Olympic Games Analytics

Relazione e progetto realizzati da Sergio Maccarrone n. matricola X81000448

Introduzione:

Il progetto nasce con l'intento di predire quante medaglie una nazione riuscirà a conquistare alle Olimpiadi estive future.

I dati utilizzati per studiare l'andamento delle nazioni alle Olimpiadi è stato scaricato dai materiali della materia "Technologies for Advanced Programming" del corso di studi in Informatica.

Componenti e tecnologie utilizzate:

- Dataset di dati contenente informazioni riguardanti le Olimpiadi estive dal 1896 al 2012 disponibile al seguente link https://github.com/salvo-nicotra/tap/blob/master/spark/dataset/olympic-games/summer.csv
- Software Spider rilasciato da Anaconda
- Linguaggio python, con le seguenti librerie:
  - Pandas: per la manipolazione e l'analisi dei dati.
  - Pearsonr: utilizzato per calcolare l'indice di Pearson.
  - Pyplot: per la creazione di grafici, nello specifico per mostrare il grafico della regressione lineare.
  - Numpy: per poter operare efficientemente sulle strutture dati.
  - Sklearn.linear_model: per utilizzare la regressione lineare.
  - Seaborn: fornisce un'interfaccia di alto livello per disegnare grafici statistici.
  - Math: per l'utilizzo di funzioni matematiche ad alto livello.

Teoria e sviluppo:

Il dataset utilizzato è un file di tipo CSV, un formato basato su file di testo che viene utilizzato per l'importazione ed esportazione di una tabella di dati.
Nel dataset sono indicati l'anno, la nazione e la tipologia di medaglia.

```
Year,Country,Medal
1896,HUN,Gold
1896,AUT,Silver
1896,GRE,Bronze
1896,GRE,Gold
1896,GRE,Silver
1896,GRE,Bronze
1896,HUN,Gold
1896,GRE,Silver
1896,GRE,Bronze
1896,AUT,Gold
1896,GRE,Silver
1896,USA,Bronze
1896,HUN,Bronze
1896,USA,Gold
```

Figura 1: *immagine riportante una parte del dataset.*

Una volta importato il file su python, sono stati individuati e stampati tutti gli stati presenti nel dataset. Dunque, l'utente potrà inserire una nazione per cui vuole mostrata la predizione.

Una volta inserita la nazione, l'applicativo creerà un istogramma che mostra il numero di medaglie conquistate dalla nazione nelle edizioni a cui ha partecipato dal 1896 al 2012.

Istogramma: *plurirettangolo avente basi proporzionali all'ampiezza delle classi e aree proporzionali alle frequenze. Utilizzato per la rappresentazione grafica di caratteri quantitativi. Dove l'altezza del rettangolo della classe i-esima deve essere proporzionale al rapporto fra la frequenza della classe e la corrispondente ampiezza.*

$$h_i \propto \frac{f_i}{x_{i+1} - x_i}$$

Equazione 1: *altezza del rettangolo della classe i-esima.*

Codice per la creazione dell'istogramma:

```
sns.catplot(x="Year", kind="count", data=specificStateDF,aspect=2);
plt.show()
```

Figura 2*: codice per la creazione dell'istogramma.*
L'asse x mostrerà gli anni di partecipazione alle Olimpiadi e l'asse y la somma di medaglie ottenute in quell'anno.

Seguendo le istruzioni del codice, dopo aver mostrato l'istogramma, è stata realizzata una tabella che mostra il numero di medaglie conquistate in uno specifico anno dalla nazione inserita. Oltre al numero totale di medaglie, viene mostrata la somma cumulativa anno per anno.

