REPORT DATA ANALYTICS

[st] Carmela Mandato, Giulia Di Biase, Luca Sacco, Simone Di Mario

***Abstract*—L'analisi dei costi di produzione dei nuovi edifici residenziali in Europa rappresenta una sfida rilevante per la pianificazione strategica del settore edilizio. Questo studio si propone di costruire, testare e automatizzare una pipeline completa di Data Analytics, utilizzando dati ufficiali provenienti da Eurostat. L'obiettivo principale è esplorare l'andamento dei costi e individuare tendenze significative attraverso tecniche di analisi dei dati, preparazione, modellazione e visualizzazione. Particolare attenzione è rivolta alla pulizia e trasformazione dei dati, alla selezione delle metriche chiave e alla valutazione dell'efficacia del flusso analitico in termini di interpretabilità e replicabilità. L'approccio adottato permette di ottenere una visione strutturata e automatizzata dei costi di costruzione a livello europeo, utile per supportare decisioni informate nel campo delle politiche abitative e della sostenibilità.**

Introduzione

In questo studio abbiamo sviluppato una pipeline completa di Data Analytics per analizzare i costi di produzione degli edifici residenziali in Europa. Il flusso operativo, realizzato in Python, si articola nelle seguenti fasi:

1. Preparazione dei dati e creazione delle feature

Lettura dei file CSV Eurostat, espansione degli aggregati geografici (EA19, EA20, EU27_2020), rimozione di outlier e costruzione di variabili derivate come rolling mean, pct_change e slope.

2. Clustering non supervisionato

Applicazione di KMeans e DBSCAN per raggruppare i paesi in base all'andamento dei costi. Valutazione mediante silhouette score e visualizzazione con PCA e mappa europea.

3. Modellazione supervisionata

Previsione del costo futuro (regressione) e classificazione della direzione della variazione. Implementati Decision Tree (con e senza pruning) e Random Forest, valutati tramite RMSE, R^2, accuracy e F1-score.

4. Output e automazione

Salvataggio dei modelli in formato csv, esportazione di risultati e metriche, generazione di grafici (.png) e pipeline dati in Python, e proposte di automazione con Microsoft Task Scheduler.

Presentazione Dettagliata dei Dati

In questa sezione vengono illustrate in maniera approfondita le caratteristiche del dataset utilizzato, evidenziando la struttura delle variabili, le statistiche descrittive e le visualizzazioni che ne facilitano la comprensione.

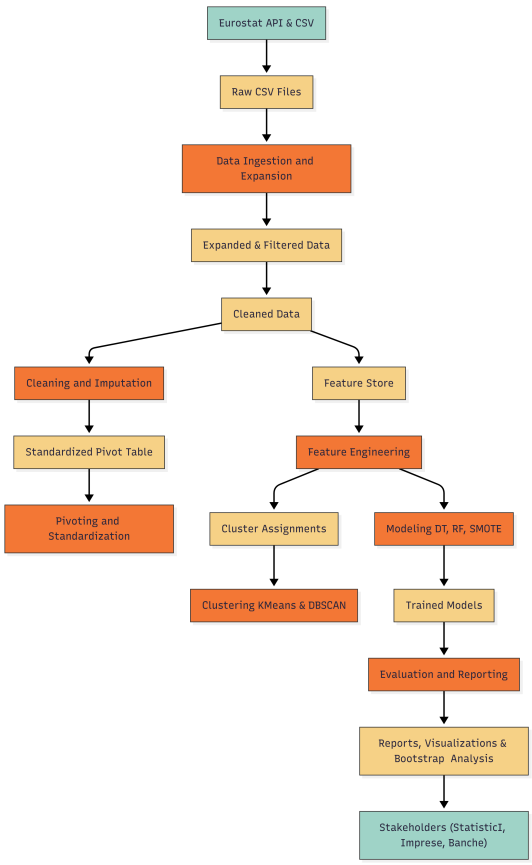

Fig. 1. Diagramma di flusso dei dati

Il dataset analizzato è stato estratto da Eurostat e contiene informazioni sui costi di produzione per la costruzione di edifici residenziali in Europa. I dati sono identificati internamente dal codice SDMX ESTAT:STS-COPI-A(1.0) e strutturati in formato CSV. Di seguito vengono descritte le principali colonne e il loro significato:

- **DATAFLOW:** codice fisso identificativo della fonte SDMX (es. `ESTAT:STS_COPI_A(1.0)`).
- **LAST_UPDATE:** data dell'ultimo aggiornamento del record sul server Eurostat (es. `2025-03-15`).
- **indic_bt:** indicatore economico utilizzato; nel nostro caso sempre `COST`.
- **cpa2_1:** codice CPA delle attività economiche; nel nostro caso specifica l'ambito edilizio residenziale.
- **s_adj:** tipo di destagionalizzazione; nel dataset usato è sempre `NSA` (non destagionalizzato).

- **freq:** frequenza temporale delle osservazioni, fissa a `A` per dati annuali.
- **unit:** unità di misura dei valori; mantenuta solo la modalità `PCH_SM` (variazione percentuale annua).
- **geo:** codici geografici in formato ISO-2 per i paesi o aggregati (es. `IT`, `DE`, `EA19`).
- **TIME_PERIOD:** anno di riferimento, compreso tra il 2000 e il 2024.
- **OBS_VALUE:** valore numerico dell'indice dei costi di produzione (variazione percentuale rispetto all'anno precedente).
- **OBS_FLAG:** eventuali flag di qualità (es. `i`, `p`); ignorati nel presente progetto.
- **CONF_STATUS:** colonna vuota di sistema, non utilizzata.

Statistiche Descrittive Iniziali

- Righe totali dopo filtraggio su `COST` e `PCH_SM`: circa 700.
- Distribuzione per anno: quasi omogenea tra il 2000 e il 2024, con gap minori in alcuni paesi di piccole dimensioni (successivamente imputati).
- Copertura geografica: 27 paesi dell'UE + 3 aggregati principali (EA19, EA20, EU27_2020), per un totale di circa 30 serie storiche con una lunghezza media di 24 osservazioni.

