**Project 5**

**Title:** Supervised machine learning approaches for predicting future events

**Learning objectives:** The overall learning objective of this project is to utilize established machine learning models to train a supervised classifier using part of the provided dataset and to evaluate its predictive power on the remaining part of the examined cancer cohort. Specifically, students will need to select an appropriate clinical endpoint and to train a classifier to predict an outcome within a point-interval for the selected clinical endpoint. Students are strongly encouraged to leverage common feature selection techniques in order to identify a reasonable and minimal set of features that would be predictive of the selected clinical outcome. Additionally, students must use model validation techniques, such as cross-validation, to increase the robustness of their trained models. Moreover, students will be required to evaluate the overall performance of their approaches including but not limited to area under the receiver operating characteristic curve, precision, sensitivity, specificity, and F1 score. Lastly, students should clearly show the predictive value of each of the features utilized by their supervised classifier and justify the use of any features without a high predictive value.

**Subject Matter:** The focus of the final project in this class will be the construction of a supervised classifier for predicting a specific clinical outcome of a cancer patient. There will be no background manuscript for this project and students are encouraged to explore both the existing literature and their provided cancer dataset. Importantly, it is anticipated that students will utilize at least some of the techniques from the prior projects in order to successfully complete project 5. Specifically, project 5 will require students to explore different clinical endpoints and to select a single clinical endpoint for predicting a patient’s outcome. For example, students may want to predict the clinical outcome of a patient for a specific interval after cancer diagnosis. Students are strongly encouraged to evaluate different supervised classifiers and to explore optimizing each of their relevant parameters. Moreover, students should examine the full set of cancer genomics and transcriptomics features in their datasets and identify a reasonable set of features to train their predictive models. All supervised models should be trained by using at least one model validation technique and students should evaluate the performance of their classifiers using standard machine learning metrics.

**Project’s datasets:** <https://www.dropbox.com/sh/57resib0deyyb2s/AABx0-QsE7hrNYQdIsoSPLgLa>

**Project’s presentation:** Team presentations on Wednesday (14-Jun) at 7:00pm to 9:59pm (TBD)

**Deadline:** All teams submit a project report (max 3 pages) by 11:59pm on Tuesday (13-Jun)

**Potential outline for approaching the project:**

* Separate the provided cancer dataset into training and testing datasets with the training dataset being utilized only for training and the testing dataset used exclusively for testing. Make sure that the testing and training datasets are similarly balanced regarding positive and negative outcomes.
* Identify and justify which clinical endpoint (*e.g*., overall survival, disease-free survival, *etc.*) and what point-interval (*e.g.*, 2-years, 5-years, 10-years) should be utilized for your dataset by considering your cancer type and the available data. If you are unsure which endpoint or point-interval is appropriate, you should compare your results for different clinical endpoints and/or distinct point-intervals.
* Perform a feature selection approach to the training data in order to identify a practical set of features that would be likely predictive of the selected clinical outcome. Students are allowed to use results from their prior projects including dimension reduction results, cancer driver genes, mutational signatures, and prognostic biomarkers.
* Students should explore at least one type of supervised classifiers (*e.g.,* support vector machine, naive Bayes classifier, random forest, *etc.*) to train a model on their training data. Student should make sure that they optimize the parameters of the classifier(s).
* All models should be trained using model validation techniques (*e.g.*, cross-validation).
* The model’s performance should be evaluated on the test dataset using standard machine learning metrics (area under the receiver operating characteristic curve, precision, sensitivity, specificity, F1 score, *etc.*).