Spotify Award Predictor

Giudice Gianluca - 830694

Grassi Marco - 829664

Descrizione dataset

Record linkage da due sorgenti dati

Dataset kaggle - Features

- id
- name
- artists
- year
- duration.ms
- acousticness
- dancebility
- energy
- instrumentalness
- valence
- liveness
- loudness
- release.date
- speechiness
- tempo
- key
- mode
- explicit
- popularity

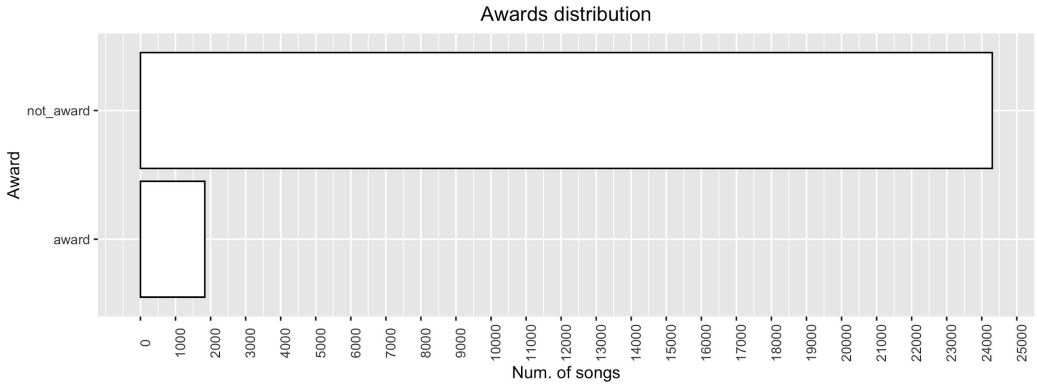
Scraping wikipedia - Label

- Certificazione vinta
  - Disco d'oro
  - Disco di platino
    - 1, 2, … N
- Nazionalità certificazione
  - Italia
  - Australia
  - USA
  - UK
  - Canada
  - Danimarca

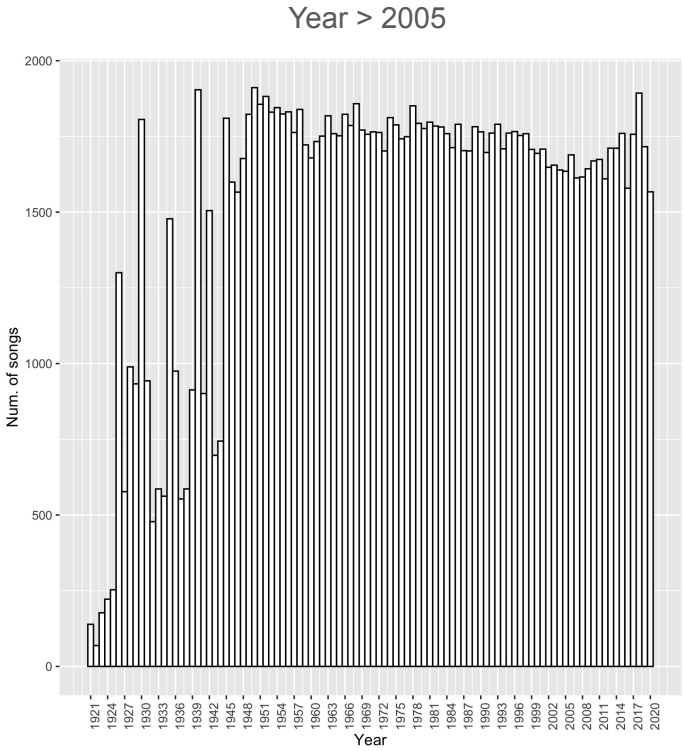
Chico	
Artista	Gué Pequeno
Featuring	Rose Villain e Luchè
Tipo album	Singolo
Pubblicazione	31 luglio 2020
Durata	3:33
Album di provenienza	*Mr. Fini*
Genere	Pop rap
Etichetta	Island
Produttore	Sixpm
Formati	Streaming
Certificazioni	
Dischi di platino	Italia (3)[1] (vendite: 210 000+)
Gué Pequeno - cronologia	
Singolo precedente	**Singolo successivo**
Saigon (2020)	*Bla Bla* (2020)
Luchè - cronologia	
Singolo precedente	**Singolo successivo**
Come me (2019)	*Maserati (Reloaded)* (2020)

