Florida Atlantic University
COLLEGE OF BUSINESS

ISM 6136 – Datamining/Predictive Analytics

Project – Data Mining/Machine Learning
20 points

Task: Perform supervised machine learning/data mining on a dataset of your own choice, using TWO different appropriate datamining algorithms learnt in the course and determine the best model. You can use either Analytic Solver (XLMiner) '<u>or</u>' RapidMiner tool.

<u>Procedure to be followed:</u>

A. Select a dataset having minimum 500 rows of data and should have at least 6 predictors (*after cleanup*).
   - Here are some examples of websites with datasets:
     - *data.gov*
     - *kdnuggets.com*
     - *kaggle.com*
   - You can also use any Generative AI tools to search for appropriate datasets or websites that have downloadable datasets:
     - Microsoft Copilot: https://copilot.microsoft.com/
     - ChatGPT: https://chat.openai.com/
     - 'Prompts' to use for the search to be efficient:

       Need dataset links for machine learning. Dataset needs to be in .xls or .csv format and should have at least 500 rows and at least 10 variables.

       Need dataset links from Kaggle.com which have at least 500 rows and at least 10 variables.

       Need datasets which have at least 500 rows and at least 10 columns to be used for supervised machine learning algorithm.

- NOTE: Use AI generated content responsibly: You are responsible for the validity of the dataset selected!! Visit the website where the dataset resides/posted and verify the selection to make sure that it not a fabricated/hallucinated result by AI.

- **Or** use any of your 'work-related' dataset (need your supervisor's approval email attached).

B. Separate out 50 or more rows as 'new data' and blank out its outcome variable column. You cannot use the same Dataset already used by someone in the class. Check the Discussion on Canvas to see if the dataset has been reserved by someone.

Follow these data mining steps:

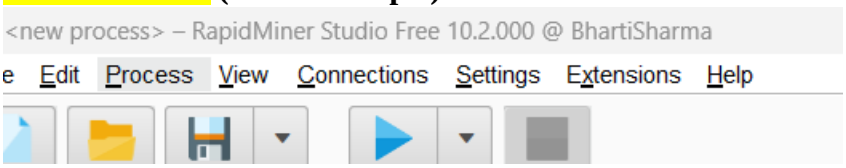
1. Explain the problem and purpose of the data mining task in at least 3 sentences (1 point)
2. Obtain the dataset for analysis: Paste the link (URL) from where you obtained the dataset. (0.5 points)
3. Explore, clean and preprocess data:
    a. Cleanup any column that is not a predictor *(here you can explain how you eliminated some of the columns that were not meaningful as a predictor)*
    b. Perform 'Missing Data Handling'
    c. For continuous numerical outcome: Explore using Scatter Plot Matrix between continuous numerical predictors and the outcome (and remove predictors that are not correlated to the outcome). *If you have categorical outcome then you do not need to perform Scatter Plot.*
    d. Do you have any categorical variables ? Do you need to convert them to dummies ? Yes or No …explain why ? (1 points)
4. Reduction of data dimension – Use **PCA technique** or **Classification/Regression Tree** technique **on any one of the models in both the algorithms** to reduce the dimensions. Select top 5 dimensions. Explain the technique you used. (3 points)
5. Partition the data accordingly (2 points)
6. Choose the data mining techniques/algorithm to apply Classify or Predict and explain the reason for selecting the two algorithms (0.5 points)
7. **Try at least 6 models** for each algorithm and select the best model by interpreting results of algorithm and explain your model selection criteria in the Model Table form as always done in class. Model Table for each of the two algorithms. (10 points)
8. Deploy the two models (best one from each algorithm) and compare and explain the predicted results (2 points)

9. Explain steps 1 through 8 in a Word document with screenshots.

10. If you use XLMINER - Upload two Excel workbooks (one for each algorithm) and the Word document on Canvas.

    If you use RAPIDMINER – Upload only the Word document with screenshots for all the model designs and their confusion matrix/R2 RMSE values and other info covering steps 1 through 8.

    NOTE: All the screenshots need to capture the date and time and your name on the Excel sign in (Analytic Solver) or if using RapidMiner need your name captured on the screenshots (below example).



Additional Notes:

After you select a dataset:

1	**Look at the Outcome column to be predicted**
	If it is categories - Classification
	If continous numerical $, height, weight, measurement - Regression
2	**What are predictors?**
	If they are all categorical columns - Naïve Bayes and all other algorithms - make sure they are 'numbers' and not text
	If one of the columns is continous numerical - then Naïve Bayes will not work - you will have to 'bin' it to make it work.