Documentazione progetto Ingegneria della Conoscenza 2021-2022

Sommario

Introduzione

Questo progetto è stato realizzato per l'esame di Ingegneria della Conoscenza dalle studentesse Nardiello Rosalba (Matricola: 717154 Mail: r.nardiello3@studenti.uniba.it) e Pizzimenti Francesca(Matricola: 716031 Mail: f.pizzimenti@studenti.uniba.it).

Il nostro sistema è in grado di prevedere se un soggetto è potenzialmente a rischio di Cancro al seno o meno a seconda dei valori riscontrati nel dataset preso in considerazione. PAUSA.

Requisiti Funzionali

Il progetto è stato scritto interamente in Python usando l'ambiente di sviluppo: Visual Studio Code; relativamente alle librerie esterne sono state installate sulla macchina:

- Pandas: usato per l'importazione del Dataset in formato .csv;
- Scikit-learn: per applicare i concetti del Machine Learning;
- Numpy: per lavorare con array multidimensionali e applicare specifiche operazioni logiche/matematiche;
- Matplotlib: per la visualizzazione dei grafici;
- Pgmpy: usato per costruire la rete bayesiana.

Ognuna delle quali può essere installata dal prompt dei comandi di windows usando il comando:
"pip install -U nome_libreria"

Manuale Utente

E' possibile scaricare il progetto dal seguente link https://github.com/FPizzimenti/ICON_Cancer

Dopo averlo scaricato si dovrà eseguire il file "ProgettoICON" su un IDE.

All’avvio del file apparirà un messaggio di benvenuto con questa frase e la visualizzazione dell’intero Dataset:

```
Benvenuto nel nostro sistema per predire se, presi dei soggetti, essi sono affetti o meno dal Cancro al Seno


     diagnosis  radius_mean  texture_mean  perimeter_mean  ...  concavity_worst  concave points_worst  symmetry_worst  fractal_dimension_worst
0            1        17.99         10.38          122.80  ...           0.7119                0.2654          0.4601                  0.11890
1            1        20.57         17.77          132.90  ...           0.2416                0.1860          0.2750                  0.08902
2            1        19.69         21.25          130.00  ...           0.4504                0.2430          0.3613                  0.08758
3            1        11.42         20.38           77.58  ...           0.6869                0.2575          0.6638                  0.17300
4            1        20.29         14.34          135.10  ...           0.4000                0.1625          0.2364                  0.07678
..         ...          ...           ...             ...  ...              ...                   ...             ...                      ...
564          1        21.56         22.39          142.00  ...           0.4107                0.2216          0.2060                  0.07115
565          1        20.13         28.25          131.20  ...           0.3215                0.1628          0.2572                  0.06637
566          1        16.60         28.08          108.30  ...           0.3403                0.1418          0.2218                  0.07820
567          1        20.60         29.33          140.10  ...           0.9387                0.2650          0.4087                  0.12400
568          0         7.76         24.54           47.92  ...           0.0000                0.0000          0.2871                  0.07039

[569 rows x 31 columns]
```

Scelte progettuali

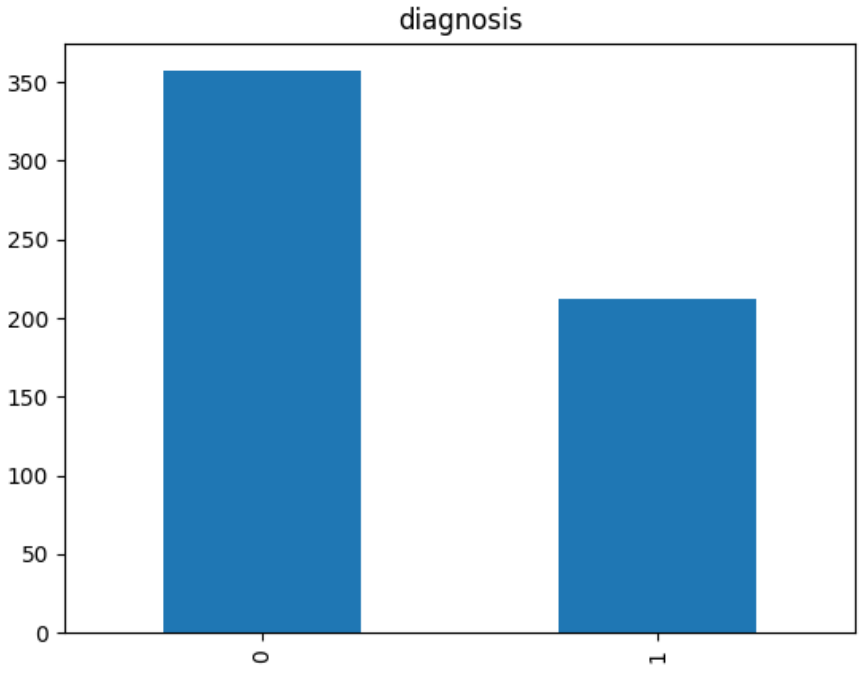
E’ stato utilizzato il Dataset dal seguente link: [Breast Cancer Prediction | Kaggle] per addestrarel il nostro sistema.

