Università di Pisa

Relazione progetto di
Distributed Data Analysis and Mining

Bank Marketing

Biviano Matteo
Currao Federica
Racioppa Arianna

Anno Accademico 2020/2021

Indice

1 Introduzione **1**

2 Data Understanding & Preparation **1**

3 Clustering **3**

3.1 KMeans 3

3.2 BisectingKMeans - GaussianMixtureModel 4

3.3 Conclusione 4

4 Classification **5**

4.1 Decision Tree 5

4.2 Multilayer Perceptron Classifier 6

4.3 Linear Support Vector Classifier 6

4.4 Conclusione 7

5 Risultati Finali **7**

1 Introduzione

Il dataset utilizzato per lo studio di approcci di data mining in ambiente distribuito, chiamato *Bank Marketing Data Set*, è reperibile al link (`https://www.kaggle.com/volodymyrgavrysh/bank-marketing-campaigns-dataset`). I dati riguardano informazioni relative a campagne di marketing condotte tramite telefonate da un istituto bancario portoghese da Maggio 2008 a Dicembre 2010. Le campagne di marketing rappresentano una strategia tipica per migliorare il proprio business. In particolare, il marketing diretto (attraverso linea telefonica fissa o mobile) viene utilizzato quando ci si vuole rivolgere a determinati segmenti di clientela al fine di raggiungere un obiettivo specifico, in questo caso la sottoscrizione di un deposito a termine. Nel dataset in esame, nel 57% dei casi lo stesso cliente è stato contattato più volte, al fine di accertare che il deposito bancario a termine fosse stato sottoscritto. Il risultato (target) è quindi un contatto binario **yes** (deposito sottoscritto), **no** (deposito non sottoscritto). L'obiettivo dello studio è duplice: realizzare uno studio non supervisionato (attraverso clustering) su coloro che hanno sottoscritto il deposito, al fine di mappare le carattestiche comuni di tali clienti e, successivamente, prevedere se il cliente sottoscriverà un deposito bancario a termine (variabile **y**) attraverso la realizzazione di modelli di classificazione.

2 Data Understanding & Preparation

Il dataset è composto da 41188 records e da 21 features contenenti informazioni sul cliente contattato, sulla chiamata effettuata e sul contesto socio-economico della banca. Le features sono riassunte nelle tabelle in Figura [1 - 2].

Nome	Descrizione	Tipologia
Job	Tipo di lavoro	Categorico: {admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown}
Marital	Stato civile	Categorico: {divorced, married, single, unknown}
Education	Livello di istruzione	Categorico: {basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown}
Default	Indica se il cliente ha il credito in default	Categorico: {no, yes, unknown}
Housing	Indica se il cliente ha richiesto un prestito immobiliare	Categorico: {no, yes, unknown}
Loan	Indica se il cliente ha richiesto un prestito per motivi personali	Categorico: {no, yes, unknown}
Contact	Metodi di contatto	Categorico: {cellular, telephone}
Month	Ultimo mese (dell'anno) di contatto	Categorico: {jan, feb, ..., nov, dec}
DayOfWeek	Ultimo giorno (della settimana) di contatto	Categorico: {mon, tue, wed, thu, fri}
Poutcome	Risultato della precedente campagna di marketing	Categorico: {failure, nonexistent, success}

Figura 1: Semantica attributi categorici

Nome	Descrizione	Tipologia	Media	Dev.Std
Age	Età dell'utente contattato	Numerico	40	10
Duration	Durata dell'ultimo contatto (in secondi)	Numerico	258	259
Campaign	Numero di contatti eseguiti per questo cliente in questa campagna	Numerico	2.57	2.77
Pdays	Numero di giorni trascorsi dall'ultimo contatto del cliente	Numerico	963	187
Previous	Numero di contatti eseguiti per questo cliente prima di questa campagna	Numerico	0.18	0.49
Emp.var.rate	Tasso di variazione dell'occupazione	Numerico (Trimestrale)	0.08	1.57
Cons.price.idx	Indice dei prezzi al consumo	Numerico (Mensile)	93.58	0.58
Cons.conf.idx	Indice di fiducia dei consumatori	Numerico (Mensile)	- 40.5	4.63
Euribor3m	Tasso euribor 3 mesi	Numerico (Giornaliero)	3.62	1.73
Nr.employed	Numero di dipendenti della banca	Numerico (Trimestrale)	4167	72

