1. To split the train and test data from the provided wine.csv. It is necessary to know how to read and write the csv file. Thus, you need to randomly split 18 instances from each type as testing dataset and totally 54 instances from the whole dataset. Then save the training dataset as train.csv and testing dataset as test.csv. (124 instances for training and 54 instances for testing.)

```
# Split the wine dataset according to its label
wine1_test = wine1.sample(n=18)
wine1_train = wine1.drop(wine1_test.index)
wine2_test = wine2.sample(n=18)
wine2_train = wine2.drop(wine2_test.index)
wine3_test = wine3.sample(n=18)
wine3_train = wine3.drop(wine3_test.index)
```

```
# Build train and test sets
(variable) test_data: Series
train = pd.concat([wine1_train, wine2_train, wine3_train])
test_data = pd.concat([wine1_test, wine2_test, wine3_test])
✓ 0.1s
```

```
# Instance and its label
print(train_data.shape)
print(test_data.shape)
✓ 0.1s
(124, 14)
(54, 14)
```

```
# Save as csv file
train_data.to_csv('train.csv')
test_data.to_csv('test.csv')
✓ 0.9s
```

2. To evaluate the posterior probabilities, you need to learn likelihood functions and prior distribution from the training dataset. Then, you should calculate the accuracy rate of the MAP detector by comparing to the label of each instance in the test data. Note that the accuracy rate will be different

depending on the random result of splitting data, but it should exceed 90% overall. (Please add corresponding comments in your code to describe how you obtain the posterior probability.)

The higher posterior probability it has, the likely the type is.

```
# Calculate prior for each wine
priors = [0,0,0]
train_total = sets_of_wine[0]+sets_of_wine[1]+sets_of_wine[2]
for i in range(3):
    priors[i] = sets_of_wine[i]/train_total
```

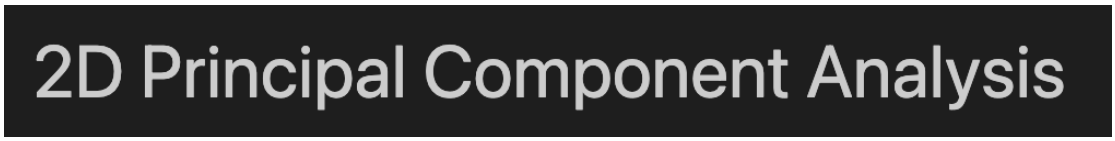
Use the priors to get the result

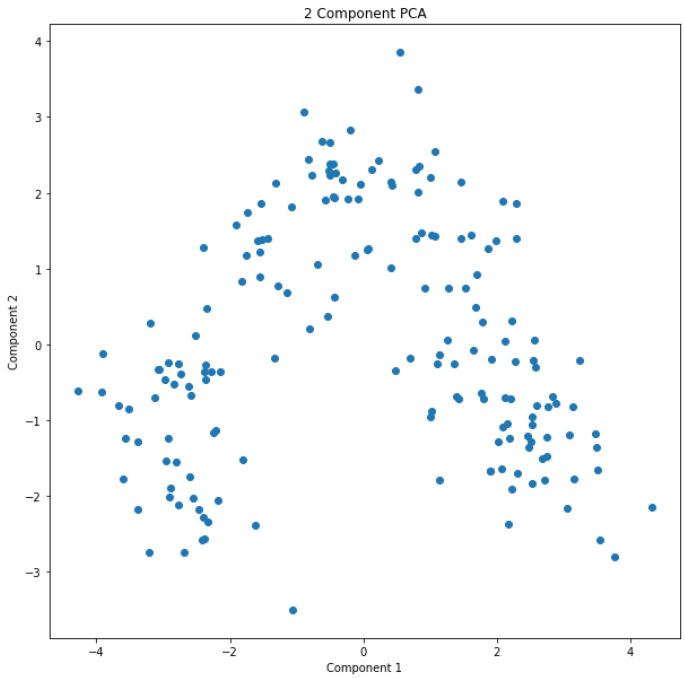
```
print('accuracy percent: {:.2%}'.format(correctly_labeled/test_data.shape[0]))
print(prediction, len(prediction))
```

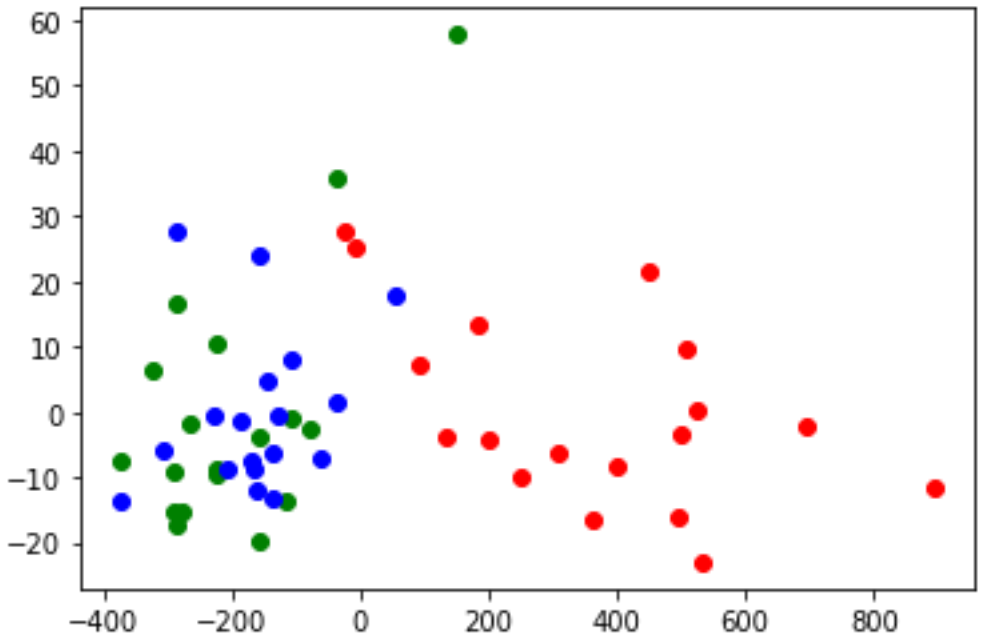
✓ 0.1s

```
accuracy percent: 98.15%
```

