Progetto data mining AA 2022-2023

NASA - Nearest Earth Object Hazard Detection

Vito Proscia mat. 735975

Table of content

Near Earth Object

Un **Near Earth Object (NEO)** è un oggetto celeste che orbita intorno al Sole e che ha un'orbita che lo porta vicino alla Terra. Questi oggetti possono essere asteroidi, comete o detriti spaziali. La loro classificazione come pericolosi o meno dipende dalla loro **dimensione**, **velocità** e **distanza dalla Terra**.
La classificazione accurata dei NEO è fondamentale non solo per prevenire eventuali collisioni con la Terra, ma il solo passaggio può **influenzare fenomeni naturali**.
Per questo motivo, la NASA e altre agenzie spaziali monitorano costantemente i NEO e cercano di classificarli in base al loro grado di pericolosità.

Obiettivo

L'obiettivo principale della sperimentazione è l'addestramento di un modello di classificazione binaria per poter predire accuratamente se un NEO può potenzialmente essere un pericolo per la terra.

Inoltre si vogliono evidenziare le differenza nell'addestramento provando varie tecniche per "risolvere" il problema dello sbilanciamento della classi

Metodologia

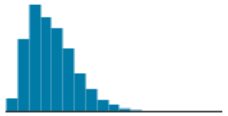
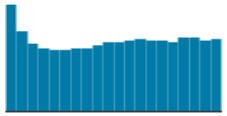
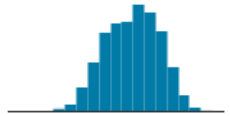
Per condurre la sperimentazione è stato usato il **CRISP-DM**, una metodologia agile standard per l'analisi dei dati e il data mining che fornisce una struttura organizzativa per gestire progetti di scoperta delle conoscenze.

Il processo CRISP-DM è iterativo, il che significa che è possibile tornare indietro a fasi precedenti se necessario per apportare modifiche o miglioramenti. Questo processo flessibile consente di affrontare i progetti di data mining in modo strutturato ed efficiente.

Dataset

Il dataset preso in esame è stato prodotto direttamente dalla NASA ed è composto da informazioni su migliaia di Near Earth Objects rilevati in decenni. Tuttavia, il dataset è **sbilanciato** poiché la maggior parte degli oggetti non rappresenta una minaccia per la Terra. Questo rende la classificazione dei NEO ancora più difficile e impegnativa.
Ne risulta un dataset fortemente sbilanciato (90.3% false e 9.7% true) per la target feature *hazardous*

Dataset

id	name	est_diameter_min	est_diameter_max	relative_velocity	miss_distance	orbiting_body	sentry_object	absolute_magnit...	hazardous
Unique Identifier for each Asteroid	Name given by NASA	Minimum Estimated Diameter in Kilometres	Maximum Estimated Diameter in Kilometres	Velocity Relative to Earth in Kmph	Distance in Kilometres missed	Planet that the asteroid orbits	Included in sentry - automated collision monitoring system	Describes intrinsic luminosity	Boolean feature that shows whether asteroid is harmful or not
 2.00m 54.3m	**27423** unique values	 0 37.9	 0 84.7	 203 237k	 6.75k 74.8m	**1** unique value	 true 0 0% false 90.8k 100%	 9.23 33.2	 true 8840 10% false 82.0k 90%
2162635	162635 (2000 SS164)	1.1982708007	2.6794149658	13569.2492241812	54839744.08284605	Earth	False	16.73	False
2277475	277475 (2005 WK4)	0.2658	0.5943468684	73588.7266634981	61438126.52395093	Earth	False	20.0	True
2512244	512244 (2015 YE18)	0.7220295577	1.6145071727	114258.6921290512	49798724.94045679	Earth	False	17.83	False
3596030	(2012 BV13)	0.096506147	0.2157943048	24764.3031380016	25434972.72075825	Earth	False	22.2	False
3667127	(2014 GE35)	0.2550086879	0.5702167609	42737.7337647264	46275567.00130072	Earth	False	20.09	True
54138696	(2021 GY23)	0.0363542322	0.0812905344	34297.5877783029	40585691.22792288	Earth	False	24.32	False
54189957	(2021 PY40)	0.1716148941	0.3837425691	27529.4723069673	29069121.41864897	Earth	False	20.95	False
54230078	(2021 XD6)	0.0053278866	0.0119135167	57544.4700827352	55115019.25807114	Earth	False	28.49	False
2088213	88213 (2001 AF2)	0.3503926411	0.7835017643	56625.2101223615	69035980.03881611	Earth	False	19.4	False
3766065	(2016 YM)	0.1058168859	0.2366137501	48425.8403287922	38355261.56076106	Earth	False	22.0	False
54049873	(2020 OT6)	0.2526707542	0.5649889822	58430.6971996129	38337496.948336646	Earth	False	20.11	True
54099949	(2020 XW4)	0.1529519353	0.3420109247	64393.9283164601	71983105.30586366	Earth	False	21.2	False
54104555	(2021 AW1)	0.0699125232	0.1563291544	38018.6152911655	52093021.60346941	Earth	False	22.9	False

