Subject Description Form

Subject Code	DSAI5102
Subject Title	Principles of Data Science
Credit Value	3
Level	5
Pre-requisite/ Co-requisite/ Exclusion	Nil
Objectives	The subject aims to: (a) Provide students with a foundation in data science, covering core concepts, fundamental steps, and mathematical principles. (b) Enable students to apply probability and statistical methods for data analysis. (c) Develop students' proficiency in implementing key machine learning algorithm. (d) Introduce transfer learning and pre-trained models. (e) Introduce algorithmic bias and data/model drift, outlining methods to quantify and identify biases, mitigation strategies, and approaches to measure and address drift effectively. (f) Introduce AI governance frameworks and ethical considerations. (g) Cultivate students' Python programming skills for data analysis, visualization, and model implementation through case studies.
Intended Learning Outcomes	Upon completion of this course, students will be able to: (a) Master the basic concepts and fundamental steps in data science. (b) Apply probability and statistical methods to analyze data and draw inferences. (c) Implement key machine learning algorithms and evaluate model performance using appropriate metrics. (d) Utilize transfer learning and pre-trained models for data analysis like classification tasks. (e) Identify and mitigate biases in data and models to ensure fair and reliable outcomes. (f) Understand data governance frameworks and ethical considerations in data science practices. (g) Develop practical skills in data analysis and model implementation using Python

Subject Synopsis/ Indicative Syllabus	**Data science introduction** *Data science Venn diagram and terminology; data type: structured versus unstructured, quantitative versus qualitative; four levels of data; five steps of data science* **Foundational Theory** *Mathematics:vectors, matrices, arithmetic symbols, logarithms, exponents, matrix multiplication, set theory* *Probability: probability axioms, conditional probability, rules' of probability, Bayes' theorem, random variables -- continuous, discrete, mixed, distribution visualizations, moments, joint/marginal/conditional distributions, Bayesian versus frequentist* *Statistics:observational/experimental data, population versus sample, data sampling and sampling distributions, bias/variance/MSE* **Statistical inference** *Point estimation; confidence interval; hypothesis test: type I/II errors, power, one-sample t-tests, Chi-square tests: categorical variables, goodness of fit, association/independence* **Machine Learning (ML) Essentials** *Supervised, unsupervised, semi-supervised ML; regression: linear, logistic; reinforcement ML; classifications: Naive Bayes, decision trees, random forests and ensembling, K-means clustering; feature extraction and PCA; under/over fitting; K folds cross-validation; grid searching* **Transfer Learning (TL) and Pretrained Models** *Pre-trained models; transfer learning: inductive/ transductive/ unsupervised; natural language processing (NLP); bidirectional encoder representations from transformers (BERT); generative pre-trained transformer (GPT); algorithmic bias; data drift* **AI governance and ethics** *Data governance introduction; ML governance introduction; architectural governance introduction* **Case studies** *COMPAS dataset; textembeddings using pretrained models and OpenAI; presenting*
Teaching/Learning Methodology	The subject will mainly be delivered through lectures and tutorials. The lectures will be conducted to introduce the theoretical background, and practical problems / scenarios will be discussed in the tutorial sessions to illustrate how the theory developed can be applied in practice. Students are encouraged to use R or Python to perform an exploratory analysis by using real-world data.

Assessment Methods in Alignment with Intended Learning Outcomes

Specific assessment methods/tasks	% weighting	Intended subject learning outcomes to be assessed (Please tick as appropriate)				
		a	b	c	d	e
1. Assignments	20%	✓	✓	✓	✓	✓
2. Project	30%			✓	✓	
3. Examination	50%	✓	✓	✓	✓	✓
Total	100%					

Explanation of the appropriateness of the assessment methods in assessing the intended learning outcomes:

This subject focuses on the mathematical foundation of data science. Many of these topics are based on theory in statistics and machine learning. **Exam-based assessment** is an appropriate assessment method, including a 50% examination.

Since this subject also emphasizes on understanding the implementation of various numerical methods in data science, a mini-project that takes a weight of 30% is appropriate for assessing the intended learning outcomes (c) and (d), in which students will be encouraged to analyze large datasets using numerical methods and communicate their findings.

A 20% worth of assignments are also included as a component of continuous assessment in order to keep students in progress.

Continuous Assessment comprises assignments, mini-project and test. A written examination is held at the end of the semester.

Student Study Effort Required

Class contact:	
▪ Lectures/Tutorials	39 Hrs.
Other student study effort:	
▪ Assignments/Projects	58 Hrs.
▪ Self-study	40 Hrs.
Total student study effort	137 Hrs.

Reading List and References	Textbooks:		
	Sinan Ozdemir	Principles of Data Science	3rd Edition, 2024, ISBN-13. 978-1837636303
	References:		
	Trevor Hastie, Robert Tibshirani, Jerome Friedman, Daniela Witten	An Introduction to Statistical Learning: with Applications in Python	2023, ISBN-13.978-3031387463
	George Casella, Roger Berger	Statistical Inference	2nd Edition,2024, ISBN-13. 978-1032593036
	Bradley Efron, Trevor Hastie	Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science (Institute of Mathematical Statistics Monographs, Series Number 6)	Student Edition, 2021, ISBN-13. 978-1108823418
	Aurelien Geron	Hands-On MachineLearning with Scikit-Learn and PyTorch: Concepts, Tools, and Techniques to Build Intelligent Systems	1st Edition, 2025, ISBN-13. 979-8341607989
	Aurelien Geron	Hands-On MachineLearning withScikit-Learn, and TensorFlow	2nd Edition, 2019, ISBN-13. 978-1492032649
	Jay Alammar, Maarten Grootendorst	Hands-On Large Language Models: Language Understanding and Generation	1st Edition, 2024, ISBN-13. 978-1098150969
	John Berryman, Albert Ziegler	Prompt Engineering for LLMs: The Art and Science of Building Large Language Model-Based Applications	1st Edition, 2024, ISBN-13. 978-1098156152
	Immanuel Trumme	Data Analysis with LLMs: Text, tables,images and sound	2025, ISBN-13. 978-1633437647