University of Suffolk

School of Engineering, Arts, Science and Technology

# Introduction to AI and Data Science
# Level: 4

| | |
|---|---|
| Assessment: | Programming Portfolio |
| Module Tutor: | Dr Adnane Ez-zizi |
| Weighting in Module: | 100% |
| Hand out: | Monday 6 February 2023 |
| Hand in: | **on or before 12 noon Wed 26 April 2023** |
| | Submission via Brightspace |

**What is required?**

- A portfolio of program solutions and supporting textual report (up to 2,000 words)

**Learning outcomes to be assessed:**

1. Demonstrate knowledge and understanding of the artificial intelligence and data science fields
2. Demonstrate the ability to develop a wide range of solutions to data problems using Python
3. Select appropriate techniques and libraries to formulate effective data solutions
4. Critically review literature in artificial intelligence and data science

**Assessment & Grading Criteria:**

1. Satisfactory knowledge and understanding of the artificial intelligence and data science fields
2. Satisfactory ability to develop a wide range of solutions to data problems using Python
3. Appropriate techniques and libraries are selected in the formulation of effective data solutions
4. Appropriate critical review of the literature in artificial intelligence and data science

**Assessment Brief.**
See the attached brief.

## ASSESSMENT BRIEF

In this assignment, you will create a portfolio of Python programs that demonstrate your understanding and ability to solve problems using search methods, Bayesian networks and machine learning. The portfolio should include a written report that explains each of your programs. As part of your assessment, you need to solve the following tasks.

### Task 1 – Search methods (40% of marks)

Write a literature review on the practical applications of search methods. Your literature review should not exceed 500 words. Implement three of the search algorithms covered in the course, where one of them should be A* search. Your program should use the search algorithms to solve one search problem of your choice, such as finding the shortest path in a map. Compare the performance of the algorithms in terms of number of nodes expanded, the maximum size of the fringe and the total running time.

**Task 1 contributes 40% of the marks for the assignment**

### Task 2 – Bayesian Networks (30% of marks)

This task requires you to analyse the relationship between the preferred modes of transportation and some socioeconomic and demographic variables. The data for this analysis was obtained from a survey, which collected information on age, sex, education, occupation, city of residence and favourite mode of travel as can be seen in Table 1. The data is stored in the "Travel_data.csv" file

Table 1. Variables contained in the travel survey data.

| Variable category | Variable | Description |
|---|---|---|
| Demographic | Sex | Biological sex: either *F* (female) or *M* (male) |
| | Age | Age: either *young* (below 30), *adult* (between 30 and 60) or *old* (older than 60). |
| Socioeconomic | Education | Highest education level: either *high* (high school) or *uni* (university degree) |
| | Occupation | Either *emp* (employee) or *self* (self-worker) |
| | Residence | Size of the city of residence: either *small* or *big*. |
| Outcome | Travel | Favourite means of transport: either *car*, *train* or *other*. |

As a first step to analyse your data with Bayesian networks, imagine that you consulted experts in the transport domain, and you came up with the Bayesian network structure shown in Figure 1.

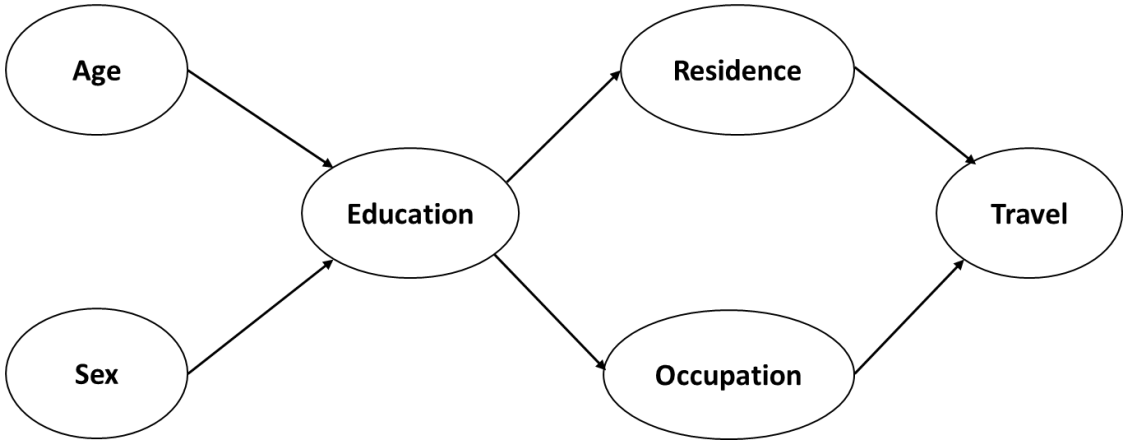


Figure 1. Bayesian network structure showing the dependencies between the travel and sociodemographic variables.

