**CANCRO AL SENO: ONTOLOGIA E PREDIZIONE TRAMITE CLASSIFICATORE**

*Giacovazzo Christian – 716696 – c.giacovazzo1@studenti.uniba.it*

*Progetto per “Ingegneria della conoscenza” 2021/2022*

**INDICE**

1. INTRODUZIONE

2. ONTOLOGIA

3. APPRENDIMENTO SUPERVISIONATO

4. APPRENDIMENTO NON SUPERVISIONATO

5. CONCLUSIONI

1. **INTRODUZIONE**

Il progetto nasce con l’intento di predire se un paziente possa ripresentare in futuro un cancro alla vescica. A tal proposito sono state adottate una serie di tecniche sia di apprendimento supervisionato sia non supervisionato. Inoltre, è stata modellata un’ontologia di dominio che offre una rappresentazione formale e concettualizzata della realtà presa in esame, affinché possa venire interrogata. L’ontologia è completa di individui d’esempio in modo tale da testarne la funzionalità.

Il dataset utilizzato contiene le seguenti features:

1. **Class** **[target]**: no-recurrence-events, recurrence-events
2. **age**: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
3. **menopause**: lt40, ge40, premeno
4. **tumor-size**: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
5. **inv-nodes**: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
6. **node-caps**: yes, no
7. **deg-malig**: 1, 2, 3
8. **breast**: left, right
9. **breast-quad**: left-up, left-low, right-up, right-low, central
10. **irradiat**: yes, no

Per la modellazione dell’ontologia è stato utilizzato Protégé, mentre il codice è stato scritto interamente in Python. Fonti del materiale utilizzato:

Dataset: <https://archive-beta.ics.uci.edu/ml/datasets/breast+cancer>

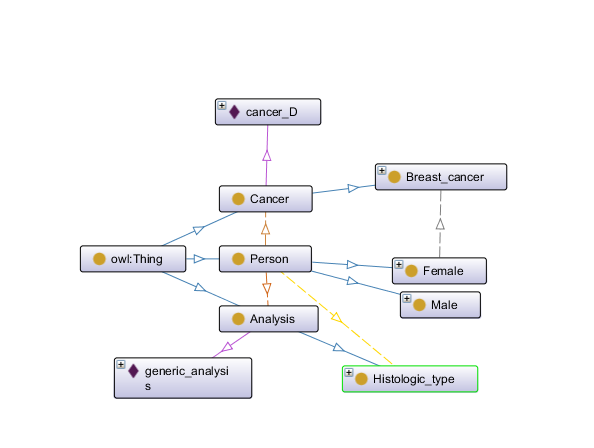
Protégé: <https://protege.stanford.edu/>

Python : <https://www.anaconda.com/products/individual>

1. **ONTOLOGIA**

Un’ ontologia è la specificazione dei significati dei simboli in un sistema informatico. La specifica formale è importante per l’interoperabilità semantica, ovvero l’abilità di basi di conoscenza differenti di operare insieme ad un livello semantico tale che i significati dei simboli sono rispettati. L’ontologia viene descritta come “una specificazione di una concettualizzazione”. Una rappresentazione formale di un insieme di conoscenze è una concettualizzazione, ossia un insieme di oggetti, concetti e relazioni fra di essi che esistono in una particolare area d’interesse. Nella modellizzazione dell’ ontologia è stato usato il programma Protégé. Sono state aggiunte delle proprietà di oggetto e di dati con un dominio di appartenenza e un loro range, mantenendo una certa coerenza con il dataset di riferimento utilizzato. È possibile interrogare l’ontologia e ottenere informazioni usando il tab di Protègé “DL query”. Si è usato come reasoner HermiT, di default in Protégé.

Di seguito la modellazione della ontologia.



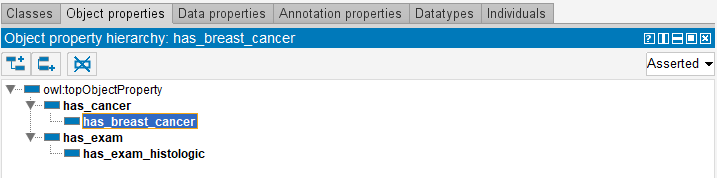
Queste entità sono in relazione tra di loro grazie a delle proprietà sia di oggetto che di data. Una object property permette di mettere in relazione due individui, siano essi di classi distinte o della stessa classe. Una data property permette di mettere in relazione un individuo con un valore di tipo primitivo. Una object property permette di mettere in relazione due individui, siano essi di classi distinte o della stessa classe. Una data property permette di mettere in relazione un individuo con un valore di tipo primitivo.

