Name	id

House Price India weka project

Report Document will contain :

*Introduction

*Dataset Preprocessing

*Classifier Evaluation

*Visualization Analysis

*Classifier Comparison

*Conclusion

*Introduction

1-dataset description:

Link:https://www.kaggle.com/datasets/mohamedafsal007/house-price-dataset-of-india/code?select=House+Price+India.csv

The "House Price India" dataset is a collection of data on house prices in India. The dataset includes information on various features of the houses, such as the number of bedrooms, bathrooms, living area, lot area, and whether or not the house has a waterfront view. The data was collected and uploaded to Kaggle by a user named Mohamed Afsal.

The dataset is in CSV format and includes 9 columns and 1,460 rows. The columns are as follows:

1. Id: A unique identifier for each house listing
2. Date: The date on which the house was listed for sale
3. Bedrooms: The number of bedrooms in the house
4. Bathrooms: The number of bathrooms in the house
5. Living Area: The area of the living space in square feet
6. Lot Area: The area of the lot in square feet
7. Water Front Present: A binary indicator (0 or 1) that indicates whether the house has a waterfront view
8. Price: The price of the house in Indian Rupees (INR)
9. Location: The location of the house in India

*Project Description: House Price Prediction using Weka on the "House Price India" dataset

Introduction:
The "House Price India" dataset is a collection of data on house prices in India that includes information on various features of the houses, such as the number of bedrooms, bathrooms, living area, lot area, and whether or not the house has a waterfront view. The dataset can be used to train machine learning models to predict the price of a house in India based on its features.

Objective:
The objective of this project is to build a machine learning model using Weka that can accurately predict the price of a house in India based on its features.

*Problem Statement: House Price Prediction using Weka on the "House Price India" dataset

The problem that we will solve is to predict the price of a house in India based on its features such as the number of bedrooms, bathrooms, living area, lot area, and whether or not the house has a waterfront view. This is a regression

problem, where we need to predict a continuous value (the price of the house) based on a set of input features.

To solve this problem using Weka, we will follow the following steps:

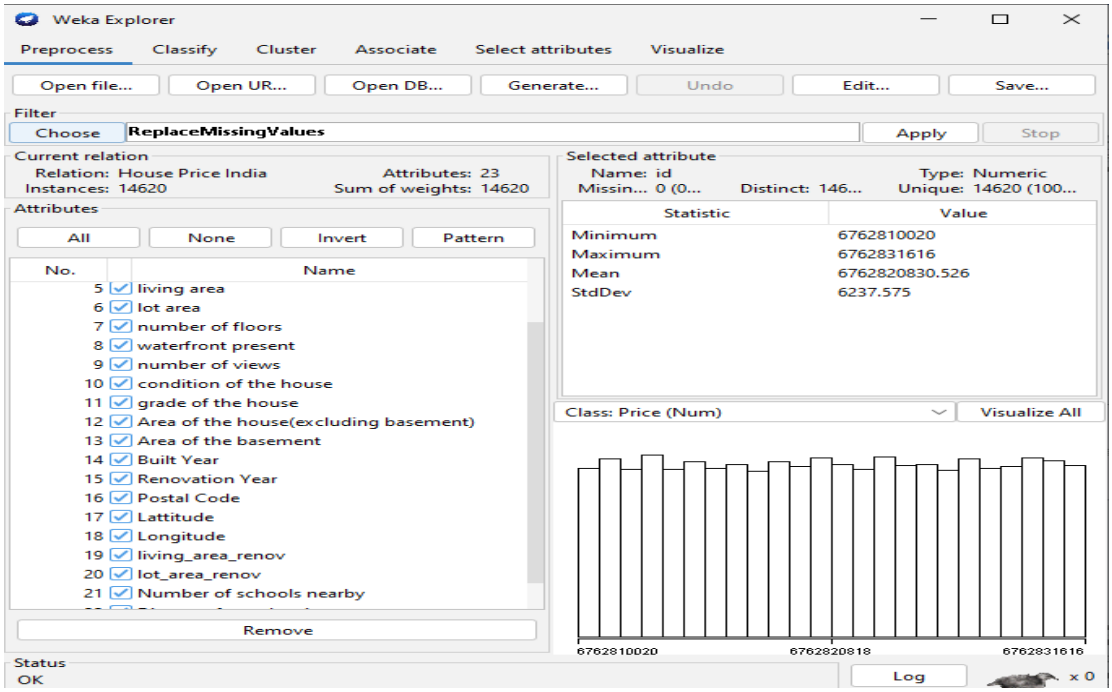
1. Data Preprocessing: We will preprocess the "House Price India" dataset using various filters in Weka such as handling missing values, data normalization, data discretization, feature selection, data transformation, and handling categorical attributes.
2. Model Selection: We will train and evaluate different regression models in Weka such as Linear Regression, Decision Tree, Random Forest, and Support Vector Machine,knn. We will choose the best-performing model based on evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).

*Data preprocessing steps:

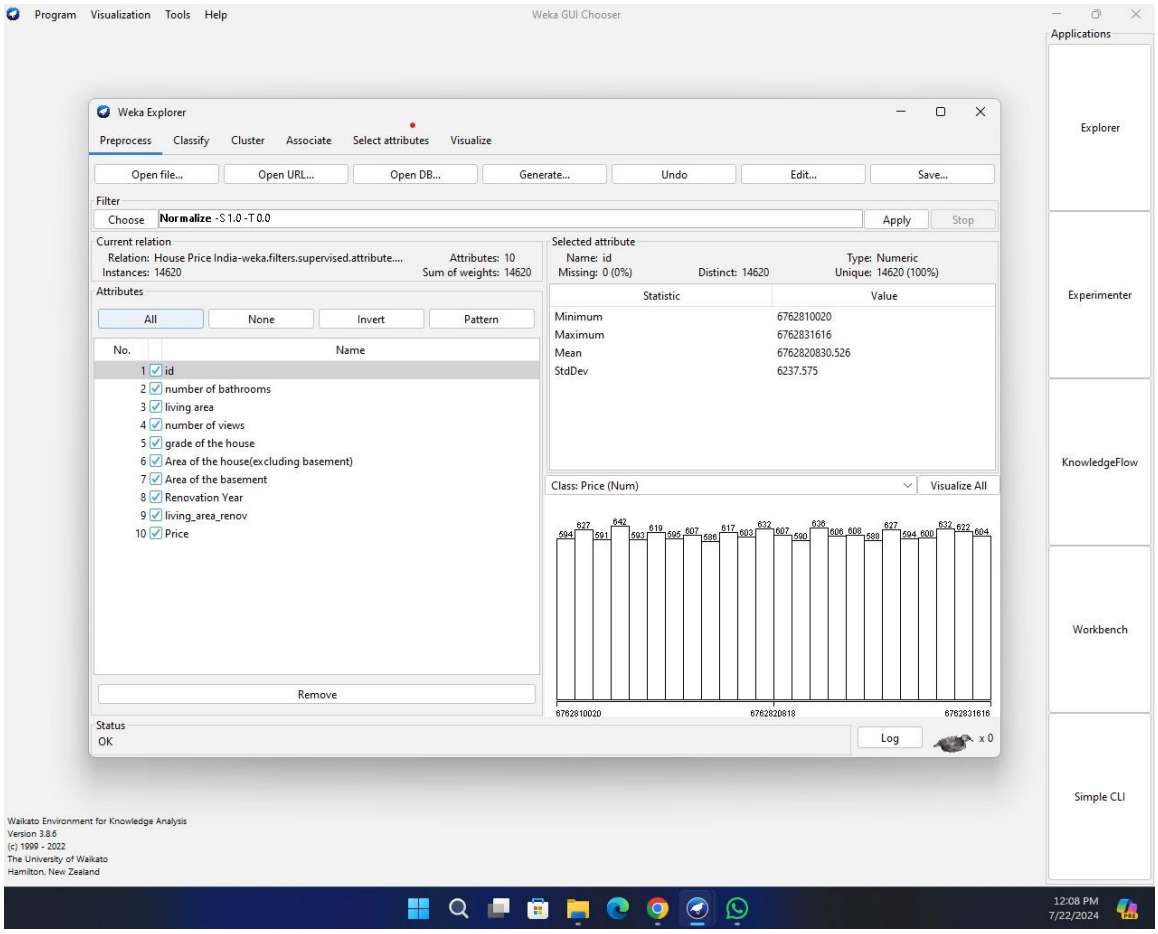
Here are some data preprocessing steps that can be performed on the "House Price India" dataset using Weka:

1-Handling missing values: The dataset may contain missing values, which can be handled by replacing them with suitable values such as the mean or median of the column, or by removing the rows or columns with missing values altogether. In Weka, you can use the

"ReplaceMissingValues" filter to replace missing values with the mean or median of the column



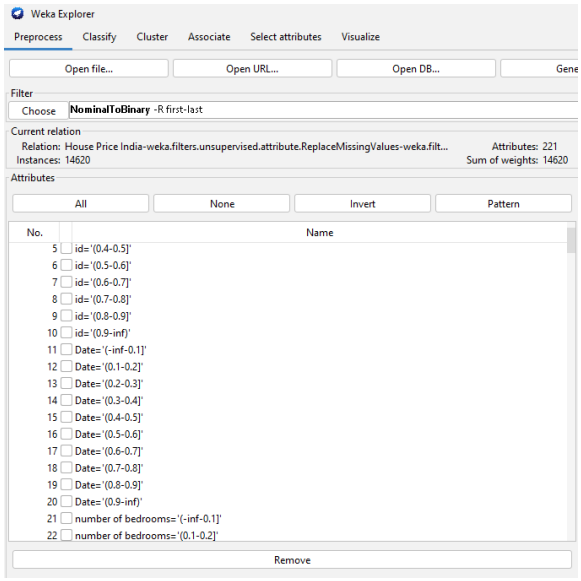
2-Normalize the data



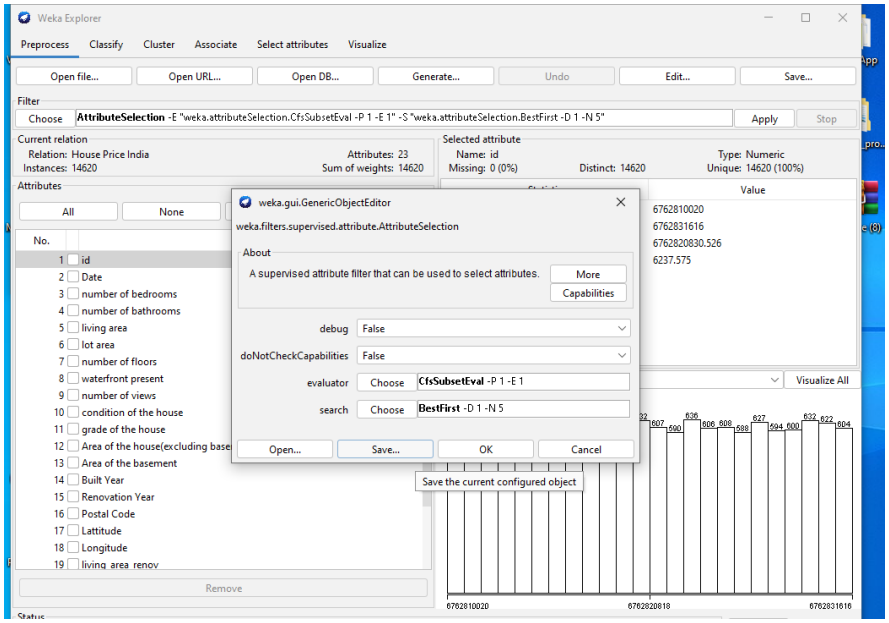
3-select important feature

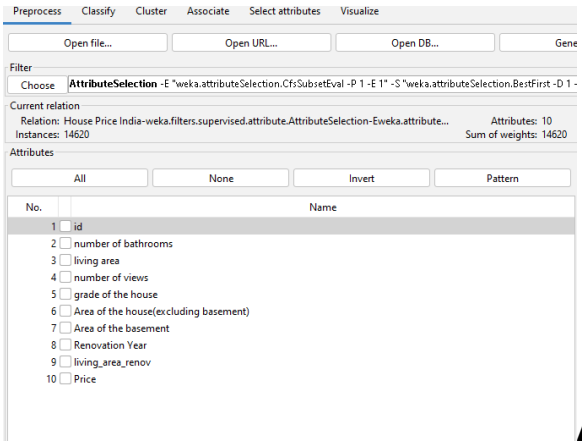
1. Click on the "Preprocess" tab in the top menu.
2. Click on the "Filter" drop-down menu and select "AttributeSelection".
3. In the "AttributeSelection" window, select the search method you want to use to select important features, such as "Ranker" or "CfsSubsetEval".

4. Set any other options for the search method you chose, such as the number of top-ranked features to keep.
5. Click "Apply" to run the filter and select the important features.
6. Save the new dataset with the important features selected by the filter.



