Project 1: House Price Prediction

Overview

This project aims to predict house prices using regression techniques. The dataset contains information on various features such as area, number of bedrooms and bathrooms, furnishing status, parking availability, transaction type, and price per square foot.

Data Preprocessing

Handled null values: Filled null values in 'Bathroom', 'Furnishing', 'Type', and 'Parking' columns with mean and mode values.

Data scaling: Applied Min-Max scaling to columns ['Area', 'BHK', 'Bathroom', 'Parking', 'Per_Sqft', 'Price'].

Outlier detection: Detected outliers using Interquartile Range (IQR) method and removed them from columns ['Area', 'BHK', 'Bathroom', 'Parking', 'Per_Sqft'].

Data Encoding:

Converted categorical columns 'Furnishing', 'Status', 'Transaction', and 'Type' into numerical using One-Hot Encoding.

Combined encoded columns with the training set.

Model Building and Evaluation

Decision Tree Regression:

Implemented GridSearchCV to find the best parameters.

Achieved performance metrics:

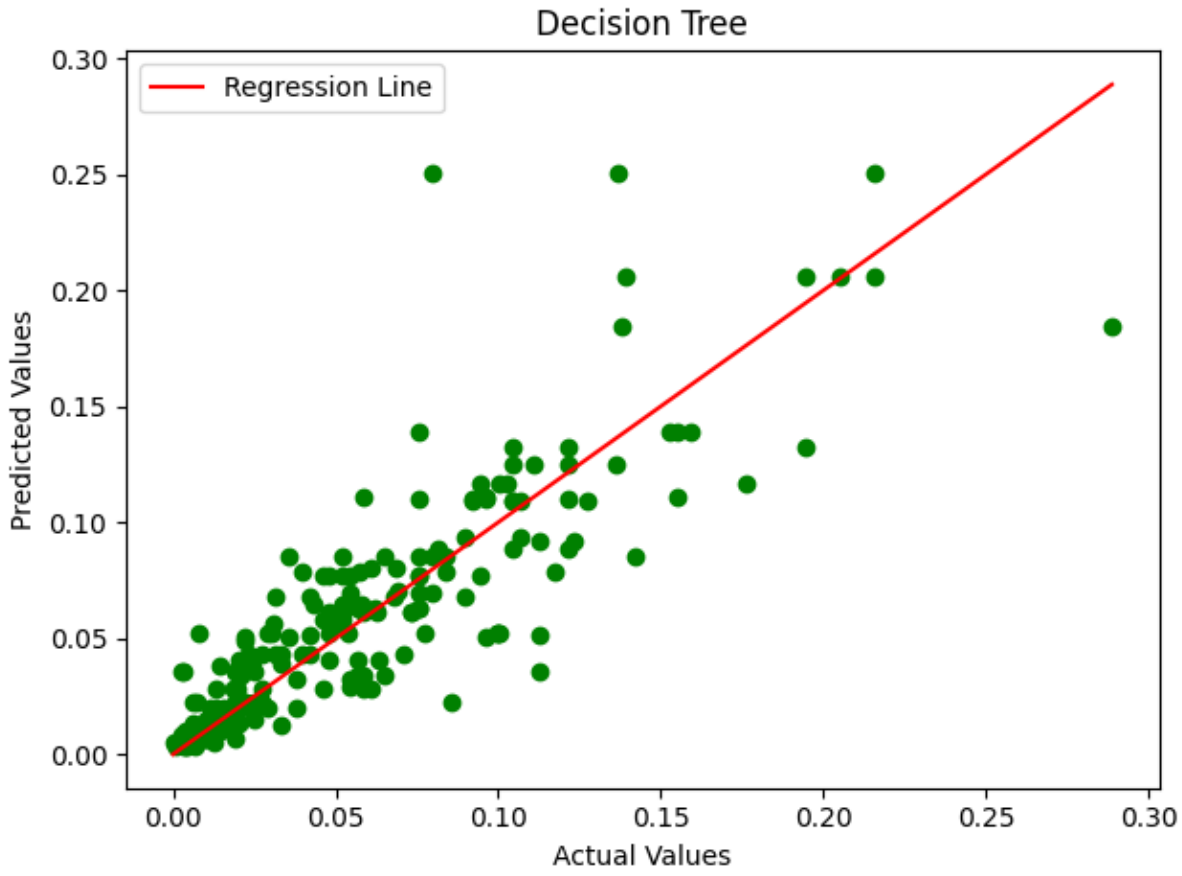
Mean Absolute Error (MAE): 0.0169

Mean Squared Error (MSE): 0.0007

Root Mean Squared Error (RMSE): 0.0269

R-squared (R2): 0.7081

Decision Tree Regression line Visualization



Linear Regression with Batch Gradient Descent:

Implemented from scratch.

