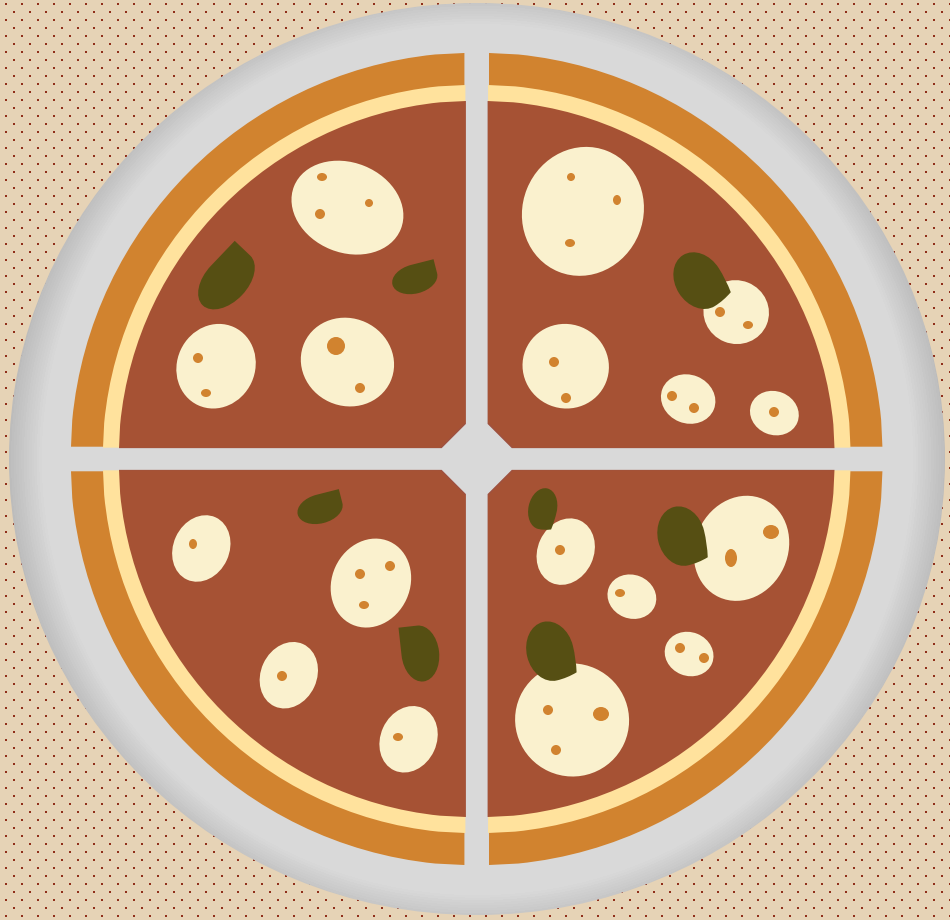
Data
Analysis

Quanto costa davvero una Pizza negli Stati Uniti?

Università di Catania

00 Indice

01 Indice

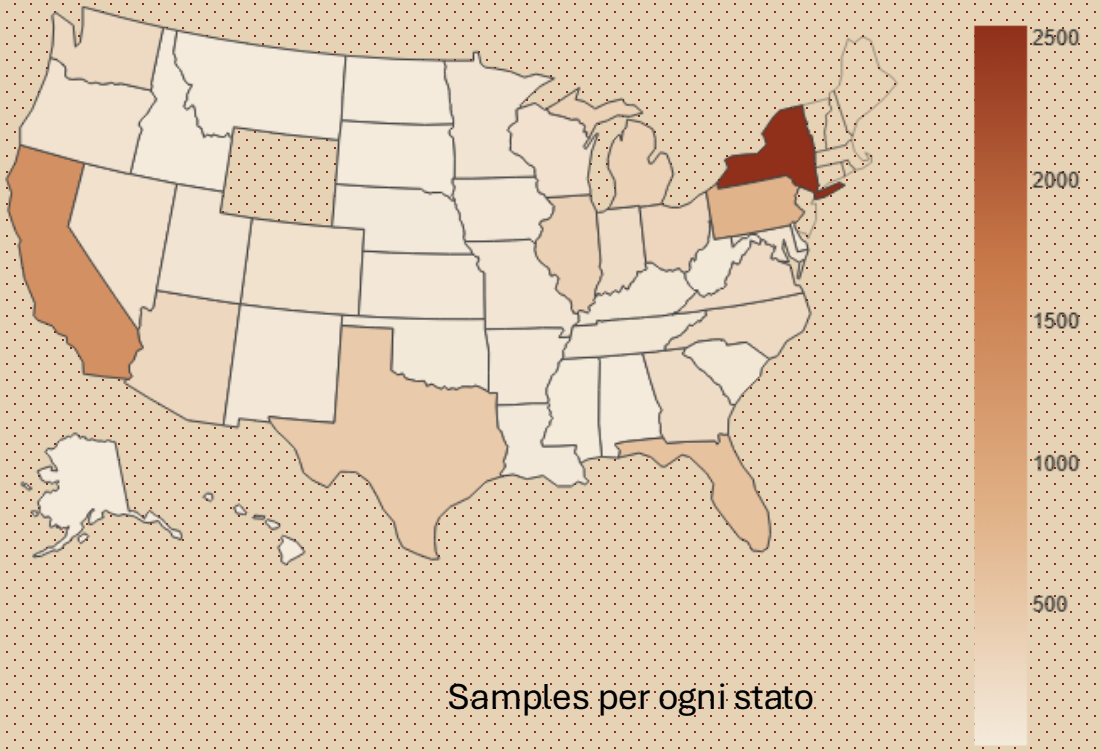


01 Il dataset

Dataset su item di menu ristoranti di pizza negli Stati Uniti

Ogni campione rappresenta un item di menu

Distribuzione geografica eterogenea delle osservazioni.



Samples per ogni stato

01 Un campione del dataset

Ogni riga del dataset rappresenta un singolo item di menu.

Le informazioni descrivono il piatto, il ristorante e il contesto geografico.

Ristorante	Shotgun Dan’s Pizza
Città	Sherwood
Stato	AR
Paese	Stati Uniti
Nome item	Cheese Pizza
Categoria	Pizza
Prezzo	$25.00
Valuta	USD

Esempio di campione

01 Preprocessing

Rimozione di una quota limitata di osservazioni (~ 2%) dovuta a outliers o non utilizzabili.

Uniformazione delle informazioni di prezzo per rendere confrontabili gli item di menu.

Associazione degli item al contesto geografico per consentire analisi territoriali dei prezzi.

01 Pulizia semantica delle categorie

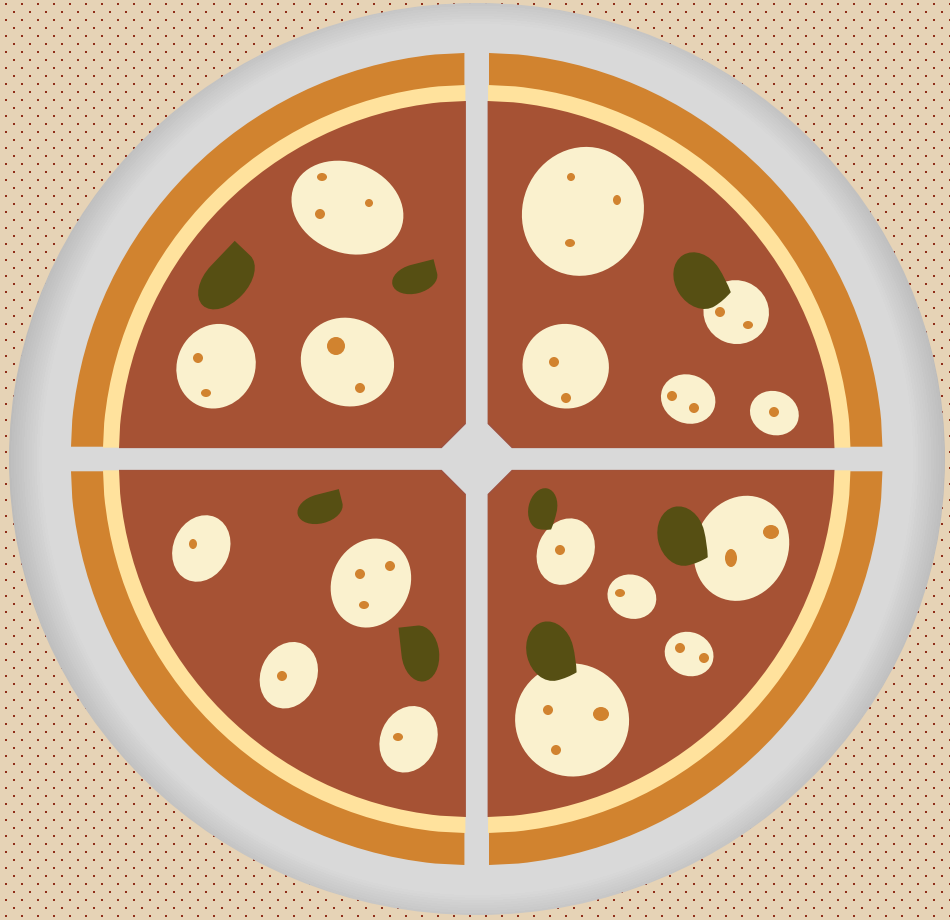
I nomi degli item di menu presentano una forte variabilità testuale, anche quando descrivono lo stesso prodotto.

