AIT-614

Project proposal

The Proposal Title: Exploration and Analysis of Olympic History Data

Author's Names: Lakshmi Prasanna Gundabolu

Course Project Advisor/Professor's Name: Dr.Duoduo Liao

University Name, Course Number and Name: George Mason University, AIT-614, Big Data Essentials.

Date: April 11, 2020

Exploration and Analysis of Olympic History Data

Introduction

The cutting-edge Olympic Games are driving worldwide games including summer and winter sports rivalries in which a huge number of competitors from around the globe take an interest in an assortment of rivalries. The Olympic Games are viewed as the world's premier games rivalry within excess of 200 countries taking an interest. The Olympic Games are held at regular intervals, with the Summer and Winter Games exchanging by happening like clockwork yet two years separated.

120 years of Olympic data, this dataset is obtained from GitHub repository. A Combination of several factors like "Name", "Age”, Height", "Weight”, “Season” , "Sport", “Event" and "Medal" and a few more are considered for analysis. The dataset consists of 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event (athlete-events). In this project, we are going to preprocess the data and perform some visualizations to know about the dataset in detail. Implementing Correlation Analysis to show which factors are influencing the Athlete’s Performance and what would be the important factors we need to consider for estimating their Losing or Winning the sport in event. Moreover, We are going to use Machine Learning techniques for better prediction and also descriptive and inferential statistics for evaluating the accuracy of our prediction. Our proposed model helps in analysing the Olympic History from the past 120 years and also implementing some inferential statistics with detailed visualizations. For Analysing this dataset, we are using PySpark and also some Big Data tools.

Related Work:

Some related work was proposed on real-time Exploratory data analysis of the Olympics participation data set, to understand the impact of historical world events on the international games.

Source: Shaswat Rungta https://github.com/srungta/120-years-of-olympics

Source of original dataset:

We have chosen the data from the Github Repository with the following

link: (n.d.). Retrieved from https://github.com/srungta/120-years-of-olympics/tree/master/data

Objectives:

The main goal of our project is to the likelihood of a participant winning a medal (any medal i.e gold, silver or bronze) in an event based on his\her physical attributes and nationality is being predicted using logistic regression, which is machine learning technique. Here are the list of all the objectives of our overall project with clear description:

- Firstly, we are going to preprocess the dataset by checking Null values. Finding the unique values for the dataset.
- Additionally, formatting the dataset by deleting inconsistent attributes which are not used for our prediction and type casting the attributes which are incompatible.
- Visualizing a dataset in such a way that everyone can understand what are the factors that affect the behavior of a participant to win.
- Analyzing and exploring the Olympic medals of the dataset by using pySpark. Pyspark is used for big data analytics, for handling the big data and for parallel processing of huge amounts of data.
- Using correlation matrix to know which attributes have strong correlation with attribute and which are affected by it.
- Performing descriptive statistics by plotting graphs like scatter and boxplot to remove outliers in the dataset.
- Predicting winners by using Linear regression and clustering techniques of Machine learning.
- Comparing efficiencies of different Prediction and Classifying techniques.
- Performing Inferential statistics for finding the accuracy of models, and also by performing hypothesis tests we can calculate how accurate is our prediction.

Proposed selected dataset:

The data set is a set of 15 features: 5 numerical and 10 categorical. This dataset has 271,116 entries in the athletes_events.csv file. Each entry is based on an individual athlete competing in an individual Olympic event.

Some of the attributes that we are using in this dataset are as follows:

Name - Name of the Athlete

Age - Age of the Athlete

Height - Height of the Athlete

Weight -Weight of the Athlete

NOC - National Olympic Committee 3-letter code

Games - Year and season

Season - Summer or Winter

Sport -Name of the Sport

Event - Name of the Event

Medal - Type of Medal (Gold, Silver, Bronze, or NA)

Description of proposed system:

Data Preprocessing and Data Cleaning is done by filling missing values, dealing with outliers and excluding the duplicated data. Data Transformation will be done by normalizing the attributes. Subsequently, exploring the data will be done by calculating correlation analysis between attributes and utilizing feature importance algorithms to scale out attributes which would be useful in predicting the winners of the sports. Prediction and Classification can be done by using Various Machine Learning algorithms. Finding the Accuracy of prediction algorithms and also visualizing the dataset to know about the facts.

