CS490/584 Data Mining
Mid Term Exam

Name _____ Grade _____
[Please Print your name, FirstName LastName]

INSTRUCTIONS:

- ***Do not begin or open this test until directed to do so!***
- ***This test is closed book, closed notes, and no computer access is allowed. However, you are allowed two letter-size (8.5X11) cards with handwritten notes on both sides.***
- ***Absolutely no collaboration. Any violations of this policy may result in a 0 on the test.***
- ***Each part often have several questions. Do the number of questions as required. Do not do more than asked for.***
- ***You must show your work where applicable. Use back of page if necessary.***
- ***The test's time limit is 75 minutes. Be sure to manage your time appropriately.***

GRADING BREAKDOWN:

	Points Possible	Points Earned
Part I	11	
Part II	68	
Part III	20	
Part IV	11	
Total	110	

Part I Matching – Select the answer that **best** fits the definition or complete the sentence. An answer may be used more than once. (1*11=11 points)

A1. Bias	A2. clustering	A3. covering	A4. discretization
B1. superset	B2. gain ratio	B3. aggression	B4. Minkowski
C1. Subset	C2. missing value	C3. unsupervised	C4. SplitInfo
D1. regression	D2. over-fitting	D3. supervised	D4. disappearance
E1. Newton	E2. medium	E3. OneR	E4. standard deviation
F1. arbitrary	F2. AQ	F3. Quinlan	F4. spread out
G1. ID3	G2. gini	G3. Search	G4. bunched up
H1. integration	H2. class	H3. Dispersion	H4. memoization
J1. regression	J2. selection	J3. bias	J4. focus
K1. coverage	K2. accuracy	K3. Cleaning	K4. transformation
L1. conversion	L2. root	L3. leaf	L4. None of the above

1. _____ learning refers to a general type of learning where the classifications of training examples (ex., play) are given. OneR is a simple example of such a learning method. In this case, the learned rule (s) is then used as a model for classification of future data with unknown _____ value.

2. ID3 uses a top-down approach to create decision tree recursively. At each step, it uses entropy measure to calculate information gain in order to select the best attribute to further divide the instances at a particular branch as long as the instances at the branch are not all of the same class and there is attribute(s) not yet used. CART, a similar system that also create decision tree, uses _____ index (or co-efficient) to select attribute.

3. Generated decision trees are often pruned to avoid _____, which describes the situation where excessive branches are generated, often for very small number of examples in the training set. Decision tree can be used either directly to classify new data item, or create production rules. A unique production rule can be generated by tracing through a path from root node of the tree to a leaf node. The class label of this rule can be found at the _____ node.

4. When searching for a target concept, Data Mining systems must make important decisions on how the target concepts should be described, the order in which search spaces will be explored, etc. This is generally referred to as the _____ of the search.

5. _____ learning refers to a general type of learning where the classifications of training examples (ex., *whether* to play a game) are given. OneR is a simple example of such a learning algorithm.

6. The style of learning that finds groupings of similar items without the help of labels is called _____. It's an example of a general class of learning referred to as _____ learning.

7. _____ is a commonly used method for smooth data by fitting them into a mathematical function.

8. _____ is a word used in the Data Mining, meaning the conversion of numeric attribute to nominal ones.

Part II. Data Analysis and Simple Algorithms

Answer any four (4) of Six (6) questions. Must show work! (ex., formulas used in calculation). (4X17 = 68)

1. **Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are shown in the table below. Calculate the approximate median. Please show work!**

Age	Frequency
1-5	100
6-15	200
16-20	300
21-50	400
51-80	500
81-110	100

2. **Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 2, 11, 12, 13, 16, 17, 20, 21, 23, 27, 29, 31, 31, 31, 35, 37, 39, 41, 45, 80**

   (a) Give the five-number summary of the data.

   (b) Draw a boxplot of this data. Be sure to mark clearly the box length, min and max whisker, and outliners, if any.

Each of the four questions (3-6) are based on the following weather data, with various modifications. However, the class attribute is always **play**. ***** **Do any 3 of the 4 ****

outlook	temperature	humidity	windy	play
overcast	64	65	true	yes
rainy	65	50	false	yes
sunny	69	70	false	yes
rainy	70	50	false	yes
sunny	70	95	false	no
rainy	71	75	true	yes
overcast	72	90	false	no
rainy	75	85	true	no
sunny	80	90	true	no
overcast	81	70	false	yes
overcast	83	50	false	yes
rainy	85	70	true	no
sunny	88	40	true	no
sunny	90	65	false	no

3. **Discretization** – Use the supervised discretization method as discussed in Witten 4.1 to convert temperature values into labels. Draw a new table with temperature values replaced by labels and show the resulting table below. Please show work (ex., sorting, break points etc.)