Pre-processing

Il **pre-processing** è uno step **fondamentale** che occorre sotto il nome di data preparation nella metodologia CRISP-DM, una delle fasi più importanti che influenza notevolmente il risultato finale della sperimentazione, consiste nell'eseguire una serie di operazioni per migliorare la qualità dei dati **in funziona dell'obiettivo finale**.

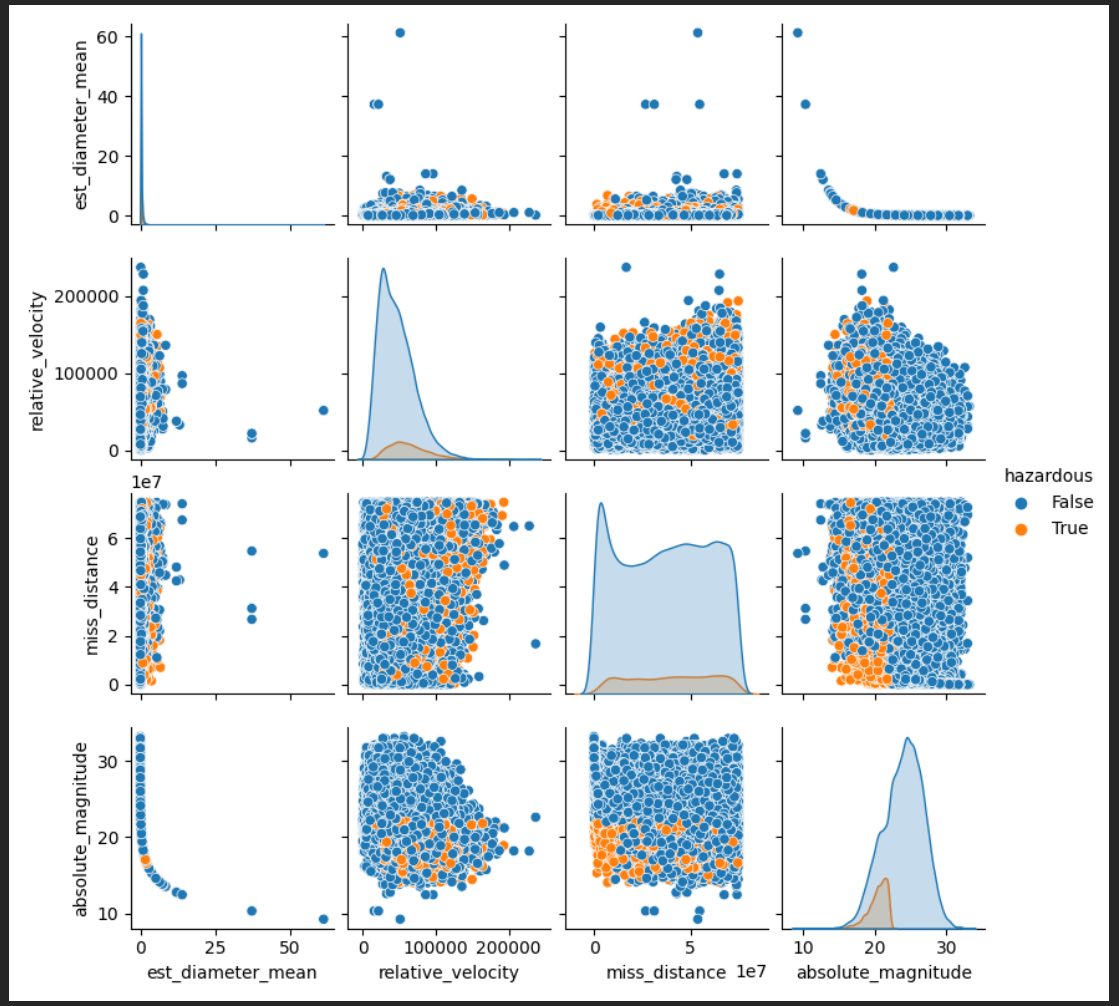
Nel nostro caso dal dataset originale si sono eliminate alcune features risultate inutili per l'analisi (id, name, orbiting_body, sentry_object)

