Introduction to Data Science

Final Projects - House Price Prediction - 22KDL

Project 1:

Objective:
In this final project, students will demonstrate their end-to-end skills in data scraping, data preprocessing, feature engineering, modeling, and evaluation. The task involves predicting house prices using machine learning models. The following requirements outline the project details.

Data Collection:

1. Scrape and collect data items from the website https://batdongsan.vn/ban-nha/ to create a comprehensive dataset.
2. Explore the website thoroughly to gather relevant information on each data item for house price prediction.
3. Determine the size of the dataset to be collected.

Data Preprocessing and Feature Engineering:

1. Select appropriate data features for training the models.
2. Perform necessary data preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features.
3. Conduct feature engineering techniques to enhance the predictive power of the selected features.

Methodology:

1. Visualize a heatmap based on the address of the listed house using kernel density estimation at a certain time.
2. Design a methodology to address the house price prediction problem.
3. Consider various approaches and techniques for data analysis and modeling.
4. Justify the chosen methodologies and explain their effectiveness.
5. Recommended models include Linear/Ridge/Lasso Regression, Decision Tree/Random Forest, and Gradient Boosting among others.
6. Encourage students to explore advanced models if feasible.
7. If deep learning models are used, prefer and recommend the PyTorch framework.
8. Choose relevant metrics for model evaluation.
9. If multiple models are employed, students should perform model benchmarks and identify the best one for the problem.

Implementation and Reporting:

1. Students are not required to implement the algorithms from scratch, but it is recommended if possible.
2. Using Jupyter Notebook for implementation and report generation.
3. The report can be written in both Vietnamese and English.
4. Include a detailed explanation of the implementation process and any challenges faced.

5. The report should demonstrate a comprehensive understanding of the selected models and techniques.

Submission:

1. Submit the Jupyter Notebook containing the report and implementation.
2. Provide the dataset used for the project as a zip file.