1: Download the following datasets

https://www.kaggle.com/pranavpandey2511/tennis-weather
https://archive.ics.uci.edu/ml/datasets/iris

2: Tennis Dataset Description

Describe the datasets, i.e., for each dataset, give these values
1. number of rows
14
2. number of columns
5
3. type of each column, point out the output
Outlook: Categorical Data
Temp: Categorical Data
Humidity: Categorical Data
Windy: Boolean
Play: Boolean
4. values each column can take, in case of categorical
Outlook: [Sunny, Overcast, Rainy]
Temp: [Cool, Hot, Mild]
Humidity: [High, Normal]
Windy(not categorical): [TRUE, FALSE]
Play(not categorical): [yes, no]
5. number of each value in each column, in case of categorical column
Outlook:
Sunny: 5
Overcast: 4
Rainy: 5
Temp:
hot: 4
mild: 6
cool: 4
Humidity:
Normal: 7
High: 7
6. mean and std dev of each column that is continuous
There is no column that is continuous in this data set
7. plot the frequency graph of each column that is continuous
There is no column that is continuous in this data set
8. if the output is categorical, split the dataset according to the output value and do 5,6,7 for each value of the output
OUTLOOK:
Sunny

Temp	Humidity	Windy	Play
hot	high	FALSE	no
hot	high	TRUE	no
mild	high	FALSE	no
cool	normal	FALSE	yes
mild	normal	TRUE	yes

5. Temp:
hot: 2
mild: 2
cool: 1
Humidity:
High: 3
Normal: 2
6 & 7 are not applicable to this data set

Overcast:

Temp	Humidity	Windy	Play
hot	high	FALSE	yes
cool	normal	TRUE	yes
mild	high	TRUE	yes
hot	normal	FALSE	yes

5. Temp:
hot: 2
mild: 1
cool: 1
Humidity:
High: 2
Normal: 2
6 & 7 are not applicable to this data set

Rainy:

Temp	Humidity	Windy	Play
mild	high	FALSE	yes
cool	normal	FALSE	yes
cool	normal	TRUE	no
mild	normal	FALSE	yes
mild	high	TRUE	no

5. Temp:
hot: 0
mild: 3
cool: 2
Humidity:
High: 2
Normal: 3
6 & 7 are not applicable to this data set

TEMP:
Hot:

Outlook	Humidity	Windy	Play
sunny	high	FALSE	no
sunny	high	TRUE	no
overcast	high	FALSE	yes
overcast	normal	FALSE	yes

5. Outlook:
sunny: 2
overcast: 2
rainy: 0
Humidity:
High: 3
Normal: 1
6 & 7 are not applicable to this data set

Mild:

Outlook	Humidity	Windy	Play
rainy	high	FALSE	yes
sunny	high	FALSE	no
rainy	normal	FALSE	yes
sunny	normal	TRUE	yes
overcast	high	TRUE	yes
rainy	high	TRUE	no

5. Outlook:
sunny: 2
overcast: 1
rainy: 3
Humidity:
High: 4
Normal: 2
6 & 7 are not applicable to this data set

Cool:

Outlook	Humidity	Windy	Play
rainy	normal	FALSE	yes
rainy	normal	TRUE	no
overcast	normal	TRUE	yes
sunny	normal	FALSE	yes

5. Outlook:
sunny: 1
overcast: 1
rainy: 2
Humidity:
High: 0
Normal: 4
6 & 7 are not applicable to this data set

HUMIDITY:
High:

Outlook	Temp	Windy	Play
sunny	hot	FALSE	no
sunny	hot	TRUE	no
overcast	hot	FALSE	yes
rainy	mild	FALSE	yes
sunny	mild	FALSE	no
overcast	mild	TRUE	yes
rainy	mild	TRUE	no

5. Outlook:
Sunny: 3
Overcast: 2
Rainy: 2
Temp:
hot: 3
mild: 4
cool: 0
6 & 7 are not applicable to this data set

Normal:

Outlook	Temp	Windy	Play
rainy	cool	FALSE	yes
rainy	cool	TRUE	no
overcast	cool	TRUE	yes
sunny	cool	FALSE	yes
rainy	mild	FALSE	yes
sunny	mild	TRUE	yes
overcast	hot	FALSE	yes

5. Outlook:
Sunny: 2
Overcast: 2
Rainy: 3
Temp:
hot: 1
mild: 2
cool: 3
6 & 7 are not applicable to this data set

2: Iris Dataset Description

Describe the datasets, i.e., for each dataset, give these values

1. number of rows

150

2. number of columns

5

3. type of each column, point out the output

Columns 1 through 4 are continuous, column 5 is categorical.

4. values each column can take, in case of categorical

Column 5 is the only categorical column. It can take 3 different values: Iris-setosa, Iris-versicolor, Iris-virginica.

