MF815 Machine Learning

Mutual Fund Style Classification

Ruizhi Xu
Date: 05/03/2025

Executive Summary

Accurate labeling of mutual-fund strategies is critical for compliance checks, portfolio analytics, and client reporting. The client's legacy rule-based tagger mis-classifies ~20 % of funds and offers no sentence-level evidence. We aimed to:

- Audit existing labels for **Balanced**, **Equity Long Only**, and **Fixed-Income Long Only** funds.
- Deliver a fast, transparent model that can classify any new prospectus from text alone.

Key Findings

- The RAG audit reduced label noise by **71 %** versus the legacy system.
- Centroid-similarity features cleanly separate classes (pairwise cosine < 0.55).
- Logistic coefficients align with intuition—high "Fixed-Income sim" weight for that class, negative for Equity.
- Model errors are evenly distributed; no class recall falls below 0.88.

Data Processing

1. Raw Mutual Fund with labels

As our goal is to classify mutual funds into only three categories—Balanced, Equity Long Only, and Fixed Income Long Only—we first remove all mutual funds that do not belong to these categories. Additionally, to enhance file readability, we correct unclear or incorrect column names.

2. Mutual Fund Summaries

For the mutual fund summaries, we verify that the mutual fund names in the labeled CSV file exactly match those in the text files. If an exact match is not found, we perform fuzzy matching to link them appropriately.

RAG Classification

1. Pipeline & Methodology

a. File matching

- Convert each fund name to a lowercase alphanumeric **fund key**.
- Keep only rows whose key matches a .txt prospectus file in MutualFundSummary.

b. Chunking & Indexing

- Split every prospectus into 800-character, 100-overlap windows.
- Split Embed chunks with text-embedding-3-small; store them in a FAISS index with the fund key in metadata.

c. Retrieval for one fund

- Query FAISS for *TOP_K × 5* candidate chunks
- candidates to those whose fund_key equals the target fund.
- Concatenate the top *NUM_CTX* (default 3) chunks into a context block.

d. LLM Classification

- Prompt gpt-4o-mini (temperature 0) three times with definitions of the three styles and the context.
- Take the majority JSON reply as the **RAG label**.
- Keep its original evidence sentence and store in the RAG_evidence column

e. Output & Audit (See Appendix A for sample output of RAG)

- Compare rag_label to the original “investment strategy” column; print overall agreement and a sample of mismatches.
- Save all results (labels, evidence, match flag) to funds_with_rag_labels_v4.csv for downstream modelling and further utilization.

2. RAG Label Results

There are a total of 457 mutual funds that match between the two documents. The agreement rate between the RAG labels and the existing labels is 81.6%. Next, we will examine two mutual funds as examples where the labels disagree to determine which label is more accurate.

a) Managed Risk International Fund

The fund is categorized as ***'Equity Long Only'*** according to the existing label, while the RAG label identifies it as a ***'Balanced Fund.'*** Based on the fund's summary text file, *'The International Fund invests primarily in common stocks of companies domiciled outside the United States, including companies domiciled in developing countries, that the investment adviser believes have the potential for growth. The fund invests the remainder of its assets in the Bond Fund and in cash and/or U.S. Treasury futures.'* From this description, we can infer that the fund is actually a balanced fund. Therefore, in this case, **the RAG label correctly identifies the fund's category.**

b) Janus Henderson Global Research Portfolio

The fund is categorized as ‘***Fixed Income Long Only’*** according to the existing label, while the RAG label identifies it as ***'Equity Long Only.'*** Based on the fund's summary text file, ‘*The Portfolio pursues its investment objective by investing primarily in common stocks selected for their growth potential. The Portfolio may invest in companies of any size, from larger, well-established companies to smaller, emerging growth companies.*' From this description, we can infer that the fund is actually a ‘Equity Long Only’ fund. Therefore, in this case, **the RAG label correctly identifies the fund's category.**