È stata applicata una procedura di clustering, con un transformer, per ricostruire delle etichette più «standard».

Veggie Pesto Pizza Wrap	Pesto Pizza
Tony's Tuna Fish Pizza	Salmon Pizza
Steak, Steak, Steak Pizza!	Pizza Steak
Specialty Gluten Free Pizza	Gluten Free Pizza

Prima e dopo il clustering

03 Indice

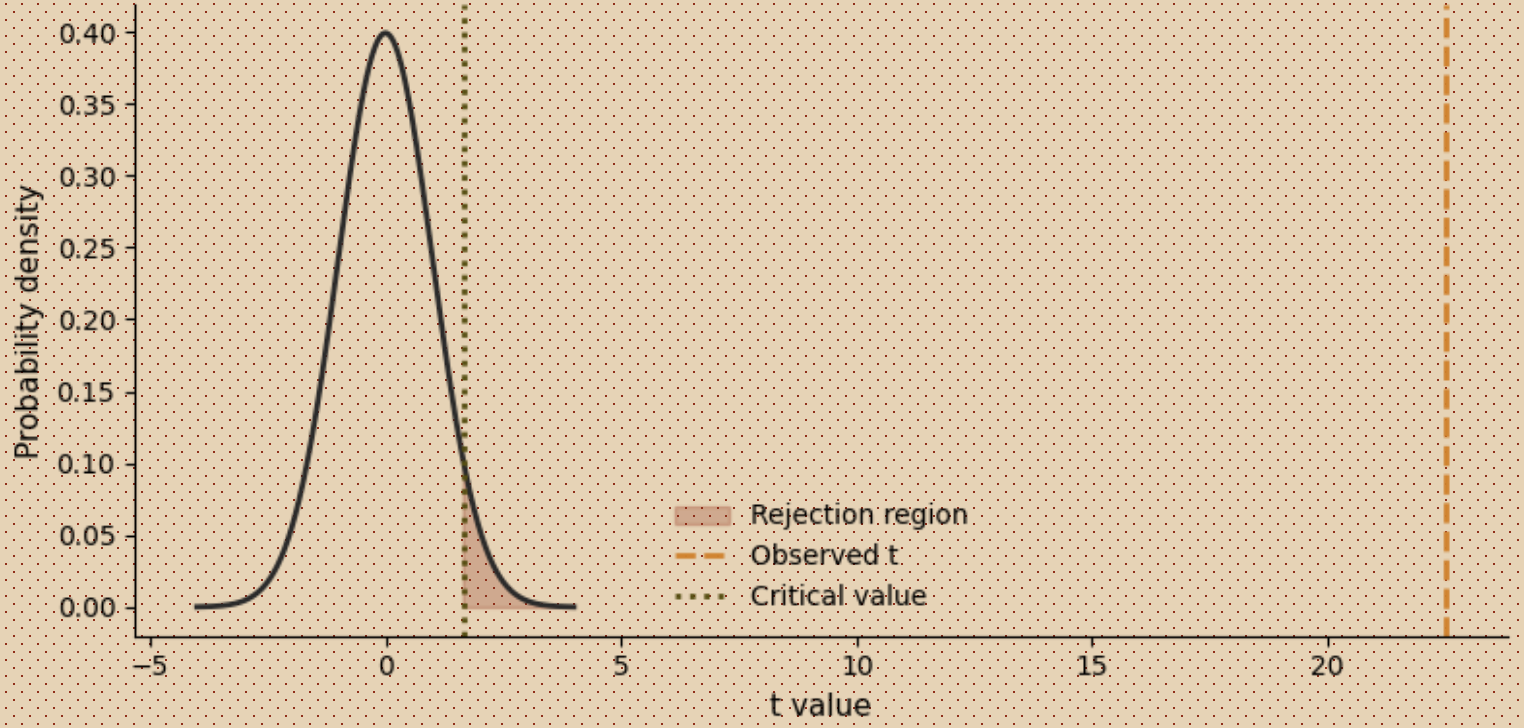


02 Prezzi in città costiere vs interne

H_0 I prezzi medi sono uguali sia nelle città costiere che interne.

H_a Le città costiere hanno prezzi, in media, più alti.

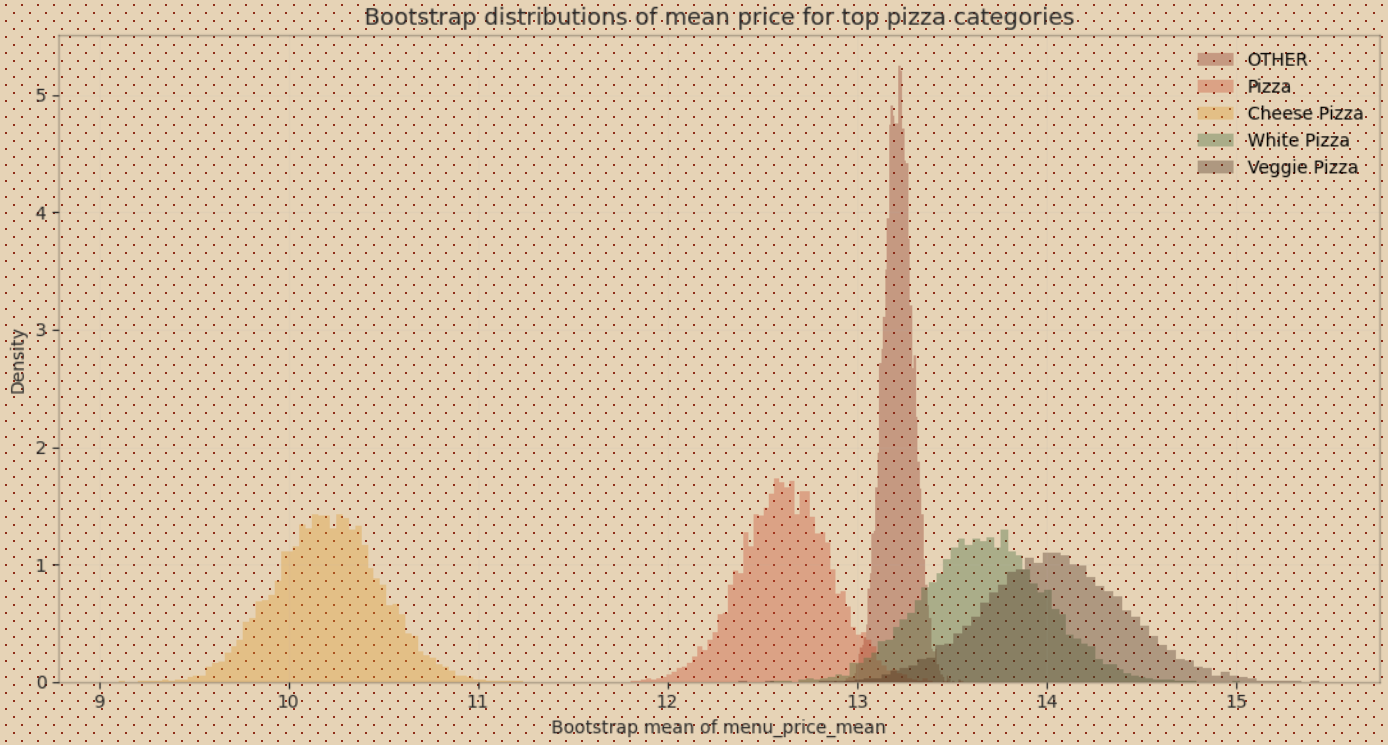
→ Evidenza statistica forte a favore di H_a $(p - value \ll 0.05)$.