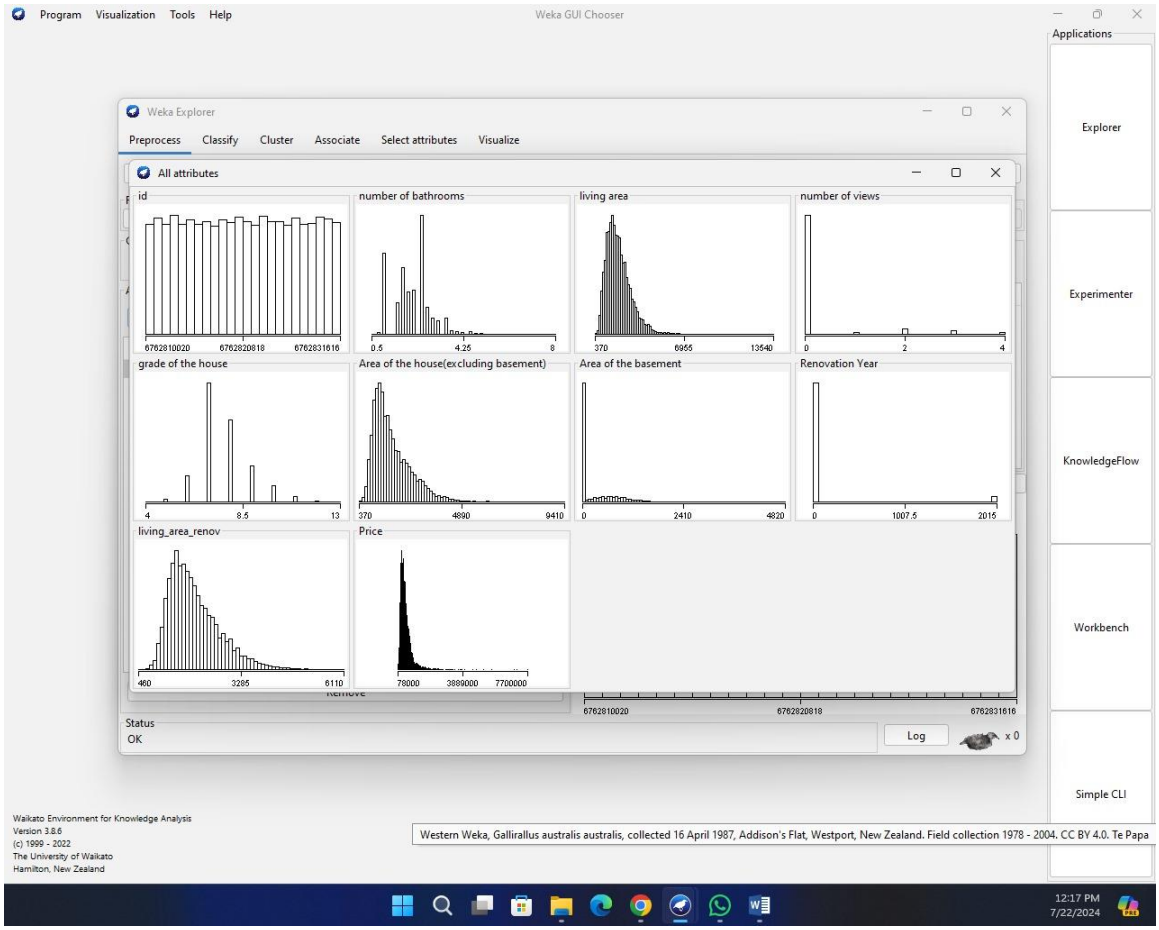
Before





After

choose visualize all from Process tab



Classifier evaluation

* Model Selection steps:

The 5 algorithms that we will review are:

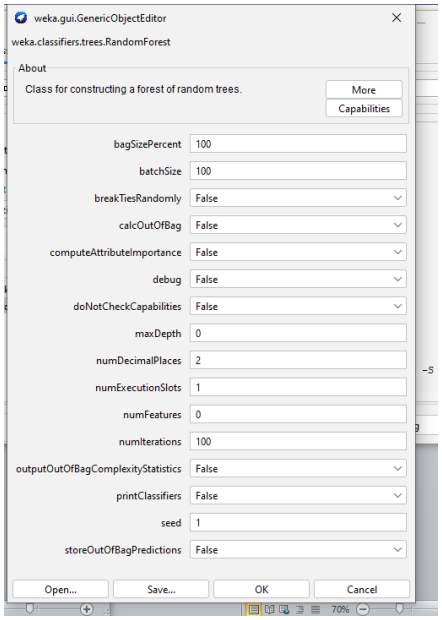
1- Random Forest
2- Linear Regression
3-Decision Tree
4-Support Vector Machines
5-kNN

1-Random Forest

Choose Random Forest algorithm:

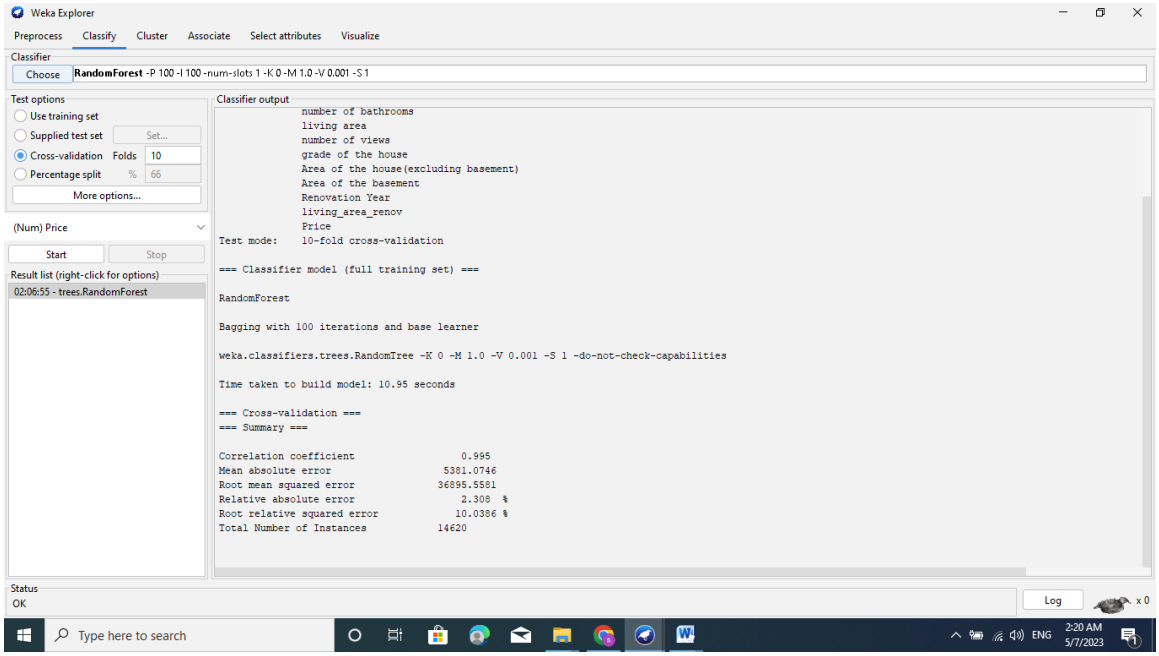
1-Click the “Choose” button and select “Random Forest” under the “trees” group.