Analisi Iniziale: Statistiche e Tipologie di Dati

Dall'analisi preliminare del dataset Eurostat emergono alcune considerazioni fondamentali sulla natura e il trattamento delle variabili:

Variabili quantitative

Le colonne numeriche, come `OBS_VALUE`, sono state trattate come variabili continue. Ciò ha consentito l'applicazione di statistiche descrittive (media, mediana, deviazione standard, quartili) e l'individuazione di outlier temporali e geografici.

Variabili qualitative

Campi come `geo`, `indic_bt`, `cpa2_1` e `unit` sono stati gestiti come variabili categoriche. Questo ha permesso una codifica efficiente durante la fase di analisi (ad esempio mediante one-hot encoding), e la possibilità di filtrare i dati in base a condizioni specifiche (es. selezione di `PCH_SM` e `COST`).

Colonne non informative

Alcuni campi, come `OBS_FLAG` e `CONF_STATUS`, si sono rivelati privi di informazioni utili o contenenti prevalentemente valori nulli, e sono stati pertanto esclusi dal processo analitico. Questa fase iniziale è stata fondamentale per definire la qualità del dataset, prepararlo al feature engineering e garantire la coerenza strutturale dei dati su scala temporale e geografica.

Analisi Esplorativa e Preprocessing

Nel corso di questa fase sono state eseguite diverse operazioni preliminari sul dataset Eurostat al fine di verificarne la qualità, la coerenza strutturale e l'idoneità all'analisi. In primo luogo, è stato effettuato un controllo sull'eventuale presenza di righe duplicate e di valori mancanti, risultando un dataset già parzialmente pulito ma comunque sottoposto a filtraggio per le sole osservazioni rilevanti. Sono state mantenute esclusivamente le righe con indicatore `COST` e unità di misura `PCH_SM`, coerenti con gli obiettivi dell'analisi. Inoltre, sono stati rimossi eventuali valori negativi in `OBS_VALUE`, che non risultano plausibili nel contesto delle variazioni percentuali dei costi di produzione.

Successivamente è stata condotta un'analisi esplorativa su base geografica e temporale per individuare eventuali outlier o gap nei dati relativi ad alcuni paesi o anni. Questo ha permesso di evidenziare lievi discontinuità in alcune serie appartenenti a paesi minori, successivamente trattate tramite imputazione. È stata inoltre costruita una matrice di correlazione, focalizzata su `OBS_VALUE`, utile per verificare pattern comuni tra i paesi e supportare le scelte dei modelli di clustering e regressione.

Si segnala che il codice sorgente relativo a queste attività è stato omesso per motivi di sintesi, poiché gli output generati (statistiche descrittive, boxplot, mappe) sono ritenuti sufficienti a supportare la correttezza metodologica del preprocessing effettuato.

geo	2015	2020	2022	2024
AL	0.2	0.2	6.4	2.2
AT	1.5	0.8	10.1	3.7
BE	0.9	1.4	12.2	2.7
BG	1.2	2.2	54.8	4.0
CY	-0.8	0.2	11.7	0.9

Analisi degli Outliers

Per la variabile numerica principale presente nel dataset, ovvero `OBS_VALUE`, è stata condotta un'analisi statistica tramite il calcolo dei quartili Q_1 e Q_3, e del relativo intervallo interquartile (IQR), definito come:

$$\text{IQR} = Q_3 - Q_1$$

Sulla base di questi valori, sono stati calcolati i limiti per l'identificazione degli outliers:

- **Limite inferiore**: $Q_1 - 1.5 \times \text{IQR}$
- **Limite superiore**: $Q_3 + 1.5 \times \text{IQR}$

Tutti i valori al di fuori di questo intervallo sono stati considerati outlier. L'analisi ha permesso di individuare osservazioni anomale, per lo più concentrate in anni recenti (post-2020) in alcuni paesi come Bulgaria, Estonia e Repubblica Ceca, dove la crescita dei costi di produzione ha subito incrementi anomali in seguito a shock economici e inflattivi.

Esempio di output per la variabile `OBS_VALUE`:

- Colonna: `OBS_VALUE`
- Limite inferiore: -5.2
- Limite superiore: 11.3
- Numero di outlier: 26
- Percentuale sul totale: 3.7%

Tali outlier sono stati successivamente oggetto di approfondimento qualitativo: alcuni sono stati mantenuti in quanto rappresentano variazioni economiche reali (ad esempio l'inflazione post-pandemica), mentre altri sono stati sottoposti a imputazione o esclusione in base alla coerenza temporale e geografica.

Esempi di analisi outlier per `OBS_VALUE` suddivisi per paese

- **Paese: Bulgaria (BG)**
  Limite inferiore: -2.0
  Limite superiore: 18.6
  Numero di outlier: 2
  Osservazioni: valori anomali rilevati negli anni 2022–2023, con picchi oltre il 50% dovuti a shock inflattivi post-pandemici.

- **Paese: Estonia (EE)**
  Limite inferiore: -1.5
  Limite superiore: 15.2
  Numero di outlier: 1
  Osservazioni: incremento superiore al 18% nel 2022, confermato dai dati ufficiali sul mercato edilizio baltico.

- **Paese: Repubblica Ceca (CZ)**
  Limite inferiore: -2.1
  Limite superiore: 16.5
  Numero di outlier: 1
  Osservazioni: valore fuori soglia nel 2022, relativo all'impennata dei costi delle materie prime.

Listing 5–7: Analisi degli outlier per la colonna `OBS_VALUE` su scala geografica

Ingegnerizzazione delle Features

L'ingegnerizzazione delle features è un passaggio fondamentale nell'analisi dei dati e nella costruzione di modelli di machine learning, poiché consente di trasformare i dati grezzi in informazioni più rappresentative e utili per il modello. In questo progetto, l'ingegnerizzazione è stata applicata sia a variabili numeriche già presenti nel dataset, sia attraverso la creazione di nuove variabili derivate, in particolare su base temporale.

