Politecnico di Bari

Politecnico di Bari

Dipartimento di Ingegneria Elettrica e dell'Informazione

Corso di Laurea Magistrale in Ingegneria dei Sistemi Medicali

DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE

TEMA D'ANNO IN BIOINFORMATICA AVANZATA, ING/INF-06

SISTEMA DI PREVISIONE DELLA DIAGNOSI CON PROGNOSTICA NEGATIVA DEL TUMORE AL SENO

Professore:
Vitoantonio Bevilacqua

Studentesse:
Benedetta **ALTAMURA**
Martina **CAFERRA**

INDUSTRIAL INFORMATICS LABORATORIO DI INFORMATICA INDUSTRIALE

Anno Accademico 2022/2023

apulian bioengineering company

INTRODUZIONE

Il cancro al seno è una delle forme più comuni di cancro nella popolazione femminile con oltre **1.300.000 casi** e **450.000 decessi** ogni anno in tutto il mondo. Esistono diversi sottotipi di cancro al seno, che differiscono per la loro biologia e le loro caratteristiche cliniche.

La classificazione dei tumori al seno si basa sulla biologia del tumore, sulle sue caratteristiche istologiche e sulle sue proprietà genetiche. Questa classificazione aiuta a identificare il trattamento più appropriato per ogni paziente e a prevedere la prognosi.

Esistono diversi tipi di tumori della mammella, tra cui le forme più comuni di tumore invasivo al seno con prognosi sfavorevole sono:

- **carcinoma duttale**: rappresenta il 20% delle neoplasie del seno.
- **carcinoma lobulare**: è la *tipologia più comune* di neoplasia alla mammella e colpisce nel 70-80% dei casi.

Le altre forme meno frequenti e con prognosi favorevole sono:

- carcinoma intraduttale in situ
- carcinoma tubulare
- carcinoma papillare
- carcinoma mucinoso
- carcinoma cribriforme

ANALISI DELL'ESPRESSIONE GENICA DIFFERENZIALE

Il campo di interesse è quello dell'analisi dell'espressione differenziale, cioè l'identificazione di geni che presentano significative differenze del loro livello di espressione tra due o più condizioni sperimentali. Si valuta cioè, se le differenze osservate tra i counts delle diverse condizioni sperimentali siano o meno statisticamente significative.

Gli approcci sperimentali per studiare su larga scala il profilo trascrizionale, definendo quali geni vengono trascritti e a quale livello in una determinata condizione, sono principalmente due: approcci basati sull'utilizzo di DNA microarray e metodi basati sul sequenziamento con tecnologie NGS dell'RNA a seguito della sua conversione in cDNA. La tecnica dell'**RNA-seq** permette il sequenziamento degli RNA messaggeri, permette di identificare le molecole di RNA e di quantificare la loro espressione in un campione biologico. Gli RNA cellulari, dopo essere stati estratti dalla cellula, vengono retrotrascritti in DNA (cDNA), sequenziato attraverso tecniche NGS.

Vengono così ottenute delle *sequenze di DNA di lunghezza variabile* in base alla tecnologia di sequenziamento utilizzata, definite **reads**. Tali sequenze (reads) vengono poi mappate su un genoma o un trascrittoma di riferimento per identificare i geni espressi nel campione in esame. Il *totale delle read allineate su un gene* (o un trascritto, o un esone), detto **count**, è una unità di misura dell'espressione del gene stesso.

Il vantaggio della tecnologia RNA-Seq sta nella possibilità di essere utilizzata anche quando non è nota la sequenza del gene in esame. Le reads sono sottoposte a controllo di qualità e pre-processing.

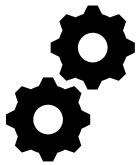
WORKFLOW 1

GDC Data Portal

TCGA-BRCA Program

STAR-Counts -> RNA-Seq

Riduzione numero trascritti



Normalizzazione dei dati

Selezione e riduzione delle features

con Test binomiale negativo

PCA

Bilanciamento dei dati

Training del classificatore e valutazione delle performance

Hidden Layers scelti dall'algoritmo genetico

La classificazione è binaria:

1. Caso migliore (Primary Diagnosis = Lobular Carcinoma) -> 0
2. Caso peggiore (Primary Diagnosis = Infiltrating Ductal Carcinoma) -> 1

Sono stati considerati tre classificatori:

- Per modellazione non lineare è stata utilizzata una Rete Neurale Artificiale (ANN) e il Decision Tree
- Per la modellazione lineare è stata utilizzata la Logistic Regression

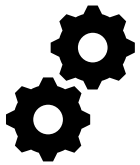
WORKFLOW 2

GDC Data Portal

TCGA-BRCA Program

STAR-Counts -> RNA-Seq

Riduzione numero trascritti



Normalizzazione dei dati

Riduzione delle features con PCA

Bilanciamento dei dati

Training del classificatore e valutazione delle performance

Hidden Layers scelti dall'algoritmo genetico

La classificazione è binaria:

1. Caso migliore (Primary Diagnosis = Lobular Carcinoma) -> 0
2. Caso peggiore (Primary Diagnosis = Infiltrating Ductal Carcinoma) -> 1

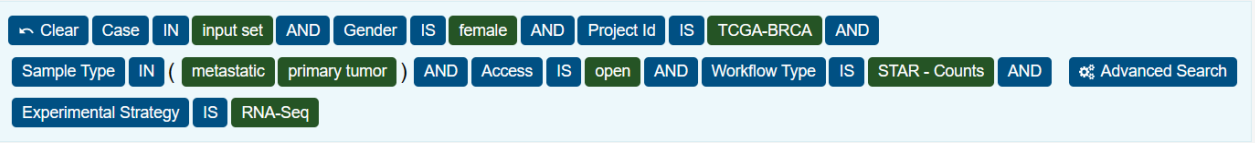
Sono stati considerati tre classificatori:

- Per modellazione non lineare è stata utilizzata una Rete Neurale Artificiale (ANN) e il Decision Tree
- Per la modellazione lineare è stata utilizzata la Logistic Regression

ACQUISIZIONE DEI DATI – GDC Data Portal

Il download dei dati dal GDC Portal è stato effettuato mediante query http utilizzando due filtri JSON, il primo creato attraverso Postman e il secondo scaricato dal GDC Portal.

FILTRO 1: permette di scaricare i dati relativi ai "counts. Considera file ad accesso open del progetto "TCGA_BRCA" che abbiano come workflow "STAR-Counts" e come approccio sperimentale la tecnica dell'RNA-Seq.

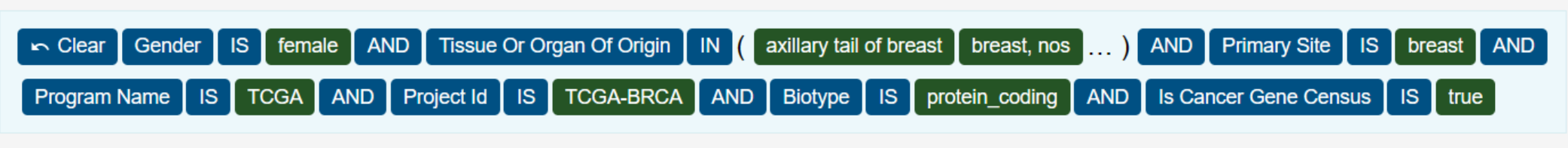



```
%% files
files_url= strcat(base_url, "files");
method= RequestMethod.POST;
filter= fileread('filtro.json');
filter= jsondecode(filter);
body= MessageBody(filter);
uri= URI(files_url);
request= RequestMessage(method, [], body);
[response, altro, altro1]= send(request, uri) ;
```

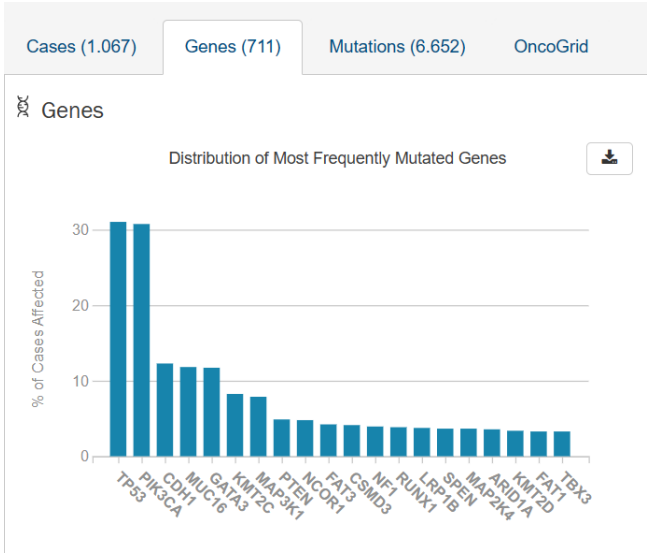
```
{
    "filters":{
        "op": "and",
        "content": [
            {"op": "=",
             "content": {
                 "field": "cases.project.project_id",
                 "value": "TCGA-BRCA"}},
            {"op": "=",
             "content": {
                 "field": "analysis.workflow_type",
                 "value": "STAR - Counts"} },
            {"op": "=",
             "content": {
                 "field": "access",
                 "value": "open"}},

            {"op": "in",
             "content": {
                 "field": "cases.samples.sample_type",
                 "value":["metastatic","primary tumor"] }},

            {"op": "in",
             "content": {
                 "field": "cases.demographic.gender",
                 "value":"female" }
            }
        ]
    },

"fields":"file_id,file_name,cases.case_id,cases.diagnoses.primary_diagnosis,cases.diagnoses.ajcc_pathologic_stage,cases.diagnoses.ajcc_pathologic_n,cases.diagnoses.ajcc_pathologic_m,cases.diagnoses.ajcc_pathologic_t,cases.diagnoses.tumor_stage,cases.diagnoses.age_at_diagnosis,cases.diagnoses.morphology,
cases.demographic.gender,cases.demographic.age_at_index,cases.demographic.race,cases.diagnoses.treatments.treatment_or_therapy",
    "size": 1000

}
```

ACQUISIZIONE DEI DATI – GDC Data Portal

FILTRO 2: Il secondo filtro JSON scaricato direttamente dal portale GDC, contenente i geni implicati in patologie tumorali. A causa dell'elevato numero di geni codificanti proteine presenti nel progetto TCGA-BRCA (circa 20.096 geni) si è resa utile una selezione dei 711 geni considerati implicati con i tumori definiti dal *Catalogue of Somatic Mutations in Cancer* (COSMIC) *Cancer Gene Census* (CGC).




```
%% files con geni di interesse
files_url= strcat(base_url, "genes");
method= RequestMethod.POST;
filter= fileread('genes.2023-01-23.json');
body_gene= MessageBody(filter);
uri= URI(files_url);
request_gene= RequestMessage(method, [], body_gene);
[response_gene, altro, altro1]= send(request_gene, uri);
```



DATASET

data_table

711x973 table

	1 ea645243-df49-4466-a255-9f:	2 001cef41-ff86-4d3f-a140-a64	3 0685edd2-ce1c-4e0e-8dda-39	4 b2aac45b-2073-4c7a-adb9-76
1 LASP1	11270	18309	27106	18998
2 HOXA11	55	8	10	71
3 CREBBP	8666	10279	8017	7937
4 ETV1	841	387	4334	1365
5 GAS7	4414	1475	3975	2928
6 CD79B	571	178	221	35
7 PAX7	1	2	0	2
8 BTK	584	398	1609	398
9 BRCA1	658	1104	735	321
10 WAS	705	348	2046	417
11 WWTR1	16909	2325	9343	4342
12 CD74	80004	61515	314340	113657
13 BIRC3	4174	1504	3134	587
14 FAS	807	323	2357	632
15 BCLAF1	8429	7746	9780	7810
16 ANK1	72	62	77	20
17 RABEP1	2391	13837	2989	8572
18 ZCCHC8	1583	1534	1657	1249
19 CUL3	5291	4477	4987	3731

data_table contiene sulle righe 711 geni e sulle colonne 973 pazienti tali per cui tutti i campi richiesti nel primo filtro fossero presenti.

Ogni cella della matrice contiene il numero di read (ovvero i counts) che mappano sugli esoni del gene in ciascuno dei campioni.