2-Click on the name of the algorithm to review the algorithm configuration.



3-Click “OK” to close the algorithm configuration.
4-Click the “Start” button to run the algorithm on the House Price India dataset.

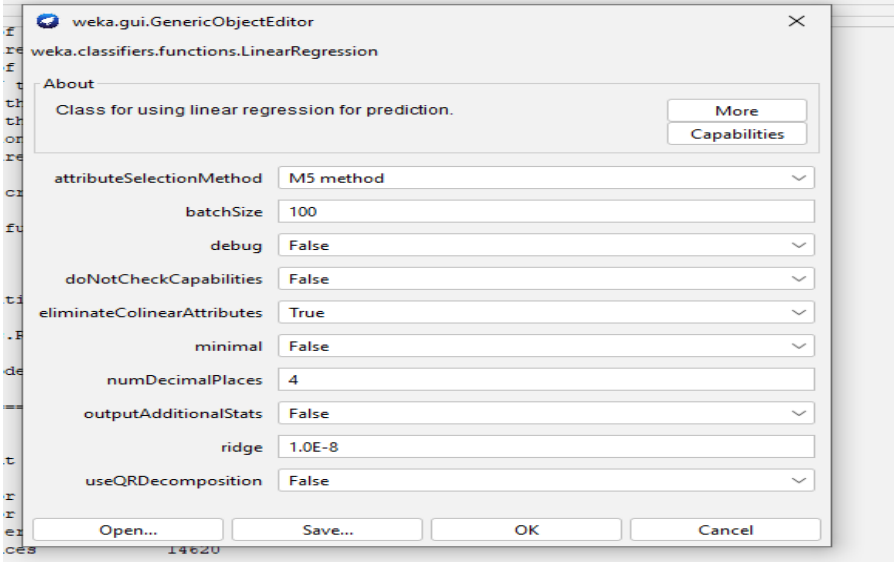
You can see that with the default configuration that linear regression achieves an RMSE of 36895.5581



2- Linear Regression

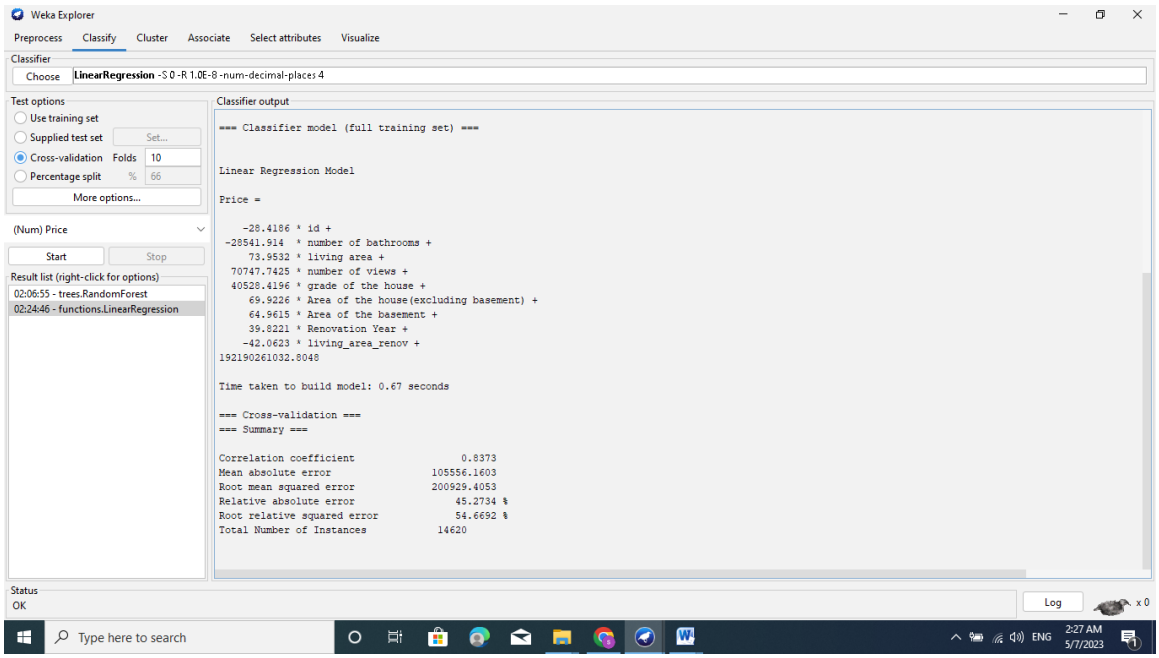
Choose the linear regression algorithm:

1-Click the “Choose” button and select “Linear Regression” under the “functions” group.
2-Click on the name of the algorithm to review the algorithm configuration.



3-Click “OK” to close the algorithm configuration.
4-Click the “Start” button to run the algorithm on the House Price India dataset.