Immagine che contiene testo

Descrizione generata automaticamente

Inoltre, per alcune entità si sono create delle istanze. Alcune ad esempio per individuare una tipologia di esame come quella istologica, altre per individuare istanze di persone a cui attribuire un cancro e le relazioni con esso.

Immagine che contiene testo

Descrizione generata automaticamente

Immagine che contiene testo

Descrizione generata automaticamenteSuccessivamente sono state formulate delle query per interrogare l’ontologia. Si è partiti con query semplici come cercare persone affette da un tumore qualsiasi o specifico al seno e a quali esami istologici si sono sottoposte, fino a query più complesse con la combinazione di più attributi inerenti a ciascuna persona.

Immagine che contiene testo

Descrizione generata automaticamenteImmagine che contiene testo

Descrizione generata automaticamente

Immagine che contiene testo

Descrizione generata automaticamente

1. **APPRENDIMENTO SUPERVISIONATO**

Con l’Apprendimento Supervisionato si cerca di costruire un modello partendo da dei dati di addestramento etichettati, con i quali si fanno previsioni su dati non disponibili o futuri.

Per ogni algoritmo di apprendimento supervisionato sono stati prodotti i seguenti grafici:

• ROC Curve

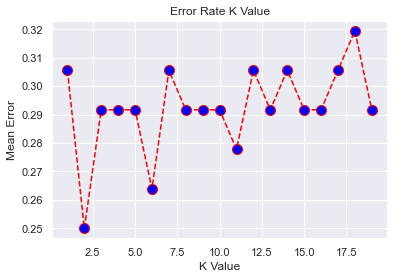
• Precision-Recall Curve

• Matrice di Confusione

Per la maggior parte degli algoritmi è stata usata la tecnica della cross-validation per rilevare possibili problemi di overfitting. A tal proposito si riportano i valori del punteggio medio (cross-val-score), della varianza, deviazione standard su cinque iterazioni.

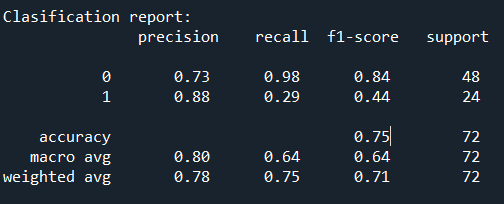
**3.1 K-nearest-neighbour**

Il KNN è un algoritmo di apprendimento supervisionato che consiste nell’individuare i k esempi più vicini a quello che si intende classificare, a quest’ultimo viene quindi attribuita la categoria “più ricorrente” tra i k esempi più vicini.



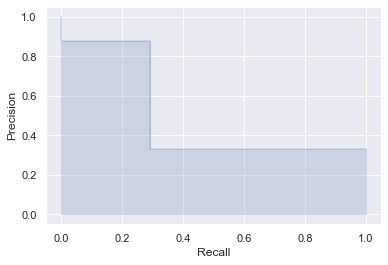
Il grafico proposto fornisce un suggerimento sulla scelta del numero di vicini per minimizzare l’errore medio. Si evince che con numero di vicini uguale a 2 l’errore medio minimo commesso è di 0.25.

Dalla classificazione dell’algoritmo sono stati prodotti i seguenti classification report.

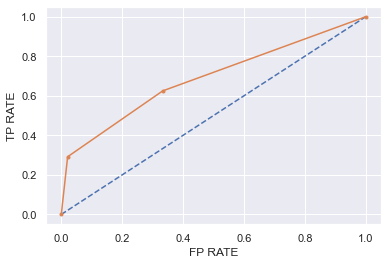
Si osservi che l’accuratezza dell’algoritmo è 0.75

Inoltre l’average precision è di 0.49

Confusion-Matrix: Precision e Recall:

Immagine che contiene piazza

Descrizione generata automaticamente



La curva ROC (Receiver Operating Characteristics ) dell'AUC ( Area Under The Curve ) è una delle metriche di valutazione più importanti per il controllo delle prestazioni di qualsiasi modello di classificazione. Il ROC è una curva di probabilità e l'AUC rappresenta la misura della separabilità. Indica quanti modelli è in grado di distinguere tra le classi. Maggiore è l'AUC, migliore è il modello nel predire 0 come 0 e 1 come 1. Nel nostro caso l’AUC è pari a 0.688.

