**Part 1: Calculations:**

1. **Suppose we have the following training data including 15 training samples.**



# Target (yes) 🡪 9

P (yes) = 9 / 15

|  |  |  |
| --- | --- | --- |
| Color(X1): in (yes) | Gender(X2): in (yes) | Price(X3): in (yes) |
| * P (R | yes) 🡪 2 / 9 * P (G | yes) 🡪 3 / 9 * P(Yellow | yes) 🡪 4 / 9 | * P (F | yes) = 6 / 9 * P (M | yes) = 3 / 9 | * P (H | yes) = 2 / 9 * P (L | yes) = 5 / 9 * P (M | yes) = 2 / 9 |

# Target (no) 🡪 6

P (no) = 6 / 15

|  |  |  |
| --- | --- | --- |
| Color(X1): in (no) | Gender(X2): in (no) | Price(X3): in (no) |
| * P (R | no) 🡪 3 / 6 * P (G | no) 🡪 2 / 6 * P(Yellow | no) 🡪 1 / 6 | * P (F | no) = 1 / 6 * P (M | no) = 5 / 6 | * P (H | no) = 2 / 6 * P (L | no) = 1 / 6 * P (M | no) = 3 / 6 |

P (yes | G, F, H) =

=

= P (G, F, H | yes) \* P (yes) + P (G, F, H | no) \* P (no)

= P (G | yes) \* P (F | yes) \* P (H | yes) \* P (yes) + P (G | no) \* P (F | no) \* P (H | no) \* P (no)

= \* \* \* + \* \* \* = + = =

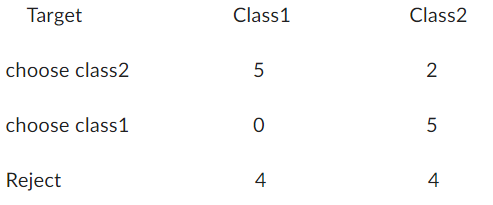
P (yes | G, F, H) = = = \* =

P (no | G, F, H) =

P (no | G, F, H) = = = \* =

The prediction when Color = G, Gender = F, Price=H is 0.80 %

2. Calculate the expected risk of three actions, and determine the rejection area of P (Class1| x)



We suppose that a1 is (choose class1) and a2 is (choose class2).

R (a1 | X) = 0 \* P (Calss1 | X) +5 \* P (Class2 | X)

R (a1 | X) = 0 P (Calss1 | X) + 5 (1 – P (Class1 | X)))

R (a1 | X) = - **5P (Calss1 | X) + 5 🡪** **equation 1**

R (a2 | X) = 5 \* P (Calss1 | X) + 2 \* P (Class2 | X)

R (a2 | X) = 5 \* P (Calss1 | X) + 2 \* (1 – P (Class1 | X))

= **2 + 3 P (Class1 | X) 🡪** **equation 2**

R (a3 | X) = 4 \* P (Calss1 | X) + 4 \* P (Class2 | X)

= 4 \* P (Calss1 | X) + 4 \* (1 – P (Class1 | X))

= 4 \* P (Calss1 | X) - 4 \* P (Class1 | X) +4 = **4** 🡪 **equation 3**

|  |  |
| --- | --- |
| We choose a1 if:  R (a1 | X) < 4 | We choose a2 if:  R (a2 | X) < 4 |
| From equation 1  - 5P (Calss1 | X) + 5 < 4  - 5 P (Calss1 | X) < -1  P (Calss1 | X) > ,  or equivalently if P(Class2 | X) < | From equation 2  2 + 3 P (Class1 | X) < 4  3 P (Class1 | X) < 2  P (Class1 | X) < ,  or equivalently if P(Class2 | X) > |

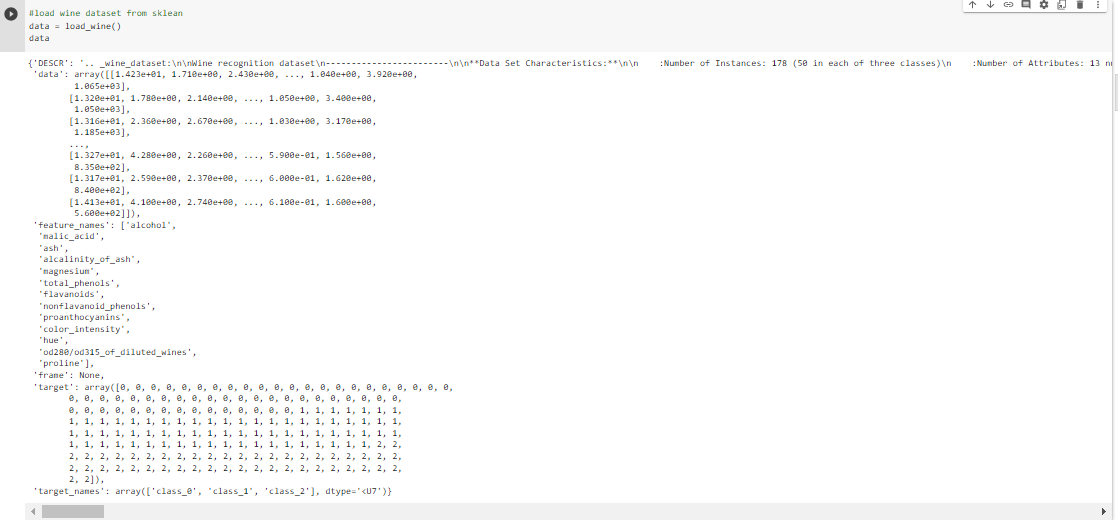
**There is no rejection area, because there is no intersection between**

