**Artificial Intelligence Capstone Project1**

Student name: 吳權祐 Student ID: 110550014

1. **Introduction**

The 2024 Paris Olympics are coming just around the corner, as one of the most renowned sporting events globally, the result of the competition is focused by the entire world. I want to see if we can predict the result of the game. Based on this motivation, I collected the historical records of men’s 100 meters as my dataset and explored how machine learning and artificial intelligence can use the features to make predictions and improve sports.

1. **Dataset**
   1. Introduction

This is a dataset on the men’s 100 meters events in the Olympics from 1948 to 2020 (without 1952 since there is no wind information from that year).

* + 1. Compositions

Each row represents a record with following attributes about a competitor in a single game.

*Name*: string , name of the competitor, unused feature

*Nation*: int , code of the nation where the competitor from according to the dictionary (sorted by nation frequency in the dataset) (Appendix 1.2).

*Weight (kg)*: float , weight of the competitor

*Height (cm)*: float , height of the competitor

*BMI*: float , calculated by the formula

*Age*: float , age of the competitor, calculated from the birthday to the first day the event start

*Year*: int , the year of the event

*Round*: int , round in the competition, range from 1 – 4 (4 means final)

*Wind (m/s)*: float , wind information of the game

*isHometown*: bool , whether the venue of the event is the hometown of the competitor or not

*Label (s)*: float , performance of the competitor in the game

* + 1. Amountand Conditions

There are totally 2238 data in the dataset, and the label (performance) of data is conditioned to be under 11.5. And the dataset consists of 921 athletes and 180 nations.

* 1. Collection

The stages of collection consist of data scraping and data processing. And all these processes are performed on my laptop using the Visual Studio Code IDE.

* + 1. Scrape

All data is scrapped from the internet, particularly from the *Olympedia* website. I wrote a *sraper.py* (Appendix 1.1) scripts using the *requests* and *BeautifulSoup* packages to extract the raw data of all competitors in each competition and the wind information of the competition separately.

* + 1. Process

Then, I used a *generate\_train\_data.ipynb* (Appendix 1.2) to process the raw data and generate the training data. The packages used in the files are *pandas, re* and *datetime*.

In this stage, it was necessary for me to preprocess the data to make the following steps easier. First, I transformed any data with null values to match the format of the other values in the same columns. Then, some of the features were extracted from the raw data, such as *Weight* and *Height* from the *Body* column. Lastly, I dropped some useless attributes, such as *Gold*, *Silver*, and *Bronze*, which are used to record who gets the medal. I also filtered some information in the data. For example, the original *Birth* data contained information about the birthplace, and I filtered it to preserve only the pure birthday.

The next step in this stage is to process the data into the format.

* 1. External source
     1. Packages
* [requests](https://requests.readthedocs.io/en/latest/) (https://requests.readthedocs.io/en/latest/#)
* [BeautifulSoup](https://requests.readthedocs.io/en/latest/) (https://requests.readthedocs.io/en/latest/#)
* [pandas](https://pandas.pydata.org/docs/index.html) (https://pandas.pydata.org/docs/index.html)
* [re](https://docs.python.org/3/library/re.html) (https://docs.python.org/3/library/re.html)
* [datetime](https://docs.python.org/3/library/datetime.html) (https://docs.python.org/3/library/datetime.html)
  + 1. Websites
* [Olympedia men’s 100 meters events page](https://www.olympedia.org/event_names/40) (https://www.olympedia.org/event\_names/40)

1. **Methods**
   1. Supervised Learning
      1. Linear Regression
      2. SVM
   2. Unsupervised learning
2. **Experiments**
   1. Experiments Setting
   2. Experiments result
3. **Discussion**
4. **References**