University of Essex
School of Computer Science and Electronic Engineering
CE802 Machine Learning

Assignment: Design and Application of a Machine Learning System for a Practical Problem

Set by:	Prof. Luca Citi (lciti@essex.ac.uk)
Type of assignment:	Individual
Distributed to students:	Week 23
Submission deadline:	Week 31 (see FASer for exact official date and time)
Submission mode:	Electronic only via FASer

Assignment objectives

This document specifies the coursework assignment to be submitted by students taking CE802. Aims of this assignment are: a) to learn to identify machine learning techniques appropriate for a particular practical problem; and b) to undertake a comparative evaluation of several machine learning procedures when applied to the specific problem.

Assignment description

1. Pilot-Study Proposal

Imagine that you work as Machine Learning independent consultant, providing scientific advisory and consulting services to companies seeking to apply data analytics to their business activities.

A healthcare provider is interested in using machine learning to identify patients who are at high risk of developing diabetes. By identifying patients at risk, the healthcare provider can proactively intervene with preventive measures such as lifestyle interventions, medication, or referral to a specialist.

The manager of the organisation has access to data collected from electronic medical records.

In the first part of your assignment, you are asked to write a detailed proposal for a pilot study to investigate whether machine learning procedures could be used to successfully solve this problem. Your report should discuss several aspects of the problem, including the following main points:

- the type of predictive task that must be performed (e.g., classification, regression, clustering, rules mining, ...);
- examples of possibly informative features that you would like to be provided with (what type of information do you think would be a good predictor of the risk of developing diabetes?);
- the learning procedure or procedures (e.g., DTs, k-NN, k-means, linear regression, Apriori, SVMs, ...) you would choose and the reason for your choice;
- how you would evaluate the performance of your system before deploying it.

You can assume that the manager has some knowledge of machine learning and you do not need to explain how the recommended learning method works. Simply discuss your recommendation and back it with sound arguments and/or references.

This document should consist of approximately 500–750 words of narrative (i.e. excluding references, pictures, and diagrams). Please report your word count on the title page. The document must be submitted in PDF format with file name `CE802_Pilot.pdf`.

2. Comparative Study

Thanks to the convincing arguments in your pilot-study proposal, the organisation decides to collect the data that you suggested and to hire you to perform the proposed study. They provide you with a training set of historical data extracted from electronic medical records and has been labelled to indicate whether the patient has eventually been diagnosed with diabetes or not. These data are available in the `CE802_P2_Data.zip` archive available from the CE802 moodle page. In this part of the assignment, you are asked to perform the following two tasks.

a) Investigate the performance of a number of machine learning procedures on this dataset. Using the data in the file `CE802_P2_Data.csv` contained in the `CE802_P2_Data.zip` archive, you are required to perform a comparative study of the following machine learning procedures:

- a Decision Tree classifier;
- at least two more ML technique to predict the target label.

You will notice that one of the features is missing for some of the instances. You are therefore required to deal with the problem of missing features before you can proceed with the prediction step. As a baseline approach you may try to discard the feature altogether and train on the remaining features. You are then encouraged to experiment with different imputation methods.

The company uses Python internally and therefore Python with scikit-learn is the required language and machine learning library for the problem. For this task, you are expected to submit a Jupyter Notebook called `CE802_P2_Notebook.ipynb` (using the file with the same name inside the zip as starting point) containing the Python code used to perform the comparative analysis and produce the results, as well as the code used to perform the predictions described in task "b" below.

b) Prediction on a hold-out test set. An additional dataset, `CE802_P2_Test.csv`, is provided inside the `CE802_P2_Data.zip` archive. Binary outcomes are withheld for this test set (i.e. the "Class" column is empty). In this second task you are required to produce class predictions of the records in the test set using one approach of your choice among those tested in task "a" (for example the one achieving the best performance). These data must not be used other than to test the algorithm trained on the training data.

As part of your submission you should submit a file called `CE802_P2_Test_Predictions.csv` in CSV format, which must be identical to `CE802_P2_Test.csv` except that the missing class is replaced with the output predictions obtained using your chosen approach. This second task will be marked based on the prediction accuracy on the test set.

3. Additional Comparative Study

Thanks to the good results obtained in the comparative study, the healthcare provider has deployed your system and is obtaining good results. Now a similar organisation would like to hire you to design a similar system but, unlike the first system, they would like you to predict not only whether the person will develop diabetes, but also by how much the patient's average blood glucose level exceeds the diagnostic threshold.

