

## **Data Collection and Preprocessing Phase**

|               |   |
|---------------|---|
| Date          | 09 July 2024  |
| Team ID       | SWTID1719935665   |
| Project Title | GeminiDecode: Multilanguage Document Extraction by Gemini Pro |
| Maximum Marks | 2 Marks   |

### **Data Collection Plan & Raw Data Sources Identification:**

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

#### **Data Collection Plan:**

| Section                     | Description  |
|-----------------------------|--|
| <b>Project Overview</b>     | The GeminiDecode project aims to develop a General Artificial Intelligence (GenAI) model named Gemini Pro, designed to accurately extract and interpret information from documents written in multiple languages. The primary objective is to enhance document processing efficiency and accessibility in multilingual environments, thus facilitating seamless information retrieval and decision-making across different sectors.  |
| <b>Data Collection Plan</b> | <ol style="list-style-type: none"> <li><b>Academic Databases:</b> Online repositories containing research papers, theses, and dissertations in multiple languages.</li> <li><b>Online Libraries:</b> Digital libraries offering a wide range of books, articles, and periodicals in various languages.</li> <li><b>Government Portals:</b> Official websites providing public documents, reports, and records.</li> <li><b>Corporate Archives:</b> Internal document repositories from partnering organizations, including reports, manuals, and memos.</li> </ol> |

|                                    |  |
|------------------------------------|--|
|                                    | <b>5. Public Datasets:</b> Open-access datasets available on platforms like Kaggle and UCI Machine Learning Repository, containing various types of documents. |
| <b>Raw Data Sources Identified</b> | <b>1. Multilingual Research Papers:</b> A collection of research papers in multiple languages sourced from academic databases.                                 |
|                                    | <b>2. Online Library Collections:</b> Digital books and articles in various languages sourced from online libraries.   |
|                                    | <b>3. Government Documents:</b> Official documents and records available on government portals.  |
|                                    | <b>4. Corporate Archives:</b> Internal documents from partnering organizations, including reports, manuals, and memos.   |
|                                    | <b>5. Public Datasets:</b> Open-access datasets available on platforms like Kaggle and UCI Machine Learning Repository, containing various types of documents. |

**Raw Data Sources:**

| <b>Source Name</b> | <b>Description</b>           | <b>Location/URL</b>   | <b>Format</b> | <b>Size</b> | <b>Access Permissions</b> |
|--------------------|------------------------------|---|---------------|-------------|---------------------------|
| Dataset 1          | Multilingual Research Papers | <a href="https://tinyurl.com/49fbcrvh">https://tinyurl.com/49fbcrvh</a> | DOCX          | 30 GB       | Public                    |

|           |                            |   |      |       |                       |
|-----------|----------------------------|---|------|-------|-----------------------|
| Dataset 2 | Online Library Collections | <a href="https://tinyurl.com/49fbcrvh">https://tinyurl.com/49fbcrvh</a> | DOCX | 20 GB | Public                |
| Dataset 3 | Government Documents       | <a href="https://tinyurl.com/49fbcrvh">https://tinyurl.com/49fbcrvh</a> | DOCX | 15 GB | Public                |
| Dataset 4 | Corporate Archives         | <a href="https://tinyurl.com/49fbcrvh">https://tinyurl.com/49fbcrvh</a> | DOCX | 10 GB | Private (with access) |
| Dataset 5 | Public Datasets            | <a href="https://tinyurl.com/49fbcrvh">https://tinyurl.com/49fbcrvh</a> | DOCX | 25 GB | Public                |