**Progress Report**

1. **Project Overview**

The Olympics serve as a global platform for countries to showcase unity, sportsmanship, and national pride through team participation. Beyond competition, the Games promote cultural exchange, diplomacy, and the pursuit of physical excellence. The success of a nation's athletes fosters a sense of pride among its citizens and underscores the power of sports to unite people and build connections worldwide.

In this project, we seek to explore the trends in countries' participation in the Olympics by studying various indicators in the underlying data.

In this project, we started off with exploratory data analysis and hypothesis testing to better understand the factors responsible for the participation of various countries in the Olympics. Then, we used those insights to model the data better and create predictive models based on various machine learning algorithms including but not limited to RandomForest, boosting models, neural networks, auto-regressive models, etc.

1. **Progress**
   1. **Dataset Generation**

First step for the project was to collect a dataset which can be used to get to the project objective. For generating a dataset for factors that influence the participation of a country in Olympics we used World Bank as our primary source. It provides data of various indicators that can be found here <https://data.worldbank.org/indicator>. We categorized the indicators that can significantly influence the countries participation and created a dataset that contains the following indicators for each country from 1960 to 2022.

(Urban\_Pop, Pop\_0\_14,Pop\_15\_64, Net\_Migration, Mortality\_Rate, Pop\_Density, Female\_Pop, Electricity, Mobile\_User, Land\_Area, GDP, GDP\_PCAP, Foreign\_Invest, Inflation, GINI, Military\_Expense, Unemployment, Labor\_Force,Labor\_Female, Literacy Rate). These indicators represent the economical, geological, cultural and social factors of a country that can highly influence their participation in Olympics games.

For the Olympic data we took ‘ 120 years of Olympic History: athletes and results’ from Kaggle <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results>. The range of data was from 1896 to 2016.

These two dataset layed a foundation to create a combined dataset that represented all the indicators and Olympic data for countries over a period of 1960 to 2022. Based on this merged dataset we were able to further investigate the trends in countries’ participation in the Olympics.

* 1. **Participation Factors Analysis**
  2. **Features Generation**
  3. **Hypothesis Testing**
  4. **Baselines Model**

1. **Remaining Work**
2. **Conclusion**
3. **Appendix**