University of South Wales
Prifysgol De Cymru

Faculty of Computing, Engineering and Science

Assessment Cover Sheet 2022-23

Module Code:	Module Title:	Module Team:
CS4S767	Data Mining	Shaily Jain Andrew Ware
Assessment Title:		**Assessment No.:**
Health Impact Factors		2
Date Set:	**Submission Date:**	**Return Date:**
21-Sep-2022 09:00	2-Dec-2022 23:59	16-Dec-2022 23:59

IT IS YOUR RESPONSIBILITY TO KEEP RECORDS OF ALL WORK SUBMITTED.

Marking and Assessment

This assignment will be marked out of **100%**.

This assignment contributes to **50%** of the total module marks.

Learning Outcomes to be assessed

As specified in the validated module descriptor https://icis.southwales.ac.uk

- Display knowledge of different data mining and Big Data tasks and appropriate models/algorithms, evaluating these with respect to their accuracy.
- Demonstrate the ability to apply data mining and Big Data concepts in appropriate contexts.

Awarded mark is only provisional: subject to change and / or confirmation by the Assessment Board.

Assessment Task

Produce a two-part report of approximately 3000 words (note your report should contain graphs, tables and code snippets that help explain what the report is saying).

Part 1 involves the analysis of a given data set to determine to what extent various characteristics (a person's gender, age, height, weight, and IQ) and lifestyle choices (the extent to which they consume alcohol, smoke and exercise) have on their health score index. Part 2 involves building a series of predictive models (informed by the work carried out in Part 1) to predict the health scores for a sample of the population. You are provided with two data files, the first, 'healthscore.csv' contains data for 5000 people relating to their characteristics and lifestyle choices (see above for details) and their health score index; the second file, 'population.csv' contains the individual characteristics and lifestyle choices (but no health score index) for 20 people.

Part 1 - Data Analytics

Produce a report that answers the questions listed below. The report, written in straightforward English, should contain appropriate graphs to help the reader understand the information. The analysis contained within the report needs to be produced using appropriate Python algorithms. (You can opt to use libraries or code your algorithms. More marks will be awarded to those who code at least some algorithms.) The report should also contain a rationale for the selection of algorithms used.

The report should answer the following questions:

1. Are there any significant differences between different population segments regarding their lifestyle choices (for example, male and female, different age groups)?
2. Which individual characteristics and lifestyle choices impact a person's health score (and to what extent)?
3. What would impact the overall population (regarding health score) if nobody consumed alcohol and did not smoke?

Part 2 - Predicting Health Scores

Use the data provided in 'healthscore.csv' to build a set of models capable of predicting the health score of an individual given a list of their characteristics and lifestyle choices. Models you may consider might include: Naive Bayes; Support Vector Machines; Tree-Based Algorithms; Regression Algorithms; various Neural Network paradigms. Your report should contain a justification for the selection of algorithms and techniques chosen. The report should also have a table showing the predicted values for each of the 20 individuals whose details are included in 'population.csv' The final table column should be labelled "Best estimate" and show your best estimate of the actual health scores. These might be determined by selecting the output of the model you think produces the best predictions or any combinations of model results. However, your report should articulate the method used for determining the values shown in the column and justify the approach you have taken.

Marking Scheme

	Fail (0/29)	Narrow Fail (30/39)	3rd Class / Pass (40/49)	Lower 2nd Class / Pass (50/59)	Upper 2nd Class / Merit (60/69)	1st Class / Distinction (70/100)
Background and introduction (10%)	☐ Missing or very superficial introduction	☐ Missing or very superficial introduction	☐ Gives a basic insight into the aim and content of the report	☐ Provides a reasonable explanation of the topic and its relevance	☐ Clear explanation of the aim, content and conclusions of the report	☐ Exceptionally clear explanation of the aim, content and conclusions of the report
Analysis (30%)	☐ Missing or very superficial	☐ Analytics techniques chosen are inappropriate ☐ Analysis performed and interpreted with many or significant errors ☐ Technical reporting of analysis contains many or significant errors ☐ Analysis performed is very basic and does not demonstrate a sufficient level of skill	☐ Minor errors in the choice of analytics techniques or justifications ☐ Analysis performed and interpreted with some errors ☐ Technical reporting of analysis contains some errors ☐ Analysis performed is of moderate complexity and demonstrates some level of skill	☐ Minor errors in the choice of analytics techniques or justifications ☐ Analysis performed and interpreted with some errors ☐ Technical reporting of analysis contains some errors ☐ Analysis performed is of moderate complexity and demonstrates some level of skill	☐ Analytics techniques are chosen appropriately, although justifications could be slightly clearer ☐ Analysis performed and interpreted with only minor errors ☐ Technical reporting of analysis contains minor errors ☐ Analysis performed is complex and demonstrates a good level of skill	☐ Analytics techniques are chosen appropriately with justifications ☐ All analyses performed and interpreted correctly ☐ Technical reporting of analysis is complete and correct ☐ Analysis performed is complex and demonstrates high level of skill
Predictions (40%)	☐ Missing or very superficial	☐ Predictive techniques chosen are inappropriate ☐ Only a single technique has been applied ☐ Predictions are not accurate	☐ Several errors in the choice of predictive techniques or justifications are unclear ☐ Only two techniques have been applied ☐ Predictions are not always accurate	☐ Minor errors in the choice of predictive techniques or justifications ☐ Two or fewer techniques have been applied ☐ Predictions are not always accurate	☐ Predictive techniques are chosen appropriately, although justifications could be slightly more apparent ☐ Three or fewer techniques have been applied ☐ Analyses performed have led to accurate predictions	☐ Predictive techniques are chosen appropriately with justifications ☐ Four or more techniques have been applied ☐ Analyses performed have led to accurate predictions
Findings and Recommendations (20%)	☐ Missing or very superficial	☐ Poor or superficial explanation of conclusions that contains many errors in answering questions	☐ Basic explanation of conclusions that answer some research questions accurately but also contain several errors	☐ Reasonable explanation of conclusions that answer most research questions accurately but also contain some errors	☐ Clear explanation of conclusions that answer research questions	☐ Thorough and concise explanation of conclusions that answer research questions effectively
Global:						