**P (Calss1 | X) > 1/5 and P (Class1 | X) < 2/3.**

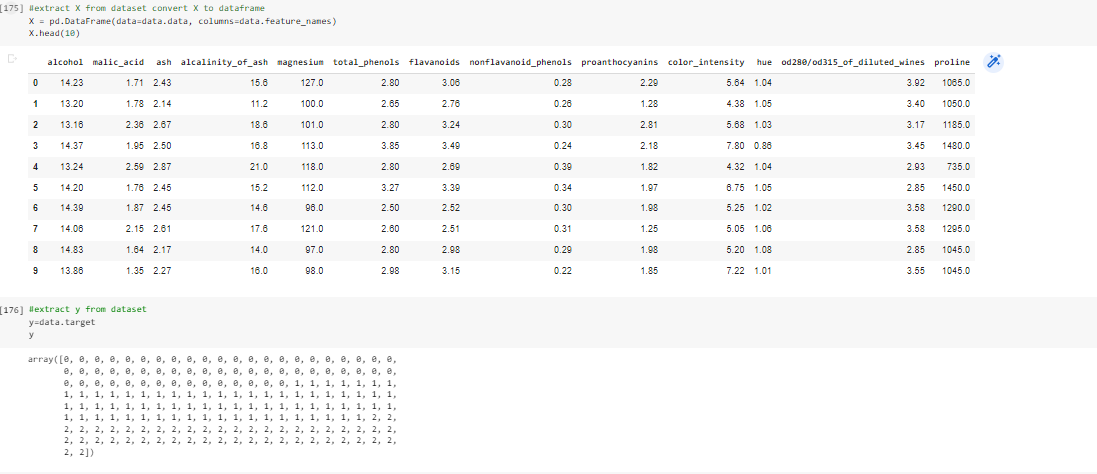
**Part 2: Programming**

1-Naïve Bayes Classifier

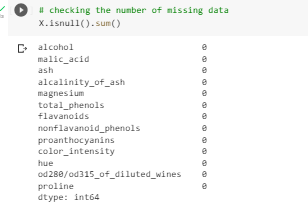
-loading the wine dataset from sklearn library



-There are 13 features and target consist of 3 classes



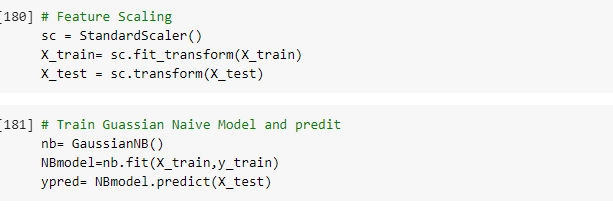
-we checked the nan values and we found there isn’t any nan values



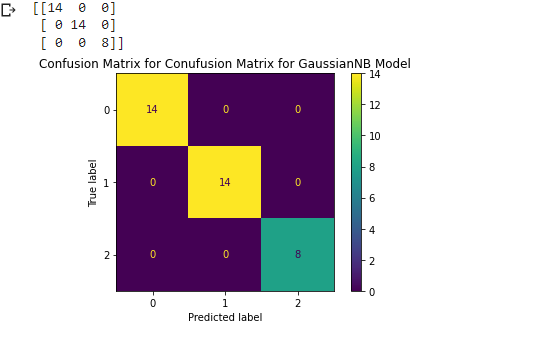
-splitting it into test and train data with 80% for training and 20% for testing

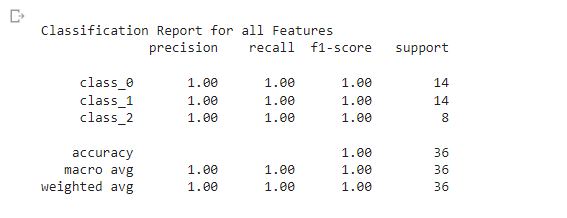


-we applied feature scaling to our data to make all features on the same scale then Gaussian Naïve Bayes Classifier for all features



* The confusion matrix and classification report for the test set for all features, because the small data the model has an over fitting.

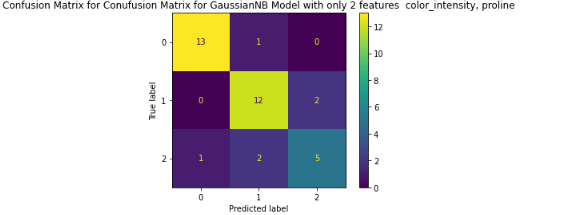


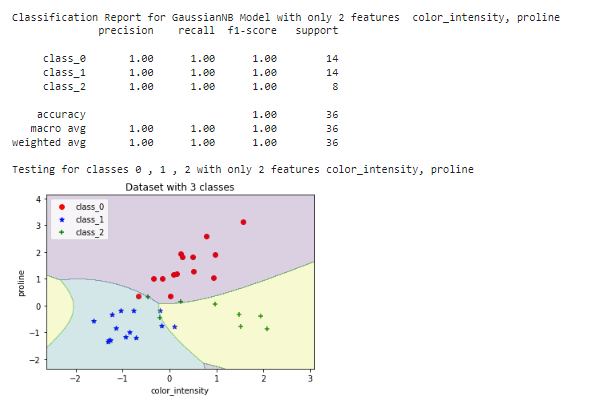


-selecting 2 features to plot the data and its boundries, here we selected the highest correlation 2 features with the target (color\_intensity, proline)



-then we trained new model with only these 2 features and plot our decision boundries and checked the confusion matrix and classification report to check the result and compare our decision bounries and our accuracy