DATASET

clinical_data

973x10 table

	1 age	2 age_at_diag	3 ajcc_pathologic_stage	4 ajcc_pat_n	5 ajcc_pathologic_t	6 ajcc_pat_m	7 race	8 primary_diag	9 diag	10 treatments
1 ea645243-df49-4466-a255-9f3d4321e357	67	24647	'Stage IA'	'N0'	'T1c'	'MX'	'black or af...	'Infiltrating duc...	'8500/3'	'no'
2 001cef41-ff86-4d3f-a140-a647ac4b10a1	60	22279	'Stage IA'	'N0 (mol+)'	'T1c'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'no'
3 0685edd2-ce1c-4e0e-8dda-393139af4223	31	11354	'Stage I'	'N0 (i-)'	'T1c'	'M0'	'white'	'Secretory carci...	'8502/3'	'yes'
4 b2aac45b-2073-4c7a-adb9-769a4fdcc111	71	26221	'Stage IIB'	'N1'	'T2'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'no'
5 0741b5db-4405-42ba-b63a-c6ee4f341480	79	28940	'Stage IIA'	'N0'	'T2'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'no'
6 bb8d42d3-ad65-4d88-ae1d-f9aadfc7962d	69	25230	'Stage IIA'	'N0 (i-)'	'T2'	'M0'	'white'	'Lobular carcin...	'8520/3'	'yes'
7 045c13ef-3db7-4adf-b0a3-23338f0479f3	49	18014	'Stage IIA'	'N1'	'T1c'	'M0'	'black or af...	'Infiltrating duc...	'8500/3'	'no'
8 cea9d8f9-e18c-4947-a461-5f712e3c1e6d	37	13817	'Stage IA'	'N0'	'T1c'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'yes'
9 0dca98b0-f43e-45b6-9a02-00092c78678c	72	26588	'Stage IIA'	'N0'	'T2'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'no'
10 2fdfd287-d13c-4910-9788-73987d45908a	73	26845	'Stage IA'	'N0 (i-)'	'T1'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'yes'
11 0fe1419e-a005-407c-8ae7-15c4c1579539	51	18788	'Stage I'	'N0'	'T1c'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'yes'
12 6cdc0d53-f813-4101-81e5-9bee68270536	46	17152	'Stage I'	'N0'	'T1c'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'no'
13 b63391a0-73f8-4544-9e94-f6529245ca2a	78	28495	'Stage IIA'	'N1a'	'T1c'	'M0'	'white'	'Lobular carcin...	'8520/3'	'yes'
14 7d9d3522-ec3b-4efe-8c8e-c0e675276ef5	67	24493	'Stage IIA'	'N1'	'T2'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'yes'
15 0bc5744c-5fa3-45bb-87d0-70a02068b392	30	11204	'Stage IIA'	'N0'	'T2'	'M0'	'black or af...	'Infiltrating duc...	'8500/3'	'yes'
16 a4903de8-6cf5-4541-8ec7-065beace8b44	61	22642	'Stage IIIC'	'N3'	'T2'	'M0'	'white'	'Lobular carcin...	'8520/3'	'yes'
17 757df7b0-4774-4493-98bf-999ded9ac86e	63	23229	'Stage IIB'	'N1'	'T2'	'M0'	'white'	'Infiltrating duc...	'8500/3'	'no'
18 67c73260-a242-4bba-87c5-d2302556dff7	50	18535	'Stage IIIA'	'N1mi'	'T3'	'M0'	'white'	'Lobular carcin...	'8520/3'	'yes'
19 35bd694d-1dd2-466f-ab27-03320614b40e	36	13458	'Stage IIA'	'N1mi'	'T1c'	'MX'	'black or af...	'Infiltrating duc...	'8500/3'	'yes'

Clinical_data: contiene sulle righe 973 pazienti e sulle colonne i dati clinici dei pazienti. Tra tutte le features presenti abbiamo selezionato «age at diagnosis» e «race» che verranno utilizzate per il classificatore.

DATASET

target

973x1 table

	1 primary_diag
1 ea645243-df49-4466-a255-9f3d4321e357	1
2 001cef41-ff86-4d3f-a140-a647ac4b10a1	1
3 0685edd2-ce1c-4e0e-8dda-393139af4223	0
4 b2aac45b-2073-4c7a-adb9-769a4fdcc111	1
5 0741b5db-4405-42ba-b63a-c6ee4f341480	1
6 bb8d42d3-ad65-4d88-ae1d-f9aadfc7962d	0
7 045c13ef-3db7-4adf-b0a3-23338f0479f3	1
8 cea9d8f9-e18c-4947-a461-5f712e3c1e6d	1
9 0dca98b0-f43e-45b6-9a02-00092c78678c	1
10 2fdfd287-d13c-4910-9788-73987d45908a	1
11 0fe1419e-a005-407c-8ae7-15c4c1579539	1
12 6cdc0d53-f813-4101-81e5-9bee68270536	1
13 b63391a0-73f8-4544-9e94-f6529245ca2a	0
14 7d9d3522-ec3b-4efe-8c8e-c0e675276ef5	1
15 0bc5744c-5fa3-45bb-87d0-70a02068b392	1

L'output scelto per il nostro classificatore riguarda la primary diagnosis del tumore al seno, in particolare la classe 1 è rappresentata dal Infiltrating ductal carcinoma (IDC) e la classe 2 dal lobular carcinoma. Sono entrambi tipi di cancro al seno, ma differiscono nel loro modo di crescita e presentazione:

- **Infiltrating Ductal Carcinoma**: è il tipo più comune di cancro al seno, che origina dalle cellule dei dotti lactiferi che portano il latte al capezzolo. IDC si diffonde attraverso i tessuti del seno infiltrandosi in profondità, dando luogo a tumori duri e spesso con una superficie irregolare.
- **Lobular Carcinoma**: questo tipo di cancro origina dalle cellule dei lobuli, che sono le unità produttrici di latte nel seno. LC si diffonde in modo diverso da IDC, infiltrandosi nei lobuli e poi nelle aree circostanti. Tumori di LC tendono ad essere più morbidi e avere una forma più regolare rispetto a IDC.

Target: è stato scelto come output la "primary diagnosis" distinguendo due casi: "Infiltrating duct carcinoma" e "Lobular carcinoma". È stata creata la table target con valori logici 0 e 1 rispettivamente per Lobular carcinoma e Infiltrating duct carcinoma.

PREPROCESSING DEI DATI

Si è reso utile un preprocessing dei dati al fine di:

1. **Eliminare le righe contenenti *valori NaN* dalla tabella clinical_data.** Il numero di pazienti si riduce da 973 a 770.

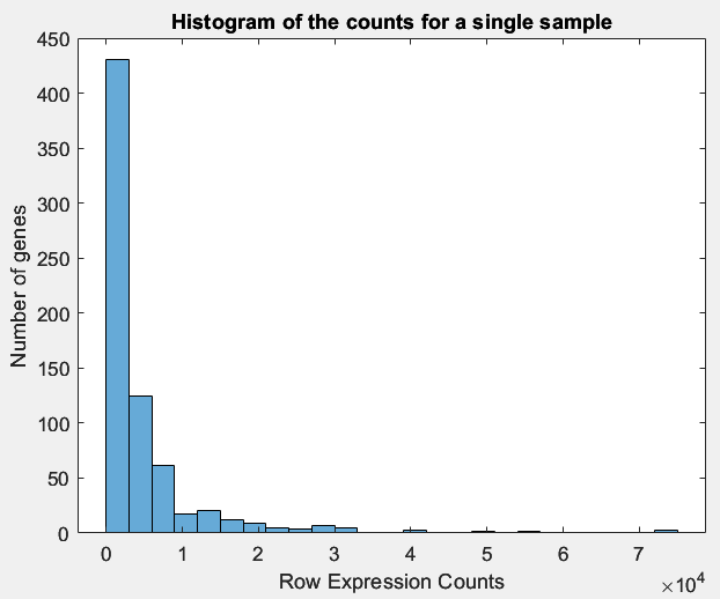
```
% elimino le righe con valori NaN
missing_values = ismissing(clinical_data);
sum_missing_values = sum(missing_values, 2);
missing_rows = find(sum_missing_values > 0);
clinical_data(missing_rows, :) = [];
data_table(:, missing_rows) = [];
```

2. **Eliminare le righe contenenti un certo numero di *counts nulli* dalla tabella data_table.** Il numero di geni si riduce da 711 a 682.

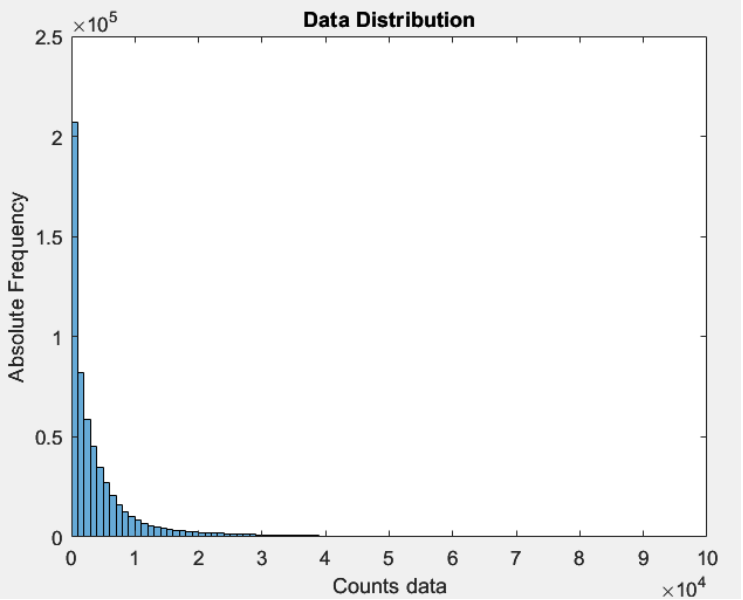
```
%Elimino le righe che hanno i counts nulli
geneData = table2array(data_table);
mask = geneData > 0;
sum_mask = sum(mask,2);
idx = sum_mask >= size(geneData,2)*80/100;
data_table(~idx,:) = [];
```

CARATTERISTICHE DEI DATI DI CONTEGGIO RNA-SEQ

Istogramma dei conteggi per un singolo campione:



Istogramma dei conteggi per tutti i campioni in analisi:



Entrambi i grafici illustrano alcune **caratteristiche comuni** dei dati di conteggio dell'RNA-seq:

- un basso numero di conteggi associati a una grande percentuale di geni
- una lunga coda destra a causa della mancanza di qualsiasi limite superiore per l'espressione
- ampia gamma dinamica

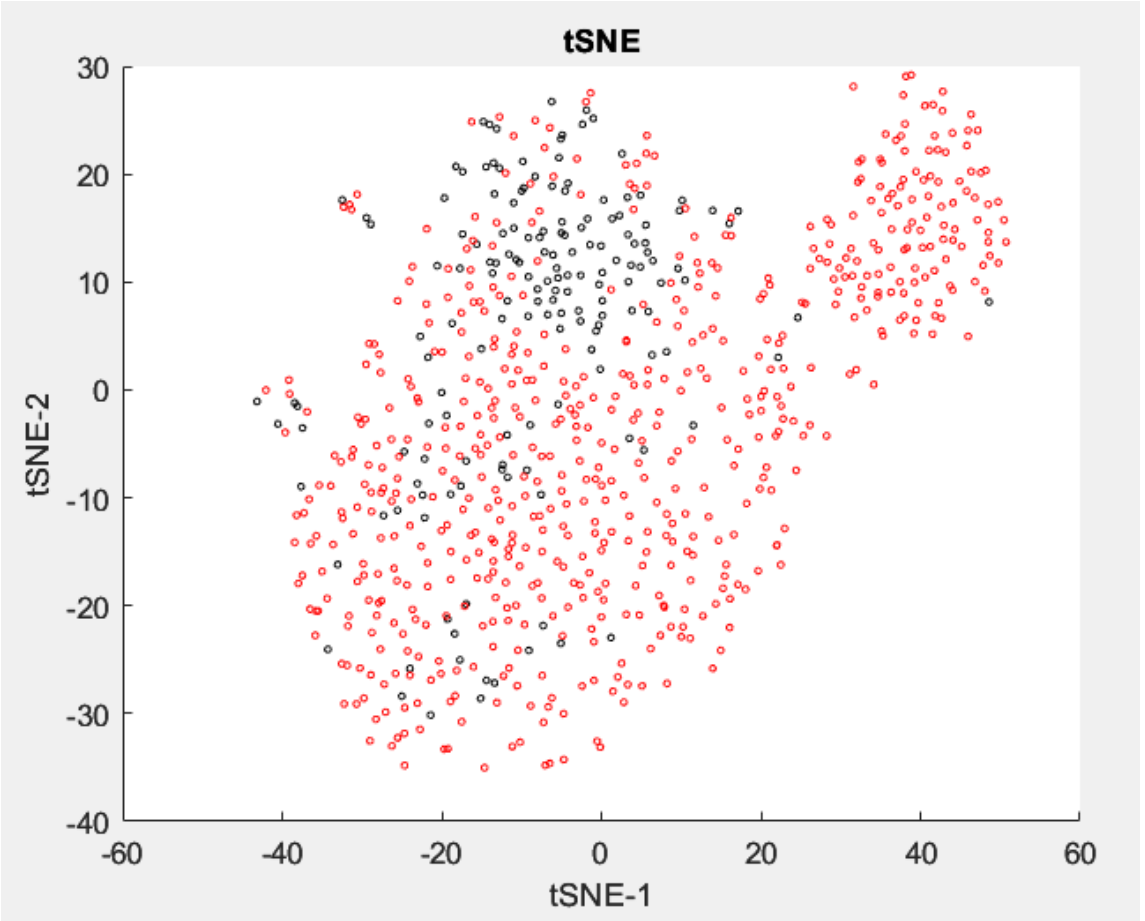
Osservando la forma dell'istogramma, vediamo che *non è distribuito normalmente*.

CARATTERISTICHE DEI DATI

La t-SNE è una tecnica di riduzione della dimensionalità adatta per incorporare dati dimensionali per la visualizzazione in uno spazio a bassa dimensione. Oggetti simili sono modellati da punti vicini e oggetti dissimili sono modellati da punti distanti con alta probabilità.