Somma cumulativa: *la somma cumulativa, o totale generale, viene utilizzata per visualizzare la somma totale dei dati man mano che crescono nel tempo. Questo consente di visualizzare il contributo totale di un determinato criterio realizzato nel tempo.*

Codice per la creazione della somma cumulativa:

```
filterDf = specificStateDF.groupby(["Year", "Country"]).size().reset_index(name="Count")
filterDf['Cumulative_Sum'] = filterDf['Count'].cumsum()
```

Figura 3: *nel seguente codice viene presa la lista contenente tutte le informazioni riguardanti le medaglie conquistate alle Olimpiadi dalla nazione specificata e viene applicato un raggruppamento per anno e nazione mostrando il numero totale di medaglie conquistate. Inoltre, la tabella viene arricchita con l'informazione della somma cumulativa che sarà utile per dei calcoli successivi.*

Una volta che è stata definita la lista contenente le informazioni riguardanti la nazione inserita, prendendo in considerazione il numero di medaglie conquistate da tutte le nazioni per ogni Olimpiade in cui la nazione inserita ha partecipato, è possibile calcolare l'andamento di essa rispetto al quartile in cui si colloca.

Quartili: *In statistica, data una distribuzione di un carattere i quartili sono quei valori che ripartiscono la popolazione in quattro parti di uguale numerosità.*

$$Q\alpha = \frac{xm + x_{m+1}}{2}$$

Equazione 6: *formula per calcolare i quartili*

Procedimento:

- *fissiamo α = 0.25, 0.5, 0.75 e calcoliamo α(n + 1)*
- *se α(n + 1) = m (media) ∈ N allora Qα = xm*
- *se α(n + 1) ∉ N allora si prende la sua parte intera, che è quel numero m ∈ N tale che m < α(n + 1) < m + 1 e infine, si applica la media aritmetica tra i due valori di posizione m e m+1*

Implementazione Calcolo quartili:

```
tableAllCountry = df.groupby(["Year", "Country"]).size().reset_index(name="Count")
mat=[]
row=filterDf.values.tolist()
for r in row:
    year=r
    table=tableAllCountry[tableAllCountry.Year == year[0]]
    tableStateSpecifiedYear=table.values.tolist()
    a=[]
    for v in tableStateSpecifiedYear:
        a.append(v)
    mat.append(a)
i=0
quartiliState=[]
for m in mat:
    resultOlimp=sorted(m , key=lambda x: x[2])
    stateValues=filterDf.Count[i]
    i=i+1
    mat1=[]
    for r in resultOlimp:
        mat1.append(r[2])
    array=np.array(mat1)
    q1=np.quantile(array, 0.25)
    q2=np.quantile(array, 0.5)
    q3=np.quantile(array, 0.75)
    if stateValues < q1:
        quartiliState.append(1)
    elif stateValues < q2:
        quartiliState.append(2)
    else:
        quartiliState.append(3)
dfQuarts=pd.DataFrame(columns=['Year','Quartile'])
i=0
for y in filterDf.Year:
    dfQuart = pd.DataFrame({"Year":[y],"Quartile":[quartiliState[i]]})
    dfQuarts.loc[i]=[y,quartiliState[i]]
    i=i+1
```

Figura 4: *attraverso le seguenti istruzioni, viene popolata una lista che contiene [anno Olimpiade, nazione, somma medaglie vinte] così da poter effettuare un confronto per ogni edizione delle Olimpiadi tra le medaglie conquistate dalla nazione inserita e quelle conquistate da tutte le nazioni che hanno partecipato. Dopo aver calcolato i quartili di ordine 0.25, 0.5, e 0.75 della lista ordinata contenente la somma delle medaglie conquistate da tutte le nazioni suddivise per anno è possibile creare una lista secondaria riempita con le posizioni della nazione inserita rispetto ai quartili. Quest'ultima lista verrà utilizzata per creare il grafico mostrante per ogni edizione delle Olimpiadi estive la posizione della nazione rispetto ai quartili.*

Prese in considerazione le due sequenze di dati “anni in cui la nazione ha partecipato” e “somma delle medaglie che essa ha conquistate anno per anno” è possibile verificare se ci sia una correlazione tra i due set di dati. Il valore che definisce tale correlazione prende il nome di covarianza.