02 Media uguale tra categorie

H_0 Tutte le medie sono uguali tra categorie.

H_a Almeno una media è diversa.



ANOVA

- $F = 22.48$
- $p \ll 0.05 \rightarrow$ rifiuto H_0

Bootstrap

conferma la separazione delle medie tra categorie

02 Regressione lineare sui prezzi

- **Quantile Regression**

$$\min_{\beta} \sum_{i} \rho_{0.5}(y_i - x_i^T \beta)$$

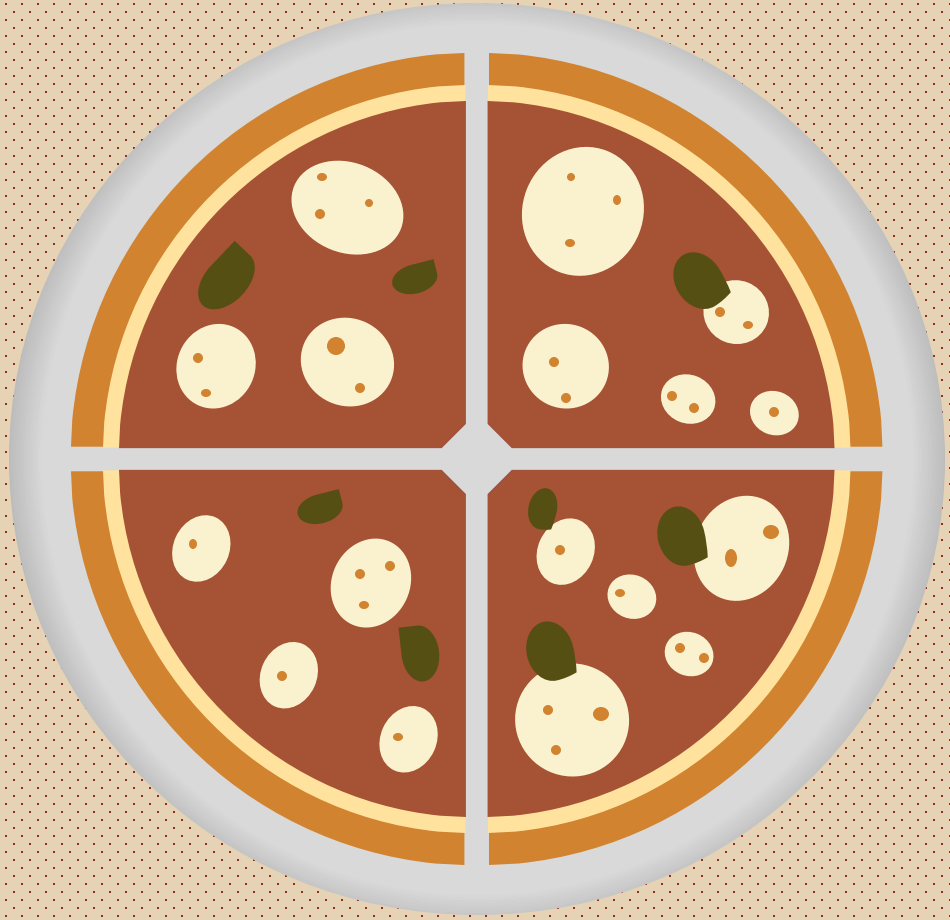
- $\beta = 0.5$

Ogni unità aggiuntiva di complessità del menu aumenta il prezzo tipico di circa 0.50 $

- $R^2 \approx 0.96$

Il modello spiega quasi tutta la variabilità osservata.

