Agritech Preprocessing

- Elaborato 1 Information Systems and Business Intelligence -

Simone Dotolo M63001503
Fabio Boccia M63001541

Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione
Università degli Studi di Napoli Federico II

Indice

Capitolo 1

Introduzione

1.1 Traccia

Google Colab per il Trattamento dei Dati su Dataset Allegato:

- Preparare un notebook in Google Colab.
- Includere codice per importare il dataset, eseguire l'analisi esplorativa dei dati, la pulizia e la trasformazione dei dati.
- Applicare tecniche di Analisi dei Dati.
- Assicurati di commentare ampiamente il codice per illustrare i vari passaggi.

1.2 Configurazione dell'ambiente di sviluppo

Per poter eseguire il notebook Colab presente nel progetto [1] è necessario caricare il file sul proprio *Google Drive*. Inoltre bisognerà caricare anche il dataset *temp_humid_data.xlsx* e per permettere la visualizzazione dei dati, cambiare la variabile *filename*, inserendo il path corretto (il path del Dataset caricato su *Google Drive*).

```
# Caricamento dei dati
filename = '/content/drive/MyDrive/Colab Notebooks/ISBI/Esercizio1/temp_humid_data.xlsx'
```

[1] https://github.com/simone-dotolo/ISBI

Capitolo 2

Preprocessing

2.1 Caricamento e visualizzazione dei dati

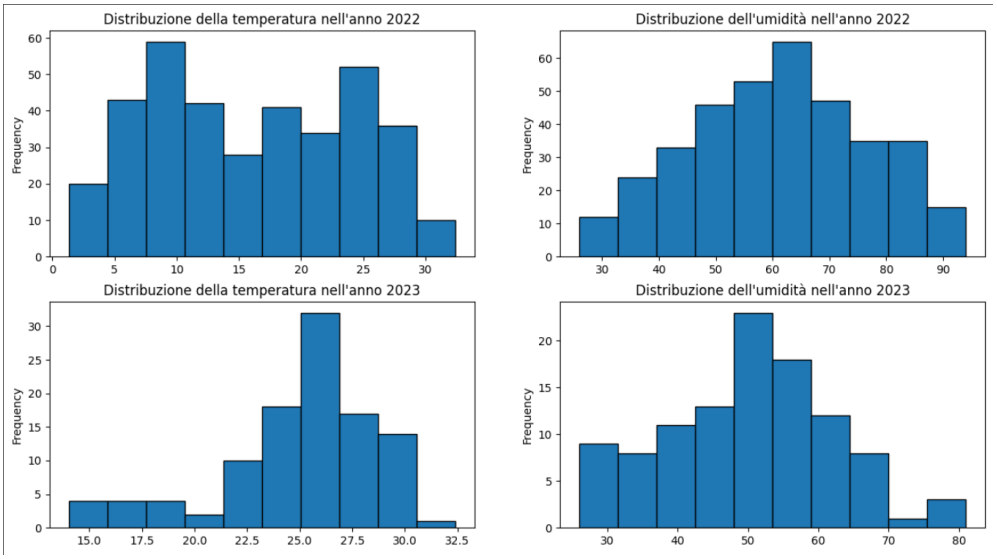
Inizialmente vengono caricati i dati in un *DataFrame* a partire dal file *temp_humid_data.xlsx* fornito. Da una prima analisi è possibile osservare che i dati sono relativi a delle misurazioni di temperatura ed umidità effettuate negli anni 2022 e 2023, e per l'anno 2023 è presente l'andamento dei parassiti presenti nel luogo dove sono state effettuate le misurazioni.

```
Dati anno 2022
         time  temperature_mean  relativehumidity_mean
0  2022-01-01             11.22                     77
1  2022-01-02              9.87                     86
2  2022-01-03              9.33                     79
3  2022-01-04             11.05                     72
4  2022-01-05             10.17                     73

Dati anno 2023
        Date  no. of Adult males  temperature_mean  relativehumidity_mean
0 2023-06-15                   1             24.62                     45
1 2023-06-16                   1             26.79                     46
2 2023-06-17                   0             26.02                     53
3 2023-06-18                   1             25.04                     48
4 2023-06-19                   0             25.09                     43
```