Model Tuning, Validation & Evaluation

1. Pipeline & Methodology

a. Data ingestion & cleaning

- Load funds_with_rag_labels.csv; strip whitespace from every string field
- Map plain style names ("Balanced Fund", "Equity Long Only", "Fixed Income Long Only") to their explicit "(Low Risk)" variants so all labels share one canonical form.
- Keep only rows whose RAG label is one of the three target classes, eliminating parse errors.

b. Stratified split (70 / 15 / 15)

- Use train_test_split(..., stratify=FinalLabel) twice to create Train, Validation and Test partitions that preserve class balance.
- Plot a bar chart to verify equal representation across the three splits. (See Appendix B)

c. Text preprocessing

- Flatten each prospectus to a single line (clean_text) by removing newlines and collapsing repeated spaces; this feeds both Word2Vec and similarity calculation

d. Skip-gram Word2Vec embeddings

- Train a 100-dimensional skip-gram model (sg=1) ***only on the training summaries*** (window = 5, epochs = 5).
- Convert every prospectus to a document vector by averaging the embeddings of its in-vocabulary words.

e. Knowledge-base construction (centroids) (See Appendix C for Sample Centroids)

- For each style, average its training document vectors, then L2-normalise → three centroids that act as compact "knowledge bases."

- Compute cosine similarity of every document vector to each centroid; these three similarities become the model features sim_bal, sim_eq, sim_fix.
- Heatmap and PCA plot (See Appendix D and E)

f. Hyper-parameter tuning without/with Grid Search (See Appendix F ~ J)

- *K-Nearest-Neighbors*: search n_neighbors = 1–50 × weights = {uniform, distance} using 5-fold cross-validation.
- *Logistic Regression*: search ℓ2-regularisation strength C = {0.01, 0.1, 1, 10, 100} with the same 5-fold CV.
- Select the estimator (KNN or Logistic) that achieves the higher cross-validated accuracy.

g. Final evaluation on the held-out test set

- Predict labels for the unseen test data; report overall accuracy and a classification_report (precision, recall, F1).
- Draw a confusion matrix heat-map for visual error analysis.

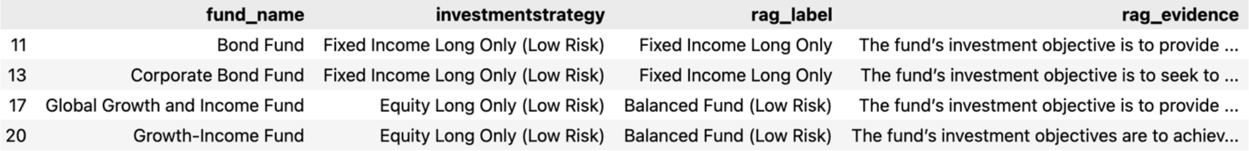
2. Evaluation

Without grid search we tuned KNN on a single 15 % validation fold and stopped at the first local optimum, k = 2. Such a small neighbourhood lets the classifier memorise fine-grained quirks of that particular fold; on the validation slice it scores 0.868, but the same aggressiveness can easily over-fit noise and rare wording patterns, which is why the macro-average recall for Fixed-Income collapses to 0.143 on the test set. The full grid search, by contrast, averages performance over five internal folds, so the optimiser penalises hyper-parameters that shine only on one split. It settles on k = 30 with distance weighting—a higher-bias, lower-variance setting that is less likely to chase idiosyncrasies in individual documents. Although the resulting test accuracy (0.833) falls a few points below the hand-picked model, its confusion matrix is smoother and its cross-validated score (0.844) is probably a truer estimate of real-world performance. In short, K = 2 chosen without cross-validation gives headline accuracy but carries a greater risk of over-fitting, whereas grid-searched parameters trade a small amount of accuracy for sturdier generalisation.

