3

**Assessment Submission Form**

|  |  |
| --- | --- |
| **Student name** | Shayan Kabiri |
| **Assessment Title** | Assignment 3 |
| **Module Code** | BSEN40870 |
| **Module Title** |  |
| **Module Coordinator** | Prof. Aoife Gowen |
| **Tutor** | Gözde Özdoğan |
| **Date Submitted** | /5/2024 |
| **Date Received** |  |
| **Grade/Mark** |  |

**A SIGNED COPY OF THIS FORM MUST ACCOMPANY ALL SUBMISSIONS FOR ASSESSMENT.**

**STUDENTS SHOULD KEEP A COPY OF ALL WORK SUBMITTED**.

**Procedures for Submission and Late Submission** There are penalties for the late submission of assessments. For further information please see the University’s Policy on Late Submission of Coursework.

**Plagiarism:** the unacknowledged inclusion of another person’s writings or ideas or works, in any formally presented work (including essays, examinations, projects, laboratory reports or presentations). The penalties associated with plagiarism designed to impose sanctions that reflect the seriousness of University’s commitment to academic integrity. Ensure that you have read UCD’s student plagiarism policy.

**Declaration of Authorship**

I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

Signed....................................................... Date ...................................................



**Title:** BSEN40870 Assignment 3: Analysis of multivariate datasets: creation and critical

evaluation of regression and classification models.

**Objective**

To create regression and classification models for the provided multivariate data using the

tools introduced in the course.

**Dataset description**  
You can download the “Assignment\_data” dataset from your Brightspace locker. This  
MATLAB workspace contains 2 sets of data and a Wavelength vector (which describes the variables in the X matrices). Set 1 contains a calibration set (“cal\_X”) of 150 samples measured at 100 different wavelengths, and a prediction set (“pred\_X”) of 50 samples, measured at the same 100 wavelengths as the calibration set. The independent variable Y represents the concentrations of 5 different components in each sample. Set 2 contains a calibration set (“cal\_X”) of 100 samples measured at 100 different wavelengths, and a prediction set (“pred\_X”) of 100 samples, measured at the same 100 wavelengths as the calibration set. The independent variable Y represents the class membership of each sample.

**Assignment Instructions**

For Set 1 of data:

* Inspect the data and explore with PCA. Identify and remove any outliers.
* Build PCR and PLSR models to predict all 5 Y variables provided
* Optimise meta-parameter selection for both PCR and PLSR
* Compare the performance of each model on the prediction set
* Apply 2 different variable selection techniques to find the most important variables
* Compare the PLSR model built on the full dataset and on the selected variables

For Set 2 of data:

* Inspect the data and explore with PCA. Identify and remove any outliers.
* Build 2 types of classification models to predict class membership
* Compare the performance of each model on the prediction set
* Apply 2 different variable selection techniques to find the most important variables
* Compare the classification model built on the full dataset and on the selected variables

**Methodology**

**Dataset 1:**

First the PCAs, the variance explained by each PCA and correlation of each response to the first five PCAs were investigated. Then the response boxplots were used to determine if there are any outliers in the dataset. Outlier datapoints were determined by quartiles method and a dataset without outliers were created.

The tuning mechanism for both Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) was based on 5-fold cross validation and fining the number of components used in the model that produces the best root mean square error (RMSE).

Two variable selection methods, minimum redundancy maximum relevance (MRMR) and F-tests feature ranking (FFR), were used to select the 10 most important variables for regression.

Models were tuned and regression metrics were produced for PCR and PLSR, for datasets with and without outliers and selected variables based on MRMR and FFR for the four responses.

**Dataset 2:**

First the PCAs, the variance explained by each PCA and degree of apparent separation by PCAs were investigated. Predictor outliers were detected using the quartiles method and a dataset without predictor outliers were created.

