**PREDICTING HOUSE PRICE USING MACHINE LEARNING**

# Definition:

The machine learning model is given the test data but without the price of the properties in order to predict the price for them given the various features for the properties. The predicted price is then compared to the actual price in the test data.

# Following Requirements:

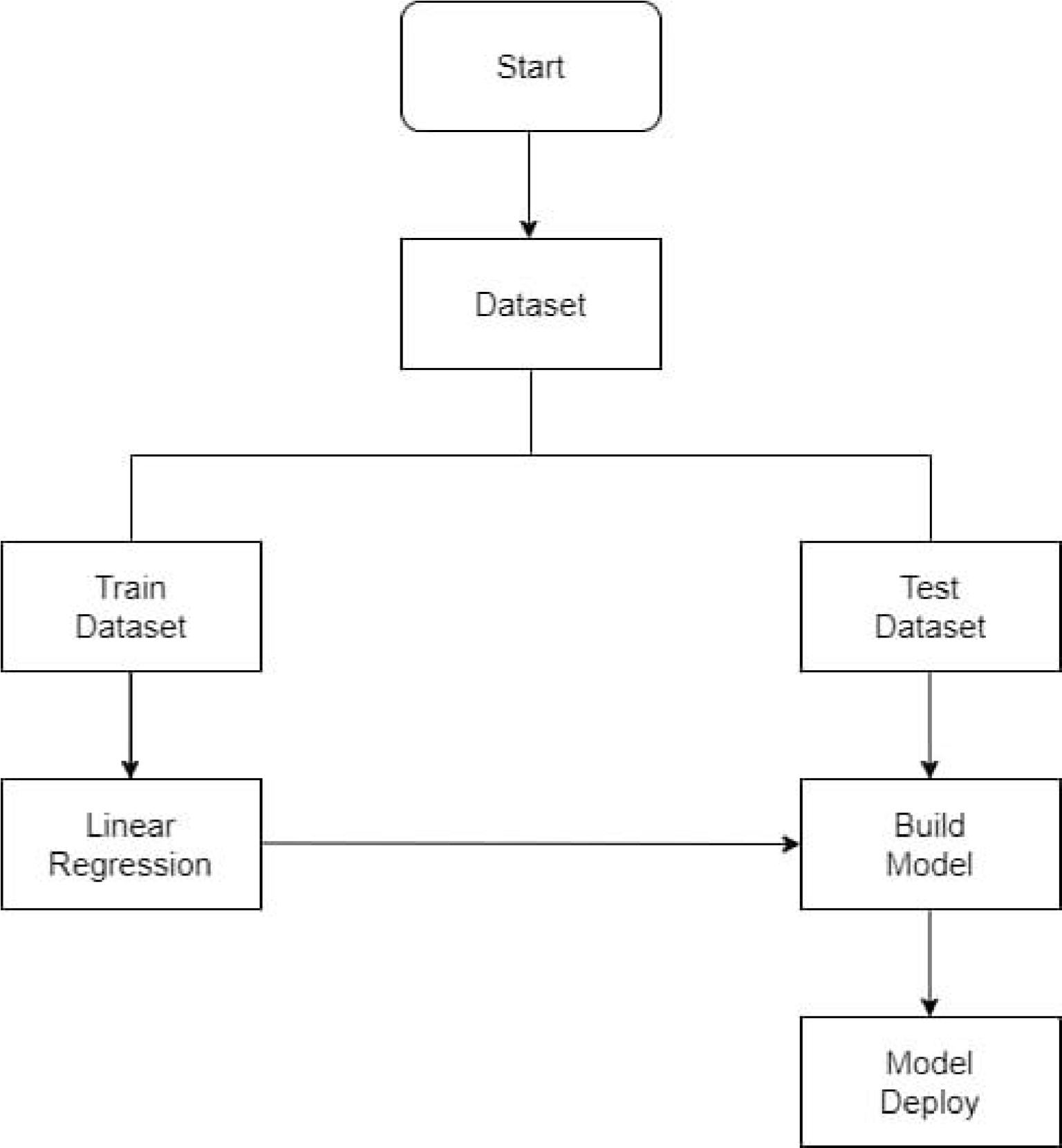
* **Data :** Real estate dataset with information on houses, including features like square footage, number of bedrooms, location, amenities, and, most importantly, the actual sale prices.
* **Programming tools :** Python is a common choice for machine learning projects. You'll need libraries such as NumPy, pandas, scikit-learn, and libraries for data visualization like Matplotlib or Seaborn.
* **Development Environment :** Jupyter Notebook or an integrated development environment (IDE) like PyCharm or Visual Studio Code for coding and experimentation.
* **Machine Learning Libraries :** Scikit-learn for machine learning algorithms and tools.XGBoost, RandomForest, or other regression models for house price prediction.
* **Data Preprocessing :** Data cleaning and preprocessing skills to handle missing values, outliers, and feature engineering.
* **Feature Selection :** Knowledge of how to choose relevant features for the prediction model.
* **Model Building :** Implement machine learning models for regression tasks. Linear Regression, Decision Trees, Random Forest, and Gradient Boosting models like XGBoost are commonly used.
* **Cross-Validation :** Implement cross-validation techniques to assess the model's generalization.
* **Hyper parameter Tuning :** Ability to fine-tune model hyperparameters for better performance.
* **Data Visualization :** Skills to create data visualizations for EDA and results presentation using libraries like Matplotlib and Seaborn.
* **Project Management :** Strong project management skills to plan and execute the project effectively.
* **Documentation :** The ability to create clear and comprehensive project documentation and reports.
* **Ethical Considerations :** Awareness of ethical concerns related to housing data, such as fairness and bias in model predictions.
* **Data Privacy :** Ensure that you have the right to use and share the data, respecting privacy regulations and property rights.
* **Continuous Learning :** Stay updated with the latest developments in machine learning and data science.