02 Indice

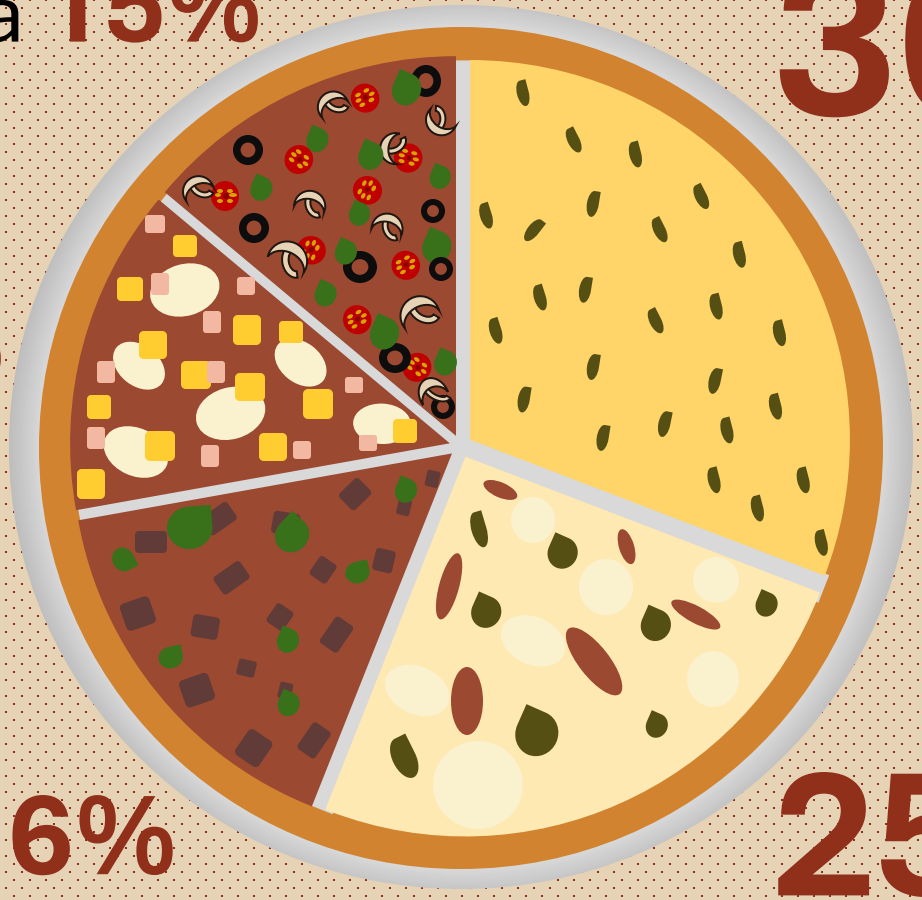


03 Quali sono le pizze più comprate?

Veggie pizza **15%**

30% Cheese Pizza

Hawaiian Pizza **14%**



BBQ Chicken Pizza **16%**

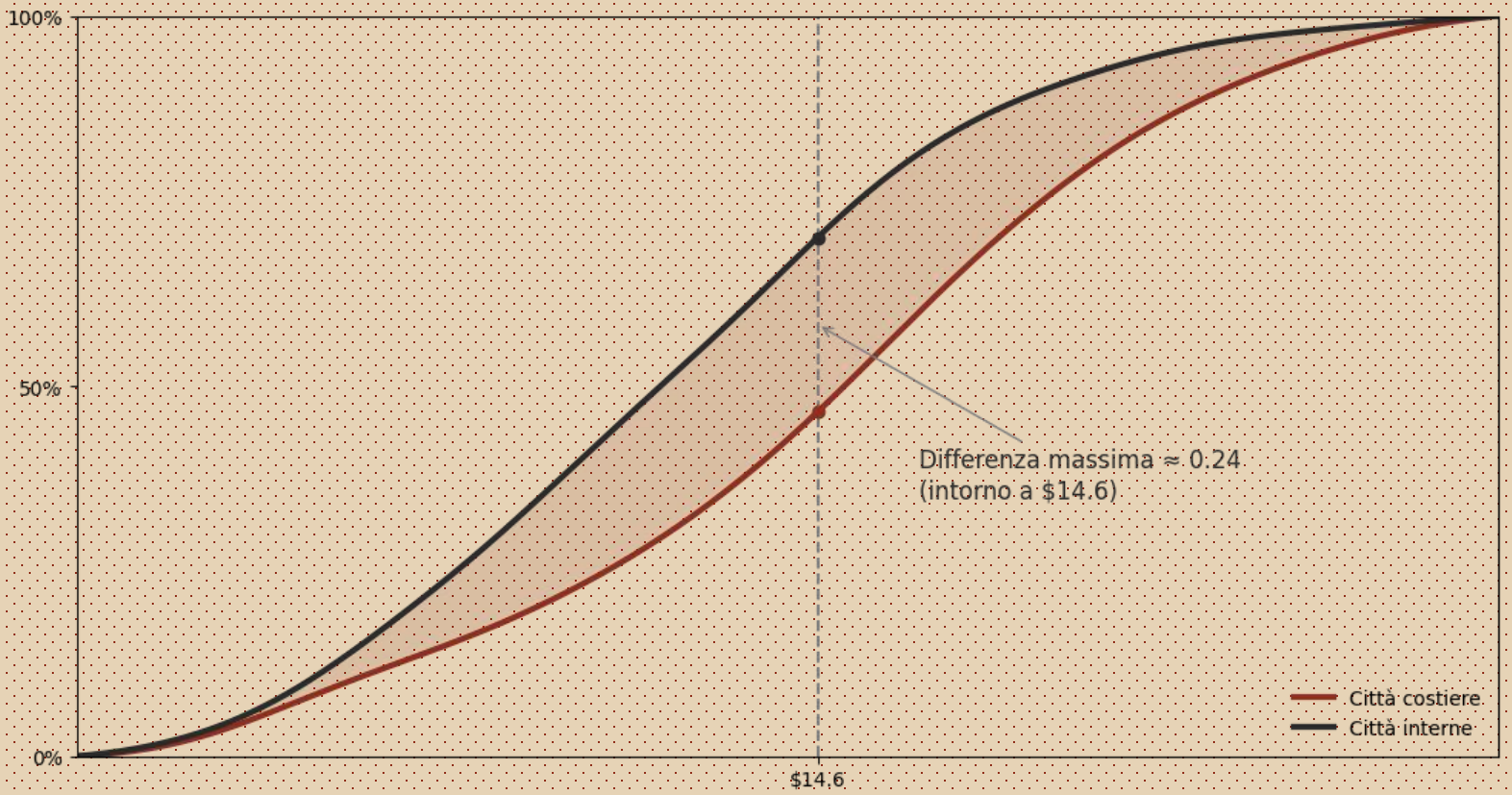
25% White pizza

03 Città Costiere VS Interne

Identificazione delle città costiere tramite formula di Haversine a 50 km dalla costa USA.

La ECDF mostra una separazione consistente, quindi una differenza sistematica.

