**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 : heart.csv

DATASET WEB LINK : https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

DATASET INFO

How many records does it contains: 14.266

How many features does it have: 14

How many different classes exist in the dataset: \*7

What are number of examples for each classes?

How many NULL values exist? (depending on the features distinctly): None

Which features are not numeric?: sex, thal, target

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

Some features of a dataset may be consists of labels, and unfortunately labels are not a machine readable form. So to overcome this obstacle, label encoding is one of the provided solutions.

Used on thal and target features.

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

Features consists of categorical data can be encoded with One Hot encoding to both improve the accuracy and transform categorical data to machine readable form.

Used on sex feature. Created sub-features of male and female

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

No reason to delete, so all columns are intact.

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

(10 Points)

Some of the rows were deleted due to incorrectly filled datas of one or more features

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

No missing values in the Dataset, so no imputation was done.

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

DISPLAY THE CORRELATION CALCULATION

Chart, treemap chart

Description automatically generated

cp, thalach, slope.

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

PUT THE SCREENSHOT OF DATA.HEAD BEFORE AND AFTER THE OPERATION

Dataset was already normalized, but in order to prove that an image file of dataset that has been twice normalized (first by creator second by author (student) ).

A picture containing text, window

Description automatically generated

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

(15 Points)

THE PREFERRED ALGORITHMS ARE

-Decision Tree Classifier:

-Random Forest Classifier:

-MLP (Multi-Layer Perceptron) Classifier:

-K Neighbors Classifier:

-Ridge Classifier:

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

(10 Points)

WITH A RANDOM SELECTION WE REACHED THE FOLLOWING RESULTS

Decision Tree Classifier:

* R2 Score: 94.1171%
* Accuracy: 98.5294%
* Mean Absolute Error: 1.4706%
* Mean Squared Error: 1.4706%
* Root Mean Squared Error: 12.1268%

Random Forest Classifier:

* R2 Score: 94.1171%
* Accuracy: 98.5294%
* Mean Absolute Error: 1.4706%
* Mean Squared Error: 1.4706%
* Root Mean Squared Error: 12.1268%

MLP (Multi-Layer Perceptron) Classifier:

* R2 Score: 37.2489 %
* Accuracy: 84.3137%
* Mean Absolute Error: 15.6863%
* Mean Squared Error: 15.6863%
* Root Mean Squared Error: 39.6059%

K Neighbor Classifier:

* R2 Score: -13.7364%
* Accuracy: 71.5686%
* Mean Absolute Error: 28.4314%
* Mean Squared Error: 28.4314%
* Root Mean Squared Error: 53.3211%

Ridge Classifier:

* R2 Score: 35.2879%
* Accuracy: 83.8235%
* Mean Absolute Error: 16.1765%
* Mean Squared Error: 16.1765%
* Root Mean Squared Error: 40.2200%

WITH 5-FOLD APPROACH I REACHED THE FOLLOWING RESULT

Decision Tree Classifier:

Scores for each fold are: [1. 0.98529412 0.98529412 1. 1. ]

Average score: 0.99

Random Forest Classifier:

Scores for each fold are: [1. 1. 0.98529412 1. 1. ]

Average score: 1.00

MLP (Multi-Layer Perceptron) Classifier:

Scores for each fold are: [0.78921569 0.8627451 0.76960784 0.85221675 0.81773399]

Average score: 0.82

K Neighbor Classifier:

Scores for each fold are: [0.77941176 0.71078431 0.72058824 0.74876847 0.73399015]

Average score: 0.74

Ridge Classifier:

Scores for each fold are: [0.83823529 0.8627451 0.85784314 0.8226601 0.85221675]

Average score: 0.85

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

(5 Points)

|  |  |
| --- | --- |
| 98.5294 | 0.99 |
| 98.5294 | 1.00 |
| 84.3137 | 0.82 |
| 71.5686 | 0.74 |
| 83.8235 | 0.85 |

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

(10 Points)

Decision Tree Classifier:

Training time: 15.716689586639404s

Random Forest Classifier:

Training time: 0.3026413917541504s

MLP Classifier

Training time: 132.50903940200806s

KNeighbor Classifier:

Training time: 0.013962030410766602s

Ridge Classifier

Training time: 0.01593613624572754s

1. Select the best 10 features from the database

(10 Points)

SHOW THE LIST OF THE FEATURES

age, male, female, thal, restecg, slope, ca, chol, oldpeak, exang

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

(50 Points)