VCDD:
riduzione dimensionale, relative visualizzazioni e classificazione

Il dataset VCDD

Il dataset utilizzato per gli esperimenti è il **Virtual Celiac Disease Diagnosis** (VCDD), un insieme di dati relativi a pazienti virtuali generato per addestrare un classificatore fuzzy per la diagnosi della celiachia.

Le informazioni presenti nel dataset sono:

- anemia, osteopenia, diarrea cronica, mancata crescita, disturbi genetici, madre celiaca, POCT, esami del sangue e class: valori booleani;
- IGA totali , TTG IGG, TTG IGA: numeri reali positivi.

Tecniche di riduzione dimensionale

L'obiettivo delle tecniche di riduzione dimensionale è quello di eseguire un mapping dallo spazio iniziale a uno spazio dimensionale inferiore.

Le tecniche che sono state utilizzate sono:

- PCA;
- Kernel PCA;
- MDS;
- t-SNE.

PCA

PCA (Principal Component Analysis) utilizza una trasformazione ortogonale per convertire una serie di osservazioni di variabili correlate in un insieme di valori linearmente non correlati, chiamati **componenti principali**.

La trasformazione è effettuata in modo tale che la prima componente principale abbia la massima varianza e ogni componente principale successiva abbia a sua volta la massima varianza e sia ortogonale alle componenti principali calcolate precedentemente.

Kernel PCA

Kernel PCA permette di generalizzare PCA in modo tale da effettuare una riduzione dimensionale non lineare. Ciò viene fatto mediante il cosiddetto “**kernel trick**”.

I kernel più comunemente usati sono:

- Gaussiano;
- Polinomiale;
- Sigmoidale.

MDS

MDS (Multi-dimensional scaling) è una tecnica esplorativa dei dati che consente di ottenere una rappresentazione di n oggetti in k dimensioni partendo da informazioni inerenti la similarità (o dissimilarità) tra ciascuna coppia di oggetti.

Si costruiscono due matrici, una contenente le distanze tra ogni coppia degli n punti nello spazio di partenza e l'altra le distanze tra le immagini nello spazio di dimensione k.

L'obiettivo è trovare la configurazione per la quale le distanze nello spazio di dimensione k siano più simili possibile alle distanze originali.

t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) è una tecnica di riduzione della dimensionalità non lineare che modella i punti in modo tale che oggetti vicini nello spazio originale risultino vicini nello spazio a dimensionalità` ridotta, ed oggetti lontani risultano lontani, cercando di preservare la struttura locale.

L'algoritmo consiste di 2 fasi:

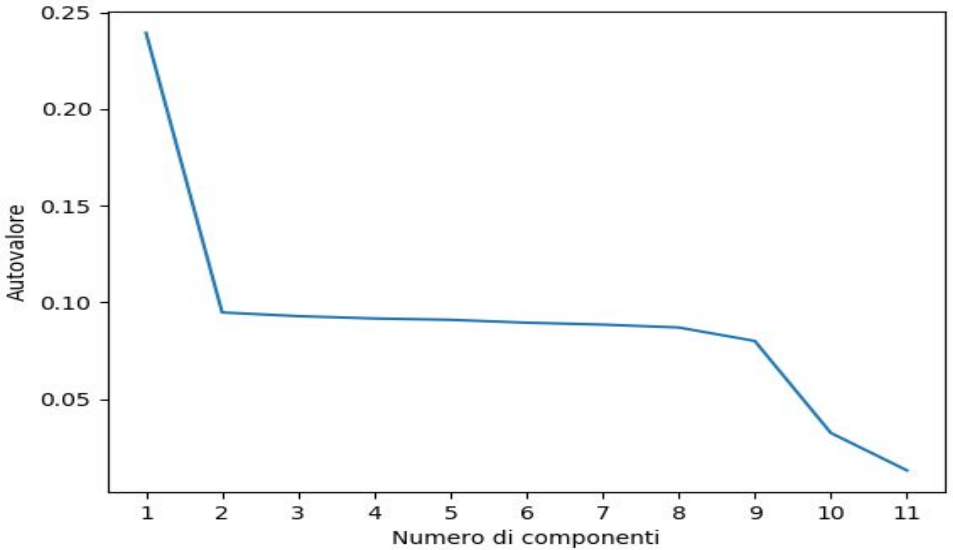
1. vengono costruite due distribuzioni di probabilità (una per lo spazio originale ed una per lo spazio a dimensionalità ridotta) che ad ogni coppia di punti associano un valore di probabilità` elevato se i due punti sono simili, basso se sono dissimili.

2. si minimizza la divergenza di Kullback-Leibler delle due distribuzioni tramite gradient descent, riorganizzando i punti nello spazio a dimensione ridotta

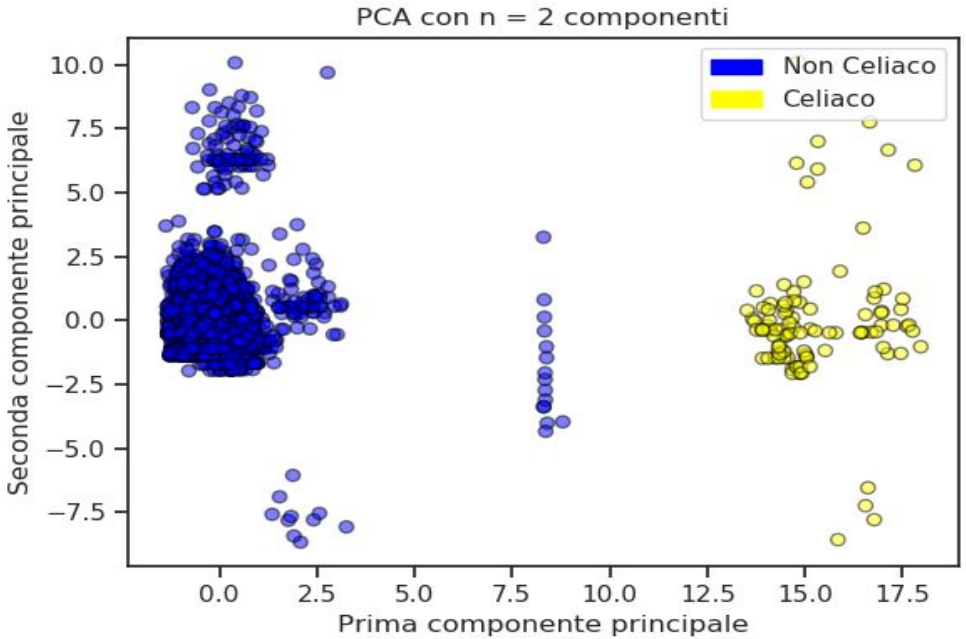
Tecniche di visualizzazione implementate

