Data Collection and Preprocessing Phase

Date	24 June 2025
Team ID	SWUID20250177148
Project Title	Machine Learning Approach for Employee Performance Prediction
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report summarizes the key quality issues found in the dataset during the Data Collection and Preprocessing phase, with severity levels and planned resolutions to ensure clean, usable data for the machine learning model.

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle Dataset	Missing values in wip, incentive, idle_time, and idle_men columns	Moderate	Used SimpleImputer with mean strategy for imputation
Kaggle Dataset	Categorical variables: quarter, department, day	Moderate	Label encoded using LabelEncoder or manual mapping
Kaggle Dataset	Date column not suitable for model use directly	Low	Extracted month from date, then dropped date column
Kaggle Dataset	Outliers in idle_time and incentive	Low	Outliers detected via boxplots, retained due to model tolerance