

## Data Collection and Preprocessing Phase

Date	5th July 2024
Team ID	740141
Project Title	Garment Workers Productivity Predictions
Maximum Marks	6 Marks

### Preprocessing Template

The images will be preprocessed by Data collection, Handling missing values, Normalization, Data aggregation, Data splitting, Visualization, Data storage, Documentation. This template provides a concise yet complete approach to preprocessing a dataset in preparation for machine learning.

Section	Description
Project Overview	Brief description of the project and its goals.
Data Collection	List of data sources and Types of data collected
Handling Missing Values	Techniques used (e.g., imputation, removal).
Normalization:	Procedures for scaling data
Data Aggregation:	Summarizing data to higher-level formats.
Data Splitting	Proportion of data split and Strategy used for cross-validation

Visualization Tools:	Tools and libraries used (e.g., Matplotlib, Seaborn).																																																																																											
Data Storage	Systems used for storing preprocessed data (e.g., SQL, NoSQL, cloud storage).																																																																																											
Documentation	Detailed documentation of all preprocessing steps for reproducibility.and A comprehensive data dictionary describing all variables and their transformations																																																																																											
Data Preprocessing Code Screenshots																																																																																												
Loading Data	<pre>df = pd.read_csv(r'C:\Users\sraa\Downloads\miniProject\garments_worker_productivity.csv') df.head()</pre>																																																																																											
Data Collection	<table><thead><tr><th></th><th>date</th><th>quarter</th><th>department</th><th>day</th><th>team</th><th>targeted_productivity</th><th>smv</th><th>wip</th><th>over_time</th><th>incentive</th><th>idle_time</th><th>idle_men</th><th>no_of_style_change</th><th>no_of_workers</th></tr></thead><tbody><tr><td>0</td><td>1/1/2015</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>8</td><td>0.80</td><td>26.16</td><td>1108.0</td><td>7080</td><td>98</td><td>0.0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>1</td><td>1/1/2015</td><td>Quarter1</td><td>finishing</td><td>Thursday</td><td>1</td><td>0.75</td><td>3.94</td><td>NaN</td><td>960</td><td>0</td><td>0.0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>2</td><td>1/1/2015</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>11</td><td>0.80</td><td>11.41</td><td>968.0</td><td>3660</td><td>50</td><td>0.0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>3</td><td>1/1/2015</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>12</td><td>0.80</td><td>11.41</td><td>968.0</td><td>3660</td><td>50</td><td>0.0</td><td>0</td><td>0</td><td>0</td></tr><tr><td>4</td><td>1/1/2015</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>6</td><td>0.80</td><td>25.90</td><td>1170.0</td><td>1920</td><td>50</td><td>0.0</td><td>0</td><td>0</td><td>0</td></tr></tbody></table>			date	quarter	department	day	team	targeted_productivity	smv	wip	over_time	incentive	idle_time	idle_men	no_of_style_change	no_of_workers	0	1/1/2015	Quarter1	sweing	Thursday	8	0.80	26.16	1108.0	7080	98	0.0	0	0	0	1	1/1/2015	Quarter1	finishing	Thursday	1	0.75	3.94	NaN	960	0	0.0	0	0	0	2	1/1/2015	Quarter1	sweing	Thursday	11	0.80	11.41	968.0	3660	50	0.0	0	0	0	3	1/1/2015	Quarter1	sweing	Thursday	12	0.80	11.41	968.0	3660	50	0.0	0	0	0	4	1/1/2015	Quarter1	sweing	Thursday	6	0.80	25.90	1170.0	1920	50	0.0	0	0	0
	date	quarter	department	day	team	targeted_productivity	smv	wip	over_time	incentive	idle_time	idle_men	no_of_style_change	no_of_workers																																																																														
0	1/1/2015	Quarter1	sweing	Thursday	8	0.80	26.16	1108.0	7080	98	0.0	0	0	0																																																																														
1	1/1/2015	Quarter1	finishing	Thursday	1	0.75	3.94	NaN	960	0	0.0	0	0	0																																																																														
2	1/1/2015	Quarter1	sweing	Thursday	11	0.80	11.41	968.0	3660	50	0.0	0	0	0																																																																														
3	1/1/2015	Quarter1	sweing	Thursday	12	0.80	11.41	968.0	3660	50	0.0	0	0	0																																																																														
4	1/1/2015	Quarter1	sweing	Thursday	6	0.80	25.90	1170.0	1920	50	0.0	0	0	0																																																																														
Handling Missing Values	<pre>df2.isnull().sum()</pre> <table><tbody><tr><td>quarter</td><td>0</td></tr><tr><td>department</td><td>0</td></tr><tr><td>day</td><td>0</td></tr><tr><td>team</td><td>0</td></tr><tr><td>targeted_productivity</td><td>0</td></tr><tr><td>smv</td><td>0</td></tr><tr><td>wip</td><td>506</td></tr><tr><td>over_time</td><td>0</td></tr><tr><td>incentive</td><td>0</td></tr><tr><td>idle_time</td><td>0</td></tr><tr><td>idle_men</td><td>0</td></tr><tr><td>no_of_style_change</td><td>0</td></tr><tr><td>no_of_workers</td><td>0</td></tr><tr><td>actual_productivity</td><td>0</td></tr><tr><td>dtype:</td><td>int64</td></tr></tbody></table>		quarter	0	department	0	day	0	team	0	targeted_productivity	0	smv	0	wip	506	over_time	0	incentive	0	idle_time	0	idle_men	0	no_of_style_change	0	no_of_workers	0	actual_productivity	0	dtype:	int64																																																												
quarter	0																																																																																											
department	0																																																																																											
day	0																																																																																											
team	0																																																																																											
targeted_productivity	0																																																																																											
smv	0																																																																																											
wip	506																																																																																											
over_time	0																																																																																											
incentive	0																																																																																											
idle_time	0																																																																																											
idle_men	0																																																																																											
no_of_style_change	0																																																																																											
no_of_workers	0																																																																																											
actual_productivity	0																																																																																											
dtype:	int64																																																																																											

