Travel Insurance

Predict whether a customer will be interested in buying travel insurance

ANDRONI STEFANO
Mat. 845811
s.androni1@campus.unimib.it

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

1 DATASET

Source

kaggle *https://www.kaggle.com/tejashvi14/travel-insurance-prediction-data*

Textual description

Il dataset 'Travel Insurance' contiene i dati raccolti da una società di viaggi riguardo la sottoscrizione o meno da parte dei propri clienti dell'assicurazione di viaggio. L'assicurazione è stata offerta ad alcuni clienti nel 2019 e i dati forniti sono stati estratti dalle prestazioni/vendite del pacchetto assicurativo durante quel periodo.

Numerical description

Il dataset conta 1 987 istanze e 10 attributi.

Features

*int>0*
Indice

Age *int>0*
Età del cliente

Employment.Type *{Government Sector, Private Sector/Self Employed}*
Il settore in cui è impiegato il cliente

GraduateOrNot *{Yes,No}*
Se il cliente è laureato o no

AnnualIncome *int>0*
Il reddito annuo del cliente in rupie indiane (arrotondato alle 50 mila rupie più vicine)

FamilyMembers *int>0*
Numero di membri nella famiglia del cliente

ChronicDisease *{1,0}*
Se il cliente soffre di malattie o condizioni gravi come diabete o ipertensione arteriosa, ecc.

FrequentFlyer *{Yes,No}*
Dati derivati basati sulla cronologia della prenotazione di biglietti aerei da parte del cliente (considerato un viaggiatore frequente se ha prenotato almeno 4 istanze diverse negli ultimi 2 anni [2017-2019])

EverTravelledAbroad *{Yes,No}*
Se il cliente ha mai viaggiato in un paese straniero (non necessariamente utilizzando i servizi dell'azienda)

TravelInsurance *{1,0}*
Se il cliente ha acquistato un pacchetto di assicurazione di viaggio durante l'offerta introduttiva tenutasi nell'anno 2019.

2 OBIETTIVO

Gli obiettivi che si vogliono raggiungere possono essere identificati in:

> L'azienda richiede di conoscere **quali tipi di clienti potrebbero essere interessati ad acquistare il pacchetto assicurativo** di viaggio in base alla cronologia del database. Il raggiungimento di questo primo obiettivo è finalizzato alla strutturazione di una campagna pubblicitaria mirata.

> L'azienda richiede di identificare **uno o più modelli di ML in grado di predire con buona capacità se un cliente acquisterà o meno il pacchetto assicurativo** di viaggio.

3 DATA PREPROCESSING

Before

$ X	: int	0 1 2 3 4 ...
$ Age	: int	31 31 34 28 ...
$ Employment.Type	: Factor	{"Government Sector", "Private Sector/Self Employed"}
$ GraduateOrNot	: Factor	{"No", "Yes"}
$ AnnualIncome	: int	400000 1250000 500000 700000 700000 …
$ FamilyMembers	: int	6 7 4 3 8 …
$ ChronicDiseases	: int	1 0 1 1 1 …
$ FrequentFlyer	: Factor	{"No", "Yes"}
$ EverTravelledAbroad	: Factor	{"No", "Yes"}
$ TravelInsurance	: int	0 0 1 0 0 …

Ridenominazione delle colonne

from $Employment.Type to $GovernmentEmployment
from $GraduateOrNot to $Graduate

Sistemazione dei valori delle colonne

$Graduate values from {"No","Yes"} to {0,1}
$FrequentFlyer values from {"No","Yes"} to {0,1}
$EverTravelledAbroad values from {"No","Yes"} to {0,1}
$GovernmentEmployment values from {"Private Sector/Self Employed", "Government Sector"} to {0,1}

Sistemazione del tipo delle colonne

$ChronicDiseases from int to factor
$TravelInsurance from int to factor

Nota $ChronicDiseases e $TravelInsurance sono variabili di tipo booleano (0:false, 1:true)

Rimozione delle colonne

$X removed

Nota $X è un attributo che contiene gli id delle istanze e non è quindi rilevante ai fini dell'analisi

After

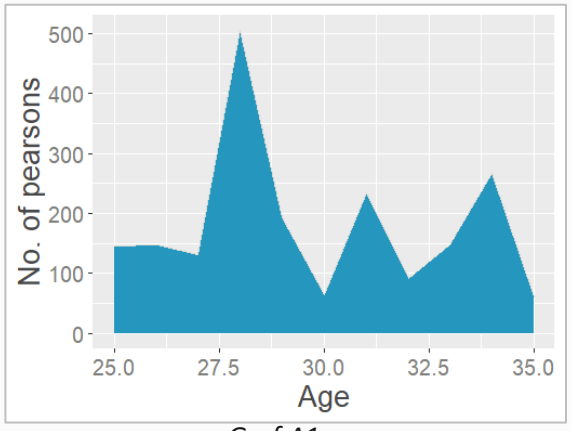
$ Age	: int	31 31 34 28 ...
$ GovernmentEmployment	: Factor	{0,1}
$ Graduate	: Factor	{0,1}
$ AnnualIncome	: int	400000 1250000 500000 700000 700000 …
$ FamilyMembers	: int	6 7 4 3 8 …
$ ChronicDiseases	: Factor	{0,1}
$ FrequentFlyer	: Factor	{0,1}
$ EverTravelledAbroad	: Factor	{0,1}
$ TravelInsurance	: Factor	{0,1}

4 EXPLORATORY ANALYSIS

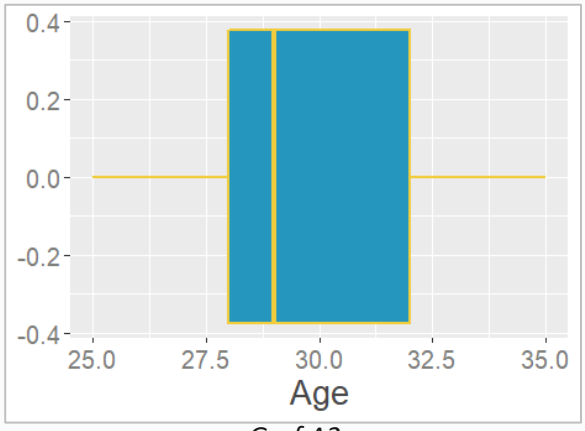
A – UNIVARIATE ANALYSIS

In questa sezione viene affrontata brevemente l'analisi univariata delle covariate.

Age



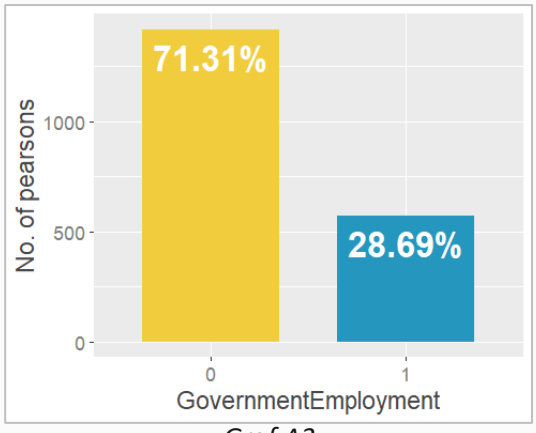
Graf.A1



Graf.A2

Il cliente più giovane ha un'età di **25** anni mentre il più anziano di **35** anni. La media dell'età dei clienti è di **30** anni, mentre la moda è di 28 anni. L'età dei clienti approssima una distribuzione trimodale e non simmetrica. La deviazione standard dell'età dei clienti è pari a 2,91 ed indica una dispersione dei valori nel set di dati non troppo grande.

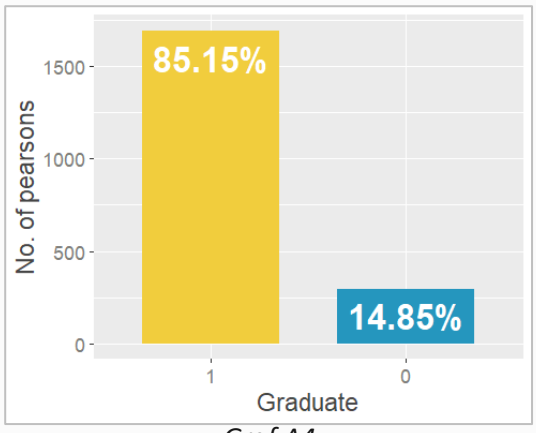
GovernmentEmployment



Graf.A3

La maggior parte dei clienti, circa il 71%, lavora nel settore privato o è un lavoratore autonomo.

