作業 2

機器學習建模

引用和重要披露給學生們：以下數據和情境完全是虛構的。它們是使用合成數據創建的，以匹配真實世界的情況。這些數據是使用一系列工具創建的，包括真實世界安全漏洞的統計模型，以及對真實世界安全漏洞的人工智能/機器學習分析。

背景說明

鑑於最近對醫療保健行業的攻擊，以及一些引人注目的數據泄露事件，FauxCura Health已聘請了Quantum.LogiGuardian（Q.LG），一家專門從事網絡安全諮詢和分析的公司。

FauxCura 相信他們的系統可能存在未被檢測到的網絡安全漏洞。作為一家關心和值得尊敬的醫療服務提供商，他們希望檢查他們的歷史網絡數據，以確定是否發生了未被檢測到的漏洞。
FauxCura 公司的安全運營中心運行一個名為 Splunk™ 的安全事件與事件管理（SIEM）平台。該平台從伺服器、桌面電腦、路由器和其他網絡設備收集大量的日誌數據，並將其匯總成報告和警報的形式，供安全人員查看，以識別需要調查的事件。

在對 FauxCura 的數據進行初步調查期間，Q.LG 能夠深入檢視
歷史數據可生成一份初始報告，提供基於特定觸發器的事件的一些頂級指標。特定事件的數據也被保留，但通常包括非常大的數據集。

希望報告數據包含足夠的信息，以便構建一個能更準確識別感興趣事件的 ML 模型。

數據概況

您正在處理的數據是從 FauxCura 的 SIEM 中提取的記錄。
records have already been processed and reduced to a summary of individual event detections that were triggered by the SIEM.

數據還從安全資訊與事件管理系統的多個其他來源和報告中進行了匯總。這意味著某些數值在系統之間可能不一致，或者數據中可能存在需要識別和清理的錯誤。

特點描述:

以下是數據集中特徵的簡要說明。不需要理解這些特徵。同時，特徵命名慣例也很重要。

這些特徵並非必須被理解為非常主觀的。過度依賴名稱的意義可能會忽略重要的數據或細節。

警報類別（分類）：此功能描述了 Splunk 創建的警報類型。這在很大程度上是主觀的，因為警報創建者可以為不同類型的事件識別自己的警報級別。數據中存在的級別可以大致總結為：

信息性：正在記錄到系統中僅供信息目的的事件，這些可能與惡意活動有關，但這是最低級別的警報。

警告：這是一個較高級別的警報，通常用於識別可能不尋常的情況。

警報：這些通常用於代表需要採取行動的安全問題的特定事件。

網路事件類型（分類）：這是 SIEM 報告認為已發生的事件類型。它可用於區分表面上正常的網路流量，到像是政策違規、威脅偵測甚至資料外洩等事情。

正常運作：在此記錄事件中並未發生特定異常–有許多原因可能導致此數據被記錄。

政策違規：已違反安全或業務政策。這可能涵蓋從嘗試在網絡上運行未經授權的軟件，到使用錯誤類型的瀏覽器訪問數據庫。

威脅偵測：已檢測到一個特定條件，此前已被識別為安全威脅。這些可能是被誤標記的正常操作，也可能包含惡意軟體或正在使用的技術。

網路互動類型（分類）：
這是另一個使用未知第3方插件的「電腦」指標，用於識別非典型的網絡互動。

這些看起來是正常的網絡流量請求。

Elevated: Requests that are attempting to access resources that require specific permissions. For example, a computer trying to log in to an administrative console or a restricted device.

