Distributed Data Analysis and Mining

A.A. 2020/2021

Biviano Matteo
Currao Federica
Racioppa Arianna

realizzazione grafica a cura di Andrea Biviano

Dataset

Bank marketing campaigns

41.188 Records

21 Features:

- 7 relative ai clienti della banca;
- 4 relative all'ultimo contatto con il cliente per la campagna di marketing in corso;
- 4 contenenti informazioni sulla campagna corrente;
- 5 relative al contesto sociale ed economico della banca.

Variabile target binaria: Sottoscrizione del deposito a termine.

Goal

Obiettivo del progetto

- Clustering dei dati relativi a clienti che hanno sottoscritto un deposito a termine.
- Classificazione della variabile target per i clusters ottenuti ed interpretazione degli stessi

Data Understanding

Panoramica degli attributi:

- 10 variabili categoriche
- 10 variabili numeriche

Numeriche:

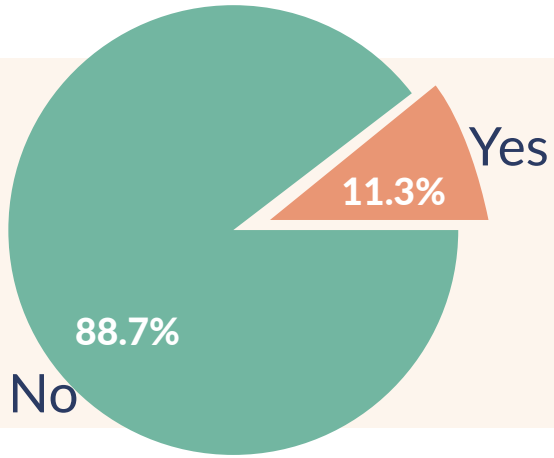
age, duration, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nremployed.

Categoriche:

job, marital, education, default, housing, loan, contact, month.

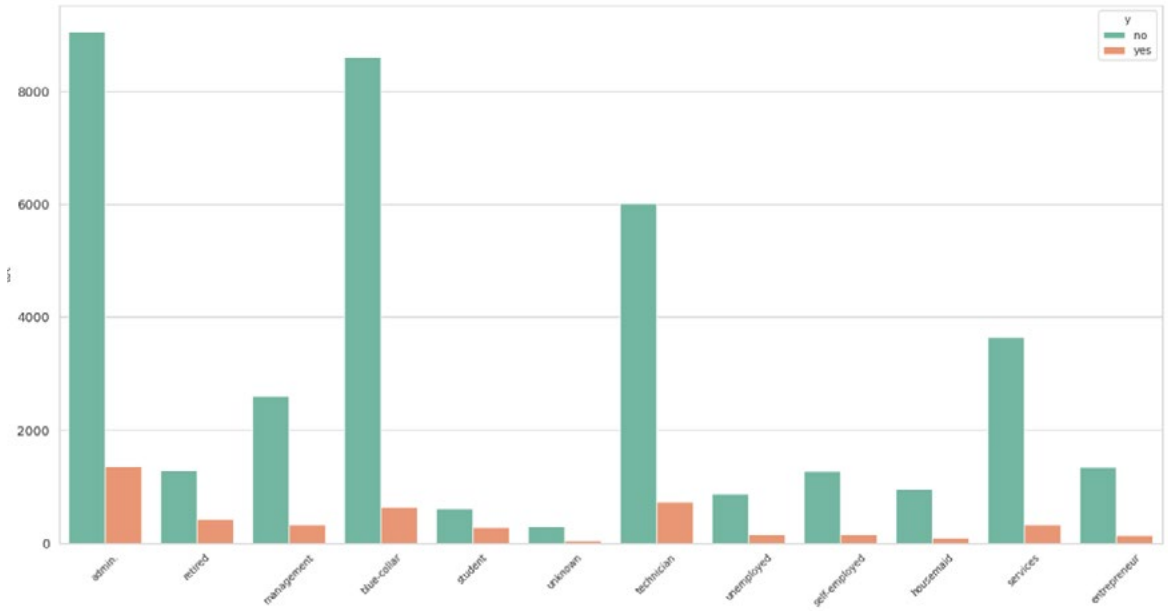
Distribuzione del target:

La variabile target «y» assume valore **«yes»** nel caso in cui il cliente chiamato decida di sottoscrivere un deposito a termine, altrimenti assume **«no»**.
Il dataset è estremamente sbilanciato.