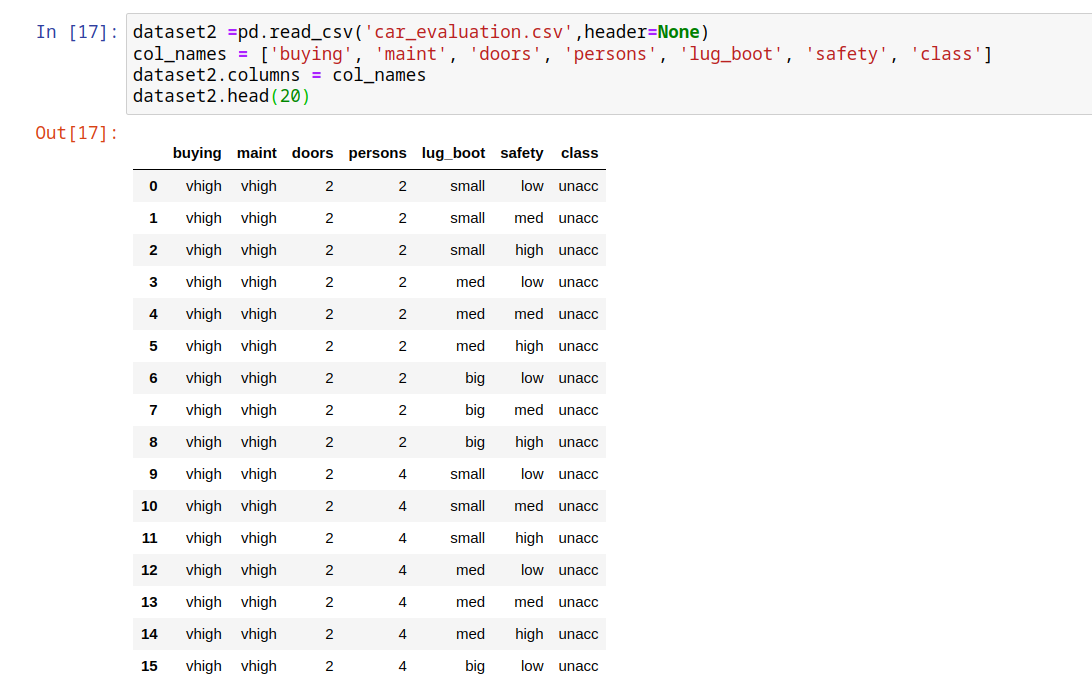


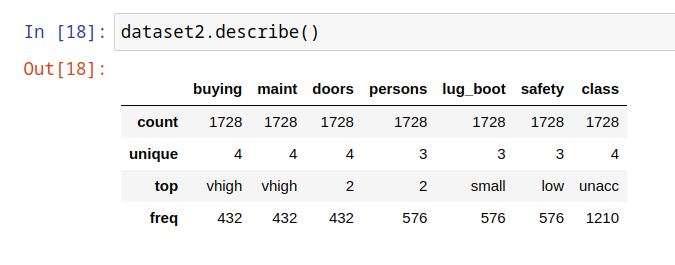


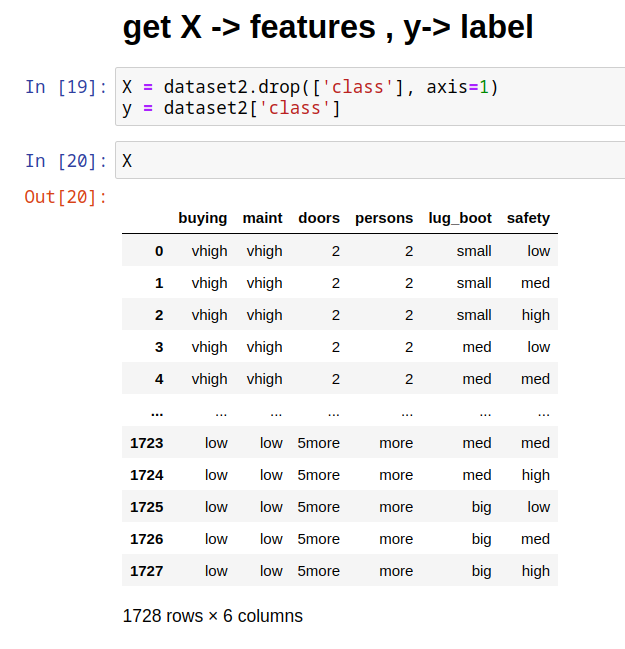
**part 2 : programming - problem 2**

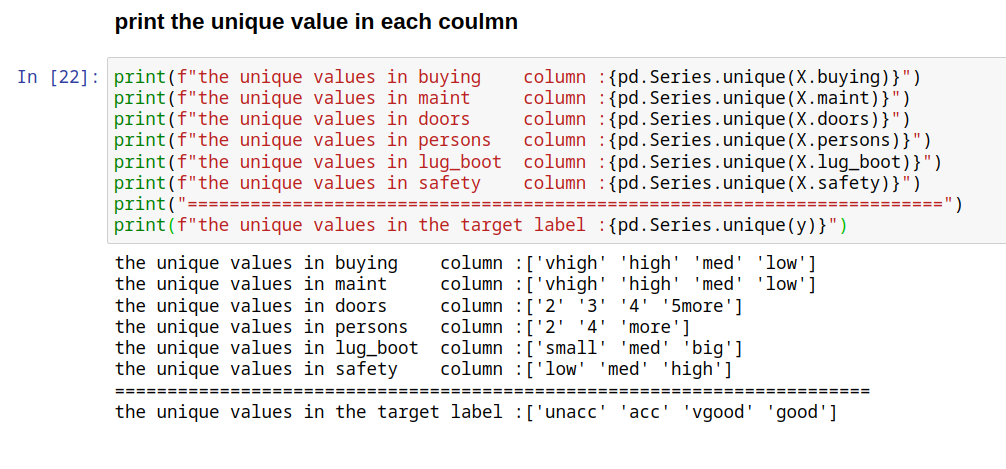
1. load the dataset , shuffle it , split it into training , validation

