Stellenbosch University

Data Science 774/874

Post-block Assignment 3

Total: [100]

Deadline: 19 May 2024, 23:55

Instructions

- This is a group assignment; you may not collaborate with anyone outside your group regarding this assignment. Your group is as per your group enrollment in this module on SUNLearn.
- Ensure your work is submitted for grading by 11:55 pm 19 May 2024. Only one member of the group is required to submit the assignment on behalf of the group. This assignment will contribute to your final mark for the module.
- Only submissions made via the provided submission link on SUNLearn will be assessed.
- The work your group submits must emanate from the combined effort of group members only. Ideas, images, and text from other individuals/sources can only be utilized if acknowledgement of this is given through adequate referencing and the resulting work remains largely your group's effort. The use of generative AI platforms is prohibited for this assignment.
- The use of data mining software such as Orange, WEKA, Rapid Miner or programming platforms such as Python or R as well as referencing software such as EndNote is required for this assignment.
- You are encouraged to practice the required presentation adequately before preparing your final deliverables for submission. Note that you are permitted only one submission for this assignment on SUNLearn.
- All group members should participate in Task A, but only group members who are registered at MEng level should proceed to Task B. Only one submission containing all the required deliverables, should be made for the group regardless of whether only Task A or Task A and B are done.

Task A – PGDip and MEng

Missing feature values need to be addressed prior to the model development phase of the CRISP-DM methodology to avoid training on incomplete data. This task assesses your ability to navigate the complexities of constructing a data pipeline to transform partially erroneous raw data to knowledge and evaluate the effectiveness of the proposed solution. The task will require you to identify a suitable publicly available dataset/s, for which there is previous research that addresses the missing feature value problem. Propose an approach to use the Naïve Bayes classifier to address the missing feature value problem in the context of categorical feature values. Implement this approach on a publicly available dataset/s and report its performance in the context of a classification problem. Compare the effectiveness of this approach against a baseline imputation approach that uses the mode value, on two different machine learning models. Justify and discuss your findings using appropriate metrics and relate your findings to previous research on data imputation using the same dataset/s. Present your findings in a written report and a video presentation. State and motivate any assumptions or scope adopted during the task.

Task B – MEng Only

This task likewise requires you to identify a suitable publicly available dataset/s, for which there is previous research that addresses the missing feature value problem. In addition to the work and deliverables of Task A, propose an approach to use the K Nearest Neighbors (KNN) classifier (use a suitable value for K and provide justification) to address the missing numerical feature value problem. An additional dataset may be used if necessary. Compare the effectiveness of this approach against a baseline numerical value imputation approach, on two different machine learning models. Justify and discuss your findings using appropriate metrics and relate the findings to previous research on data imputation using the same dataset/s. Consolidate your findings from Task A and B into a flow chart or decision tree that can be used for context-based selection of an appropriate method for handling missing values - pay particular attention to the prediction outcomes of instances with imputed feature values. Present your findings in a written report and a video presentation. State and motivate any assumptions or scope adopted during the task.

Deliverables

Expected deliverables differ depending on the task done i.e. Task A and/or Task B, as outlined below. Note that PowerPoint presentation files should NOT be submitted for this assignment.

Task A Deliverables

Your submission for this assignment must include the following:

- A video presentation (maximum 50MB file size and maximum 10 minutes recording length). Each group member should state their name and surname the first time they present during the recording.
- A report (maximum 2 pages, single line spacing, font size 11, IEEE referencing style and maximum 10 references) that outlines the tasks performed, tools and data used as well as the findings deduced.

- A zip file (maximum 20MB) containing the software project or code used in Task A and any other supplementary material related to the task.
- A text file containing all group member names, surnames, SUN numbers and levels of registration (PGDip/MEng).

Task B Deliverables

Your submission for this assignment must include the following:

- A video presentation (maximum 20MB file size and maximum 5 minutes recording length). Each group member should state their name and surname the first time they present during the recording.
- A report (maximum 2 pages, single line spacing, font size 11, IEEE referencing style and maximum 10 references) that outlines the tasks performed, tools and data used as well as the findings deduced.
- A zip file (maximum 10MB) containing the software project or code used in Task B and any other supplementary material related to the task.
- A text file containing all group member names, surnames, SUN numbers and levels registration (MEng).

Naming Convention

Your deliverables should be named using the following convention: A3videoTNumber.mp4, A3reportTNumber.docx, A3evidenceTNumber.zip, and A3membersTNumber.txt, where Number is replaced with your Group number and T is replaced by either A or B, for Task A and B, respectively. Other video formats (besides mp4) are permitted provided they can be played on the latest version of VLC media player.

Assessment Criteria

- Admin – how well the assignment instructions have been followed - 5%.
- Content – relevant approach design and correct use of and adequate motivation for the data science concepts and methods for the given task – 70%.
- Quality – well edited and formatted report – 5%.
- Group effort – demonstrated group effort and quality of explanations by all group members in presentation – 20%.

Assessment Guide

The primary deliverable that will guide the assessment is the video presentation. Other submitted deliverables should complement the presentation and serve as evidence of work authenticity. Individual group member marks may be lower than the awarded group mark in the event of poor participation in the presentation and a negative peer-review from other group members. A separate mark will be awarded for Task A and Task B. The marks for PGDip students will be based on Task A. The marks of MEng students will be weighted as follows: 0.6 x Task A + 0.4 x Task B.

Grade [Weight] → Criteria ↓	Poor [0 - 0.3]	Fair [0.4 - 0.6]	Good [0.7 - 1]	Maximum Mark

Admin	Deliverable requirements poorly met e.g. missing deliverable and ad hoc file names	Some deliverable requirements met but some outstanding e.g. file sizes marginally exceeding the specified limit.	All deliverable requirements met, and submission made timeously.	5%
Content	Incorrect application of data science concepts and erroneous interpretations. Incomplete submission.	Acceptable application of data science concepts with missing; vague motivation vague task design.	Clear articulation of utilized data science concepts/methods with firm understanding demonstrated and supported by deliverables.	70%
Quality	Poorly structured and compiled report.	Report has some grammatical or referencing errors.	Well-written report with meticulous referencing.	5%
Group effort	Poor group participation e.g. missing group members and disjointed/incomplete presentation.	Group members demonstrate unequal levels of understanding during presentation and some repetition/redundancy.	All group members demonstrate thorough understanding and ownership of work. Well-coordinated presentation.	20%