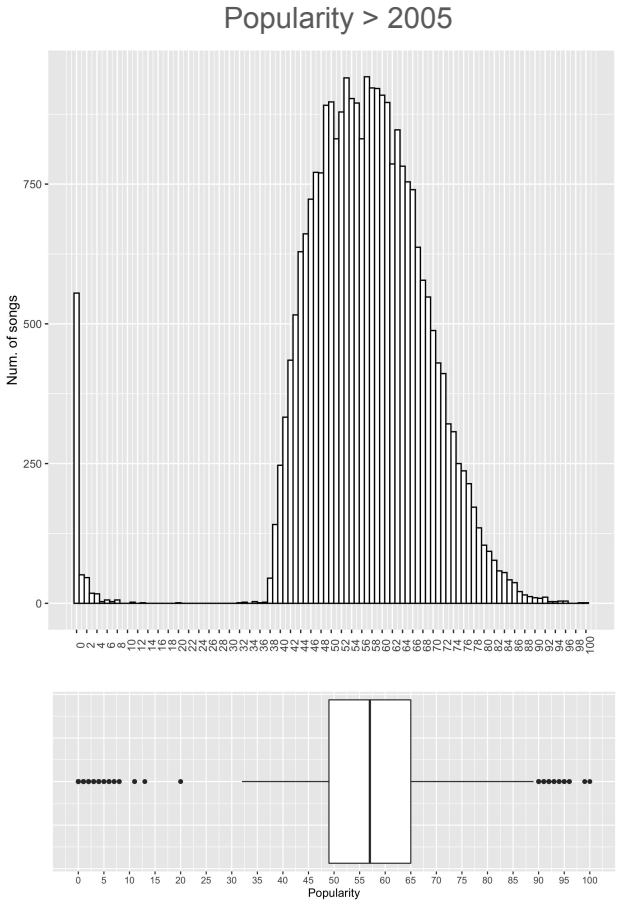
Dataset sbilanciato

Strategia adottata - Undersampling



Assunzioni dataset

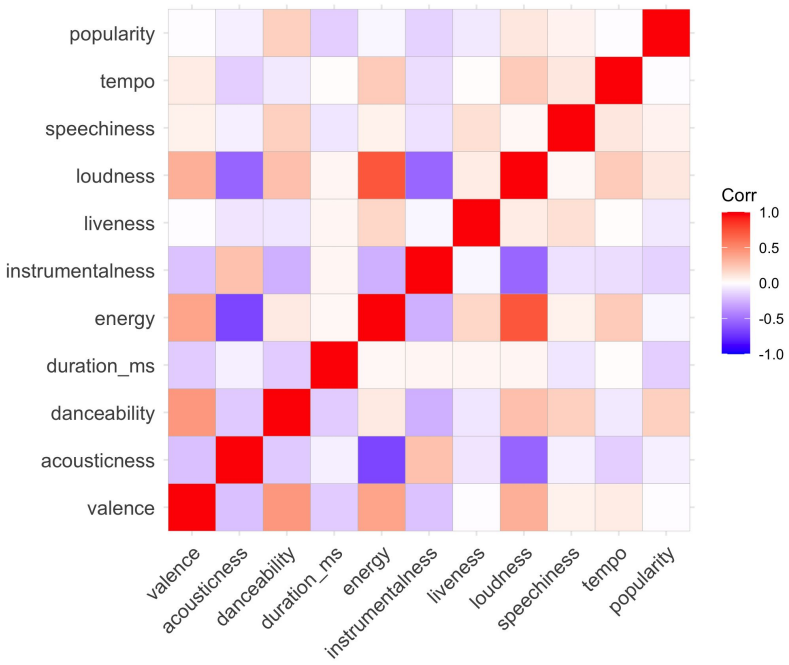


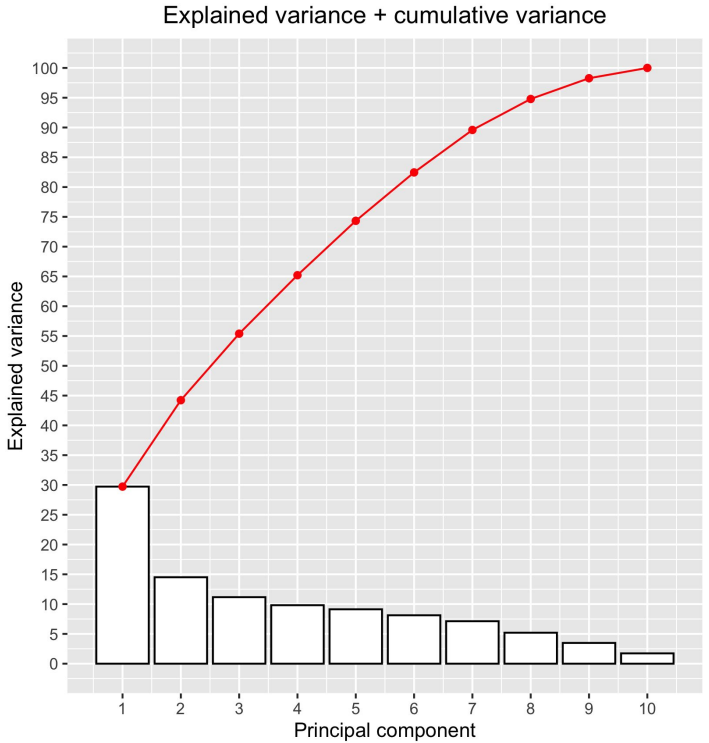



valence
acousticness
danceability
duration_ms
energy
instrumentalness
liveness
loudness
speechiness
tempo
popularity
award
not_award

