# Artificial Intelligence & Machine Learning – Task 1

## 1 Introduction

- This report documents the completion of Task 1: building a Linear Regression model to predict California housing prices.
- The project utilized the scikit-learn library for the complete machine learning pipeline, from data ingestion to model serialization.

## 2 Exploratory Data Analysis (EDA)

- We analyzed the dataset structure and distributions before modeling.

### 2.1 Distributions and Correlations

- **Target Distribution:**
    - A histogram analysis of the MedianHouseValue revealed a right-skewed distribution with a noticeable cap at the maximum value (5.0), indicating censored data.
- **Correlation:**
    - The heatmap demonstrated that MedInc (Median Income) has the strongest positive correlation with house prices, whereas location-based features like Latitude showed weaker individual correlations.

## 3 Methodology

### 3.1 Data Splitting

- The data was split into training and testing sets using a standard ratio:
    - **Train Size:** 80%
    - **Test Size:** 20%
    - **Random State:** 42 (ensuring reproducibility)

### 3.2 Model Architecture

- A **Linear Regression** algorithm was trained on the dataset.
- This model fits a linear equation to the observed data by minimizing the residual sum of squares between the observed and predicted targets.

# 4 Evaluation Results

## 4.1 Performance Metrics

- The model was evaluated on the unseen test set ($n \approx 4{,}128$ samples).

| Metric | Score |
|---|---|
| Mean Absolute Error (MAE) | 0.5332 |
| Root Mean Squared Error (RMwSE) | 0.7456 |
| R-Squared ($R^2$) | 0.5758 |

Table 1: Final Model Metrics

## 4.2 Feature Importance Analysis

- We extracted the coefficients to understand which features most influenced the predictions.
  - **Highest Positive Impact:**
    - AveBedrms and AveRooms showed significant coefficients, indicating a strong relationship with value.
  - **Negative Impact:**
    - Latitude and Longitude had negative coefficients, reflecting the geographic variance in California housing prices (e.g., inland areas vs. coastal).

# 5 Conclusion & Deliverables

- The model achieved an $R^2$ score of 0.576, successfully establishing a baseline for price prediction.
  - **ModelArtifact:**
    - Thefinaltrainedmodelhasbeenserializedandsavedashouse_price_model.pkl.
  - **Visualizations:**
    - Plots for the correlation heatmap, target distribution, and actual vs. predicted prices were generated to validate the results.