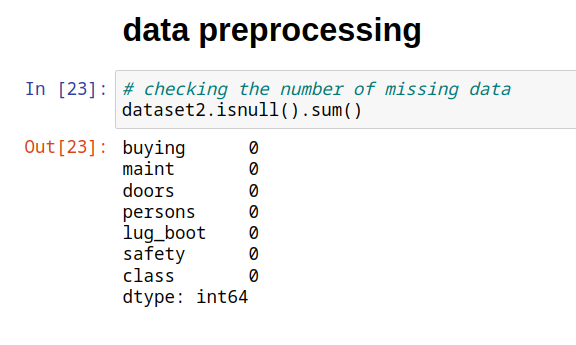
**load the dataset**



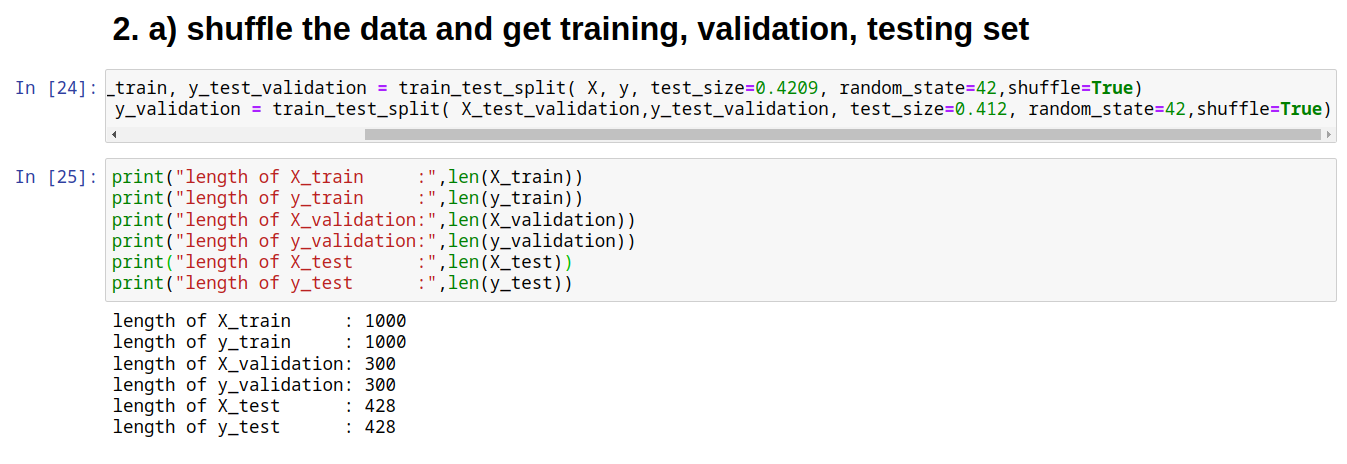


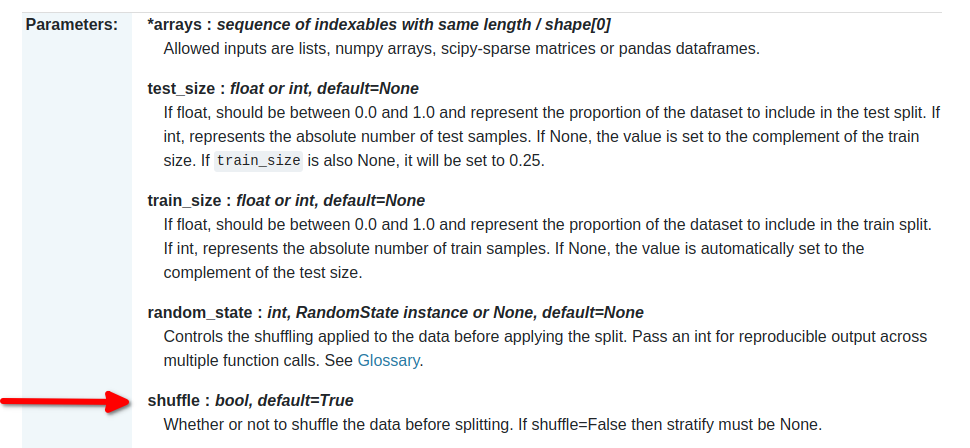




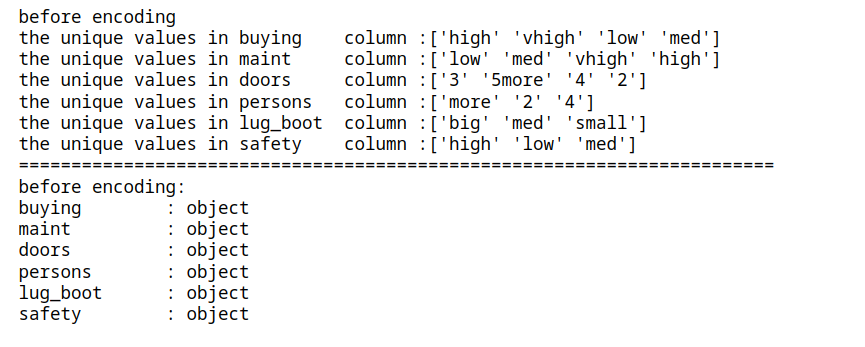


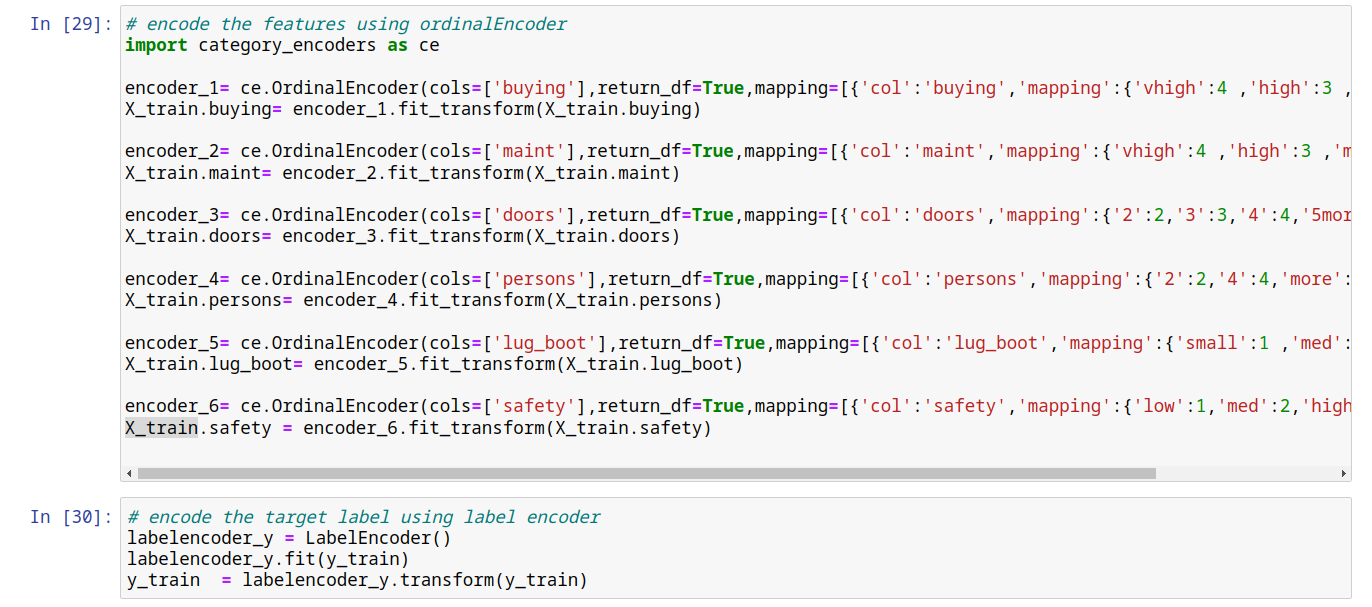
**we used train\_test\_split and set shuffle = True to shuffle the data , and then we split the dataset to training , validation and testing.**

