CS-451 REVIEW

PRESENTATION ON

DATA SCIENCE

By:
ORUGANTI MONIK PAPARAO
Y22CS139

AGENDA....

OF THE DAY....

- Abstract
- Introduction
- Key Concepts
- Dataset Details
- Implementation
- Related Clicks
- Conclusion

ABSTRACT

- This learning journey covers essential concepts in Artificial Intelligence (AI) and Data Science, focusing on both theory and practical applications. It includes an introduction to AI, core machine learning concepts, and Python programming fundamentals such as data structures, control flow, functions, exception handling, and object-oriented programming.
- The course also emphasizes data analysis using Python libraries like NumPy, Pandas, and Matplotlib for data manipulation, statistics, and visualization, along with SQL for efficient database management.
- Advanced machine learning techniques—linear and logistic regression, Naive Bayes, hypothesis testing, and evaluation metrics—are applied to real-world datasets including wine quality, housing prices, stock market trends, movie ratings, and more.

CONTD..

- The project presents a House Price Prediction Web Application developed using Streamlit and Machine Learning techniques. The system utilizes a housing dataset to train a Linear Regression model for predicting house prices based on key features such as area, bedrooms, bathrooms, floors, parking, and total rooms. The application provides an interactive interface where users can input property details and receive an estimated price.
- In addition to predictions, the app offers data exploration and visualization tools. Users can analyze scatter plots (e.g., Rooms vs Price, Area vs Price) and generate custom plots to understand feature relationships. A correlation heatmap is included to highlight dependencies among numerical variables. The dataset viewer allows sorting and interactive inspection of the data.

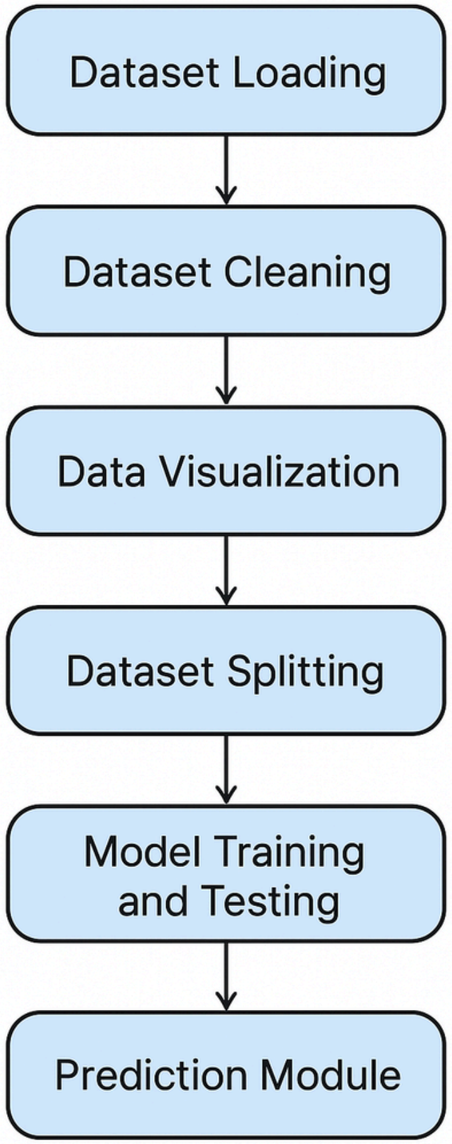
INTRODUCTION

- In recent years, the application of machine learning in real estate has gained significant importance, as it enables accurate prediction of property prices based on various influencing factors.
- This internship provided a strong foundation in Artificial Intelligence, Data Science, and Machine Learning using Python and SQL. It covers essential Python programming concepts, data structures, object-oriented programming, and data analysis with NumPy, Pandas, Matplotlib, and exploratory data analysis.
- Machine learning modules introduce regression techniques, logistic regression, Naive Bayes, and model evaluation methods.
- The concepts are applied through real-time tasks such as house price prediction ensuring both theoretical knowledge and hands-on experience.

CONTD...

- House price prediction is a challenging task because it depends on multiple features such as area, number of bedrooms, bathrooms, floors, parking facilities, and other attributes.
- Traditional estimation methods are often subjective, whereas machine learning provides a data-driven and reliable approach.
- This project focuses on developing a House Price Prediction Web Application using Streamlit and a Linear Regression model. The dataset is preprocessed, cleaned, and visualized to identify trends and relationships. The system is designed to split the dataset into training and testing sets, ensuring the model is well-evaluated before deployment.
- Users can input property details through an interactive interface, and the application instantly predicts the estimated price.
- Additionally, the project integrates data visualization features such as scatter plots and correlation heatmaps, allowing users to explore the dataset and better understand the factors that influence house pricing.