Covarianza: *quando si analizza un set di dati in cui si hanno più caratteri per un soggetto la correlazione ci permette di capire se esiste una relazione tra questi diversi caratteri. Per valutare la correlazione si può confrontare la variazione dei dati rispetto ai rispettivi valori medi del carattere. Supponendo di avere due soli caratteri X e Y, tanto più i prodotti $(xi - \bar{x})(yi - \bar{y})$ hanno concordanza di segno, tanto più i dati considerati hanno forte dipendenza. Anche nel caso in cui ci sia sempre discordanza di segno si evidenzia una forte dipendenza. Quando invece i segni sono sempre differenti, sicuramente non esiste una relazione tra i dati. Dunque, la covarianza indica se due serie di dati sono correlate positivamente, negativamente o non lo sono affatto. Più il valore assoluto del risultato è grande, più i dati sono correlati, mentre se il valore è uguale a 0 le due serie si dicono statisticamente incorrelate.*

Codice per il calcolo della covarianza:

```
#Calcolo Covarianza
covariance= np.cov(countlist,yearlist)[0][1]
print('- Covarianza: {}'.format(covariance))
```

Figura 5: *attraverso queste istruzioni, viene calcolata la varianza. Il metodo utilizzato è “cov()” definito nella libreria numpy, esso prende in input due serie di dati (somma totale medaglie conquistate in ogni partecipazione; anni di partecipazione) e ritorna il valore della covarianza tra esse.*

Lo step successivo è stato l’applicazione della regressione lineare ai dati che sono stati acquisiti e filtrati.

Regressione lineare: *date due sequenze di dati entrambi di numerosità n, per verificare se le due sequenze hanno un legame funzionale del tipo y = f(x) bisogna determinare una funzione f(x) tale da descrivere la relazione tra le due sequenze. Per definirla, possono essere applicate diverse tipologie di analisi che prendono il nome di analisi di regressione. Nel nostro caso è stata utilizzata la regressione lineare semplice dove la funzione f è una retta del seguente tipo f(x) = mx + q. La retta in questione può essere ottenuta attraverso l’utilizzo del “Metodo dei minimi quadrati” :*

$$g(m,q) = \sum_{i+1}^{n}[mx_i + q - y_i]^2$$

Equazione 2*: funzione residuo con f(x)=mxi + q.*

Si cerca f(x) tale che sia minima la funzione residuo “Equazione 2”, con i che va da 0 a n. Questa funzione rappresenta la somma dei quadrati delle distanze tra i dati sperimentali y(i) e quelli calcolati con la funzione f(xi).
Da questa formula possiamo trovare le incognite della nostra retta, m e q.

$$\text{Dove } m = \frac{c_{xy}}{s_x^2} \quad \text{e} \quad q = \bar{y} - \frac{c_{xy}}{s_x^2}\bar{x}$$

Equazione 3 e 4

Implementazione della regressione lineare:

```
#Preparazione per applicare la regressione

#Converto in array
yearlist=np.array(yearlist,dtype = int)
countlist=np.array(countlist,dtype = int)

#Preaprazione degli array
X = np.reshape(yearlist, (-1, 1))
Y = np.reshape(countlist, (-1, 1))

#Regressione lineare con apprendimento dal dataframe relativo alla nazione inserita
regsr=LinearRegression()
regsr.fit(X,Y)
```

Figura 6: *nelle seguenti istruzioni, dopo aver definito una lista contenente gli anni in cui la nazione ha partecipato alle Olimpiadi e una lista contenente le medaglie conquistate anno per anno, esse vengono convertite in due array e gli viene applicato un reshape così da modificare la struttura senza variarne il contenuto. I risultati del reshape vengono salvati in due array X e Y che saranno i dati da correlare. Infine, le due strutture vengono passate all'algoritmo di regressione lineare per apprendere dai dati, questa operazione si effettua attraverso l'utilizzo della funzione fit.*
Dunque, avremo:

- ***X: lista degli anni in cui la nazionale ha partecipato.***
- ***Y: lista delle medaglie conquistate in quell'anno.***

```
to_predict_x= [inputYear]
to_predict_x= np.array(to_predict_x).reshape(-1,1)

predicted_y= regsr.predict(to_predict_x)

print("Totale medaglie Previste da  {} nel {}: {}\n".format(inputStates,inputYear,int(predicted_y)))


#Coefficienti per la retta di regessione

m= regsr.coef_
print("Inclinazione (m): ",m)

c= regsr.intercept_
print("Intercetta (c): ",c)

#Visualizzazione grafico con retta di regressione

plt.title('Medaglie Grafico-Stato')
plt.xlabel('Anni')
plt.ylabel('Medaglie')
plt.scatter(X,Y,color="green")

new_y=[ m*i+c for i in np.append(X,to_predict_x)]
new_y=np.array(new_y).reshape(-1,1)
plt.plot(np.append(X,to_predict_x),new_y,color="blue")
plt.show()
```

Figura 7: *una volta che l'algoritmo ha imparato dai dati, l'utente dovrà inserire l'anno per cui desidera prevedere il numero di medaglie conquistate dalla nazione. L'applicativo dunque, avrà l'incognita x da utilizzare all'interno della funzione che ha creato, così da predire l'incognita y che sarà il numero di medaglie conquistate. Sfruttando la libreria utilizzata per la regressione lineare vengono definiti e mostrati i due coefficienti m e q. Attraverso i dati determinati dall'algoritmo di regressione lineare viene mostrato anche un grafico mostrante l'andamento della retta di regressione.*

Applicando la regressione lineare alle sequenze di dati precedentemente definite, è stata determinata la retta che meglio le approssima, ma non il grado di approssimazione, noto anche come *indice di Pearson*.

Indice di Pearson: *l'indice di correlazione di Pearson (anche detto coefficiente di correlazione lineare) tra due variabili statistiche è un indice che esprime un'eventuale relazione di linearità tra esse. Esso fornisce un valore compreso tra -1 e +1, dove +1 corrisponde alla perfetta correlazione lineare positiva, 0 corrisponde a un'assenza di correlazione lineare e -1 corrisponde alla perfetta correlazione lineare negativa.*

Codice per il calcolo dell'indice di Pearson:

```
yearlist = filterDf['Year'].tolist()
countlist = filterDf['Count'].tolist()
corr, _ = pearsonr(countlist,yearlist)
print('\n Indice Pearson: %.3f' % corr)
```

Figura 8: *dopo essere state definite le due liste contenenti anni di partecipazione e le medaglie conquistate, è stata utilizzata la funzione pearsonr della libreria scipy.stats a cui sono state passate le due liste.*

Infine, sfruttando i dati analizzati e predetti, vengono definite: media; quartili; varianza; scarto quadratico medio; scarto quadratico assoluto;

Media: *cioè la somma dei valori numerici divisa per il numero di valori considerati.*

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

Equazione 5: *media aritmetica*

Implementazione calcolo media:

```
#Calcolo differenza anni

lastSum=sortedDf['Cumulative_Sum'].iloc[-1]
countYear = len(yearlist)
l= list(range(countYear))
base = l[-1]+2
#+2 perché +1 è la differenza tra quelle che a cui ha partecipato e quella predetta e l'altro +1 è per il fatto che l'indice parte da 0

#MEDIA

avg = lastSum / base
print("- Media : {} \n".format(avg))
```

Figura 9: *dopo aver aggiunto la riga [anno inserito dall'utente e predizione delle medaglie conquistate] al dataframe filtrato per la nazione inserita, si ricava l'ultima somma cumulativa e il numero di Olimpiadi a cui la nazione ha partecipato, a questo dato viene aggiunto un 1 così da considerare l'olimpiade predetta. Queste istruzioni ci permettono di calcolare la media così definita: numero totale di medaglie/il numero di Olimpiadi a cui ha partecipato (si considera che dal 2012 [ultimo anno di cui si hanno dati] all'anno inserito dall'utente la nazione non abbia partecipato a nessuna Olimpiade).*

Varianza: *la varianza di una variabile statistica X è una funzione che fornisce una misura della variabilità dei valori assunti dalla variabile stessa. Nello specifico, essa definisce la misura di quanto i valori si discostino quadraticamente dalla media aritmetica. Tanto è più grande il suo valore, tanto più i dati sono distanti dalla media.*

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Equazione 7: *formula per calcolare la varianza.*