Prima di passare direttamente all’apprendimento supervisionato abbiamo voluto verificare che il Dataset fosse ben bilanciato cosicché da riuscire ad avere ottimi risultati durante l’apprendimento che vedremo più avanti.

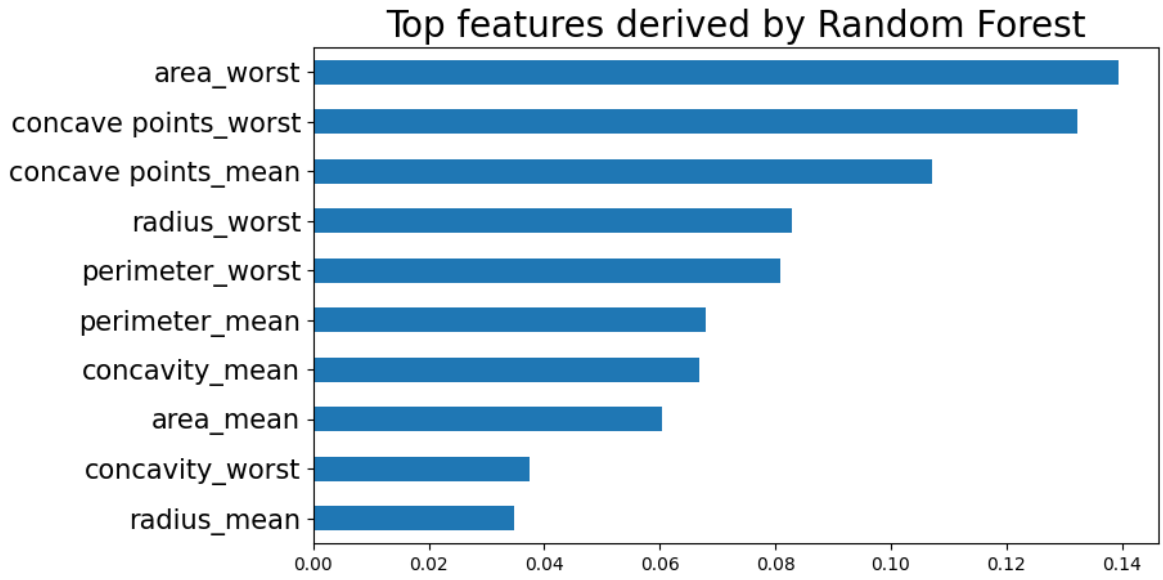
I risultati ottenuti sono i seguenti:

```
Pazienti non malati di cancro al seno: 357 (% 62.74)
Pazienti malati di cancro al seno: 212 (% 37.26)
```

Successivamente vengono visualizzati i seguenti risultati in un grafico:



Dopodiché viene visualizzato una ulteriore tabelle contenente le features più importanti derivanti dal Random Forest:



Apprendimento Supervisionato

Per questo tipo di apprendimento abbiamo usato varie metriche per poi identificare quale fosse quella più adatta al nostro dataset. Le metriche da noi utilizzate sono:

KNN:

```
KNN  accuracy:  0.9333333333333333
KNN  precision:  0.8636363636363636
KNN  recall:  0.95
KNN  F1:  0.9047619047619048
```

SVM:

```
SVM  accuracy:  0.95
SVM  precision:  0.8636363636363636
SVM  recall:  1.0
SVM  F1:  0.9268292682926829
```