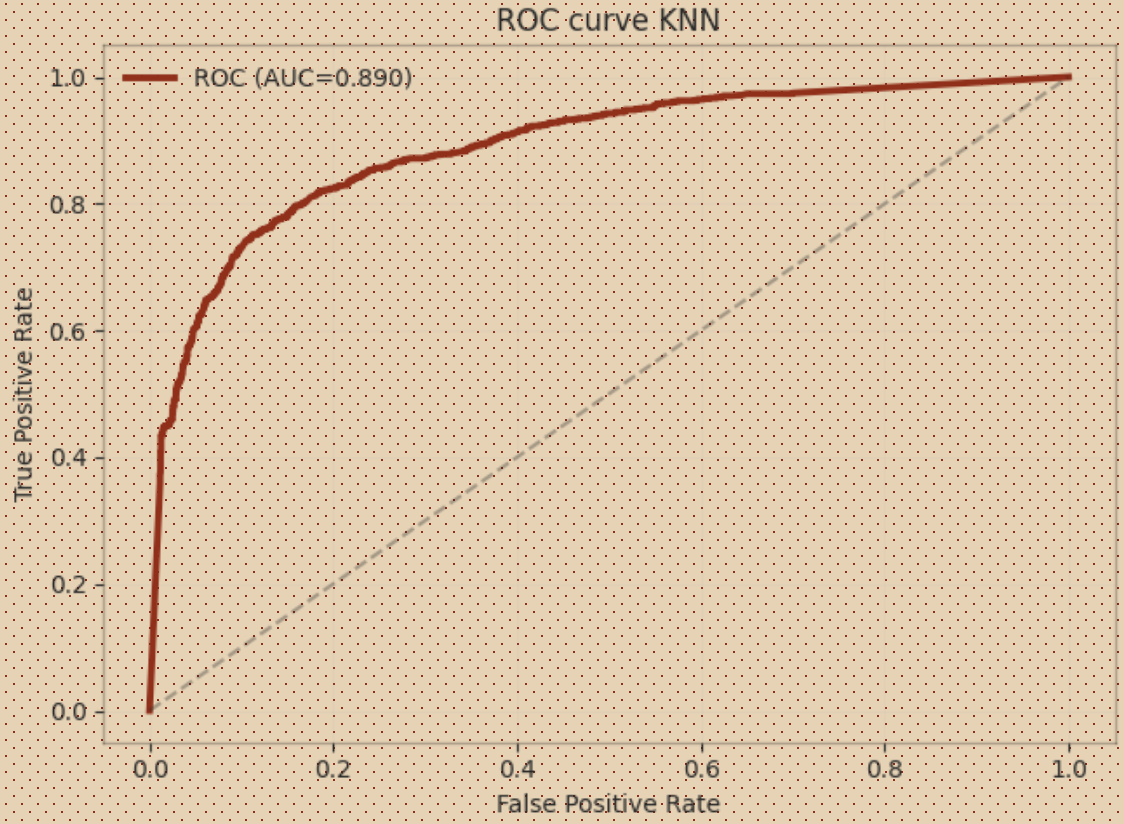
ECDF su minimo dei prezzi di menù

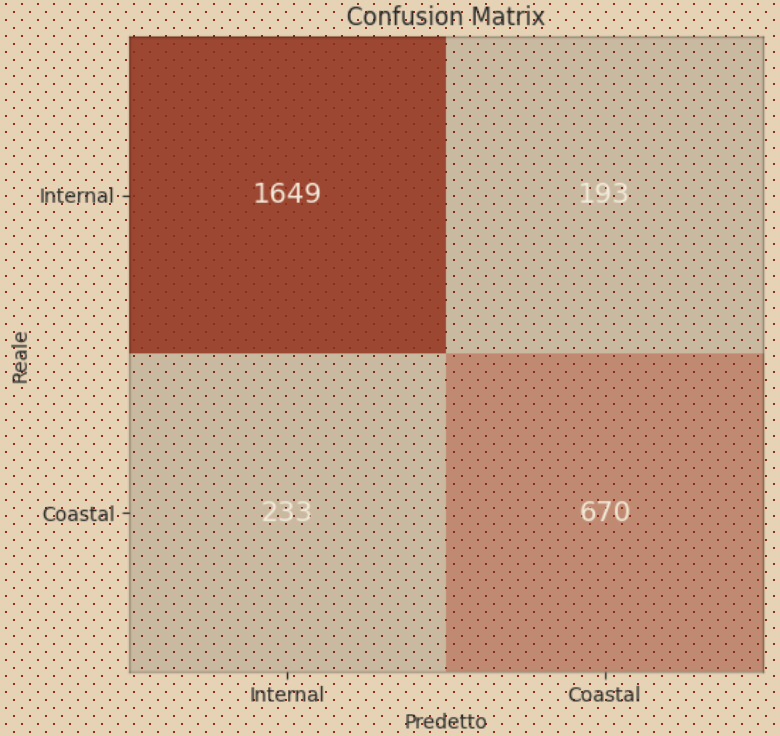


03 KNN: geografia e prezzo

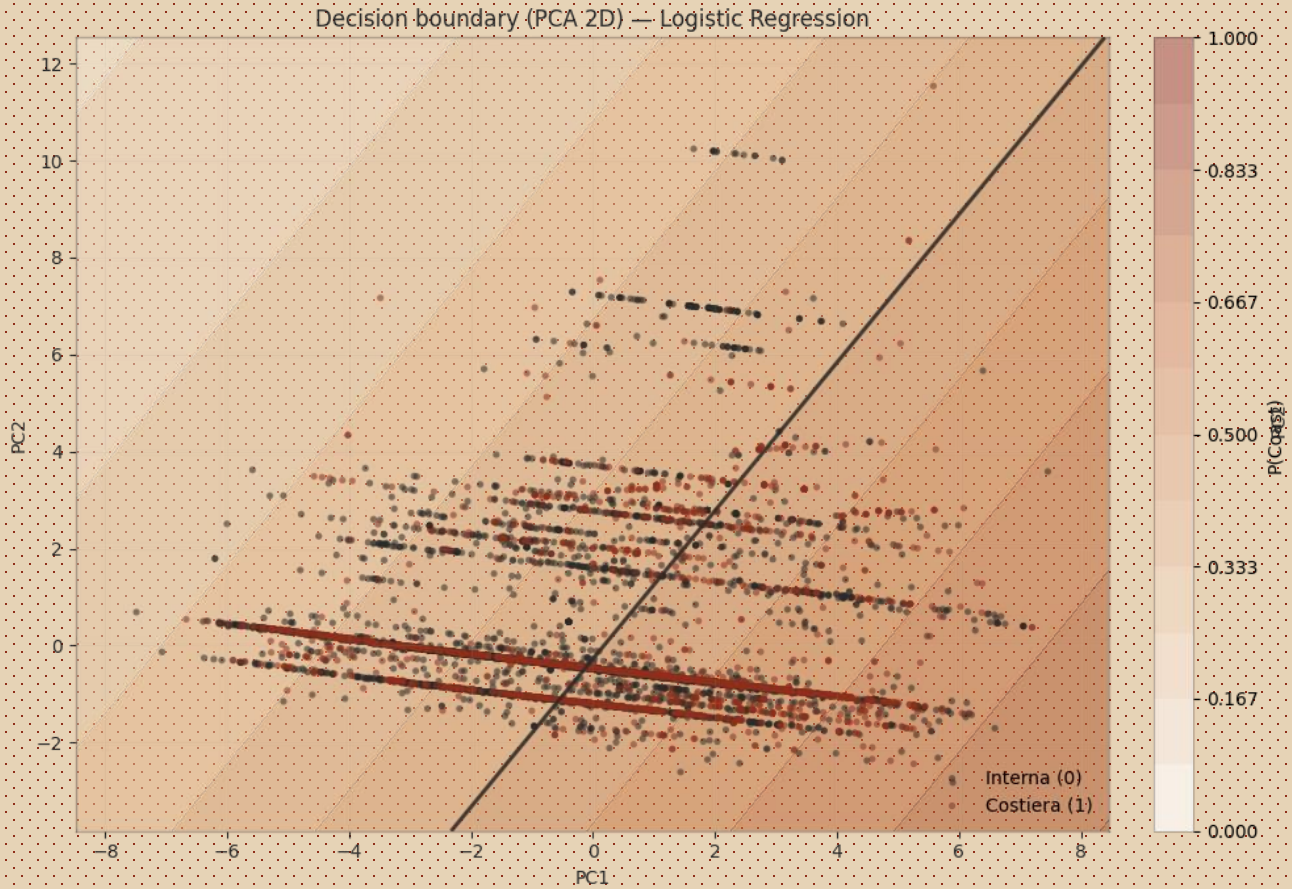
Il modello cattura efficacemente i pattern locali nei dati (AUC = 0.894).

