Project Journal

LIKHITHA KONASALE PRAKASH [X23160691]

Dataset Name: King-county housing dataset

1.Summary

I worked with the King-county housing dataset for predicting the house prices by applying decision tree and xgboost algorithms. The work also included performing data analysis techniques along with an interactive streamlit application.

2. Tasks and Responsibilities

2.1 Dataset Selection

- *Dataset used:* King-county housing data
- *Source of the dataset:* Kaggle

2.2 Data Preprocessing and Transformation

- preprocessed data was stored to MongoDB database and queries were performed.
- The data was transformed by manipulating the columns, creating new columns and converting the datatypes of features.
- The work utilized tools and libraries like python, numpy, pandas

2.3 Database Integration

Database used: PostgreSQL
Process: The transformed data was programatically stored to postgresql

2.4 Analysis and Visualization

Data visualizations were done to get more insights about the data. Heatmap, barplot, scatter plot, and line graph were performed to visualize the data effeciently by utilizing matplotlib, seaborn and plotly libraries.

2.5 Algorithm Application

Decision tree and XGBoost algorithms were applied on the preprocessed data to predict the house prices.

2.6 Streamlit Application

An interactive application was built by using streamlit. The application predicts the house price based on the particular input given by the user using the algorithm.

2.5 Documentation and Reporting

The report for the entire part of King-county dataset is been authored by me.

3. Time Log

Data selection – 2 hours

Data preprocessing and transformation – 7 hours

Database integration – 4 hours

Analysis and visualization – 6 hours

Algorithm application – 5 hours

Streamlit application – 8 hours

Report and Presentation – 5 hours

4. Skills gained

This project helped me to gain more experience on programming languages. It also enhanced my knowledge on data analysis, using databases and problem-solving skills.