- Scatterplot 2D per dati bidimensionali
- Scatterplot 2D per dati tridimensionali
- Scatterplot 3D interattivo
- Plot di Draftman

Scree-plot relativo alle componenti principali della PCA per il dataset VCDD



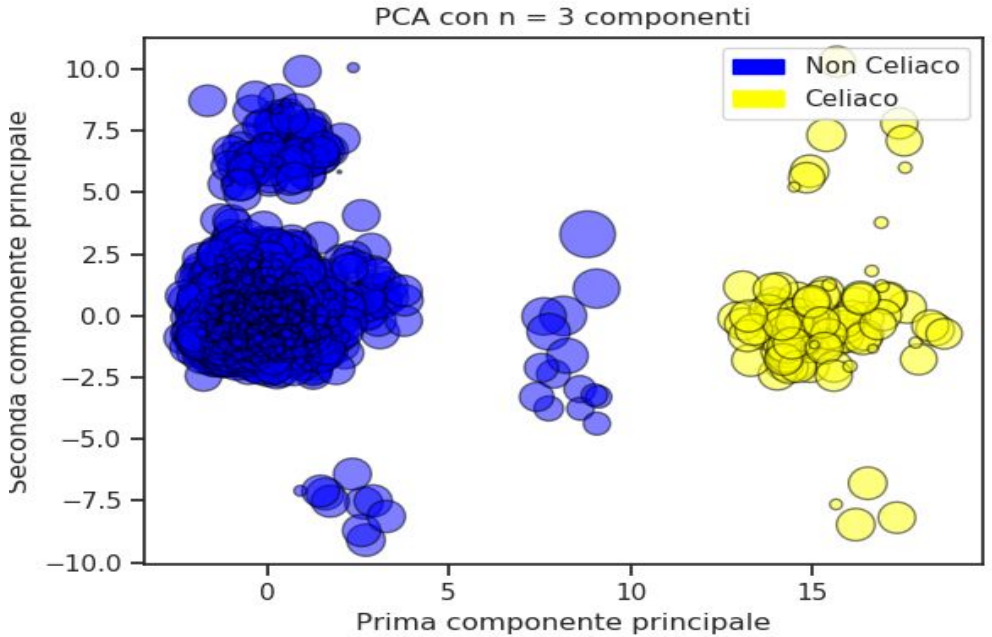
Scatterplot 2D bidimensionale applicato su PCA



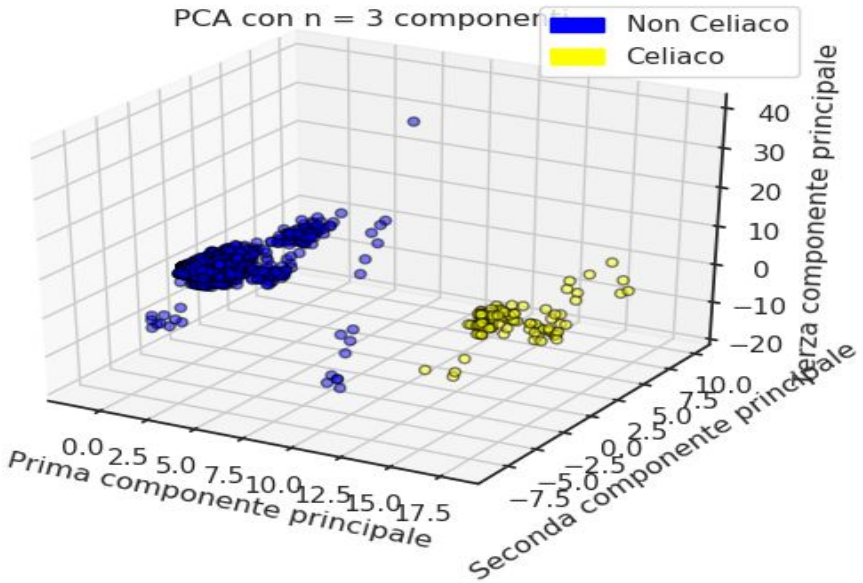
Lista dei pazienti corrispondenti ai punti allineati del cluster centrale

	Anemia	Osteopenia	Diarrea Cronica	Mancata Crescita	Disturbi Genetici	Madre Celiaca	POCT	IGA totali	TTG IGG	TTG_IGA	Esami del sangue
Paziente 1	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.19	0.00	NaN	0.0
Paziente 2	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.05	0.32	NaN	0.0
Paziente 3	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.23	0.51	NaN	0.0
Paziente 4	1.0	1.0	0.0	0.0	0.0	0.0	2.0	0.18	2.40	NaN	0.0
Paziente 5	0.0	1.0	1.0	0.0	0.0	0.0	2.0	0.24	0.00	NaN	0.0
Paziente 6	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.07	2.09	NaN	0.0
Paziente 7	1.0	1.0	0.0	0.0	0.0	0.0	2.0	0.08	0.00	NaN	0.0
Paziente 8	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.22	0.00	NaN	0.0
Paziente 9	1.0	1.0	0.0	0.0	0.0	0.0	2.0	0.04	0.78	NaN	0.0
Paziente 10	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.16	1.87	NaN	0.0
Paziente 11	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.07	1.64	NaN	0.0
Paziente 12	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.16	0.70	NaN	0.0
Paziente 13	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.13	4.01	NaN	0.0
Paziente 14	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.25	2.44	NaN	0.0
Paziente 15	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.24	0.00	NaN	0.0

