# Progetto Ingegneria della conoscenza

## Intelligenza Artificiale nella lotta al cancro al seno.

Studente: Urselli Gabriele
(776148)
Email: g.urselli1@studenti.uniba.it

# INDICE

## Sommario

# 1. INTRODUZIONE AL PROGETTO

L'obiettivo di questo progetto è capire se il tumore al seno di una paziente possa ripresentarsi nel tempo. Per farlo è stato necessaria l'Intelligenza Artificiale in due modi: prima addestrando degli algoritmi a riconoscere i segnali di rischio (usando tecniche supervisionate e non), e poi creando un'ontologia necessaria a organizzare le informazioni in modo che siano leggibili anche da altri sistemi e facilmente consultabili.

I dati analizzati riguardano informazioni cliniche specifiche, come l'età della paziente, la dimensione del tumore e i trattamenti effettuati, con l'obiettivo finale di prevedere se ci sarà o meno una recidiva.

Il dataset utilizzato è consultabile sul link:
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

e prevede la seguente Tavola delle Variabili

**Variables Table**

| Variable Name | Role | Type | Demographic | Description | Units | Missing Values |
|---|---|---|---|---|---|---|
| Class | Target | Binary | | no-recurrence-events, recurrence-events | | no |
| age | Feature | Categorical | Age | 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99 | years | no |
| menopause | Feature | Categorical | | lt40, ge40, premeno | | no |
| tumor-size | Feature | Categorical | | 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 | | no |
| inv-nodes | Feature | Categorical | | 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 | | no |
| node-caps | Feature | Binary | | yes, no | | yes |
| deg-malig | Feature | Integer | | 1, 2, 3 | | no |
| breast | Feature | Binary | | left, right | | no |
| breast-quad | Feature | Categorical | | left-up, left-low, right-up, right-low, central | | yes |
| irradiat | Feature | Binary | | yes, no | | no |

# 2. APPRENDIMENTO SUPERVISIONATO

L'**Apprendimento Supervisionato** mira alla costruzione di modelli predittivi basati su dataset etichettati. La disponibilità di segnali di output noti permette all'algoritmo di apprendere le relazioni tra feature e classi, rendendo possibile la classificazione di nuovi campioni. Data la natura discreta del target, sono state adottate tecniche di classificazione statistica.

### *Bilanciamento del Dataset*

L'analisi preliminare ha evidenziato una distribuzione sbilanciata tra le classi (rapporto circa 2:1). Per ovviare a tale disparità e prevenire bias verso la classe maggioritaria, è stata applicata la tecnica **SMOTE**. L'algoritmo ha permesso di riequilibrare il dataset attraverso la generazione sintetica di campioni per la classe minoritaria, producendo effetti positivi sulla capacità discriminante dei modelli, in particolare nel caso dell'algoritmo **KNN**.

### *Validazione e Metriche di Performance*

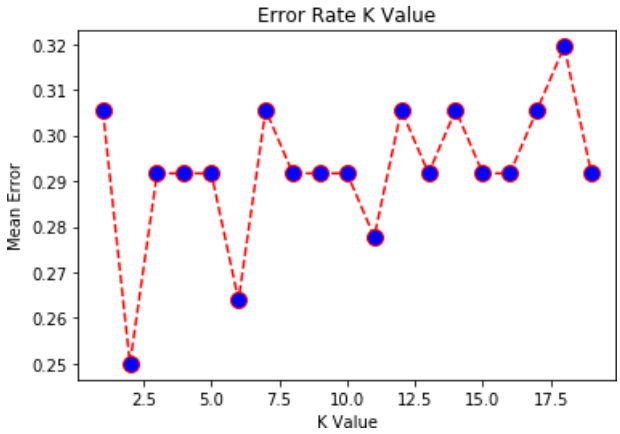
Al fine di garantire la capacità di generalizzazione e prevenire il **sovra-adattamento** (overfitting), è stata utilizzata la **cross-validation** (10-fold). Per ogni algoritmo sono stati estratti i valori di accuratezza media (*cross-val-score*), varianza e deviazione standard.

La valutazione delle prestazioni è stata supportata dalla produzione della seguente reportistica grafica:

- **ROC Curve** e **Precision-Recall Curve**;
- **Matrice di Confusione** per il dettaglio degli errori;
- **Bar Chart** relativi alla variabilità dei risultati (varianza e deviazione standard)

## 2.1 K-NEAREST-NEIGHBOUR

L'algoritmo **K-Nearest Neighbour** è un classificatore supervisionato che assegna un campione alla categoria prevalente tra i suoi k vicini più prossimi, basandosi su una misura di distanza nello spazio delle feature.




Il grafico dell'errore medio fornisce l'indicazione per la selezione del parametro k ottimale. Dall'analisi si evince che il valore k=2 minimizza l'errore medio, attestandolo a 0.25.

Dalla fase di test sono stati generati i seguenti *classification report*, confrontando le prestazioni prima e dopo l'applicazione della tecnica di bilanciamento **SMOTE**:

PRE-SMOTE

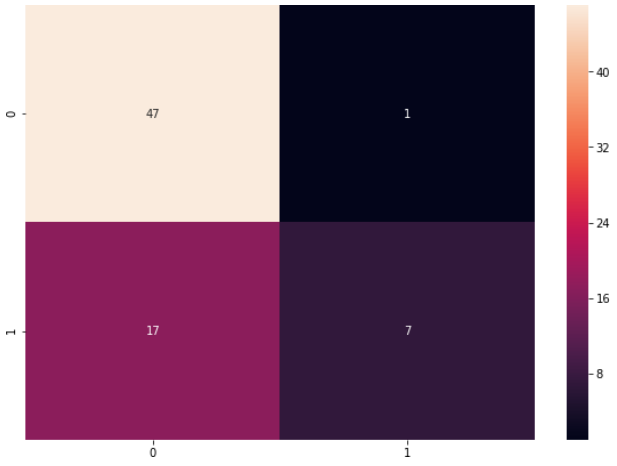
Clasification report:

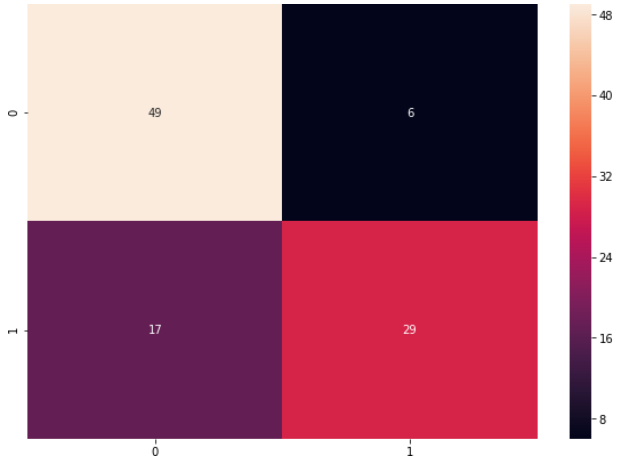
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.73 | 0.98 | 0.84 | 48 |
| 1 | 0.88 | 0.29 | 0.44 | 24 |
| micro avg | 0.75 | 0.75 | 0.75 | 72 |
| macro avg | 0.80 | 0.64 | 0.64 | 72 |
| weighted avg | 0.78 | 0.75 | 0.71 | 72 |

