**UTRGV High Scholars**

**Research Proposal Guidelines**

**Research proposals are due on or before 5:00 pm on June 11, 2018**. It is to be emailed to [mkhitar.hobosyan@utrgv.edu](mailto:mkhitar.hobosyan@utrgv.edu) no later than 5 pm on June 11, 2018. It should be two to three pages long and written in clear English using a Times Roman font 12, with one inch margins.

**Title of the Project:** Machine Learning Through Cross Validation on NCAA March Madness Results

**Name of the Student:** Wenhan Li

**E-mail of the Student:** wenhanli1010@gmail.com

**Name of the Faculty Mentor:** Alejandro Martinez

**Department of the Mentor:** School of Mathematical and Statistical Sciences

**Abstract:** This project attempts to predict the probabilities of match results (noted as a win or lose) in National Collegiate Athletic Association (NCAA) March Madness through simulated machine learning by a ten-fold cross-validation. The program, written in Python, will process and “learn” data from match results of March Madness from past years to predict correlations between data and the probabilities of results for upcoming seasons. The program compares different data analysis models to determine the integrity across models and evaluate their performance.

**Introduction**: A 250-500 word description of the background of the research project; include at least 4-5 key peer-reviewed original research papers on which the project is based on; clearly state what has been done and why your proposed research is new and original; also describe significance of the proposed research and any potential practical applications; references within the text need to be cited within brackets with the last name of the author and year of publication as (Grewal and Grewal, 2012)

With more advanced mechanics, data analysis has become more efficient and competent as driven by the increasing demand of giant data. While some data, specifically numerical data, can be represented by an algorithm, such as the fluctuation of economic exchanges, others cannot. Data like human behavior cannot be represented through an algorithm, but they can be recorded and analyzed, and through machine learning, which uses computation to recognize patterns and regularities in data, can be represented in “useful approximations” (Alpaydin, 2014). Often, machine learning is achieved using cross-validation.

Cross-validation is a statistical method of comparing and validating learning algorithms through the division of data into sections for learning and training. While the field is relatively new, it has grown quickly along with the growth of technology. The simplest form of cross-validation is the *k*-fold cross-validation, which divides the data into *k* sets, as equally as possible, and repeats *k* iterations of training and validating, using *k-1* sets to train and one set for validation each iteration (Refaeilzadeh, 2016). Other forms of cross-validation are also available, such as the “leave one out” method, which repeats *n* iterations of training and validating, where *n* is the cardinality of the set of data, with each iteration trains *n-1* data sets and is validated by only one data. In addition, cross-validation must also address accuracy and validity of models derived, which must not over-fit the data, where the model represents the data but does not function in predictions, and must represent the data.

This project aims to gain a deeper understanding of machine learning through cross-validation by attempting to predict the probability of win/loss of NCAA March Madness games. March Madness is a NCAA basketball league that includes 68 teams in a bracketed single-elimination tournament that is hosted annually around March (Wonderopolis). On a predictive analytics competition website named Kaggle, there is a competition by Google Cloud and NCAA that challenges competitors to 1) build and test models based on past data on win/loss records and 2) predict future outcomes of matchups (Kaggle). The competition has a prize of $50,000 for the past year, but it is closed at the moment. This project is aimed to challenge this competition in order to gain a better understanding of the process behind machine learning.

**Objectives**: state one or two clear objectives of your proposed research project

To gain a better understanding of machine learning through cross-validation.

To predict the probability of win/loss of future NCAA matchups through past data.

To compare learning models for integrity and validity.

**Hypothesis**: should be clearly articulated and should be testable

The mean log loss for logistic classification will be minimal.

**Materials and Methods**: briefly describe the methods, models or theories that will be used to conduct the proposed research; cite any references to the methods to be used; state exactly what data will be collected and how

A computer, data provided by Kaggle, Python, and Git will be used for the project. The dataset includes results from past March Madness matchups

The project will test the following models:

1. Logistic Classification
2. Support Vector Machines
3. Linear Discriminant Analysis
4. Decision Trees
5. Naïve Bayes
6. K Nearest Neighbors

**Experimental/Simulations Design and Data Analysis**: clearly describe the design of the experiment(s)/simulation(s) to be conducted including the number of treatments and replications; state exactly how the data collected will be analyzed

The project focuses on cross-validation to score different workflows. Workflows will take the form of

coder->imputer->filter->feature engineering->feature selector->model training->cross validation score

This is to ensure the separation of different jobs in creating an effective model, and the sections in the workflow will run through various combinations. The models will be trained and the data and scores will be presented in different forms of visualization through codes written in Python.

**Expected Results**: briefly describe the results you expect from your research/experiments or theoretical predictions

We expect to find the best workflows and evaluation them with the 2018 data, as well as the data that will be generated next year. We expect the logistic classification model to perform the best.

**List of References Cited**: Webpages cannot be used as citation source unless they are officially organizational (e.g. nasa.gov, in which case the full link should be given). Theses that are not published and available online, cannot be cited. Please chronologically list all the references cited in the research proposal including the name(s) of the author(s), year of publication, title of the study, name of the journal, volume of the journal and page numbers exactly as follows:

Alpaydin, Ethem. Introduction to Machine Learning. MIT Press, 2014.

Git, git-scm.com/.

“Google Cloud & NCAA® ML Competition 2018-Men's | Kaggle.” Kaggle, [www.kaggle.com/c/mens-machine-learning-competition-2018](http://www.kaggle.com/c/mens-machine-learning-competition-2018).

Refaeilzadeh P., Tang L., Liu H. (2016) Cross-Validation. In: Liu L., Özsu M. (eds) Encyclopedia of Database Systems. Springer, New York, NY

“What Is March Madness?” Wonderopolis, wonderopolis.org/wonder/what-is-march-madness.