3D Position Estimation using WiFi Fingerprinting

Dataset Description

This dataset contains measurements of WiFi signal strengths (RSSI - Received Signal Strength Indicator) collected at various known locations inside a building. Each row represents a single scan or sample point taken at a specific indoor location and includes:

Data Type	Description
*_signal (100 features)	RSSI values from up to 100 distinct WiFi Access Points (APs), in dBm (typically negative values, e.g., -45 to -100). A value of 0 may indicate no signal received from that AP.
x, y, z	True 3D coordinates of the sample location (in meters). These are the **regression targets**.
Additional features	Other metadata such as mean, std, min_db, max_db, Floor, etc., which can be ignored for this task.

- **Total samples:** 1984
- **Signal features:** 100 (0_signal to 99_signal)
- **Target variables:** x, y, z (3D location)

Objective: Your goal is to **predict the 3D coordinates (x, y, z)** of a device using only the available WiFi signal strength values as input features.

Evaluation Metric: Model performance will be evaluated based on the mean Euclidean distance between the predicted and actual locations:

$$\boldsymbol{E = \frac{\sum_{i=1}^{N} \sqrt{(xi - \widehat{xi})^2 + (yi - \widehat{yi})^2 + (zi - \widehat{zi})^2}}{N}}$$

Hints:

Preprocessing

- RSSI values are usually negative; higher values (e.g., -40 dBm) indicate stronger signals.
- Signal values of 0 may indicate missing data (no signal received).
- Try filtering out APs (features) that are rarely active or always zero.
- Normalize data using StandardScaler or MinMaxScaler before model training.
- Use tools like SelectKBest, RFE, or model-based feature importance (RandomForest.feature_importances_) for ranking.
- Handle missing or zero-valued signal features appropriately.
- Apply normalization or standardization to signal features.

Model Training

- Use one or more regression models to predict the x, y, and z values.
- Suggested models: Linear Regression, KNN, Random Forest, XGBoost, etc.

Feature Selection / Ranking

- Use **feature importance or ranking** techniques to identify the most relevant signal features.
- Your final solution should aim to achieve the **lowest possible Euclidean error with the minimum number of features**.
- **Do not use dimensionality reduction techniques such as PCA or UMAP.**

Performance Reporting

- Report mean Euclidean distance error on both the training and test sets.
- Clearly state the number of features used in your final model.

Visualization (Optional but Recommended)

- Visualize predicted vs actual 3D coordinates (scatter plots).
- Plot feature importance scores.