5. number of each value in each column, in case of categorical column

There are 50 of each column 5 categorical value.

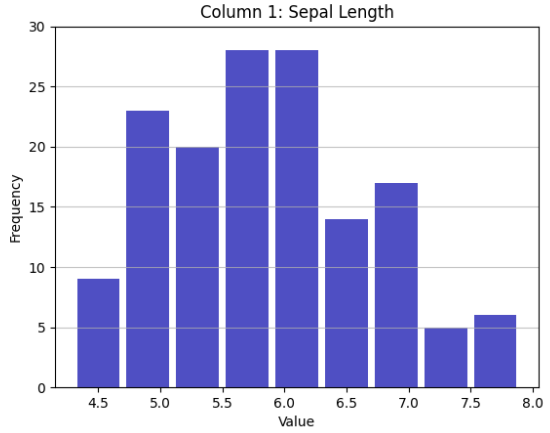
6. mean and std dev of each column that is continuous

Standard Deviation:
0.825
0.432
1.759
0.761

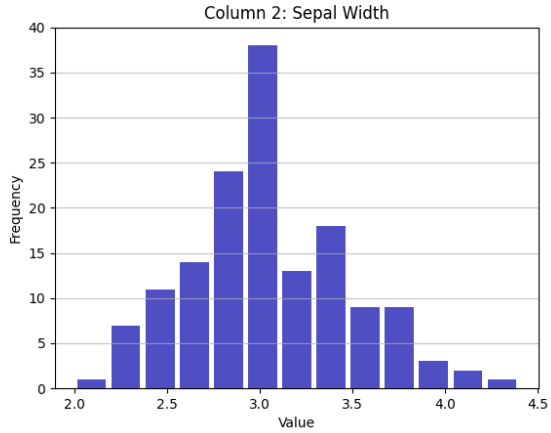
Mean:
5.843
3.054
3.759
1.199

7. plot the frequency graph of each column that is continuous

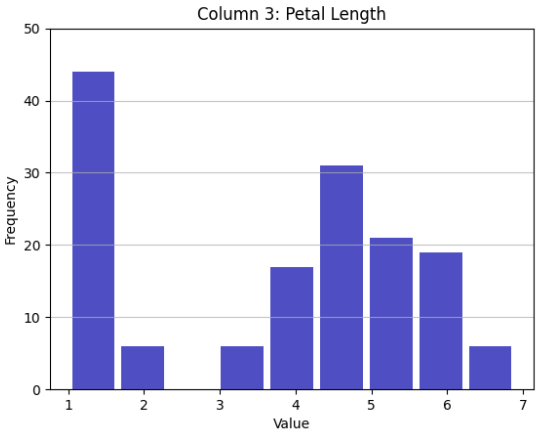
Column 1:



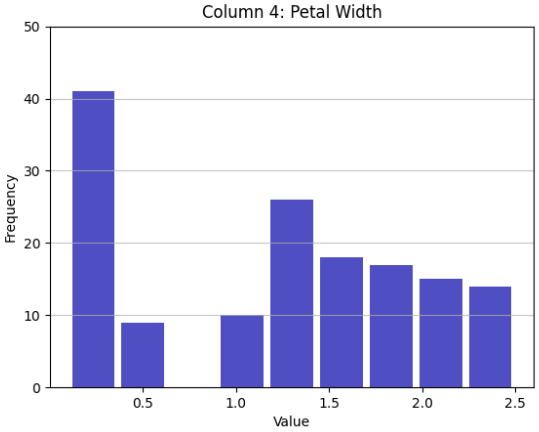
Column 2:



Column 3:



Column 4:



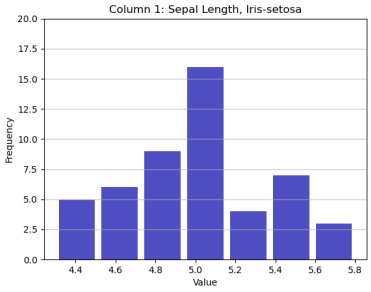
8. if the output is categorical, split the dataset according to the output value and do 5,6,7 for each value of the output
Part 5 is not applicable for a single categorical column.

