**Q1: K-NN**

a. Systematically vary the value of Split and the value of K.

split_set = [0.2,0.4,0.6,0.7,0.8]

k_set = [2, 3, 6, 9, 12,15,18,21,24,27]

For Iris Data set processing, the results are as follows.

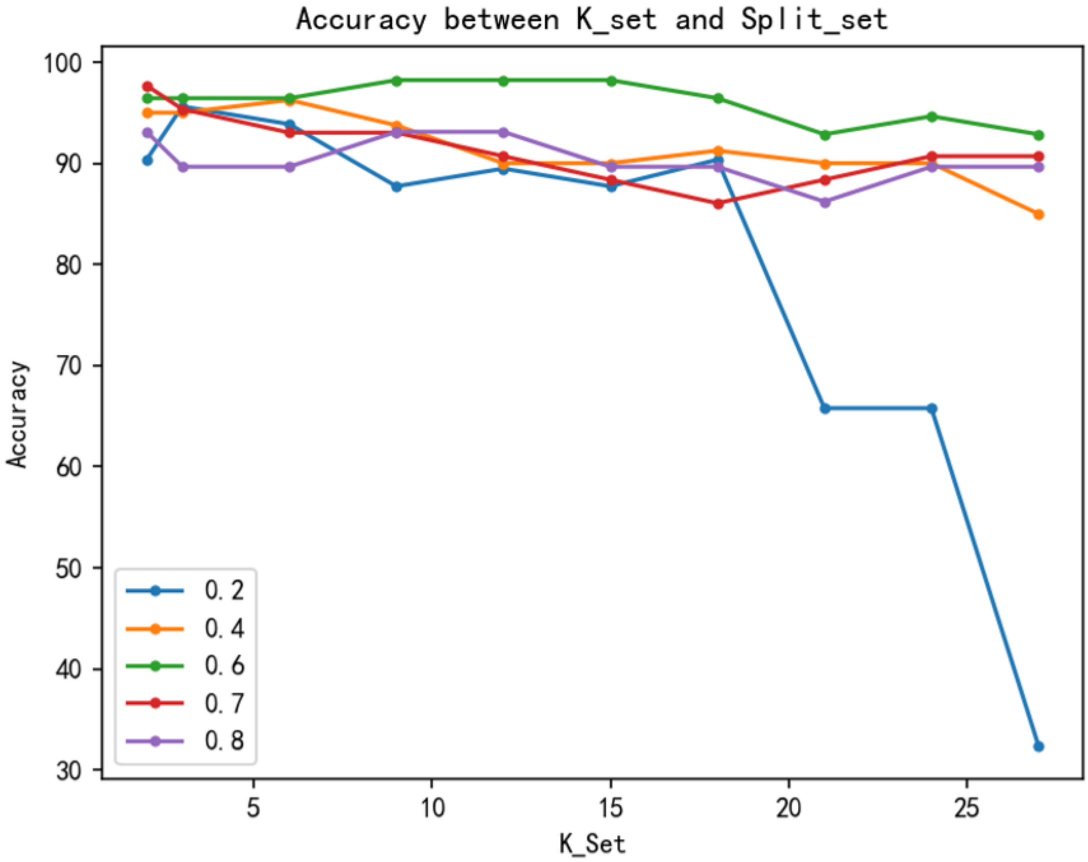


Figure 1: Accuracy between K set and Split_set

When split = 0.2 is chosen, 20 percent of the training data and 80 percent of the test data. The accuracy results obtained from the calculation are lower. Due to the model's small training set, the prediction results' accuracy will be lower compared to other models with more training sets.

However, a higher split value is not better. When split = 0.6, 60 percent of the training set and 40 percent of the test set. The model achieves the highest prediction accuracy. As the value of split was increased again, the accuracy did not improve with the increase. Providing more training sets helps improve the model's accuracy, but higher is not better when the number of sets is constant.

Considering the X-axis, the accuracy does not always increase as the value of K increases. Instead, the accuracy decreases after increasing to a certain level. When split = 0.2, the excessive increase in the value of K leads to the need for more neighbors to be labeled. However, because the training set is so small, there is often not so much data, which leads to a decrease in accuracy, eventually down to close to 30%.

b. k-fold cross-validation

Select split = 0.6, and apply k-fold cross-validation, k = 5. The k values in the cross-validated k-NN are [2, 3, 6,9,12,15,18,21,24,27]. Here is the plot.

