**Project Proposal**

*Sai Teja Patibandla - U00876137*

*Naveen Mogallapu - U00893266*

**Problem/Application: House Rent Prediction System**

The problem we aim to address in this project is the prediction of house rent prices. Housing prices have been on the rise, impacting the general public and real estate developers. Accurate price prediction is essential for making informed decisions about investments in the real estate market. We will employ regression models to optimize prediction accuracy based on various housing characteristics. This project seeks to provide a valuable tool for both the public and developers to forecast housing values effectively.

**Methods to be Tested/Implemented:**

**Prediction Models:**

**Random Forest:** We will implement a Random Forest regression model, which leverages multiple decision trees and their ensemble effect to improve prediction accuracy. This method is known for its robustness and ability to handle complex datasets.

**Lasso Regression:** Lasso Regression will be employed to select the most relevant features and reduce model complexity. It helps in preventing overfitting and provides a simpler, more interpretable model.

**XGBoost:** Extreme Gradient Boosting (XGBoost) is a powerful machine learning algorithm known for its efficiency and scalability. It will be used to enhance prediction accuracy by iteratively improving model performance.

**Data Cleaning and Pre-processing:**

Before proceeding with analysis, it is preferred that the loaded data undergoes thorough cleaning and pre-processing. This involves addressing missing values and handling outliers. A model trained on high-quality, meticulously cleaned data is expected to yield superior performance compared to one trained on data of lower quality.

**Exploratory Data Analysis (EDA):**

During this phase, we will analyze the data to uncover valuable insights regarding feature patterns and relationships. Moreover, data visualization will play a pivotal role in our data analysis, enhancing our understanding of the dataset.

**Data Sets to be Used:**

<https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset?select=House_Rent_Dataset.csv>

**Potential Challenges for Implementation:**

* Handling Outliers: Managing outliers poses a potential challenge, especially considering the dataset's size. It is possible that certain columns may contain a significantly higher number of outliers compared to others within the dataset.
* Visualization Challenges: Visualizing large datasets in a meaningful way can be difficult. Using correct already available Python libraries can help overcome this.

**Expected Deliverables:**

* Exploratory Data Analysis (EDA) to understand the dataset and identify patterns.
* Visualizations and graphical representations for data insights.
* A regression model for estimating house rent prices in India.
* Implementation of the Random Forest, Lasso Regression, and XGBoost algorithms.
* Evaluation metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RSME), and R-squared (R2) for model performance assessment.