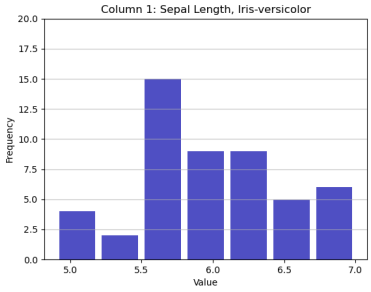
Part 6:
Standard Deviation:
St. Dev. of Column 1: Sepal Length, Iris-setosa: 0.349
St. Dev. of Column 1: Sepal Length, Iris-versicolor: 0.511
St. Dev. of Column 1: Sepal Length, Iris-virginica: 0.629
St. Dev. of Column 2: Sepal Width, Iris-setosa: 0.377
St. Dev. of Column 2: Sepal Width, Iris-versicolor: 0.311
St. Dev. of Column 2: Sepal Width, Iris-virginica: 0.319
St. Dev. of Column 3: Petal Length, Iris-setosa: 0.172
St. Dev. of Column 3: Petal Length, Iris-versicolor: 0.465
St. Dev. of Column 3: Petal Length, Iris-virginica: 0.546
St. Dev. of Column 4: Petal Width, Iris-setosa: 0.106
St. Dev. of Column 4: Petal Width, Iris-versicolor: 0.196

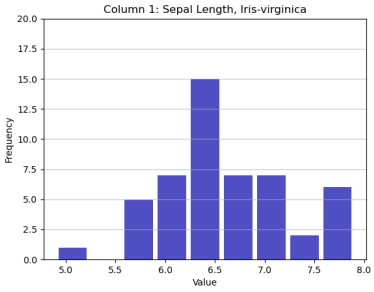
St. Dev. of Column 4: Petal Width, Iris-virginica: 0.272

Mean:
Mean of Column 1: Sepal Length, Iris-setosa: 5.006
Mean of Column 1: Sepal Length, Iris-versicolor: 5.936
Mean of Column 1: Sepal Length, Iris-virginica: 6.588
Mean of Column 2: Sepal Width, Iris-setosa: 3.418
Mean of Column 2: Sepal Width, Iris-versicolor: 2.77
Mean of Column 2: Sepal Width, Iris-virginica: 2.974
Mean of Column 3: Petal Length, Iris-setosa: 1.464
Mean of Column 3: Petal Length, Iris-versicolor: 4.26
Mean of Column 3: Petal Length, Iris-virginica: 5.552
Mean of Column 4: Petal Width, Iris-setosa: 0.244
Mean of Column 4: Petal Width, Iris-versicolor: 1.326
Mean of Column 4: Petal Width, Iris-virginica: 2.026

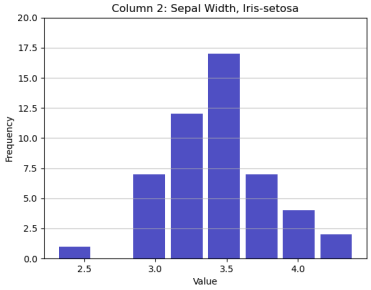
Part 7: Column 1, separated:

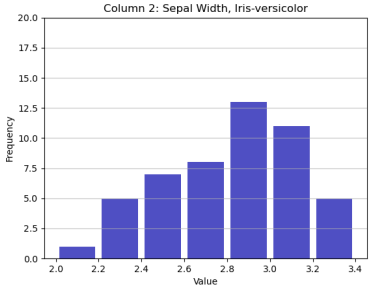


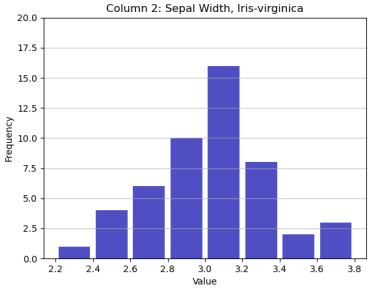




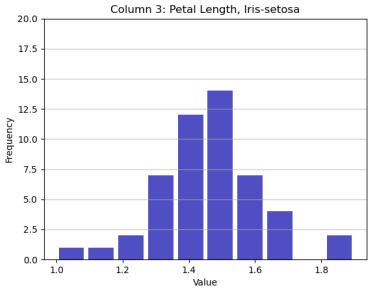
Column 2, separated:

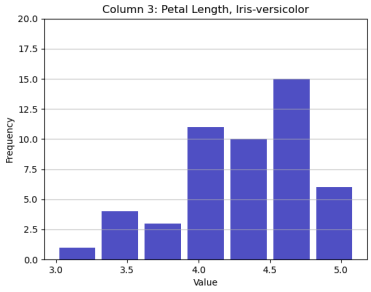


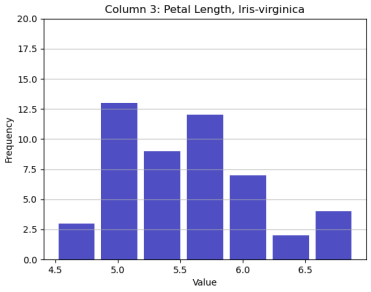




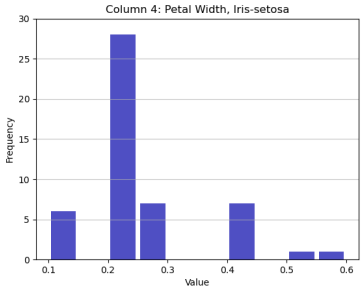
Column 3, separated:

