



Sketching data analysis with INFN Cloud

Uso e sviluppo di applicazioni
e servizi su INFN Cloud

Emidio Giorgio

Sketching data analysis with INFN Cloud

- Use case
- Implementazioni



Lo use case

- Simulare uno use case scientifico
- Data una tabella con migliaia di entries relative a pazienti e relative *variabili predittrici* (*età, BMI, etc*), individua una matrice di correlazione tra queste variabili e l'essere diabetici
 - Dati reali, anonimizzati
 - circa metà delle entries è formata da persone sane



Input data e analysis framework



- Pandas : è un framework, basato su python, open source, per la manipolazione e l'analisi dei dati
 - è solo uno dei tanti ambienti disponibili

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|-------------|---------|---------------|---------------|---------|------|--------------------------|-----|---------|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

```
1 import pandas as pd
2 import pandas_profiling as pp
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5 import plotly.offline as py
6 import plotly.graph_objs as go
7 import plotly.tools as tls
8 import plotly.express as px
9 import numpy as np
10
11
12 def df_to_plotly(df):
13     return {'z': df.values.tolist(),
14           'x': df.columns.tolist(),
15           'y': df.index.tolist()}
16
17 df = pd.read_csv('diabetes.csv')
18 pp.ProfileReport(df)
19
20 color = sns.color_palette()
21 get_ipython().run_line_magic('matplotlib', 'inline')
22 py.init_notebook_mode(connected=True)
23 dfNew = df.corr()
24 fig = go.Figure(data=go.Heatmap(df_to_plotly(dfNew)))
25 fig.show()
26 fig.write_image('heatmap.png')
27
```

Le implementazioni proposte



- Sono state preparate tre implementazioni
- Jupyter : https://baltig.infn.it/corsi_formazione_ccr/clueapp22/-/wikis/Jupyter-with-persitence-for-Notebooks
- Docker-compose:
https://baltig.infn.it/corsi_formazione_ccr/clueapp22/-/wikis/Docker-Compose
- Kubernetes:
https://baltig.infn.it/corsi_formazione_ccr/clueapp22/-/wikis/Kubernetes

Referenze



- [Data analysis in python](#)
- [Pandas](#)