CS 5990 (Advanced Data Mining) - Assignment #1

Maximum Points: 100 pts.

Bronco ID: |__|__|__|__|__|__|__|__|__|

Last Name: ______________________

First Name: ______________________

Note 1: Your submission header must have the format as shown in the above-enclosed rounded rectangle.
Note 2: Homework is to be **done individually**. You may discuss the homework problems with your fellow students, but you are NOT allowed to copy – either in part or in whole – anyone else's answers.
Note 3: Your deliverable should be a **.pdf file** submitted through **Gradescope** by the deadline. Do not forget to **assign a page to each of your answers** when making a submission. In addition, source code (.py files) should be added to an **online repository (e.g., GitHub)** to be downloaded and executed later.
Note 4: All submitted materials must be legible. Figures/diagrams must be of good quality.
Note 5: Please use and check the Canvas discussion for further instructions, questions, answers, and hints. The bold words/sentences provide information for a complete or accurate answer.

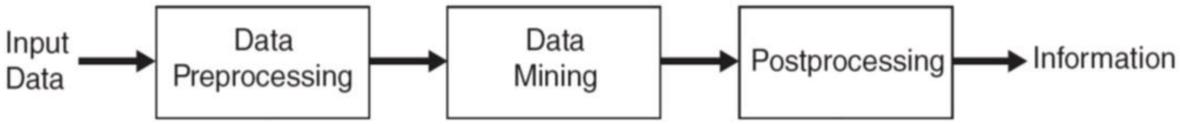
1. [12 points – 2 points each] Answer whether each of the following activities is a data mining task. **Justify** your answer.

   a. Dividing the customers of a company according to their gender.
   b. Monitoring seismic waves for earthquake activities.
   c. Computing the total sales of a company.
   d. Predicting the outcomes of tossing a (fair) pair of dice.
   e. Predicting the future stock price of a company using historical records.
   f. Monitoring the heart rate of a patient for abnormalities.

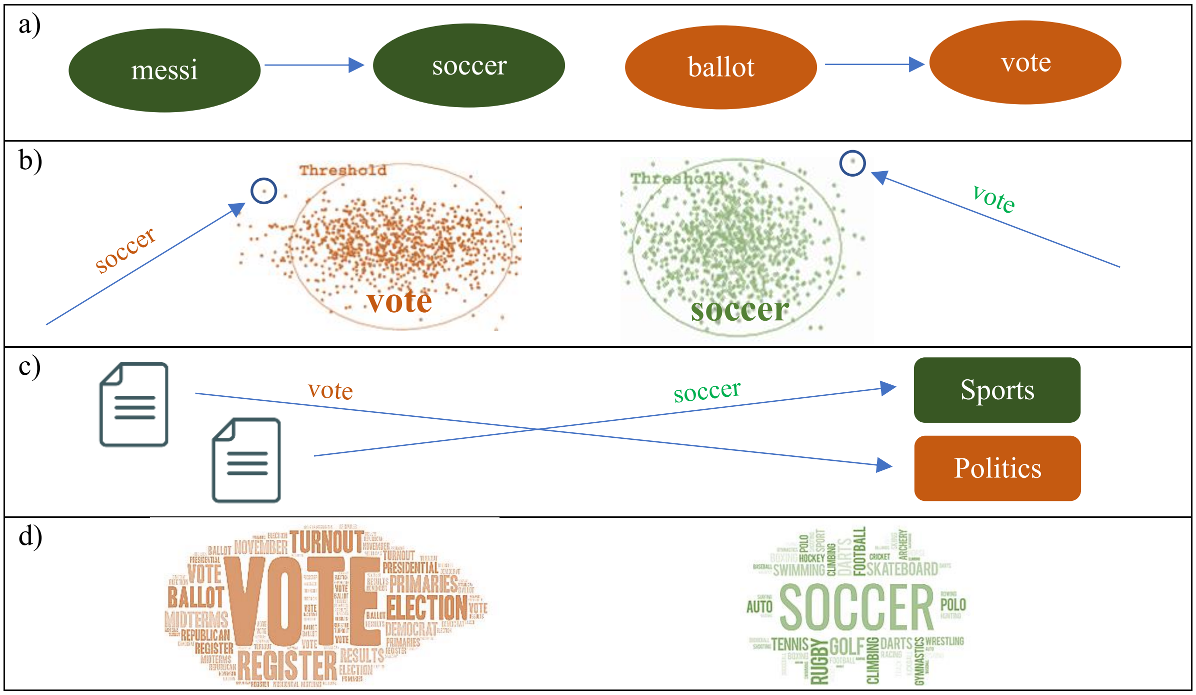
2. [12 points – 2 points each] Classify the following attributes as **discrete**, or **continuous**. Also classify them as **nominal**, **ordinal**, **interval**, or **ratio**.