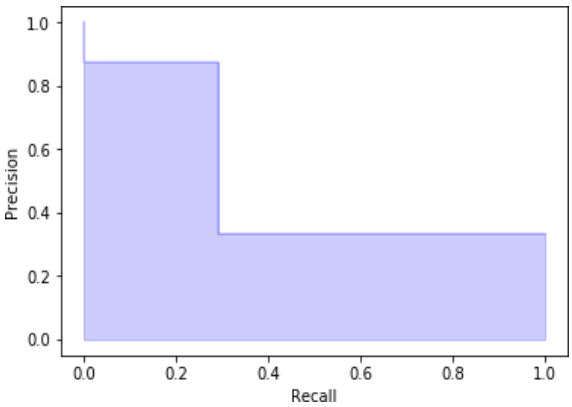
POST-SMOTE

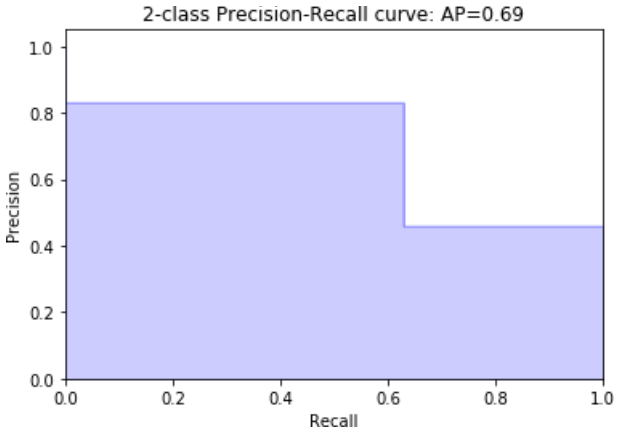
Clasification report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.74 | 0.89 | 0.81 | 55 |
| 1 | 0.83 | 0.63 | 0.72 | 46 |
| micro avg | 0.77 | 0.77 | 0.77 | 101 |
| macro avg | 0.79 | 0.76 | 0.76 | 101 |
| weighted avg | 0.78 | 0.77 | 0.77 | 101 |

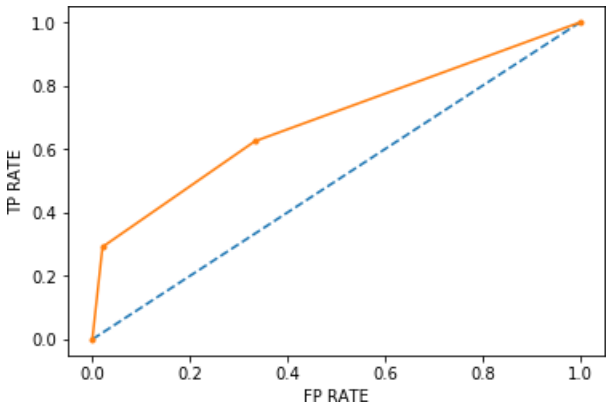

47
1
17
7
0
1
0
1
40
32
24
16
8

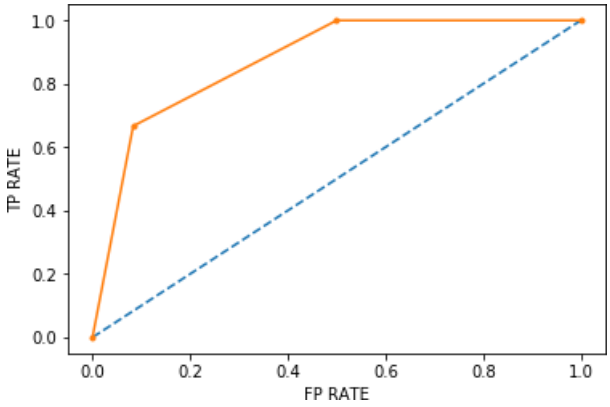

49
6
17
29
0
1
0
1
48
40
32
24
16
8



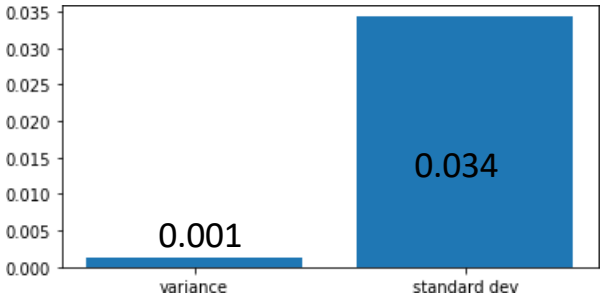



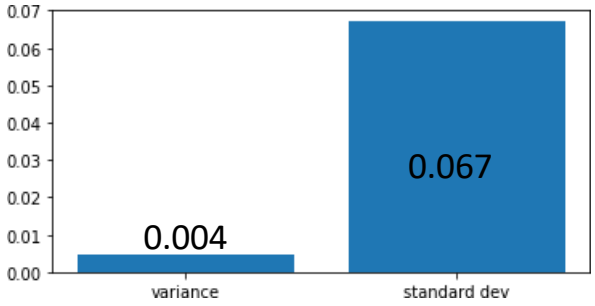

L'analisi comparativa delle curve Precision-Recall evidenzia un netto miglioramento della precisione in corrispondenza di elevati valori di recall a seguito dell'ottimizzazione del dataset. Tale incremento si riflette sull'Average Precision (AP) che evolve da un valore iniziale di 0.491 a un incremento di +0.20 post-ottimizzazione, con un'accuratezza complessiva del 0.75.




La **curva ROC** (*Receiver Operating Characteristics*) e il relativo valore **AUC** (*Area Under The Curve*) sono stati utilizzati per valutare la capacità discriminante del modello. L'AUC misura il grado di separabilità tra le classi: un valore maggiore indica una superiore capacità del classificatore nel distinguere correttamente tra pazienti con e senza recidiva. Nel caso in esame, l'AUC iniziale di 0.688 ha registrato un incremento di +0.187 a seguito dell'integrazione dell'algoritmo SMOTE.




Infine, l'applicazione della **cross-validation** sul classificatore KNN ha prodotto i seguenti parametri statistici, confermando la solidità del modello dopo il bilanciamento:

Configurazione Pre-SMOTE:

- CV Scores Mean: 0.717
- CV Score Variance: 0.0011
- CV Score Dev. Standard: 0.034

Configurazione Post-SMOTE:

- CV Scores Mean: 0.774
- CV Score Variance: 0.0045
- CV Score Dev. Standard: 0.067

# 2.2 RANDOM FOREST

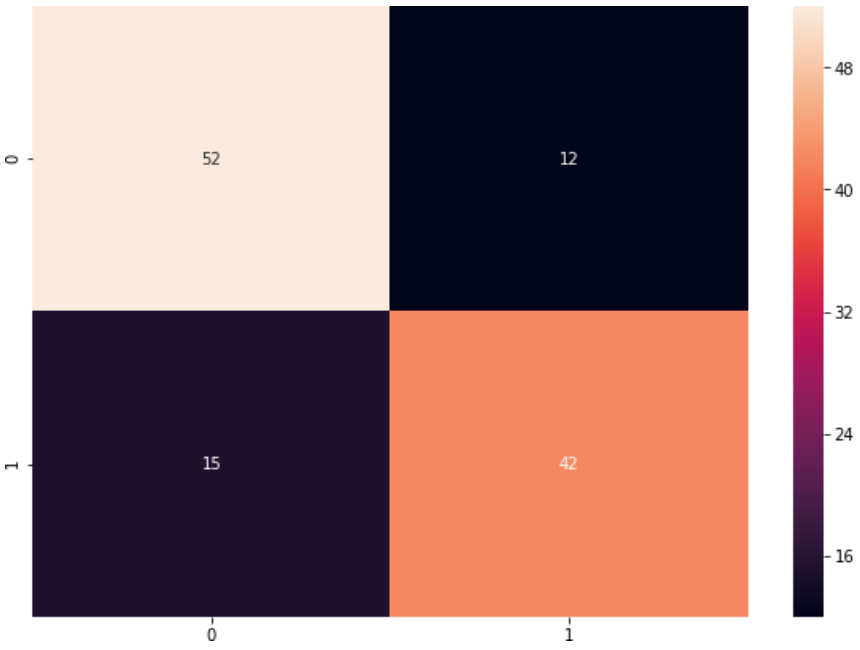
Il modello **Random Forest** si configura come un'architettura d'insieme composta da una moltitudine di **alberi di decisione**, dove ogni unità genera una previsione specifica. L'obiettivo centrale risiede nella combinazione di tali risultati individuali per formulare una sintesi complessiva riferita a ogni singolo esempio analizzato.