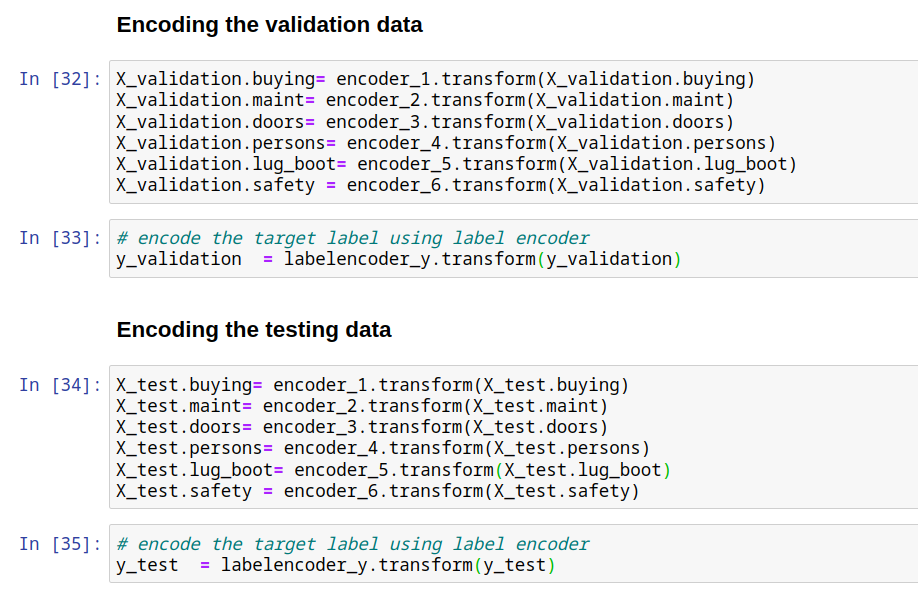


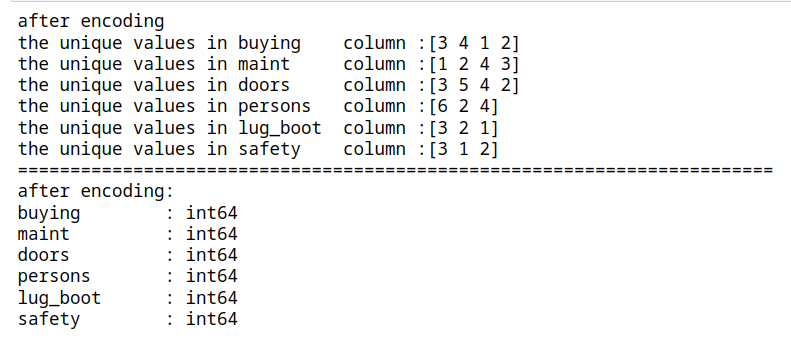


1. transform the string values into numeric

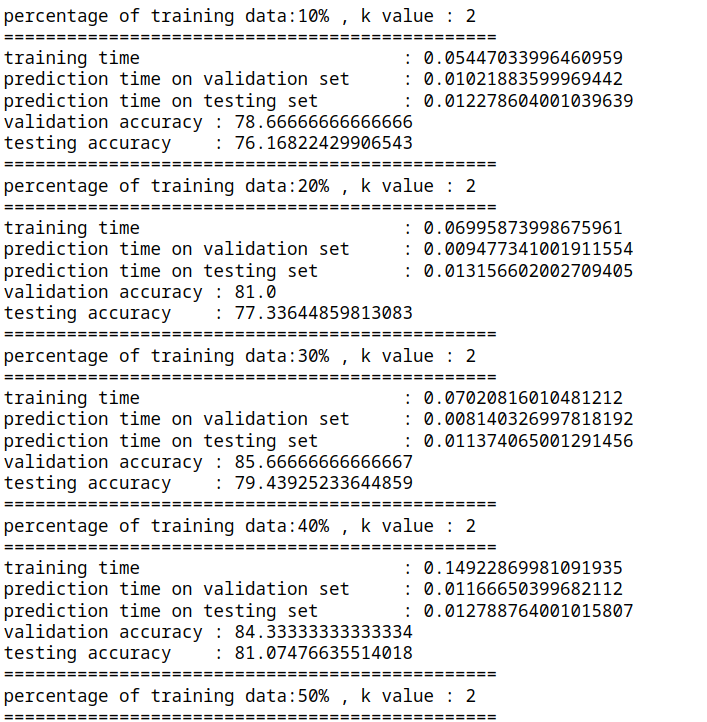


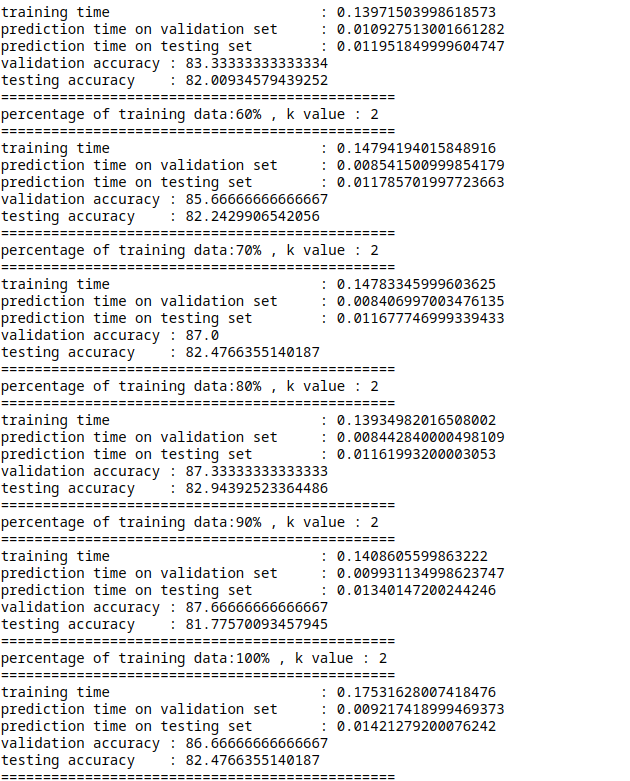


