**Title:**

TN Marginal Workers Assessment with Machine Learnings

**Introduction:**

Preprocessing data for product demand prediction with machine learning is a crucial step to ensure that your model can learn effectively from the data because it ensures data quality and prepares it for effective model training, resulting in accurate predictions, which are essential for business decision-making and customer satisfaction. Here are some common preprocessing steps:

**I) Data Selection and loading:**

Data selection and loading are vital steps in the data preparation process for machine learning. Data selection involves choosing the right dataset that aligns with your objectives, ensuring it contains relevant information. Data loading refers to the process of importing the chosen dataset into your machine learning environment, such as a Python notebook or database. This step includes reading data from files (e.g., CSV, Excel) or databases, making it ready for further analysis. Accurate data selection and loading set the foundation for successful machine learning tasks by providing the right information for model training and evaluation.

**Code:**

# Data selection and loading

import pandas

df=pandas.read\_csv("DDW\_B06SC\_3300\_State\_TAMIL\_NADU-2011.csv")

print(df.tail())

print(df.head(8))

print(df.info())

print(df.describe())

**II) Data Cleaning:**

- Data often contains missing values, outliers, and inconsistencies that need to be addressed. Missing values can be filled in through imputation or removed if necessary. Outliers can be adjusted or flagged.

- Importance: Cleaning data ensures that the model is not influenced by errors or outliers, leading to more robust and accurate predictions.

**Code:**

# Data cleaning and handling missing values

print(df.drop\_duplicates())

print(df.dropna())

print(df.isna().any())

print(df.isna().sum())

cf=df.copy()

**III) Feature Selection:**

- In demand prediction, there can be numerous features (attributes) associated with table code,state code,district code,area name rural or urban area,age group ,worked period,industrial catogory and more. Feature selection involves identifying the most relevant features for prediction.

- Importance: Selecting the right features reduces model complexity, improves performance, and speeds up training.

**Code:**

# Feature selection

x=df[["Age group","work period"]]

y=df["Units Sold"]

**IV) Train-Validation-Test Split:**

- The data is typically split into three sets: training, validation, and test sets. The training set is used to train the model, the validation set helps tune hyperparameters, and the test set evaluates the model's performance. The data is divided into training, validation, and test sets, typically with a split like 70% training, 15% validation, and 15% test. This split is crucial for model development and evaluation.

- Importance: Separating data into different sets allows for unbiased model evaluation and helps prevent overfitting. Proper data splitting ensures that the model can be trained, tuned, and evaluated effectively without introducing bias.

**Code:**

# Data splitting

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.25)

**V) Feature Scaling:**

- Numerical features may have different scales, which can affect the performance of some machine learning algorithms. Feature scaling, using methods like Min-Max scaling or standardization, helps to normalize feature values.

- Importance: Feature scaling ensures that all features contribute equally to the model's learning process, preventing any one feature from dominating.

**Code:**

# Feature scaling

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

x\_train\_scaled = scaler.fit\_transform(x\_train)

x\_test\_scaled = scaler.transform(x\_test)

**Conclusion:**

In summary, preprocessing is a critical phase in TN Marginal Workers Assessment with machine learning. Proper data preparation ensures that the model is trained on high-quality, relevant data, which ultimately results in more accurate predictions. Each preprocessing step plays a crucial role in addressing different data challenges and improving the model's ability to forecast marginal workers.