Progetto d'Esame per Mining II

Predict Students' Dropout and Academic Success

Link al dataset:
https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success_

Link al notebook:
https://github.com/PaoloCarlevero/mining_II_project_work

Alessandra Barillaro
Paolo Carlevero
Cristiano Degrandis

Data: 09/07/2024

Data Exploration e Data preprocessing

#	Column	Non-Null Count	Dtype
0	Marital Status	4424 non-null	int64
1	Application mode	4424 non-null	int64
2	Application order	4424 non-null	int64
3	Course	4424 non-null	int64
4	Daytime/evening attendance	4424 non-null	int64
5	Previous qualification	4424 non-null	int64
6	Previous qualification (grade)	4424 non-null	float64
7	Nacionality	4424 non-null	int64
8	Mother's qualification	4424 non-null	int64
9	Father's qualification	4424 non-null	int64
10	Mother's occupation	4424 non-null	int64
11	Father's occupation	4424 non-null	int64
12	Admission grade	4424 non-null	float64
13	Displaced	4424 non-null	int64
14	Educational special needs	4424 non-null	int64
15	Debtor	4424 non-null	int64
16	Tuition fees up to date	4424 non-null	int64
17	Gender	4424 non-null	int64
18	Scholarship holder	4424 non-null	int64
19	Age at enrollment	4424 non-null	int64
20	International	4424 non-null	int64
21	Curricular units 1st sem (credited)	4424 non-null	int64
22	Curricular units 1st sem (enrolled)	4424 non-null	int64
23	Curricular units 1st sem (evaluations)	4424 non-null	int64
24	Curricular units 1st sem (approved)	4424 non-null	int64
25	Curricular units 1st sem (grade)	4424 non-null	float64
26	Curricular units 1st sem (without evaluations)	4424 non-null	int64
27	Curricular units 2nd sem (credited)	4424 non-null	int64
28	Curricular units 2nd sem (enrolled)	4424 non-null	int64
29	Curricular units 2nd sem (evaluations)	4424 non-null	int64
30	Curricular units 2nd sem (approved)	4424 non-null	int64
31	Curricular units 2nd sem (grade)	4424 non-null	float64
32	Curricular units 2nd sem (without evaluations)	4424 non-null	int64
33	Unemployment rate	4424 non-null	float64
34	Inflation rate	4424 non-null	float64
35	GDP	4424 non-null	float64

- Data ingestion attraverso API
- X_df= 36 variabili con 4424 record
- y_df= Graduate/Enrolled/Droput

