Research Proposal Outline – Ruth Allison ra22895

Sections required in a standard research proposal for the Computing department:

Project title:

Predicting international foundation student outcomes based on demographic characteristics and study behaviour, using classification algorithms

Significance/contribution to the discipline/research problem

Many factors are involved in predicting international student outcomes, these can be demographic characteristics or study behaviours. Students studying a foundation year before progression to a UK Higher Education Institute come from a large number of educational backgrounds, with different expectations and attitudes. Students can be split into three broad classes: those at risk of failure, those likely to pass but may need extra support to gain high grades, and those who are likely to gain high grades but may become disengaged through lack of challenge.

The ability to classify students allows for targeted support, for weaker students, and the opportunity to stretch and challenge stronger students. Early intervention can prevent issues escalating, improving outcomes.

A large amount of demographic data is collected but the size of the datasets and the number of variables makes analysis difficult. This type of problem is a good fit for the capabilities of machine learning algorithms.

Research question

Can machine learning algorithms accurately categorise students based on their demographic characteristics and study behaviours?

Aims and objectives

To create a machine learning model using classification algorithms to predict outcomes for foundation students studying to enter higher education in the UK.

Key literature related to the project

Chaubey, G. et al. (2022) Customer purchasing behavior prediction using machine learning classification techniques. *Journal of Ambient Intelligence and Humanized Computing [Preprint].* DOI: https://doi.org/10.1007/s12652-022-03837-6

Javadpour, A. et al. (2021) Improving the Efficiency of Customer's Credit Rating with Machine Learning in Big Data Cloud Computing. *Wireless Personal Communications* 121(4): 2699–2718. DOI: https://doi.org/10.1007/s11277-021-08844-y

Katyayan, A., et al. (2022) 'Analysis of Unsupervised Machine Learning Techniques for Customer Segmentation'. In: Chen, J.IZ., Wang, H., Du, KL., Suma, V. (eds) *Machine Learning and Autonomous Systems. Smart Innovation, Systems and Technologies,* vol 269. Springer, Singapore. DOI: https://doi.org/10.1007/978-981-16-7996-4_35

Rusli, N. et al. (2023) 'A Comparative Study of Machine Learning Classification Models on Customer Behavior Data'. *Communications in Computer and Information Science*. 222–231. DOI: https://doi.org/10.1007/978-981-99-0405-1_16

Siva Subramanian, R. et al. (2023) 'Enhancing Customer Prediction Using Machine Learning with Feature Selection Approaches'. *Lecture Notes in Networks and Systems*. 45–57. DOI: https://doi.org/10.1007/978-981-19-7402-1_4

Methodology/development strategy/research design

The research will be primary in nature, utilising existing demographic data from student applications to study with Kaplan International Pathways, in addition to information about attendance and early assessment scores.

Models used: **need to make a final decision, likely to be naïve Bayes, ANN & at least one ensemble method**

Evaluation metrics: Accuracy, Precision, Recall, F1 score

Ethical considerations and risk assessment

Anonymous data from previous academic years will be used. No un-anonymised data will be stored. Data will be securely stored in encrypted files.

The risk of stereotyping students will be mitigated by the explicit statement that the predictions are the potential outcomes and not certainties

Approval for use of existing data has been granted by the appropriate committee, following input from the legal team

Description of artefact(s) that will be created

A machine learning model that will classify students into one of three groups: 'at risk of failing', 'pass', and 'stretch and challenge'

Timeline of proposed activities

March – May 2024	Data collection & pre-processing
May – July 2024	Model testing & evaluation
August – September 2024	Write up