Università degli Studi di Milano– Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Anno Accademico 2021 – 2022

Contraceptive Method Choice

Giacomo Savazzi 845372 Raffaele Cerizza 845512

22 Febbraio 2022

Analisi del Dominio di Riferimento

Target → Contraceptive Method Choice:

- *No Use*
- *Short Term*
- *Long Term*

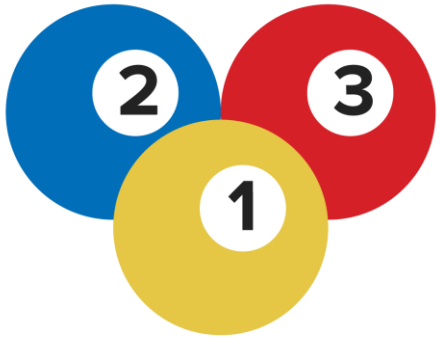
Covariate di riferimento:

- *Wife Age → numeric*
- *Wife Education → factor (High, Mid-High, Mid-Low, Low)*
- *Husband Education → factor (High, Mid-High, Mid-Low, Low)*
- *Number of Children → numeric*
- *Wife Religion → factor (Islam, Non-Islam)*
- *Wife is Working → factor (Yes, No)*
- *Husband Occupation → factor (High, Mid-High, Mid-Low, Low)*
- *Living Index → factor (High, Mid-High, Mid-Low, Low)*
- *Media Exposure → factor (Good, Not-Good)*

Obiettivi dell'Elaborato

Problema binario:

- *No Use*
- *Use (Short Term o Long Term)*

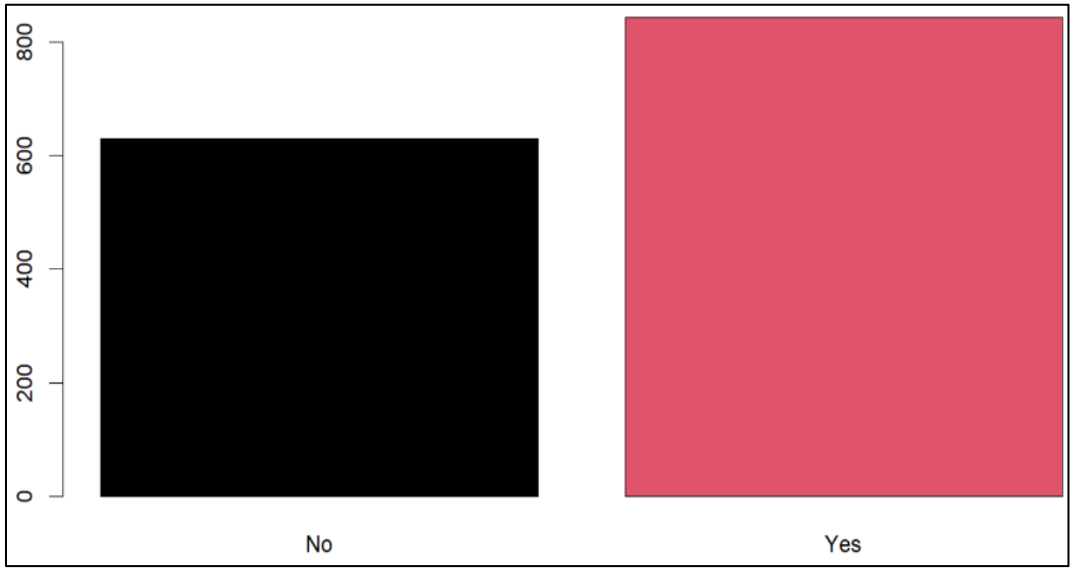


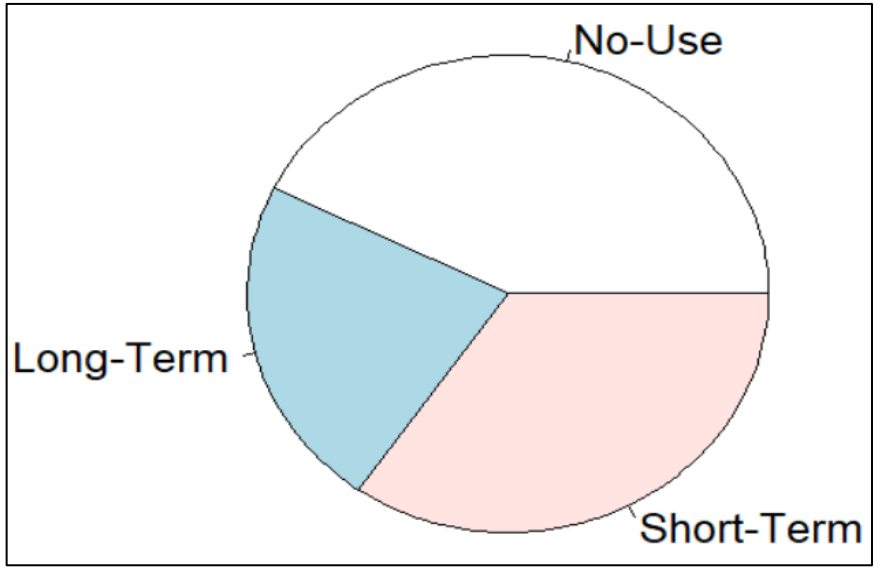
Problema multi-classe:

- *No Use*
- *Short Term*
- *Long Term*

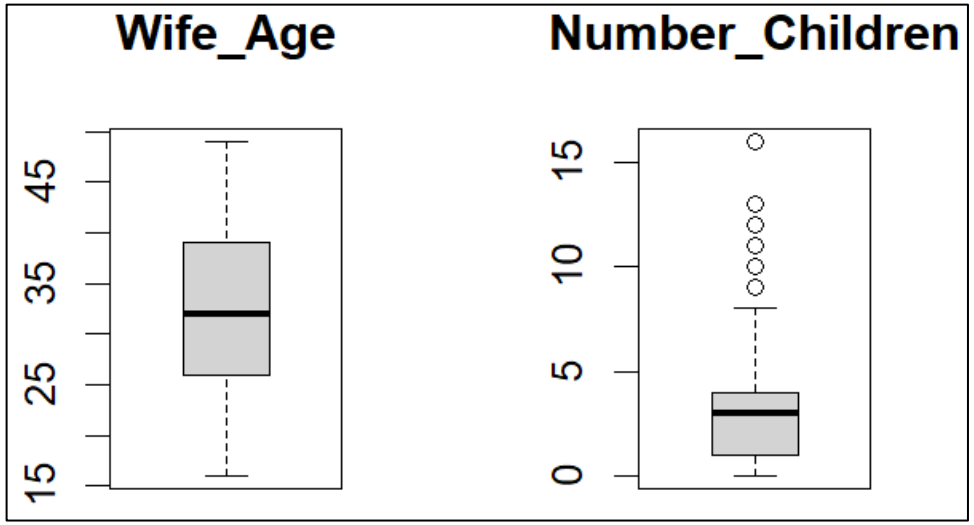
Analisi Esplorativa

Dataset → 1473 istanze in 10 variabili
Target → Contraceptive_Is_Used



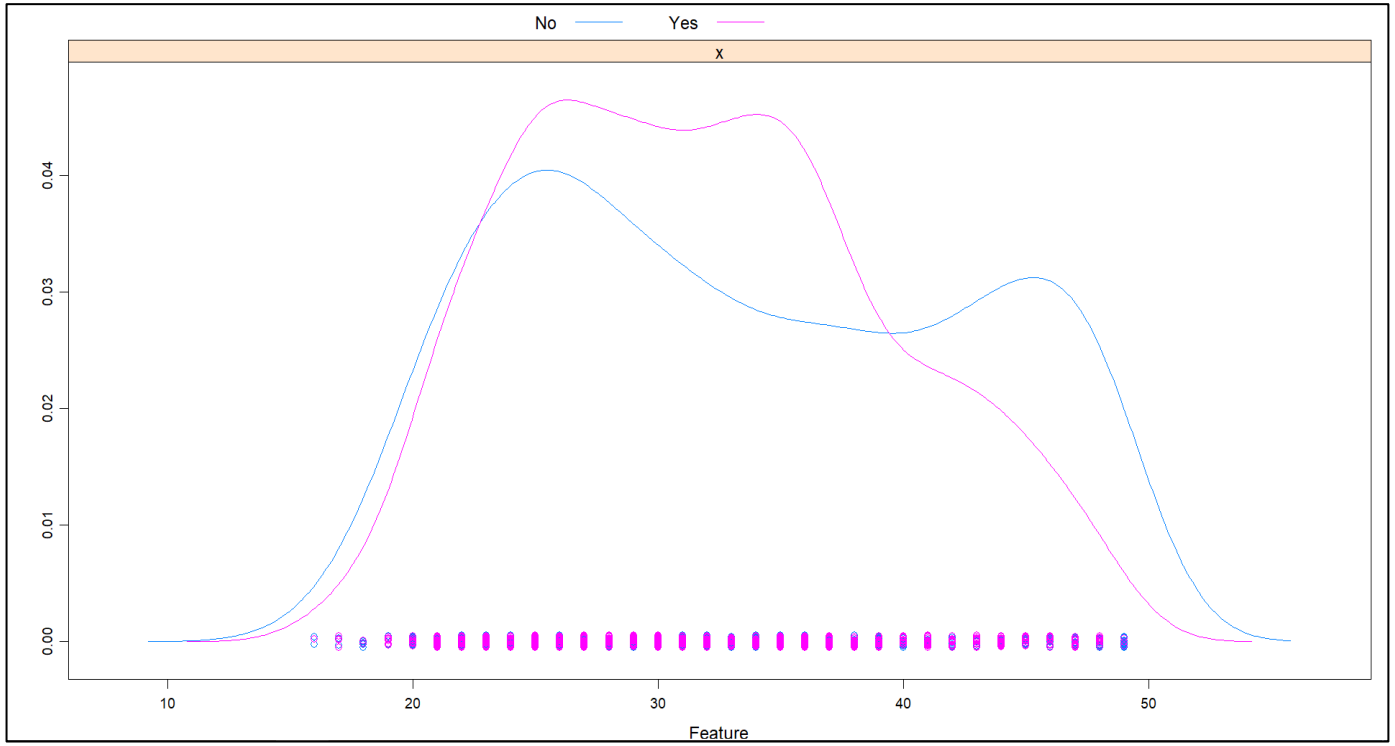


Analisi Esplorativa – Variabili Numeriche

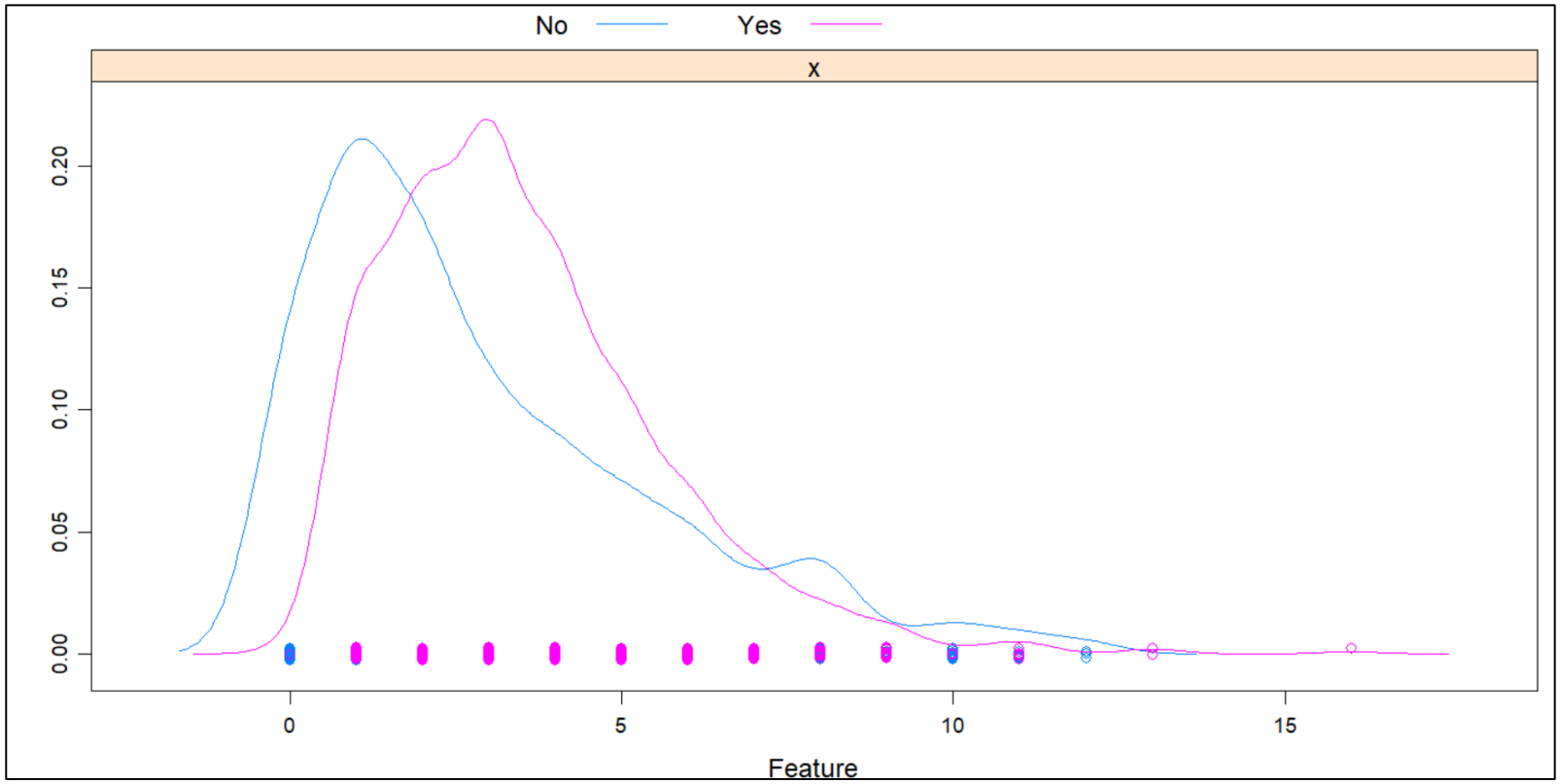


Boxplot

Feature plot Wife Age



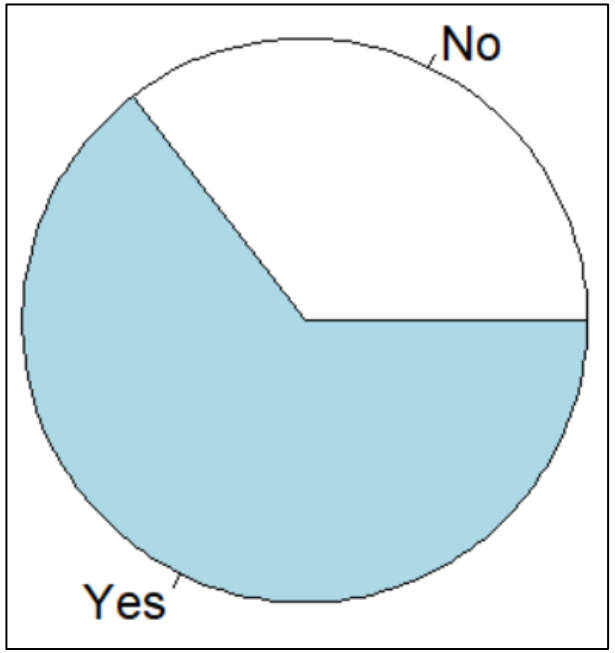
Analisi Esplorativa – Variabili Numeriche