Learning rate: 0.5, Number of iterations: 1000.

Performance metrics:

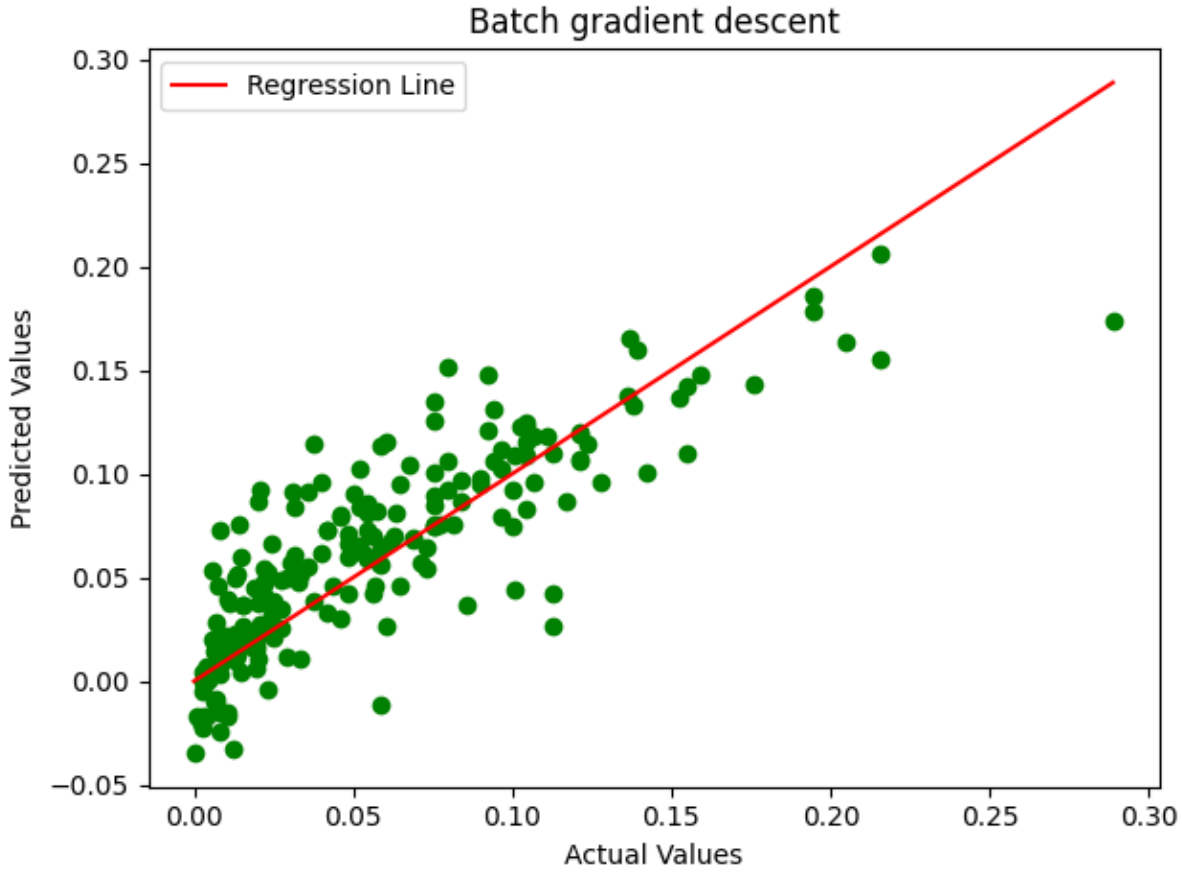
MAE: 0.0210

MSE: 0.0008

RMSE: 0.0285

R2: 0.6709

Batch GD Regression line Visualization



Polynomial Regression (Degree=2):

Applied polynomial regression with degree 2.