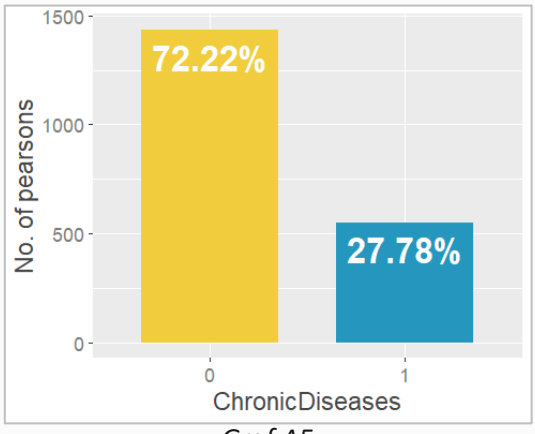
Graduate



Graf.A4

La maggior parte dei clienti, circa l'85%, ha conseguito una laurea.

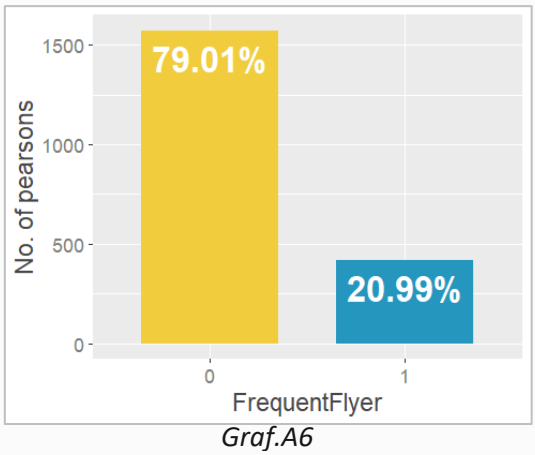
ChronicDisease



Graf.A5

Quasi il 28% dei clienti soffre di malattie o condizioni gravi (come diabete o ipertensione arteriosa, ecc).

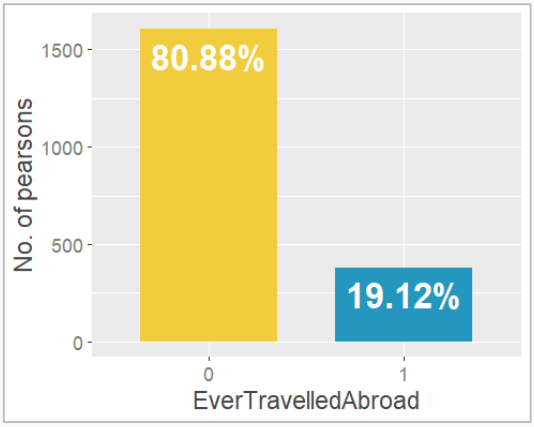
FrequentFlyer



Graf.A6

Circa il 21% dei clienti viene considerato un viaggiatore frequente (ha prenotato almeno 4 istanze diverse negli ultimi 2 anni [2017-2019]).

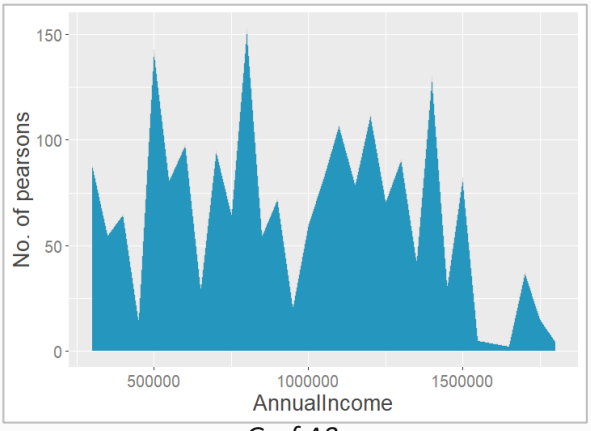
EverTravelledAbroad



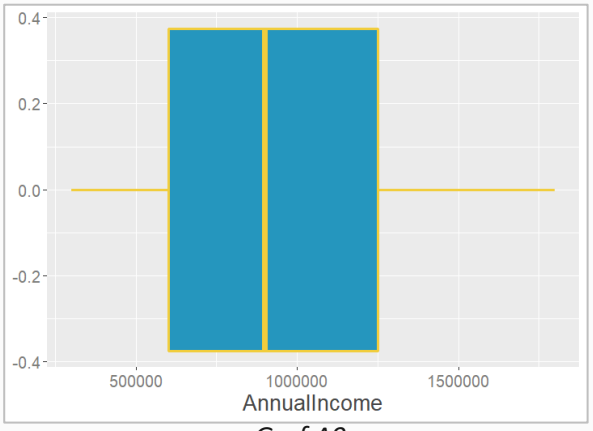
Graf.A7

Circa il 19% dei clienti ha viaggiato in un paese straniero (non necessariamente utilizzando i servizi dell'azienda).

AnnualIncome



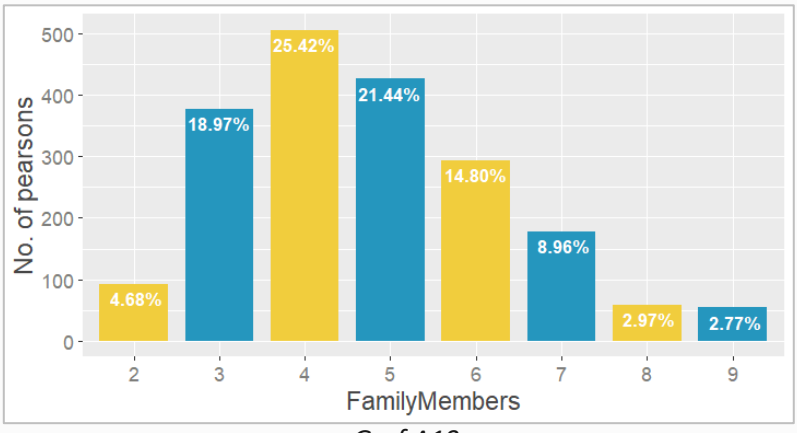
Graf.A8



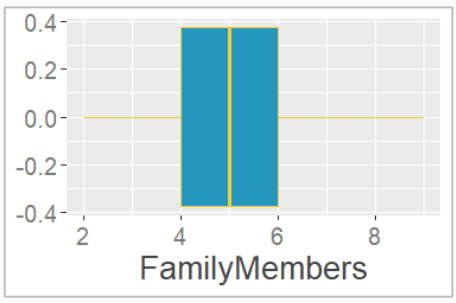
Graf.A9

Il reddito annuale medio dei clienti è di **932 763** rupie, mentre la mediana è pari 900 000 rupie. Il reddito annuale minimo è di **300 000** rupie, mentre quello massimo è pari a **1 800 000** rupie. La deviazione standard della covariata considerata è pari a 376 855 ed indica una forte dispersione dei valori nel set di dati. La distribuzione può considerarsi simmetrica.

FamilyMembers



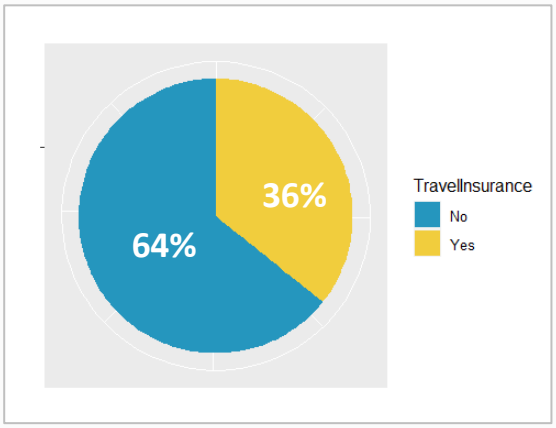
Graf.A10



Graf.A11

La famiglia più piccola è composta da **2** componenti, mentre quella più numerosa da **9** componenti. La media della dimensione della famiglia dei clienti è pari a **4,75**, mentre la mediana è pari a 5. La moda è pari a 4. La deviazione standard è pari a 1,6. La distribuzione, come confermano i dati e i grafici, approssima una normale.