Abbiamo selezionato come numero di dimensioni 2 e come metrica di ottimizzazione la minimizzazione della divergenza di Kullback-Leibler

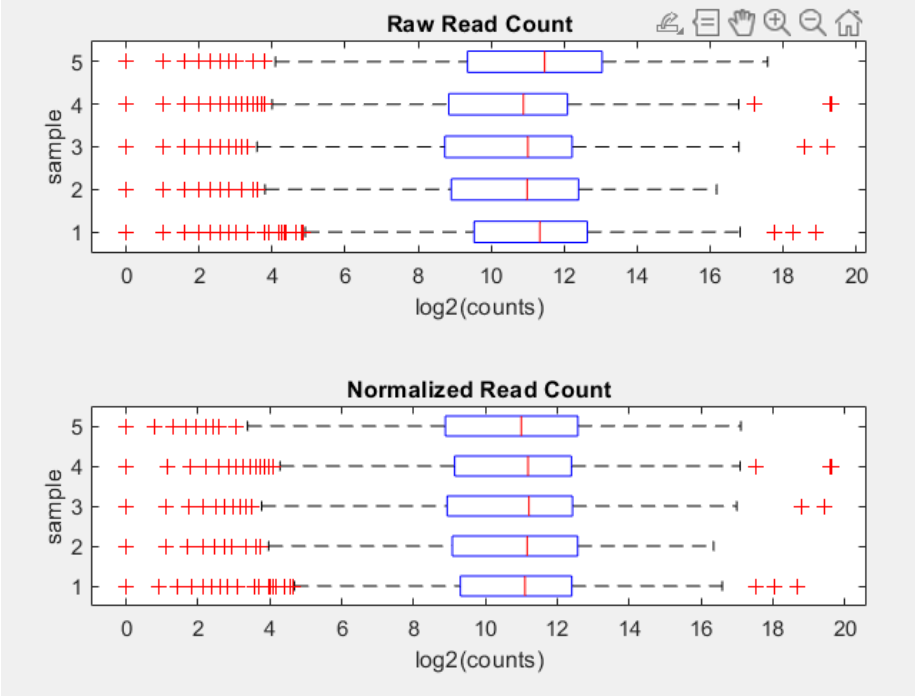
```
genDataEmbedding = tsne(data_table_array', ...
                        'Algorithm', 'exact', ...
                        'NumDimensions', 2);
```



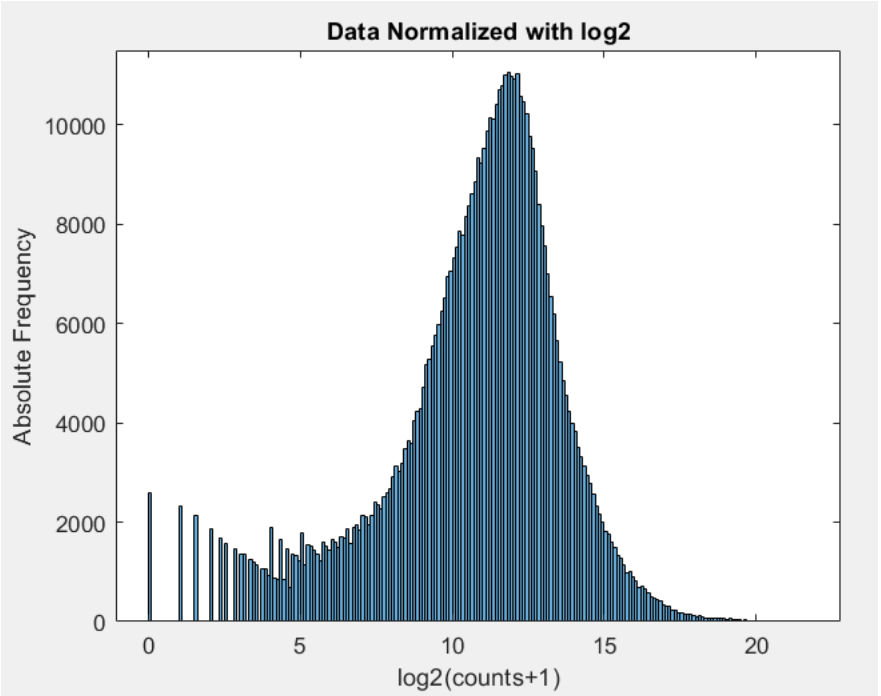
NORMALIZZAZIONE DEI DATI

I conteggi iniziali presenti nella Matrice dei Conteggi non sono valori direttamente interpretabili come livelli di espressione e non possono essere utilizzati direttamente per le analisi successive. È necessaria, dunque, la normalizzazione dei dati.

1. **Normalizzazione con Size Factor**: l'obiettivo è portare tutti i valori dei conteggi su una scala comune, rendendoli comparabili.



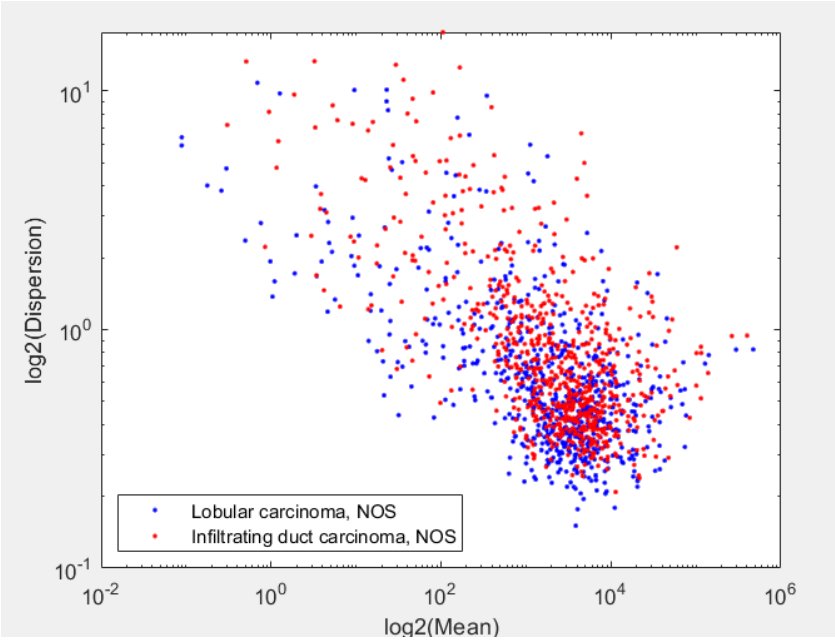
2. **Normalizzazione con log2**: permette di rendere i dati più uniformi e omogenei.



WORKFLOW 1

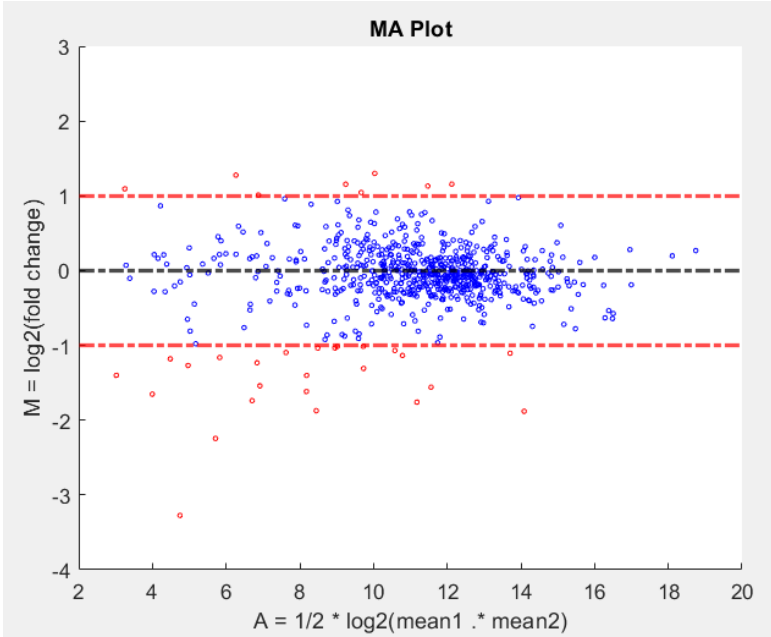
Media, Dispersione E Fold Change Dei Conteggi Normalizzati Con Size Factor

È possibile tracciare i valori di dispersione empirica rispetto alla media dei conteggi normalizzati in una scala logaritmica:




```
% compute the mean and the log2FC
meanBase = (mean1 +  mean2) / 2;
foldChange = mean1 ./ mean2;
log2FC = log2(foldChange);
```

L'MA plot consente di visualizzare la differenza di espressione tra i geni e di identificare quelli che hanno una differenza significativa di espressione tra due classi:



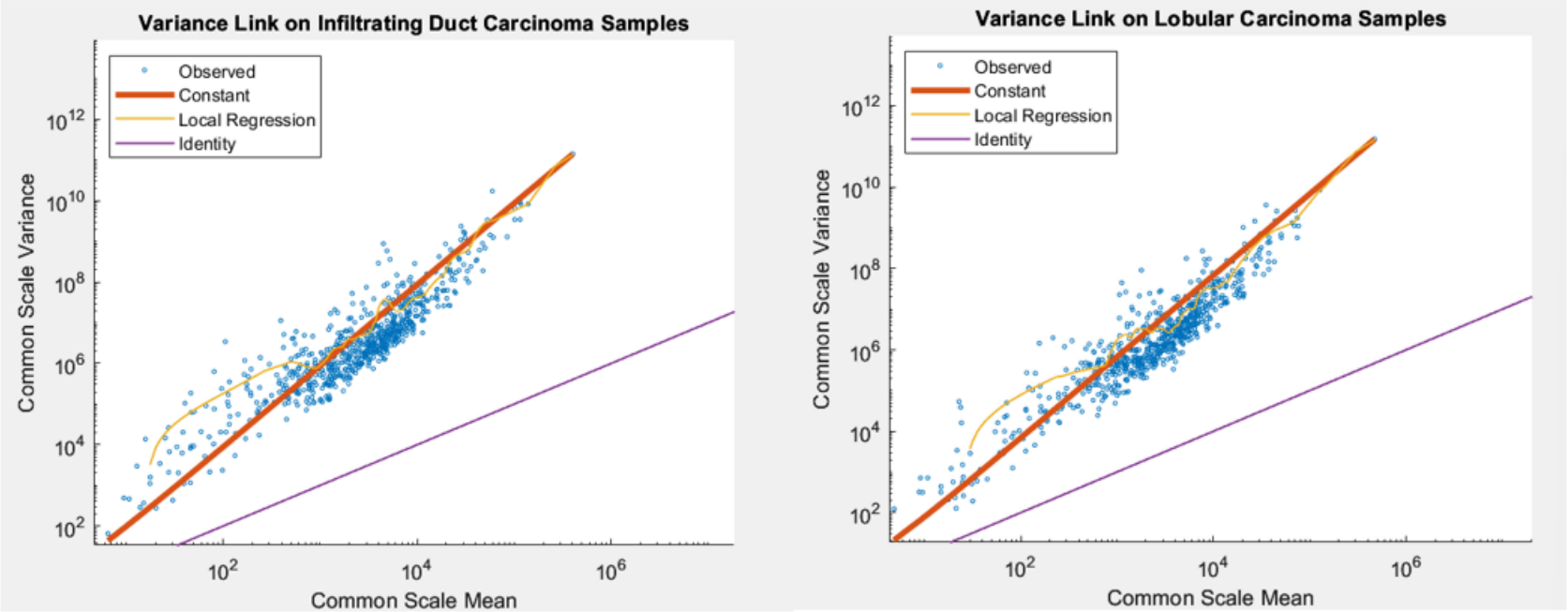
I dati rappresentati in rosso corrispondono ai geni differenzialmente espressi nelle due classi: i geni up-regolati sono quelli in cui il $log2FC > 1$ e i geni down-regolati sono quelli in cui il $log2FC < -1$.

WORKFLOW 1

Inferenza dell'espressione differenziale con un modello binomiale negativo

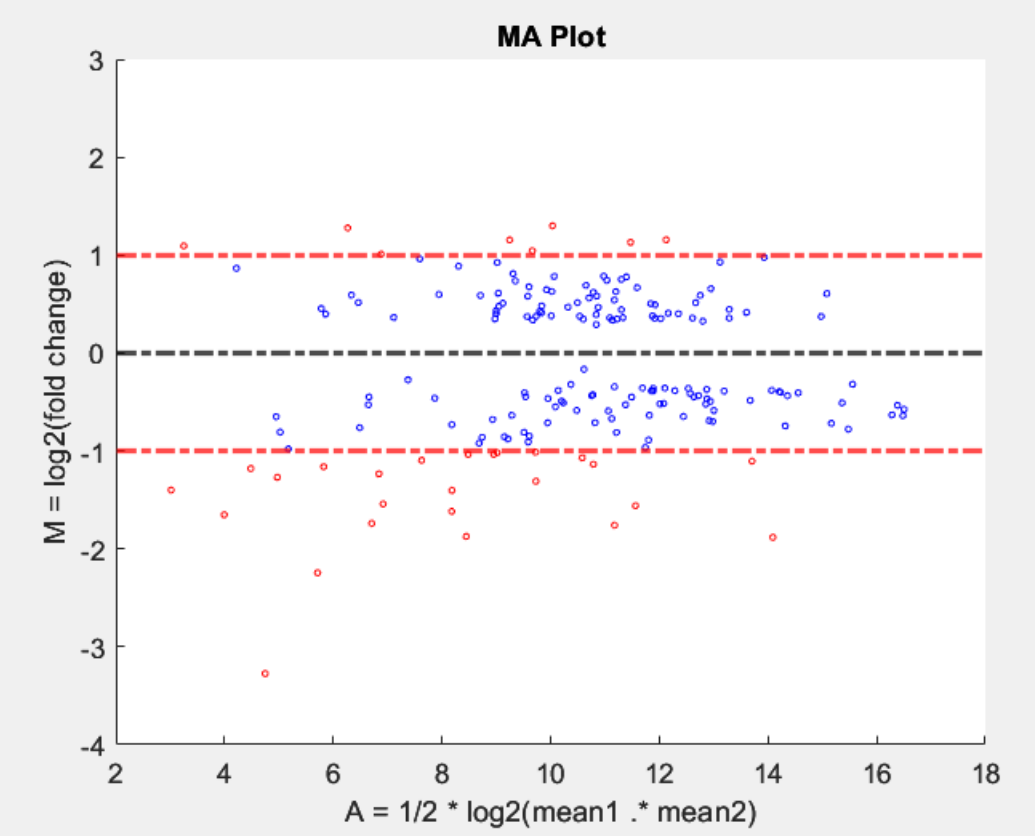
La distribuzione che è stata dimostrata meglio modellare la variabilità sperimentale dei geni è la distribuzione binomiale negativa. In questa distribuzione la varianza σ^2 è funzione della media μ e dipende da un parametro φ detto dispersione: $\sigma^2 = \mu + \varphi\mu^2$.

