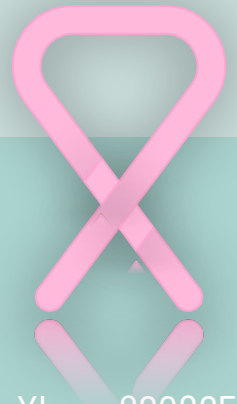
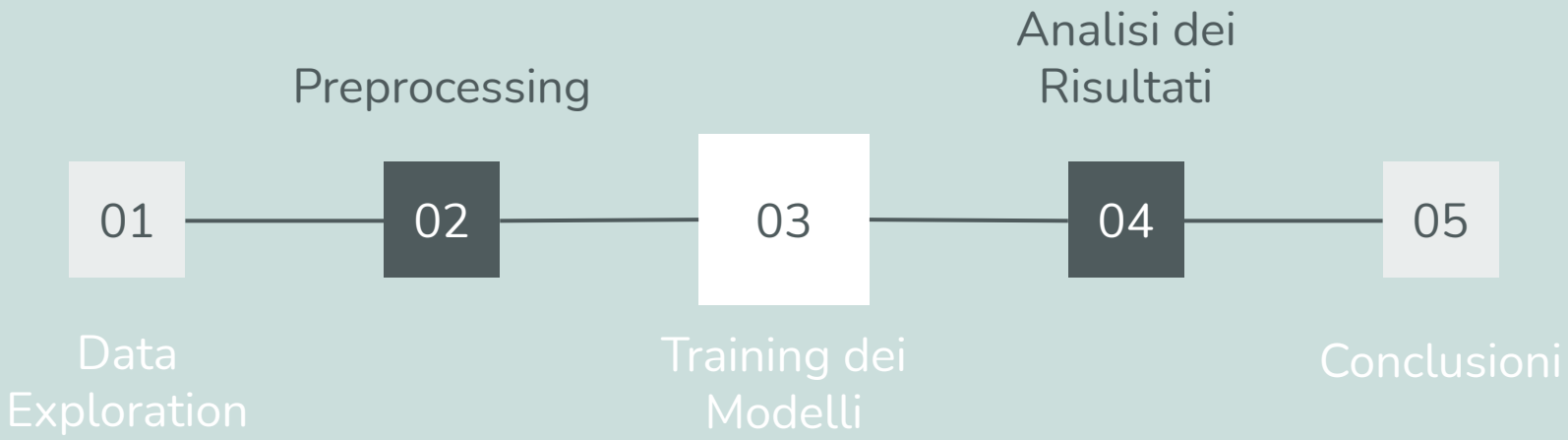
BREAST CANCER CLASSIFICATION



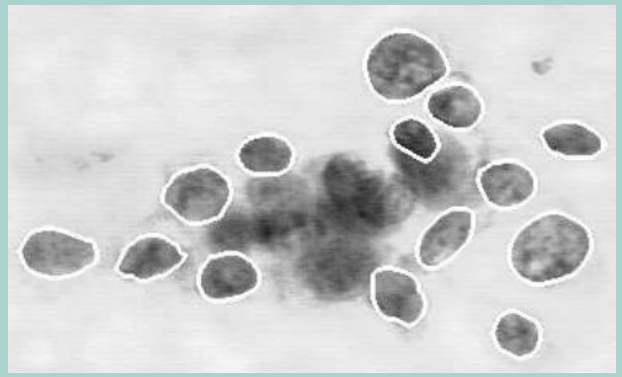
A.Xhyra 829865
M.Marino 829707
P.Tropeano 829757

Machine Learning, Febbraio 2021

Pipeline di lavoro



Breast Cancer Wisconsin (Diagnostic) Data Set



Composizione:

- 569 osservazioni
- 30 feature

*Contorni nuclei cellule tumorali ricavati tramite tecnica di active contour

Feature geometriche

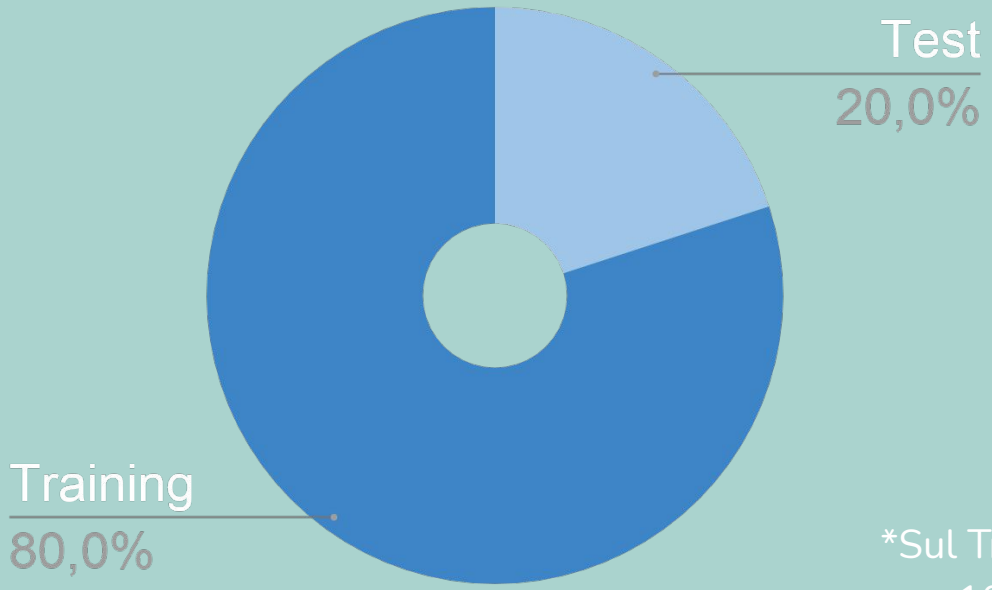
concave points, symmetry,
fractal dimension, radius, texture,
perimeter, area, smoothness,
compactness, concavity

mean value + standard error + worst value

62.7% B

37.3% M

Partizionamento dataset



*Sul Training set viene eseguita 5 volte una 10-fold cross-validation per diminuire la varianza dei risultati.