TravelInsurance



Graf.A12

Circa il 36% dei clienti ha scelto di stipulare l'assicurazione di viaggio.

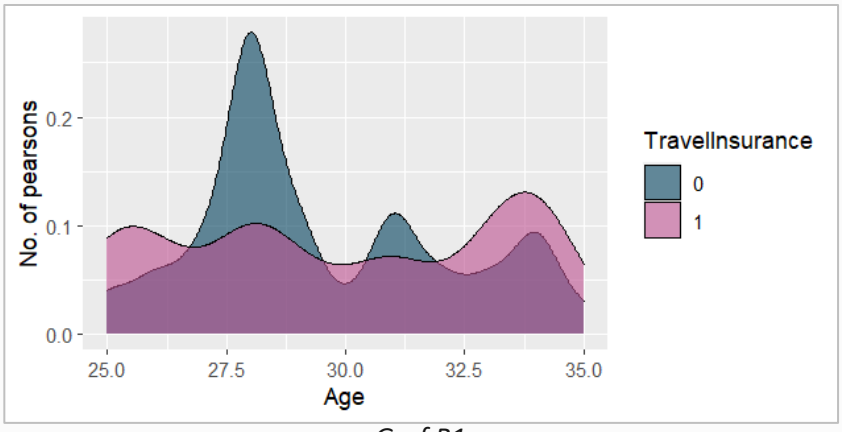
Note

> TravelInsurance è la **variabile target** per il problema considerato
> La variabile target è **binaria**
> La distribuzione della variabile target evidenzia che **il problema non è sbilanciato**, ma nemmeno perfettamente bilanciato (alcuni algoritmi, come Naive Bayes, soffrono di un forte sbilanciamento della variabile target).

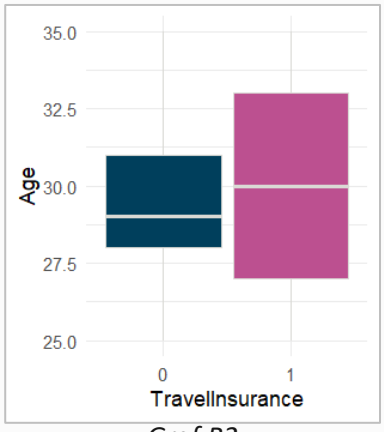
B – MULTIVARIATE & DISCRIMINATORY ANALYSIS

In questa sezione viene brevemente affrontata l'analisi multivariata delle covariate con particolar attenzione all'**analisi della discriminabilità delle classi della variabile target** (viene analizzata la variabile target in relazione con ognuna delle altre covariate).

TravelInsurance | Age



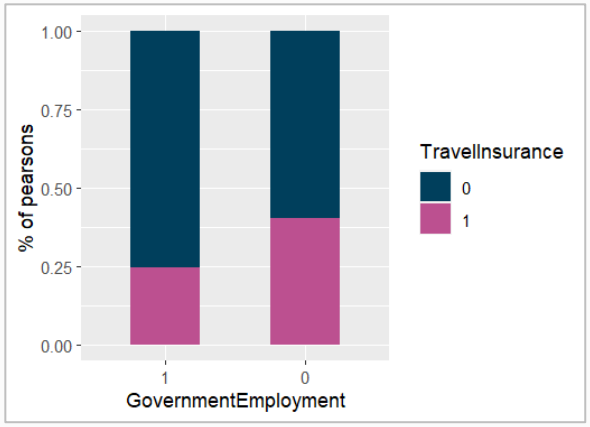
Graf.B1



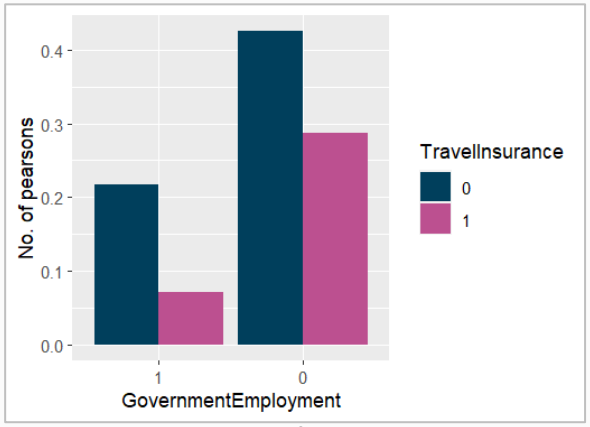
Graf.B2

La distribuzione dell'età dei clienti in base alla stipulazione o meno dell'assicurazione di viaggio evidenzia un comportamento diverso in differenti fasce di età. In particolar modo nelle fasce di anni 27-29 e 31-32 i clienti sono meno propensi all'acquisto dell'assicurazione di viaggio. Il comportamento è opposto nelle fasce di età complementari. Quindi l'attributo Age discrimina discretamente le classi dell'attributo target.

TravelInsurance | GovernmentEmployment



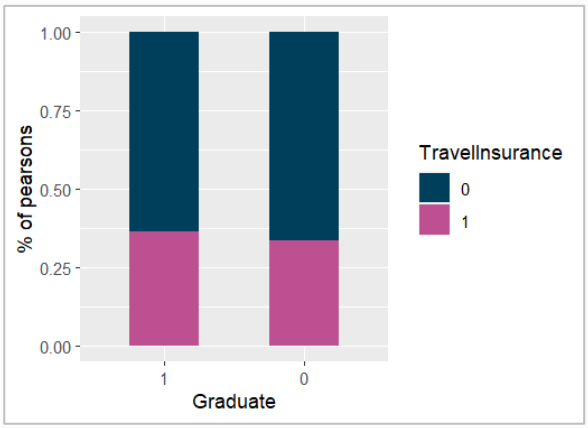
Graf.B3



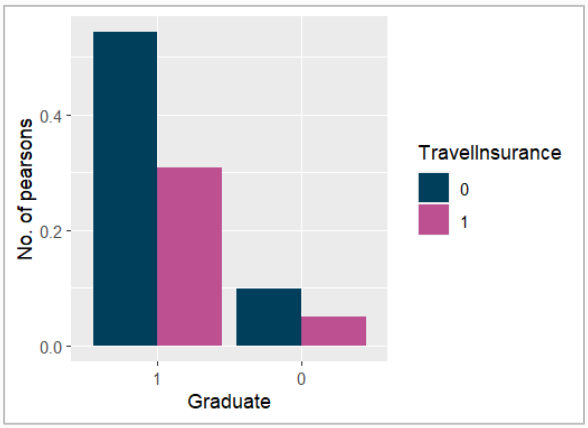
Graf.B4

I dipendenti pubblici sono meno propensi a stipulare l'assicurazione di viaggio rispetto ai dipendenti del settore privato e ai lavoratori autonomi. Tuttavia, la differenza è minima e quindi l'attributo GovernmentEmploment da solo non permette di discriminare le classi.

TravelInsurance | Graduate



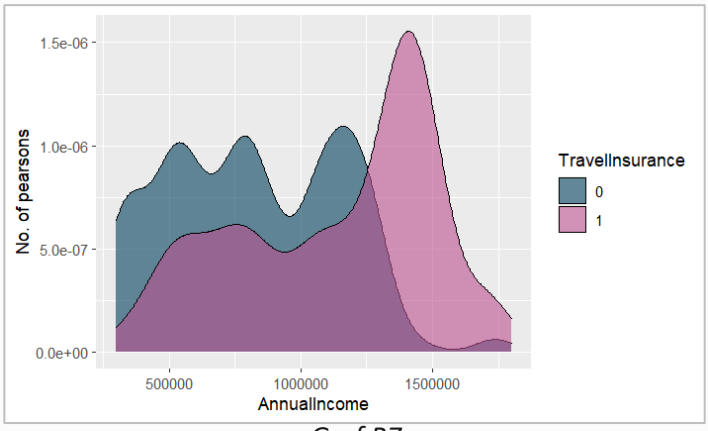
Graf.B5



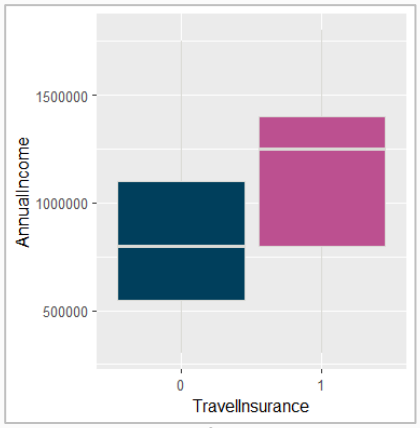
Graf.B6

Le persone laureate sono leggermente più propense a stipulare l'assicurazione di viaggio. Tuttavia, la differenza è minima e quindi l'attributo Graduate da solo non permette di discriminare le classi.