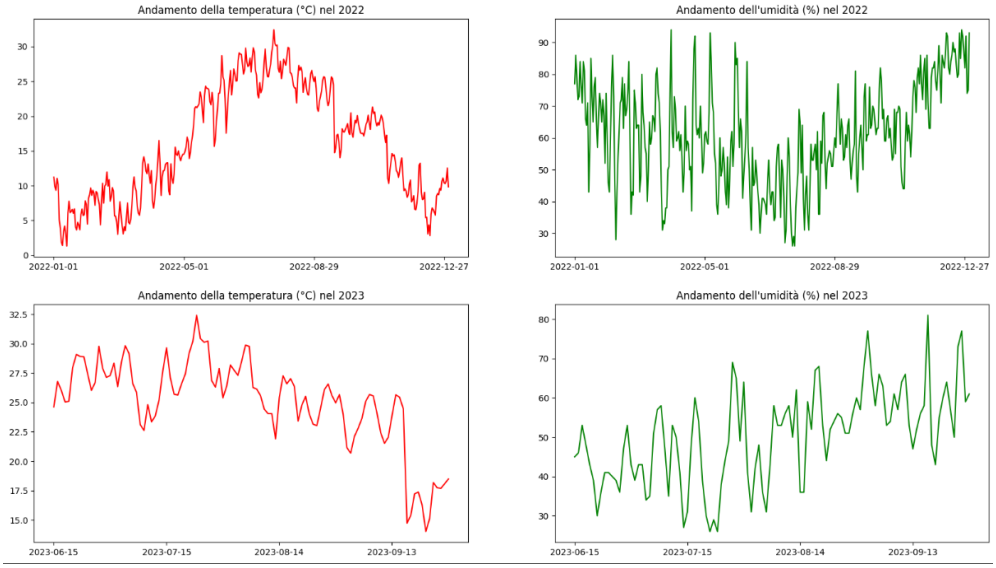
Sono poi state visualizzate alcune statistiche relative ai dati. Per l'anno 2022 abbiamo 365 misurazioni, mentre per l'anno 2023 abbiamo 106 misurazioni (ogni misurazione è relativa ad un giorno). Inoltre per ogni colonna vengono stampati i valori massimi, i valori minimi, le medie, le deviazioni standard ed i percentili.

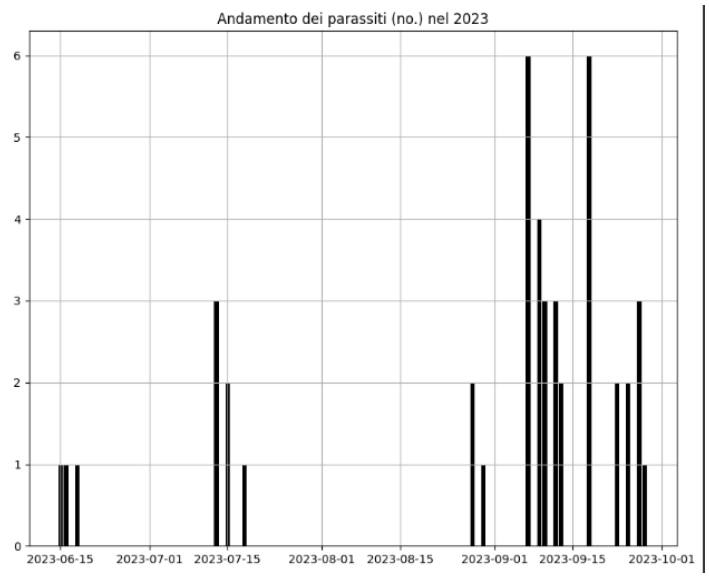
```
Dati anno 2022
       temperature_mean  relativehumidity_mean
count        365.000000             365.000000
mean          16.038740              61.249315
std            7.965726              15.660750
min            1.330000              26.000000
25%            9.150000              50.000000
50%           15.410000              61.000000
75%           23.410000              72.000000
max           32.410000              94.000000

Dati anno 2023
       no. of Adult males  temperature_mean  relativehumidity_mean
count          106.000000        106.000000             106.000000
mean             0.415094         25.015566              50.283019
std              1.120101          3.768792              11.928162
min              0.000000         14.030000              26.000000
25%              0.000000         23.700000              41.250000
50%              0.000000         25.640000              51.500000
75%              0.000000         27.297500              58.000000
max              6.000000         32.410000              81.000000
```

