AY2025 Computer Based Assessment (CBA)

Machine Learning Detection of Misinformation

Misinformation is false or inaccurate information, regardless of the intent. The 2024 World Economic Forum Global Risks Report [1] ranked misinformation and disinformation as the most severe risk in the short term, with AI-generated misinformation and disinformation highlighted.

The Singapore government has taken measures such as the Protection from Online Falsehoods and Manipulation Act (POFMA) to protect both the public and Singapore's interests against misinformation[2].

To investigate the usefulness of Machine Learning models to detect misinformation, a data sample (misinformation2.csv) is provided[3]. The last column "is_misinformation" (0: No; 1: Yes) is the target Y variable.

1. Explore the data (without models) and report 3 notable findings.

2. Is additional data preparation or data cleaning required before using models? Explain. [Note: You will perform these actions (if any) before using models.]

3. Using 70-30 train-test, execute (a) Logistic Regression and (b) CART to compare testset errors. Display the results in a table:

Model	Model Complexity	False Positive Rate	False Negative Rate	Overall Error
Logistic Regression	*State the number of X variables.*			
CART	*State number of terminal nodes.*			

4. Conduct additional analysis to enhance quality of work and state your findings.

5. Write an executive summary that reports the most important findings in less than 300 words.

---

[1] https://www.weforum.org/publications/global-risks-report-2024/

[2] https://www.gov.sg/explainers/singapore-fight-against-misinformation

[3] Data Dictionary/Definitions are not provided. You may make reasonable assumptions about the data.