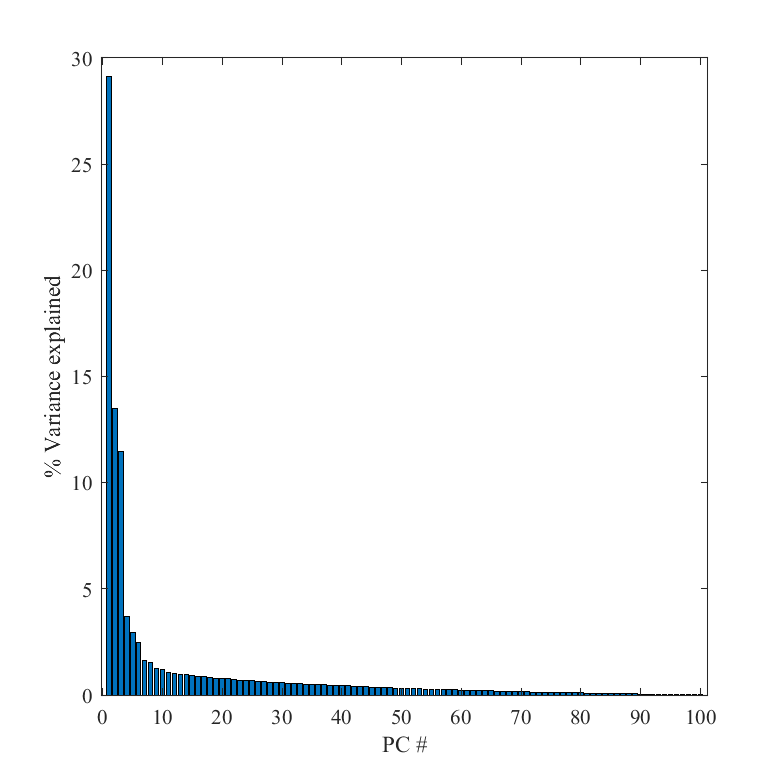
Responses were converted to binary and a PCR and PLSR were adapted to act as classifiers. Predicted responses closer to 1 were converted to 1 and predicted responses closer to 0 were converted to 0.

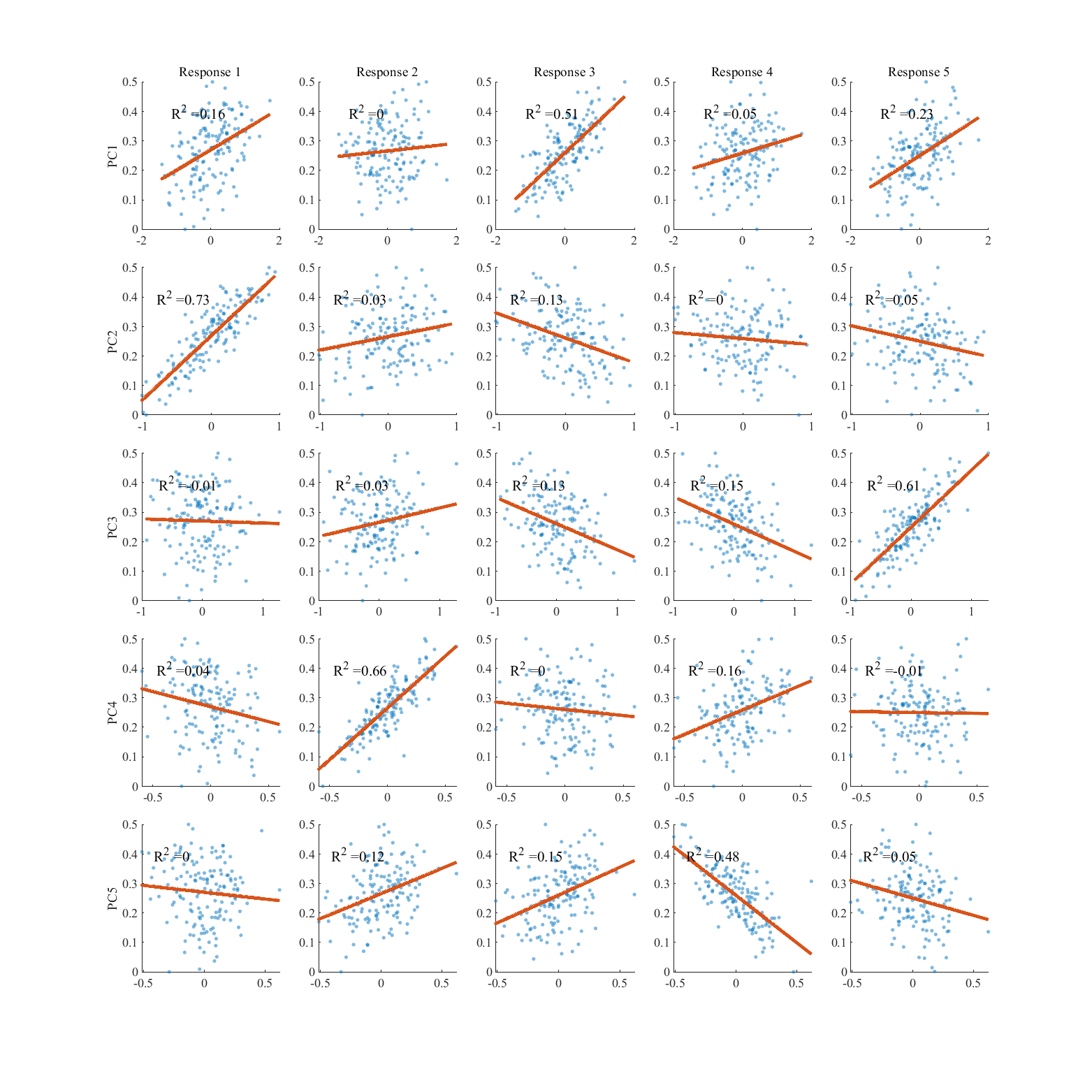
Two variable selection methods, minimum redundancy maximum relevance (MRMR) and F-tests feature ranking (FFR), were used to select the 10 most important variables for classification.