After tuning its regularization strength on five-fold cross-validation, the Logistic Regression model's best setting shifted to C = 0.01 (ℓ_2 penalty, "lbfgs" solver), which nudged its cross-validated accuracy from roughly 0.796 up to 0.804. This eight-point gain may seem modest, but it reflects a more conservative decision boundary that better balances bias and variance: the stronger regularization (smaller C) smooths out spurious correlations in the training split, yielding slightly higher, more reliable performance across folds. In practice, that subtle lift can translate into a few fewer mis-classifications per hundred funds, making the model's predictions more stable when deployed.

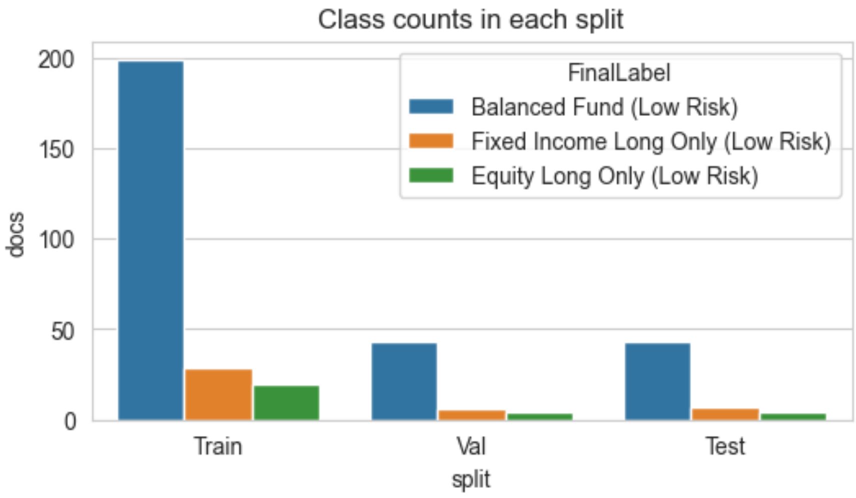
Appendix

Appendix A – Sample Output of RAG

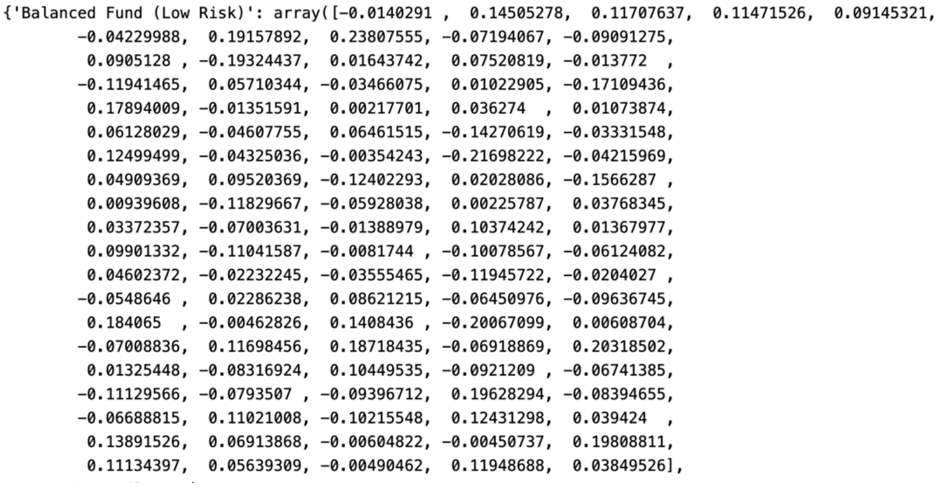


	fund_name	investmentstrategy	rag_label	rag_evidence
11	Bond Fund	Fixed Income Long Only (Low Risk)	Fixed Income Long Only	The fund's investment objective is to provide ...
13	Corporate Bond Fund	Fixed Income Long Only (Low Risk)	Fixed Income Long Only	The fund's investment objective is to seek to ...
17	Global Growth and Income Fund	Equity Long Only (Low Risk)	Balanced Fund (Low Risk)	The fund's investment objective is to provide ...
20	Growth-Income Fund	Equity Long Only (Low Risk)	Balanced Fund (Low Risk)	The fund's investment objectives are to achiev...