Decision Tree:

```
Decision Tree  accuracy:  0.9416666666666667
Decision Tree  precision:  0.9545454545454546
Decision Tree  recall:  0.8936170212765957
Decision Tree  F1:  0.9230769230769231
```

Random Forest:

```
Random Forest  accuracy:  0.9583333333333334
Random Forest  precision:  0.9090909090909091
Random Forest  recall:  0.975609756097561
Random Forest  F1:  0.9411764705882352
```

Dopo aver effettuato il calcolo di queste metriche, abbiamo eseguito il K-Fold cross validation per verificare quale di esse fosse la più attendibile; essendo il dataset scelto non abbastanza equilibrato, abbiamo usato lo Stratified K-Fold con 5 Fold. I risultati ottenuti sono i seguenti:

```
KNN K-fold: 0.928 (0.022)
SVM K-fold: 0.946 (0.019)
Decision Tree K-fold: 0.916 (0.021)
Random Forest K-fold: 0.963 (0.021)
```

In base a ciò abbiamo riscontrato che il classificatore più attendibile è il Random Forest.

Apprendimento non Supervisionato

Per lo sviluppo dell'apprendimento non supervisionato abbiamo scelto di implementare una rete bayesiana utilizzando come metodo di scoring il K2score e come stimatore il MaximumLikeliHoodEstimator. Per verificare che la rete fosse effettivamente creata abbiamo fatto visualizzare i nodi e gli archi della rete:

```
Nodi della rete:
 ['diagnosis', 'perimeter_se', 'texture_mean', 'radius_mean', 'perimeter_mean', 'radius_worst', 'area_mean', 'texture_worst', 'radius_s
e', 'area_se', 'perimeter_worst', 'area_worst', 'compactness_worst', 'concavity_worst']

Archi della rete:
 [('diagnosis', 'perimeter_se'), ('diagnosis', 'texture_mean'), ('perimeter_se', 'area_se'), ('texture_mean', 'texture_worst'), ('radiu
s_mean', 'perimeter_mean'), ('radius_mean', 'radius_worst'), ('radius_mean', 'area_mean'), ('radius_worst', 'perimeter_worst'), ('radiu
s_worst', 'area_worst'), ('radius_worst', 'diagnosis'), ('radius_se', 'perimeter_se'), ('radius_se', 'radius_mean'), ('radius_se', 'are
a_mean'), ('compactness_worst', 'concavity_worst'), ('compactness_worst', 'diagnosis'), ('compactness_worst', 'texture_worst'), ('compa
ctness_worst', 'area_mean'), ('compactness_worst', 'texture_mean'), ('concavity_worst', 'texture_mean')]
```

Infine per effettuare la predizione attraverso la rete bayesiana abbiamo passato due query, una contenente i dati di una persona con cancro al seno ed una senza di esso (differenti dai dati già presenti nel dataset).

Query:

- Soggetto potenzialmente senza il cancro al seno:

```
# Soggetto potenzialmente senza cancro al seno
notCancer= data.query(variables = ['diagnosis'],
                      evidence = { 'radius_mean':14,'texture_mean':16,'perimeter_mean':91,'area_mean':334,
                                   'radius_se':0,'perimeter_se':1,'area_se':2,'radius_worst':16,'texture_worst':25,
                                   'perimeter_worst':106, 'area_worst':520,'compactness_worst':0,'concavity_worst':0 })
```

- Soggetto potenzialmente con il cancro al seno:

```
# Soggetto potenzialmente con cancro al seno
cancer = data.query(variables = ['diagnosis'],
                    evidence={'radius_mean': 15,'texture_mean':23,'perimeter_mean':80,'area_mean':301,
                    'radius_se':0,'perimeter_se':4,'area_se':30, 'radius_worst':15,'texture_worst':32,
                    'perimeter_worst':170,'area_worst':406, 'compactness_worst':0,'concavity_worst':0})
```

I risultati ottenuti dalle seguenti query sono (**0**= Senza Cancro al Seno; **1**= Con il Cancro al Seno) :

- Probabilità soggetto potenzialmente senza il cancro al seno:

```
Probabilità per un soggetto potenzialmente senza cancro al seno:
+--------------+------------------+
| diagnosis    |   phi(diagnosis) |
+==============+==================+
| diagnosis(0) |           0.9639 |
+--------------+------------------+
| diagnosis(1) |           0.0361 |
+--------------+------------------+
```

