FIU | Knight Foundation School of Computing and Information Sciences

Fall 2025: CAP5610 – HW 3

In HW3, you will learn and apply tree-based machine learning methods for Classification and Regression and use SHAP to interpret model results.

Algorithms to Use:

- Decision Trees
- Random Forests
- Gradient Boosting Machines (GBM)
- XGBoost
- LightGBM
- CatBoost

Data for Classifiers:

- Gene Expression: lncRNA_5_Cancers.csv (available under Module 2)
- Cancer Types: KIRC, LUAD, LUSC, PRAD, THCA

Data for Regressor:

- Download from: https://drive.google.com/file/d/1-y21VY3--PITlPqRG82Wy-nTHPj2ZPUV/view?
- Data is also available under Module 2
- GDSC2 drug screening data for 13 drugs across different cancer cell lines. The number of cell lines ranges from 961 to 963 for each drug.
- LN_IC50 (Log normalized IC50) is the target variable, which represents the drug doses.
- Each row can be uniquely identified by a cell line name and a drug name pair.
- The remaining columns are the gene expression values for the cell line.
- Information about drugs, cell lines, and screening methods can be found at the GDSC (Genomics of Drug Sensitivity in Cancer) official website: https://www.cancerrxgene.org/

Task 1: [25 points] Find the best Tree-based classifier applying **ALL** the algorithms listed above. Use the metrics **Accuracy** and **F1 score**.

Task 2: [25 points] Apply SHAP on the best classifier found in Task 1. (a) Find each cancer-specific 10 significant features. (b) Show force plots for one patient (ID: TCGA-39-5011-01A) for five cancer types.

Task 3: [25 points] Find the best Tree-based regressor applying **ALL** the algorithms listed above. Use the metrics **MAE, MSE, RMSE,** and **R^2**.

Task 4: [25 points] Apply SHAP on the best regressor found in Task 3. (a) Find each drug-specific 10 significant features. (b) Find top 10 features for the drug-cell line pair with the least prediction error.

You must submit the following items in CANVAS:

- Report (MS word or PDF)

  - Describe the algorithms/approaches/tools used: (a) What it is or What it does, (b) How it does, and (c) Application.
  - Describe results: (a) Put Figure/Table number and Title: On top of the table, and bottom of the figure. (b) Describe the figure and table. (c) Your observation about the figure and table. (d) Conclusion.
- Source code (*.py or Jupyter notebook)
  - Must be well organized (comments, indentation, …)
- File name: HW3_lastName

You must submit the files **SEPERATELY**. DO NOT compress into a ZIP file. If you fail to provide all required information or files, you may be given zero score without grading.

Deadline:

The deadline is **11:59pm Wednesday, October 8, 2025**. Late assignments will not be accepted.