	Città interne	Città costiere
Precision	0.88	0.78
Recall	0.90	0.74

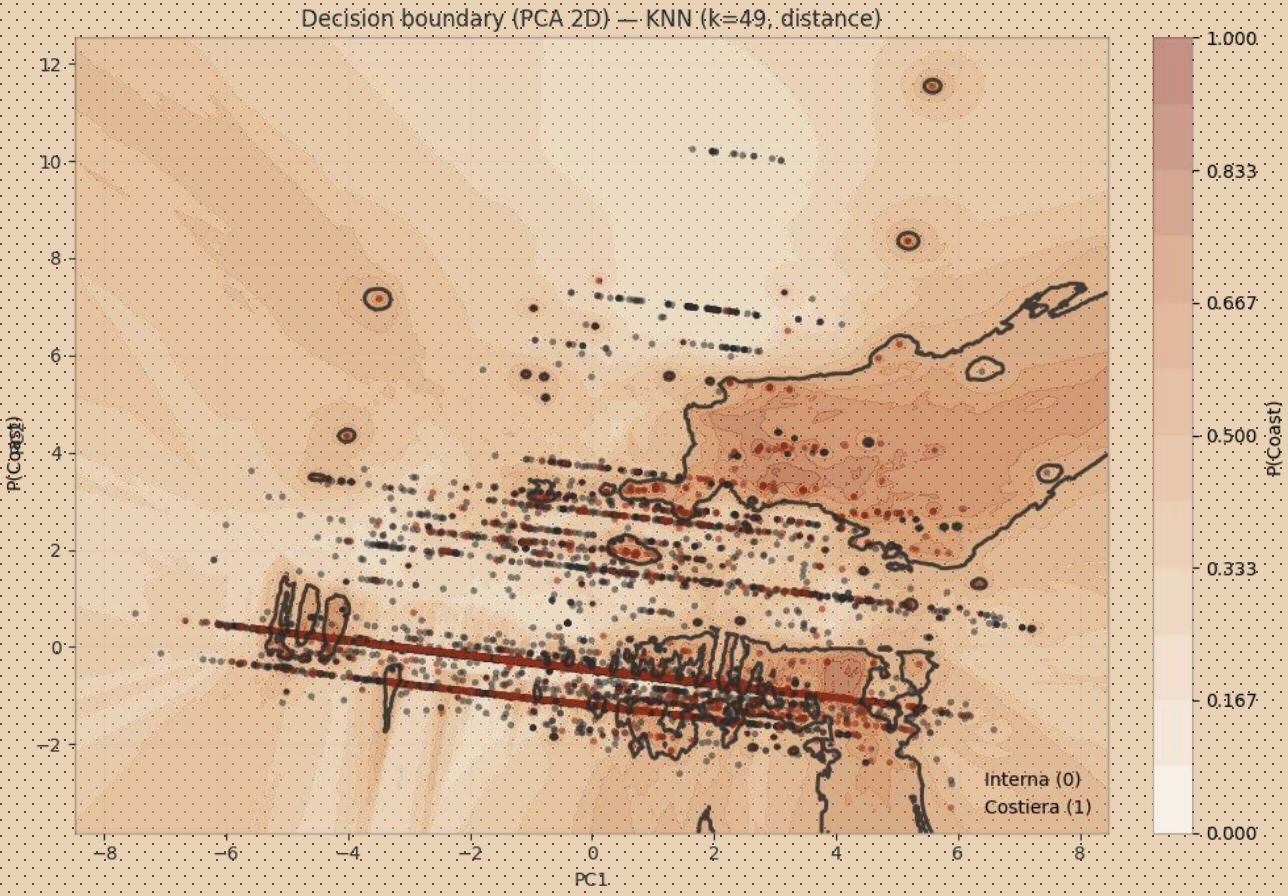




03 Decision Boundaries



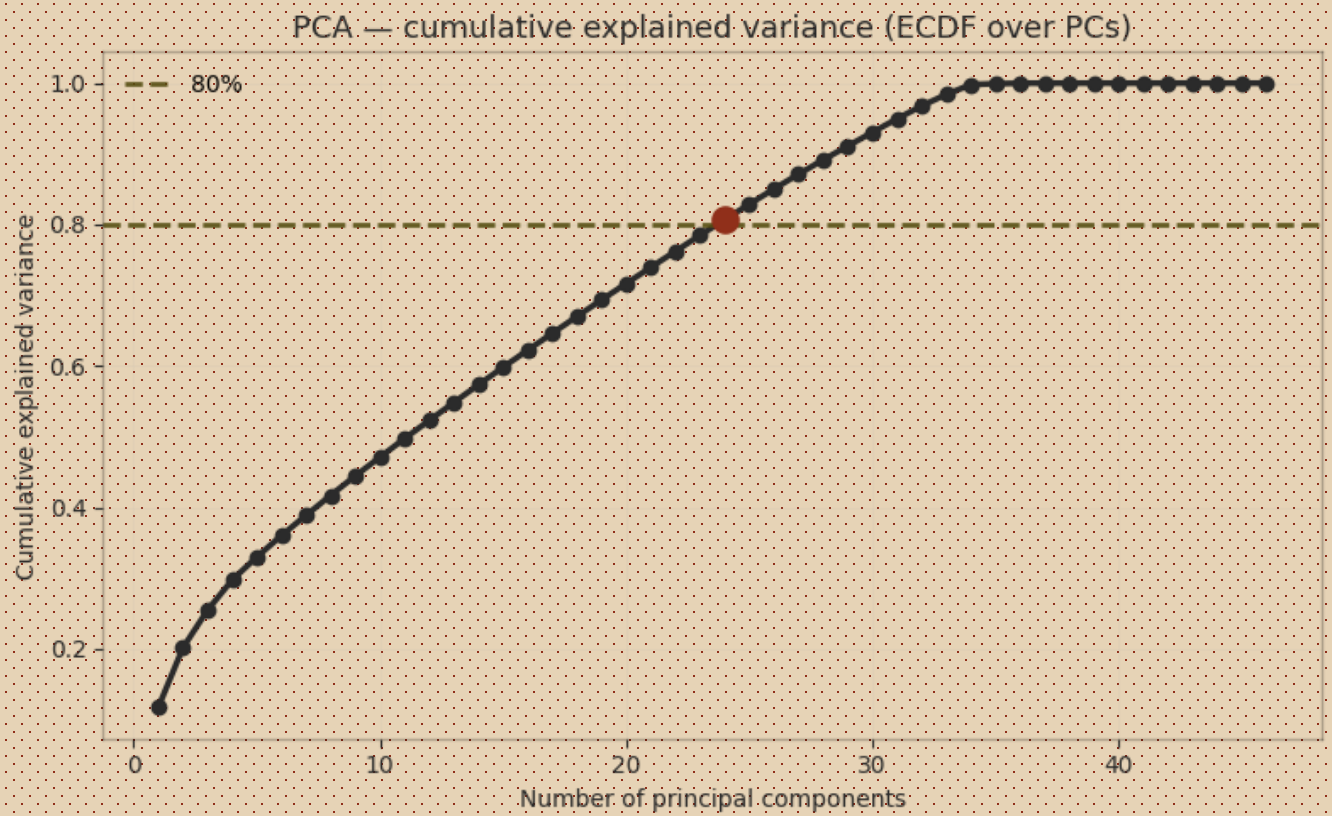
Regressore logistico

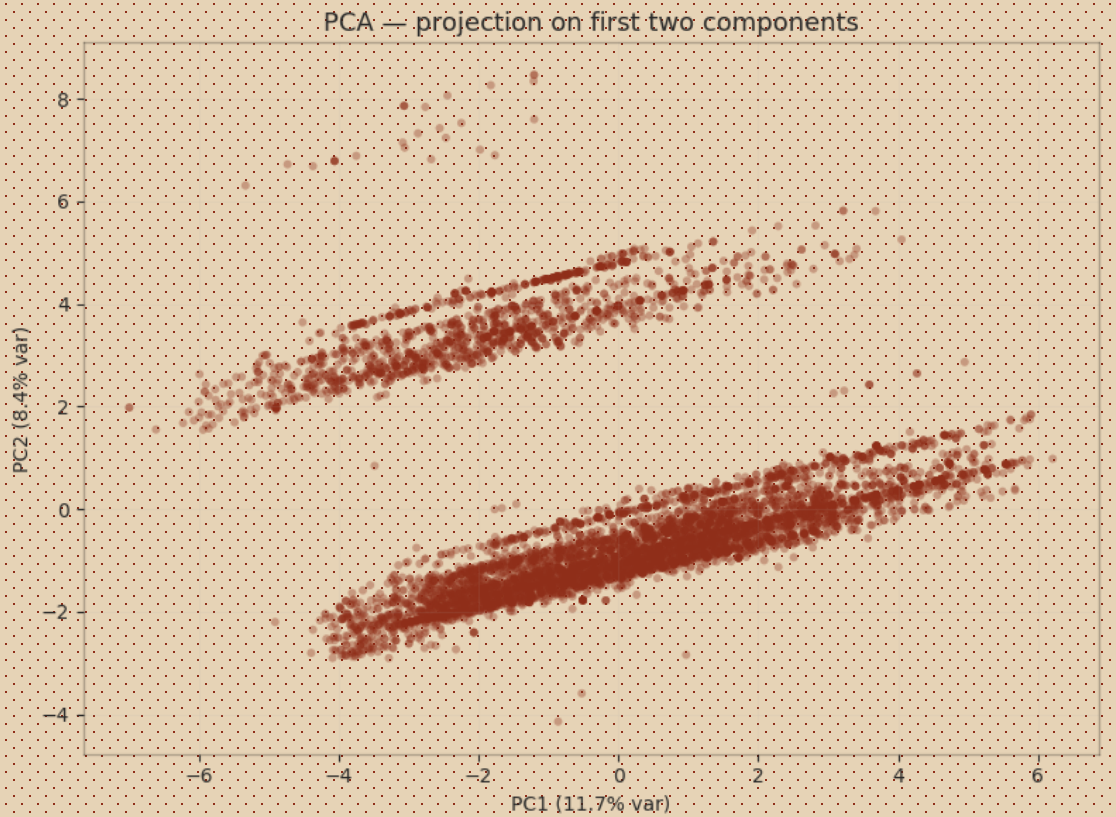


KNN

04 Indice

04 PCA – Riduzione della dimensionalità





PC1 – 11.7 %

Livello dei prezzi dell’item.

PC2 – 8.4 %

Ampiezza del range di prezzo del ristorante.

04 Clustering – Gaussian Mixture Model

BIC

Scelta del numero di Cluster.