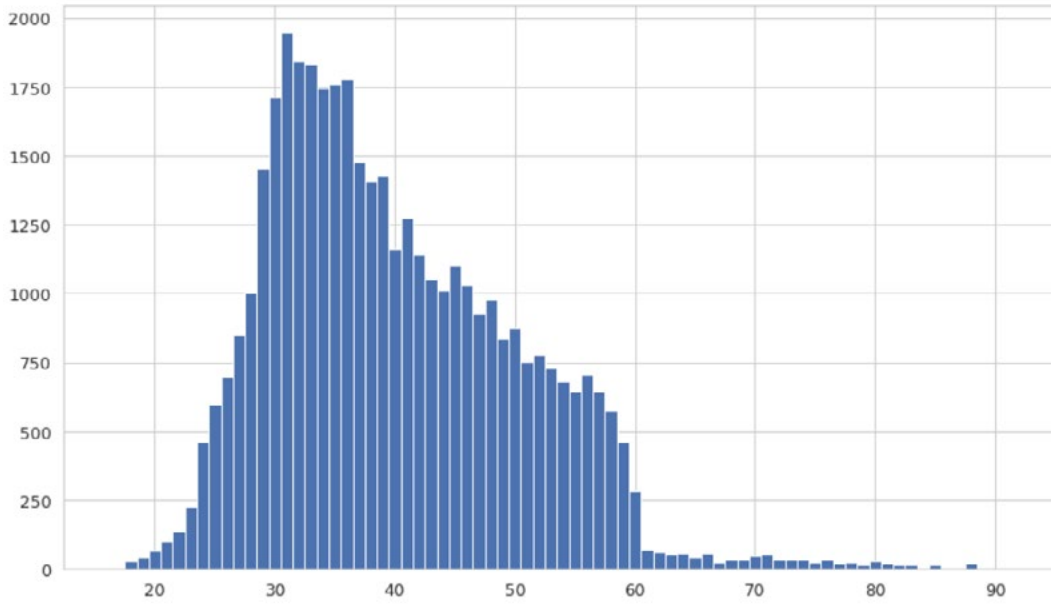


Distribuzione delle variabili

Rappresentazione della variabile target in funzione al lavoro del cliente contattato



Distribuzione dell'età dei clienti contattati

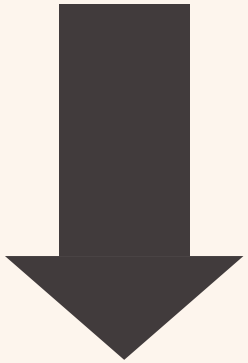


Data Cleaning e Preparation

- Missing Value
- Rimozione di dati duplicati
- Normalizzazione

Dataset iniziale:

41188 records

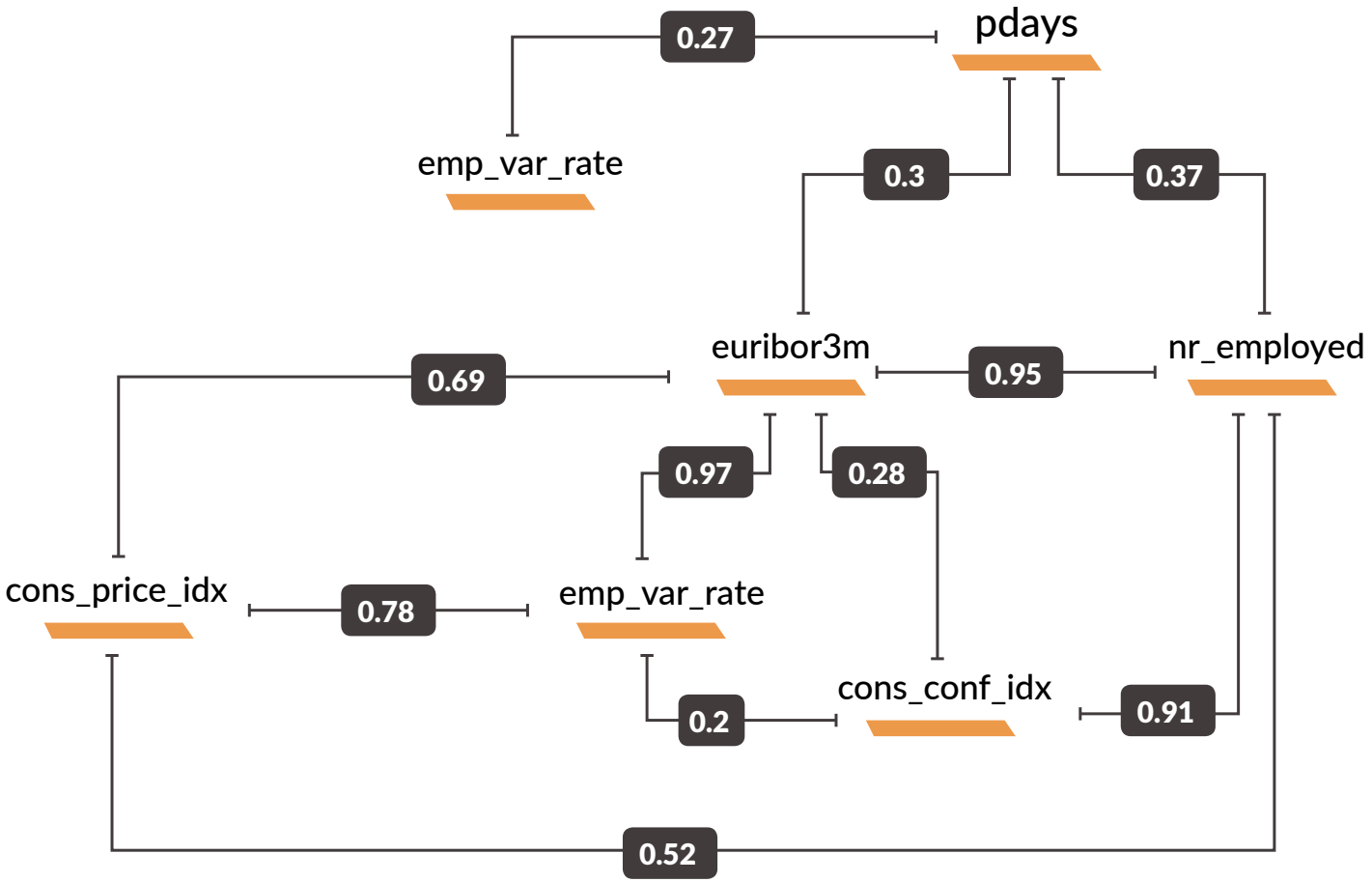


Dataset finale:

41172 records

Data preparation

Correlazione - Pearson



Clustering

METODOLOGIA

1) Features usate: “age”, “duration”, “campaign”, “previous“
2) Clustering effettuato solo sui dati con variabile target **«yes»**
3) Scelta del K attraverso l’Elbow method
4) Valutazione del clustering attraverso la silhouette

MODELLI

K-Means:

K: 6 Silhouette: 0.45

BisectingKMeans:

K: 7 Silhouette: 0.35

GaussianMixtureModel:

K: 6 Silhouette: 0.33

K-Means

KMEANS K=6						
	0	1	2	3	4	5
N° Records	788	1563	414	623	1006	245
Age	17-45	18-42	57-98	21-59	42-66	19-86
Duration	63-1624	37-1141	63-1962	207-4199	73-1640	64-1472
Campaign	1.74	1.82	1.79	3.34	2.036	1.73
Previous	1.25	0	0.83	0.049	0.18	3
Job	Admin (33.7%), technician (20.2%), student (13%)	Admin (33%), Technician (17%), blue-collar (13%)	Retired (71%)	Admin (26.6%), blue-collar (25.2%), technician (16.8%)	Admin (26.73%), blue-collar (17%), technician (13.8%)	Admin (34.2%), technician (15%)
Marital	Single (54%), married (39.7%)	Irrilevante	Married (75%)	Married (58%), single (30%)	Married (74.2%)	Married (59.6%), single (29%)
Education	Degree (40.2%), high school (26%)	Degree (43.1%), high school (24.6%)	Basic 4y (41.3%), degree (16.6%)	Degree (29.3%), high school (24.8%)	Degree (31%)	Degree (46.12%), high school (18%)

Cliente "tipo": admin o technician sposato con livello di istruzione elevato (laurea)

