**2020 Fall Machine Learning Final Project**

**1. Outline**

* Select a dataset and build classification(or regression) models.
* If necessary, modifying or preprocessing the dataset is allowed.
* Build as much machine learning models as possible (more than 4 classification or regression models at minimum + more than 1 clustering model). Using models that were not covered in class is also allowed.
* Analyze all aspects of the model on the selected dataset.

**2. Limitations on the Dataset**

Select a dataset from any public data repositories or use dataset provided from any machine learning competitions.

**Public Dataset Repositories**

* UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/>)
* KD Nugget (<https://www.kdnuggets.com/datasets/index.html>)
* Kaggle (<https://www.kaggle.com/>)
* Google Dataset Search (<https://toolbox.google.com/datasetsearch>)

**Conditions on Dataset Selection**

* At least more than 10 features (attributes)
* At least more than 5,000 instances

**Sample Datasets**

* Adult Data set (<https://archive.ics.uci.edu/ml/datasets/Adult>) : 14 features, about 50,000 instances
* Mushroom Data set (<https://archive.ics.uci.edu/ml/datasets/Mushroom>) : 22 features, about 8,000 instances
* Video Game Sales (<https://www.kaggle.com/gregorut/videogamesales>) : 11 features, about 16,000 instances
* Mobile App Store (7200 apps) (<https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps>) : 17 features, about 7,000 instances

**3. Tools**

Machine Learning Tools

* Python-based Jupyter Notebook environment (recommended) : Scikit-Learn Library (recommended), other machine learning libraries that are python supported, your own code implementation of machine learning algorithms
* Using GUI-based machine learning tools (Weka, Orange, RapidMiner etc) are not allowed.

**4. Project Progress and Writeup(report)**

Your project report should include:

* Overall summary of the experiment
* Description of your selected dataset and data URL
* Experiment design and methods (Describe the project progress as detailed as possible)
  + List your selected machine learning models and the describe the reasons behind your selection
  + If preprocessing dataset is necessary, you may apply various preprocessing techniques such as normalization, regularization, discretization and missing value handling.
  + You should apply more than one feature selection technique.
  + Describe your trial and errors during the project progress (Your report should describe your process of modifying model parameters apart from only using default models. The report should also describe the process of selecting the optimal model through validation process.)
  + Apply clustering to your dataset and create clusters. Then, compare the resulting clusters and the labels of the classification. Your report should describe this process and describe the differences or similarities.
* Experiment Result Analysis
  + The dataset should be divided into train, validation and test set or be divided into train and test set. For the latter case, the validation process should be done through cross-validation. The final model evaluation must be done with test dataset.
  + Analyze all models’ performance on the test dataset. If there are differences among the model’s performance, analyze the model itself if the model can be analyzed. If the model cannot be analyzed (meaning black-box models), you should provide the analysis by inferring through the dataset features.
  + Compare the model’s performance before and after applying feature selection techniques. Analyze or infer the reason behind the difference and include the analysis in your report.
  + The experiment design and model performance analysis should include description from perspective of overfitting.
* Conclusion

**5. Submission Materials**

Submit project report and your program code

* The submitted code should be executable (no errors) for grading.
* When working in Jupyter Notebook environment, it is possible to integrate the project report with the code. (The project report should be written in Markdown cell)
* When working in other environments, the program code and the project report (pdf or doc) should be zipped together and submitted. (There is no specified format for the project report)

**6. Project Interview**

* There is not enough time to interview total 80 students one at a time, so we will replace the project interview with your self-video answering questions on the project. The project interview date will be notified later. The questions regarding the final project will be uploaded to Blackboard, and you will have a limited time to record your answers and upload to Blackboard. The video should include your face and your voice answering the project questions.
* There may be additional interviews if necessary. Students conducting an additional interview will be notified.

**7. Submission Deadline**

* 11:59PM, Sunday, December 20, 2020
* Submit through Blackboard

**8. Help Session**

* If you are experiencing difficulties in choosing a dataset or in making progress for the project, please email the TAs through [koreaAIassistant@gmail.com](mailto:koreaAIassistant@gmail.com). Also email us whenever you have any questions regarding the project.