Università degli Studi di Milano Bicocca

Scuola di Scienze

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di laurea magistrale in Informatica

FETAL HEALTH

PIAZZA MARCO - 829588

CAZZANIGA ELISA - 829914

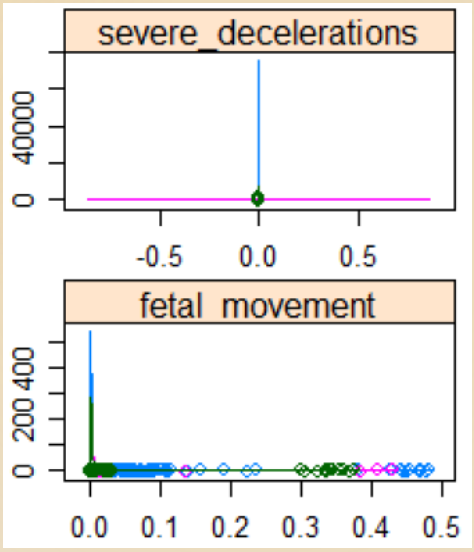
MACHINE LEARNING – Anno Accademico 2020/2021

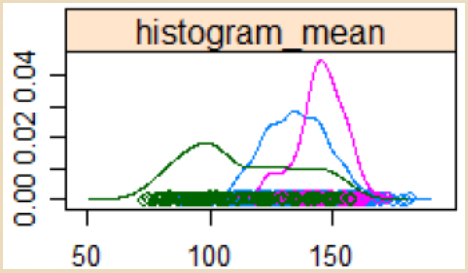
DOMINIO DI RIFERIMENTO

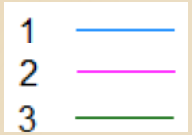
- https://www.kaggle.com/andrewmvd/fetal-health-classification
- 2126 registrazioni di caratteristiche estratte da esami di CTG (cardiotocografia)utilizzati per valutare la salute fetale
- Classificati da tre ostetriche esperte in 3 classi:
  - Normale
  - Sospetto
  - Patologico

ANALISI DELLE COVARIATE

- 22 covariate
- Analisi univariata → distribuzione + presenza di eventuali outliers