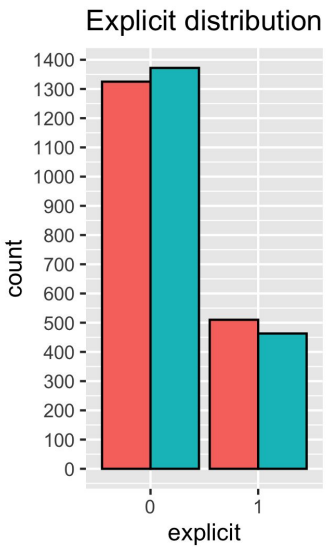
Principal components analysis

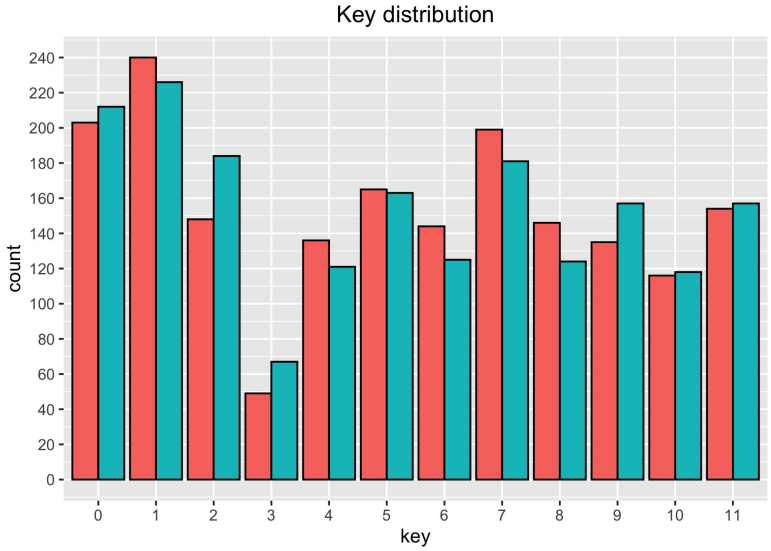
7 Componenti principali utilizzate

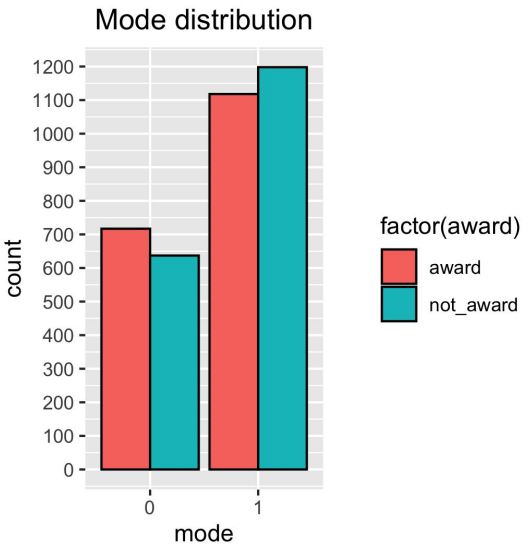




