T.C.

RECEP TAYYIP ERDOGAN UNIVERSITY

FACULTY OF ENGINEERING AND ARCHITECTURE

COMPUTER ENGINEERING DEPARTMENT

Data Mining

2024-2025

Student Name Surname

Hüseyin Yaşar

201401014

Topic

Using and Analyzing Data Mining Methods

Data Sets Used

Iris

Mall Customers,

Real Estate Valuation

RİZE 2024

1. Question 1:

1.1. Introduction

This study uses the Iris dataset. The goal is to classify flowers by their leaves and compare Random Forest and Decision Tree. The dataset is balanced. Balanced datasets help to check and compare how well algorithms work.

1.1.1. Why Iris Dataset?

The Iris dataset was chosen for this study because it is simple, easy to understand, and used a lot in machine learning problems. It is a good choice for learning and testing machine learning methods.

1.1.2. Why Decision Tree?

- The Decision Tree algorithm gives quick results in training and prediction, making it useful for small datasets.
- Its visual structure makes it easy to understand and explain.
- The Decision Tree is simple, easy to use, and easy to understand because of its low complexity.

1.1.3. Why Random Forest?

- It combines predictions from many trees to give balanced and reliable results.
- It is not affected by noise in the dataset.
- It provides higher classification accuracy compared to a Decision Tree.

1.2. Dataset: Iris

The dataset was created by Ronald A. Fisher. It includes the classification of different flower types. The dataset has 3 classes and a total of 150 samples. Each class has the same number of samples, making it a balanced dataset. It has 4 features. [1]

1.3. Data Splitting Steps

Since the Iris dataset is small, I split the data into 70% training set and 30% test set.

In this study, the Decision Tree algorithm was used to classify flower types in the Iris dataset. The performance of the Decision Tree model was evaluated using metrics like Accuracy, Precision, Recall, and F1-Score.

- **Accuracy**: The model correctly classified 91.11% of the test data. This shows that the model's overall performance is high.
- **Precision**: The model made 91.55% correct positive predictions. This means the model's predictions are reliable.
- **Recall**: The model correctly identified 91.11% of the actual classes. This shows that the model has a low number of false negatives.
- **F1-Score**: The balance between Precision and Recall is 91.07%. This shows that the model is balanced in both accuracy and reliability.

The prediction performance of the Decision Tree algorithm is shown in **Figure 1**.

