INTEL UNNATI INDUSTRIAL TRAINING PROGRAM 2024

Name:Shubham Prajapati

College:Parul Institute of Technology

Enrollment No.2203051050549

Email id:2203051050549@paruluniversity.ac.in

Problem Statement-12

KNOWLEDGE REPRESENTATION AND INSIGHTS GENERATIONFROM STRUCTUREE DATASETS

Report

Indroduction:Clearly explaining the problem and objectives

Definition:**Knowledge Representation and Insights Generation from Structured Datasets is** the process of organizing structured data in a way that enables systems to extract meaningful insights, which can then be used to inform decisions and strategies across various fields.

Objective

- The objective is to develop methods and techniques for:

1. Knowledge Representation: Structuring and organizing data so that computer systems can understand and process it efficiently.
2. Insights Generation: Analyzing structured data to extract meaningful patterns, trends, and actionable insights.

Data Set Description

Source:

1. *Title: Boston Housing Data*
2. *Sources: (a) Origin: This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. (b) Creator: Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.*
3. *(c) Date: July 7, 1993*

Website: Kaggle:Boston Housing Dataset

Continue

Dataset Description

The dataset includes various attributes related to housing values in the suburbs of Boston. Each attribute provides specific information that can influence real estate pricing and decision-making.

Attribute Information:

1. *CRIM: Per capita crime rate by town*
2. *ZN: Proportion of residential land zoned for lots over 25,000 sq.ft.*
3. *INDUS: Proportion of non-retail business acres per town*
4. *CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise)*
5. *NOX: Nitric oxides concentration (parts per 10 million)*
6. *RM: Average number of rooms per dwelling*
7. *AGE: Proportion of owner-occupied units built prior to 1940*
8. *DIS: Weighted distances to five Boston employment centers*
9. *RAD: Index of accessibility to radial highways*
10. *TAX: Full-value property-tax rate per $10,000*
11. *PTRATIO: Pupil-teacher ratio by town*
12. *B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town*
13. *LSTAT: Percentage of lower status of the population*
14. *MEDV: Median value of owner-occupied homes in $1000's*

Methodology

Methods used:Algorithm for knowledge representation (i.e data visualization,graph based representation,Regression Analysis).

Techniques for pattern identification

1.Statistical Analysis

- **Descriptive Statistics:** Summarize basic features of the data such as mean, median, standard deviation, and distribution.
- **Correlation Analysis:** Measure the strength and direction of relationships between variables using correlation coefficients.

2. Data Visualization

- **Scatter Plots:** Visualize relationships between two continuous variables.
- **Heatmaps:** Show the correlation matrix to identify which variables are strongly correlated.
- **Histograms:** Display the distribution of individual variables.

Regression Analysis

- **RandomForestRegressor:** Model the relationship between a dependent variable and one or more independent variables.

Result and Discussion

we present the results of our analysis on the Boston Housing dataset and discuss the implications of these findings. The goal is to interpret the results in a way that provides actionable insights for maximizing company profits in real estate.

We aimed to identify key factors affecting housing prices in Boston suburbs using statistical analysis and machine learning techniques. Specifically, we utilized RandomForestRegressor to predict the median value of homes and exploratory data analysis (EDA) to uncover significant patterns in the data.

Result

Data Summary

Provide an overview of the dataset with key statistics.

Example: "The Boston Housing dataset includes 506 observations with 14 attributes. Key statistics for the median value of homes (MEDV) show a mean of $22,532 and a standard deviation of $9,197. Other important features include crime rate (CRIM), average number of rooms (RM), and nitric oxide concentration (NOX)."

Exploratory Data Analysis (EDA)

Highlight the initial findings with visual aids.

- ***Correlation Analysis:*** *"A correlation matrix revealed that the average number of rooms per dwelling (RM) has a strong positive correlation (0.7) with MEDV, while the percentage of lower status population (LSTAT) has a strong negative correlation (-0.74) with MEDV."*

Conclusion: Summarizing Findings and Suggesting Future Work

- we analyzed the Boston Housing dataset to identify key factors influencing housing prices and developed a predictive model for real estate valuation. The main findings from our analysis are as follows:

1. **Key Predictors of Housing Prices:**
    1. **Average Number of Rooms (RM):** The number of rooms in a dwelling is a strong positive predictor of housing prices. Homes with more rooms tend to have higher median values.
    2. **Percentage of Lower Status Population (LSTAT):** There is a strong negative correlation between LSTAT and housing prices, indicating that higher percentages of lower status population are associated with lower home values.
    3. **Crime Rate (CRIM):** Higher crime rates negatively impact housing prices.
    4. **Proximity to the Charles River (CHAS):** Homes located near the Charles River tend to have higher values.
2. **Model Performance:**
    1. The RandomForestRegressor model achieved an R-squared value of 0.74 and a Mean Squared Error (MSE) of 21.48 on the test set. This indicates that the model explains a significant portion of the variance in housing prices, although there is room for improvement.
3. **Insights for Real Estate Investment:**
    1. Investments should focus on properties with potential for expansion (more rooms).
    2. Consideration of socioeconomic factors is crucial, as areas with lower percentages of lower status population are more desirable.
    3. Proximity to natural amenities and lower crime rates are important factors in property valuation.

