Task: Supervised Learning on European Topology 6 Paths Dataset

1. Data Exploration and Preprocessing

Dataset Overview:

The dataset, sourced from '/content/DataSet_EU_3k_5k.xlsx', was loaded into Pandas for initial analysis. It consists of both numerical and categorical features related to European Topology 6 Paths.

Handling Missing Values:

Missing values were imputed with the mean of each respective column to ensure completeness of the dataset.

Feature Scaling:

Numerical features were normalized to zero mean and unit variance using standard scaling techniques to mitigate scale differences among features.

Visualization:

Visual exploratory data analysis (EDA) was performed to understand relationships between various features and the target variable 'GSNR_1'. Scatter plots were generated for key features like 'Power_1', 'NLI_1', 'ASE_1', 'frequency_1', 'No. Spans', and 'Total Distance(m)' to visualize their correlations with 'GSNR_1'.

2. Feature Engineering

Feature Creation:

No new features were explicitly created, as the focus was primarily on leveraging existing features.

Feature Importance:

Feature importance analysis was conducted, particularly using Random Forest models, to identify which features significantly influence the prediction of 'GSNR_1'. This analysis aids in understanding which features contribute most to the model's performance.

3. Feature Selection

Methodology:

SelectKBest method with f_regression scoring was applied to select the top features based on their correlation with the target variable. This step ensures that only the most relevant features are used for model training, potentially improving both accuracy and computational efficiency.

Feature Discard:

Less significant features, including the 'frequency_1' feature, were discarded from the final model inputs to streamline model complexity and enhance interpretability.

4. Model Selection and Training

Implemented Models:

- **Linear Regression:** Utilized as a baseline model for predicting 'GSNR_1'.
- **Decision Trees:** Applied for its ability to handle non-linear relationships and feature interactions.
- **Random Forest:** Employed for ensemble learning to improve robustness and generalization.
- **Gradient Boosting:** Utilized to enhance predictive performance through iterative model training.

Training and Evaluation:

Data was split into training and testing sets to evaluate each model's performance. Cross-validation techniques were employed to ensure reliable performance metrics.

5. Model Evaluation

Performance Metrics:

- **Mean Squared Error (MSE):** Quantifies the average squared difference between predicted and actual values.
- **R-squared (R2):** Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Mean Absolute Error (MAE):** Represents the average absolute difference between predicted and actual values.

Comparative Analysis:

Models were evaluated based on these metrics to identify the best-performing algorithm for predicting 'GSNR_1'. Results were compared across models to determine strengths and weaknesses in predictive accuracy and computational efficiency.

6. Reporting

Analysis Summary:

- Detailed insights into model performance and feature importance were provided.
- Visualizations such as scatter plots and bar charts were used to illustrate key findings.
- Recommendations were made based on the analysis to optimize model parameters and improve predictive accuracy.

7. Conclusion

Summary of Findings:

- **Best Performing Model:** Random Forest demonstrated superior performance in predicting 'GSNR_1', achieving the lowest MSE and highest R2 score among the models evaluated.
- **Feature Insights:** 'Power_1', 'NLI_1', 'ASE_1', 'No. Spans', and 'Total Distance(m)' were identified as critical features influencing 'GSNR_1' predictions.
- **Recommendations:** Further fine-tuning of model parameters and exploration of advanced ensemble techniques could potentially enhance model performance.