Assignment 1

Data Cleaning and Visualisation

Scenario

You have been provided an export from DCE's incident response team's security information and event management (SIEM) system. The incident response team extracted alert data from their SIEM platform and have provided a .CSV file (MLData2023.csv), with 500,000 event records, of which approximately 3,000 have been 'tagged' as malicious.

The goal is to integrate machine learning into their Security Information and Event Management (SIEM) platform so that suspicious events can be investigated in real-time. *security data.*

Data description

Each event record is a snapshot triggered by an individual network 'packet'. The exact triggering conditions for the snapshot are unknown. But it is known that multiple packets are exchanged in a 'TCP conversation' between the source and the target before an event is triggered and a record created. It is also known that each event record is anomalous in some way (the SIEM logs many events that may be suspicious).

A very small proportion of the data are known to be corrupted by their source systems and some data are incomplete or incorrectly tagged. The incident response team indicated this is likely to be less than a few hundred records. A list of the relevant features in the data is given below.

Assembled Payload Size (continuous)	The total size of the inbound suspicious payload. Note: This would contain the data sent by the attacker in the "TCP conversation" up until the event was triggered
DYNRiskA Score (continuous)	An un-tested in-built risk score assigned by a new SIEM plug-in
IPV6 Traffic (binary)	A flag indicating whether the triggering packet was using IPV6 or IPV4 protocols (True = IPV6)
Response Size (continuous)	The total size of the reply data in the TCP conversation prior to the triggering packet
Source Ping Time (ms) (continuous)	The 'ping' time to the IP address which triggered the event record. This is affected by network structure, number of 'hops' and even physical distances.

	E.g.: • < 1 ms is typically local to the device • 1-5ms is usually located in the local network • 5-50ms is often geographically local to a country • ~100-250ms is trans-continental to servers • 250+ may be trans-continental to a small network. *Note, these are estimates only and many factors can influence ping times.*
Operating System (Categorical)	A limited 'guess' as to the operating system that generated the inbound suspicious connection. This is not accurate, but it should be somewhat consistent for each 'connection'
Connection State (Categorical)	An indication of the TCP connection state at the time the packet was triggered.
Connection Rate (continuous)	The number of connections per second by the inbound suspicious connection made prior to the event record creation
Ingress Router (Binary)	DCE has two main network connections to the 'world'. This field indicates which connection the events arrived through
Server Response Packet Time (ms) (continuous)	An estimation of the time from when the payload was sent to when the reply packet was generated. This may indicate server processing time/load for the event
Packet Size (continuous)	The size of the triggering packet
Packet TTL (continuous)	The time-to-live of the previous inbound packet. TTL can be a measure of how many 'hops' (routers) a packet has traversed before arriving at our network.
Source IP Concurrent Connection (Continuous)	How many concurrent connections were open from the source IP at the time the event was triggered
Class (Binary)	Indicates if the event was confirmed malicious, i.e. 0 = Non-malicious, 1 = Malicious

The raw data for the above variables are contained in the **MLData2023.csv** file.

Objectives

The data were gathered over a period of time and processed by several systems in order to associate specific events with confirmed malicious activities. However, the number of confirmed malicious events was very low, with these events accounting for less than 1% of all logged network events.

Because the events associated with malicious traffic are quite rare, rate of 'false negatives' and 'false positives' are important.

Your initial goals will be to

- Perform some basic exploratory data analysis
- Clean the file and prepare it for Machine Learning (ML)
- Perform an initial Principal Component Analysis (PCA) of the data.
- Identify features that may be useful for ML algorithms
- Create a brief report to the rest of the research team on your findings

Task

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 that is unique to you.

```
# You may need to change/include the path of your working directory
dat <- read.csv("MLData2023.csv", stringsAsFactors = TRUE)

# Separate samples of non-malicious and malicious events
dat.class0 <- dat %>% filter(Class == 0) # non-malicious
dat.class1 <- dat %>% filter(Class == 1) # malicious

# Randomly select 300 samples from each class, then combine them to form a working dataset
set.seed(Enter your student ID here)
rand.class0 <- dat.class0[sample(1:nrow(dat.class0), size = 300, replace = FALSE),]
rand.class1 <- dat.class1[sample(1:nrow(dat.class1), size = 300, replace = FALSE),]

# Your sub-sample of 600 observations
mydata <- rbind(rand.class0, rand.class1)

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

Use the *str(.)* command to check that the data type for each feature is correctly specified. Address the issue if this is not the case.

You are to clean and perform basic data analysis on the relevant features in **mydata**, and as well as principal component analysis (PCA) on the continuous variables. This is to be done using "R". You will report on your findings.

Part 1 – Exploratory Data Analysis and Data Cleaning

(i) For each of your **categorical** or **binary** variables, determine the number (%) of instances for each of their categories and summarise them in a table as follows. **State all percentages in 1 decimal places.**

Categorical Feature	Category	N (%)
Feature 1	Category 1	10 (10.0%)
	Category 2	30 (30.0%)
	Category 3	50 (50.0%)
	Missing	10 (10.0%)
Feature 2 (Binary)	YES	75 (75.0%)
	NO	25 (25.0%)
	Missing	0 (0.0%)
…	…	…
Feature *k*	Category 1	25 (25.0%)
	Category 2	25 (25.0%)
	Category 3	15 (15.0%)
	Category 4	30 (30.0%)
	Missing	5 (5.0%)

(ii) Summarise each of your continuous/numeric variables in a table as follows. State all values, except N, to **2 decimal places**.

Continuous Feature	Number (%) missing	Min	Max	Mean	Median	Skewness
Feature 1						
Feature2						
….	….	….	….	….	….	….
Feature k						

Note: The tables for subparts (i) and (ii) should be based on the original sub-sample of 600 observations, not the cleaned version.

(iii) Examine the results in sub-parts (i) and (ii). Are there any invalid categories/values for the categorical variables? Is there any evidence of outliers for any of the continuous/numeric variables? If so, how many and what percentage are there?

Part 2 – Perform PCA and Visualise Data

(i) For all the observations that you have deemed to be invalid/outliers in Part 1 (iii), mask them by replacing them with NAs using the *replace(.)* command in **R**.

(ii) Export your "cleaned" data as follows. This file will need to be submitted along with you report.

