Weekly Report

# 1. Introduction

This report highlights the feature scaling strategy applied to a structured dataset in the context of a data preprocessing pipeline. Feature scaling is a critical preprocessing step, particularly in machine learning workflows that use algorithms sensitive to the scale of features.

# 2. Objective of Feature Scaling

The objective of feature scaling is to normalize the numerical data to a common scale without distorting the differences in the ranges of values. This step ensures that no single feature dominates the learning algorithm due to its magnitude.

# 3. Scaling Technique Used

The technique used in the code is `StandardScaler` from the `sklearn.preprocessing` library. StandardScaler standardizes the features by removing the mean and scaling to unit variance, applying the following transformation:  
  
 z = (x - μ) / σ  
where:  
• x is the original value  
• μ is the mean of the feature  
• σ is the standard deviation

Only numeric features were selected for scaling. The target variable 'RSI' was excluded from the scaling and later merged back to the scaled dataset for further analysis. This is a common practice to avoid data leakage during model training.

# 4. Code Snippet

Below is the exact Python code used for scaling:  
  
from sklearn.preprocessing import StandardScaler  
features = df.drop(columns=['RSI'])  
numeric\_features = features.select\_dtypes(include=['number'])  
scaler = StandardScaler()  
scaled\_numeric = scaler.fit\_transform(numeric\_features)  
scaled\_df = pd.DataFrame(scaled\_numeric, columns=numeric\_features.columns)  
scaled\_df['RSI'] = df['RSI']

# 5. Outcomes and Insights

After applying feature scaling, correlation analysis was conducted between scaled features and the target variable 'RSI'. This step revealed the most influential predictors, assisting in feature selection for downstream modeling.

# 6. Future Work

The next step in this project will involve model development using the scaled dataset. Potential models to be explored include Linear Regression, Decision Trees, and more advanced techniques like Random Forests and Gradient Boosting. Hyperparameter tuning and cross-validation will be implemented to ensure optimal performance and generalization.

Additionally, pipelines will be developed to automate preprocessing and model training, ensuring scalability and reproducibility of the experiments.