Target	
Graduate	0.499322
Dropout	0.321203
Enrolled	0.179476

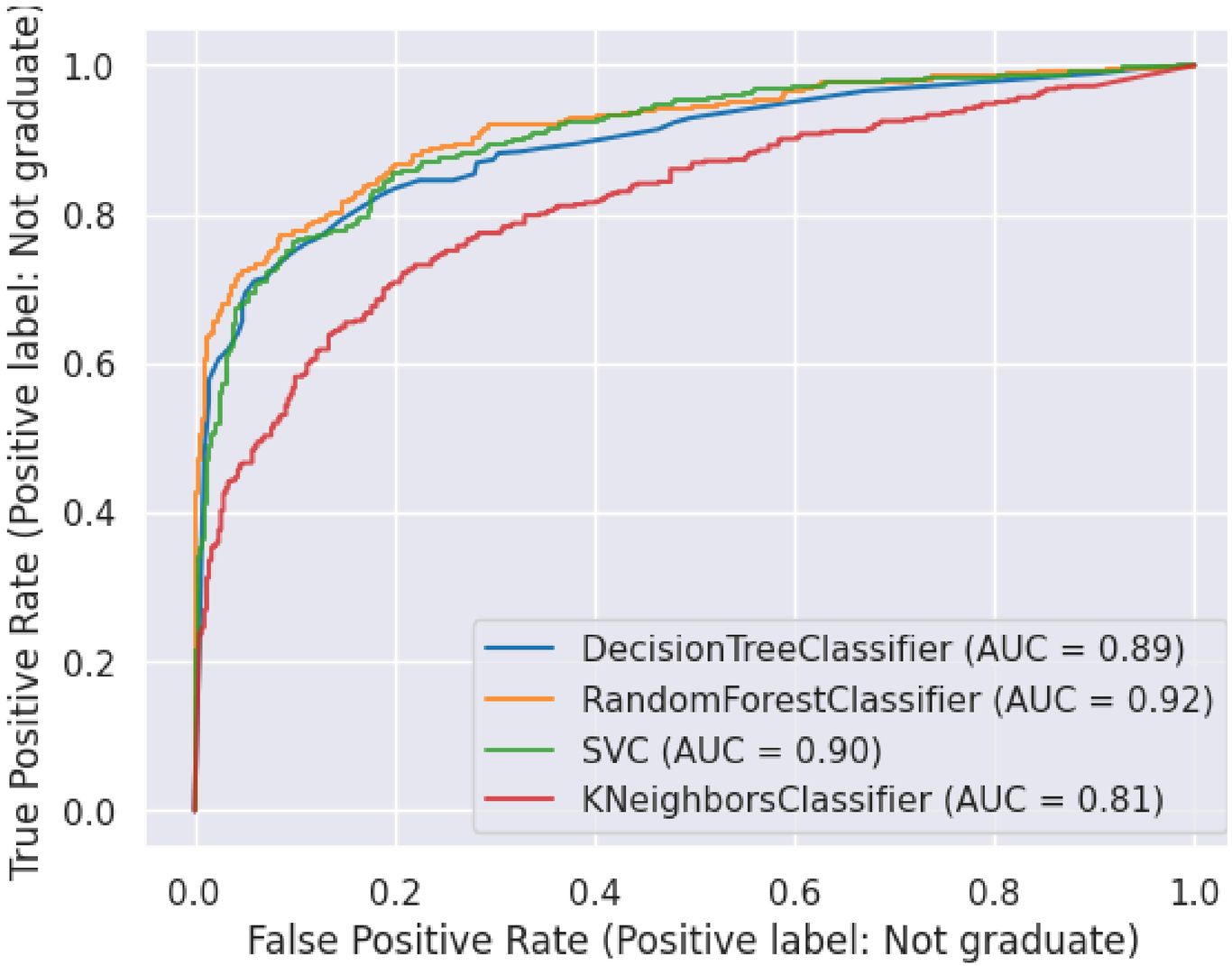
- Dropout e Enrolled vengono ricodificate come “Not graduate”
- Alcune variabili devono essere decodificate per essere analizzate meglio

Data Exploration e Data preprocessing

- Viene creata una funzione ‘def get_column_encoding(column_name: str, variables=variables)‘ che restituisce un dizionario con coppie chiave-valore corrispondenti alla codifica della colonna specificata.
- Vengono eliminate le colonne in ecesso per evitare di creare ridondanza e viene diminuita la cardinalità di determiante variabili per ridurre il numero di categorie uniche in una feature categorica e per prevenire overfitting
- Vengono iterate e decodificate attraverso un ciclo ‘for col in colms_to_decode: cols_decode[col] = get_column_encoding(col)‘
- Il ‘column preprocessing‘ permette di separare le colonne in due gruppi, categoriche (per il ‘One-Hot Encoding‘) e ordinali. Si usa un Column Tranformer che applica un ’Encoder‘ alle colonne categoriche e un‘altro (‘Ordinal Encoder‘) alle trasformando il dataframe X e preparandolo per l'uso dei modelli di classificazione successivi.

CLASSIFICAZIONE

ROC CURVE


True Positive Rate (Positive label: Not graduate)
False Positive Rate (Positive label: Not graduate)
0.0
0.2
0.4
0.6
0.8
1.0
DecisionTreeClassifier (AUC = 0.89)
RandomForestClassifier (AUC = 0.92)
SVC (AUC = 0.90)
KNeighborsClassifier (AUC = 0.81)

DECISION TREE

MODELLO DI DEFAULT

	precision	recall	f1-score	support
Graduate	0.76	0.75	0.75	418
Not graduate	0.78	0.79	0.78	467
accuracy			0.77	885
macro avg	0.77	0.77	0.77	885
weighted avg	0.77	0.77	0.77	885