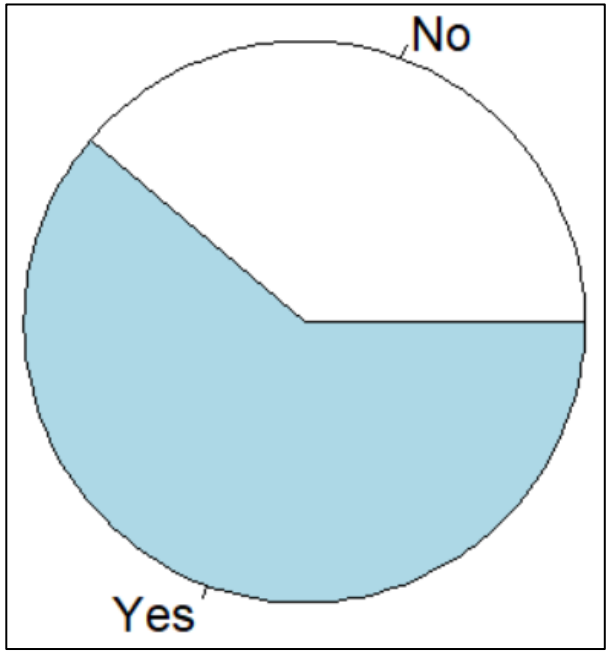
Feature plot Number of Children

```
> cor(cmc[, c(1,4)]) # 0.5401259
                 Wife_Age Number_Children
Wife_Age        1.0000000        0.5401259
Number_Children 0.5401259        1.0000000
```

Analisi Esplorativa – Variabili Categoriche

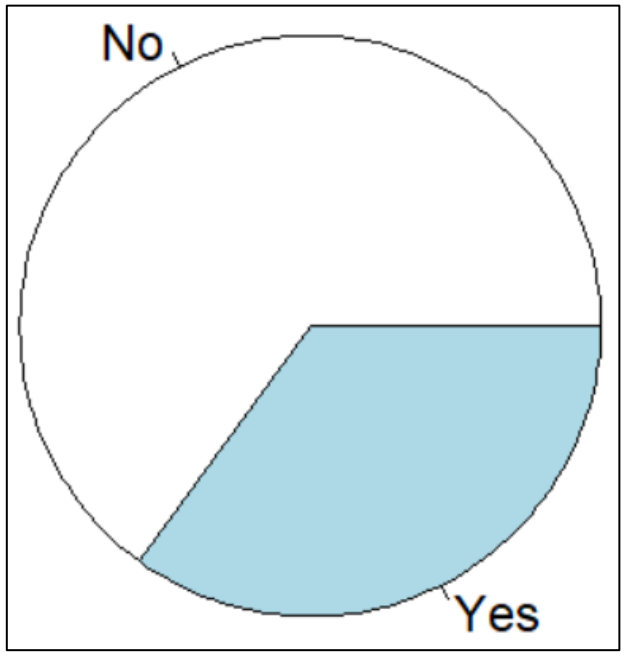


Utilizzo del contraccettivo da parte di mogli con medio alto o alto livello di educazione



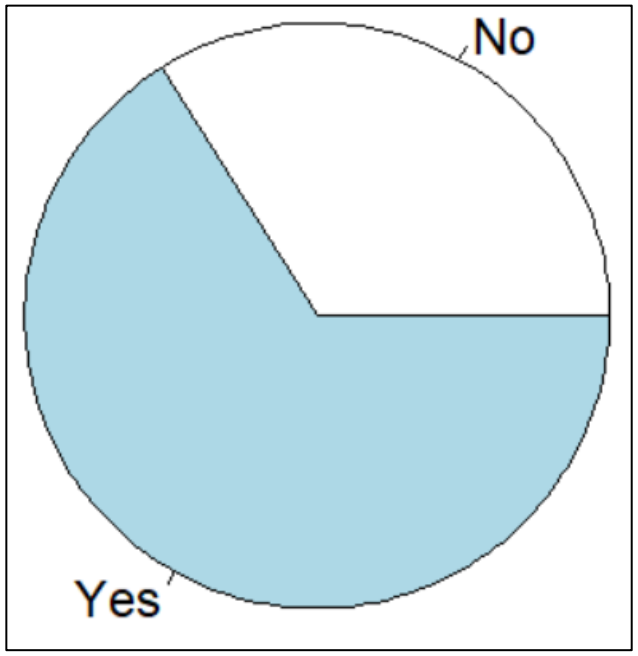
Utilizzo del contraccettivo da parte di coppie con alto o medio alto livello di vita

Analisi Esplorativa – Variabili Categoriche



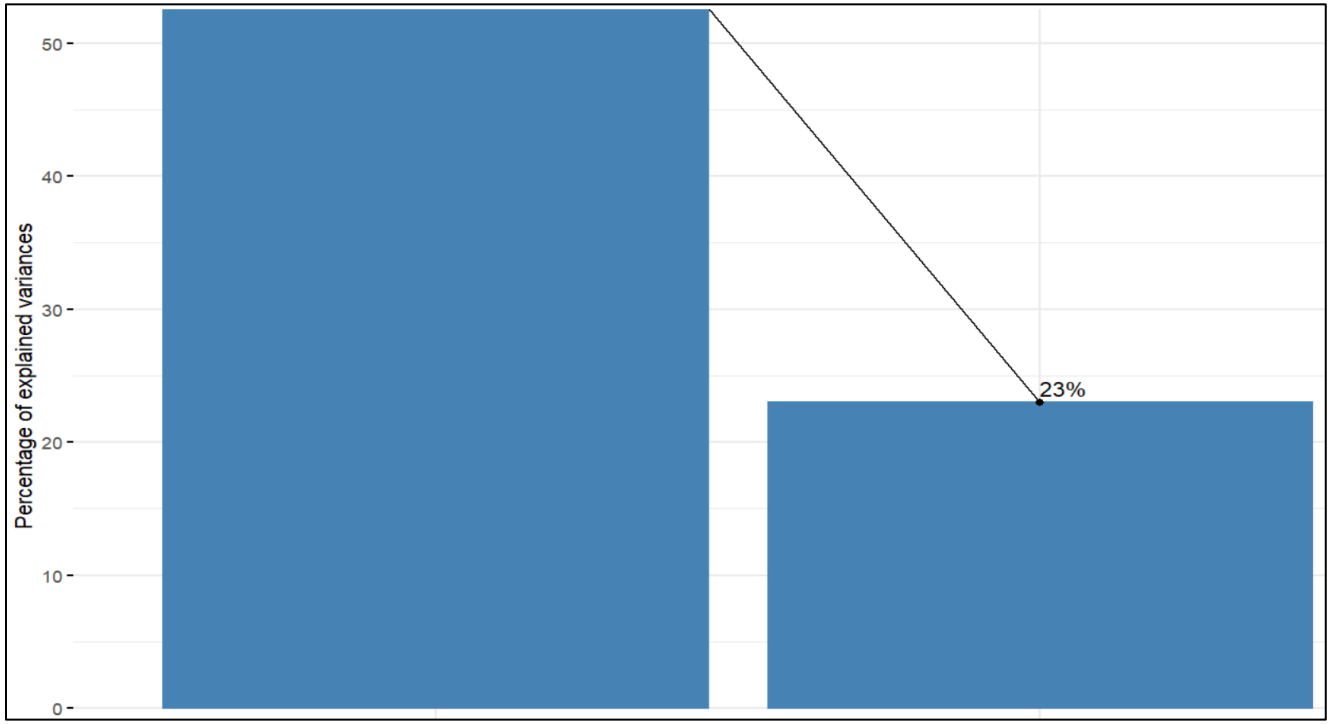
Utilizzo del contraccettivo da parte di mogli con basso o medio basso livello di educazione, e basso o medio basso standard di vita