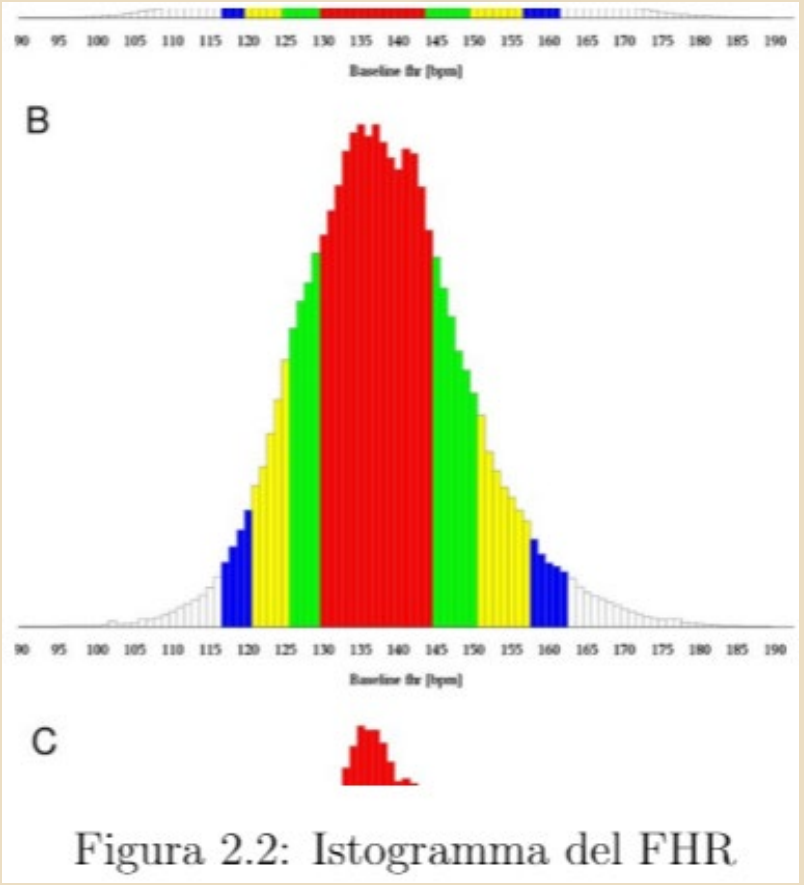
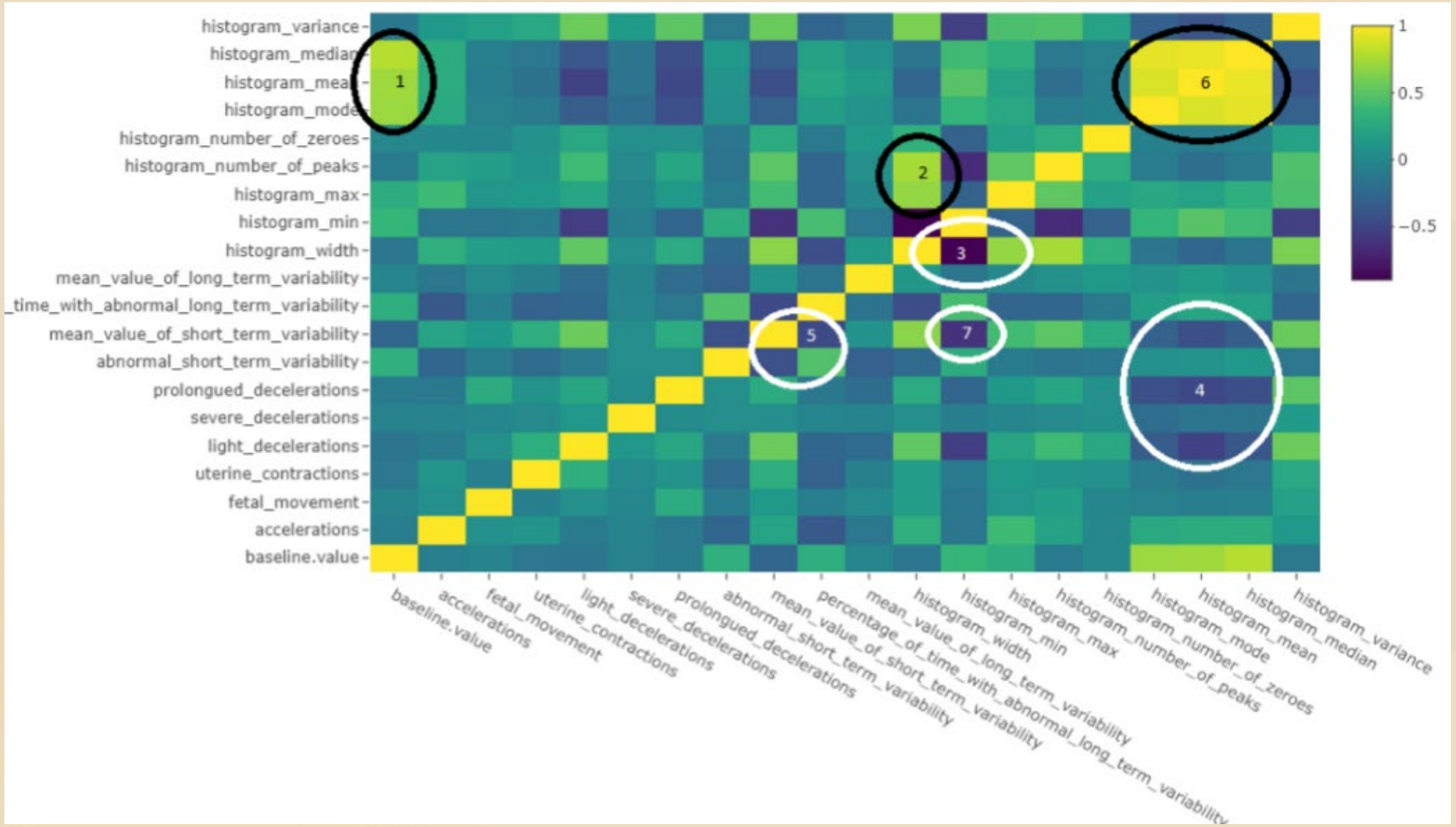


