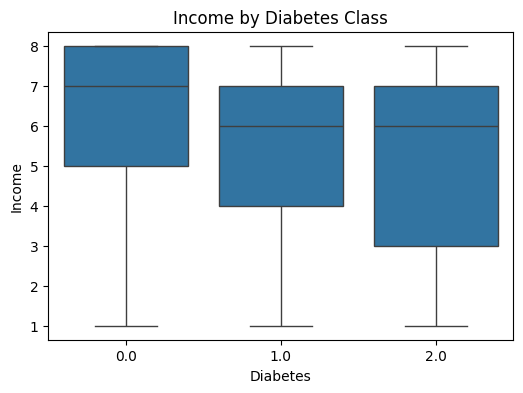
* Mental health by diabetes class
  + Class 2 appears to experience a larger number of days in poor health compared to class 0



* Age by diabetes class
  + The median age seems to increase from class 0 to class 2.



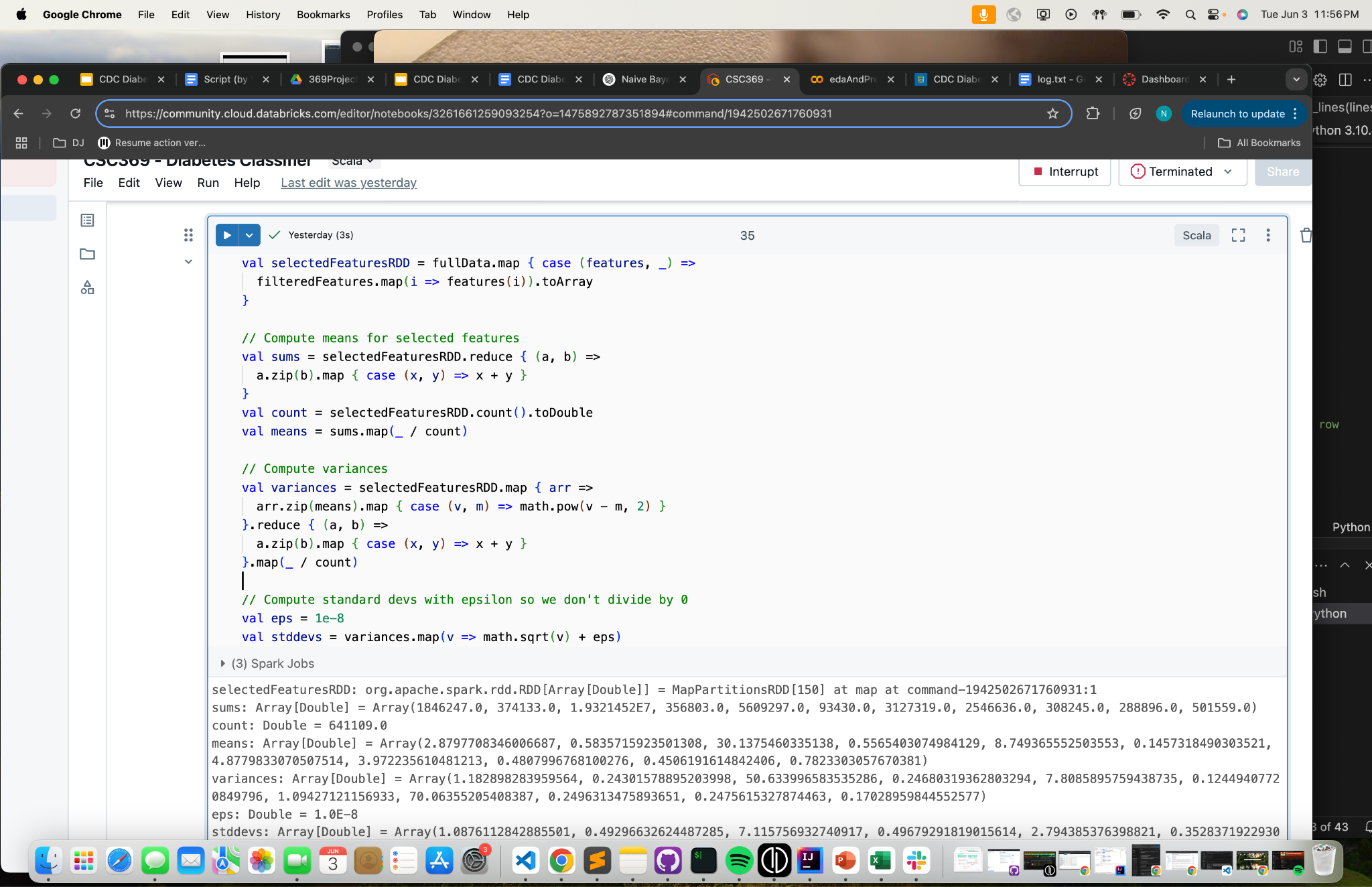
* Income by diabetes class
  + The median income decreases from class 0 to class 2.



