Data Collection and Preprocessing Phase

Date	24 June 2025
Team ID	SWUID20250176341
Project Title	Machine Learning Approach for Employee Performance Prediction
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan:

Section	Description
Project Overview	The machine learning project aims to predict employee performance based on workplace and departmental data. Using features such as SMV, incentives, overtime, and departmental info, the objective is to build a model that forecasts actual productivity. This enables better HR decision-making and resource optimization.
Data Collection Plan	• Search for datasets related to employee productivity and performance evaluation in manufacturing sectors. • Prioritize datasets containing both numerical and categorical features like team, department, SMV, and productivity scores.
Raw Data Sources Identified	The raw data source for this project is a dataset obtained from **Kaggle**, specifically the *Garments Worker Productivity* dataset. It includes fields such as department, team, targeted and actual productivity, incentives, overtime, and work-in-progress, providing a reliable base for machine learning analysis.

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	Contains garment worker data including productivity, incentives, and departmental info.	https://www.kaggle.com/datasets/utkarshsarbahi/productivity-prediction-of-garment-employees	CSV	93 kB	Public