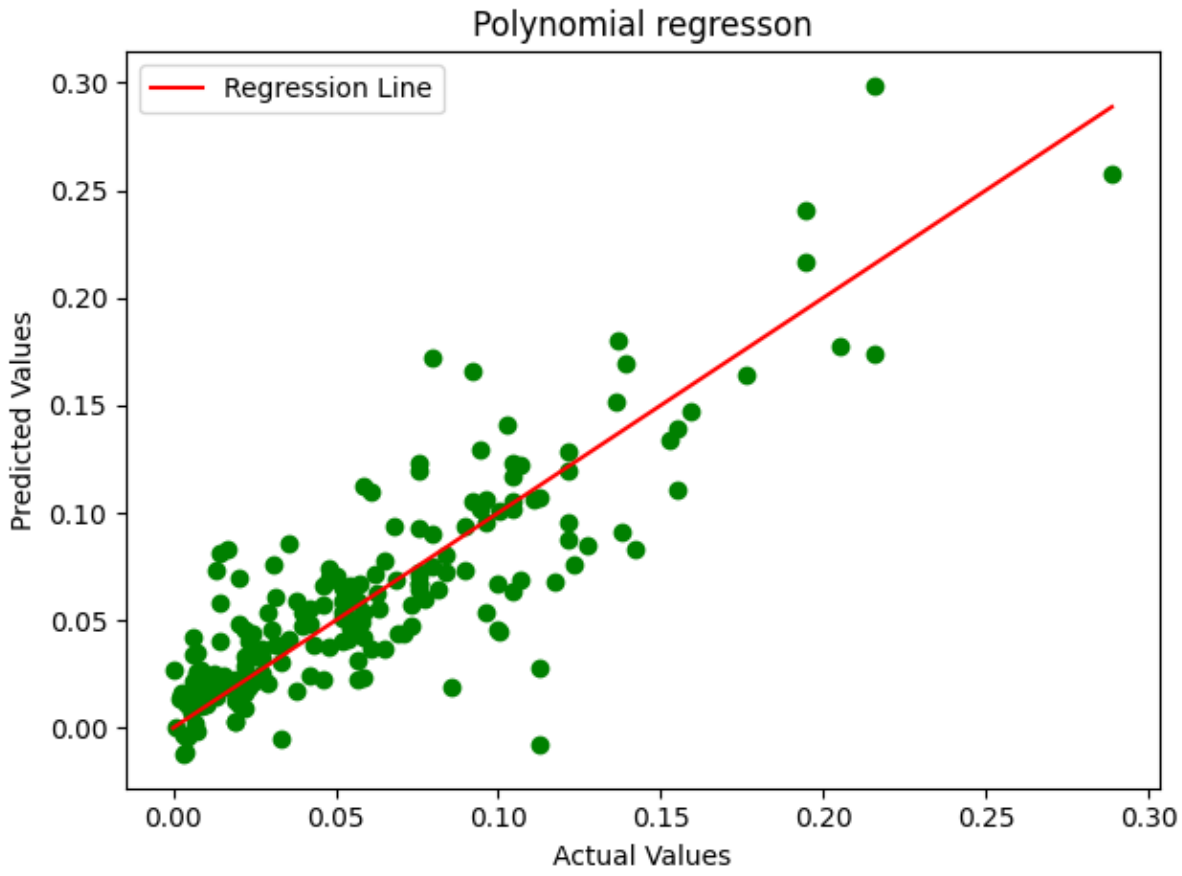
Performance metrics:

MAE: 0.0181

MSE: 0.0008

RMSE: 0.0280

R2: 0.7719



Conclusion

The models demonstrate low errors (MAE, MSE, RMSE), indicating good accuracy in predicting house prices. The R2 score suggests that the models explain approximately 70.8% to 77.2% of the variance in the target variable, implying a moderately good to strong fit to the data.

Future Work

Explore other regression techniques such as Random Forest Regression, Support Vector Regression, etc.

Experiment with feature engineering to improve model performance.

Collect additional relevant features to enhance prediction accuracy.

This concludes the technical documentation for the House Price Prediction project.

Project 2: Breast Cancer Classification

Overview

This project aims to classify breast cancer cases using machine learning techniques. The dataset contains various features such as age, menopause status, tumor size, number of involved lymph nodes, breast quadrant, presence of metastasis, patient's medical history, and diagnosis result.

Data Preprocessing

Renamed Columns:

Renamed columns to standardized names ['S/N', 'Year', 'Age', 'Menopause', 'Tumor_Size', 'Inv_Nodes', 'Breast', 'Metastasis', 'Breast_Quadrant', 'History', 'Diagnosis_Result'].

Dropped Columns:

Removed 'S/N' and 'Year' columns.

Handling Categorical Error:

Corrected category error in the 'Breast_Quadrant' column.