可疑：通常這些是意外的、來自意外來源或具有意外使用模式的高級網絡事件。

異常:
網絡互動可能不典型，但可能與安全事件無關。

Critical: 一種永不應該發生的網路狀況。這可能是指示攻擊狀況的互動，或嚴重的設備停機或故障。

未知：互動狀態未知

資料傳輸量（出和入）（數值）：量化在網路上傳輸的資料量。無論是進入網路還是離開網路，都會給出數值。

TransactionsPerSession（整數）：設備之間以及它們正在通信的服務之間交換的交易數量。

網路存取頻率（整數）：衡量網路埠口存取的頻繁程度，異常頻率可能表示未經授權的存取嘗試或掃描。

UserActivityLevel（數字）：一個生成的指標，顯示使用者在他們所連接的系統上的活躍程度。
較高的分數通常表示更多的活動。

系統存取率（整數）：
一個生成的指標，顯示公司核心系統被訪問的頻率。

SessionIntegrityCheck (Logical): 一個指示會話是否已正確開啟、通訊並關閉的標誌，所有底層網絡協議和信號都正確使用。

ResourceUtilizationFlag（邏輯）：當伺服器或網路設備的資源利用率異常高時會觸發的標誌。這可能包括某些設備的記憶體消耗過多、反應時間過慢或大型網路傳輸。

安全風險等級（數字）：
一個由第 3 方「AI」插件創建的計算指標，可以根據未知的參數和條件識別安全風險。

回應時間（毫秒）（數值）：衡量回應網絡請求或事件所需的時間。這是網絡資源或事件發生後，對應的回覆封包返回之間的時間。

分類（分類）：
事件的最終分類。在標註為“正常”和“惡意”時，可以假定已經被合理準確地識別。

以上變數的原始數據包含在 HealthCareData_2024.csv 檔案中。

草堆中的針筒

數據是在一段時間內收集的，並通過多個系統處理，以將特定事件與確認的惡意活動聯繫起來。然而，確認的惡意事件數量非常少，這些事件大約佔所有記錄的網絡事件的 4%。

儘管惡意事件相當罕見，但識別惡意事件極為重要。

目標

你是被 Q.LG 聘請來擔任數據科學家的人，負責檢視數據並提供洞察。你的目標將是

- 整理資料檔案，為機器學習（ML）做準備
- 推薦一個機器學習演算法，可以提供最準確的惡意事件檢測。
- 創建一份關於您的研究發現的簡報

您的工作

您的工作是開發檢測演算法，以提供最準確的事件檢測。您無需擔心 SIEM 外掛程式或軟體整合的細節，也就是說，您的任務是專注於使用 R 準確分類惡意事件。

您將測試和評估兩種機器學習演算法（每種在兩種情境下），以確定哪種監督式學習模型最適合所描述的任務。

Task

您需要導入並清理與上一個任務中使用的相同 HealthCareData_2024.csv 檔案。然後運行、調整並評估兩種監督式機器學習演算法（每種演算法使用兩種類型的訓練數據），以確定最準確的惡意事件分類方式。

第一部分 – 一般數據準備和清理

a) 將 HealthCareData_2024.csv 匯入 R Studio。這個版本與 Assignment 1 相同。

b) 在 R Studio 中編寫適當的代碼，準備並清理 HealthCareData_2024 資料集，步驟如下：i. 根據 Assignment 1 收到的反饋清理整個資料集。

ii. 對於特徵 NetworkInteractionType，合併“Regular”和將「未知」類別一起合併為「其他」類別。提示：使用 forcats::fct_collapse(.) 函數。

iii. 使用 thena.omit(.) 函數僅選擇完整案例，並將數據集命名為 dat.cleaned。

請在您的報告中簡要概述準備和清潔過程，並解釋您認為上述步驟為何是必要的。

c) 使用以下代碼生成兩個訓練數據集（一個不平衡的 mydata.ub.train 和一個平衡的 mydata.b.train），以及測試集（mydata.test）。請確保將您的學生證號碼輸入到命令 set.seed(.) 中。

```
# 分開正常和惡意事件的樣本
dat.class0 <- dat.cleaned %>% filter(Classification == "正常") # 正常 dat.class1
<- dat.cleaned %>% filter(Classification == "惡意") # 惡意

# 隨機選擇 9600 個非惡意和 400 個惡意樣本，使用您的學生證號碼，然後將它們結合起來形成一個工作數據集
set.seed(輸入您的學生證號碼)
rows.train0 <- sample(1:nrow(dat.class0), size = 9600, replace = FALSE)
rows.train1 <- sample(1:nrow(dat.class1), size = 400, replace = FALSE)

# 您的 10000 個「不平衡」訓練樣本
train.class0 <- dat.class0[rows.train0,] # 非惡意樣本 train.class1
<- dat.class1[rows.train1,] # 惡意樣本 mydata.ub.train <-
rbind(train.class0, train.class1)

# 您的 19200 '平衡' 訓練樣本，即每個 9600 個正常和惡意樣本。

set.seed(輸入您的學生證號碼)
```

```
train.class1_2 <- train.class1[sample(1:nrow(train.class1), size = 9600,
replace = TRUE),]
mydata.b.train <-
rbind(train.class0,
train.class1_2)
# 您的測試樣本
test.class0 <- dat.class0[-rows.train0,]
test.class1 <- dat.class1[-rows.train1,]
mydata.test <- rbind(test.class0, test.class1)
```

