

**Financial Intelligence Unit - India (FIU-IND)**

**Approach Note – RE Clustering**

**CONFIDENTIALITY CLAUSE**

This document is the property of Financial Intelligence Unit - India under Project FINnet 2.0. All ideas and information contained within the document is the intellectual property of FIU-IND. These documents are not for general distribution and are meant for use solely by the person/persons to whom it is specifically issued to. Copying or unauthorized distribution of these documents, in any form or means including electronic, mechanical, photocopying or otherwise is illegal.

DOCUMENT CONTROL

|  |  |
| --- | --- |
| Document Version Control Note | |
| Name of the Document | Approach Note - RE Clustering |
| Document Version Number | 1.0 |
| Effective Date | 20-06-2022 |
| Approved By | FIU-IND |
| This Revision Supersedes | NA |
| Document Classification | Restricted |
| Distribution List | FIU-IND, PMU, MSP |
| Access Level | Read Only |

REVISION HISTORY

| Version | Release Date | Prepared By | Reviewed/  Approved By | Description of changes made | Section of document Impacted |
| --- | --- | --- | --- | --- | --- |
| 0.1 | 20-06-2022 | Abhijeet Kaushal | Robin Singh/Kuldeep Rana | Initial Draft |  |
| 1.0 | 22-06-2022 | Kuldeep Rana | Chaitanya Shukla Sir | Review Comments given by Sir (Removal of PTR and CCR Amount and Number in Peer Group Thresholds. Addition of No of Branches) |  |

REFERENCE DOCUMENTS

| S. No. | Document Name | Document Description |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

1. **Methodology of RE Clustering:**

* RE Clustering will be initiated on the bases of RE Type Id i.e. on the basis of category like Banking, Mutual Fund etc.
  + In 1.0, RECATGEORY is represented as a String of 5 Characters. A few example are,
    - BADCB – District Cooperative Banks
    - BAFOR – Foreign Banks
    - INBRO – Share Brokers
    - INTRU – Trustees to Trust Deeds
  + In 2.0, RE Type is a String of 2 Characters. A few example are,
    - BA – Banks
    - BF – Brokerage Firms
    - FI – FI/NBFC/Others
* Each segregated RE will be clustered on the basis of their filed Branch Count, KYC Count, Amount and Report differentiated on the basis of STR,CTR,NTR & CBWTR
* Each RE will be assigned RE\_CLUSTER\_ID on the bases of RE Type and Cluster ID

1. **Assumptions:**

* 'RE Cluster Creation' should run only once in a Year wherein entire Clustering exercise and RE\_CLUSTER\_ID assignment to RE's would be done. This in line with the expectation that frequent Updates to RE\_CLUSTER\_ID of existing RE's would not be done since it will have a cascading effect.
* The above point was discussed on 21/06. Sir wanted the Frequency of Clustering Exercise to be flexible such that it could be changed later and we can lower the Frequency of complete RE Clustering.
* However, for the current Implementation Sir has agreed for Annual run and showcase the results. After seeing results and few dry-runs Sir will confirm for the Frequency.
* 'RE\_CLUSTER\_ID Assignment' would be run every month and only new REs (REs which have not been assigned any RE\_CLUSTER\_ID) would be considered and a RE\_CLUSTER\_ID will be assigned.
* No of Cluster Groups would be based on ML Technique (of K Means).

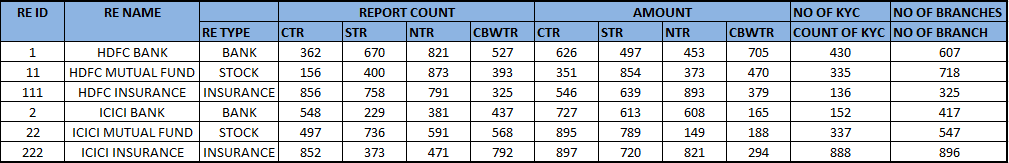
1. **Implementation Steps:**

| STEPS | Description |
| --- | --- |
|  |  |
| Step 1 | Extracting the data of the sum of the report count, amount, branch count and KYC filled by RE during some specific time frame on the basis of the Report Type. |
| Step 2 | Clustering will be initiated on the final data where optimal number of clusters will be chosen on the basis of AI/ML and below Column Values   * RE TYPE * KYC COUNT * BRANCH COUNT * STR\_REPORT\_COUNT * STR\_AMOUNT * CTR\_REPORT\_COUNT * CTR\_AMOUNT * NTR\_REPORT\_COUNT * NTR\_AMOUNT * CBWTR\_REPORT\_COUNT * CBWTR\_AMOUNT |

Table Structure for the RE Clustering is embedded along with:

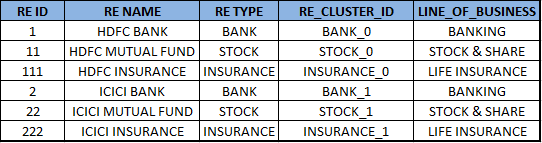
****

**3.1 Extracted Data**



* 1. **Clustering**
* K means (unsupervised machine learning algorithms) clustering will be used on the output of the data. K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids and then perform clustering.
* For each Retype id, there will be multiple clusters, the optimal numbers of cluster of each category will be chosen on the bases silhouette score. (Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1)

Output of sample clustering will be as follows:



**One RE belonging to multiple peer groups:**

**Pro:**

1. Same RE under different category will have different behaviour and attributes.

For example, HDFC is one Reporting Entity but different RE Types like Banks, Mutual Funds, Insurance will have different RE\_CLUSTER\_ID so that it justifies and the characteristics of RE and its Type.

2. RE ID along with RE Type will be the bases of Clustering.

3- One RE should not be clubbed as Single Entity as average outcome can be different as there may be extreme behaviour on their different category