Understanding AI (771763_B23_T2)

Formative Assignment 2:
A Coding Exercise

Deadline (for students on the *Thursday* stream):
Wednesday 20th March, 2024 at 2pm

Formative Assignment

As this is a formative assignment, it will not count towards your final grade on this module. We will assess your submission and provide you with feedback on what you did well and areas that you could improve.

Submission Instructions

Please upload the following files as part of your submission:

1. Your written report as a single PDF file.
2. Your code as a Jupyter Notebook file.

These files need to be submitted on the Canvas assignment page corresponding to your stream.

Context

For this assignment, you will use Supervised and Unsupervised machine learning models to analyse the country life expectancies dataset that you looked at in the workshops in Weeks 6 and 7. This dataset can be downloaded from Canvas at the following link:

https://canvas.hull.ac.uk/courses/67474/modules/items/1012704

Two of the exercises in the final Summative Portfolio of Work assignment will involve coding-based tasks similar to this. The feedback you receive on this formative assignment is therefore designed to guide you in how to complete these exercises for the summative assignment.

The Task

Write up a report on your analysis of the life expectancies in different countries. Your report should cover the following:

- **Part I: Supervised learning models to predict life expectancy.**

  In the first part of your report you will explore how we can use supervised learning models to predict the life expectancy in a country. Please address the following questions:

a. Compare simple linear regression models that predict a country’s life expectancy from a single input feature, using the following variables as the input feature in each case: (i) GDP; (ii) Adult Mortality; (iii) infant deaths. Which of these single input features provides the best prediction for life expectancy and why?
b. You will find that the relationship between life expectancy and GDP is non-linear, i.e. it does not follow a straight line. Explore the non-linearity of this relationship by fitting polynomial regression models to predict life expectancy from the GDP. Compare the results of these fits for different values of the polynomial degree from 2 to 5 (inclusive). What degree of polynomial gives the best fit to this data?
c. Explore regression models that use multiple input features (the GDP, Adult Mortality and infant deaths) to predict a country’s life expectancy. Compare a multiple-linear regression model with at least one other regression model with multiple input features that you have encountered in your studies.
d. Out of all the regression models that you have considered, which is the best model to use for predicting the life expectancy in a country?

- **Part II: Unsupervised learning models to identify clustering patterns.**

  In the second part of your report you will explore how we can use unsupervised learning models to identify clusters of countries based on the life expectancy dataset. Please address the following questions:

  a. Using the *k*-Means clustering algorithm, identify clusters of countries based on their life expectancy versus GDP. Compare these results to clusters identified based on the countries’ life expectancy versus Adult Mortality, also using the *k*-Means clustering algorithm. What is the optimal number of clusters to use in each case, and why? Comparing the clusters identified using life expectancy versus GDP and life expectancy versus Adult Mortality, which produces the best clustering, and why? What is your interpretation of the clusters identified in each case?
  b. Using the life expectancy versus GDP of each country, compare the clusterings of countries predicted by the *k*-Means clustering, Agglomerative Hierarchical clustering, and DBSCAN models. Which of these three models produces the best clustering for this data? Are the countries generally assigned to the same cluster in each model?

Your report should be written in the style of a scientific paper, including an introduction to describe the context of the problem, a description of the methodology that you used, a discussion of the results of your analysis supported by appropriate figures, and a summary of your key conclusions.

You will also need to submit the code that you used to analyse the data and produce your results, which needs to be uploaded as a single Jupyter Notebook file. Your code needs to include suitable comments that describe what the code is doing at each step.

Maximum word count: 1500 words.

Grading Criteria

Since this is a formative assignment, you will not receive a grade for your submission. However, we will provide you with feedback on what you have done well and areas that you could improve upon. See below for an outline of the criteria that we will consider in our feedback:

Criteria	Points to consider
Quality & Structure of the Written Report.	• Is the report written clearly and concisely in an appropriate scientific style? • Does the report follow a logical structure, with a clear beginning, middle and end? • Are the figures presented clearly, with suitable axis labels and figure captions?
Analysis.	• Does the report include results from all the regression models requested in Part I of the Task? Are these results supported by suitable figures? • Does the report include results from all the clustering models requested in Part II of the Task? Are these results supported by suitable figures? • In Part II, does the report correctly identify the optimal number of clusters to use in the *k*-Means algorithm, and is this justified using a suitable method? • Does the report use appropriate evaluation metrics to compare the results of different models?
Coding.	• Does the Jupyter Notebook include all of the code needed to reproduce the results and figures presented in the written report? • Is the code written clearly, with a logical structure and sufficient comments that make it easy to follow what the code is doing? • Does the code make use of functions to improve the readability of the code?