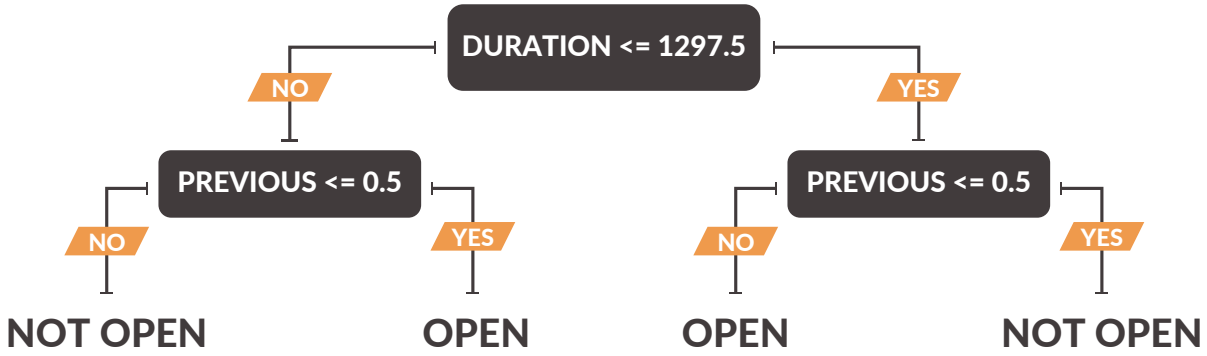
Classification

METODOLOGIA

1) Inserimento dei record con target negativo nei clusters ottenuti dall'analisi precedente
2) Features usate: "age", "duration", "campaign", "pdays", "previous", "emp_var_rate", "cons_price_idx", "cons_conf_idx", "euribor3m", "nr_employed"
3) Scelta degli iperparametri tramite CrossValidation (DecisionTree) e GridSearch (MLP, LSVC)

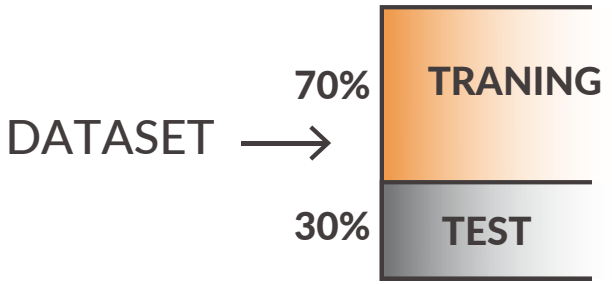
MODELLI

	Accuracy	F1-score	Precision	Recall
DECISION TREE	0.55	0.55	0.56	0.56
MULTILAYER PERCEPTRON	0.56	0.40	0.31	0.56
LINEARSVC	0.60	0.48	0.77	0.60



Il **miglior modello** rispetto all'F1-score è il **Decision Tree**.

Methodology Comparison



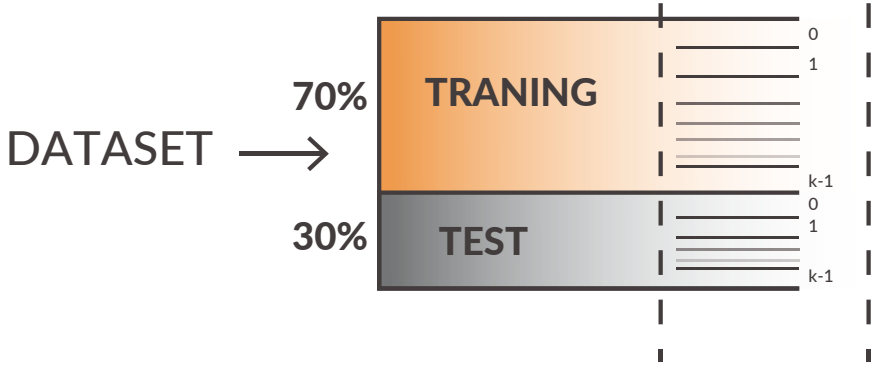
DECISION TREE + CROSS VALIDATION

[Metrics_DT]

- Accuracy_DT
- F1_DT
- Recally_DT
- Precision_DT

[1] KMEANS

[2] DECISION TREE + CROSSVALIDATION




```
predictions_labels = [ ]
for i in [0,k) ;
   op = CrossValidation(TRANING[i])
   model = DecisionTree(op,TRANING[i])
   model.predict(TEST[i])  ⟶ predictions_labels
```

[Metrics_predictions_labels]

- Accuracy_predictions_labels
- F1_predictions_labels
- Recally_predictions_labels
- Precision_predictions_labels

Risultati classificazione

DecizioneTree *vs* KMeans+DecisioneTree

	Accuracy	Precision	Recall	F1 Score
MetricsDT	0.855	False:0.894 True:0.211	False:0.954 True:0.088	0.855
MetricsPredictionsLabels	0.908	False:0.931 True:0.624	False:0.969 True:0.417	0.908

Il modello creato risulta classificare meglio dell'applicazione triviale del Decisione Tree