**Description:** In this project, you will work on a regression task to predict house prices using Apache Spark. The dataset contains information about various features of houses and their corresponding sale prices. The data is posted [here](https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data). Your goal is to build a regression model that accurately predicts the sale price of a house based on its features.

**Tasks:**

1. **Data Exploration and Preprocessing:**
   * Load the house price dataset into Spark RDDs or DataFrames.
   * Explore the dataset to understand its structure, features, and distribution.
   * Preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features.
2. **Feature Engineering:**
   * Create additional features from the house data that might be predictive of house prices, such as the age of the house, proximity to amenities, or neighborhood characteristics.
   * Perform feature selection to identify the most relevant features for modeling.
3. **Model Development:**
   * Split the dataset into training and testing sets.
   * Build regression models using machine learning algorithms supported by Spark, such as linear regression, decision trees, random forests, or gradient-boosted trees.
   * Tune model hyperparameters using techniques like grid search or random search to improve performance.
4. **Model Evaluation:**
   * Evaluate the performance of each model using appropriate metrics (e.g., mean squared error, mean absolute error, R-squared) on the test dataset.
   * Compare the performance of different models and select the best-performing one.
5. **Model Interpretation:**
   * Interpret the trained model to understand the factors influencing house prices.
   * Analyze feature importance to identify the most significant predictors of house prices.
6. **Performance Optimization:**
   * Optimize Spark job configurations and parallelism settings for better scalability and efficiency.
   * Experiment with different optimization techniques to improve model training time and resource utilization.
7. **Reporting and Visualization:**
   * Summarize the findings and insights from the house price prediction analysis in a comprehensive report.
   * Create visualizations (e.g., scatter plots, regression plots) to illustrate key findings and model performance.

**Deliverables:**

1. Jupyter notebook or Python script containing the Spark code for data preprocessing, modeling, and evaluation.
2. Report documenting the entire house price prediction process, including data exploration, feature engineering, model development, evaluation results, and optimization strategies.
3. Visualizations generated during the analysis, preferably in formats suitable for presentation.
4. Presentation slides summarizing the key findings, insights, and recommendations for predicting house prices.

**Additional Notes:**

* Ensure that the project follows best practices in data analysis, including proper documentation, code readability, and reproducibility.
* Experiment with different regression algorithms and hyperparameter settings to find the most effective approach for house price prediction.
* Consider the business context and market trends when interpreting model results and proposing recommendations.
* Discuss with team members or peers to discuss ideas, share insights, and validate findings.