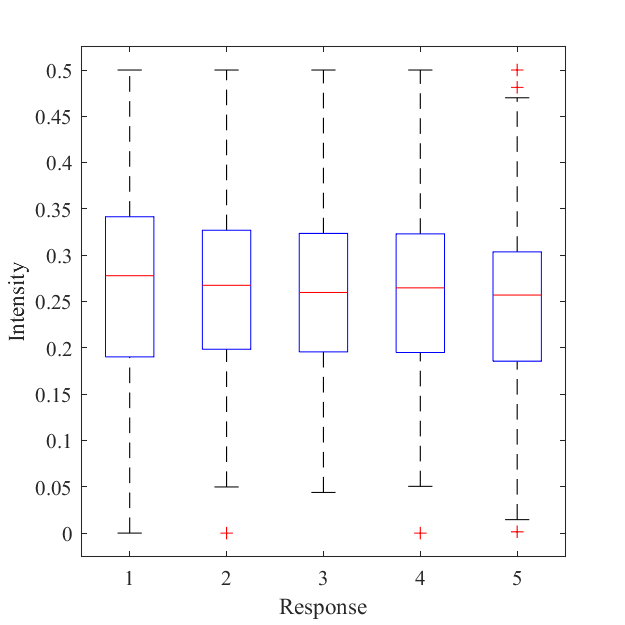
Models were tuned and classification metrics were produced for PCR and PLSR, for datasets with and without outliers and selected variables based on MRMR and FFR for response classes.