TravelInsurance | AnnualIncome



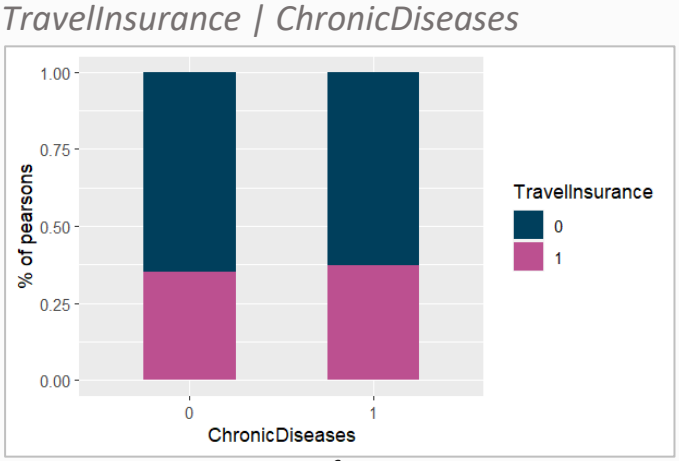
Graf.B7



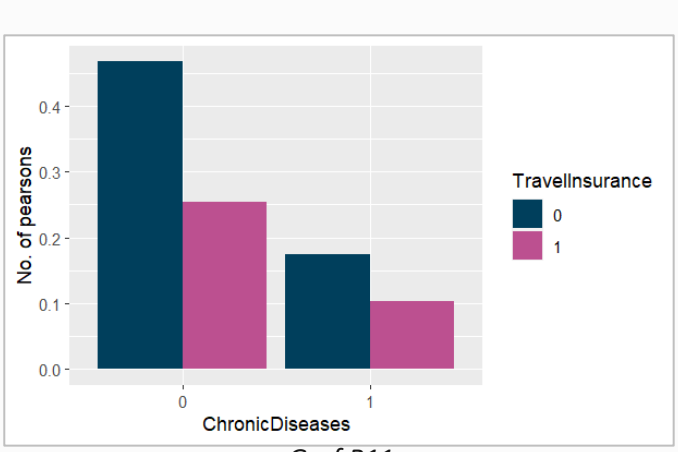
Graf.B17

La distribuzione del reddito dei clienti in base alla stipulazione o meno dell'assicurazione di viaggio evidenzia un comportamento diverso in differenti fasce di reddito. In particolar modo i clienti con un reddito maggiore di 1 300 000 rupie sono molto più propensi all'acquisto dell'assicurazione di viaggio. Il comportamento è opposto nella fascia di reddito complementare. Quindi l'attributo AnnualIncome discrimina discretamente le classi dell'attributo target.

TravelInsurance | ChronicDiseases



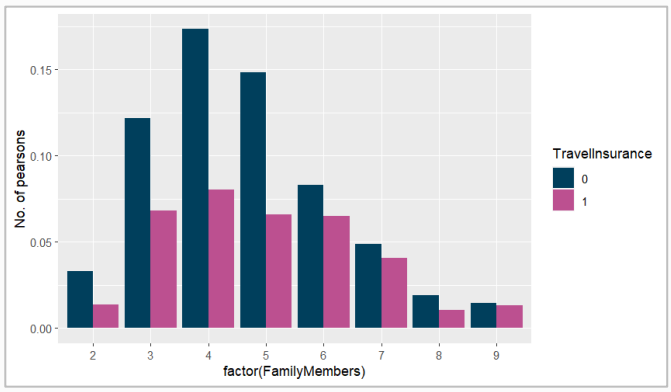
Graf.B10



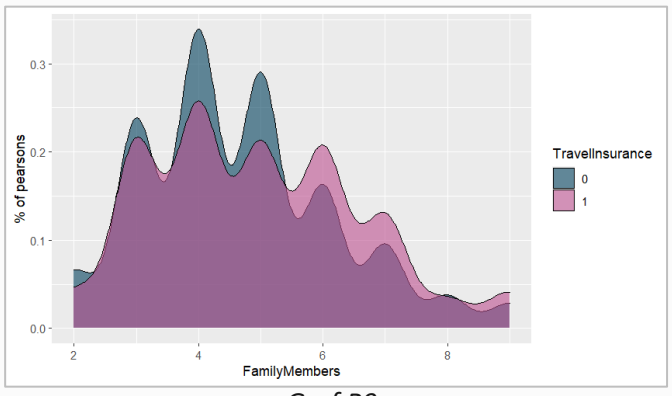
Graf.B11

ChronicDiseases non sembra avere un impatto sulla discriminabilità delle classi della variabile target.

TravelInsurance | FamilyMembers



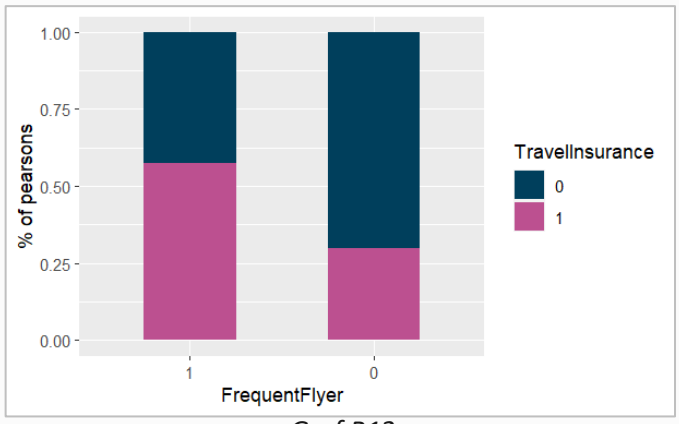
Graf.B8



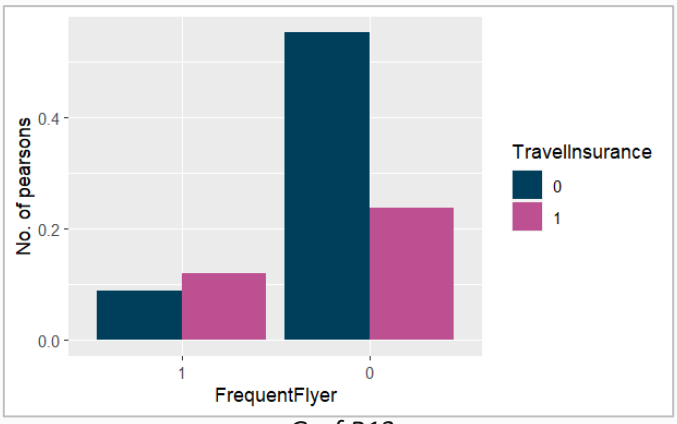
Graf.B9

La distribuzione della dimensione della famiglia dei clienti in base alla stipulazione o meno dell'assicurazione di viaggio evidenzia un comportamento diverso in differenti fasce. In particolar modo i clienti con una famiglia di dimensioni maggiore di 5 sono leggermente più propensi all'acquisto dell'assicurazione di viaggio. Il comportamento è opposto nella fascia di reddito complementare. Quindi l'attributo AnnualIncome discrimina, anche se solo in maniera debole, le classi dell'attributo target (le classi sono per lo più sovrapposte).

