Scaling Techniques Documentation

In this project, Scaling is a crucial step in data preprocessing for machine learning models, especially for algorithms that rely on distance-based calculations (e.g., K-Means, logistic regression). Scaling ensures that all numerical features are in the same range, which helps improve model performance and convergence.

We applied **StandardScaler** to standardize the numerical features in the dataset. StandardScaler subtracts the mean and scales the data to unit variance. This method ensures that each feature has a mean of 0 and a standard deviation of 1.

Mention which features were scaled. For example:

In our dataset, we applied StandardScaler to the following features:

MonthlyCharges

Tenure

**Code Snippet** in python:



StandardScaler was chosen because it standardizes the features by removing the mean and scaling to unit variance, which is appropriate for many machine learning algorithms, especially those based on distance measurements.

The reason why scale is important is it ensures that no feature dominates others due to differences in magnitude. Algorithms such as K-Means, logistic regression, and neural networks are sensitive to the range of features. It improves model convergence during training.

Other scaling techniques such as MinMaxScaler (scaling to a fixed range like 0-1) and RobustScaler (which is robust to outliers) were considered but not applied here. Since the features MonthlyCharges and tenure did not have extreme outliers, StandardScaler was the most appropriate technique.

In this project, we applied **StandardScaler** to standardize the MonthlyCharges and tenure features. This step ensures that these numerical features are scaled appropriately, contributing to the improved performance of subsequent machine learning models.