# Flow chart :



* **Data Collection :** The machine learning model is given the test data but without the price of the properties in order to predict the price for them given the various features for the properties. The predicted price is then compared to the actual price in the test data.
* **Preprocessing :** Data preprocessing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, greater is the reliance on the produced results. Incomplete, noisy, and inconsistent data are the properties of large real-world datasets. Data preprocessing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise and resolving inconsistencies.
* **Data Analysis :** Data Analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information by informing conclusions and supporting decision making.
* **Application of algorithm :** Machine learning is a buzzword for today’s technology, and it is growing very rapidly day by day. We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc. Below are some most trending real-world applications of Machine Learning:
* **Evaluating the models :** Machine Learning Model does not require hardcoded algorithms. We feed a large amount of data to the model and the model tries to figure out the features on its own to make future predictions. So we must also use some techniques to determine the predictive power of the model.

**Block Diagram :**

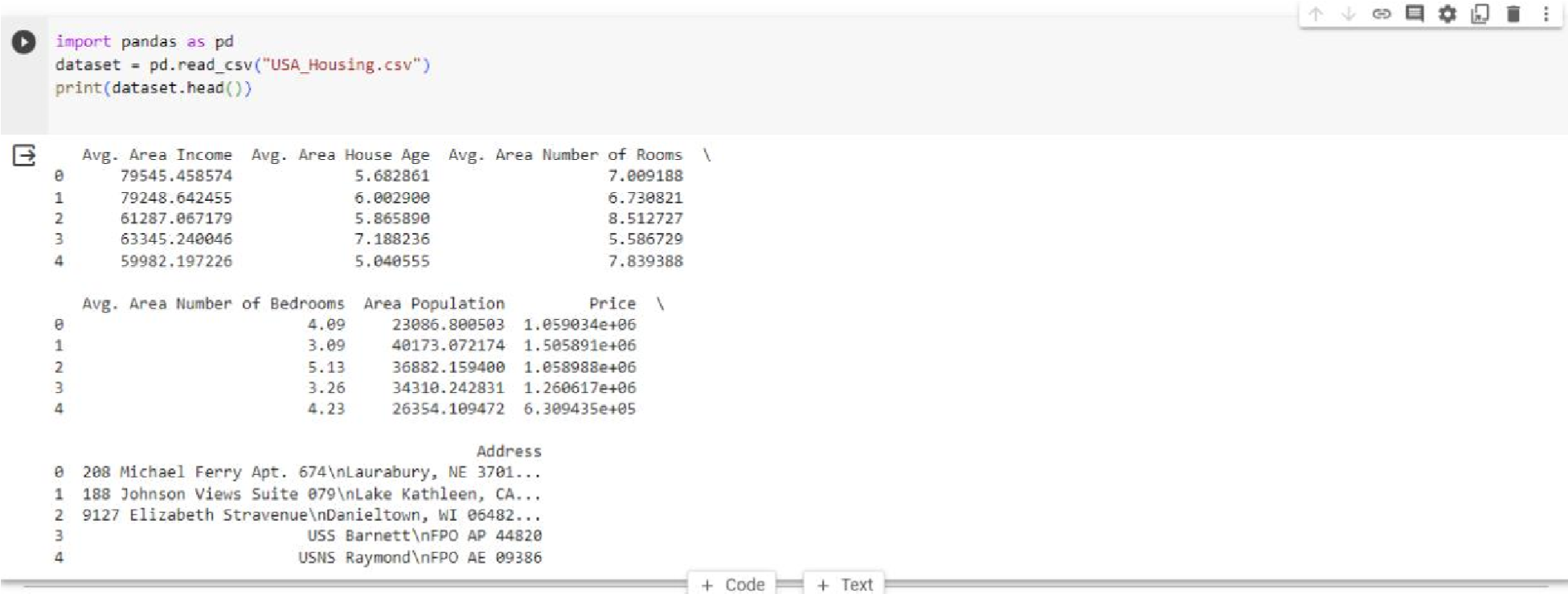


We started building our project by loading the dataset, Data processing and exploratory data analysis in Google colab Notebook.

# Loading the Dataset:

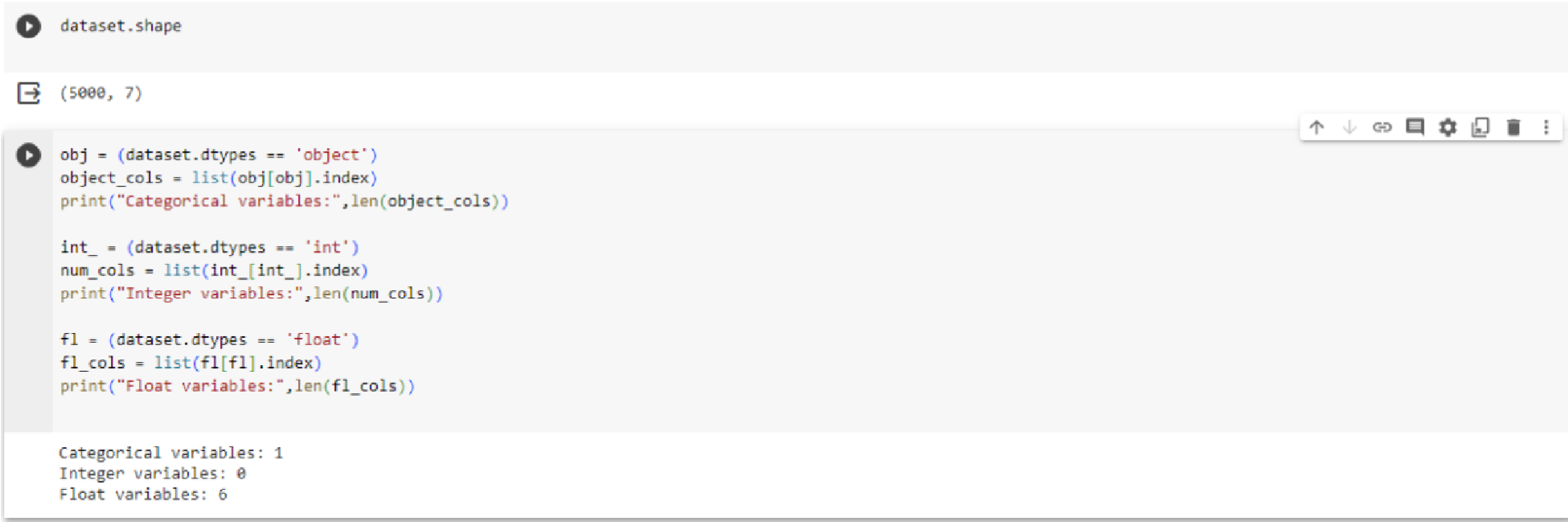
Downloaded the **USA\_Housing.csv** dataset from the Kaggle

We loaded a train.csv dataset using the **pandas** library.



