# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 8 July 2024 |
| Team ID | SWTID1720078167 |
| Project Title | Rice Type Classification using CNN |
| Maximum Marks | 2 Marks |

**Data Quality Report Template**

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

| Data Source | Data Quality Issue | Severity | Resolution Plan |
|---|---|---|---|
| Dataset (Kaggle) | Missing Images | Moderate | Check for missing or broken links in the dataset and re-download any missing images. Use a script to verify the integrity of each image file. |
| Dataset (Kaggle) | Image Size Inconsistency | High | Resize all images to a standard size (e.g., 224x224 pixels) to ensure uniformity for model training. Use OpenCV or PIL for resizing. |
| Dataset (Kaggle) | Image Quality (Blurriness, Noise) | Moderate | Apply denoising filters (e.g., Gaussian Blur) to improve image quality. Detect and remove |

| | | | or enhance blurry images using sharpness metrics. |
|---|---|---|---|
| Dataset (Kaggle) | Duplicate Images | Low | Identify and remove duplicate images using hash functions or similarity metrics. Use libraries like imagehash or scikit-image for detection. |
| Dataset (Kaggle) | Color Variations | Low | Normalize color distributions by converting images to a uniform color space (e.g., HSV) and applying histogram equalization. |