**Assignment 6.1**

In this assignment, you will apply your knowledge and understanding of the topics presented in the module readings and materials.

**Instructions:**

1. Answer the assignment questions on the following pages.
2. Create a single document that combines your solutions to all question prompts.

**Final Format for Submission:**

Submit one assignment file using these guidelines:

* The file must be either a **Microsoft Word** or a **PDF** document file.
* **Do not** combine and submit files into a zipped compressed folder.
* Please use the following naming conventions for your **Word** or **PDF** assignment file:

File Naming: **LastName\_FirstName\_Assignment<Number>.pdf**

Example: **Smith\_James\_Assignment5.1.pdf**

* Answer all parts of a question in one place and answer questions in the order they appear in the assignment.
* For programming answers using **R**, it is recommended that the answers are written in R Markdown and ‘knitted’ to a Word/PDF file.
  + Do not print data frames in your submission, if you want to make a point about data you can use head(df) to print the first few rows.
  + Submit the code used to answer the questions in the assignment with your name on it, answers without code and appropriate results will not get full credit.
  + It is not a professional practice, but in case of difficulty, you can take screenshots of code and outputs and submit them in a Word/PDF file.
* Maximum number of pages should be 15. Any submissions that exceed 15 pages will **not be graded.**

**Grading and Scoring:**

* Use common sense to gauge the expectations of the answer to the number of points assigned to the question. See the Scoring Rubric in Blackboard for details.

# **Assignment 6.1 – Questions**

1. (15 points)

The table below lists a dataset that was used to create a nearest neighbor model that predicts whether observations belong to class A or B.

| X1 | X2 | X3 | Class |
| --- | --- | --- | --- |
| 1 | 2 | 30 | A |
| 2 | 3 | 40 | B |
| 3 | 4 | 50 | A |
| 4 | 5 | 60 | B |
| 5 | 6 | 70 | B |
| 6 | 7 | 80 | A |

Assuming that the model uses Euclidean distance to find the nearest neighbor, what prediction will the model return for each of the following query instances.

| X1 | X2 | X3 | Class |
| --- | --- | --- | --- |
| 9 | 4 | 15 | ? |
| 7 | 8 | 45 | ? |
| 8 | 7 | 75 | ? |

1. Perform the analysis WITHOUT doing any preprocessing of the data. (5 points)
2. Perform the analysis WITH appropriate preprocessing of the data. (8 points)
3. Do you observe any differences between the predictions with and without preprocessing? (2 points)

2. (30 points)

Table

Description automatically generated

**Answer the following questions without performing any preprocessing of the training data.**

1. What target level would a nearest neighbor model using Euclidean distance return for the following email: “machine learning for free”? (10 points)
2. What target level would a *k*-NN model with k = 3 and using Euclidean distance return for the same query? (5 points)
3. What target level would a *k*-NN model with *k* = 3 and using Manhattan distance return for the same query? (5 points)
4. There are a lot of zero entries in the spam bag-of-words dataset. This is indicative of sparse data and is typical for text analytics. Cosine similarity is often a good choice when dealing with sparse non-binary data. What target level would a 3-NN model using cosine similarity return for the query? (10 points)

3. (25 points)

Text

Description automatically generated

Table

Description automatically generated

1. Create a naive Bayes model that uses probability density functions to model the descriptive features in this dataset (assume that all the descriptive features are normally distributed). Show clearly the distribution of each of the predictors conditional on status and the prior probabilities of the status. (15 points)
2. What prediction will the naive Bayes model return for the following query? SS-IN = 222, SED-IN = 4.5, COND-IN = 1518, SS-OUT = 74, SED-OUT = 0.25, COND-OUT = 1642. (10 points)

4. (20 points)

Table

Description automatically generated

1. What target level will a Naive Bayes model predict for the following query document: “machine learning is fun”? (8 points)
2. What target level will a Naive Bayes model predict for the following query document: “christmas family fun”? (4 points)
3. What target level will a Naive Bayes model predict for the following query document: “learning is family fun”? (8 points)