KEY CONCEPTS OF IMPLEMENTATION



1. ***Dataset Loading*** – Importing the dataset into the system so it can be used for analysis and model building.
2. ***Dataset Cleaning*** – Handling missing values, duplicates, and errors to ensure data quality.
3. ***Data Visualization*** – Using charts and plots to understand data patterns, trends, and relationships.
4. ***Dataset Splitting*** – Dividing data into training and testing sets for proper evaluation of the model.
5. ***Model Training and Testing*** – Building the machine learning model on training data and validating its performance on test data.
6. ***Prediction Module*** – Making predictions on new or unseen data.

DATASET DETAILS

Dataset Soure:
https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

Dataset Attributes:
area, floors, bedrooms, bathrooms, total rooms, parking, price

Dataset Size:

545 Rows.

Dataset....

	A	B	C	D	E	F	G
1	area	floors	bedrooms	bathrooms	total rooms	parking	price
2	7420	3	4	2	10	2	1330000
3	8960	4	4	4	12	3	1225000
4	9960	2	3	2	9	2	1225000
5	7500	2	4	2	10	3	1221500
6	7420	2	4	1	9	2	1141000
7	7500	1	3	3	10	2	1085000
8	8580	4	4	3	11	2	1015000
9	16200	2	5	3	12	0	1015000
10	8100	2	4	1	9	2	987000
11	5750	4	3	2	9	1	980000
12	13200	2	3	1	8	2	980000

IMPLEMENTATION DETAILS

STEP-1 DATASET LOADING.....

- *The first step in any data science or machine learning project is loading the dataset. In this project, the dataset consists of house-related features such as area, number of floors, bedrooms, bathrooms, total rooms, parking availability, and the target variable price.*
- *The dataset is represented in the form of a NumPy array and then converted into a Pandas DataFrame for easier handling and analysis.*
- *Once loaded, the dataset can be inspected to understand the type of features, the range of values, and to ensure it is suitable for further processing like cleaning, visualization, and splitting for model training and testing.*

STEP-2 DATASET CLEANING....

- *Dataset cleaning is an essential step to ensure the quality and reliability of the data before applying machine learning models. Raw datasets often contain missing values, duplicate records, inconsistent entries, or outliers that may affect the accuracy of predictions.*
- *In this project, the dataset was carefully checked for invalid or illogical entries such as negative values for numerical features (e.g., floors, bathrooms, bedrooms, parking) or mismatches where the sum of bedrooms and bathrooms exceeded the total number of rooms. Such inconsistencies were handled by applying validation rules and correcting or removing erroneous records.*
- *Through dataset cleaning, the data was standardized and made consistent, providing a strong foundation for accurate analysis, visualization, and model training.*

STEP-3 DATA VISUALIZATION....

- *Data visualization is a crucial step in understanding the relationship between different features in the dataset and the target variable.*
- *In this project, scatter plots were used to analyze how factors such as total rooms, area, bathrooms, bedrooms, floors, and parking relate to house prices.*
- *These plots help in identifying patterns, trends, and potential outliers in the data.*
- *Additionally, a correlation heatmap was generated to measure the strength of relationships between numerical features.*
- *The heatmap provides a clear overview of which features are most strongly correlated with the target variable (price) and with each other.*
- *This step is essential for selecting meaningful features for model training and gaining insights into the dataset.*

STEP-4 DATA SPLITTING....

- *After preparing and visualizing the dataset, the next step is splitting the data into training and testing sets.*
- *The training set is used to teach the machine learning model how to identify patterns and relationships, while the testing set is used to evaluate the model's performance on unseen data.*
- *In this project, the dataset was divided into 80% training data and 20% testing data using the train_test_split function from Scikit-learn.*
- *This ensures that the model is trained effectively while still being tested on a separate portion of data to measure its accuracy and generalization ability.*

STEP-5 MODEL TRAINING AND TESTING....

- *Once the dataset was split into training and testing sets, the machine learning model was built using Linear Regression, which is one of the most widely used regression techniques.*
- *Linear Regression is suitable for this project as it establishes a linear relationship between the independent variables (features such as area, bedrooms, bathrooms, floors, parking, etc.) and the dependent variable (house price).*
- *The model was trained using the training dataset, where it learned the relationship between the input features and the target output.*
- *After training, the model was tested using the testing dataset to evaluate its performance.*
- *This testing step ensures that the model is not only accurate on the training data but also generalizes well to new, unseen data.*

STEP-6 PREDICTION MODULE....

- *The prediction module is the final stage of the project where the trained Linear Regression model is used to predict house prices based on user-provided inputs.*
- *A simple interface was designed where users can enter details such as area, number of floors, bedrooms, bathrooms, total rooms, and parking availability.*
- *The system validates the inputs to ensure correctness, such as preventing negative values and checking that the sum of bedrooms and bathrooms does not exceed the total number of rooms.*
- *Once valid inputs are provided, the data is converted into a DataFrame and passed to the trained model, which generates a price prediction.*
- *This module allows real-time interaction with the machine learning model, giving users a practical way to estimate house prices based on specific features.*

MODULES USED

1. ***For User Interface***: *streamlit*
2. **For *Data Loading, Data Cleaning***: *pandas, numpy*
3. **For *Data Visualization***: *matplotlib*
4. **For *Heat Map***: *seaborn*
5. **For Linear Regression (ML Model)** : *scikit-learn*

For Deployment of the Project used the Free Deployment Service Provied by Streamlit.*

RESULT CLICKS...

