Project Title:  
Name:

Abstract:

In this project, we aimed to analyze and predict house prices using a dataset containing various housing features. The primary goal was to build a predictive model that could estimate house prices based on the given features, which can be valuable for real estate professionals, homeowners, and potential home buyers. The project involved several steps, including exploratory data analysis, data preprocessing, feature engineering and selection, model selection and implementation, model evaluation, feature importance analysis, and model interpretation.

We implemented and trained multiple machine learning algorithms, including Linear Regression, Decision Trees, and Random Forest, and fine-tuned their hyperparameters to achieve the best performance. Based on the evaluation metrics and cross-validation results, the Tuned Random Forest model emerged as the best-performing model. The analysis of feature importance revealed that living area, overall quality, and total square footage of the house were among the most important factors influencing house prices.

The insights gained from this project can help better understand the factors that impact house prices and enable stakeholders to make informed decisions. Furthermore, the project identified potential areas for improvement, such as feature engineering, feature selection, hyperparameter tuning, and exploring alternative machine learning models, which can be considered for refining the model and achieving better performance in the future.

ML Problem specification:

In this project, we focused on a supervised learning problem: predicting house prices using a dataset of various housing features. The dataset contained both numerical and categorical features describing different aspects of the houses, such as the size, location, and quality of materials used. The target variable was 'SalePrice', which represented the property's sale price in dollars.

The dataset was split into a training set and a test set, allowing us to train and evaluate the performance of our machine learning models. We performed exploratory data analysis, data preprocessing, feature engineering, and feature selection to prepare the data for modeling.

During the preprocessing phase, we handled missing values, encoded categorical variables, and standardized numerical features as needed. We also created new features and applied correlation analysis to select the most relevant ones for our models.

To solve this supervised learning problem, we implemented and trained multiple machine learning algorithms, including Linear Regression, Decision Trees, and Random Forest. We evaluated their performance on the test set using Mean Absolute Error (MAE) as the performance metric and employed cross-validation techniques to ensure the models' generalization to new data.

The project's findings and insights include the identification of the best-performing model (Tuned Random Forest) and the most important factors influencing house prices. The insights can help stakeholders better understand the factors that impact house prices and make informed decisions.

Design and Milestones:  
In this project, we used a housing dataset containing various features describing the properties, including size, location, and quality of materials. The data was structured as a features x observations matrix, with a target column 'SalePrice' representing the property's sale price in dollars.

We used Python and its libraries, including Pandas, NumPy, Seaborn, and scikit-learn, to preprocess and analyze the data. We performed exploratory data analysis to visualize the distribution of house prices and their relationships with other features. During the data preprocessing phase, we handled missing values, encoded categorical variables, and standardized numerical features as needed.

We employed feature engineering to create new features and applied correlation analysis for feature selection, identifying the most relevant features for our models. We then implemented and trained multiple machine learning algorithms, including Linear Regression, Decision Trees, and Random Forest, using the scikit-learn library.

To fine-tune the hyperparameters of the models, we used techniques such as Grid Search and Randomized Search, ensuring optimal performance. We evaluated multiple models and also considered ensemble methods like Random Forest, which combines the output of multiple decision trees.

To obtain reliable performance estimates, we split the dataset into training, validation, and test sets. The training set was used to train the models, while the validation set was utilized for hyperparameter tuning and model selection. Finally, the test set was employed to evaluate the performance of the selected model and report its accuracy using Mean Absolute Error (MAE) as the performance metric.

The project's findings and insights include the identification of the best-performing model (Tuned Random Forest) and the most important factors influencing house prices. The insights can help stakeholders better understand the factors that impact house prices and make informed decisions.