Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2523145

1 Introduction

We made a model that, from a 7 dimensional feature vector, attempts to predict a 4 dimensional value vector of the medals. We take both the year and the country, along with other features, as inputs.

We used multiple models — a tree, and a neural network — to approximate a

In section 2, we discuss the data, and what we did with

In section 3, we present our models

In section 4, we analyze the performance of our models and conclude our report.

1.1 Model Assumptions

We assume that time series is not critical

2 Data

Before beginning any analysis of the data, we had to clean and organize the data provided for this problem. This consisted of four distinct parts: (1) creating a list of countries to consider, (2) cleaning the data, (3) sorting and reducing the data to see what is valuable, and (4) organizing the data for analysis.

2.1 Considered Countries

Due to a myriad of historic and geopolitical reasons, countries once prevalent at the Olympics (e.g., the Soviet Union) no longer exist, others have split (e.g., Czechoslovakia), split and then re-merged (e.g., East and West Germany), and yet others have changed their names (e.g., Rhodesia to Zimbabwe). Moreover, there are several multinational teams, such as the mixed team (MIX) in the 1896, 1900, and 1904 Olympics, the West Indies Federation (WIF) in the 1960 Olympics, and the Unified Team (EUN), consisting of multiple former Soviet states, at the 1992 Olympics. To create a working list of states to consider for our analysis, without prejudice to current or past geopolitical events, for our analysis we only consider countries for each unique three-letter country code (NOC) in the `summerOly_athletes.csv` file. For the purposes of this report, we use ”country” to refer to those states.

The primary drawbacks of this approach is that it ignores the continuity in the athletic ability of several countries (e.g., Ukraine, Czechia), but such changes to the geographical base from which to draw athletes intrinsically kneecaps any attempt at continuity in these situations. Perhaps, with substantially more time and data, this could be competently addressed, but for the analysis we present we consider them as separate nations.

2.2 Cleaning Data

In the provided data sets, there were several issues which needed to be addressed. Immediately noticeable were several issues in the `summerOly_athletes.csv` file, including commas in the `Events` category and several nonsensical `Team` names. For instance, the Austrian sailer Harald Musil supposedly competed for the team `30. Februrar` and `May-by 1960`. To circumvent these issues, we did not consider the `Team` data in our analysis, and relied purely on the NOC country acronyms.

The NOC acronyms, however, were missing from the `summerOly_medal_counts.csv` and `summerOly_hosts.csv` files. To redress this, we simply mapped each country name to the respective NOC in Python. Luckily, the rest of the data that we used in our analysis (see sections 2.4 and 3 below) was already clean.

2.3 Data Pruning

After some initial investigations of the data, we noted that the results for the early Olympics (pre-WW1) were markedly different from those later on. Some notable differences included the host country fielding substantially more athletes than the visiting nations, and substantially fewer countries competeting. As such, we felt it prudent to ignore all data before 1920.

Similarly, the existence of one-off multinational teams (notably WIF and EUN, as mentioned above), would serve little use in any prediction of future medals, so we likewise chose to drop any teams which have only competed for one year between 1920 and 2024, inclusive. See Table 1 for a list of the countries that were pruned. While the reason for not including most are self-evident — clearly, Newfoundland is not a sovereign state sending athletes to the Olympics — three are particularly interesting. The teams “Individual Neutral Athletes,” “Unified Team,” and “Russian Olympic Committee” are all primarily or wholly Russian. After pruning, we were left with 218 nations to consider.

NOC	Name
AIN	Individual Neutral Athletes
ANZ	Australasia
BOH	Bohemia
CRT	Crete
EUN	Unified Team
NBO	North Borneo
NFL	Newfoundland
RHO	Rhodesia
ROC	Russian Olympic Committee
SAA	Saar
UAR	United Arab Republic
UNK	Unknown
WIF	West Indies Federation
YMD	South Yemen

Table 1: Countries that were dropped for competing a) only once or b) only before 1920

2.4 Data Organization

The basic strategy for our models is to map each country-year pair to their score — measured in amount of gold, silver, bronze, and total medals. Considering each unique country-year pair in our data set (e.g., we are not considering the Soviet Union in 2024 to be a data point), we arrive at a total of $N = 3083$ separate data points.

To build our models, we first organized our data into features (or $\boldsymbol{x}_i$ vectors) and values (or $\boldsymbol{y}_i$ vectors).

$$\boldsymbol{x}_i = \begin{bmatrix} \text{Country} \\ \text{Year} \\ \text{\# Athletes} \\ \text{Host} \\ \text{\# Events} \\ \text{\# Sports} \\ \text{\# Disciplines} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{y}_i = \begin{bmatrix} \text{\# Golds} \\ \text{\# Silvers} \\ \text{\# Bronzes} \\ \text{\# Total Medals} \end{bmatrix}.$$