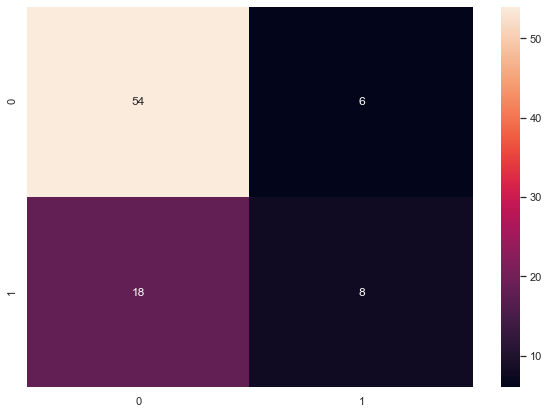
Per quanto riguarda la cross validation sul classificatore del KNN i dati ottenuti sono:

Immagine che contiene testo

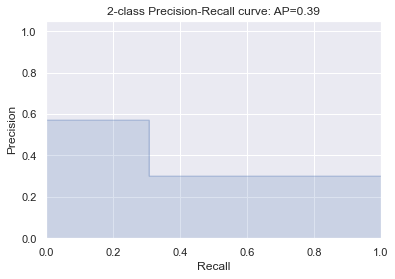
Descrizione generata automaticamente

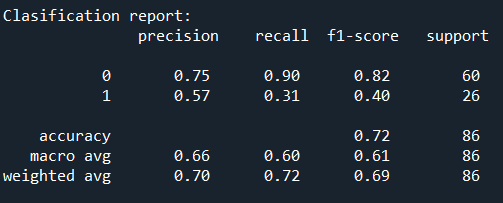
**3.2 Random Forest**

Il random forest è un modello composito costituito da molti alberi di decisione, ognuno dei quali fornisce una predizione. Esse vengono poi, combinate allo scopo di ottenere una previsione complessiva della foresta per un dato esempio. La predizione di ciascun albero può essere ottenuta o, attraverso la media delle predizioni di un albero per ogni esempio o, attraverso un meccanismo di votazione in cui tutti gli alberi votano la propria classificazione più probabile e l’esempio col maggior numero di voti sarà scelto come predizione finale.



Average precision = 0.39





Accuratezza = 0.72

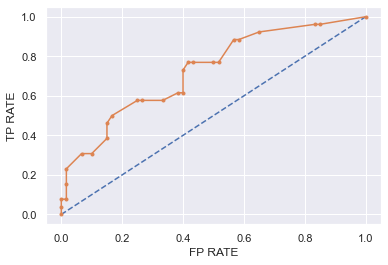
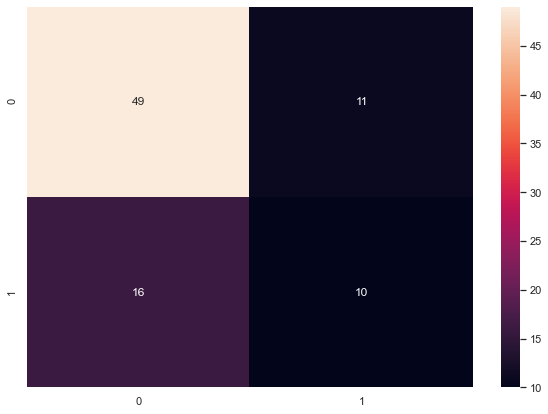
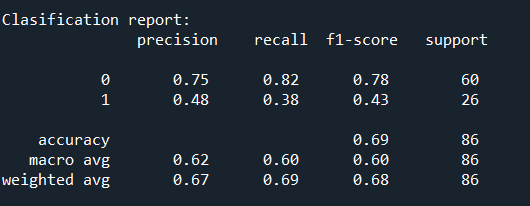
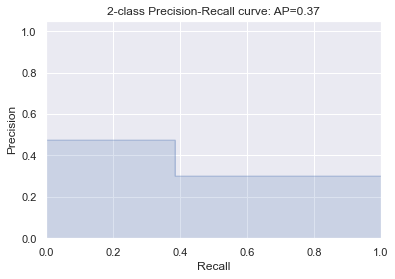
L’algoritmo Random Forest, secondo la curva ROC, è in grado di differenziare meglio le due classi rispetto al Knn; il valore AUC infatti è pari a 0.725.