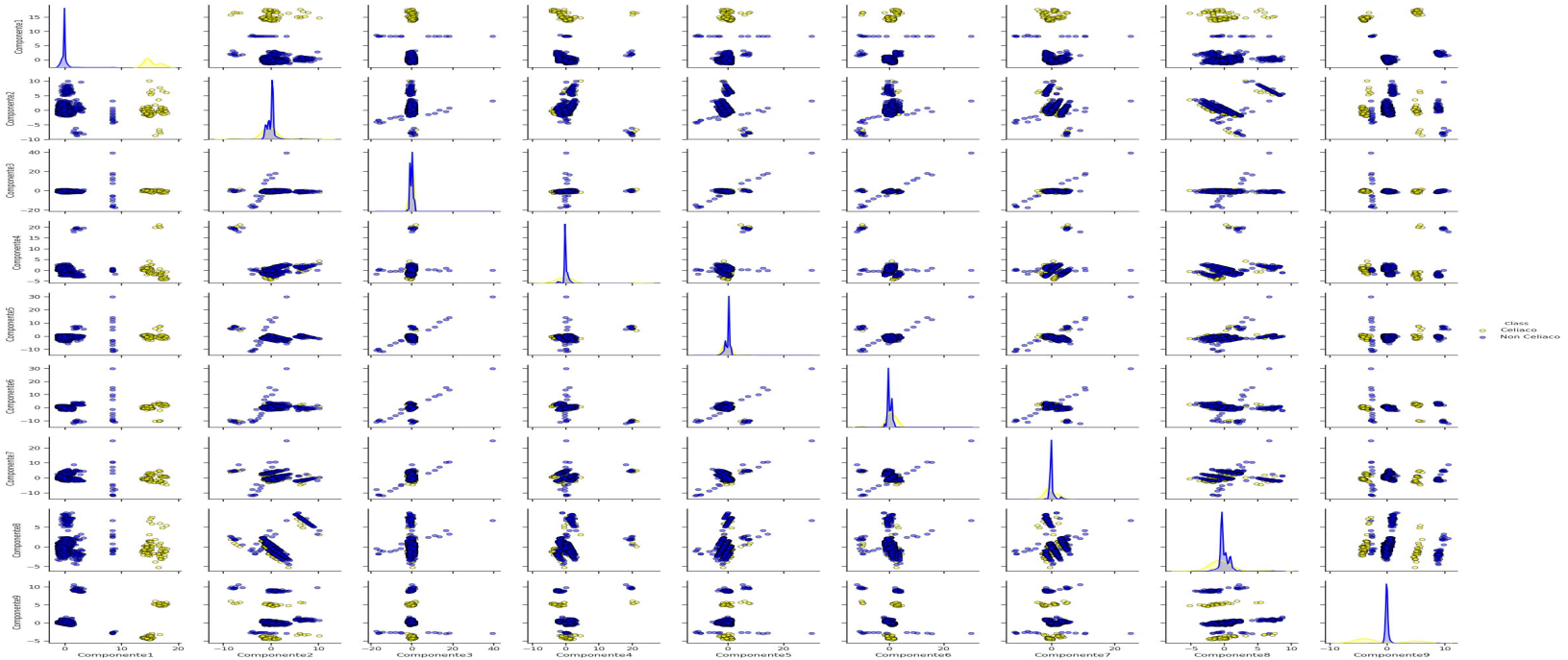
Scatterplot 2D tridimensionale applicato su PCA



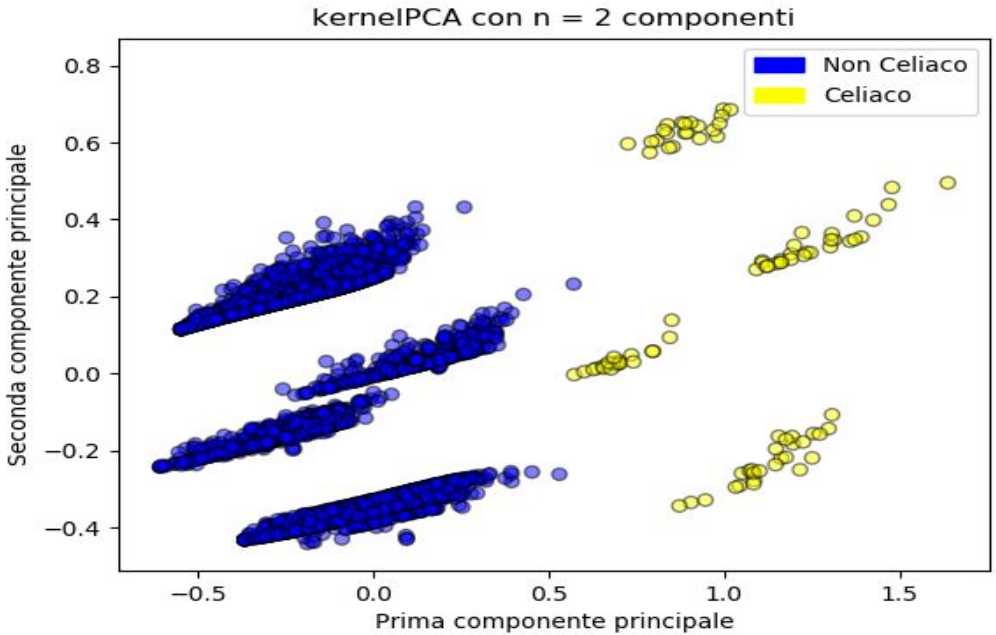
Scatterplot 3D interattivo applicato su PCA