They provide you with a training set of historical data containing features of each patient as well as the extent by which the patient's blood glucose level exceeds the diagnostic threshold if untreated.

These data are available in the `CE802_P3_Data.zip` archive available from the CE802 moodle page. In this part of the assignment, you are asked to perform the following two tasks.

a) Investigate the performance of a number of machine learning procedures on this dataset. Using the data in the file `CE802_P3_Data.csv` contained in the `CE802_P3_Data.zip` archive, you are required to perform a comparative study of the following machine learning procedures:

- Linear Regression;
- at least two more ML technique to predict the target value.

This company too uses Python internally and therefore Python with scikit-learn is the required language and machine learning library for the problem. For this task, you are expected to submit a Jupyter Notebook called `CE802_P3_Notebook.ipynb` (using the file with the same name inside the zip as starting point) containing the Python code used to perform the comparative analysis and produce the results as well as the code used to perform the predictions described in task "b" below.

b) Prediction on a hold-out test set. An additional dataset, `CE802_P3_Test.csv`, is provided inside the `CE802_P3_Data.zip` archive. Target values are withheld for this test set (i.e. the "Target" column is empty). In this second task you are required to produce predictions of the records in the test set using one approach of your choice among those tested in task "a" (for example the one achieving the best performance). These data must not be used other than to test the algorithm trained on the training data.

As part of your submission you should submit a file called `CE802_P3_Test_Predictions.csv` in CSV format, which must be identical to `CE802_P3_Test.csv` except that the missing "Target" column is replaced with the output predictions obtained using your chosen approach. This second task will be marked based on the root mean squared error on the test set.

4. Report on the Investigation

After conducting the studies in parts 2 and 3, you are asked to write a report containing an account of your investigation. There should be a brief summary of the experiments performed followed by one or more tables and/or graphs summarizing the performance of the different solutions. Any numerical data that you include should be in a suitable graphical or tabular form. The rest of the report should concentrate on your interpretation of the results and what they tell you about the relative strengths and weaknesses of the alternative methods when applied to the given data.

This document should consist of approximately 750–1500 words of narrative (i.e. excluding references, pictures, and diagrams). Please report your word count on the title page. The document must be submitted in PDF format with file name `CE802_Report.pdf`.

Suggested material

`Scikit-learn` online documentation and tutorials: `https://scikit-learn.org`.
Lecture notes on machine learning and Lab notes on `scikit-learn`, `pandas` (to read/write CSV files): see CE802 moodle page.

Submission

Your work must be submitted to the university's online coursework submission system at the address `https://faser.essex.ac.uk/` by the deadline given on the system. No other mode of submission