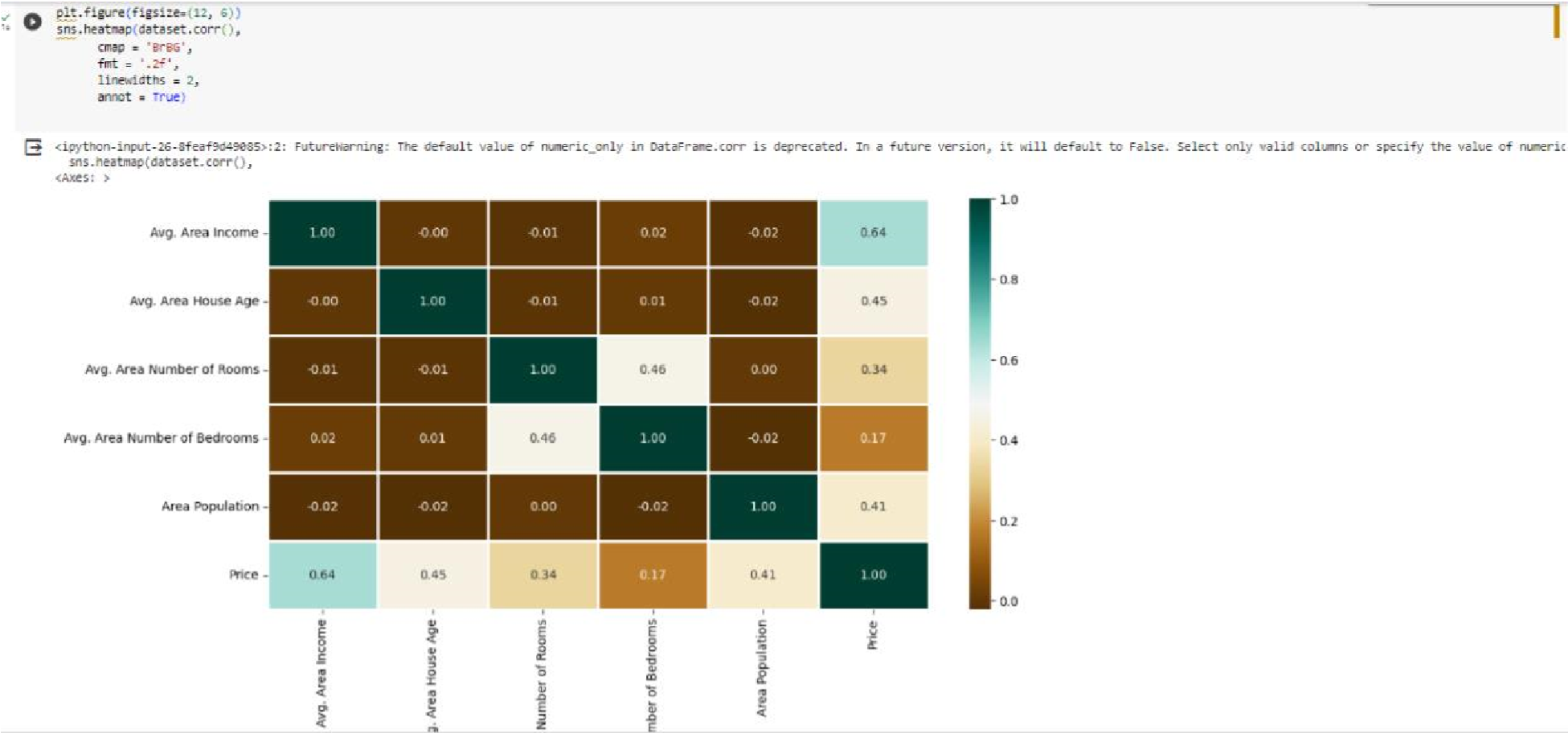
# Data Processing:

Now, we categorize the features depending on their datatype (int, float, object) and then calculate the number of them.



# Exploratory Data Analysis :

EDA refers to the deep analysis of data so as to discover different patterns and spot anomalies. Before making inferences from data it is essential to examine all your variables.Using Seaborn libraries.

