Data Collection and Preprocessing Phase

Date	15 July 2024
Team ID	SWTID1720108643
Project Title	Garment worker productivity prediction
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Building a Robust Foundation: Data Collection and Source Identification

This project prioritizes a meticulous data strategy to ensure accurate and reliable insights into garment worker productivity. Our Data Collection Plan outlines a comprehensive approach, encompassing public datasets from platforms like Kaggle, relevant academic research identified through Google Scholar searches, and exploration of open-source repositories on GitHub. This multi-pronged approach aims to gather a diverse and informative dataset.

Furthermore, our Raw Data Sources report meticulously details the identified sources, including details about their content and potential limitations. For instance, we highlight missing values in the "WIP" (Work in Progress) column concentrated in the Finishing department, requiring imputation. Additionally, features like "Idle_men," "idle_time," and "no_of_style_change" exhibit a high frequency of zeros, which will be investigated further.

Data Collection Plan Template

Section	Description
Project Overview	This project aims to develop a machine learning model that can accurately predict garment worker productivity. This will allow garment manufacturers to: • **Optimize Production Processes:** By predicting worker output, manufacturers can allocate resources more effectively, minimize idle time, and ensure timely completion of production goals. • **Improve Worker Well-Being:** Identifying factors influencing productivity can lead to adjustments in working conditions, incentive structures, and training programs,

	ultimately fostering a positive work environment for garment workers.
Data Collection Plan	To construct a robust dataset for our garment worker productivity prediction model, we will employ a multi-source approach: • **Public Datasets:** We will leverage public repositories like Kaggle, which offer datasets specifically tailored to garment worker productivity. • **Academic Research:** Relevant research papers identified through Google Scholar searches will be reviewed to explore the data employed in similar studies. This will provide insights into best practices for data collection and feature selection in this domain. • **Open-Source Repositories:** We will investigate open-source repositories on GitHub, where machine learning practitioners often share valuable datasets related to their projects. By exploring these resources, we can potentially access datasets that have been preprocessed and readied for machine learning tasks. This multifaceted data collection strategy will ensure we gather a rich and diverse dataset, encompassing a wide range of factors influencing garment worker productivity. This comprehensive data foundation will be crucial for building and training an accurate and effective prediction model.
Raw Data Sources Identified	Kaggle dataset: UC irvine machine learning repo:

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions

Dataset 1	The dataset Is having 1197 rows and 15 columns and each column containing information about time and factors that affect productivity.	https://www.kaggle.com/code/dmitriiannenkov/garment-worker-productivity-prediction	CSV	19kb	Public
Dataset 2	Description of the data in this source.	Link of Dataset 2	Excel	YY GB	Private (with access)
Dataset2	800 rows and 20 factors which will be affecting the worker productivity.There are no so many null valu	https://archive.ics.uci.edu/dataset/597/productivity+prediction+of+garment+employees	csv	16kb	Public