Continue

- **Suggestions for Future Work**
- While the current study provides valuable insights, there are several areas for future research and improvement:

1. **Incorporating Additional Variables:**
    1. Integrate more comprehensive datasets that include recent market trends, economic indicators, and other relevant factors such as school quality, proximity to public transport, and local amenities.
2. **Advanced Modeling Techniques:**
    1. Explore more sophisticated machine learning models such as Random Forest, Gradient Boosting, or Neural Networks, which may capture complex relationships better than linear regression.
    2. Perform feature engineering to create new variables that might better represent the underlying patterns in the data.

Features Offered

1.Property Appreciation Prediction:

•Predict the potential appreciation of a property over various time horizons (e.g., 1 year, 5 years, 10 years).

2.Property Value Estimation:

•Estimate the current market value of properties based on recent sales and market trends.

3.Neighborhood Analysis:

•Assess the quality of the neighborhood, including factors such as crime rates, school quality, and proximity to amenities.

4.Market Trend Analysis:

•Analyze local market trends and economic indicators to understand the overall market sentiment and potential future changes.

5.Geospatial Insights:

•Provide insights based on location, such as proximity to infrastructure projects, public transportation, and future development plans.

PROCESS FLOW

❑ Problem Definition

•Define the problem
•Set objectives

❑ Data Collection

•Identify data sources
•Gather data

❑ Data Preprocessing

•Clean data
•Handle missing values
•Feature engineering
•Normalize/scale data

❑ Exploratory Data Analysis (EDA)

•Visualize data
•Analyze correlations
•Generate insights

Continue

❑ Model Selection

•Consider algorithms (Linear Regression, Decision Trees, etc.)
•Select appropriate models

❑Model Training

•Split data into training and test sets
•Perform cross-validation
•Tune hyperparameters

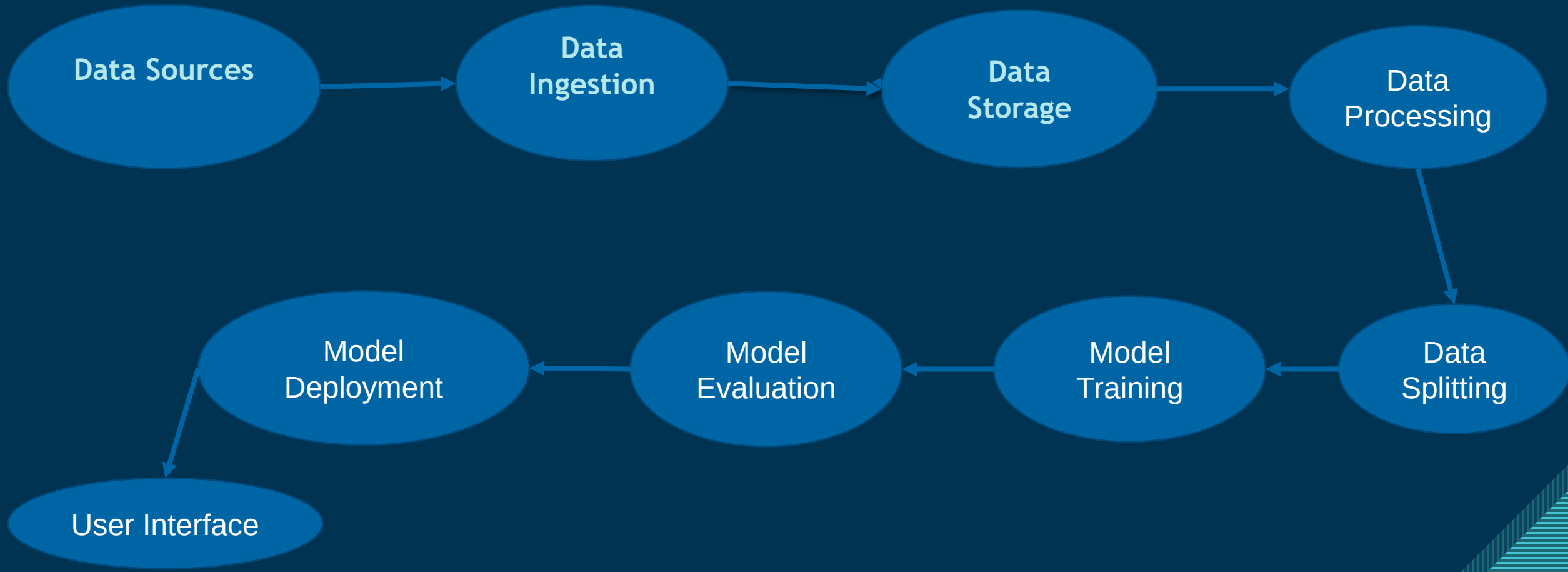
❑ Model Evaluation

•Evaluate on training set
•Evaluate on test set
•Compare model performance

❑ Model Deployment

•Choose deployment platform (AWS, GCP, Azure)
•Develop API for predictions
•Monitor and maintain model

Architecture Diagram



Thank You