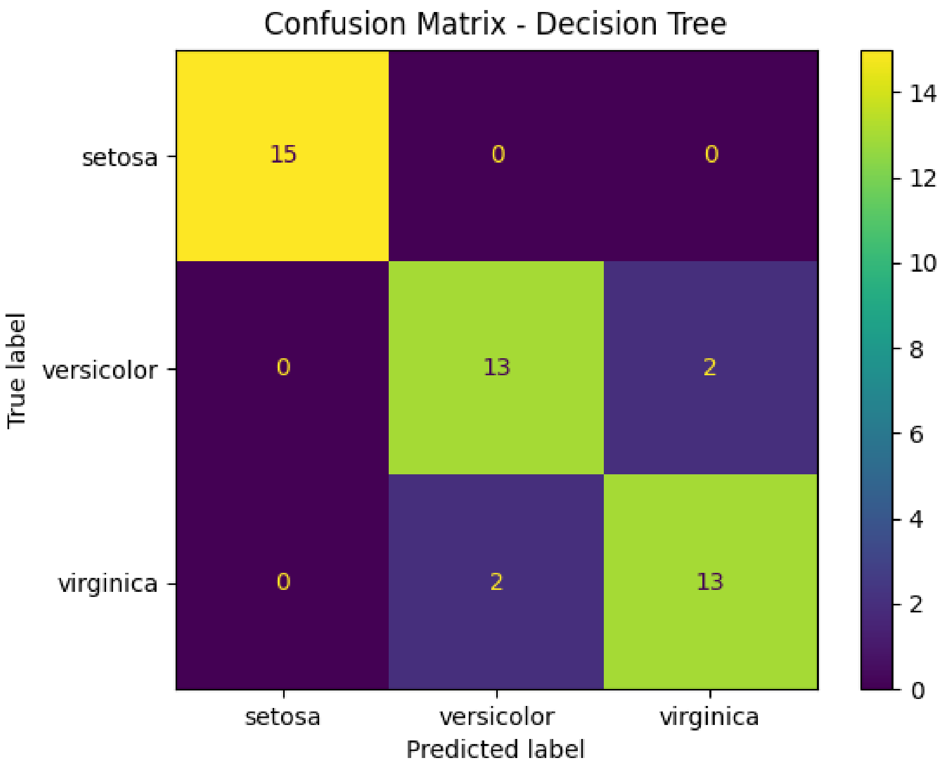


Figure 1: Confusion Matrix - Decision Tree

- **Setosa**: All examples (15) in this class were correctly predicted.
- **Versicolor**: Out of 15 examples, 13 were correctly predicted. 2 examples were classified as "Virginica."

- **Virginica**: Out of 15 examples, 13 were correctly predicted. 2 examples were classified as "Versicolor."

The Decision Tree correctly predicted the Setosa class clearly, but there was confusion between the Versicolor and Virginica classes. This is because the features of these classes are very similar. The structure of the Decision Tree is shown in **Figure 2.**

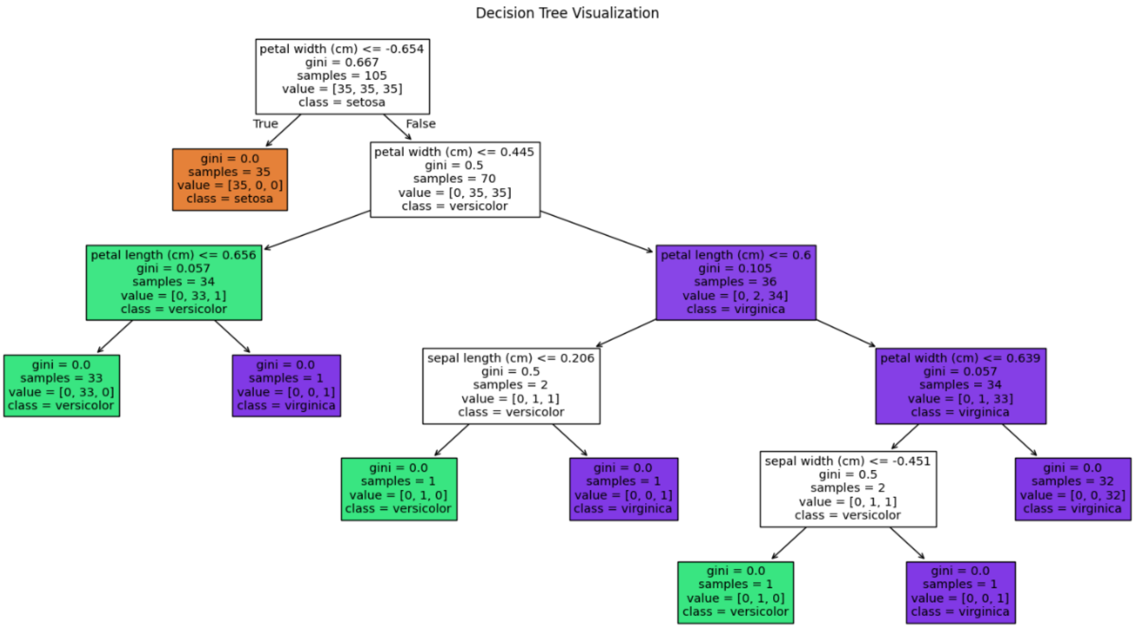


Figure 2: Decision Tree Visualization

The Decision Tree algorithm showed very successful performance on a small dataset like the Iris dataset. The model was clearly successful in distinguishing the Setosa class.

1.4. Random Forest Classifier

In this study, the Random Forest algorithm was used on the Iris dataset. The model showed good performance in classification. However, its accuracy is slightly lower compared to the Decision Tree.

- **Accuracy**: The model correctly classified 88.89% of the test data.
- **Precision**: The model made 89.31% correct positive predictions.
- **Recall**: The model correctly identified 88.89% of the actual classes.
- **F1-Score**: The model showed a performance of 88.77% between Precision and Recall.

The prediction performance of the Random Forest algorithm is shown in **Figure 3**.

