**CA1 DATA SCIENCE AND**

**MACHINE LEARNING PORTFOLIO LOG**

**Date: February 28th 2025**

**Prepared by:**

Kirubel Temesgen

Student Number: C00260396

**Course:** Data Science and Machine Learning

**Supervisor:** Greg Doyle

## K-Means Clustering: CRISP-DM Process

## Phase 1: Business Understanding

**Objective**

The goal of this project is to use K-Means Clustering to group emails into meaningful clusters without predefined labels, in order to analyse patterns in potential spam emails and separate legitimate communication.

Why K-Means?

* K-Means was chosen to discover natural patterns in email content.
* Helps determine whether spam-like emails form distinct clusters.
* Allows us to analyse internal vs. external emails without explicit labels.
  + - However, K-Means does not inherently classify spam, so a binary classification rule was added to explicitly label emails as spam (1) or not spam (0).

## Phase 2: Data Understanding

**Dataset: Enron Email Dataset**

* **Total Emails:** ~500,000 (Used 10% sample for efficiency).
* **Key Fields:** From, Subject, Body, To.

**Initial Observations**

| **Observation** | **Impact on Clustering** |
| --- | --- |
| Most emails are **internal (enron.com)** | Harder to detect external spam since most emails are company-related. |
| Frequent words include **business-related terms** (e.g., "meeting", "update", "agreement") | Business reports were mistakenly grouped with spam. |
| Spam-like words (e.g., "win", "free", "offer") **appear but are less frequent** | Needed to expand the spam-word list. |

**Adjustments Based on Data Understanding**

| **Issue** | **Adjustment Made** | **Outcome** |
| --- | --- | --- |
| Business reports may be **grouped as spam** | **Filtered financial terms** before clustering | **More accurate spam clustering** |
| High overlap between **internal & external emails** | Added "**Is\_External\_Sender**" feature | **Better differentiation** between internal & external emails |
| Common spam words appeared in **legitimate promotional emails** | Expanded spam keyword list | Better classification of actual spam emails |

## Phase 3: Data Preparation

**Preprocessing Steps**

Text Cleaning: Lowercased, removed special characters, stopwords, numbers.  
 Feature Engineering:

* TF-IDF Vectorization (Bigrams & Unigrams).
* Spam Word Count Feature (Counts occurrences of known spam-like words).
* External Email Indicator (Flags if the sender is from outside Enron).
* Binary Spam Label: Emails classified as Spam (1) or Not Spam (0) based on the number of spam words and whether the sender is external.

**Adjustments & Improvements**

| **Adjustment** | **Reason** | **Outcome** |
| --- | --- | --- |
| Added **Spam\_Words\_Subject & Spam\_Words\_Body** features | Improve separation between **legitimate vs. spam emails** | **Higher cluster coherence** |
| Adjusted **K value from 2 to 3** | Better representation of **internal vs. external emails** | **More meaningful segmentation** |
| Expanded **Spam Keyword List** | Some spam words were missing (e.g., "investment", "guarantee", "unsubscribe") | **Better spam detection** |

**Spam Classification Rules**

* An email is classified as spam (1) if:
  + Spam\_Words\_Body > 2 OR Spam\_Words\_Subject > 2
  + AND Is\_External\_Sender == 1 (Only external senders are considered spam)
* Otherwise, it is classified as Not Spam (0).

**Adjustments & Improvements**

| **Adjustment** | **Reason** | **Outcome** |
| --- | --- | --- |
| Implemented **binary spam classification** | CRISP-DM requires clear labels | Clear distinction between spam (1) and non-spam (0) |
| Adjusted **spam word threshold to 2 instead of 3** | Captures more potential spam emails | Increased recall of spam detection |
| **Excluded internal Enron emails** from spam detection | Internal emails are mostly legitimate | Reduced false positives |
| Introduced **spam word filtering for newsletters & reports** | Financial reports were being classified as spam | Improved classification accuracy |

## Phase 4: Modeling (K-Means Clustering)

**Choosing K (Number of Clusters)**

* **Elbow Method:** Suggested **K=3** (**Optimal balance in WCSS**).
* **Silhouette Score:** Highest at **K=3** (**Indicating well-separated clusters**).

These methods validated that clustering produced meaningful results.

