Faculty of Engineering and Technology
Continuous Assessment

Title: Application of Machine Learning Methods to Real-World Data

Module Name:	Machine Learning and Data Mining
Module Code:	7021DATSCI
Level:	7
Credit Rating:	20
Programme:	MSc Data Science
Type of assessment:	Coursework
Weighting:	50%
Max. mark available:	100

Lecturer:	Dr Ivan Olier-Caparroso
Contact details:	I.A.OlierCaparroso@ljmu.ac.uk

Resource requirements: Desktop/Laptop computer, Module Notes, Python, Microsoft Word, Library Resources, Internet.

Important dates:

Hand-out date: 13 February 2023

1st coursework element – Model predictions

Hand-in date: 21 March 2023, 23.55

Hand-in method: Standard Canvas submission

2nd coursework element – Report

Hand-in date: 19 April 2023, 23.55

Hand-in method: Turnitin, accessed from the Canvas module page.

Feedback date: 3 May 2023

Feedback method: On Canvas.

Introduction

This coursework provides experience in using the methods developed theoretically in class. In particular, you will be provided with a real-world problem and be asked to provide a solution using data mining.

Coursework format

This coursework requires you to work in pairs. You should send me an email (I.A.OlierCaparroso@ljmu.ac.uk) indicating the name of your teammate, copying her/his name in, and a team name. Please do this as soon as possible but no later than Friday, 17th of February. After that date, students without a team will be randomly assigned to one.

You are required to submit a brief report that summarises the work done and the predictions of the final model you develop that you consider is the best possible one.

This is not a prescriptive coursework with a clear path to the solution. Instead, it requires you to conceive, code and test several approaches before you reach a final solution. Also, training machine learning (ML) models requires certain amount of computing time that may further slow your progress. Therefore, it is extremely unwise to leave the work to the last minute. It is also expected that a significant amount of the work is carried out during the subsequent IT lab sessions.

As part of the coursework assessment, a leaderboard will be created. This requires you to submit your predictions in a standard file format. See details in the "**What you need to submit**" section.

Details of the real-world problem

Currently, more than 5 billion mobile devices with several sensors (e.g., accelerometer and GPS) are in use, capturing detailed, continuous, and objective measurements on different aspects of our life, including physical activity. Such widespread smartphone adoption provides unparalleled potential for data collection to study human behaviour and health. Smartphones, with appropriate storage, strong processors, and wireless transmission, may collect massive amounts of data about huge groups of people over long periods of time without the use of extra hardware or instruments.

With this coursework, you are required to propose a solution using ML to perform human activity recognition (HAR). HAR is a process aimed at the classification of human actions in a given period of time based on discrete measurements (acceleration, rotation speed, geographical coordinates, etc.) made by personal digital devices. In order to make an informed decision, you should develop and test several ML models before suggesting a final approach to solve the task. You will be supplied with a database that consists of data collected from 36 different users performing six types of human activities (ascending and descending stairs, sitting, walking, jogging, and standing) for specific periods of time. These data were acquired from accelerometers, which are able of detecting the orientation of the device measuring the acceleration along the three different dimensions. They were collected using a sample rate of 20 Hz (1 sample every 50 millisecond) that is equivalent to 20 samples per second.

The coursework's database consists of synthetic data generated from an extract of data collected by the WISDM Lab (https://www.cis.fordham.edu/wisdm/). Note that the coursework's database is unique hence there is not an equivalent one available elsewhere from the Internet. Therefore, you will certainly be at risk of plagiarism if your submitted coursework mainly uses code already published (see more details under the Academic Misconduct section of this assignment). However, you can browse the Internet to find ideas on how to solve the coursework. If you have any doubt, please discuss with me the situation as soon as possible but always before submission.

Database description

The coursework's database will be distributed via a Community Kaggle Competition (https://www.kaggle.com/competitions/7021datsci-challenge-2023) whose access is restricted to students enrolled in this module. It is the official database to be used in this coursework, and no other variant that could be available elsewhere is allowed to be used.