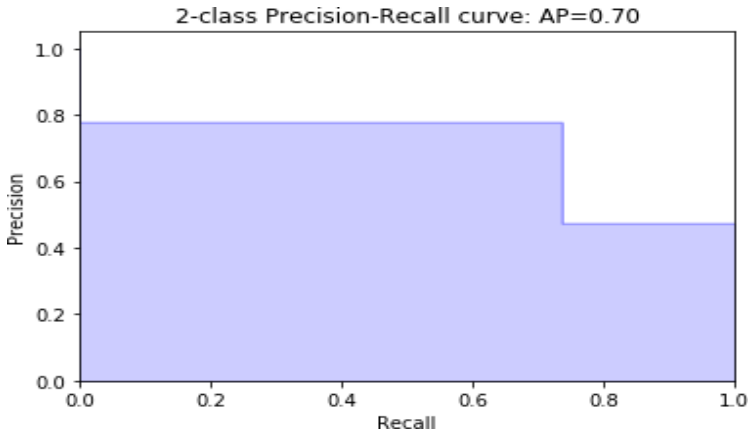
Per arrivare alla risposta definitiva, la foresta "interroga" tutti i suoi alberi e usa uno di questi due metodi:

- **La Media (per i numeri):** Se bisogna indovinare un valore (es. il prezzo di una casa), si sommano i risultati di tutti gli alberi e si fa la media.
- **La Maggioranza (per le categorie):** Se bisogna scegliere tra diverse opzioni (es. "Sì" o "No"), ogni albero vota la sua preferenza. La scelta che riceve più voti vince, proprio come in un'elezione.

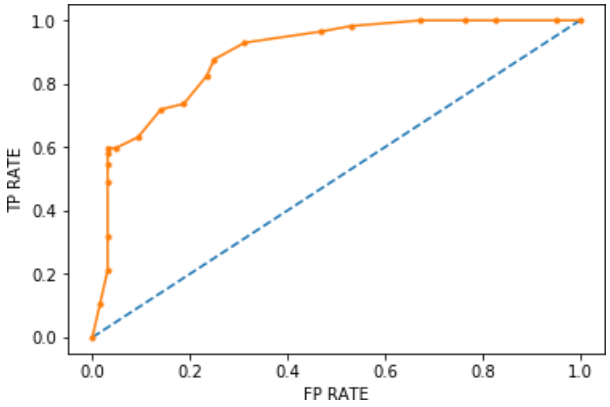
Clasification report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.81 | 0.79 | 64 |
| 1 | 0.78 | 0.74 | 0.76 | 57 |
| micro avg | 0.78 | 0.78 | 0.78 | 121 |
| macro avg | 0.78 | 0.77 | 0.78 | 121 |
| weighted avg | 0.78 | 0.78 | 0.78 | 121 |

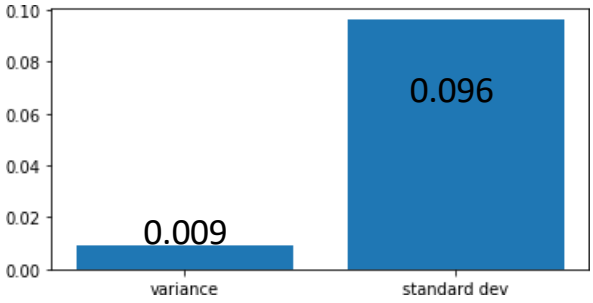

52
12
15
42
0
1
0
1
48
40
32
24
16




L' Average-Precision (AP) è pari a 0.697, l'accuratezza è uguale a 0.777.




L'algoritmo Random Forest, secondo la curva ROC, è in grado di differenziare abbastanza bene le due classi, infatti il valore AUC è 0.89.




Per quanto riguarda la cross validation (con cv=5) sul classificatore i dati ottenuti sono:

cv_scores mean: 0.7271341463414634
cv_score variance: 0.009199620761451524
cv_score dev standard: 0.09591465352828797

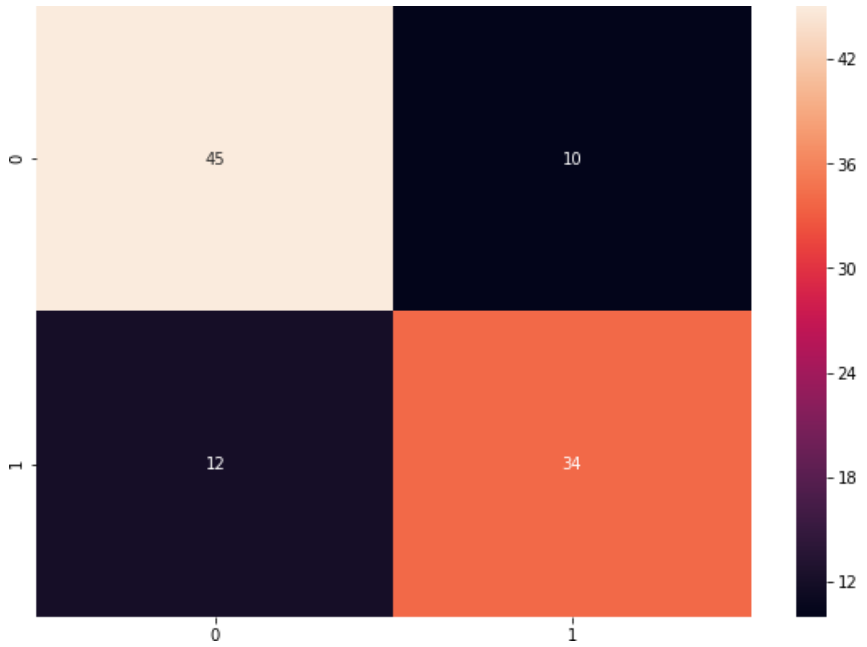
# 2.3 SUPPORT-VECTOR MACHINES

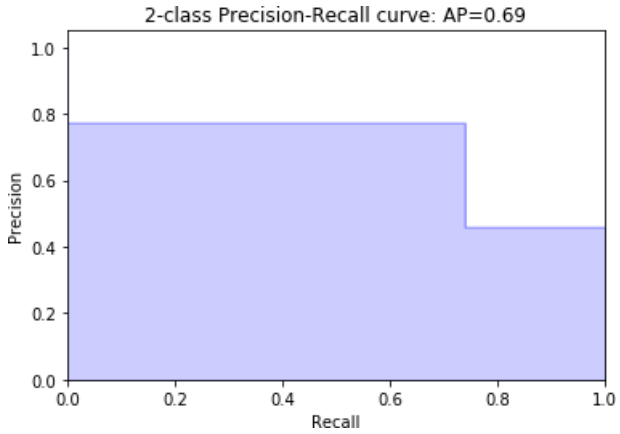
Il modello **SVM (Support Vector Machine)** proietta i dati come punti all'interno di un iperpiano. L'obiettivo principale è tracciare un confine di separazione tra le diverse categorie, cercando di creare un corridoio (chiamato "margine") il più ampio possibile tra i gruppi.

Quando viene inserito un nuovo dato, questo viene posizionato nello stesso spazio. L'appartenenza a una categoria viene stabilita semplicemente osservando in quale zona, rispetto al confine tracciato, va a cadere il punto.

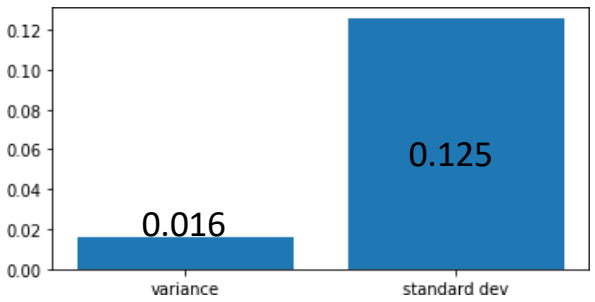
Clasification report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.79 | 0.82 | 0.80 | 55 |
| 1 | 0.77 | 0.74 | 0.76 | 46 |
| micro avg | 0.78 | 0.78 | 0.78 | 101 |
| macro avg | 0.78 | 0.78 | 0.78 | 101 |
| weighted avg | 0.78 | 0.78 | 0.78 | 101 |






Il grafico Precision-Recall indica una buona tenuta della precisione rispetto al recall, con un valore di Average Precision (AP) di 0.69 e un'Accuracy generale dello 0.78.




Per quanto riguarda la Cross-Validation (eseguita con 5 sottogruppi o "fold"), i risultati mostrano la stabilità del classificatore:

cv_scores mean: 0.6647560975609756
cv_score variance: 0.01570633551457466
cv_score dev standard: 0.125324919766879

# 2.4 MULTINOMIAL NAIVE BAYES

I classificatori **Naive Bayes** sono strumenti probabilistici che utilizzano il principio di Bayes per assegnare una categoria ai dati. La particolarità di questi modelli è l'assunzione secondo cui ogni singola caratteristica di un esempio sia totalmente indipendente dalle altre.