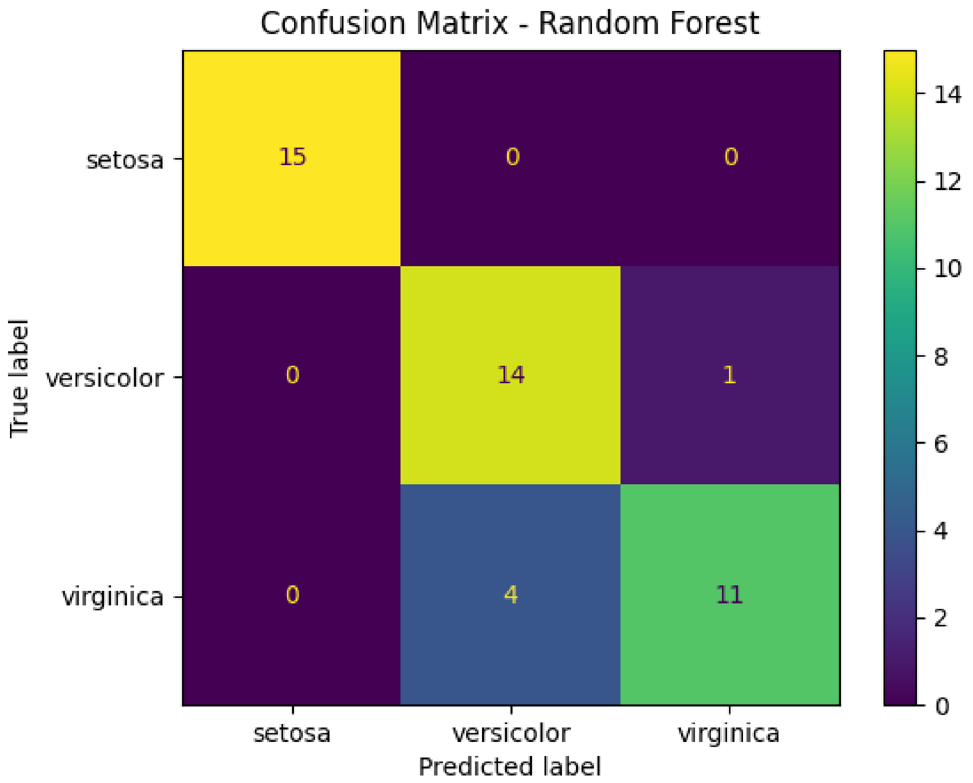


Figure 3: Confusion Matrix - Random Forest

- **Setosa**: All examples (15) in this class were correctly predicted.
- **Versicolor**: Out of 15 examples, 14 were correctly predicted. 1 example was classified as "Virginica."
- **Virginica**: Out of 15 examples, 11 were correctly predicted. 4 examples were classified as "Versicolor."

Random Forest is a model that combines multiple decision trees, making it a powerful algorithm for generalization. However, since this study uses a small dataset, the Random Forest algorithm is not very suitable for this dataset.

1.5. Comparison of Decision Tree and Random Forest

The Decision Tree is a model that works quickly and effectively on small datasets. Its visual structure makes it easy to understand. Random Forest usually performs better on larger datasets. However, in this study, due to the simple structure of the dataset, the Decision Tree gave better results. The Decision Tree achieved an accuracy of 91.11%, while Random Forest reached an accuracy of 88.89%.

2. Question 2:

2.1. Introduction

In this study, the Mall Customers dataset was used. The K-Means clustering algorithm was applied to evaluate features such as age, annual income, and spending score in the dataset.

2.1.1. Why Mall Customers Dataset?

The Mall Customers dataset was chosen in this study because it is clean, small, and easy to process.

2.1.2. Why K-Means?

The K-Means algorithm is a simple and understandable algorithm based on Euclidean distance. Its simple structure allows it to produce quick results on small and medium-sized datasets. Additionally, the results are practical for visualization and interpretation. [2]

2.1.3. Why Elbow Method?

- It is a good method to determine the number of clusters.
- By calculating WCSS for different K values, it shows the optimal separation between clusters.

2.2. Dataset: Mall Customers

This dataset includes spending score data derived from shopping mall customers' gender, age, annual income, and spending amount or frequency. The Mall Customers dataset contains a total of 200 samples. Since there is no missing or lost data, the dataset is balanced. [3]

2.2.1. Features

- **CustomerID**: A unique identification number assigned to customers. It cannot be used for clustering analysis as it does not provide meaningful data.
- **Gender**: A demographic variable that can take values Male or Female and can be used to understand customers' shopping habits.