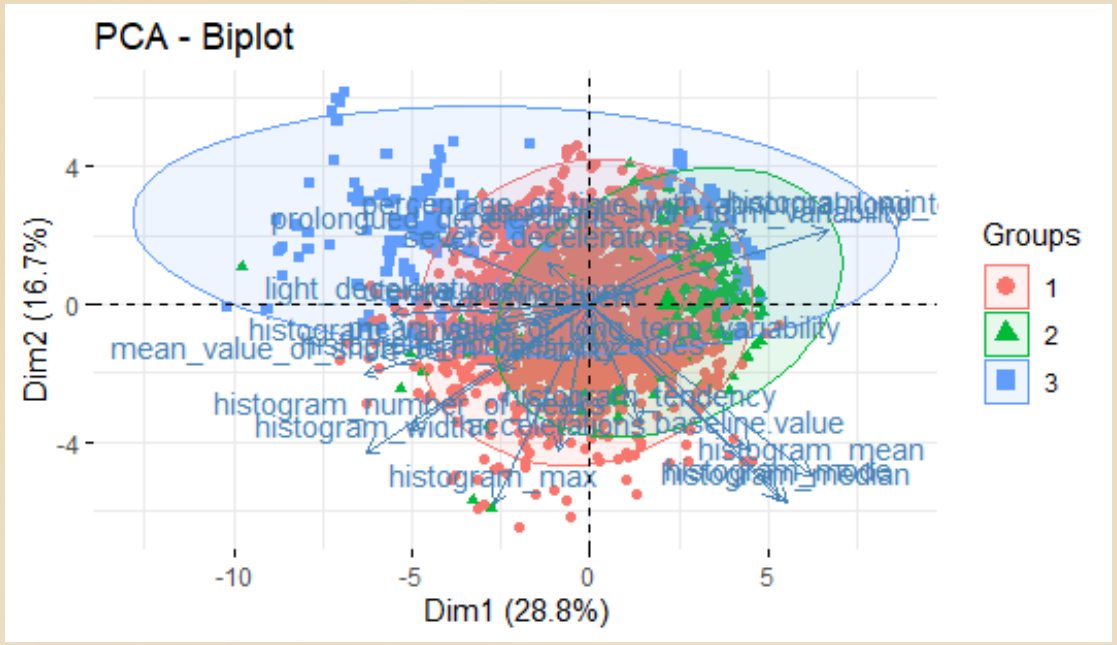
Figura 2.2: Istogramma del FHR

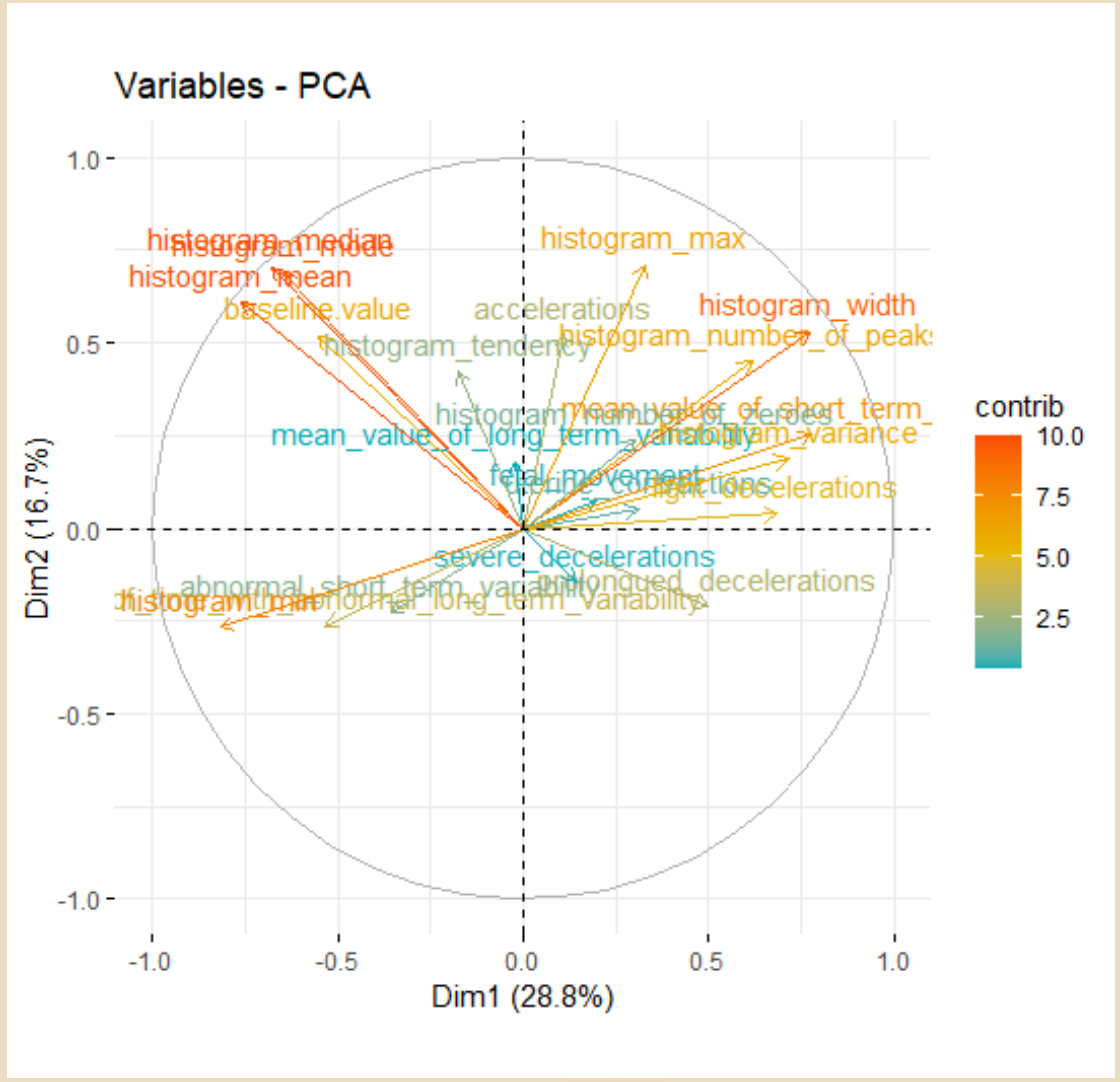


- Analisi Multivariata → correlazioni tra i vari attributi

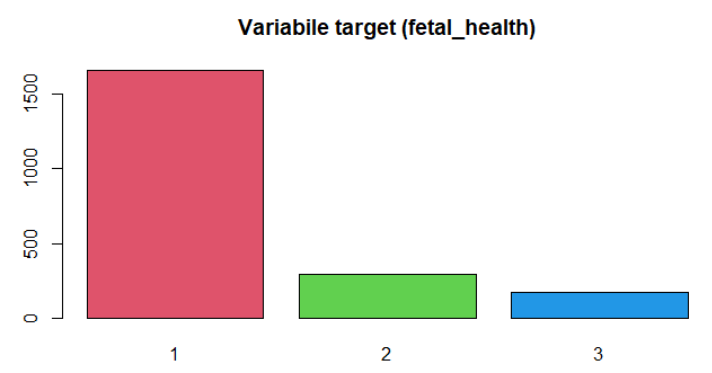
PRINCIPAL COMPONENT ANALYSIS

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	6.058351e+00	2.884929e+01	28.84929
Dim.2	3.507779e+00	1.670371e+01	45.55300
Dim.3	1.824323e+00	8.687251e+00	54.24025
Dim.4	1.496719e+00	7.127233e+00	61.36749
Dim.5	1.218359e+00	5.801710e+00	67.16920
Dim.6	1.020172e+00	4.857964e+00	72.02716
Dim.7	9.817713e-01	4.675101e+00	76.70226
Dim.8	9.265325e-01	4.412059e+00	81.11432
Dim.9	7.624281e-01	3.630610e+00	84.74493





PROBLEMATICA DATASET SBILANCIATO



- Originale
- Oversampling
- Undersampling
- Classificazione basata sui pesi

Decision tree

	1	2	3
1	459	21	10
2	16	79	1
3	12	2	42

	1	2	3
1	414	12	3
2	57	86	1
3	16	4	49

	1	2	3
1	418	13	6
2	40	82	2
3	29	7	45

	1	2	3
1	300	5	0
2	104	92	3
3	83	5	50

	1	2	3
Accuracy	0.8713	0.8715	0.88434