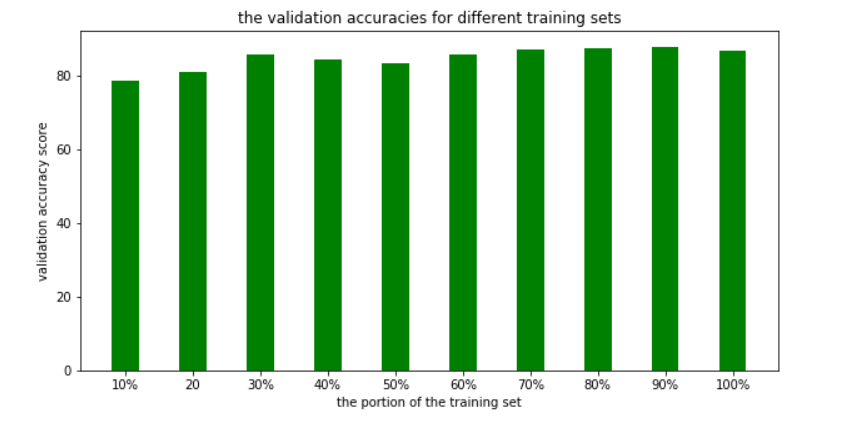


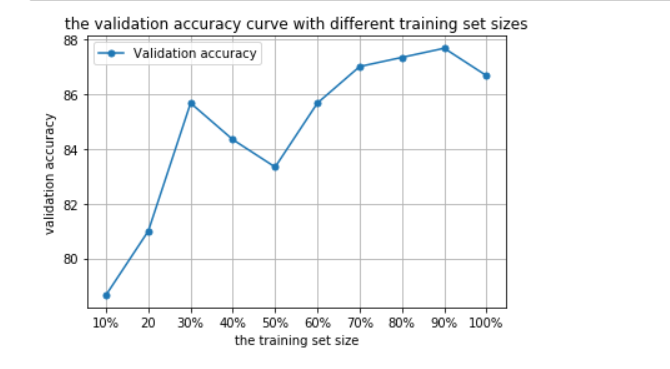


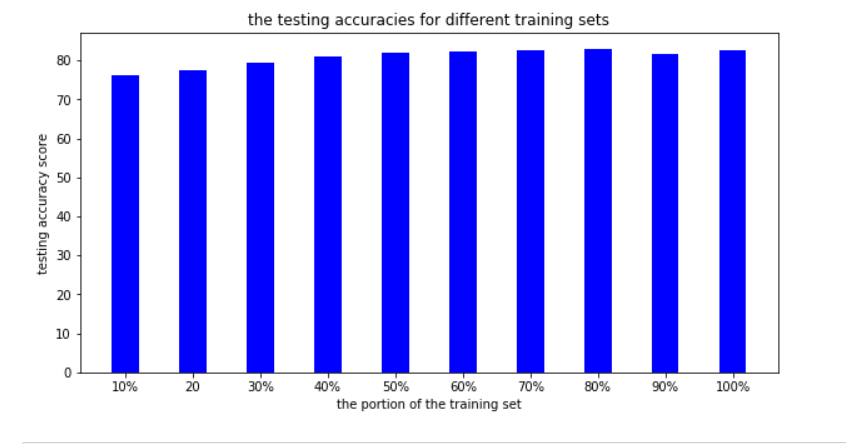
1. the impact of training sample size with a fixed nymber of K =2