1017 istanze analizzate



Utilizzo del contraccettivo da parte di mogli con alto o medio alto livello di educazione, e alto o medio alto standard di vita

Principal Component Analysis (PCA)



Percentuale di varianza spiegata

Correlazione tra dimensioni e variabili originali

```
> cmc.pca$var$cor
                    Dim.1      Dim.2
Wife_Age        0.8775323  0.4795175
Number_Children 0.8775323 -0.4795175
```

Modelli di Apprendimento scelti

Alberi Decisionali & Reti Neurali

Motivazioni:

- *Modelli di apprendimento supervisionato*
- *Gestione delle variabili categoriche*
- *Classificazione di pattern complessi*
- *Tolleranti alla presenza di outliers*
- *Alberi Decisionali facilmente interpretabili*

Tipologie:

- *Alberi Decisionali di tipo CART*
- *Reti Neurali con funzione logistica*

Cross-Validation

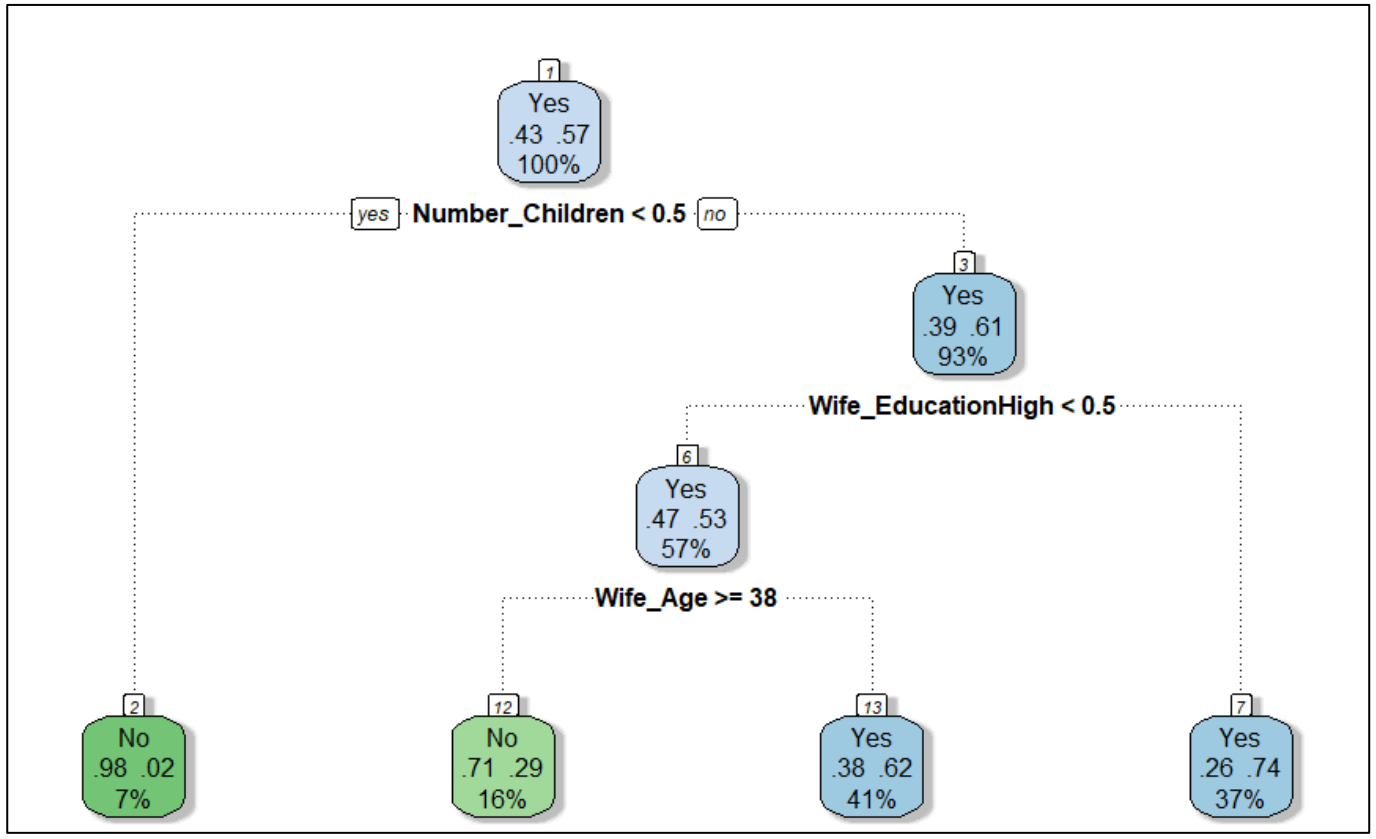
Problema per il testing:

- *Le istanze disponibili non sono molte: 1473*

Soluzione:

- *Cross-Validation*
- *In particolare per ogni modello sono state eseguite:*
  - *10-Fold CV sull'intero dataset per il problema binario*
  - *10-Fold CV sull'intero dataset per il problema multi-classe*
  - *10-Fold CV ripetuta 3 volte su train set per il problema binario*
  - *10-Fold CV ripetuta 3 volte su train set per il problema multi-classe*
- *I risultati migliori sono stati ottenuti con la 10-Fold CV sull'intero dataset. Viene ora riportato il confronto fra questi modelli.*

