A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Choijamts Bayartai |
| *Student Number* | 2024150 |
| *Module Title* | Machine Learning |
| *Assessment Title* | CA1 Project |
| *Assessment Due Date* | 6th March 2024 |
| *Date of Submission* | th April 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Housing Price Prediction Using Machine Learning Models**

**Introduction**

This dataset was generated using fictitious data on housing costs in Paris, an urban setting.

A key component of the economy, the real estate market influences personal financial decisions and reflects overall economic situations. Policy makers, real estate brokers, investors, and homeowners depend on accurate housing price forecasts. Using a dataset of past sales data, this project attempts to use machine learning to forecast property prices based on a range of characteristics, including location, size, and number of rooms.

The goal of this research is to offer a more reliable, data-driven method of estimating prices that can take into account a variety of factors and adjust to shifting market conditions.

**Justification for Using Linear Regression and Random Forest**

1. For this study, a baseline for prediction accuracy was established through the use of linear regression. It works very well when there is a linear relationship between the variables. Stakeholders may quickly comprehend how input parameters like square meters, number of rooms, and accessibility to important facilities impact property pricing because of its high interpretability.
2. Random Forest can handle non-linear connections between features and is strong against overfitting, it was chosen as a supplement to Linear Regression. It is especially helpful in complicated situations where several variables interact to affect the result, as is frequently the case with predict pricing.
3. This project combines the benefits of a more sophisticated ensemble approach with the simplicity of a linear model to provide forecasts that are precise, dependable, and comprehensible by utilizing both Random Forest and Linear Regression. This dual methodology guarantees that the prediction models are both reliable and understandable by enabling cross-validation of results and greater insights into the factors influencing real estate prices.

**Characterization of data**

Investigate the dataset to learn about its characteristics, structure, and goal variable ( Housing price in Paris). This entails looking at feature distributions, descriptive statistics, missing value detection, and using plots like correlation matrices, scatter plots, and histograms to visualize the relationships between variables.

df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 10000 entries, 0 to 9999

Data columns (total 17 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 squareMeters 10000 non-null int64

1 numberOfRooms 10000 non-null int64

2 hasYard 10000 non-null int64

3 hasPool 10000 non-null int64

4 floors 10000 non-null int64

5 cityCode 10000 non-null int64

6 cityPartRange 10000 non-null int64

7 numPrevOwners 10000 non-null int64

8 made 10000 non-null int64

9 isNewBuilt 10000 non-null int64

10 hasStormProtector 10000 non-null int64

11 basement 10000 non-null int64

12 attic 10000 non-null int64

13 garage 10000 non-null int64

14 hasStorageRoom 10000 non-null int64

15 hasGuestRoom 10000 non-null int64

16 price 10000 non-null float64

dtypes: float64(1), int64(16)

memory usage: 1.3 MB

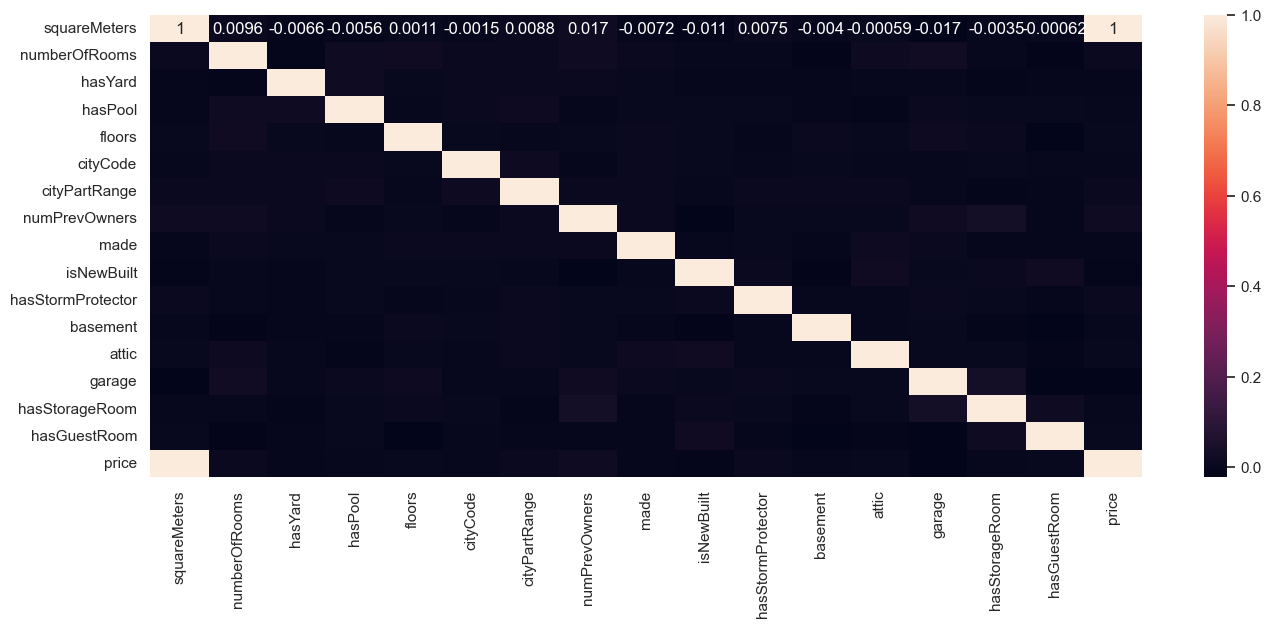
Dataset contains **10000** rows and **17** features (**columns**)