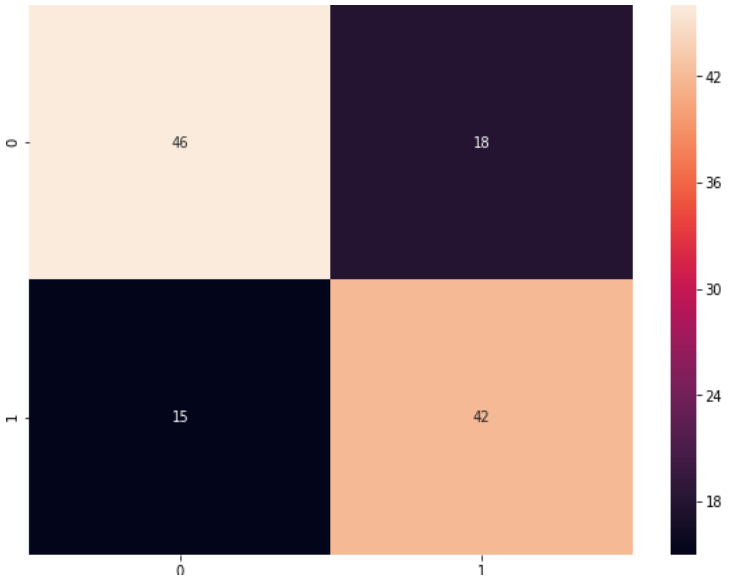
Nel modello di tipo multinomiale, i dati vengono interpretati come conteggi o frequenze di eventi. In questo contesto:

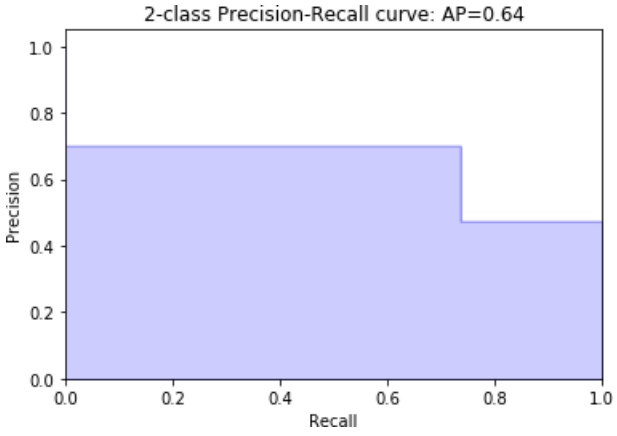
- Gli esempi vengono visti come risultati generati da una distribuzione particolare (polinomiale).
- Ogni caratteristica rappresenta la probabilità specifica che un determinato evento si verifichi all'interno del sistema.

Questo metodo è molto efficace per gestire dati che possono essere contati, rendendolo ideale per compiti come l'organizzazione automatica di documenti o messaggi.

Clasification report:

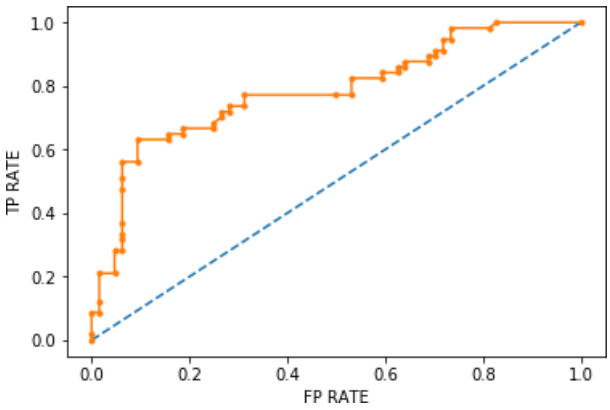
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.75 | 0.72 | 0.74 | 64 |
| 1 | 0.70 | 0.74 | 0.72 | 57 |
| micro avg | 0.73 | 0.73 | 0.73 | 121 |
| macro avg | 0.73 | 0.73 | 0.73 | 121 |
| weighted avg | 0.73 | 0.73 | 0.73 | 121 |



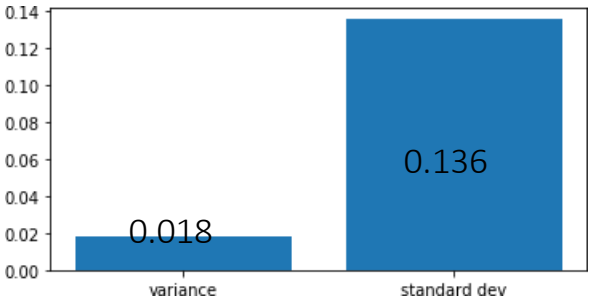



L'algoritmo Multinomial Naive Bayes mostra un'accuratezza del 73% e una precisione media (average precision) di 0.639.

Rispetto ai modelli analizzati in precedenza, questa soluzione presenta una curva di precisione più contenuta rispetto al recall. Anche l'integrazione della tecnica SMOTE, utilizzata per bilanciare i dati, non produce miglioramenti significativi, portando l'average precision a un valore pressoché identico di 0.640.




Per l'AUC invece abbiamo un valore pari a 0.785.




Per quanto riguarda la cross validation (con cv=5) sul classificatore i dati ottenuti sono:

cv_scores mean ---> 0.6444512195121951
cv_score variance ---> 0.018386875371802495
cv_score dev standard ---> 0.13559821301109576

# 2.5 NEURAL NETWORK

La rete neurale è un sistema di calcolo ispirato al funzionamento del cervello biologico. Questa struttura si basa sul "connessionismo", ovvero sull'uso di numerosi neuroni artificiali interconnessi che elaborano le informazioni in parallelo.

Si tratta di un sistema dinamico e adattivo: durante la fase di apprendimento, la rete modifica i propri collegamenti interni in risposta ai dati che riceve, ottimizzando progressivamente le proprie prestazioni.

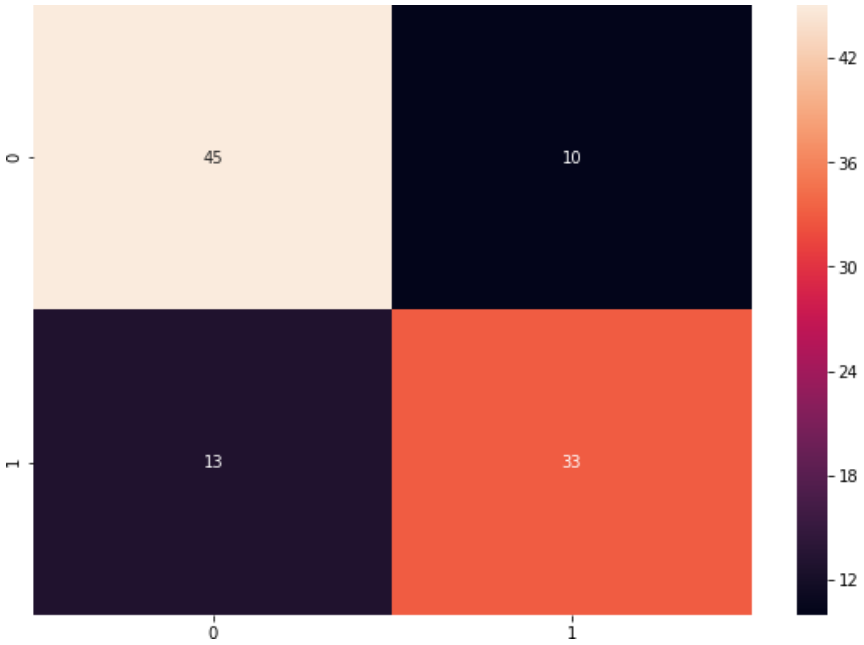
Nello specifico, il modello utilizzato segue un'architettura sequenziale a tre livelli:

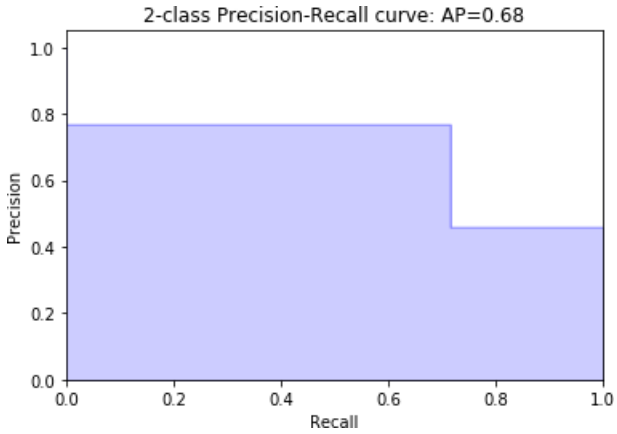
- Livello di Input: composto da 41 ingressi, corrispondenti alle caratteristiche fornite al sistema.
- Livello Nascosto (Hidden Layer): formato da 17 neuroni che elaborano i segnali provenienti dall'input.
- Livello di Output: costituito da un singolo neurone.