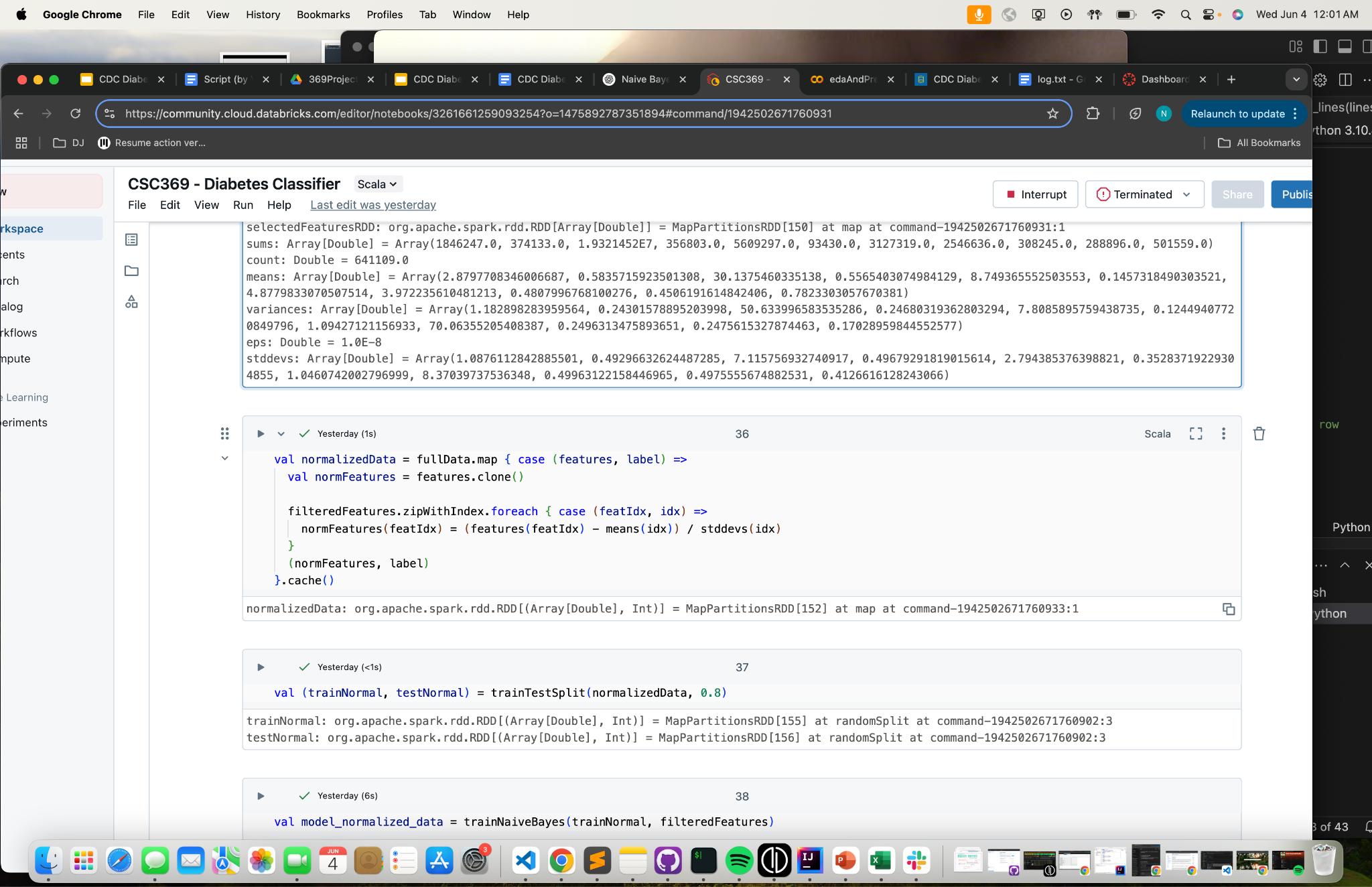
**Data Preprocessing**

I. Feature Normalization

Since the features explored were either binary or categorical with different numeric scales, we normalized each feature by standardizing them so that each feature has a mean of 0 and standard deviation of 1.