Alberi Decisionali – Problema Binario



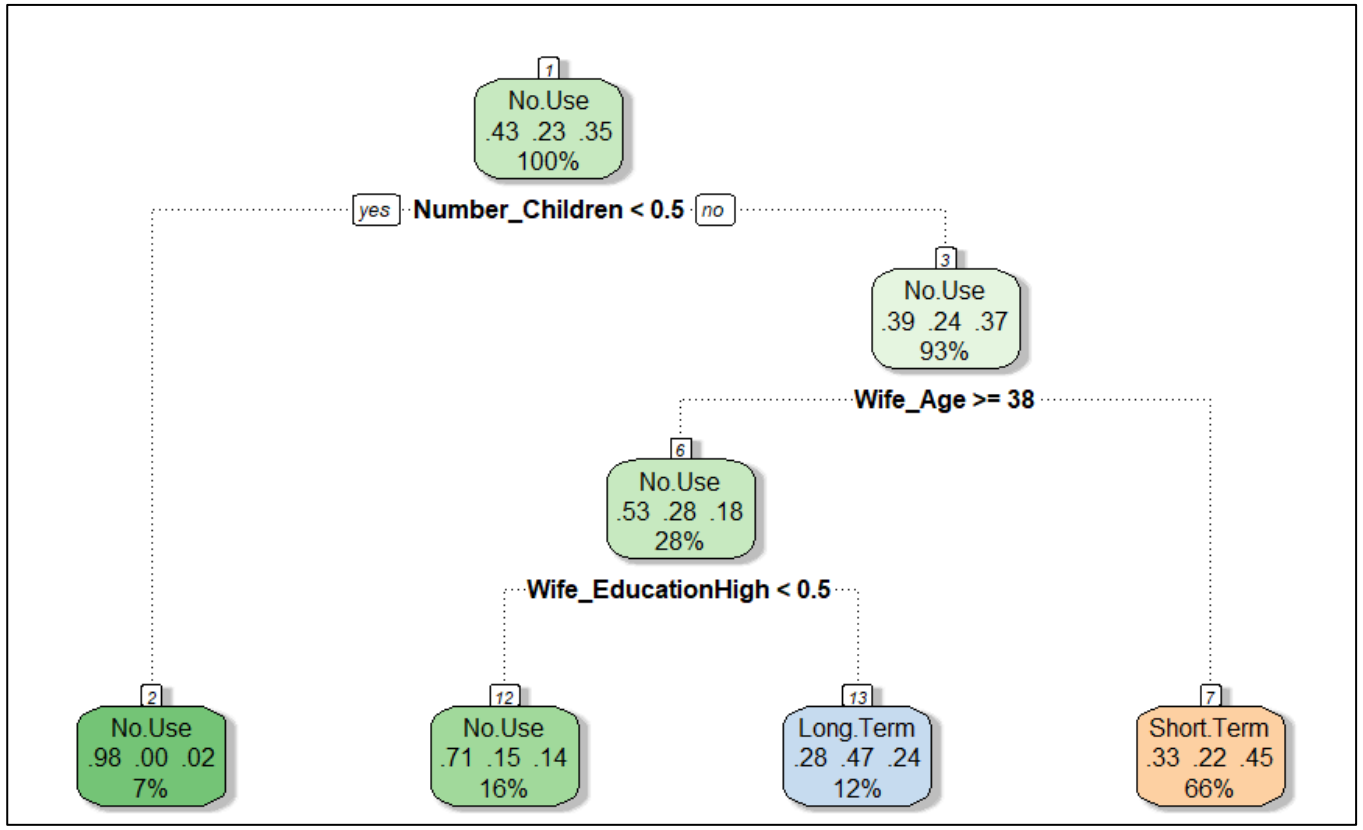
Albero Decisionale su dataset binario

rpart variable importance

	Overall
Wife_Age	100.000
Number_Children	93.536
Wife_EducationHigh	68.201
Media_ExposureNot-Good	49.427
Husband_EducationHigh	31.579
Wife_EducationMid-High	12.130
Husband_OccupationMid-Low	2.744
Living_IndexHigh	0.000
`Husband_OccupationMid-Low`	0.000
Wife_Is_WorkingNo	0.000
`Husband_EducationMid-Low`	0.000
`Wife_EducationMid-High`	0.000
`Media_ExposureNot-Good`	0.000
Husband_OccupationHigh	0.000
`Husband_EducationMid-High`	0.000
`Living_IndexMid-High`	0.000
`Husband_OccupationMid-High`	0.000
`Living_IndexMid-Low`	0.000
Wife_ReligionIslam	0.000
`Wife_EducationMid-Low`	0.000

Var importance

Alberi Decisionali – Problema Multi-Classe



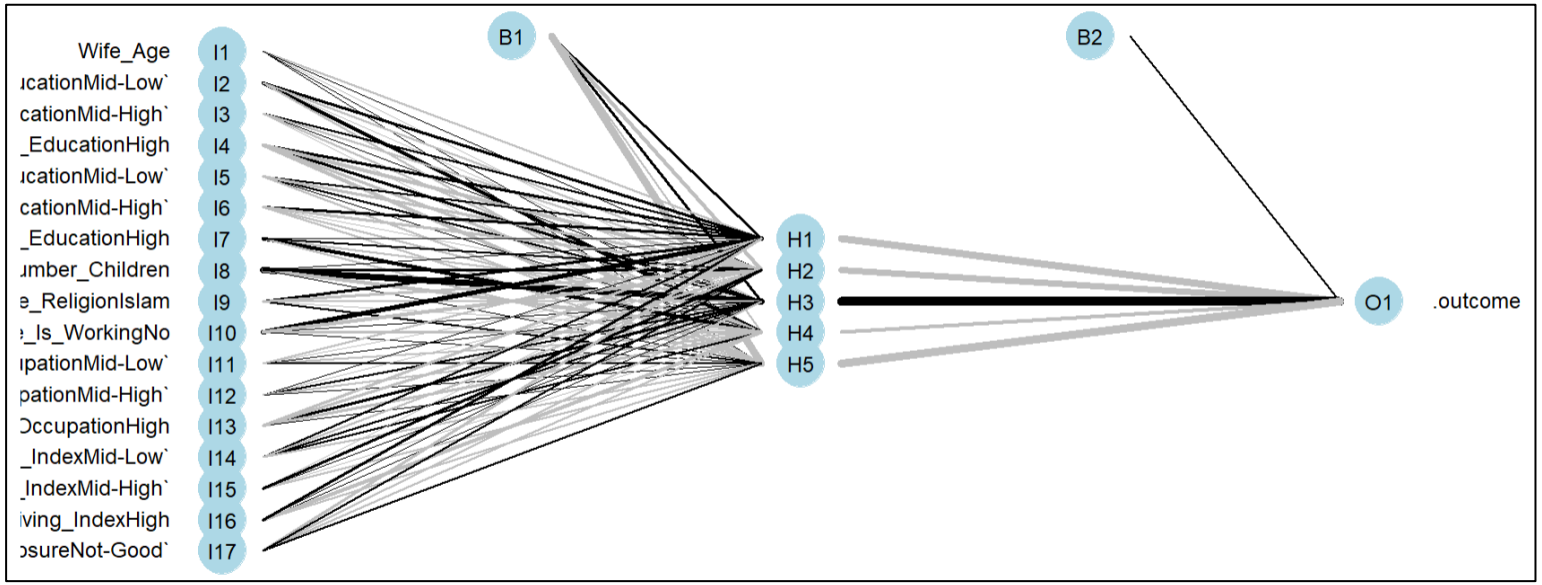
Albero Decisionale su dataset binario

rpart variable importance

	Overall
Wife_EducationHigh	100.00
Wife_Age	62.70
Husband_EducationHigh	51.44
Number_Children	51.16
Living_IndexHigh	38.79
Media_ExposureNot-Good	27.26
Husband_EducationMid-Low	12.02
`Wife_EducationMid-Low`	0.00
`Living_IndexMid-Low`	0.00
Husband_OccupationHigh	0.00
`Husband_EducationMid-Low`	0.00
`Wife_EducationMid-High`	0.00
`Husband_OccupationMid-High`	0.00
`Husband_OccupationMid-Low`	0.00
`Living_IndexMid-High`	0.00
Wife_ReligionIslam	0.00
`Media_ExposureNot-Good`	0.00
`Husband_EducationMid-High`	0.00
Wife_Is_WorkingNo	0.00

