CIS 419/519: Applied Machine Learning	Spring 2023
Homework 1	
Handed Out: September 15	*Due: September 27*

Name: Jingxuan Bao

PennKey: bjx

PennID: 82897132

Note: This document is a read-only file. To create an editable version click on Menu in the top left corner of your screen and choose the Copy Project option.

1 Multiple Choice & Written Questions

1. (a) i. Error increase. Based on the loss function, if $\lambda_0 \to \infty$, in order to minimize the loss, θ_0 tends to be ignored.
   The decision boundary (linear) should pass through the origin. Based on the figure of x1 and x2 distribution, the two classifications could not be totally split.

   ii. Error increase. Based on the loss function, if $\lambda_1 \to \infty$, in order to minimize the loss, θ_1 tends to be ignored.
   The decision boundary (linear) should perpendicular to x2 axis. Based on the figure of x1 and x2 distribution, the two classifications could be split.

   iii. Error same. Based on the loss function, if $\lambda_2 \to \infty$, in order to minimize the loss, θ_2 tends to be ignored.
   The decision boundary (linear) should perpendicular to x1 axis. Based on the figure of x1 and x2 distribution, the two classifications could not be totally split.

   (b) i. θ_0 tends to be zero because the size of the two classifications is equal.

   ii. θ_0 tends to be less than zero because the size of label 1 increased.

2. (a) The decision boundary should be a perpendicular line passing through the middle of the line across these two points.

   (b) Yes, if there is a dataset like every point holds the same label as its nearest neighbor, there should also be such a boundary.

   (c) There should be a constant model family that predicts the label just based on the majority number of labels in the dataset, because as $k \to \infty$, all the data are each other's k-nearest neighbors.

   (d) Increase k, increase bias, decrease variance. As (c), if $k \to \infty$, the classification result will be decided by the majority label in the dataset. So, the feature of test

data can be ignored. As (b), if $k \to 1$, the label only be decided by its nearest neighbour, so model overfit.

(e) Not compute the majority vote, but set a threshold like t. For example, if we want to make a prediction, do not check the majority (higher than $k/2$?), but check the other fraction k/t. Fro k nearest neighbors, if the true label number more than k/t, predict as true label. And through changing the value of t, we could influence the true positive rate for our KNN model.

3. (a)

$$H(D) = -(\frac{1}{2}log_2(\frac{1}{2}) + \frac{1}{2}log_2(\frac{1}{2})) = 1$$

$$H(D|Weather) = -(\frac{3}{8}log_2(1)+\frac{3}{8}(\frac{1}{3}log_2(\frac{1}{3})+\frac{2}{3}log_2(\frac{2}{3}))+\frac{2}{8}(\frac{1}{3}log_2(\frac{1}{3})+\frac{2}{3}log_2(\frac{2}{3})))$$

$$IG(Weather) = H(D) - H(D|Weather) = 0.77$$

$$H(D|WT) = -(\frac{3}{8}(\frac{2}{3}log_2(\frac{2}{3})+\frac{1}{3}log_2(\frac{1}{3}))+\frac{3}{8}(\frac{1}{3}log_2(\frac{1}{3})+\frac{2}{3}log_2(\frac{2}{3}))+\frac{2}{8}(\frac{1}{2}log_2(\frac{1}{2})+\frac{1}{2}log_2(\frac{1}{2})))$$