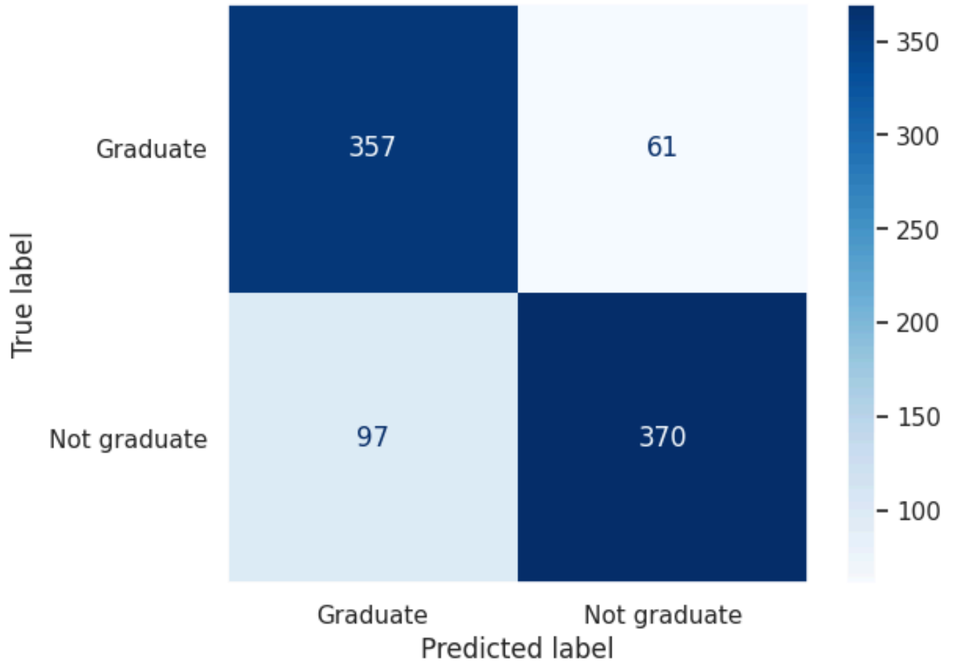
L'argoritmo struttura un albero in cui ogni nodo interno rappresenta un test su un carattere/attributo che attraverso Gini o Entropy migliora la divisione in classi.

RISULTATI DECISION TREE CLASSIFIER:

- PRECISION= 0,786
- RECALL= 0,85
- F1 score passa da 0,75 e 0,78 del modello di default a 0,82 nel modello fatto con greedsearchcv.
- ACCURACY aumenta da 0,77 a 0,82

BEST MODEL CON METODO GRIDSEARCHCV

	precision	recall	f1-score	support
Graduate	0.79	0.85	0.82	418
Not graduate	0.86	0.79	0.82	467
accuracy			0.82	885
macro avg	0.82	0.82	0.82	885
weighted avg	0.82	0.82	0.82	885



RANDOM FOREST

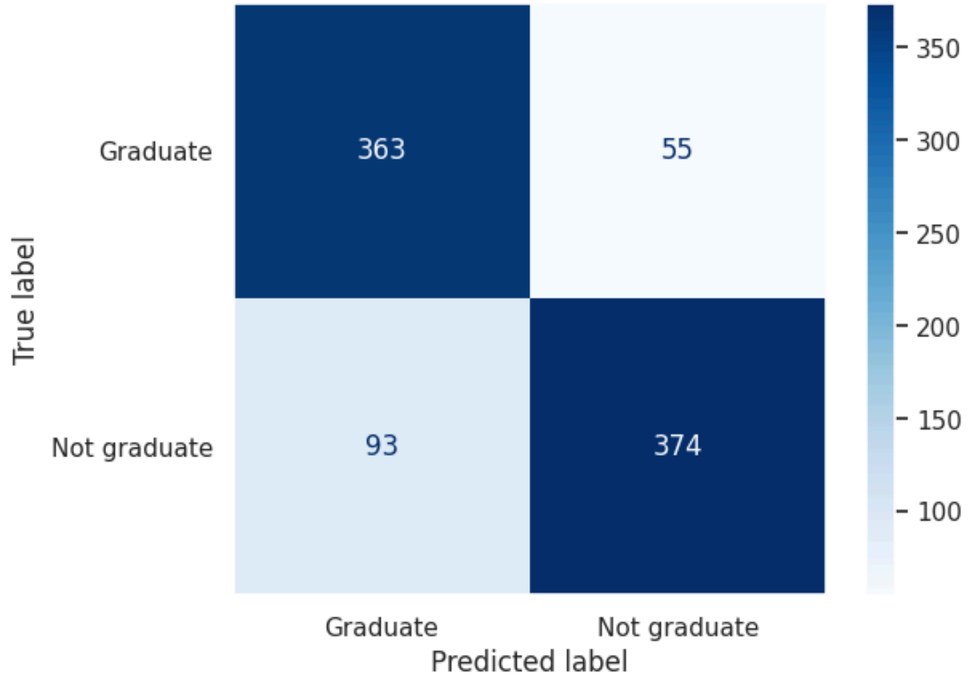
MODELLO DI DEFAULT

	precision	recall	f1-score	support
Graduate	0.79	0.89	0.84	418
Not graduate	0.89	0.79	0.84	467
accuracy			0.84	885
macro avg	0.84	0.84	0.84	885
weighted avg	0.84	0.84	0.84	885

