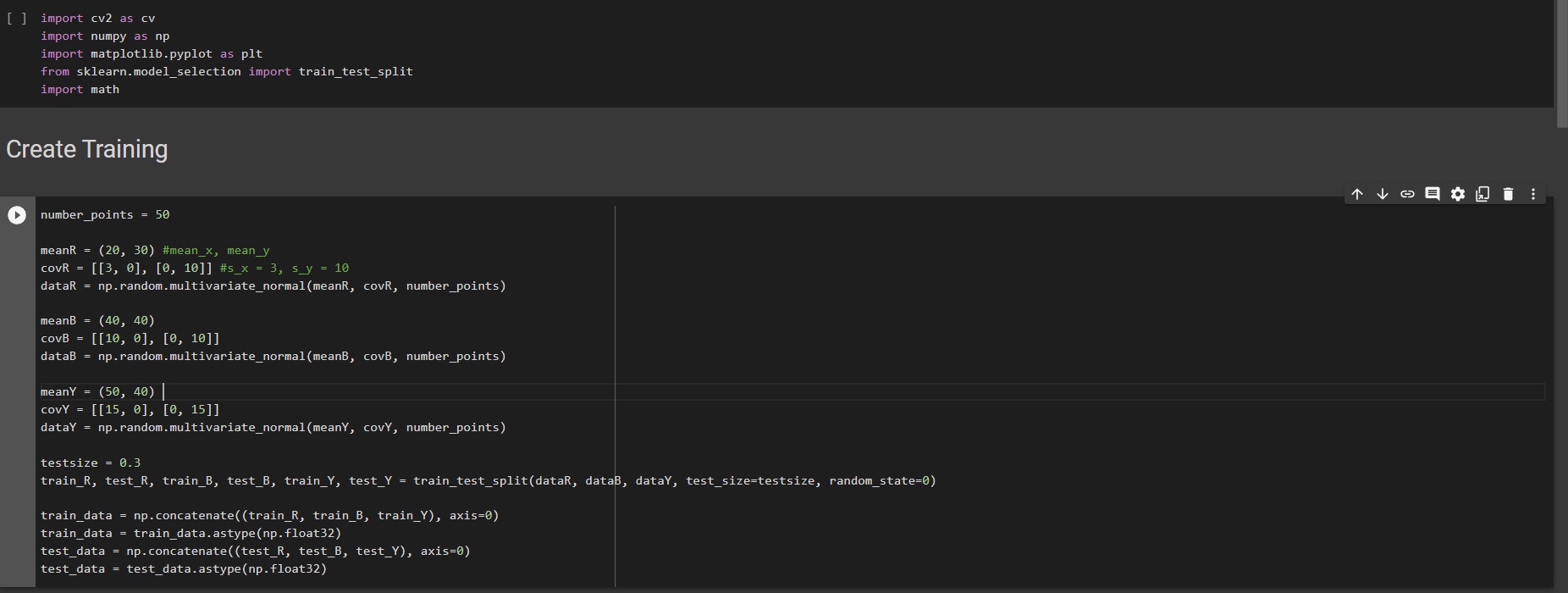
**CPE383 Machine Learning: Quiz2**

1. 1.5 hr. Min Distance classifier on 3 Gaussian Classes. Modify your KNN program from quiz 1 with 3 classes to create 50 random points for each class: red, blue, and yellow from a 2D Gaussian distribution (see Gaussian Data.ipynb) with means: (20, 30), (40, 40), (50, 40) and (s\_x, s\_y) of (3, 10), (10, 10), (15, 15), for red, blue, and yellow, respectively. Use 70% of the dataset as training data and 30% as testing data.
2. 5 pts. Plot the 3 classes using the training data.