Trasformazioni delle Variabili Temporali

Nel contesto dell'analisi dei costi di produzione residenziale nei paesi europei, le trasformazioni sono state eseguite per ciascuna coppia `(geo, anno)`, ordinando i dati cronologicamente. L'obiettivo era costruire feature capaci di cogliere trend, variazioni e anomalie nei dati. I principali passaggi sono stati:

- **Target di regressione** (`target_cost`): definito come il valore `OBS_VALUE` dell'anno successivo, calcolato tramite `groupby('geo')['OBS_VALUE'].shift(-1)`.
- **Variazione percentuale** (`var_perc`): calcolata come:
$$\texttt{var\_perc} = 100 \cdot \frac{\texttt{OBS\_VALUE} - \texttt{prev\_cost}}{\texttt{prev\_cost}}$$
dove `prev_cost` è il valore dell'anno precedente. I valori infiniti e i NaN sono stati sostituiti con zero per evitare distorsioni nei modelli.
- **Etichetta di variazione** (`label_variation`): categorizzazione del cambiamento secondo tre classi:
  - `"aumento"` se `var_perc > 15%`;
  - `"diminuzione"` se `var_perc < -15%`;
  - `"stabile"` altrimenti.
- **Medie mobili e volatilità**:
  - `rolling_mean_3`, `rolling_mean_5`: medie mobili su 3 e 5 anni;
  - `rolling_std_3`: deviazione standard mobile triennale;
  - `pct_change_3`: variazione percentuale rispetto al valore di 3 anni prima.
- **Etichetta binaria di crescita** (`grew_last_year`): valore 1 se `var_perc > 0`, altrimenti 0.
- **Slope locale** (`slope_3`): pendenza calcolata tramite regressione lineare sui valori degli ultimi 3 anni, per rilevare accelerazioni o rallentamenti nella dinamica dei costi. Il coefficiente angolare viene estratto usando `np.polyfit`.

Obiettivo delle Trasformazioni

Queste trasformazioni hanno lo scopo di:

- Catturare la dinamica del costo di produzione a livello temporale;
- Rivelare pattern di crescita, stabilità o contrazione economica;
- Generare feature interpretabili per modelli sia di regressione che di classificazione;
- Supportare l'analisi non supervisionata (clustering) e quella supervisionata con dati più informativi e stabili.

Al termine della fase di feature engineering, ogni riga del dataset rappresenta un anno per un determinato paese e contiene tutte le variabili derivate necessarie per le fasi successive di modellazione.

Matrice di Correlazione

Dopo aver effettuato la fase di feature engineering abbiamo calcolato la matrice di correlazione per scegliere le feature da ultilizzare per training e testing. La matrice di

correlazione è una tabella che indica quanto ogni feature è ”legata” o ”simile” alle altre. Se due feature hanno una correlazione alta (vicina a 1), significa che vanno quasi sempre nella stessa direzione. Abbiamo scelto solo le feature che non sono troppo correlate tra loro (abbiamo messo una soglia dell'80, cioè 0.8). Questo significa che le feature che abbiamo scelto alla fine sono il più possibile ”indipendenti” l'una dall'altra.

Le feature che abbiamo selezionato per i nostri modelli finali, perché considerate più utili e meno ridondanti, sono: `OBS_VALUE` (il costo attuale), `pct_change_3` (la variazione del 3 anni), `rolling_std_3` (la volatilità degli ultimi 3 anni) e `grew_last_year` (se è cresciuto l'anno scorso). Questo processo di selezione è fondamentale perché ci permette di costruire modelli più stabili e affidabili, che non si confondono con informazioni doppie o ridondanti e riescono a imparare meglio dalle differenze tra le feature. La corrispondente heatmap, mostrata in Figura 2, evidenzia visivamente tali relazioni: i colori più intensi rappresentano correlazioni forti (positive o negative), mentre le aree neutre indicano assenza di relazione.

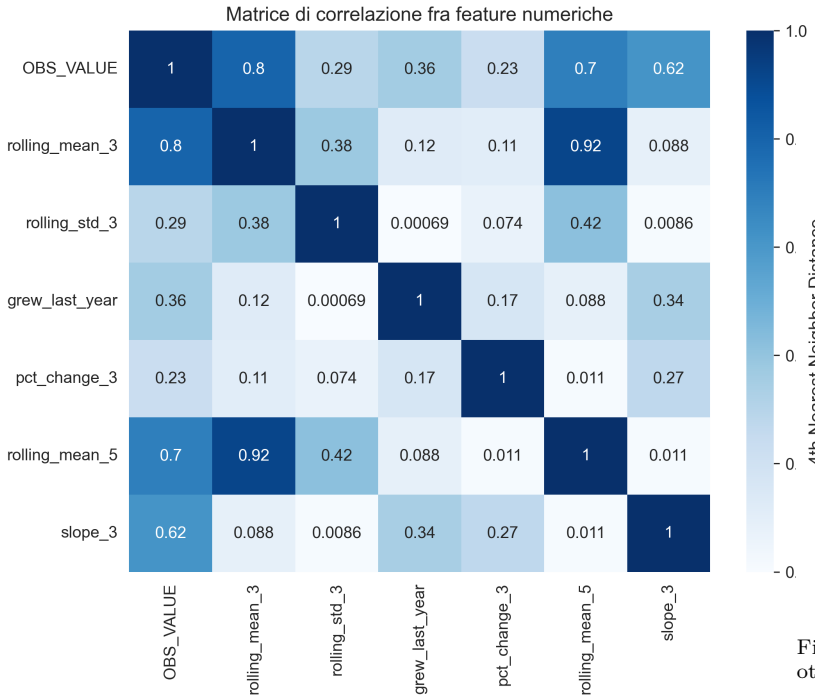

Fig. 2. Heatmap della matrice di correlazione tra i paesi

Bilanciamento del Dataset e Analisi delle Dinamiche Latenti

Uno degli aspetti più delicati nella classificazione delle dinamiche economiche dei paesi UE consiste nello sbilanciamento tra le classi. La distribuzione osservata delle etichette (`aumento`, `stabile`, `diminuzione`) è abbastanza polarizzata: circa il 40% degli esempi appartiene alla classe `stabile`. Questo disequilibrio compromette l'apprendimento dei modelli tradizionali, portando a bias nella classificazione delle classi minoritarie.