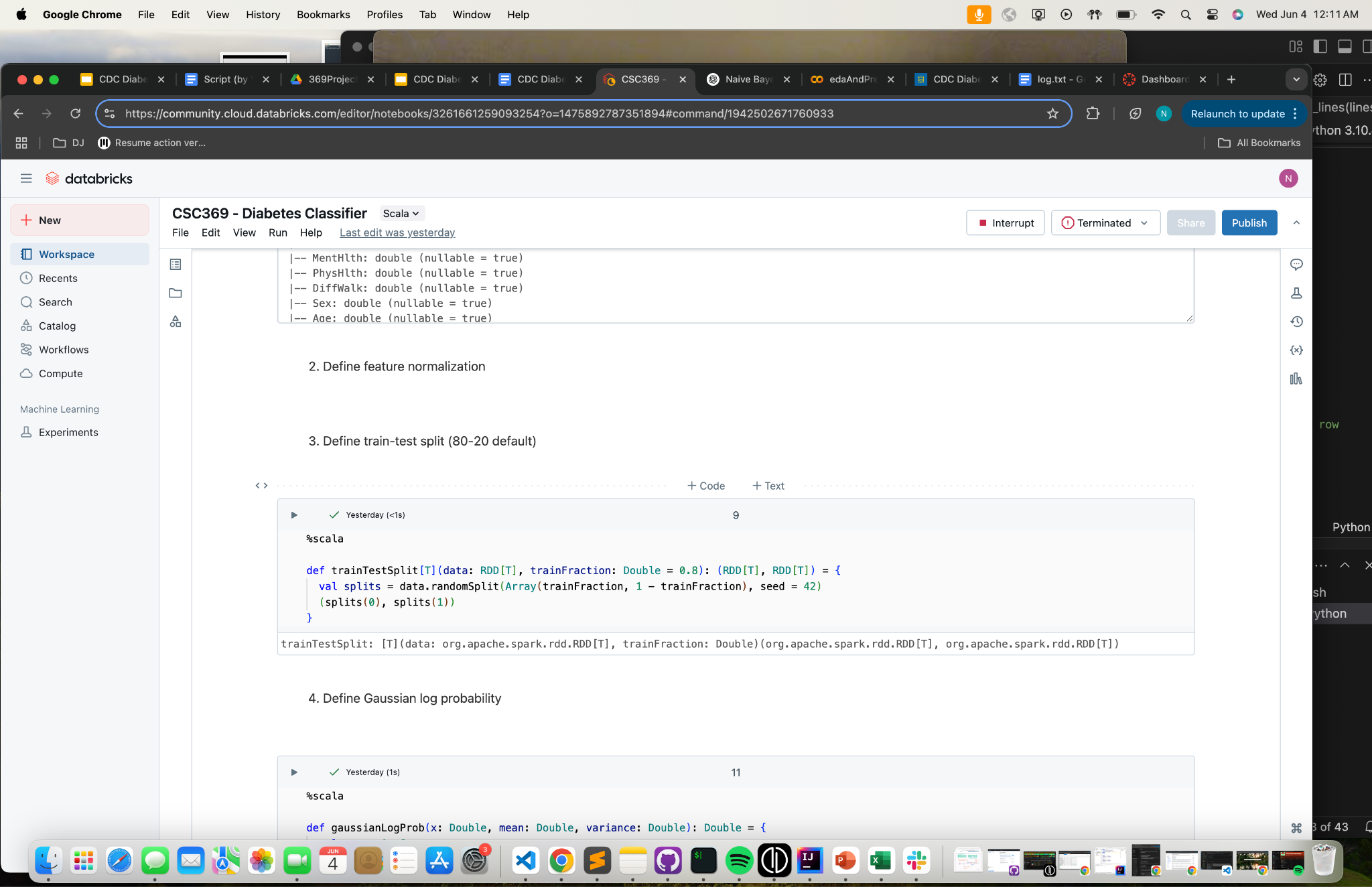
1. Specific features were selected from the list mentioned previously and stored in the selectedFeaturesRDD.
2. Means were calculated for each feature array.
3. Variances were calculated for each feature array.
4. The standard deviation was computed using the variance, and epsilon was added to prevent division by zero errors.