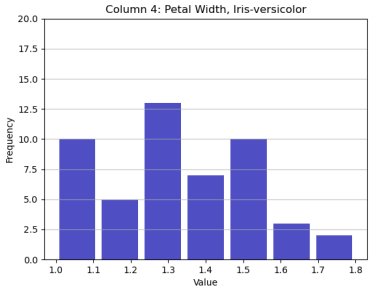


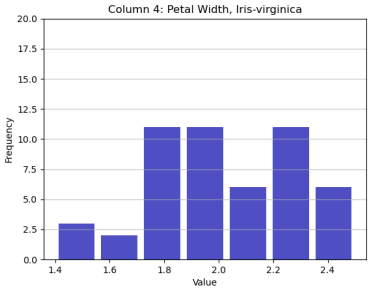




Column 4, separated:







3: Make the table needed for naive bayesian for each data set.

Iris Dataset:
Due to the uniformity of iris.data, all data points are out of 50.

		Setosa	Versicolor	Virginica
Q1	SL	28	3	1
	SW	1	21	11
	PL	37	0	0
	PW	34	0	0
Q2	SL	21	18	2
	SW	1	13	10
	PL	13	25	0
	PW	16	15	0
Q3	SL	1	18	16
	SW	16	14	21
	PL	0	24	9
	PW	0	34	5
Q4	SL	0	11	31
	SW	32	2	8
	PL	0	1	41
	PW	0	1	45

Q1-Q4 were calculated with inclusive lower bound and exclusive upper bound. Quartiles used for respective columns:

	SL	SW	PL	PW
Q1	5.1	2.8	1.6	0.3
Q2	5.8	3.0	4.35	1.3
Q3	6.4	3.3	5.1	1.8

Tennis Dataset:

		Play Tennis	
		Yes	No
	Features	9/14 = 0.642	5/14 = 0.357
Outlook	Sunny	2/9 = 0.222	3/5 = 0.6
	Overcast	4/9 = 0.444	0/5 = 0
	Rainy	3/9 = 0.333	2/5 = 0.4
Humidity	High	3/9 = 0.333	4/5 = 0.8
	Normal	6/9 = 0.666	1/5 = 0.2
Temp	Hot	2/9 = 0.222	2/5 = 0.4
	Mild	4/9 = 0.444	2/5 = 0.4
	Cool	3/9 = 0.333	1/5 = 0.2
Wind	TRUE	3/9 = 0.333	3/5 = 0.6
	FALSE	6/9 = 0.666	2/5 = 0.4

4: For the iris data set, give the formula and compute the naive Bayes' output for $< 5.5, 3.0, 1.9, 0.4 >$

Assuming events are independent:
Since the denominator is proportional, it can be left out.

P(Setosa $| < 5.5, 3.0, 1.9, 0.4 >$):
$P(Q2 - SL|Setosa)P(Q3 - SW|Setosa)P(Q2 - PL|Setosa)P(Q2 - SW|Setosa)P(Setosa)$
$(21/50)(16/50)(13/50)(16/50)(50/150) = 0.0037$

P(Versicolor $| < 5.5, 3.0, 1.9, 0.4 >$):
$P(Q2 - SL|Versicolor)P(Q3 - SW|Versicolor)P(Q2 - PL|Versicolor)P(Q2 - SW|Versicolor)P(Versicolor)$
$(18/50)(14/50)(25/50)(15/50)(50/150) = 0.0050$

P(Virginica $| < 5.5, 3.0, 1.9, 0.4 >$):

$P(Q2 - SL|Virginica)P(Q3 - SW|Virginica)P(Q2 - PL|Virginica)P(Q2 - SW|Virginica)P(Virginica)$
$(2/50)(21/50)(0/50)(0/50)(50/150) = 0.0000$

Output: Versicolor

5: For the weather-tennis data set, give the formula and compute the naive Bayes output for <sunny, mild, normal>

Assuming all features are independent:
P(play = yes | outlook = sunny, temp = mild, humidity = mild)
$= \frac{P(sunny|yes)P(mild|yes)P(normal|yes)P(yes)}{P(sunny)P(mild)P(normal)} = \frac{(0.642)(0.222)(0.444)(0.666)}{(0.357)(0.5)(0.4285)} = 0.552$ or 55%
AND
P(play = no | outlook = sunny, temp = mild, humidity = mild)
$= \frac{P(sunny|no)P(mild|no)P(normal|no)P(no)}{P(sunny)P(mild)P(normal)} = \frac{(0.357)(0.6)(0.2)(0.4)}{(0.357)(0.5)(0.4285)} = 0.224$ or 22%
Ouput: yes