Student Assignment Brief

This document is intended for Coventry University Group students for their own use in completing their assessed work for this module. It must not be passed to third parties or posted on any website. If you require this document in an alternative format, please contact your Module Leader.

Contents:

The work you submit for this assignment must be your own independent work, or in the case of a group assignment your own groups' work. More information is available in the 'Assignment Task' section of this assignment brief.

Assignment Information

Module Name: Principles of Data Science

Module Code: 7144CEM

Assignment Title: Individual/Group Portfolio

Assignment Due: Friday 03/11/2023 at 6pm UK time

Assignment Credit: 10 credits

Word Count (or equivalent): 2000 words equivalent (not including reference list or output)

Assignment Type: Written

Percentage Grade (Applied Core Assessment). You will be provided with an overall grade between 0% and 100%. You have one opportunity to pass the assignment at or above 40%.

Assignment Task

This *Portfolio* involves analysing a given multivariate dataset using linear regression models and multivariate statistical methods. You are encouraged to explore the topic, use your initiative, and show some originality, within the time available.

There are two tasks in this portfolio: a group task and an individual task.

(1) For the group task, each group of three or four students will work together on a single report. Only one member of the group should submit one group report (e.g. as a single Microsoft Word document) on behalf of the whole group. Please clearly state (in your report) how each group member has contributed to this group task.

(2) For the individual task, please submit one individual report (e.g. as a single Microsoft Word document or a single PDF document), clearly organised by subtask.

Make sure you read each task through carefully. Aim to demonstrate your understanding of the topics and relevant module learning outcomes.

Dataset: Pima Indians Diabetes

The Pima Indians Diabetes Dataset involves predicting the onset of diabetes within 5 years in Pima Indians given medical details. So, there is a binary (2-class) target variable. The number of observations for each class is not balanced. There are 768 observations with 8 input variables and 1 output variable (diabetes and no diabetes). Missing values are believed to be encoded with zero values. The variable (column) names are as follows:

1. Number of times pregnant.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg).
4. Triceps skinfold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (weight in kg/(height in m)^2).
7. Diabetes pedigree function.
8. Age (years).
9. Class variable (0 or 1).

To select you own sample from the available dataset:

Please use the last three digits of your student ID number (or one of your group members if you work on task group) to take your own sample before any further analysis. For example, if your student ID is 546123, you should use set.seed function as below to fix your sample.

```
set.seed(123)
library(tidyverse)
# Read Pima Indians Diabetes Dataset as csv file
data = read_csv('Pima_Indians_Diabetes_Dataset1.csv')
# To take stratified random samples
strata_sample = data %>%
  group_by(Target) %>%
  slice_sample(n=200) %>%
  ungroup()
```

```
View(strata_sample)
# Select you own filename instead of 'your_dataset.csv' for your sample
write_csv(strata_sample, 'your_dataset.csv')
```

Task 1 (Group). Multivariate Statistical Analysis (MLO2)

This is a group task. Please be clear in your group report about how each group member has individually contributed to this group task. You may find the R package "factoextra" useful for this task. You must interpret and evaluate your results, not only write R code and build plots. Make sure it is clear which R code has produced which plots.

(1) Use R to carry out Principal Component Analysis (PCA) on the first **8 input variables** from your sample of the Pima Indians Diabetes Dataset.

   (a) Produce, interpret, and evaluate relevant plots such as scree plot, loadings plot, and biplot using PC1 and PC2. You can consider only the first three principal components for loadings plot.

   (b) Produce and interpret a biplot using PC2 and PC3 as the axes.

[15 marks]

(2) Use R to carry out Factor Analysis (FA) on the first **8 input variables** from your sample of the Pima Indians Diabetes Dataset.

   (a) By considering appropriate number of factors obtain Loadings, Communalities, and Specific Variances. Interpret and evaluate different factors that you have obtained.
   (b) Obtain and interpret the rotated version of the loadings.

[15 marks]

(3) Use R to carry out Cluster Analysis on the **8 input variables** from your sample of the Pima Indians Diabetes Dataset. You may find the agnes() function in the R package "cluster" useful, especially the agglomerative coefficient used for comparing clustering.

   (a) Using all 8 variables, cluster the observations (rows) and compare results using different distance metrics (Manhattan, Euclidean, etc) and hierarchical clustering methods (single linkage, Ward's method, etc) in a small table. Produce, interpret and evaluate at most two relevant dendrograms in your report.

   (b) Cluster the diabetes measurements (columns) using a combination of distance metric and hierarchical clustering method that gives the "best" result. Produce, interpret and evaluate the (one) corresponding dendrogram.

[15 marks]

(4) Briefly discuss any insights that PCA, FA and Cluster Analysis have in common and those that appear to conflict.

[5 marks]

Task 2 (Individual). Exploratory Data Analysis and Linear Models (MLO2)

This is an individual task. Please make sure it is all your own work. To save you a lot of work, please use the R package "olsrr" (where appropriate) to help carry out this task (see https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html*).*

In this task, use all diabetes variables to predict using linear models (regression) in R. We are primarily concerned with critically assessing any linear models proposed, and with model selection (which predictors to include in any final linear models recommended).

