**2021-2022 SPRING SEMESTER**

**CSE443 MACHINE LEARNING LECTURE PROJECT-1**

In this project you need to apply all seen topics in our Lecture. You need to submit all related files with your professor by sending a mail to [o.sahingoz@iku.edu.tr](mailto:o.sahingoz@iku.edu.tr) **(sharing a cloud folder)**

Each dataset can only be used for only one user. (First submitted students is accepted, the other(s) should change the dataset)

Upper Limit of the Project is greater than 100. Writing a paper is optional. If you write and submit the paper, you will get +50 Points (up to).

1. Find a dataset from Internet with at least 10.000 data in it.

* Kaggle [**https://www.kaggle.com/datasets**](https://www.kaggle.com/datasets)
* UCI Machine Learning Repository [**https://archive.ics.uci.edu/ml/datasets.php**](https://archive.ics.uci.edu/ml/datasets.php)
* [**https://www.v7labs.com/blog/best-free-datasets-for-machine-learning**](https://www.v7labs.com/blog/best-free-datasets-for-machine-learning)
* https://imerit.net/blog/the-60-best-free-datasets-for-machine-learning-all-pbm/
* [**Google Dataset Search**](https://toolbox.google.com/datasetsearch):
* [**CMU Libraries**](https://guides.library.cmu.edu/machine-learning/datasets)**:**Discover high-quality datasets thanks to the collection of Huajin Wang, at CMU

1. Show related information about the dataset. (How many records does it have? What are the features? Types of the features?.... etc.)

(15 Points)

* Dataset should contain at least 15 features in it.

DATASET NAME: German House Prices

DATASET WEB LINK: https://www.kaggle.com/datasets/scriptsultan/german-house-prices?resource=download

DATASET INFO:

**It contains 10553 records.**

My dataset has 26 features initially but the first column is stating the number of entries so it is meaning that we have 25 actual features.

How many different classes exist in the dataset?

Type, Free\_of\_Relation, Furnishing\_quality, Condition, Heating, Energy\_source, Energy\_certificate

Energy\_certificate\_type, Energy\_efficiency\_class, State, City, Place, Garagetype

What are number of examples for each classes?

Type-->Unique 11

Free\_of\_Relation --> Unique 722

Furnishing\_quality --> Unique 4

Condition --> Unique 10

Heating --> Unique 13

Energy\_source --> Unique 115

Energy\_certificate --> Unique 3

Energy\_certificate\_type --> Unique 2

Energy\_efficiency\_class --> Unique 9

State --> Unique 16

City -->Unique 537

Place --> Unique 5092

Garagetype --> Unique 7

**How many NULL values exist? (depending on the features distinctly)**

* Price: 0, Type: 402, Living\_space: 0, Lot: 0, Usable\_area: 4984, Free\_of\_Relation: 3569, Rooms: 0,
* Bedrooms: 3674, Bathrooms: 1801, Floors: 2664, Year\_built: 694, Furnishing\_quality: 2726,
* Year\_renovated: 5203, Condition: 323, Heating: 584, Energy\_source: 1227, Energy\_certificate: 755,
* Energy\_certificate\_type: 3526, Energy\_consumption: 8119, Energy\_efficiency\_class: 4819, State: 1, City: 1,
* Place: 290, Garages: 1960, Garagetype: 1960

Which features are not numeric?

Column1: Numeric, Price: Numeric, Type: Not Numeric

Living\_space: Numeric, Lot: Numeric,Usable\_area: Numeric

Free\_of\_Relation: object ,Rooms: Numeric, Bedrooms: Numeric

Bathrooms: Numeric, Floors: Numeric, Year\_built: Numeric

Furnishing\_quality: Not Numeric, Year\_renovated: Numeric

Condidition: Not Numeric, Heating: Not Numeric, Energy\_source: Not Numeric

Energy\_certificate: Not Numeric, Energy\_certificate\_type: Not Numeric

Energy\_Consumption: Numeric, Energy\_efficency\_class: Not Numeric

State: Not Numeric, City: Not Numeric, Place: Not Numeric

Garages: Numeric, Garagetype: Not Numeric

1. Text

   Description automatically generatedUse **Label Encoding** for at least one of the features (Explain your reason “why do make this operation?”) (10 Points)

We are using label encoding because it assigns a unique number to each class of data so we can identify easily.

