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 Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed machine learning analysis and predictions.

Data Collection Plan Template

Section	Description
Project Overview	The machine learning project aims to predict employee productivity based on operational and production-related factors. Using a dataset containing features such as SMV, overtime, incentive, idle time, and number of workers, the objective is to build a predictive model that helps HR teams and managers make data-driven decisions regarding performance support and task allocation.
Data Collection Plan	• Searched for real-world garment production or employee productivity datasets • Prioritized structured datasets with categorical and numerical features • Required a dataset with timestamps, team configurations, and actual productivity ratings • Source selection based on quality, structure, and feature richness for modeling
Raw Data Sources Identified	The dataset is taken from the Kaggle in CSV form

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	The dataset comprises garment worker details such as team, targeted productivity, SMV, WIP, incentive, idle time, and Actual productivity scores	https://www.kaggle.com/datasets/anmolkumar/garments-worker-productivity	CSV	95kb	Public