Figura 2: Semantica attributi numerici

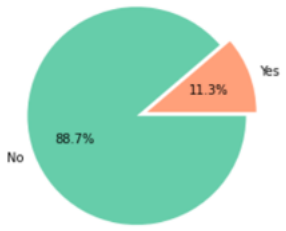


Figura 3: Distribuzione Target

Il dataset si presenta sbilanciato in quanto solo 4640 (11.3%) clienti hanno sottoscritto il deposito, come mostrato in Figura [3]. Dall'analisi della distribuzione delle variabili rappresentate in Figura [4] si può osservare che l'età segue una distribuzione simile ad una normale, mentre per la feature **Job** è possibile vedere che il valore più diffuso è "admin". L'attributo **Contact** non è stato incluso nelle analisi successive perchè risultato non discriminante, in quanto composto da valori con quasi uguale distribuzione. Infine, per l'attributo **Pdays** è stato sostituito il valore "999" (il quale indica che il cliente non è mai stato contattato precedentemente), di cui è quasi interamente composto, con "-1"; al fine di evitare alterazioni nella normalizzazione e nelle analisi successive.

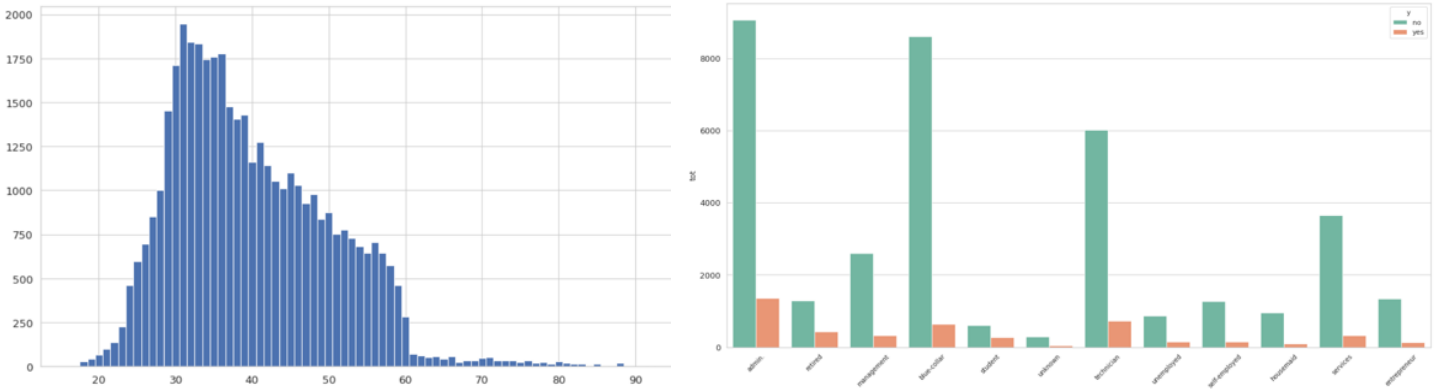


Figura 4: Dstribuzione Age - Job

Come è possibile osservare dalla tabella in Figura [1], molte delle features categoriche possiedono come valore "unknown". Questi valori sono stati identificati come *Missing Values*, ma è stato deciso di non eliminarli poichè una buona percentuale di essi possiede come valore target "yes", la cui rimozione avrebbe appesantito lo sbilanciamento del dataset. Inoltre, questi valori non sono stati sostituiti con altri delle rispettive classi perchè mantenere la categoria "unknown" sarebbe potuto risultare utile in fase di clustering. Per quanto riguarda gli **outliers**, è stato osservato che questi erano relativi principalmente all'attributo **Age**, riguardanti i soggetti al di sopra dei 70 anni (come visibile in Figura [4]). Gli outliers individuati non sono stati eliminati per ragioni analoghe a quanto fatto per il trattamento dei *Missing Values*. Sono stati, inoltre, individuati e rimossi 12 record duplicati. Sul dataset risultante è stata calcolata la correlazione tra tutte le features, rappresentata dalla heatmap in Figura [5]. Gli attributi che presentano maggior correlazione sono quelli di carattere socio-economico di natura trimestrale.