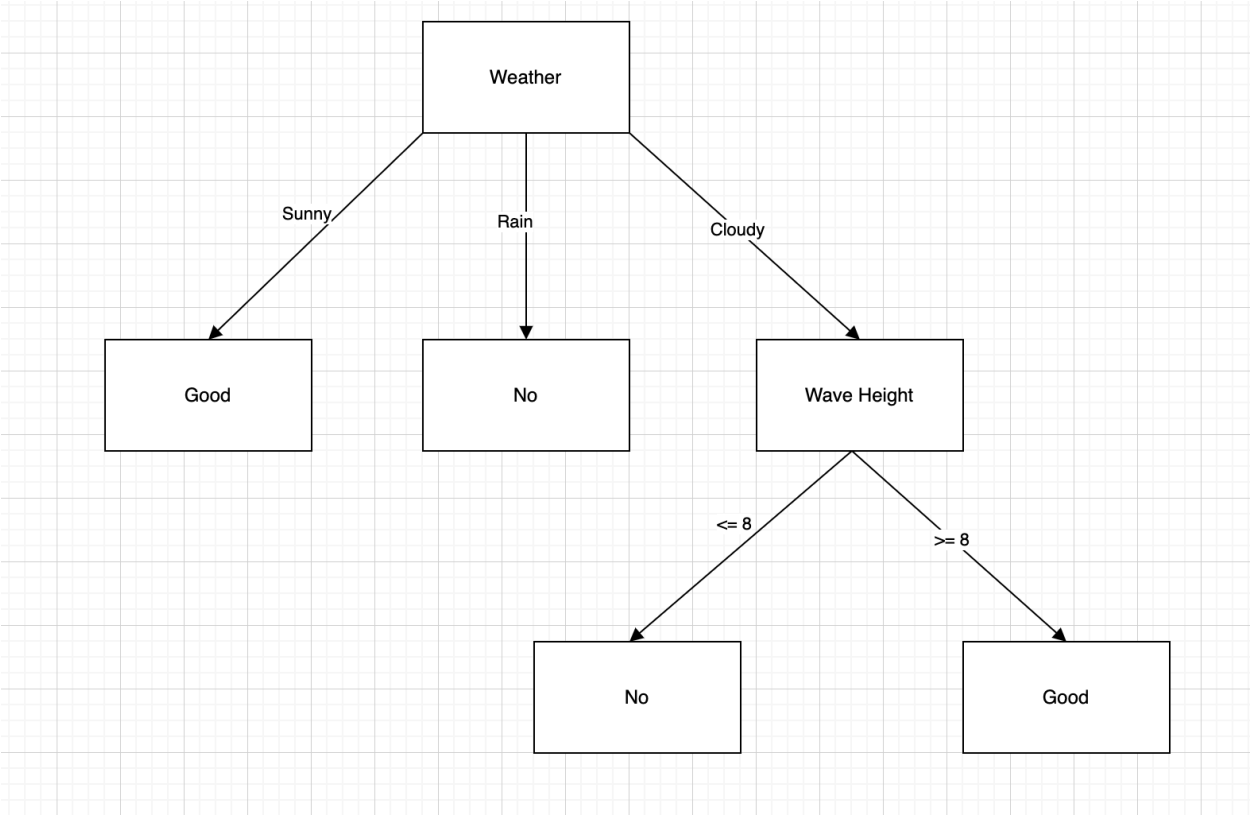
$$IG(WT) = H(D) - H(D|WT) = 0.061$$

$$H(D|WH = -(\frac{5}{8}(\frac{4}{5}log_2(\frac{4}{5}) + \frac{1}{5}log_2(\frac{1}{5})) + \frac{3}{8}log_2(1))$$

$$IG(WH) = H(D) - H(D|WH) = 0.55$$

So $IG(Weather)$ is the highest, we choose Weather as the root node.

(b)



(c) Based on your decision tree graph in (b), yes

(d) First, because the ID3 selects the attributes that maximize the information gain. The locally optimal step can not guarantee the long term optimal. And ID3 may also overfit the training set, so the prediction for unseen data patterns may not optimal.

4. We can set a threshold like t to split data using $x_j >= t$, and $x_j < t$. And for all features, we calculate the new formula of information gain to split data. The model could be also tuned by selecting the different values of t to calculate information gain. We choose the t with the highest value of IG.

5.