Precision	0.9367	0.8229	0.75000
Recall	0.9425	0.7745	0.79245
F-Measure	0.9396	0.7980	0.77064

	1	2	3
Accuracy	0.8767	0.8679	0.94529
Precision	0.9650	0.5972	0.71014
Recall	0.8501	0.8431	0.92453
F-Measure	0.9039	0.6992	0.80328

	1	2	3
Accuracy	0.8679	0.8631	0.89397
Precision	0.9565	0.6613	0.55556
Recall	0.8583	0.8039	0.84906
F-Measure	0.9048	0.7257	0.67164

	1	2	3
Accuracy	0.7919	0.8519	0.89700
Precision	0.9836	0.4623	0.36232
Recall	0.6160	0.9020	0.94340
F-Measure	0.7576	0.6113	0.52356

DECISION TREE

- Albero normale
- Albero puned
- Confronto utilizzo dataset originario e dataset ottenuto tramite PCA
- Confronto tra Gini Index e Information Gain

	1	2	3
1	459	21	10
2	16	79	1
3	12	2	42

	1	2	3
1	472	34	16
2	7	67	1
3	8	1	36

	1	2	3
Accuracy	0.8713	0.8715	0.88434
Precision	0.9367	0.8229	0.7500
Recall	0.9425	0.7745	0.79245
F-Measure	0.9396	0.7980	0.77064

	1	2	3
Accuracy	0.8233	0.8210	0.83198
Precision	0.9042	0.8933	0.8000
Recall	0.9692	0.6569	0.67925
F-Measure	0.9356	0.7571	0.73469