Immagine che contiene testo

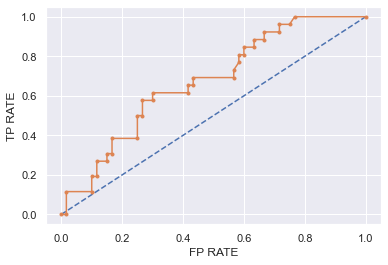
Descrizione generata automaticamenteInvece, per quanto riguarda la cross-validation i dati sono:

**3.3 Multinomial Naive Bayes**

I classificatori “Naive Bayes” sono una famiglia di semplici classificatori probabilistici basati sull’applicazione del teorema di Bayes con una forte assunzione di indipedenza tra le feature. Con un modello di eventi multinomiali, gli esempi (vettori di feature) rappresentano le frequenze con cui certi eventi sono stati generati da una distribuzione polinomiale (p\_1, …, p\_n) dove p\_i è la probabilità che l’evento i si verifichi.



Il Multinomial Naive Bayes propone un’accuratezza di 0.69 e average precision di 0.37.

Come anche osservato dal grafico a sinistra, tra gli algoritmi trattati fin ora, questo è quello che propone una precision (al variare della recall) leggermente più bassa.

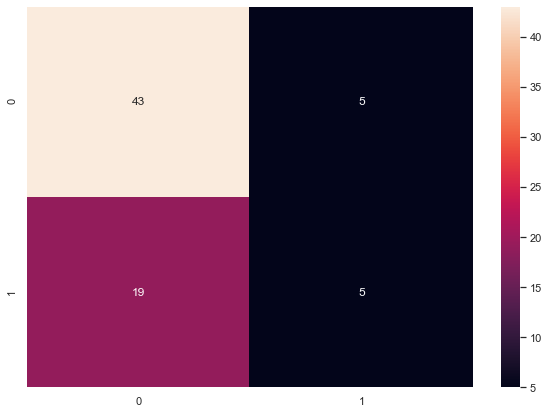
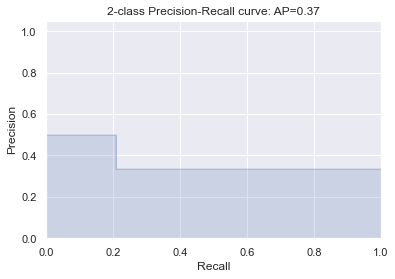
Per quanto riguarda l’AUC, esso è 0.673.

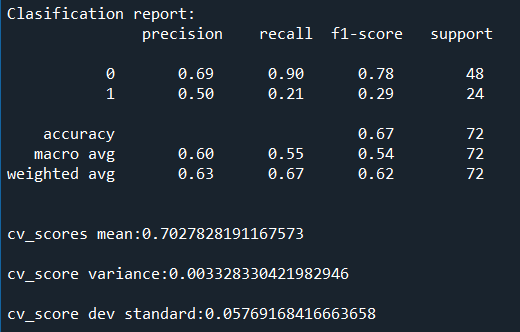
Immagine che contiene testo

Descrizione generata automaticamentePer quanto riguarda la cross-validation i dati sono:

**3.4 Support-Vector Machines**

Un modello SVM è una rappresentazione degli esempi come punti nello spazio, mappati in modo tale che gli esempi appartenenti alle due diverse categorie siano chiaramente separati da uno spazio il più possibile ampio. I nuovi esempi sono quindi mappati nello stesso spazio e la predizione della categoria alla quale appartengono viene fatta sulla base del lato nel quale ricade.



