CS 6316 HW 2
Bryan Chen bc2vf

1. Problem 2.2

a) There is a high correlation between a person's age, gender, and especially weight-to-height ratio on the remaining life expectancy, as older people have lower remaining life expectancy. This model certainly has useful practical value as people wish to predict their remaining life expectancy, so it is appropriate to use predictive learning in this case.

b) It is common to consider that a low family income may lead to more problems in life because the need of all family members can hardly be all met. The marriage length also has a big influence. Statistics show that a couple with about 8 years' marriage is more likely to have divorce. It is appropriate to use predictive learning in this case.

c) A person's gender could be a factor that influences the remaining life expectancy, but it is not the very important one. A cell phone number has little correlation on remaining life expectancy. Thus, predictive learning is not appropriate to be used in this case.

d) It is very hard to predict a stock index's closing price very accurately since the price is influenced by a number of factors, and this is hard to be concluded solely by looking at the daily changing rate. The next day closing price may be independent of the last-day closing price. Thus, it is not appropriate to use predictive learning in this case, but it would be appropriate if there are more relevant inputs.

e) Years of data of recent stock market trend is an important reference when predicting the following price change of the stock. Many economics try to predict the stock market so this model has very high practical value. So, it is appropriate to use predictive learning to estimate a data-analytic model in this case.

f) The older a patient is, the higher the possibility for them to be diagnosed the presence of Alzheimer's disease, and the brain scan information is one of the most important factors that influence the possibility of Alzheimer's disease, since psychological condition influences a lot on patient's physiological condition. The dependency can be estimated from data and this estimated model has useful practical value, so it is appropriate to use predictive learning in this case.

2. Problem 2.7

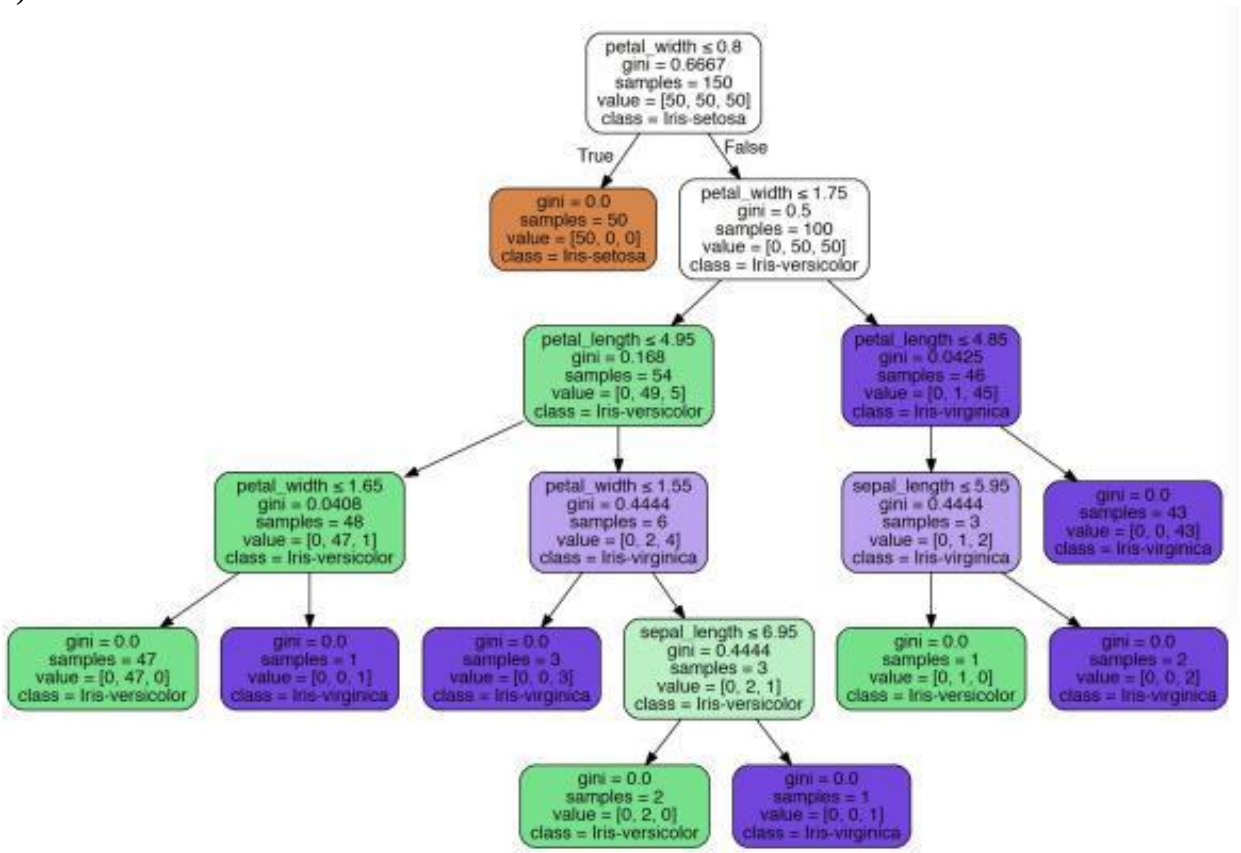
The optimal number of neighbors is 17. The cross validation score of the optimal number is 0.68627 and the mean square error is 0.31373. The accuracy of the classifier is 68%. Obesity rate is not a good factor to predict the election resul and other factors should be taken into consideration for a good model.

