Data Preprocessing Documentation

1. Introduction

This document provides an overview of the data preprocessing steps applied to the Customer Churn dataset. The preprocessing includes splitting the dataset into training and testing sets, encoding categorical variables, and scaling numeric features. The goal is to prepare the data for subsequent analysis or modeling.

2. Code Explanation

The following code performs the preprocessing tasks:

```
# Load the dataset
df = pd.read_csv(r'D:\WIL PROGRAM\Customer_Churn_data.csv')

# Splitting the data into 80% training and 20% testing
train_set, test_set = train_test_split(df, test_size=0.2, random_state=42)

# Perform one-hot encoding (Encoding Categorical Variables)
train_set_encoded = pd.get_dummies(train_set, drop_first=True)
test_set_encoded = pd.get_dummies(test_set, drop_first=True)

# Scale numeric features
numeric_features = ['tenure', 'MonthlyCharges']
scaler = StandardScaler()
train_set_encoded[numeric_features] = scaler.fit_transform(train_set_encoded[numeric_features])
test_set_encoded[numeric_features] = scaler.transform(test_set_encoded[numeric_features])
```

```
# Optionally, save the preprocessed datasets to CSV files

train_set_encoded.to_csv('train_set_encoded.csv', index=False)

test_set_encoded.to_csv('test_set_encoded.csv', index=False)
```

3. Scaling Technique

The StandardScaler is used to standardize numeric features. It transforms the data such that each feature has a mean of 0 and a standard deviation of 1. This helps in improving the performance and stability of machine learning algorithms. The scaler is fitted on the training data and then applied to both the training and test datasets to ensure consistency.

4. Benefits

Scaling ensures that numeric features contribute equally to the model's learning process, preventing features with larger ranges from dominating the results. It improves the performance and convergence of many machine learning algorithms, particularly those that use distance metrics, such as K-Means clustering and gradient-based methods.