Variabili categoriche



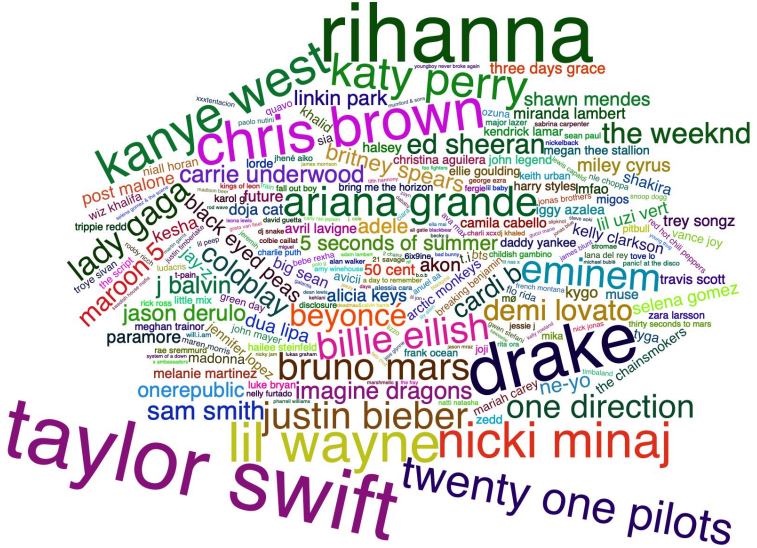




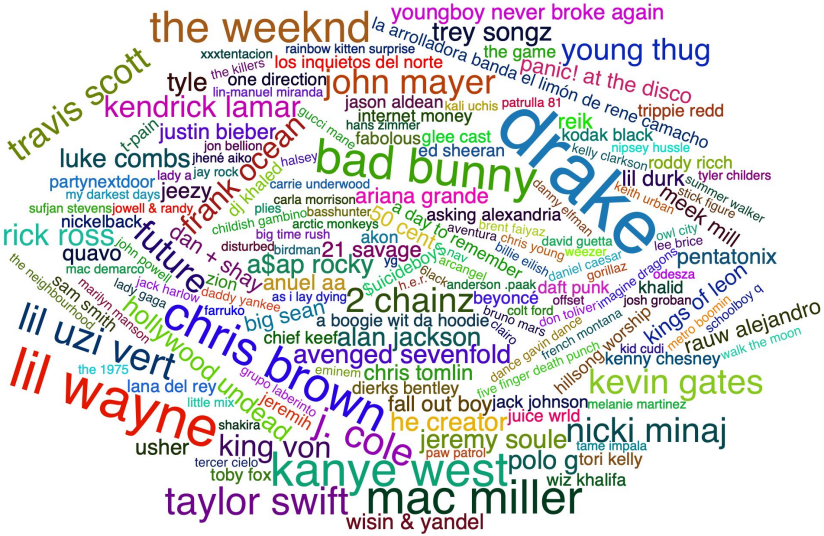
Rappresentazione one-hot-encoding artisti