Var importance

Reti Neurali – Problema Binario

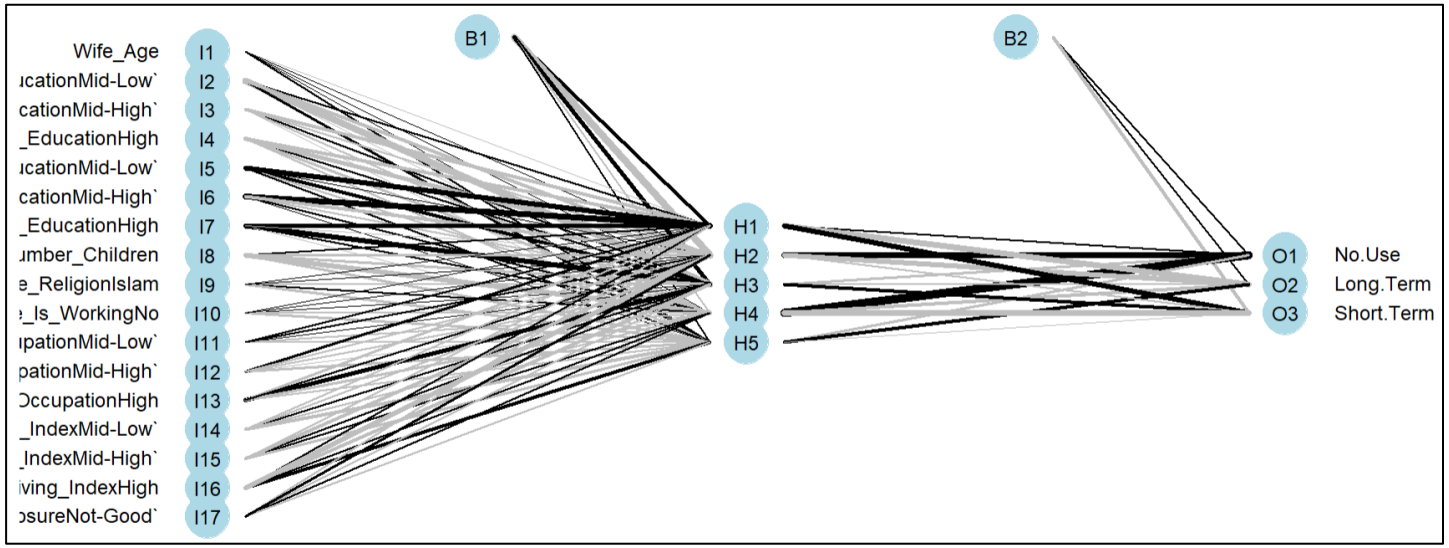


Rete Neurale su dataset binario

Var Importance

	Overall
Wife_EducationHigh	100.000
Number_Children	98.833
Wife_Is_WorkingNo	67.284
Husband_OccupationHigh	58.881
Wife_EducationMid-Low	47.182
Media_ExposureNot-Good	46.861
Wife_ReligionIslam	46.595
Living_IndexHigh	46.295
Husband_EducationMid-High	36.264
Husband_EducationMid-Low	34.676
Husband_EducationHigh	34.204
Living_IndexMid-Low	21.578
Living_IndexMid-High	19.622
Wife_EducationMid-High	16.687
Husband_OccupationMid-Low	12.809
Husband_OccupationMid-High	3.824
Wife_Age	0.000

Reti Neurali – Problema Multi-Classe



Rete Neurale su dataset multi-classe

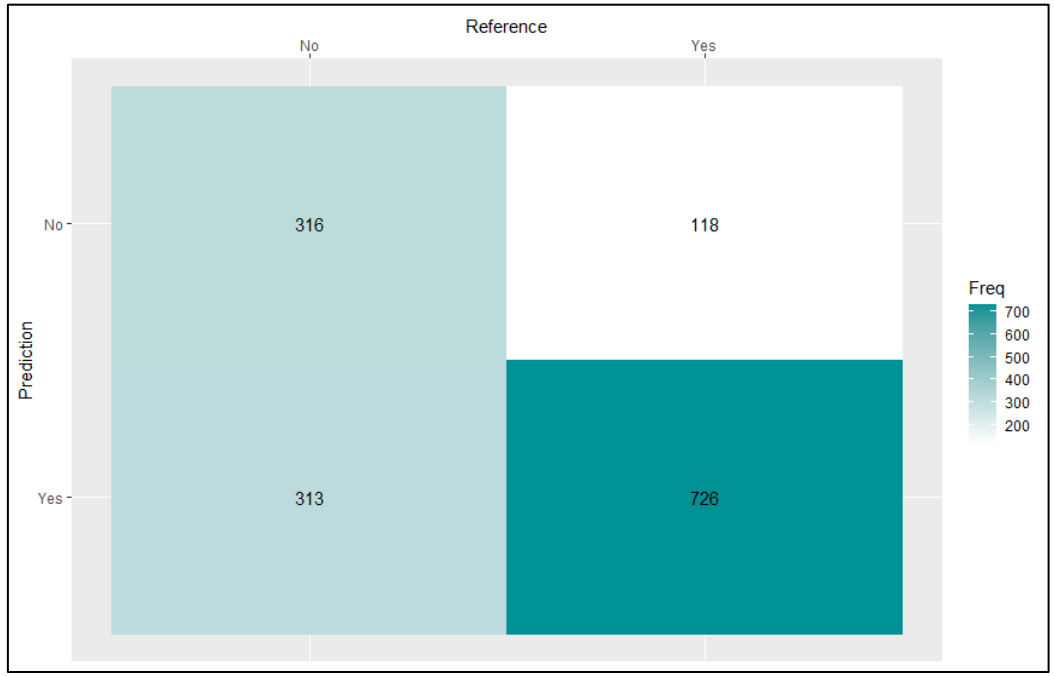
Var Importance

variables are sorted by maximum importance across the classe

	Overall	No.Use	Long.Term	Short.Term
Living_IndexHigh	100.00	100.00	100.00	100.00
Number_Children	95.51	95.51	95.51	95.51
Wife_EducationMid-Low	94.75	94.75	94.75	94.75
Husband_OccupationHigh	78.24	78.24	78.24	78.24
Wife_EducationHigh	73.58	73.58	73.58	73.58
Husband_EducationMid-High	71.30	71.30	71.30	71.30
Husband_EducationHigh	67.64	67.64	67.64	67.64
Husband_EducationMid-Low	66.22	66.22	66.22	66.22
Living_IndexMid-High	59.30	59.30	59.30	59.30
Living_IndexMid-Low	44.14	44.14	44.14	44.14
Wife_EducationMid-High	34.16	34.16	34.16	34.16
Husband_OccupationMid-High	32.09	32.09	32.09	32.09
Media_ExposureNot-Good	30.05	30.05	30.05	30.05
Wife_ReligionIslam	28.00	28.00	28.00	28.00
Wife_Is_WorkingNo	21.60	21.60	21.60	21.60
Husband_OccupationMid-Low	11.08	11.08	11.08	11.08
Wife_Age	0.00	0.00	0.00	0.00

