MACHINE LEARNING PROJECT FINAL REPORT

**Baran BAHTİYAR 19066449920**

**1** Computer Engineering, Engineering Faculty, Eskisehir Technical University, Eskisehir, Turkiye

**Burak DOĞANAY**. **30998066696**

**2** Computer Engineering, Engineering Faculty, Eskisehir Technical University, Eskisehir, Turkiye.

ABSTRACT

In this project, we are asked to experiment with a real-world dataset and use machine learning algorithms to find patterns in the data. The aim of this assignment is to determine the model that gives the most accurate prediction and to make the best prediction using this model. For this, we found and organized a suitable data set with suitable features and real values. While editing our data set, we performed operations such as deleting outliers and variable transformation. We researched the most suitable algorithm for house price prediction and decided that it is linear regression and decision tree regression. That's why we trained our data set with these methods and obtained two separate models. We evaluated the performance of the models using metrics such as mean absolute error (MAE), mean square error (MSE), and variance score. Finally, we wrote a code that takes the necessary appropriate values ​​to use this model as input and outputs the house price prediction.

**Keywords:** dataset, linear regression, model, train

1. INTRODUCTION

For a machine learning project to be successful, it is very important to choose the right data set. Features such as the number of bathrooms and beds, the size of the house and the city where the house is located are very important in determining the price of a house, so we used a data set with these features. We used methods such as outlier handling and variable transformation when editing this data set. We deleted extreme values ​​that could mislead the algorithm and converted categorical values ​​into numerical values.

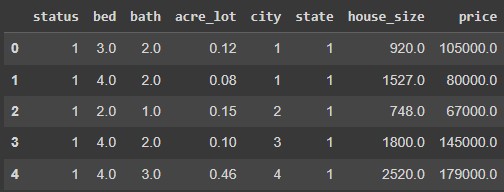
The next step in the project was to choose the right algorithm. The most important factor here is that the algorithm is suitable for the subject. Since our project topic is house price prediction, we chose the algorithm as linear regression. We trained our data set with the linear regression method and obtained a model that can predict house prices. The last stage in the project is to write code that can run this model. We wrote a code that takes the necessary values ​​to run this model as input and gives the house price appropriate to these values ​​as output.

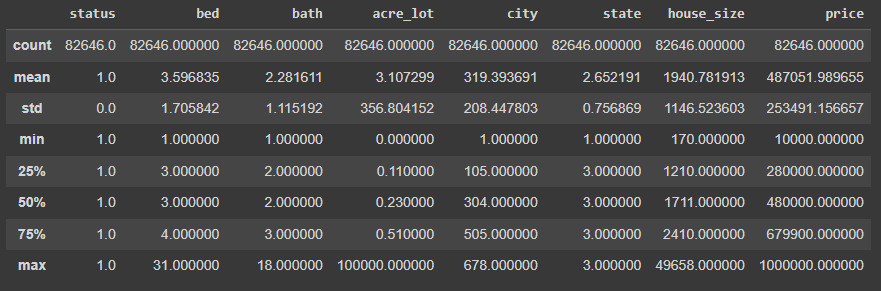
2. MAIN PART OF PROJECT

In this section, we will explain the parts of our project in detail.

2.1. Data Set

The first stage of a machine learning project is to select a dataset. It is very important to choose the right data set because the data set forms the basis of the project. The most important factor in choosing the right data set is that it is suitable for the project subject. Since the subject of our project is house price prediction, we chose data that includes the number of beds, number of bathrooms, acre lot size, city, state and house size.





2.1.1. Feature Engineering

We used data transformation and data standardization techniques when organizing our data set.

2.1.1.1 Data Transformation

This involves converting the data into a format that is suitable for analysis. This may involve changing the data type, scaling the data, or normalizing the data. So, we converted categorical data into numerical values.

