

# MODULAR PROGRAMME

# COURSEWORK ASSESSMENT SPECIFICATION

## Module Details

|  |  |  |
| --- | --- | --- |
| Module Code UFMFJR-15-M | Run 22Jan | Module Title  Advanced Statistics |
| Module Leader  Deirdre Toher | Module Coordinator | Module Tutors Deirdre Toher |
| Component and Element Number  Coursework CW1 | | Weighting: (% of the Module's assessment)  100% |

## Dates

|  |  |
| --- | --- |
| Date Issued to Students | Date to be Returned to Students |
| Submission Place Blackboard | Submission Date  09/05/2023 |
| Submission Time 14:00 |

## Deliverables

|  |
| --- |
| One markdown file that can be compiled to word. |

## Module Leader Signature

|  |
| --- |
| D Toher |

## Brief

You are working with a team who have been tasked to create a classification tool for use by a company doing chemical analysis.

They wish to know if they need to measure all variables, or if a subset of variables could achieve the same out-of-sample classification performance. Being able to use a subset of variables would mean faster processing times; but they would not wish to sacrifice classification performance for the sake of marginal speed gains.

### Your Task

Your role in this team is to create and build a markdown file (either using quarto or rmarkdown) to

1. Check for missing data;
2. Check for outliers and perform exploratory data analysis;
3. Create a training, test and validation split – with at least 15% of the observations to be in the validation set;
4. Investigate a variety of classification approaches and recommend the optimal one for this dataset;
5. Evaluate the performance of that approach on the validation dataset.

### Your Report

Therefore your short (no more than 2000 words) report should concentrate on describing the process, how decisions are made so that others in your team can explain it to the client.

### Your Data

You will have an individual dataset from blackboard (based on your student ID – called STUDENTID.csv) to create a model and test your model on.

This dataset of 2500 observations will have five different groups, labelled A-E.

### Include

You should include full appropriate Exploratory Data Analysis and Descriptive Statistics to highlight what, if any, data checking should be carried out prior to conducting the analysis.

You may select **any appropriate combination** of dimension reduction and classification techniques that we have covered.

You should report on your in-sample training and test classification performance and then also report on your out-of-sample validation performance and what, if any, differences are observed.

**Please structure your code so that the data is in a subdirectory called “data”, your .Rmd/ .qmd file should be in a subdirectory called “scripts” so that your relative file locations are correct.**

## **What will be marked?**

|  |  |  |
| --- | --- | --- |
| **Exploratory Data Analysis** | Correct and appropriate code and interpretation of output – including highlighting to the client what they need to check before proceeding on their data | 10 |
| **Code for dimension reduction and classification** | Correct and appropriate code and comments  See breakdown of marking allocation below: | 30 |
|  | Code for dimension reduction and / or variable selection | 10 |
|  | Code for classification process | 10 |
|  | Comments on code | 10 |
| **Explanation of process** | Your report explaining the implementation of the process, including what decisions were made and what evidence was used to inform those decisions.  Clarity of explanation is important. | 50 |
|  | Explanation of your implementation in a manner that a non-statistician could follow. | 20 |
|  | Decision making process – how evidence informed your decisions; for example on what classification approach to take. | 20 |
|  | General standard of English and presentation of your findings in a professional manner when your markdown file is compiled to word. | 10 |
| **Reproducibility of process** | Does your code run appropriately and make predictions on validation data?  Will your code run on a new dataset of the same format (called externalvalidation.csv) and be able to make predictions? | 10 |

*You should verify the word count prior to submission – the easiest way to do this is to compile a version that does not display your code to a Word document and use that.*

## **How will this be marked?**

## See the rubric below for guidelines about expectations as to what is deemed Excellent work all the way to Unacceptable work.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Rating** | | | | |
|  | **Excellent** | **Very Good** | **Satisfactory** | **Questionable** | **Unacceptable** |
| Description | Exemplary solution which demonstrates full comprehension of the skill. No errors. You have a clearly interpreted solution in highly articulate and appropriate Statistical and English language. | Cogent solution which demonstrates good comprehension of the skill. Your overall strategy was apparent and effective. Errors are present but insignificant. You have interpreted the solution in understandable Statistical and English language. | Poor solution which demonstrates little to no comprehension of the skill. The strategy was unclear or ineffective. Errors are striking. You misinterpreted the solution in part or used unclear Statistical and/or English language making the results unclear. | You have submitted an incomplete solution which demonstrates partial comprehension of the skill. The strategy you used had the potential to be effective. Errors are significant. You have interpreted the solution incompletely or misused Statistical and/or English language. | This is a poor solution which demonstrates little to no comprehension of the skill. Your approach strategy was unclear or ineffective, without having a potential to be effective. Errors are striking. You have misinterpreted your solution completely or used unclear Statistical and English language. |