**Final Decision:** Used **K=3** based on both methods.

**Model Training**

* Algorithm: K-Means
* Binary Classification: Added Spam\_Label (0 = Not Spam, 1 = Spam).
* Features Used:
  + TF-IDF (Processed\_Subject & Processed\_Body)
  + Spam Word Count (in both Subject & Body)
  + External Sender Flag
  + Binary Classification
* Random State: 42 (to ensure reproducibility).

## Phase 5: Evaluation & Adjustments

**Final Cluster Analysis**

| **Cluster** | **Total Emails** | **Avg. Spam Words (Subject)** | **Avg. Spam Words (Body)** | **% External Emails** |
| --- | --- | --- | --- | --- |
| **0** | 34,457 | 0.00 | 0.00 | 12.30% |
| **1** | 1,956 | 0.11 | 3.65 | 60.02% |
| **2** | 15,327 | 0.04 | 1.23 | 24.17% |

**Observations & Adjustments**

| **Issue** | **Adjustment** | **Outcome** |
| --- | --- | --- |
| Cluster 0 had mostly **internal, valid emails** | No change needed | N/A |
| Cluster 1 had **highest % of external emails and spam words** | Confirmed it correctly captures spam-like emails | Improved Spam Detection |
| Cluster 2 contained **business reports mistakenly clustered** | **Removed finance-related terms** | Clearer separation of financial emails |

**Final Spam Classification Results**

* **Spam (1): 2,356 emails (4.5% of dataset)**
* **Not Spam (0): 49,384 emails (95.5% of dataset)**

| **Metric** | **Value** |
| --- | --- |
| **Total Emails** | **51,740** |
| **Spam Emails (1)** | **2,356 (4.5%)** |
| **Non-Spam Emails (0)** | **49,384 (95.5%)** |

**Sample Classification Results**

**Spam Emails (Example)**

| **Processed\_Subject** | **Spam Words in Subject** | **Spam Words in Body** | **External Sender?** | **Spam Label** |
| --- | --- | --- | --- | --- |
| "Save up to 50% off original prices" | 0 | 3 | Yes | 1 (Spam) |
| "Limited Time Offer from Morgan Stanley" | 1 | 4 | Yes | 1 (Spam) |
| "Your football knowledge could win you a million" | 1 | 3 | Yes | 1 (Spam) |

**Non-Spam Emails (Example)**

| **Processed\_Subject** | **Spam Words in Subject** | **Spam Words in Body** | **External Sender?** | **Spam Label** |
| --- | --- | --- | --- | --- |
| "Enron HPL Actuals for September" | 0 | 0 | No | 0 (Not Spam) |
| "Summary of Commission Meeting" | 0 | 1 | No | 0 (Not Spam) |
| "Market Energy Update" | 0 | 1 | Yes | 0 (Not Spam - Likely Report) |

## Phase 6: Deployment & Future Work

**Key Takeaways**

* Successfully implemented binary spam classification.
* Spam emails are primarily from external senders.
* Business reports were effectively filtered out.
* Spam classification now meets project specification.

**Future Enhancements**

* Further refine spam keyword detection (expand finance/business exclusion list).
* Experiment with threshold values.
* Test Cosine Similarity instead of Euclidean Distance for clustering.

## References

Comment *et al.* (2025) *K means clustering - introduction*, *GeeksforGeeks*. Available at: <https://www.geeksforgeeks.org/k-means-clustering-introduction/> (Accessed: 08 March 2025).

*W3schools.com* (no date) *Python*. Available at: <https://www.w3schools.com/python/python_ml_k-means.asp> (Accessed: 14 March 2025).

*Clustering text documents using K-means* (no date) *scikit*. Available at: <https://scikit-learn.org/stable/auto_examples/text/plot_document_clustering.html> (Accessed: 21 March 2025).

Jakevdp (no date) *Pythondatasciencehandbook/notebooks/05.11-K-means.ipynb at master · JAKEVDP/Pythondatasciencehandbook*, *GitHub*. Available at: <https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.11-K-Means.ipynb> (Accessed: 18 March 2025).

*K-means clustering algorithm in machine learning* (no date) *Tutorialspoint*. Available at: <https://www.tutorialspoint.com/machine_learning/machine_learning_k_means_clustering.html> (Accessed: 18 March 2025).