**Data Science Project 7PAM2002-0509-2023**

**Semester C 2023**

**Logbook (Activities and GitHub submissions)**

**Student Name and ID: Dhanuja Tharinda Memmenda Arachchi (21062872)**

**Project Title: Fraud Detection using Retrieval Augmented Generation**

**Supervisor: Dr.** **Vito Graffagnino**

**Student GitHub URL:** [**https://github.com/21062872/fraud-detection-rag**](https://github.com/21062872/fraud-detection-rag.git)

**Number of versions of the code submitted on GitHub: 12**

**User documentation has been submitted on GitHub: YES**

**Student GitHub URL has been shared with markers: YES**

**Log of Activities**

**Must record attendance at lectures and supervisions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Week** | **Date** | **Activity**  **incl. lectures & supervisions** | **Reason if not attend lecture or supervision** | **Weekly project progress.**  **How lecture/supervision was helpful to your project.** |
| 1 | 14/05/2024 | Lecture 1 |  | This lecture provided an overview of the Data Science project module and focused on selecting the project. |
| 2 | 21/05/2024 | Lecture 2 |  | This lecture covered how to frame research questions, plan the project, and manage data effectively. |
| 3 | 31/05/2024 | Supervision 1 |  | During the initial meeting with my supervisor, I discussed the project title and research question. The supervisor provided insights on how to choose the research questions, dataset what guidelines to follow. |
| 4 | 3/06/2024 | Lecture 3 |  | Attended the lecture on Data Ethics Part 1 |
| 5 | 10/06/2024 | Supervision 2 |  | We discussed the progress of project planning and addressed concerns that arose when selecting a dataset. We also talked about the importance of ethical considerations and choosing a dataset that adheres to both university policy and GDPR guidelines. |
| 6 | 17/06/2024 | Lecture 4 and Supervision 3 |  | Lecture on Data Ethics Part 2 Conducted the PDM Plan presentation to the supervisor and the group |
| 7 | 03/07/2024 | Quiz |  | Data Ethics quiz. |
| 8 | 15/07/2024 | Supervision 4 |  | We discussed the progress of the project and the final report, making sure we are staying aligned with the project timeline |
| 9 | 15/07/2024 | Lecture 5 |  | Lecture on Final Project Report |
| 10 | 26/07/2024 | Supervision 5 |  |  |
| 11 |  |  |  |  |
| 12 |  |  |  |  |
| 13 |  |  |  |  |
| 14 |  |  |  |  |
| 15 |  |  |  |  |
| 16 |  |  |  |  |

**Log of GitHub Submissions**

**Record the versions of code and user documentation submitted on GitHub**

|  |  |  |
| --- | --- | --- |
| **Date** | **Filename and version submitted to GitHub** | **Description of code and/or documentation submitted (what has been added since the previous version).** |
| 11/06/2024 | code\_files/README.md data\_files/bs140513\_032310.csv | Initial commit of the artefacts to repository |
| 11/06/2024 | documents/pdm\_plan\_v\_1.0.docx | Added pdm plan draft version to GIT repository |
| 16/06/2024 | documents/pdm\_plan\_v\_1.1.docx | Added final version of pdm plan |
| 30/06/2024 | code\_files/main.ipynb | Added main code file to the repository |
| 02/07/2024 | code\_files/main.ipynb | Added EDA part to the code. Contains plots |
| 11/07/2024 | code\_files/main.ipynb | Added more pre-processing and feature engineering. |
| 14/07/2024 | code\_files/main.ipynb | Added experiments done for Random Forest classifier model |
| 31/07/2024 | code\_files/main.ipynb | Added Logistic regression code and updated the logbook |
| 06/08/2024 | code\_files/main.ipynb | Added code for XGBoost method. Carried out multiple experiments |
| 12/08/2024 | code\_files/main.ipynb | RAG experiments added. Focused on retriever related experiments along with embeddings. |
| 13/08/2024 | code\_files/main.ipynb | Rag experiments contd. Generation part was added. Prototype model was created and tested. |
| 19/08/2024 | code\_files/main.ipynb | RAG model finalization and more experiments done on embeddings and similarity search |
| 20/08/2024 | Documents/ [Final Project Report - 21062872.docx](https://github.com/21062872/fraud-detection-rag/blob/main/documents/Final%20Project%20Report%20-%2021062872.docx) | Adding initial version of project report with background discussion till research question and objectives |
| 21/08/2024 | Documents/ [Final Project Report - 21062872.docx](https://github.com/21062872/fraud-detection-rag/blob/main/documents/Final%20Project%20Report%20-%2021062872.docx) | Adding Literature review. Discussed relevant topics, justification and literature about topics |
| 22/08/2024 | Documents/ [Final Project Report - 21062872.docx](https://github.com/21062872/fraud-detection-rag/blob/main/documents/Final%20Project%20Report%20-%2021062872.docx) | Writing methodology section along with model configurations, architecture and customizations |
| 24/08/2024 | code\_files/main.ipynb | Adding relationship plots |
| 25/08/2024 | Documents/ [Final Project Report - 21062872.docx](https://github.com/21062872/fraud-detection-rag/blob/main/documents/Final%20Project%20Report%20-%2021062872.docx) | Results section completion along with analysis. Comparative analysis done along with results evaluations |
| 25/08/2024 | code\_files/main.ipynb | Finalizing code structure |
| 26/08/2024 | code\_files/main.ipynb | Added experiments for BERT vs OPEN AI embeddings |
| 26/08/2024 | Documents/[Enhancing Fraud Detection by Ensemble Machine Learning and RAG by Tharinda Arachchi.](https://github.com/21062872/fraud-detection-rag/blob/main/documents/Enhancing%20Fraud%20Detection%20by%20Ensemble%20Machine%20Learning%20and%20RAG%20by%20Tharinda%20Arachchi.docx)docx | Finalizing project report with conclusion, limitation and future work, Adding styles and code |