1. The normalized data is saved as (feature, label) pairs, where feature vectors have been normalized using the corresponding means and sttdevs arrays.

II. Train-test split

We divided our data into training and test sets using the standard 80% training, 20% test split as defined below.



**Model Selection & Fitting**

We fit a Gaussian Naive Bayes Classifier to predict diabetes diagnosis from the following variables:

* General health score
* High blood pressure (0/1)
* BMI
* High cholesterol (0/1)
* Presence of heart disease / attack (0/1)
* Mental health score
* Smoker (0/1)
* Age
* Education level
* Sex

These features were chosen because they had higher correlations with the target variable relative to other factors explored. We made sure to include key demographics like age, education level, and sex so that there was a balance between health predictors and other types of predictors in our model.

In general, for each class (0 = Healthy, 1 = Prediabetic, 2 = Diabetic), we compute the following:

* = Predicted class (0 / 1 / 2)
* = Prior probability of class y
* = Log posterior
* = Likelihood of feature given it belongs to class

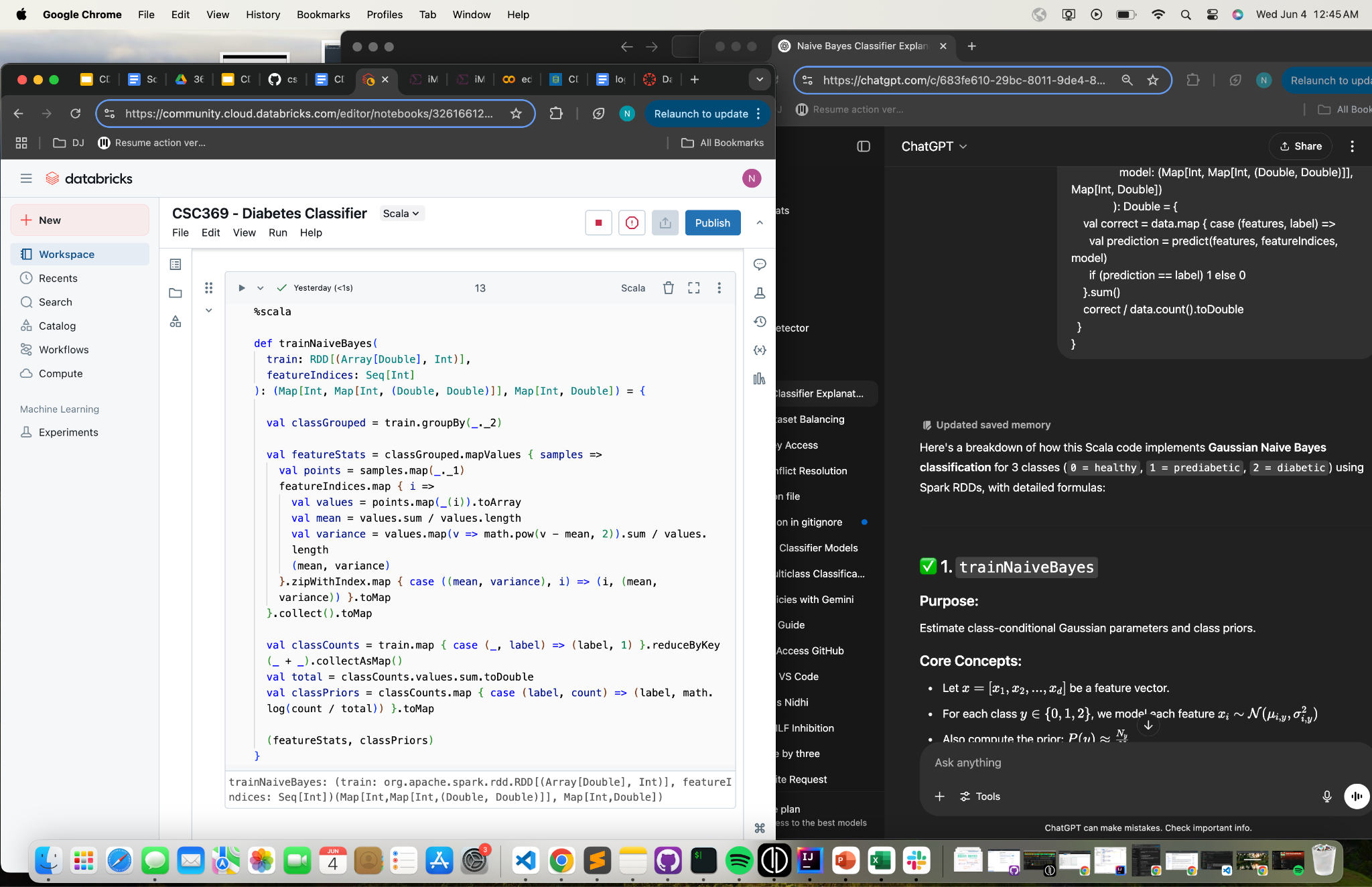
Note that is modeled by a Gaussian distribution, where

