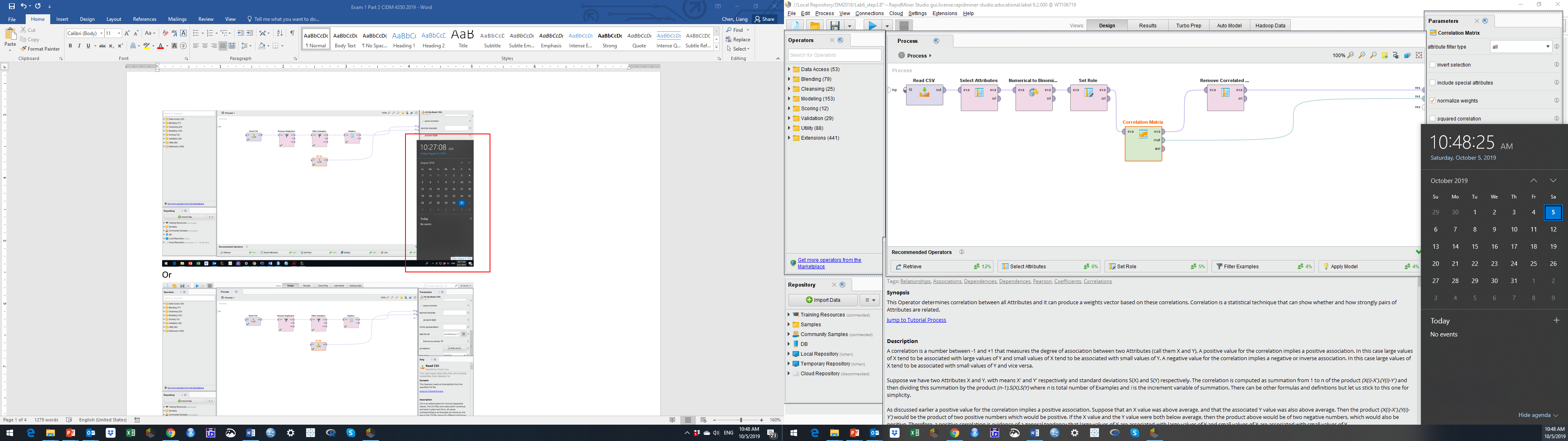
**CIDM 6355 Data Mining Methods Exam 1 Part 2 Submission**

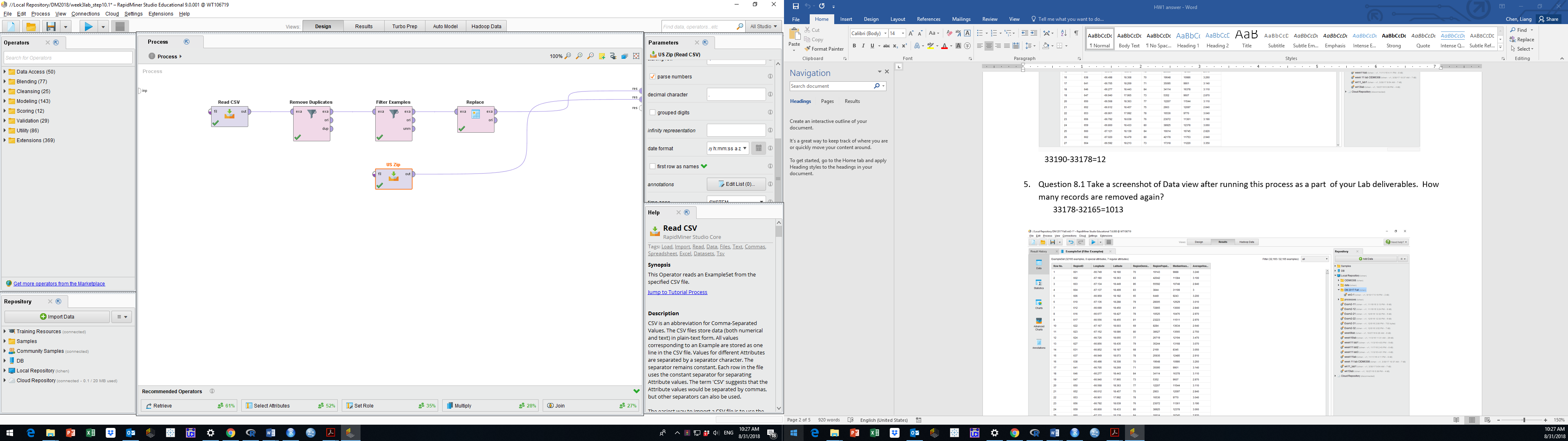
**(50 points in total; due 11:59 pm CDT, March 4, 2024)**

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.