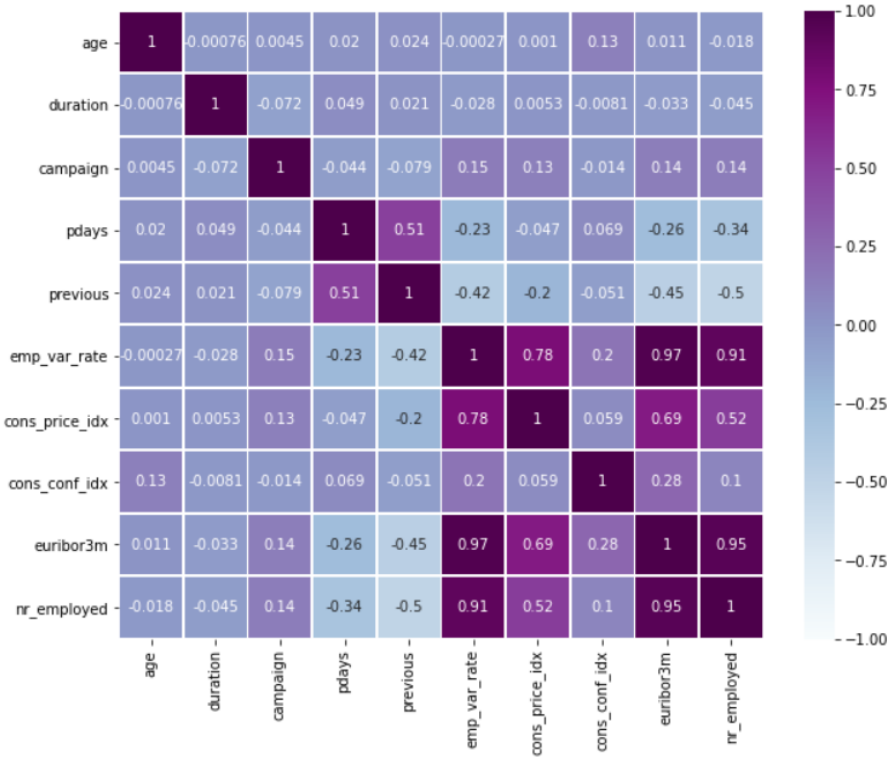


Figura 5: Correlazione

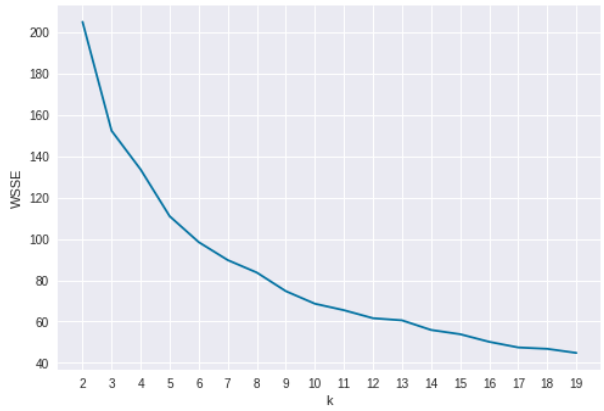
3 Clustering

In questa sezione vengono analizzati i dati attraverso diversi modelli di clustering, al fine di individuare le caratteristiche comuni dei clienti che hanno stipulato un deposito a termine. Per effettuare il clustering, dal dataset risultate dalla fase precedente (Sezione [2]) sono stati estratti i record con valore target $\mathbf{y = yes}$, usando come features numeriche discriminatorie le seguenti: **Age**, **Duration**, **Campaign**, **Previous**. È stato deciso di escludere da questa analisi gli attributi socio-economici a causa della loro natura trimestrale in quanto avrebbero "guidato" il raggruppamento dei dati per trimestre, andando quindi ad alterare i risultati. Questi dati sono stati poi normalizzati attraverso **MinMax**. I modelli utilizzati sono: **KMeans**, **BisectingKMeans** e **GaussianMixtureModel**.

3.1 KMeans

Al fine di identificare il miglior parametro **k** per il *K-Means*, è stato usato il metodo *Elbow* calcolando il WSSSE per $k \in [2, 20]$. Il grafico in Figura [6] (a sinistra) mostra le variazioni del WSSSE rispetto al **k** considerato, mentre

a destra viene riportato il confronto di WSSSE e Silhouette nell'intervallo di valori all'interno del quale sarebbe identificabile il valore ottimo. Il valore considerato ottimale è risultato essere **k=6**, per il quale sono stati ottenuti i clusters discussi in Sezione [3.3]



K	WSSSE	SILHOUETTE
6	98.42	0.448
7	89.73	0.434
8	83.71	0.438
9	74.69	0.394

Figura 6: KMeans: WSSE - Silhouette

3.2 BisectingKMeans - GaussianMixtureModel

Anche per il *BisectingKMeans* l'individuazione del k ottimale è stata effettuata calcolando il WSSSE al variare del numero di clusters, i cui risultati sono visibili in Figura [7]. Il valore ottimale individuato è stato per $\mathbf{k = 7}$, con il quale è stato ottenuto un WSSSE = 7274.9 ed una Silhouette = 0.35. Il *Gaussian Mixture Model*, invece, calcola una probablità di appartenenza dei records al cluster e non possiede la funzione di costo, quindi non potendo considerare il WSSSE per individuare il numero di clusters ottimale, è stata utilizzata solamente la Silhouette. Il valore migliore è stato ottenuto per $\mathbf{k = 6}$, con Silhouette = 0.331.