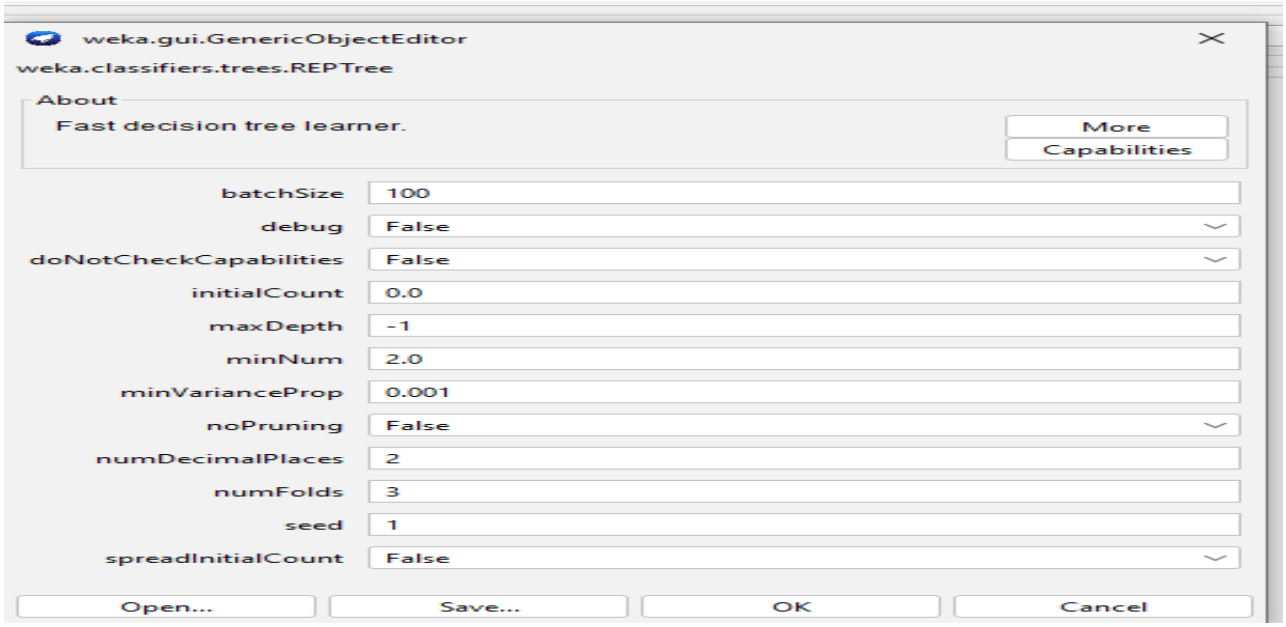
You can see that with the default configuration that linear regression achieves an RMSE of 200929.4053



3-Decision Tree

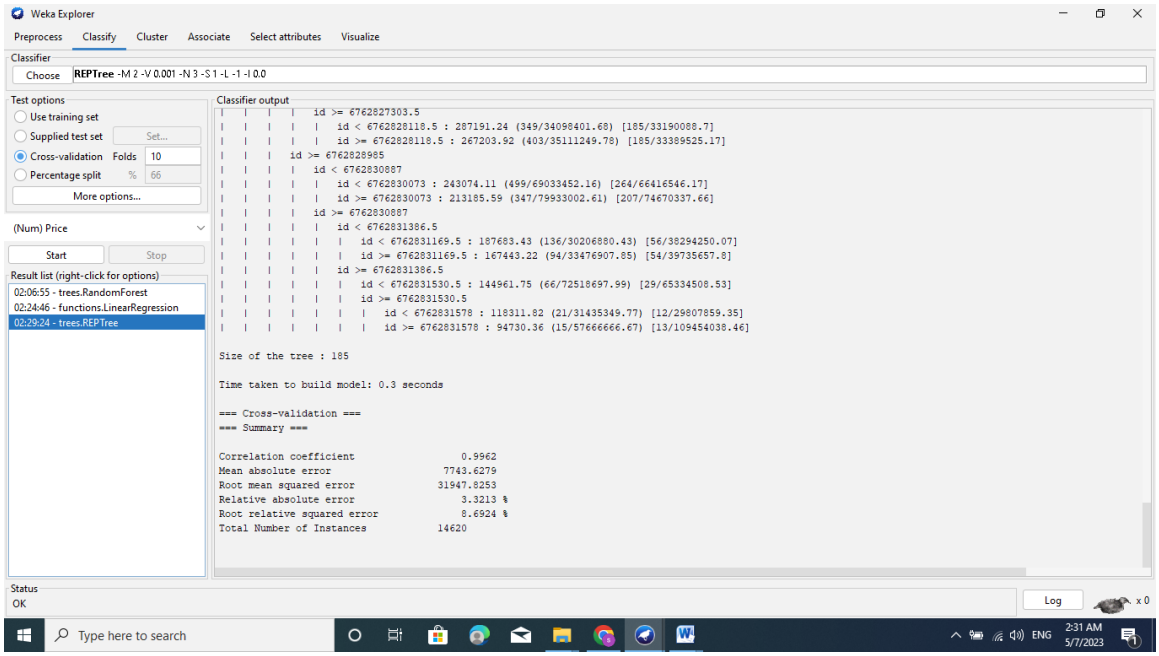
Choose the decision tree algorithm:

1. Click the “Choose” button and select “REPTree” under the “trees” group.
2. Click on the name of the algorithm to review the algorithm configuration.



3-Click “OK” to close the algorithm configuration.
4-Click the “Start” button to run the algorithm on the House Price India dataset.

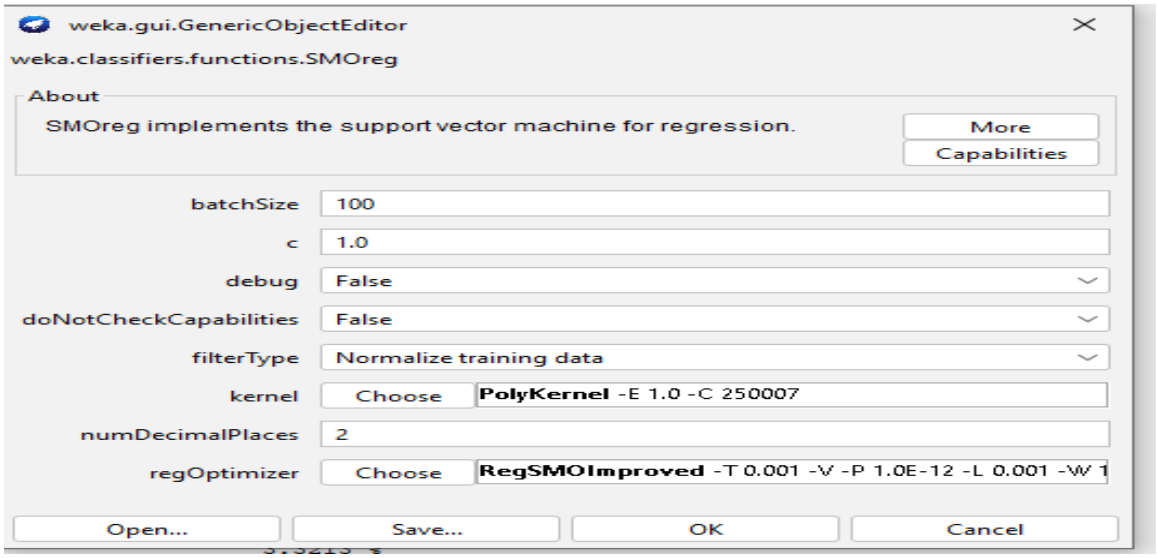
You can see that with the default configuration that decision tree algorithm achieves an RMSE of 31947.8253