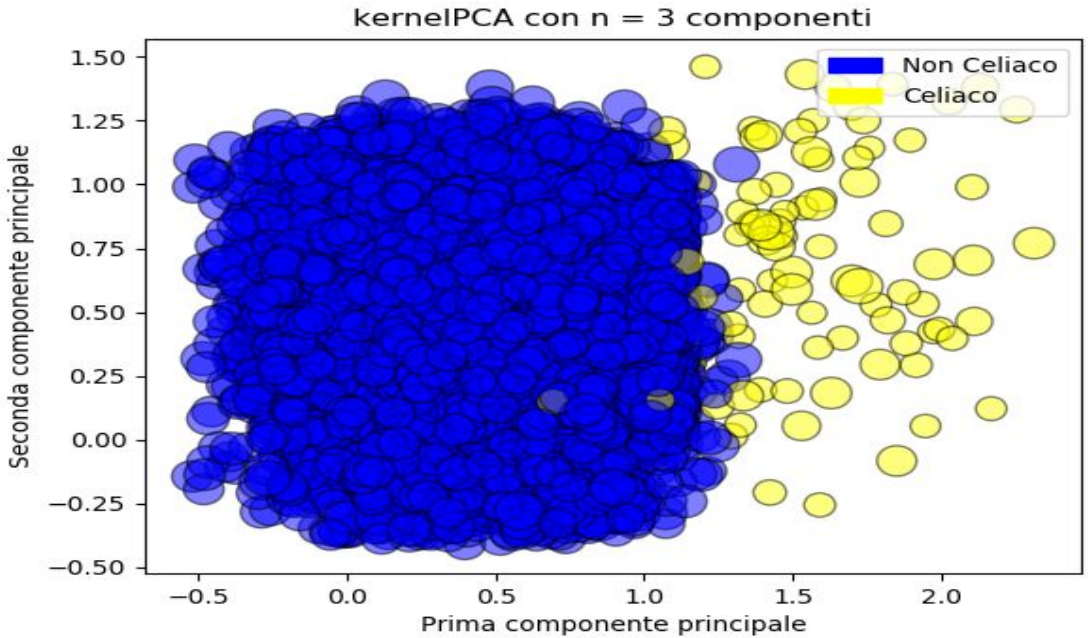
Plot di Draftman applicato su PCA



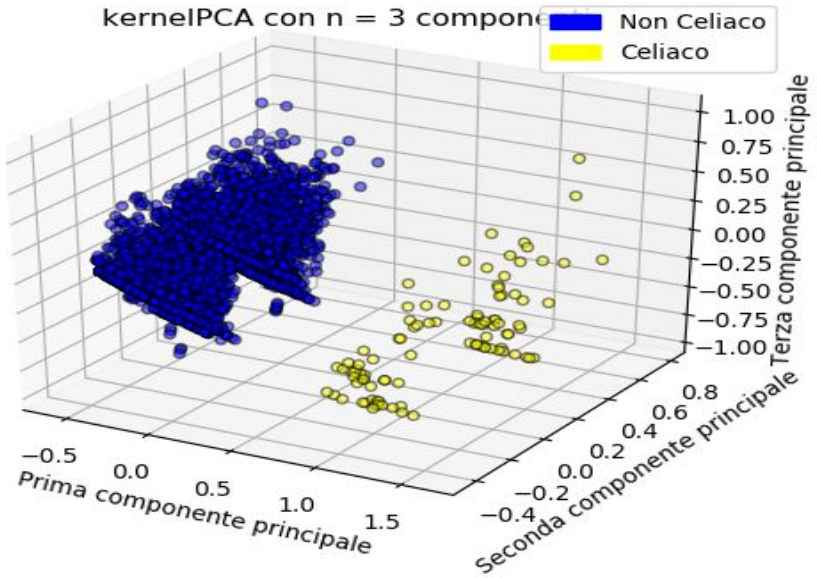
Scatterplot 2D bidimensionale applicato su kernelPCA



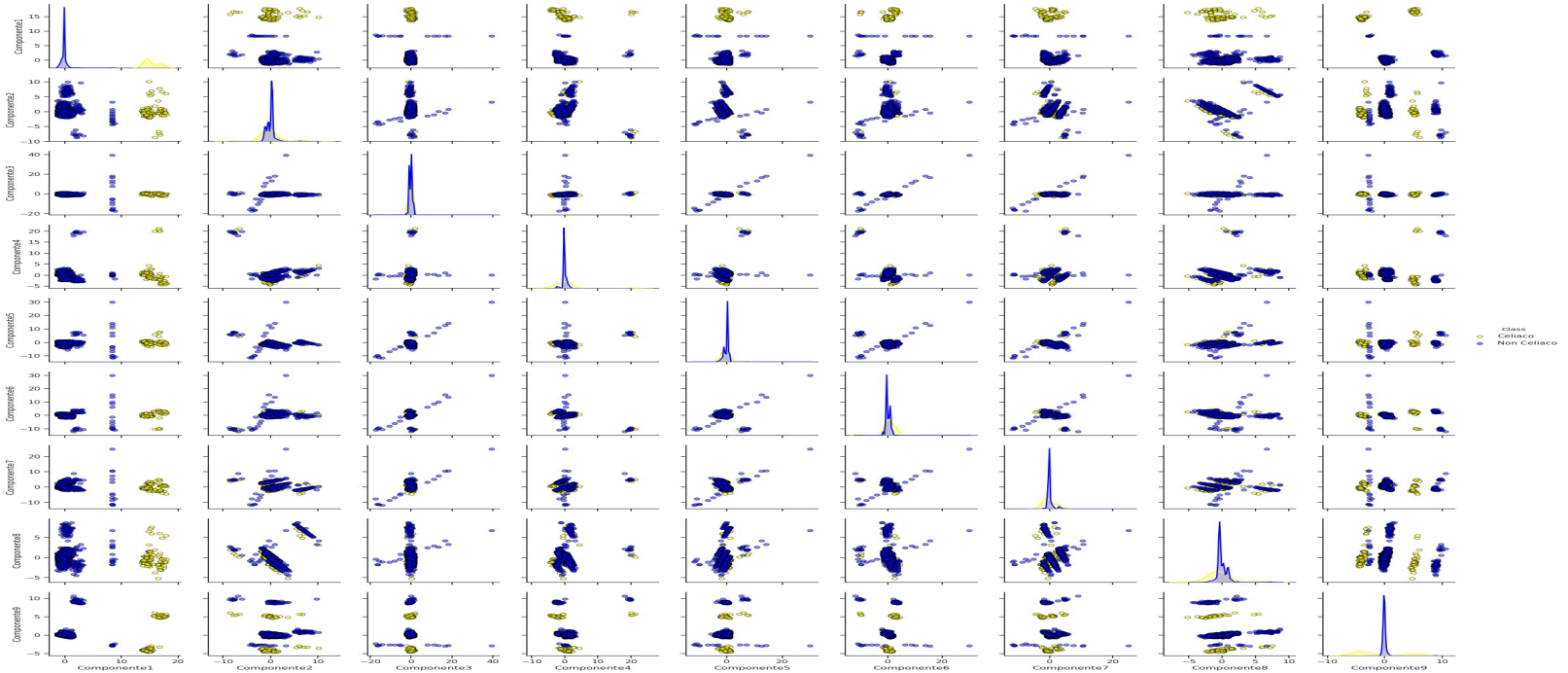
Scatterplot 2D tridimensionale applicato su kernelPCA