L' nbintest permette di identificare statisticamente quali sono i **geni differenzialmente espressi** tra due condizione (Infiltrating Ductal Carcinoma/ Lobular Carcinoma), permettendo di effettuare una **prima riduzione delle feature in modo significativo**.



Con 'VarianceLink' si specifica la relazione tra media e varianza → utilizziamo VarianceLink' = **Constant**

WORKFLOW 1




```
% find up-regulated genes
up = geneTableSig.log2FC > 1;
upGenes = sortrows(geneTableSig(up,:),'log2FC','descend');
numberSigGenesUp = sum(up);
fprintf('There are %d Up-regulated genes\n', numberSigGenesUp);

% find down-regulated genes
down = geneTableSig.log2FC < -1;
downGenes = sortrows(geneTableSig(down,:),'log2FC','ascend');
numberSigGenesDown = sum(down);
fprintf('There are %d Down-regulated genes\n', numberSigGenesDown);
```

```
There are 183 significant genes on 682
There are 8 Up-regulated genes
There are 25 Down-regulated genes
```

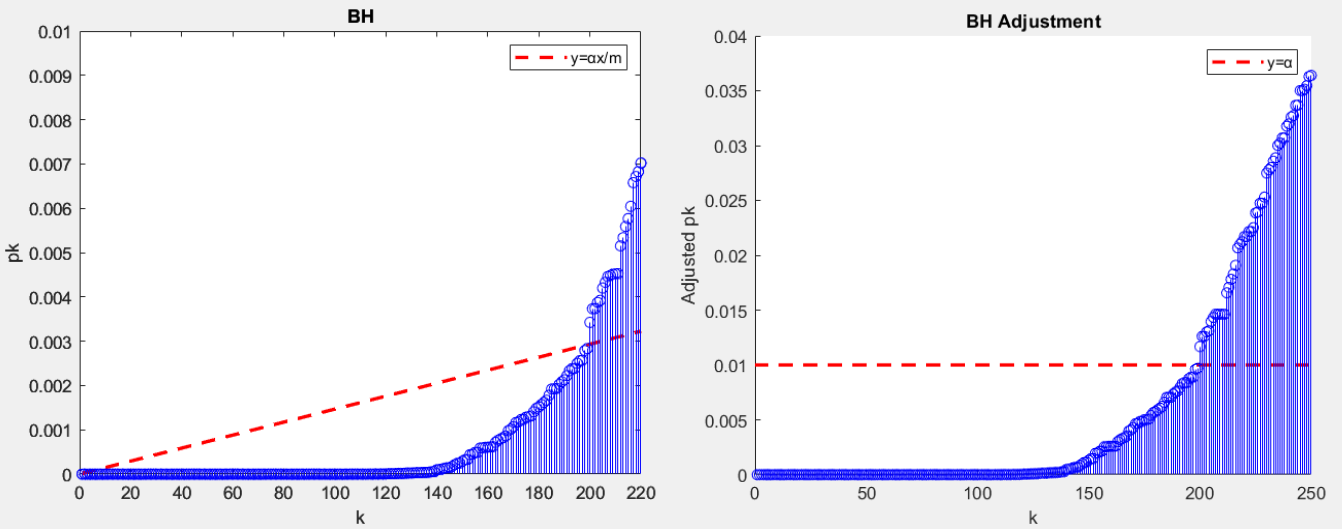
WORKFLOW 1

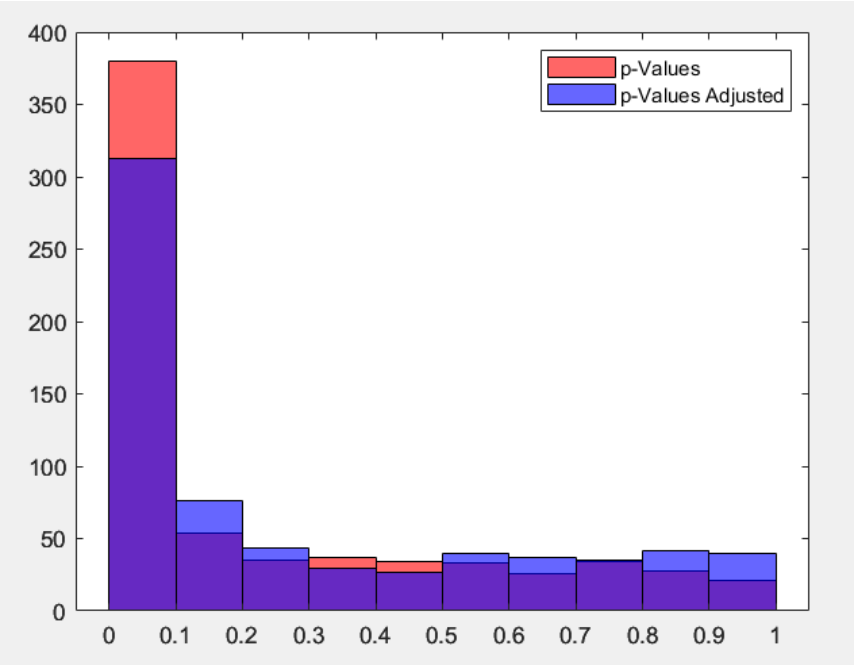
Aggiustamento dei p-Values

Durante l'esecuzione di un gran numero di test simultanei, la probabilità di ottenere un risultato significativo semplicemente a causa del caso aumenta con il numero di test.

Applichiamo la correzione di Benjamini – Hochberg →

```
padj = mafdr(tConstant.pValue,'BHFDR',true,'Showplot',true);
```





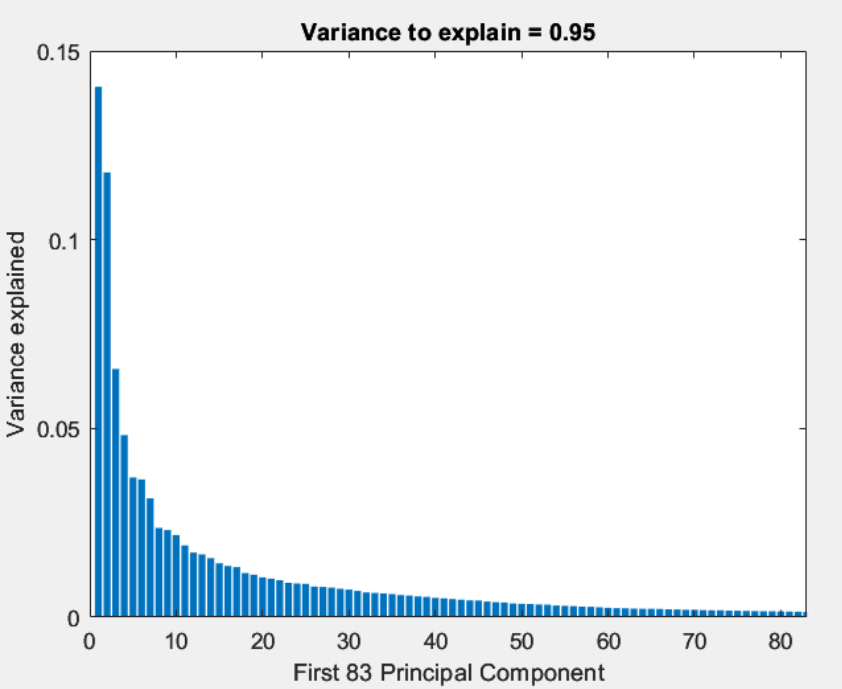
Viene determinato il valore K tale che tutti i p-value sotto la curva di coefficiente angolare α/m corrispondono alle ipotesi verificate. Ovvero tutti i p-Values al di sotto della linea rossa sono associati ai geni significativi.
Il valore α (livello di significatività) è fissato a 0.01.

WORKFLOW 1

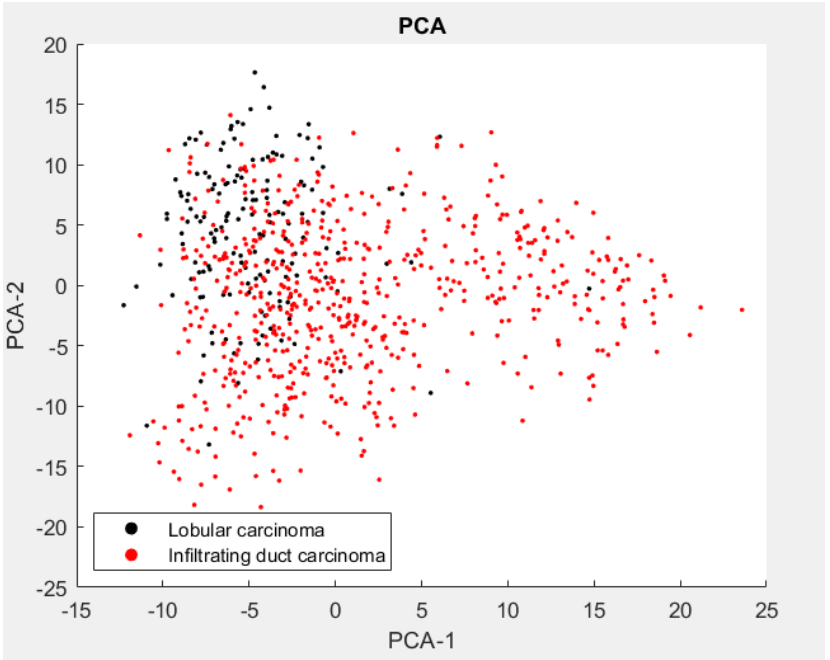
PCA

A questo punto è possibile scegliere se ridurre ulteriormente il campione implementando la PCA oppure se mantenere il campione contenente i geni significativi. Nel caso il cui viene scelto di ridurre il campione è necessario porre la variabile doPCAgenisig=true mentre se si sceglie di non ridurre poniamo la variabile doPCAgenisig= false.

Fissata la percentuale di varianza pari al 95% e data in input la matrice dei counts con i geni significativi, normalizzata con il log2, si ottiene un numero di Principal Component pari a 83.




```
For preserving 95.00% of variance, you have to use 83 PC
```



WORKFLOW 1

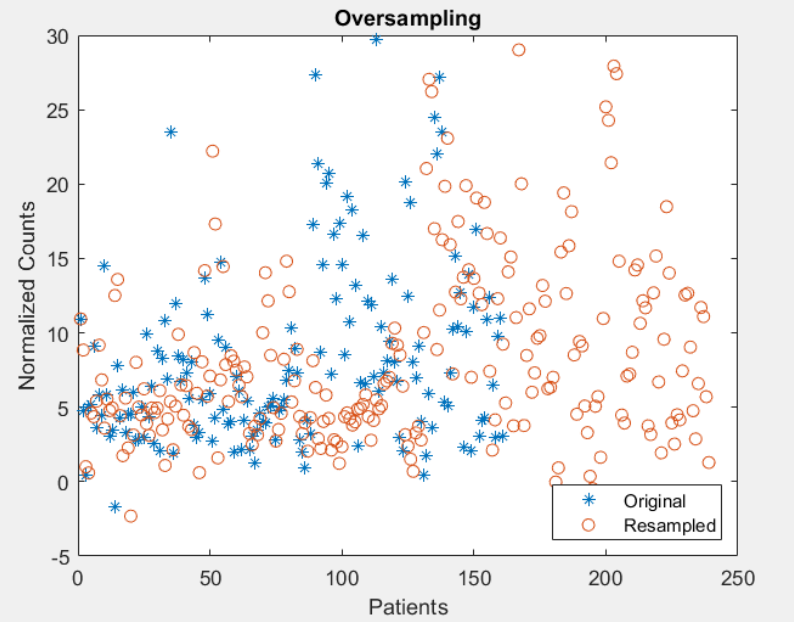
BILANCIAMENTO

Su 770 pazienti totali abbiamo che:

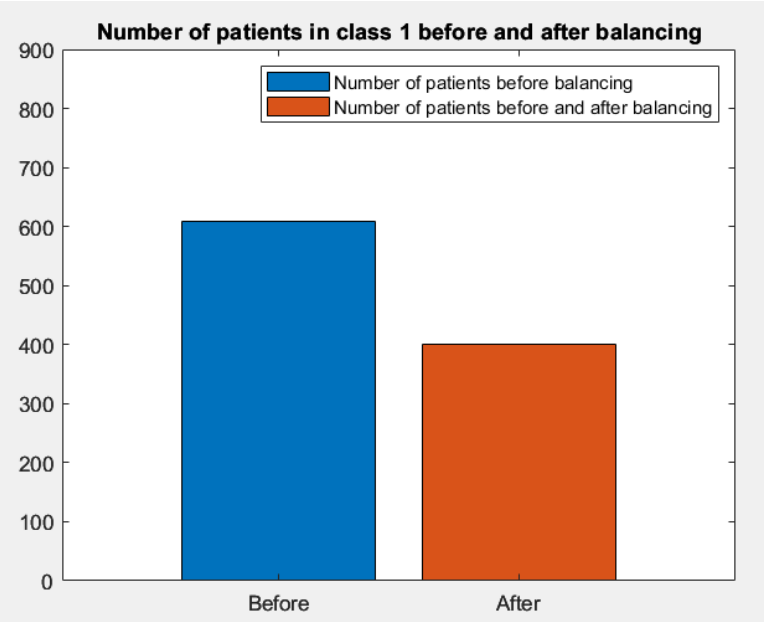
- 609 pazienti appartenenti alla classe Infiltrating duct carcinoma, NOS
- 161 pazienti appartenenti alla classe Lobular carcinoma, NOS

Con il bilanciamento entrambe le classi presentano un numero di pazienti pari a 400 per un totale di 800 pazienti.