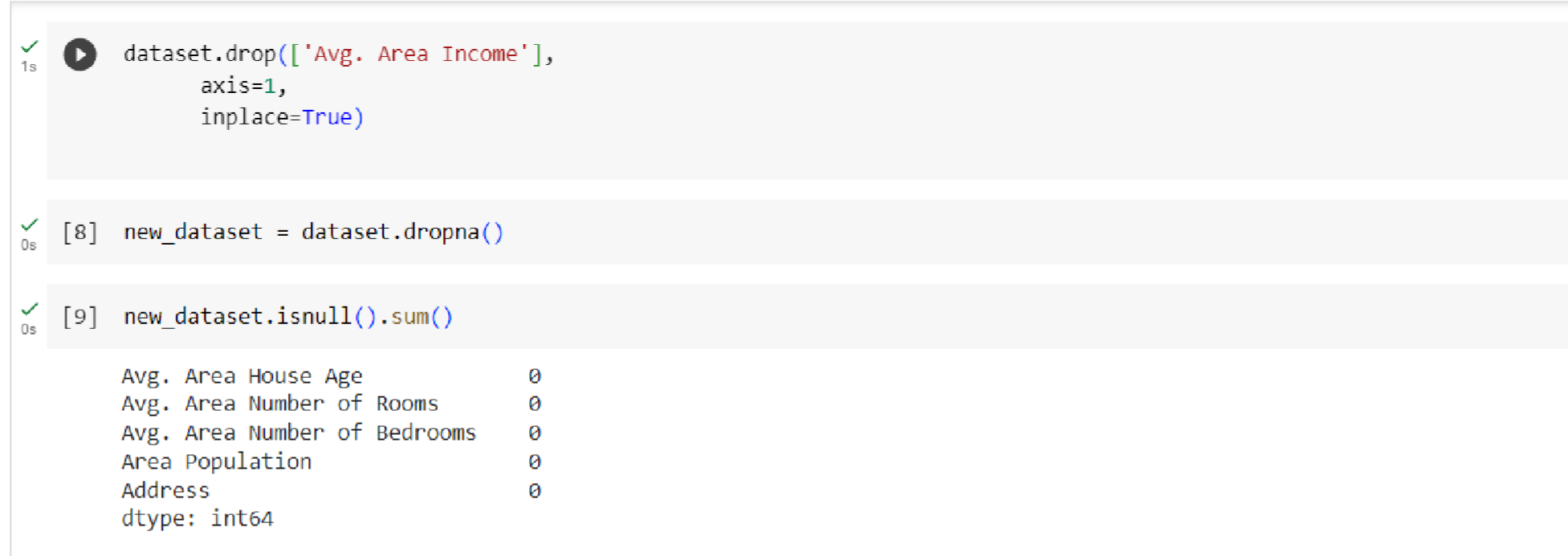


We further building our project by loading the Data cleansing, Splitting Data set into Testing and Training, Model and accuracy like Random forest and Linear regression or Support Vector Machine in Google colab Notebook.

**Data Cleansing:**

Downloaded the **USA\_Housing.csv** dataset from the Kaggle

We perform data cleansing operation



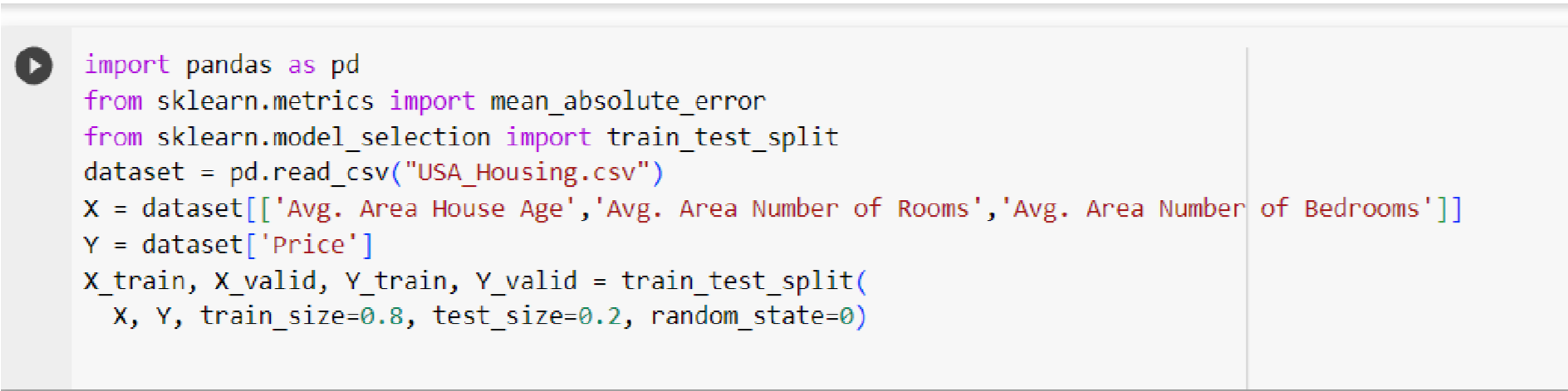
# Splitting Data Sets :

