**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.

**Extension of the Project** can be done by implementing more prediction models like SVR etc. and implementing advanced algorithms in the data mining field like KNN and Gaussian Naive Bayes which are discussed in the class and implementing Advanced Feature engineering methods which make use of data in implementing features like price per square foot. Can make use of GUI which will provide an interface for users to get rent prediction and can use another Dataset to make the problem bigger like the average income of a person which will influence choosing a House for rent

**Prediction Models:**

**Neural Networks:** Deep learning models, especially neural networks, can capture intricate patterns in data. We'll implement a simple feed-forward neural network using frameworks like TensorFlow or PyTorch.

**Support Vector Regression (SVR):** SVR is effective in handling non-linear relationships. It can provide another perspective compared to tree-based models.

**K-Nearest Neighbors (KNN) for House Rent Prediction:**

K-Nearest Neighbors (KNN) is a straightforward yet potent algorithm used extensively in regression problems such as predicting house rents. In the context of house rent prediction, KNN works by identifying the K nearest houses based on their features, such as location, number of bedrooms, or square footage. The rent of the new house (for which a prediction is needed) is estimated based on the average or median rent of these K nearest houses. The underlying assumption is that similar houses (in terms of features) in the dataset will have similar rents.

**Gaussian Naive Bayes for House Rent Prediction:**

Gaussian Naive Bayes is a probabilistic algorithm typically used in regression and classification tasks. For predicting house rents, this algorithm calculates the likelihood of a particular house's features given its rent, using the Gaussian distribution. By integrating the probabilities associated with all features of a house (e.g., location, number of rooms, amenities), it computes the probable rent range. The Gaussian Naive Bayes model is particularly useful when house features have a Gaussian (or near-Gaussian) distribution and are relatively independent. For instance, the distribution of house sizes in square feet or the distribution of houses in various neighbourhoods might be assumed to follow a Gaussian pattern. Using these distributions, the model can then predict rent values for new or unlisted houses.

**Feature Engineering:**

Derived features like price per square foot, distance to the city center, etc. could be beneficial.

Using clustering algorithms (like KMeans) to create location-based clusters as features.

* Implementing advanced features like Interaction Features: Combine two or more features. For example, bedrooms \* and bathrooms might capture the effect of the total number of rooms.

**GUI:**

Comprehensive GUI: Develop a user-friendly interface where users can input house features and get a predicted rent.

Historical Data: Allow users to view past rental trends based on regions or house features.

Demographic Data: Integrate data about the average income, population density, or employment rates of the property's location.

**Responsibility:**

Combined: researching python libraries that will make Data Mining algorithms in Python easy

Naveen: Data cleaning, Pre-processing Data, Prediction models, python implementation, GUI

SaiTeja: EDA KNN, Bayes, Neural networks, SVR, Testing, python implementation of KNN and Bayes