**Applied Machine Learning: Project Proposal (Group 12)**

*Project Topic:*

* Powerlifting Competition – Predicting Winners, Comparative Performance Analysis, and Weight Class Recommendation
* We intend to build a model to predict winners or top performers based on the available historic data. This would help enthusiasts, coaches and athletes understand their competition and prepare; accordingly, we can utilize logistic regression (Supervised Learning) for this task or Gradient Boosting algorithms like XGBoost, LightGBM for a higher predictive accuracy.
* To calculate and compare the performance of athletes across different weight classes and divisions, we will use a linear or multiple regression (Supervised Learning) to predict performance scores; this comparative analysis aims to provide valuable insights into the strengths and weaknesses of the given athlete performance metrics and help practitioners in the field of powerlifting make informed decisions about the suitable metric in different scenarios.
* Again, to develop a model that recommends the best suited weight class based on an athlete’s attributes such as weight, age, body composition etc., a classification model (Supervised Learning) will be used.

*Dataset:*

* The Powerlifting dataset contains details about both powerlifting competitions (meets) and the athletes who participated in them. The dataset consists of 1048576 records and 37 attributes.  
  *(Link for reference* - <https://www.kaggle.com/datasets/open-powerlifting/powerlifting-database/data>)
* The dataset includes categorical features such as sex, event, equipment, age class etc. and numerical features like age, body weight, weight class, squat weight etc.
* The data contains information about the athletes' performances in different powerlifting events (Squat, Bench, Deadlift), including the weights lifted and their overall totals. Also, the “place” column indicates the ranking or position of the athlete in the competition.
* Information about the competition, such as the event date, location, and the name of the meet, is also available.

*Method:*

* After the initial data cleaning and pre-processing including handling missing or null values, encoding categorical features and scaling numerical features, we would proceed to perform an Exploratory Data Analysis (EDA)
* Since lifts are performed equipped and un-equipped, we would perform an equipment analysis to understand the most preferred equipment. Further, we would perform a gender and weight analysis to understand how the body weight is spread out for our data for each gender.
* Finally, we perform a winner analysis by age category, and plot appropriate visualizations for variations of athlete age over time and number of athletes YoY.