Now, we categorize the features depending on Categorial features and Calculate the No. of Categorial features of them.



# Splitting Dataset into Training and Testing :

X and Y splitting (i.e. Y is the SalePrice column and the rest of the other columns are X)



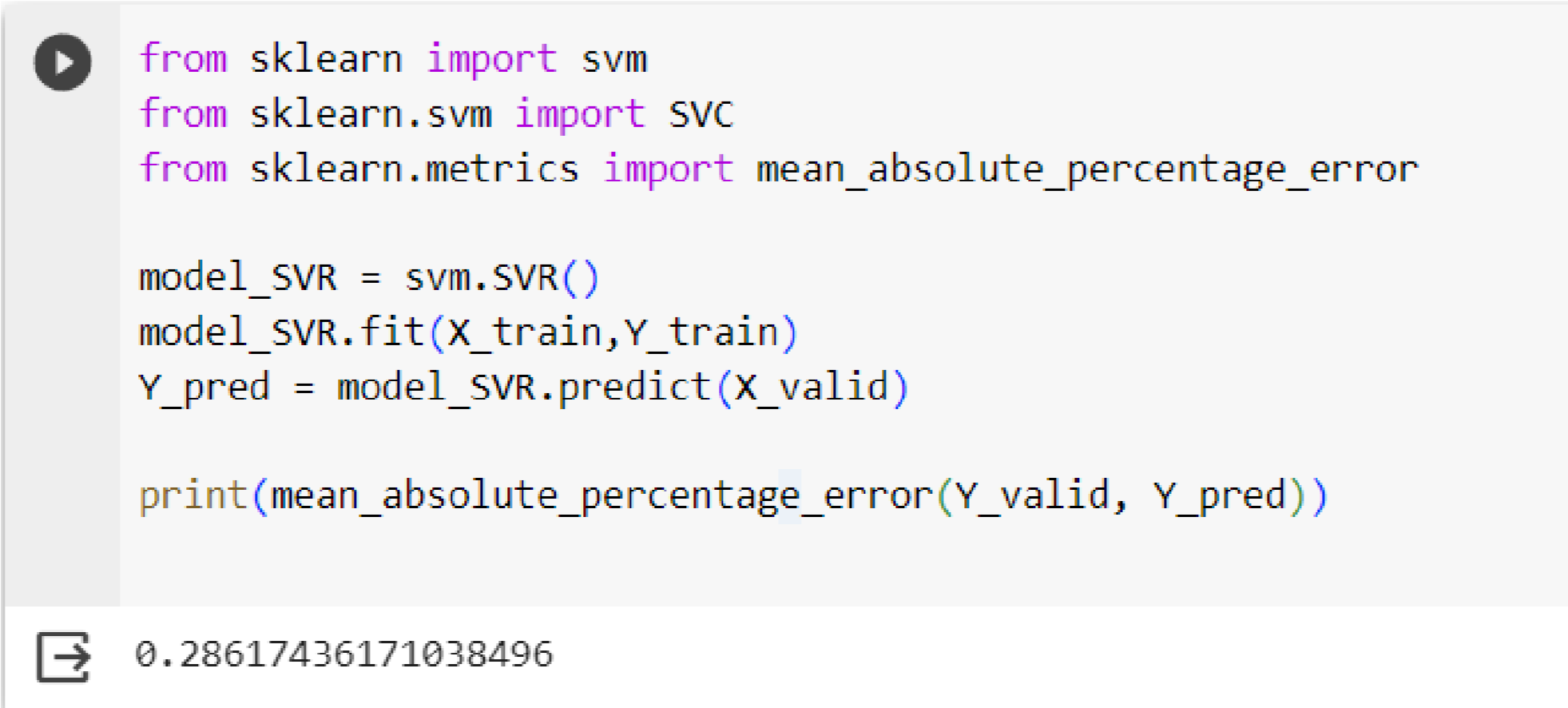
# Model And Accuracy :

