Data Collection and Preprocessing Phase

Date: 31 July 2025
SkillWallet ID: SWUID20240141492
Project Title: Employee Performance Prediction using Machine Learning

---

Maximum Marks: 2 Marks

Data Collection Plan & Raw Data Sources Identification Report

Data Collection Plan:

Project Overview:

This machine learning project aims to predict employee performance scores using historical productivity data. The dataset includes features such as **quarter, department, targeted productivity, overtime, incentive, idle time, and team size**, enabling accurate performance prediction. This model will assist HR teams in identifying high and low performers and improving workforce management.

Data Collection Plan:

- Search for datasets related to **employee productivity and HR performance metrics**.
- Prioritize datasets with **detailed workforce-related features** (productivity scores, overtime, idle time).
- Select data from **reliable sources** like Kaggle HR and productivity datasets.
- Ensure datasets contain **enough samples for model training and validation**.

Raw Data Sources Identified:

The raw dataset for this project is sourced from **Kaggle**, focusing on garment manufacturing employee productivity records, which can be generalized for employee performance modeling.

Raw Data Sources Report:

Source Name	Description	Name	Format	Size	Access Permissions
Kaggle Dataset	Contains employee work records including **quarter, department, targeted productivity, SMV, overtime, incentive, idle time, and performance scores**.	Garment Worker Productivity Dataset	CSV	~6 MB	Free Public Access