Data Collection and Preprocessing Phase

|  |  |
| --- | --- |
| Date | 12 July 2024 |
| Team ID | 740022 |
| Project Title | Lymphography Classification Using ML |
| Maximum Marks | 2 Marks |

# Data Collection Plan & Raw Data Sources Identification Report:

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** | **Descripion** |
| Project Overview | The machine learning project aims to predict the condition of lymph nodes. It’s so significant that it is used to accurately sense the stage in cancer progression,which decides the treatment to be adopted. Using a dataset with features such as class,lymphatics,by pass the objective is to build a model that accurately sense the stage in cancer progression,facilitating efficient and informed decision-making in the lending process. |
| Data Collection Plan | This detailed description outlines a comprehensive approach to collecting and preparing data for lymphography classification, ensuring transparency, compliance, and readiness for subsequent stages of machine learning model development |
| Raw Data Sources Identified | The raw data sources for this project include datasets obtained from Kaggle & UCI, the popular platforms for data science competitions and repositories. The provided sample data represents a subset of the collected information, encompassing variables such as class, |

|  |  |
| --- | --- |
|  | by pass, block of affere, extravasates and changes in node details for machine learning analysis |

**Raw Data Sources Report:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source Name** | **Description** | **Location/URL** | **Format** | **Size** | **Access Permissions** |
| Kaggle Dataset | The dataset comprises annotated medical images aimed at classifying conditions of the lymphatic system. It includes diverse lymphography images labeled with categories such as Normal, Metastasis, and Primary Tumor, crucial for automated diagnosis. | <https://www.kaggle.com/datasets/naifislam/lymphography-dataset> | CSV | 5.62kB | Public |
| UCI | It consists of lymphography data essential for medical imaging research. It includes features detailing lymphatic system status and imaging findings related to lymph node conditions. | <https://archive.ics.uci.edu/dataset/63/lymphography> | CSV | 2.98kB | Public |