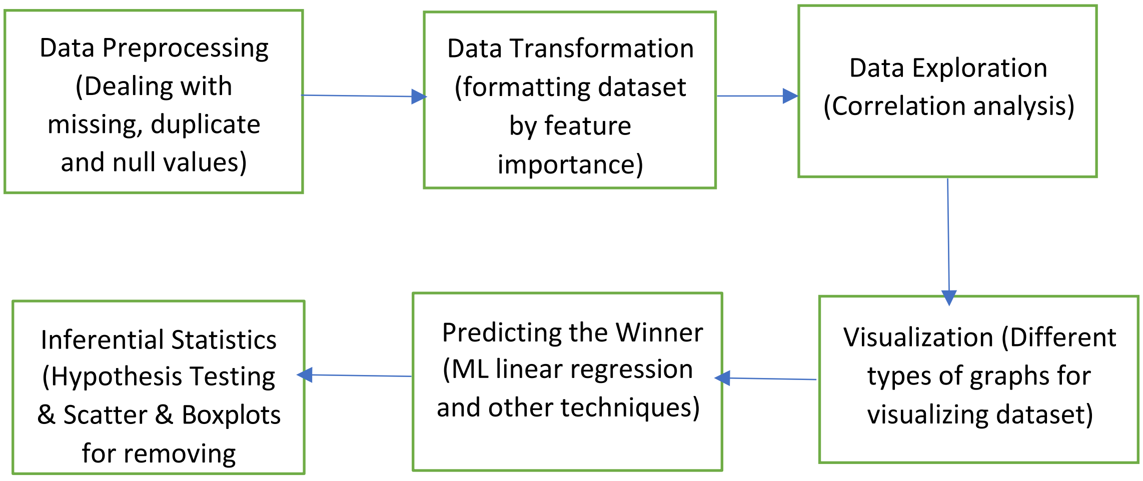
Additionally, We are using some BigData Tools like Pyspark for Exploring the dataset and some statistical analysis using pyspark will also be done.

Moreover we are using nltk technique word cloud for categorical variables in the dataset, which displays the

Exploring the olympic medals will be done by using pySpark as it is faster in analysing and retrieving the data from the huge dataset.

We are also using NoSQL for retrieving the important data in the dataset.

Conceptual diagram for our proposed system:


Data Preprocessing (Dealing with missing, duplicate and null values)
Data Transformation (formatting dataset by feature importance)
Data Exploration (Correlation analysis)
Visualization (Different types of graphs for visualizing dataset)
Predicting the Winner (ML linear regression and other techniques)
Inferential Statistics (Hypothesis Testing & Scatter & Boxplots for removing

Data Visualization:

We are using some plots to visualize the data. This includes Bar graphs, Density plots and other graphs if necessary for better visualization of the dataset. Visualizing the dataset using all the features for finding the correlation between them. Visualization helps with knowing some facts about the dataset.

Data analysis:

We need to preprocess the dataset before analysing it. For this, we will follow some steps like data preprocessing, Data exploration and Data Transformation. In Data Analysis we mainly have two methods, qualitative and quantitative. Inferential statistics comes under qualitative analysis methods as it performs hypothesis testing and compares the accuracies of different techniques for predicting the grades. Descriptive Analytics comes under the quantitative analytics method which calculates the central tendency of every attribute in the dataset.

Descriptive analytics:

This shows how the data has been dispersed with attributes. We can calculate mean, median and mode of numerical attributes in our dataset which we call as central tendency. It helps to get an in-depth understanding of variables in our data. For large numbers of statistics, the better way is to keep them in graphs like scatter plot for visualization.