Funzione resample → Aggiunta di 239 campioni fittizi per la classe Lobular carcinoma



Funzione datasample → Riduzione dei pazienti della classe Infiltrating duct carcinoma da 609 a 400



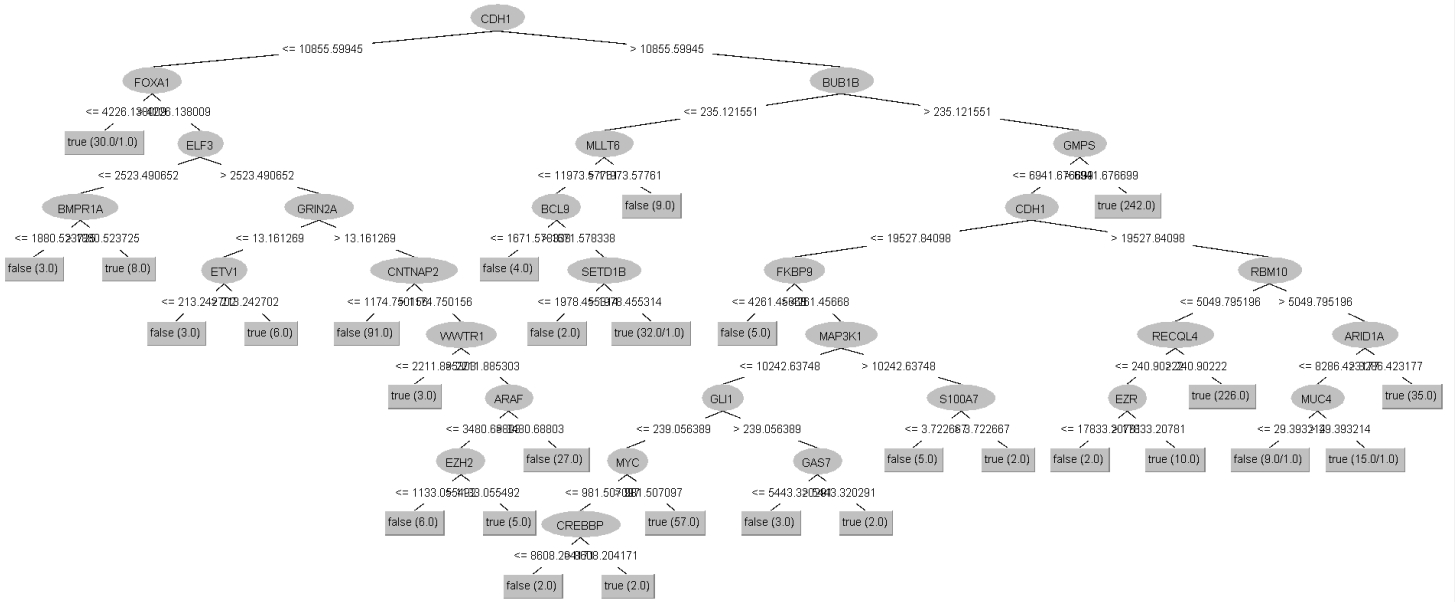
SISTEMA A REGOLE (WEKA)

La classificazione con un sistema a regole è un metodo di classificazione in cui un insieme di regole viene utilizzato per assegnare una classe ad un'istanza di dati. Questo metodo di classificazione è anche noto come sistema di classificazione a regole d'inferenza.

Utilizzando il dataset normalizzato con il size factor e selezionando come output la colonna ‘primary diag’ che contiene valori di verità true/false è stato prodotto il seguente albero decisionale selezionando come algoritmo “J48 Algorithm for Decision Tree” :

Il nodo radice è rappresentato dal gene **CDH1** e la discriminazione avviene sulla base dei valori di conteggio.

A seguito dell’analisi genica differenziale del workflow1, dopo aver individuato i geni significativi ponendo, per ogni gene, $padj < 0.01$ e ordinando in ordine crescente i geni significativi in base al valore del padj, vediamo che anche in questo caso il gene CDH1 è posto in cima alla lista.



SISTEMA A REGOLE (WEKA)

I risultati dell'albero decisionale mostrano che le istanze classificate correttamente sono circa l'87% (valore di accuratezza) mentre quelle classificate erroneamente sono circa il 12%. La AUC (area sotto la curva ROC) è pari all'81%.

Correctly Classified Instances	741	87.5887 %
Incorrectly Classified Instances	105	12.4113 %
Kappa statistic	0.6193	
Mean absolute error	0.125	
Root mean squared error	0.3414	
Relative absolute error	38.3727 %	
Root relative squared error	84.6327 %	
Total Number of Instances	846	

Su 683 pazienti appartenenti alla classe 1 vengono classificati correttamente 620 pazienti mentre su 173 pazienti appartenenti alla classe 2 vengono classificati correttamente 121 pazienti. È visibile come l'algoritmo sbaglia maggiormente sui falsi positivi che sono 52 su 173 totali piuttosto che sui falsi negativi che sono 53 su 683 totali.

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,921    0,301    0,923      0,921    0,922      0,619    0,813     0,915     true
                 0,699    0,079    0,695      0,699    0,697      0,619    0,813     0,581     false
Weighted Avg.    0,876    0,255    0,876      0,876    0,876      0,619    0,813     0,846

=== Confusion Matrix ===

   a   b   <-- classified as
 620  53 |   a = true
  52 121 |   b = false
```

SISTEMA A REGOLE (WEKA)

Andando a rimuovere dalla matrice il gene CDH1 vediamo che le prestazioni della rete risultano peggiori rispetto al caso precedente:

```
Correctly Classified Instances         520               80.1233 %
Incorrectly Classified Instances       129               19.8767 %
Kappa statistic                          0.2968
Mean absolute error                      0.1999
Root mean squared error                  0.4402
Relative absolute error                 68.3804 %
Root relative squared error            115.2877 %
Total Number of Instances              649

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,888    0,600    0,873      0,888    0,880      0,297    0,673     0,879     true
                 0,400    0,112    0,434      0,400    0,416      0,297    0,673     0,316     false
Weighted Avg.    0,801    0,514    0,795      0,801    0,798      0,297    0,673     0,780

=== Confusion Matrix ===

   a   b   <-- classified as
 474  60 |   a = true
  69  46 |   b = false
```

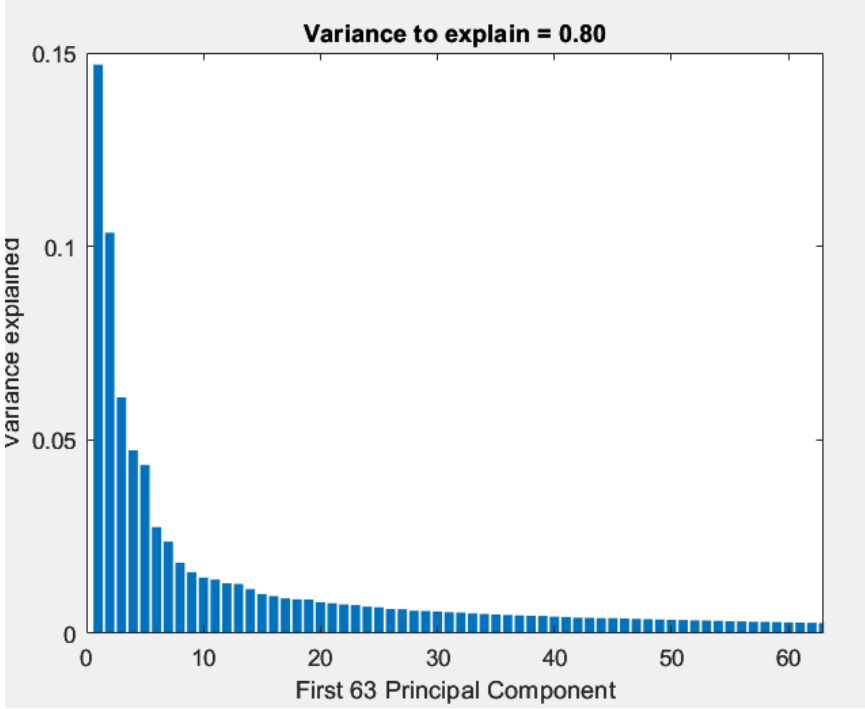
Il nodo radice è rappresentato dal gene GMPS che è presente sempre nella lista di geni significativi trovati nel workflow1.

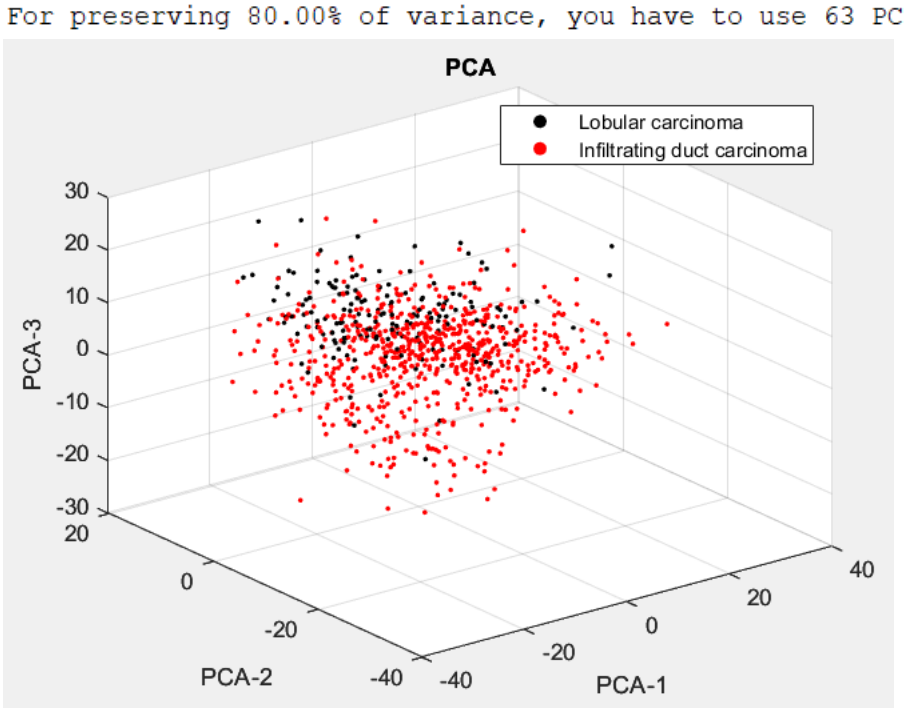
WORKFLOW 2

PCA

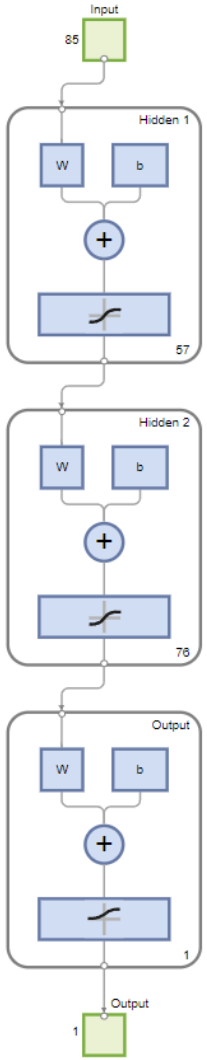
Partendo dai dati normalizzati prima con il Size Factor e in seguito con il log2 effettuiamo la PCA, ovvero una tecnica di riduzione della dimensionalità che consente di eseguire un mapping dallo spazio iniziale ad uno spazio di dimensione inferiore.

La PCA è utilizzata per ridurre il numero di features. Ogni Principal Component spiega una percentuale di varianza del dataset e si sceglie il numero minimo di PC che ci permette di preservare la percentuale di varianza fissata a 80%. Si ottengono 63 PC.





CLASSIFICATORE BINARIO- RETE NEURALE ARTIFICIALE



Una **rete neurale artificiale** è un algoritmo di machine learning utilizzato per la classificazione binaria.

È composta da più strati di neuroni artificiali, ognuno dei quali elabora i dati in ingresso e produce un output. Gli strati di neuroni sono fully connected, ovvero tutti i neuroni sono connessi a tutti i neuroni degli strati precedenti e successivi.

La rete neurale viene addestrata utilizzando un insieme di dati di addestramento. Durante il processo di addestramento, la rete neurale fa previsioni sulle etichette di classe (positivo o negativo) per i dati di addestramento, e la precisione delle previsioni viene valutata utilizzando una funzione di perdita.

Il numero dei neuroni negli hidden layers è stato calcolato attraverso l'algoritmo genetico massimizzando la funzione di fitness posta uguale all'accuratezza del test.