Successivamente si passa alla visualizzazione delle distribuzioni di ogni colonna. Tutte le distribuzioni risultano essere unimodali, fatta eccezione per la distribuzione delle temperatura nell'anno 2022. A differenza delle misurazioni di temperatura dell'anno 2023, che ricoprono solo un periodo di circa 3 mesi a partire dal mese di Giugno, le misurazioni di temperatura dell'anno 2022 ricoprono un anno intero. Per questo motivo è chiaro il motivo per il quale sono presenti due picchi: un picco è relativo ai mesi Primaverili-Estivi mentre l'altro picco è relativo ai mesi Autunnali-Invernali.



Mediante un plot sono stati visualizzati gli andamenti della temperatura, dell'umidità e dell'andamento dei parassiti negli anni 2022 e 2023.





2.2 Processing ed analisi dei dati

Per il processing dei dati, è stata prima verificata la presenza di eventuali righe con valori nulli, ma il riscontro è stato negativo.

```
Valori mancanti per dati 2022
time                    0
temperature_mean        0
relativehumidity_mean   0
dtype: int64

Valori mancanti per dati 2023
Date                    0
no. of Adult males      0
temperature_mean        0
relativehumidity_mean   0
dtype: int64
```

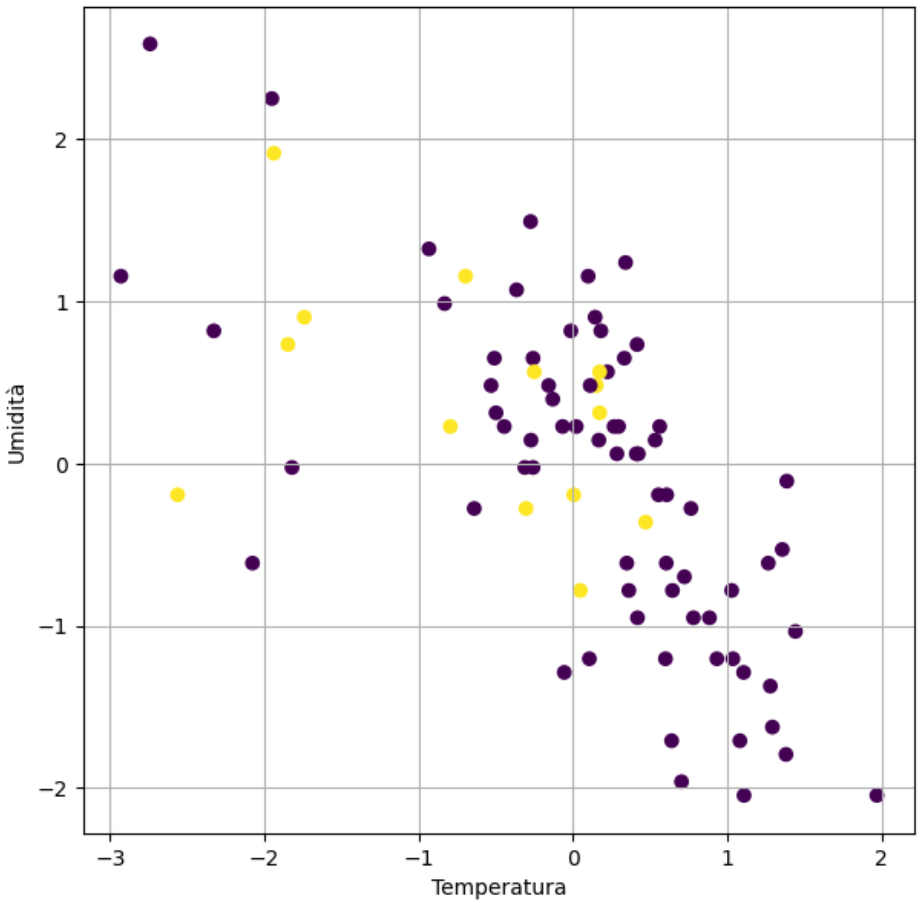
É stata poi effettuata la standardizzazione dei dati, rendendo ogni colonna relativa a temperatura ed umidità, a media nulla e varianza unitaria. Sono state poi visualizzate nuovamente le statistiche.

```
Dati anno 2022 normalizzati (media nulla e varianza unitaria)
         temperatura       umidità
count  3.650000e+02  3.650000e+02
mean  -1.168015e-16 -1.265350e-16
std    1.001373e+00  1.001373e+00
min   -1.849038e+00 -2.253896e+00
25%   -8.659845e-01 -7.192987e-01
50%   -7.903897e-02 -1.594159e-02
75%    9.266422e-01  6.874155e-01
max    2.058034e+00  2.094130e+00

Dati anno 2023 (media nulla e varianza unitaria)
         temperatura       umidità
count  1.060000e+02  1.060000e+02
mean  -6.933657e-16  2.272815e-16
std    1.004751e+00  1.004751e+00
min   -2.928725e+00 -2.045443e+00
25%   -3.507267e-01 -7.608827e-01
50%    1.664726e-01  1.025106e-01
75%    6.083580e-01  6.500282e-01
max    1.971338e+00  2.587398e+00
```

