ENGM 4676/6676 – Machine Learning for Engineers

Assignment #2: Machine Learning with Scikit-Learn

Building on the foundation established in Assignment #1, this assignment will further develop your skills in machine learning techniques by utilizing Scikit-Learn. You will work with the engineering dataset you selected in Assignment #1. Note that you have the option to change this dataset if you prefer; please refer to Question 0 for more details.

Question 0: Revisiting Your Dataset (0 points, or -5 points if excluded)

To ensure a smooth transition from Assignment #1, start by including your answer to question #1 from the assignment #1 where you described your dataset. If you've decided to switch to a different dataset, please provide a detailed description along with the appropriate citation. Additionally, share your reasons for making the change.

Question 1: Utilizing Machine Learning (15 points)

Experiment with three machine learning models of your choosing **(5 points each)** using your selected dataset. Feel free to explore a mixture of regression and classification methods or focus solely on various classification techniques according to your preference.

For each model:

1. Train your models by applying an appropriate training/test split.
2. If applicable to that model, explore how using regularization can improve it.
3. For each model, create visuals that highlight interesting aspects of your model and the results.
4. For classification models, obtain the classification report and the confusion matrix. Comment on the results.
5. If applicable, calculate and visualize feature importance.

For all models, show your code with comments.

Question 2: Model Comparison (5 points)

Once you've explored your three models, it's time to put them head-to-head. Summarize what you found for each one, and then compare them to each other. Based on your findings, which model would you choose to use, and why? If any, identify the noticeable weaknesses of each model.

Evaluation metric:

For each question, you will be evaluated based on your presentation quality, results quality, code quality, visualization quality, clarity, completeness of answers, and code comments. A rating of 1 to 5 will be used as follows (Guideline):

No Answer (0)

Limited (1): Poor quality and incomplete answers.

Basic (2): Below-average quality and partially complete answers.

Adequate (3): Satisfactory quality.

Proficient (4): Good quality.

Excellent (5): Excellent quality.

Submission Guidelines:

Write a Jupyter Notebook (.ipynb file) to perform the tasks mentioned in each question.

Include comments in your code to explain the steps you are taking and the rationale behind your decisions. Ensure that your code is well-organized and follows best practices for coding style.

If you encounter any challenges or limitations during the assignment, document them along with your approach to overcoming them.

Submission Deadline: October 25th, 2024, by 11:59pm.

Please submit:

1- **Jupyter Notebook file (with the output)**
2- **A PDF print of that Jupyter Notebook file.**

Please name them as ***"[Your Banner]_[Your Name]_A3. IPYNB"*** and ***"[Your Banner][Your Name]_A2".PDF***

- **No PDF:** If you do not include the PDF we will assess your assignment based on the ipynb file alone. Please note that your marks will negatively impact even further if we are unable to reproduce the results from your ipynb file or if the file doesn't contain the outputs. **We will not upload and mount the dataset you utilized.**

- **No ipynb**: **You will lose 5 marks**, not providing the ipynb or the source code files will prevent us from verifying the legitimacy of your work, consequently resulting in a mark deduction.

Late submissions without an approved reason will incur a penalty of 2 mark per day.

If you have any questions, please ask them during the tutorial or use the discussion board.

Good luck!