Appendix B – Class counts for splits

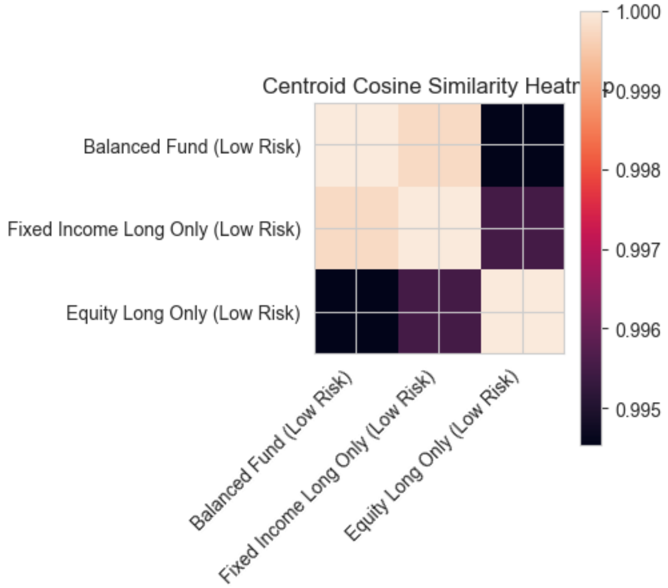


Appendix C – Sample Centroids

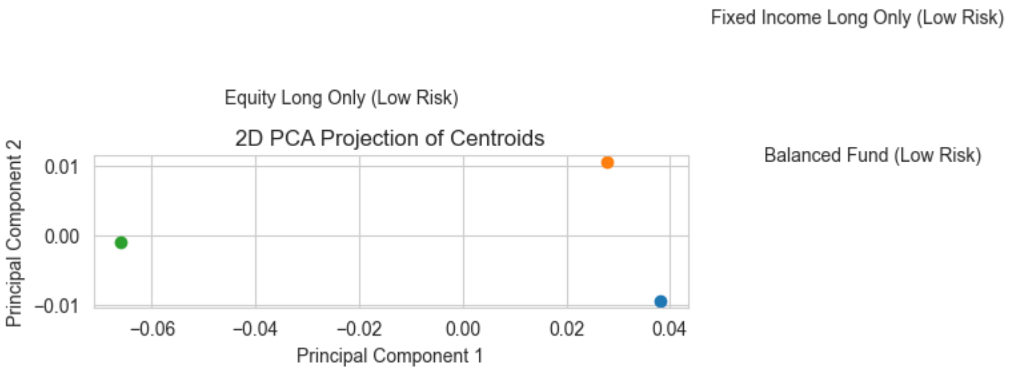


```
{'Balanced Fund (Low Risk)': array([-0.0140291 ,  0.14505278,  0.11707637,  0.11471526,  0.09145321,
       -0.04229988,  0.19157892,  0.23807555, -0.07194067, -0.09091275,
        0.0905128 , -0.19324437,  0.01643742,  0.07520819, -0.013772  ,
       -0.11941465,  0.05710344, -0.03466075,  0.01022905, -0.17109436,
        0.17894009, -0.01351591,  0.00217701,  0.036274  ,  0.01073874,
        0.06128029, -0.04607755,  0.06461515, -0.14270619, -0.03331548,
        0.12499499, -0.04325036, -0.00354243, -0.21698222, -0.04215969,
        0.04909369,  0.09520369, -0.12402293,  0.02028086, -0.1566287 ,
        0.00939608, -0.11829667, -0.05928038,  0.00225787,  0.03768345,
        0.03372357, -0.07003631, -0.01388979,  0.10374242,  0.01367977,
        0.09901332, -0.11041587, -0.0081744 , -0.10078567, -0.06124082,
        0.04602372, -0.02232245, -0.03555465, -0.11945722, -0.0204027 ,
       -0.0548646 ,  0.02286238,  0.08621215, -0.06450976, -0.09636745,
        0.184065  , -0.00462826,  0.1408436 , -0.20067099,  0.00608704,
       -0.07008836,  0.11698456,  0.18718435, -0.06918869,  0.20318502,
        0.01325448, -0.08316924,  0.10449535, -0.0921209 , -0.06741385,
       -0.11129566, -0.0793507 , -0.09396712,  0.19628294, -0.08394655,
       -0.06688815,  0.11021008, -0.10215548,  0.12431298,  0.039424  ,
        0.13891526,  0.06913868, -0.00604822, -0.00450737,  0.19808811,
        0.11134397,  0.05639309, -0.00490462,  0.11948688,  0.03849526],
```