$$f_{\hat{\beta}} = \hat{\beta}^T x$$

$$f_{\hat{\beta}} = x^T \hat{\beta} = x^T (X^T X)^{-1} X^T Y$$

$$Y = (y_1, y_2, ..., y_n)^T$$

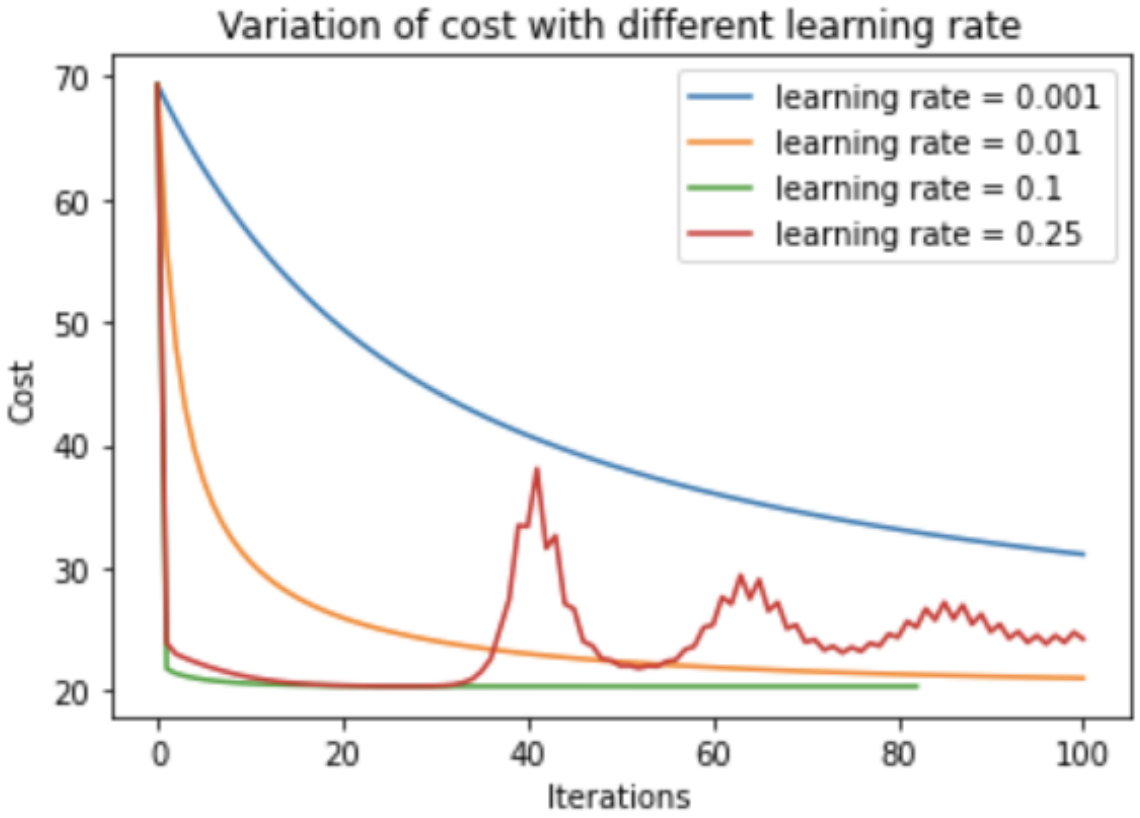
$$F = (X^T X)^{-1} X^T$$

$$f_{\hat{\beta}} = \sum_{i=1}^{n} k_i(x; X) y_i$$