- Ogni canzone ha associato un insieme di artisti
- Si vuole tenere conto di questa informazione
- Rappresentazione one-hot-encoding

Word cloud classe positiva



Word cloud classe negativa



Modelli

Iperparametri

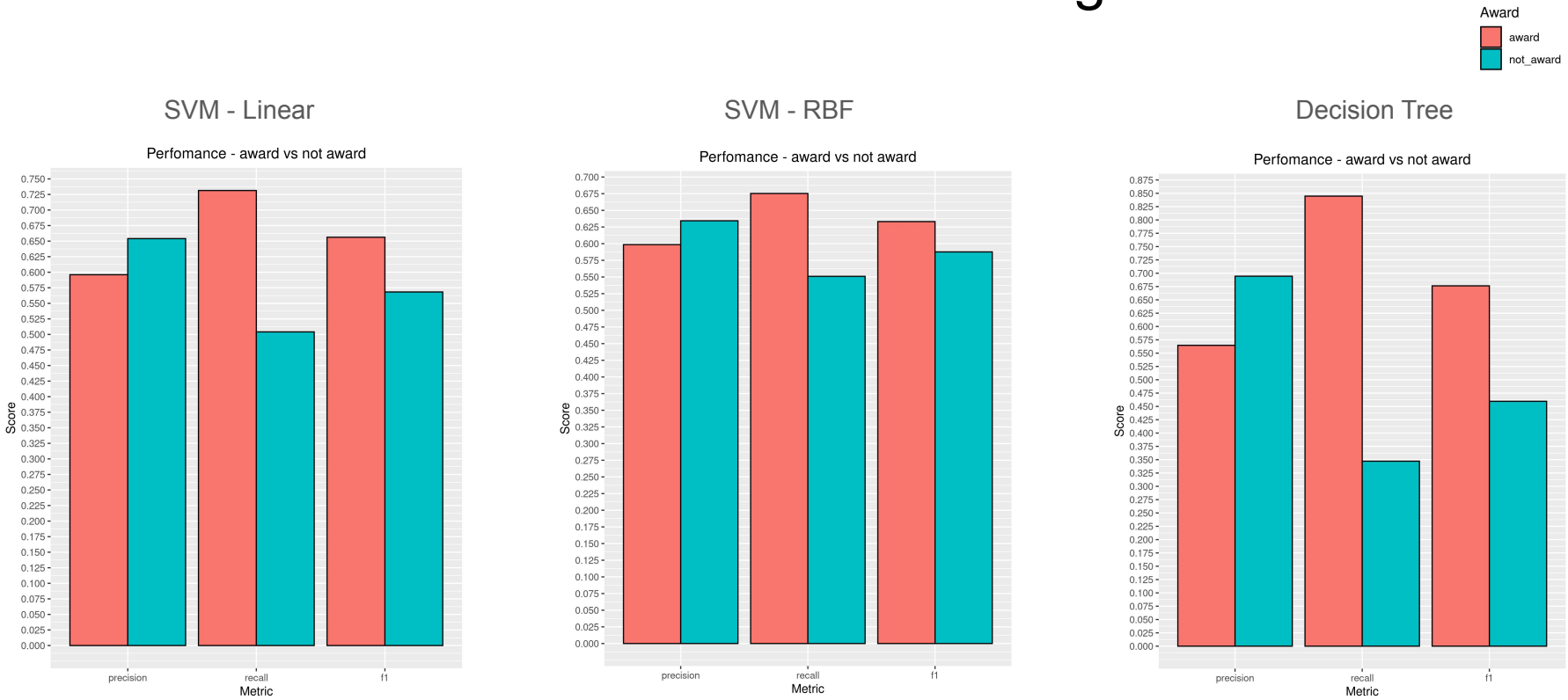
- Nested 10-folds cross validation
  - Per ogni fold tuning iperparametri per trovare quelli ottimali
- Grid search per model selection