Essendo una classificazione di tipo "single-class", il risultato finale è un valore numerico compreso tra 0 e 1. Questo valore indica se un esempio appartiene o meno alla categoria "recurrence-events".

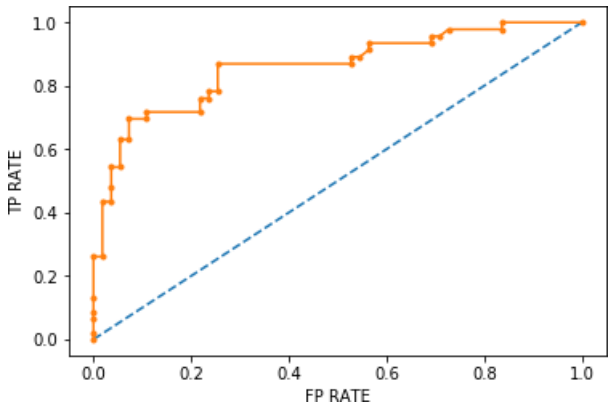
Clasification report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.82 | 0.80 | 55 |
| 1 | 0.77 | 0.72 | 0.74 | 46 |
| micro avg | 0.77 | 0.77 | 0.77 | 101 |
| macro avg | 0.77 | 0.77 | 0.77 | 101 |
| weighted avg | 0.77 | 0.77 | 0.77 | 101 |

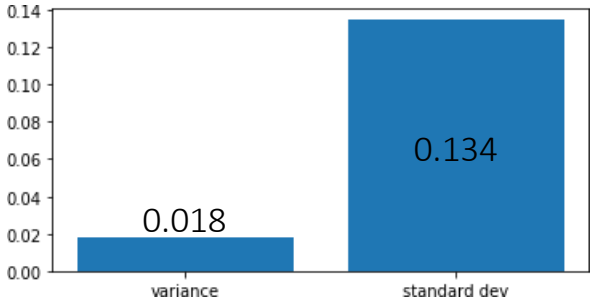



L’average precision risulta pari a 0.679, con un’accuratezza di 0.77 in linea con le altre tecniche trattate e una F1 di 0.742.




Nella rete neurale, l'AUC è pari a 0.860.




Per quanto riguarda la cross validation (con cv=5) sul classificatore i dati ottenuti sono:

cv_scores mean: 0.639567903086635
cv_score variance: 0.018067486318166993
cv_score dev standard: 0.13441535000946503

# 2.6 TABELLA RIASSUNTIVA

| *ALGORITMO* | *ACCURATEZZA* | *VARIANZA* | *DEV.STANDARD* | *F1* | *AVERAGE-PRECISION* | *AUC* |
|---|---|---|---|---|---|---|
| K-NN | 0.75<br><br>0.772(con SMOTE) | 0.001<br><br>0.005(con SMOTE) | 0.034<br><br>0.067(con SMOTE) | 0.438<br><br>0.716 (con SMOTE) | 0.491<br><br>0.691(con SMOTE) | 0.688<br><br>0.875 (con SMOTE) |
| Random forest | 0.777 | 0.009 | 0.096 | 0.757 | 0.697 | 0.890 |
| Support-vector machines | 0.782 | 0.016 | 0.125 | 0.756 | 0.690 | / |
| Multinomial naive Bayes | 0.727 | 0.018 | 0.136 | 0.718 | 0.640 | 0.785 |
| Neural network | 0.772 | 0.018 | 0.134 | 0.742 | 0.679 | 0.860 |

In sintesi, l'analisi evidenzia che il Random Forest è il modello più bilanciato e performante, grazie al punteggio F1 (0.757) e all'AUC (0.890) più elevati del gruppo. Le SVM offrono l'accuratezza massima (0.782), mentre il K-NN dimostra l'importanza cruciale del bilanciamento dei dati, migliorando drasticamente solo dopo l'applicazione di SMOTE.

I modelli probabilistici come il Naive Bayes e la Rete Neurale si posizionano su livelli medi, con la Rete Neurale che mostra una buona solidità (F1 0.742) ma una stabilità leggermente inferiore rispetto ai modelli basati su alberi o vettori di supporto.

# 3. APPRENDIMENTO NON SUPERVISIONATO

L'apprendimento non supervisionato è una branca del machine learning in cui il sistema analizza informazioni prive di etichette o categorie predefinite. In questo scenario, l'obiettivo è esplorarne la struttura per identificare autonomamente schemi e caratteristiche comuni, permettendo al sistema di organizzare i dati in gruppi coerenti. Poiché le classi non sono note in anticipo, il modello deve "imparare" a distinguere gli esempi basandosi solo sulle somiglianze intrinseche presenti nei dati di input.

## 3.1 K-MEANS

Esso è un metodo di clustering partizionale che suddivide gli oggetti in un numero prestabilito di gruppi, indicato con K (nel nostro caso K=2). Il funzionamento si basa sulla rappresentazione dei dati come vettori in uno spazio geometrico, dove ogni cluster è identificato da un punto centrale chiamato centroide. Il processo è di tipo iterativo e si articola in quattro fasi principali:

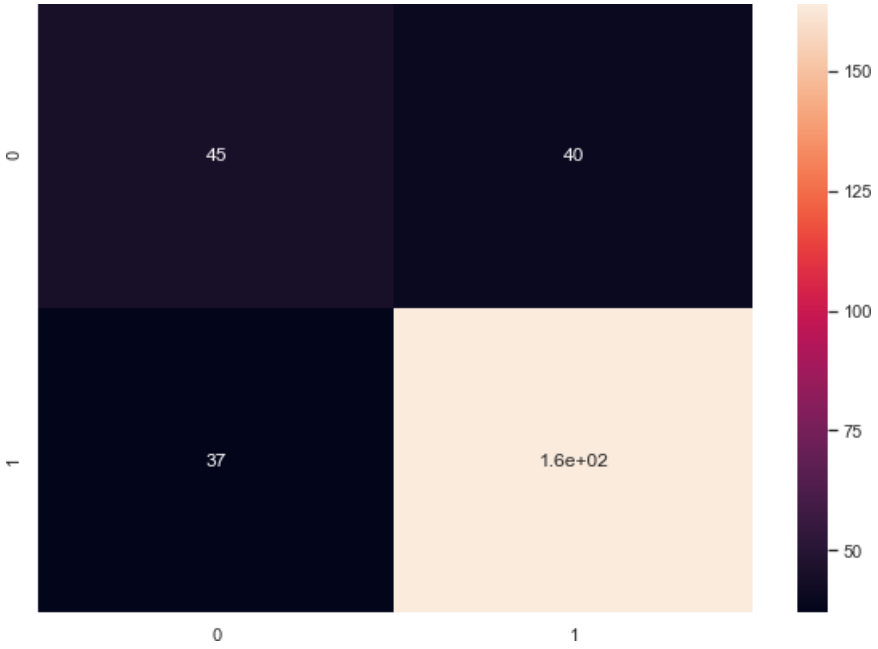
- Inizializzazione: il sistema crea K partizioni iniziali assegnando i punti d'ingresso a gruppi casuali o definiti tramite euristiche.
- Calcolo dei centroidi: viene individuato il centro geometrico di ogni gruppo basandosi sulla posizione attuale dei punti assegnati.
- Riassegnazione: ogni dato viene associato al cluster il cui centroide risulta geometricamente più vicino nello spazio vettoriale.
- Aggiornamento e convergenza: i centroidi vengono ricalcolati in base alla nuova disposizione dei punti e il ciclo si ripete fino a quando le posizioni non diventano stabili.

