# CS 7641: Project Proposal

By: Jaswanth Sai Pyneni, Vishal Vijayakumar, Hemanth Chittanuru, Kavin Krishnan, Rafael Hanashiro

## Introduction

The scope of our project is to analyze and model PUBG data. PUBG is a video game where 100 players battle on a game map by rummaging for weapons and tools and fighting until there is only one surviving player or team. We will be analyzing and modeling data of over 65,000 games to build an engine that can predict the final placement of any given player based on in-game stats and starting player ratings.

## Methods

We will be performing Principal Component Analysis (PCA) to reduce the effect of the Curse of Dimensionality. This phenomenon occurs when there are too many features in the dataset. This makes regression computationally more expensive, and there is a higher risk of overfitting to the training dataset. PCA converts a large set of features to a smaller set of uncorrelated features. Research indicates that using PCA improves prediction accuracy when dealing with high-dimensionality data [1]. Our training dataset has 29 features, and we propose that using PCA would improve our accuracy and performance.

Next, we will use Linear Discriminant Analysis (LDA) for a similar purpose. Unlike PCA, LDA is a supervised algorithm that maximizes the separability of classifications by finding the axes that maximize the distance between classes. At the same time, it minimizes the variation within the different classifications [2]. LDA is very effective when working with multiple classes. However, PCA is more optimal when the sample size of data is very small. LDA also has poor performance when it is difficult to linearly separate data.

We will be using artificial neural networks as one of our supervised learning approaches. Since this a regression problem, we want to create a function approximator to predict rankings. This is a perfect opportunity to use neural networks because it is well-established that given a sufficient number of hidden units and an applicable activation function, neural networks can serve as universal function approximators [3]. Neural networks also have the advantage of being able to create non-linear decision boundaries by stacking nonlinearities (non-linear activation functions), thus producing complex approximations that fit the data better.

Lastly, we will be using gradient boosting as one of our supervised learning models. Boosting methods are ensemble learners, which combine weak learners to form more powerful models. Gradient boosting typically uses decision trees as weak learners, and iteratively move against the gradient of the loss function to parametrize subsequent trees [4]. Lately, the XGBoost algorithm has gained a lot of popularity, as it explores second order methods (hessian updates) to better minimize loss. This results in faster convergence, which we plan on exploring.

## Results

By using these algorithms, we are trying to build a model that will predict the likely outcome/final placement of any given player, with a 0 denoting last place and 1 denoting first place. As such, we hope to utilize these algorithms to build an effective regression model to figure out the best strategy to win.

References:

1. Howley, T., Madden, M. G., O'Connell, M.-L., & Ryder, A. G. (2006). The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data. *Applications and Innovations in Intelligent Systems XIII*, 209–222. doi: 10.1007/1-84628-224-1_16
2. Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. AI Communications, 30(2), 169–190. doi: 10.3233/aic-170729
3. Hornik, K., Stinchcombe, M. B., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366. doi: 10.1016/0893-6080(89)90020-8
4. Friedman, J. H. "Greedy Function Approximation: A Gradient Boosting Machine." (Feb. 1999a)