SVM - RBF kernel		**SVM - Linear kernel**	**Decision tree**
C	$sigma$	C	cp
0.001	0.00001	0.000001	**0.01580381**
0.01	0.0001	0.00001	0.03487738
0.1	0.001	0.0001	0.17983651
1	0.01	0.001	
10	**0.1**	0.01	
		0.1	
		1	
		10	

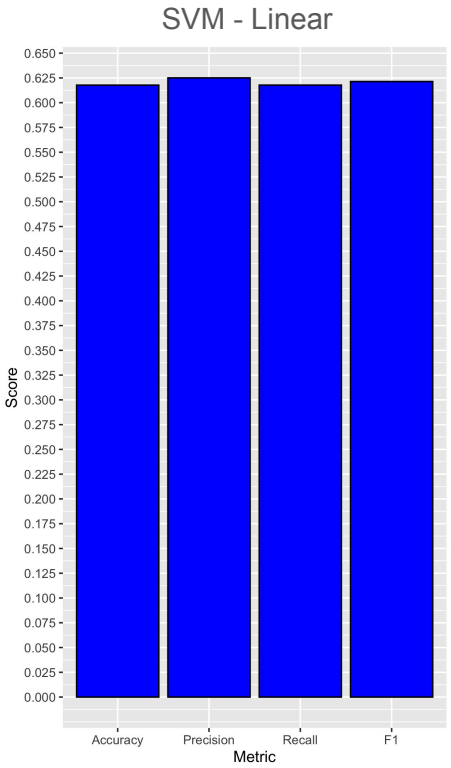
Table 3.1: Tabella iperparametri utilizzati per grid search.

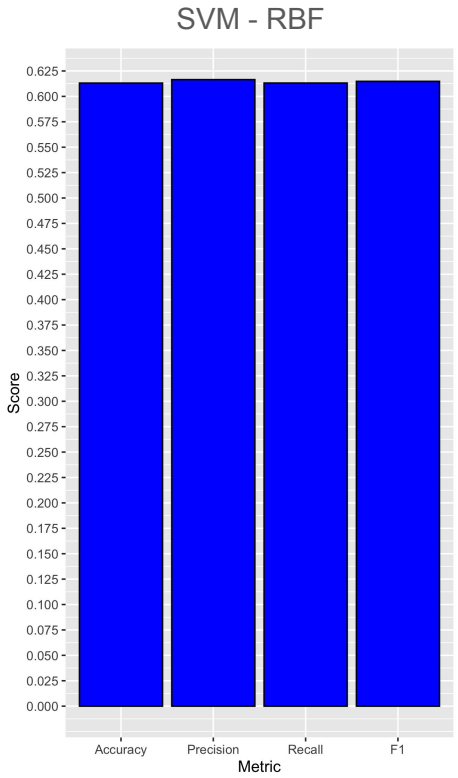
Performance

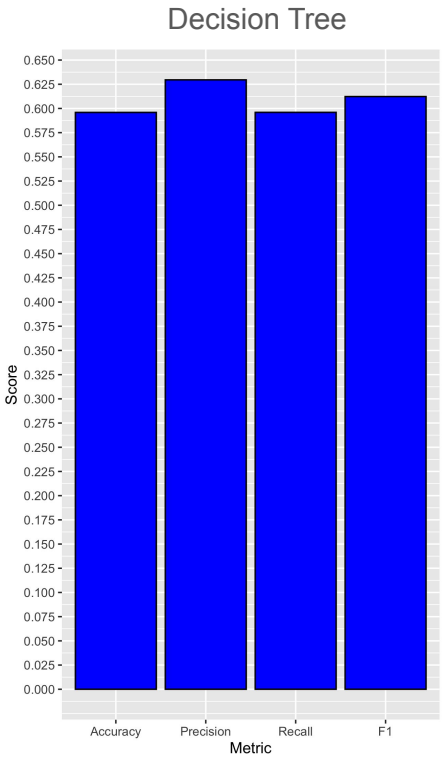
Performance classificatori - macro average



