BIS2216 Data Mining and Knowledge Discovery Fundamentals
Semester August 2020
Coursework (15% of Total Assessment)

This is an individual work.

Use ***student_performance.xlxs*** dataset for this assessment.

You are required to build models to predict the result of students based on their demographics, social and school related attributes. Detailed attributes information as follows:

- gender - student's sex (binary: 'F' - female or 'M' - male)
- age - student's age (numeric: from 15 to 22)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - primary education (5th to 9th grade), 3 - secondary education or 4 - higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - primary education (5th to 9th grade), 3 - secondary education or 4 - higher education)
- Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- internet - Internet access at home (binary: yes or no)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)
- result - final grade (numeric: from 0 to 20, output **target**)

For modelling techniques, you must use Decision Tree and Regression, then select the best model. You can choose to build many different Decision Tree and Regression models with different parameters testing. Suggestion: build 7-10 models, then compare their performance to choose the best model.

Present your answer according to the requirements listed below, i.e., you DO NOT need to present other information other than the report content required as follows:

<table>
<tr><th>No.</th><th>Content to Report</th></tr>
<tr><td>1.</td><td>Data preparation: Which attributes you performed pre-processing? State your rationale. For each attribute you pre-processed, which method and how did you do it? NOTE: if you do not perform any data preparation, then just leave this section blank.

Present your answers in a tabular form as follows:
<table>
<tr><th>Attribute processed</th><th>Rationale for processing</th><th>Methods for processing</th></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
</table></td></tr>
</table>

2. Present your modelling methods and the model performance in a tabular form as follows.

For example:

No.	Modelling technique	Partition ratio	Other preparation methods applied	Model performance (misclassification /error)
1.	Decision Tree			
2.	Decision Tree			
3.	Regression			
4.	Regression			
…..	…..	…..		

3. Screenshot the Model Comparison Node's `Fit Statistics` result using SAS Mining. For example:

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	HPDMForest2	HPDMFore...	HP Forest	Status	Status	0.013931
	Tree	Tree	1 DT_NoVa...	Status	Status	0.017234
	HPSVM2	HPSVM2	HP SVM	Status	Status	0.017665
	HPBNC3	HPBNC3	HP BN Cla...	Status	Status	0.017665
	Tree2	Tree2	2 DT_Valid...	Status	Status	0.017952
	Reg	Reg	Regression	Status	Status	0.018096

Note: You do not need to show the complete screenshot (due to too lengthy), just ensure you have at least the columns as shown in the above example.

4. Present the best model.

5. Interpret the best mode selected.

6. Present the screenshot of the whole modelling process diagram in SAS Miner.

Submission due date: by the **22nd November 2020 (Sunday)**

Submission format:

- Include the above stated 6 requirements' content into a **PDF file (maximum 2 pages)**. Use your ID and name as the filename, for example, *11232612JohnSmith.pdf*
- Submit through the eLearn portal.