dataset.norm

Realizzazione dataset

➔ Normalizzazione feature

dataset.std

Realizzazione dataset

➔ Standardizzazione feature

dataset.pca

Realizzazione dataset

➔ Standardizzazione feature

➔ Proiezione sulle componenti principali

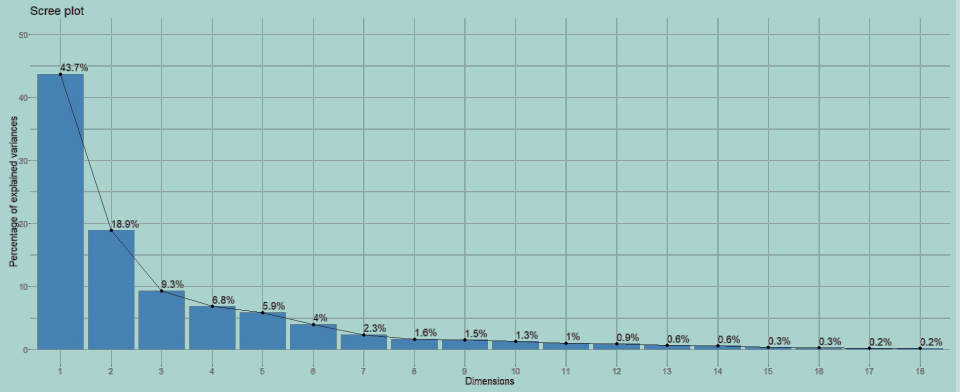
dataset.corr

Realizzazione dataset

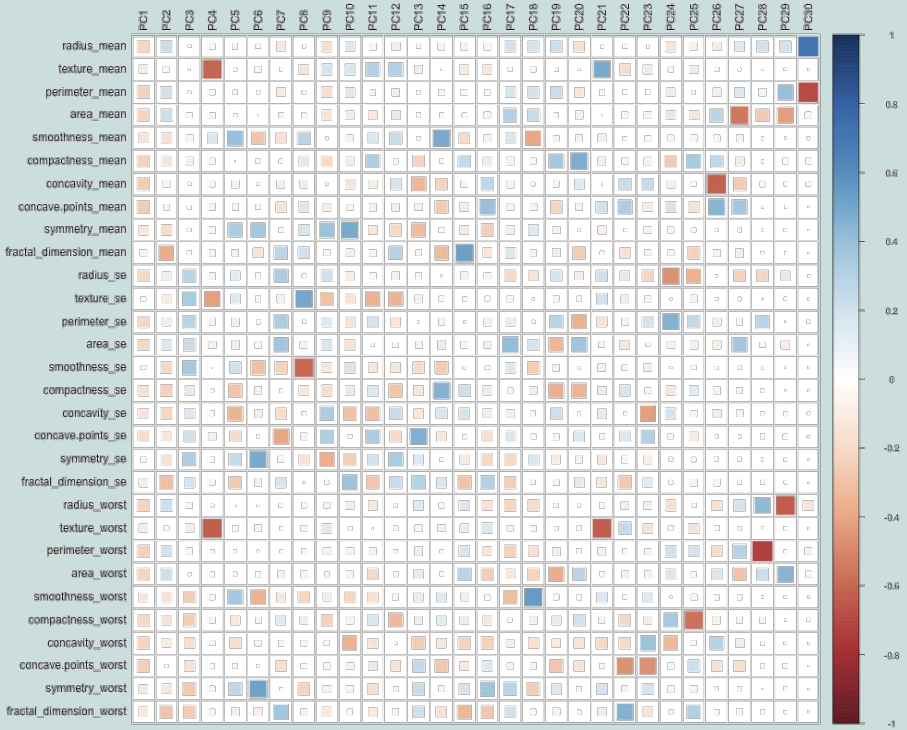
➔ Standardizzazione feature

➔ rimozione feature fortemente correlate

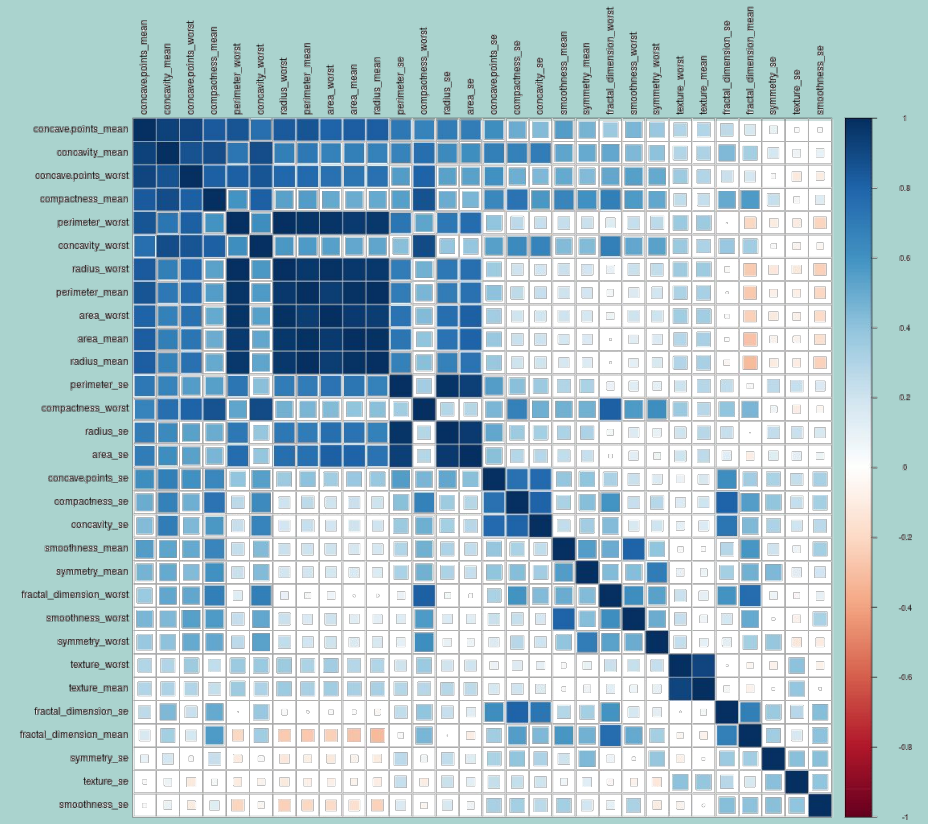
PCA



- 99% della varianza spiegata con 17 componenti principali.
- Le prime due componenti spiegano il 63% della varianza
- Le ultime 19 componenti spiegano meno dell'1% della varianza
- Alcune componenti sono correlate a specifiche feature.



Analisi di correlazione



Rimosse 13 feature fortemente correlate
(corr > 0.85)

concavity mean, concave.points mean, compactness mean, concave.points worst, concavity worst, perimeter worst, radius worst, perimeter mean, area worst, radius mean, perimeter se, area se, texture mean

Modelli

"There's no Free Lunch"

Naive Bayes

Baseline

- indipendenza stocastica feature;
- no tuning di iperparametri.

SVM

Complessità intermedia

- ricerca migliore iperpiano separatore;
- Grid Search per un totale di 40 combinazioni di iperparametri.

Neural Network

Maggiore complessità

- scarsa spiegabilità del modello;
- Grid Search per un totale di 48 combinazioni di iperparametri.

