**Data mining Final Project**

1.Results of Data Mining have limitations in terms of:

D. poor data

2.(T/F) Data sampling techniques can be used to analyze a small subset of data prior to Data mining

True

3.Data Mining relies on:

C. Computational Efficiency of the algorithm

4.The data model (algorithm) selected in data mining must be:

A. Scalable

5.One of the fundamentals purposes of performing Data Mining is to:

D. Develop an optimized method for big data computation.

6.Which of the following Venn diagrams closely represents the relationship between various disciplines and sub-disciplines.

B

7. (T/F) The odds ratio can be computed from probability values and represents another way of describing it.

True

8. (T/F) KNN methods are common in practical applications of the Data Mining and can be used for solving both Classification and Regression types of problem.

True

9. Compare and contrast TF-IDF and BoW technique/algorithm for application where one could be better applied than the other. Cite two applications each of the these could be beneficial.

* The bag-of-words model converts text into fixed-length vectors by counting how many times each word appears. Can be used in Natural processing language (NLP) and information retrieval.
* TFIDF works by proportionally increasing the number of times a word appears in the document but is counterbalanced by the number of documents in which it is present. Can be used in connection with information retrieval and shows how frequently an expression (term, word) occurs in a document
* The bag-of-words model is commonly used in methods of [document classification](https://en.wikipedia.org/wiki/Document_classification) where the (frequency of) occurrence of each word is used as a [feature](https://en.wikipedia.org/wiki/Feature_(machine_learning)) for training a [classifier](https://en.wikipedia.org/wiki/Statistical_classification)

10. (T/F): Baysian classifier cab be used to solve problems related to:

C. Categorization

11.In k-Means clustering algorithms:

B. Efficiency depends on proximity of values to each other

12. (T/F): One of the advantages of logit is easy to train the model on limited number of observations

True

13. Describe The IBk set of algorithms in Weka. The purpose of algorithms, what specific algorithm within it could you relate with what we discussed in the class:

Weka provides a graphical user interface for exploring and experimenting with machine learning algorithms on datasets, without you having to worry about the mathematics or the programming. The IBk algorithm does not build a model, instead it generates a prediction for a test instance just-in-time. The IBk algorithm uses a distance measure to locate k “close” instances in the training data for each test instance and uses those selected instances to make a prediction.

We can relate IBk set algorithms with k-nearest neighbor algorithm (KNN)

14. List the key steps taken to run the text mining algorithm for assignments? What mining algorithm did you choose? Any issues encountered while setting up the Database and or running the program? What did you conclude from your analysis of the Movie Review Dataset.

The first step is to read data into the workspace, clean the dataset such that ensuring your data is free from unused columns, missing values, and duplicates

Then, perform sentiment analysis, this is where, we have tokenization, converting all characters to lowercase, removing stop words, removing punctuation and stemming of the words. Then we perform data analysis and visualization such as wordcount. Finally, select data in x and y values and use logistics regression to fit and predict data.

15. Write a piece of code (R/Python) that will do the following

a. Generate a random number from 10-40 total 10

b. Add them to a list/vector

c. assign values between 0.1 -0.10 i.e., first of the 10 numbers get 0.1, second gets 0.2,

etc.

d. Generate a histogram with numbers as X-axis and assigned values as Y-axis

f. Do you see any possible trend in the graphs? (Individual answers and graphs will vary)

g. Insert the graphics in the submission document