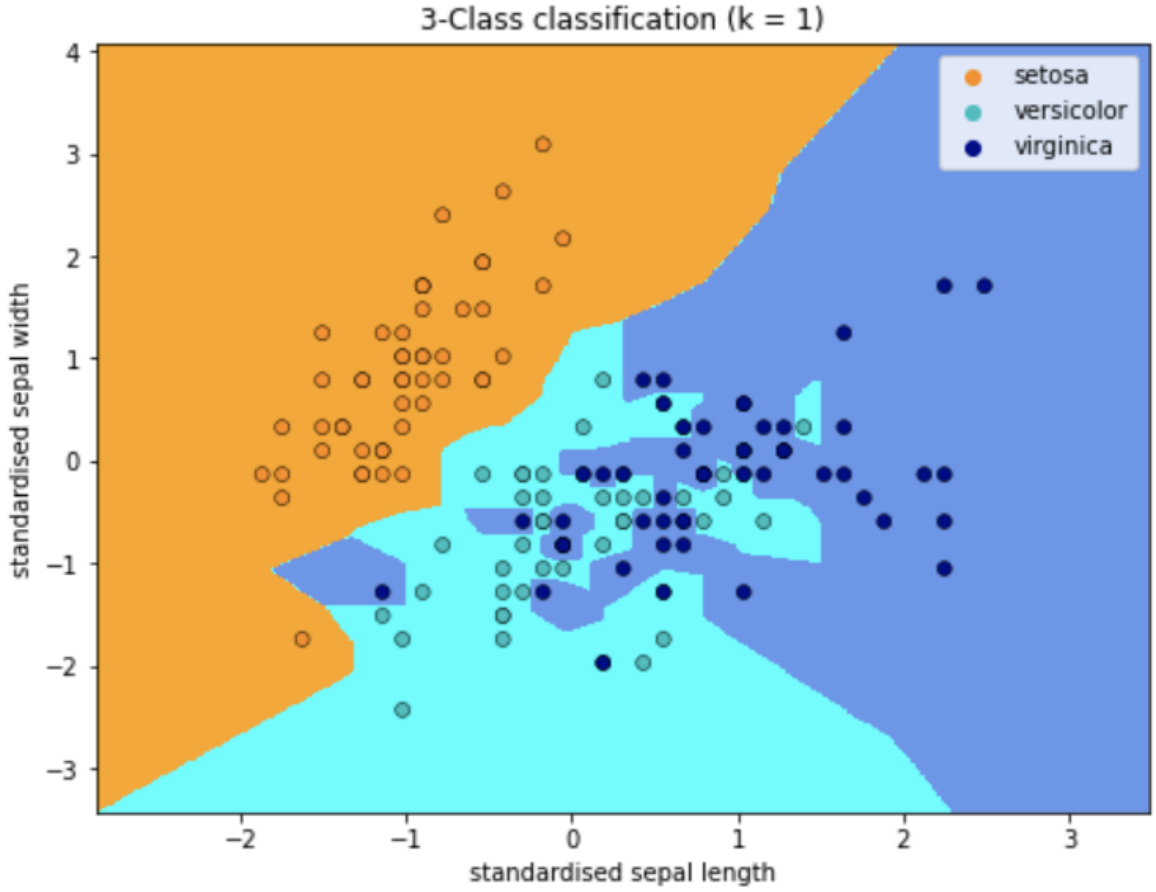
Let $k_i = x_i^T f_i$, where f_i is the ith column of F. If we introduce p_i, which $(n, 1)$ vector where only the ith is one, and others are zero. $f_i = (X^T X)^{-1} X^T p_i$, $k_i = x_i^T (X^T X)^{-1} X^T p_i$

