Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	738306
Project Title	Machine Learning Approach For Employee Performance Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Cleanse and preprocess the collected data to handle missing values, outliers, and encode categorical variables. Ensure data quality and consistency for accurate model training.

Section	Description
Data Overview	The Dataset has 15 features and 1197 observations.
Univariate Analysis	Univariate analysis is a statistical method used to analyze one variable at a time. It provides insights into the distribution, central tendency, and variability of the variable, typically through summary statistics, visualizations like histograms or box plots, and measures like mean, median, mode, and standard deviation.
Bivariate Analysis	Bivariate analysis is a statistical method used to analyze the relationship between two variables. It helps to understand how changes in one variable are associated with changes in another. This analysis typically involves techniques such as correlation analysis, scatter plots, and regression analysis.
Multivariate Analysis	Multivariate analysis is a statistical technique used to understand relationships between multiple variables simultaneously. It explores how these variables interact and affect each other within a dataset, enabling researchers to uncover complex patterns and draw meaningful conclusions.

Outliers and Anomalies	Outliers and anomalies are data points that deviate significantly from the rest of the dataset. Outliers are typically legitimate data points but lie far from the majority, while anomalies are often errors or rare occurrences. Both can distort statistical analysis and machine learning models, so identifying and handling them appropriately is crucial for accurate data analysis.

Data Preprocessing Code Screenshots

Loading Data

```
#importing the dependencies
import pandas as pd
import numpy as np

#import dataset to the pandas dataframe
data=pd.read_csv('/content/garments_worker_productivity.csv')
```

Handling Missing Data

```
#checking for null values
data.isnull().sum()

#replace the null values with mean
data['wip'].fillna(data['wip'].mean(),inplace=True)
```

Data Transformation

```
#encoding the categorical values
#importing label encoder
from sklearn.preprocessing import LabelEncoder

Encoder=LabelEncoder()

data['quarter']=Encoder.fit_transform(data['quarter'])
data['department']=Encoder.fit_transform(data['department'])
data['day']=Encoder.fit_transform(data['day'])

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

x_train_scaled = scaler.fit_transform(x_train)
x_test_scaled = scaler.transform(x_test)

x_train_scaled=x_train
x_test_scaled=x_test
```

Feature Engineering	`data['quarter']=Encoder.fit_transform(data['quarter'])`
	`data['department']=Encoder.fit_transform(data['department'])`
	`data['day']=Encoder.fit_transform(data['day'])`

Save Processed Data	`#import dataset to the pandas dataframe` `data=pd.read_csv('/content/garments_worker_productivity.csv')`