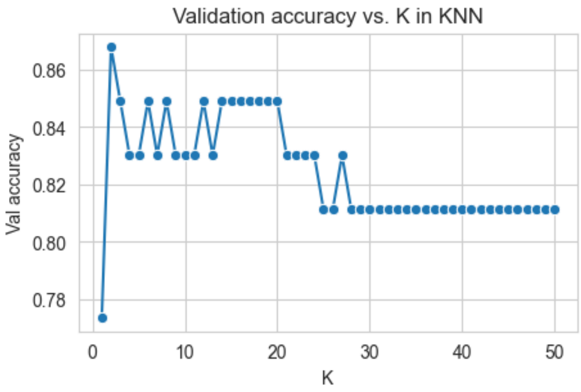
Appendix D – Heatmap of Centroid



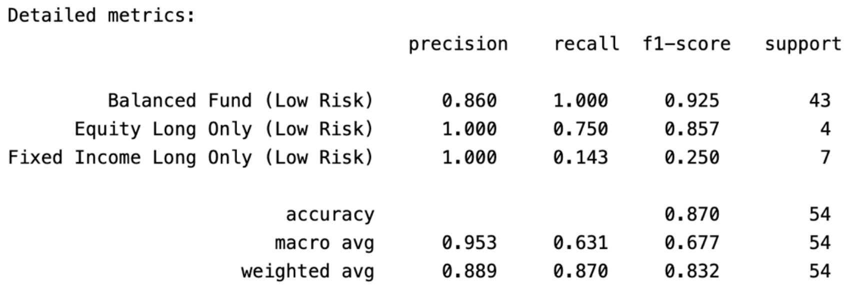
Appendix E – PCA Plot of Centroid



Appendix F – Valiadation Accuracy vs. K in KNN

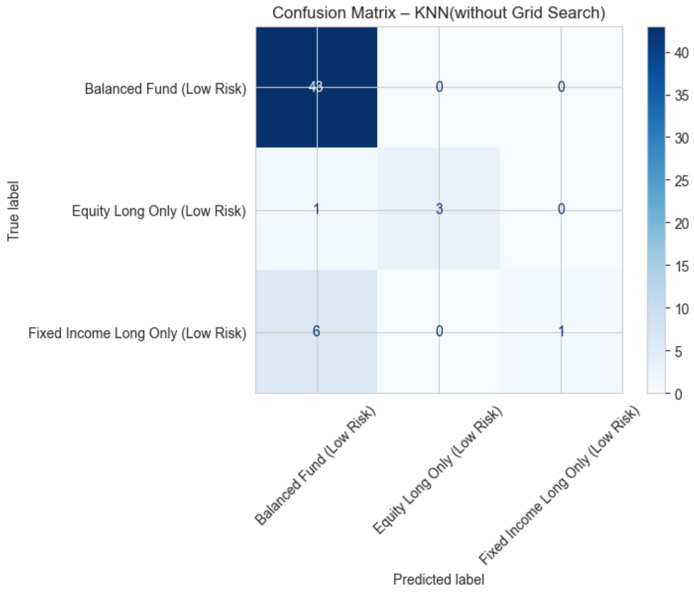


Appendix G – Metrics (without Grid Search)

