CMSC 435 Assignment 1

Fall 2022
(individual work; 5 pts total)

Peer-reviewed scientific articles are high quality sources that are reviewed, accordingly improved/corrected, and approved by scientists before publication. This assignment provides an opportunity to read and discuss a peer-reviewed scientific article related to the core topics of our data science class.

We ask you to write about 1-page long report of an article entitled "Top 10 algorithms in data mining" by Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, McLachlan, Ng, Liu, Yu, Zhou, Steinbach, Hand and Steinberg. Your review should consist of two main parts:

PART 1. (1 pt) Presentation of **your** criticisms of this article that covers **your** opinions and insights about the ideas presented by the authors. As a few options how to proceed, you can discuss which parts and why were particularly useful/interesting, what and why was deficient, and/or present relevant insights that were missed by the authors. Make sure that you do not duplicate statements that are included in the article. (answer length limit: up to ½ page)

PART 2. Answers to the following four questions:

Question 2.1. (1 pt) Why is the Naive Bayes algorithm called **naive**? What is the meaning of the word **Bayes**? (answer length limit: up to 3 sentences).

Question 2.2. (1 pt) **List** the algorithm(s) discussed in this article that generate **decision trees**. Which of these algorithms was published/released **first**? (answer length limit: 2 sentences).

Question 2.3. (1 pt) **Briefly explain** the meaning of k in the kNN algorithm. **Could** k be set to 1? **What would be the prediction** for the 7NN algorithm with the majority vote where the 7 neighbors belong to the following classes: $class_1$, $class_1$, $class_3$, $class_2$, $class_1$, $class_2$, $class_3$? (answer length limit: up to 3 sentences).

Question 2.4. (1 pt) Which **five of the nine algorithms** would be the most suitable choices to be used as the "weak learners" by the **AdaBoost** algorithm? Hint: these five are called classifiers or supervised algorithms (answer length limit: 1 sentence).

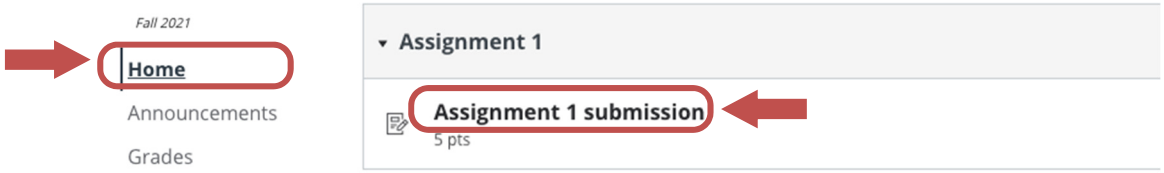
Notes

- You **do not need to understand all concepts and information covered in this article** but you need to learn enough to complete the review. Feel free to use other external resources to answer the questions (collaboration with other students is not allowed). A useful resource that I recommend is a post by Raymond Li at http://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html
- Use a separate and **clearly marked paragraph** for each of the five sub-parts of the solution and number them accordingly (i.e., 1, 2.1, 2.2, 2.3 and 2.4). There will be **deductions** if this is not followed.
- In the first part we ask for your **personal** criticism. This section must have substance and must not be based on existing sources.
- Read the questions carefully and make sure to answer each part of each question.
- Observe the limits on the length of the answers.
- The review should be typed single-spaced, using 12-point font size (Times New Roman font would be a good choice) and with standard margins. Convert the file into the **pdf** format for submission.
- Fun fact: we will learn many of these algorithms in details in the class including C4.5, k-means, SVM, and Apriori.