Our target variable (independent variables) is “**price**”

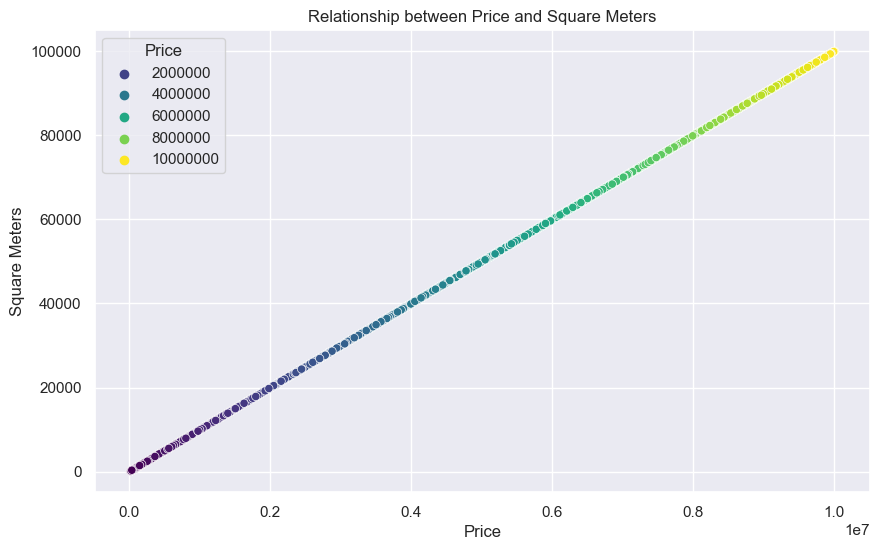
**Explain code**

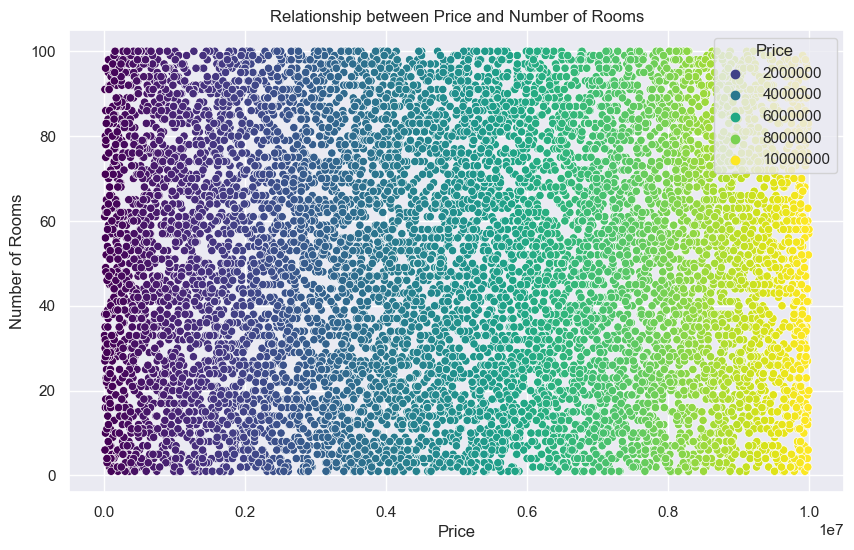
1. Important Libraries Imports
2. Loading Data The code puts data into a Data Frame called “df” from a CSV file called "ParisHousing.csv."
3. Preliminary data analysis .This entails looking over the loaded data to figure out its structure, helping in determining which pre-processing stages are required:

* The DataFrame's initial few rows are displayed by df.head(), giving a brief overview of its columns and values.
* The statistical summaries of the numeric columns and comprehensive information about the data types are provided by df.info() and df.describe().
* df.isnull().Each column's missing values are checked using sum().

