Task 1

Describe the difference between classification and clustering?

Task 2

Describe what is entropy?

Task 3

Describe and compare the following "feature selection measures" or called "splitting criteria": information gain, gain ratio, and Gini index?

Task 4

Given training instances and their attributes, construct the following three decision trees by hand, and then implement the three decision trees using Python Decision Tree models:

- ID3: information gain
- C4.5: gain ratio
- CART: gini index

Question 1: We have
(1) 6 training instances and 6 testing instances
(2)3 attributes: (a) 2-value attribute (Home/Away), (b) 2-value attribute (In/Out), (c) 4-value attribute (NBC/ESPN/FOX/ABC)

	Date	University	Is Home/Away?	Is Opponent in AP Top 25 at Preseason?	Media	Label: Win/Lose
1	9/2/17	Temple	Home	Out	1-NBC	Win
2	9/9/17	Georgia	Home	In	1-NBC	Lose
3	9/16/17	Boston College	Away	Out	2-ESPN	Win
4	9/23/17	Michigan State	Away	Out	3-FOX	Win
5	9/30/17	Miami Ohio	Home	Out	1-NBC	Win
6	10/7/17	North Carolina	Away	Out	4-ABC	Win

7	10/19/17	USC	Home	In	1-NBC	?
8	10/25/17	North Carolina State	Home	Out	1-NBC	?
9	11/4/17	Wake Forest	Home	Out	1-NBC	?
10	11/12/17	Miami Florida	Away	In	4-ABC	?
11	11/18/17	Navy	Home	Out	1-NBC	?
12	11/26/17	Stanford	Away	In	4-ABC	?

Question 2: We have
(1) 14 training instances and 1 testing instance
(2) 4 attributes: (a) 3-value attribute (Sunny/Overcast/Rainy), (b) 3-value attribute (Hot/Mild/Cool), (c) 2-value attribute (High/Normal), (d) 2-value attribute (True/False)

ID	Date	Outlook	Temperature	Humidity	Windy	Label: Play?
1	9/1/17	Sunny	Hot	High	"False"	No
2	9/8/17	Sunny	Hot	High	"True"	No
3	9/15/17	Overcast	Hot	High	"False"	Yes
4	9/22/17	Rainy	Mild	High	"False"	Yes
5	9/29/17	Rainy	Cool	Normal	"False"	Yes
6	10/1/17	Rainy	Cool	Normal	"True"	No
7	10/8/17	Overcast	Cool	Normal	"True"	Yes
8	10/15/17	Sunny	Mild	High	"False"	No
9	10/22/17	Sunny	Cool	Normal	"False"	Yes

10	10/29/17	Rainy	Mild	Normal	"False"	Yes
11	11/1/17	Sunny	Mild	Normal	"True"	Yes
12	11/8/17	Overcast	Mild	High	"True"	Yes
13	11/15/17	Overcast	Hot	Normal	"False"	Yes
14	11/22/17	Rainy	Mild	High	"True"	No
15	11/29/17	Rainy	Hot	High	"False"	?

Task 5

Given a university's football game data for the last two seasons, please construct three classification models to predict game results on games, and evaluate the model performance. Here, the three classification models are ID3, C4.5, and Naïve Bayes.

- Data
  - Each data object (or called instance) is a game. We have three attributes: (1) "Is Home/Away?", a 2-value attribute ("Home", "Away"), (2) "Is Opponent in AP Top 25 at Preseason?", a 2-value attribute ("In", "Out"), (3) "Media", a 5-value attribute ("1-NBC", "2- ESPN", "3-FOX", "4-ABC", "5-CBS"). The label "Win/Lose" is binary ("Win", "Lose").
- Training set
  - 24 games. Please use game ID 1-24 to construct classification models.
- Testing set
  - 12 games. Please use your classification models to predict labels of game ID 25-36 and evaluate the performance of the classification models.
- Predictive labels
  - Suppose "Win" is the positive label and "Lose" is the negative label. Keep it in mind when you use Precision and Recall to evaluate the models.
- Stop criteria of decision tree models
  - We stop splitting instances into child nodes when one of the criteria is satisfied: (1) All features have been used; (2) Information Gain or Gain Ratio will be zero with any feature that has not yet been used.
- Prediction criteria
  - If the node is not pure, we use the majority of this node for prediction: For example, if we have 5 positives and 1 negatives, we predict the testing case at

this node to be a positive. (2) If the node has a balance (half/half labels), e.g., 2 positives and 2 negatives, we use the majority of the root node (the entire dataset) for prediction.