Normalization	<pre>df = pd.read_csv(r'C:\Users\sra\Downloads\miniProject\garments_worker_productivity.csv') df.head()</pre>
Data Aggregation	<pre>import numpy as np  # Sample data data = np.array([[1, 10],                  [2, 20],                  [3, 30],                  [4, 40],                  [5, 50]])  # Manual Min-Max normalization min_vals = np.min(data, axis=0) # Compute minimum values for each column max_vals = np.max(data, axis=0) # Compute maximum values for each column  # Normalize data normalized_data = (data - min_vals) / (max_vals - min_vals)  print("Original Data:") print(data) print("\nNormalized Data:") print(normalized_data)</pre>
Data Splitting	<pre>from sklearn.model_selection import train_test_split x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.30, random_state=42)</pre> <pre>print(x_train.shape) print(x_test.shape) print(y_train.shape) print(y_test.shape)</pre> <pre>(823, 12) (353, 12) (823,) (353,)</pre> <pre>from sklearn.metrics import mean_squared_error from sklearn.metrics import mean_absolute_error from math import sqrt from sklearn.metrics import mean_absolute_percentage_error</pre>
Visualization Tools:	<pre>plt.figure(figsize=(10,5)) p = sns.boxplot(data = df6, orient = 'v',width=0.8) plt.xticks(rotation=90)</pre> <pre>(array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9]),  [Text(0, 0, 'team_number'),   Text(1, 0, 'time_allocated'),   Text(2, 0, 'unfinished_items'),   Text(3, 0, 'over_time'),   Text(4, 0, 'incentive'),   Text(5, 0, 'idle_time'),   Text(6, 0, 'idle_men'),   Text(7, 0, 'style_change'),   Text(8, 0, 'no_of_workers'),   Text(9, 0, 'actual_productivity')])</pre>
Data Storage	<pre># Create DataFrame df = pd.DataFrame(data)  # Save to CSV df.to_csv('data.csv', index=False)  print("Data saved to data.csv")</pre>

## Documentation

```
df = pd.read_csv(r'C:\Users\sraira\Downloads\miniProject\garments_worker_productivity.csv')  
df.head()
```