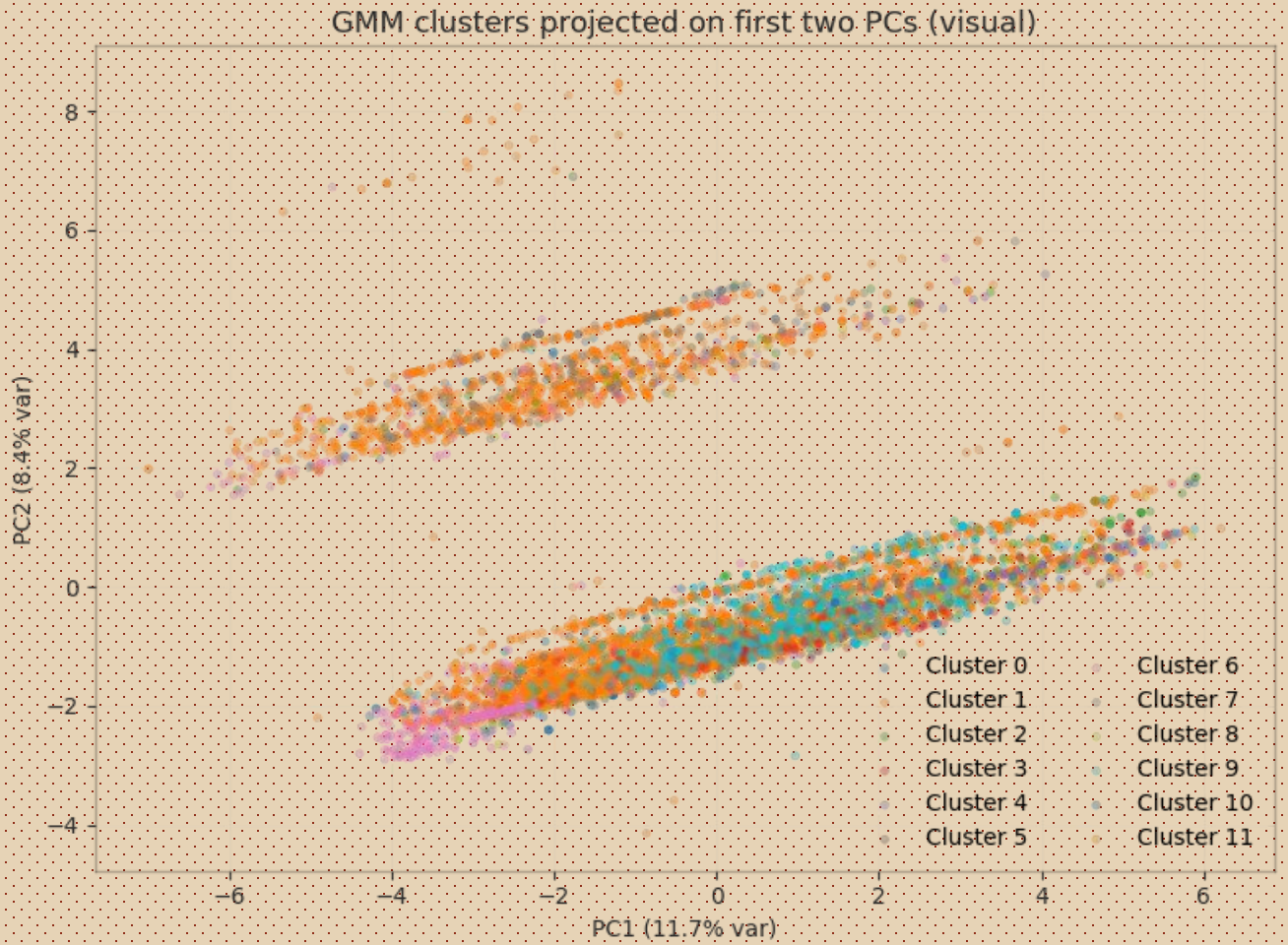
$B = -2 \log L + p \log n$

Interpretazione

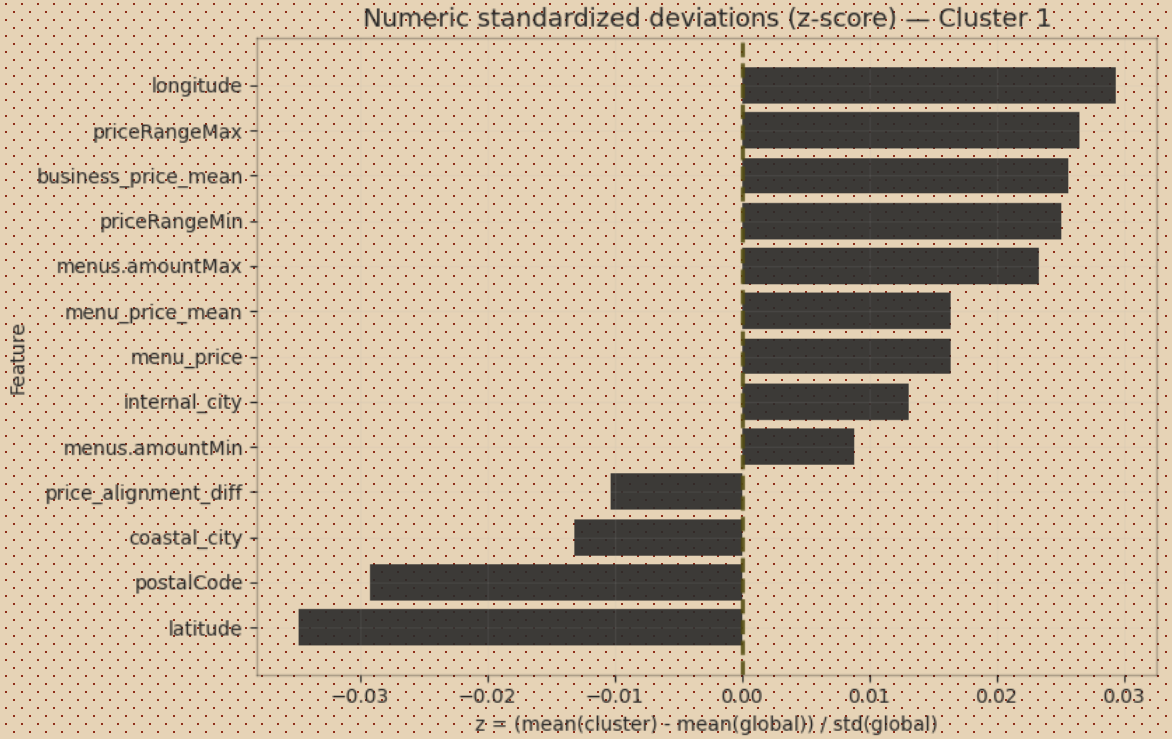
I cluster sono molto specializzati tra loro, in quanto le dimensioni hanno una variabilità elevata.

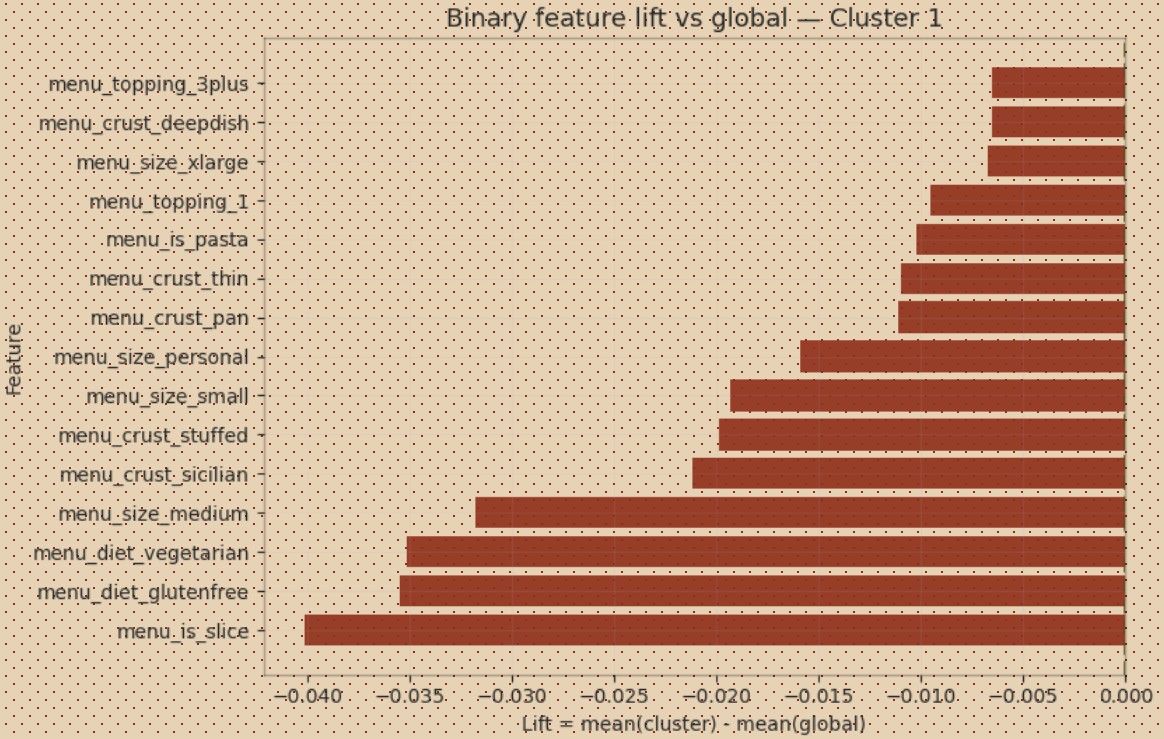
Top 5 cluster

1. 6 817
2. 734
3. 407
4. 398
5. 207



04 Clustering –Cluster Maggioritario





Caratteristiche principali

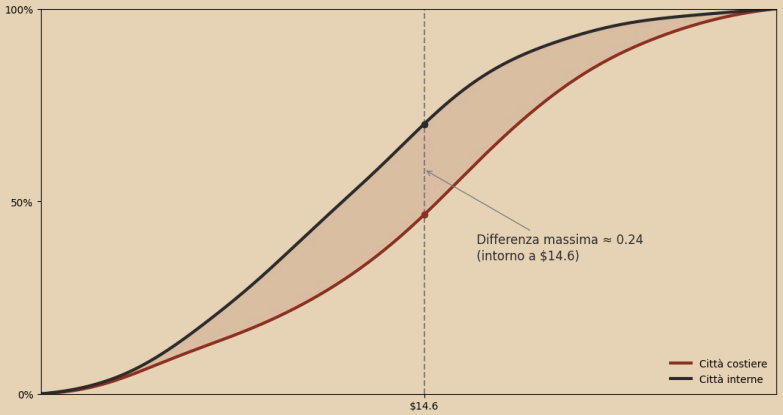
Questo cluster ha un’assenza di forti specializzazioni nel menu. I prezzi e range sono vicini alla media globale.