L’algoritmo riporta una accuratezza di 0.67 e una average precision di 0.37

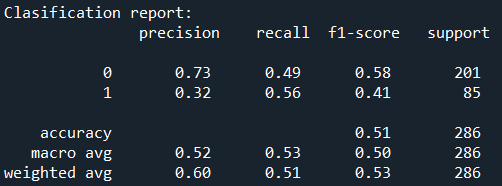
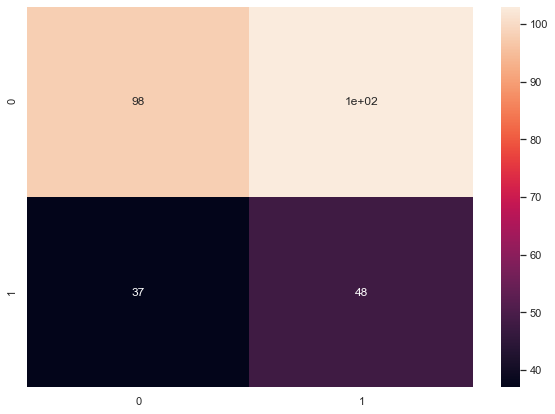
I dati sulla cross validation sono a sinistra.

Questo algoritmo sembra avere il comportamento peggiore in questo dataset, in quanto ha accuratezza e average precision molto bassi rispetto agli altri esaminati.

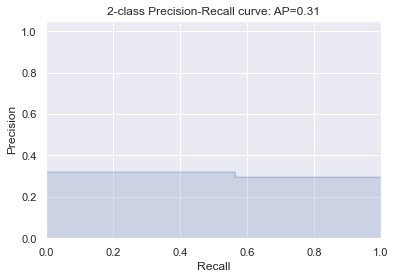
**4. APPRENDIMENTO NON SUPERVISIONATO**

L'apprendimento non supervisionato è una tecnica di apprendimento automatico che consiste nel fornire al sistema una serie di input che esso riclassificherà ed organizzerà sulla base di caratteristiche comuni per cercare di effettuare ragionamenti e previsioni sugli input successivi. Al contrario dell'apprendimento supervisionato, durante l'apprendimento vengono forniti all’algoritmo solo esempi non etichettati, in quanto le classi non sono note a priori ma devono essere apprese automaticamente.

**4.1 K-Means**

L'algoritmo K-means è un algoritmo di hard-clustering partizionale che permette di suddividere un insieme di oggetti in K gruppi (nel nostro caso K=2) sulla base dei loro attributi. Si assume che gli attributi degli oggetti possano essere rappresentati come vettori, e che quindi formano uno spazio vettoriale. Ogni cluster viene identificato mediante un centroide. L'algoritmo segue una procedura iterativa. Inizialmente crea K partizioni e assegna ad ogni partizione i punti d'ingresso (casualmente o usando alcune informazioni euristiche). Successivamente calcola il centroide di ogni gruppo. Costruisce quindi una nuova partizione associando ogni punto d'ingresso al cluster il cui centroide è più vicino ad esso. Dopodichè vengono ricalcolati i centroidi per i nuovi cluster. Si itera il processo finché l'algoritmo non converge. 

L’accuratezza è 0.51, la più bassa di tutti gli algoritmi usati. È evidente inoltre dalla matrice di confusione che l’algoritmo trova un elevata quantità di falsi positivi.

Come si vede dal grafico, anche l’average precision è la più bassa rilevata: 0.31. Tuttavia la Precision si mantiene quasi costante rispetto al Recall.

Il risultato è tutto sommato prevedibile in quanto in questo algoritmo non si è usata la cross validation.

1. **CONCLUSIONI**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuratezza** | **Varianza** | **Deviazione standard** | **F1-score** | **Average precision** | **AUC** |
| **Knn** | 0.75 | 0.001 | 0.034 | 0.44 | 0.49 | 0.688 |
| **Random forest** | 0.72 | 0.005 | 0.069 | 0.40 | 0.39 | 0.725 |
| **Multinomial naive Bayes** | 0.69 | 0.016 | 0.128 | 0.43 | 0.37 | 0.673 |
| **SVM** | 0.67 | 0.003 | 0.058 | 0.29 | 0.37 | // |
| **K-means** | 0.51 | // | // | 0.41 | 0.31 | // |

Dalla tabella riassuntiva si evince come, per il dataset utilizzato, l’accuratezza della classificazione migliore si ha nel Knn. I risultati poco soddisfacenti per gli altri algoritmi sono riconducibili allo sbilanciamento del dataset. Una possibile soluzione potrebbe essere quella di usare tecniche di ricampionamento, per esempio l’algoritmo SMOTE.