

Maths for Data Science

**CT4031**

**2023 - 2024**

Zainab Loukil Kamal Zidan

School of Business, Computing, and Social Science

## University of Gloucestershire 2023

All rights reserved. No part of this publication may be reproduced, stored or transmitted in any form or by any means, including – but not limited to – photocopy, recording, or any information storage and retrieval system, without the specific prior written permission of University of Gloucestershire.

# Learning Outcomes

On successful completion of this module, students should be able to:

* Demonstrate an understanding of mathematical and statistical concepts.
* Develop code to visualise, analyse and manipulate data.
* Demonstrate a sound understanding in probability theory.
* Demonstrate knowledge of statistical inference, distribution and hypothesis testing.
* Apply skills relevant for academic progression and career development within the sector.

# Scheme of work

|  |  |  |
| --- | --- | --- |
| **Date w/c** | **Topic(s)** | **Practical work(s)** |
| Week 1 | Introduction to Data Science | Recap - Python I (variables, conditional statements and loops) |
| Week 2 | Maths, Data Science and Computer Science Applications | Recap - Python II (data types, collections, file I/O, string manipulation, functions) |
| Week 3 | Data Cleaning | Practical using Python |
| Week 4 | EDA and Visualisation | Practical using Python |
| Week 5 | Probability Theory – Part 1 – Introduction | Practical using Python |
| Week 6 | Probability Theory – Part 2 – Uncertainty and Bayes Classifiers | Practical using Python |
| Week 7 | Probability Theory – Part 3 – Bayesian Net | Practical using Python |
| Week 8 | Probability Theory – Part 4 – Markov chains | Practical using Python |
| Week 9 | Hypothesis and Statistical Tests | Practical using Python |
| Week 10 | Introduction to Mathematical Proof | Practical using Python |
| Week 11 | Introduction to Data Mining | Practical using Python |
| Week 12 | Assignment workshop | Assignment workshop |

**Assessment 1**

|  |  |
| --- | --- |
| **1. Module code and title** | **CT4031 Maths for Data Science** |
| **2. Module tutor** | Zainab Loukil |
| **3. Tutor with responsibility for this Assessment** | Zainab Loukil  **This is your first point of contact.** |
| **4. Assignment** | 001: 100% Coursework: Individual, portfolio: 2,000 words or equivalent. You will be penalised according to the **Academic Regulations for Taught Provision** if you exceed the size limit. |
| **5. Submission deadline**  Your attention is drawn to the penalties for late submission; see Undergraduate Modular Handbook. | **17/05/2024 – 15:00**  Your attention is drawn to the penalties for late submission; **see Academic Regulations for Taught Provision.** |
| **6. Arrangements for submission** | **Moodle** |
| **7. Date and location for return of work** | **Written feedback and provisional mark will be within 20 working days.** |
| **8. Students with Disabilities** | Alternative assessment arrangements may be made, where appropriate, for students with disabilities. However, these will only be implemented upon the advice of the Disability Advisor. Students wishing to be considered for alternative assessment arrangements must give notification of the disability (with evidence) to the Disability Advisor by the published deadlines. |
| **9. University Regulations for Assessment** | All assessments are subject to the **Academic Regulations for Taught Provision**. These include regulations relating to Errors of Attribution and Assessment Offences. In exercising their judgement, Examiners may penalise any work where the standard of English, numeracy or presentation adversely affects the quality of the work, or where the work submitted exceeds the published size or time limits, or where the work  fails to follow normal academic conventions for acknowledging sources. |

## The requirements for Assessment 1:

Android, a mobile operating system that is widely used across the globe, has become a target for malware due to its significant impact, open-source code, and ability to download apps from third-party sources without centralized control. Despite including security measures, recent news regarding Android's vulnerabilities and malicious activities highlights the importance of enhancing its security through continued development of frameworks and methods.

To combat malware attacks, researchers and developers have suggested various security solutions that leverage static analysis, dynamic analysis, and artificial intelligence. Data science has emerged as a promising field in cybersecurity, as data-driven analytical models can provide valuable insights to predict and prevent malicious activities.

AndroiHypo, Telecommunication company, proposes utilizing network layer features as the foundation for machine learning models to effectively detect malware applications, using open datasets from the research community. In this context, you have been hired by AndroiHypo as a data scientist. Your role is to investigate the given dataset, analyse it and draw conclusions.

After collecting the data, AndroiHypo has compiled the dataset to support their studies and now it is time to make data analysis magic. While studying the dataset, the company has proposed two hypotheses:

* The **probability** of having benign traffic **given,** the Domain Name System (DNS) query times is more than 5 times **and** the number of Transmission Control Protocol (TCP) packets is more than 40, is **at least 9%**.
* There is a massive traffic volume bytes difference between benign and malicious traffic types.

Using the dataset provided and the hypotheses presented by AndroiHypo agency, write a technical report addressing the following requirements:

- **Dataset Analysis and Pre-Processing, containing (25%)**:

· An explanation and analysis of the provided dataset;

· A list of problems encountered when manipulating the dataset;

· A description of the steps taken to clean the dataset.

- **Dataset Visualisation and proposed hypotheses (25%)**:

· Discussion related to the hypotheses proposed by the agency using at least two different types of graphs (e.g., boxplot, scatter plots or histogram).

- **Hypothesis testing (30%)**

· An analysis and evaluation of the hypotheses proposed by the agency applying statistical tests to support your arguments.

- **List of references using the Harvard referencing format (10%)**.

- **Appendix containing the Python code used to demonstrate actual use of the language in solution implementation (10%)**.

Your submission must be done via **Moodle**.