MIS772 Predictive Analytics – Sample exam only

Your exam will take place in the form of an online quiz.
You will receive more information about the date, time and the link to the quiz.
Exam consists of 5 sections, each of 2-3 questions.
The following sample questions relate to the case study provided in a separate document.
Under each question you will have a text box, where you'd be asked to type in your answer.
Follow the online exam instructions exactly!

Please use the attached case study for answering Sections 1 to 5 – Link to the PDF file.pdf

Section 1 of 5 (Question 1 to 3 - Concepts)

1) Identify two different methods of **k-NN model optimisation**.

In the space provided, describe how they are used and explain their main differences.

- *M1. First Method … (2 marks)*
- *M2. Second Method … (2 marks)*

Identify two significant differences between M1 and M2:

- *AA. First Difference … (1 mark)*
- *AB. Second Difference … (1 mark)*

<Answer Box>

2) Identify two different approaches to **dealing with multicollinearities**, explain their use and their limitations in the space provided.

Approach A:

- *AD. It is… (1 mark)*
- *AU. It is used to… (1 mark)*
- *AL. Its main limitations are… (1 mark)*

Approach B:

- *BD. It is… (1 mark)*
- *BU. It is used to… (1 mark)*
- *BL. Its main limitations are… (1 mark)*

<Answer Box>

3) In the context of model **Validation**, define these concepts and explain how they are used in the development of a classifier: *CA. **Holdout validation**, CB. **k-Fold Cross-Validation**, CC. **Bootstrapping Validation**, CD. **LOOCV**.*

Provide the following information:

- *Concept CA is defined as … and is used to … (2 marks)*
- *Concept CB is defined as … and is used to … (2 marks)*

- *Concept CC is defined as … and is used to … (2 marks)*
- *Concept CD is defined as … and is used to … (2 marks)*

<Answer Box>

Total for section 1 of 5: 6 + 6 + 8 = 20 Marks

Section 2 of 5 (Question 4 to 6 - Statistical Models)

Consider the business case explained at the top of this paper.
While developing a predictive model, your colleague suggested that the label attribute can be effectively estimated with **linear regression** using a subset of the relevant predictors. She created such a model in RapidMiner and the results have been provided to you (in your answers refer to Figures 2A, 2B, ...).

4) *Considering the table of coefficients included in the figures:*

- *IA. The observed **coefficients** indicate that (make your answer specific) … (3 marks)*
- *IB. The observed **p-values** indicate that (make your answer specific) ... (3 marks)*
- *IC. The observed **tolerance values** indicate that (make your answer specific) ...(3 marks)*
- *O. The reported **table of coefficients** indicates that this specific model is … (1 mark)*

<Answer Box>

5) Describe the distribution of residuals and the chart of predicted vs. actual label values, included in Figures 2A, 2B, ...

- *AA. The **shape of residuals distribution** is (make your answer specific) ... (3 marks)*
- *AB. **Outliers** are (make your answer specific) … (3 marks)*
- *AC. Based on the **regression assumptions**, this model is … (4 marks)*

<Answer Box>

6) Based on the reported performance indicators and other relevant figures, give two suggestions on how to improve this specific model, justify.

- *CA. **The first suggested change** … (5 marks)*
- *CB. **The second suggested change** … (5 marks)*

<Answer Box>

Total for section 2 of 5: 10 + 10 + 10 = 30 Marks

Section 3 of 5 (Question 7 to 9 - Machine Learning Models)

Consider the business case explained at the top of this paper.
Your team has decided to explore the data using cluster analysis (see Figures 3A, 3B, …) and then construct a classifier. They have tried ***three different models***. Now they have asked you to select the best performing model based on the supplied evaluation results (in your answers refer to Figures 4A, 4B, ...).

7) Interpret the clustering system and the way it was developed.

