# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 28 June 2025 |
| Team ID | NONE |
| Project Title | Employee Performance Prediction using Machine Learning |

### 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 productivity based on historical work-related data. Using a dataset with features such as overtime, idle time, department, team size, and quarter, the objective is to build a regression model that accurately forecasts the actual_productivity of employees. This enables organizations to identify performance trends early, support underperforming teams, and make data-driven decisions for resource allocation and training. |
| Data Collection Plan | ● Identified and downloaded the "Garment Employee Productivity" dataset from Kaggle, a reliable source for real-world industrial datasets.<br>● Verified dataset completeness, structure, and relevance to the problem of productivity prediction.<br>● Selected features such as over_time, idle_time, department, team, and actual_productivity for modeling.<br>● Ensured the dataset was in CSV format and compatible with Python-based data analysis tools (pandas, scikit-learn).<br>● Documented data types, missing values, and categorical variables for preprocessing planning. |

| | |
|---|---|
| Raw Data Sources Identified | The raw data source for this project is the "Garment Employee Productivity" dataset obtained from Kaggle, a widely used platform for real-world machine learning datasets. The dataset contains detailed information on employee work patterns in garment manufacturing units, including features such as over_time, idle_time, department, team_size, quarter, and the target variable actual_productivity. This dataset was selected for its relevance to workforce performance analysis and its suitability for regression-based productivity prediction. |

**Raw Data Sources Report:**

| Source Name | Description | Location/URL | Format | Size | Access Permissions |
|---|---|---|---|---|---|
| Kaggle Dataset | Contains detailed information on employee work patterns in garment manufacturing units, including features such as<br>over_time,idle_time,department,team_size,quarter,and the target variable<br>actual_productivity. | [https://www.kaggle.com/datasets/utkarshsarbahi/productivity-prediction-of-garment-employees?select=garments_worker_productivity.csv](https://www.kaggle.com/datasets/utkarshsarbahi/productivity-prediction-of-garment-employees?select=garments_worker_productivity.csv) | CSV | 94 kB | Public |