Dataset:

We used the "ParisHousing.csv" dataset for this analysis. The dataset contains information about housing prices, including features such as square meters, bedrooms, bathrooms, city code, the year built and many more. It also includes a continuous real-valued label "Price" indicating what the predicted value of a house is based on the inputs given.

Data Preprocessing:

Upon acquiring the data, we initiated preprocessing by excluding non-continuous input features, ensuring a streamlined focus for model creation. For both Naive Bayes Classifier (NBC) and Stochastic Gradient Descent (SGD), missing values were handled by removing rows with any empty columns. The dataset was then randomly split, allocating 20% for testing and 80% for training. This split was consistently applied to both Linear Regression and Ordinary Least Squares (OLS) Linear Regression models. Regression and OLS Linear Regression.

Evaluation Metrics for Linear Regression:

1. Training Data:
   1. MSE: 8276810.25
   2. MAE: 2330.47
   3. R2 Score: 0.99999955
2. Test Data:
   1. MSE: 8254728.47
   2. MAE: 2346.07
   3. R2 Score: 0.99999958

The stellar performance of the Linear Regression model is evident in the high R-squared scores, affirming its robust fit to the training data. The relatively low MAE, especially considering the average price of 4,993,447.53, underscores the model's accuracy. The potential impact of outliers on the seemingly high MSE necessitates further investigation.

**Evaluation Metrics for Ordinary Least Squares Linear Regression (OLS):**

For both OLS and SGD Regressor models, a deep copy of the training and test data was created after a random 20-80% data split. Notably, for OLS, a column of 1’s was added at the beginning of the data, aligning with Assignment 5 instructions.

W Vector:

The input parameters of the model were generally between the range -0.2 to 0.2, while the bias term was consistently around and a little higher than -0.4. showing that the model is fitted closely to the training data (but possibly overfitted).

Evaluation Metrics for OLS Linear Regression:

1. Training Data:
   1. MSE: 8152235.24
   2. MAE: 2313.97
   3. R2 Score: 0.99999956
2. Test Data:
   1. MSE: 8180301.42
   2. MAE: 2332.79
   3. R2 Score: 0.9999995

Similar to Linear Regression, the OLS model boasts a high R-squared score, indicating a good fit to the training data. It's worth noting that the mean absolute error (MAE) is relatively low, especially considering that the average price is 4,993,447.53. The presence of outliers in the features might contribute to the seemingly high MSE, a factor to be further investigated.

In summary, our models exhibit exceptional predictive performance, as evidenced by the outlined evaluation metrics. The high R-squared scores affirm their proficiency, while ongoing exploration into outliers promises to further enhance overall model effectiveness. These insights offer a solid foundation for refining our models and optimizing their predictive capabilities.