- Probabilità soggetto potenzialmente con il cancro al seno:

```
Probabilità per un soggetto potenzialmente con cancro al seno:
+--------------+------------------+
| diagnosis    |   phi(diagnosis) |
+==============+==================+
| diagnosis(0) |           0.0441 |
+--------------+------------------+
| diagnosis(1) |           0.9559 |
+--------------+------------------+
```

Per verificare quali features all'interno del dataset fossero più rilevanti nella rete bayesiana abbiamo effettuato dei test su i soggetti presi in considerazione e abbiamo riscontrato che cambiando solo i valori di due features le percentuali cambiano drasticamente.

Query:

- Test su un soggetto potenzialmente senza cancro al seno:

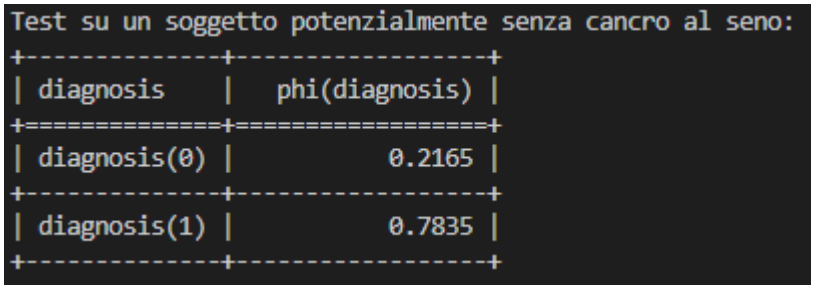
```
# Test su Soggetto potenzialmente senza cancro al seno
TestnotCancer= data.query(variables = ['diagnosis'],
                          evidence = { 'radius_mean':14,'texture_mean':16,'perimeter_mean':91,'area_mean':334,
                                       'radius_se':0,'perimeter_se':5,'area_se':20,'radius_worst':16,'texture_worst':25,
                                       'perimeter_worst':106, 'area_worst':520,'compactness_worst':0,'concavity_worst':0 })
```

- Test su un soggetto potenzialmente con il cancro al seno:

```
# Test su Soggetto potenzialmente con cancro al seno
Testcancer = data.query(variables = ['diagnosis'],
                        evidence={'radius_mean': 15,'texture_mean':23,'perimeter_mean':80,'area_mean':301,
                        'radius_se':0,'perimeter_se':1,'area_se':10, 'radius_worst':15,'texture_worst':32,
                        'perimeter_worst':170,'area_worst':406, 'compactness_worst':0,'concavity_worst':0})
```

I risultati ottenuti dalle seguenti query sono (**0**= Senza Cancro al Seno; **1**= Con il Cancro al Seno) :

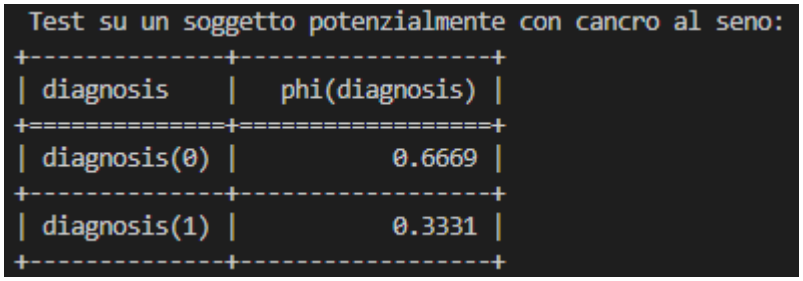
- Probabilità test soggetto potenzialmente senza il cancro al seno:



Test su un soggetto potenzialmente senza cancro al seno:

diagnosis	phi(diagnosis)
diagnosis(0)	0.2165
diagnosis(1)	0.7835

- Probabilità test soggetto potenzialmente con il cancro al seno:



Test su un soggetto potenzialmente con cancro al seno:

diagnosis	phi(diagnosis)
diagnosis(0)	0.6669
diagnosis(1)	0.3331

Da questi test è emerso che le features più rilevanti sono : perimeter_se e area_se ovvero il perimetro e l'area della zona dove è probabile che possa esserci il cancro.