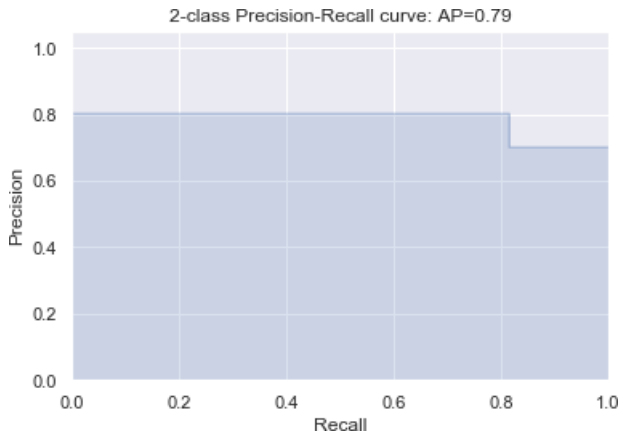
Questo approccio permette di scoprire relazioni nascoste nei dati senza alcuna supervisione umana, rendendo il K-means uno strumento fondamentale per l'analisi esplorativa dei dati.

```
Clasification report:
              precision    recall  f1-score   support

           0       0.55      0.53      0.54        85
           1       0.80      0.82      0.81       201

   micro avg       0.73      0.73      0.73       286
   macro avg       0.68      0.67      0.67       286
weighted avg       0.73      0.73      0.73       286
```




L'analisi tramite **K-Means** evidenzia che la struttura spaziale dei dati permette di raggruppare con discreta coerenza le istanze appartenenti alla categoria **'no-recurrence-events'**. Al contrario, si riscontra una significativa sovrapposizione geometrica per i casi di **'recurrence-events'**, che l'algoritmo non riesce a isolare in un cluster distinto.

Nonostante questa parziale sovrapposizione nelle distribuzioni, le metriche aggregate mostrano risultati interessanti:

- **Accuratezza (allineamento ai cluster):** 73% (0.731)
- **Average Precision:** 0.785

Tali valori confermano che, sebbene il K-Means operi in modalità **non supervisionata** (ovvero senza la guida delle etichette), i centroidi individuati riflettono in buona parte la realtà clinica del dataset. La difficoltà nel distinguere le ricorrenze suggerisce che tali eventi non formano un gruppo isolato, ma condividono caratteristiche di prossimità spaziale con i casi di non-ricorrenza.

# 4. **ONTOLOGIE**

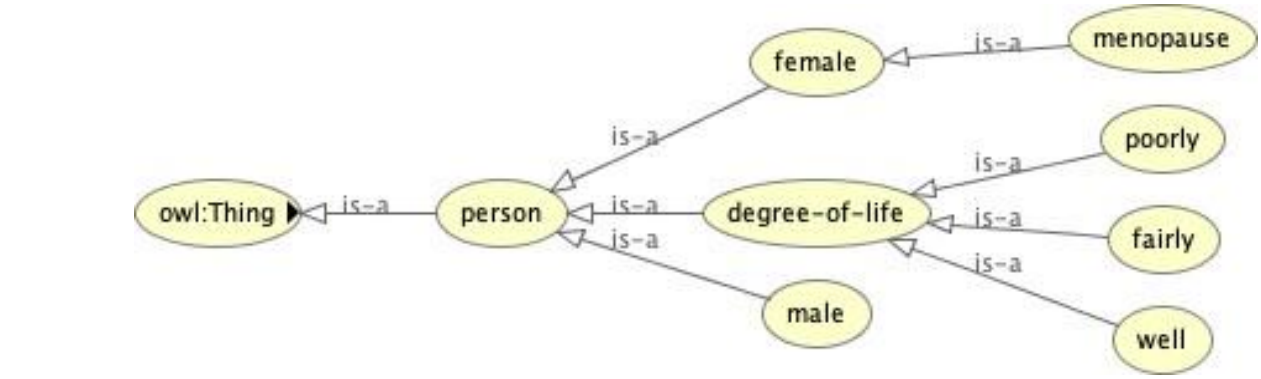
Un’ontologia definisce formalmente il significato dei simboli all’interno di un sistema informatico. Tale specifica risulta fondamentale per garantire l’interoperabilità semantica, ovvero la capacità di basi di conoscenza eterogenee di interagire preservando il significato dei dati. In quanto "specificazione di una concettualizzazione", l'ontologia rappresenta formalmente oggetti, concetti e relazioni propri di uno specifico dominio di interesse.

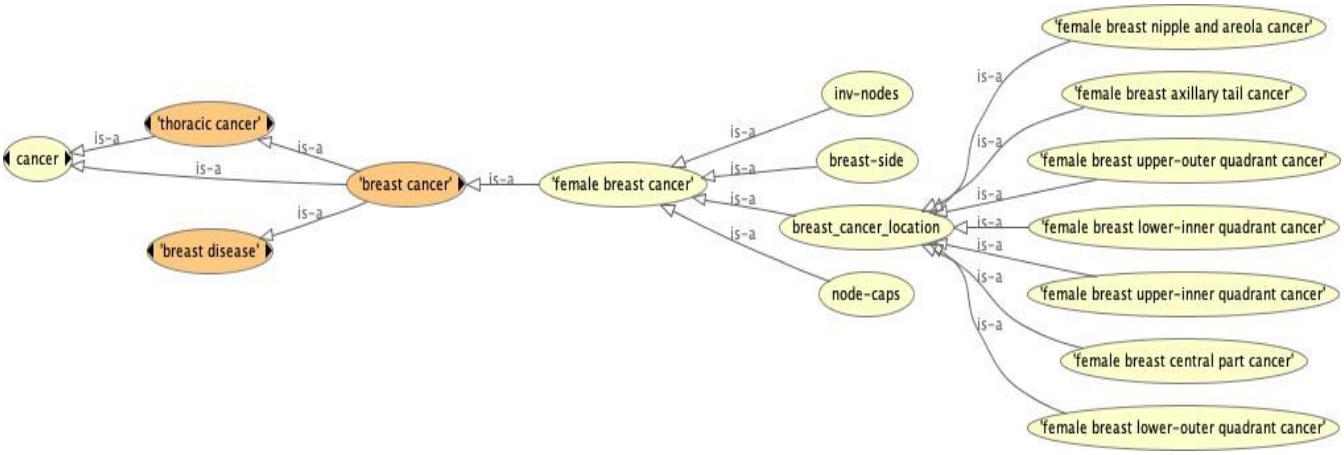
La modellizzazione è stata effettuata tramite il software Protégé. Il lavoro ha previsto l'integrazione di un'ontologia medica preesistente relativa alle malattie (DOID), al fine di stabilire connessioni coerenti. Per colmare le lacune informative del dataset principale, è stato aggiunto un dataset esterno sui tumori primari, selezionando le caratteristiche più rilevanti.

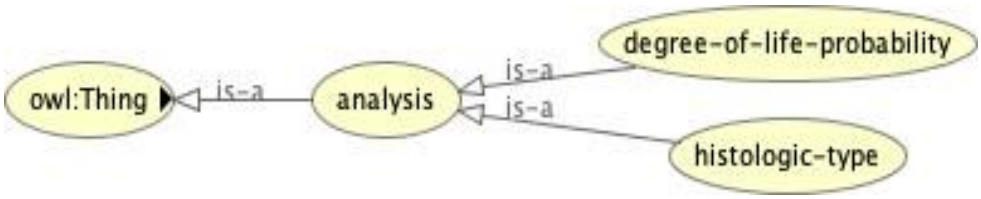
Questa integrazione ha generato un nuovo ecosistema informativo comprendente le patologie e le proprietà derivanti dall'unione delle diverse fonti.

L’interrogazione dell'ontologia permette di estrarre conoscenze precedentemente non disponibili. Attraverso lo strumento DL Query di Protégé e l'utilizzo del ragionatore HermiT, è stata calcolata la probabilità di benessere per pazienti di sesso femminile affette da cancro in base all'età. Tale metodologia consente di arricchire il dataset originale con una nuova variabile relativa alla probabilità dello stato di salute in funzione della fascia anagrafica.