Tempi di training

Naive Bayes

	Tempo (s)
No parallelizzazione	3.3145 ±0.0841
Parallelizzazione	1.3865 ±0.1718

Con Bayes si eseguono due training: assumendo distribuzione normale delle feature e stimando la distribuzione.

Fase di cross-validation ≈ 1.7 s.

SVM

	Tempo (s)
No parallelizzazione	59.9235 ±0.4066
Parallelizzazione	10.5640 ±0.7849

Con SVM si testano tramite Grid Search 40 combinazioni di parametri.

Fase di cross-validation ≈ 1.5 s.

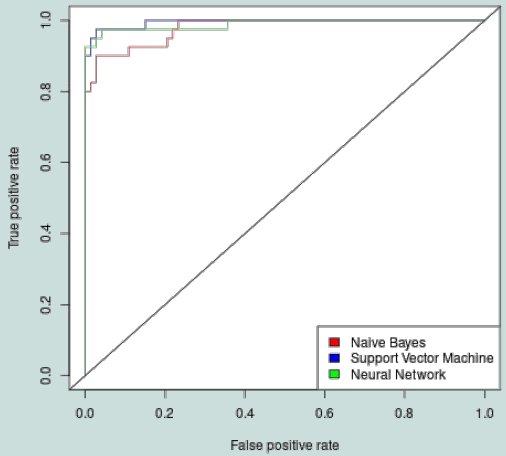
Neural Network

	Tempo (s)
No parallelizzazione	353.3030 ±0.5657
Parallelizzazione	58.0615 ±0.2807

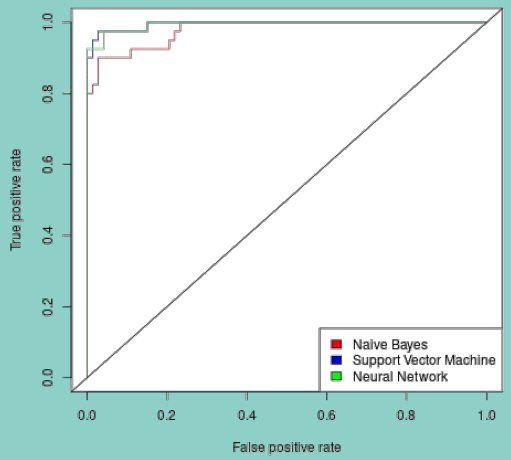
Con Neural Network si testano tramite Grid Search 48 combinazioni di parametri.

Fase di cross-validation ≈ 7.4 s

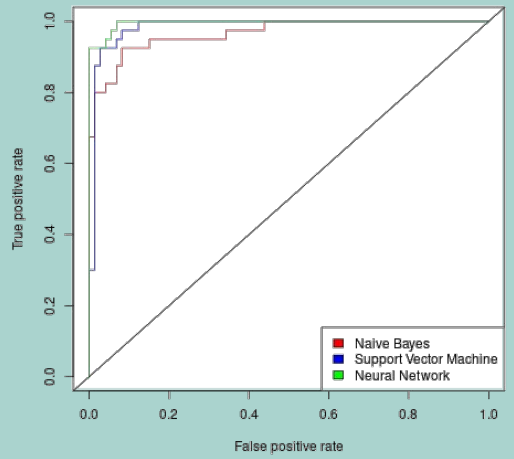
Risultati ottenuti



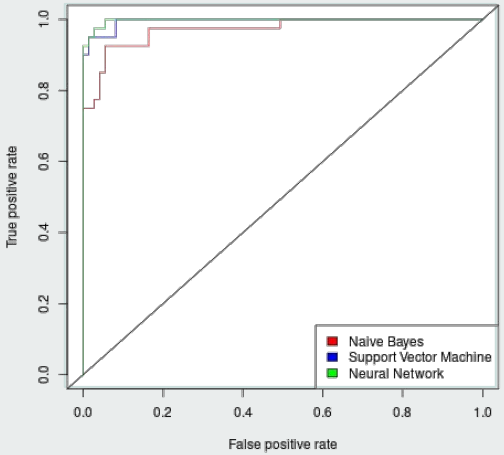
testset.norm



testset.std



testset.pca



testset.corr

Risultati Naive Bayes

Miglior risultato: dataset.corr

Train

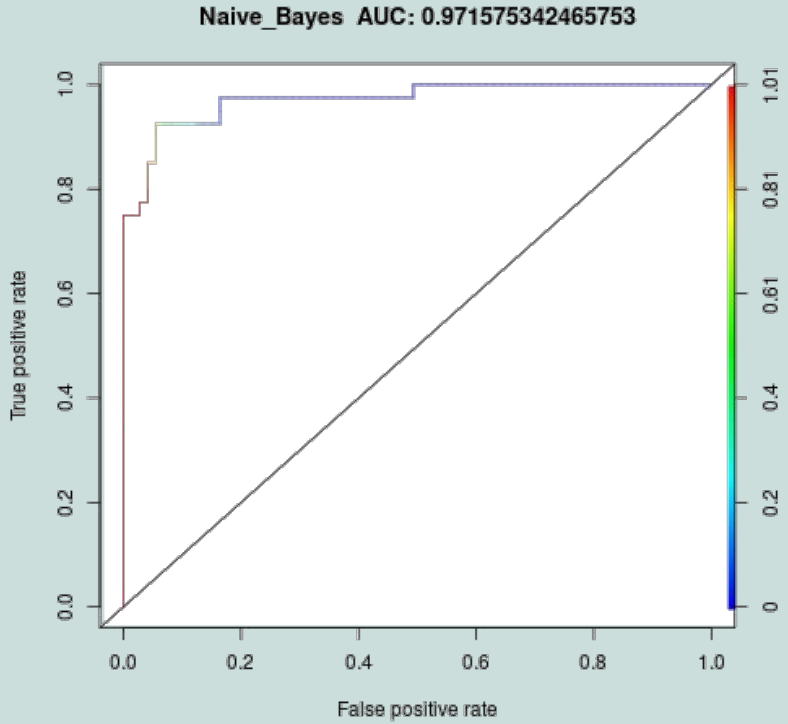
	Reference	
Predicted	M	B
M	152	15
B	20	209

Test

	Reference	
Predicted	M	B
M	37	4
B	**3**	69

Accuracy	**0.9381**
Accuracy CI 95%	**(0.8765; 0.9747)**
Precision	0.9024
Recall	**0.9250**
F1	0.9136
Sensitivity	**0.9250**
Specificity	0.9452

- **Parametri:** distribuzione delle feature stimata a partire dalle osservazioni.



