# CS 634 Data Mining
# Final Term Project

**Dr. Yasser Abduallah**
**Department of Computer Science**
**New Jersey Institute of Technology**

# Table of Contents

# General Submission Rules

- Embed your last name and first name in your project file name. For example, if your name is John Smith, your file name should read: smith_john_finaltermproj.zip

- Your project will automatically lose **10** points if the above submission rules are violated.

- Submit your project file in Canvas under Final Term Project Submission Site before the due time. The project file in Canvas is considered as the final version.

- No late project is accepted. A project is late if it is not submitted in Canvas before the due time. Zero points will be given to the late project.

- **NOTE: Pay attention to the project description front page in Canvas because it may have additional information and requirements, etc..**

# Project Grading

- ❖ The grades will be posted on Canvas when they are completed.

- ❖ Note: There is a limit on the file size in Canvas and in NJIT’s email box. So, keep your project file small to avoid any problem that may occur when submitting the file in Canvas.

- ❖ The project file must contain the **all source code (Jupyter and Python files) in running state and data sets** and documentation including **screenshots**. The screenshots are used to demonstrate the running situation of your program, particularly how the program executes and produces output based on different input data and user-specified parameter values, if applicable.

- ❖Project documentation/report should state how to run your program, any required packages and how to install them as if someone without any knowledge of can follow your instructions to replicate and run your program.

- ❖ The code should be running without any editing by the TA or me.

- Github & Jupyter Notebook.
    - After you finished your code in development, tested it, made sure it worked, and prepared the report (meaning all heavy lifting job was done 😊), Create a Github repository in https://github.com/, or use your current account if it was created with njit email. Your account must be with your NJIT email, not your personal email.
    - Load your project to the repository.
    - Create Jupyter notebook for your work to show the output, for more info visit https://jupyter.org/
    - Give me ya54@njit.edu access as a collaborator to your repository. (If we have a grader, you give him/her access too).
    - Add Github link to your repository to your report.

      NOTE: If you need help with Github and/or Jupyter book, let me know.

  NOTE: Jupyter notebook is the classic notebook so that you use to import and run your Python work you did above.

- Copying and sharing code with peers is prohibited and will result in 0 point for all parties that are involved.

# Final Term Project

- This is a single person project. *Do not share or copy code from your peers.*
- Make sure to follow the submission rules when submitting your project.
- The Appendix will include a list of helping documents, sites, tools, and resources to help you implement your project. Make sure to read and use these documentation and resources.

# Supervised Data Mining (Classification) Binary Classification Only

- Implement 3 different classification algorithms in Python. One of them is Random Forest , the second one is from the deep learning list in the "Appendix → Additional Option: Deep Learning", and the third is from the list of algorithms in "Appendix → Additional Option: Algorithms" on 1 dataset of your choice (each of the three algorithms must run on the same dataset).

  You may also make it fun to try to solve one of the existing problems:

  ie: Quora Insincere Questions Classification, Predicting Diabetes from Medical Records

  **NOTE: This is not from scratch implementation, just use the existing libraries to implement the algorithms, but the performance metrics must be calculated manually. You may use "confusion_matrix" library to get TP, TN, FP, FN ONLY, then calculate the FPR, FNR, etc.… using the formulas from the slides.**
- Sources of data are listed in the Appendix "Additional Option: Sources of Data" or use your own.
- Your final term project documentation must clearly indicate the algorithms and dataset you used in the project.
- In addition to the general submission rules and grading, include the websites where the software and complete dataset can be downloaded.
- You must present experimental results that show the comparison of classification performance between the algorithms used in your project.
- In evaluating classification performance, students must use the 10-fold cross validation method. You must show the statistics as discussed in the "Evaluating Classifiers" module to include all parameters that were introduced: TP, TF, FP, FN, TSS, HSS, etc.. for each run of the 10-folds and also for overall as an average of all 10-folds execution.
- Provide the result of the metrics in tabular format listing all details for easier visualization (for each fold and average). Your Juypter Notebook should also show the result in tabular format.
- Provide a discussion about your results. Which algorithm performs better and why? Justify your answer.

**NOTE: If any thing is not clear, ambiguous, or doesn't make sense, consult with me right away and don't wait to the last minute.**

# Appendix

## **Additional Option**: General Sources of Algorithms/Software

- https://scikit-learn.org/stable/

## **Additional Option:** Algorithms:

- Select two classification algorithms from the following:

I. Algorithm (Support Vector Machines)
II. Algorithm (Decision Trees)
III. Algorithm (KNN, K-Nearest Neighbor)
IV. Algorithm (Bayesian Networks)
V. Algorithm (Naïve Bayes)

## **Additional Option:** Deep Learning:

I. Algorithm (LSTM)
II. Algorithm (Bidirectional-LSTM)
III. Algorithm (GRU)
IV. Algorithm (Conv1D)

## **Additional Option:** Sources of Data

1. http://aws.amazon.com/datasets
2. https://archive.ics.uci.edu/ml/index.php
3. And more you can find or your own…