Data Collection and Preprocessing Phase

Date	27 May 2025
Team ID	SWUID20240006489
Project Title	Gemini Decode: Multilanguage Document Extraction by Gemini Pro
Maximum Marks	2 Marks

Data Quality Report

Overview: The Data Quality Report summarizes data quality issues from the document extraction process, including severity levels and resolution plans. It aims to systematically identify and rectify data discrepancies, ensuring high accuracy and efficiency in data extraction from multilingual documents.

Data Source	Data Quality Issue	Severity	Resolution Plan
Extracted Invoices Dataset	Missing text in specific fields due to OCR limitations	High	Enhance OCR model accuracy by training on a diverse dataset, and implement post-OCR validation checks.
Extracted Legal Documents Dataset	Misclassification of document types	Moderate	Improve the classification algorithm by incorporating additional features and refining the training data.

Extracted Financial Statements Dataset	Inconsistent numerical data formats	Low	Standardize numerical data formats using data preprocessing scripts before analysis.
Extracted Medical Records Dataset	Incorrect language detection for multilingual documents	High	Integrate advanced language detection algorithms and cross-validate with known language segments.

Extracted Business Reports Dataset	Duplicate data entries due to repeated scans	Moderate	Implement deduplication techniques to identify and merge duplicate entries based on unique identifiers.
Extracted Research Papers Dataset	Incomplete metadata extraction	Low	Automate metadata extraction processes and cross-reference with existing bibliographic databases.
Extracted Contracts Dataset	Data entry errors in extracted text	High	Use automated error detection tools to identify and correct data entry errors, and employ human verification for critical sections.
Extracted Receipts Dataset	Outliers in numerical data due to scanning artifacts	Moderate	Apply statistical methods to detect and correct outliers, such as z-score analysis or the IQR method.
Extracted Surveys Dataset	Missing responses in survey data	High	Use imputation techniques to fill missing responses, and ensure completeness by prompting users for mandatory fields during survey collection.
Extracted Documents Dataset	Data integrity issues due to format variations	Moderate	Establish data integrity constraints and conduct regular audits to ensure consistency across different formats.