BEST MODEL CON METODO GRIDSEARCHCV

	precision	recall	f1-score	support
Graduate	0.80	0.87	0.83	418
Not graduate	0.87	0.80	0.83	467
accuracy			0.83	885
macro avg	0.83	0.83	0.83	885
weighted avg	0.84	0.83	0.83	885

Durante la classificazione vengono combinati diversi alberi decisionali tra di loro riducendo il rischio di overfitting, RF si basa sul concetto di enseble learning



RISULTATI RANDOM FOREST CLASSIFIER:

- PRECISION= 0,796
- RECALL= 0,868
- F1 SCORE e ACCURACY diminuiscono da 0,84 a 0,83 passando dal modello di default al miglior modello, perciò le performance sembrano anche se di poco peggiorare

K-NEAREST NEIGHBORS

MODELLO DI DEFAULT

Modello di partenza:

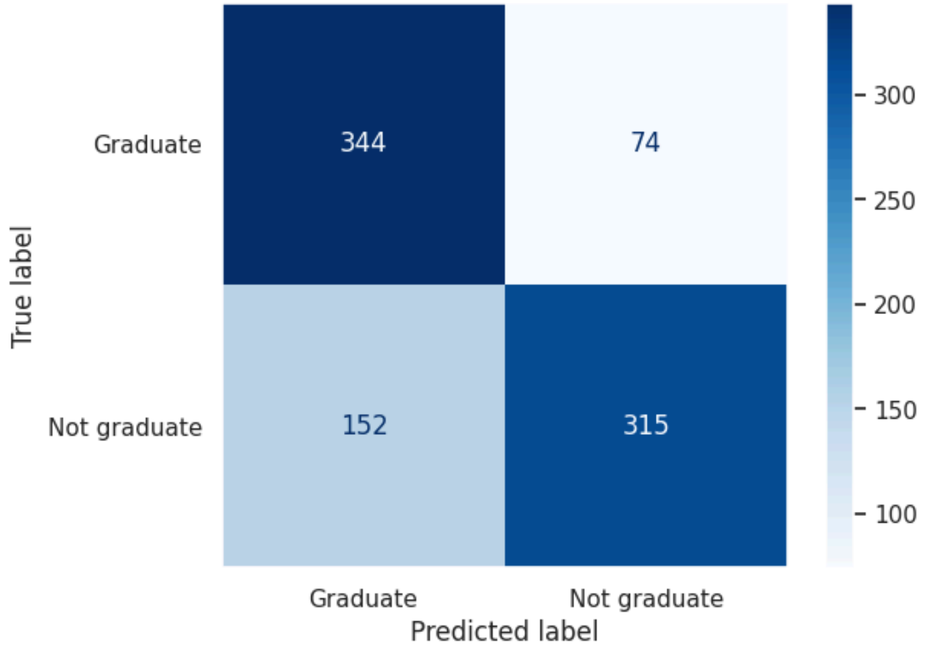
	precision	recall	f1-score	support
Graduate	0.68	0.78	0.73	418
Not graduate	0.77	0.67	0.72	467
accuracy			0.72	885
macro avg	0.73	0.73	0.72	885
weighted avg	0.73	0.72	0.72	885

BEST MODEL CON METODO GRIDSEARCHCV

Risultati del miglior modello:

	precision	recall	f1-score	support
Graduate	0.69	0.82	0.75	418
Not graduate	0.81	0.67	0.74	467
accuracy			0.74	885
macro avg	0.75	0.75	0.74	885
weighted avg	0.75	0.74	0.74	885

KNN definisce un parametro K, iperparametro che considera i punti vicini durante la classificazione. Calcola la distanza tra il punto di partenza e gli altri nel set.