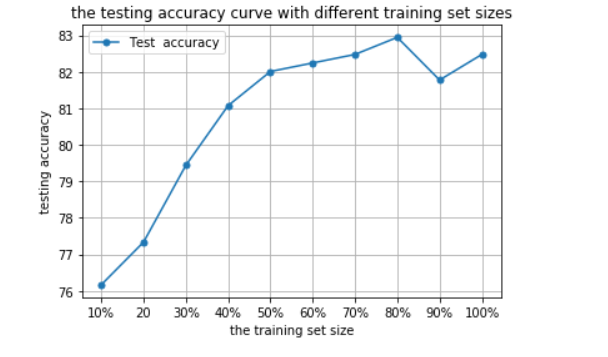


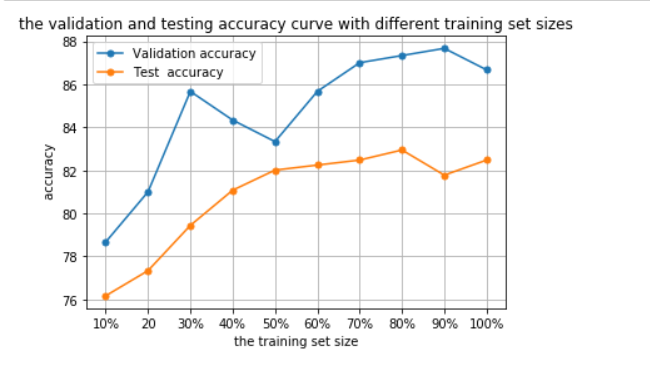




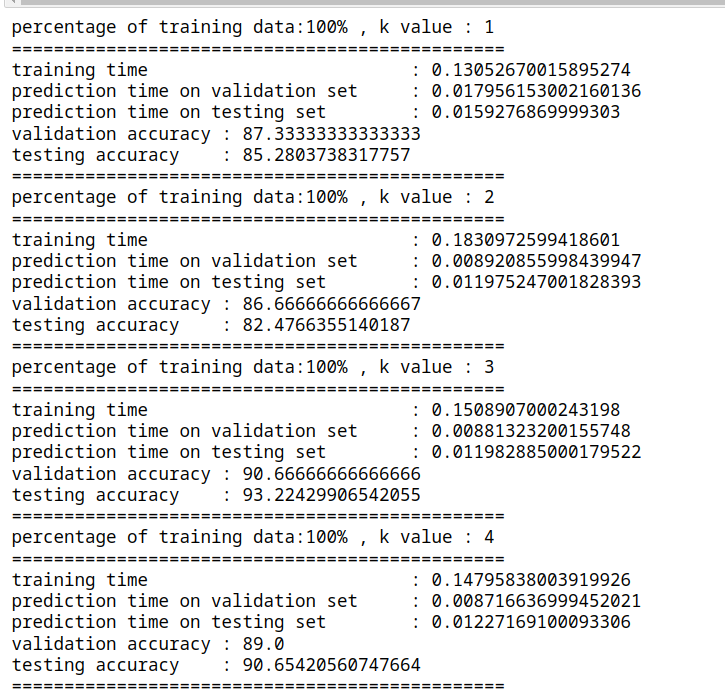


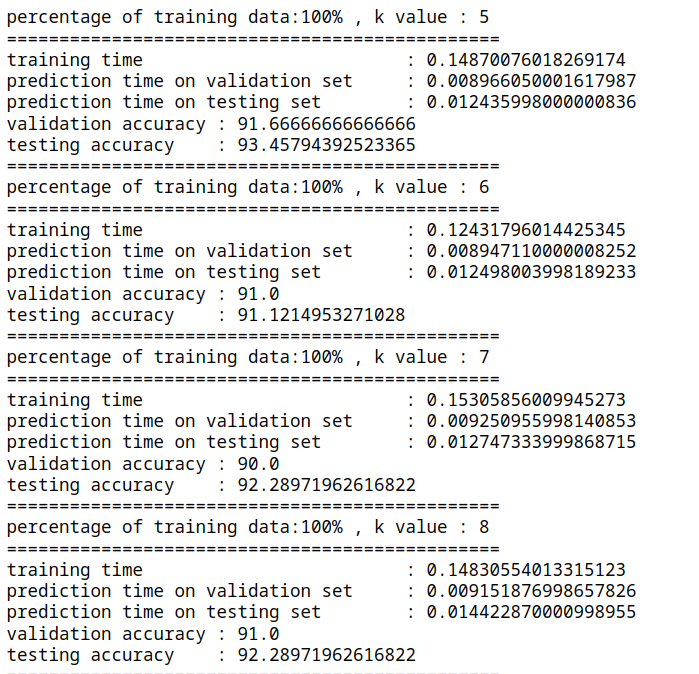


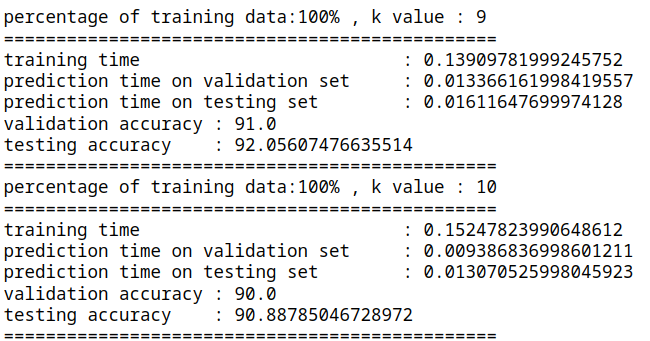


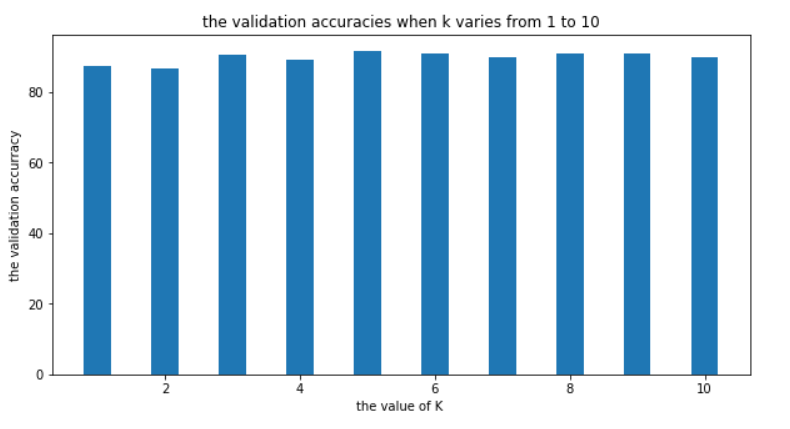


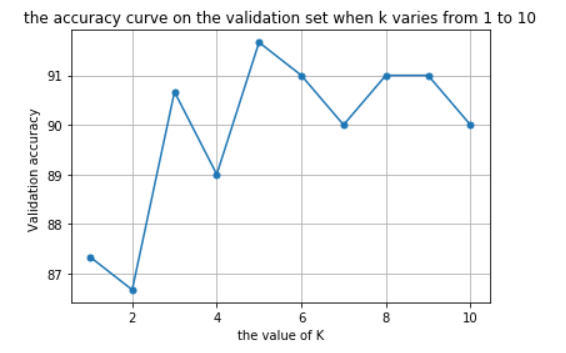
1. the impact of changing the K with fixed number of training sample size = 100%