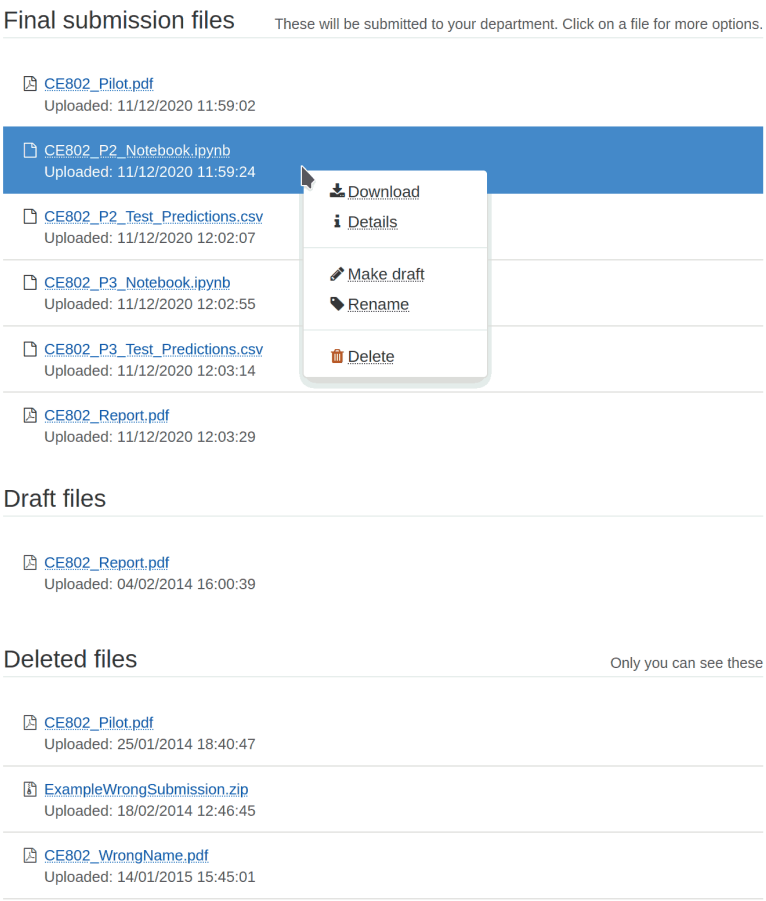

Figure 1: Example of how the submission should look like on FASER.

is acceptable. You are strongly advised to submit a draft submission well before the deadline, and then update it up to the deadline.
You are required to submit the following files on FASER:

1. The file `CE802_Pilot.pdf` with the pilot study document exported to **PDF format** (no doc or docx, etc.);
2. The Jupyter Notebook file `CE802_P2_Notebook.ipynb` used to perform the comparative study in part 2a and also to make the predictions in part 2b;
3. The file `CE802_P2_Test_Predictions.csv` with your predictions as described in part 2b;
4. The Jupyter Notebook file `CE802_P3_Notebook.ipynb` used to perform the comparative study in part 3a and also to make the predictions in part 3b;
5. The file `CE802_P3_Test_Predictions.csv` with your predictions as described in part 3b;
6. The file `CE802_Report.pdf` with the comparative study report exported to **PDF format** (no doc or docx, etc.).

DO NOT submit a zip file but each file individually on FASER. If you submit any draft versions please mark them as draft before the submission and make sure that your final submission contains only the six files above, as shown in Figure 1.
DO NOT WAIT UNTIL CLOSE TO THE DEADLINE TO MAKE YOUR FIRST SUBMISSION. Difficulties with the submission system will not be accepted as an excuse for a missing submission.

Marking criteria

This assignment is worth 50% of the module mark and will be assessed based on:

- Pilot-Study Proposal
  - Correctness of identified type of predictive task 2%
  - Validity of examples of possibly informative features 3%
  - Appropriateness of learning procedure(s) suggested 3%
  - Correctness of evaluation methods suggested 4%
  - Overall clarity of presentation 4%
- Comparative Study – Task a
  - Correctness and completeness of investigation 11%
  - Quality of the code and comments 5%
- Comparative Study – Task b
  - Accuracy of predictions 15%
- Additional Comparative Study – Task a
  - Correctness and completeness of investigation 11%
  - Quality of the code and comments 5%
- Additional Comparative Study – Task b
  - Accuracy of predictions 15%
- Report on the Investigation
  - Quality of presentation of the methods followed in part 2 ..... 4%
  - Quality of presentation and discussion of results of part 2 ..... 4%
  - Quality of presentation of the methods followed in part 3 ..... 4%
  - Quality of presentation and discussion of results of part 3 ..... 4%
  - Justification of conclusions drawn 4%
- Others
  - Compliance with submission instructions (e.g., FASER submission and file formats) 2%

For each requirement, the following scale will be used to guide the marking. *Poor*: Nothing relevant submitted. *Unsatisfactory*: Work is seriously flawed, displaying inadequate knowledge, major lack of understanding, irrelevance or incoherence. *Borderline*: Significant gaps in knowledge but some understanding of fundamental concepts. *Fair*: Minor gaps in knowledge but reasonable understanding of fundamental concepts. *Good*: Substantially complete and correct knowledge but not going significantly beyond what was taught. *Very good*: Comprehensive knowledge demonstrating very good depth with clear insight into links between theory and practice. *Excellent*: Shows very good understanding supported by evidence that the student has gone beyond what was taught by extra study or creative thought. Work at the top-end of this range is of *Outstanding* quality.

Problems

If any general problems arise in the assignment, please start a thread on the moodle forum. For specific inquiries, please email `lciti@essex.ac.uk`.

Late Submission and Plagiarism

Please refer to the Postgraduate Students' Handbook for details of the Departmental policy regarding late submission and University regulations regarding plagiarism. In particular, it is not acceptable to submit text, code or results that were developed by other people (including your fellow students). Please keep in mind that this is an **individual assessment** and therefore working jointly on any aspect of the submission is not allowed.

Revision 1.0
07/03/2023
Luca Citi