- *MB. This clustering system was visualised using SVD, which is a more general form of PCA, explain how to **interpret the Cumulative Variance Plot** (here called Cumulative Proportion of Singular Values) … (2 marks)*
- *M1. Select and describe **two clusters using the centroid chart** … (3 marks)*
- *M2. Interpret the **cluster scatter plot** … (3 marks)*

<Answer Box>

8) The following four operators were used in the process. Very briefly explain the **aims** (what) and **the specific reasons** (why) of using the following operators in that process:

- *O1. The aim of "**Normalize**" is to … because … (1 marks)*
- *O2. The aim of "**Replace Missing Values**" is to … because … (1 marks)*
- *O3. The aim of "**Filter Examples**" is to … because … (1 marks)*
- *O4. The aim of "**SMOTE**" is to … because … (1 marks)*

<Answer Box>

9) Suggest which of the currently used models has the best performance and why. Then, suggest a single most important change to the RapidMiner process for each of the three models to improve its performance.

- *MB. Currently the **best** and the **worst** performing models are … because … (2 marks)*
- *M1. The performance of the **best** model can be improved by … (3 marks)*
- *M2. The performance of the **worst** model can be improved by … (3 marks)*

<Answer Box>

Total for section 3 of 5: 8 + 4 + 8 = 20 Marks

Section 4 of 5 (Question 10 to 11 - Modelling)

Consider the business case explained at the top of this paper.
It has been suggested that the text attributes included in the provided examples could offer additional insights. Develop a RapidMiner process responsible for ***Outlier Detection*** using a ***mix of structured and text attributes***. Add operators to facilitate ***outlier elimination***. Extend your process to ensure that all of its models can be ***saved for later deployment*** (do not include the actual deployment model). Do not attach any drawings of the process, just answer the following questions.

10) Consider ***the main process***, then ...

- PP. Provide a list of operators responsible for ***data pre-processing***, for each identify its name ... role … and in brackets their main parameters (4 marks)
- TP. Provide a single operator (in the main process) responsible for ***text processing***, for each identify its name ... role … and in brackets their main parameters (4 marks)
- DC. Provide a list of operators for ***outlier detection and elimination***, for each identify its name ... role … and in brackets their main parameters (4 marks)
- DR. Explain **how you would store the results of the trained model so it can be deployed by the organisation** (4 marks)

<Answer Box>

11) Provide details of the sub-process responsible for ***parsing all text fields*** of the examples.

- TA. ***Provide a list of operators*** responsible for text processing within the sub-process of the "Process Documents from Data" operator, for each identify its name ... role … and in brackets their main parameters (if any). (10 marks)
- TB. ***Explain the parameters*** of the "Process Documents from Data" operator (do not describe the operators of its internal sub-process). (4 marks)

<Answer Box>

Total for section 4 of 5: 16 + 14 = 30 Marks

Section 5 of 5 (Question 12 to 13 - Evaluation and Optimisation)

Consider the business case explained at the top of this paper.
Your colleague developed a RapidMiner process which produced optimisation charts listed in Figures 5A, 5B... Unfortunately, she left the company and your job is to quickly explain and then replicate her work.

12) Explain the design of a workflow that must be placed within the operator “Optimize Parameters (Grid)”, where you will have to use holdout validation for the sake of efficiency.

- *WF. **Explain the logic** of the workflow in a short paragraph. (4 marks)*
- *OP. **Provide a list of operators** used in this workflow, for each identify its name ... role ... and in brackets the main parameters (6 marks)*

<Answer Box>

13) Explain the hyper-parameters which must have been used to generate the optimum performance measurements for the model, refer to the plots presented in Figures 5A, 5B, ...

- *O. **Identify the performance measurements** tracked and logged by the optimiser. (2 marks)*
- *TA. **Identify the operators and their parameters** that were logged, explain why it was important to experiment with them. (4 marks)*
- *TB. **Identify the specific parameter values** that would result in the model optimum performance, justify your answer. (4 marks)*

<Answer Box>

Total for section 5 of 5: 10 + 10 = 20 Marks

Exam Total: 20 + 30 + 20 + 30 + 20 = 120 Marks