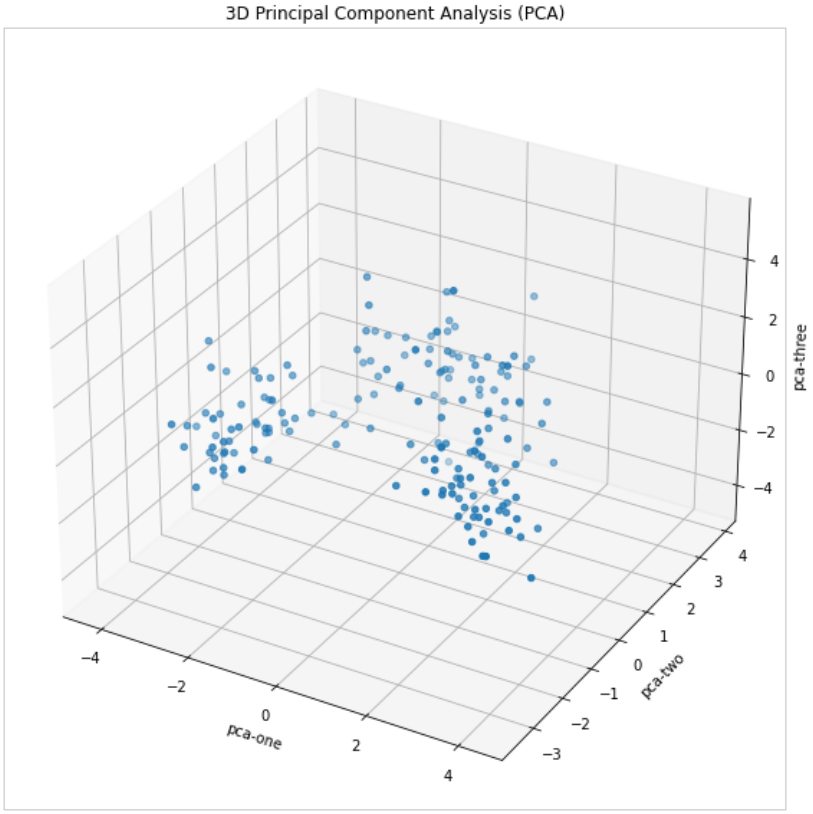
3. **Please plot the visualized result of testing data in your report. (You can directly use the built-in PCA function to get visualized result.)**

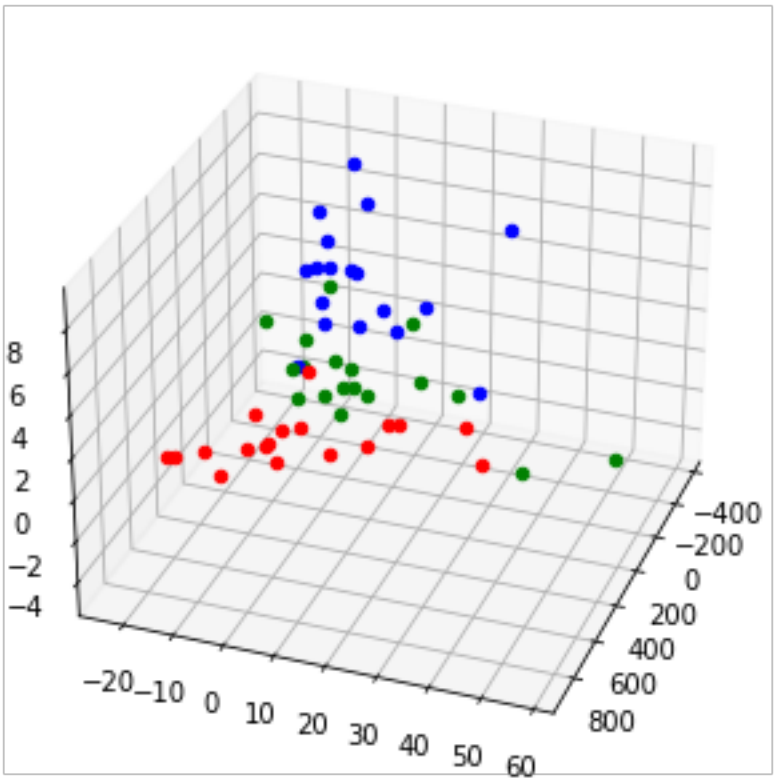






3D Principal Component Analysis





4. Please discuss the effect of prior distribution on the posterior probabilities in your report.

You can think of posterior probability as an adjustment on prior probability: Posterior probability = prior probability + new evidence (called likelihood).

For example, historical data suggests that around 60% of students who start college will graduate within 6 years. This is the prior probability. However, you think that figure is actually much lower, so set out to collect new data. The evidence you collect suggests that the true figure is actually closer to 50%; This is the posterior probability.