Final Project Proposal

Math 244, Spring 2025

Julia, Vishnu, and Mia

Dataset 1: Link

Introduction to data set: This data set includes statistics from the 2024 Division 1 and 3 Men's and Women's Ultimate Frisbee Championships. The statistics were found on USA Ultimate, the non-profit organization serving as the governing body for ultimate in the United States, and were taken from a data visualization titled "USA Ultimate 2024 Nationals Stats Dashboard", which was created by Ben Ayres. The data set includes 1665 rows which each correspond to an individual player, and it includes 15 variables which categorize the players by Division, Gender, and Team, and provide game statistics for each player.

Research question: Ultimate frisbee is a sport growing in popularity at the collegiate level and within the Vassar student body as well. The data set has 15 variables and over 1500 observations. While we do not have a concrete hypothesis in mind yet, the project would include a prediction aspect where we train the model under supervised conditions to be able to predict variables like a player's plus/minus score (AKA individual impact) based on other variables such as scores, assists and defensive plays. Such a model would be applicable to our teams at Vassar as we could test it in my currently developing AI model that can collect individual and team statistics to see if it can identify the most impactful player. Other comparisons that would be interesting to explore are between the men's and women's divisions and the division 1 and division 3 level.

Dataset 2: Link

Introduction to data set: The data set is the Employment Scam Aegean data set (EMSCAD) from the University of Aegean in Greece. This is a data set of 18k online job postings from 2012-2014 with rich information about job title, location, description, and requirements. The majority of jobs are real but ~800 are fraudulent. The data was scraped from the web and manually annotated. Some categorical variables of interest include telecommuting availability, the presence/absence of a company logo, industry, and required education. There are many open-ended text variables regarding job description, benefits, and qualifications.

Research question: Fraudulent job postings are becoming more common and intersect with issues of data privacy and illegal data harvesting. This data set has a rich combination of textual and categorical data for building a classification model. It has a large number (18,000) observations, so we have enough data to split the data set into training/testing data. The size and contents of the data set provides an exciting opportunity to explore more complex machine learning techniques to build a classification model that identifies job postings as real or fraudulent. This would be a supervised learning model, as the data set contains a variable that indicates if the job is fraudulent. We do not have a concrete hypothesis, since this project would be classification-based, but we believe that the richness of the data set will enable us to build a robust, accurate classification model.