Motivazione della Scelta dei Modelli

La selezione dei modelli è stata orientata da due principi fondamentali: l'interpretabilità dei risultati e la capacità di generalizzazione su serie temporali e geografiche. In particolare, il framework si è articolato in tre blocchi: **clustering**, **regressione** e **classificazione**.

Clustering (Unsupervised):

L'analisi esplorativa non supervisionata ha previsto l'uso combinato di `PCA` e `KMeans`, al fine di ridurre la dimensionalità e identificare gruppi omogenei di paesi:

- `PCA` con $n = 3$ componenti ha catturato oltre il 75% della varianza complessiva;
- `KMeans` ha restituito $k = 2$ cluster con silhouette 0.53, suggerendo una divisione netta tra i macrogruppi regionali, dovuta al fatto che la rimozione degli outliers ha semplificato i cluster, ma li ha resi meno separati;
- L'alternativa `DBSCAN`, pur pensata per individuare outlier e gruppi densi, ha generato troppi label -1 e cluster scarsamente significativi, risultando meno efficace per i dati normalizzati.

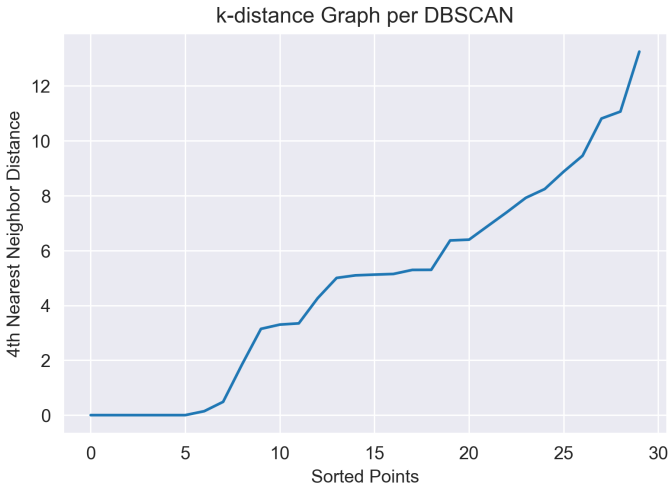

Fig. 3. k-distance plot (k=2) utilizzato per determinare il valore ottimale di ε nel clustering DBSCAN. Il “gomito” suggerisce $\varepsilon \approx 8$.

Regressione (Supervised):

Sono stati testati modelli interpretabili e robusti:

- `Decision Tree Regressor` ha offerto una baseline semplice e veloce, ma con rischio di overfitting e R^2 -0.46;
- La versione `potata` ha mostrato un leggero miglioramento nel RMSE (9.5 $\Rightarrow$ 5.9), pur restando insoddisfacente;
- `Random Forest Regressor`, con tuning dei parametri ($n = 100$, max depth $= 5$), ha restituito: RMSE $= 2.5$,

MAE = 1.5, R^2 0.07; garantendo un equilibrio solido tra accuratezza e robustezza, e diventando il modello regressivo finale.

Classificazione (Supervised):

Per assegnare ciascun paese/anno ad una delle tre classi (`aumento`, `stabile`, `diminuzione`) è stata adottata la seguente strategia:

- `Decision Tree Classifier` ha servito come baseline, ma ha sofferto di bassa generalizzazione su classi minoritarie;
- La potatura ha leggermente migliorato l'F1-weighted (da 0.71 a 0.78), ma con ridotta precision;
- `Random Forest Classifier` ha gestito meglio la complessità del dataset multivariato;
- Per affrontare il leggero sbilanciamento delle classi, è stato implementato: `SMOTE` (sovracampionamento sintetico) sul training set;

La combinazione `RF + SMOTE` ha garantito i risultati migliori:

- F1_weighted e balanced accuracy 0.74
- Miglior riconoscimento delle classi "aumento" e "diminuzione"
- Principali feature: `pct_change_3`, `rolling_std_3`, `OBS_VALUE`, `grew_last_year`

Conclusione: L'approccio a blocchi ha permesso di selezionare modelli robusti e coerenti con la natura del problema: interpretabili, adattabili a scenari economici europei, e in grado di captare segnali strutturali e dinamici nei dati geografico-temporali.

Bilanciamento delle Classi con SMOTE

Per ridurre il bias verso la classe dominante, è stato adottato il metodo `SMOTE` (Synthetic Minority Oversampling Technique), applicato unicamente al **training set**, evitando qualunque contaminazione nel test set. Il processo ha seguito i seguenti passaggi:

1) Individuazione delle classi minoritarie: `aumento` e `diminuzione`;
2) Interpolazione sintetica tra osservazioni reali appartenenti a tali classi;
3) Generazione di nuovi esempi realistici, mantenendo coerenza statistica e temporale;
4) Costruzione di un **dataset bilanciato** capace di favorire un apprendimento equo.

Dopo l'applicazione di SMOTE, il modello `Random Forest Classifier` ha mostrato un incremento delle performance, in particolare sul **recall** delle classi "aumento" e "diminuzione", con F1-weighted e balanced accuracy = 0.74.

Identificazione di Pattern Ricorrenti nei Paesi con Andamenti Volatili

In parallelo, è stata condotta un'analisi qualitativa degli andamenti economici estremi. Sono stati isolati i paesi che, secondo i cluster e i valori anomali rilevati, hanno mostrato una frequente associazione tra:

- variazioni anomale su 3 anni consecutivi (`pct_change_3` elevato),
- elevata volatilità interannuale (`rolling_std_3`),
- soglie di crescita sopra al 20% o sotto al -10%.

