**作業6: 眾裡尋他千百度,那模型就在燈火欄珊處**

**現在我們就要開始來使用江湖中傳說的文字探勘絕學啦!**

**比如：CHATGPT**

**為了確保大家的學習品質,因此課程中一半的內容將做為作業的一部份;另一部份才是讓大家帶回去練習的作業唷!那我們就開始吧!!**

**1. 請問啟spam.xls檔案! 並開始進行分析,開始前請針對需要的欄位進行必要之前置處理. (10%)**

* 流程圖：

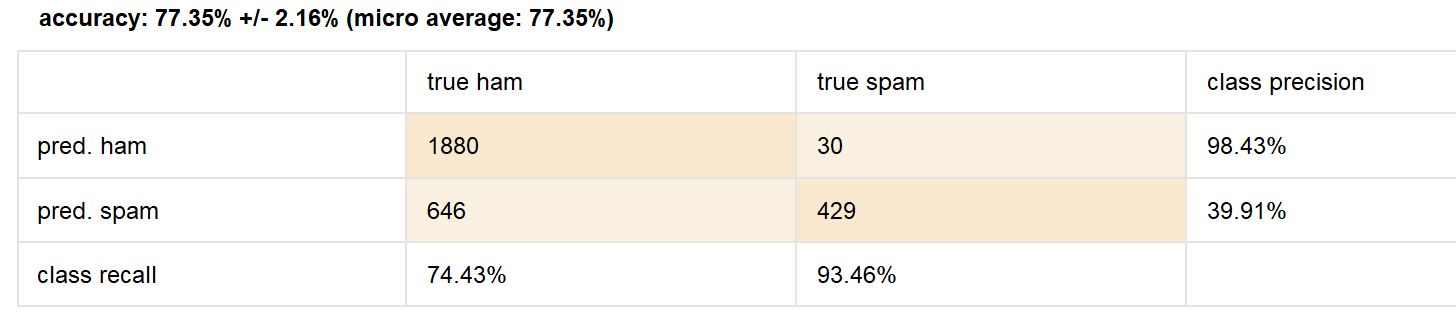
**A picture containing text, diagram, plan

Description automatically generated**

* 必要之前處理：
* Read CSV：讀取資料，刪除att 欄位，att對資料分析沒有意義
* Filter Examples: 刪除v1是ham”””是資料
* Set Role: 給 v1 貼標籤
* Nominal to Text：將資料轉換為文本
* Split Data: 將資料切割成測試集和驗證集

**2. 接下來,我們要針對文字進行分析,因此有一些必要的處理程序,請說明一下您的步驟(10%),並請說明下列分析結果的意涵: TF-IDF (5%)**

* 必要的處理程序
* Process Documents from Data: 對測試集進行處理文本文檔
  + Tokenize: 將文本拆分成單詞
  + Transform Cases: 將文本轉換成不同的大小寫形式
  + Filter Stopwords: 用於過濾停用詞
* Cross Validation（交叉驗證）：使用Naive Bayes模型並看模型效率
* Process Documents from Data 2: 對驗證集進行處理文本文檔，類似Process Documents from Data，需要將測試集的wor給驗證集
* Apply Model 2: 將驗證集套用模型
* Performance 2：衡量驗證集的正確率評
* 測試集的準確率達到77.35%