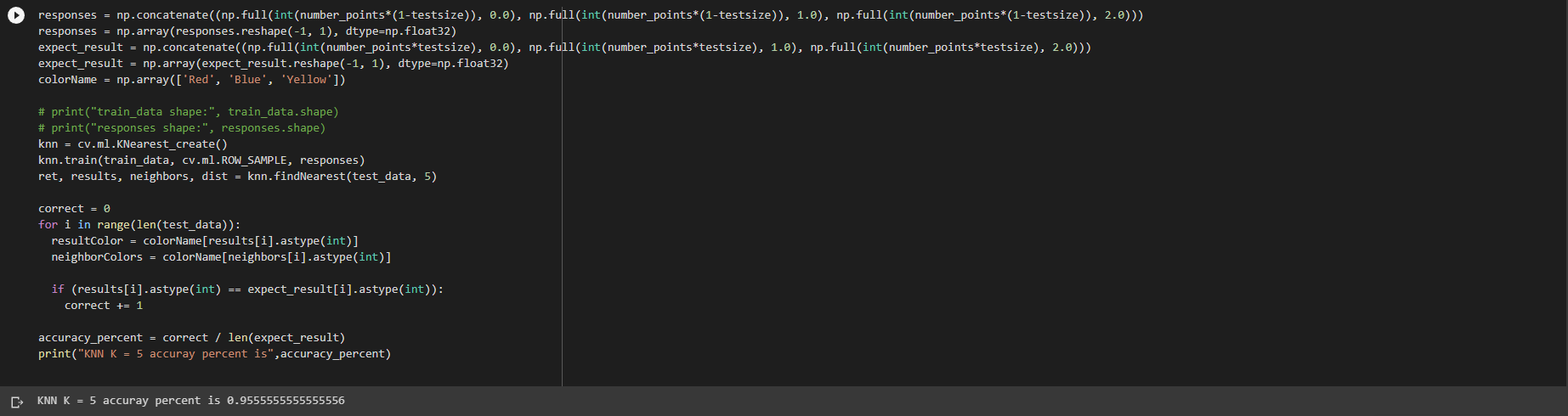
Graphical user interface

Description automatically generated with medium confidence

Background pattern

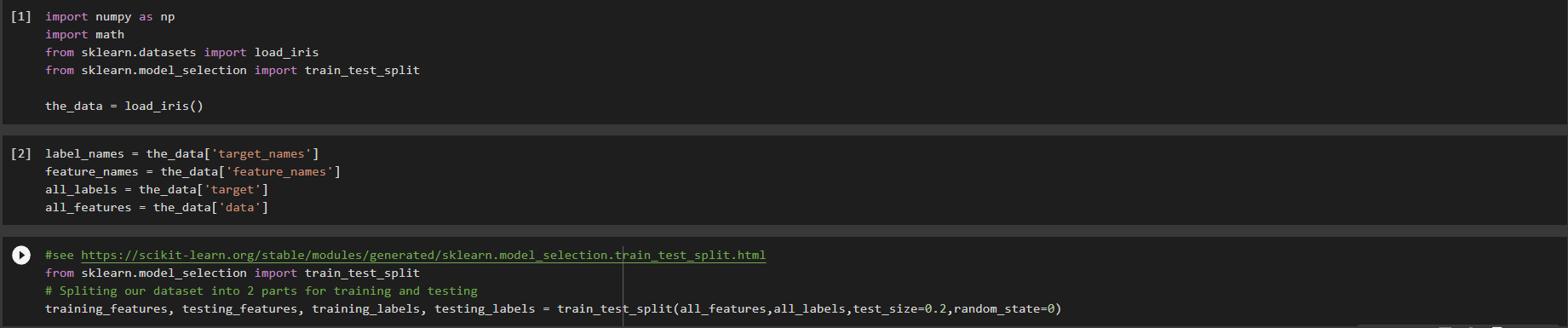
Description automatically generated with low confidence

1. 5 pts. Using KNN with K = 5, report the total accuracy of the testing data.