   a) Brightness as measured by a light meter.
   b) Brightness as measured by people's judgments.
   c) Density of a substance in grams per cubic meter.
   d) Time of each day in the meaning of a 12-hour analog clock.
   e) CPP bronco IDs.
   f) Customer satisfaction ratings.

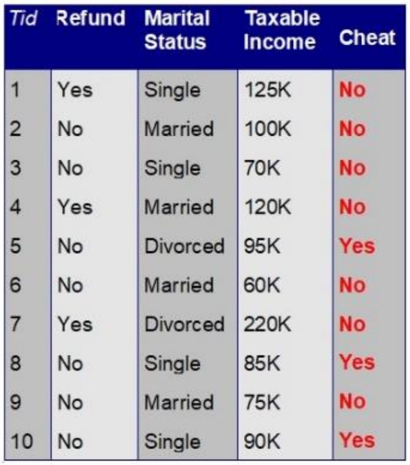
3. [9 points] Explain where **visualization**, **dimensionality reduction**, and **machine learning** techniques are applied during the 3 main phases of the KDD process shown below. Justify the importance of these techniques to produce useful information at the end of the process.



4. [8 points] Suppose that you are employed as a data mining consultant for company that provides a Web search engine. Use the illustrations below to explain how **clustering**, **classification**, **association rule mining**, and **anomaly detection** can be applied to help the engine. You will need to figure out **which data mining technique** corresponds to **which illustration** and what **kind of results is being provided**.



5. [13 points] Analyze the dataset below and answer the proposed questions:



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

a. [3 points]. What is the **most likely task** that data scientists are trying to accomplish and **for whom**?

b. [2 points]. **In general**, what is a feature and how would you **exemplify** it with **this data**?

c. [2 points]. **In general**, what is a feature value and how would you **exemplify** it with **this data**?

d. [2 points]. **In general**, what is dimensionality and how would you **exemplify** it with **this data**?

e. [2 points]. **In general**, what is an instance and how would you **exemplify** it with **this data**?

f. [2 points]. **In general**, what is a class and how would you **exemplify** it with **this data**?

6. [16 points – 2 points for each] For each of the following vectors, x and y, calculate the indicated similarity or the distance measures. Show your **math** for full mark.

   (a) x = (1 1 0 0 0), y = (0 0 0 1 1). Jaccard, Cosine, Euclidean, Correlation.

   (b) x = (0 1 0 1 1), y = (1 0 1 0 0). Jaccard, Cosine, Euclidean, Correlation.

7. [10 points] Given vectors u = (2, k) and v = (3, -2), find the value of k such that vectors are

   (a) perpendicular (b) parallel

   Show your **math** for full mark.

8. [20 points] Complete the python program (similarity.py) to find and output the two most similar documents from the cleaned_documents.csv dataset based on their cosine similarity. The output should follow this format: "The most similar documents are document 10 and document 100 with cosine similarity = x."

Important Note: Answers to all questions should be written clearly, concisely, and unmistakably delineated. You may submit it multiple times until the deadline (the last submission will be considered).

NO LATE ASSIGNMENTS WILL BE ACCEPTED. ALWAYS SUBMIT WHATEVER YOU HAVE COMPLETED FOR PARTIAL CREDIT BEFORE THE DEADLINE!