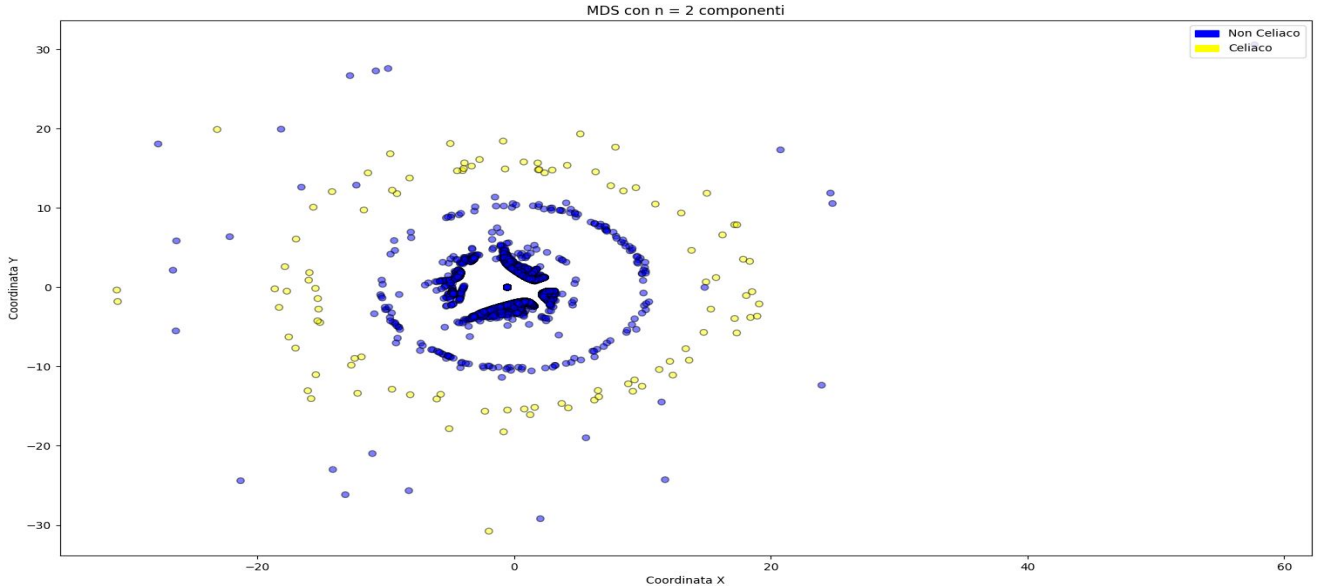
Scatterplot 3D interattivo applicato su kernelPCA



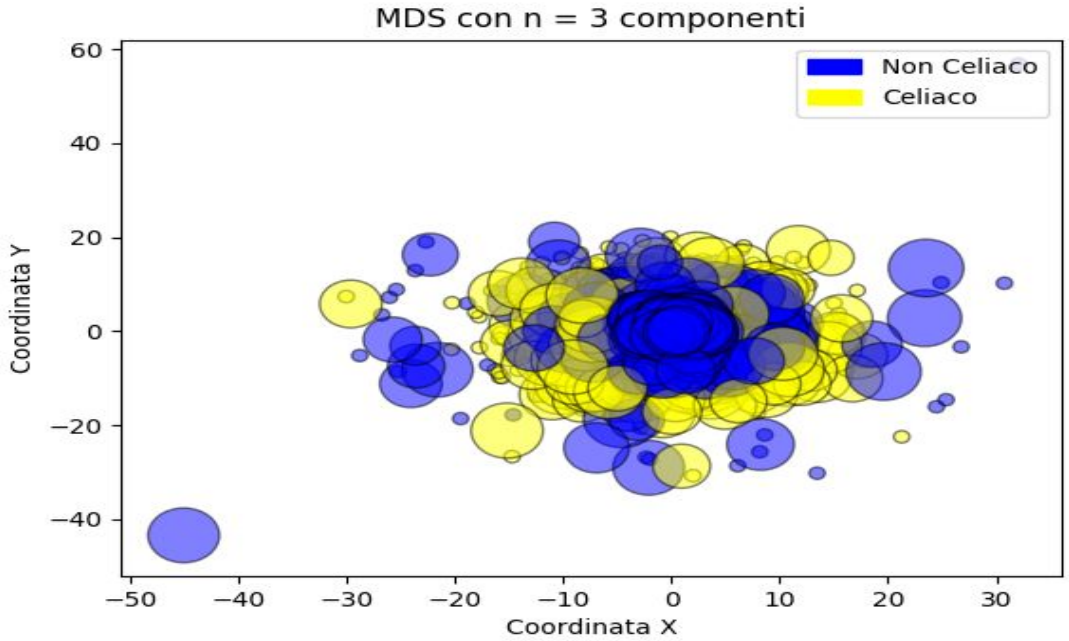
Plot di Draftman applicato su kernelPCA



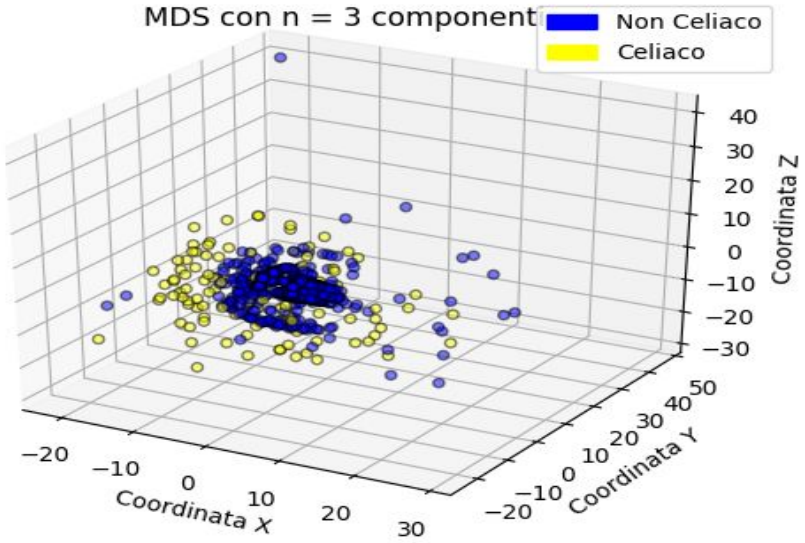
Scatterplot 2D bidimensionale applicato su MDS