Handling Missing Values:

Removed rows containing '#' values in the dataset.

Encoding Categorical Variables:

Assigned numerical values (0 and 1) to 'Breast' column (left and right).

Data Type Conversion:

Converted data types from object to int for columns ['Inv_Nodes', 'Breast', 'Tumor_Size', 'Metastasis', 'History'].

Data Splitting

Split the data into training and test sets.

Model Building and Evaluation

Random Forest Classifier:

Used GridSearchCV to find the best parameters.

Achieved performance metrics:

Accuracy: 90%

Precision: 93%

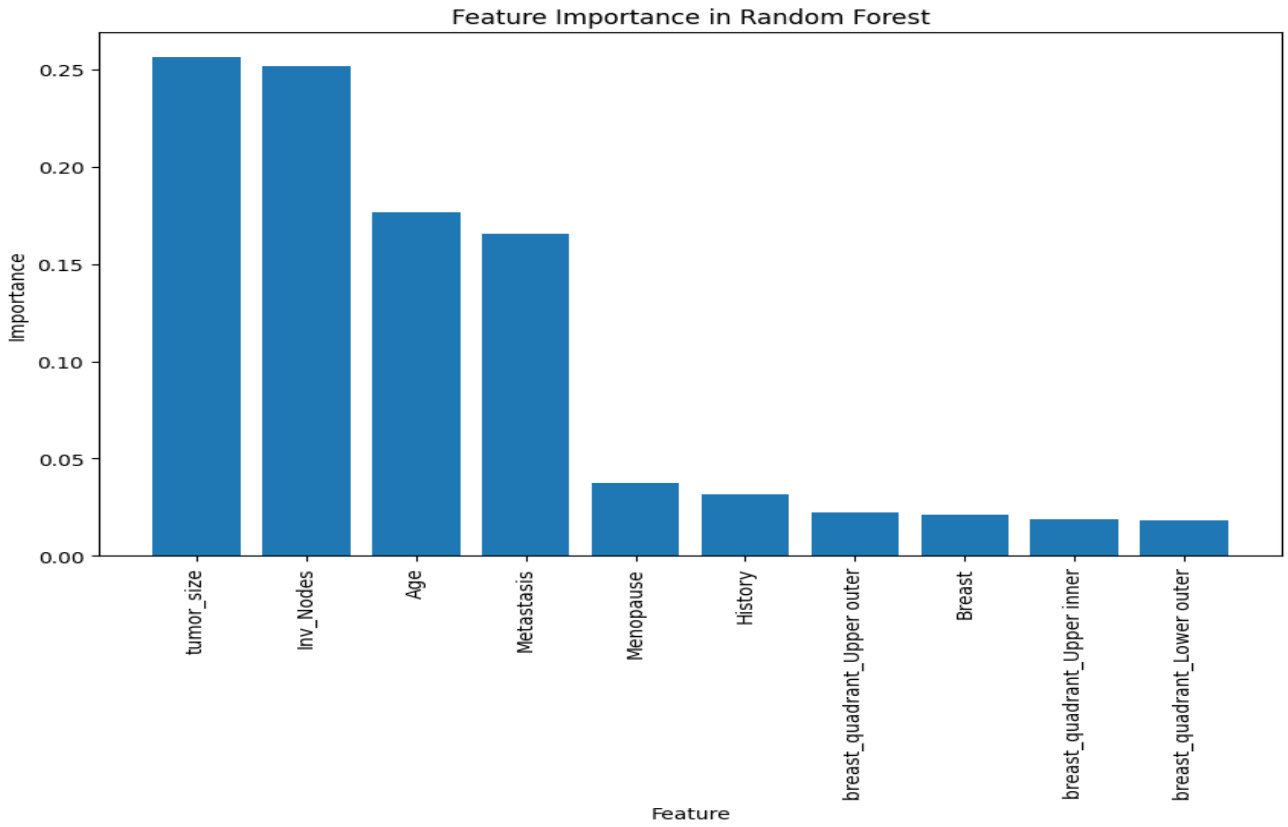
Recall: 89%

F1 Score: 90%

Obtained feature importance scores and identified top features.

Visualized feature importance using a bar plot.

Feature Importance Bar Plot



Logistic Regression:

Implemented logistic regression classifier.

Performance metrics:

Accuracy: 87%

Precision: 89%

Recall: 87%

F1 Score: 87%

Conclusion

The Random Forest classifier outperformed Logistic Regression with higher accuracy, precision, recall, and F1 score. Feature importance analysis revealed significant predictors of breast cancer classification. Overall, the models show promising results in predicting breast cancer cases.