RISULTATI KNN:

- PRECISION= 0,693
- RECALL= 0,823
- F1 SCORE aumenta di poco passando dal modello di dafault al modello con greedsearchcv da 0,73 e 0,72 a 0,75 e 0,74.
- ACCURACY aumenta da 0,72 a 0,74

SUPPORT VECTOR CLASSIFIER

MODELLO DI DEFAULT

Risultati

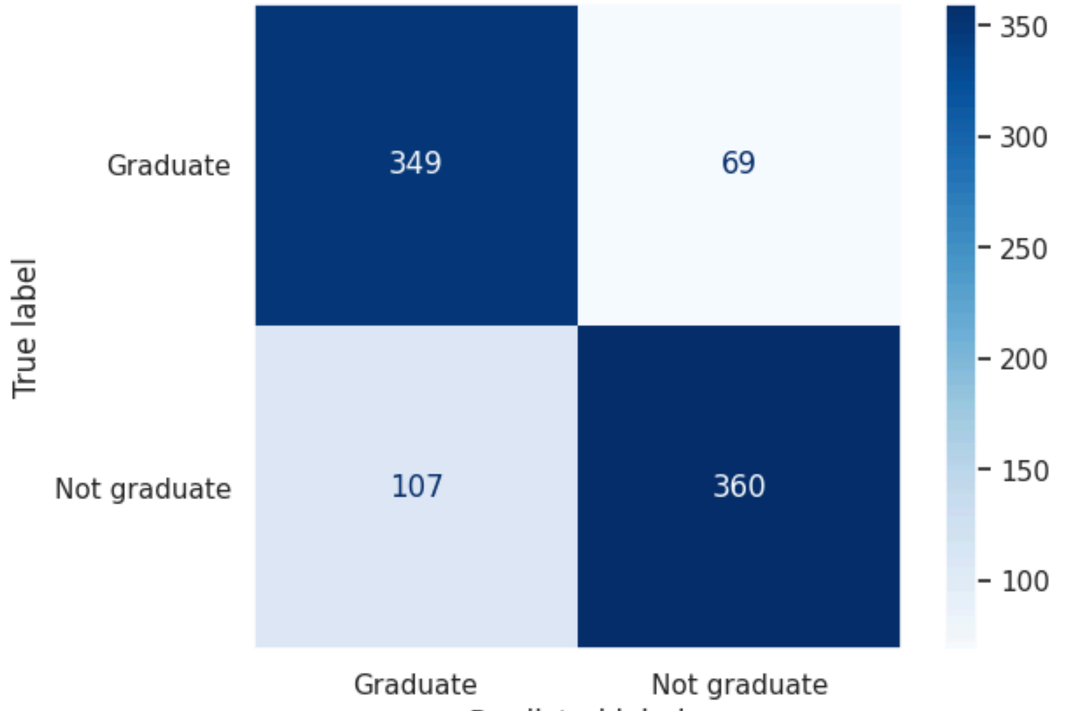
	precision	recall	f1-score	support
Graduate	0.78	0.87	0.82	418
Not graduate	0.87	0.78	0.82	467
accuracy			0.82	885
macro avg	0.82	0.82	0.82	885
weighted avg	0.82	0.82	0.82	885

BEST MODEL CON METODO RANDOMSEARCHCV

Risultati con tweak dei parametri

	precision	recall	f1-score	support
Graduate	0.78	0.84	0.81	418
Not graduate	0.85	0.78	0.81	467
accuracy			0.81	885
macro avg	0.81	0.81	0.81	885
weighted avg	0.81	0.81	0.81	885

SVC cerca il margine migliore per separare le classi, in questo caso per margine si intende la distanza tra l'iperpiano e i punti di addestramento più vicini.