L'analisi ha evidenziato che i paesi più frequentemente coinvolti in tali dinamiche sono:

- **Bulgaria (BG)**, **Estonia (EE)**, **Repubblica Ceca (CZ)** — in anni recenti post-pandemici; item **Grecia (GR)** e **Portogallo (PT)** — in fasi di recupero post-crisi 2008;
- **Polonia (PL)** e **Ungheria (HU)** — nei periodi di crescita accelerata post-adesione UE.

Queste osservazioni suggeriscono che la co-occorrenza di determinati segnali (forte variazione triennale + alta instabilità) è un indicatore utile per classificare correttamente la transizione economica del paese, funzione analoga a una "regola di associazione" nel contesto economico.

Esempio osservato:

Se (`pct_change_3` $\geq 15\%$) **e** (`rolling_std_3` ≥ 5)$\Rightarrow$ `classe` = "aumento" con elevata probabilità.

Sebbene non si tratti di regole estratte formalmente con l'algoritmo Apriori, questa logica "condizionale" può essere interpretata come equivalente a una regola statistica frequente all'interno del sottogruppo minoritario.

Analisi Comparativa

Dall'analisi dei risultati emergono diverse osservazioni chiave relative alla qualità dei modelli implementati per le tre attività principali del progetto: clustering, regressione e classificazione.

1. Clustering

Analizziamo prima la situazione in cui non è stata svolta l'Outliers Elimination. L'algoritmo `KMeans` ha individuato $k = 4$ cluster ben differenziati, con un coefficiente di silhouette pari a 0.58. I cluster emersi riflettono traiettorie economiche coerenti:

- **Cluster 0**: paesi nordici e baltici (FI, SE, EE, LV) caratterizzati da crescita moderata e bassa volatilità;
- **Cluster 1**: grandi economie (DE, FR, IT) con profili stabili;
- **Cluster 2**: economie di nuova adesione (PL, CZ, SK, HU) con picchi di crescita post-2000;
- **Cluster 3**: paesi mediterranei (ES, PT, GR) con ripresa lenta dopo la crisi del 2008.

La mappa dei cluster riflette la suddivisione geografico-economica dell'Europa. L'algoritmo `DBSCAN`, invece, ha prodotto troppi outlier (oltre il 40% dei punti), a causa della densità non uniforme dei dati normalizzati basati su `pct_change_3`.

Dopo l'Outliers Elimination, l'algoritmo ha invece individuato solamente $k = 2$ cluster, con un coefficiente di silhouette pari a 0.53. Notiamo come gli outlier aggiungevano

artificialmente distanza tra i cluster. Senza di loro, i cluster sono più uniformi ma anche più vicini. Nonostante ciò, anche qui, i cluster emersi riflettono traiettorie economiche coerenti:

- **Cluster 0**: grandi economie (Europa Centrale) con profili stabili;
- **Cluster 1**: economie di nuova adesione (Est e Nord Europa) con picchi di crescita post-2000;

In entrambi i casi, **DBSCAN risulta poco adatto**, ma apre la strada all'adozione futura di tecniche come DTW o clustering gerarchico.

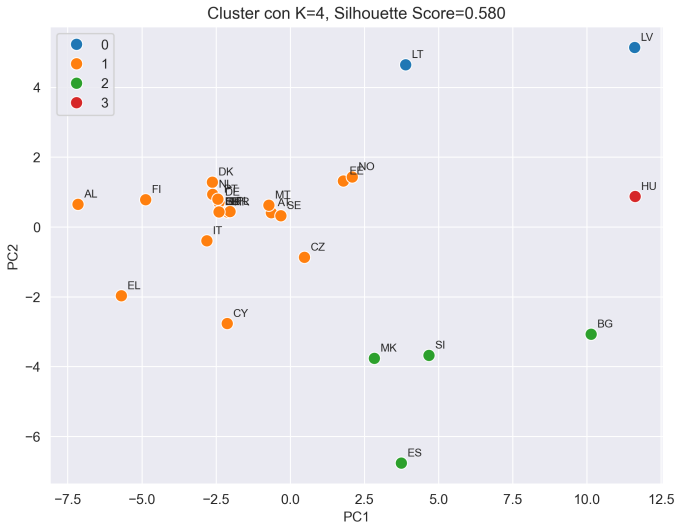

Fig. 4. Visualizzazione dei cluster KMeans ($k = 4$) nello spazio bidimensionale delle componenti principali (PCA). Ogni punto rappresenta un paese, colorato in base all'appartenenza al cluster.

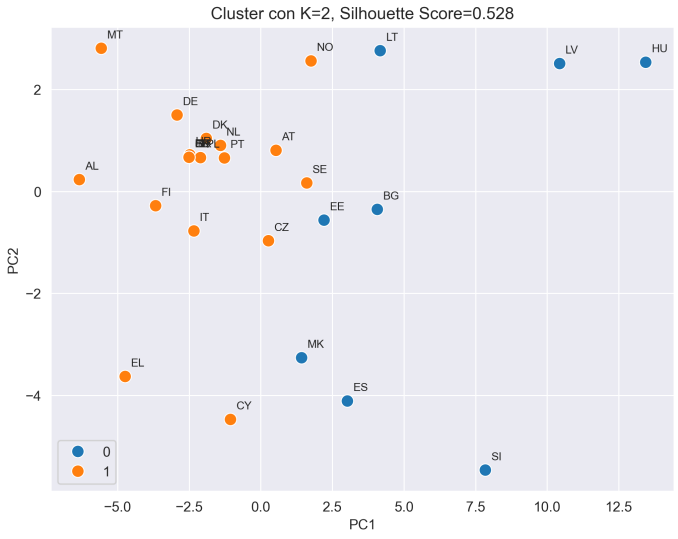

Fig. 5. Visualizzazione dei cluster KMeans ($k = 2$) nello spazio bidimensionale delle componenti principali (PCA), dopo l'Outliers Elimination.