- **Age**: A numerical variable that helps in segmenting customers based on their age groups.
- **Annual Income**: A variable representing the customer's annual income (in thousand dollars) and allows segmentation based on income levels.
- **Spending Score**: A score that evaluates the customer's shopping frequency, intensity, and loyalty, ranging from 1 (very low) to 100 (very high).

2.2.2. Choosing K

To ensure the clustering algorithm works correctly, it is important to determine the ideal number of clusters. We will use the Elbow Method and Silhouette Score to find the ideal number of clusters.

2.2.2.1 Elbow Method

The Elbow Method calculates WCSS values and identifies an "elbow point" on the decreasing curve, which indicates the optimal number of clusters. The graph needed to select K is shown in **Figure 4**.

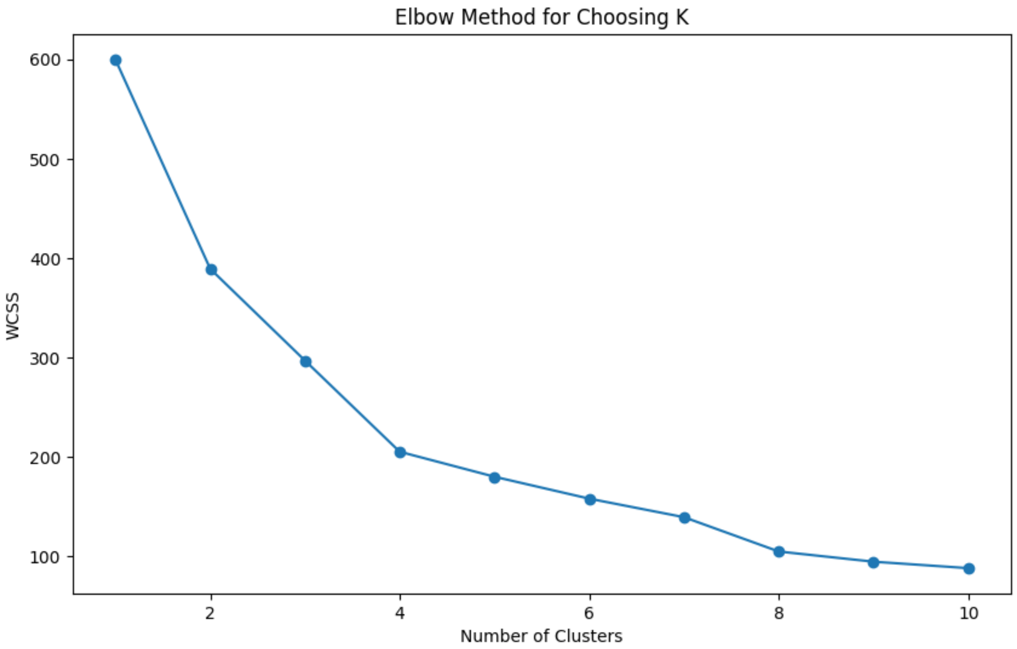


Figure 4: Elbow Method for Choosing K

The elbow point is often based on interpretation, but it is usually the first point where the rapid decrease slows down. In the graph shown in Figure X, the first point where the WCSS decrease slows is at 4. To make a more accurate choice for K, we will also look at the Silhouette Score graph, as it is better to consider different methods.

