5G Wave Propagation Analysis & Predictive Modeling Report

1. Executive Summary

This project analyzes a 5G Wave dataset to understand the environmental and spatial factors influencing signal strength) `Received_Power_dBm`. (Through iterative exploration and machine learning, we developed a "Pruned" predictive model that achieves over **98 accuracy**) $R^2 = 0.9870$ (using only the top 5 most significant features).

2. Phase I: Exploratory Data Analysis (EDA)

2.1Distribution of LOS vs. NLOS

We began by analyzing the `LOS_NLOS_Flag`.

- **Logic :**Understanding if the data is balanced is critical for model bias.
- **Findings :**The dataset is nearly balanced (~52% NLOS, ~48% LOS).
- **Visualization Strategy :**A bar chart was used with labels mapped as **0 NLOS** and **1LOS** to improve interpretability.

2.2Distance and Shadowing Distributions

- **Distance :**Histogram analysis showed a spread up to 120 meters. We identified this as the primary driver of path loss.
- **Shadowing :**Histogram and Boxplots showed the data is centered at ~0 dB but contains significant outliers (-11 dB to +10 dB). These represent sudden environmental obstructions.

2.3Condition-Based Power Analysis

We compared `Received_Power_dBm` across LOS/NLOS conditions using boxplots relative to the global mean -93.47 dBm.

Conclusion : While LOS (1) has higher power peaks, the medians are surprisingly similar, suggesting that 5G reflections (multipath) keep NLOS signals viable.

3. Phase II: Correlation & Feature Engineering

3.1The Correlation Heatmap

We used a Pearson correlation matrix to quantify relationships:

- **Strong Negative :**`Distance_m` (-0.793)- proving path loss is the dominant factor.
- **Moderate Positive :**`Shadowing_dB` $(+0.543)$- identifying the primary source of signal variance.
- **Weak/Near-Zero :**`Humidity` ,`Temperature` ,and `Blockage_Events`.
- **Interpretation :**We concluded that weather factors in this specific dataset are "noise" rather than signal drivers.

3.2Scatter Analysis & Path Loss Trend

We plotted a scatter of **Distance vs. Power** and calculated a linear regression line:

- **Equation**: $y = -0.1910x - 83.45$
- **Finding :**For every 1-meter increase, power drops by 0.19 dBm. The "spread" around the line confirmed the need for non-linear modeling to capture shadowing effects.

4. Phase III: Machine Learning Development

4.1Initial Model: Random Forest

We chose Random Forest to handle the non-linear relationships and interactions between spatial coordinates.

- **Performance**: $R^2 = 0.9776$
- **Insight :**Distance and Shadowing accounted for nearly all "Feature Importance".

4.2 Optimized Model: Gradient Boosting

To improve accuracy, we moved to Boosting (Gradient Boosting/XGBoost logic), which learns sequentially from errors.

- **Performance**: R^2 improved to 0.9848
- **MAE**: 0.34 dBm.

5. Phase IV: Feature Engineering & Pruning

5.1 The Pruning Strategy

We asked :*Can we achieve the same results with fewer variables?* We removed "noisy" features (Humidity, Temperature, Blockage, etc.) and kept only the Top 5.

- **The Selected Top 5 :**`Distance_m` ,`Shadowing_dB` ,`Rx_Position_y` , `Tx_Position_y` ,and `Tx_Position_z`.

5.2 Results of the Pruned Model

Surprisingly, pruning the model **increased** performance:

- **Final R2 Score**: 0.9870
- **Final MAE**: 0.3229 $\boldsymbol{dBm}$
- **Conclusion :**By removing irrelevant environmental data, we reduced "overfitting" and created a more robust, efficient model for real-time 5G signal prediction.

6. Technical Takeaways for 5G Deployment

1. **Distance Management :**The single most critical factor; 5G cells must be carefully spaced (approx. 86m radius for -100dBm threshold).
2. **Shadowing Resilience :**Network planning must include a "Fade Margin" of at least 10 dB to account for the shadowing outliers identified in the boxplots.
3. **Model Utility :**This model can now be used as a digital twin to predict power levels at any coordinate x, y, z without performing manual field tests.