Strathmore University

Institute of Mathematical Sciences

Bachelor of Science in Statistics and Data Science

DAT 2102: INFORMATION SECURITY, GOVERNANCE & THE CLOUD

CLOUD COMPUTING SEMESTER PROJECT

Group 2A Start Date: 19th June 2025 **Due Date:** 10th July 2025

Group 2B Start Date: 20th June 2025 **Due Date:** 11th July 2025

Kamusi Data is an organisation of data scientists currently developing 3 projects on:

- Classifying Mushroom Species using decision-tree algorithm and/or any similar classification algorithm [Dataset: Mushroom Classification]
- Predicting Costs for Diamonds using linear regression algorithm and/or any similar regression algorithm [Dataset: Diamonds]
- Clustering Mall Customers using k-means algorithm and/or any similar clustering algorithm. [Dataset: Mall Customer Segmentation Data (kaggle.com)]

They intend to develop these as fully functional applications that run on a cloud platform as a convenient way for their end users to access these projects.

Instructions:

1. **In your previously defined groups,** implement the project as assigned by the instructor.
2. Use a similar technological approach to the part 2 of the project preliminaries exercise to create a cloud-based application that implements a machine learning model. This project will follow the data mining approach. **The business objective and output are fully dependent on what you have set as a group.**
3. Include an element of creativity in the interface of the application: use of colours, images, appealing fonts are highly encouraged and will be assessed during presentation.

Once done, **develop a slide deck/presentation** with these items:

- Brief background of the case study selected which would discuss the objective for the project, for example: vividly describe a furniture store whose owner would want to predict the costs of different types of wood. This case study can be based on a real-life example or a fictional case prepared as a group

- Description of the dataset used in the project: rows, columns, brief descriptive statistics etc.
- Data cleaning techniques used e.g. dropping columns, log transformation for normalisation
- Modelling approach applied: algorithm used, train/test split ratio, algorithm and model validation approach/results e.g. k-fold cross validation.
- A link to your developed application. Ensure the app is functioning as expected at the time of submission (*non-functional applications will receive a zero score on the functionality score for this assessment*)
- Conclusion to business objective set i.e. does the wood cost predictor application achieve what the store owner wanted? What features can be added to this application to ensure full functionality?

This slide deck will be uploaded to e-learning before the set due date for your class group. Only one person should submit the presentation on behalf of the project group.

Feel free to use any available resources for your work, however the rules against plagiarism apply. **Do not directly use ChatGPT or any other AI-generative text applications to develop your work (especially for the slide deck)**, this will be penalised strictly when detected.