2.2.2.2. Silhouette Score

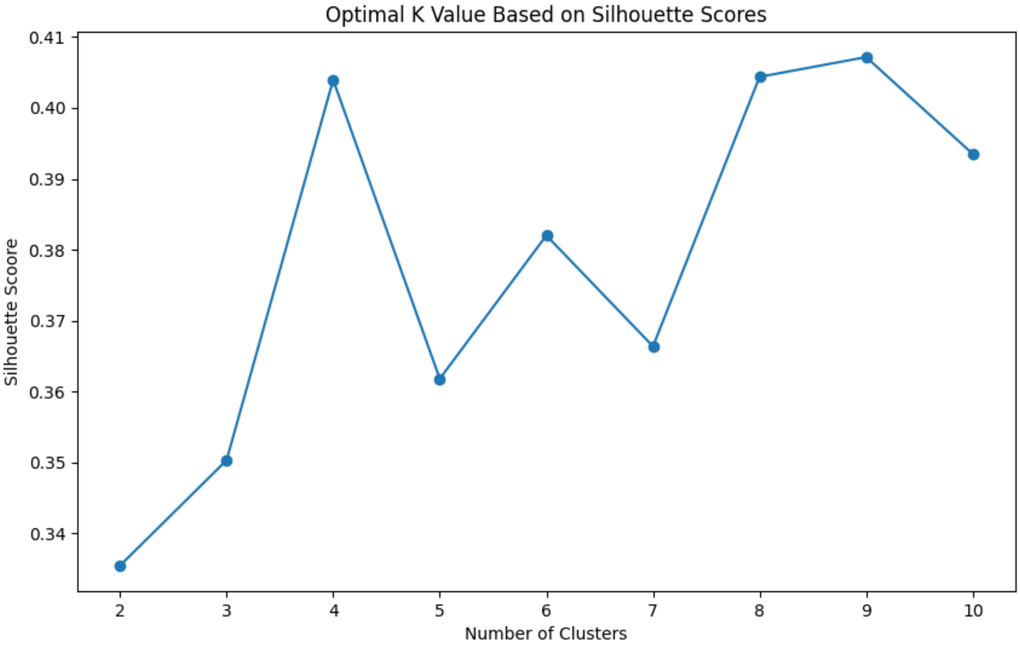


Figure 5: Optimal K Value Based on Silhouette Scores

In the Silhouette Score graph shown in **Figure 5**, the K value with the highest score is selected, but other methods should also be considered. Although the highest value is at K=9, considering the Elbow Method graph, the best choice is K=4. It is also clear that there is not much difference between the scores of K=4 and K=9.

2.3. Result

Both graphs indicate that the optimal K value is 4. Although the Silhouette Score shows maximum performance at K=9, considering the Elbow Method graph, K=4 was the correct choice.

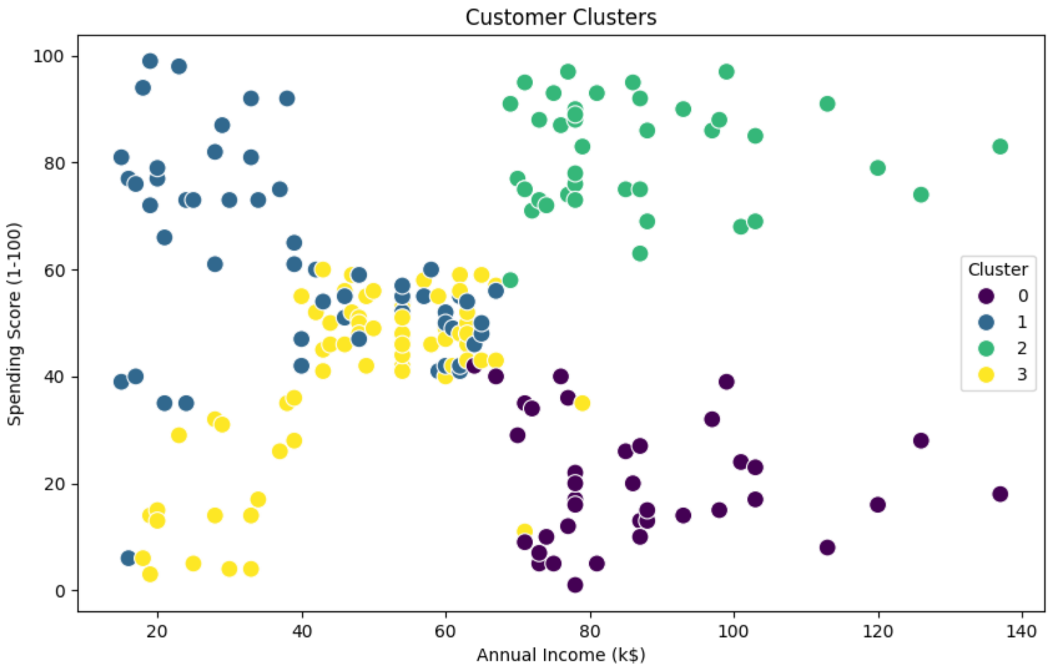


Figure 6: Customer Clusters

The graph in **Figure 6** visualizes customer clusters based on Annual Income and Spending Score features. Each color represents a different cluster.