4-support vector machine

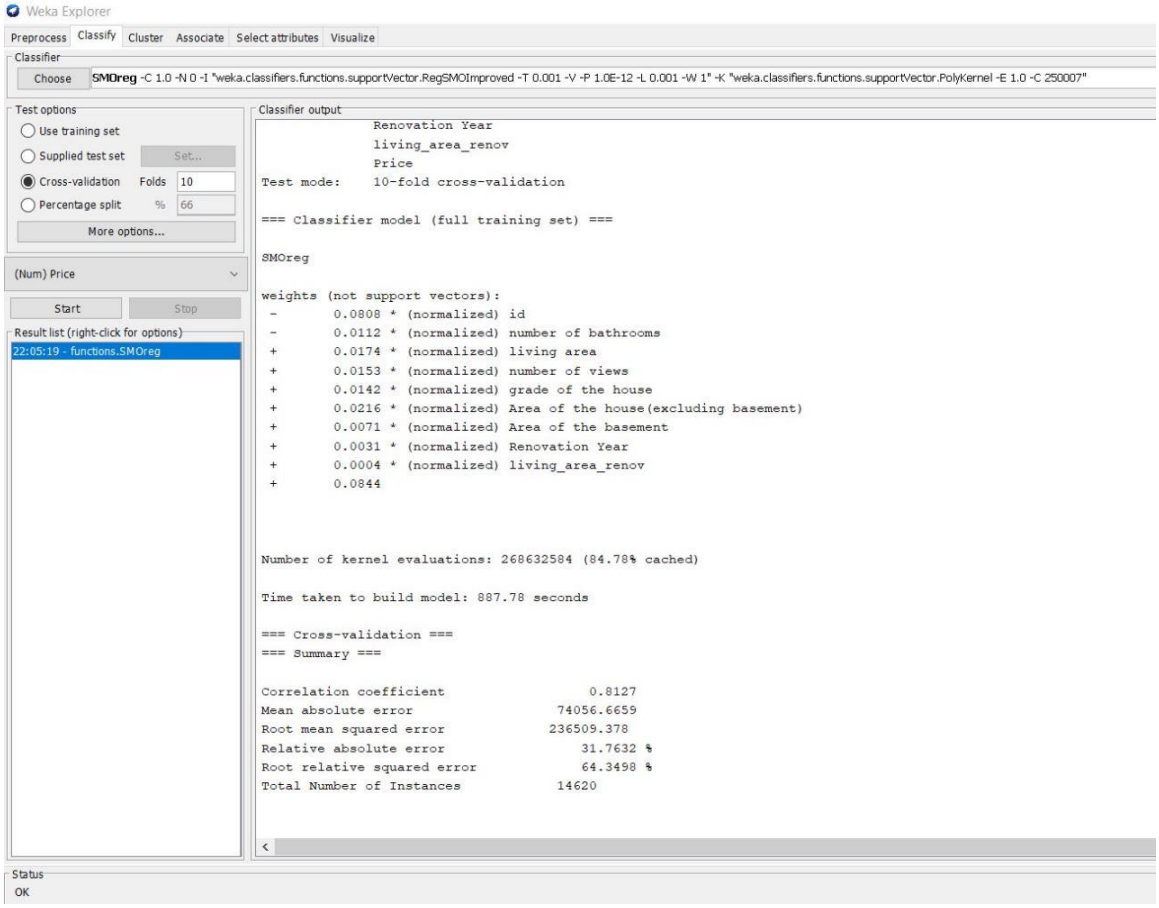
Choose the SVR algorithm:

1-Click the “Choose” button and select “SMOreg” under the “function” group.
2-Click on the name of the algorithm to review the algorithm configuration.

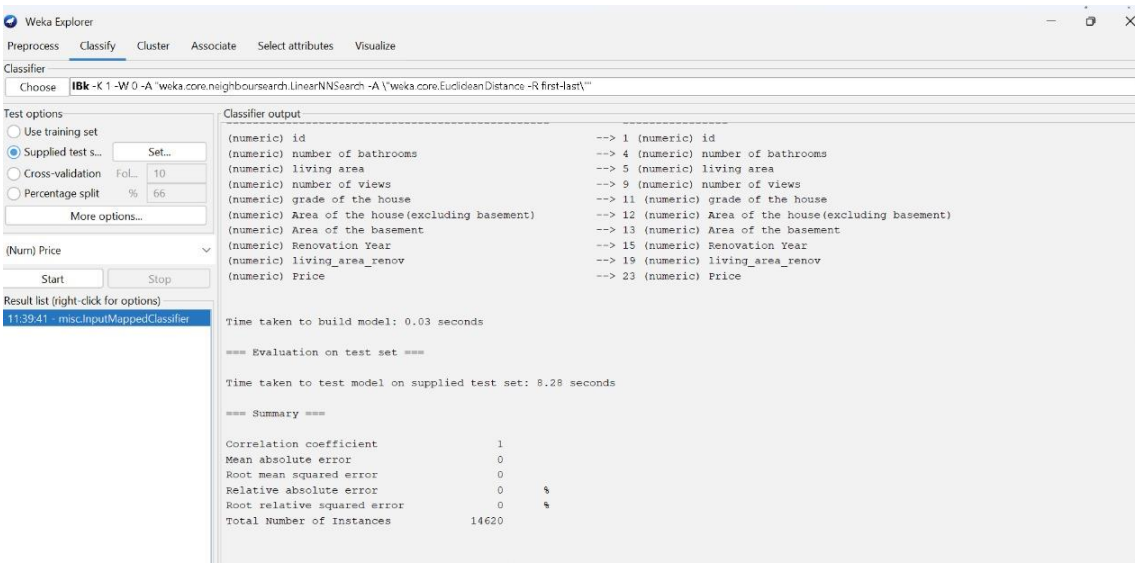


3-Click “OK” to close the algorithm configuration.
4-Click the “Start” button to run the algorithm on the House Price India dataset