TravelInsurance | FrequentFlyer



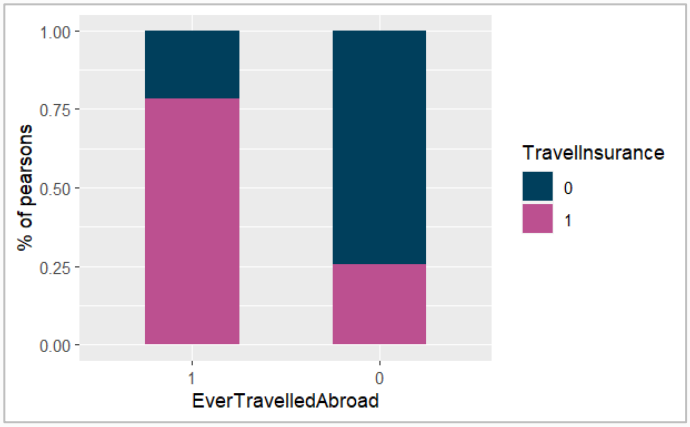
Graf.B12



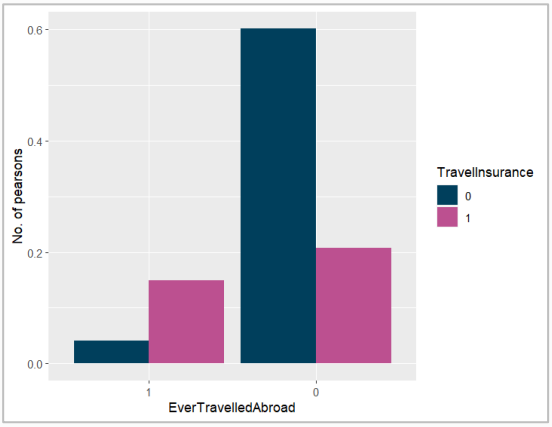
Graf.B13

I viaggiatori frequenti sono più propensi a stipulare l'assicurazione di viaggio rispetto agli altri viaggiatori. La covariata considerata permette, anche se solo debolmente, di discriminare la classi dell'attributo target.

TravelInsurance | EverTravelledAbroad



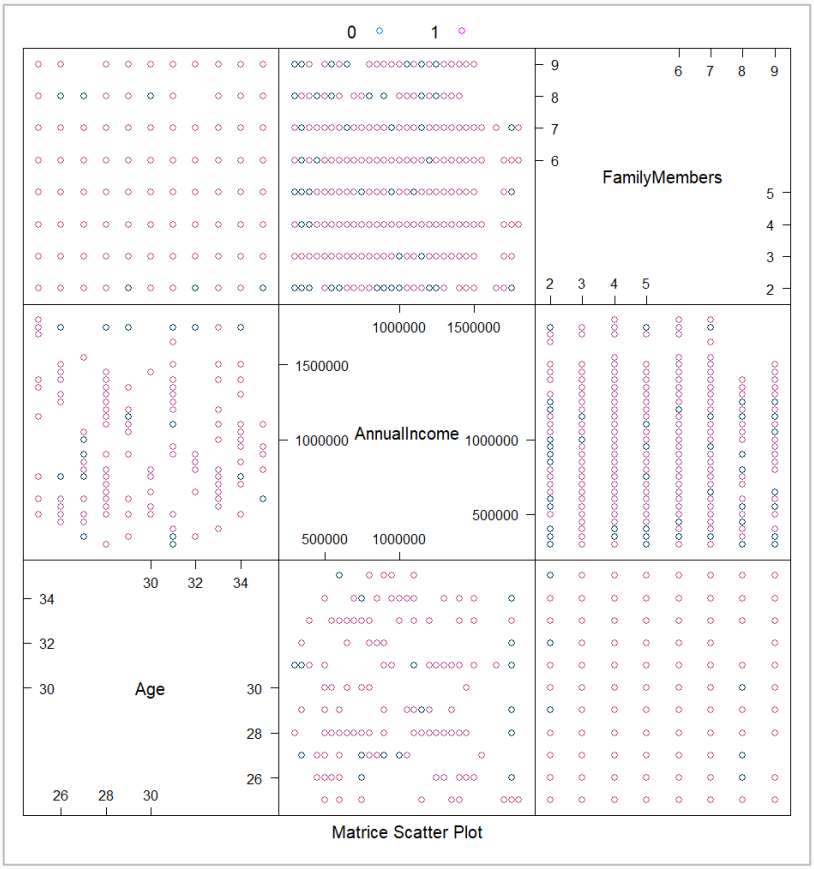
Graf.B14



Graf.B15

I viaggiatori che hanno viaggiato anche in paesi stranieri sono molto più propensi ad acquistare l'assicurazione di viaggio rispetto agli altri viaggiatori. La covariata considerata permette quindi di discriminare la classi dell'attributo target.

*TravelInsurance | *Covariate numeriche**



Graf.B16

Le covariate numeriche, prese a coppie (nello spazio bidimensionale), non sembrano rendere migliore la discriminabilità delle classi dell'attributo target. (si sovvrappongono → usare jitter(?))

DIFFICOLTÀ DEL PROBLEMA

I valori delle covariate rappresentano/separano abbastanza bene le classi della variabile target: probabilmente non sarà necessario effettuare nessuna trasformazione dello spazio di input e ci si può aspettare un modello con una buona capacità predittiva.

C – FAMD

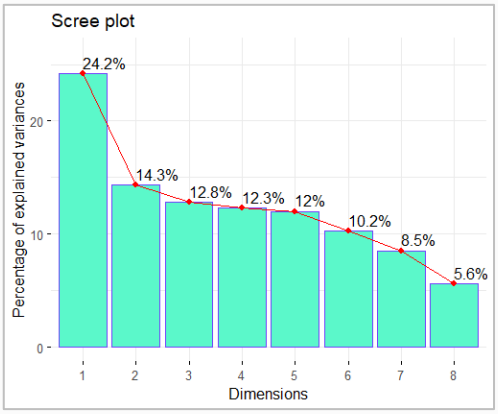
L'analisi delle componenti principali (PCA) è il metodo di riduzione delle variabili più indicato solo quando le variabili sono continue e non è conveniente quindi utilizzarla nel caso sotto studio. Tra le strategie più adatte nel caso in cui le variabili siano sia categoriche (/binarie) che continue, è stata scelta la FAMD. **FAMD** (Factor analysis of mixed data) è un metodo componente principale dedicato all'analisi di un set di dati contenente variabili sia quantitative che qualitative, che funge da PCA per variabili quantitative e da MCA per variabili qualitative.

A scopo esplorativo, FAMD è stata applicata con ncp (numero delle dimensioni/componenti da mantenere nel risultato finale) pari al numero delle dimensioni/attributi originali (lo spazio di input viene mappato in uno spazio delle componenti grande tanto quanto lo spazio originale).

Nota Analisi utile se si rimane nel nuovo spazio (Feature Extraction)

Varianza spiegata per dimensioni

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.9344499	24.180624	24.18062
Dim.2	1.1452486	14.315607	38.49623
Dim.3	1.0243566	12.804457	51.30069
Dim.4	0.9857286	12.321607	63.62230
Dim.5	0.9611335	12.014169	75.63646
Dim.6	0.8191641	10.239551	85.87602
Dim.7	0.6818238	8.522798	94.39881
Dim.8	0.4480949	5.601186	100.00000



Graf.C1

Criteri di scelta delle componenti

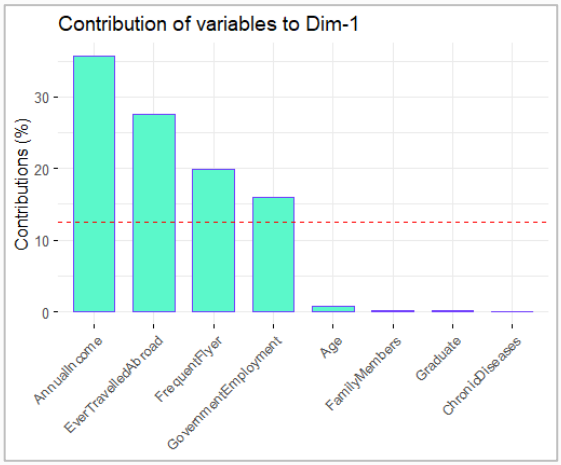
(1) Autovalori > 1 → Scelte le prime 3 dimensioni

(2) Varianza spiegata > 75% → Scelte le prime 5 dimensioni

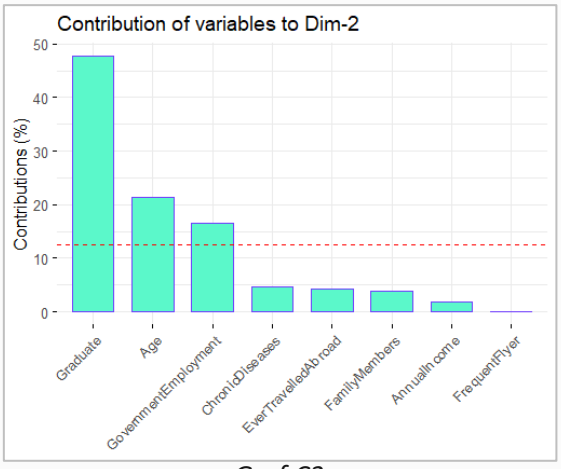
Nota Analisi utile se si rimane nello spazio originale (Feature Selection)

Contributo di varianza delle variabili orginali per le dimensioni

La prima e la seconda dimensione (Dim.1 e Dim.2) sono quelle che spiegano più varianza (rispettivamente 24,18% e 14,32%). Risulta interessante quindi analizzare la contribuzione delle variabili/attributi originali per queste dimensioni.



