**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 15 March 2024 |
| Team ID | 740050 |
| Project Title | YOLOChemDetect safeguarding with Automated Drug Name Detection |
| Maximum Marks | 2 Marks |

**Data Collection Plan & Raw Data Sources Identification Template**

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 Template**

|  |  |
| --- | --- |
| **Section** | **Description** |
| Project Overview | Automate drug name detection to enhance YOLOChemDetect’s accuracy. |
| Data Collection Plan | Data will be collected from medical documents, clinical trials, and pharmaceutical sources. |
| Raw Data Sources Identified | 1. **Medical Datasets**: Public datasets with drug information.  2. **Clinical Trials**: Drug-related reports.  3. **Pharmaceutical Websites**: Data from drug manufacturers.  4. **EHR**: Anonymized hospital records. |

**Raw Data Sources Template**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source Name** | **Description** | **Location/URL** | **Format** | **Size** | **Access Permissions** |
| Dataset | A collection of medical documents containing drug names and related information. | https://www.kaggle.com/datasets/pkdarabi/The-Drug-Name-Detection-Dataset | image | 58MB | Private (with access) |