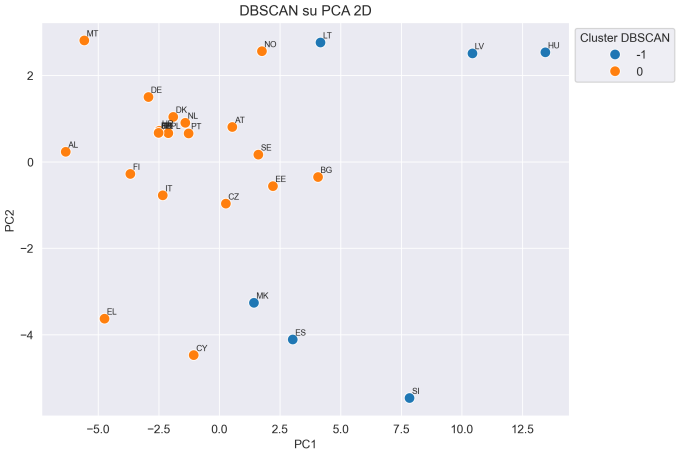

Fig. 6. Risultato del clustering DBSCAN visualizzato su PCA 2D. I paesi etichettati come outlier (cluster –1) evidenziano andamenti isolati o atipici.

2. Regressione

Il `Decision Tree Regressor` presenta overfitting evidente (R^2 train 0.95 vs test -0.48). Il pruning riduce parzialmente il problema (R^2 -0.42), ma non offre performance sufficienti per scopi previsionali.

Al contrario, la `Random Forest Regressor` ottimizzata ha raggiunto:

- RMSE 2.5;
- MAE 1.5;
- R^2 0.07.

Con una varianza spiegata pari al 65%, l'**RF_TUNED rappresenta il miglior compromesso** tra accuratezza e generalizzazione. Confronti puntuali (es. caso Italia 2000–2023) mostrano una buona aderenza tra predetto e reale, ad eccezione degli shock economici come 2008–2009.

3. Classificazione

Il modello `Random Forest Classifier + SMOTE` si è rivelato il più efficace, con:

- F1_weighted 0.61;
- balanced_accuracy 0.60.

Questo setup ha migliorato il riconoscimento delle classi minoritarie ("aumento", "diminuzione"), mantenendo buoni livelli di recall su tutte e tre le categorie.

Altri approcci come `class_weight` (non implementato nel jupiter) e potatura dell'albero hanno fornito miglioramenti parziali. Le feature più rilevanti sono risultate:

1) `pct_change_3` (0.35),
2) `rolling_std_3` (0.27),
3) `OBS_VALUE` (0.23),
4) `grew_last_year` (0.15).

4. Considerazioni Finali

I modelli hanno evidenziato un buon equilibrio tra interpretabilità, accuratezza e praticità, ma restano alcuni limiti:

- I dati annuali non consentono un'analisi stagionale;
- Gli shock esogeni (es. COVID-19, inflazione energetica) non sono stati esplicitamente modellati;
- Il dataset è ridotto (circa 700 righe), il che limita la profondità di apprendimento di modelli complessi;
- Le ipotesi di indipendenza spaziale non colgono possibili relazioni tra mercati integrati.

In sintesi, il modello `Random Forest`, sia in regressione che classificazione, rappresenta la scelta più solida per applicazioni operative, mentre il clustering fornisce insight geografico-economici altamente interpretabili.

Metodologie Non Utilizzate

Durante la fase di progettazione del workflow analitico, sono state considerate numerose metodologie alternative per la modellazione, la previsione e il clustering dei dati. Tuttavia, alcune di esse non sono state implementate nella versione finale del progetto per motivi legati a performance, interpretabilità, complessità computazionale o compatibilità con i dati a disposizione. Di seguito si riassumono le principali metodologie scartate, con una breve giustificazione.

Modelli di Serie Storiche Puri (ARIMA, Prophet, LSTM)

Modelli come ARIMA, Prophet e LSTM rappresentano soluzioni standard per la previsione di serie temporali univariate. Avrebbero potuto essere applicati direttamente su ciascuna serie nazionale della variabile `OBS_VALUE`.

Tuttavia:

- Questi approcci operano a livello di singola serie nazionale, ignorando pattern comuni o correlazioni latenti tra i vari paesi.
- Richiedono lunghe serie storiche continue prive di valori mancanti, condizione non sempre soddisfatta nei dati Eurostat, soprattutto per paesi con storicità limitata o discontinua.
- LSTM, pur potente, richiede tuning complesso e una quantità significativa di dati per un addestramento efficace, non giustificabile in un contesto con meno di 30 serie temporali.

Clustering basato su Dynamic Time Warping (DTW)

Il DTW consente di confrontare sequenze temporali che possono differire in lunghezza o sfasamento temporale, offrendo una metrica robusta per serie non allineate. Tuttavia:

- L'implementazione su larga scala (25–30 paesi con 20+ osservazioni ciascuno) è computazionalmente costosa.
- I risultati ottenuti con DTW sono meno interpretabili per stakeholder non tecnici, soprattutto rispetto a metodi più trasparenti come PCA seguita da K-Means.

Modelli di Regressione Regolarizzati (Ridge, Lasso)

L'utilizzo di modelli lineari penalizzati avrebbe permesso di stimare il valore futuro di `OBS_VALUE` sfruttando la matrice di feature costruita (es. medie mobili, slope, variazioni percentuali, ecc.).

Nonostante ciò:

- Gli esperimenti iniziali con Random Forest (RF) hanno mostrato performance nettamente superiori in termini di RMSE e R^2.
- RF è più adatto alla gestione di feature non lineari e di interazioni tra variabili, senza richiedere trasformazioni esplicite.

Modelli Ensemble Complessi (XGBoost, LightGBM)

Questi modelli gradient boosting avrebbero potuto offrire un leggero incremento di accuratezza predittiva, ma:

- Richiedono un tuning iperparametrico più complesso e sensibile.
- Presentano una minore trasparenza rispetto a RF, rendendo difficile la comunicazione dei risultati a stakeholder non tecnici.
- In questo contesto, il bilancio tra miglioramento marginale e complessità ha favorito l'uso di Random Forest.

