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:	60%
Max. mark available:	100

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

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

Important dates:

Hand-out date:	03 March 2025
Hand-in date:	04 April 2025, 17.00
Hand-in method:	Turnitin, accessed from the Canvas module page.
Feedback date:	22 April 2025
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.Olier@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 Monday, 10th March 2025. 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. In addition, 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 be 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-2025) 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.
- 'signals_kaggle' file – contains the time-series data corresponding to the kaggle 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'.
- metadata_kaggle.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 "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

Outcome

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

Community Kaggle Competition (https://www.kaggle.com/competitions/7021datsci-challenge-2025)

This assignment has an associated *Kaggle* competition. The competition is already open and will close on the day of your coursework's submission deadline. To register for the competition, follow the invitation link, which has been sent to you via email. To participate, create a predictions file for each ML model you wish to compete with and submit it to Kaggle using the following link: https://www.kaggle.com/competitions/7021datsci-challenge-2025.

Kaggle manages two leaderboards based on model performances measured using the accuracy. One leaderboard is public and will be visible to you at all times on Kaggle's website, while the other is private and will only be available during IT lab sessions. When you submit a prediction file, Kaggle will split it into two halves and calculate the accuracy for each of them. One will be used for the public leaderboard, and the other for the private leaderboard. The winning team will be the one with the average position between the public and the private leaderboards when the competition is closed.

What you need to submit

On canvas, you will have three separate assignments: 1) to submit your final report, 2) to submit your model predictions, and 3) to submit your code (as Jupyter Notebooks).

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. **ML 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 of the results:** 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.

It is expected the report length to be between 2000 and 4000 words. You must submit one file only in PDF/DOC/DOCX format only (preferably PDF) that contains your report. If you submit your report in more than one file, only the first one of your most recent submissions will 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 model predictions

You must run your model(s) on the supplied Kaggle's data subset and submit the predicted outcome. You must use the file format as in the *predictions_specimen.csv* file, which is available in the same folder on Canvas. Your file name must be "**predictions_XXXX.csv**", where XXXX should contain your team's name.

Important: You must submit the predictions file corresponding to your best ranked model to Canvas. Note that you won't be awarded any marks associated with the *Model predictions* assessment component if you fail to submit this file to Canvas or to submit it using an incorrect format.

Submission of the Jupyter Notebooks (Report's appendix)

It is expected that you use Jupyter Notebooks to implement your code. You must submit your code to Canvas as a Jupyter Notebook (ipynb) file. Due to the nature of this assignment, it is acknowledged that you might have implemented your code using several Jupyter Notebooks. You are required to submit at least one file, which you consider a representative sample of your work. The expectation is that it can be executed without errors and should demonstrate its relevance 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.

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 65
- Model predictions – up to 30
- Code listing – up to 5

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

Using AI-assisted Technologies

Where you use AI-assisted technologies (e.g. ChatGPT) in any stage of the development of this coursework (e.g. literature review, coding, writing, etc), these technologies should only be used to help you with your code development and to improve readability and language of your report, and not to replace your tasks as producing insights, analysing and interpreting data or drawing conclusions. These technologies should be used with your oversight and control. You are ultimately responsible and accountable for the contents of your report.

You must disclose in your report the use of any AI-assisted technologies in the form of a short paragraph at the beginning of your report, clearly indicating for what they were used. Failing to disclose their use could put you at risk of academic misconduct (please refer to the "Academic Misconduct" section for further details)

Appendix – Marking rubric

Rubric used to assess the report (marks)

