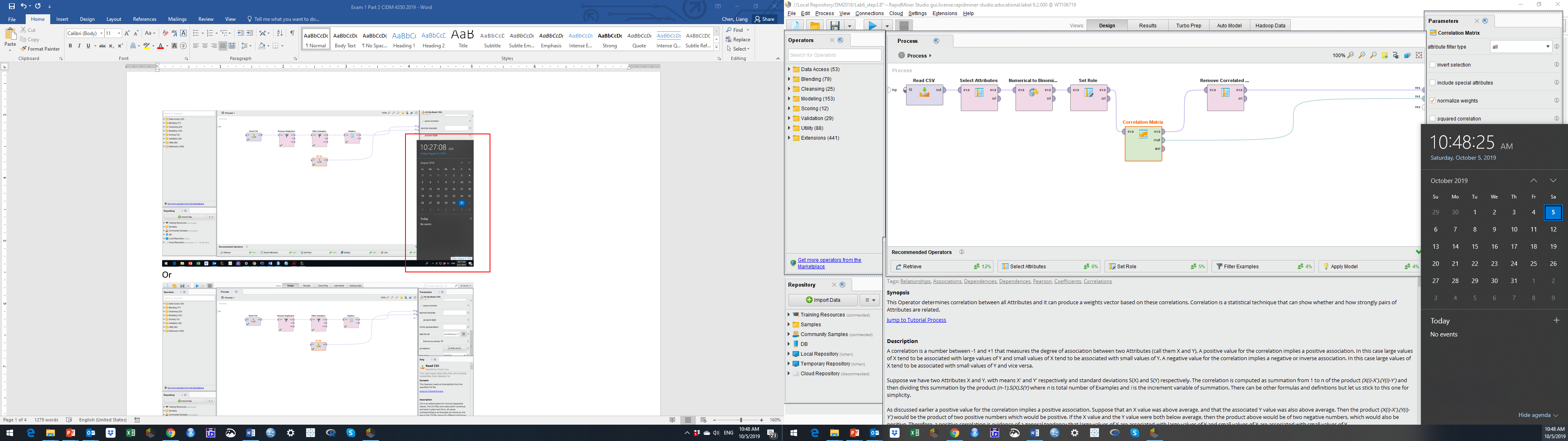
**CIDM 6355 Data Mining Methods Exam 1 Part 2 Submission**

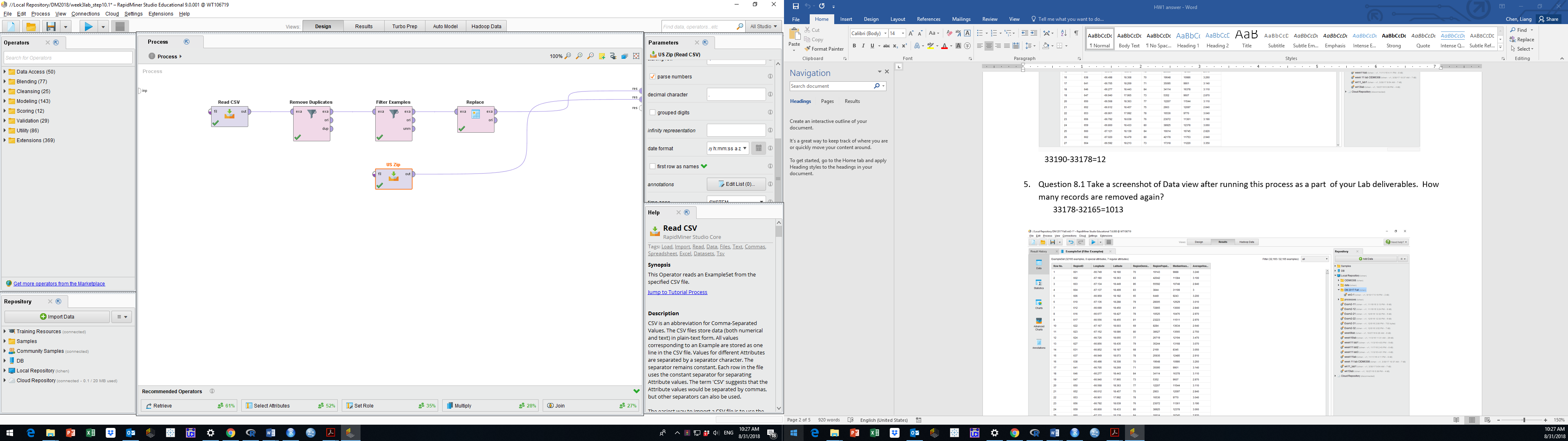
**(50 points in total; due 11:59 pm CDT, March 4, 2024)**