4. **OneR** – use oneR algorithm to select the best one rule, considering among three candidate attributes: outlook, temperature, and windy. Note that (a) The select will be based on the table you created in question 3 where temperature is already discretized, and (b) the humidity attribute will not be considered for this problem. Please show work!

5. **Bayes** – Use basic (or modified if necessary) Naïve Bayes method to determine the classification for a new day using the weather training data minus the temperature column. Here, we will consider the humidity attribute as a continuous-valued attribute with a Gaussian distribution. You may want to compute P(humidity=70|play=yes) and P(humidity=70|play=no) separately first before using them to determine P(play=yes|day) and P(play=no|day). Use 3.14 for pi. 2.72 for e, and 2 decimal places for mean and standard deviation. Please show work! (16 points)

Day

Outlook	Humidity	Windy	Play
Sunny	70	True	?

6. **PRISM** - Apply the PRISM algorithm on the following modified weather data to create ***one*** classification rule for the "Yes" class. How many instances does this rule cover? Please show work! (16 points)

outlook	humidity	windy	play
sunny	low	true	no
rainy	low	false	yes
rainy	low	false	yes
overcast	low	false	yes
overcast	low	true	yes
sunny	low	false	no
sunny	mid	false	yes
overcast	mid	false	yes
rainy	mid	true	no
rainy	mid	true	yes
rainy	high	true	no
overcast	high	false	no
sunny	high	true	no
sunny	high	false	no

Part III. ID3 - We are to create an ID3 decision tree based on the training data shown in the table on

the next page. Suppose the calculation for root is already done and the attribute <u>Windy was</u> selected, perhaps erroneously. Do each of the following (2+16+2=20 points)

1. **Draw the partial tree with *windy* as root and indicate the data elements in each branch by using their ID numbers.**

2. **Continue the process to generate the partial tree for the windy=false branch using ID3 algorithm. Note that you do not have to show actual calculation for simple cases such as when the class split ratios are (1, 1), (N, 0). Specifically,**

   Info([N,0]) = Info([1,0]) = 0
   Info([N,N]) = Info([1,1]) = 1
   Info([2,3]) = Info([3,2]) = 0.971
   Info([1,2]) = Info([2,1]) = 0.918

   Note: the only entropy you will probably need to calculate and show work is Info([3,1]) .

3. **How many rules can we get from the generated partial tree? Show one of them.**

ID	Outlook	temperature	humidity	windy	play
1	overcast	Hot	high	false	yes
2	overcast	Hot	normal	false	no
3	Rainy	Mild	high	false	no
4	Rainy	Cool	normal	false	no
5	Rainy	Mild	normal	false	no
6	Sunny	Hot	high	false	no
7	Sunny	Cool	normal	false	yes
8	overcast	Cool	normal	true	no
9	overcast	Mild	high	true	yes
10	Rainy	Cool	normal	true	no
11	Rainy	Mild	high	true	no
12	Sunny	Hot	high	true	no
13	Sunny	Mild	high	true	no
14	Sunny	Mild	normal	true	no

Part IV. Bonus Round - Know your Tech

Fill in each black with the letter that best completes the sentence. (11X1 pts)

A. AMD	B. Boring	C. ChatGPT	D. Deep Mind
C. Elon Musk	F. Jamie Dimon	G. Jenson Huang	H. Lisa Su
I. IBM	J. Nvidia	K. OpenAI	L. Sam Altman
M. Satya Nadella	N. SpaceX	O. Sundai Pichai	P. Tesla
P. Tim Cook	Q. Jeff Bezos	R. X	S. none of the above

.1 As evident from its failed rollout of the Gemini AI (Exhibit B), α-β (formally Google), under the leadership of _____, has squandered its lead in AI which it held for a long time, especially after acquiring _____ , the creator of Alpha-Go.

.2 Microsoft, on the other hand, under the bold (no pun intended) and visionary leadership of _____, (exhibit A), is striving in the current generative AI frenzy, especially after acquiring 49% of _____, creator of _____, which is led by its embattled CEO _____.

.3 _____ corporation, the lead "shovel maker" during the current generative AI "Gold Rush", is led by _____, who, incidentally, is a distance cousin of _____, a female CEO of _____, another top semiconductor company.

.4 Among many of the companies that Elon Musk help to create and led, the _____corporation is perhaps the least known. Many of its projects involve digging underground tunnels in major cities, including the somewhat famous Vegas Loop Tunnel.

Exhibit A ***Exhibit B***