請注意，在主數據集中，惡意事件的百分比大約為 4%。這種分佈大致上由不平衡的數據表示。平衡的數據是通過對少數類別進行自助採樣生成的。這裡的想法是確保訓練模型不偏向於多數類別，即正常事件。

第 2 部分 – 比較不同機器學習演算法的表現

a) 隨機選擇兩個監督式學習建模演算法，通過運行以下代碼來進行對比測試。請確保將您的學生證號碼輸入到 commandset.seed(.) 中。您的兩個機器學習方法由 myModels 給出。

```
set.seed(輸入您的學生證號碼)
models.list1 <- c("邏輯岭迴歸",
"邏輯 LASSO 迴歸", "邏輯彈性網絡迴歸")
models.list2 <- c("分類樹",

                  "Bagging Tree"
"Random Forest") myModels <-
c(sample(models.list1, size =
1),
sample(models.list2, size = 1))
%>% myModels %>% data.frame
```

對於您的兩種機器學習建模方法，您將需要：

b) 在 R 中對兩個訓練集運行 ML 演算法，以分類作為結果變數。

c) 執行超參數調整以優化模型：

- 概述您在每個機器學習建模方法中的超參數調整/搜索策略。報告超參數調整的搜索範圍、使用的 k–fold 交叉驗證、重複交叉驗證的次數（如果適用），以及最終的最佳調整參數值和相關的交叉驗證統計數據（即交叉驗證結果、表格和圖表），如適用。如果您使用了重複交叉驗證，則需要至少 2 次重複。

- 如果您選擇的樹模型是 Bagging，您必須調整 nbagg、cp 和 minsplit 超參數，每個至少設定 3 個值。

- 如果您選擇的樹模型是隨機森林，您必須調整 num.trees 和 mtry 超參數，每個至少設定 3 個值。
- 請務必使用您的學生證號設置隨機種子。

d) 評估您從平衡和不平衡訓練集中得出的兩個機器學習模型在測試集上的預測性能。提供混淆矩陣，並在項目背景下報告和解釋以下指標。

- 整體準確度
- 精確度
- Recall
- F1–score

確保您在研究的背景下定義上述每個指標。提示：在 R Studio 的幫助菜單中使用 confusionMatrix(.) 函數，了解如何獲取精確度、召回率和 F1 分數指標。

e) 提供一個關於您最終推薦模型的簡短說明，並解釋您為什麼選擇了它。這包括解釋您在做出這個決定時使用了哪些指標，以及為什麼。如果決策很接近，可以考慮簡潔性，並在較小程度上考慮可解釋性。如果有助於您的論點，您可以在附錄中概述您的懲罰模型估計。

請提交

將你的研究結果整理成一份報告（最多 5 頁），必要時引用相關來源。
呈現數據是如何以及為什麼被「清理和準備」，機器學習模型是如何調整的，並提供相關的 CV結果。最後，呈現它們在不平衡和平衡情況下彼此表現的方式。您可以使用適當的圖表和圖像來幫助讀者理解您的研究結果。所有表格和圖形應適當地加上標題，並在正文中引用。