Imbalanced learning

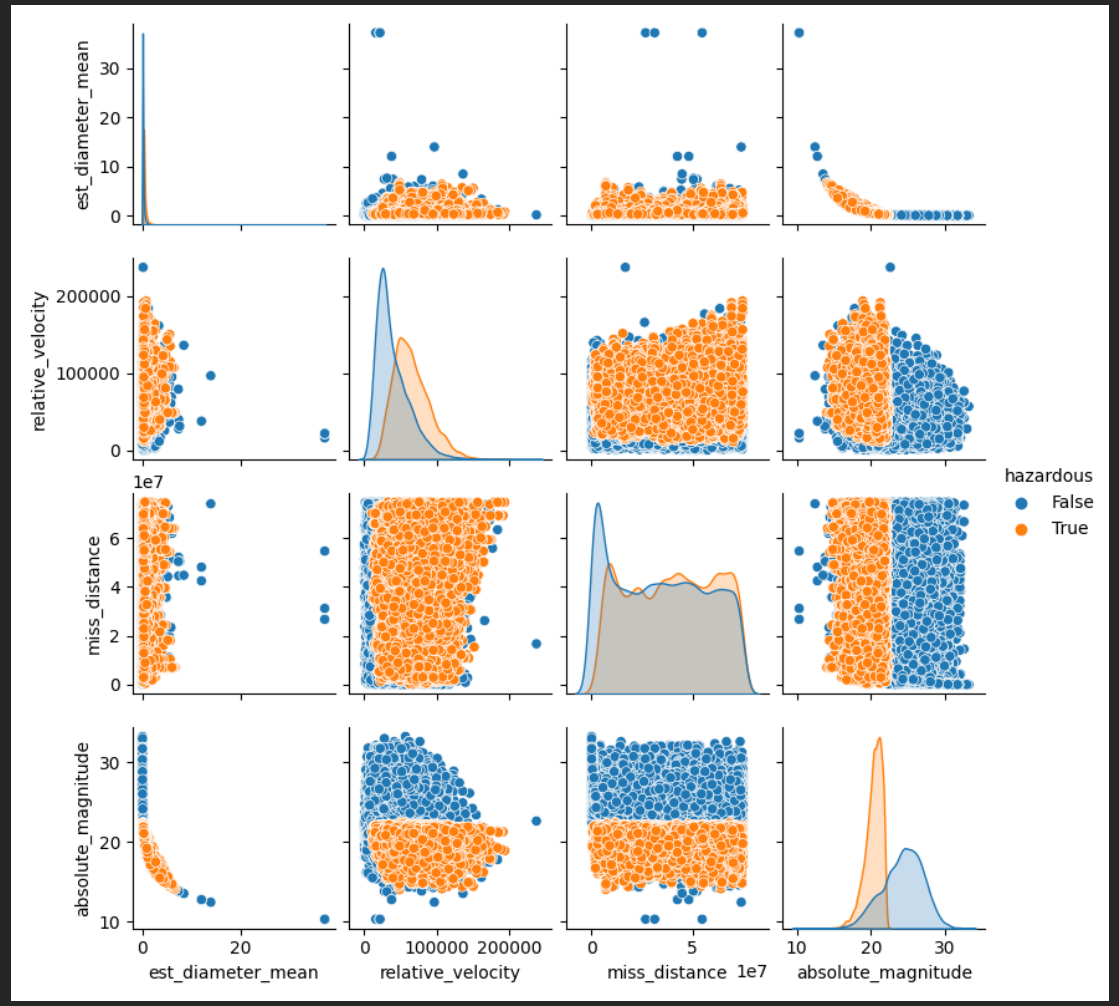
Avendo un dataset fortemente sbilanciato è stato necessario ricorrere a delle tecniche che "migliorassero la situazione". In particolare si sono usati due metodi che trattano il problema da due punti di vista differenti:

- **Resampling** (data), in particolare si è usato *SMOTEENN* ed *ADASYN*
- **Cost-sensitive learning** (algorithm), si è applicata una matrice di costo che va a definire il costo di una classificazione sbagliata nel calcolo dell'errore in fase di learning

Imbalanced learning



Dataset originale



Resampling

Modelli

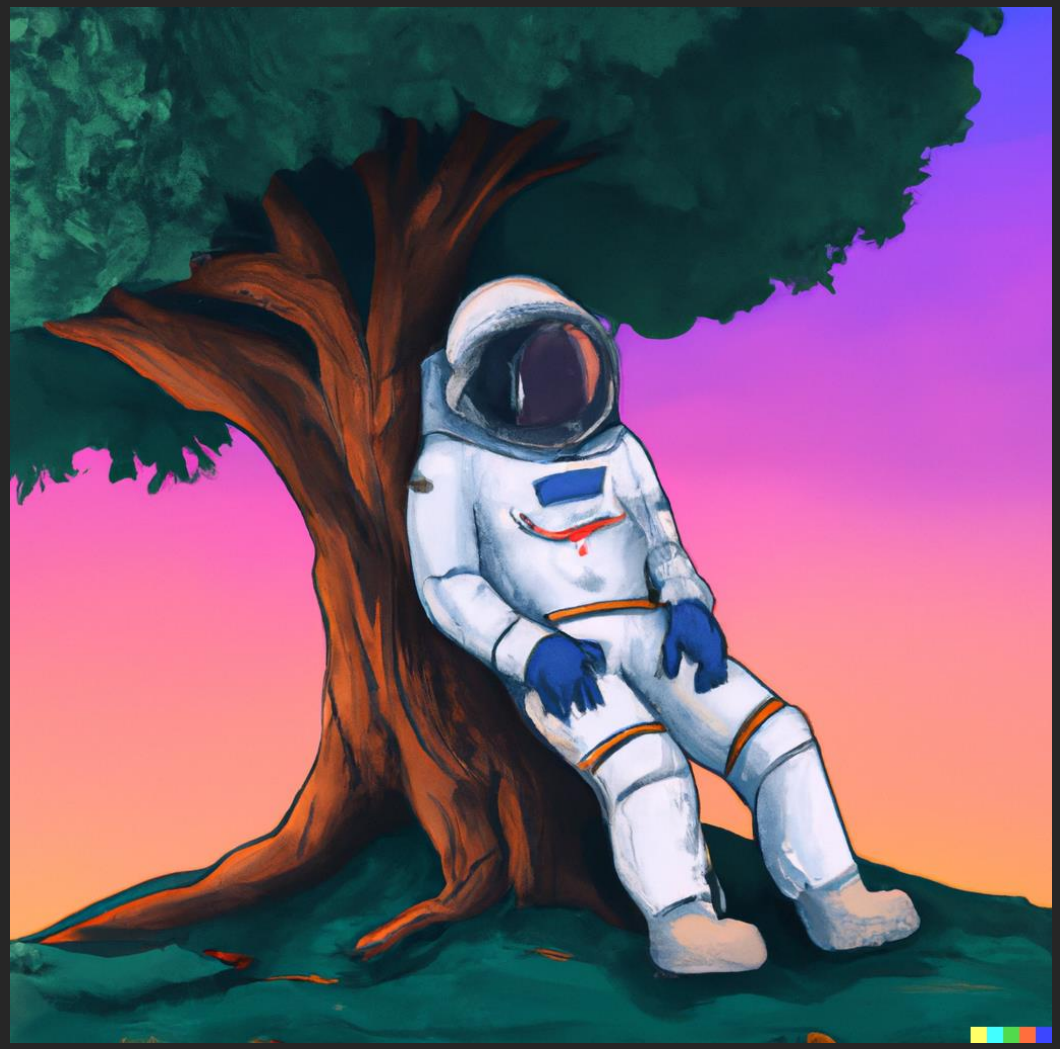
Per il task di classificazione si sono presi in considerazione principalmente due modelli