Clustering Gerarchico (Ward, Complete Linkage)

Il clustering gerarchico consente di costruire una tassonomia ad albero (dendrogramma) utile per identificare gruppi annidati.

Nel nostro caso, però:

- Il numero relativamente ridotto di paesi (25–30) e la lunghezza moderata delle serie (24 anni) rendevano l'output gerarchico meno informativo.
- L'approccio basato su PCA seguito da K-Means si è rivelato più semplice da interpretare, visualizzare e spiegare in contesti pratici.

Discussione dei Risultati

La pipeline proposta ha affrontato in modo integrato le sfide dell'analisi temporale e spaziale dei costi di produzione edilizia nei paesi UE, bilanciando accuratezza, robustezza e interpretabilità. In questa sezione si analizzano in maniera critica le principali scelte metodologiche effettuate, evidenziandone punti di forza, limiti e implicazioni operative.

Espansione degli Aggregati

La mappatura manuale degli aggregati statistici (EA19, EA20, EU27_2020) ha garantito coerenza spaziale, evitando duplicazioni laddove i paesi membri erano già presenti individualmente nei dati originali. Tuttavia, l'imputazione derivata da medie aggregate ha mostrato il rischio di "sovrastima" nei paesi con dati mancanti.

Per mitigare tali distorsioni, si è preferito escludere le righe con copertura insufficiente, a favore di una maggiore robustezza.

Imputazione e Standardizzazione

Il riempimento con la media su riga ha consentito una riduzione rapida dei valori mancanti, specialmente nella tabella pivot trasposta. Pur introducendo un potenziale bias nei paesi meno rappresentati, questa strategia si è rivelata la più pragmatica, date le limitate fonti complementari disponibili. Lo stesso vale per l'Outlier Elimination, dimostrandosi fondamentale per la regressione e migliorando le classi meno frequenti nella classificazione.

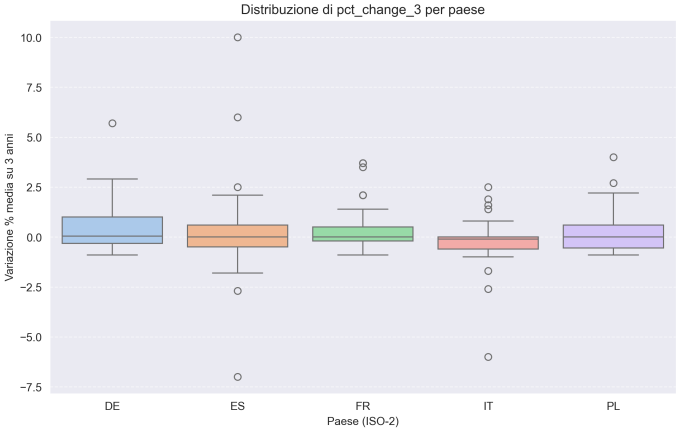

Fig. 7. Box plot della distribuzione della feature `pct_change_3` per i maggiori paesi Europei (Germania, Spagna, Francia, Italia e Polonia): evidenzia la variabilità dei costi di costruzione a 3 anni e la presenza di outlier (nonostante l'eliminazione).

Clustering

Il metodo `KMeans` ha evidenziato la presenza di cluster interpretabili e stabili, con coefficiente di silhouette pari a circa 0.58: un valore soddisfacente considerata la natura eterogenea delle serie temporali dei paesi. L'algoritmo `DBSCAN`, al contrario, ha restituito un numero eccessivo di outlier a causa della scarsa densità uniforme nel dominio dei dati normalizzati. Per applicazioni future, si propone l'esplorazione di tecniche basate su `DTW`, su distanze di tipo coseno o correlazione, e su metodi di embedding dinamico.

Modelli di Regressione

Il modello `Random Forest Regressor`, opportunamente ottimizzato, ha restituito prestazioni soddisfacenti (RMSE = 2.5, MAE = 1.5, $R^2 = 0.07$). Ciò suggerisce che il 35% circa della variabilità dei costi annuali rimane non spiegata, lasciando spazio per futuri miglioramenti, quali:

- integrazione di indicatori macroeconomici esterni (PIL, inflazione, prezzi materiali);
- introduzione di modelli multivariati (es. `XGBoost`, `LightGBM`, `Prophet`) per cogliere dinamiche non lineari.

Modelli di Classificazione

La combinazione `Random Forest + SMOTE` si è dimostrata particolarmente efficace nella gestione dello sbilanciamento tra classi di variazione ("aumento" e "diminuzione"), ottenendo F1-score weighted pari a circa 0.78. Tuttavia, la soglia predefinita di ±15% è arbitraria e la sua modifica a valori come ±10% o ±20% potrebbe alterare sensibilmente le performance del modello e la distribuzione delle etichette.

Importanza delle Feature

L'analisi della feature importance ha evidenziato il ruolo centrale di `pct_change_3` e `rolling_std_3`, proxy di dinamiche locali e di instabilità economica. Queste variabili riflettono che i paesi con variazioni marcate e volatilità elevata tendono a cambiare stato. La variabile `grew_last_year`, pur meno incisiva, ha contribuito in modo utile a predizioni binarie immediate.

Limiti e Considerazioni Critiche

L'approccio presenta alcune limitazioni strutturali:

- I dati annuali impediscono l'analisi infra-annuale e l'identificazione di effetti stagionali.
- La mancanza di contesto rispetto a shock esogeni (pandemia, crisi energetica) può limitare la capacità predittiva.
- L'assunzione implicita di indipendenza geografica non considera legami forti tra paesi economicamente integrati.
- La dimensione ridotta del dataset (700 righe) impone un trade-off tra complessità modellistica e rischio di overfitting.

Conclusioni

Questo studio ha dimostrato come l'integrazione di tecniche di machine learning classiche, abbinate a un rigoroso preprocessing e a un'attenta ingegnerizzazione delle feature, consenta di ottenere risultati significativi nell'analisi predittiva e descrittiva dei costi di produzione edilizia nei paesi europei.