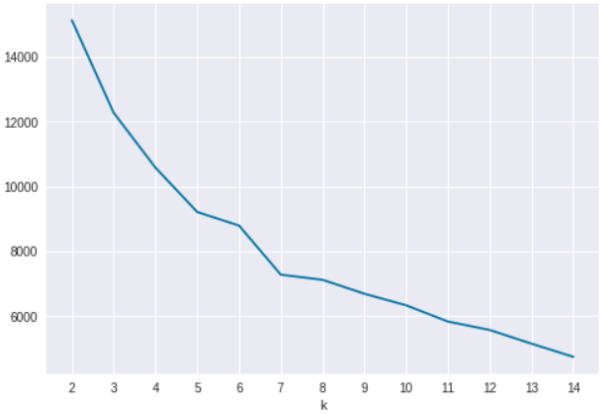


Figura 7: BisectingKMeans: WSSE

3.3 Conclusione

	Silhouette
KMeans	0.448
BisectingKMeans	0.35
Gaussian Model	0.331

Figura 8: Confronti Silhouette

Il KMeans è stato scelto come miglior modello di clustering sulla base del valore di Silhouette e della significatività dei clusters ottenuti. In Figura [8] sono riassunti i valori di Silhouette per i modelli considerati, mentre in Figura [9] è stata riportata la composizione dei clusters ottenuti tramite il KMeans. Come è possibile osservare, i cluster 0, 1 e 3 sono simili tra loro in termini di età (fascia medio bassa), lavoro (*"admin", "blue-collar"*) e livello d'istruzione; si differenziano invece per range di durata delle chiamate, simile nei clusters 0 ed 1, ma molto più alto nel cluster 3 (chiamate fino a circa 20 minuti). Per il cluster 3 risulta, inoltre, più alto il numero di contatti avuti con il cliente per la campagna corrente. I cluster 4 e 5 sono simili tra loro, ma si differenziano per il range d'età, poichè nel cluster 5 è molto più ampio e per *"prevoius"*, molto più basso nel cluster 4, infatti in questo cluster i clienti non sono stati precedentemente contattati. Il cluster 2 raggruppa i clienti con fascia d'età medio-alta, quindi per la maggior parte pensionati, con un livello d'istruzione basso, i quali sono stati contattati più volte sia in passato che nella campagna attuale.

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%)
Martial	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%)

Figura 9: Composizione dei Clusters ottenuti con il KMeans

Dall'analisi sui clusters ottenuti tramite *KMeans* è stato possibile individuare il **cliente "tipo"** che stipula un deposito a termine, ovvero un admin o technician sposato e appartenente ad una fascia d'età medio-bassa e con un livello d'istruzione alto (laurea); inoltre la durata delle chiamate per chi stipula il contratto è in media sempre superiore ai 5 minuti ed il fatto che siano stati contattati più volte nella campagna precedente non influisce sulla risposta del cliente, al contrario, il fatto che siano stati contattati più volte nella campagna attuale risulta significativo.

4 Classification

Una volta terminata la fase esplorativa, ai clusters ottenuti nella sezione precedente sono stati aggiunti i record con target **y = no** sulla base della somiglianza tra record e cluster. Ogni cluster risultante è stato trattato come dataset, diviso successivamente in training set e test set al fine di analizzare i modelli di classificazione. Attraverso la *Feature Importance*, sono stati selezionati per questa fase, i seguenti attributi: "age","duration", "campaign", "pdays", "previous", "emp_var_rate", "cons_price_idx", "cons_conf_idx", "euribor3m", "nr_employed". I modelli che sono stati analizzati sono: Decision Tree, Multilayer Perceptron (MLP) e Linear Support Vector Classification (LSVC).

