**Comparison of different ML models**

**Overview**

This project aims to identify the best machine learning model for predicting house prices per unit area using a structured dataset. Multiple regression algorithms were compared to evaluate their performance and accuracy in predicting this specific target variable.

**Dataset**

The dataset was taken from Kaggle and includes information about various factors affecting house prices. These factors comprise house age, distance to the nearest MRT (Mass Rapid Transit) station, the number of convenience stores in the vicinity, house price per unit area, and other relevant features.

**Models Used**

**Linear Regression**: A basic regression model that assumes a linear relationship between the input features and the target variable.

**Decision Tree Regression**: A model that splits the data into branches to make predictions based on decision rules from the features.

**Random Forest Regression**: An ensemble method that combines multiple decision trees to improve predictive performance.

**Gradient Boosting Regression**: An ensemble technique that builds models sequentially to correct the errors of previous models.

**Ensemble methods** are techniques in machine learning that combine the predictions of multiple models to produce a single, more accurate prediction.

### Data Preprocessing

The dataset was split into training and testing sets using an 80-20 split ratio. Standard scaling was applied to the features to ensure they have a mean of 0 and a standard deviation of 1.

**Why standardised X data?**

Standardizing only the feature data (X) is important because it ensures that each feature contributes equally to the model's training. Standardization transforms the data to have a mean of 0 and a standard deviation of 1, making it easier for the model to learn from the data.

**Model Training and Evaluation**

Each model was trained on the scaled training data and evaluated on the test data. The performance metrics used were Mean Absolute Error (MAE) and R² Score.

**Results**

Based on the comparison, Linear Regression emerged as the best-performing model, demonstrating the lowest MAE and the highest R² Score. This indicates that Linear Regression provided the most accurate predictions for the task, surpassing the performance of the other algorithms evaluated.

**Visualisation**

Metrics are visualized using bar graphs. The actual vs. predicted values for each model were plotted to visually assess the performance and residuals.

**Conclusion**

This project provided a comprehensive comparison of various machine learning models for predicting diabetes progression. Linear Regression was identified as the best model for this task, based on the evaluation metrics.