Albero di classificazione C4.5

Rappresenta una serie di **decisioni basate su attributi una classe** a un determinato dato di input. Ogni ramo dell'albero rappresenta una scelta decisionale che porta ad un'etichetta di classe finale.

Naive Bayes

Si basa sull'assunzione di indipendenza condizionale tra le caratteristiche del dato in ingresso per determinare la probabilità di appartenenza a una determinata classe. Utilizza il **teorema di Bayes** per calcolare la probabilità a posteriori delle classi dato un insieme di caratteristiche.



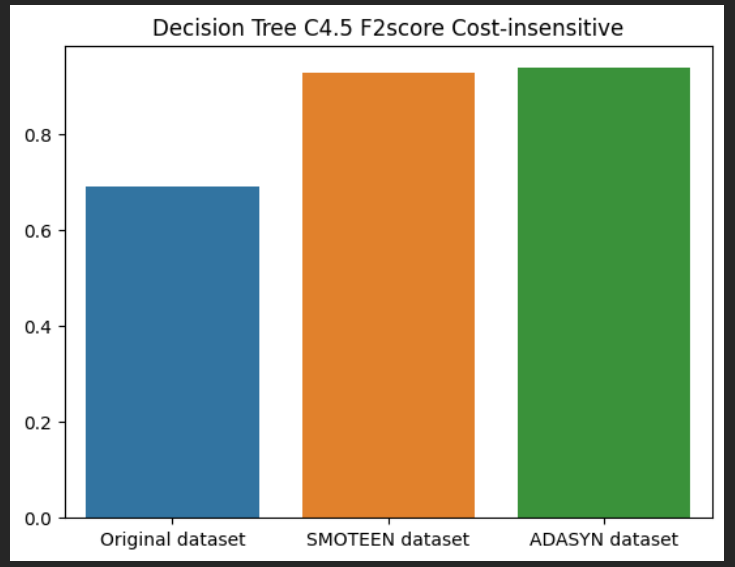
Risultati

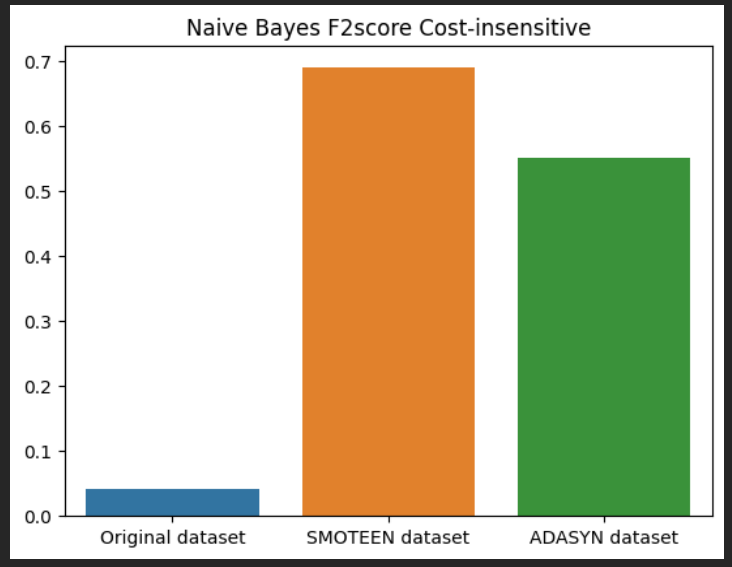
I risultati ottenuti dai vari modelli con le diverse combinazioni mostrano chiaramente le differenze tra i vari metodi di pre-processing applicati, in particolare si sono considerate principalmente l'*f-score* e *AUC ROC* per la valutazione dei modelli, in quanto **non dipendo dalla distribuzione della classi** tanto quanto le metriche classiche (*precision*, *recall*, *accuracy*, ...).