This exam is open book, open slides, and open notes, but you are not allowed to collaborate nor discuss with anyone else. Sharing your screenshots, RM processes, R script, or answers with other students are considered as cheating for this exam. Should you have any question or unclarity, please contact with the instructor. Please put all your deliverables in a word document and submit it to WT class before the deadline; Make sure that all your screenshots include dates and time [see the examples as below]; otherwise, a penalty of 50% of your grade will be applied. Please type your name below to indicate whether or not you have understood and complied with such requirements in this exam.

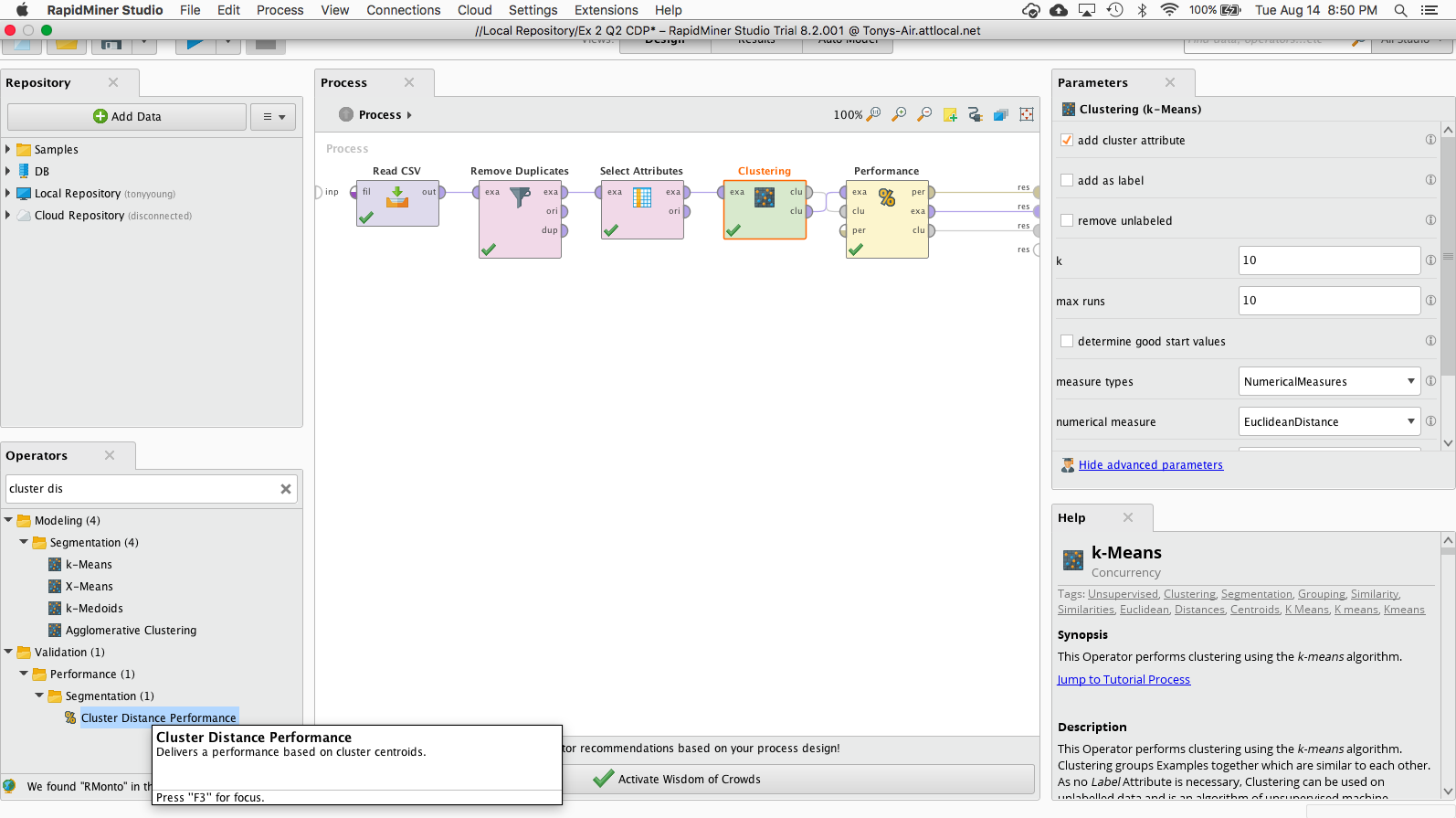
Your name (First Last): Mehnaz Afrose



Or



Or on Mac



**PLEASE MAKE SURE THAT YOU TYPE YOUR NAME AT THE FIRST PAGE; OTHERWISE, YOUR SUBMISSIN WILL NOT BE GRADED AND A ZERO POINT WILL BE ASSIGNED.**

2 Screenshots in RM (6 points for each: 3 pts for your screenshot and 3 pts for your description/discussion).

Screenshot 1 with description (6 pts): A screenshot of your decision tree graph with date and time at Step 2.3 and briefly describe your model. Your description must include root node, split nodes, and leaf nodes.

A screenshot of a computer

Description automatically generated

**Description:**  This decision tree model includes one root nod, with other five split nods or internal nods. It has seven leaf nodes. Among these seven leaf nods, only one leaf nod is completely homogeneous, which has “yes” feature. This homogeneous nod has 14 items. The leaf node with the biggest size (the highest number of items) is not a homogeneous nod. It is featured in “no”. This heterogeneous nod has 953 items, among which the ratio of being “no” is 73.76%. In this decision tree model, the measures of OpenPrice, SellerRating and Duration play the determining role in classifying whether Competitive should be "yes", or "no".

Screenshot 2 with your discussion (6 pts): A screenshot of your RapidMiner Process (the flow chart in your design mode) with date and time and briefly discuss why the operator Nominal to Numerical must be used in your process.