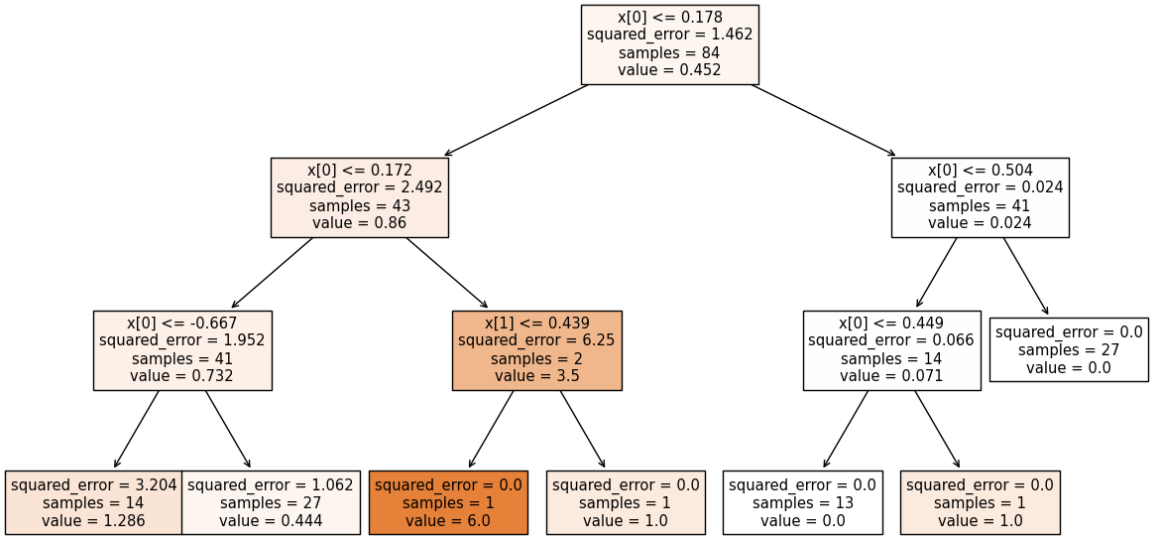
Per l'analisi dei dati, sono stati presi in considerazione i dati relativi all'anno 2023. L'obiettivo è quello di utilizzare temperatura ed umidità per predirre quanti parassiti sono presenti. Inizialmente è stata effettuata una divisione sui dati, in modo da ottenere un Dataset per l'addestramento (80% dei dati) ed un Dataset per il testing (20% dei dati).

```
Dimensione del dataset di training 84
Dimensione del dataset di test 22
```

Sui dati relativi all'addestramento è stato realizzato uno scatter plot, in cui per ogni punto $(temperatura, umidità)$ viene visualizzato un puntino di colore giallo, se per quel punto sono presenti parassiti, mentre un puntino di colore viola se non sono presenti parassiti.