House Price Prediction App

Navigation

Do Predict | Graphs | See Data | Correlation Heatmap | About

Do Predict Section

Predict House Price

Enter details of the house below:

Enter area: 6936.94

Enter floors: 2

Enter bedrooms: 3

Enter bathrooms: 1

Enter total rooms: 9

Enter parking: 1

Predict

Predicted House Price: 733,269.97 Rupees

House Price Prediction App

Navigation

Do Predict | Graphs | See Data | Correlation Heatmap | About

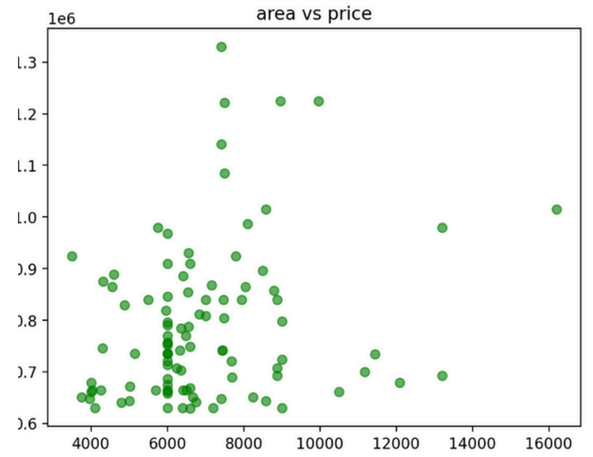
Graphs Section

Graphs and Visualizations

Choose a graph: Custom Scatter Plot

Select X-axis: area

Select Y-axis: price



RESULT CLICKS...

House Price Prediction App

Navigation

Do Predict | Graphs | See Data | Correlation Heatmap | About

Dataset Viewer

Explore the housing dataset below:

Sort by column

area

Order

Ascending | Descending

	↑ area	floors	bedrooms	bathrooms	total rooms	parking	price
13	3500	2	4	2	10	2	924000
84	3760	2	3	1	8	2	651000
87	3960	1	3	1	8	2	647500
70	4000	2	3	2	9	0	679000
81	4000	2	3	2	9	1	661500
74	4040	2	3	1	8	1	665000
95	4100	3	3	2	9	2	630000
75	4260	2	4	2	10	0	665000
48	4300	2	3	2	9	1	745500
20	4320	2	3	1	8	2	875000

See Data Section

Heatmap Section

House Price Prediction App

Navigation

Do Predict | Graphs | See Data | Correlation Heatmap | About

Correlation Heatmap

RESULT CLICKS...

🏠 House Price Prediction App

Navigation

○ Do Predict ○ Graphs ○ See Data ○ Correlation Heatmap ◉ About

ℹ️ About this App

This **House Price Prediction App** helps users estimate house prices based on multiple features.

Features:

- Predict house prices using **Linear Regression**
- Explore graphs like **Rooms vs Price, Area vs Price etc...**
- Sort & view the dataset easily
- Check correlations between variables

Importance:

- Useful for **buyers** to know fair price
- Helpful for **real estate agents** in decision-making
- Good for **students & researchers** to learn Data Science & ML

Developed with ❤️ using Streamlit, Scikit-Learn, Pandas, and Matplotlib

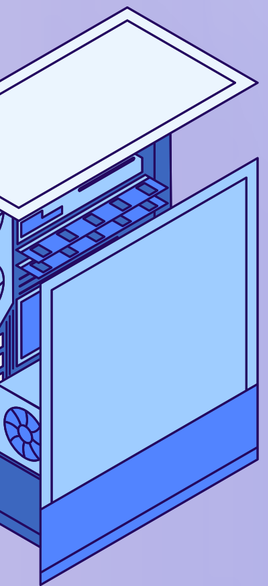
- Hosted on: Streamlit Cloud
- Project URL: https://omprhousepriceprediction.streamlit.app/
- Availability: Public
- Languages Used: Python
- Github Code Link: https://github.com/omonikpaparao/house_price_prediction/blob/main/sam.py

About Section

Short Description

CONCLUSION

- The work demonstrates a complete Data Science workflow, including dataset loading, cleaning, visualization, splitting, training, and prediction.
- Linear Regression was applied as a supervised learning approach to establish the relationship between house features and price.
- Exploratory Data Analysis (EDA) using scatter plots and correlation heatmaps helped uncover important feature dependencies.
- The integration of statistical techniques, feature engineering, and predictive modeling reflects the practical use of AI and Data Science concepts.
- Overall, the study highlights how data-driven approaches can be effectively applied for accurate house price estimation.




Thank
You