Risultati SVM

Miglior risultato: dataset.corr

Train

	Reference	
Predicted	M	B
M	166	5
B	6	279

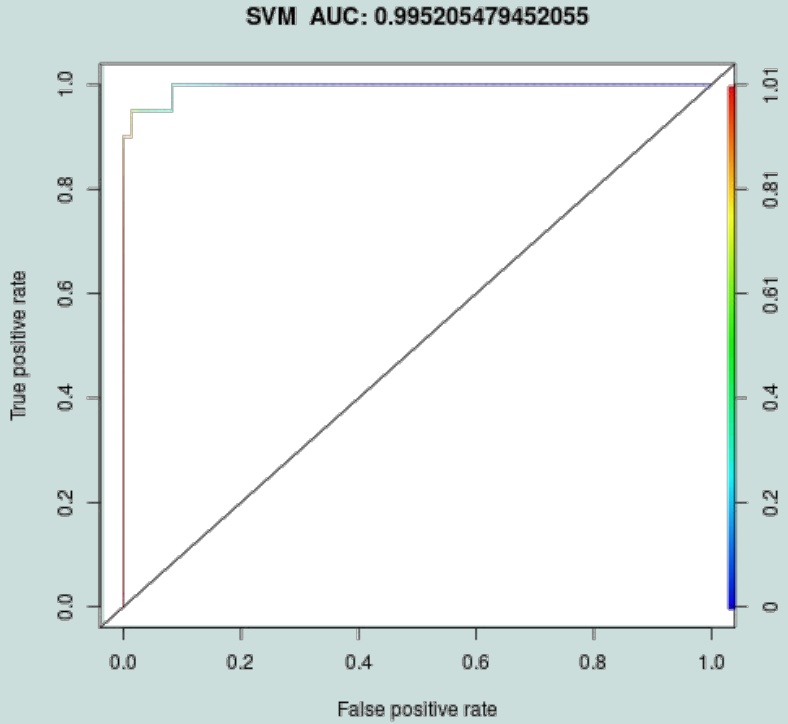
Test

	Reference	
Predicted	M	B
M	38	1
B	**2**	72

Accuracy	**0.9735**
Accuracy CI 95%	**(0.9244; 0.9945)**
Precision	0.9744
Recall	**0.9500**
F1	0.9620
Sensitivity	**0.9500**
Specificity	0.9863

Parametri:

- polinomio di primo grado;
- scale = 0.0625;
- C = 0.75.



Risultati Neural Network

Miglior risultato: dataset.corr

Train

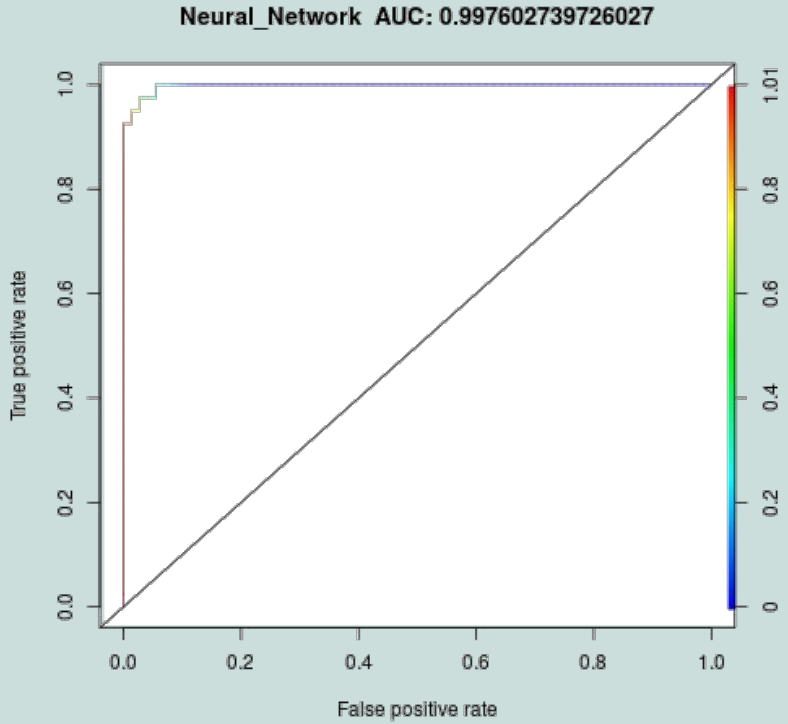
	Reference	
Predicted	M	B
M	168	0
B	4	284

Test

	Reference	
Predicted	M	B
M	39	2
B	**1**	71

Accuracy	**0.9735**
Accuracy CI 95%	**(0.9244; 0.9945)**
Precision	0.9512
Recall	**0.9750**
F1	0.9630
Sensitivity	**0.9750**
Specificity	0.9726

- **Parametri:** un layer nascosto con 4 neuroni. Rete fully connected.



Conclusioni

Performance migliori ottenute su dataset.corr

dataset.norm e dataset.std hanno avuto le stesse performance (esclusa la rete neurale che ha performato meglio sul dataset normalizzato)

I risultati di dataset.pca hanno ottenuto le performance più basse

Neural Network e SVM hanno sempre performato meglio di Naive Bayes

Neural Network ha avuto una Recall migliore di SVM a parità di Accuracy

Conclusioni

AMBITO INFORMATICO

Dataset utilizzato:

- Condizioni ideali
- Assenza di rumore
- Assenza di problemi nella raccolta dati
- Individuazione perfetta dei nuclei.

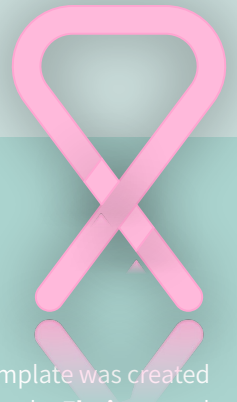
AMBITO MEDICO E DELLA RICERCA

Esigenze del medico:

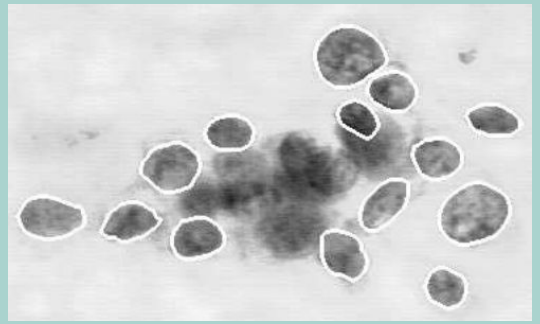
- Assenza di falsi negativi
- Non immediatezza dei risultati
- Semplicità di utilizzo
- Correttezza dei risultati

Nel 2018 sono stati contati oltre 9.6 milioni di casi di tumore al seno.
Tramite la ricerca possiamo contribuire ad individuare e trattare in tempo i pazienti.