State	obesity rate	2004	Prediction	[Ground Truth] 2000
Alabama	0.301	R	R	R
Alaska	0.273	R	R	R
Arizona	0.233	R	D	R
Arkansas	0.281	R	R	R

California	0.231	D	D	D
Colorado	0.21	R	D	R
Connecticut	0.208	D	D	D
Delaware	0.221	D	D	D
D.C.	0.259	D	R	D
Florida	0.233	R	D	R
Georgia	0.275	R	R	R
Hawaii	0.207	D	D	D
Idaho	0.246	R	D	R
Illinois	0.253	D	R	D
Indiana	0.275	R	R	R
Iowa	0.263	R	R	D
Kansas	0.258	R	R	R
Kentucky	0.284	R	R	R
Louisiana	0.295	R	R	R
Maine	0.237	D	D	D
Maryland	0.252	D	D	D
Massachusetts	0.209	D	D	D
Michigan	0.277	D	R	D
Minnesota	0.248	D	D	D
Mississippi	0.344	R	R	R
Missouri	0.274	R	R	R
Montana	0.217	R	D	R
Nebraska	0.265	R	R	R
Nevada	0.236	R	D	R
New Hampshire	0.236	D	D	R
New Jersey	0.229	D	D	D
New Mexico	0.233	R	D	D
New York	0.235	D	D	D
North Carolina	0.271	R	R	R
North Dakota	0.259	R	R	R
Ohio	0.269	R	R	R
Oklahoma	0.281	R	R	R
Oregon	0.25	D	D	D
Pennsylvania	0.257	D	R	D
Rhode Island	0.214	D	D	D
South Carolina	0.292	R	R	R
South Dakota	0.261	R	R	R
Tennessee	0.29	R	R	R
Texas	0.272	R	R	R
Utah	0.218	R	D	R
Vermont	0.211	D	D	D
Virginia	0.252	R	D	R
Washington	0.245	D	D	D
West Virginia	0.306	R	R	R
Wisconsin	0.255	D	R	D
Wyoming	0.24	R	D	R

3. Decision Tree

a)



b) The petal_width attribute was used as the first decision node of this tree generated by the decision tree classifier

c)

E(class)

= P(Iris-setosa)E(Iris-setosa) + P(Iris-versicolor)E(Iris-versicolor) + P(Iris-virginica)E(Iris-virginica)

= -((50/150)log_2(50/150)) * 3

= 1.5850

E(class, pw0.8)

= P(low)E(low) + P(high)E(high)

= (50/150) * [-((50/50)log_2(50/50))-0-0]+ (100/150) * [0-((50/100)log_2(50/100))-((50/100)log_2(50/100))]

= 0.6667

G(class, pw0.8)= 1.5850-0.6667 = 0.918

E(class, pw1.75)

= P(low)E(low) + P(high)E(high)

= (104/150) * [-((50/104)log_2(50/104))-((49/104)log_2(49/104))-((5/104)log_2(5/104))] + (46/150) * [0-((1/46)log_2(1/46))-((45/46)log_2(45/46))]

= 0.8992
G(class, pw1.75)= 1.5850–0.8992= 0.686

E(class, pl4.95)
= P(low)E(low) + P(high)E(high)
= (104/150) * [-((50/104)log_2(50/104)) -((48/104)log_2(48/104)) - ((6/104)log_2(6/104))] + (46/150) * [0-((2/46)log_2(2/46))-((44/46)log_2(44/46))]
= 0.9529
G(class, pl4.95)= 1.5850–0.9529= 0.632