Di seguito viene riportata come è stata modellata l'ontologia:



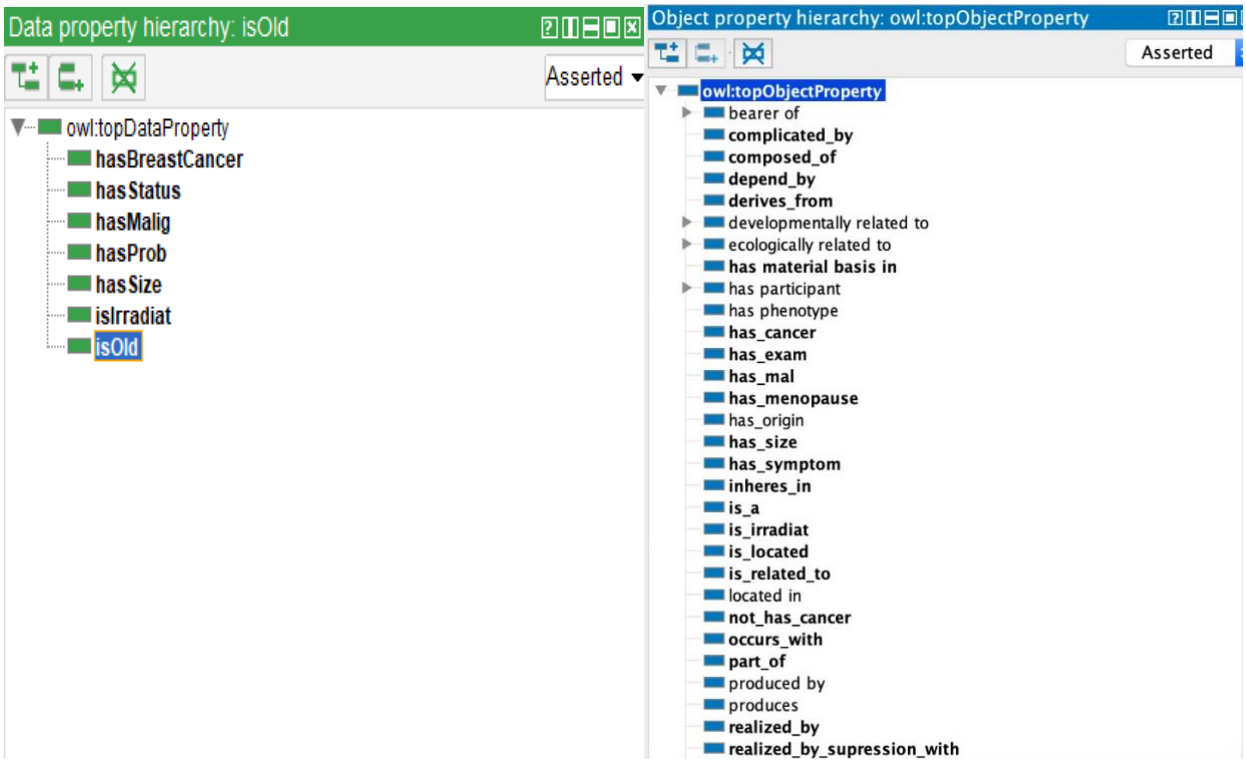





Le connessioni tra le diverse entità dell'ontologia vengono stabilite attraverso l'impiego di specifiche proprietà, suddivise in due tipologie principali:

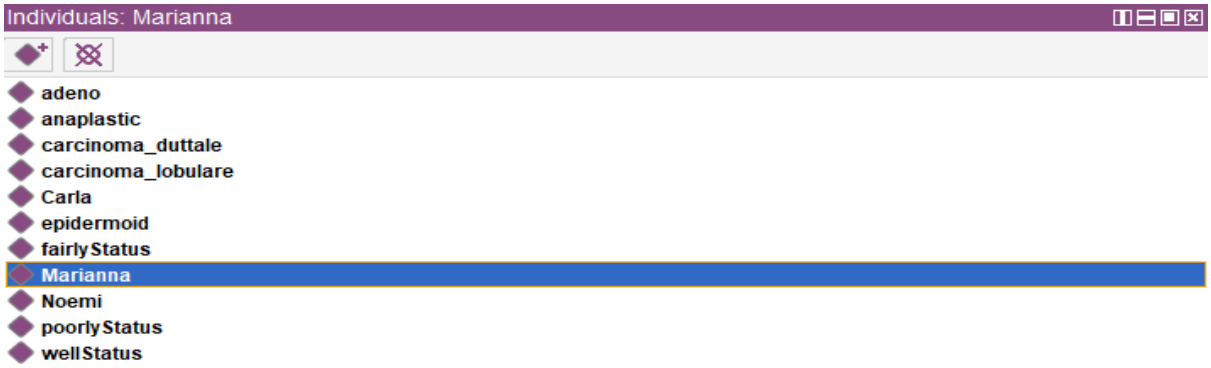
- **Object Property**: questa tipologia di relazione collega tra loro due individui, i quali possono appartenere alla medesima classe o a classi differenti.
- **Data Property**: questa tipologia permette di associare un individuo a un valore di tipo primitivo (come un numero, una stringa o una data).

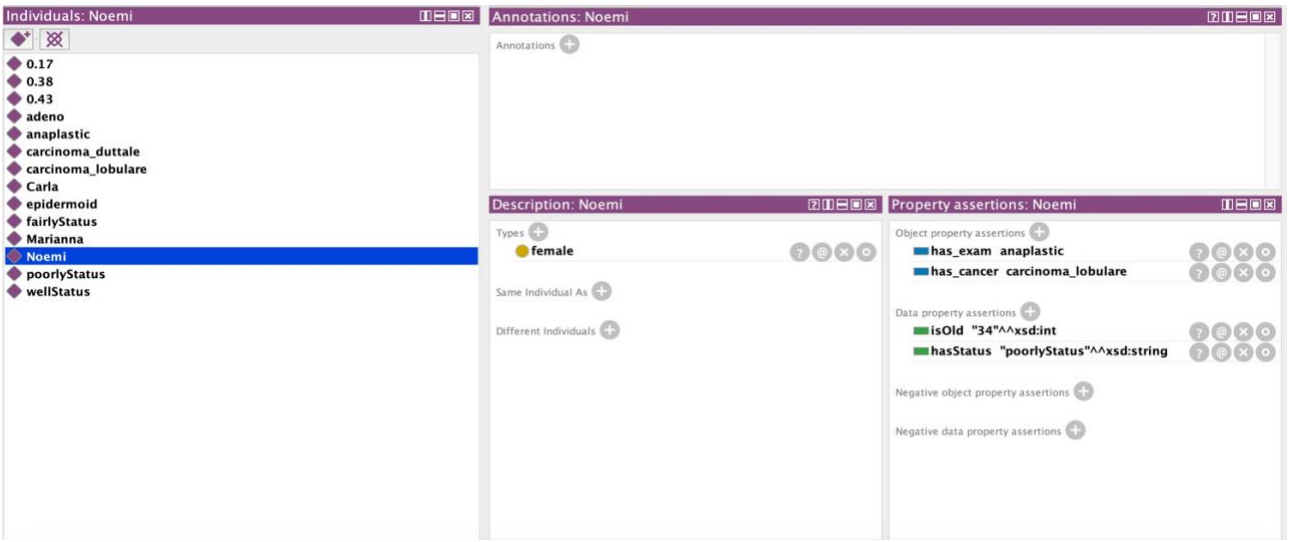



Il sistema include la creazione di istanze per inserire dati concreti nell'ontologia:

- Esami: identificazione di procedure specifiche, come l'esame istologico.
- Pazienti: rappresentazione di singoli individui a cui associare patologie e relative relazioni cliniche.

Questi elementi permettono di trasformare i concetti astratti in casi reali analizzabili dal ragionatore.

