Coursework 1 specification for 2022-23

Data Analytics ECS784U/P,
Revised on 08/02/2023,
Dr Anthony Constantinou, Dr Neville Kitson.

1. Important Dates

- Release date: end of Week 2, **Friday** 3rd **February** 2023 at **12:00 noon**.
- Submission deadline: mid-Week 8, **Wednesday 15th March** 2023 at **10:00AM**.
- Late submission deadline (cumulative penalty applies): Within 7 days after deadline.

General information:

i. Students will sometimes upload their coursework and not hit the submit button. Make sure you fully complete the submission process.

ii. A penalty will be applied automatically by the system for late submissions.
   a. Lecturers cannot remove the penalty!
   b. Penalties can only be challenged via submission of an Extenuating Circumstances (EC) form which can be found on your Student Support page. All the information you need to know is on that page, including how to submit an EC claim along with the deadline dates and full guidelines.
   c. Deadline extensions can only be granted through approval of an EC claim
   d. If you submit an EC form, your case will be reviewed by a panel. When the panel reaches a decision, they will inform both you and the module organiser.
   e. If you miss both the submission deadline and the late submission deadline, you will automatically receive a score of 0.

iii. Submissions via e-mail are not accepted.

iv. The School requires that we set the deadline during a weekday at 10:00 AM.

v. For more details on submission regulations, please refer to your relevant student handbook.

2. Coursework overview and deliverables

The submission involves two files: a data analytic report (see Deliverable 1) and a Jupyter notebook (see Deliverable 2).

- You should address a data-related problem in your professional field or a field you are interested in (e.g., healthcare, sports, bioinformatics, gaming, finance, etc). If you are motivated by the subject matter, the project will be more fun for you, and you will likely produce a better report.

- Once you determine the area that interests you the most, you should search for a suitable data set or collate the data set yourself (see Section 5 for possible data sources).

- You should apply a minimum of **TWO** data analytic techniques (i.e. machine learning algorithms) of your choice to your data, from those covered in this course up to and including Week 5. The aim is to *learn* two models and contrast their performance on your input data. You are allowed to test more than TWO data analytic techniques if you wish (e.g., using multiple techniques to learn a model, or learning more than two models), but this is not a requirement and will not necessarily improve your mark. Remember to use the page limit wisely against the marking criteria (see below). The algorithms you can choose from are:

    - Linear, non-linear and logistic regression,
    - Support vector classification or regression,
    - Decision trees,
    - KNN,
    - k-means,
    - GMMs.

Deliverable 1: Technical report.

- The report shall have a maximum length of **7 pages** including references. Pages exceeding the 7-page limit will **NOT** be marked.

- Font size should be **not lower than 11**, and page margins should be **not lower than 2cm**. There are no other formatting requirements; e.g., it can have a single-column or a two-column format.

- Reports should be written with a technical audience in mind. It should be concise and clear, adopting the same style you would use in writing a scientific report or project dissertation.

- Some of the components your report should include:

i. Problem statement and hypothesis.
ii. A review of relevant literature
iii. Description of your data set and how it was obtained, including a sample of the data presented in a figure, along with pointers to your data sources.
iv. Description of any data pre-processing steps you took (if any).

v. What you have learnt by exploring the data; you may include some visualisations if necessary.
vi. How you chose which features to use in your analysis.
vii. Details of your modelling process, including how you selected your data analytic methods, as well as how you determined the optimal model through validation.
viii. Your challenges and successes.
ix. Key findings.
x. Possible extensions or business applications of your project.

Deliverable 2: Jupyter notebook:

- You must submit your Jupyter notebook Python code as a separate **PDF** file. This is needed so that we can quickly refer to your code outputs while marking your report. Please do not forget to add some section headings and comments around your code, similar to those added to the notebooks used in the labs.

    a. In Windows, you can generate a PDF file by right clicking and selecting ‘Print’ the Jupyter notebook loaded in your browser, and then you should be given an option to save it as a PDF file.

    b. Do **NOT** copy-and-paste your notebook’s code into a word document, as this approach will not preserve the notebook’s format.

    c. You do **NOT** need to submit your data set nor the actual .ipynb file. These might be requested at a later stage, if and only if we would like to review your code and/or data in greater depth.

3. Marking criteria

This coursework contributes **60%** towards your total module mark.

Criterion	Part of report	Weighting	Evidenced by (at least)
#1	Introduction to the project and background information	10%	Problem statement and hypothesis; project aims; concepts communicated; clarity.
#2	Literature review	5%	Subject placed in the context of literature; a minimum of 5 references to journal papers, conference papers, or books (web references do not count).
#3	Data processing	10%	Data source; description of data; any pre-processing steps; any feature selection or dimensionality reduction methods.
#4	Learning methods	10%	Brief description of the two data analytic methods; justification of the methods selected.
#5	Analysis, testing, results	15%	Includes possible cross-validation or any other approach to assess learning accuracy; documentation of testing; analysis of the strengths and weaknesses.
#6	Concluding remarks	20%	Discuss limitations and achievements; possible future improvements or directions; conclusion based on results.
#7	Quality of report	20%	Clarity; organisation; quality of the writing; quality and clarity of tables and figures; ease of understanding of the presentation of ideas.
#8	Jupyter notebook	10%	Jupyter notebook presented clearly with some comments.

Please note: We do realise that each project is different. The marking scheme shown above represents a simplified version of the marking scheme we use to assess project dissertations, and applies to any data analytic project.

4. Timetable

The coursework lasts for almost 6 weeks; from end of Week 2 to mid-Week 8. To ensure the coursework runs smoothly, be careful not to deviate much from the timetable below:

Weeks 2 and 3: Determine your project area and data set.

What is the question you hope to answer? What data are you planning to use to answer that question? What do you know about the data so far? Why did you choose this topic?

