Assignment 1

Note: You can do manual calculations, use R, use Python, or use any other software (e.g., Weka, Excel, JMP) to answer the questions unless otherwise noted. As explained in submission guidelines below, please provide your "final answers" first, and then provide an appendix, additional file(s), or screenshot(s) that show how you obtained your answers.

Problem 1. Install Weka (10 points)

- Complete WEKA installation using the attached installation instruction.
- Start Weka and start Explorer.
- Open the *iris.arff* dataset. Provide a snapshot of the resulting screen in your submission.

Problem 2. Install JMP Pro (10 points)

- Complete JMP installation using the attached installation instruction.
- Follow the instruction and capture the last screen and include it in your submission file.

Problem 3 (**24 points)** Consider the dataset *a*1.*csv* which is posted along with this assignment. It has 100 instances and 5 attributes.

(1). Calculate the mean, median, and standard deviation (sample) of the attribute *A*5.
(2). Determine Q1, Q2, and Q3 of *A*5.
(3). Detect outliers using the IQR method, which we discussed in the class, and show the *A*5 values of the detected outliers. When detecting outliers, use only the *A*5 values.
(4). Plot the boxplot of the attribute *A*5. In your boxplot, you need to show outliers separately.

Note: You may use any tool to determine mean, median, standard deviation, Q1, Q2, and Q3. However, when you detect outliers, you must do it manually using the method we discussed in the class.

Problem 4 (14 points). Consider the following dataset that has some information about 10 people.

ID	job	marital	education	default	housing	loan	contact
P1	unemployed	married	primary	no	no	no	cellular
P2	services	married	secondary	no	yes	yes	cellular
P3	management	single	tertiary	no	no	no	cellular
P4	management	married	tertiary	no	yes	yes	unknown

P5	blue-collar	single	secondary	no	yes	no	unknown
P6	management	single	tertiary	no	no	yes	cellular
P7	self-employed	married	tertiary	no	yes	no	cellular
P8	technician	married	secondary	no	yes	no	cellular
P9	entrepreneur	married	tertiary	no	yes	no	unknown
P10	services	married	primary	no	yes	yes	cellular

Calculate the distance between P4 and P5, *d*(P4, P5), and the distance between P4 and P6, *d*(P4, P6). Is P4 closer to P5 or P6? Here, all attributes are nominal attributes.

Problem 5 (14 points). Consider the following dataset with two objects.

Object	A1	A2	A3	A4
O1	43	14	32	21
O2	25	20	15	14

(1). Calculate the distance between O1 and O2 using the Manhattan distance.
(2). Calculate the distance between O1 and O2 using the Euclidean distance.

Problem 6 (14 points). Consider the following dataset with two objects.

Object	A1	A2	A3	A4
O1	2	second	silver	small
O2	4	first	gold	large

Here, all attributes are ordinal attributes and ranks of their values are shown below (lowest rank on the left):

A1: {1, 2, 3, 4}
A2: {first, second, third}
A3: {bronze, silver, gold}
A4: {small, medium, large}

Calculate the distance between O1 and O2 using the method that we discussed in the class. Use the Euclidean distance measure.

Problem 7 (14 points). Consider the following dataset.

Document	apple	orange	banana	pear	lemon	tomato	grape	berry	pineapple	mango
D1	5	1	2	1	3	2	4	2	3	1
D2	2	1	0	0	4	2	3	3	1	1
D3	2	0	1	0	3	1	0	2	3	2
D4	1	3	4	3	0	3	5	0	4	0

Calculate the similarity between D1 and D2, *cos*(D1, D2), and the similarity between D1 and D3, *cos*(D1, D3), using the cosine similarity measure. Is D1 closer to D2 or D3? You must calculate the cosine similarity yourself (i.e., you must not use a built-in function of a software).

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. Make sure the filename includes your name.**
- **Your submission should be organized well and easy to follow. Clearly list which question you are answering and provide the answer requested (one or more numbers, a screenshot, a plot, an explanation, whatever the question asks you to do). If a question contains multiple parts, you should clearly provide your answer for each part.**
- **If a problem requires you to work out formulas or calculations, you should provide this detailed work either in an Appendix in your Word/PDF or as a separate file. All separate files should include your name in the filename.**
- **If you need more than 3-4 separate files, you might ZIP/RAR these files. Please still upload the Word/PDF separately. The separate Word/PDF greatly simplifies the facilitator's ability to comment on your assignment.**
- **If your facilitator tells you to submit the files differently than the above guidelines, you are expected to respect your facilitator's wishes starting on the next assignment.**
- **Facilitators can deduct up to 20% if you fail to follow these requirements (more if the questions are not actually answered).**
- **Facilitators can deduct 5% for each day the assignment is late.**
- **Unless your facilitator or the professor agrees, your assignment will not be graded if it is more than 3 days late (e.g., no credit will be given after Friday at 6 AM Boston time). The professor will usually ask the facilitator to make the decision but in rare cases (<1% of the time) has overridden a facilitator. Do not expect the professor to override in most cases.**