Next, you need to build and use your Bayesian network system understand how people use different modes of transportation. You should do this using the pgmpy package in Python (https://pgmpy.org/). Here are the specific steps you need to follow:

1. *Create the Bayesian network:*
    a. Write code to build the Bayesian network as shown in Figure 1. You do not need to worry about the conditional probability tables for now.
    b. Visualise the resulting network structure using the network package.
2. *Learn the network structure from the data:*
    a. Use one of the structure learning algorithms to estimate the structure of the Bayesian network based on the travel survey data. Examples of algorithms available in pgmpy include Peter and Clark (PC) algorithm, Hill-Climb Search and Chow-Liu algorithm.
    b. Compare the learned model structure to the expert-based one (shown in Figure 1) and note any differences.
    c. Evaluate the learned structure using the F1-score.
3. *Learn the parameters of the Bayesian network from the data:*
    a. Estimate the conditional probability tables (CPTs) of all nodes in the expert-based Bayesian network using the travel survey data.
4. *Inference:*
    a. Use the expert-based Bayesian network with the learned CPTs to answer a few interesting queries of your choice. Examples of queries you could answer include: Are there any differences between the age groups in the usage of travel modes? What about the difference between females and males? You could also imagine an individual profile and then predict the most likely preferred mode of travel for that profile.
    b. Run the same queries using the Bayesian network that you learned in Step 2. For this, you will need to learn the parameters of the Bayesian

network as you did in Step 3. Are there any differences with the results found with the expert-based Bayesian network?

c. Determine which Bayesian network, expert-based or estimated, leads to the most accurate predictions of the favourite mode of travel, using the available survey data?

**Task 2 contributes 30% of the marks for the assignment**

## Task 3 – Machine Learning (30% of marks)

In this task, you are required to build a machine learning model to predict the sentiment of tweets. You are given a dataset divided into a training set of 200,000 tweets and a test set of 20,000 tweets. Each set contains two columns: "Text" (representing the tweets) and "Sentiment" (positive or negative). The dataset can be found in the "Twitter_data.zip" file. Here are the steps that you need to follow:

1. *Data Pre-processing:* Clean and pre-process the data for modelling. When creating the the document-term matrix using CountVectorizer, you need to restrict the number of features to a number close to 3,000 by adding the option "max_features= 3000" (for more information, see https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html). If you have a small RAM capacity, you might want to reduced it to around 2000.

2. *Data Modelling:* Tune, train and test a Naive Bayes model on the pre-processed dataset. If you are using GridSearchCV, it is more memory efficient to use n_jobs=1.

3. *Model Evaluation:* Evaluate the performance of the model using different metrics such as accuracy, precision, recall, and F1-score.

4. *Model Comparison:* Compare the performance of the Naive Bayes model with another machine learning model of your choice.

**Task 3 contributes 30% of the marks for the assignment**

**Important Remark:** Please note that your grade will depend on how well you present your code, report and your references.

## REQUIREMENTS AND SUBMISSION INSTRUCTIONS

1. Please attempt all the portfolio tasks.

2. Your portfolio should include:

    a. The Python code that you used to solve the tasks. Please include separate script(s) or jupyter notebook(s) for each task. For example, you could include three notebooks named “Task1.ipynb”, ““Task2.ipynb” and ““Task3.ipynb”” corresponding to the three tasks described above. Place all the package/library imports at the top your scripts. Document all your code using comments as if future developers will be using it and will need to understand and maintain it.
    b. A report of up to 2,000 words that describes the problems you are solving, your approach and the results obtained. It is recommended that you included code snippets and figures to showcase your solutions. Your report should be well-organised and easy to navigate, and should include the list of references and resources that you used.
    c. If you used any extra dataset, please include it with your submission.

3. Submit your report as a Word or a PDF file. If you write your report in Pages or another application, convert it to PDF/Word when you have finished and check that all diagrams and tables are visible. Submit your scripts as a single Zip file.