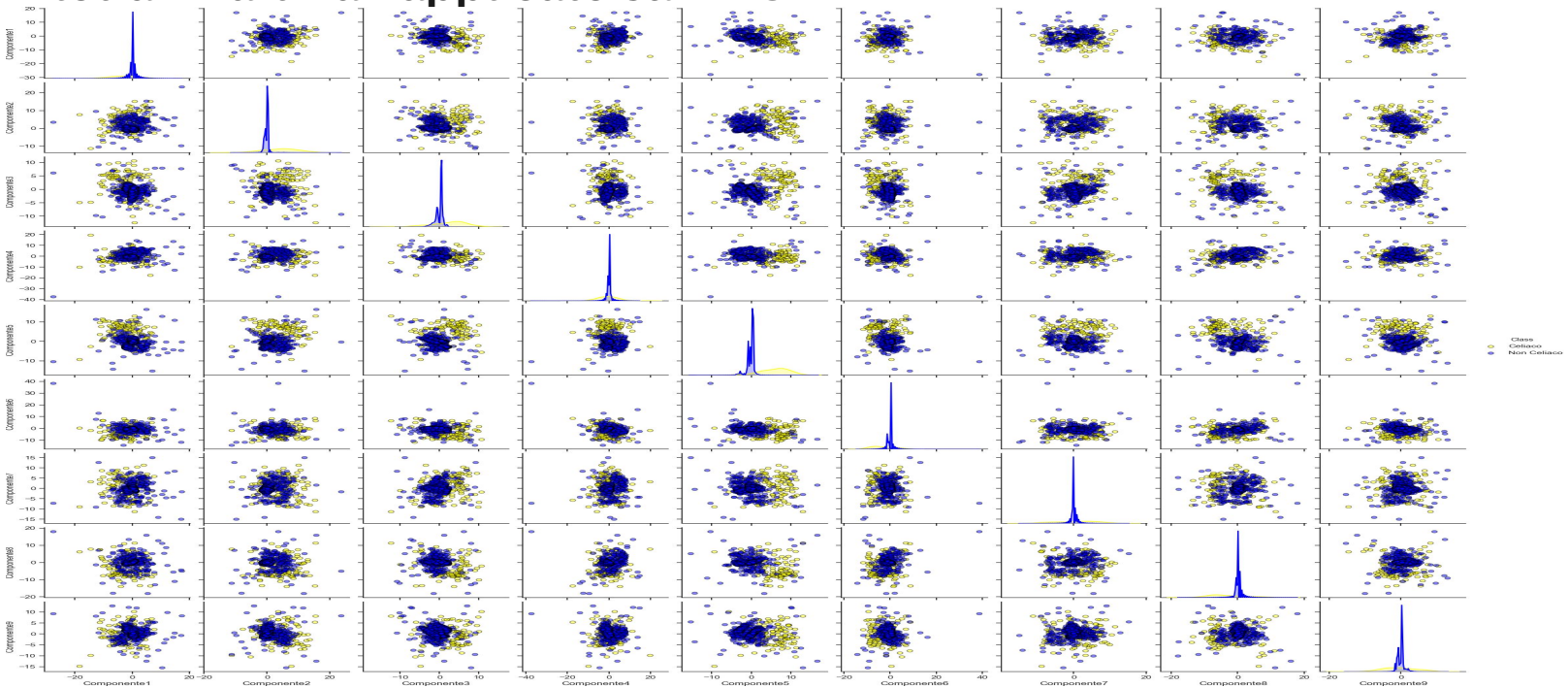
Scatterplot 2D tridimensionale applicato su MDS



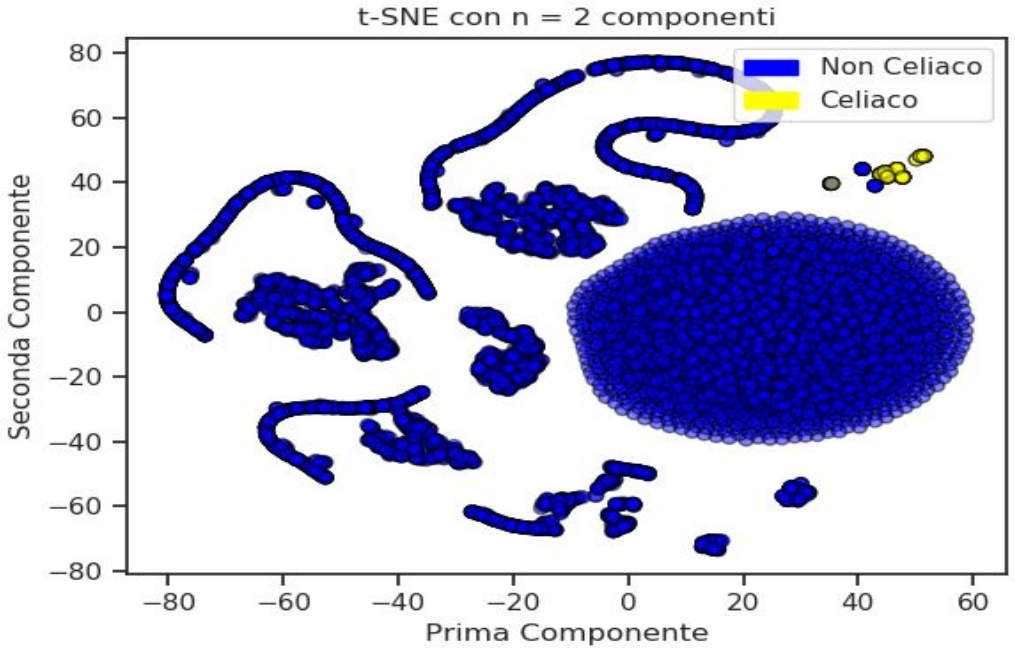
Scatterplot 3D interattivo applicato su MDS