- **Cluster 0**: This group has high income ($80k-$140k) but a low spending score (0-40).
- **Cluster 1**: This group has low income ($20k-$40k) but a high spending score (60-100).
- **Cluster 2**: This group has high income ($70k-$140k) and a high spending score (60-100).
- **Cluster 3**: This group has medium income ($40k-$80k) and a medium spending score (40-60).

In this study, customer segmentation was performed using the Mall Customers dataset and the K-Means algorithm. Customers were divided into meaningful clusters based on annual income, age, and spending score. The optimal number of clusters was determined using the Elbow Method and Silhouette Score, and the results were visualized and interpreted.

3. Question

3.1. Introduction

In this study, the "Real Estate Valuation" dataset was used. Linear, Ridge, and Lasso regression models were compared. Predictions made using different housing features in the dataset were evaluated using metrics such as MAE, MSE, and R², and the approach with the best performance was identified.

3.1.1 Why Real Estate Valuation Dataset?

The Real Estate Valuation dataset was chosen in this study because it is commonly used in machine learning problems and has a simple and easy-to-understand structure. This makes it easy to perform data preprocessing, model building, and evaluation.

3.2. Dataset: Real Estate Valuation

This dataset includes features such as age, distance to the metro, and the number of markets, along with the price per unit area of houses in Taiwan. The dataset contains 414 samples and 7 features.

3.2.1. Data Splitting Steps

The dataset was split into 60% training, 20% validation, and 20% test sets.

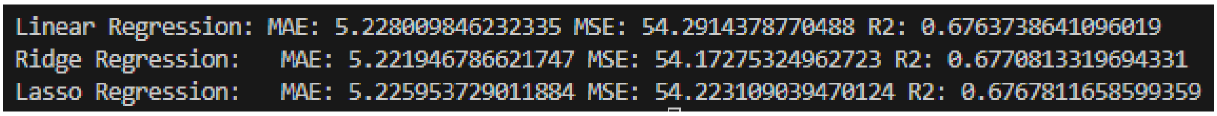
To ensure reproducible results and prevent overfitting, I set the `random_state` value to a fixed number.

3.2.2. Features

- **Transaction_Date**: Indicates the transaction date of the real estate.
- **House_Age**: Specifies the age of the building.
- **Distance_MRT**: Indicates the distance (in meters) from the real estate to the nearest metro station.
- **Convenience_Stores**: Represents the number of convenience stores around the real estate.
- **Latitude**: Indicates the latitude of the real estate's geographic coordinates.
- **Longitude**: Indicates the longitude of the real estate's geographic coordinates.
- **Price**: Target variable; specifies the price per unit area of the real estate (1000 Taiwan dollars/m²).

3.3. Result

The model's result is shown in Figure 7.



```
Linear Regression: MAE: 5.228009846232335 MSE: 54.2914378770488 R2: 0.6763738641096019
Ridge Regression:   MAE: 5.221946786621747 MSE: 54.17275324962723 R2: 0.6770813319694331
Lasso Regression:   MAE: 5.225953729011884 MSE: 54.223109039470124 R2: 0.6767811658599359
```

Figure 7: Output

3.3.1. Linear Regression

In this study, the Linear Regression model was used as a basic approach to predict housing prices in the Real Estate Valuation dataset. The model aimed to capture the linear relationships between the independent variables in the data and the target variable (Price). According to the results, the Linear Regression model demonstrated basic prediction performance in terms of Mean Absolute Error (MAE) and Mean Squared Error (MSE).

- **MAE**: The predicted prices deviate from the actual prices by an average of 5.23 units. This error can be considered acceptable.
- **MSE**: The mean of the squared errors is 54.29, indicating that larger errors have a significant impact on the model.
- **R²**: The model explained 67.67% of the variance in the target variable. This shows that the linear relationships in the dataset were reasonably captured by the model.

3.3.2. Ridge Regression

In this study, the Ridge Regression model was applied to predict housing prices in the Real Estate Valuation dataset. Ridge Regression aims to prevent overfitting by using a regularization method.