4. The name of the files should match the pattern SXXXXXX.pdf (or .doc(x)) and SXXXXXX.zip where sXXXXXX is your UoS student number. Therefore, **your submission should include two files**: the report file in PDF or doc(x) formats as well as the script files stored in a zip file. Ensure that the report document is also marked with your UoS number. Your name should not appear anywhere in the files.

5. The report should not exceed 2000 words. If your submission exceeds the word count by up to 10% then there will be no penalty applied. Submissions that exceed the word count by more than 10% will be applied a fixed penalty of 5 percentage points (i.e., 5 marks). In all cases, the penalised mark will not be reduced below a pass level, assuming the work merits a pass. Tables, diagrams (including associated legends), appendices, reference lists, tables of contents, footnotes, and endnotes are excluded from the word count however should be used appropriately. It is for the Module Leader to decide if there is an excessive or inappropriate use of components excluded from the word count.

6. All bibliographies must be formatted according to the University of Suffolk Harvard Style. More information about citation and referencing is available here: https://libguides.uos.ac.uk/academic/referencing/Harvard

7. Submit your files to the Brightspace module for ‘Introduction to AI and Data Science’ under ‘Assessment’ > ‘Submission Folder for Introduction to AI and Data Science’ (https://brightspace.uos.ac.uk/d2l/lms/dropbox/admin/mark/folder_submissions_users.d2l?db=36604&ou=55972)

8. Cite your references carefully – remember that the University of Suffolk has strict rules dealing with plagiarism and the university reserves the right to call any student to a viva examination of any piece of assessed work.

**Level 4**

In accordance with the FHEQ, at the end of Level 5 students will be expected to have developed sound knowledge and critical understanding of the well-established concepts and principles in their field of study, and will have learned to apply those concepts and principles more widely outside the context in which they were first studied. They will have knowledge of the main methods of enquiry in the subject area, and ability to critically evaluate different approaches to problem solving. They will possess an understanding of the limits of their knowledge, and how this influences their analyses and interpretations. They will be able to effectively communicate information, arguments and analysis in a variety of forms to specialist and non-specialist audiences, and deploy key techniques of the discipline effectively.

| | | **Assessment category** | | | | |
|---|---|---|---|---|---|---|
| | | **Knowledge and critical understanding of well-established concepts and principles of the subject(s)** | **Cognitive and intellectual skills** | **Application of theory to practice (for courses with a professional practice element)** | **Reading and referencing** | **Presentation, style and structure** Work that significantly exceeds the specified word limit may be penalized |
| **Pass mark, demonstrating achievement of all associated learning outcomes** | **90%-100%** | Excellent work showing flawless understanding of the established concepts and principles of the subject(s). | Insightful application of excellent critical, analytical and evaluative skills to demonstrate exceptional ability to express arguments fully supported relevant evidence. Shows outstanding independent thinking through its original expression, and evidences both self-awareness and a deep and comprehensive understanding of the subjects' key stances and knowledge boundaries. | Sophisticated application of theory to practice, demonstrating insightful selection of theory and flawless application to practice | Insightful and effective use of a carefully selected range of relevant reading, including research-informed literature where relevant. Consistently accurate application of referencing. | Exemplary presentation of work that is fluent and flawless throughout. |
| | **80%-89%** | High quality work showing fluent understanding of the established concepts and principles of the subject(s). | Use of excellent critical, analytical and evaluative skills in order to develop highly logical and coherent judgements / arguments, supported by a range of relevant evidence. Evidence of independent thinking and creativity. Critiques a variety of stances meaningfully, and effectively expresses the limits of their knowledge. | Excellent application of theory to practice, with all links fully appropriate and meaningfully applied. | Consistent and balanced engagement with a refined range of relevant reading, including research-informed literature where relevant. Consistently accurate application of referencing. | Highly effective presentation of work that is coherently structured and clearly expressed throughout. |
| | **70% – 79%** | Commendable work showing detailed understanding of the established concepts and principles of the subject(s). | Use of effective critical, analytical and evaluative skills in order to develop logical and coherent judgements / arguments, supported by a range of relevant evidence. Clear evidence of originality. Explicit discussion of other stances and a strong awareness of the limits of their knowledge. | Effective application of theory to practice, with the student making highly appropriate and carefully expressed links between the two. | Consistent engagement with a wide range of relevant reading, including research-informed literature where relevant. Consistently accurate application of referencing. | Well-formed presentation of work that is coherently structured and clearly expressed throughout. |