Implementazione calcolo varianza:

```
#VARIANZA

lst = sortedDf['Count'].tolist()
variance = np.var(lst)
print("- Varianza: {} \n".format(variance))
```

Figura 10: *anche qui si sfrutta un metodo definito nella libreria Numpy, il metodo in questione è il "var()", che preso in input la lista ordinata delle medaglie conquistate suddivise per anni di partecipazione, ne ritorna la varianza.*

Scarto quadratico medio: *lo scarto quadratico medio è un indice di dispersione statistico, vale a dire una stima della variabilità di una popolazione di dati. È uno dei modi per esprimere la dispersione dei dati intorno ad un indice di posizione.*

$$s = \sqrt{s^2}$$

Equazione 8: *formula per il calcolo della varianza.*

Implementazione calcolo scarto quadratico medio:

```
#Scarto quadratico medio
print("- Scarto quadratico medio (DEVIAZIONE STANDARD): {} \n".format(math.sqrt(variance)))
```

Figura 11: *Per calcolare lo scarto quadratico medio è stata applicata la radice quadrata alla varianza.*

Scarto medio assoluto: Lo scarto medio assoluto è un indice di dispersione che misura la distanza dalla media aritmetica.

$$s.a. = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

Equazione 9: *Formula per il calcolo dello scarto medio assoluto.*

Conclusioni finali:

Analisi dei dati che si ricavano inserendo come nazione “ITA” e come anno “2020”.

Informazioni riguardanti le medaglie vinte dall’Italia alle Olimpiadi estive:

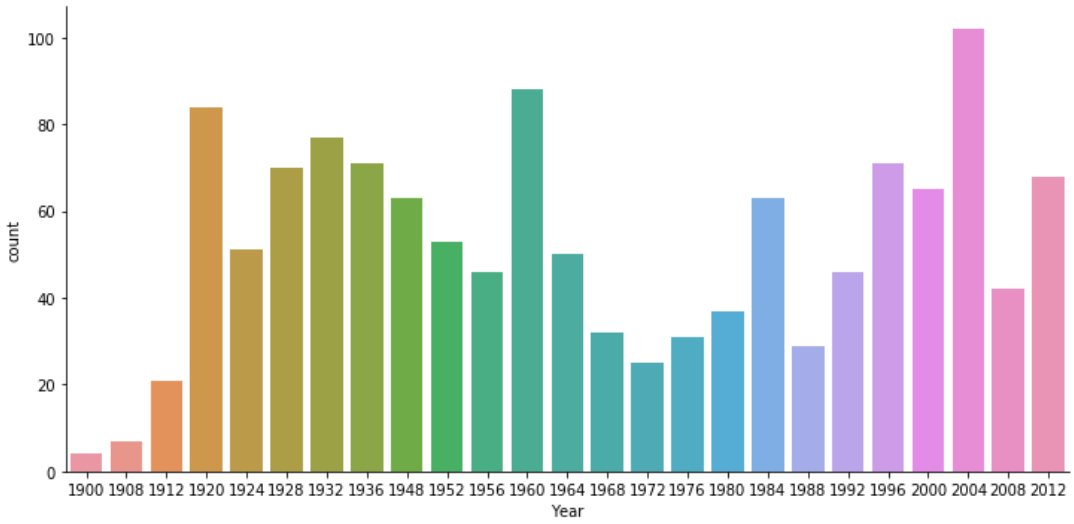

Figura 12: *Rappresentazione attraverso istogramma delle Olimpiadi estive a cui l’Italia ha partecipato e le medaglie conquistate in ogni edizione.*

Year	Country	Count	Cumulative_Sum
1900	ITA	4	4
1908	ITA	7	11
1912	ITA	21	32
1920	ITA	84	116
1924	ITA	51	167
1928	ITA	70	237
1932	ITA	77	314
1936	ITA	71	385
1948	ITA	63	448
1952	ITA	53	501
1956	ITA	46	547
1960	ITA	88	635
1964	ITA	50	685
1968	ITA	32	717
1972	ITA	25	742
1976	ITA	31	773
1980	ITA	37	810
1984	ITA	63	873
1988	ITA	29	902
1992	ITA	46	948
1996	ITA	71	1019
2000	ITA	65	1084
2004	ITA	102	1186
2008	ITA	42	1228
2012	ITA	68	1296