We take the log of this likelihood for this model.

This formula for y-hat works because we want to pick the most probable class for each predicted label, which should be the highest posterior probability.

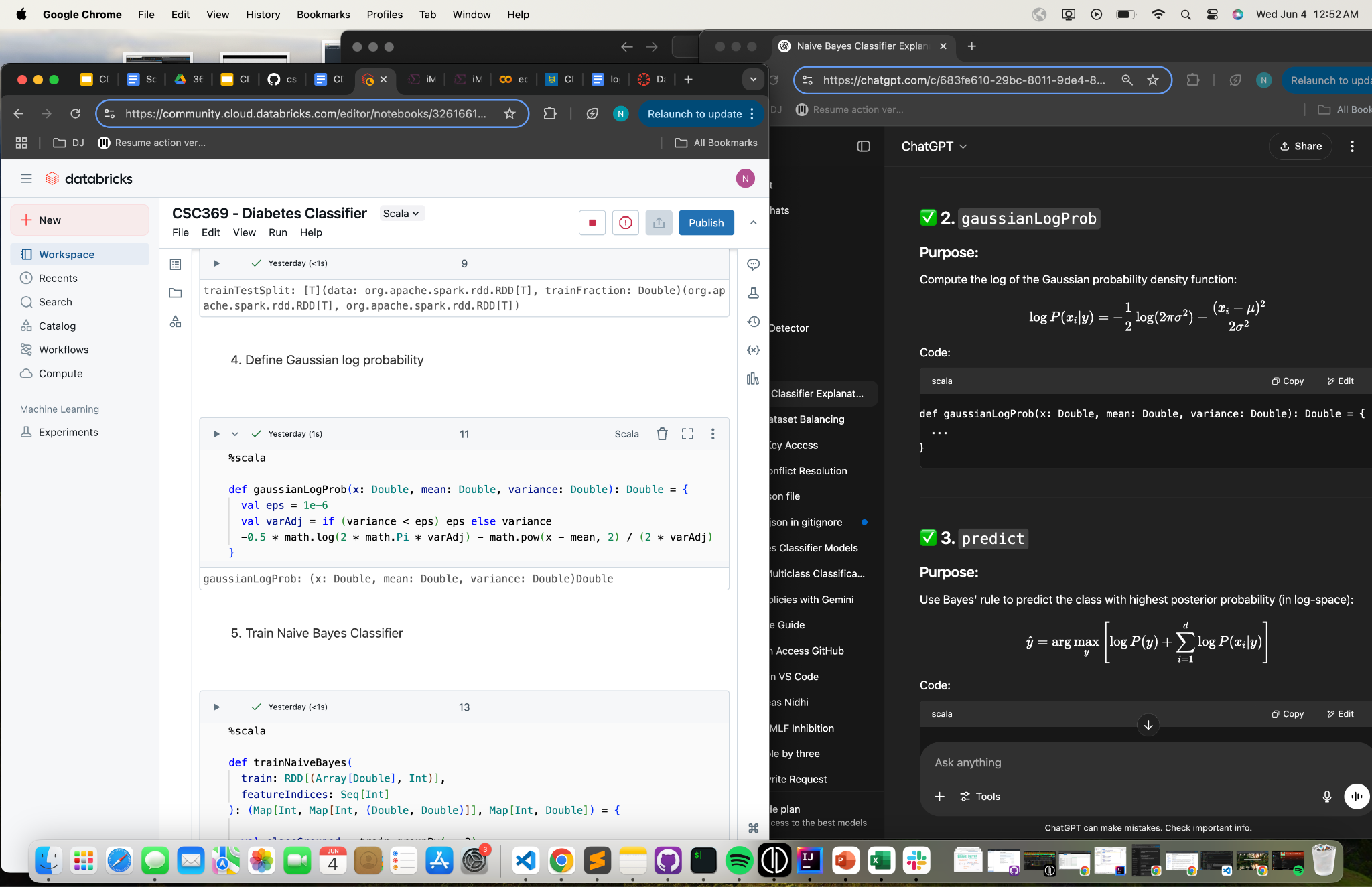
We followed these steps to build and run our Naive Bayes Classifier under the Gaussian assumption:

1. Train Naive Bayes Classifier

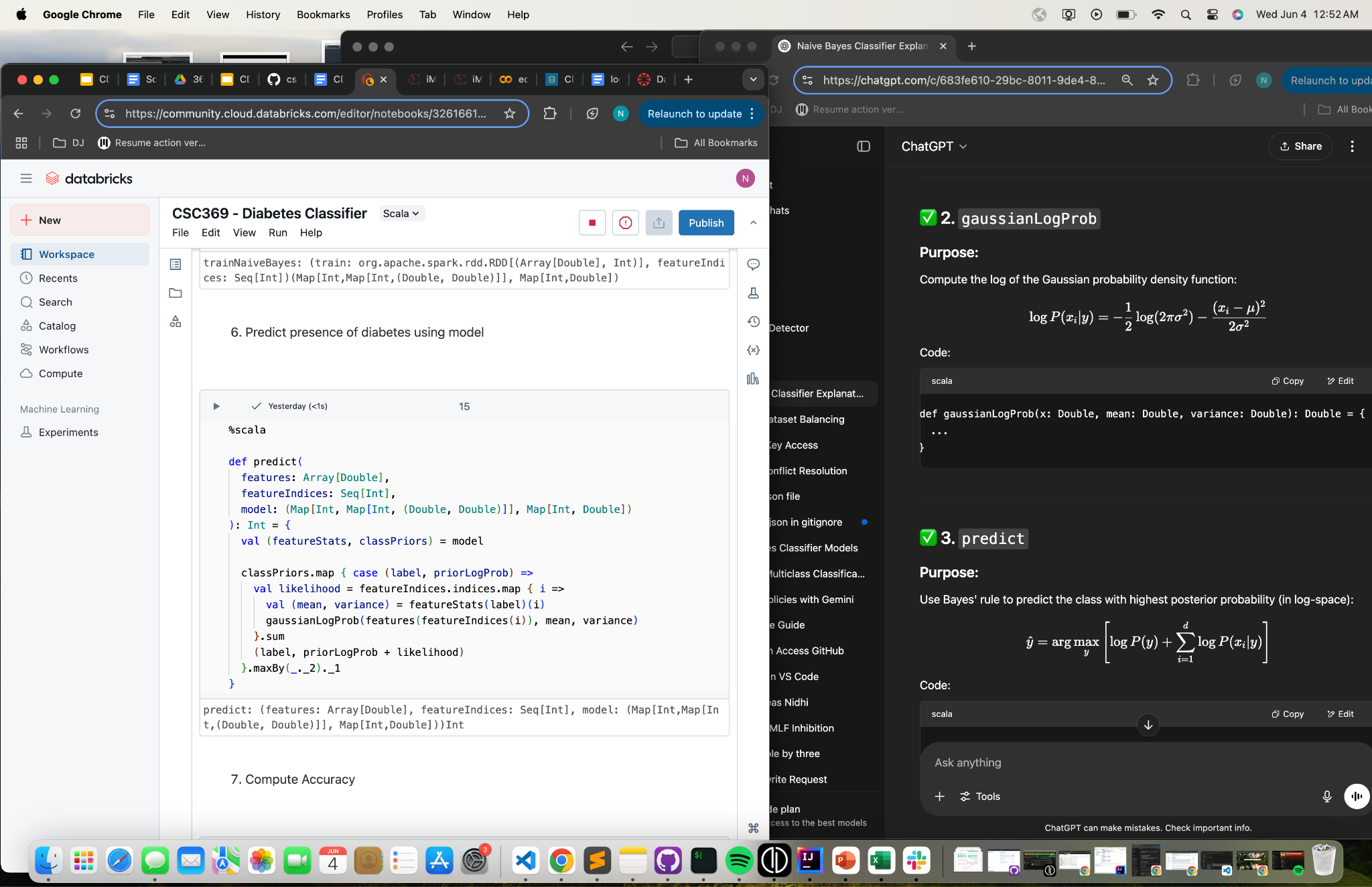


* For each class y = {0, 1, 2} we model each feature using the normal distribution.
  + This is done by storing mean and variance arrays for each class y.
* Compute log prior probabilities using classCounts and mapping them to each class.
* The function returns the transformed features and log priors.

