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 **(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).

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

- Kaggle **https://www.kaggle.com/datasets**
- UCI Machine Learning Repository **https://archive.ics.uci.edu/ml/datasets.php**
- **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:
- CMU Libraries: Discover high-quality datasets thanks to the collection of Huajin Wang, at CMU

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

(20 Points)

- Dataset should contain at least 15 features in it.

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

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

e) Analyze the Missing Values

a. **Delete some columns** (Explain your reason "why do make this operation?") (10 Points)

b. **Delete some rows** (Explain your reason "why do make this operation?") (10 Points)

c. **Impute** some missing data (Explain your reason "why do make this operation?") (15 Points)

f) Train your new dataset **at least 5 different Machine Learning algorithms** (20 Points)

g) Use **5-fold** approach to measure the performance of the system (10 Points)

h) Put their results to **a table** to make a comparison (5 Points)

i) Calculate **the training time** for all of them (10 Points)

j) Write **a Conference paper** to Show all your reached results. **(OPTIONAL)** (50 Points)