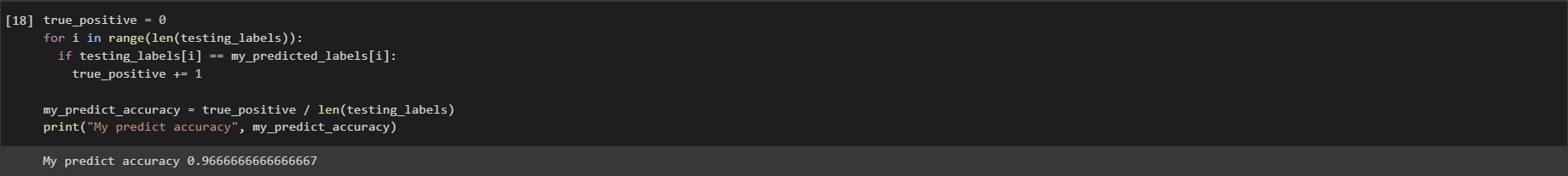


1. Graphical user interface, text

   Description automatically generatedUsing the minimum distance classifier: 5 pts. Report the training data cluster mean for each class of red, blue, and yellow.
2. A picture containing graphical user interface

   Description automatically generatedReport the total accuracy of the testing data using this classifier.
3. 4 hrs. Code Naive Bayes from Scratch. Write a program to read the Iris dataset, split into 2 parts: training and testing just like it was done in the example. Then write your own code to:
4. 5 pts. Find the mean and standard deviation for each of the 4 features of each of the 3 classes from the training data. μ\_ik and σ\_ik for i = 1..4, k = 1..3. You will get pdfs P(x\_i | c\_k) for each class using the Gaussian distribution equation with μ\_ik and σ\_ik for i = 1..4, k = 1..3. This gets you pdfs: P(x\_1, x\_2, x\_3, x\_4 | c\_1), P(x\_1, x\_2, x\_3, x\_4 | c\_2), P(x\_1, x\_2, x\_3, x\_4 | c\_3).
5. Text

   Description automatically generated5 pts. Find the P(c\_k) by counting the percent frequency of each class in your training data. Now we have P(c\_k | x\_1, x\_2, x\_3, x\_4) ∝ P(x\_1, x\_2, x\_3, x\_4 | c\_k) \* P(c\_k).
6. A picture containing background pattern

   Description automatically generated5 pts. Then for each (x\_1, x\_2, x\_3, x\_4) in your test data: find the class k of 1, 2, or 3 for which P(c\_k | x\_1, x\_2, x\_3, x\_4) is maximum, put that k into array my\_predicted\_labels
7. 5 pts. Calculate and print the accuracy score from your implementation of Naive Bayes from scratch
8. Text

   Description automatically generated5 pts. Use sklearn's GaussianNB classifier to report the accuracy score. Compare your result to sklearn’s.
9. 1 hr. Try the digits datasets. Change the “**Naive Bayes and KNN Iris and Cancer.ipynb**” program to allow the user to also select the digits dataset by entering “digits”, in addition to “iris” and “cancer”. Present the output results for both Naive Bayes and KNN classifiers using Sklearn. What is the best value of K in KNN?

Graphical user interface, text, application

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generatedText

Description automatically generatedGraphical user interface

Description automatically generated with low confidence

Text

Description automatically generatedGraphical user interface

Description automatically generated with low confidenceText

Description automatically generated

Ans: The best K is 1 with 98.89% accuracy

1. 10 pts. 1.5 hr. Normalize data option. Add an option to “Naive Bayes and KNN Iris and Cancer.ipynb” to ask the user whether to normalize the dataset by converting each feature into a Z-distribution by making mean = 0, and standard deviation = 1. For this problem, compare the accuracy results for the breast cancer dataset on the sklearn's KNN classifier using normalized vs. unnormalized data.

Graphical user interface

Description automatically generated Text

Description automatically generated Text

Description automatically generated Graphical user interface, text

Description automatically generated A screenshot of a computer

Description automatically generated with medium confidence Text

Description automatically generated Shape

Description automatically generated with low confidence Text

Description automatically generated Graphical user interface, application

Description automatically generated with medium confidence Text

Description automatically generated Text

Description automatically generated

Ans: The best K of normalized data is 16, different from unnormalized that is 9