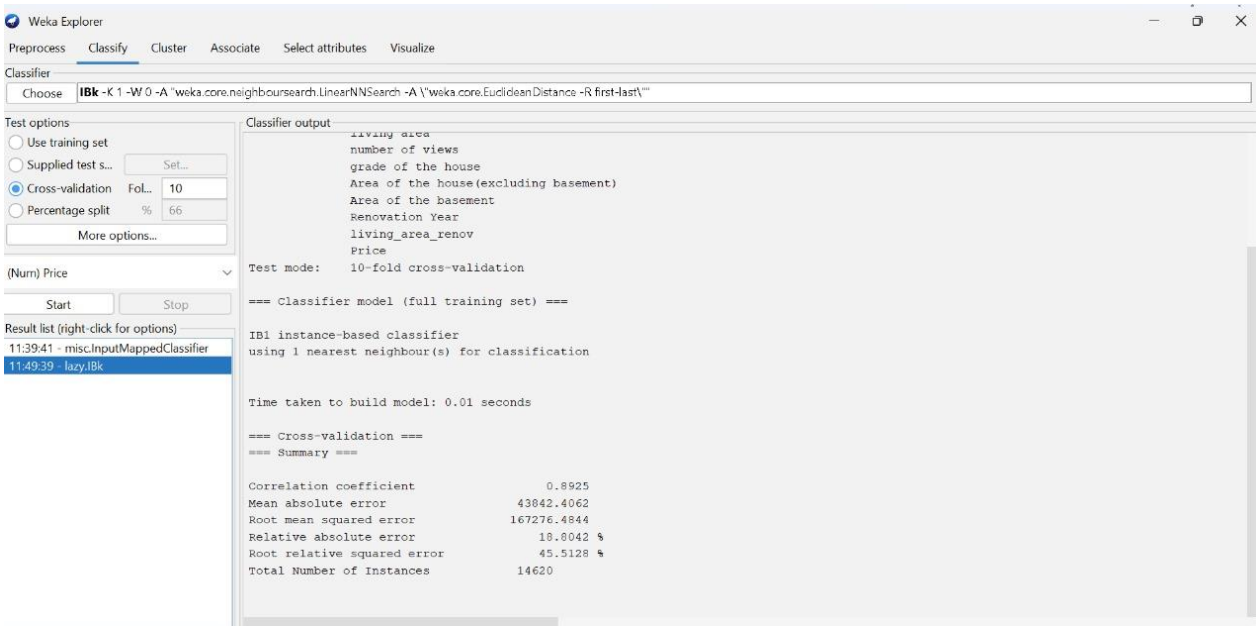
You can see that with the default configuration that SVR algorithm achieves an RMSE of 236509.38



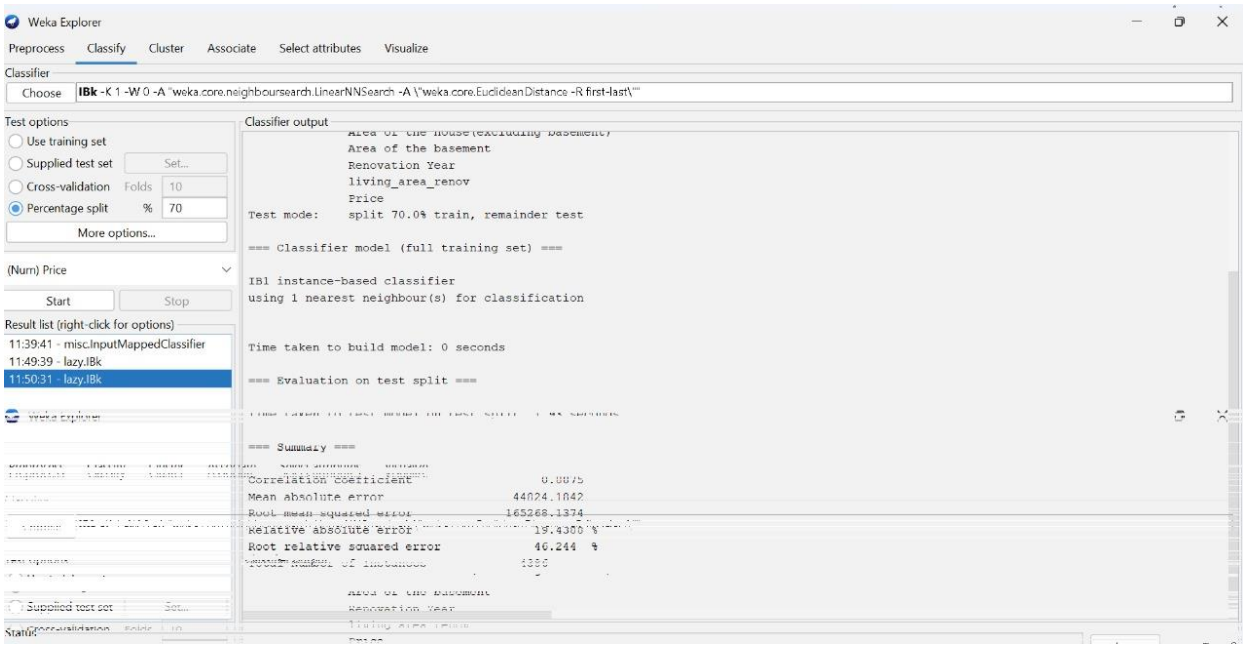
5-KNN(Supplied test set)



KNN(10-fold cross-validation)

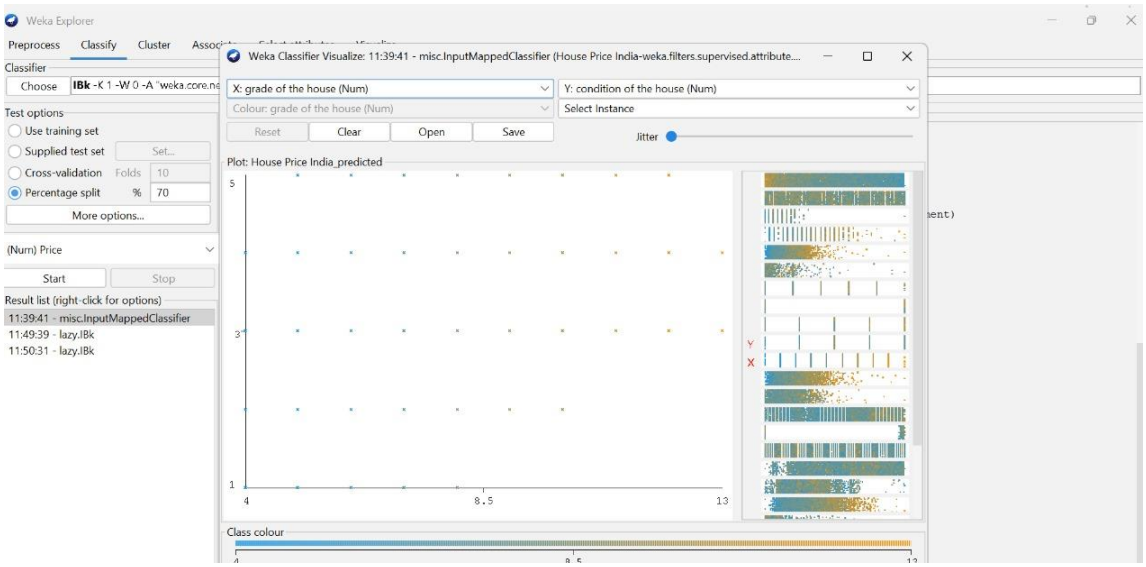


Percentage split (70% training, 30% testing)