Criteria	[full marks]	[89 – 80]	[79 – 70]	[69 – 60]	[59 – 50]	[49 – 40]	[39 – 30]	[29 – 20]	[no marks]
Problem **[5]**	Extraordinary with several paths to a solution which are innovative and beyond the current state-of-the-art. **[5/5]**	Excellent with several paths to a solution which are sophisticated and convincing. **[4/5]**		Good with congruent and consistent paths to a solution. **[3/5]**		Descriptive with unsophisticated paths to a solution. **[2/5]**	Erroneous, insufficient and/or inappropriate description. **[1/5]**		Missing or unrelated to the problem. **[0/5]**
Exploratory data analysis **[20]**	Exploratory analysis is extraordinary. Analysis is performed using a vast range of models. **[20/20]**	Exploratory analysis is outstanding. Analysis is performed using a vast range of models. **[18/20]**	Exploratory analysis is Excellent. Analysis is performed using a handful of models. **[15/20]**	Exploratory analysis is fluent. Analysis is performed using a handful of models. **[13/20]**	Exploratory analysis is good. Analysis is performed using a handful of models. **[11/20]**	Exploratory analysis is adequate. Analysis is performed using a few models. **[9/20]**	Inadequate details of the exploratory analysis is provided and supported. Analysis is very limited. **[7/20]**	Erroneous details on the exploratory analysis is provided and supported. Analysis is very limited. **[5/20]**	Exploratory analysis is missing or very few details are provided. Very limited evidence that the analysis was performed. **[0/20]**
Modelling **[30]**	Description of the results is extraordinary.	Description of the results is outstanding.	Description of the results is Excellent.	Description of the results is fluent.	Description of the results is good.	Description of the results is adequate.	Inadequate details of the results are	Erroneous details on the results are	Results are missing or very few details are

	Analysis is performed using a vast range of models. Models are validated and compared in many ways. **[30/30]**	Analysis is performed using a vast range of models. Models are validated and compared in many ways. **[26/30]**	Analysis is performed using a handful of models. Models are validated and compared in several ways. **[23/30]**	Analysis is performed using a handful of models. Models are validated and compared in several ways. **[20/30]**	Analysis is performed using a handful of models. Models are validated and compared in several ways. **[17/30]**	Analysis is performed using a few models. Limited model validation and comparison. **[14/30]**	provided and supported. Analysis is limited to one model. Models are not properly validated and compared. **[11/30]**	provided and supported. Analysis is limited to one model. Models are not properly validated and compared. **[8/30]**	provided. Very limited evidence that the analysis was performed. **[0/30]**
Discussion **[10]**	Reflection of the work done is exceptional, critical and clearly distinctive. Suggestions for future work are insightful. **[10/10]**	Reflection of the work done is outstanding and critical. Suggestions for future work are insightful. **[8/10]**	Reflection of the work done is excellent and critical. Suggestions for future work are insightful. **[7/10]**	Reflection of the work done is precise. Suggestions for future work are credible. **[6/10]**	Reflection of the work done is coherent. Suggestions for future work are credible. **[5/10]**	Reflection of the work done is adequate. Suggestions for future work are limited. **[4/10]**	Reflection of the work done is imprecise and limited. Suggestions for future work are inadequate. **[3/10]**	Reflection of the work done is ambiguous, incoherent, irrelevant and/or erroneous. **[2/10]**	Reflection of the work done is missing, or unrelated to the results or the problem. **[0/10]**

Rubric used to assess final model predictions (marks)

Criteria	[full marks]	[79-70]	[69-60]	[59-50]	[49-40]	[39-20]	[No marks]
Model performance on the Kaggle's test set as measured using the accuracy **[30]**	Submitted prediction file correctly formatted. Private leaderboard position 1. **[30/30]**	Submitted prediction file correctly formatted. Private leaderboard position 2. **[24/30]**	Submitted prediction file correctly formatted. Private leaderboard position 3 or 4. **[20/30]**	Submitted prediction file correctly formatted. Private leaderboard position 5 or 6. **[17/30]**	Submitted prediction file correctly formatted. Private leaderboard position 7 or 8. **[14/30]**	Submitted prediction file correctly formatted. Private leaderboard position 9 or below. **[10/30]**	Submitted prediction file wrongly formatted or no prediction file submitted. **[0/30]**

Rubric used to assess the code (marks)

Criterion	Full marks	No marks
Code listings **[5]**	Code is attached, free of errors, and seems to provide a clear path to a solution to the problem. **[5/5]**	Code is missing, with errors, or so limited as to provide no clear path to a solution to the problem. **[0/5]**