Inferential analytics:

In Inferential Statistics, we perform hypothesis tests and also we can compare the average performance between the genders in the dataset by considering the t test for differences between groups. The analysis is useful in program outcome evaluation. By comparing the accuracies of different models we can easily find which model performs better for the dataset.

Advanced analytics (ML, DL, etc.):

In this project, we are going to use a Linear regression Model to predict the likelihood of participants winning the medal.

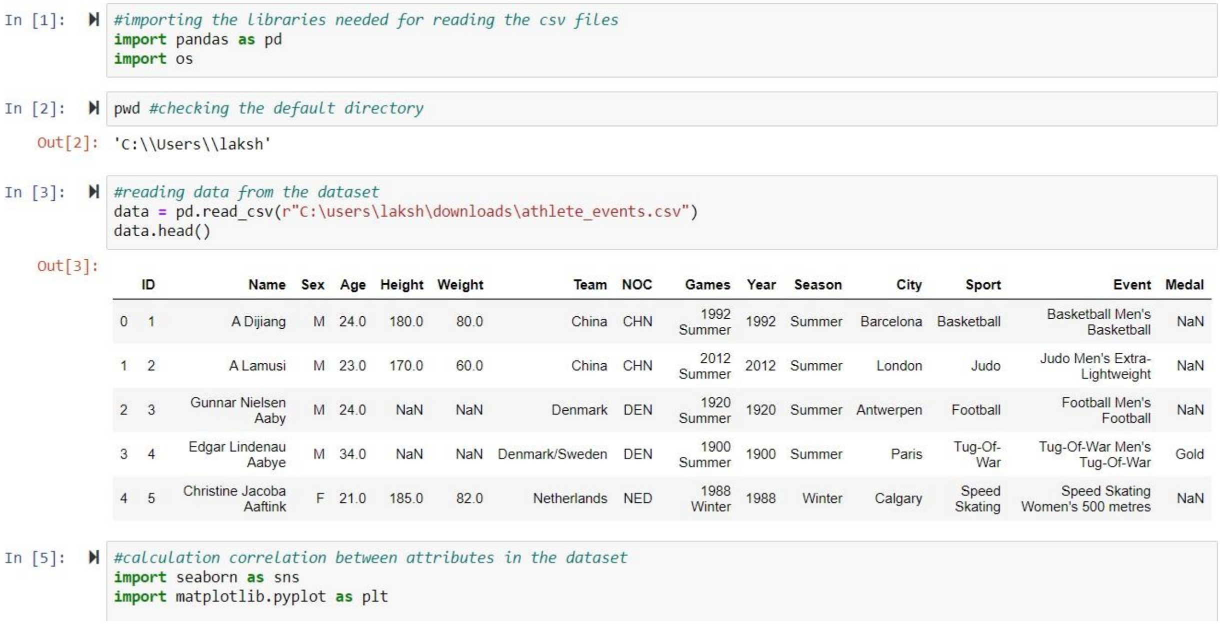
Proposed development platforms:

Frameworks: Jupyter notebook (python and R kernels), Tableau for visualization, Pyspark and NoSQL,NLTK, Machine Learning regression techniques.

Languages: Python and R

Hardware: Laptops with configuration Intel I7 8th gen processor, 512 SSD and 16 GB RAM, University Labs with high computational capacity and storage.

Show a piece of data as an appendix:



```
In [1]:  #importing the libraries needed for reading the csv files
         import pandas as pd
         import os
```

```
In [2]:  pwd #checking the default directory

Out[2]:  'C:\\Users\\laksh'
```

```
In [3]:  #reading data from the dataset
         data = pd.read_csv(r"C:\users\laksh\downloads\athlete_events.csv")
         data.head()
```

Out[3]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

```
In [5]:  #calculation correlation between attributes in the dataset
         import seaborn as sns
         import matplotlib.pyplot as plt
```