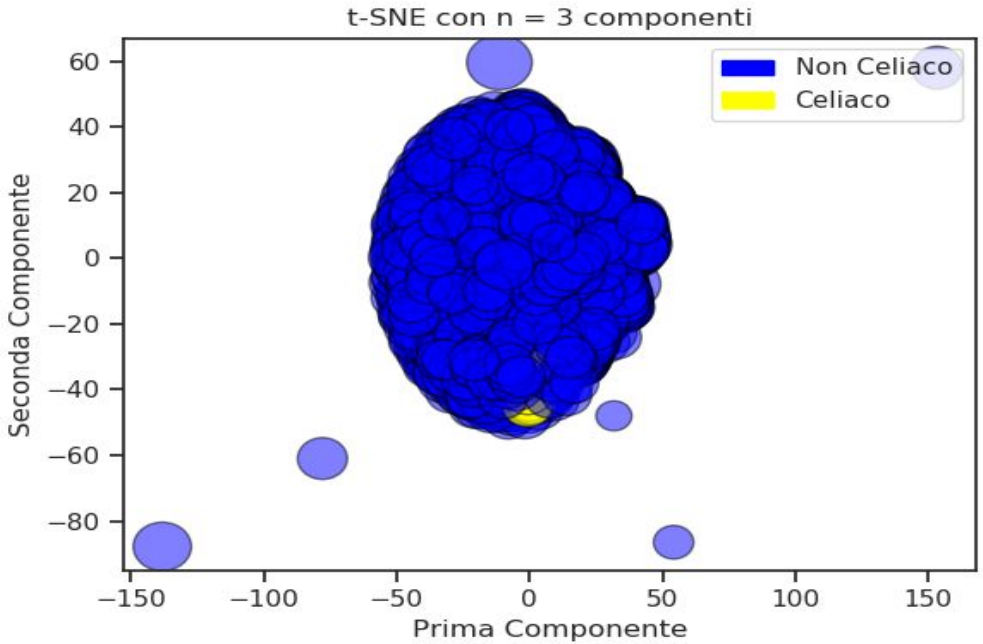
Plot di Draftman applicato su MDS



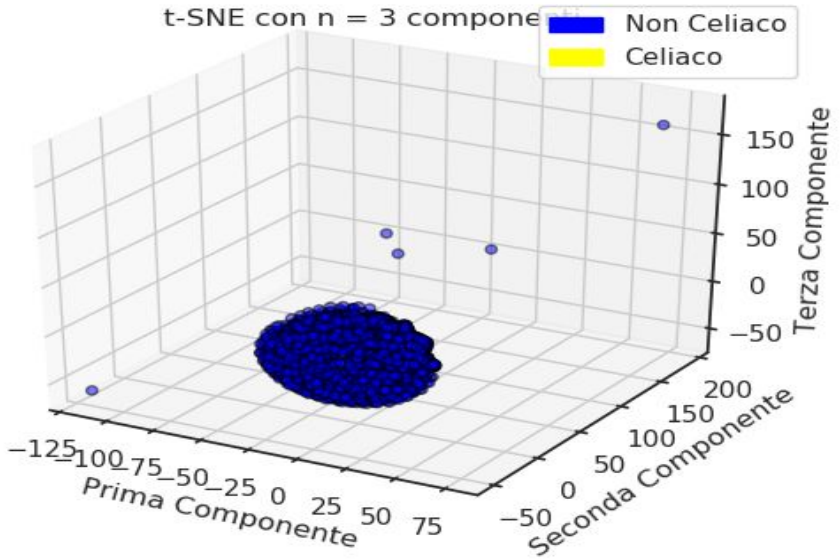
Scatterplot 2D bidimensionale applicato su t-SNE



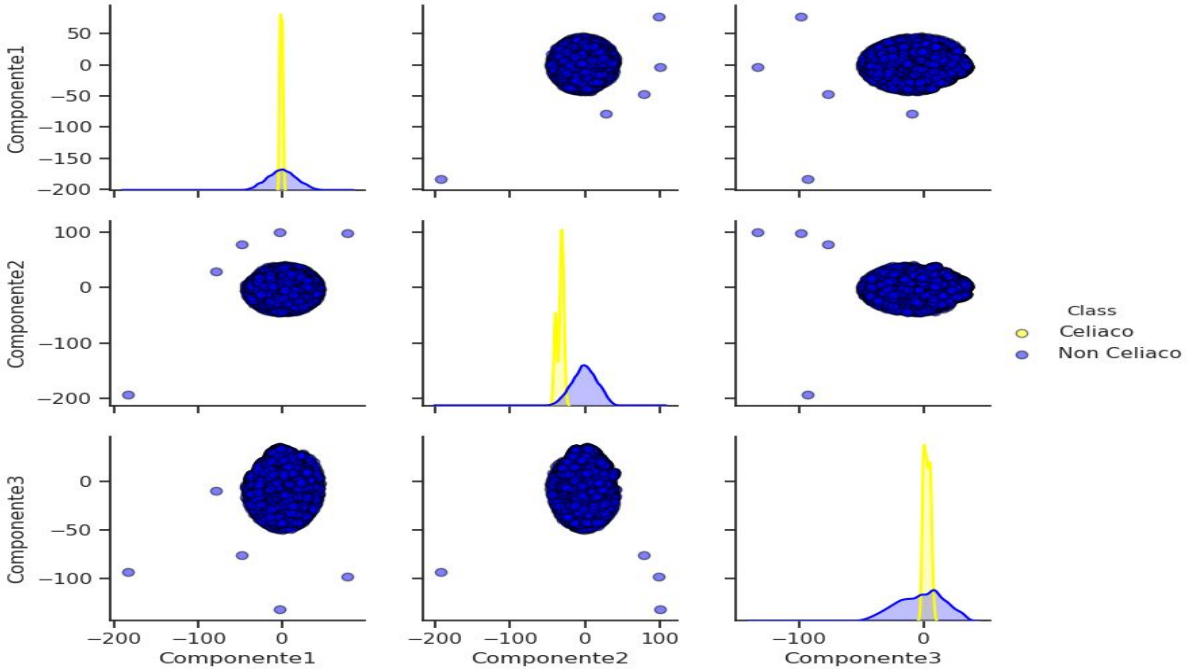
Scatterplot 2D tridimensionale applicato su t-SNE



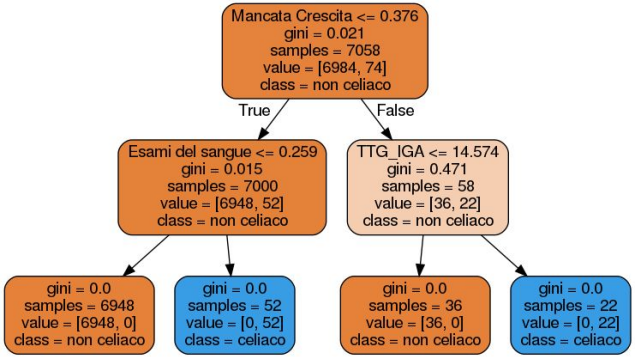
Scatterplot 3D interattivo applicato su t-SNE