**Results & Short Discussion**

**Dataset 1:**

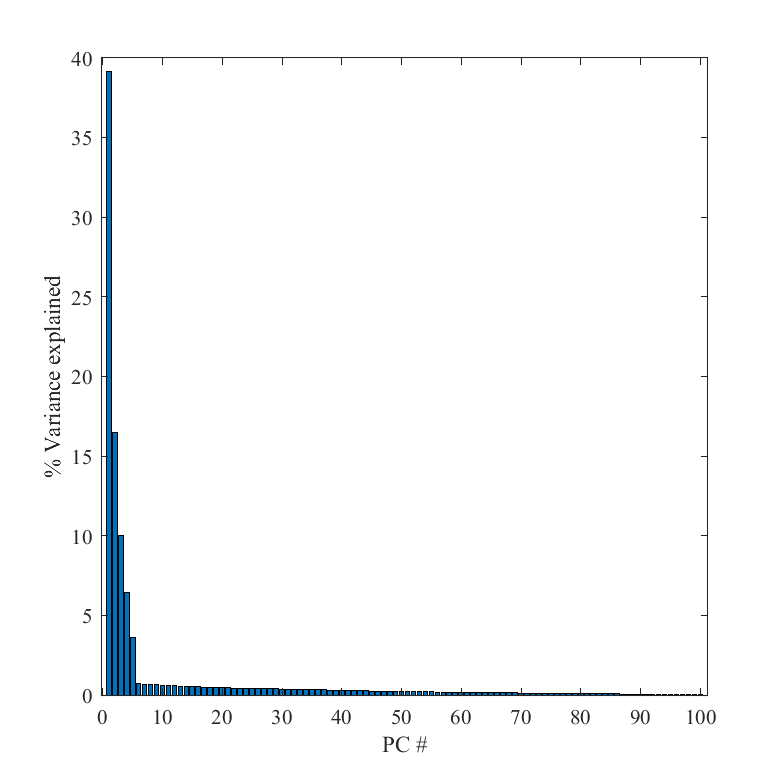
****

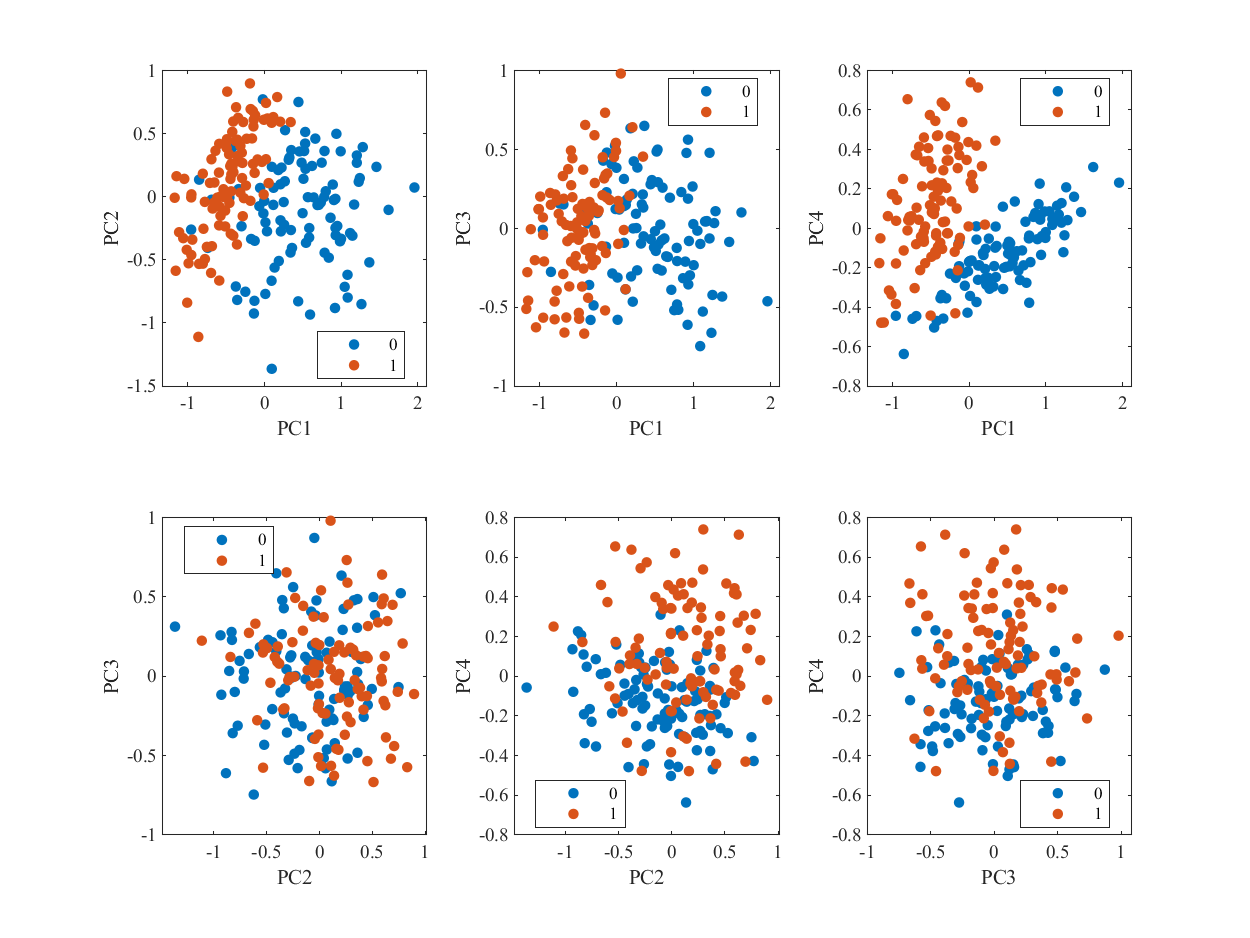
****

****

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Response | R2 test PCR | R2 test PCR WOO | R2 test PLSR | R2 test PLSR WOO | R2 test PLSR FFS | R2 test PLSR MRMR | |
| 1 | 0.94 | 0.94 | 0.9 | 0.92 | 0.92 | 0.88 |  |
| 2 | 0.84 | 0.83 | 0.85 | 0.85 | 0.66 | 0.74 |  |
| 3 | 0.97 | 0.96 | 0.97 | 0.97 | 0.92 | 0.89 |  |
| 4 | 0.86 | 0.87 | 0.87 | 0.86 | 0.41 | 0.42 |  |
| 5 | 0.95 | 0.95 | 0.95 | 0.96 | 0.87 | 0.81 |  |

**Dataset 2:**

****

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Accuracy for WOO dataset | Accuracy for FFR | Accuracy for MRMR |
| PCA Classifier | 0.995 | 0.995 | 0.945 | 0.935 |
| PLS Classifier | 0.895 | 0.94 | 0.935 | 0.955 |

**Conclusions**

**Appendix**