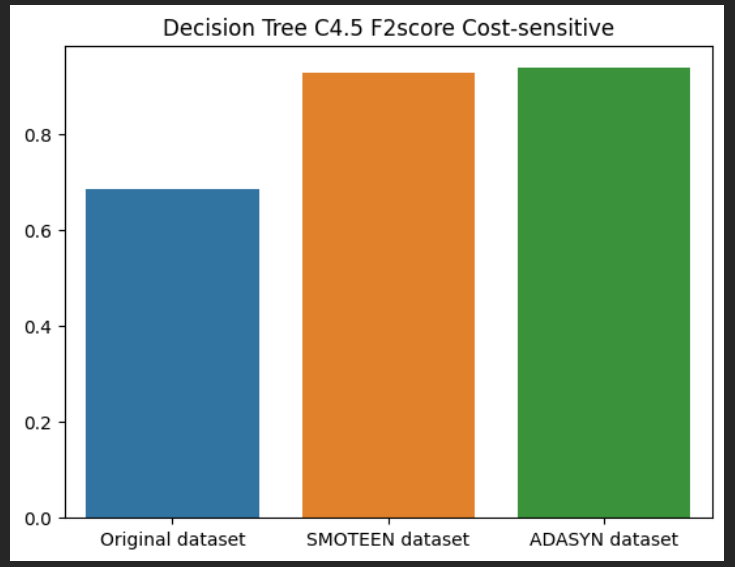
Si sono presi in esame i risultati dei modelli allenati con il dataset originale, il dataset *SMOTEENN* ed il dataset *ADASYN*, a cui sono stati applicati algoritmi cost-insensitive e cost-sensitive

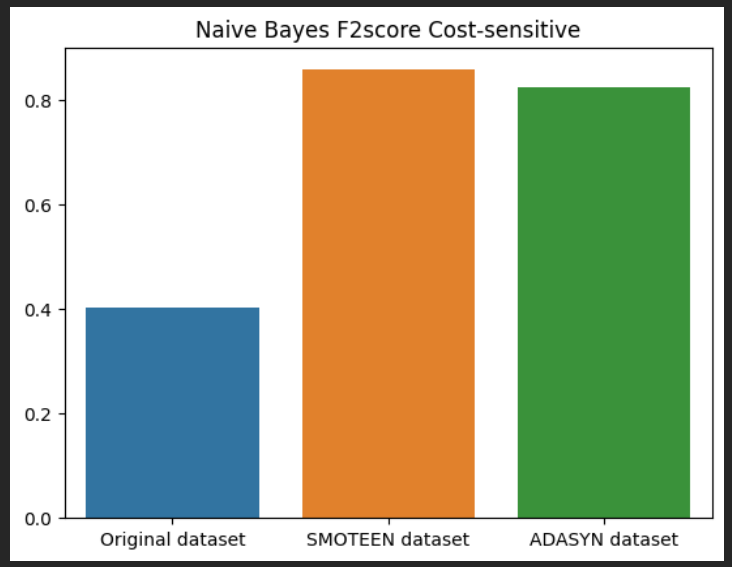
Risultati

Possiamo subito apprezzare il notevole miglioramento dovuto principalmente all'aver eseguito il modelli sui dataset che hanno subìto il ribilanciamento delle classi sia con SMOTEENN che con ADASYN.

