LEARNING
AUTONOMOUS
TA MINING
DEE
MACHINE
LEARNING
RTIFICIAL
NTELLIGENCE
ALGORITHM

Principal Component Analysis

TASK

For your video presentation, you must demonstrate your PCA analysis on the continuous features of the WACY-COM dataset and interpret the results. **Submit the recording via the Panopto link on Canvas.** Please ensure you follow the instructions carefully.

The due date for this assessment is **Friday of Week 7 on 11 April 2025 before midnight**.

Perform PCA and Visualise Data

(i) First, copy the code below to a R script. Enter your student ID into the command **set.seed(.)** and run the whole code. The code will create a sub-sample of 400 that is unique to you.

```
#You may need to change/include the path of your working directory

#Import the dataset into R Studio.
dat <- read.csv("WACY-COM.csv", na.strings=NA, stringsAsFactors=TRUE)

set.seed(Enter your student ID here)

#Randomly select 400 rows
selected.rows <- sample(1:nrow(dat),size=400,replace=FALSE)

#Your sub-sample of 400 observations
mydata <- dat[selected.rows,]

dim(mydata)  #check the dimension of your sub-sample
```

(ii) Extract only the **numeric features** and the **APT feature** from the WACY-COM dataset and store them as a data frame/tibble. Refer to Assignment 1 for the feature description if needed.

(iii) Clean the extracted data based on the feedback received from Assignment 1.

(iv) Remove the incomplete cases to make it usable in "R" for PCA.

(v) Perform PCA using *prcomp(.)* in R, but **only on the numeric features** (i.e. ignore APT in this step).

- Explain why you believe the data should or should not be scaled, i.e. standardised, when performing PCA.

- Display and describe the individual and cumulative proportions of variance (**3 decimal places**) explained by each of the principal components.
- Outline how many principal components are adequate to explain at least 50% of the variability in your data.
- Display and interpret the coefficients (or loadings) to **3 decimal places** for PC1, PC2 and PC3. Describe which features (based on the loadings) are the key drivers for each of these three principal components.

(vi) Create and display the biplot for PC1 vs. PC2 to visualise the PCA results in the first two dimensions. Colour-code the points based on the APT feature. Explain the biplot by commenting on the PCA plot and the loadings plot individually, and then both plots combined (see Slides 28–29 of Module 3 notes). Finally, comment on and justify which (if any) features can help distinguish APT activity.

(vii) Based on the results from parts (v) and (vi), describe

- **whether PC1 or PC2 (choose one) best assists in classifying APT.** Hint: Project all points in the PCA plot onto the PC1 axis (i.e. consider the PC1 scores only) and assess whether there is a clear separation between known and unknown APT actors. Then, project onto the PC2 axis (i.e. consider the PC2 scores only) and evaluate whether the separation is better than in PC1. You can access the PCA scores for PC1 and PC2 via *mypca$x*, assuming *mypca* contains your PCA results from *prcomp(.)*.
- the key features in this dimension that can drive this process (Hint: based on your decision above, examine the loadings from part (v) of your chosen PC and choose those whose absolute loading (i.e. disregard the sign) is greater than 0.3).

Video Presentation Checklist

1. In your video presentation, you **must**
   a. Run your code corresponding to parts (i) to (vii) above
   b. Display the relevant output
   c. Interpret the output

2. Your video presentation **must** include a camera shot of yourself in the video capture, unless there is an exceptional reason and is supported by a Learning Assessment Plan (LAP). **20% is automatically deducted from your final mark if this is not included in your video presentation. If you choose to record with another application, you must make sure that this feature is included.**

3. Your video presentation **must** be between 6-8 minutes long.

Marking Rubrics

Criteria	Fail <0-49%	Pass 50-59%	Credit 60-69%	Distinction 70-79%	High Distinction 80-100%
Working Code (7%)	Code does not run or contains major flaws, preventing meaningful PCA analysis. Little to no documentation.	Code has significant errors or omissions that affect PCA output. Poor documentation and some redundancy.	Code has a few errors and/or does not fully achieve intended PCA and relevant analyses. Documentation is present but could be improved.	Code runs with minor issues but still performs PCA and relevant tasks correctly. Minimal redundancy and good documentation.	Code runs flawlessly, correctly performs PCA and relevant tasks, and produces meaningful outputs. No errors, redundant code, or inefficiencies.
Interpretation of results (18%)	Fails to interpret the PCA results meaningfully or provides incorrect conclusions.	Interpretation is vague, lacks depth, and/or has major inaccuracies or errors.	Provides a basic interpretation with some inaccuracies or missing key insights.	Provides a strong and mostly accurate interpretation of PCA results with minor omissions or inaccuracies.	Provides an in-depth, clear, and accurate interpretation of PCA results, including the significance of principal components and key loadings. Justifies conclusions with evidence.
Presentation skills (7%)	The presentation is unclear. The presenter made an attempt at expression, but the pace and tone need improvement to better engage the audience.	The presentation lacks structure. Presenter made a good attempt, but the expression, pace, and tone could be improved.	The presentation is understandable and delivered at a good pace. However, there is minimal confidence in the presentation style.	Clear and structured presentation with minor pacing or engagement issues. Presenter was fluent and displayed good confidence.	The presenter was dynamic, natural, and persuasive, with an appropriate tone. Delivery was clear, confident, and well-structured, with effective pacing and engagement that maintained a high level of confidence throughout.
Timing (3%)	Presentation is less than 4 minutes or more than 11 minutes.	Presentation is between 4 and 5 minutes, or between 10 and 11 minutes	Presentation is between 5 and 6 minutes, or between 9 and 10 minutes	Presentation is between 8 and 9 minutes	Presentation is between 6 and 8 minutes

Academic Misconduct

Edith Cowan University regards academic misconduct of any form as unacceptable. Academic misconduct, which includes but is not limited to, plagiarism; unauthorised collaboration; cheating in examinations; theft of other student's work; collusion; inadequate and incorrect referencing; will be dealt with in accordance with the ECU Rule 40 Academic Misconduct (including Plagiarism) Policy. Ensure that you are familiar with the Academic Misconduct Rules.

Assignment Extensions

Instructions to apply for extensions are available on the ECU Online Extension Request and Tracking System to formally lodge your assignment extension request. The link is also available on Canvas in the Assignment section.

Normal work commitments, family commitments and extra-curricular activities are not accepted as grounds for granting you an extension of time because you are expected to plan ahead for your assessment due dates.

Where the assignment is submitted not more than 7 days late, the penalty shall, for each day that it is late, be 5% of the maximum assessment available for the assignment. Where the assignment is more than 7 days late, a mark of zero shall be awarded.