# Predicting House Prices with Regression Analysis

**Name: Induja Rajendrakumar**

**Programme: MSc Data Science**

**Student ID: 23047729**

**Module: UFCF9Y-60-M – Csct Masters Project**

## Background and Problem Statement:

Housing prices are a crucial indicator of economic health and accurately predicting them is of great interest to both buyers and sellers. Traditional methods, such as linear regression, often fail to capture the complex, non-linear relationships present in housing data, leading to significant inaccuracies. The advent of machine learning, specifically models like XGBoost and ensemble techniques, offers a significant improvement in predictive accuracy by effectively handling these complexities. This project explores the development and application of advanced machine learning models, including a hybrid model that combines Ridge Regression and XGBoost, to enhance the accuracy of house price predictions, addressing the limitations of traditional methods and providing a more reliable solution for real estate analytics.

## Aim and Scope:

The aim of this project is to develop a robust predictive model for house prices by integrating traditional regression techniques with advanced machine learning algorithms. Specifically, the project will focus on creating a hybrid model that combines Ridge Regression with XGBoost, leveraging their respective strengths in regularization and handling non-linear relationships. An ensemble approach will also be explored to further enhance prediction accuracy. Additionally, the project aims to develop a Streamlit application to facilitate the interactive exploration of the dataset, model training, and visualization of results, making the predictive models accessible and practical for real-world application. This scope will ensure the creation of a tool that can be utilized by real estate professionals to improve decision-making processes.

## Project Objectives:

**Predict Sale Prices:** Estimate the sale price for each house in the test set by generating predictions using advanced machine learning models, including Ridge Regression, XGBoost, and an ensemble model that combines the strengths of these techniques to improve robustness and accuracy.

**Model Implementation and Comparison:** Develop and compare multiple models, including Ridge Regression, Lasso, Decision Trees, Random Forest, XGBoost, and an ensemble model, evaluating their performance using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²). The comparison will provide insights into the effectiveness of each model in capturing the complexities of the housing data.

**Visualization Techniques:** Employ various visualization techniques, including scatter plots, histograms, box plots, heatmaps, and feature importance plots, to analyze, interpret, and communicate the findings effectively. These visualizations will be critical in understanding the underlying data patterns and model behaviors.

**Streamlit Application Development:** Create a Streamlit application that facilitates interactive exploration of the dataset, model training, evaluation, and visualization of results, providing a user-friendly interface for engaging with the data and predictive models. This application will allow users to interact with the model outputs and explore different scenarios dynamically.

## Project Outcomes:

The project aims to develop an accurate and reliable predictive model for house prices by integrating traditional regression methods with advanced machine learning algorithms. The final output will include a hybrid model combining Ridge Regression and XGBoost, and an ensemble model that provides the most accurate predictions. Additionally, a Streamlit application will be developed to allow real estate professionals and stakeholders to interact with the predictive models, visualize results, and generate predictions in a user-friendly environment. This tool will enable better decision-making by providing transparent and interpretable predictions that account for various factors influencing house prices.

## Literature Review:

The literature review in this project covers the evolution of housing price prediction models, starting from traditional regression techniques to advanced machine learning approaches. Madhuri, Anuradha, and Pujitha (2019) provided insights into various regression models for house price prediction, highlighting their strengths and limitations in different scenarios. Hassanpour (2019) emphasized the importance of feature engineering techniques in enhancing model performance, demonstrating how preprocessing steps like handling missing values and encoding categorical variables can significantly improve prediction accuracy. Arumugam et al. (2022) conducted a comparative analysis of regression models, concluding that advanced techniques like Ridge and Lasso regression outperform basic linear models in predictive accuracy. Spiekermann et al. (2020) discussed the ethical considerations in algorithmic decision-making for housing, stressing the need for fairness and transparency in predictive models. De Cock (2011) introduced the Ames Iowa dataset as a comprehensive alternative to the Boston housing data, providing a robust dataset for evaluating and benchmarking regression models.