Figura 13: *Rappresentazione tabellare delle Olimpiadi estive a cui l’Italia ha partecipato e le medaglie conquistate in ogni edizione.*

Quartili:

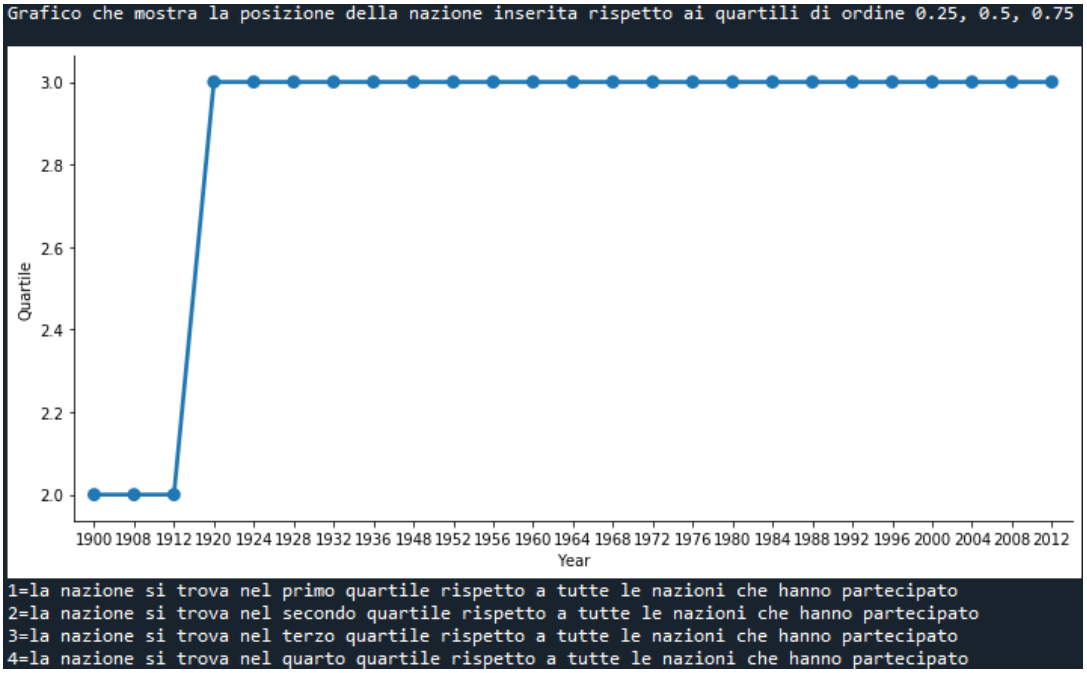

Figura 14

La *figura 14* mostra il grafico contenente la posizione dell'Italia anno per anno rispetto ai quartili calcolati sul numero totale di medaglie conquistate da tutte le nazioni in una edizione. Dall'immagine si può notare che l'Italia nelle prime 3 edizioni delle Olimpiadi estive a cui ha partecipato ha conquistato un numero inferiore alla mediana (quartile di ordine 2) ma superiore al 25%(quartile di ordine 1) delle nazioni partecipanti. Mentre dal 1920 in poi si è avuta una crescita poiché l'Italia ha conquistato un numero superiore alla mediana, dunque, essa ha avuto un andamento superiore al 50% delle nazioni partecipanti.

Covarianza:

Figura 15

Nella *figura 15,* è mostrata la covarianza tra gli anni di partecipazione alle Olimpiadi estive e le medaglie conquistate in ogni edizione. Analizzandone il risultato è possibile affermare che:

- le due serie di dati non sono statisticamente incorrelate, poiché la covarianza è diversa da 0.
- le due serie sono fortemente correlate, poiché il valore della covarianza è grande.
- le due serie di dati sono correlate positivamente, poiché il valore della covarianza è >0.

