Project Report - Phase 2: Data Collection and Preprocessing

Author: Aryan Rai Date: August 5, 2025 Project: Employee Performance Prediction using a Machine Learning Approach

2.1. Data Sourcing and Initial Exploration

The foundation of this project is the garments_worker_productivity.csv dataset. This dataset contains 1,187 records and 16 distinct attributes related to the work of employees in a garment factory. An initial exploratory data analysis (EDA) was conducted to understand the structure, data types, and statistical properties of the dataset. This involved examining the distribution of key variables and identifying any immediate data quality issues.

2.2. Data Cleaning and Transformation

To prepare the data for machine learning, several cleaning and transformation steps were performed:

- **Standardization of Categorical Data: The department column contained inconsistencies such as "sweing" and "finishing ", which were corrected to "sewing" and "finishing" respectively. This ensures that the model treats these categories correctly.**
- **Handling of Missing Values: We will identify and handle any missing data points. This may involve filling in missing values with the mean, median, or mode of the respective features, or, in some cases, dropping records that have a significant amount of missing information. This ensures that the dataset is complete and ready for model training.**
- **Feature Engineering: The notebook Employee_Productivity_Analysis.ipynb may also include feature engineering, where new features are created from existing ones to improve the model's predictive power. For example, a new feature could be created by combining 'years of experience' and 'age' to create a more meaningful metric.**

2.3. Categorical Data Encoding

Many machine learning algorithms require numerical input. The raw employee data may contain categorical features like 'Department,' 'Job Role,' or 'Education Level.' We will convert these into a numerical format using techniques like One-Hot Encoding or Label Encoding, which are available in scikit-learn. This step is crucial for the model to interpret the data correctly.

2.4. Final Dataset Preparation

After the preprocessing steps, the dataset was finalized and prepared for the model development phase. The data was split into features (X) and the target variable (y, which is actual_productivity). This clean and structured dataset formed the basis for training and evaluating the machine learning models in the next phase.