4.1 Decision Tree

Per scegliere la configurazione di parametri con cui eseguire il Decision Tree su ogni cluster, è stata effettuata una Grid Search Cross Validation con k_fold = 3, attraverso la quale sono stati testati i seguenti gruppi di iperparametri: **impurity = ["entropy", "gini"]**; **maxDepth = [2, 5, 10, 15]** e **maxBins = [2, 5, 10, 15]**. In Figura [10] vengono mostrati per ogni clusters le configurazioni ottime utilizzate e i risultati ottenuti dai classificatori.

	Parameters	Accuracy	Precision	Recall	F1 Score
0	impurity: *entropy* , maxDepth: *5* , maxBins: *15*	0.834	0.765	0.785	0.834
1	impurity: *entropy* , maxDepth: *5* , maxBins: *15*	0.923	0.735	0.682	0.923
2	impurity: *entropy* , maxDepth: *15* , maxBins: *15*	0.656	0.655	0.645	0.656
3	impurity: *entropy* , maxDepth: *2* , maxBins: *2*	0.553	0.552	0.552	0.553
4	impurity: *entropy* , maxDepth: *5* , maxBins: *5*	0.921	0.735	0.592	0.921
5	impurity: *gini* , maxDepth: *5* , maxBins: *5*	0.719	0.745	0.695	0.719

Figura 10: Decision Tree Classifier

4.2 Multilayer Perceptron Classifier

In questo caso, la Grid Search è stata eseguita con i seguenti gruppi di iperparametri: **maxIters = [100, 150, 200, 250]**, **blockSize = [32, 64, 128]**, **layers = [[10,6, 4, 2], [10, 8, 6, 4, 2], [10, 8, 2]]**. Per tutti i clusters è stata ottenuta come configurazione migliore: {'maxIter': 250, 'layers': [10, 8, 2], 'blockSize': 128}. I risultati ottenuti dal classificatore sono riportati in Figura [11].

	Accuracy	Weighted Precision	Weighted Recall	F1 Score
0	0.764	0.583	0.764	0.661
1	0.919	0.844	0.919	0.880
2	0.596	0.355	0.596	0.445
3	0.558	0.312	0.558	0.40
4	0.916	0.840	0.916	0.877
5	0.538	0.289	0.538	0.376

Figura 11: Multilayer Perceptron Classifier

4.3 Linear Support Vector Classifier

Per quest'ultimo classificatore sono stati testati i seguenti iperparametri: **numIterations = [100, 150, 200, 250]**, **regParm = [0.01, 0.02, 0.04, 0.06, 0.08]**, **tol = [1e-03, 1e-04, 1e-05, 1e-06, 1e-07]**. La configurazione migliore ottenuta dalla Grid Search per tutti i clusters è stata {'maxIter': 250, 'regParam': 0.08, 'tol': 1e-07}. In Figura [12] vengono mostrati i risultati ottenuti sull'esecuzione del classificatore.

	Accuracy	Weighted Precision	Weighted Recall	F1 Score
0	0.825	0.814	0.825	0.806
1	0.919	0.844	0.919	0.880
2	0.727	0.725	0.727	0.718
3	0.598	0.766	0.598	0.483
4	0.916	0.857	0.916	0.877
5	0.615	0.693	0.615	0.549

Figura 12: Linear Support Vector Classifier

4.4 Conclusione

Il modello migliore, in termini di F1 Score, è risultato essere il Decision Tree, del quale in Figura [13] viene riportato un esempio di classificazione, appartenente al cluster 4.