```
%RETE NEURALE ARTIFICIALE
hiddenLayer= GA_ANN(i,x_train{i},x_test{i},t_train{i},t_test{i})
%hiddenLayer=[5 10];
trainFcn = 'traingdx';
performFcn = 'crossentropy';
net = patternnet(hiddenLayer,trainFcn,performFcn);
net.layers{end}.transferFcn = 'logsig';
net.divideFcn = 'dividerand';
net.divideParam.trainRatio = 0.75;
net.divideParam.valRatio = 0.25;
net.divideParam.testRatio = 0.00;
net = configure(net,x_train{1},t_train{1});
net.trainParam.epochs = 500;
net.trainParam.lr = 1e-3;
net.trainParam.max_fail = 100;
net=init(net);
```

ALGORITMO GENETICO PER LA RICERCA DEGLI HIDDEN LAYERS

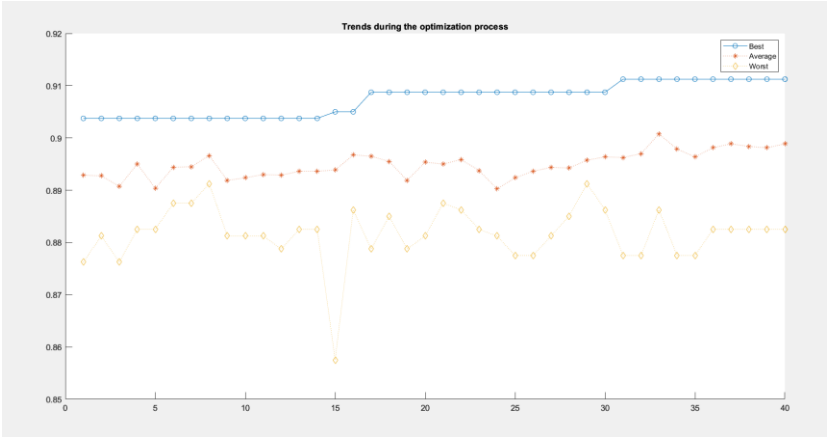
L'algoritmo genetico si basa sull'idea della competizione tra individui di una popolazione in un ambiente dalle risorse limitate che quindi causa la selezione naturale del più "adatto" tra gli individui, secondo una funzione di fitness.

I meccanismi implementati sono:

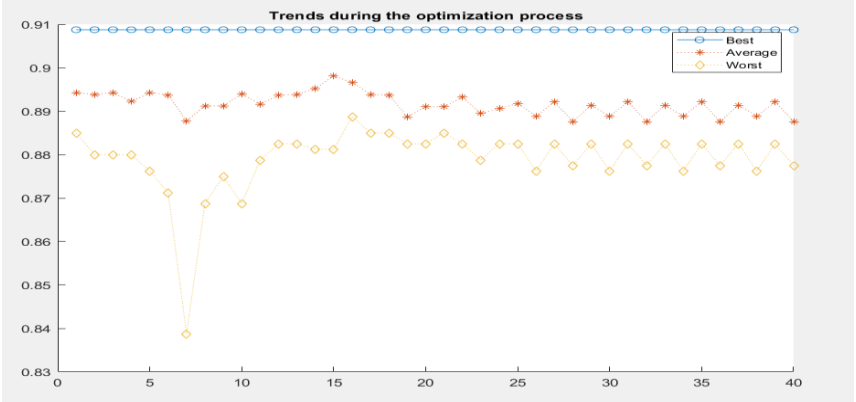
- Selezione della specie
- Crossover a singolo punto
- Mutazione puntiforme

Parametri:

- Pop_size= 10
- Epoche= 40
- Cromosoma fatto da 2 geni rispettivamente da 7 bit
- Pm= 0.2
- L'individuo migliore va in automatico nella pop successiva
- Funzione di fitness → accuratezza calcolata sul test set




```
miglior individuo trovato all'epoca 31
hiddenlayer=    57    76
```




```
miglior individuo trovato all'epoca 1
hiddenlayer=    55    98
```

CLASSIFICATORE BINARIO- REGRESSIONE LOGISTICA

Relazione ipotizzata per la **Regressione Logistica**:

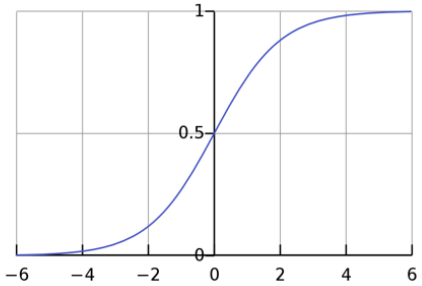
$$h_\theta(\boldsymbol{x}) = g(\theta_0 + x_1\theta_1 + x_2\theta_2 + \cdots + x_n\theta_n) = g(X\theta)$$

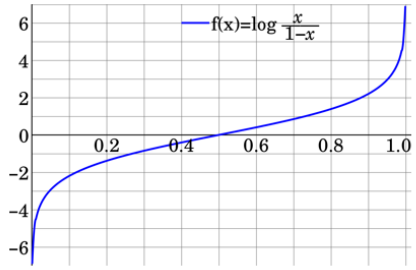
Dove $g(z)$ è una funzione logistica (una sigmoide) con funzione inversa $f(x)$

Nota che $g(z)$: $\mathbb{R}\rightarrow$ (0,1), mentre $f(x)$:(0,1) $\rightarrow\mathbb{R}$.

$$g(z) = \frac{1}{1+e^{-z}}$$

$$f(x) = \log\left(\frac{x}{1-x}\right)$$

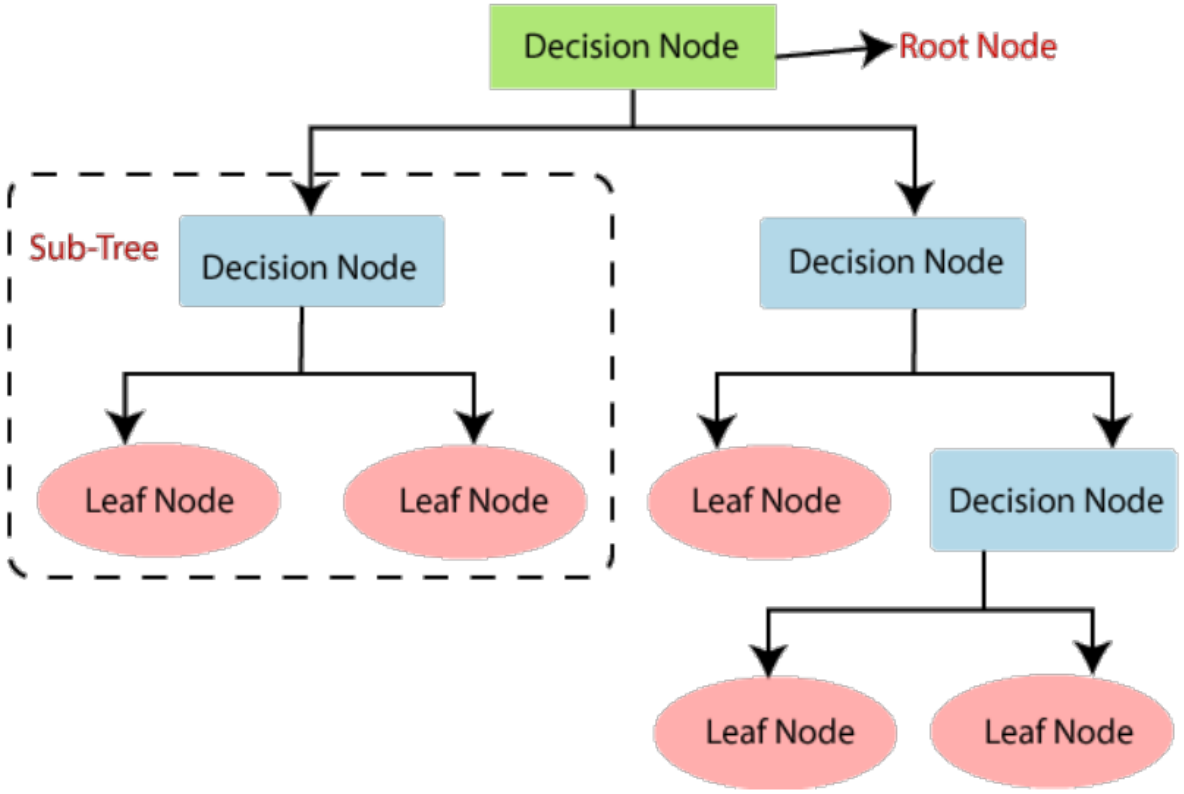




La funzione logistica restituisce un valore compreso tra 0 e 1 che viene utilizzato per assegnare una probabilità all'appartenenza di un'osservazione a una determinata categoria. Questa probabilità viene quindi utilizzata per prevedere la classe di una nuova osservazione.

CLASSIFICATORE BINARIO- ALBERO DECISIONALE

Un albero decisionale è uno strumento di supporto alle decisioni che utilizza un modello ad albero delle decisioni e delle loro possibili conseguenze, compresi i risultati degli eventi casuali, i costi delle risorse e l'utilità. È un modo per visualizzare un algoritmo che contiene solo istruzioni di controllo condizionale.



Un albero decisionale è una struttura simile a un diagramma di flusso in cui ogni nodo interno rappresenta un "test" su un attributo (ad esempio, se il lancio di una moneta esce testa o croce), ogni ramo rappresenta il risultato del test e ogni nodo foglia rappresenta un'etichetta di classe (decisione presa dopo aver calcolato tutti gli attributi). I percorsi dalla radice alla foglia rappresentano regole di classificazione (o regressione).

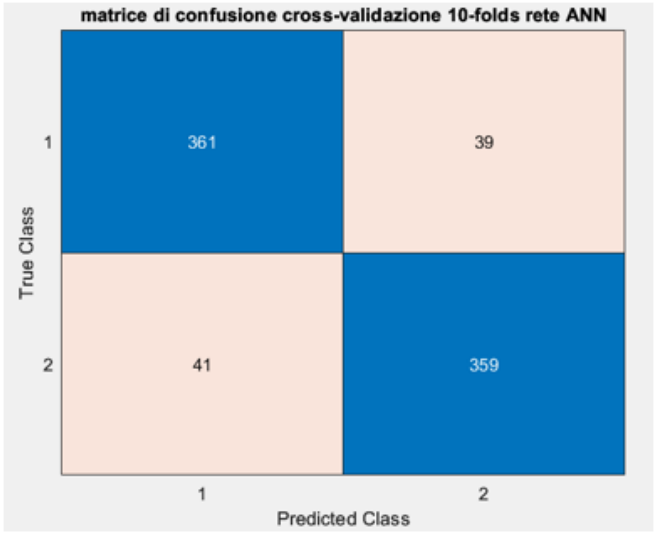
RISULTATI - WORKFLOW 1

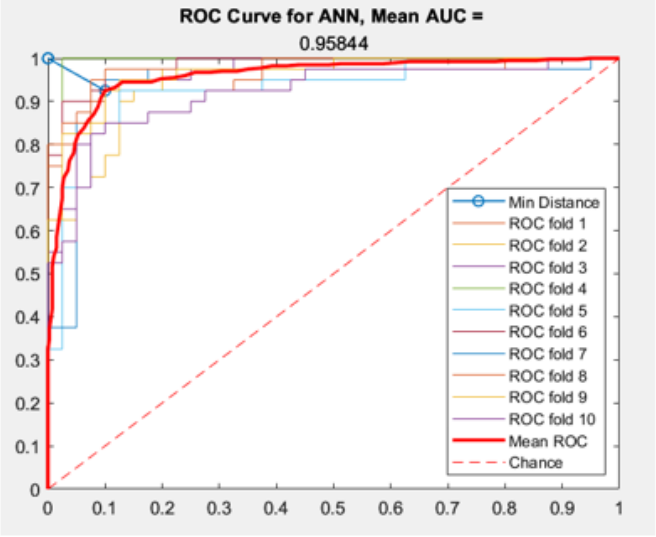
L'addestramento è fatto su una matrice (Xs) 85x800, in cui 800 sono i casi (400 positivi + 400 negativi) e 85 sono le features (83 geni significativi con PCA + 2 features cliniche, ovvero razza ed età alla diagnosi).

È stata implementata la cross validazione con k=10.

I risultati riportati per la ANN sono quelli ottenuti applicando l'algoritmo genetico per il calcolo degli hidden layers.

RETE NEURALE ARTIFICIALE



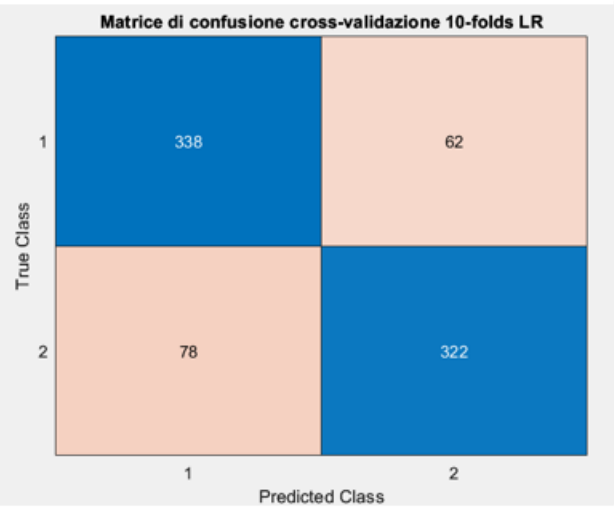


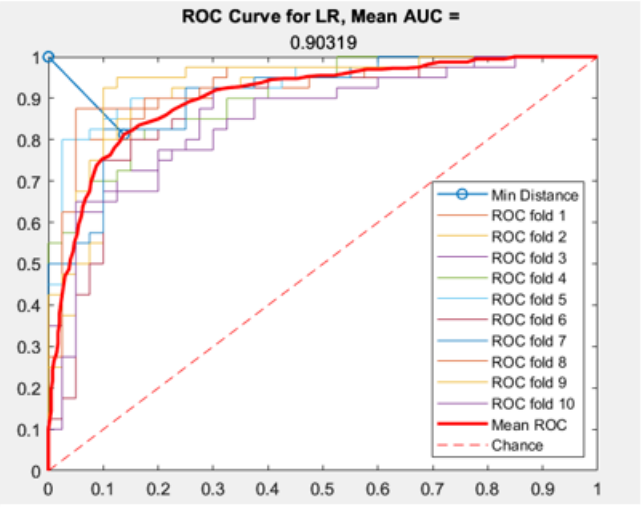
I risultati sono:

```
Risultati rete neurale artificiale con cross-validazione 10-folds:
Accuratezza media: 90.00
Precisione media: 90.37
Miss Rate media: 10.25
Recall media: 89.75
Area media sottesa alla curva roc: 95.84

Il best cut-off per l'ANN è: 0.45
Si ottiene in corrispondenza della cordinata x: 0.10 e della coordinata y:0.93
```

RISULTATI - WORKFLOW 1

REGRESSIONE LOGISTICA



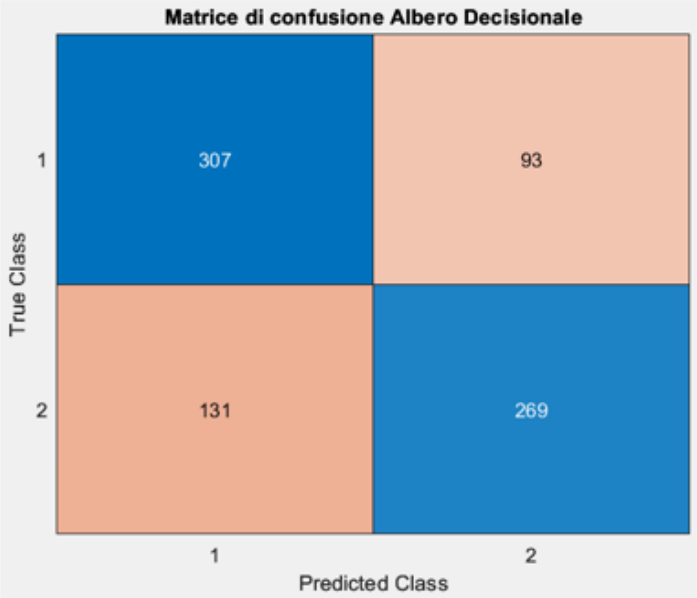


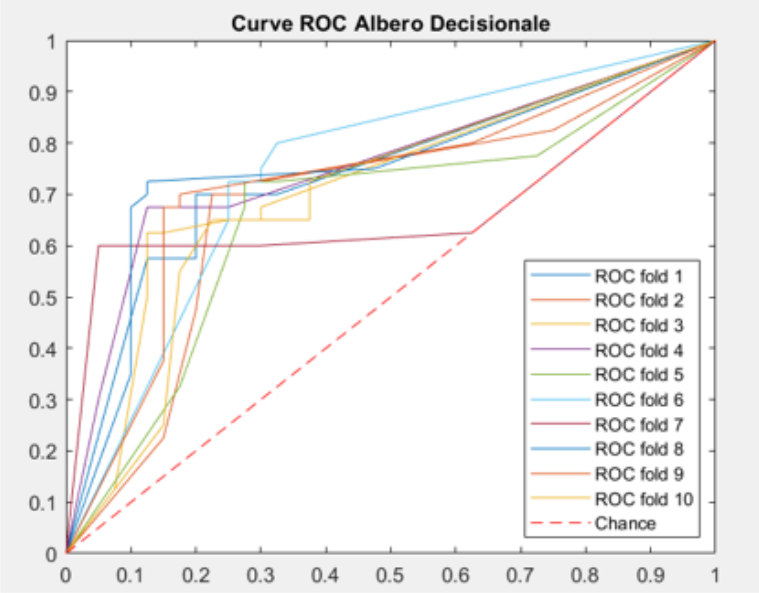
I risultati sono:

```
Risultati regressione logistica con cross-validazione 10-folds:
Accuratezza media: 82.50
Precisione media: 83.99
Miss Rate media: 19.50
Recall media: 80.50
Area media sottesa alla curva roc: 90.32

Il best cut-off per la LR è: 0.52
Si ottiene in corrispondenza della cordinata x: 0.14 e della coordinata y:0.81
```

RISULTATI - WORKFLOW 1

ALBERO DECISIONALE



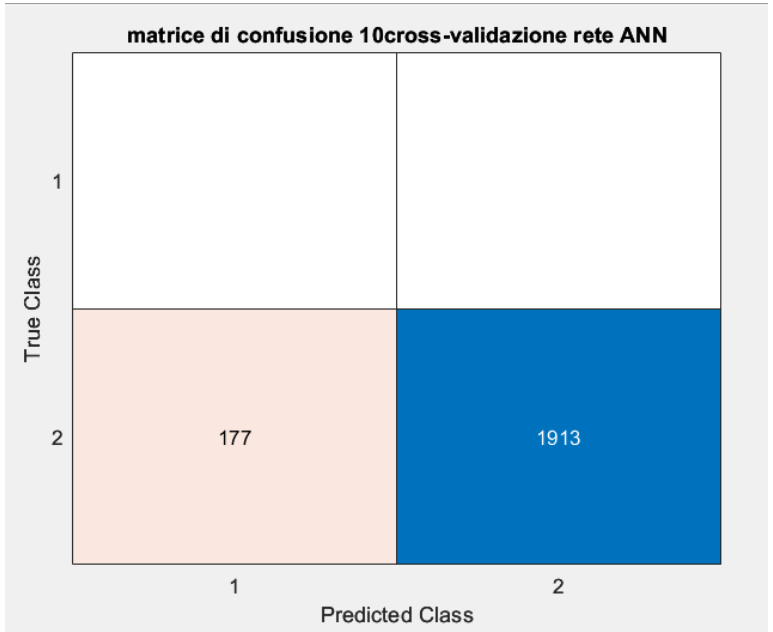


I risultati sono:

```
Risultati Albero Decisionale con cross-validazione 10-folds:
Accuratezza media Albero Decisionale: 72.00
Precisione media Albero Decisionale: 74.67
Miss Rate media Albero Decisionale: 32.75
Recall media Albero Decisionale: 67.25
Area media sottesa alla curva roc Albero Decisionale: 71.34
```

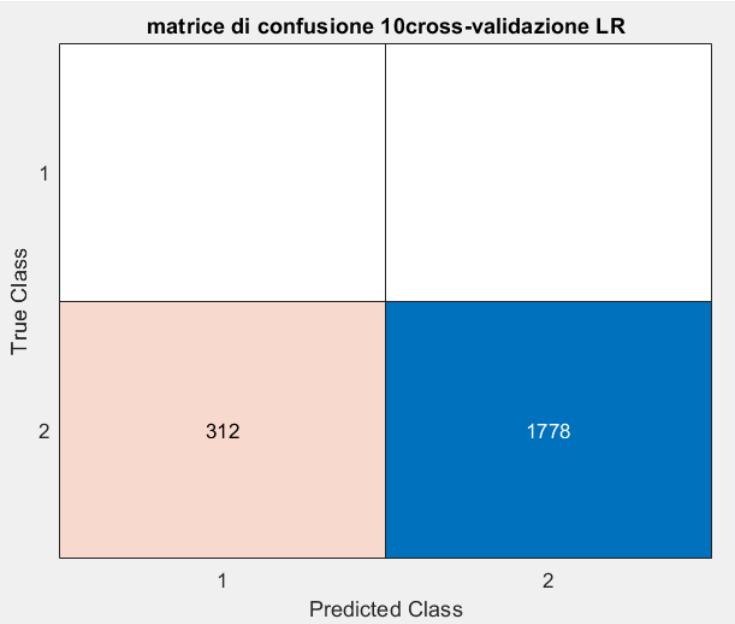
RISULTATI - WORKFLOW 1 (209 valori esclusi durante il bilanciamento)

RETE NEURALE



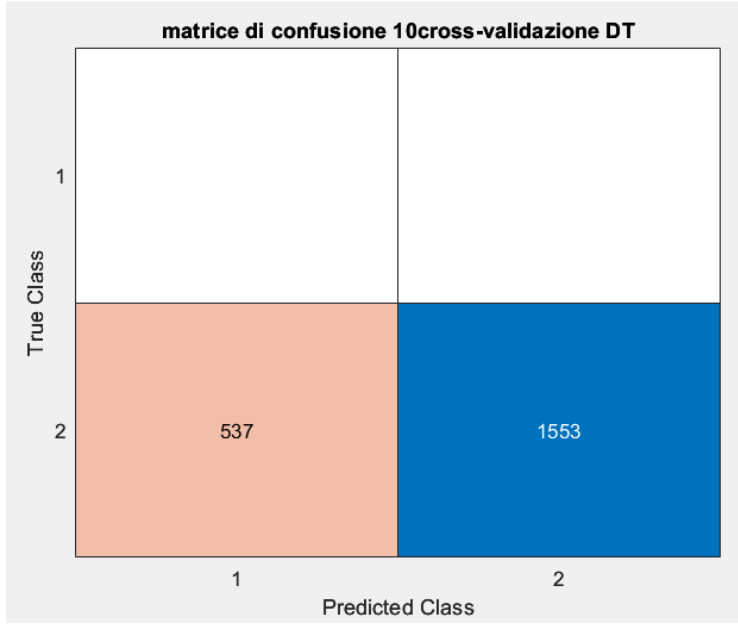
Risultati rete neurale artificiale con 10 cross-validazioni:
Accuratezza media: 91.53
Precisione media: 100.00
Miss Rate media: 8.47
Recall media: 91.53

REGRESSIONE LOGISTICA



Risultati regressione logistica con 10 cross-validazioni:
Accuratezza media: 85.07
Precisione media: 100.00
Miss Rate media: 14.93
Recall media: 85.07

ALBERO DECISIONALE



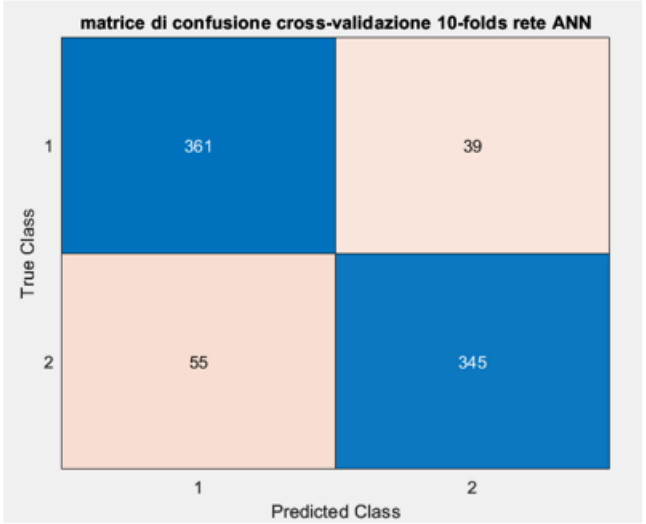
Risultati albero decisionale con 10 cross-validazioni:
Accuratezza media: 74.31
Precisione media: 100.00
Miss Rate media: 25.69
Recall media: 74.31

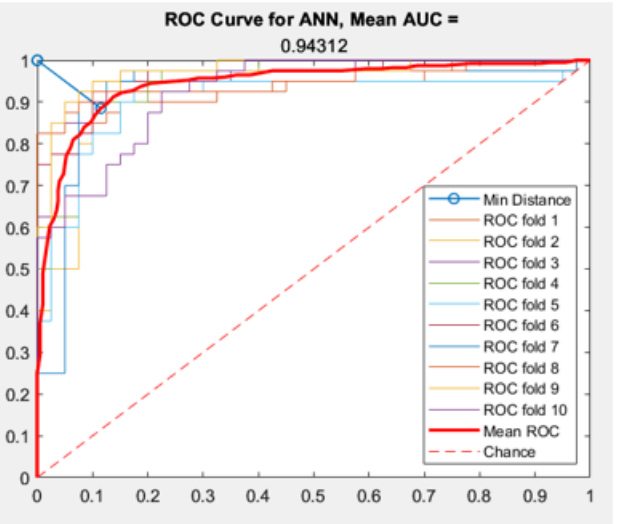
RISULTATI - WORKFLOW 2

Abbiamo riportato i risultato considerando il campione bilanciato. L'addestramento è fatto su una matrice 67x800, in cui 800 sono i casi (400 positivi + 400 negativi) e 67 sono le features (65 componenti della PCA + 2 features cliniche). I risultati riportati per la ANN sono quelli ottenuti applicando a hidden layers l'algoritmo genetico.

È stata implementata la cross validazione con k=10.

RETE NEURALE ARTIFICIALE



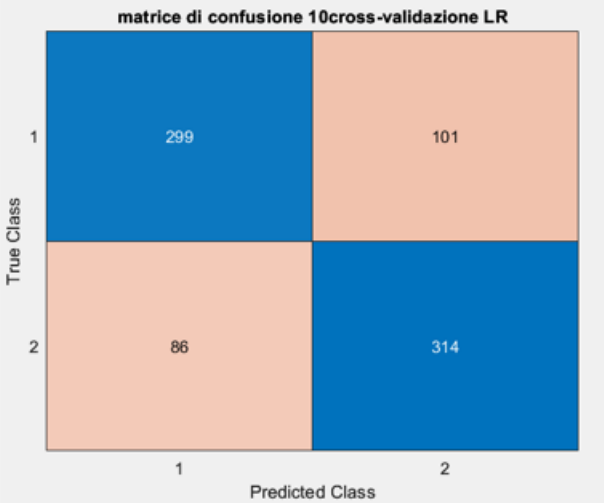


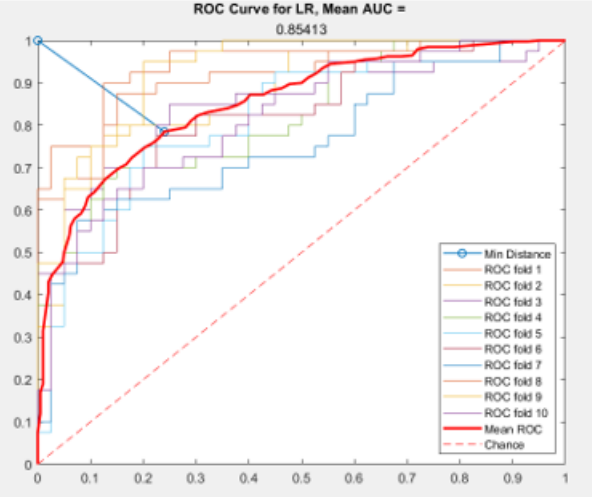
I risultati sono:

```
Risultati rete neurale artificiale con cross-validazione 10-folds:
Accuratezza media: 88.25
Precisione media: 90.02
Miss Rate media: 13.75
Recall media: 86.25
Area media sottesa alla curva roc: 94.31

Il best cut-off per l'ANN è: 0.44
Si ottiene in corrispondenza della cordinata x: 0.11 e della coordinata y:0.89
```

RISULTATI - WORKFLOW 2

REGRESSIONE LOGISTICA





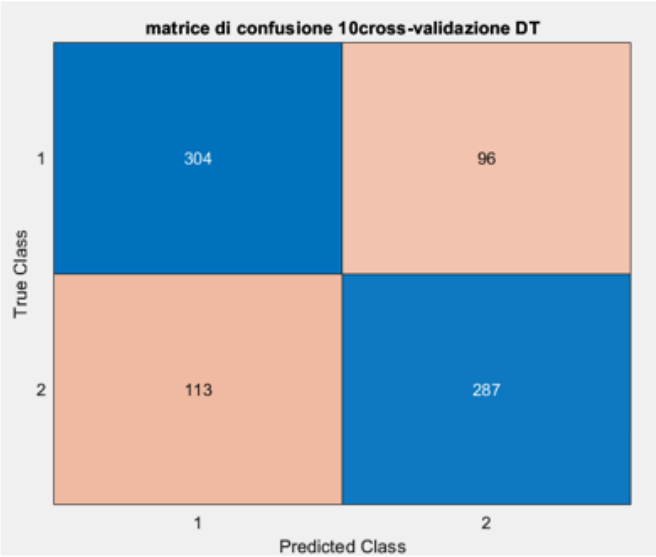
I risultati sono:

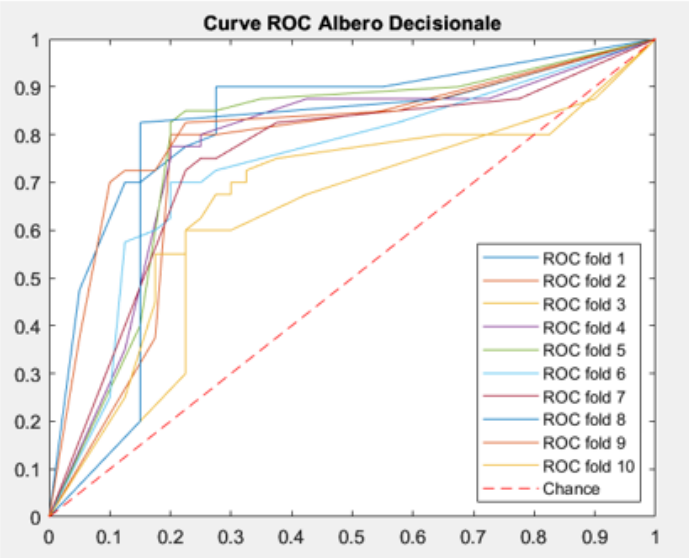
```
Risultati regressione logistica con 10 cross-validazioni:
Accuratezza media: 76.62
Precisione media: 76.04
Miss Rate media: 21.50
Recall media: 78.50
Area media sottesa alla curva roc: 85.41
```

```
Il best cut-off per la LR è: 0.51
Si ottiene in corrispondenza della cordinata x: 0.24 e della coordinata y:0.79
```

RISULTATI - WORKFLOW 2

ALBERO DECISIONALE



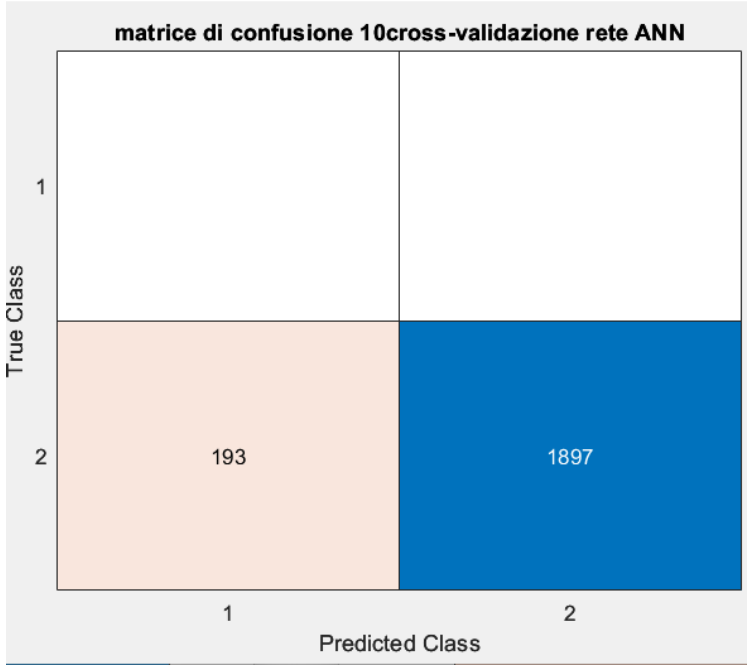


sultati sono:

```
Risultati albero decisionale con 10 cross-validazioni:
Accuratezza media: 77.50
Precisione media: 79.26
Miss Rate media: 24.00
Recall media: 76.00
Area media sottesa alla curva roc: 76.39
```

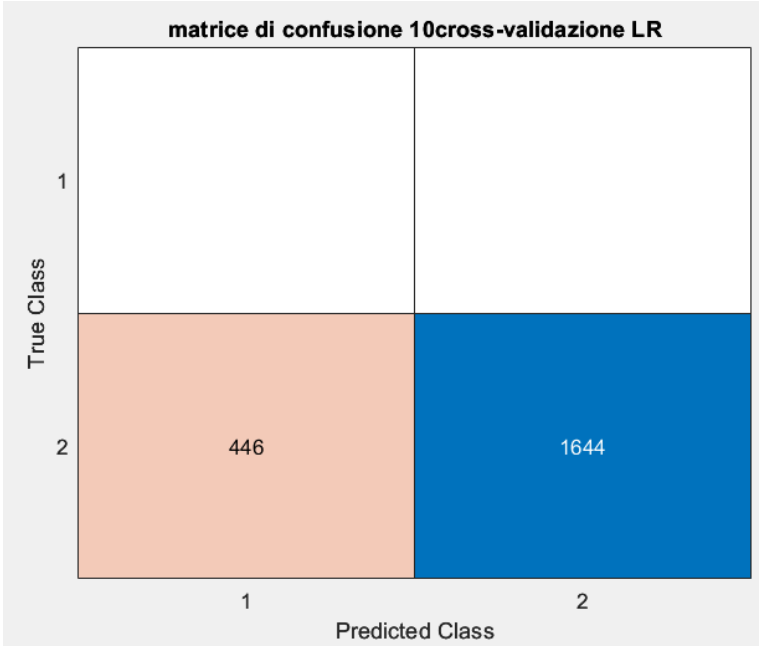
RISULTATI - WORKFLOW 2 (209 valori esclusi durante il bilanciamento)

RETE NEURALE



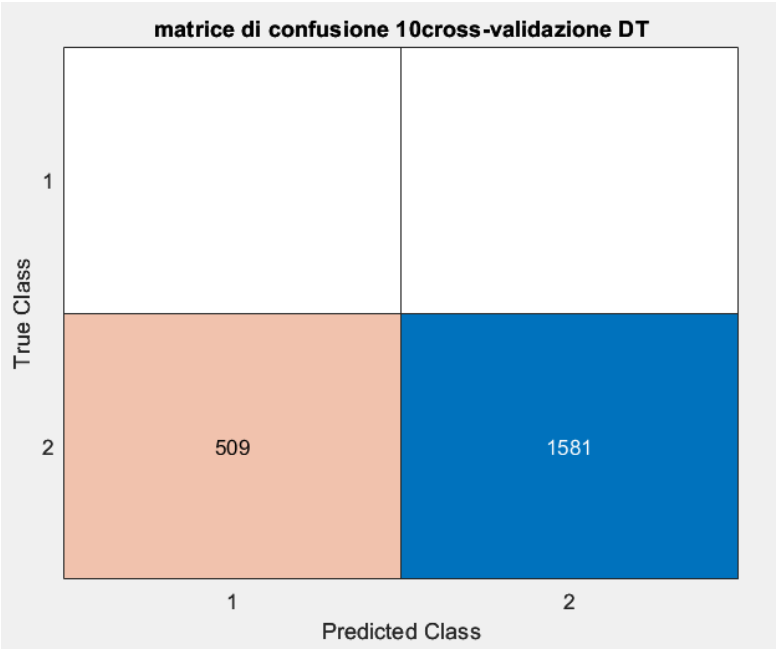

```
Risultati rete neurale artificiale con 10 cross-validazioni:
Accuratezza media: 90.77
Precisione media: 100.00
Miss Rate media: 9.23
Recall media: 90.77
```

REGRESSIONE LOGISTICA




```
Risultati regressione logistica con 10 cross-validazioni:
Accuratezza media: 78.66
Precisione media: 100.00
Miss Rate media: 21.34
Recall media: 78.66
```

ALBERO DECISIONALE




```
Risultati albero decisionale con 10 cross-validazioni:
Accuratezza media: 75.65
Precisione media: 100.00
Miss Rate media: 24.35
Recall media: 75.65
```

CONCLUSIONI - RISULTATI GENERALI

L'ANN è un modello di apprendimento automatico più complesso e flessibile rispetto alla regressione logistica e all'albero decisionale. La capacità di apprendere modelli non lineari complessi rende l'ANN una scelta ottima per l'analisi di dati biologici, dove le relazioni tra i geni possono essere molto complesse. In secondo luogo, l'ANN è particolarmente adatto per la classificazione di grandi quantità di dati.

Inoltre, la PCA riduce la dimensionalità del dataset, il che può migliorare le prestazioni del modello riducendo la complessità del problema.

Infine, il bilanciamento del campione può avere un effetto significativo sulle prestazioni del modello, poiché i modelli di apprendimento automatico spesso funzionano meglio quando le classi sono bilanciate.

Workflow 1	RETE ANN	REGRESSIONE LOGISTICA	ALBERO DECISIONALE
ACCURATEZZA MEDIA	90.00	82.50	72.00
PRECISIONE MEDIA	90.37	83.99	74.67
MISS RATE MEDIA	10.25	19.50	32.75
RECALL MEDIA	89.75	80.50	67.25
AUCROC	95.84	90.32	71.34

Workflow 2	RETE ANN	REGRESSIONE LOGISTICA	ALBERO DECISIONALE
ACCURATEZZA MEDIA	88.25	76.62	73.88
PRECISIONE MEDIA	90.02	76.04	75.16
MISS RATE MEDIA	13.75	21.50	28.25
RECALL MEDIA	86.25	78.50	71.75
AUCROC	94.31	85.41	74.58

GRAZIE PER L'ATTENZIONE