**Capstone Project for Data Science - Tinyuka Session**

**Project Overview**

In this capstone project, you will apply your data science and machine learning skills to develop a predictive model for house prices. This project will involve data cleaning, exploratory data analysis (EDA), feature engineering, model training, and evaluation. You will use a real-world dataset to build a model that can accurately predict house prices based on various features.

**Instruction Prompts**

**Project Introduction**

**Objective:** Develop a machine learning model to predict house prices using a dataset containing various house-related features.

**Dataset:** You will use the "House Prices - Advanced Regression Techniques" dataset from Kaggle (or any other relevant house price dataset).

**Project Phases**

**Phase 1: Data Collection and Preparation**

Task 1.1: Download the dataset from Kaggle or another source.

Task 1.2: Load the dataset into a Pandas DataFrame.

Task 1.3: Inspect the dataset for missing values and handle them appropriately.

Task 1.4: Perform data cleaning to ensure the dataset is ready for analysis.

**Phase 2: Exploratory Data Analysis (EDA)**

Task 2.1: Conduct exploratory data analysis to understand the distribution of features and the target variable (house prices).

Task 2.2: Visualize the relationships between features and the target variable using scatter plots, histograms, and box plots.

Task 2.3: Identify and handle outliers in the dataset.

**Phase 3: Feature Engineering**

Task 3.1: Create new features that may help improve the model's performance.

Task 3.2: Encode categorical variables using one-hot or label encoding techniques.

Task 3.3: Normalize or standardize numerical features as needed.

**Phase 4: Model Training and Evaluation**

Task 4.1: Split the dataset into training and testing sets.

Task 4.2: Choose and justify the selection of machine learning algorithms (e.g., Linear Regression, Decision Tree, Random Forest, Gradient Boosting).

Task 4.3: Train multiple models and evaluate their performance using appropriate metrics (e.g., RMSE, MAE, R²).

Task 4.4: Perform hyperparameter tuning to optimize the model's performance.

Task 4.5: Select and evaluate the best-performing model on the testing set.

**Phase 5: Model Interpretation and Reporting**

Task 5.1: Interpret the results of the best-performing model and explain the importance of critical features.

Task 5.2: Create visualizations to support your findings and model interpretations.

Task 5.3: Write a comprehensive report summarizing the project, including the methodology, results, and conclusions.

**Deliverables**

**Code:** Submit the complete code used for data preparation, EDA, feature engineering, model training, and evaluation (Jupyter Notebook format)

**Report:** Submit a detailed report (PDF or Jupyter Notebook format) documenting your approach, findings, and conclusions. The report should include visualizations and a clear explanation of your steps.

**Presentation:** Prepare a brief presentation (5-10 slides) summarizing your project and key findings. You may be asked to present this to your peers or instructors.

Submit: [AltSchool of Data Science Tinyuka 2023 Capstone Project](https://forms.gle/o4h2nihFqvgZX5dr7)