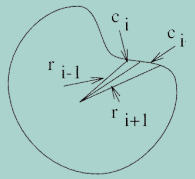
GRAZIE PER L'ATTENZIONE



Feature

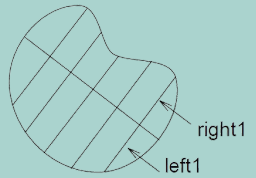


1. Radius: average of the length of the line segments given by the nucleus centroid and each contour point.
2. Perimeter: total distance between the contour points.
3. Area: #pixels within the boundary plus a half of the perimeter pixels.
4. Compactness: perimeter^2/area.



5. Smoothness: difference between a radial line length and the average length of the two lines surrounding it.

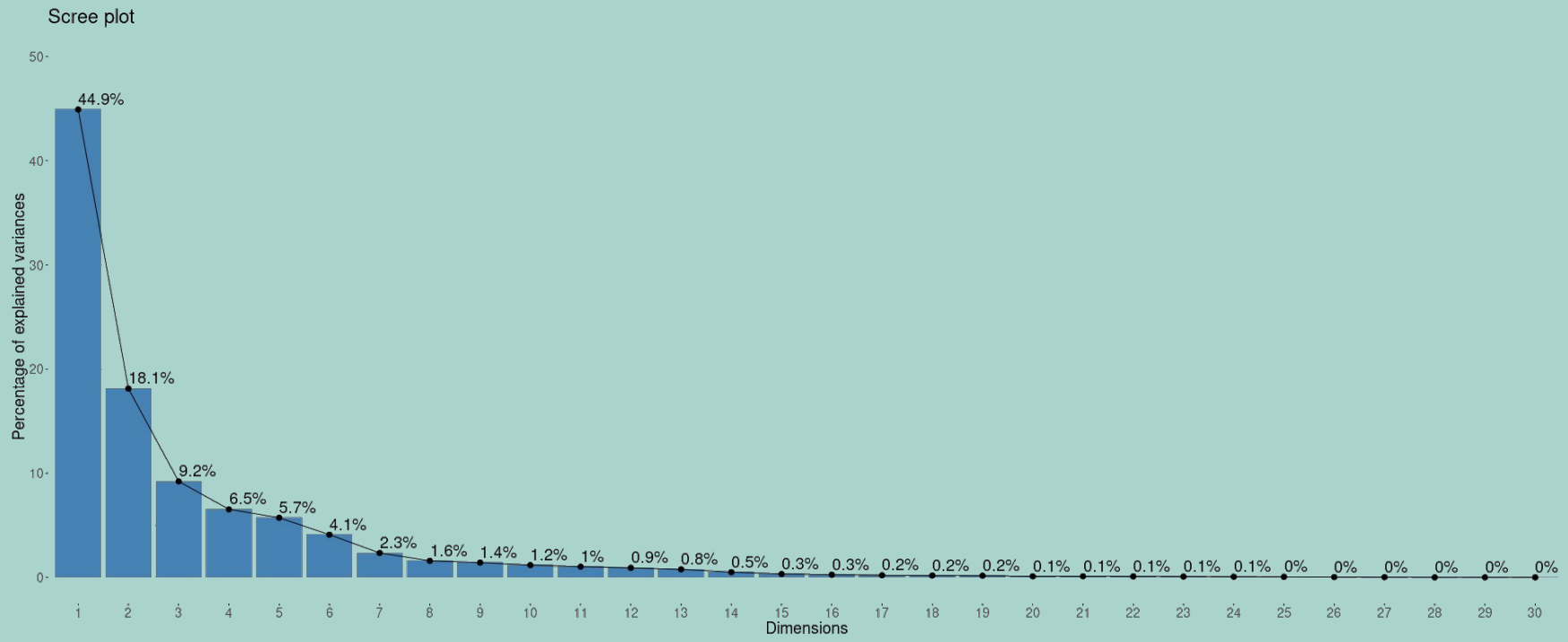
6. Concavity: chords between non adjacent contour points are drawn, creating a more serrated shape. Concavity is the average distance of snake points within the chords.
7. Concave points: #snake points that lie within the chords drawn to measure concavity.



8. Symmetry: relative difference between the lengths of each pair of segments perpendicular to the major axis.

9. Fractal dimension: Mandelbrot fractal dimension
10. Texture: variance of the gray scale intensity values of pixels

PCA


Scree plot
Percentage of explained variances
Dimensions
50
40
30
20
10
0
44.9%
18.1%
9.2%
6.5%
5.7%
4.1%
2.3%
1.6%
1.4%
1.2%
1%
0.9%
0.8%
0.5%
0.3%
0.3%
0.2%
0.2%
0.2%
0.1%
0.1%
0.1%
0.1%
0.1%
0%
0%
0%
0%
0%
0%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

Risultati ottenuti

Test.norm

	Bayes	SVM	NN
Accuracy	**0.9115**	**0.9646**	**0.9646**
Accuracy CI 95%	**(0.8433; 0.9567)**	**(0.9118; 0.9903)**	**(0.9118; 0.9903)**
Precision	0.8571	0.9500	0.9286
Recall	**0.9000**	**0.9500**	**0.97500**
F1	0.8780	0.9500	0.9512
Sensitivity	**0.9000**	**0.9500**	**0.9750**
Specificity	0.9178	0.9726	0.9589

Test.std

	Bayes	SVM	NN
Accuracy	**0.9115**	**0.9646**	**0.9558**
Accuracy CI 95%	**(0.8433; 0.9567)**	**(0.9118; 0.9903)**	**(0.8998; 0.9855)**
Precision	0.8751	0.9500	0.9487
Recall	**0.9000**	**0.9500**	**0.9250**
F1	0.8780	0.9500	0.9367
Sensitivity	**0.9000**	**0.9500**	**0.9250**
Specificity	0.9178	0.9726	0.9726

Test.PCA

	Bayes	SVM	NN
Accuracy	**0.9027**	**0.9558**	**0.9558**
Accuracy CI 95%	**(0.8325; 0.9504)**	**(0.8998; 0.9855)**	**(0.8998; 0.9855)**
Precision	0.8718	0.9487	0.9487
Recall	**0.8500**	**0.9250**	**0.9250**
F1	0.8608	0.9367	0.9367
Sensitivity	**0.8500**	**0.9250**	**0.9250**
Specificity	0.9315	0.9726	0.9726