Ruolo nel clustering

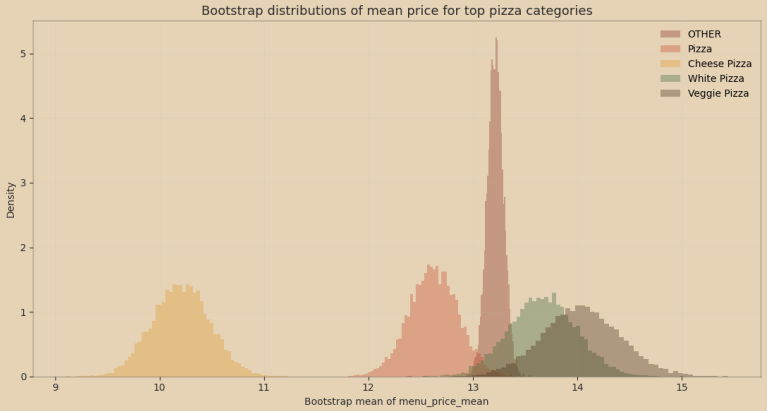
Cluster di riferimento utilizzato per confrontare e interpretare i cluster più piccoli e specializzati.

00 Indice

05 Riepilogo - Quanto costa davvero una pizza negli USA?

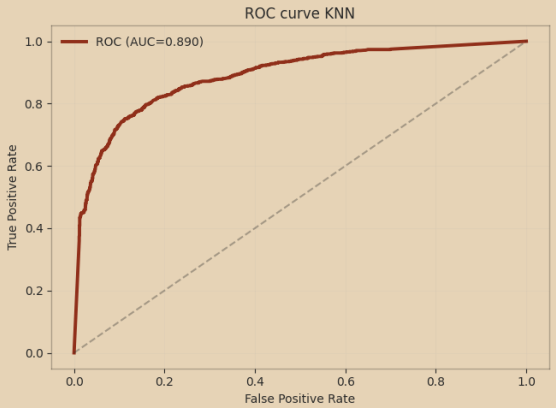


Città costiere vs interne

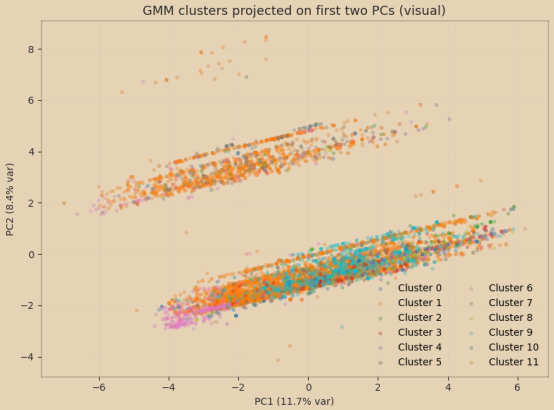


Media dei prezzi delle categorie diversa

Associazione forte tra prezzo e menù



I modelli catturano i pattern geografici



Esiste un cluster maggioritario

05 Un esempio reale

Fake

Item di menù generato a partire dai dati.

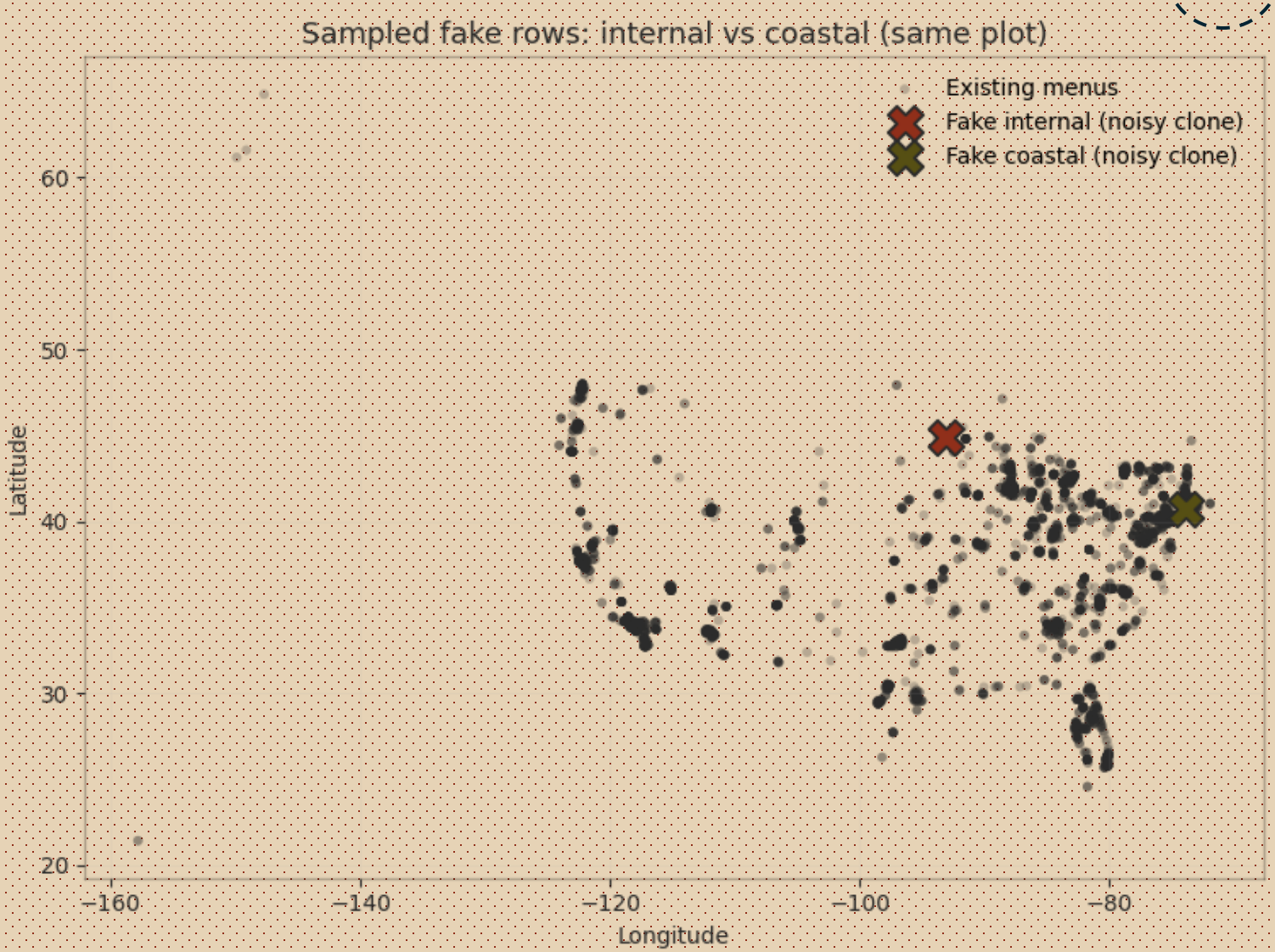
15%

Livello di incertezza.

KNN

Il KNN ha classificato:

- Costiera: 87%
- Interna: 97%



THANKS!

Emanuele Galiano

Corso di Fondamenti di
Analisi dei dati

Università
di Catania