## Testing And Evaluation:

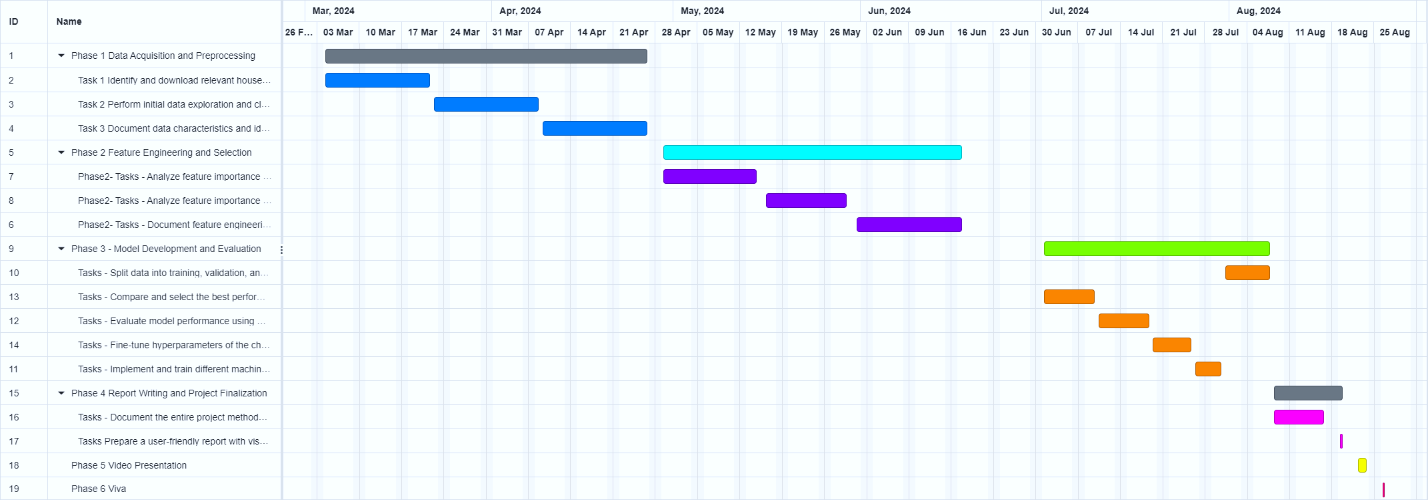
The house price forecasting models will be rigorously evaluated using a comprehensive approach. The dataset will be divided into training and testing sets to ensure the generalizability of the models. Performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²) will be used to assess the accuracy of each model. A comparison will be made between the models developed, including the hybrid Ridge-XGBoost model and the ensemble approach, to determine the most effective method for predicting house prices. Feature importance will also be analyzed to understand the contribution of different variables to the model's predictions, ensuring that the models provide interpretable and actionable insights for users.

## Ethical Considerations:

This project will address ethical considerations related to data handling, algorithmic fairness, and transparency in predictive modeling. The dataset used is publicly available, and all steps will be taken to ensure that the models developed are fair and unbiased. The project will also focus on making the models transparent and interpretable, providing users with insights into how predictions are made and ensuring that the models are used responsibly in real estate decision-making. Special attention will be given to mitigating the risk of algorithmic bias, particularly in how different demographic and geographic factors are handled in the predictive models.

## Project Plan:

Gantt Chart:



## References:

* Arumugam, S.R., et al., 2022. Comparative Analysis of Regression Models for House Price Prediction. *Journal of Property Research*, 38(1), pp.48-70.
* Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle. <https://kaggle.com/competitions/house-prices-advanced-regression-techniques>
* Competition, K., 2020. House Prices-Advanced Regression Techniques. <https://www.kaggle.com/code/juliencs/a-study-on-regression-applied-to-the-ames-dataset>
* Rawool, A.G., Rogye, D.V., Rane, S.G. and Bharadi, V.A., 2021. House price prediction using machine learning. *Int. J. Res. Appl. Sci. Eng. Technol*, *9*, pp.686-692. <https://www.irejournals.com/paper-details/1702692>
* Rana, V.S., Mondal, J., Sharma, A. and Kashyap, I., 2020, December. House Price Prediction Using Optimal Regression Techniques. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 203-208). IEEE - <https://www.researchgate.net/publication/349802688_House_Price_Prediction_Using_Optimal_Regression_Techniques>
* Pandiri, V.S., 2017. Machine Learning Models to Predict House Prices based on Home Features - <https://scholarworks.calstate.edu/downloads/fx719m836>
* Ho, W.K., Tang, B.S. and Wong, S.W., 2021. Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), pp.48-70 - <https://www.researchgate.net/publication/346308101_Predicting_property_prices_with_machine_learning_algorithms>
* Özöğür Akyüz, S., Eygi Erdogan, B., Yıldız, Ö. and Karadayı Ataş, P., 2023. A novel hybrid house price prediction model. *Computational Economics*, *62*(3), pp.1215-1232 - <https://www.researchgate.net/publication/363613155_A_Novel_Hybrid_House_Price_Prediction_Model>

Signature