* 驗證集的準確率達到74.66%

**A screenshot of a computer

Description automatically generated with low confidence**

**3. 接下來,我們要開發製做一個垃圾郵件過濾器,請找出您能找出正確率最高的分類機制,並略加說明(15%)**

* 將模型改成SVM，準確率提高，精確率和召回率都70%以上：

**A screenshot of a computer

Description automatically generated with medium confidence**

* 測試集的準確率達到96.35%：

**A picture containing text, screenshot, font, number

Description automatically generated**

* 驗證集的準確率達到96.09%：

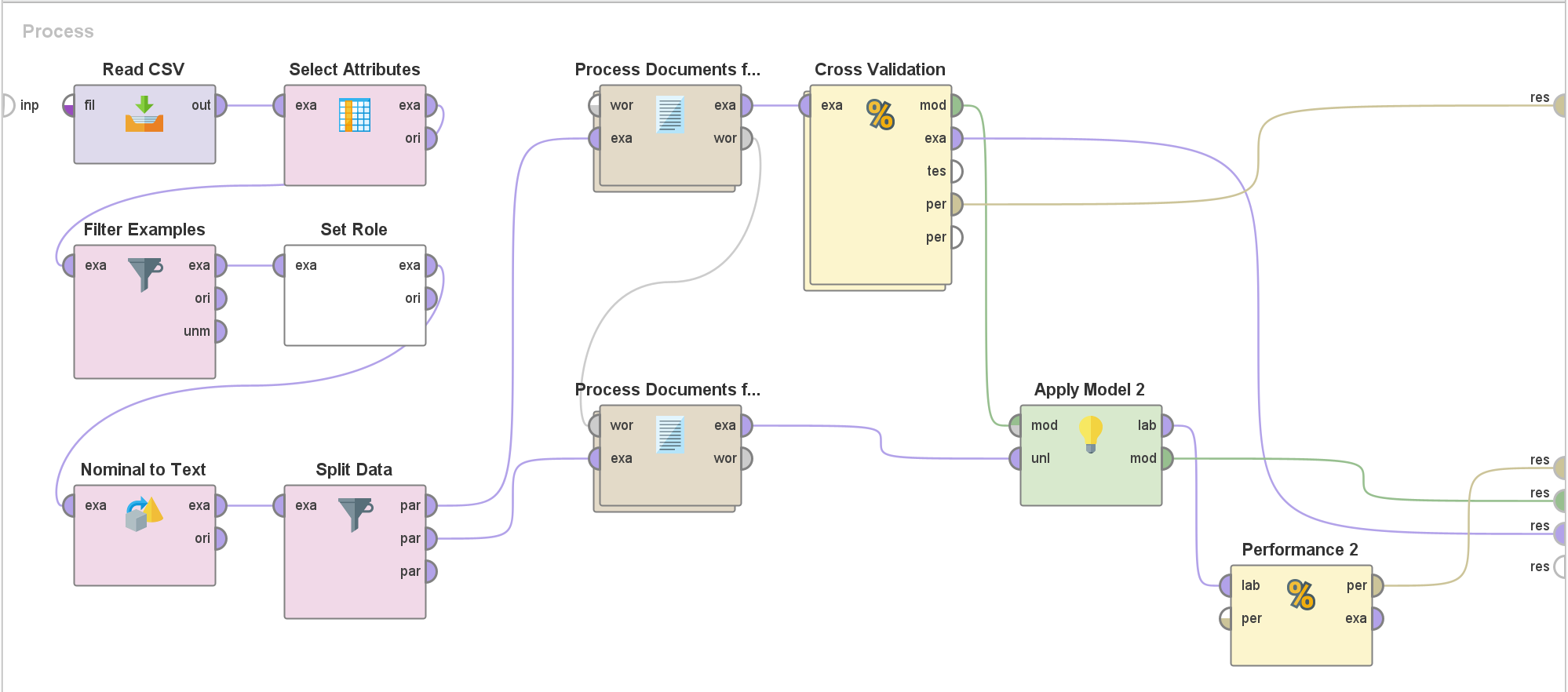
**A picture containing text, screenshot, font, number

Description automatically generated**

**\*註:因為有正確率評估的關係,請記得要使用訓練/測試集,又或是Cross-Validation**

**4. 接下來,承上述步驟,請使用Womens Clothing E-Commerce Reviews.csv檔案進行分析,在這邊我們希望您可以借由分析評論來預測評分或是您認為值得分析的議題. 另外也請您自行思考一下還有什麼分析可以進行並說明結果唷 (60%)。**

* **流程圖：**

****

* **必要之前處理：**
* 在讀取資料之前，先使用 clean() 函數刪除資料中不必要的換行符號
* Read CSV：進行資料讀取，將 Recommended IND 類別改為二元
* Select Attributes: 只留下Clothing ID、Recommended IND、Review Text 欄位
* Filter Examples: 刪除缺漏值
* Set Role: 給 Recommended IND貼標籤
* Nominal to Text：將資料轉換為文本
* Split Data: 將資料切割成測試集和驗證集
* **必要的處理程序：**
* Process Documents from Data: 對測試集進行處理文本文檔
  + Tokenize: 將文本拆分成單詞
  + Transform Cases: 將文本轉換成不同的大小寫形式
  + Filter Stopwords: 用於過濾停用詞
* Cross Validation（交叉驗證）：使用SVM模型並看模型效率
* Process Documents from Data 2: 對驗證集進行處理文本文檔，類似Process Documents from Data，需要將測試集的wor給驗證集
* Apply Model 2: 將驗證集套用模型
* 測試集的準確率達到 86.16%，精確率和召回率都60%以上：

A screenshot of a graph

Description automatically generated with low confidence

* 驗證集的準確率達到 85.92%：

**A picture containing text, screenshot, font, number

Description automatically generated**

* **可進行分析的議題：**
* 產品評分分析：通過分析Rating欄位，可以了解產品的評分分佈情況，分類受歡迎和不受歡迎的產品。
* 產品推薦率分析：分析Recommended IND欄位，可以計算產品的推薦率，了解哪些產品更受客戶推薦，哪些產品不太受歡迎。
* 年齡段偏好分析：通過分析Age欄位，可以了解不同年齡段客戶對產品的評價和偏好，可能發現特定年齡段的消費者更傾向於購買某些產品。
* 用戶反饋分析：通過分析Positive Feedback Count欄位，可以了解哪些評價受到其他客戶的積極反饋，識別出受歡迎的產品和具有高度影響力的評價。