Performance classificatori - macro average

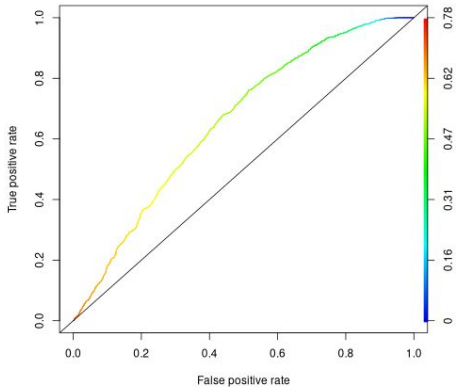


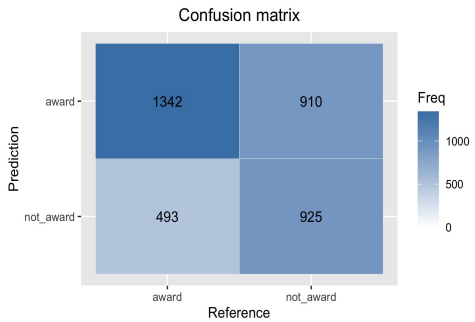




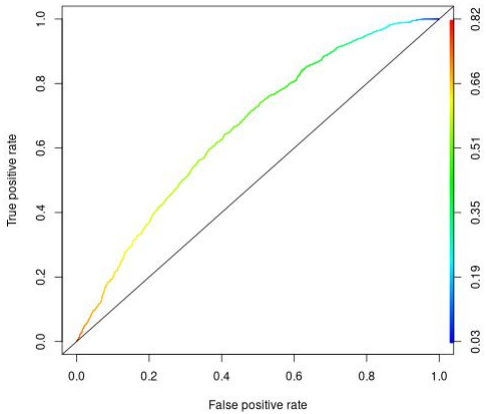
Curve ROC - Classe positiva

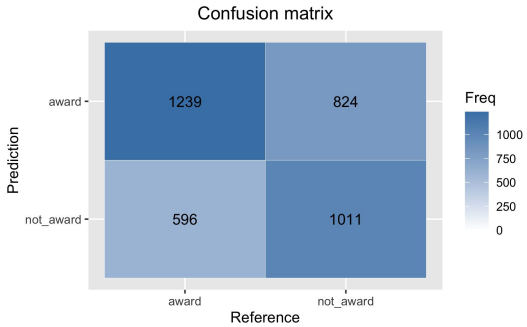
SVM - Linear



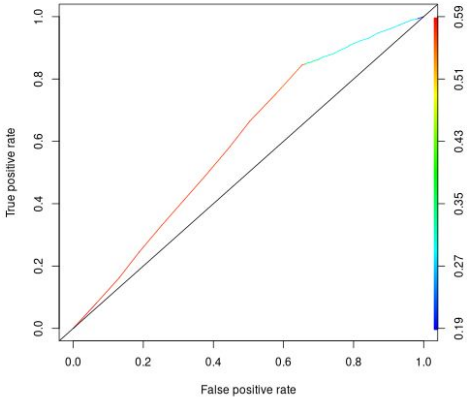


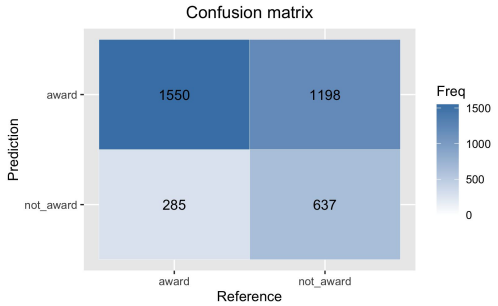
SVM - RBF





Decision Tree





Confronto secondo metrica ROC