Case Study

Question 1: What type of machine learning problem is this (e.g., regression, classification, clustering)?

Answer: This is a regression problem since the objective is to predict a continuous numerical variable (median house value) based on input features.

Question 2: What is the target variable in this case, and why is it important for prediction?

Answer: The target variable is "median_house_value," and it's essential for prediction because it represents the value, we aim to predict for each house based on the input features.

Question 3: What are some potential sources of data preprocessing that may be required for this dataset?

Answer: Data preprocessing includes handling missing values (using **dropna()**), applying logarithmic transformations to specific columns, encoding categorical variables using one-hot encoding, and creating new features like "bedroom_ratio" and "household_rooms."

Question 4: How would you handle missing values, if any, in this dataset?

Answer: In the presence of missing data, a straightforward approach is implemented: rows with missing values are removed, as guessing the missing values is deemed inappropriate.

Question 5: How can you encode the categorical variable "Neighborhood" for machine learning?

Answer: The categorical variable "ocean_proximity" is encoded using one-hot encoding, converting it into binary columns that the machine learning model can interpret.

Question 6: What are some potential features you could engineer from the existing data to improve model performance?

Answer: Feature engineering possibilities include creating new variables such as "bedroom_ratio" and "household_rooms," which may enhance the model's capacity to capture meaningful relationships within the data.

Question 7: What machine learning algorithms could you consider for this regression problem, and why?

Answer: Three machine learning algorithms were contemplated:

- **Linear Regression:** Selected for its interpretability and simplicity.
- **Random Forest Regressor:** Chosen for its capacity to handle non-linear relationships and outliers effectively.
- **Artificial Neural Network (ANN):** Employed to explore the potential benefits of deep learning, which can capture intricate patterns in the data.

Question 8: How would you split the dataset into training and testing sets? What is the purpose of this split?

Answer: The dataset is divided into training and testing sets using the **train_test_split()** method. This division is essential to train the model on a subset of the data and subsequently evaluate its performance on unseen data, ensuring generalization.

Question 9: What evaluation metrics would you use to assess the performance of your regression model?

Answer: Multiple evaluation metrics are used, including R-squared (R2) score, Mean Squared Error (MSE), and Mean Absolute Error (MAE), to assess the performance of the models.

Question 10: How would you handle outliers in the data, if any?

Answer: While this code does not explicitly address outliers, managing outliers often requires specialized techniques. Handling them would entail further investigation and potential adjustments to the model.

Question 11: Could you briefly outline the steps involved in building and evaluating a machine learning model for predicting house prices?

Answer: The key steps encompass data preprocessing, feature engineering, model selection, hyperparameter tuning, and evaluation. The model's performance is assessed using various metrics, providing an overview of its effectiveness.

Question 12: What are some potential challenges you might encounter when working with this dataset, and how would you address them?

Answer: Challenges may include addressing outliers, selecting relevant features, and optimizing the hyperparameters of the models. Additionally, ensuring that the neural network architecture is well-suited to the data can pose challenges. Addressing these issues would require careful consideration and potentially further experimentation.