You may discover during your data exploration that you do not have the data necessary to answer your project's question. You may decide to change the research question to address in the project. You should aim to finalise any changes as soon as possible.

Our advice is to spend your time during the first week wisely doing some research on the data sources and the data analytic methods covered in the labs, depending on the problem you are trying to address. Researching appropriate data sets and determining what data analytic method to use is part of the coursework. Various data sources are provided for reference at the end of this document.

Week 4: Data processing.

What data have you gathered, and how did you gather it? What steps have you taken to explore the data? Which areas of the data have you 'cleaned' (if any)? If your data may not need cleaning, explain why. What insights have you gained from your exploration (visualisations are optional here – consider page limit)? Will you be able to answer your question with these data, or do you need to gather more data (or perhaps adjust the project aims)? How might you use modelling to answer your question?

Weeks 5 and 6: Apply two data analytic methods and analyse results.

Which are the two data analytic methods you have selected and why? Is this a supervised or an unsupervised learning problem? Is this a classification or a regression problem? What do you understand about the two data analytic methods and why are they appropriate in answering the project question/s? Apply them to your data and start exploring the results. Generate plots that help explain the results.

Week 7: Have produced your first draft.

At a minimum, this should include a) literature review and background information on your selected topic, b) narrative of what you have done so far, c) visualisations of the results.

Week 8: Finalise draft.

Remember we have no lectures or labs on Week 7, so use that week to complete the coursework.

5. Data sources

Using public data is the most common choice. If you have access to private data, that is also an option, though you will have to be careful about what results you can release to us. Some sources of publicly available data are listed below (you don`t have to use these sources).

- **Kaggle**
  https://www.kaggle.com/
  Over 50,000 public data sets for machine learning.

- **UK Covid Data**
  https://coronavirus.data.gov.uk/
  Official UK COVID data

- **Data.gov**
  http://data.gov
  This is the resource for most government-related data.

- **Socrata**
  http://www.socrata.com/resources/
  Socrata is a good place to explore government-related data. Furthermore, it provides some visualization tools for exploring data.

- **UN3ta**
  https://data.un.org/
  UN data is an Internet-based data service which brings UN statistical databases.

- **European Union Open Data Portal**
  http://open-data.europa.eu/en/data/
  This site provides a lot of data from European Union institutions.

- **Data.gov.uk**
  http://data.gov.uk/
  This site of the UK Government includes the British National Bibliography: metadata on all UK books and publications since 1950.

- **The CIA World Factbook**
  https://www.cia.gov/library/publications/the-world-factbook/
  This site of the Central Intelligence Agency provides a lot of information on history, population, economy, government, infrastructure, and military of 267 countries.

- **US Census Bureau**
  http://www.census.gov/data.html
  This site provides information about US citizens covering population data, geographic data, and education.

- **Health Data**
  Healthdata.gov
  https://www.healthdata.gov/
  This site provides medical data about epidemiology and population statistics.

- **NHS Health and Social Care Information Centre**
  http://www.hscic.gov.uk/home
  Health datasets from the UK National Health Service.

- **Social Data**
  Facebook Graph
  https://developers.facebook.com/docs/graph-api
  Facebook provides this API which allows you to query the huge amount of information that users are sharing with the world.

- **Topsy**
  http://topsy.com/
  Topsy provides a searchable database of public tweets going back to 2006 as well as several tools to analyze the conversations.

- **Google Trends**
  http://www.google.com/trends/explore
  Statistics on search volume (as a proportion of total search) for any given term, since 2004.

- **Likebutton**
  http://likebutton.com/
  Mines Facebook's public data--globally and from your own network--to give an overview of what people "Like" at the moment.

- **Amazon Web Services public datasets**
  http://aws.amazon.com/datasets
  The public data sets on Amazon Web Services provide a centralized repository of public data sets. An interesting dataset is the 1000 Genome Project, an attempt to build the most comprehensive database of human genetic information. Also a NASA database of satellite imagery of Earth is available.

- **DBPedia**
  http://wiki.dbpedia.org
  Wikipedia contains millions of pieces of data, structured and unstructured, on every subject. DBPedia is an ambitious project to catalogue and create a public, freely distributable database allowing anyone to analyze this data.

- **Freebase**
  http://www.freebase.com/
  This community database provides information about several topics, with over 45 million entries.

- **Gapminder**
  http://www.gapminder.org/data/
  This site provides data coming from the World Health Organization and World Bank covering economic, medical, and social statistics from around the world.

- **Google Finance**
  https://www.google.com/finance
  Forty years' worth of stock market data, updated in real time.

- **National Climatic Data Center**
  http://www.ncdc.noaa.gov/data-access/quick-links#loc-clim
  Huge collection of environmental, meteorological, and climate data sets from the US National Climatic Data Center. The world's largest archive of weather data.

- **WeatherBase**
  http://www.weatherbase.com/
  This site provides climate averages, forecasts, and current conditions for over 40,000 cities worldwide.

- **Wunderground**
  http://www.wunderground.com/
  This site provides climatic data from satellites and weather stations, allowing you to get all information about the temperature, wind, and other climatic measurements.

- **Football datasets**
  http://www.football-data.co.uk/
  This site provides historical data for football matches around the world.

- **Pro-Football-Reference**
  http://www.pro-football-reference.com/
  This site provides data about football and several other sports.

- **New York Times**
  http://developer.nytimes.com/docs
  Searchable, indexed archive of news articles going back to 1851.

- **Google Books Ngrams**
  http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
  This source searches and analyses the full text of any of the millions of books digitized as part of the Google Books project.

- **Million Song Data Set**
  http://aws.amazon.com/datasets/6468931156960467
  Metadata on over a million songs and pieces of music. Part of Amazon Web Services.