E(class, pw1.65)
= P(low)E(low) + P(high)E(high)
= (102/150) * [-((50/102)log_2(50/102))-((48/102)log_2(48/102))-((4/102)log_2(4/102))] + (48/150) * [0-((2/48)log_2(2/48))-((46/48)log_2(46/48))]
= 0.8954
G(class, pw1.65)= 1.5850-0.8954= 0.690

E(class, pw1.55)
= P(low)E(low) + P(high)E(high)
= (98/150) * [-((50/98)log_2(50/98))-((45/98)log_2(45/98))-((3/98)log_2(3/98))] + (52/150) * [0-((5/52)log_2(5/52))-((47/52)log_2(47/52))]
= 0.9194
G(class, pw1.55)= 1.5850-0.9194 = 0.666

E(class, sl6.95)
= P(low)E(low) + P(high)E(high)
= (137/150) * [-((50/137)log_2(49/137))-((49/137)log_2(49/137))-((38/137)log_2(38/137))] + (13/150) * [0-((1/13)log_2(1/13))-((12/13)log_2(12/13))]
= 1.4816
G(class, sl6.95)= 1.5850-1.4816= 0.103

E(class, pl4.85)
= P(low)E(low) + P(high)E(high)
= (99/150) * [-((50/99)log_2(50/99))-((46/99)log_2(46/99))-((3/99)log_2(3/99))] + (51/150) * [0-((4/51)log_2(4/51))-((47/51)log_2(47/51))]
= 0.9034
G(class, pl4.85)= 1.5850-0.9034 = 0.6816

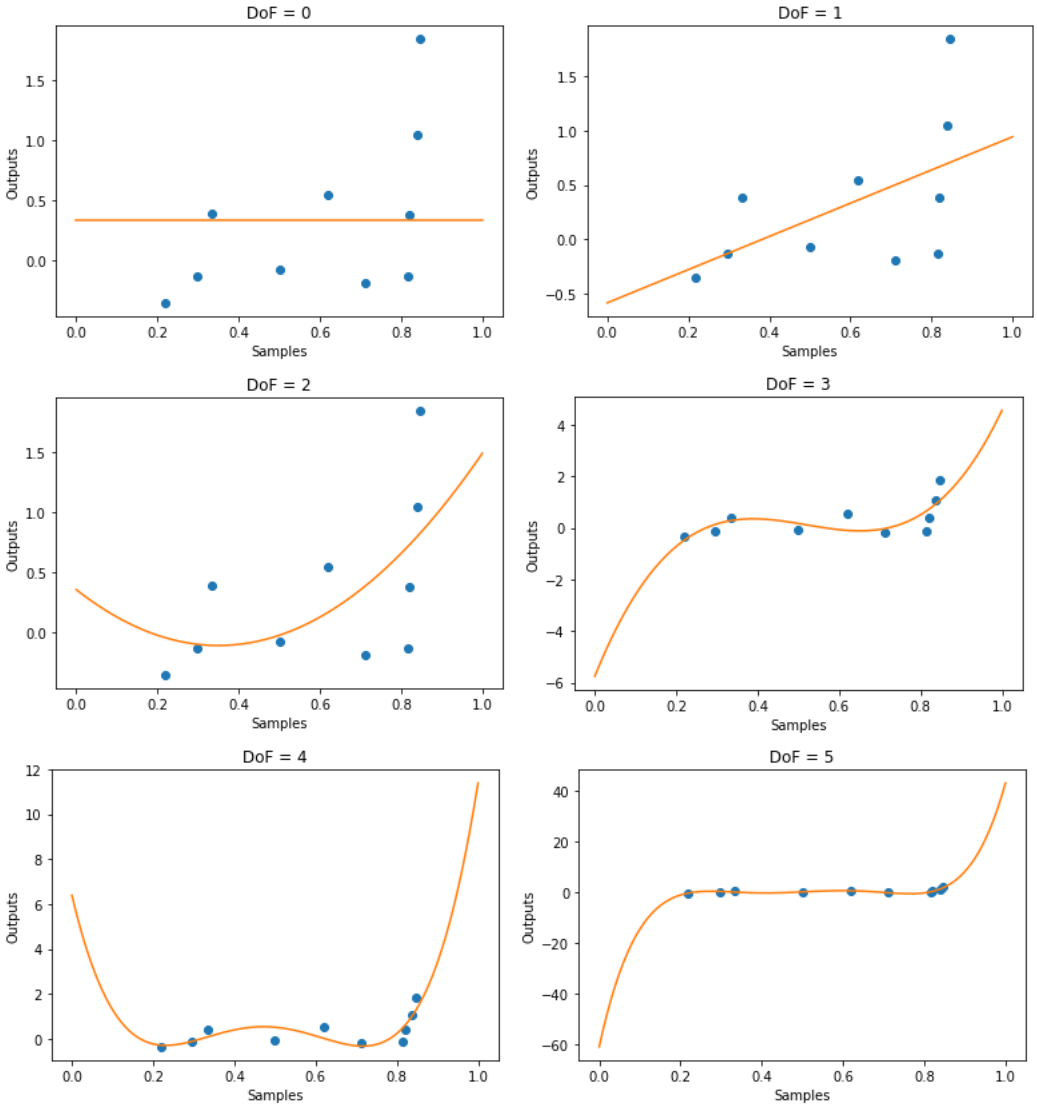
E(class, pl5.95)
= P(low)E(low) + P(high)E(high)
= (83/150) * [-((50/83)log_2(50/83))-((26/83)log_2(26/83))-((7/83)log_2(7/83))] + (67/150) * [0-((24/67)log_2(24/67))-((43/67)log_2(43/67))]
= 1.1209
G(class, pl5.95)= 1.5850-1.1209 = 0.4641