(1) Use R to build a *scatter matrix* using ggpairs(), including the class variable from Task 1 with value of *0 and 1* (either 0=this is not diabetes case or 1=it is a diabetes case). Identify and discuss any groups of strongly correlated variables. Justify which single-predictor linear model would "best" predict Diabetes pedigree function and fit this linear model to your sample of the dataset.

[10 marks]

(2) We wish to compare and critically assess several linear models to predict the variable of Diabetes pedigree function as below.

(a) Briefly explain the acronym "AIC" used to compare models.

(b) Consider the following linear models to predict Diabetes pedigree function for your taken sample. Compare these models and summarise your results in a compact table. *Justify your selection of linear models and critically assess any linear models you fit to the dataset, including a discussion of any relevant diagnostic plots*.

Model #1: The linear model to predict Diabetes pedigree function using Diastolic blood pressure and Body mass index as predictor variables.

Model #2: The best two-predictor linear model to predict Diabetes pedigree function using any of the remaining 8 variables.

Model #3: The best four-predictor linear model to predict Diabetes pedigree function using any of the remaining 8 variables.

[30 marks]

(3) Critically assess your conclusions from fitting linear models to the dataset and draw comparisons with the results from applying FA, PCA and Cluster Analysis to the dataset (from your group task). Discuss how you could communicate your conclusions to the participants represented in your dataset (not data scientists or medical experts).

[10 marks]

Submission Instructions:

There are two tasks in this portfolio: a group task and an individual task.

(1) For the group task, each group of three or four students will work together on a single report. Only one member of the group should submit one group report (e.g. as a single Microsoft Word document or a single PDF document) on behalf of the whole group. Please clearly state (in your report) how each group member has contributed to this group task.
(2) For the individual task, please submit one individual report (e.g. as a single Microsoft Word document or a single PDF document), clearly organised by subtask. **No collaboration with other students is permitted for individual task.**

Make sure you clearly reference any sources you have used using APA style referencing. Please include both in-text citations and a list of references for each task (where relevant).

Make sure you include your R code and relevant output/plots directly in each report. You must not submit a zip file. You must not submit a Jupyter notebook (but you can print a Jupyter notebook to a PDF file and submit the PDF file).

Do not use screenshots to include R code or text output into your report. Simply copy-and-paste R code and text output into your report. Submission is online via Aula using the two submissions box provided. *Do not leave uploading too late.*

Marking and Feedback

How will my assignment be marked?

Your assignment will be marked by the module team.

How will I receive my grades and feedback?

Provisional marks will be released once internally moderated. Feedback will be provided by the module team alongside grades release. Students will be able to access their feedback via Aula/Turnitin. Your provisional marks and feedback should be available within 2 weeks (10 working days).

What will I be marked against?

Details of the marking criteria for this task can be found at the bottom of this assignment brief.

Assessed Module Learning Outcomes

The Learning Outcomes for this module align to the marking criteria which can be found at the end of this brief. Ensure you understand the marking criteria to ensure successful achievement of the assessment task. The following module learning outcomes are assessed in this task:

MLO2. Investigate, develop, combine and critically assess approaches and solutions to challenges in data analysis, statistical modelling, and communication of results, both individually and as part of a team.

Assignment Support and Academic Integrity

If you have any questions about this assignment please see the Student Guidance on Coursework for more information.

Spelling, Punctuation, and Grammar:

You are expected to use effective, accurate, and appropriate language within this assessment task.

Academic Integrity:

The work you submit must be your own, or in the case of groupwork, that of your group. All sources of information need to be acknowledged and attributed; therefore, you must provide references for all sources of information and acknowledge any tools used in the production of your work, including Artificial Intelligence (AI). We use detection software and make routine checks for evidence of academic misconduct.

Definitions of academic misconduct, including plagiarism, self-plagiarism, and collusion can be found on the Student Portal. All cases of suspected academic misconduct are referred for investigation, the outcomes of which can have profound consequences to your studies. For more information on academic integrity please visit the Academic and Research Integrity section of the Student Portal.

Support for Students with Disabilities or Additional Needs:

If you have a disability, long-term health condition, specific learning difference, mental health diagnosis or symptoms and have discussed your support needs with health and wellbeing you may be able to access support that will help with your studies.

If you feel you may benefit from additional support, but have not disclosed a disability to the University, or have disclosed but are yet to discuss your support needs it is important to let us know so we can provide the right support for your circumstances. Visit the Student Portal to find out more.

Unable to Submit on Time?

The University wants you to do your best. However, we know that sometimes events happen which mean that you cannot submit your assessment by the deadline or sit a scheduled exam. If you think this might be the case, guidance on understanding what counts as an extenuating circumstance, and how to apply is available on the Student Portal.