Albero normale vs pruned (dataset originario)

SUPPORT VECTOR MACHINE

- Tuning svm per scegliere costo ottimale
- Confronto utilizzo dataset originario e dataset ottenuto tramite PCA
- Confronto tra varie forme di kernel

	1	2	3
1	486	49	0
2	1	52	1
3	0	1	52

	1	2	3
1	502	42	1
2	3	52	15
3	0	0	32

	1	2	3
Accuracy	0.8409	0.75305	0.98972
Precision	0.9084	0.96296	0.98113
Recall	0.9979	0.50980	0.98113
F-Measure	0.9476	0.66667	0.98113

	1	2	3
Accuracy	0.8456	0.76032	0.83333
Precision	0.9211	0.74286	1.00000
Recall	0.9941	0.55319	0.66667
F-Measure	0.9562	0.63415	0.80000

Confronto tra dataset originale e dataset risultante dalla PCA

ESPERIMENTO

- Obiettivo: trovare il modello migliore
- Dataset ottenuto tramite PCA e pesato
- Utilizzo 10-Fold cross validation

	1	2	3
1	492	28	4
2	16	57	6
3	5	6	48

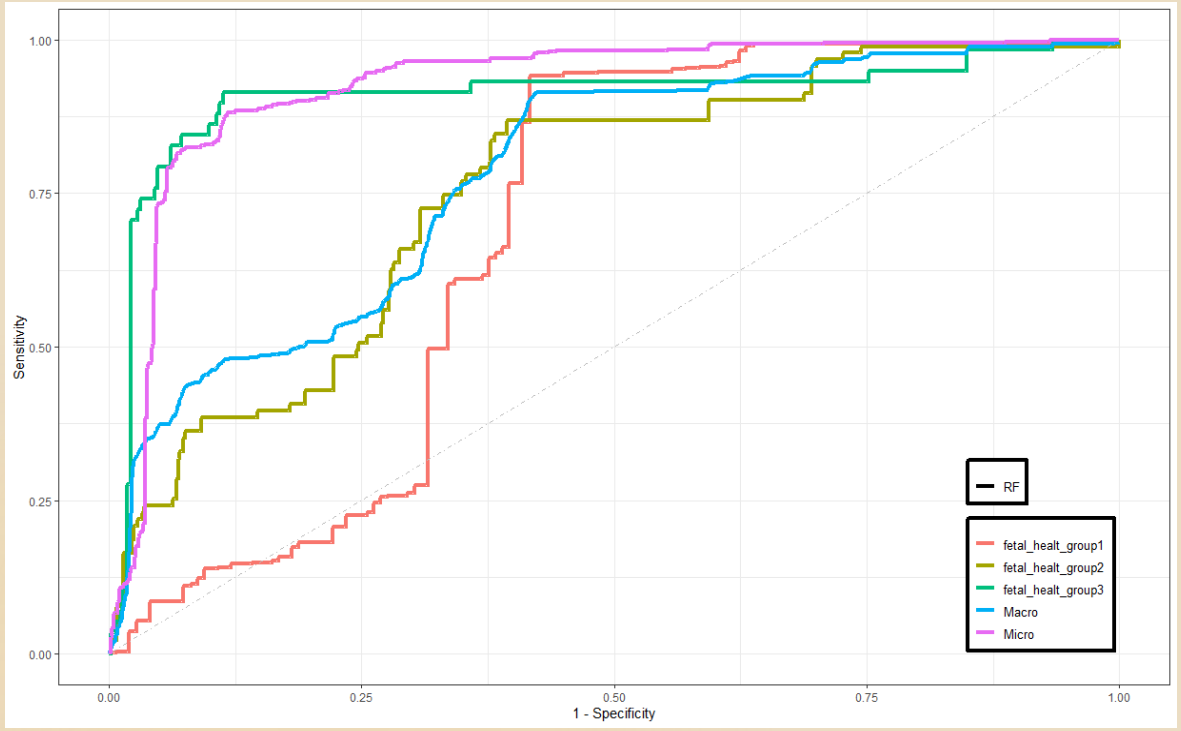
	1	2	3
1	483	56	7
2	26	22	10
3	4	13	41

	1	2	3
Accuracy	0.8721	0.7939	0.90469
Precision	0.9389	0.7215	0.81356
Recall	0.9591	0.6264	0.82759
F-Measure	0.9489	0.6706	0.82051

	1	2	3
Accuracy	0.7594	0.58936	0.83938
Precision	0.8846	0.37931	0.70690
Recall	0.9415	0.24176	0.70690
F-Measure	0.9122	0.29530	0.70690

