Name of the project : Building a machine learning model

Presented\_by : Group 6

Presented\_to : Keshav Dhir

Summary of Project :

We used the dataset team.csv in our project.Before starts the work on building a model ,we first imported some important libraries.

Following are some libraries for building a model :

* Pandas library: It is used for working data sets and data analysis.

import pandas as pd

* Matplotlib library: It is data visualization and graphical plotting library.

import matplotlib.pyplot as plt

* Seaborn library: It is a library for making statistical graphs.

import seaborn as sns

Loading a dataset:

We used pandas to read and explore our data in team.csv file that containing our data.

teams = pd.read\_csv(r"D:\Project\ALL CSV\teams.csv")

Perform EDA:

library has various functions such as head(), tail(), describe(), info(), shape () .

* teams.head() is used to fetch first five values.
* teams.shape() is used to get dimension of the dataset.
* print(teams.dtypes) is used to view to type of the datasets.
* teams.isnull().sum() is used to get the sum of null values.

After observing the dataset we cleaned data ,remove missing values, handling the missing values and duplicate records using various pandas functions to build a machine learning model. We also remove attributes from the dataset such as “height”, “weight”, “events” which are not essential for our predictions.

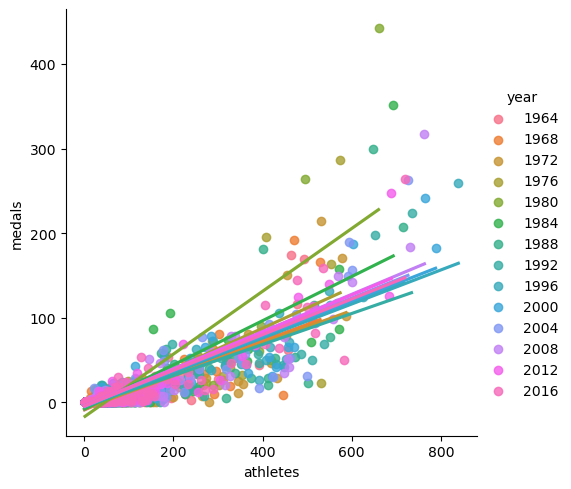
teams = teams[["team", "country", "year", "athletes", "age", "prev\_medals", "medals"]]

Building a model :

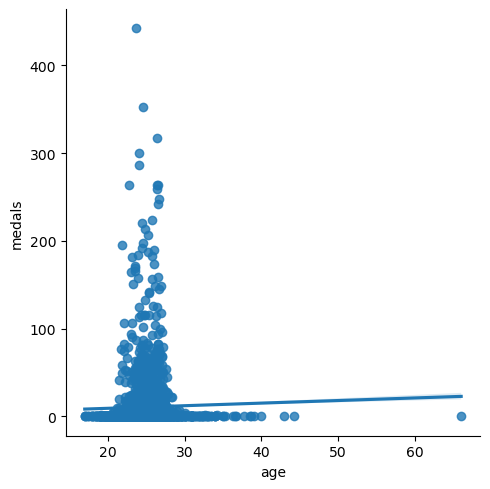
In teams dataset ,there are some numeric variables such as “year”, “athletes”, “age”, “prev\_medals”.We used [“athletes”,”medals”] and [“age”,“medals”] as continuous variables.Further we analysed the relationship between these variables using matplotlib.pyplot library.

Following are interesting patterns were seen in the dataset:

CODE:- sns.lmplot(x='athletes', y='medals', data=teams, fit\_reg=True, ci=None,hue='year')



As all the datapoints lies along a linear line and we see there is a strong linear relationship between the number of “athletes” and the number of “medals”.So,this showed that if the number of athletes increase the number of medals also increase.



This graph showed if age increased from age group 20-30 the number of medals also increased.

We used linear regression algorithm because our dataset met all assumptions of it.The data was split into training and test dataset with 80% and 20% split approximately.The training data was trained using fit() function which lead to the creation of best fit line.

train =data1[data1["year"]<2012].copy()

test = data1[data1["year"]>=2012].copy()

from sklearn.linear\_model import LinearRegression

predictor = ["athletes", "prev\_medals"]

target ="medals"

We defined predictors are "athletes","prev\_medals" and target is "medals" column.Then we test data was provided and target variables are predicted.