Project Report - Phase 2: Data Collection and Preprocessing

Author: Arpit Jain Date: August 20, 2025 Project: Employee Productivity Prediction

2.1 Data Sourcing and Initial Exploration

The project is built on the garments_worker_productivity.csv dataset, which contains 1,197 records and 15 attributes describing various aspects of employee performance in a garment factory. An exploratory data analysis (EDA) was conducted to:

- **Understand the dataset's structure, data types, and statistical properties.**
- **Examine the distribution of key variables.**
- **Identify potential data quality issues requiring remediation.**

This initial exploration guided the subsequent data preprocessing strategy.

2.2 Data Cleaning and Transformation

To ensure the dataset was suitable for machine learning, the following preprocessing steps were applied:

- **Standardization of Categorical Data: The department column included inconsistencies such as "sweing" and "finishing ". These were corrected to "sewing" and "finishing" to maintain categorical integrity.**
- **Handling of Missing Values: The wip (work in progress) column contained a substantial number of missing entries. To preserve dataset quality, this feature was dropped from the final model input.**
- **Feature Engineering: The date column was transformed to derive a new month feature, capturing potential seasonal effects on productivity. The original date column was then removed.**

2.3 Categorical Data Encoding

Since machine learning models require numerical inputs, categorical attributes (quarter, department, and day) were converted into numerical form. Label Encoding was applied, mapping each unique category within a feature to an integer value. This ensured the dataset was fully compatible with the modeling algorithms.

2.4 Final Dataset Preparation

After cleaning and encoding, the dataset was finalized for model development. Key steps included:

- **Splitting the dataset into features (X) and the target variable (y = actual_productivity).**
- **Ensuring a clean, structured dataset to support effective training, validation, and evaluation of multiple machine learning models in the subsequent phase.**