1. Compute Gaussian log probabilities () from x’s, means, and variances.



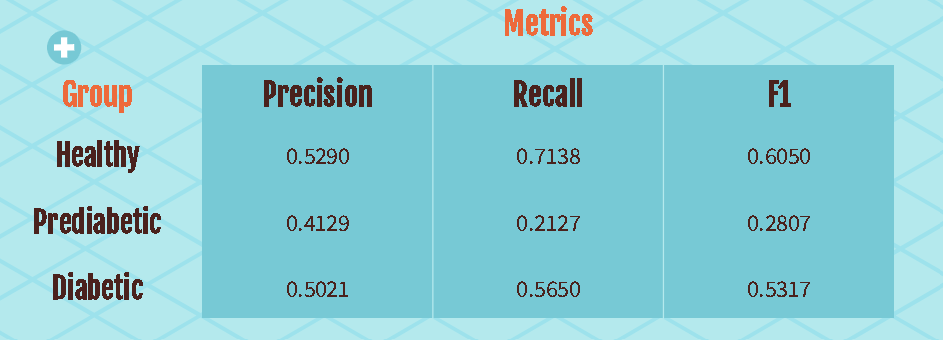
1. Use Bayes’ Rule (formula for ) to predict class with highest log posterior probability.



1. Compute accuracy as proportion of correct predictions.

**Results**

To evaluate the performance of our multi-class classification model, we computed the precision, recall, and F1-score for each diabetes class (non-diabetic(0), pre-diabetic (1), diabetic (2)). The metrics are as follows:



The model performs best on class 0, with a high recall indicating that it correctly identifies the majority of non-diabetic individuals. This suggests the classifier is well-tuned to detect the absence of diabetes.

Performance of the classifier for class 1 was substantially weaker as evidenced by the fact that the model only managed to correctly identify about 20% of true cases of pre-diabetic patients, resulting a low F1 score. This suggests that predicting cases of pre-diabetic patients may be difficult. The classifier yielded stronger results when predicting on class 2 (diabetic patients). The precision and recall are fairly balanced and the model is able to detect more than 50% of true positive cases.

**Key Takeaways**

The data shows that most patients have pretty similar backgrounds when it comes to demographics and socioeconomic status. Because there isn’t much difference here, these factors might not be very useful for predicting who will develop diabetes.

However, health issues like high cholesterol and high blood pressure seem to be important risk factors for diabetes. This means keeping an eye on these conditions could help in spotting and preventing diabetes.

**Challenges**

One significant challenge we encountered was that Naive Bayes was not well-suited for our dataset. The relatively small sample size made it difficult for the model to learn accurate patterns. Additionally, Naive Bayes relies on certain assumptions, such as that the features are independent and follow a normal distribution. These assumptions did not hold true for our data. For instance, it is reasonable to expect correlations between features such as high cholesterol and dietary habits (e.g., fruit and vegetable intake), which violates the independence assumption.

Another issue was class imbalance. The dataset contained a much larger number of healthy individuals compared to those with prediabetes or diabetes. As a result, the model struggled to correctly classify the less-represented classes, leading to poor performance in identifying diabetic and prediabetic.

**Conclusion**

In this project, we explored the relationship between various health, lifestyle, and demographic factors and a likelihood of receiving a diabetes diagnosis. Through our analysis of over 250,000 instances from a CDC-collected dataset, we identified key health indicators, such as high blood pressure and high cholesterol as significant predictors of diabetes.

Our modeling approach using a Naive Bayes classifier highlighted both the promise and limitations of predictive modeling in health data. While we were able to incorporate a diverse range of features and normalize them effectively, the model’s performance was constrained by key challenges, such as the assumptions of Naive Bayes and the imbalance in class labels.

Despite these limitations, our findings reinforce the importance of monitoring specific health conditions as potential early warning signs for diabetes. Future work could explore more flexible classification models and address class imbalance through resampling techniques or improve evaluation metrics. By refining these approaches, we can work toward more accurate and equitable predictive models in healthcare.