Confronto tra Modelli – Matrici di Confusione

Problema Binario



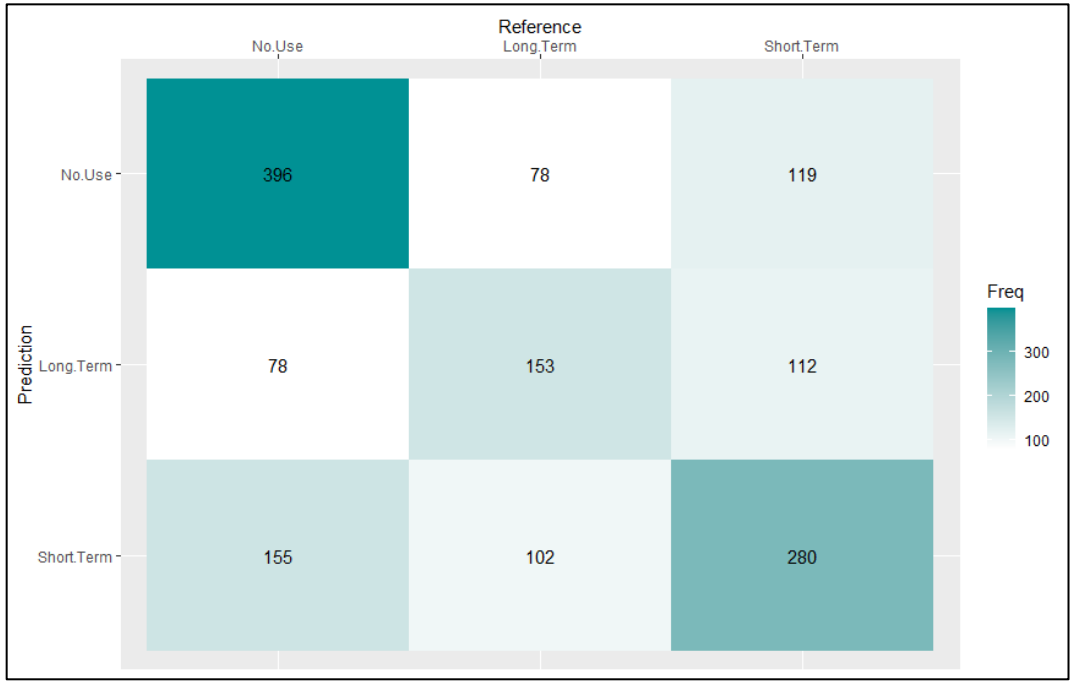
Accuracy: 0.7074
Albero Decisionale

Accuracy: 0.7101
Rete Neurale

Confronto tra Modelli – Matrici di Confusione

Problema Multi-Classe

Accuracy: 0.5261
Albero Decisionale



Accuracy: 0.5628
Rete Neurale

Confronto tra Modelli – Performance

Problema Binario

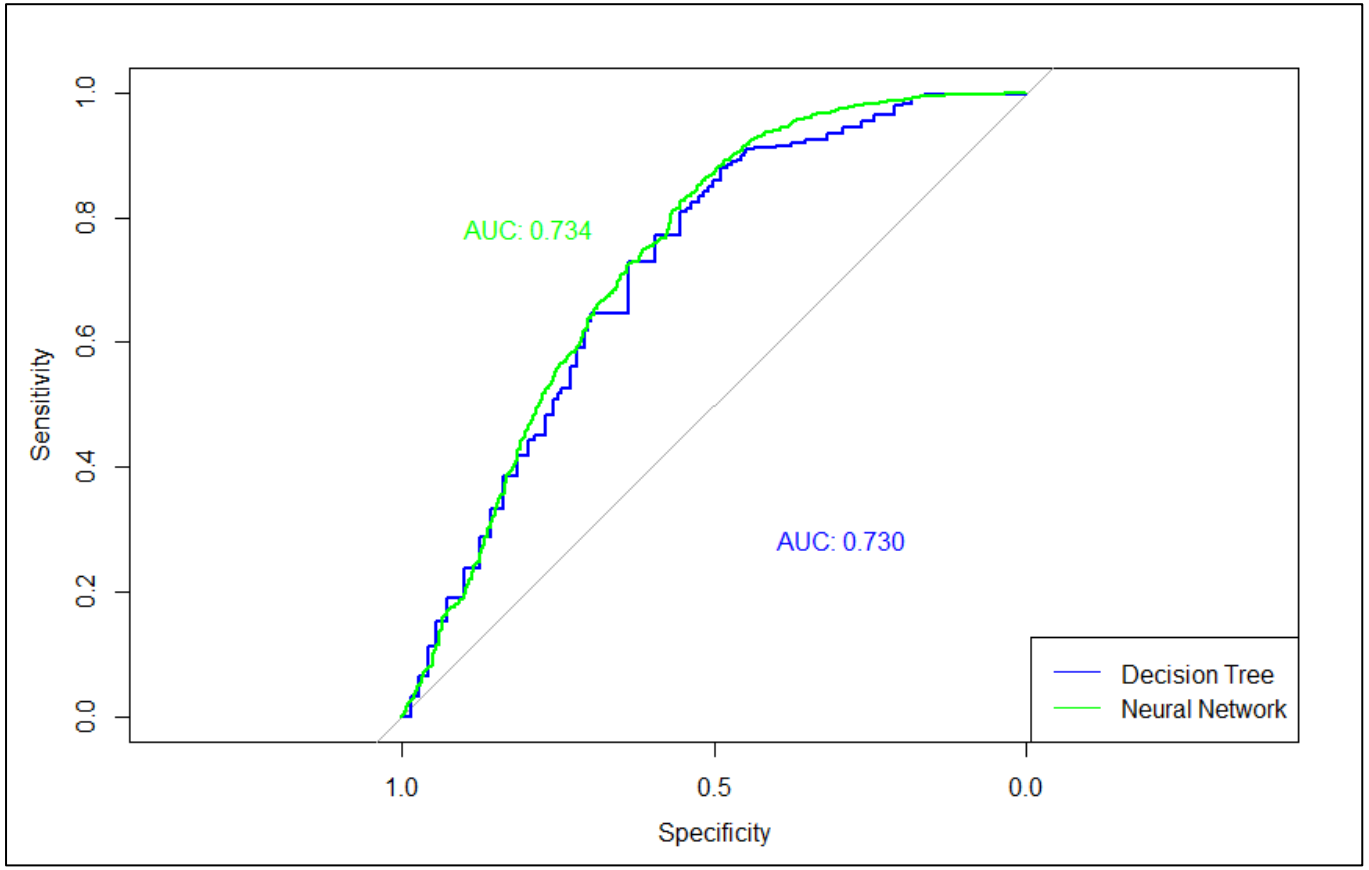
	DT binario	NN binario
Accuracy	0.7074 ± 0.0231	0.7101 ± 0.0231
Precision No	0.7281106	0.7186147
Precision Yes	0.6987488	0.7062315
Precision Macro Average	0.7134297	0.7124231
Recall No	0.5023847	0.5278219
Recall Yes	0.8601896	0.8459716
Recall Macro Average	0.6812872	0.6868968
F1-Measure No	0.5945437	0.6086159
F1-Measure Yes	0.7711099	0.7698113
F1-Measure Macro Average	0.6828268	0.6892136

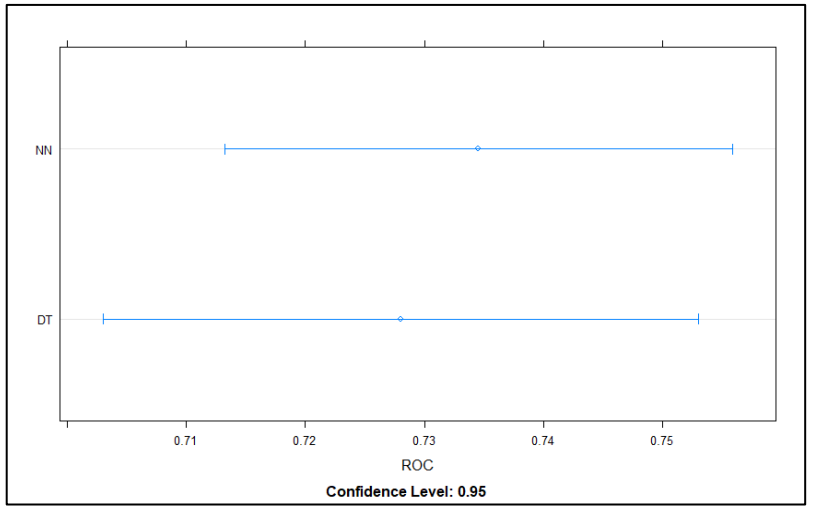
Confronto tra Modelli – Performance

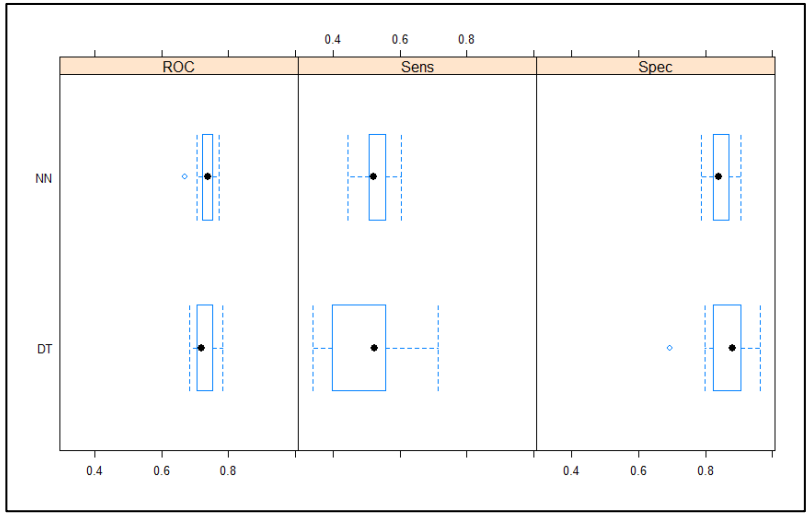
Problema Multi-Classe

	DT multi-classe	NN multi-classe
Accuracy	0.5261 ± 0.0258	0.5628 ± 0.0255
Precision No Use	0.6457143	0.6677909
Precision Long Term	0.4974874	0.4460641
Precision Short Term	0.4499332	0.5214153
Precision Macro Average	0.531045	0.5450901
Recall No Use	0.5389507	0.6295707
Recall Long Term	0.2972973	0.4594595
Recall Short Term	0.6594912	0.5479452
Recall Macro Average	0.4985797	0.5456585
F1-Measure No Use	0.5875217	0.6481178
F1-Measure Long Term	0.3721805	0.4526627
F1-Measure Short Term	0.5349206	0.5343511
F1-Measure Macro Average	0.4982076	0.5450439

