1. Technical Specification Double-side Page

2. **TECHNICAL SCOPE:** Summarize the mock-up devised during the EXPLORE phase: how have you addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.

Unsupervised clustering for fraud detection typically involves using machine learning algorithms to identify patterns or anomalies in a dataset without needing labelled data. In our solution, we initially undertook an exploratory data analysis and cleaning of both datasets. This is crucial to understand the features and their distributions and to manage anomalies such as missing data, outliers, and duplicate entries. This was followed by a feature engineering step and applying clustering algorithms to define clusters based on feature similarities. Once the clusters are identified we used the fraud user dataset to define rules to identify potential fraudulent users.

Figure 1 presents the high-level workflow of our solution. Initially, datasets 1 and 2 are provided as an input to the model. This dataset is unlabelled, meaning the model needs to gain prior knowledge about the data. We conducted an exploratory data analysis to learn about the dataset. This was followed by a feature engineering step where we used dimensionality reduction methods (section 3) to retain the most relevant features. Once the feature engineering is completed, then we apply three clustering algorithms (section 3) to find the most suitable clustering algorithm for the selected feature distribution. And finally, we conduct a pattern recognition in each top cluster to derive fraud rules by using the fraudulent user data as a lookup dataset.

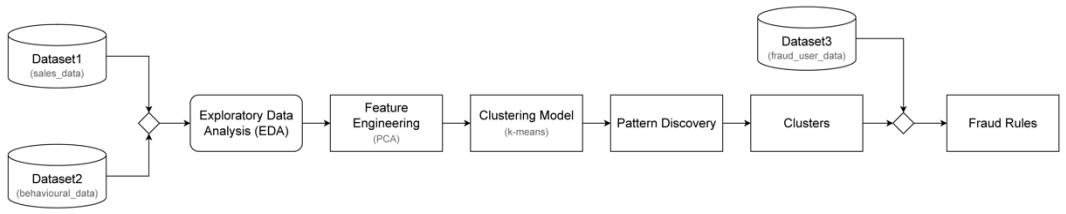

Figure 1: High-level Workflow of the Proposed Solution

3. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools identified to accomplish the challenge/Theme Challenges. Show clear understanding of the used REACH dataset/s and addressed challenge/Theme Challenges.

Challenge: UNSUPERVISED CLUSTERING FOR FRAUD DETECTION (REACH-2022-READYMADE-ALMERYS_2.2)

Tools: FedEHR Anonymizer (the tool is already used with the provided dataset)

Programming Language and Environment: Python 3.9, Jupyter Notebooks

Algorithms:

- Feature Engineering (Dimensionality Reduction)
  - Variable Ranking - a feature extraction method that assigns ranks or scores to each feature based on their relevance or importance for a specific task.
  - Correlation Heatmap - a feature extraction method that visually represents the pairwise correlation between different features in a dataset using a colour-coded matrix.
  - Principal Component Analysis (PCA) - a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional representation while preserving essential information. By identifying the principal components, which capture the maximum variance in the data, PCA enables the reduction of feature space complexity.
  - Independent Component Analysis (ICA) – this method observe data in a linear mixture of independent sources and aims to recover the sources by estimating a mixing matrix.
- Clustering Models
  - K-means Clustering – an unsupervised machine learning algorithm used to partition a dataset into distinct clusters based on similarity. It starts by randomly selecting k initial cluster centroids and assigns each data point to the nearest centroid based on their feature similarity, using a similarity metric.
  - Hierarchical Clustering - a clustering model of grouping data points into hierarchical clusters based on their similarity or distance.

Datasets: Dataset1 (sales data) and Dataset2 (behavioural data) for the unsupervised clustering and Dataset3 (fraudulent user data) to derive fraudulent user rules

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

4. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Discuss whether the solution can truly cope with humongous and increasing datasets and how flexible it is to adapt to other related domains and integrate into Data Value Chains (DVC).

Scalability in terms of:

1. Dataset size
   To support larger datasets we have the flexibility of using
   - Sampling and subsetting: this reduce the computational requirements while still providing insights into the clustering patterns. However, this approach may sacrifice some level of accuracy.
   - Parallel processing: algorithms (i.e., Hierarchical clustering) can be parallelized to take advantage of distributed computing resources. By distributing the computation across multiple processors or machines, the processing time can be significantly reduced.

2. Domain: our current implementation would be generalised in a way where this could be used with any domain of dataset where the requirement is similar to the current problem domain.

5. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Describe the security level of the proposed solution, i.e. how authentication, authorization policies, encryption or other approaches are used to keep data secure. Explain how will be compliant with the current data legislations concerning security and privacy (e.g. GDPR).

The solution will be secure and we are currently implementing several security methods to ensure this:

1.) We will ensure that the uploaded data will be encrypted before it is being processed by our NLP/AI models. To do this, we will use the AES encryption method. In this regard, we are collaborating with the Glasgow University Software Services under Dr. Tim Stoerer.

2.) Only certain individuals (backend software developers/data scientists) will be able to access the data. To ensure this, we included this rule in our employee NDA policies and have an overall authorization policy within the business. In terms of the project, only 5 out of 30 employees currently have access to the data. .

3.) We will use real-time database monitoring as well as database firewalls to ensure that the database is secure from cyber security threats. In addition, we will also encrypt the data within the database at rest and at transit for extra security.

The solution will be compliant with current data regulations by ensuring that data owners have full transparency/control of the data. The Explainable Artificial Intelligence models will be able to transparently show how the data is processed. Further, all other requirements of the GDPR and similar rules are being upheld by the solution.

6. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process planed for the final product. Technologically, which are the potential risks in all the phases of the project (design of the solution, development, testing, deployment...) and indicate mitigation plans to still fulfil the challenge/Theme Challenges and data provider requirements.

The quality process planned for the solution will be a constant feedback and improvement loop. Technologically, these are the following risks of the project:

1.) AI risk during the design of the solution is that the required solution by the challenge owner might not be technologically feasible. F.e. our AI models cannot technically process the data the way the challenge owner likes. To mitigate this, we will develop the most technically advanced AI models possible and will adjust them to the given datasets. We will also communicate with the challenge owners.

2.) The test risks are code issues. To mitigate these we will conduct constant tests of the code, resolve bugs and use best coding practices. Another issue will be cyber attacks during this phase. To mitigate it, we will ensure that the source code is encrypted and can only be used by authorised people.

3.) Risks that happen during this time include the changing of requirements, false KPIs as well as data security issues. To mitigate these risks, we will communicate with and adjust the tech to the pre-set requirements. We will also implement measurable/feasible KPIs. Lastly, we will ensure that the uploaded data is encrypted.

4.) Technical risks during the deployment include the wrong deployment environment and overall failure of the code during the deployment. To mitigate these risks, we will first test our software on the development environment before

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

deploying it. We will also run tests on the features before deploying them and will test the code again.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981