ID Training	Date Data:	Opponent	Is_Home_or_Away	Is_Opponent_in_AP25_Preseason	Media	Label
1	9/5/15	Texas	Home	Out	1-NBC	Win
2	9/12/15	Virginia	Away	Out	4-ABC	Win
3	9/19/15	GeorgiaTech	Home	In	1-NBC	Win
4	9/26/15	UMass	Home	Out	1-NBC	Win
5	10/3/15	Clemson	Away	In	4-ABC	Lose
6	10/10/15	Navy	Home	Out	1-NBC	Win
7	10/17/15	USC	Home	In	1-NBC	Win
8	10/31/15	Temple	Away	Out	4-ABC	Win
9	11/7/15	PITT	Away	Out	4-ABC	Win
10	11/14/15	WakeForest	Home	Out	1-NBC	Win
11	11/21/15	BostonCollege	Away	Out	1-NBC	Win
12	11/28/15	Stanford	Away	In	3-FOX	Lose
13	9/4/16	Texas	Away	Out	4-ABC	Lose
14	9/10/16	Nevada	Home	Out	1-NBC	Win
15	9/17/16	MichiganState	Home	Out	1-NBC	Lose
16	9/24/16	Duke	Home	Out	1-NBC	Lose

17	10/1/16	Syracuse	Home	Out	2-ESPN	Win
18	10/8/16	NorthCarolinaState	Away	Out	4-ABC	Lose
19	10/15/16	Stanford	Home	In	1-NBC	Lose
20	10/29/16	MiamiFlorida	Home	Out	1-NBC	Win
21	11/5/16	Navy	Home	Out	5-CBS	Lose
22	11/12/16	Army	Home	Out	1-NBC	Win
23	11/19/16	VirginiaTech	Home	In	1-NBC	Lose
24	11/26/16	USC	Away	In	4-ABC	Lose

Testing Data

ID	Date	Opponent	Is_Home_or_Away	Is_Opponent_in_AP25_Preseason	Media	Label
25	9/2/17	Temple	Home	Out	1-NBC	Win
26	9/9/17	Georgia	Home	In	1-NBC	Lose
27	9/16/17	BostonCollege	Away	Out	2-ESPN	Win
28	9/23/17	MichiganState	Away	Out	3-FOX	Win
29	9/30/17	MiamiOhio	Home	Out	1-NBC	Win
30	10/7/17	NorthCarolina	Away	Out	4-ABC	Win
31	10/21/17	USC	Home	In	1-NBC	Win
32	10/28/17	NorthCarolinaState	Home	Out	1-NBC	Win
33	11/4/17	WakeForest	Home	Out	1-NBC	Win
34	11/11/17	MiamiFlorida	Away	In	4-ABC	Lose
35	11/18/17	Navy	Home	Out	1-NBC	Win
36	11/25/17	Stanford	Away	In	4-ABC	Lose

Question 1: ID3 model, a decision tree model using “Information Gain”
(1) Programming: Use ID3 to construct a decision tree based on the training set (24 games). Use the tree to predict labels of instances in the testing set (12 games) based on their attributes. Calculate Accuracy, Precision, Recall, and F1 score on the testing result.
(2) Attach a figure of your decision tree (either hand- or electronically drawn) and write down prediction label of the 12 testing games as well as evaluation result in the PDF.

Question 2: C4.5 model, a decision tree model using “Gain Ratio”
(1) Programming: Use C4.5 to construct a decision tree based on the training set (24 games). Use the tree to predict labels of instances in the testing set (12 games) based on their attributes. Calculate Accuracy, Precision, Recall, and F1 score on the testing result.
(2) Attach a figure of your decision tree (either hand- or electronically drawn) and write down prediction label of the 12 testing games as well as evaluation result in the PDF.

Question 3: which model is the best, which model performs the worst? Can you explain why?

Please submit a report (PDF or word) that includes a link to your code, your answers/results, and your explanations or interpretations (if any).