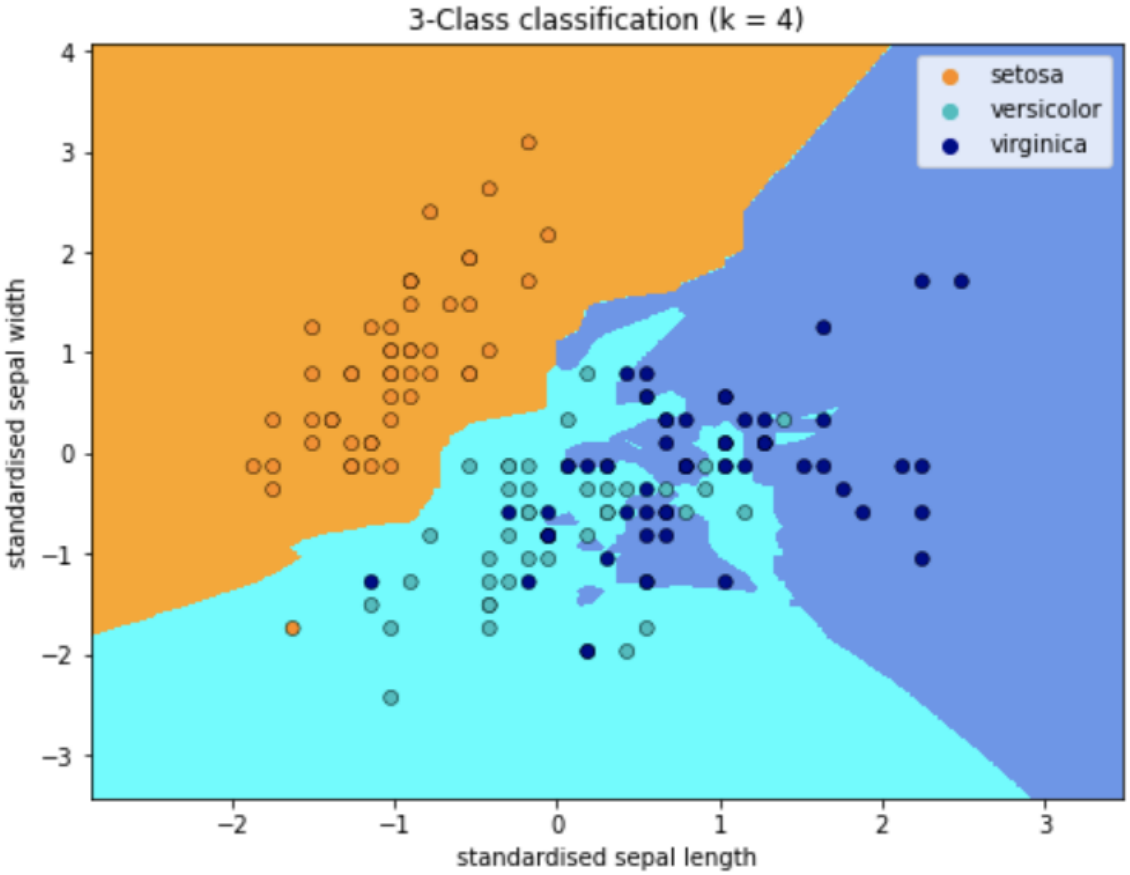
2 Python Programming Questions

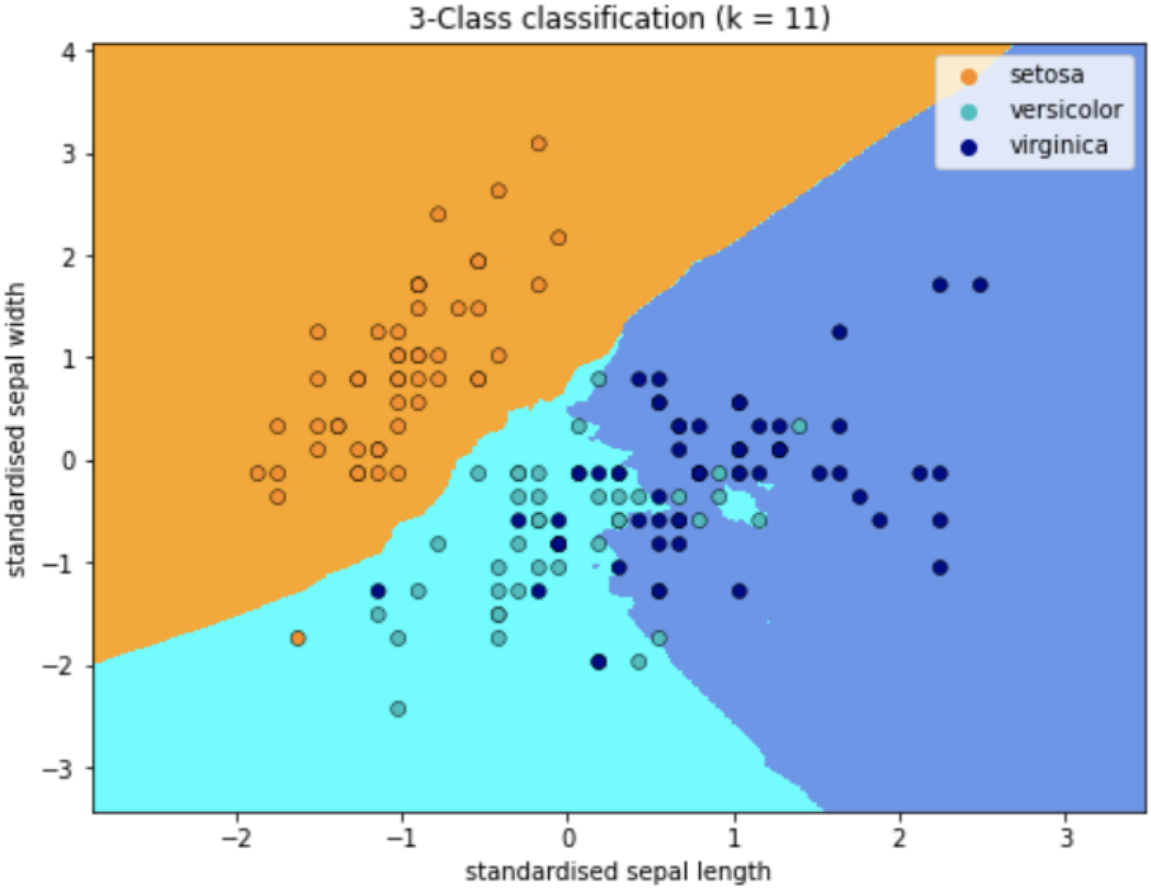
Q1.2:

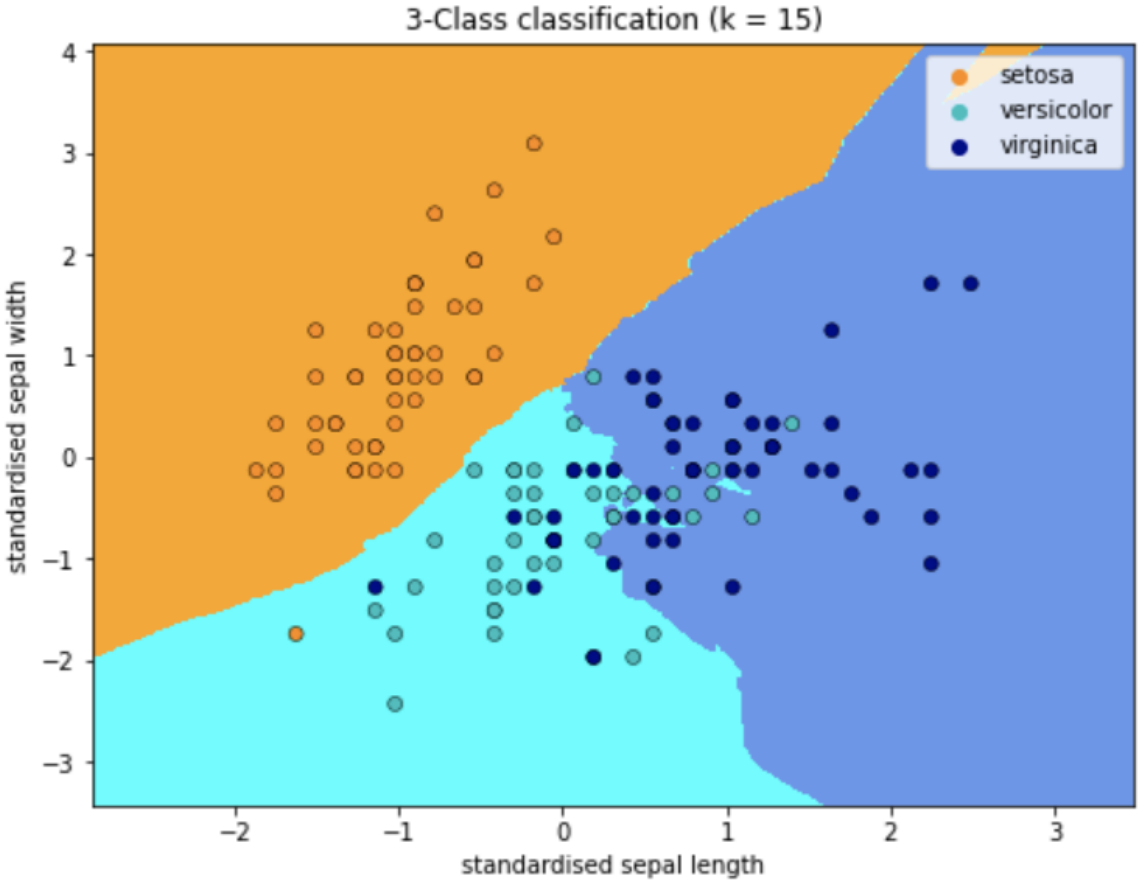


Q2.3:

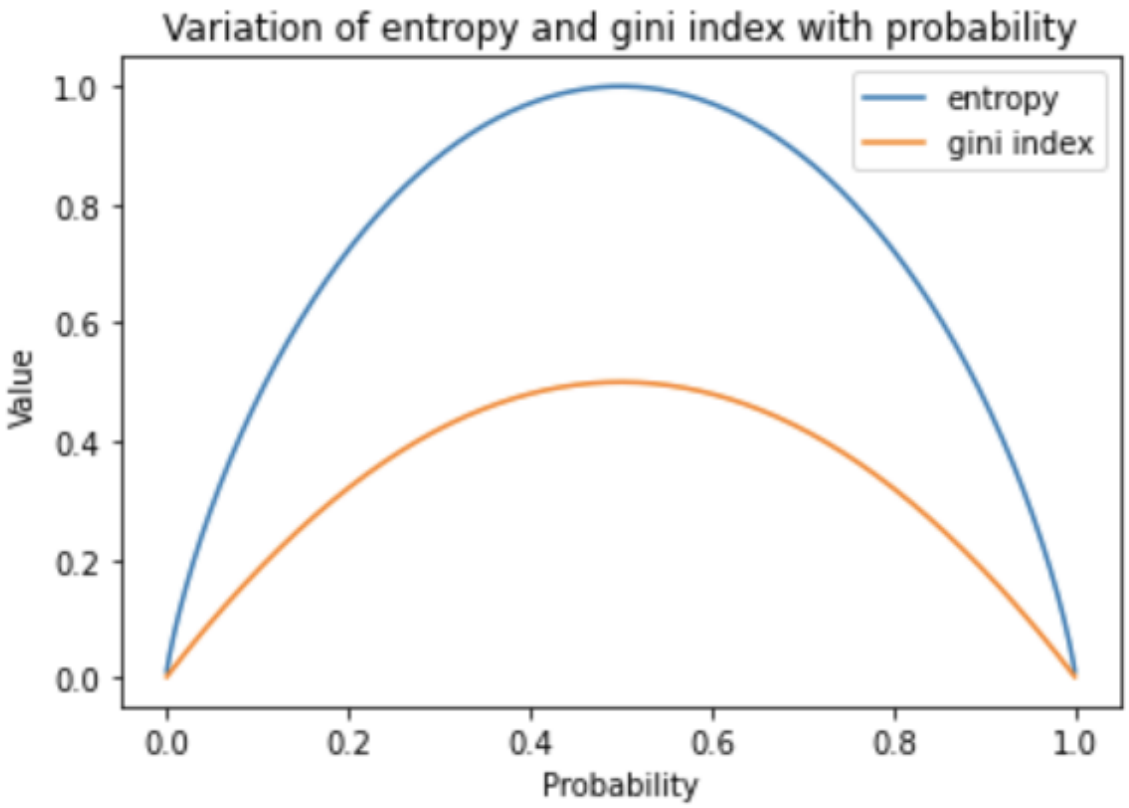

3-Class classification (k = 1)
setosa
versicolor
virginica
standardised sepal width
standardised sepal length
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3


3-Class classification (k = 4)
setosa
versicolor
virginica
standardised sepal width
standardised sepal length
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3


3-Class classification (k = 11)
setosa
versicolor
virginica
standardised sepal width
standardised sepal length
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3


3-Class classification (k = 15)
setosa
versicolor
virginica
standardised sepal width
standardised sepal length
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

Q3.1.2:



Q4.5:

S.No.	Features	Best CCP Alpha	Mean Cross-validation F1 Score	Cross-validation F1 Score Confidence Interval
1	Set 1	0.0005333333333333333	0.3378363962488977	[0.24181988 0.43385291]
2	Set 2	0.0005955704448166759	0.3559483053024465	[0.26915733 0.44273928]
3	Set 3	0.0017006802721088437	0.364028358029422	[0.29792652 0.4301302]

Q5.3:

S.No.	Features	Best Alpha	Mean Cross-validation F1 Score	Cross-validation F1 Score Confidence Interval
1	Set 1	0.005	0.34689300373807785	[-0.0745084 0.76829441]
2	Set 2	0.02	0.31685887416126474	[-0.06894015 0.7026579]
3	Set 3	0.01	0.37546933943138877	[-0.00433926 0.75527794]