Future Work

Experiment with other classification algorithms such as Support Vector Machines, Gradient Boosting, etc.

Explore additional feature engineering techniques to improve model performance.

Gather more data to enhance the robustness of the models.

Project 3: Wine Clustering and Dimensionality reduction

Overview

This project aims to cluster wines based on their chemical properties using the K-means clustering algorithm. The dataset contains various chemical attributes of wines including alcohol content, malic acid, ash, magnesium, total phenols, flavonoids, proanthocyanins, color intensity, and more.

Data Preprocessing

Scaled Data: Applied Min-Max scaling to normalize the data.

Elbow Method: Utilized the Elbow Method to determine the optimal number of clusters by plotting the within-cluster sum of squares errors against different values of k.

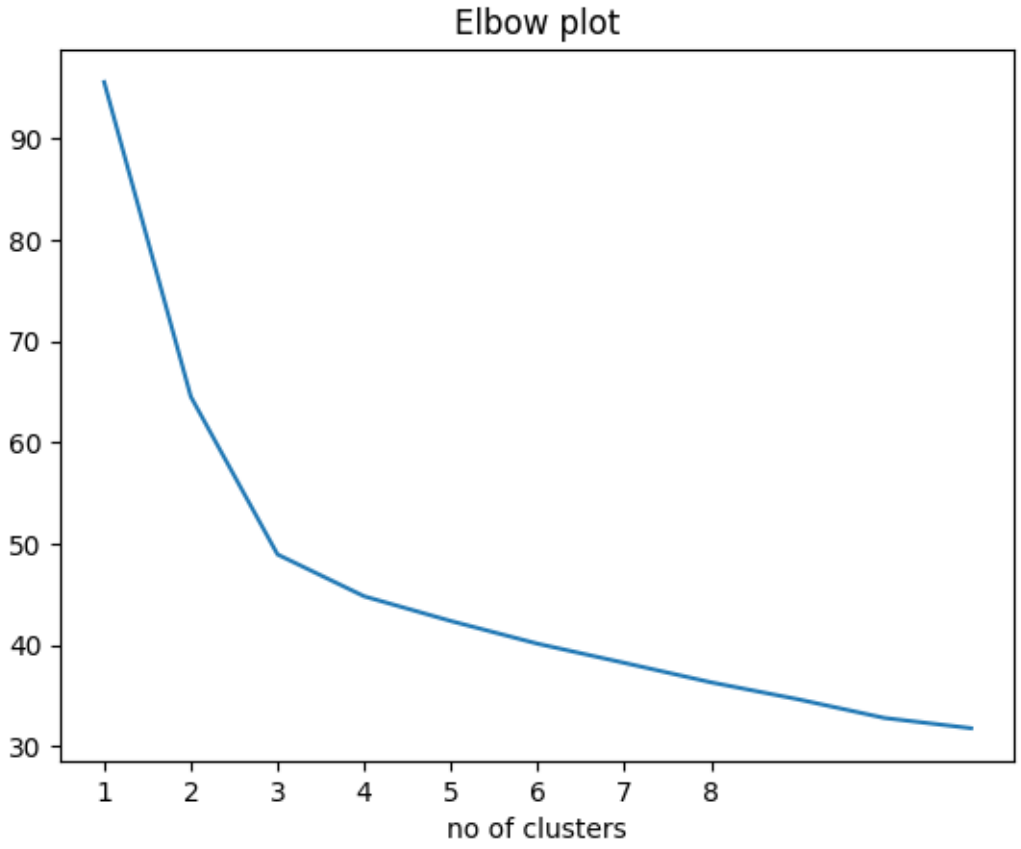
Model Building and Evaluation

K-means Clustering:

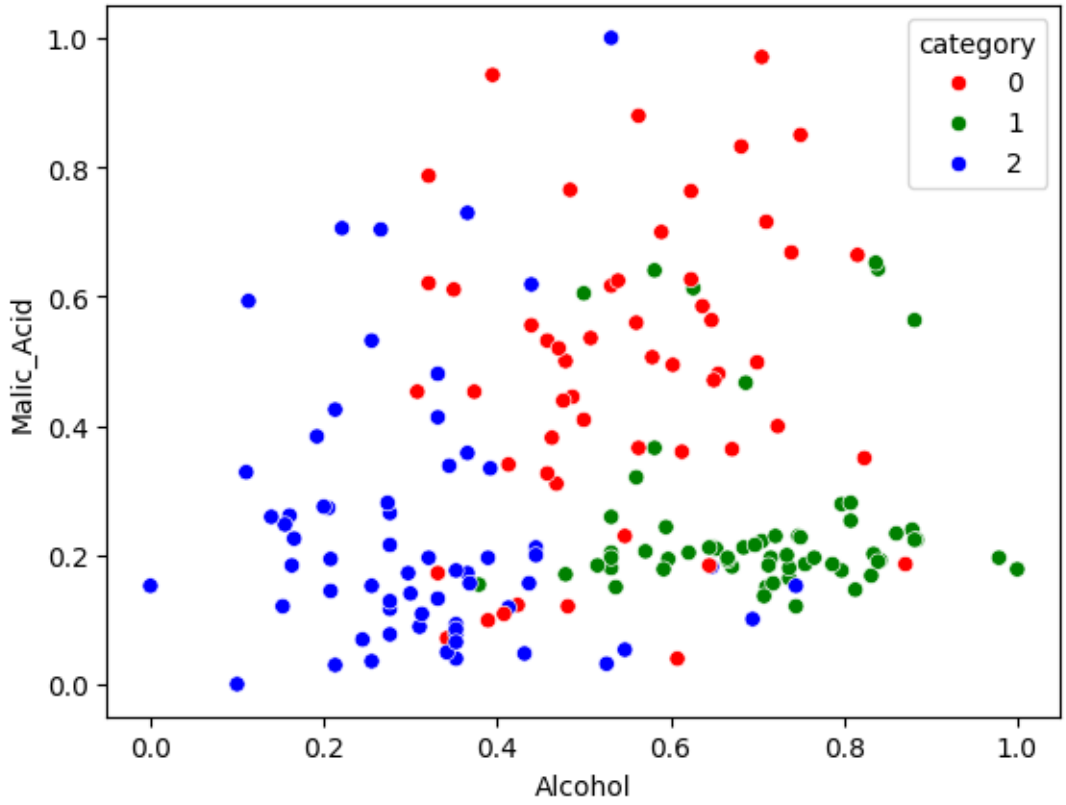
Implemented K-means clustering algorithm.

Found the optimal number of clusters using the Elbow Method.

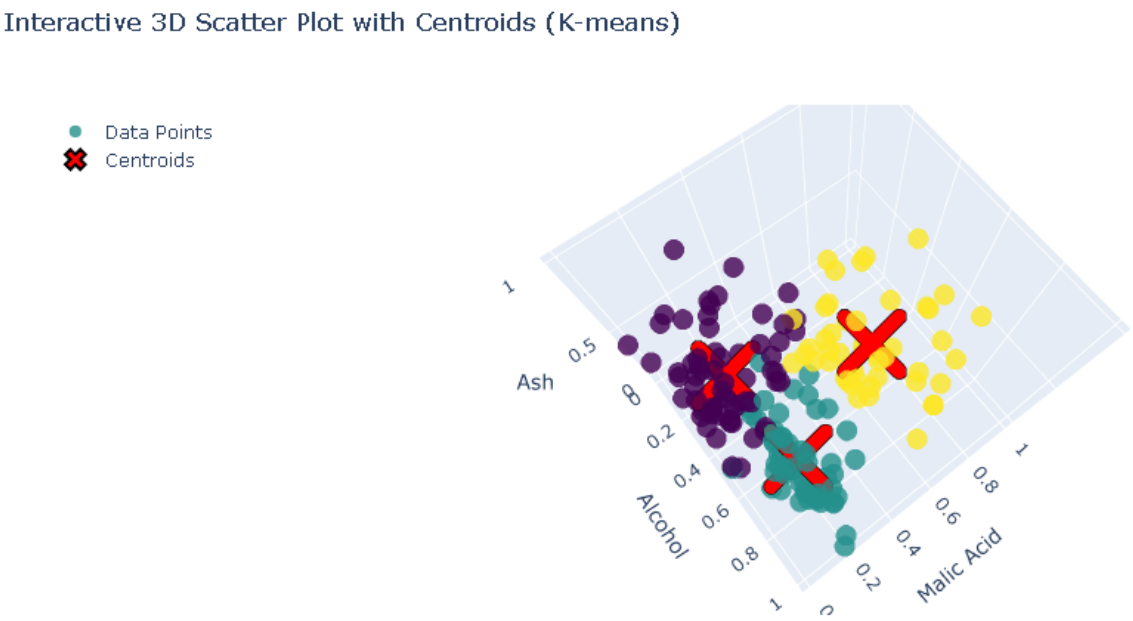
Visualized clustering results using elbow method :



Plotted 2D scatter plot using Seaborn library.