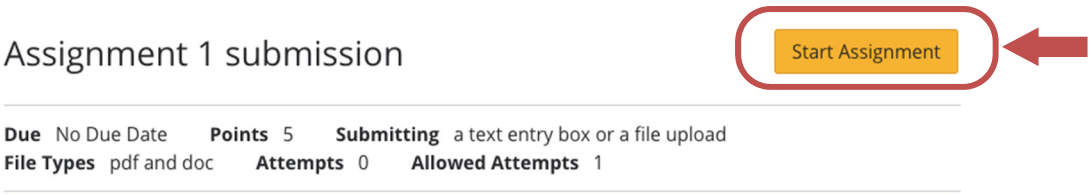
Due Date

Your assignment must be received by 12:30pm, September 8 (Thursday), 2022. Submission of the pdf file must be done via the class web page in Canvas. Follow the 5-step instructions that are included below.

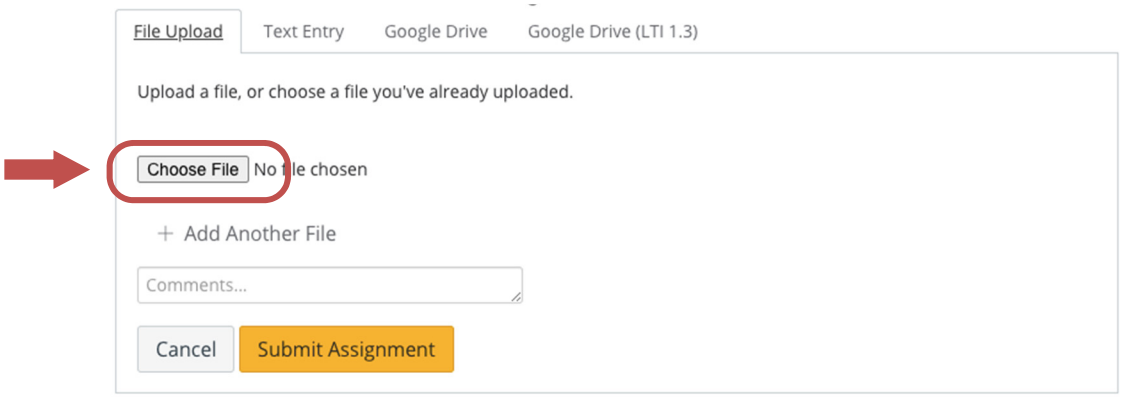
1. Go to the "Home" section, locate the "Assignment 1 submission" field and select it by clicking on the assignment title.



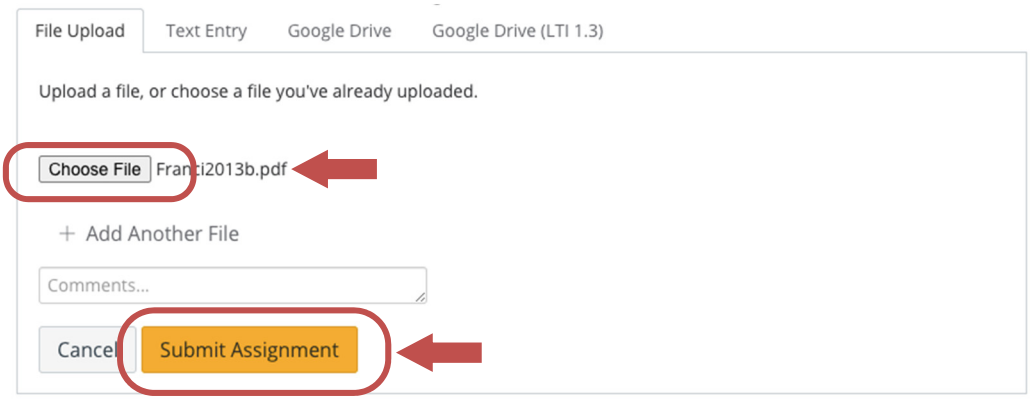
2. Select “Start Assignment”



3. You can select your submission file(s) by clicking on “Choose file”.



4. [Important!] Your file(s) will be submitted only after you click the “Submit Assignment” button.



5. Make sure your submission was completed and the correct file was sent.