Text

Description automatically generated

1. Use **One Hot encoding** for at least one of the features (Explain your reason “why do make this operation?”) (10 Points)

A picture containing calendar

Description automatically generatedWe use One Hot encoding because our training data becomes much more informative and descriptive with just one hot encoding, and it can be effortlessly resampled. We can more easily determine a probability for our values by using numeric numbers.

Text

Description automatically generated

1. Analyze the Missing Values
   1. **Delete some columns** (Explain your reason “why do make this operation?”) (10 Points)

General consensus online indicates that if your values are valid at about %95 you can ignore and if your data percent is %60 valid it can be imputed the base minimum amount of acceptable imputation range. So, I am basing my operations upon these claims.

Text

Description automatically generated

* 1. **Delete some rows** (Explain your reason “why do make this operation?”)

(10 Points)

We delete rows because there are some redundant data that we have to get rid of but they are not so common in those features so instead of deleting the whole column we delete rows.

Text

Description automatically generated

* 1. **Impute** some missing data (Explain your reason “why do make this operation?”) (10 Points)

We impute data because there is nan values in our dataset and to make our algorithms more efficient we fill those nan values with mean or fill them with unknown.

We fill the numeric values of their mean and categorical values as unknown.

Text

Description automatically generated

1. Find the best correlated Features in the Dataset (10 Points)

DISPLAY THE CORRELATION CALCULATION

Text

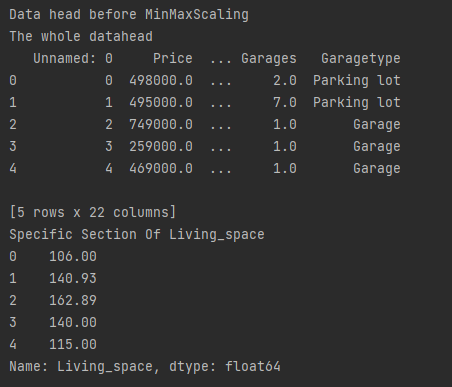
Description automatically generated

Text

Description automatically generated with medium confidence

1. Execute a Normalization/Scaling in the Dataset (10 Points)

A screenshot of a computer

Description automatically generated with medium confidenceText

Description automatically generatedPUT THE SCREENSHOT OF DATA.HEAD BEFORE AND AFTER THE OPERATION

I transformed Living\_spaces values and created another column with them using MinMaxScaling I called this column NormalizedSpace

I have also calculated the mean, standard deviation, min, max values of Living Space before and after the MinMaxScaling.

1. Train your new dataset **at least 5 different Machine Learning algorithms**

(15 Points)

THE PREFERRED ALGORITHMS ARE:

KNN classification

Random Forest

Linear Regression

Fully Connected Neural Network

KNN Regression

1. Use **5-fold** approach to measure the performance of the system

(10 Points)

WITH A RANDOM SELECTION WE REACHED THE FOLLOWING RESULTS

Random Selection:

Score: 0.00982669287118099

Mean 0.00982669287118099

Standard Deviation: 0.0

WITH 5-FOLD APPROACH I REACHED THE FOLLOWING RESULT

Five Fold calculation: [0. 0.00803571 0.00536193 0.00625559 0.02144772]

Total time: 0.209416

1. Put their results to **a table** to make a comparison

(5 Points)

A picture containing chart

Description automatically generated

A picture containing timeline

Description automatically generatedA picture containing graphical user interface

Description automatically generatedA picture containing timeline

Description automatically generated

A picture containing website

Description automatically generated

1. Calculate **the training time** for all of them

KNN Regression Between Living Space & Garages

Total time: 0.004147

Fully Connected Neural Network:

Total time: 61.051750

Linear Regression:

Total time: 0.007292

Random Forest:

Total time: 0.348042

KNN classification Prediction for key-mode: [1972. 1972. 1977. ... 1972. 1972. 1850.]

Total time: 0.050511

Random Selection:

Five Fold calculation: [0. 0.00803571 0.00536193 0.00625559 0.02144772]

Total time: 0.209416

(10 Points)

1. Select the best 10 features from the database

(10 Points)

SHOW THE LIST OF THE FEATURES

Price

Type

Living\_space

Lot

Rooms

State

City

Place

Condition

Year\_built

1. Write **a Conference paper** to Show all your reached results. **(OPTIONAL**)

(50 Points)