Created a 3D interactive scatter plot using Plotly library.



Dimensionality Reduction and Visualization

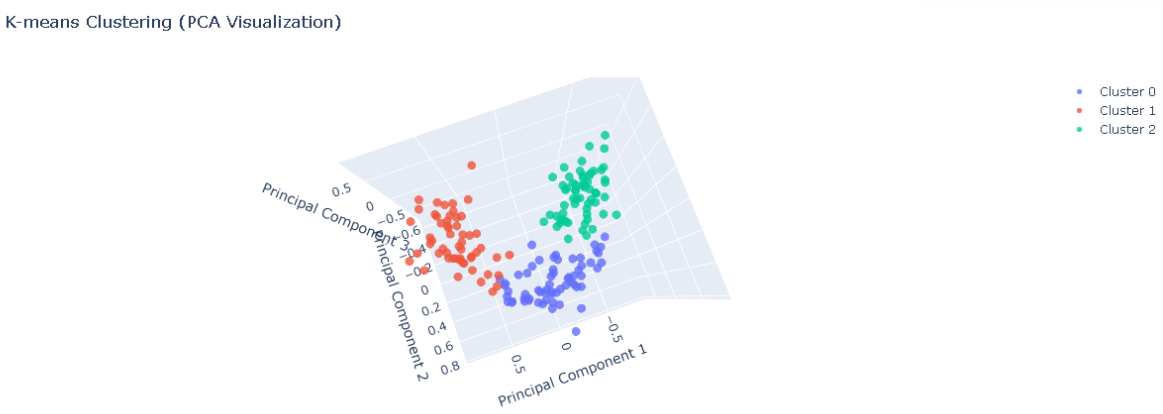
- PCA (Principal Component Analysis):
    - Performed PCA for dimensionality reduction.
    - Visualized in 3D using the first three principal components.

Interactive 3D Scatter Plot with Plotly

- Utilized Plotly library to create an interactive 3D scatter plot of the clustered data.
- Plotted the data using the first three principal components obtained from PCA.
- Each data point is colored according to its assigned cluster label.
- The interactive plot allows for zooming, panning, and hovering over data points to view additional information.

Interactive 3D Scatter Plot with Plotly

Interactive 3D Scatter Plot with Plotly



Conclusion

The K-means clustering algorithm successfully grouped wines into distinct clusters based on their chemical properties. The optimal number of clusters was determined using the Elbow Method. Visualizations provided insights into the clustering structure of the dataset.

Future Work

Experiment with different clustering algorithms such as Hierarchical Clustering, DBSCAN, etc.

Explore additional features or feature engineering techniques to improve clustering performance.

Conduct further analysis to understand the characteristics of each cluster and their implications.

This concludes the technical documentation for the Wine Clustering project. Let me know if you need further assistance or if there are additional details to include!