| Level 5 | | Knowledge and critical understanding | Cognitive and intellectual skills | Application of theory to practice | Reading and referencing | Presentation, style and structure |
|---|---|---|---|---|---|---|
| Pass mark, demonstrating achievement of all associated learning outcomes | 60% – 69% | Work of solid quality showing competent and consistent understanding of the established concepts and principles of the subject(s). | Use of sound critical, analytical and evaluative skills in order to develop logical and coherent | Sound application of theory to practice, with the student making appropriate, well-developed and articulated links between the two. | Engagement with a wide range of relevant reading. Sound application of referencing, with no inaccuracies or inconsistencies. | Competent presentation of work in terms of structure and clarity of expression. |
| | 50% – 59% | Adequate work showing understanding of the established concepts and principles of the subject(s), but lacking depth and breadth. | Evidence of use of evaluation and critical analysis to support the development of logical and coherent judgements / arguments, supported by relevant evidence. An awareness of other stances and of the limits of their knowledge. | Consistent and accurate application of theory to practice, with the student making appropriate links between the two. | Engagement with an appropriate range of reading beyond essential texts. Referencing may show minor inaccuracies or inconsistencies. | Work is structured in a largely coherent manner and is for the most part clearly expressed. |
| | 40% – 49% | Simple factual approach showing limited understanding of the established concepts and principles of the subject(s). Narrow or misguided selection of material, with elements missing or inaccurate. | Limited and inconsistent use of evaluation and critical analysis to support emerging judgements or arguments, although not always logical or coherent and with inaccuracies. Limited awareness of other stances and the limits of their knowledge. | Relevant theoretical knowledge and understanding applied in practice, but with students not always making logical links between the two. | Evidence of reading, largely confined to essential texts, but mainly reliant on taught elements. Referencing may show inaccuracies and/or inconsistencies. | Ordered presentation in which relevant ideas / concepts are reasonably expressed. |
| Marginal fail | 35% - 30% | Weak work showing limited but fragmentary understanding of the established concepts and principles of the subject(s), for example through inaccuracies, inclusion of irrelevant material and/or absence of appropriate information. | Largely descriptive work, with very little effort made to use evaluation and critical analysis to develop judgements or arguments. Information accepted uncritically, with unsubstantiated opinions evident. | Limited understanding of the application of theory to practice, with the student often not making appropriate links between the two. | Poor engagement with essential texts and no evidence of wider reading. Heavily reliant on taught elements. Inconsistent and weak use of referencing. | Work is loosely, and at times incoherently, structured, with information and ideas often poorly expressed. |

| Level 5 | | Knowledge and critical understanding | Cognitive and intellectual skills | Application of theory to practice | Reading and referencing | Presentation, style and structure |
|---|---|---|---|---|---|---|
| Fail | 20% – 34% | Unsatisfactory work showing weak and flawed understanding of the established concepts and principles of the subject(s), for example through serious inaccuracies, inclusion of a significant amount of irrelevant material and/or absence of appropriate information. | Descriptive work with no effort made to use evaluation or critical analysis to develop judgements or arguments. Views expressed are often illogical, invalid or irrelevant. Minimal or no use of evidence to back up views. | Weak understanding of the application of theory to practice, with only occasional evidence of the student making appropriate links between the two. | Limited evidence of reading and/or reliance on inappropriate sources. Limited engagement with taught elements. Very poor use of referencing. | Work is poorly presented in a disjointed and incoherent manner. Information and ideas are very poorly expressed, with weak English and/or inappropriate style. |
| | < 20% | Highly unsatisfactory work showing major gaps in understanding of the established concepts and principles of the subject(s). Inclusion of largely irrelevant material, absence of appropriate information and significant inaccuracies. | Work is largely irrelevant or inaccurate, characterised by descriptive text and unsubstantiated generalisations. Minimal or no use of evidence to back up views. | Very weak theoretical knowledge and understanding, with no evidence of appropriate application in practice. | No evidence of reading or engagement with taught elements. Absent or incoherent referencing. | Work is extremely disorganised, with much of the content confusingly expressed. Very poor English and/or very inappropriate style. |