**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 12 July 2024 |
| Team ID | SWTID1720077079 |
| Project Title | Wild Blueberry Yield Prediction |
| Maximum Marks | 2 Marks |

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

For our ML-based blueberry yield prediction system, we sourced a high-quality dataset from Kaggle, consisting of an 87 KB CSV file. This dataset includes crucial features like weather conditions, soil properties, and agricultural practices. Our data collection plan ensures meticulous curation, supporting accurate analysis and informed decision-making to optimize blueberry yields.

**Data Collection Plan**

|  |  |
| --- | --- |
| **Section** | **Description** |
| Project Overview | A machine learning-based system to accurately predict blueberry yields, addressing the challenges faced by farmers in yield estimation. |
| Data Collection Plan | Obtains a dataset from Kaggle |
| Raw Data Sources Identified | CSV file from Kaggle (87 kb)  The dataset includes 777 entries with 18 columns detailing blueberry yield factors such as clone size, pollinator counts, temperature ranges, rainy days, fruit set rate, mass, seed count, and overall yield. |

**Raw Data Sources**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source Name** | **Description** | **Location/URL** | **Format** | **Size** | **Access Permissions** |
| Kaggle | A csv file detailing blueberry yield factors such as clone size, pollinator counts, temperature ranges, rainy days, fruit set rate, mass, seed count, and overall yield. | https://www.kaggle.com/datasets/saurabhshahane/wild-blueberry-yield-prediction | CSV | 87 KB | Public |