Graf.C2

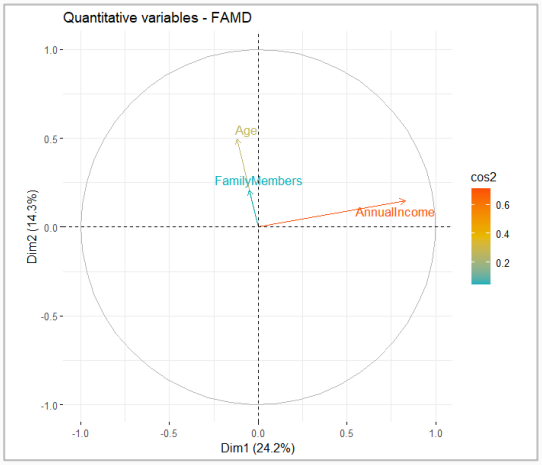


Graf.C3

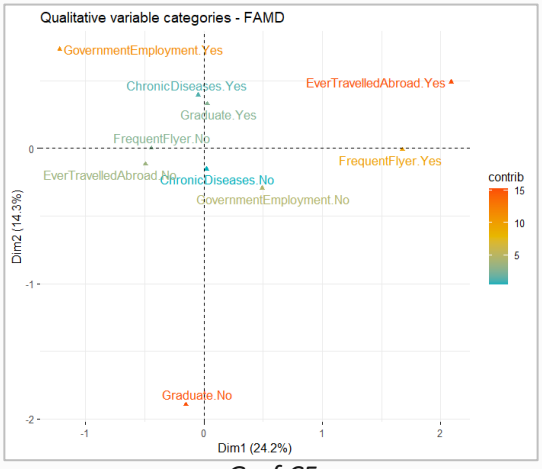
Le variabili che contribuiscono maggiormente alla <u>prima dimensione</u> sono: AnnualIncome (35,63%), EverTravelledAbroad (27,48%), FrequentFlyer (19,92%) e GovernmentEmployment (15,98%).

Le variabili che contribuiscono maggiormente alla <u>seconda dimensione</u> sono: Graduate (47,66%), Age (21,31%) e GovernmentEmployment (16,44%).

Variabili nel nuovo spazio



Graf.C4



Graf.C5

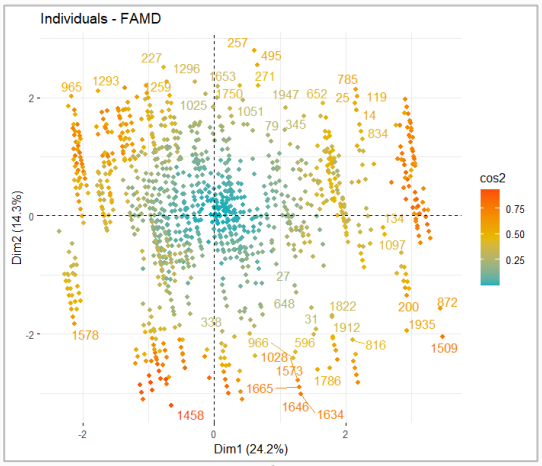
<u>Variabili quantitative</u>

Tra le covariate quantitative Age e FamilyMembers c'è un'alta correlazione (sono addirittura sovrapposte). Tra queste, la covariata Age ha una qualità migliore (è più lontana dall'origine). Complessivamente Age e AnnualIncome sono comunque qualitativamente buone.

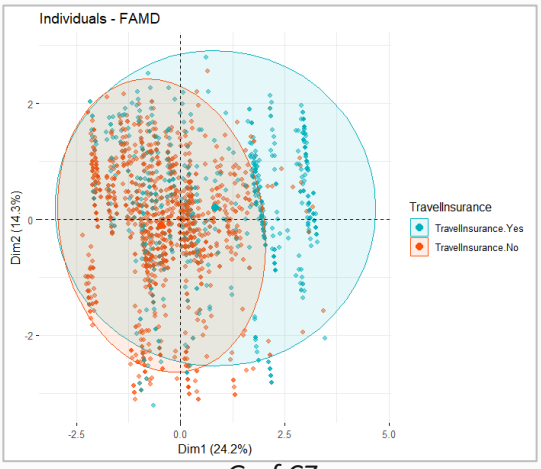
<u>Variabili qualitative</u>

La covariata Graduate.No ha un contributo importante nel polo negativo della seconda dimensione. Le covariate EverTravelledAbroad.Yes e FrequentFlyer.Yes hanno un importante contributo nel polo positivo della prima dimensione.

Individui



Graf.C6



Graf.C7

Gli individui con profili simili sono vicini l'uno all'altro sulla mappa. Più l'individuo contribuisce in modo qualitativo al calcolo delle componenti principali, più il cos2 sarà maggiore (rosso).

COMMENTO

FAMD è stata **utilizzata unicamente a scopo esplorativo**: non viene effettuata nessuna feature reduction in quanto lo spazio di input non è grande (non ci sono problemi di efficienza), anche rispetto al numero di istanze (non ci sono problemi di efficacia). Inoltre, il risultato della FAMD mostra che una riduzione dello spazio di input non è "così tanto" migliorativa nel semplificare il problema. Se volessimo infatti mantenere una varianza cumulata "alta", dovremmo mantenere quasi tutte le covariate dello spazio originale.

5 MODELLI PREDITTIVI

0 – NOTE

VALUTAZIONE DELLE PERFORMANCE

Repeated k-Fold Cross-Validation (k=10, repeats=3)
La valutazione delle performance del modello è stata effettuata utilizzando Repeated k-Fold Cross-Validation in quanto il dataset è di piccole dimensioni e una divisione in train e set avrebbe potuto essere non rappresentativa.

Micro Average
L'aggregazione delle misure di performance è stata effettuata seguendo una Micro/Weighted-Average in quanto il problema non è sbilanciato ed entrambe le classi della variabile target sono importanti.

Fmeasure
Fmeasure viene calcolata senza cambiare il parametro Beta in quanto il problema non è sbilanciato e nessuna classe è più importante dell'altra.

Nota
La somma totale di righe e colonne della matrice complessiva è uguale a 3 volte il numero di istanze del dataset in quanto la 10-Fold Cross-Validation viene ripetuta 3 volte.

TUNING

Repeated k-Fold Cross-Validation (k=10, repeats=3)
Il tuning dell'iperparametro C per SVM è stato effettuato utilizzando Repeated k-Fold Cross-Validation e utilizzando come criterio di confronto l'accuracy ottenuta.

A – DECISION TREE

SCELTA DEL MODELLO

I motivi che hanno portato alla scelta del modello sono:

> Può essere usato per problemi di classificazione

> La variabile target è binaria e le features sono numeriche o discrete (per lo più discrete)

> Dall'analisi risulta che alcune features permettono di discriminare le classi della variabile target

> Il modello è computazionalmente semplice

> Le scelte effettuate dal modello, di interesse per il raggiungimento dell'obiettivo del problema, sono facilmente comprensibli all'uomo

Model1: Baseline Model – 0R

Viene creato un primo modello baseline 'ingenuo' in modo tale da poter confrontare le performance (in questo caso solo l'accuracy, per semplicità) con il modello trainato e poter quindi valutare se la complessità computazionale di quest'ultimo porta ad un effettivo vantaggio.