A screenshot of a computer

Description automatically generated

**Why the operator Nominal to Numerical must be used:** There is a presence of polynomial attributes in my dataset. Neural Net operator doesn’t have sufficient capabilities to handle polynomial attributes. Polynomial attributes are not supported in Nural Net operator. That is why we need to use Nominal to Numerical operator. The “Nominal to Numerical" operator replaces the categorical or polynomial values with numerical values, and then we can continue data analysis or modeling process.

5 Screenshots in R: Each screenshot must include complete and correct code description or comments (20 pts in total and 4 points for each).

Screenshot 3: A screenshot of your R codes with date and time to show how you import and prepare the data for modeling and prediction, that is, Steps 6.1-6.3.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Screenshot 4: A screenshot of your R codes with date and time to show Step 6.4.1-6.4.3. Requirements: your screenshot must clearly include all the R codes for your decision tree model and the output of 6.4.3.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Screenshot 5: A screenshot of your R codes with date and time to show Step 6.5.1-6.5.3. Requirements: your screenshot must clearly include all the R codes for your NB model and the output of 6.5.3.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Screenshot 6: A screenshot of your R codes with date and time to show Step 6.6.1-6.6.4. Requirements: your screenshot must clearly include all the R codes for your logistic regression model and the output of 6.6.4.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Screenshot 7: A screenshot of your R codes with date and time to show Step 6.7.1-6.7.4. Requirements: your screenshot must clearly include all the R codes for your NN model and the output of 6.7.4.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Step 7:** **Comparative Analysis** (18 points)