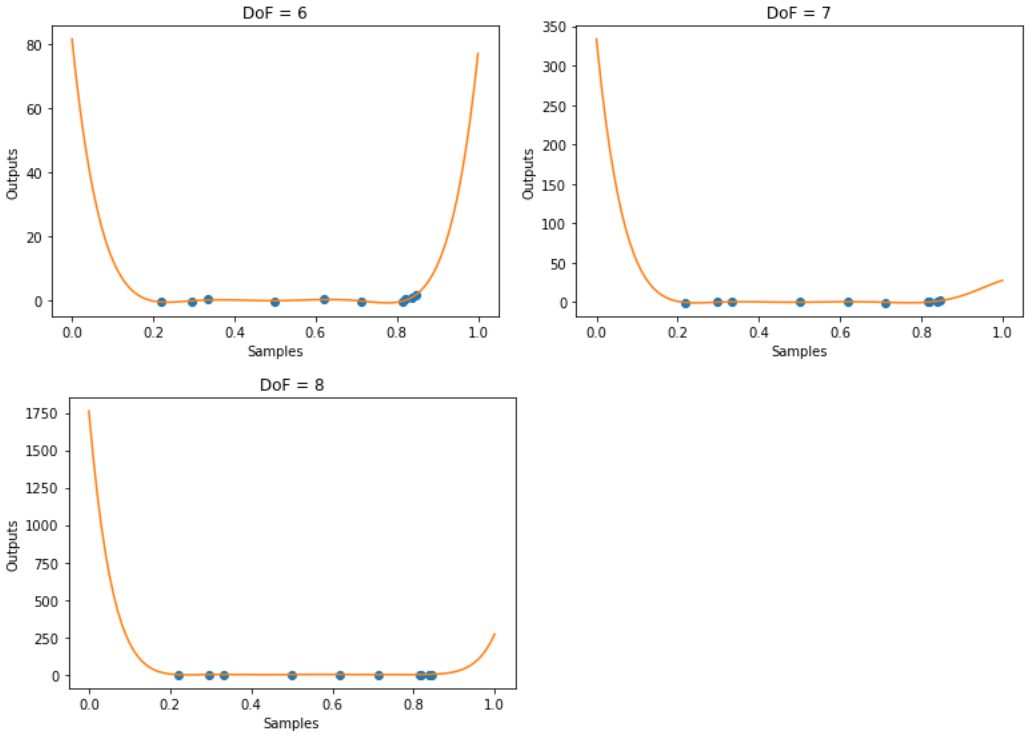
d) True

e) The best classification is the first classification, which is to see whether petal_width is smaller than 0.8 and the gain is 0.918, which is very close to 1, which is the highest among the 8 calculations.

4. Problem 2.11 b

	m	DoF	Remp	r	Rpen
0	0	1	0.418	1.101	0.460
1	1	2	0.291	1.069	0.312
2	2	3	0.275	1.065	0.293
3	3	4	0.199	1.047	0.208
4	4	5	0.153	1.036	0.159
5	5	6	0.044	1.010	0.045
6	6	7	0.020	1.005	0.020
7	7	8	0.010	1.002	0.010
8	8	9	0.009	1.002	0.009


DoF = 0
DoF = 1
DoF = 2
DoF = 3
DoF = 4
DoF = 5
Outputs
Samples



5. Prediction Accuracy using Resampling

Fold # Optimal m	Prediction	accuracy
0	0	0.51
1	3	2.37
2	0	0.19
3	0	2.75
4	0	0.38