As we have to train the model to determine the continuous values, so we will be using these models like

* Support Vector Machine ( SVM )
* Linear Regression
* Random Forest Regressor

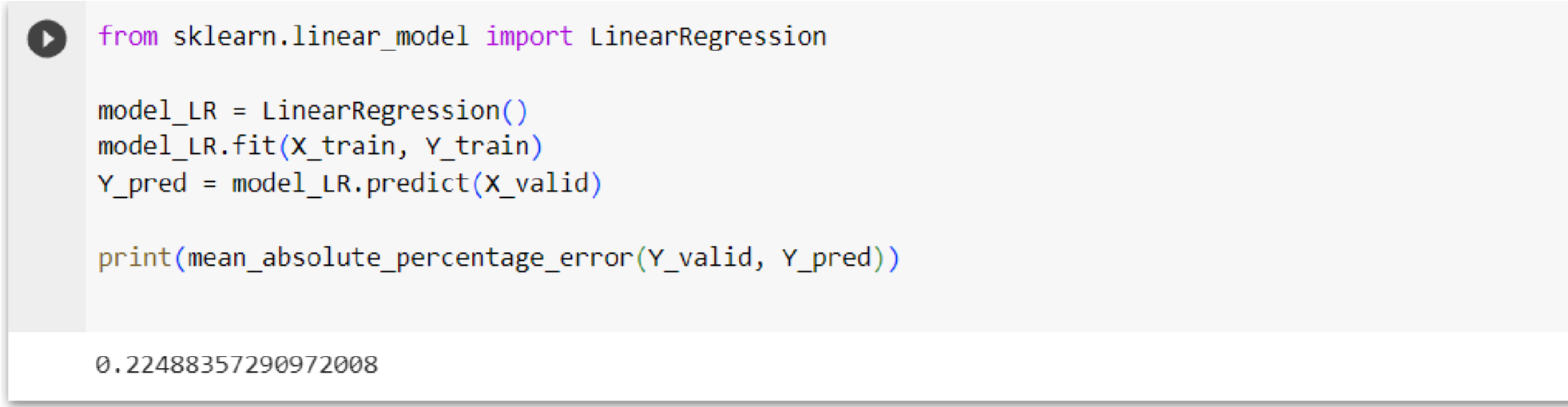
# Support Vector Machine (SVM):

SVM can be used for both regression and classification model. It finds the hyperplane in the n-dimensional plane



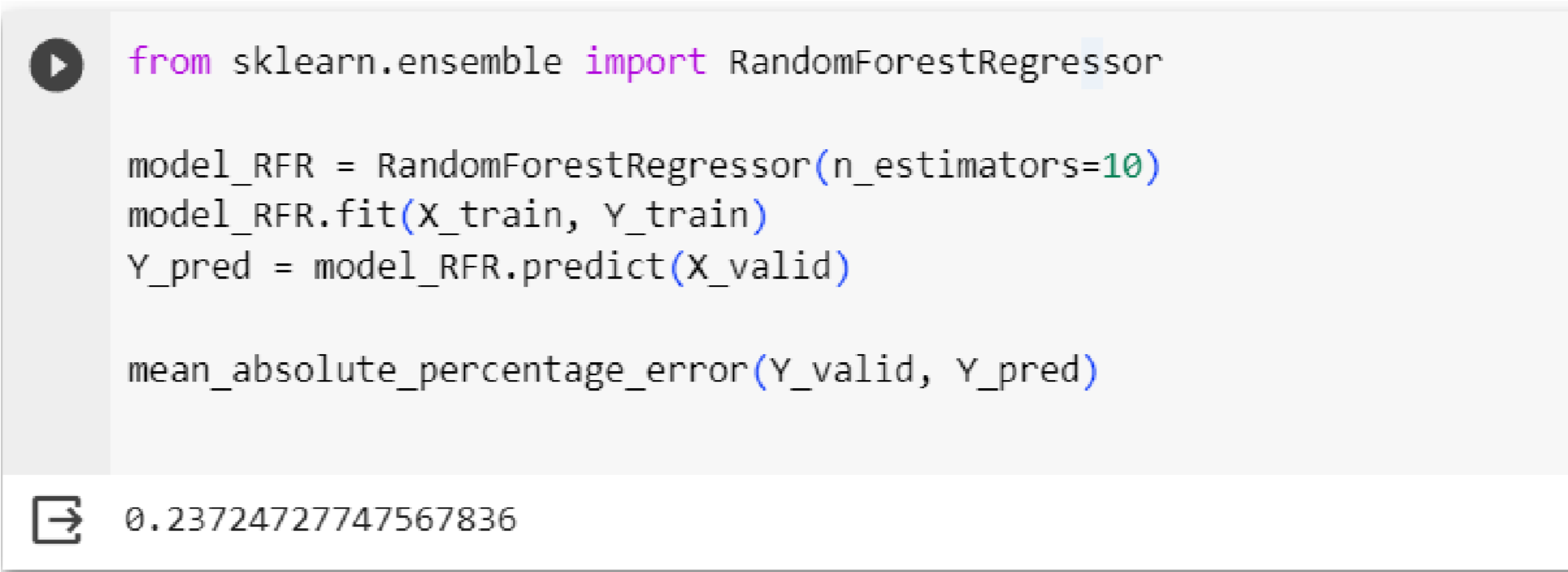
# Linear Regression :