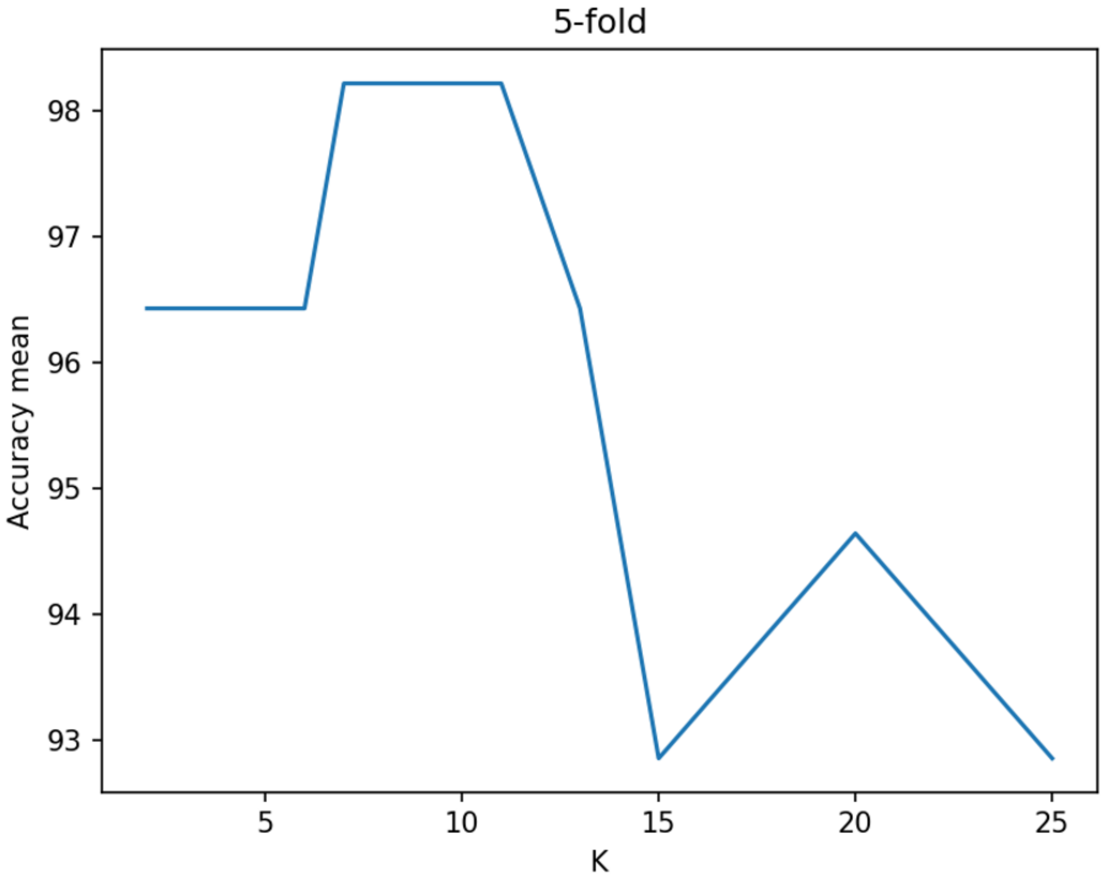
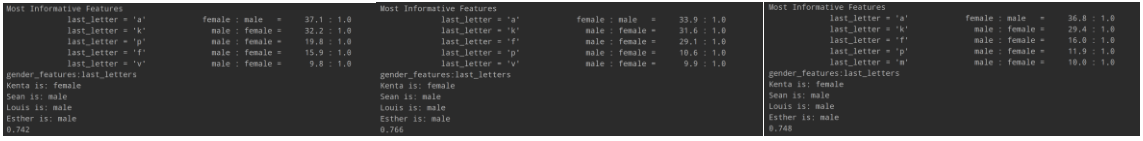


Figure 2: 5-fold

We see that the accuracy rate increases with increasing k until 7, 9 and 11, when it peaks, after which the accuracy rate decreases as k increases again. After reaching k = 15, the accuracy increases slightly, after which the accuracy decreases again as k increases. Therefore, in this data test, the accuracy should be the highest for values around k = 7, 9 and 11. However, due to the limited data samples, uncontrolled expansion of the value of k, which increases both the number of neighbors, will lead to the number of neighbors in the prediction being more significant than the number of training samples, making it with each sample is a neighbor, but lead to the accuracy of the results decreased, and even the program run error.