Risultati Naive Bayes

Dataset: dataset.norm

Train

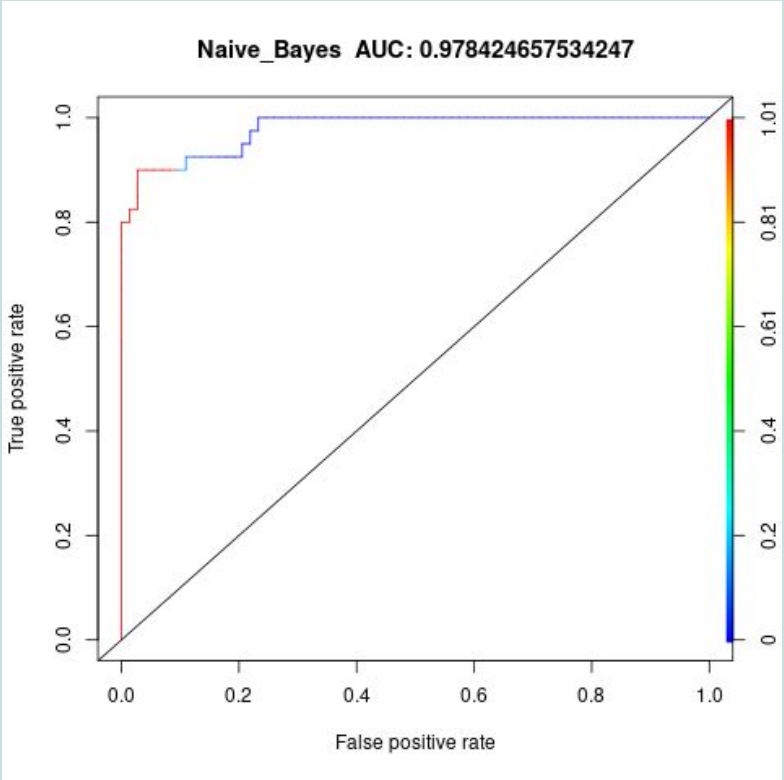
	Reference	
Predicted	M	B
M	161	10
B	11	274

Test

	Reference	
Predicted	M	B
M	36	6
B	**4**	67

Accuracy	**0.9115**
Accuracy CI 95%	**(0.8433; 0.9567)**
Precision	0.8571
Recall	**0.9000**
F1	0.8780
Sensitivity	**0.9000**
Specificity	0.9178

- **Parametri:** distribuzione delle feature stimata a partire dalle osservazioni.



Risultati SVM

Dataset: dataset.norm

Train

	Reference	
Predicted	M	B
M	167	1
B	5	283

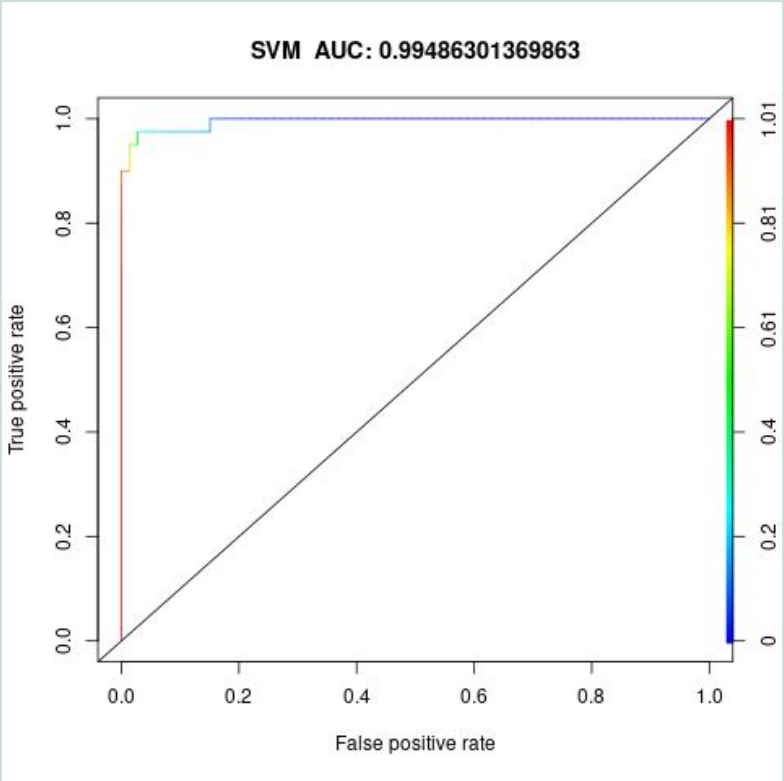
Test

	Reference	
Predicted	M	B
M	38	2
B	**2**	71

Accuracy	**0.9646**
Accuracy CI 95%	**(0.9118; 0.9903)**
Precision	0.9500
Recall	**0.9500**
F1	0.9500
Sensitivity	**0.9500**
Specificity	0.9726

Parametri:

- polinomio di primo grado;
- scale = 0.0333;
- C = 0.75.



Risultati Neural Network

Dataset: dataset.norm

Train

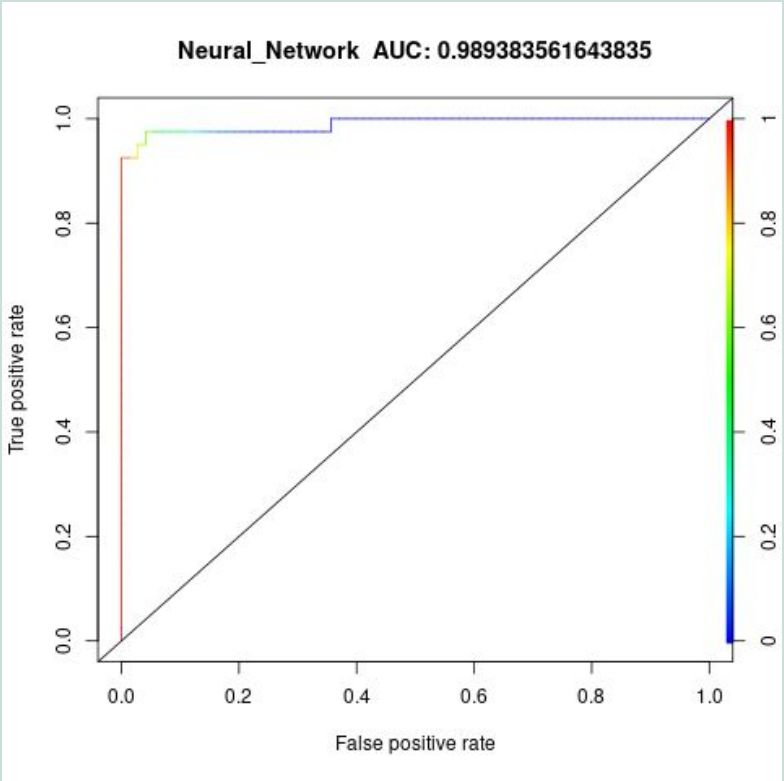
	Reference	
Predicted	M	B
M	168	4
B	4	280

Test

	Reference	
Predicted	M	B
M	39	3
B	**1**	70

Accuracy	**0.9646**
Accuracy CI 95%	**(0.9118; 0.9903)**
Precision	0.9286
Recall	**0.9750**
F1	0.9512
Sensitivity	**0.9750**
Specificity	0.9589

- **Parametri:** un layer nascosto con 3 neuroni. Rete fully connected.