**2.1.1.2 Data Standardization**

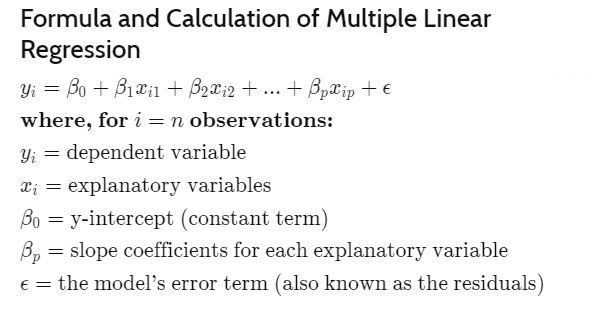
This involves bringing the data into a consistent format. This may involve converting the data to a common unit of measurement or converting the data to a common time format. Accordingly, we deleted the extreme values ​​in the data set.

2.2 Algorithms

We researched the algorithms that would be most suitable for our topic to use in the project. As a result of our research, we have established multiple linear regression and decision tree regression algorithms.

2.2.1 Multiple Linear Regression

Multiple linear regression is a statistical method used to analyze the relationship between one or more independent variables and a single dependent variable. Unlike single linear regression, which considers only one independent variable, multiple linear regression takes more than one predictor variable into account.



(Figure 1)

Figure 1: Formulization of Multiple Linear Regression

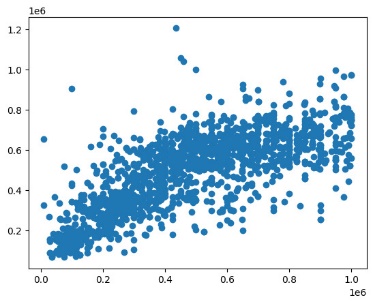
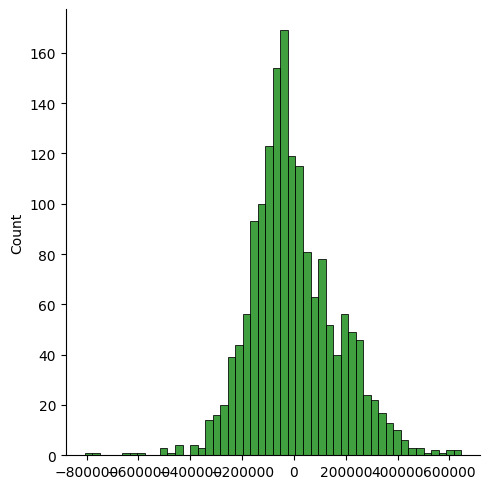
When we trained our model with this method, we obtained these results:

Mean absolute error = 129462.66

Mean squared error = 28356146210.52

Median absolute error = 102977.84

Explain variance score = 0.56

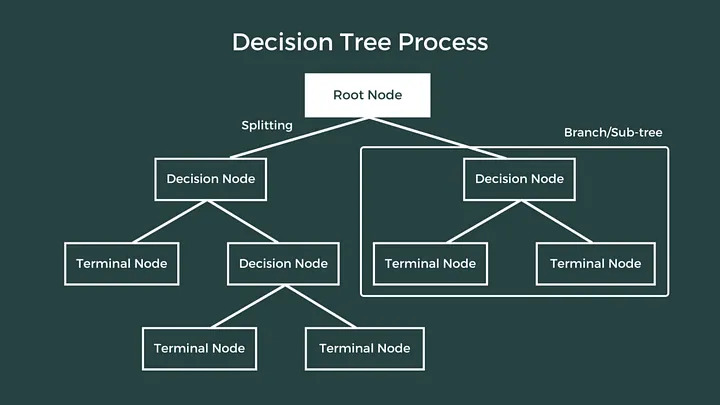
(Figure 2) (Figure 3)

Figure 2: The result graph of the multiple linear regression model we trained.

Figure 3: The error chart of the multiple linear regression model we trained.

2.2.2 Decision Tree Regression

Decision Tree Regression is a machine learning algorithm used for predicting a continuous target variable based on one or more input features. Unlike classification trees that predict categorical outcomes, decision tree regression models are designed for numeric or continuous predictions.