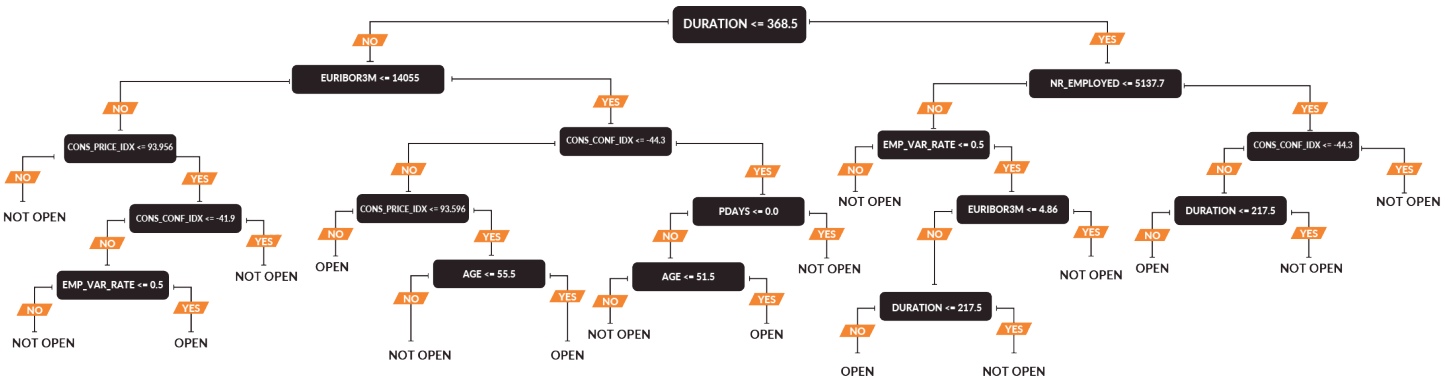


Figura 13: Decision Tree cluster 4

5 Risultati Finali

Per valutare l'efficienza del modello di classificazione creato sulla base del clustering, è stato effettuato un confronto con l'applicazione triviale del **DecisionTree** all'intero dataset originale. La struttura del confronto viene effettuata per fasi, visibili nelle Figure [14 - 15]. Il dataset è stato, in prima istanza, diviso in *Training* (70%) e *Test* (30%). Successivamente è stata effettuata, come visibile in Figura [14], la *3-Fold Cross Validation* sul training, in modo da trovare la configurazione di iperparametri ottima per valutare il **DecisionTree** sul test ottenendo i valori delle metriche [**Metrics_DT**].

Nella seconda fase sono stati eseguiti i seguenti passaggi:

1. Dai dati del training sono stati estratti i record con target positivo e su di essi è stato eseguito il clustering (*KMeans*);
2. I record con target negativo sono stati aggiunti ai clusters ottenuti al passo precedente;
3. Il test è stato anch'esso diviso in k clusters sulla base dei risultati ottenuti per il training, in modo tale da avere (come mostrato in Figura [15]) sia Training che Test, iniziali, divisi in k clusters ed ottenere k coppie della forma: <training[i], test[i]> per i = 0, ..., k-1;

4. Per ogni **training[i]** è stata effettuata la *Cross Validation* ottenendo così i parametri utilizzati per eseguire il corrispettivo DecisionTree successivamente eseguito sul **test[i]**;

5. Dall'esecuzione dei k DecisionTree è stata ottenuta una serie di coppie <label_originale, label_predetta>, le quali sono state inserite nella lista "predictions_labels", utilizzata per calcolare le metriche complessive del modello [**Metrics_predictions_labels**].

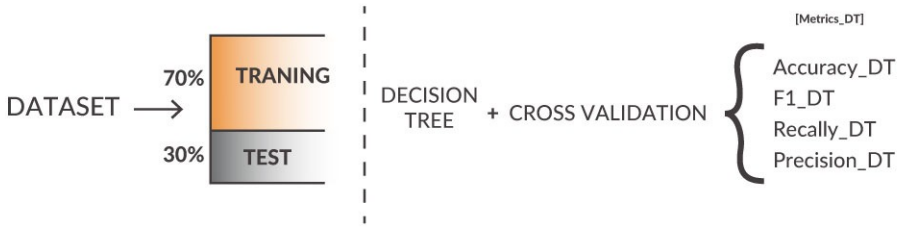


Figura 14: DecisionTree

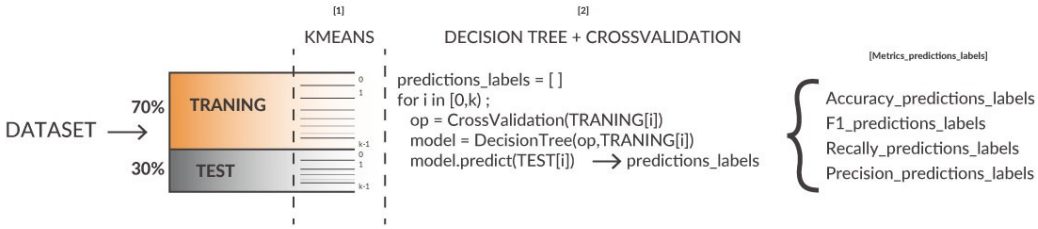


Figura 15: Clustering DecisionTree

Confrontando i risultati dei due modelli riportati in Figura [16] è possibile osservare che in termini di performance il modello *"KMeans + DecisionTree"* risulta essere migliore rispetto all'uso del classificatore applicato in maniera standard.

	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

Figura 16: Confronto finale