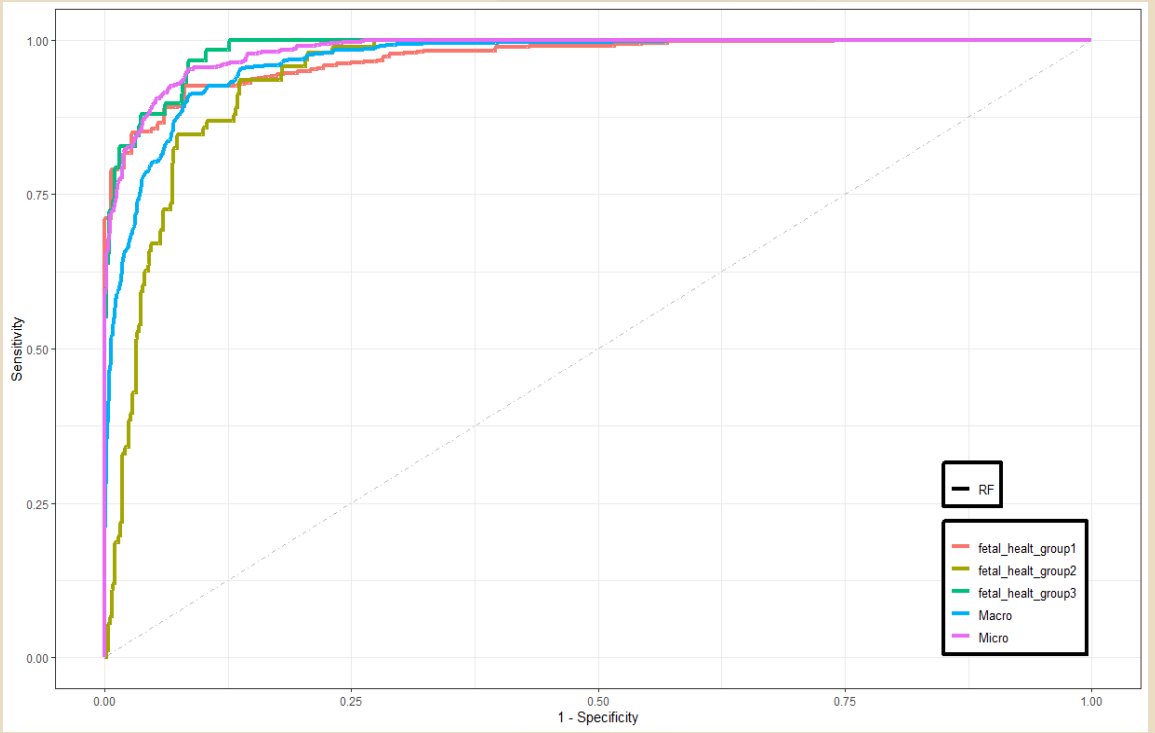
Confronto tra risultati ottenuti da SVM e decision tree