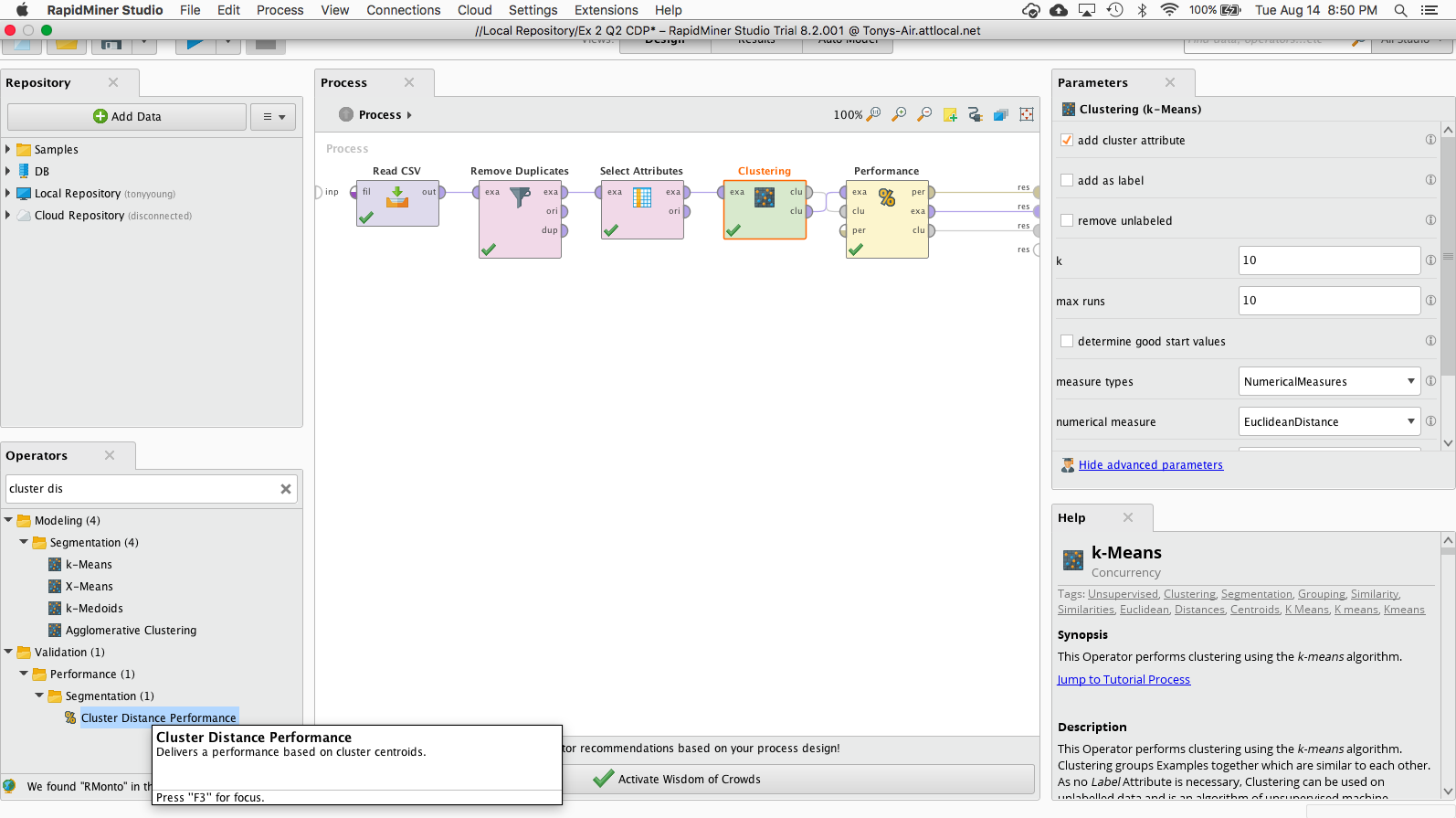
Your name: Mariam Adegbindin

l

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

The root node is 'OpenPrice'. It's the very first node that splits the data into two paths based on whether the price is less than or equal to 1.245 or greater. The split nodes include 'OpenPrice' at the top level (also the root), 'SellerRating' at the second level (which splits again based on different thresholds), and 'Duration' at the third level for one of the branches.

These nodes are where decisions are made to further branch out based on the conditions met. The leaf nodes are the end points with outcomes "yes" or "no". They are the nodes that do not split any further and provide the final prediction of the decision tree. For instance, in the path where 'OpenPrice' is less than or equal to 1.245 and 'Duration' is less than or equal to 6, the leaf node would be the "yes" outcome. Similarly, for 'OpenPrice' greater than 1.245 and 'SellerRating' less than or equal to 17.500, the leaf node would also be a "yes" outcome.

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

The "Nominal to Numerical" operator is crucial for converting data into a format that algorithms requiring numerical values can process. Machine learning methods, such as neural networks, are unable to process nominal or categorical data directly, as this type of data reflects categories rather than numerical values.

The "Nominal to Numerical" operator prepares the dataset for the "Neural Net" operator to effectively train a model. Without this stage, the neural network would struggle to process nominal data, perhaps leading to mistakes or hindering model development.

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

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

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

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

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

**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).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | ID | 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 | yes | | 6 | no | yes | yes | no | yes | yes | yes | no | | 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)

RM\_DT and R\_DT:

Joint predict "yes" for 6 records (IDs 2, 4, 8, 10, 15, 16)

Joint predict "no" for 6 records (IDs 1, 7, 9, 14, 17, 20)

RM\_NB with R\_NB:

Joint predict "yes” for 9 records (IDs 2, 3, 6, 11, 12, 15, 16, 18, 19)

Joint predict "no “for 8 records (IDs 1, 5, 7, 9, 13, 14, 17, 20)

RM\_LR with R\_LR**:**

Joint predict "yes" for 11 records (IDs 2, 3, 6, 8, 10, 11, 12, 15, 16, 18, 19)

Joint "no" for 9 records (IDs 1, 4, 5, 7, 9, 13, 14, 17, 20)

RM\_NN with R\_NN:

Joint "yes" for 8 records (IDs 3, 4, 5, 8, 10, 13, 14, 20)

Joint "no" for 9 records (IDs 1, 2, 6, 7, 9, 12, 15, 16, 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).

“Yes” Predictions (ID = 15, 4)

“No” Predictions (ID= 17, 7, 1)

7.3.4 Bonus Question: 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.

When a set of attributes for a record (e.g., "yes" or "no") firmly indicates a specific class, it is possible for all models to readily recognize this indication and reach a consensus regarding the prediction. For example, if the feature values of ID 4 and ID 15 are universally associated with a "yes" outcome across all models, it indicates that these records possess distinct attributes that are consistent with the "yes" prediction. This may encompass specific numerical features assigned high values or categorical features assigned to particular groups that are universally interpreted as reliable indicators of the "yes" class.

On the other hand, models may interpret records with feature values that are more ambiguous or lie within the overlap of class distributions in varying ways, contingent upon their approach to data separation and complexity. As an illustration, records such as ID 3 and ID 19 are subject to divergent predictions among the models due to the lack of clear distinction between their feature values, which categorizes them into distinct classes.

Certain models exhibit greater sensitivity to data disturbance or outliers. Depending on how they manage outliers, models may generate predicted values for outlier records that vary. For instance, models that possess robustness to anomalies, such as random forests, may generate accurate predictions even if ID 6 is an outlier, whereas others, such as naive Bayes, may fail to do so.