Il modello 0R viene creato assegnando a tutte le istanze 0, in quanto la maggior parte delle persone non acquista l'assicurazione di viaggio.

PERFORMANCE EVALUATION

		Predicted	
		0	1
True	0	1277	0
	1	710	0

accuracy = 0,64 *Not bad considering the "dummy" prediction*

Model2: Better Baseline Model

L'analisi esplorativa ha evidenziato alcune variabili in grado di discriminare bene le classi della variabile target. In particolar modo è stata rilevata una forte correlazione tra la variabile EverTravelledAbroad e la variabile target. Viene quindi creato un altro modello baseline che assegna a tutte le istanze con valore della variabile EverTravelledAbroad pari a 1 il valore 1, 0 altrimenti.

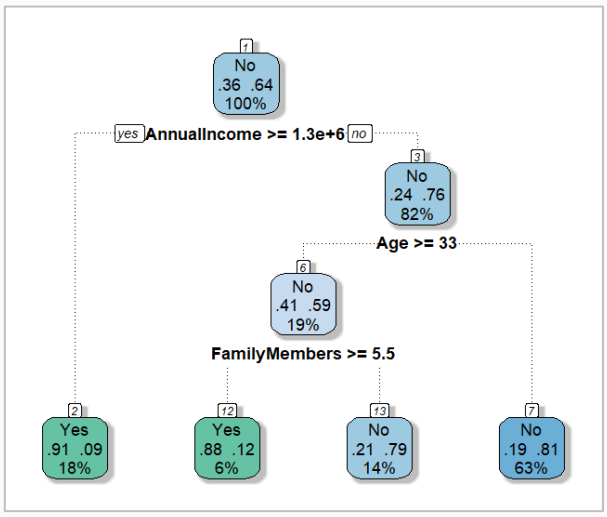
PERFORMANCE EVALUATION

		Predicted	
		0	1
True	0	1195	82
	1	412	298

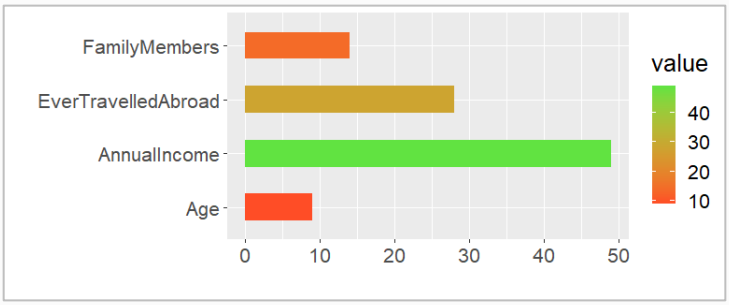
accuracy = 0,75 *Not bad considering the "dummy" prediction*

Model3a: CART [Gini Index]

Metodo	Classificazione
Attributi per indurre	Tutti (la colonna contenente gli id è già stata esclusa durante l'analisi)
Splitting Criteria	Gini Index



Graf.A1

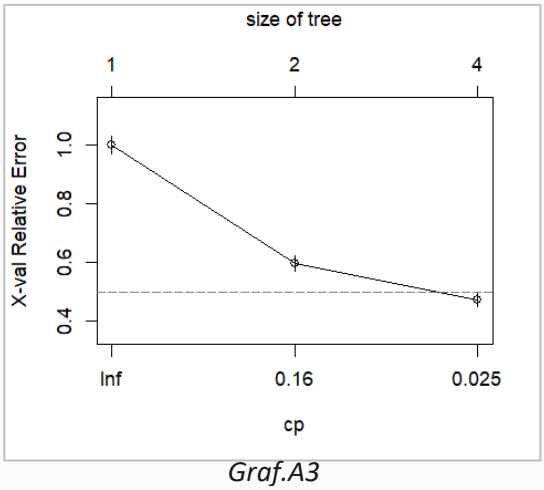


Graf.A2

L'albero ottenuto è formato da 3 split e 7 nodi totali. Le variabili più importanti per il modello sono in ordine decrescente per importanza: AnnualIncome, EverTravelledAbroad, FamilyMembers e Age. Quest'ultimo risultato conferma ciò che era stato predetto durante l'analisi esplorativa.

PRUNING DECISION TREE

Con l'obiettivo di evitare l'overfitting, seguendo il principio del Occam's razor, si valuta se è conveniente tagliare l'albero. La scelta viene effettuata studiando quanto l'aggiunta di un nodo permette di diminuire l'errore relativo di classificazione.



Graf.A3

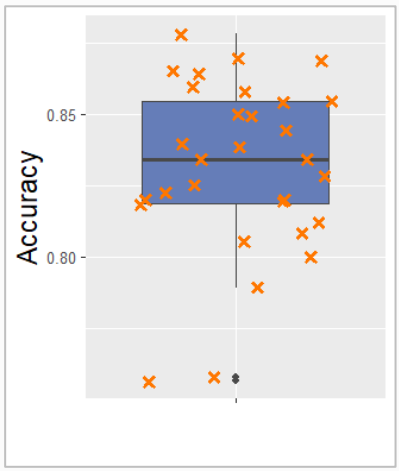
	CP	nsplit	rel error	xerror	xstd
1	0.404225	0	1.00000	1.00000	0.030086
2	0.061972	1	0.59577	0.59577	0.025700
3	0.010000	3	0.47183	0.47183	0.023506

L'errore minimo occorre quando l'albero è di dimensione 4 (dimensione attuale) e quindi non viene effettuato nessun taglio.

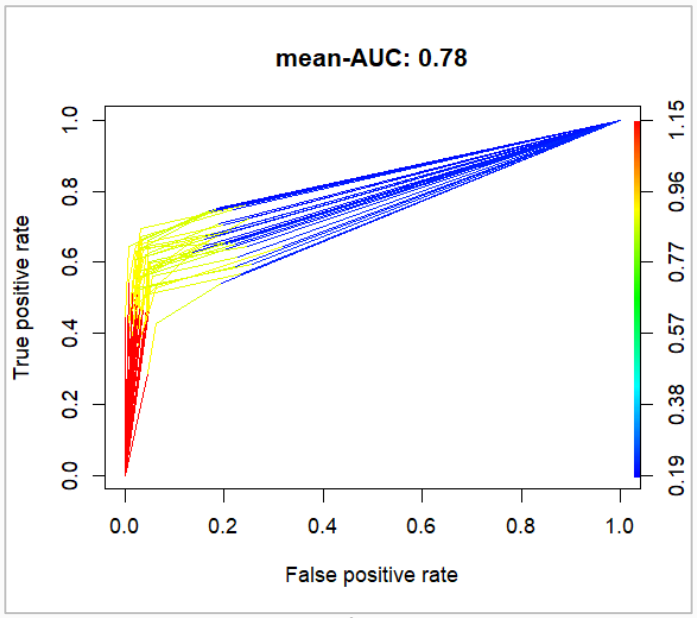
PERFORMANCE EVALUATION

		Predicted	
		0	1
True	0	1260	135
	1	870	3696

Performance		[95% CI]		Micro Average
accuracy	**0,83**	*0,82*	*0,84*	
precision [yes]	0,90	*0,89*	*0,92*	**0,84**
precision [no]	0,81	*0,78*	*0,82*	
recall [yes]	0,59	*0,57*	*0,62*	**0,83**
recall [no]	0,96	*0,96*	*0,97*	
fmeasure [yes]	0,71	*0,69*	*0,71*	**0,82**
fmeasure [no]	0,88	*0,87*	*0,89*	
AUC	**0,78**	*0,77*	*0,78*	



Graf.A4

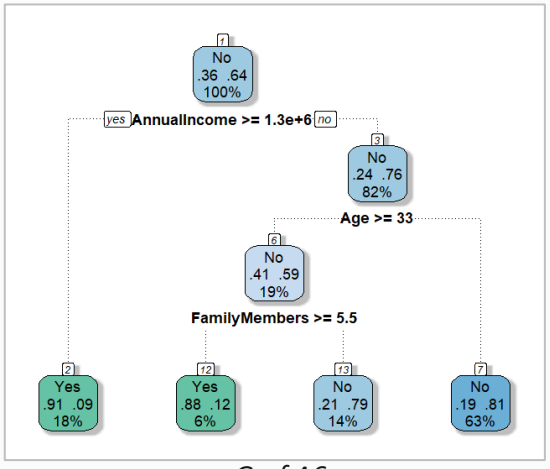


Graf.A5

Le performance del modello si possono ritenere complessivamente buone, sicuramente migliori rispetto a quelle dei modelli baseline.

