Data Collection and Preprocessing Phase

Date	18 July 2024
Team ID	SWTID1721319573
Project Title	Blueberry Yield Prediction
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
Blueberries Dataset	Missing values in columns 'clonesize' and 'honeybee'	High	Fill missing values using forward fill method.
Blueberries Dataset	Inconsistent data types for 'harvestdate'	Moderate	Convert 'harvestdate' to a consistent datetime format.
Blueberries Dataset	Outliers in 'yield' column	High	Identify and treat outliers using IQR method.
Blueberries Dataset	Duplicate rows	Low	Remove duplicate rows using drop_duplicates() method.
Blueberries Dataset	Mixed units in 'temperature' columns	Moderate	Standardize all temperature values to a single unit (e.g., Celsius).

Blueberries Dataset	Incorrect values in 'soil_quality' (negative values)	High	Replace negative values with the median of the column.
Blueberries Dataset	Typographical errors in categorical data	Moderate	Standardize categorical data using a predefined mapping.
Blueberries Dataset	Missing entries in 'weather_conditions'	Low	Fill missing values with the most frequent category.