## Q2: Bayes Classifiers

a. The accuracy scores for the original classifier using the last_letter feature as follows.



```
Most Informative Features
             last_letter = 'a'            female : male   =     37.1 : 1.0
             last_letter = 'k'              male : female =     32.2 : 1.0
             last_letter = 'p'              male : female =     19.8 : 1.0
             last_letter = 'f'              male : female =     15.9 : 1.0
             last_letter = 'v'              male : female =      9.8 : 1.0
gender_features:last_letters
Kenta is: female
Sean is: male
Louis is: male
Esther is: male
0.742
```

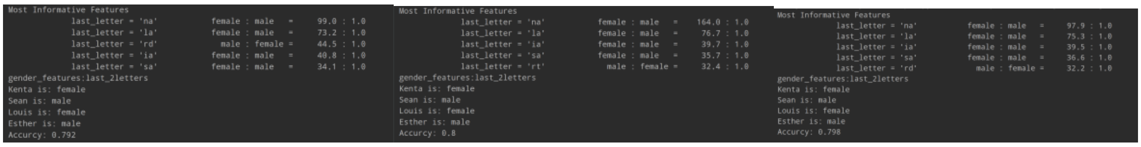
```
Most Informative Features
             last_letter = 'a'            female : male   =     33.9 : 1.0
             last_letter = 'k'              male : female =     31.6 : 1.0
             last_letter = 'f'              male : female =     29.1 : 1.0
             last_letter = 'p'              male : female =     10.6 : 1.0
             last_letter = 'v'              male : female =      9.9 : 1.0
gender_features:last_letters
Kenta is: female
Sean is: male
Louis is: male
Esther is: male
0.766
```

```
Most Informative Features
             last_letter = 'a'            female : male   =     36.8 : 1.0
             last_letter = 'k'              male : female =     29.4 : 1.0
             last_letter = 'f'              male : female =     16.0 : 1.0
             last_letter = 'p'              male : female =     11.9 : 1.0
             last_letter = 'm'              male : female =     10.0 : 1.0
gender_features:last_letters
Kenta is: female
Sean is: male
Louis is: male
Esther is: male
0.748
```

Figure 3: Last_letter Feature

I ran them three times and obtained three accuracy values, which are 0.742, 0.766 and 0.748. The mean value is 0.752. We can see that the last letter 'k,' 'f,' 'p,' and 'v' are predicted to be male to a greater extent. Apart from that, 'a' are more likely to be predicted as female. However, we can see that it incorrectly classifies “Kenta” as female and “Esther” as male. “Sean” is more similar to neutral. And the classification of “Louis” is correct.

b. I have defined a new method that takes the last two letters of the name



```
Most Informative Features
             last_letter = 'na'           female : male   =     99.0 : 1.0
             last_letter = 'la'           female : male   =     73.2 : 1.0
             last_letter = 'rd'             male : female =     44.5 : 1.0
             last_letter = 'ia'           female : male   =     40.8 : 1.0
             last_letter = 'sa'           female : male   =     34.1 : 1.0
gender_features:last_2letters
Kenta is: female
Sean is: male
Louis is: female
Esther is: male
Accury: 0.792
```

```
Most Informative Features
             last_letter = 'na'           female : male   =    164.0 : 1.0
             last_letter = 'la'           female : male   =     76.7 : 1.0
             last_letter = 'ia'           female : male   =     39.7 : 1.0
             last_letter = 'sa'           female : male   =     35.7 : 1.0
             last_letter = 'rt'             male : female =     32.4 : 1.0
gender_features:last_2letters
Kenta is: female
Sean is: male
Louis is: female
Esther is: male
Accury: 0.8
```

```
Most Informative Features
             last_letter = 'na'           female : male   =     97.9 : 1.0
             last_letter = 'la'           female : male   =     75.3 : 1.0
             last_letter = 'ia'           female : male   =     39.5 : 1.0
             last_letter = 'sa'           female : male   =     36.6 : 1.0
             last_letter = 'rd'             male : female =     32.2 : 1.0
gender_features:last_2letters
Kenta is: female
Sean is: male
Louis is: female
Esther is: male
Accury: 0.798
```

Figure 4: Last_2letters Feature

I ran them three times and obtained three accuracy values, which are 0.792, 0.8 and 0.798. The mean value is 0.807. In the combination of the last two letters, 'na,' 'la,' and 'ia' are predicted to be female to a greater extent. Besides, 'us' and 'rt' are classified as male to a higher degree. We can see that It also has a similar incorrectness to the previous feature classification. It similarly classifies “Kenta” as female and “Esther” as male. However, it also creates a new error by classifying “Louis” as female.

In terms of the mean value of accuracy, the last two letters of the name were more accurate than the last. Male names mostly end in accented consonants (b c d g k p q t), -an, -ian, -in, -on, -r, -ah, -as, -o, -os, -us. while female names ending in -a, -ia, -ie, -ine, -e, -es, -is, -x, -z are predominant. Therefore, when classifying male names, the accuracy of taking the last letter is higher than taking the last two letters, as can be seen in the classification of “Louis”. When classifying names similar to Kenta and Esther, both appear to have been incorrectly predicted. The combination of taking the last two letters does not reflect a higher accuracy than taking the last letter due to the influence of names ending in "a" and "r."

Adding combinatorial features that fit the name pattern will improve the accuracy of Naive Bayes. Using the last two letters as features accuracy is more accurate than taking only the last letter, and the former has a more robust combinations.