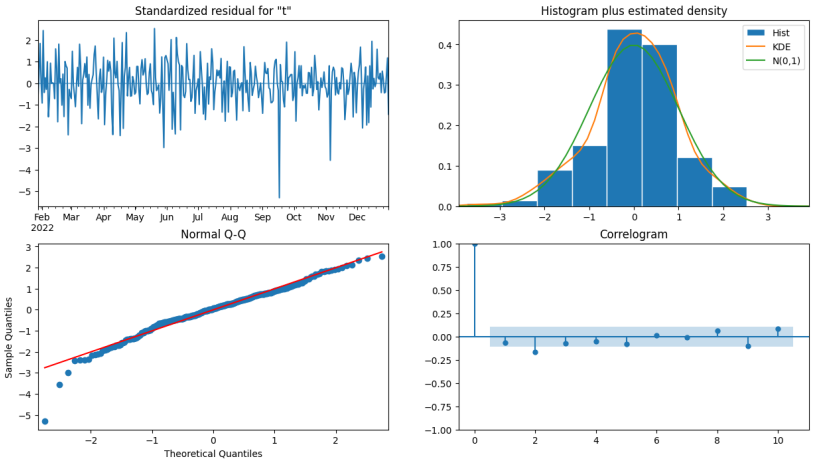


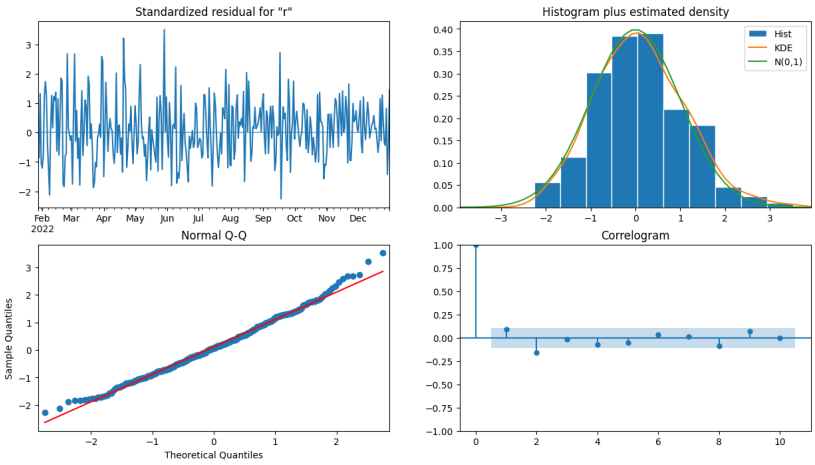
Infine è stato addestrato con i dati a disposizione un *Decison Tree Regressor*.