The database contains information about 36 users who performed six different human activities, including ascending and descending stairs, sitting, walking, jogging, and standing. The data was collected using accelerometers, which measure acceleration in three dimensions and determine device orientation. The time series data was divided into 10-second snippets and statistical features or metadata were extracted from each snippet. You will have access to both the metadata and the time series, so you can choose to build ML models based on either one. Additionally, you may consider extracting additional features from the time series.

Description of the files and folders

- 'signals.csv' file – contains the time-series data organised by users and snippets.
- 'signals_test' file – contains the time-series data corresponding to the test subset. It follows the same structure as the 'signals.csv' file.
- metadata.csv –contains features extracted from the signals and the target column (activity)
- metadata_test.csv – contains the same columns as 'metadata.csv' but without the target column.
- predictions_example.csv - A sample of the predictions file using the correct format for the Kaggle competition.

1. The "signals.csv" file has the following columns:

   - user_snippet: User id + snippet id.
   - timestamp: in milliseconds.
   - x-axis: The acceleration in the x direction as measured by the phone's accelerometer. Values are floating-point between -20 and 20. A value of 10 = 1g = 9.81 m/s^2, and 0 = no acceleration. The acceleration recorded includes gravitational acceleration toward the centre of the Earth, so that when the phone is at rest on a flat surface the vertical axis will register +-10.
   - y-axis: same as x-axis, but along y axis.
   - z-axis: same as x-axis, but along z axis.

2. The "signals_test.csv " file has a similar structure as the "signals.csv" file.

3. The "metadata.csv" file is a CSV file with the following columns:
   - user_snippet – the identifier of the user who acquired the data (integer) + the snippet identifier (integer).
   - activity – the activity that the user carried out at the corresponding snippet. This is the target, which could be one of the following class labels:
     - walking
     - jogging
     - sitting
     - standing

    - upstairs
    - downstairs
- Extracted features – the rest 30 columns correspond to 10 features extracted from the x, y and z acceleration time series. Each column name has the format D-axis__FEATURE, where D is either x, y or z; and FEATURE is one of the following:
    - sum_values
    - median
    - mean
    - length
    - standard_deviation
    - variance
    - root_mean_square
    - maximum
    - absolute_maximum
    - minimum

4. The "metadata_test.csv" file has all the columns as "metadata.csv" but the "activity" column.

Outcome

The main coursework outcome is to predict the user activity at a given time snippet.

What you need to submit

On canvas, you will have two separate assignments: 1) to submit your model predictions, and 2) to submit your final report:

Submission of model predictions

You must run your model(s) on the supplied test data and submit the predicted outcome using the file format as in the *predictions_example.csv* file, which is available in the same folder on Canvas. Your file name must be "**predictions_XXXX.csv**", where XXXX is your team's name. Important: Note that you won't be awarded any marks associated with the *Model predictions* assessment component if you fail to submit this file or to submit it using an incorrect format.

Your submitted final predictions will be used to generate a leaderboard based on model performance on the test data as measured using the model accuracy. In order to avoid overfitting the test set, this is randomly split into two subsets of similar size. The accuracy score is calculated on both splits. The student with the highest average accuracy on the test splits will win the competition.

One prediction file will be used for each team. However, you are allowed to submit as many prediction files as you want through Canvas. The only one to be used for the leaderboard will be the most recent submission after the submission system is shut down.

Note that you will also receive marks associated with your position on the board.

There is a *Community Kaggle Competition* associated with this assessment that you could use to test your predictions and to see how they compare against your classmates. I strongly recommend the use of this facility. See more information in the *Community Kaggle Competition* section.

Submission of the final report

You must produce a report that summarises your main results. Although the main content of your report should be the presentation and discussion of your results, you should also describe how you addressed the task, alternative solutions, possible reasons for success/failure. Also, a brief reflection on your work should be included. You must list your code as an appendix.