```
#Write to a csv file.
write.csv(mydata,"mydata.csv")
```

** Do not read the data back in and use them for PCA **

(iii) Extract only the data for the **numeric features** in **mydata**, along with **Class,** and store them as a separate data frame/tibble. Then, **filter the incomplete cases (i.e. any rows with NAs)** and perform PCA using *prcomp(.)* in R, but **only on the numeric features (i.e. exclude Class)**.

- Outline why you believe the data should or should not be scaled, i.e. standardised, when performing PCA.
- Outline the individual and cumulative proportions of variance (**3 decimal places**) explained by each of the first 4 components.
- Outline how many principal components (PCs) are adequate to explain at least 50% of the variability in your data.
- Outline the coefficients (or loadings) to **3 decimal places** for PC1, PC2 and PC3, and describe which features (based on the loadings) are the key drivers for each of these three PCs.

(iv) Create a biplot for PC1 vs PC2 to help visualise the results of your PCA in the first two dimensions. Colour code the points with the variable **Class**. Write a paragraph to explain what your biplots are showing. That is, comment on the PCA plot, the loading plot individually, and then both plots combined (see Slides 28-29 of Module 3 notes) and outline and justify which (if any) of the features can help to distinguish Malicious events.

(v) Based on the results from parts (iii) to (iv), describe which dimension (have to choose one) can assist with the identification of Malicious events (Hint: project all the points in the PCA plot to PC1 axis and see whether there is good separation between the points for Malicious and Non-Malicious events. Then project to PC2 axis and see if there is separation between Malicious and Non-Malicious events, and whether it is better than the projection to PC1.

What to Submit

1. A single report (**not exceeding 5 pages, does not include cover page, contents page and reference page, if there is any**) containing:
    a. summary tables of all the variables in the dataset;
    b. a list of data issues (if any);
    c. your implementation of PCA and interpretation of the results, i.e. variances explained, and the contribution of each feature for PC1, PC2 and PC3;
    d. PC1 vs PC2 biplot and its interpretation;
    e. your explanation of selection and contribution of the features with respect to possible identification of Malicious events.

If you use any references in your analysis or discussion outside of the notes provided in the unit, you must cite your sources.

2. The dataset containing your sub-sample of 600 observations, i.e., **mydata**.
3. A copy of your R code.

The report must be submitted through **TURNITIN** and checked for originality. The R code and data file are to be submitted separately via a Canvas submission link.

Note that no marks will be given if the results you have provided cannot be confirmed by your code.

Marking Criteria

Criterion	**Contribution to assignment mark**
Correct implementation of descriptive analysis, data cleaning and PCA in R • Working code • Masking of invalid/outliers done correctly • External sources of code in **in APA 7 referencing style** (if applicable) • Good documentation/commentary	20%
Correct identification of missing and/or invalid observations in the data with justifications.	10%
Accurate specification and interpretation of the contribution of principal components and its loading coefficients. • Explain why you should scale the observations when running PCA. • Outline the individual and cumulative proportion of variance explained, and comment on the number of components required to explain at least 50% of the variance. • Outline the loadings (to specified decimal place) and comment as to their contribution to its respective PC. • Tabulation of results – no screenshot	15%
Accurate biplot, with appropriate interpretation presented • 2-d with clear labels • Interpretation of each biplot i) PCA plot – Clustering? Separation? ii) Loadings plot – vectors (features) and its relation to each of the dimension, and as well as to each other. • PCA + Loadings plot – Do any of the features appears to be able to assist with the classification of Malicious events and how?	25%

Appropriate selection of dimension for the identification Malicious events with justification. • Choose a dimension, i.e. PC or PC2 and justify why it's the best for classifying Malicious events	10%
Presentation and communication skills – Tables (no screenshots) and figures are well presented and appropriately captioned and are referenced in text. Report, analysis and overall narrative is well-articulated and communicated. • All figures and tables should be labelled/captioned appropriately and referenced in text. The labels in the plots should be clear. • Solutions should be in the order that the questions were posed in the assignment. • Spelling and grammatical errors should be kept to a minimum. • Overall narrative – all interpretation should be in the context of the study.	20%
Total	100%

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 for Academic Misconduct (including Plagiarism) Policy. Ensure that you are familiar with the Academic Misconduct Rules.

Assignment Extensions

Applications for extensions must be completed using the ECU Application for Extension form, which can be accessed online.

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

Please submit applications for extensions via email to both your tutor and the Unit Coordinator.

Where the assignment is submitted no 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.