Scaling Techniques Documentation

Stage 2 – Data Preparation Deliverable

1. Purpose of Scaling

Scaling was applied to the dataset to ensure that all numerical features contribute equally to the predictive model. Since machine learning models—especially **Artificial Neural Networks (ANNs)**—are sensitive to feature magnitude, scaling helps:

- Improve model convergence
- Prevent dominance of large-scale variables
- Enhance overall predictive performance

2. Scaling Technique Used

The **StandardScaler** technique from the scikit-learn library was applied.

StandardScaler standardises numerical features by transforming them to have:

- Mean = 0
- Standard Deviation = 1

This method is well suited for ANN-based models, which assume normally distributed input features.

3. Features Scaled

Scaling was applied **only to numerical features**, including (but not limited to):

- Tenure
- Monthly Charges
- Total Charges
- Usage-related numerical attributes

Categorical variables were **encoded prior to scaling** and not directly scaled.

4. Scaling Process

To avoid data leakage:

- The scaler was **fitted only on the training dataset**
- The same scaler was then **applied to both training and testing datasets**

This ensures the testing data remains unseen during training.

5. Code Snippet – Scaling Implementation

from sklearn.preprocessing import StandardScaler

```
# Initialize the scaler

scaler = StandardScaler()

# Fit scaler on training data and transform

X_train_scaled = scaler.fit_transform(X_train)

# Apply the same scaler to testing data

X_test_scaled = scaler.transform(X_test)
```

6. Output of Scaling

After scaling:

- All numerical features are on a comparable scale
- Feature distributions are centred around zero
- Variance is normalised across features

This significantly improves:

- ANN training stability
- Gradient descent optimisation
- Model accuracy and recall performance

7. Files Generated

The scaled data was used to produce:

- training_set.csv
- testing_set.csv

Both files contain:

- Encoded categorical variables
- Scaled numerical features
- Target variable (Churn)

8. Justification for Technique Selection

StandardScaler was chosen because:

- It performs well with neural networks
- It preserves relationships between values
- It is widely accepted in predictive analytics workflows

9. Conclusion

The applied scaling technique ensures that the dataset is fully prepared for predictive modelling. By standardising numerical attributes and preventing data leakage, the data preparation process meets all **Stage 2 ACS requirements** and supports accurate churn prediction.