Administration of Assessment

Module Leader Name: Dr Omid Chatrabgoun

Module Leader Email: ad8337@coventry.ac.uk

Assignment Category: Written

Attempt Type: Standard

Component Code: Por

Assessment Marking Criteria

Mark band	Outcome	Guidelines
90-100% Distinction	Meets learning outcomes	Distinction - Exceptional work with very high degree of rigour, creativity and critical/analytic skills. Mastery of knowledge and subject-specific theories with originality and autonomy. Demonstrates exceptional ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Innovative research with exceptional ability in the utilisation of research methodologies. Demonstrates, creativity, originality and outstanding problem-solving skills. Work completed with very high degree of accuracy, proficiency and autonomy. Exceptional communication and expression demonstrated throughout. Student evidences the full range of technical and/or artistic skills. Work pushes the boundaries of the discipline and may be strongly considered for external publication/dissemination/presentation.
80-89% Distinction		Distinction - Outstanding work with high degree of rigour, creativity and critical/analytic skills. Near mastery of knowledge and subject-specific theories with originality and autonomy. Demonstrates outstanding ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Innovative research with outstanding ability in the utilisation of research methodologies. Work consistently demonstrates creativity, originality and outstanding problem-solving skills. Work completed with high degree of accuracy, proficiency and autonomy. Outstanding communication and expression demonstrated throughout. Student demonstrates a very wide range of technical and/or artistic skills. With some amendments, the work may be considered for external publication/dissemination/presentation
70-79% Distinction		Distinction - Excellent work undertaken with rigour, creativity and critical/analytic skills. Excellent degree of knowledge and subject-specific theories with originality and autonomy demonstrated. The work exhibits excellent ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Innovative research with excellent ability in the utilisation of research methodologies. Work demonstrates creativity, originality and excellent problem-solving skills. Work completed with very consistent levels of accuracy, proficiency and autonomy. Excellent communication and expression demonstrated throughout. Student demonstrates a very wide range of technical and/or artistic skills.

60-69% Merit		Merit - Very good work often undertaken with rigour, creativity and critical/analytic skills. Very good degree of knowledge and subject-specific theories with some originality and autonomy demonstrated. The work often exhibits the ability to fully analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Very good research evidence and shows very good ability in the utilisation of research methodologies. Work demonstrates creativity, originality and problem-solving skills. Work completed with very consistent levels of accuracy, proficiency and autonomy. Very good communication and expression demonstrated throughout. Student demonstrates a wide range of technical and/or artistic skills.
50-59% Pass		Pass - Good work undertaken with some creativity and critical/analytic skills. Demonstrates knowledge and subject-specific theories with some originality and autonomy demonstrated. The work exhibits the ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Good research and shows some ability in the utilisation of research methodologies. Work demonstrates problem-solving skills and is completed with some level of accuracy, proficiency and autonomy. Satisfactory communication and expression demonstrated throughout. Student demonstrates some of the technical and/or artistic skills.
40-49% Pass		Pass - Assessment demonstrates some advanced knowledge and understanding of the subject informed by current practice, scholarship and research. Work may be incomplete with some irrelevant material present. Sometimes demonstrates the ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Acceptable research with evidence of basic ability in the utilisation of research methodologies. Demonstrates some originality, creativity and problem-solving skills but often with inconsistencies. Expression and presentation sufficient for accuracy and proficiency. Sufficient communication and expression with professional skill set. Student demonstrates some technical and/or artistic skills.
30-39% Fail	Fails to achieve learning outcomes	Fail - Very limited understanding of relevant theories, concepts and issues with deficiencies in rigour and analysis. Some relevant material may be present but be informed from very limited sources. Fundamental errors and some misunderstanding likely to be present. Demonstrates limited ability to analyse and apply concepts within the complexities and uncertainties of the subject/discipline. Limited research scope and ability in the utilisation of research methodologies. Limited originality, creativity, and struggles with problem-solving skills. Expression and presentation insufficient for accuracy and proficiency. Insufficient communication and expression and with deficiencies in professional skill set. Student demonstrates deficiencies in the range of technical and/or artistic skills.

20-29% Fail -		Fail - Clear failure demonstrating little understanding of relevant theories, concepts, issues and only a vague knowledge of the area. Little relevant material may be present and informed from very limited sources. Serious and fundamental errors and virtually no evidence of relevant research. Fundamental errors and misunderstandings likely to be present. Little or no research with no evidence of utilisation of research methodologies. No originality, creativity, and struggles with problem-solving skills. Expression and presentation insufficient for accuracy and proficiency. Insufficient communication and expression and with serious deficiencies in professional skill set. Student has clear deficiencies in range of technical and/or artistic skills.
0-19% Fail		Fail - Clear failure demonstrating no understanding of relevant theories, concepts, issues and no understanding of area. Little or no relevant material may be present and informed from minimal sources. No evidence of ability in the utilisation of research methodologies. No evidence of originality, creativity, and problem-solving skills. Expression and presentation deficient for accuracy and proficiency. Insufficient communication and expression and with deficiencies in professional skill set. Student has clear deficiencies in range of technical and/or artistic skills.