(Figure 4)

Figure 4: Decision tree process schema

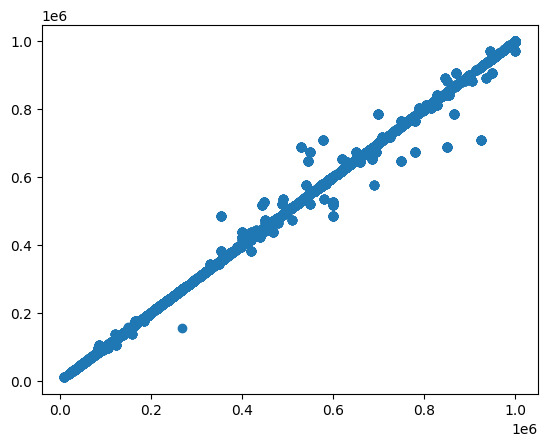
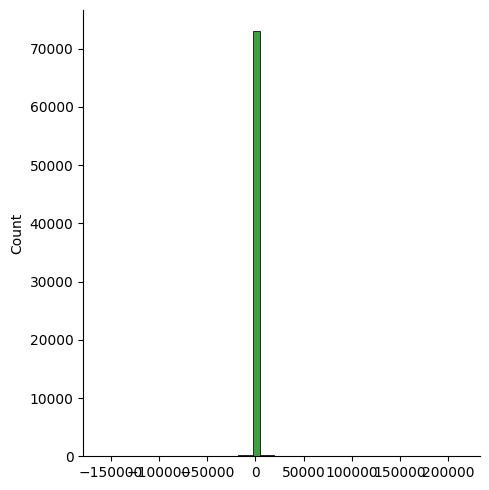
When we trained our model with this method, we obtained these results:

Mean absolute error = 3663.42

Mean squared error = 838976923.14

Median absolute error = 0.0

Explain variance score = 0.99

(Figure 5) (Figure 6)

Figure 5: The result graph of the decision tree regression model we trained.

Figure 6: The error chart of the decision tree regression model we trained.

2.3 Libraries

In this section, we explained the libraries that we used for data set manipulation and model creation.

2.3.1 Joblib

Joblib is a set of tools to provide lightweight pipelining in Python. Joblib is optimized to be fast and robust on large data in particular and has specific optimizations for numpy arrays

2.3.2 Pandas

Pandas is a fast, powerful, flexible and easy to use open-source data analysis and manipulation tool,

built on top of the Python programming language.

2.3.3 Sklearn

Scikit-learn provides a wide range of tools for working with machine learning algorithms, including regression models, classification models, clustering, dimensionality reduction, and more.

2.3.4 Seaborn

A statistical data visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics.

2.3.5 NumPy

A library for numerical operations in Python, providing support for large, multi-dimensional arrays and matrices.

2.3.6 google.colab

İt is a module specifically designed for Google Colab (Colaboratory), the cloud-based Jupyter notebook environment provided by Google. It provides functionality that is specific to the Colab environment, allowing users to interact with Google Colab features and settings.

2.3.7 Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

**CONFLICT OF INTEREST**

There is no conflict of interest between the authors of this article.

The entire project was carried out collaboratively, without sharing tasks.

**ETHICAL STATEMENT**

While preparing the project, We did not take any action that could be considered as plagiarism.

**AI DECLERATION**

It helped us pay attention to spelling rules in report writings. It helped us solve errors we encountered while coding. It helped us find the necessary libraries. We took ideas while preparing the report.

**REFERENCES**

[1] <https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data>

[2] <https://www.investopedia.com/terms/m/mlr.asp>

[3] <https://www.udemy.com/course/machinelearning/>

[4] <https://joblib.readthedocs.io/en/stable/>

[5] <https://pandas.pydata.org/>

[6] <https://scikit-learn.org/stable/>

[7] <https://www.w3schools.com/python/numpy/numpy_random_seaborn.asp>

[8] <https://www.w3schools.com/python/numpy/numpy_intro.asp>

[9] <https://www.w3schools.com/python/matplotlib_pyplot.asp>