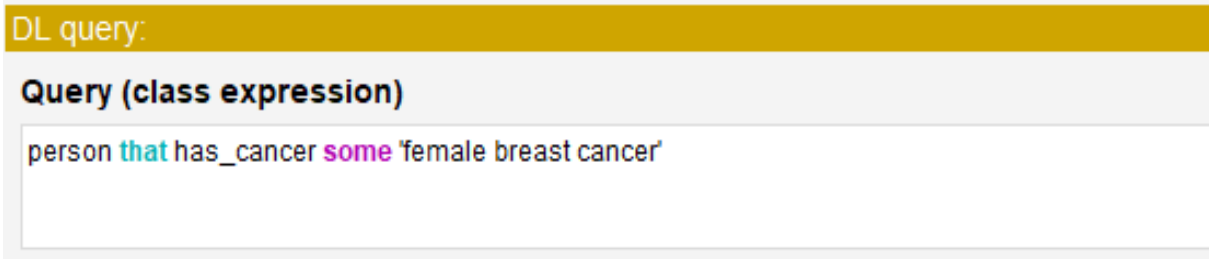





L'interrogazione dell'ontologia è avvenuta attraverso la formulazione di query con livelli di complessità crescente. Il processo è iniziato con ricerche dirette, come l'individuazione di persone affette da tumore al seno e dei relativi esami istologici effettuati. Successivamente, sono state elaborate query più articolate per generare elenchi di individui caratterizzati da uno specifico stato di salute, calcolato combinando molteplici attributi personali.

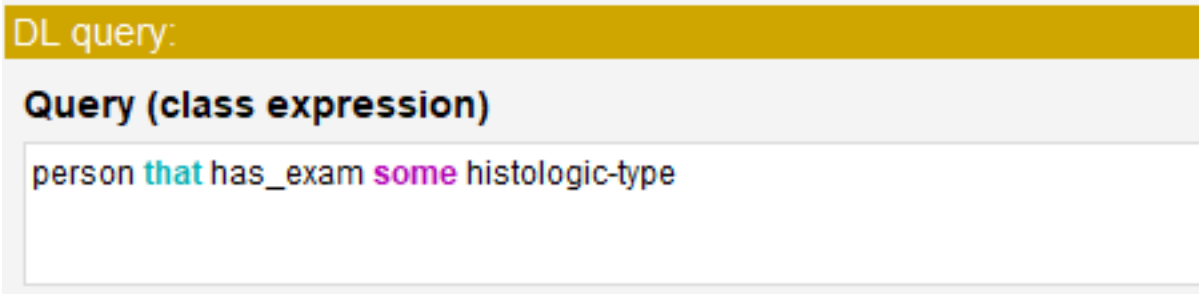
Le query formulate sono le seguenti:

1. Query: Ottenere tutte le persone che sono affette da un qualsiasi tumore al seno

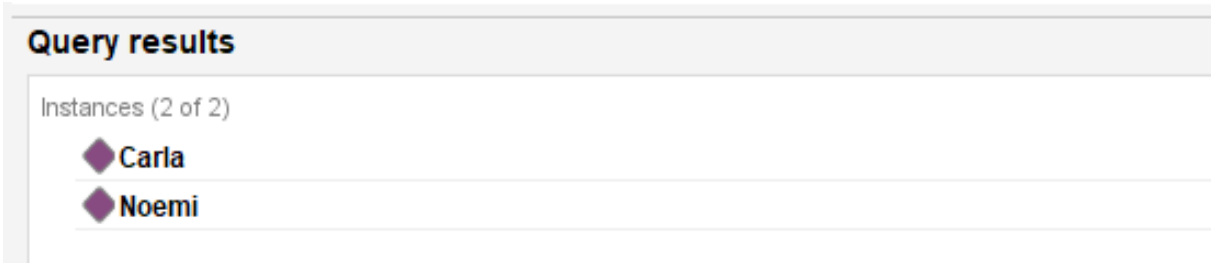



Result Set

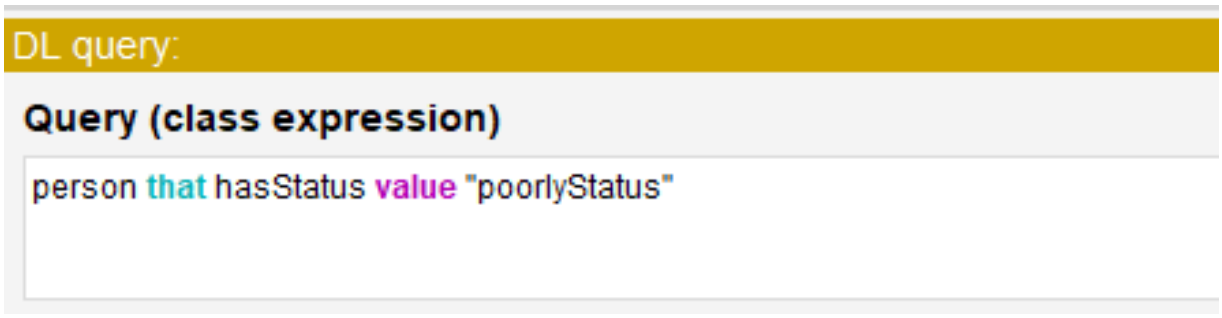
2. Query: Ottenere tutte le persone che si sono sottoposte ad un esame istologico



Result Set

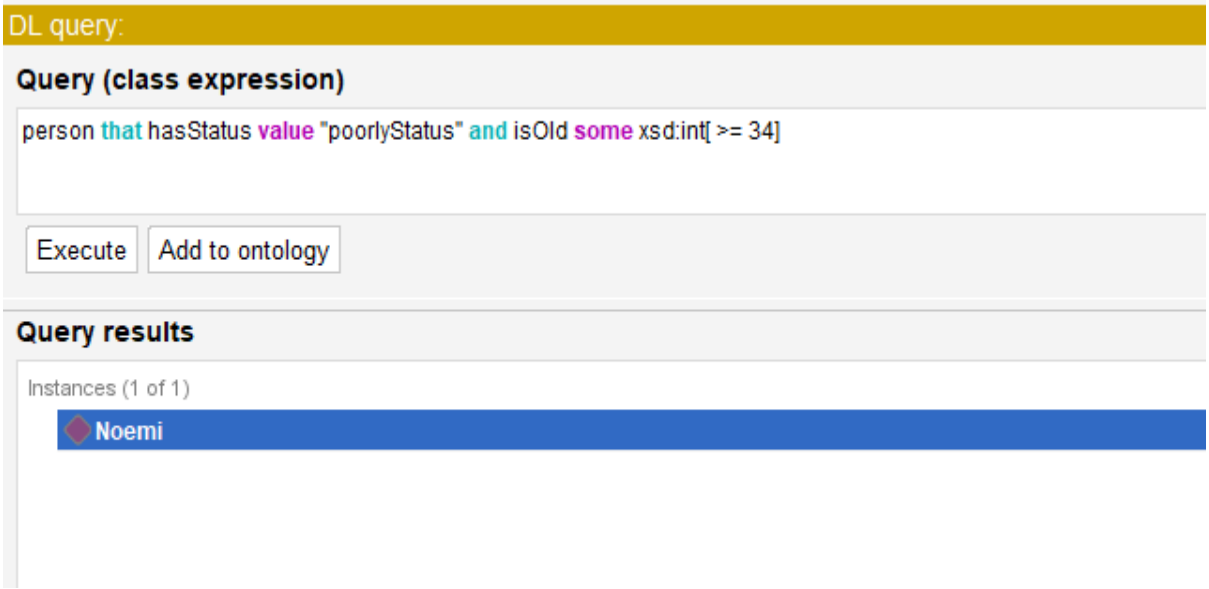



3. Query: Ottenere tutte le persone con uno stato di salute basso




Result Set


4. Query: Ottenere tutte le persone con età maggiore di 34 con uno stato di salute basso



Nello specifico, le ultime operazioni hanno sfruttato le capacità del **reasoner** per l'inferenza automatica:

- L'analisi si è basata esclusivamente sull'età del soggetto e sulla relativa probabilità associata allo stato di salute.
- Il **reasoner** ha combinato autonomamente questi due parametri per attribuire lo stato di salute a ogni individuo.
- L'elaborazione ha permesso di ottenere classificazioni non esplicitate direttamente, ma derivate logicamente dalla struttura dei dati.