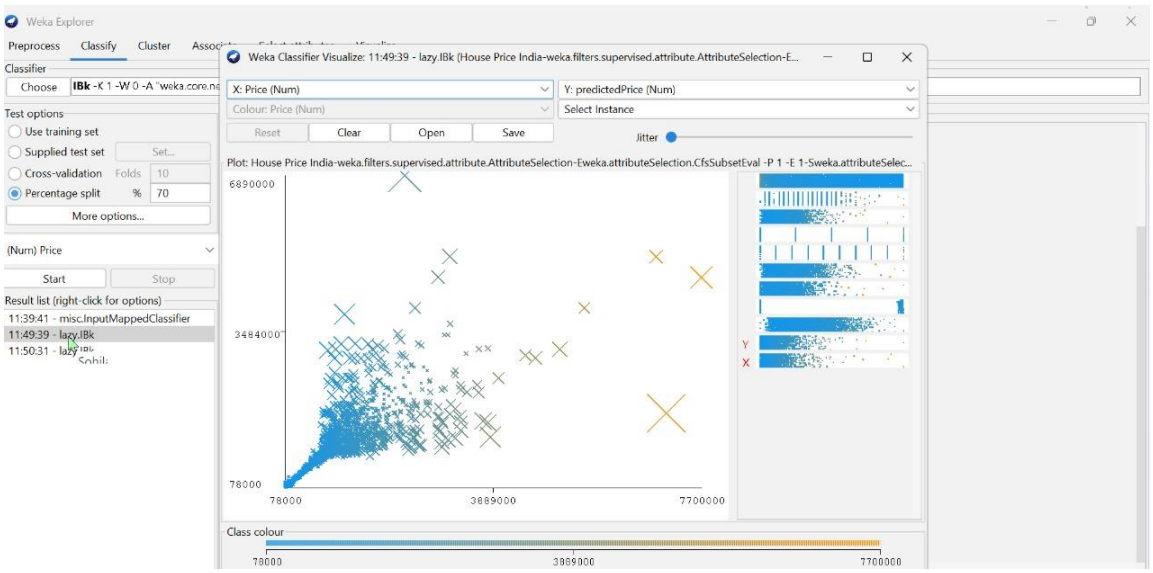


Visualization Analysis

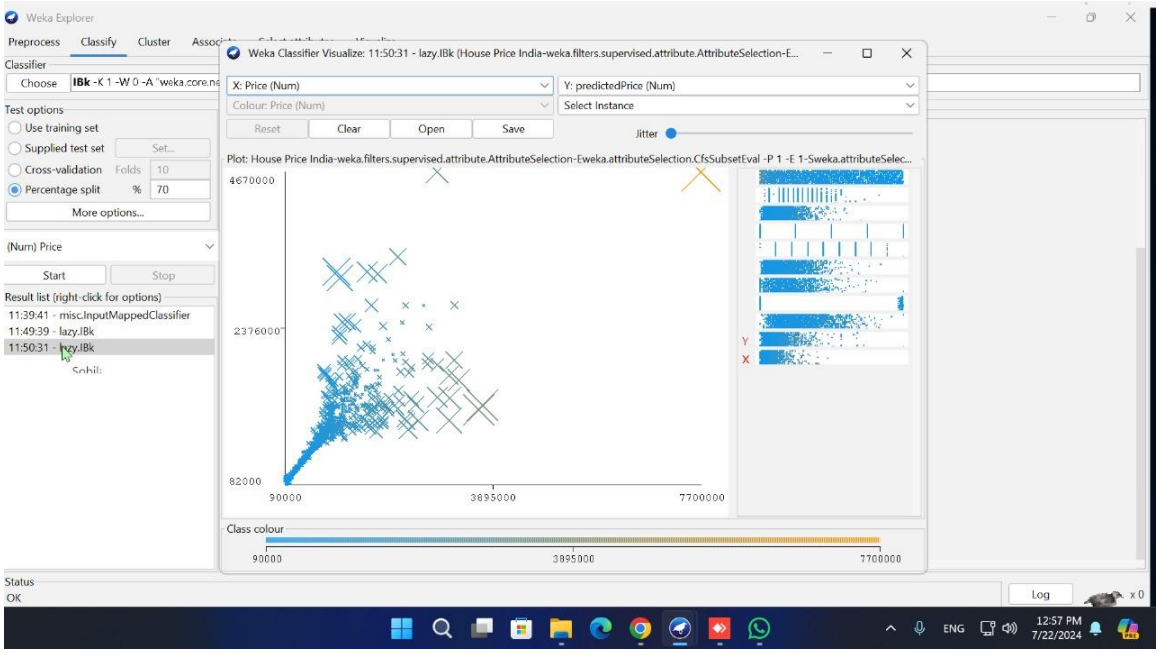
KNN(Supplied test set)



KNN(10-fold cross-validation)



Percentage split (70% training, 30% testing)



The objective for applying the mining task for this dataset

The objective of applying the mining task to this dataset is to develop an accurate model to predict house prices in India. This model should be able to identify the most important

factors that influence house prices and provide accurate predictions for new instances.

Classifier Comparison

The comparison of the results between different algorithms for the used mining task.

	1-Random Forest	2-Linear Regression	3-Decision Tree	4-Support Vector Machines	5-KNN(Supplied test set)	KNN(10-fold cross-validation)	KNN(70% training, 30% testing)
RMSE	36895.5581	200929.4053	31947.8253	236509.38	0	167276.4844	265268.1374

So the best algorithm can we applied in this data is (Random Forest)

And for KNN compare the best one can applied is KNN use Percentage split(70% training, 30% testing)