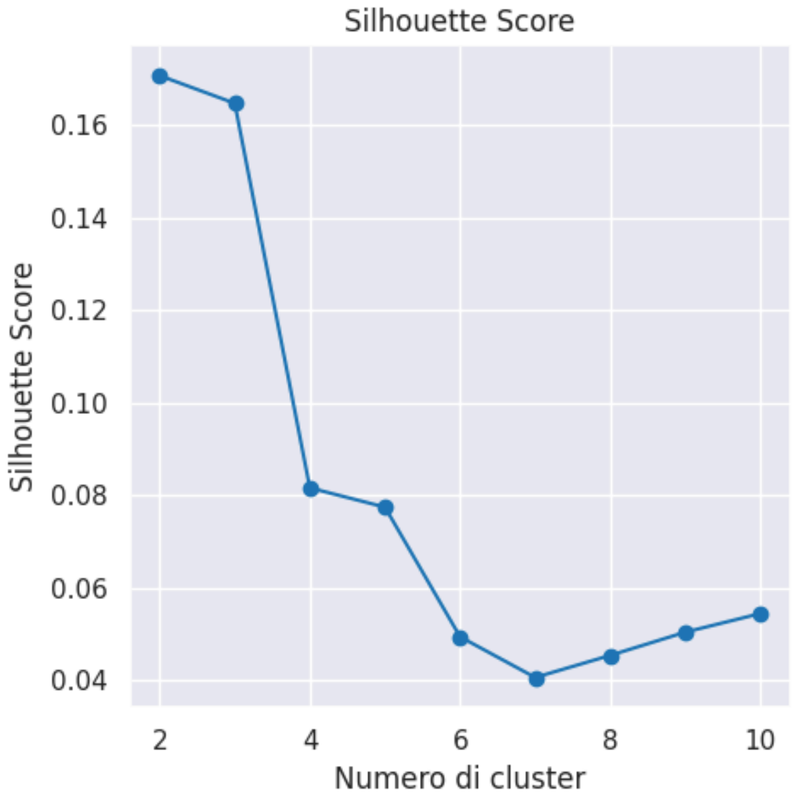


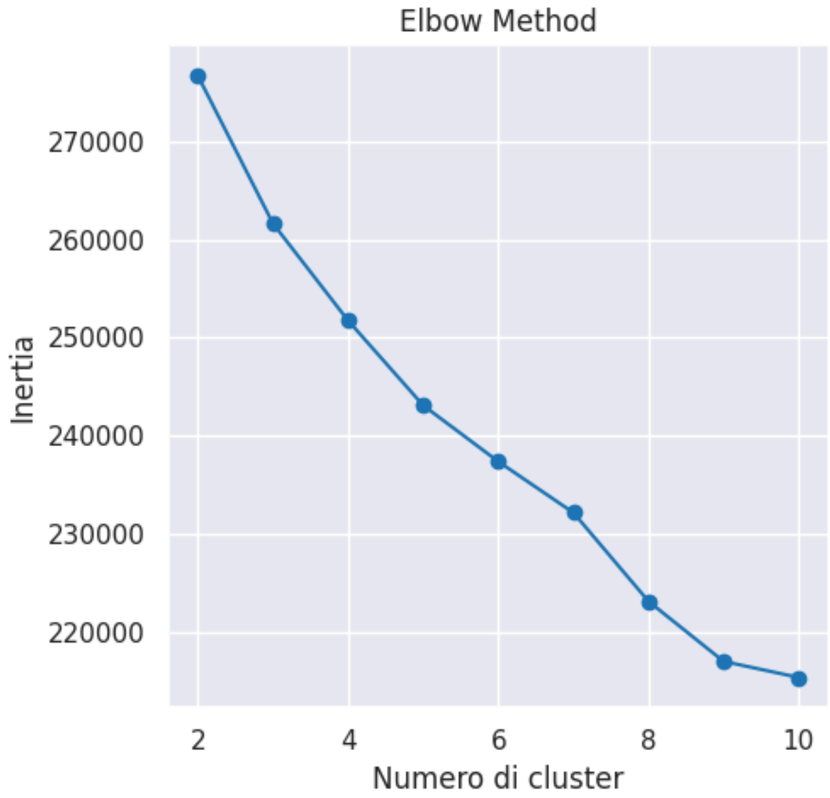
RISULTATI SUPPORT VECTOR CLASSIFIER:

- **PRECISION= 0,765**
- **RECALL= 0,835**
- **F1 SCORE diminuisce passando dal modello di dafault al modello di greedsearchcv da 0,82 a 0,81**
- **ACCURACY diminuisce da 0,82 a 0,81**
- **sembra esserci un peggioramento passando dal modello di default al miglior modello**

CLUSTERING

K-MEANS





K-Means è un algoritmo di clustering pensato per raggruppare i dati in cluster, raggruppamenti, basati sulla loro somiglianza senza utilizzare informazioni sulla variabile di output che si desidera predire, in questo caso il drop-out scolastico.

RISULTATI:

Sia il silhouette score sia l'elbow method identificano 2 cluster, tuttavia con un Adjusted Rand Index di 0.0593 i cluster non sembrerebbero molto simili. Le osservazioni potrebbero essere state assegnati in modo diverso nei cluster e casualmente, senza una chiara struttura.

Considerazioni Finali e Sviluppi Futuri

- Dataset si presta ad utilizzare maggiormente modelli di classificazione.
- Riduzione variabili: provare a fare analisi componenti principali, durante la fase di data cleaning.
- Per migliorare la classificazione: eliminare features meno importanti.
- Utilizzare e implementare l’algoritmo della Regressione Logistica, data la natura delle variabili.
- Provare ad implementare una soluzione di clustering come k-modes o k-prototype, quindi invece di convertire i dati categoriali in numerici provare a fare lo scaling.