CURVA ROC

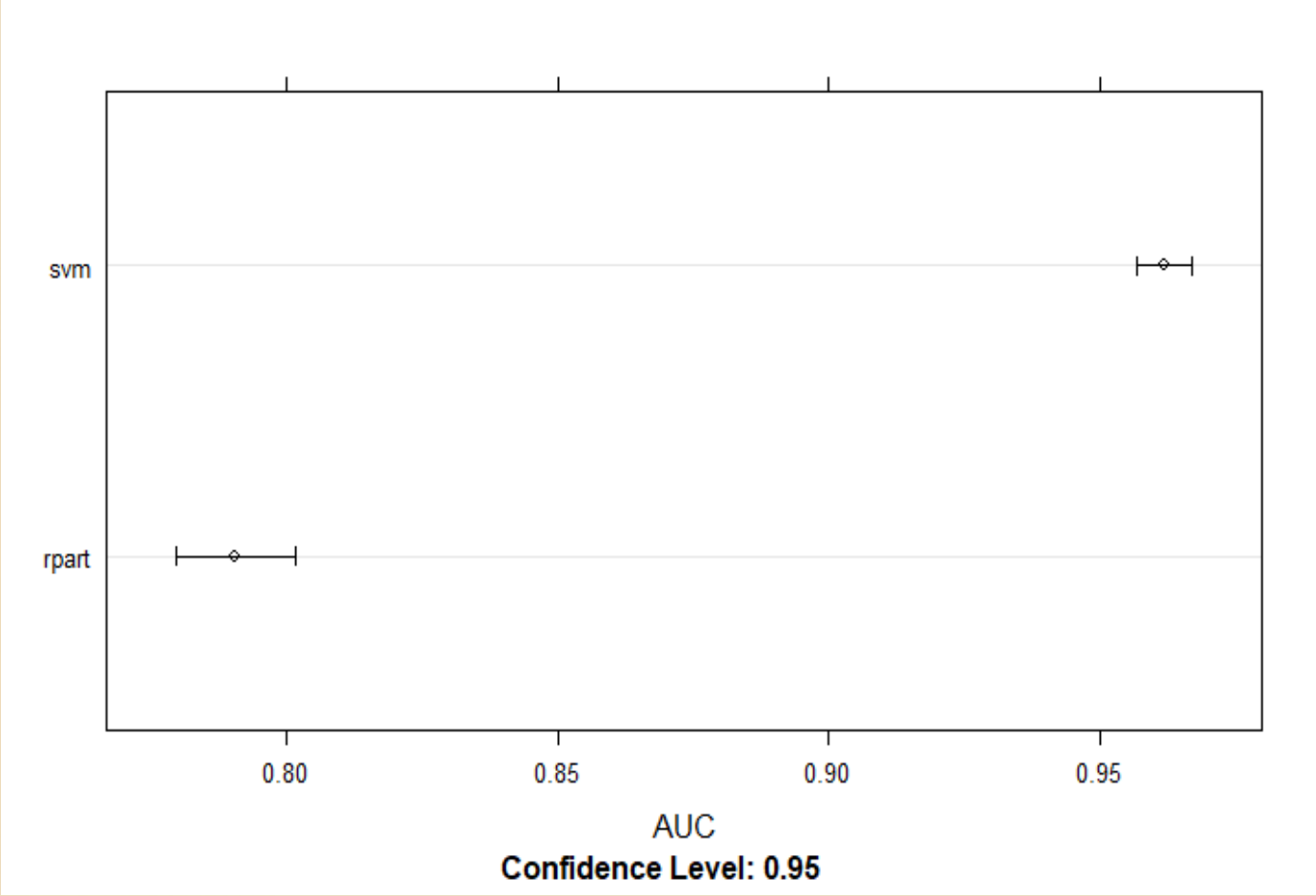


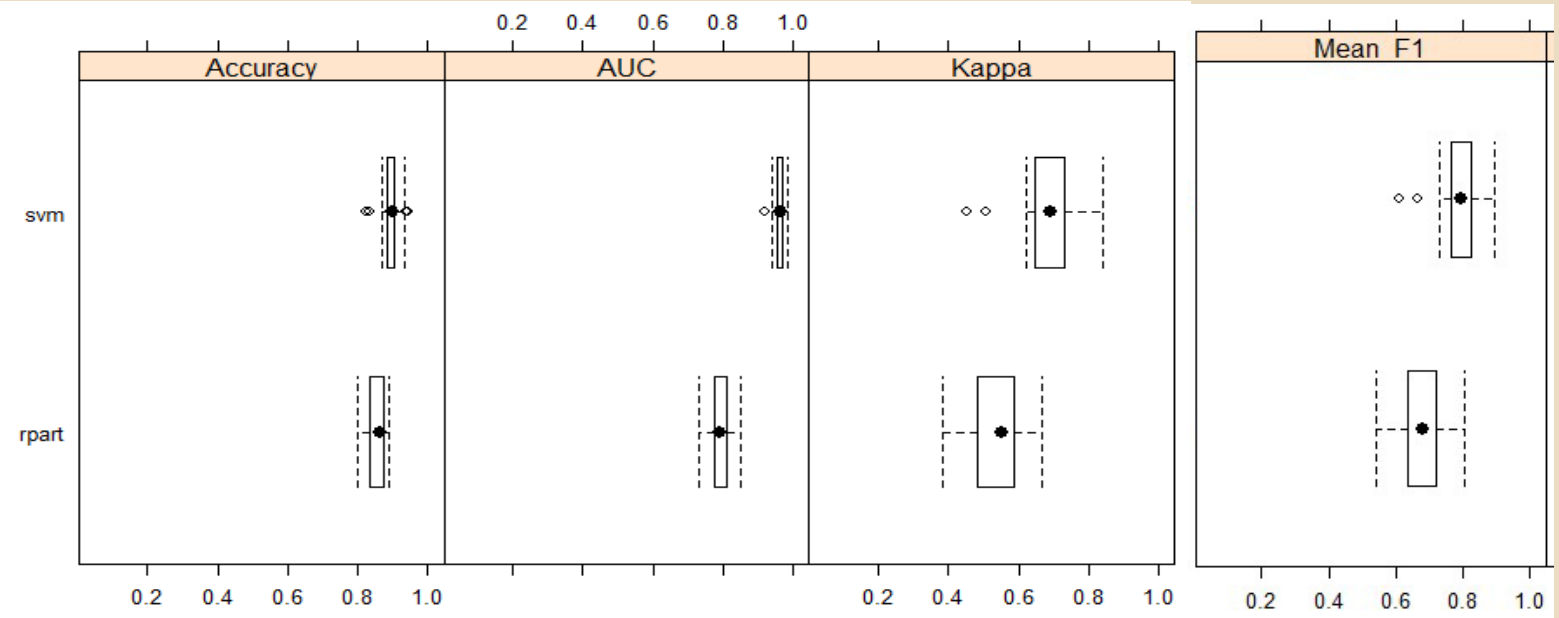
Decision tree + PCA + weight

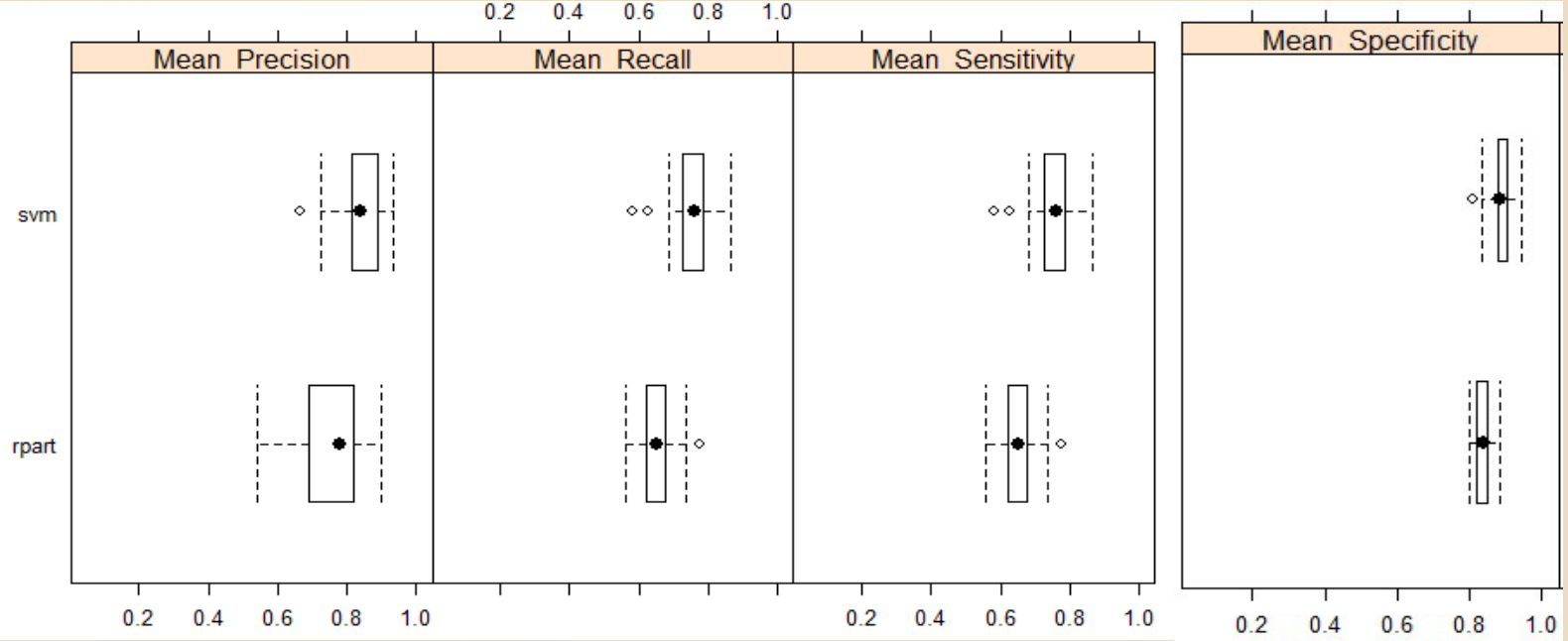
Support Vector Machine (Polynomial)
+ PCA + weight



ANALISI DEI RISULTATI







- Tempistiche:

	Everything	FinalModel
svm	176.48	0.11
rpart	8.19	0.02

CONCLUSIONI

- Utilizzo PCA per semplificare modelli e ridurre overfitting
- SVM → performance migliori, timing maggiore
- Decision tree → performance leggermente peggiori, timinig minore
- Utilizzo classificazione basata sui pesi per diminuire i falsi negativi (Migliorare predizione classe 3)
- Modello migliore valutando campo di applicazione: SVM