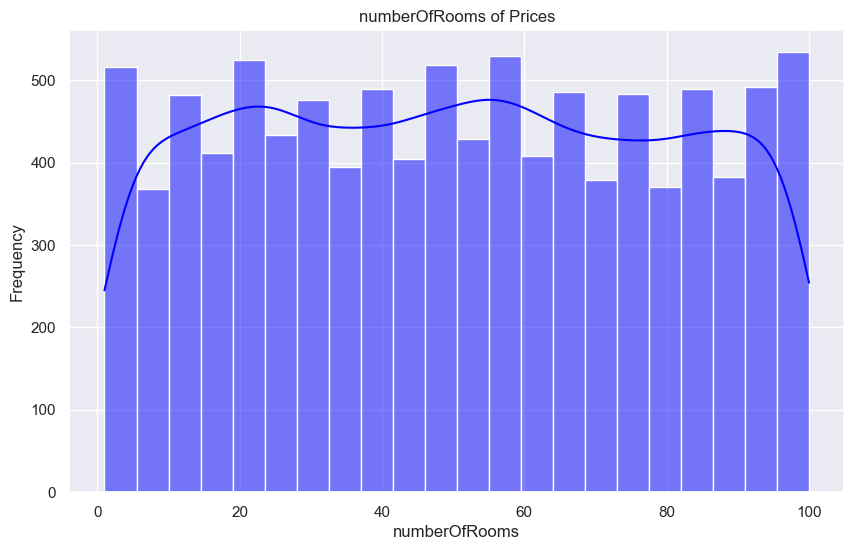
1. Visualization of Data visualizations are used to improve data understanding:  
     
   * + Correlations between numerical features are displayed in heatmaps.
       - * 

* The correlations between pricing and other variables, such as square meters and number of rooms, may be shown with the use of scatter plots.





* The distribution of features such as price and number of rooms is shown via histograms





1. Data is ready for modelling using machine learning:

Encoding of categorical data is suggested but not shown clearly.  
To assess the model's performance on data that hasn't been seen yet, data is divided into training and test sets. For durable model assessment and efficient learning, train\_test\_split with a 25% test size and a fixed random\_state offers an ideal balance between training and testing the model.

1. Training and Evaluating Models  
   In this part, machine learning models are trained and their performance is assessed. Hyperparameter tuning is the act of varying the parameters of the algorithm that governs the training process in order to determine which combination produces the best results based on a certain metric (e.g., RMSE, R2 score). This procedure is essential since not all datasets or certain problems will benefit from the default settings.  
     
   Both a Random Forest Regressor and a Linear Regression model are trained.  
   The Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared scores are used to assess models.  
   To evaluate the resilience of the model, cross-validation is utilized.
2. Model Performance Comparison  
   Lastly, a visual comparison of the models' performances is made:  
     
   The RMSE scores from cross-validation of various models are compared using a bar plot, which gives a clear visual representation of the best performing model.

**Conclusion**

The performance of linear regression is remarkably strong, with an almost perfect R2 score of 1.000000. This suggests that all of the response data variability around its mean is explained by the model. The RMSE and MAE are substantially lower than those of the Random Forest, It appears that Linear Regression offered a more accurate fit. The model is consistent and has good generalization across various data subsets, as seen by the near RMSE values obtained from both cross- validation and the model. The error measures (MAE, MSE, and RMSE) for Random Forest are significantly greater than those for Linear Regression, despite the fact that Random Forest also exhibits a high R2 Score around 1.  
The model fits the data well, however it does not forecast as correctly as Linear Regression, particularly on this dataset, as indicated by the higher values in these measures.

Dataset link : https://www.kaggle.com/datasets/mssmartypants/paris-housing-price-prediction/data

GitHub link : https://github.com/Cho1joo/-CA1---Machine-Learning

**Reference:**

1. Kaushik, S. (2023). *House Price Prediction: A Simple Guide with Scikit-Learn and Linear Regression*. [online] Medium. Available at: https://medium.com/@kaushiksimran827/house-price-prediction-a-simple-guide-with-scikit-learn-and-linear-regression-f91a27b9d650 [Accessed 9 Apr. 2024].
2. paula (2023). *Machine Learning Models for Precise Predictive Analytics*. [online] Stefanini. Available at: https://stefanini.com/en/insights/news/machine-learning-models-for-precise-predictive-analytics#:~:text=Some%20of%20the%20popular%20machine.