Confronto tra Modelli – Curve ROC Problema Binario

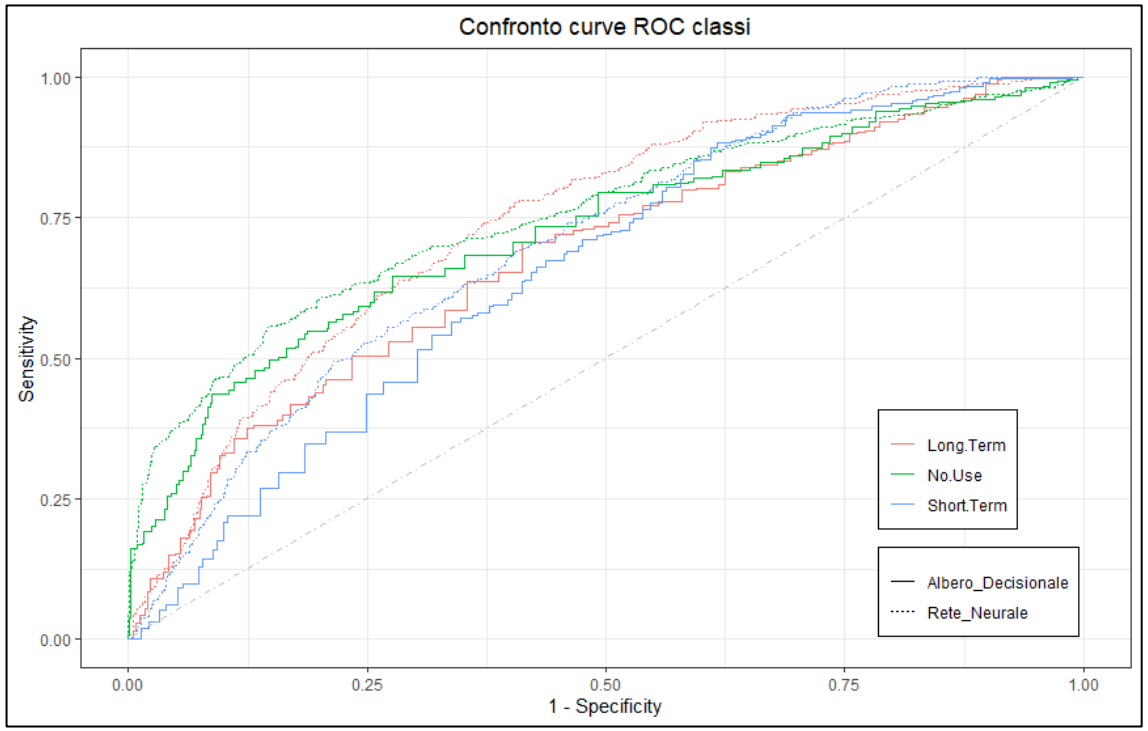


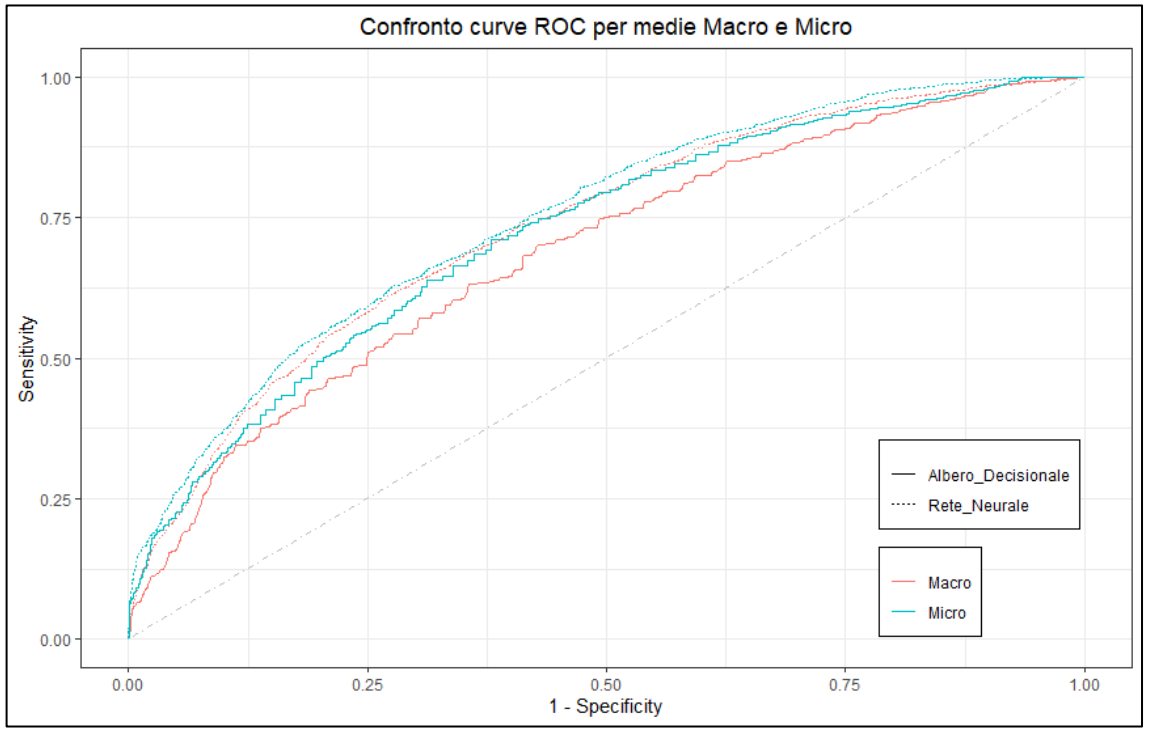




Curve ROC classe Yes

Confronto tra Modelli – Curve ROC Problema Multi-Classe





Confronto tra Modelli – AUC

Problema Binario

	DT binario	NN binario
AUC No	0.7301	0.7343
AUC Yes	0.7301	0.7343

Problema Multi-Classe

	DT multi-classe	NN multi-classe
AUC No Use	0.717548	0.7549032
AUC Long Term	0.6654286	0.7409646
AUC Short Term	0.6614156	0.7004345
AUC Macro Average	0.6814631	0.7320958
AUC Micro Average	0.7169151	0.7480347

Confronto tra Modelli – Tempi Computazione

	Everything (s)	Final (s)	Prediction (s)
DT binario	0.76	0.01	NA
NN binario	8.76	0.17	NA
DT multi-classe	1.02	0.02	NA
NN multi-classe	14.89	0.59	NA

Conclusioni

- Le Reti Neurali hanno registrato performance predittive migliori
- Gli Alberi Decisionali hanno beneficiato di tempi di computazione migliori
- Le performance predittive dei modelli sono complessivamente modeste:
  - Dataset sbilanciato
  - Poche istanze
- Possibili rimedi:
  - Aumentare il numero di istanze
  - Aggiungere ulteriori attributi quantitativi e qualitativi

GRAZIE PER L’ATTENZIONE