Plot di Draftman applicato su t-SNE

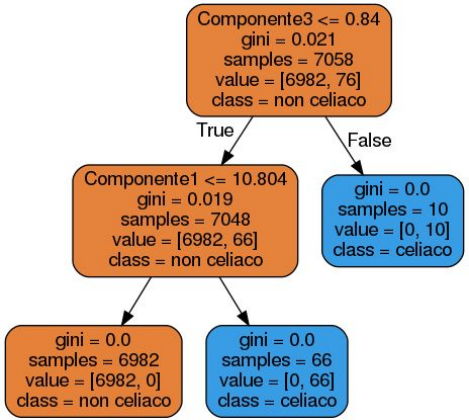


Alberi decisionali: VCDD non ridotto



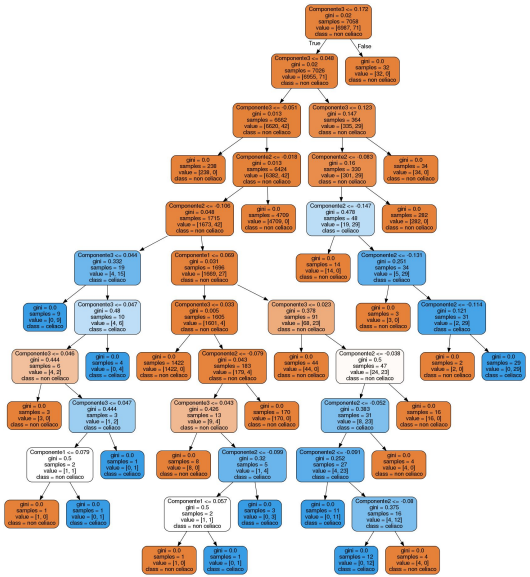
Accuratezza: 100%

Alberi decisionali: VCDD ridotto (PCA)



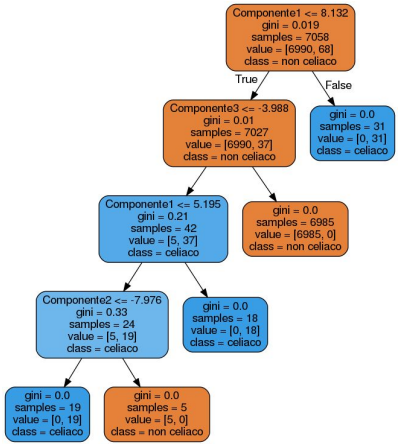
Accuratezza: 100%

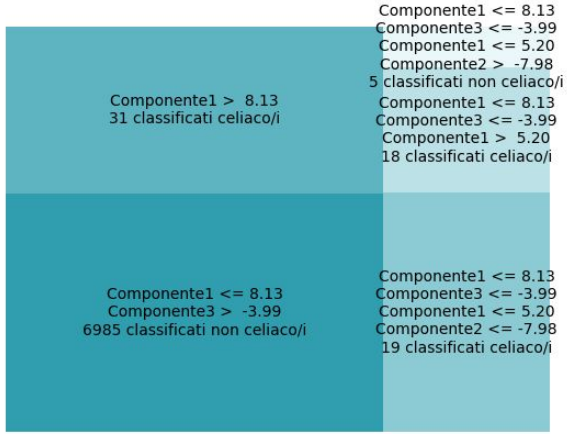
Alberi decisionali: VCDD ridotto (kernel PCA)



Accuratezza: 99%

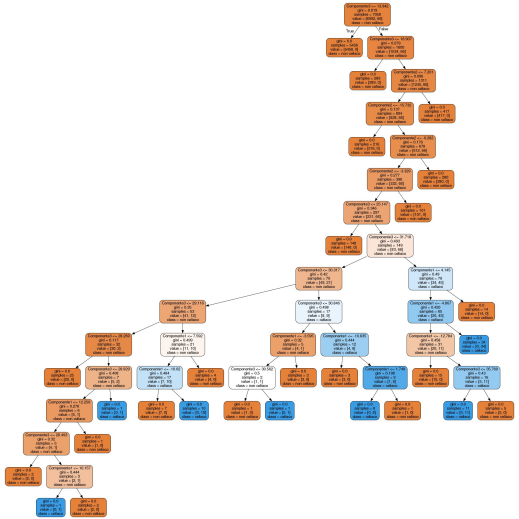
Alberi decisionali: VCDD ridotto (MDS)





Accuratezza: 99%

Alberi decisionali: VCDD ridotto (t-SNE)



Accuratezza: 99%

Conclusioni

- La scatterplot 2D il più delle volte è più che sufficiente
- La tecnica più efficace su questo dataset risulta essere la PCA
- Dall'analisi delle componenti della PCA risulta che le caratteristiche più influenti sono: esami del sangue, TTG IGA e POCT per la prima componente, madre celiaca ed IGA totali per la seconda

	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7	PCA8	PCA9	PCA10	PCA11
Anemia	0.02	-0.38	0.31	0.23	-0.68	0.33	-0.10	0.35	0.04	-0.05	-0.02
Osteopenia	-0.02	-0.12	-0.69	0.26	0.15	0.62	0.16	0.07	-0.10	-0.07	-0.02
Diarrea Cronica	0.11	-0.13	-0.20	-0.40	-0.21	-0.24	0.75	0.30	0.11	0.01	-0.01
Mancata Crescita	0.24	0.07	-0.06	-0.23	-0.03	0.27	-0.14	-0.18	0.87	-0.01	0.05
Disturbi Genetici	0.09	-0.27	0.01	0.73	0.24	-0.40	0.19	0.06	0.36	-0.00	0.02
Madre Celiaca	0.06	0.65	-0.08	0.13	0.03	-0.04	-0.15	0.72	0.09	-0.00	0.00
POCT	0.54	-0.05	-0.01	0.03	0.01	0.05	-0.04	0.01	-0.17	0.73	0.37
IGA totali	-0.01	0.56	0.03	0.35	-0.44	0.06	0.37	-0.47	-0.02	0.04	0.02
TTG IGG	-0.00	0.07	0.62	0.01	0.46	0.47	0.42	0.10	-0.00	-0.00	-0.00
TTG_IGA	0.54	0.02	0.03	0.00	0.01	-0.03	-0.01	-0.04	-0.18	-0.67	0.47
Esami del sangue	0.58	0.01	0.02	0.02	0.02	-0.01	-0.03	-0.05	-0.13	-0.04	-0.80