I suggest structuring your report as follows:

1. **Methodological approach:** Description of approaches used to solve the problem (indicating final and alternative approaches, methods, portions of the data used, etc)
2. **Exploratory data analysis:** Present the results of any pre-processing steps to get the data right for modelling (e.g. further feature extractions, data normalisation, feature selection, etc), data visualisation, clustering, etc.
3. **Modelling**: Present the results of the ML models. It is expected the use of several ML algorithms. You should report results of the hyperparameter tuning, model performance and further results that could give further insights about the quality of the implemented models.
4. **Discussion:** You should discuss the results of the exploratory analysis and modelling You should explain reasons for success/failure of the considered approaches and insights for future improvements.

Appendix 1 – Code listings as a Jupyter Notebook (IPYNB file)

Excluding appendices, it is expected the report length to be between 2000 and 3000 words. You should use tables and figures to enrich the content of your report. You must submit one file using PDF/DOC/DOCX format only (preferably PDF) that contains the report without the code. No other file is accepted. If you submit more than one files, only the first most recent file be marked. You can use any word processor, provided that you manage to export your report to any of the acceptable file formats.

Submission of the Jupyter Notebooks (Report's appendix)

The appendix contains the code you used to perform the analysis. It consists of one or more Jupyter Notebooks (IPYNB files). The expectation is that they can be executed free of errors and should show they are related to the coursework. If you submit more than one Notebook file, please clearly indicate in the first notebook cell the sequence in which they should be run.

Community Kaggle Competition (https://www.kaggle.com/competitions/7021datsci-challenge-2023)

This assignment has an associated *Community Kaggle* competition. You can make use of this facility to test as many prediction files as you want. The competition is already open and will close at the same time as the deadline for submitting the prediction files is hit. Invitation to join the competition will be shared during the IT lab sessions. Important: Note that submitting predictions to Kaggle does not constitute a formal submission to any of the assessment components. You must still submit your prediction file through Canvas. The *Community Kaggle Competition* is available to help you to decide which prediction file you should submit based on how your team ranks against your classmates. Also, the only valid leaderboard will be the one generated from the prediction files submitted to Canvas. You might expect small differences between both leaderboards, the one generated by Kaggle and the one generated from your submissions due to floating-point precision differences.

Assessment Criteria

The coursework is 50% of the assessment for this module. It will be marked out of 100. The breakdown of the marks available is as follows:

- Report – up to 70
- Model predictions – up to 30

Report and model predictions are group assessment components. Please refer to the Appendix for the marking rubric to be used to assess your coursework.

Extenuating Circumstances

If something serious happens that means that you will not be able to complete this assignment, you need to contact the module leader as soon as possible. There are a number of things that can be done to help, such as extensions, waivers and alternative assessments, but we can only arrange this if you tell us. To ensure that the system is not abused, you will need to provide some evidence of the problem.

More guidance is available at https://www.ljmu.ac.uk/about-us/public-information/student-regulations/guidance-policy-and-process

Any coursework submitted late without the prior agreement of the module leader will receive 0 marks.

Academic Misconduct

The University defines Academic Misconduct as 'any case of deliberate, premeditated cheating, collusion, plagiarism or falsification of information, in an attempt to deceive and gain an unfair advantage in assessment'. This includes attempting to gain marks as part of a team without making a contribution. The Faculty takes Academic Misconduct very seriously and any suspected cases will be investigated through the University's standard policy (https://www.ljmu.ac.uk/about-us/public-information/student-regulations/appeals-and-complaints). If you are found guilty, you may be expelled from the University with no award.

It is your responsibility to ensure that you understand what constitutes Academic Misconduct and to ensure that you do not break the rules. If you are unclear about what is required, please ask.