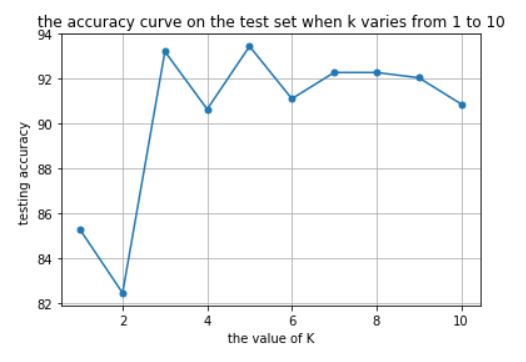


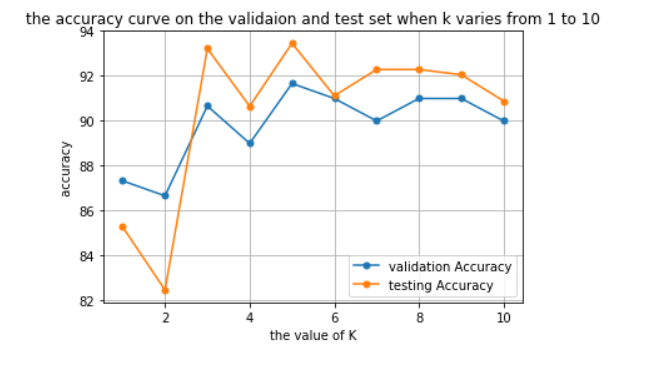


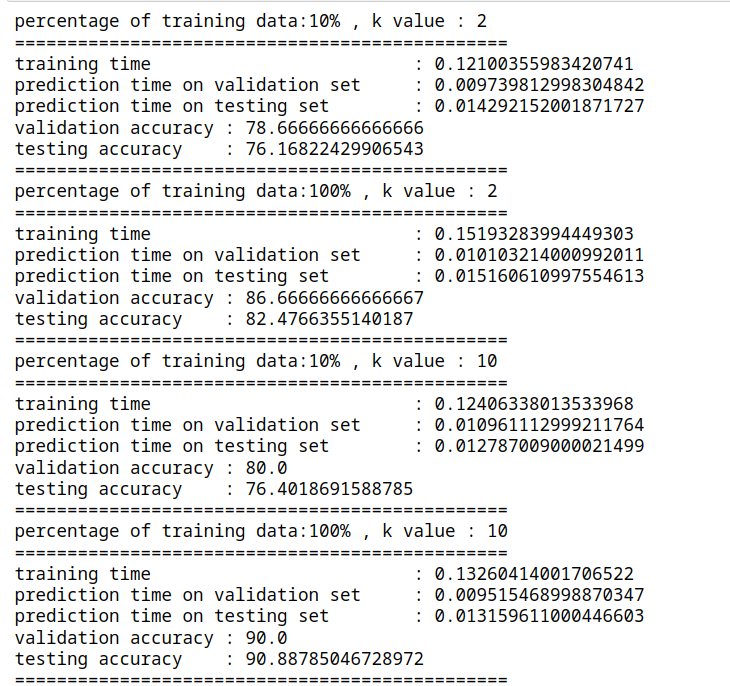


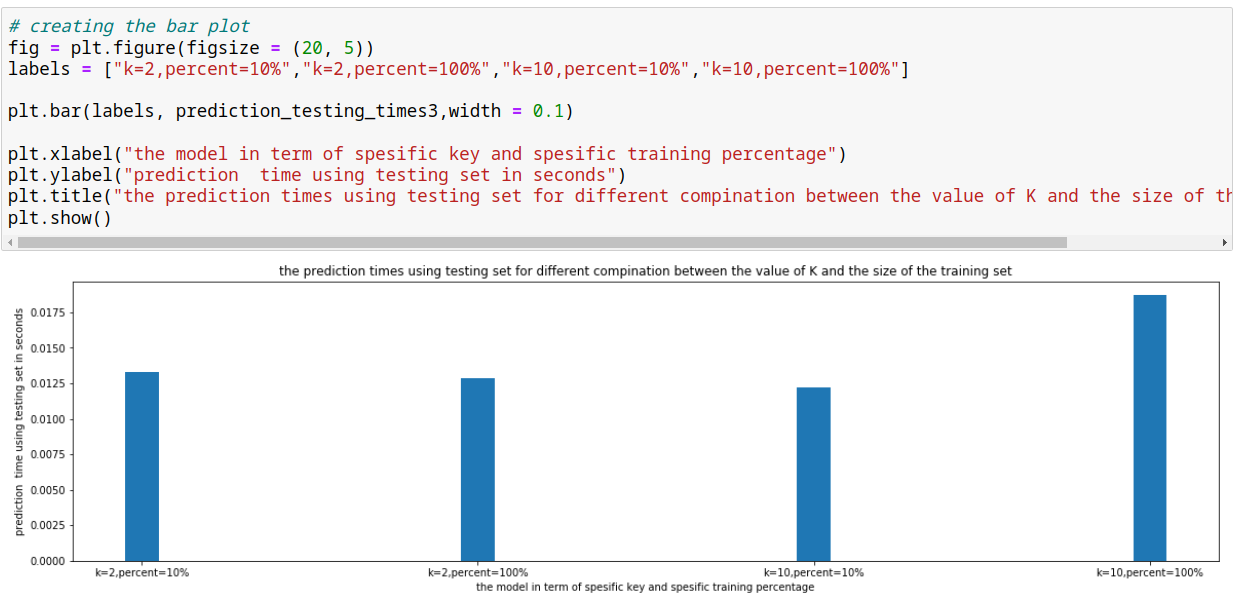


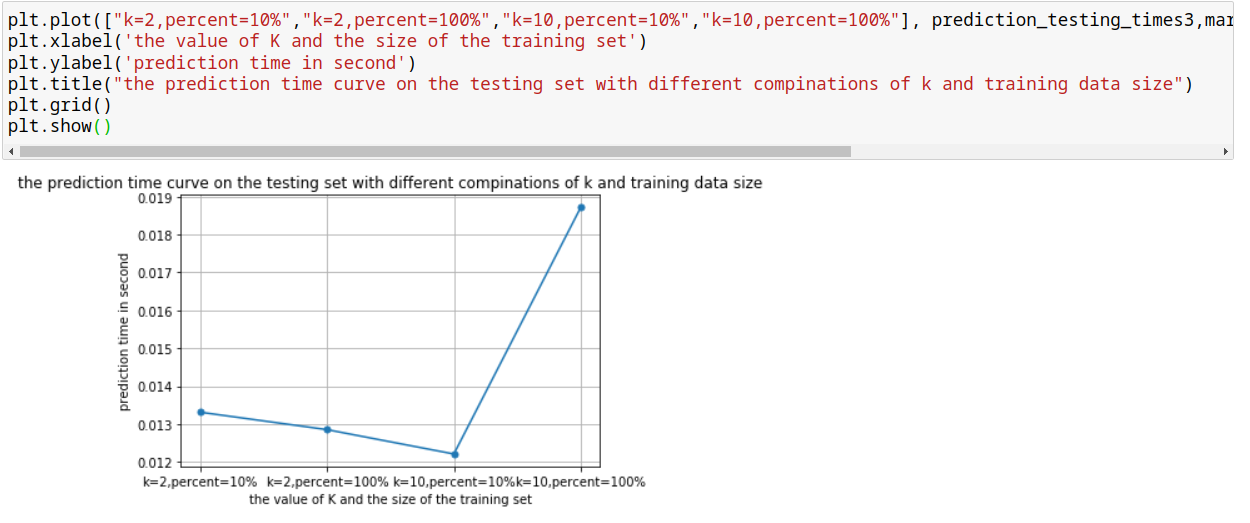






1. the impact of using different compination of k and traning samples 





1. **the conclusion**

**for the experiment, c)**the size of the training set affects the validation accuracy; the more the training set size the more the validation accuracy.

the size of the training set also affects the test accuracy, the testing accuracy increases by increasing the size of the training set.

**for the experiment, d )**the value of K affects the accuracy of the model, we tried different numbers of k with 100% of training set , the best model was the model with k = 5.

**for the experiment, e )**the usage of different combinations between the k value and the size of the training set affects the training time ,and also affects the accuracy which is, the more the training set size the more time the model takes to train.

but for accuracy, the model with higher k values and a higher training dataset is the best.