Risultati Naive Bayes

Dataset: dataset.std

Train

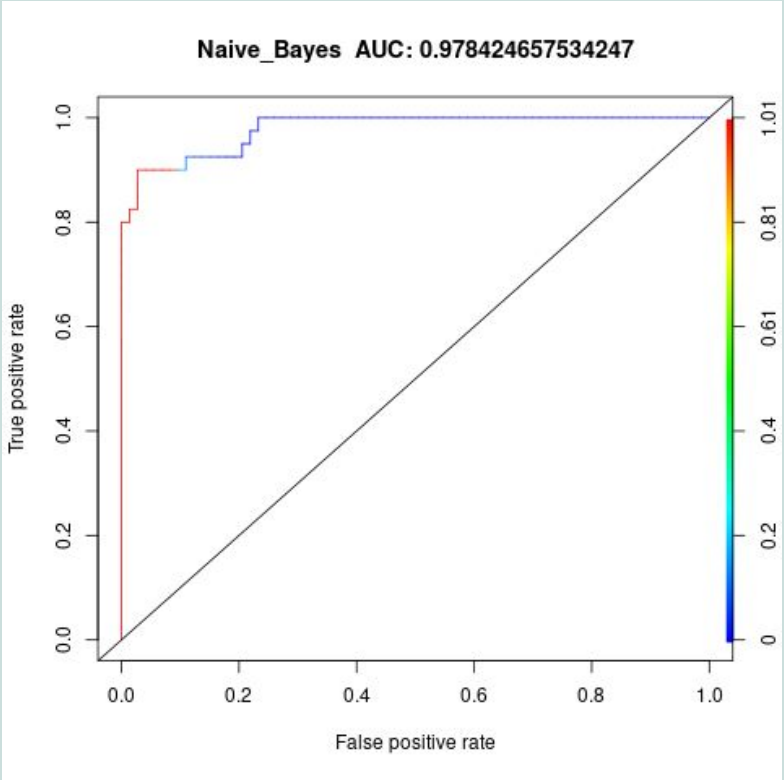
	Reference	
Predicted	M	B
M	161	10
B	11	274

Test

	Reference	
Predicted	M	B
M	36	6
B	**4**	67

Accuracy	**0.9115**
Accuracy CI 95%	**(0.8433; 0.9567)**
Precision	0.8571
Recall	**0.9000**
F1	0.8780
Sensitivity	**0.9000**
Specificity	0.9178

- **Parametri:** distribuzione delle feature stimata a partire dalle osservazioni.



Risultati SVM

Dataset: dataset.std

Train

	Reference	
Predicted	M	B
M	167	1
B	5	283

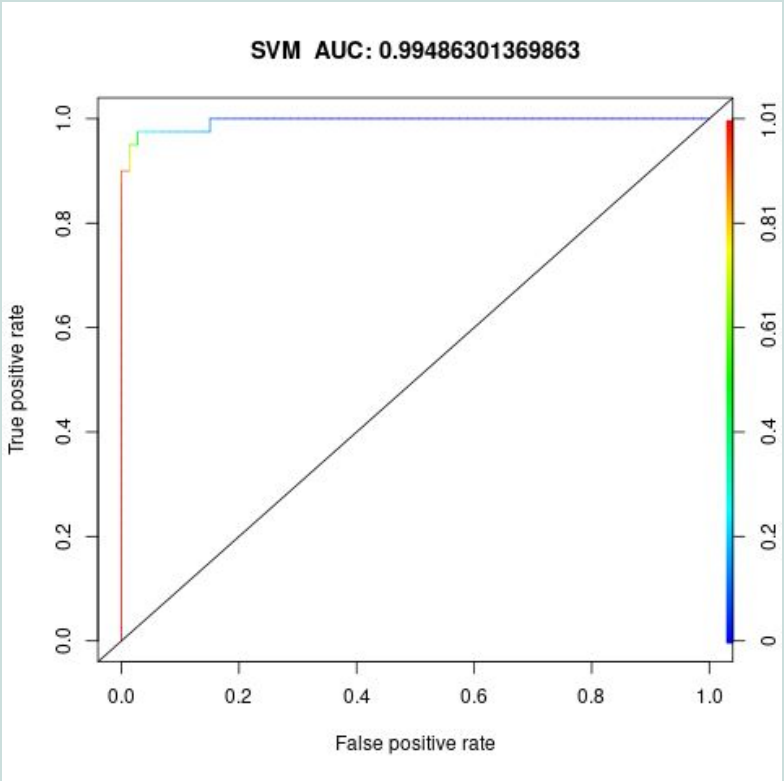
Test

	Reference	
Predicted	M	B
M	38	2
B	**2**	71

Accuracy	**0.9646**
Accuracy CI 95%	**(0.9118; 0.9903)**
Precision	0.9500
Recall	**0.9500**
F1	0.9500
Sensitivity	**0.9500**
Specificity	0.9726

Parametri:

- polinomio di primo grado;
- scale = 0.0333;
- C = 0.75.



Risultati Neural Network

Dataset: dataset.std

Train

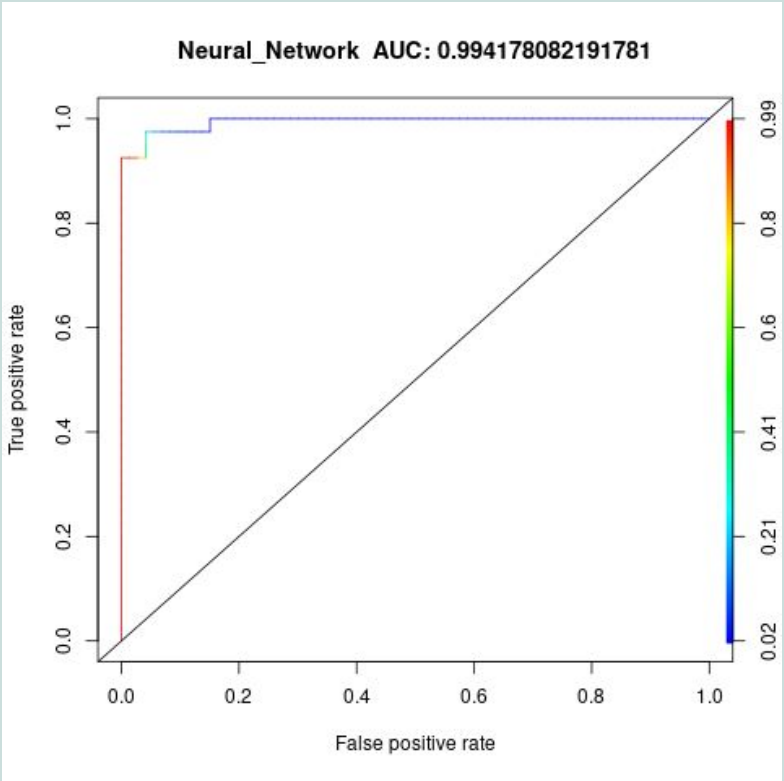
	Reference	
Predicted	M	B
M	168	0
B	4	284

