**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 15 July 2024 |
| Team ID | 740713 |
| Project Title | Genetic Classification Of An Individual by Using Machine Learning |
| 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 | This project aims to classify individuals based on their genetic information using machine learning techniques. The objective is to build a model that can accurately categorize genetic data into predefined classes, aiding in personalized medicine and genetic research. |
| Data Collection Plan | Kaggle |
| Raw Data Sources Identified | 1000 Genomes Project  Genome Aggregation Database (gnomAD)  ClinVar |

**Raw Data Sources Template**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source Name** | **Description** | **Location/URL** | **Format** | **Size** | **Access Permissions** |
| Dataset 1 | Predict whether a variant will have conflicting clinical classifications. | <https://www.kaggle.com/datasets/kevinarvai/clinvar-conflicting> | CSV | 4MB | Public |