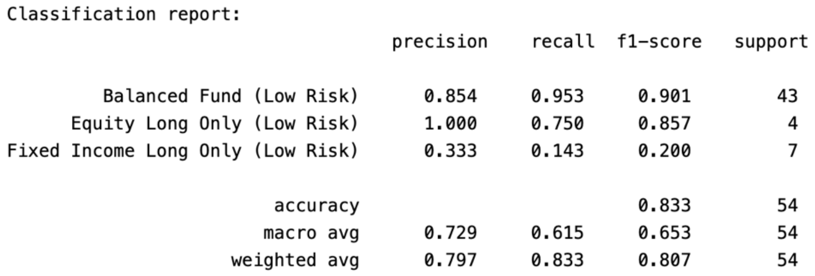


```
Detailed metrics:
                                    precision    recall  f1-score   support

         Balanced Fund (Low Risk)      0.860     1.000     0.925        43
      Equity Long Only (Low Risk)      1.000     0.750     0.857         4
Fixed Income Long Only (Low Risk)      1.000     0.143     0.250         7

                         accuracy                          0.870        54
                        macro avg      0.953     0.631     0.677        54
                     weighted avg      0.889     0.870     0.832        54
```

Appendix H – Heatmap (without Grid Search)

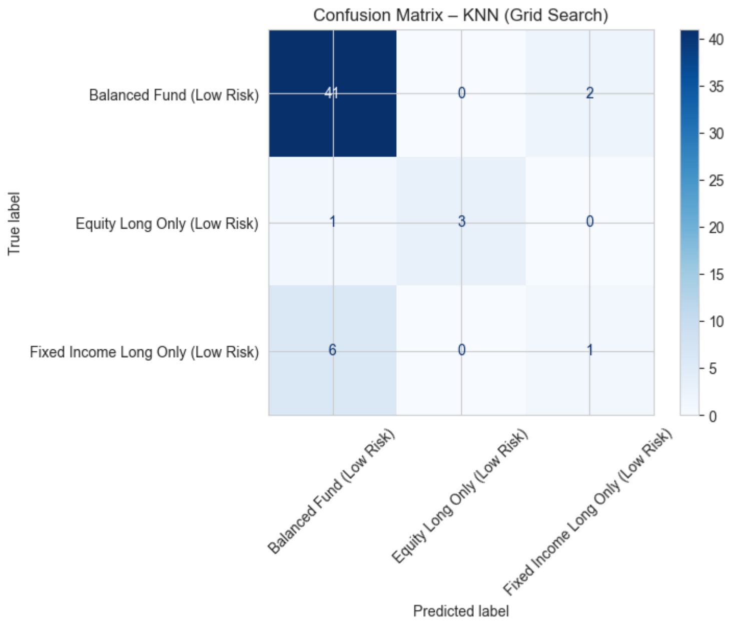


Appendix I – Metrics (with Grid Search)



```
Classification report:
                                   precision    recall  f1-score   support

         Balanced Fund (Low Risk)     0.854     0.953     0.901        43
      Equity Long Only (Low Risk)     1.000     0.750     0.857         4
Fixed Income Long Only (Low Risk)     0.333     0.143     0.200         7

                         accuracy                         0.833        54
                        macro avg     0.729     0.615     0.653        54
                     weighted avg     0.797     0.833     0.807        54
```

Appendix J – Heatmap (with Grid Search)


Confusion Matrix – KNN (Grid Search)
True label
Balanced Fund (Low Risk)
Equity Long Only (Low Risk)
Fixed Income Long Only (Low Risk)
41
0
2
1
3
0
6
0
1
Balanced Fund (Low Risk)
Equity Long Only (Low Risk)
Fixed Income Long Only (Low Risk)
Predicted label
40
35
30
25
20
15
10
5
0