7.3. Please include the following deliverables in your submission:

7.3.1. Please copy and paste the provided table into your submission (your table must be accessible; screenshot or image is not accepted). Ensure that the table includes the predicted results of 20 records using 8 different methods (8 pts in total and each column is worth 1 point).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | RM\_DT | RM\_NB | RM\_LR | RM\_NN | R\_DT | R\_NB | R\_LR | R\_NN |
| 1 | No | no | no | no | No | no | No | No |
| 2 | Yes | yes | yes | yes | Yes | yes | Yes | No |
| 3 | No | yes | yes | yes | Yes | yes | Yes | Yes |
| 4 | Yes | yes | yes | yes | Yes | yes | Yes | Yes |
| 5 | Yes | no | no | yes | No | no | No | No |
| 6 | No | yes | yes | no | Yes | yes | Yes | Yes |
| 7 | No | no | no | no | No | no | No | No |
| 8 | Yes | no | yes | yes | Yes | no | No | Yes |
| 9 | No | no | no | yes | No | no | no | No |
| 10 | Yes | no | yes | yes | Yes | no | no | Yes |
| 11 | No | yes | yes | yes | Yes | yes | yes | Yes |
| 12 | Yes | yes | yes | yes | No | yes | yes | No |
| 13 | Yes | no | no | no | No | no | no | Yes |
| 14 | No | no | no | yes | No | no | no | Yes |
| 15 | Yes | yes | yes | yes | Yes | yes | yes | Yes |
| 16 | Yes | yes | yes | yes | Yes | yes | yes | No |
| 17 | No | no | no | no | No | no | no | No |
| 18 | Yes | yes | yes | yes | No | yes | yes | Yes |
| 19 | No | yes | yes | yes | Yes | yes | yes | Yes |
| 20 | No | no | no | yes | No | no | no | Yes |

7.3.2. Discuss the number of records predicted to be "yes" or “no” by each method in the RM and R datasets. For example, among the 20 records, RM\_DT and R\_DT jointly predict "yes" for 6 records (ID =2, 4, 8, 10, 15, 16), and jointly predict "no" for 6 records (ID =1, 7, 9, 14, 17, 20). (8 pts: 2 pts for each pair of methods)

**Discussion:**

Among the 20 records, RM\_DT and R\_DT jointly predict "yes" for 6 records (ID: 2, 8, 10, 15, 16), and jointly predict "no" for 6 records (ID =1, 7, 9, 14, 17, 20).

Among the 20 records, RM\_NB and R\_NB jointly predict "yes" for 10 records (ID: 2, 3, 4, 6, 11, 12, 15, 16, 18, 19), and jointly predict "no" for 10 records (ID =1, 5, 7, 8, 9, 10, 13, 14, 17, 20).

Among the 20 records, RM\_LR and R\_LR jointly predict "yes" for 10 records (ID: 2, 3, 4, 6, 11, 12, 15, 16, 18, 19), and jointly predict "no" for 8 records (ID = 1, 5, 7, 9, 13, 14, 17, 20).

Among the 20 records, RM\_NN and R\_NN jointly predict "yes" for 10 records (ID: 3, 4, 8, 10, 11, 14, 15, 18, 19, 20), and jointly predict "no" for 3 records (ID = 1, 7, 17).

7.3.3. Finally, provide an analysis of the number of records that all eight models predict as "yes" and the number of records that all eight models predict as "no." For example, all eight models jointly predict as "yes" for 2 records (ID =4, 15), and "no" for 3 records (ID =1, 7, 17). (2 pts).

**Answer:** All eight models jointly predict as "yes" for 2 records (ID =4, 15), and "no" for 3 records (ID =1, 7, 17).

7.3.4. Bonus Question (5 bonus pts): Discuss why some records receive consistent predictions across all

methods, while others yield varying results. Please include specific records and their corresponding

values in your discussion.

**Answer:**

As we can see from the result table is that all eight models jointly predict as "yes" for 2 records (ID =4, 15) and "no" for 3 records (ID =1, 7, 17). This indicates that these records have clear, straightforward patterns that all eight of the models can recognize easily. The features for these records work well with the all the eight models, making predictions stable and reliable. Essentially, these are instances where the data characteristics align smoothly with the models, resulting in consistent and accurate predictions.

On the other hand, other records yield varying results. This can happen for various reasons. Each model has its own assumptions and limitations. Say for example, Decision trees capture complex interactions, while Naive Bayes assumes independence. Logistic regression models linear relationships. Neural networks are highly flexible but prone to overfitting. High bias models (underfitting) or High variance models (overfitting) model may yield varying results. These levels of complexity and differences of the models can cause predicting different results using the same dataset.

Also, there can be a presence of noise or outliers, there are non-linear relationships between features and the target variable, these may yield varying results.