# Summary of the datasets:

|  |  |  |
| --- | --- | --- |
| **File Name** | **Number of Rows** | **Number of Columns** |
| US\_input\_data\_20240903\_With\_Headings.csv | 15284 | 42 |
| EMEA\_input\_sample\_data\_20240903\_WithHeadings.csv | 76668 | 56 |
| EMEA\_additional\_sample\_data\_With\_headings.csv | 64343 | 37 |
| EMEA\_output\_sample\_dat\_03\_09\_With\_Headings.csv | 34896 | 11 |

# Current findings

**Data Set EMEA**

|  |  |
| --- | --- |
| **CLIENT\_ASSET\_CLASS\_NAME** | **Count** |
| Equities | 44806 |
| Fixed Income | 21136 |
| Hedge Funds | 5302 |
| Mutual Funds | 2893 |
| FX, PM, Other Commodities | 1193 |
| Money Market Funds | 784 |
| Funds - Other | 442 |
| Alternatives | 11 |
| Warrants | 1 |

**Data set US**

|  |  |  |
| --- | --- | --- |
| **ASSET\_CLASS\_LEVEL1** | **Count** | |
| Na | | 10967 |
| Fixed Income | | 2252 |
| Equities | | 1235 |
| Funds | | 634 |
| Options | | 86 |
| Offshore Managed Funds | | 37 |
| Currency | | 21 |
| Cash And Cash Equivalents | | 17 |
| Alternative Investments | | 12 |
| Fx | | 10 |
| Others | | 8 |
| Swp | | 4 |

# Key takeaways from meeting

If no market instrument reference like ISIN,CUSIP,SEDOL etc is missing, the asset it is treated as Non advisory and these are internal products.

Internal products are not tagged by Blackrock, SMART tags these assets based done Asset Classification Product Guidelines

Blackrock rules overrides SMART classification rules

SMART output data is the most comprehensive data set that includes Balckrock and user overrides.

# Documentation to self

**Drop Columns** with more than 90% missing values to maintain relevant features.

**Categorical columns:** Missing values were filled with "Unknown".

**Numerical columns:** Missing values were filled with the median. Model Approach

**Used:**

**Initial Asset Class Tagging**:

We used a simple rule-based tagging model based on keywords extracted from relevant columns (e.g., blackrock\_asset\_class\_level\_1, blackrock\_asset\_class\_level\_2, etc.) in the training dataset. This mapped detailed classifications to broader categories such as "Hedge Fund", "Alternatives", "Equities", and "Fixed Income."

**Refined Tagging**:

To improve tagging for the US input dataset, we expanded the classification to consider more columns (e.g., CLIENT\_ASSET\_CLASS, CLIENT\_PRODUCT\_TYPE, HEDGE\_FUND\_STRATEGY, and others).

We applied a **rule-based mapping** using keyword searches within these columns to assign a more specific tag (e.g., entries with "Swp" were tagged as "Swaps", "Options" were correctly identified, etc.).

**Handling "Other" Cases**:

Entries with incomplete or ambiguous classifications (like "Unknown" or missing data) were tagged as "Other."

**Potential Next Steps:**

**Further Refinement**: The tagging process could be enhanced using more sophisticated techniques, such as Natural Language Processing (NLP) to handle edge cases or applying machine learning models to predict asset classes based on available features.

**Imputation**: Missing values in key columns could be imputed using predictive models, which might improve the classification further.

# Decision Tree Workflow and results

Patterns and Insights

1. **Product Types and Descriptive Terms:**

* **Structured Products** often include words like **"phoenix"**, **"dual currency"**, and terms like **"callable"** or **"linked"**, indicating complex financial instruments.
* **Single Line Fixed Income** products commonly include terms such as **"bonds"**, **"term"**, and specific coupon rates (e.g., **"5%"**, **"3%"**), pointing to predictable income structures.

1. **Issuer and Product Correlation:**

* Certain issuers, such as **HSBC** or **BlackRock**, are more frequently associated with structured or alternative products. This might suggest a pattern where the issuer is tied to the type of financial product.
* **"HSBC"** often correlates with **Structured Products**, while **"BlackRock"** is more associated with **Alternatives** or **Hedge Funds**.

1. **Country/Region-Specific Patterns:**

* **Latin America** and **Emerging Markets** appear frequently in conjunction with fixed-income products, especially **emerging market debt**. This suggests that the region plays a role in identifying product types.
* The region **Others** (often meaning a non-specific or global market) appears more in hedge funds and alternative investment products.

**4. Numerical Associations:**

* Specific numbers, such as **10**, **15**, **25**, and **75**, are repeated in the dataset, especially linked to bond terms or percentages. These may represent coupon rates, terms of maturity, or thresholds used in structuring financial products.
* For instance, products containing **"25"** or **"75"** may indicate bond coupon rates, maturity periods, or specific pricing strategies, which can be associated with **Single Line Fixed Income** products.

**5. Multi-Level Asset Classes:**

* **Hedge Fund Strategies** and **Private Equity** products tend to have multi-level classifications, where they appear under broad categories such as **Alternatives** but are sub-classified based on specific strategies like **"macro"** or **"multi-strategy"**.
* The repetition of "Others" across categories suggests that some products, though complex, are broadly grouped into a high-level classification due to lack of more detailed tagging.

**6. Year-Based Patterns:**

* Products with years like **2024** in the description often align with maturities and callable features, frequently appearing in **Structured Products**. This pattern reflects a focus on specific future timelines in structured finance.

**7. Common Terms and Product Type Correlation:**

* Words like **"notes"**, **"shs"** (shares), and **"series"** frequently occur in various product types but tend to appear more in **Single Line Fixed Income** and **Structured Products** categories. These terms indicate asset-backed or security-based financial instruments.
* Descriptions containing **"dual"** or **"currency"** are more likely to be linked with **Structured Products**, especially those tied to foreign exchange or precious metals.

**8. Specific Terms for Risk Classification:**

* Terms such as **"callable"**, **"fixed"**, and **"linked"** are frequently associated with products that might have specific risk-return profiles, particularly in structured finance. This could indicate the type of risk the product is managing (interest rate risk, currency risk, etc.).

**9.Currency Terms:**

* The presence of currency abbreviations like **USD**, **GBP**, and **EUR** often indicates specific product types. For example:
  + **"GBP"** is frequently linked with **Structured Products**.
  + **"USD"** and **"EUR"** tend to appear in various structured or fixed-income products.

**10. Fixed Income and Rates:**

* The term **"fixed"** is strongly associated with **Single Line Fixed Income** products, often linked to terms such as **"rate"** and **"term"**.
* Other terms like **"shs" (shares)** and **"notes"** are common in fixed-income product descriptions.

**11. Dates and Series:**

* The presence of numbers and dates, like **"2024"**, and terms like **"series"** and **"global"**, suggest both **Structured Products** and **Single Line Fixed Income** products. These terms often occur in product names or descriptions related to bonds or securities.

**12. Thresholds:**

* Specific numeric thresholds, such as values like **75** or **25**, play a role in distinguishing between different product categories.

**13. Issuer-Specific Patterns:**

* Company names or abbreviations, such as **"HSBC"**, appear to influence the categorization, especially for structured products.

**Summary of Additional Insights:**

* **Currency-linked products** are prevalent in structured finance.
* **Region-specific** products, especially from Latin America and emerging markets, lean toward fixed-income categories.
* **Issuer names** (e.g., HSBC, BlackRock) provide strong hints for the categorization of products.
* **Specific numbers and dates** (e.g., 2024, 75, 25) have critical importance for categorization in terms of bond maturities and fixed-income structures.