***Coding***

* + - * You should use Spyder or Jupyter Notebook for your IDE.
      * You should work with the libraries following libraries:
        + For neural networks: *Tensorflow/Keras****,***
        + for other techniques *scikit-learn*.

You might possibly use other libraries to stretch your work, but those studied should form the core of your work.

* + - * Your code should be well commented, including details of any reused code.

***The*** ***Task***

In this work, you are expected to:

1. Define the problem domain and dataset(s) (you are free to choose the problem domain and the dataset that you want to investigate).
2. Define questions and analysis tasks (a brief **overview of the domain, the questions** that are being asked, a list of your **objectives** and the expected output(s) of your analysis).
3. Perform an initial investigation of the dataset and the characteristics of the data. Develop a plan as to how you might transform the data to make it useable.
4. Develop a plan as to which artificial intelligence techniques you might use and what sorts of potential observations these can lead to, and how you will evaluate these.
5. You must use choose from the following models: perceptron, decision trees, linear regression, support vector machines, random forest, k-nearest neighbour, naïve Bayes, neural networks as well as unsupervised techniques k-means and GMM, and principal component analysis. Most supervised models have both classification and regression variants. You are encouraged to work with neural networks. An additional technique from outside the taught module content might be applied for comparison purposes, if this is done it should be clearly indicated and well justified.
6. Split your dataset (train/validate/test, some datasets come pre-split). If you have a holdout test set then you most likely don’t want to use this until the near the end of your work.
7. Perform the analysis. Get the data ready for analysis, carry out your analysis/modelling as needed, validate your results and communicate observations, iterating through this process. Analytical operations can include data processing to an extent that is needed (not all datasets are messy) to prepare a useful and robust dataset to work within, and data derivation (such as feature engineering).

* You might establish a baseline result first, computing metrics on training and validation sets, analyse errors, work on succeeding iterations, and alternative models. (If initial baseline results are amazing and there are no errors is the problem too easy?)
* Generally, be close to your data (visualise the dataset, collect summary statistics, look at errors, analyse how different parameters affect performance, try out different model variants).

***Report***

Graphical illustration of your results is expected (perhaps training/testing error curves, confusion matrices, algorithm outputs, etc), as well as results. Figures and Tables should be numbered and described. Following the above analytical process, make sure that in your report you answer the following questions where appropriate (this is \*\*not\*\* a report structure):

* What is your dataset, problem domain?
* Is your problem classification or regression?
* Did you have any missing, corrupt or misleading data? If so, how did you cope it?
* Have you omitted some data? If so, why?
* Did you apply techniques to understand your dataset?
* How did you encode the input variables?
* What models did you use?
* What are the criteria for selecting model performance evaluation tools?
* What were your outputs?
* Did you have any problems or difficulties working with the dataset?

You should present the results clearly and concisely and provide a discussion of the results, with conclusions related to problem being addressed. The conclusions section might propose some further work based on the results of this coursework.

We hope that you will have a lot of work to report, maybe more than you can fit into the page limit. In this case you will need to display good editorial judgement as to what to report: what was most important, what was most interesting.

***Datasets***

You are free to choose the domain and the dataset that you want to investigate. Here are some suggestions and sources for datasets.

You ***cannot*** use the datasets that come with scikit-learn, or others used in the exercises. Other heavily used datasets make bad choices.

Some of these sources will come with code. Whilst you may use this code if referenced, you get little credit for this.

**General:**

* Kaggle is Google’s online data community, and contains thousands of datasets.
* UCI Machine Learning Repository. The University of California, Irvine has a collection of several hundred datasets (some of them a little small).

Some other possible sources of data:

**Images:** Labelme, ImageNet, LSUN, Google’s Open Images, COCO

**Text:** Project Gutenberg

**Clinical:** MIMIC, The World Health Organisation (<https://www.who.int>)

**General:** FiveThirtyEight (https://data.fivethirtyeight.com)

There are many, many other sources of data available.

**Note:** You are not necessarily being marked on how good the results are. What matters is that you try something sensible and clearly describe the problem, method, what you did, and what the results were. Don't pick a dataset that is way too hard for your experiments. Don’t pick a dataset that is too straightforward (too small) to produce interesting results. Be careful not to do foolish things like data snooping, testing on your training data, including plots with unlabelled axes, using undefined symbols in equations. Do sensible cross-checks like running your models several times, varying your random seed, leaving out small parts of your data, adding a few noisy points, etc. to make sure everything still works reasonably well. If you pick something you think is interesting it will make the process of getting it to work more rewarding.

***Coding & Referencing***

This is, in large part, a coding assignment. If you use code (or other materials) written by someone else, you should ***cite*** that code (or other material) in Harvard format. Your code itself should reflect sources in your comments. If you do not cite work appropriately you will have committed academic misconduct. Making superficial changes to the code does not make it yours. You are also expected to make a coding contribution, so if you use a large amount of code written by someone else, and cite it appropriately, your contribution will be low and your work marked accordingly.