### Project Plan

Project Title:Fraudulent Transactions detection using Retrieval Augmented Generation (RAG)

Research Question:   
How effective is Retrieval-Augmented Generation (RAG) in detecting anomalies in payment transactions compared to traditional machine learning methods?

### Project Objectives:

Assess the capability of RAG in detecting payment anomalies.  
Compare the effectiveness of RAG with traditional anomaly detection techniques.  
Develop a RAG-based model tailored for the detection of fraudulent transactions.

### Summary of Project and Background:

Payment anomalies, such as fraudulent transactions, pose significant risks to financial institutions and their customers. Traditional anomaly detection methods often rely on predefined rules or supervised learning models that may fail to capture novel or evolving patterns of fraud. Retrieval-Augmented Generation (RAG) combines retrieval-based and generative models, leveraging large datasets to enhance the detection of anomalies.  
In response to these limitations, there is a growing interest in more advanced detection techniques that can effectively capture complex and subtle anomalies associated with fraudulent transactions. One promising approach is Retrieval-Augmented Generation (RAG), which represents an innovative fusion of retrieval-based and generative models.  
The retrieval component of RAG leverages a vector database to extract pertinent information from extensive datasets of historical payment transactions. This retrieval process ensures that the model has access to a rich repository of relevant examples and patterns. On the other hand, the generative aspect of RAG synthesizes this retrieved information to make accurate predictions about whether a transaction is fraudulent or legitimate. By combining these two approaches, RAG harnesses the power of large-scale data to potentially outperform traditional anomaly detection methods.  
The research project focuses on collecting a diverse dataset of payment transactions, encompassing legitimate and fraudulent cases. Rigorous preprocessing, including data cleaning, normalization, and feature engineering, will enhance data quality and extract valuable insights like transaction frequency and aggregated amounts. The project will implement the Retrieval-Augmented Generation (RAG) model, evaluating its performance against traditional methods like logistic regression and decision trees. Metrics such as precision, recall, and F1 score will gauge RAG's effectiveness in improving fraud detection accuracy and adapting to evolving fraud patterns. The goal is to minimize financial losses, maintain customer trust, and address the growing complexity of fraudulent activities in payment transactions.  
  
Reference List:  
Correa Bahnsen, A., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134-142. doi:https://doi.org/10.1016/j.eswa.2015.12.030  
  
Douze, M., Alexandr, M., Zilliz, G., Deng, C., Fair, J., Gergely, M., Fair, S., Pierre-Emmanuel, M., Fair, M., Lomeli, M., Hosseini, M., Labs, S. and Kyutai, H. (n.d.). THE FAISS LIBRARY. [online] Available at: https://arxiv.org/pdf/2401.08281  
  
Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv.org. doi:https://doi.org/10.48550/arXiv.2005.11401

### Task List and Project Timeline

‌

|  |  |
| --- | --- |
| **Task** | **Description** |
| Literature Review and Background Research | Conduct an extensive review of current literature. |
| PDM Plan Presentation | Preparing the project and data management plan and presentation. Presentation to take place on 17th June. |
| Data Collection and Preprocessing | Gather and preprocess necessary data for analysis and model development. |
| Explorative Data Analysis (EDA) | Perform exploratory analysis to gain insights into the dataset and identify patterns. |
| Feature Engineering | Engineer relevant features from collected data to improve model performance. |
| Data Ethics Quiz | Preparation for Data Ethics quiz to be carried out. Quiz is scheduled on 03rd July. |
| Model Development and Testing | Implement and train baseline models such as logistic regression and decision trees. |
| RAG Model Implementation | Implement the Retrieval-Augmented Generation (RAG) model framework |
| Model Training and Validation | Train the RAG model on the dataset and validate its performance using cross validation techniques. |
| Model Optimizations | Fine tuning the model to result promising results |
| Comparative Analysis | Conduct a comparative analysis between the RAG model and baseline models. |
| Final Review | Conduct a final review of the project, ensuring all aspects are complete and meet requirements. |
| Final Project Report + Logbook Submission | Preparing the final project report and maintaining the logbook. Submission to take place on 29th August |
| Viva | Preparation and participation to Viva. |

#### Timeline:

### Data Management Plan

Data Overview:  
The dataset used in this project is sourced from BankSim, comprising an aggregated sample of transactional data from a Spanish bank. Its primary purpose is to facilitate research in fraud detection through synthetic data generation.  
  
Data Collection:   
The dataset is available for download as a .csv file from the Kaggle repository (Link: [Kaggle BankSim Dataset](https://www.kaggle.com/datasets/ealaxi/banksim1/data))

#### Summary of data: This dataset contains historical data on 594,643 transactions from 4,112 unique users. The target variable (fraud) identifies fraudulent payments, while the remaining seven columns include a time step identifier, personal information about the payer (such as an identifier, age, and gender), and transaction details (including merchant, category, and amount). Out of all the transactions in the dataset, only 7,200 records are identified as fraud, resulting in a highly imbalanced dataset. Document Control:

GitHub Repository: [fraud-detection-rag](https://github.com/21062872/fraud-detection-rag.git)  
GitHub repository will be utilized for version control and documentation of the project. All code, data, and documentation files will be stored in the repository, ensuring accessibility and traceability of changes throughout the project's lifecycle.  
The following conventions for file naming will be employed within the repository:  
Code Files: module-name\_version\_number.py (e.g., data\_preprocessing\_v1.0.py)  
Data Files: dataset\_name\_version\_number.csv (e.g., transaction\_data\_v1.0.csv)  
Documentation: document\_name\_version\_number.docx (e.g., project\_plan\_v1.0.docx)  
Reports/Visualizations: report\_name\_version\_number.pdf (e.g., evaluation\_report\_v1.0.pdf)  
Version Control:  
I will be using the ‘main’ branch to store stable versions of the code and the ‘development’ branch for ongoing development. Once the development of a particular file is complete and tested, it will be merged into the main branch. A comprehensive README file will be included in the GitHub repository, providing detailed instructions and information about the code.  
  
Security and storage: Backups of the project repository will be performed weekly to ensure data integrity and minimize the risk of data loss. These will be stored securely on Google cloud, with restricted access.

Ethical Requirements:   
The dataset does not include personal information or disclose legal and private customer transactions, ensuring compliance with GDPR requirements by not involving the processing of personal data. Its synthetic nature guarantees the maintenance of privacy and confidentiality.  
Additionally, the dataset adheres to UH Ethical Policies as it is anonymized, ensuring ethical compliance.  
The dataset is openly accessible for academic and research purposes, authorized for use in research by BankSim.Furthermore, the dataset was collected, synthesized, and validated using statistical and Social Network Analysis (SNA) techniques by the original creators of the dataset. The dataset does not contain any personal or sensitive information, ensuring that ethical standards were maintained during the data collection process.