Linear Regression predicts the final output-dependent value based on the given independent features. Like, here we have to predict SalePrice depending on features like MSSubClass.



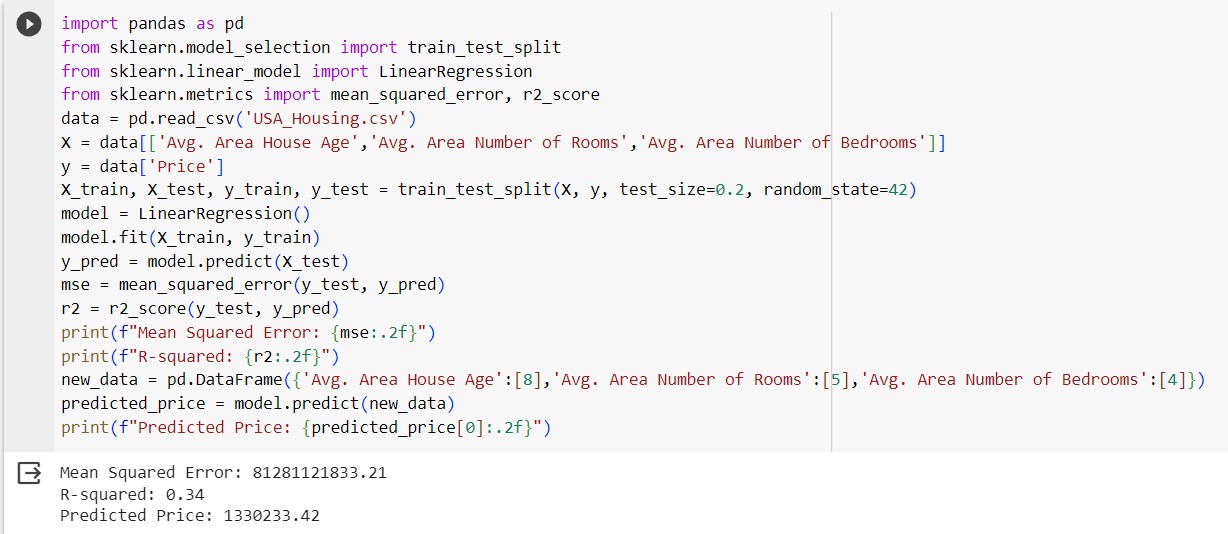
# Random Forest Regressor :

Random Forest is an ensemble technique that uses multiple of decision trees and can be used for both regression and classification tasks.



# Evaluation :

The machine learning model is given the test data but without the price of the properties in order to predict the price for them given the various features for the properties. The predicted price is then compared to the actual price in the test data.



**Team Members:**

K.Santhosh

J.Saravanakumar

M.Sangara Sequvar

C.Mohan

**JP COLLEGE OF ENGINEERING**