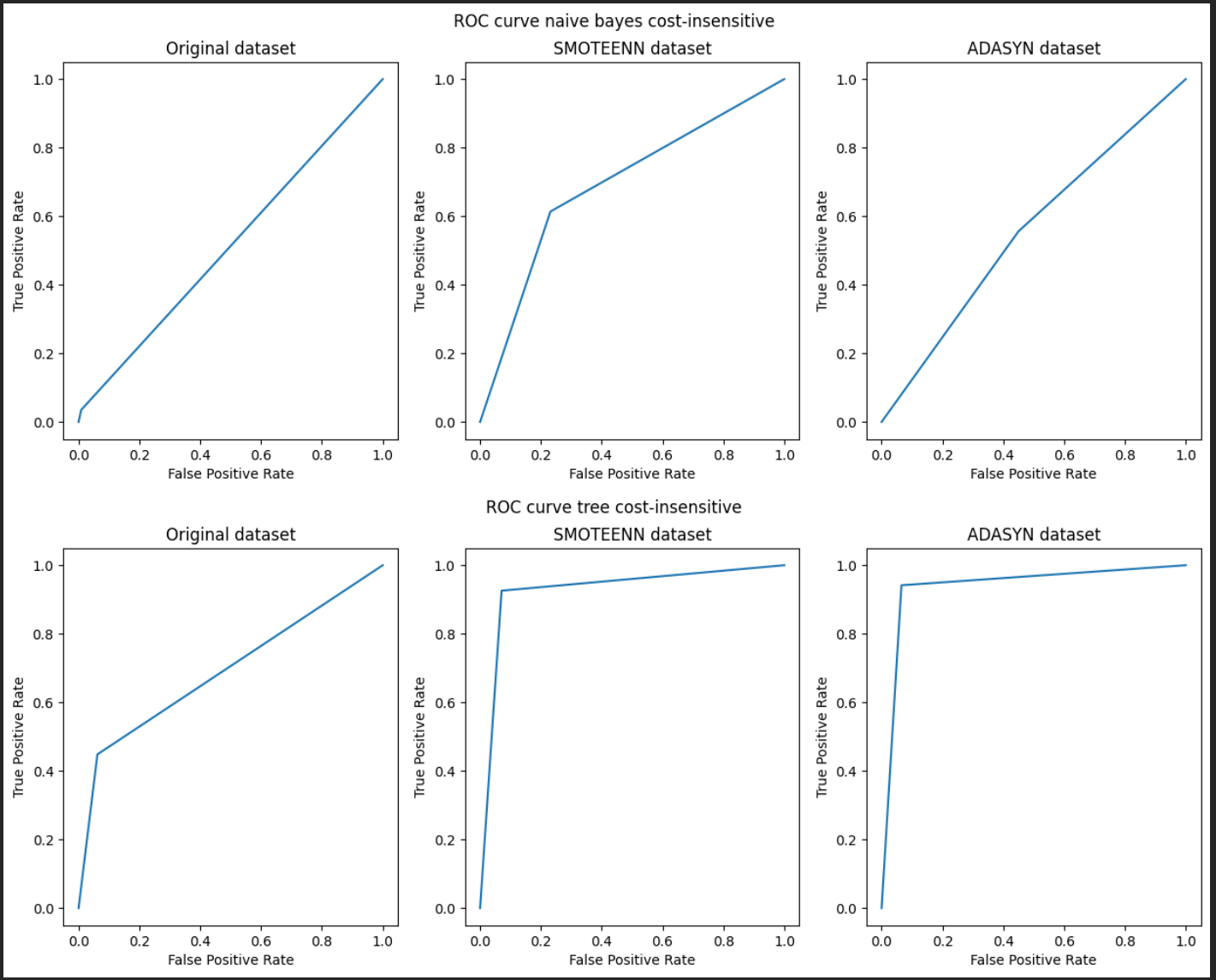








Risultati


ROC curve naive bayes cost-insensitive
Original dataset
SMOTEENN dataset
ADASYN dataset
ROC curve tree cost-insensitive
Original dataset
SMOTEENN dataset
ADASYN dataset
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Conclusioni

In conclusione possiamo dire che il modello che è risultato migliore ai fini dell'obiettivo è il Classifiation Tree C4.5, avendo molti indicatori superiori al 90%.
Inoltre possiamo considerare cruciale il miglioramento ottenuto, in ogni caso, effettuando operazioni di **resampling** sui dati, andando a ribilanciare le classi.

Tuttavia, ci sono ancora sfide e possibilità di **miglioramento**, che possono essere affrontate attraverso ulteriori ricerche.