對於這個任務，最終建議採用哪種機器學習建模方法？

您的最終報告應該看起來專業，包括適當的標題和副標題，應引用事實並按照 APA–7 格式引用來源資料。

您的提交必須包括以下內容：

- 您的報告（不超過 5 頁，不包括封面/目錄/參考文獻/附錄頁）。報告必須通過 TURNITIN 提交並檢查原創性。

- 一份您的 R code 副本，應該通過另一個提交鏈接單獨提交，與報告分開。

確保保留每個訓練集的副本和一個測試集（以 .csv 格式），以防日後需要。

請注意，如果您提供的結果無法通過您的代碼進行確認，將不會給予任何分數。此外，超過5頁限制的所有頁面將不會被閱讀或審查。

評分標準

標準	貢獻至 作業分數
在 R 中準確實施數據清理和每個監督式機器學習算法。 • 嚴格關於程式碼（1）程式碼是否從頭到尾都能運作？（2）結果是否可再現？ (3) APA 7 引用風格中的外部程式碼來源（如適用）（4) 所有步驟都執行正確嗎？(5) 是否有良好的文件記錄？ 備註：代碼至少 80%（不包括上面提供給您的部分）必須與單元內容一致。否則，此部分將被判為零分。	20%
數據清理和準備的解釋。 • 對應於第 1 部分 b) • 簡要概述子部分 (i) 的原因。 • 提供合併類別的理由，即子部分(ii)。	10%
選定建模方法的概要、超參數調整和搜索策略、在訓練集中的相應性能評估（即 CV 結果、表格和圖表），以及最佳調整超參數值。 • 懲罰 logistic 迴歸模型 – 概述 lambda 和 alpha 的值範圍（如果是彈性網）。繪製/列出交叉驗證結果。概述最佳值。	20%

<table>
<tr><td>超參數的選擇。如果需要，請概述您的模型選擇參數。

• 樹模型 – 概述超參數的範圍（bagging 和 RF）。製作表格，例如列出頂級組合和最佳的OOB錯誤分類率，或繪製CV結果圖（例如分類樹）。</td><td></td></tr>
<tr><td>呈現、解釋和比較所選機器學習演算法的性能指標（即混淆矩陣、準確率、精確率、召回率和F1分數）。對推薦的建模方法進行說明。

• 提供測試集中的混淆矩陣（頻率、比例）。總共應該有 4 個。

<table><tr><td></td><td colspan="2">Actual</td></tr><tr><td>預測</td><td>Yes</td><td>No</td></tr><tr><td>Yes</td><td>Freq1（靈敏度 %）</td><td>Freq2（誤報率 %）</td></tr><tr><td>No</td><td>Freq3 (False 負面 %)</td><td>Freq4 (特異性 %)</td></tr></table>
• 概述並解釋該研究中的指標（包括準確度、精確度、召回率和F1分數）。
• 解釋您使用了哪些指標來幫助您決定最佳模型，以及原因。</td><td>30%</td></tr>
<tr><td>報告結構和呈現（包括表格和圖表，適當時應採用 APA7 格式進行引文和參考文獻引用）。報告應清晰、邏輯嚴謹，結構良好，大部分免於溝通、拼寫和語法錯誤。

• 整體結構、呈現方式和敘述。
• 參考資料
• 表格和圖表清晰，並且在文中有適當的標籤和引用。

• 不要截取 R 輸出的螢幕截圖，除非是圖表。
• 拼寫和語法。</td><td>20%</td></tr>
</table>

學術不端

伊迪斯考恩大學視任何形式的學術不端行為為不可接受。學術不端行為包括但不限於抄襲、未經授權的合作、考試作弊、盜用他人作品、串通行為。

包括但不限於不足和不正確的引用；將根據 ECU 規則 40 學術不端行為（包括抄襲）政策處理。請確保您熟悉學術不端行為規則。

作業延期

在 ECU 在線延期申請和跟蹤系統上提供了申請延期的指引，以正式提交您的作業延期請求。該鏈接也可在 Canvas 的作業部分找到。

正常的工作承諾、家庭承諾和課外活動並不被視為延長時間的理由，因為預期您應該提前計劃好您的評估截止日期。

在作業遲交不超過 7 天的情況下，每遲交一天的懲罰將為作業可獲得的最高評分的 5%。若作業遲交超過 7 天，將被給予零分。