# IAI5101: Foundations of Machine Learning for Scientists & Engineers

## Winter 2022

## Final Assignment

**Submission Deadline: 15th April 2022 on Brightspace.**

This final assignment should be **completed individually**. Upon completion, present your result in one submission, including the answers generated **(Note: not more than 15 pages**).

### Part A: Data preparation (25 points)

1) A dataset has 500,000 records and 40 variables with 15% of the values missing, spread randomly throughout the records and variables. You have decided to remove the records with missing values. About how many records would you expect to be removed? (5 points)
2) Some women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, were tested for diabetes according to World Health Organization criteria. Given a sample of the data for their *age* and *BMI* in Table 1 below, answer the following (20 points):
   a) Normalize the two attributes based on *z-score* normalization
   b) Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated?

*Table 1: Age and BMI of Pima Indian Women*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Age* | 24 | 55 | 35 | 26 | 23 | 52 | 25 | 24 | 63 | 31 |
| *BMI* | 30.2 | 25.1 | 35.8 | 47.9 | 26.4 | 35.6 | 34.3 | 25.9 | 32.4 | 43.3 |
| | | | | | | | | | | |
| *Age* | 33 | 45 | 59 | 44 | 29 | 21 | 21 | 51 | 42 | 21 |
| *BMI* | 43.1 | 30.9 | 30.1 | 27.6 | 41.3 | 23.2 | 25.4 | 36.6 | 29.3 | 22.1 |

### Part B: Classification (25 points)

Table 2 presents dataset containing details of activities at a waste water treatment plant for 13 days. Each day is described in terms of six descriptive features that are generated from different sensors at the plant.

- *SS-IN - solids coming into the plant per day*
- *SED-IN - sediment coming into the plant per day*
- *COND-IN - electrical conductivity of the water coming into the plant.*
- *SS-OUT, SED-OUT & COND-OUT - water flowing out of the plant*
- *STATUS – this is the target. Gives the current situation at the plant*

*Table 2: Waste Water Treatment*

| ID | SS -IN | SED -IN | COND -IN | SS -OUT | SED -OUT | COND -OUT | STATUS |
|---|---|---|---|---|---|---|---|
| 1 | 168 | 3 | 1,814 | 15 | 0.001 | 1,879 | ok |
| 2 | 156 | 3 | 1,358 | 14 | 0.01 | 1,425 | ok |
| 3 | 176 | 3.5 | 2,200 | 16 | 0.005 | 2,140 | ok |
| 4 | 256 | 3 | 2,070 | 27 | 0.2 | 2,700 | ok |
| 5 | 230 | 5 | 1,410 | 131 | 3.5 | 1,575 | settler |
| 6 | 116 | 3 | 1,238 | 104 | 0.06 | 1,221 | settler |
| 7 | 242 | 7 | 1,315 | 104 | 0.01 | 1,434 | settler |
| 8 | 242 | 4.5 | 1,183 | 78 | 0.02 | 1,374 | settler |
| 9 | 174 | 2.5 | 1,110 | 73 | 1.5 | 1,256 | settler |
| 10 | 1,004 | 35 | 1,218 | 81 | 1,172 | 33.3 | solids |
| 11 | 1,228 | 46 | 1,889 | 82.4 | 1,932 | 43.1 | solids |
| 12 | 964 | 17 | 2,120 | 20 | 1,030 | 1,966 | solids |
| 13 | 2,008 | 32 | 1,257 | 13 | 1,038 | 1,289 | solids |

Assuming the dataset is normalized, create a Naive Bayes model to predict the output for following query:

- ***SS-IN = 222, SED-IN = 4.5, COND-IN = 1,518, SS-OUT = 74 SED-OUT = 0.25, COND-OUT = 1,642***

### Part C: Clustering (25 points)

1) Consider the following "data" to be clustered: **10 20 40 80 85 121 160 168 195.** For each part of the problem, assume that Euclidean distance will be used to measure the distance between the data points. Use hierarchical

agglomerative clustering with single linkage to cluster the data. Draw a dendrogram to illustrate your clustering and include a vertical axis with numerical labels indicating the height of each parental node in the dendrogram (10 points)

2) If DBSCAN algorithm is applied with similarity threshold of 0.8 (using the similarity matrix) **or** *eps* of 0.2 (for the dissimilarity matrix), and $MinPts \geq 2$ (required density), what are core, border, and noise points in the set of points $p_i$ given in table 3. Explain (15 points).

*Table 3*

| | p1 | p2 | p3 | p4 | p5 |
|---|---|---|---|---|---|
| p1 | 1.00 | 0.10 | 0.41 | 0.55 | 0.35 |
| p2 | 0.10 | 1.00 | 0.64 | 0.47 | 0.98 |
| p3 | 0.41 | 0.64 | 1.00 | 0.44 | 0.85 |
| p4 | 0.55 | 0.47 | 0.44 | 1.00 | 0.76 |
| p5 | 0.35 | 0.98 | 0.85 | 0.76 | 1.00 |

**Part D: Performance Evaluation (25 points)**

1) Two models are applied to a dataset that has been partitioned. Model A is considerably more accurate than model B on the training data, but slightly less accurate than model B on the validation data. Which model are you more likely to consider for final deployment? (10 points)
2) A large number of insurance records are to be examined to develop a model for predicting fraudulent claims. Of the claims in the historical database, 2% were judged to be fraudulent. A sample is taken to develop a model, and oversampling is used to provide a balanced sample in light of the very low response rate. When applied to this sample (n = 1600), the model ends up correctly classifying 620 frauds, and 540 non-frauds. It missed 180 frauds, and classified 260 records incorrectly as frauds when they were not. (15 points)
    a. Produce the confusion matrix for the sample as it stands.
    b. Find the adjusted misclassification rate (adjusting for the oversampling).
    c. What percentage of new records would you expect to be classified as fraudulent?