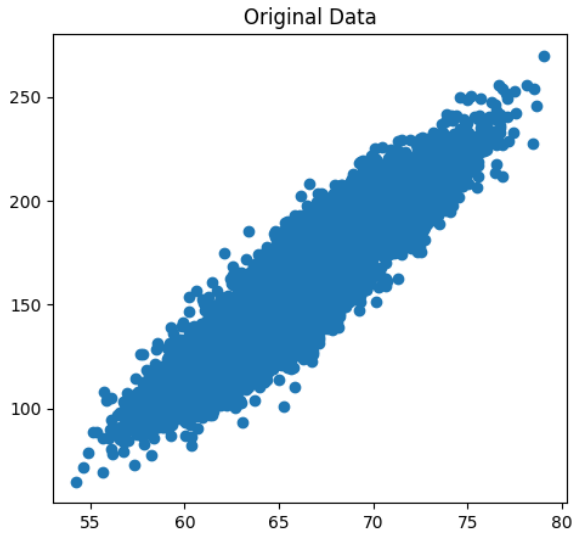
Project 4: Anomaly Detection

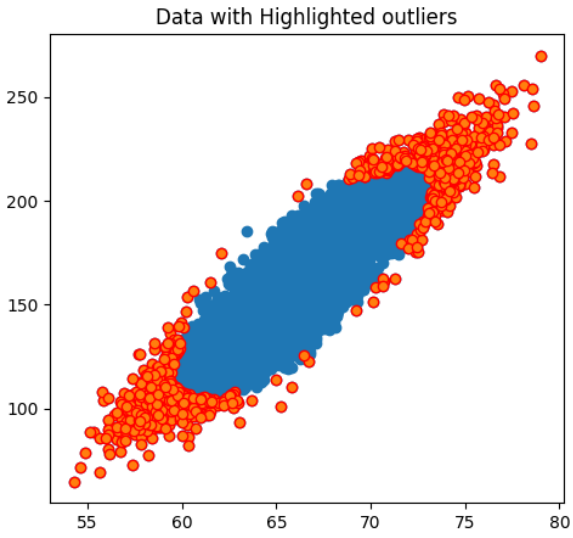
Overview

This project focuses anomaly detection using Isolation Forest and DBSCAN on two datasets: one containing height and weight data and the other containing circular data.

Isolation Forest for Anomaly Detection

- Implemented Isolation Forest algorithm directly on the height and weight dataset for Anomaly Detection.
- Isolation Forest identifies outliers (anomalies) in the data by isolating them in the tree structure.

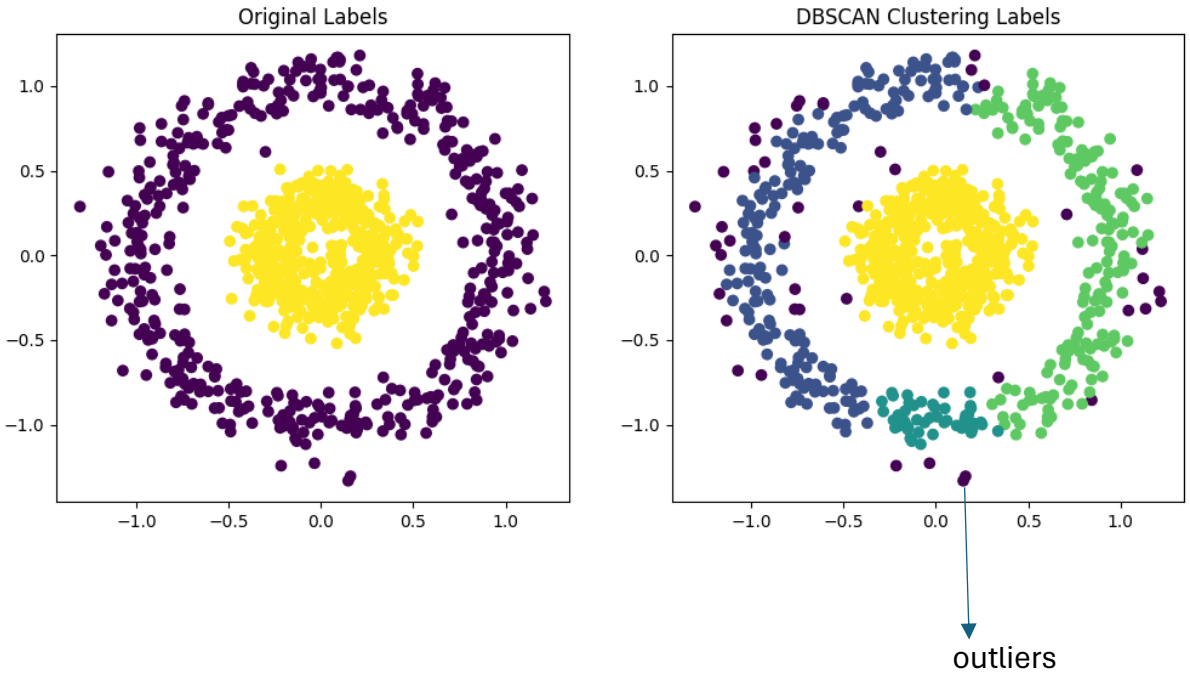




DBSCAN for Anomaly Detection on Circular Data

- Created a circular dataset for anomaly detection using DBSCAN.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that groups together closely packed points while marking points in low-density regions as outliers (anomalies).

- Visualized the anomalies detected by DBSCAN in the circular dataset. Outliers



Conclusion

- Isolation Forest anomalies in the height and weight dataset.
- DBSCAN successfully detected anomalies in the circular dataset, highlighting points that deviate from the overall circular pattern.

Future Work

- Explore different anomaly detection algorithms and assess their effectiveness on various types of data.
- Conduct further analysis to understand the characteristics of anomalies and their potential implications.