Model3b: CART [Information Gain]

Metodo	Classificazione
Attributi per indurre	Tutti (la colonna contenente gli id è già stata esclusa durante l'analisi)
Splitting Criteria	Information Gain



Graf.A6

L'albero ottenuto è utilizzando Information Gain come criterio di split è completamente identico a quello appena analizzato: non ha senso quindi effettuare una valutazione anche di questo modello.

B – SVM

SCELTA DEL MODELLO

I motivi che hanno portato alla scelta del modello sono:

> Può essere usato per problemi di classificazione

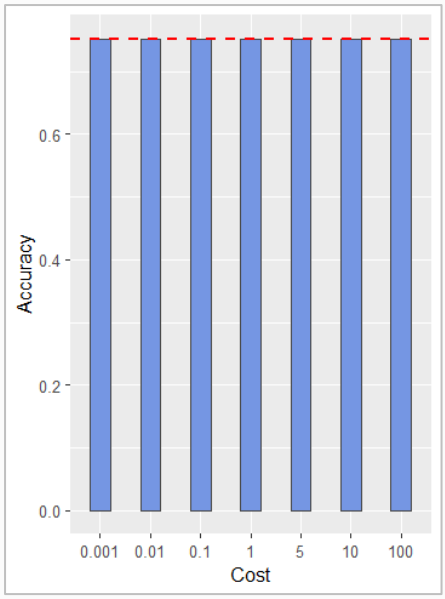
> Può essere usato anche per problemi non lineari (kernel trick)

Model4a: SVM [Linear]

Metodo	Classificazione
Attributi per indurre	Tutti (la colonna contenente gli id è già stata esclusa durante l'analisi)
Kernel	Linear

TUNING

Cost	Accuracy
0,001	0,7515518
0,01	0,7513840
0,1	0,7513840
1	0,7513840
5	0,7513840
10	0,7513840
100	0,7513840



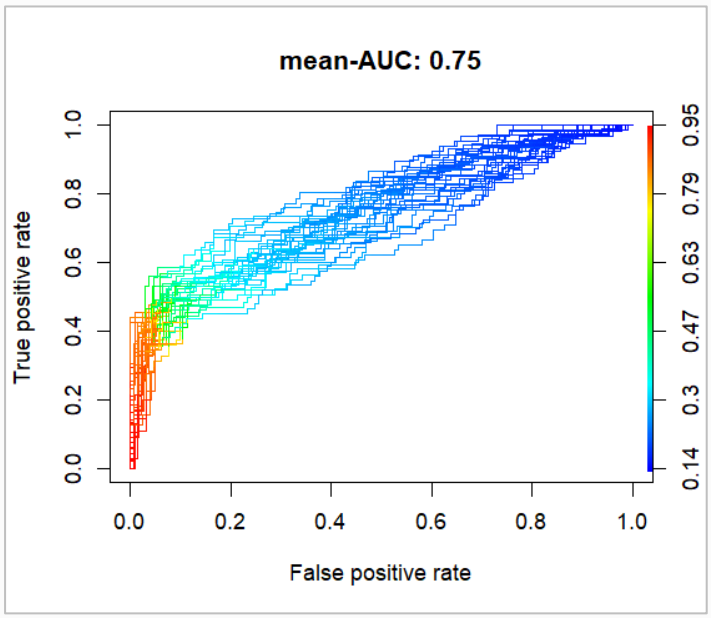
Graf.B1

L'accuracy migliore è ottenuta con iperparametro c=0,001.

PERFORMANCE EVALUATION

		Predicted	
		0	1
True	0	894	244
	1	1236	3587

Performance		[95% CI]			
accuracy	**0,75**	*0,74*	*0,76*		
				Micro Average	
precision [yes]	0,79	*0,76*	*0,81*	**0,76**	
precision [no]	0,74	*0,73*	*0,75*		
recall [yes]	0,42	*0,40*	*0,43*	**0,75**	
recall [no]	0,94	*0,93*	*0,94*		
fmeasure [yes]	0,55	*0,53*	*0,56*	**0,73**	
fmeasure [no]	0,83	*0,82*	*0,84*		
AUC	**0,75**	*0,73*	*0,76*		



BoxPlot Accuracy +

Graf.B2

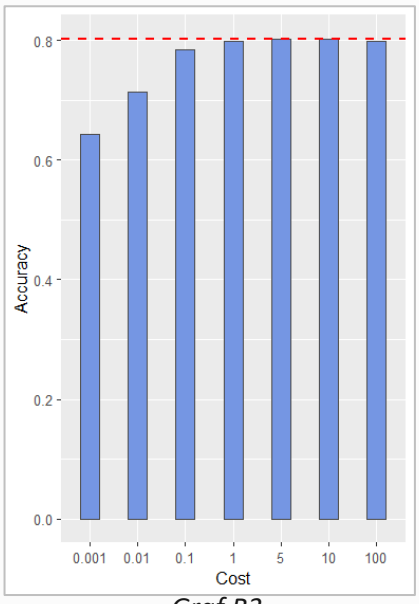
Model4b: SVM [Polynomial] (Tuning C)

Metodo	Classificazione
Attributi per indurre	Tutti (la colonna contenente gli id è già stata esclusa durante l'analisi)
Kernel	Polynomial

TUNING (C)

Gli altri iperparametri per il kernel polinomiale sono stati lasciati con il loro valore di default (*degree* = 3, *gamma* = 1/(data dimension), *coef0* = 0).

Cost	Accuracy
0,001	0,6426774
0,01	0,7144774
0,1	0,7845999
1	0,7985237
5	0,8020466
10	0,8010401
100	0,8015434



Graf.B3

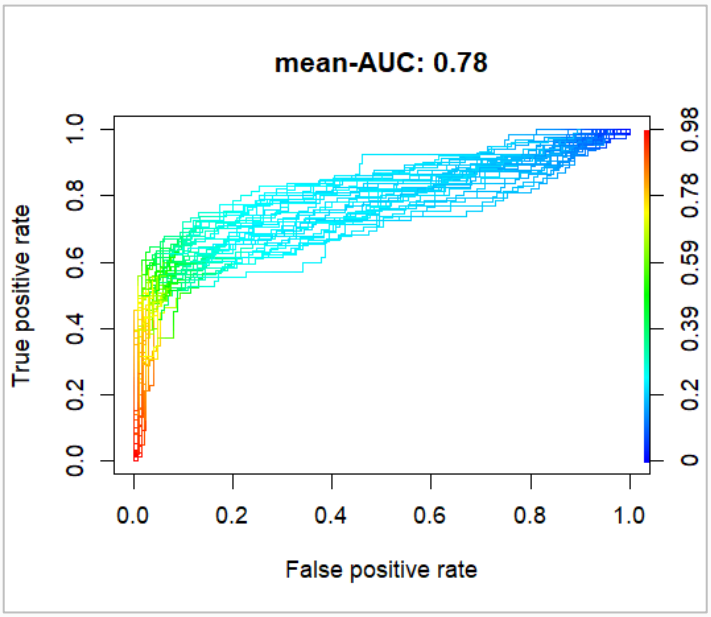
L'accuracy migliore è ottenuta con iperparametro c=5.

PERFORMANCE EVALUATION

		Predicted	
		0	1
True	0	1151	208
	1	979	3623

Performance		[95% CI]			
accuracy	**0,80**	*0,79*	*0,81*		
				Micro Average	
precision [yes]	0,85	*0,83*	*0,87*	**0,81**	
precision [no]	0,79	*0,77*	*0,80*		
recall [yes]	0,54	*0,52*	*0,56*	**0,80**	
recall [no]	0,86	*0,85*	*0,87*		
fmeasure [yes]	0,66	*0,64*	*0,68*	**0,79**	
fmeasure [no]	0,86	*0,85*	*0,87*		
AUC	**0,78**	*0,77*	*0,8*		

BoxPlot Accuracy +



Graf.B4

…