For more information you are directed to following the University web pages:

- Information regarding ***academic misconduct***: https://www.ljmu.ac.uk/about-us/public-information/student-regulations/appeals-and-complaints
- Information on ***study skills***: https://www2.ljmu.ac.uk/studysupport/
- Information regarding ***referencing***: https://www2.ljmu.ac.uk/studysupport/69049.htm

Appendix – Marking rubric

Rubric used to assess final model predictions (marks)

Criteria	30 out of 30	25 out of 30	20 out of 30	15 out of 30	10 out of 30	5 out of 30	0 out of 30
Model performance on the test set as measured using the area under the ROC curve (AUC)	Submitted prediction file correctly formatted. Leaderboard position 1.	Submitted prediction file correctly formatted. Leaderboard position 2.	Submitted prediction file correctly formatted. Leaderboard position 3 or 4.	Submitted prediction file correctly formatted. Leaderboard position 5 or 6.	Submitted prediction file correctly formatted. Leaderboard position 7 to 10.	Submitted prediction file correctly formatted. Leaderboard position 11 or below.	No prediction file was submitted, or it was not usable (i.e. wrongly formatted)

Rubric used to assess the report (in %)

Criteria	100 - 90	89 – 80	79 – 70	69 – 60	59 – 50	49 – 40	39 – 30	29 – 20	19 – 0
Problem (10%)	Extraordinary with several paths to a solution which are innovative and beyond the current state-of-the-art.	Outstanding with several creative paths to a solution, with one at least beyond the current state-of-the-art.	Excellent with several paths to a solution which are sophisticated and convincing.	Fluent with credible and precise paths to a solution.	Good with congruent and consistent paths to a solution.	Descriptive with unsophisticated paths to a solution.	Inadequate and/or contradictory. Paths to a solution are vaguely described.	Erroneous, insufficient and/or inappropriate description.	Missing or unrelated to the problem.
Exploratory data analysis (20%)	Exploratory analysis is extraordinary. Analysis is performed using a vast range of models.	Exploratory analysis is outstanding. Analysis is performed using a vast range of models.	Exploratory analysis is Excellent. Analysis is performed using a handful of models.	Exploratory analysis is fluent. Analysis is performed using a handful of models.	Exploratory analysis is good. Analysis is performed using a handful of models.	Exploratory analysis is adequate. Analysis is performed using a few models.	Inadequate details of the exploratory analysis is provided and supported. Analysis is limited to one model.	Erroneous details on the exploratory analysis is provided and supported. Analysis is limited to one model.	Exploratory analysis is missing or very few details are provided. Very limited evidence that the

									analysis was performed.
Modelling (35%)	Description of the results is extraordinary. Analysis is performed using a vast range of models. Models are validated and compared in many ways.	Description of the results is outstanding. Analysis is performed using a vast range of models. Models are validated and compared in many ways.	Description of the results is Excellent. Analysis is performed using a handful of models. Models are validated and compared in several ways.	Description of the results is fluent. Analysis is performed using a handful of models. Models are validated and compared in several ways.	Description of the results is good. Analysis is performed using a handful of models. Models are validated and compared in several ways.	Description of the results is adequate. Analysis is performed using a few models. Limited model validation and comparison.	Inadequate details of the results are provided and supported. Analysis is limited to one model. Models are not properly validated and compared.	Erroneous details on the results are provided and supported. Analysis is limited to one model. Models are not properly validated and compared.	Results are missing or very few details are provided. Very limited evidence that the analysis was performed.
Discussion of the results (30%)	Discussion of the results is exceptional and clearly distinctive.	Discussion of the results is outstanding and insightful.	Discussion of the results is excellent critical.	Discussion of the results is credible and precise.	Discussion of the results is accurate and coherent.	Discussion of the results is adequate.	Discussion of the results is imprecise, limited and/or inadequate.	Discussion of the results is ambiguous, incoherent, irrelevant and/or erroneous.	Discussion of the results is missing, or unrelated to the results or the problem.
Code listings (5%)	Code is attached and seems to provide a clear path to a solution to the problem.	(no marks available)	(no marks available)	(no marks available)	(no marks available)	(no marks available)	(no marks available)	(no marks available)	Code is missing, with errors, or so limited as to provide no clear path to a solution to the problem.