Grafico della regressione lineare:

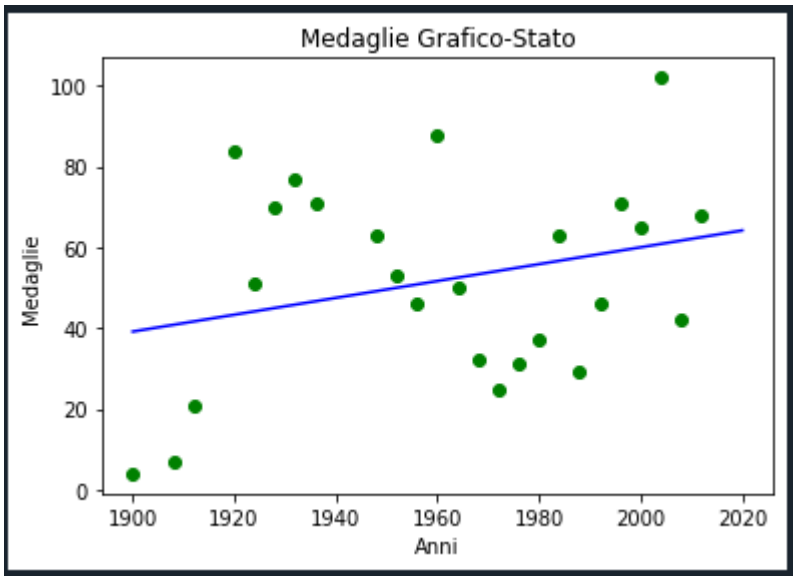

Figura 16: *rappresentazione della retta che meglio approssima i dati.*

Dalla *figura 16*, si evince che l'andamento delle *conquiste olimpiche* italiane è molto imprevedibile, esse non seguono un andamento lineare. Analizzando il grafico si intuisce anche che dalle prime edizioni dell'Olimpiadi fino agli anni 30 c'è stato un netto miglioramento e l'andamento è cresciuto, mentre nei 30 anni successivi l'andamento è andato verso il basso. Dagli anni 70 in poi l'andamento si è alternato. Inoltre, l'applicativo ha predetto che l'Italia nelle olimpiadi del 2020 conquisterà 63 medaglie.

Indice di Pearson:

```
Indice Pearson: 0.281
```

Figura 17

La *figura* 17 mostra il grado di approssimazione tra le due serie di dati. Analizzando il valore dell'indice si può notare che:
- il valore è maggiore di 0, dunque, la retta che approssima le due sequenze di dati è ascendente (come si evince dalla *figura 16*).
- il valore è minore di 0.9, dunque, i dati si allontanano da un andamento rettilineo.

Inoltre, si può affermare che il legame tra il numero di medaglie conquistate dall'Italia e le edizioni a cui ha partecipato hanno un legame molto basso, ciò significa che indipendentemente dall'anno in cui essa partecipa il numero di medaglie conquistate è molto imprevedibile.

Varianza:

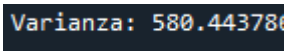

Figura 18

Figura 19

Nella *figura 18* viene mostrato il valore della varianza relativa alla lista delle medaglie conquistate dall'Italia. Analizzando l'indice di varianza si può affermare che essendone risultato un valore molto alto, i dati appartenenti alla lista sopracitata sono molto distanti dalla media (*figura 19*) e quindi non omogenei.

Scarto quadratico medio:

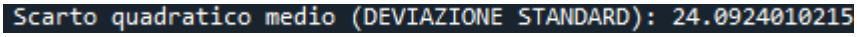

Figura 20

Nella *figura 20* è presente il valore assunto dalla deviazione standard per la lista delle medaglie conquistate dall'Italia. Essa indica quanto, in media, ognuno dei valori nella distribuzione devia dalla media. Il concetto è simile a quello della varianza, la differenza principale è che il valore della deviazione standard è espresso utilizzando la stessa unità della media, dunque, nel nostro caso, è possibile affermare che in media il numero di medaglie conquistate dall'Italia si allontana di 24,092 (unità di misura: medaglie conquistate in una edizione) medaglie.