Test

	Reference	
Predicted	M	B
M	37	2
B	**3**	71

Accuracy	**0.9558**
Accuracy CI 95%	**(0.8998; 0.9855)**
Precision	0.9487
Recall	**0.9250**
F1	0.9367
Sensitivity	**0.9250**
Specificity	0.9726

- **Parametri:** un layer nascosto con 3 neuroni seguito da un layer nascosto con un neurone. Rete fully connected.



Risultati Naive Bayes

Dataset: dataset.pca

Train

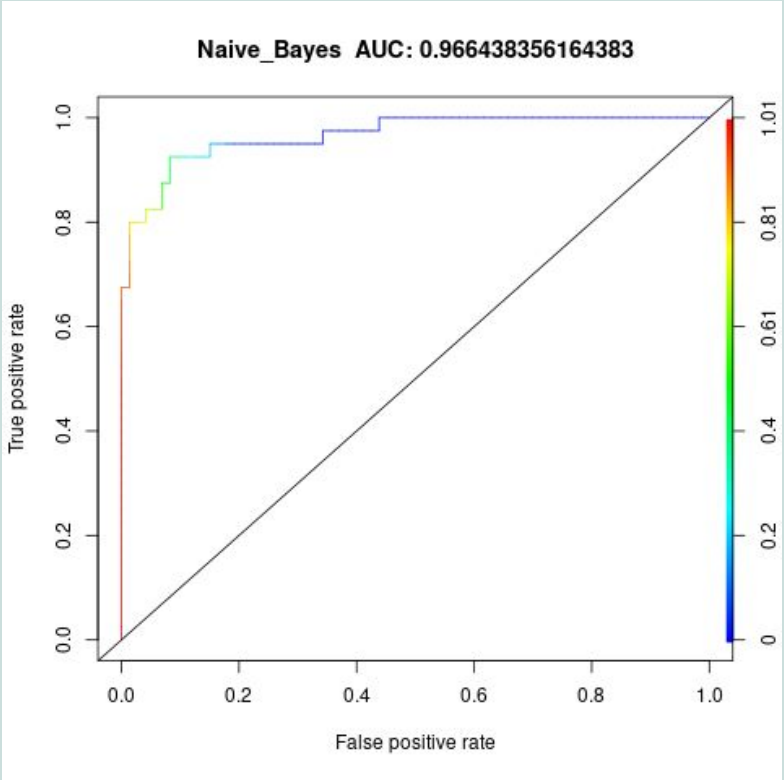
	Reference	
Predicted	M	B
M	154	5
B	18	279

Test

	Reference	
Predicted	M	B
M	36	5
B	**6**	68

Accuracy	**0.9027**
Accuracy CI 95%	**(0.8325; 0.9504)**
Precision	0.8718
Recall	**0.8500**
F1	0.8608
Sensitivity	**0.8500**
Specificity	0.9315

- **Parametri:** distribuzione delle feature stimata a partire dalle osservazioni.



Risultati SVM

Dataset: dataset.pca

Train

	Reference	
Predicted	M	B
M	166	0
B	6	284

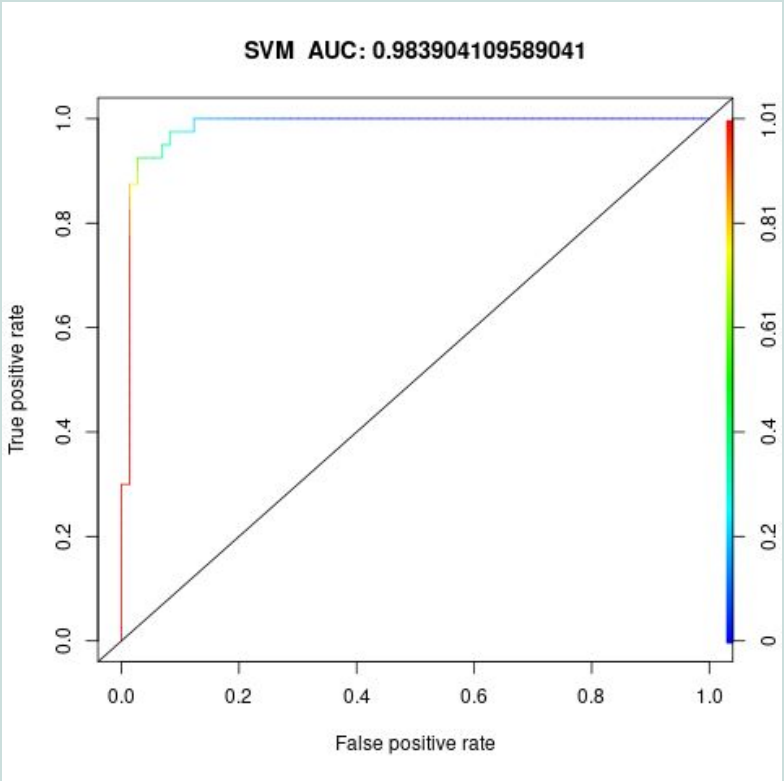
Test

	Reference	
Predicted	M	B
M	37	2
B	**3**	71

Accuracy	**0.9558**
Accuracy CI 95%	**(0.8998; 0.9855)**
Precision	0.9487
Recall	**0.9250**
F1	0.9367
Sensitivity	**0.9250**
Specificity	0.9726

Parametri:

- polinomio di secondo grado;
- scale = 0.0625;
- C = 1.5.



Risultati Neural Network

Dataset: dataset.pca

Train

	Reference	
Predicted	M	B
M	168	0
B	4	284

Test

	Reference	
Predicted	M	B
M	37	2
B	**3**	71

Accuracy	**0.9558**
Accuracy CI 95%	**(0.8998; 0.9855)**
Precision	0.9487
Recall	**0.9250**
F1	0.9367
Sensitivity	**0.9250**
Specificity	0.9726

- **Parametri:** un layer nascosto con 4 neuroni seguito da un layer nascosto con un neurone. Rete fully connected.