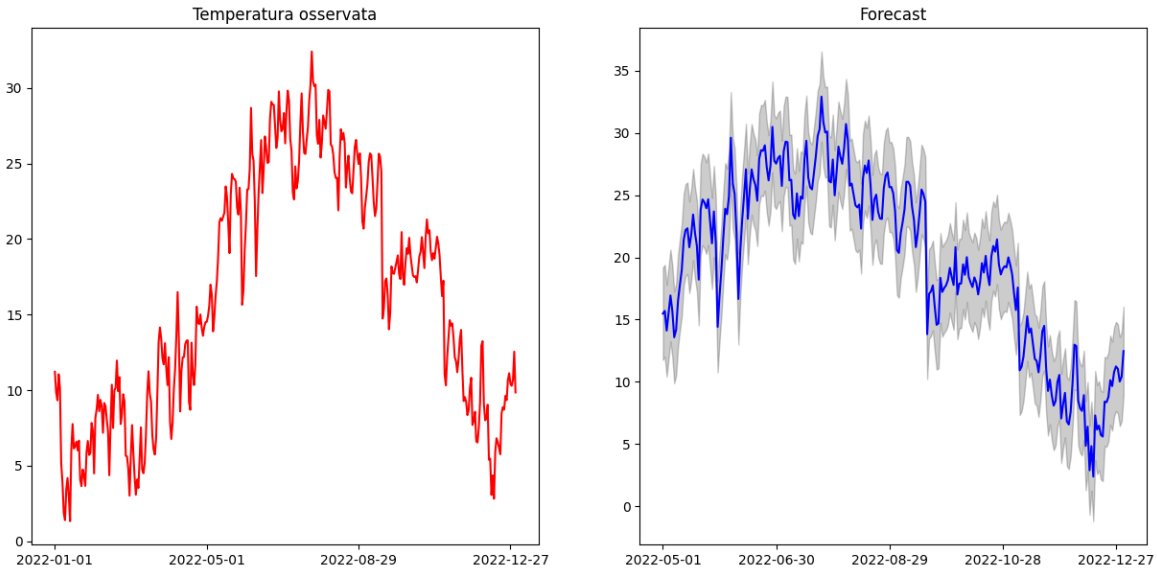
2.3 Analisi delle serie storiche

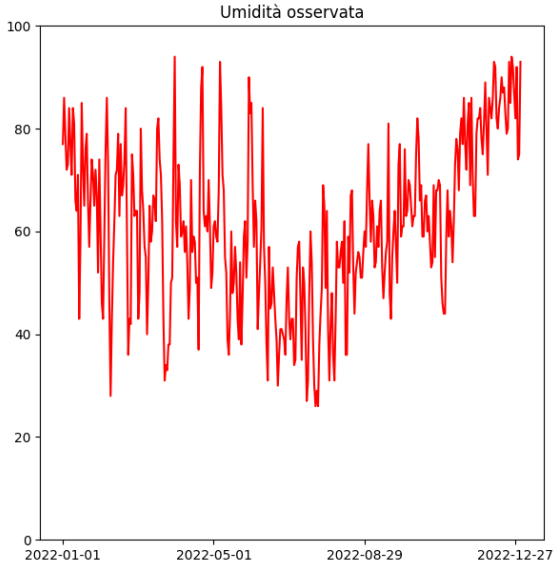
Per effettuare un'analisi della serie relativa alle temperature ed all'umidità dell'anno 2022, è stato utilizzato un modello SARIMAX, i cui parametri sono stati scelti mediante una *grid search*. Una volta trovato il modello migliore (quello con l'AIC più basso), sono state visualizzate delle statistiche relative ai residui delle previsioni, in particolare è possibile osservare che la distribuzione dei residui è pressocchè gaussiana con media nulla e varianza unitaria.

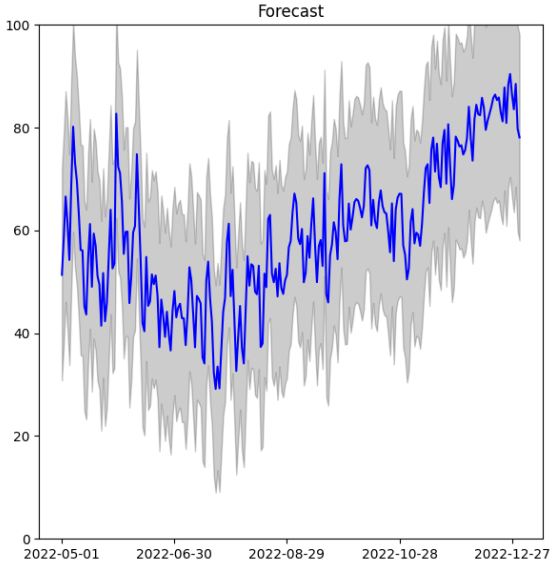




Infine sono state generate delle previsioni su temperatura ed umidità con i relativi intervalli di confidenza.




Umidità osservata
100
80
60
40
20
0
2022-01-01
2022-05-01
2022-08-29
2022-12-27


Forecast
100
80
60
40
20
0
2022-05-01
2022-06-30
2022-08-29
2022-10-28
2022-12-27