According to the results, the Ridge Regression model showed some improvement with slightly lower error values and a higher R² score compared to Linear Regression.

- **MAE**: The predicted prices deviate from the actual prices by an average of 5.221 units. It produced lower errors compared to Linear Regression. This shows that the regularization method reduced overfitting and provided better generalization.
- **MSE**: The mean of the squared errors is 54.17. The mean squared error is smaller than that of Linear Regression, indicating better generalization by the model.

- **R²**: The model explained 67.7% of the variance in the target variable. This shows that it was more successful in explaining the target variable compared to Linear Regression.

3.3.3. Lasso Regression

In this study, the Lasso Regression model was used to predict housing prices in the Real Estate Valuation dataset. Lasso Regression aims to create a simpler and more interpretable model by using a regularization method that reduces the coefficients of some features to zero.

- **MAE**: The predicted prices deviate from the actual prices by an average of 5.225 units. The values are very close to those of the other models.
- **MSE**: The mean of the squared errors is 54.22. The values are very close to those of the other models.
- **R²**: The model explained 67.67% of the variance in the target variable. Similar results were obtained compared to the other models.

Both Ridge and Lasso models, with their regularization methods, provided better generalization by showing lower error values and slightly higher R² scores compared to Linear Regression.

The relationship between Distance and Price is shown in the graph in Figure X.

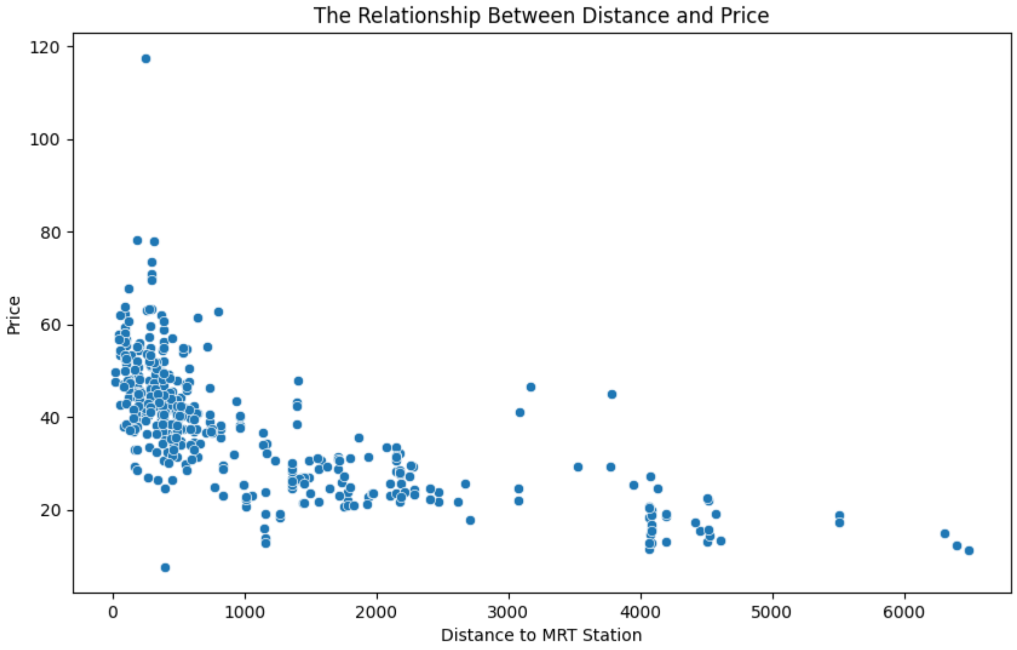


Figure 8: The Relationshi Between Distance and Price

In the graph, it can be seen that as the distance to the MRT station increases, the prices decrease, which supports the hypothesis that proximity to public transportation increases the value of real estate.

4. References

[1] Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7(2), 179–188.

[2] Takaoğlu, M., & Takaoğlu, F. (2023). K-Means ve Hiyerarşik Kümeleme Algoritmasının WEKA ve MATLAB Platformlarında Karşılaştırılması.

[3] Aslantaş, G., Gençgül, M., Rumelli, M., Özsaraç, M., vd. (2023). Customer Segmentation Using K-Means Clustering Algorithm and RFM Model. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen Ve Mühendislik Dergisi, 25(74), 491-503.