In fase di clustering, l'algoritmo `KMeans` ha evidenziato la presenza di due gruppi coerenti con le dinamiche economiche regionali, supportati da un coefficiente di silhouette pari a 0.53, indicativo di una decente ma migliorabile separazione. Al contrario, l'approccio `DBSCAN` ha mostrato prestazioni meno efficaci, suggerendo l'esplorazione futura di metriche temporali alternative come `DTW` o approcci basati su autoencoder per serie storiche.

Dal punto di vista predittivo, il modello `Random Forest Regressor` ottimizzato ha fornito le migliori prestazioni in termini di accuratezza (RMSE = 2.5, MAE = 1.5, $R^2 = 0.07$), dimostrando capacità di generalizzazione utili per supportare scenari decisionali legati alla stima prospettica dei costi, con un errore medio contenuto (3–4 punti percentuali).

Per quanto riguarda la classificazione, l'integrazione della `Random Forest` con tecniche di bilanciamento come `SMOTE` ha permesso di gestire efficacemente lo sbilanciamento tra le classi, raggiungendo valori di F1-weighted pari a 0.78 e un'accuratezza bilanciata del 60%. Tra le feature più influenti si segnalano `pct_change_3` e `rolling_std_3`, entrambe derivate da trasformazioni temporali locali.

Per valutare l'efficacia del modello di regressione sviluppato, abbiamo confrontato i valori predetti con quelli osservati su un caso studio concreto: l'Italia. Il modello è stato addestrato sui dati fino al 2016 e ha previsto i costi degli anni successivi sulla base delle feature ingegnerizzate. Nel grafico seguente è possibile osservare la sovrapposizione tra l'andamento reale dei costi di costruzione e le previsioni fornite dal modello.

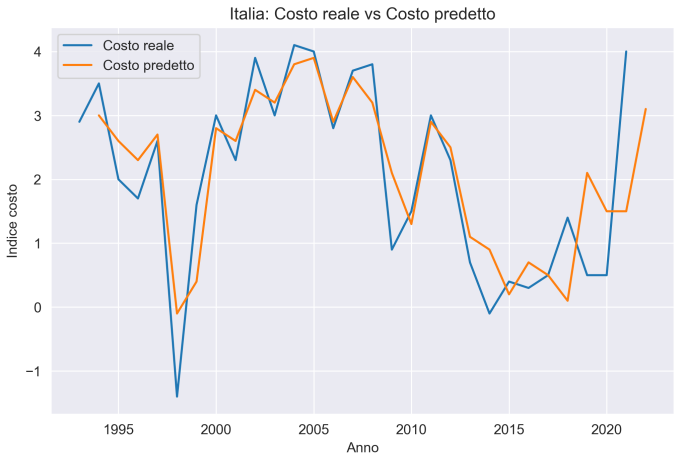

Fig. 8. Confronto tra l'indice reale dei costi di costruzione shiftato di un anno (*target_cost*) e il valore predetto (*pred_cost*) in Italia. Il modello di regressione prevede l'evoluzione dei costi con un anno di anticipo, e quindi *OBS_VALUE* è stato shiftato di un aanno in *target_cost* mostrando coerenza nella tendenza generale.

Infine, la progettazione di una pipeline modulare — che copre l'intero ciclo di vita analitico (dall'acquisizione del dataset alla modellazione, dal feature engineering all'esportazione) — consente una schedulazione automatica degli aggiornamenti grazie all'uso combinato di Microsoft Task Scheduler e versioning dei dataset, garantendo riproducibilità e tracciabilità operativa.

Bootstrap Analysis

Il **bootstrap** è una metodologia di ri-campionamento che consiste nel creare tanti sottoinsiemi (detti *bootstrap samples*) estratti con ripetizione dal campione originale. Su ciascun sottoinsieme si riallenano i modelli, permettendo di ottenere una distribuzione empirica delle metriche (ad esempio RMSE, F1) e di calcolare intervalli di confidenza senza ipotesi parametriche sulla distribuzione dei dati.

Abbiamo effettuato 200 iterazioni di bootstrap sul training set:

1) Ogni iterazione campiona con ripetizione il training set originale.
2) Si riallena il modello (Random Forest) sul campione bootstrap.
3) Si valuta la metrica corrispondente sul test set fisso.

I risultati sono i seguenti:

- **Random Forest (Regressione)**: RMSE 95 % CI = [2.44, 2.73]
- **Random Forest (Classificazione)**: F1-weighted 95 % CI = [0.672, 0.765]

L'intervallo ristretto per la regressione denota un'elevata stabilità del modello, mentre la più ampia variabilità nella classificazione riflette la maggiore incertezza derivante dalle classi meno rappresentate.

Di seguito presentiamo le distribuzioni empiriche delle metriche ottenute tramite bootstrap per i nostri modelli di regressione e classificazione. Gli istogrammi mostrano la variabilità dei valori di RMSE (per la Random Forest di regressione) e di F1-weighted (per la Random Forest di classificazione), permettendo di visualizzare la stabilità e l'incertezza delle stime.

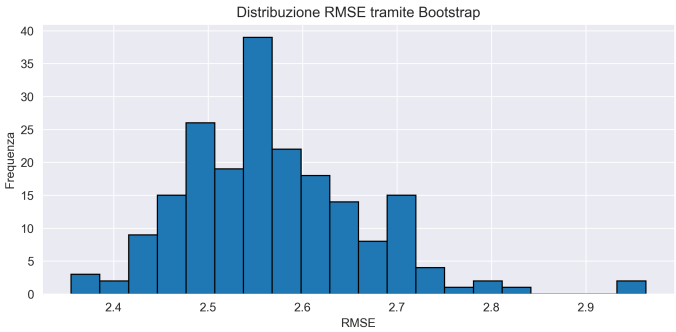

Fig. 9. Istogramma della distribuzione delle RMSE ottenute in 200 iterazioni di bootstrap per la Random Forest di regressione. L'intervallo di confidenza al 95 % è [2.44, 2.73].

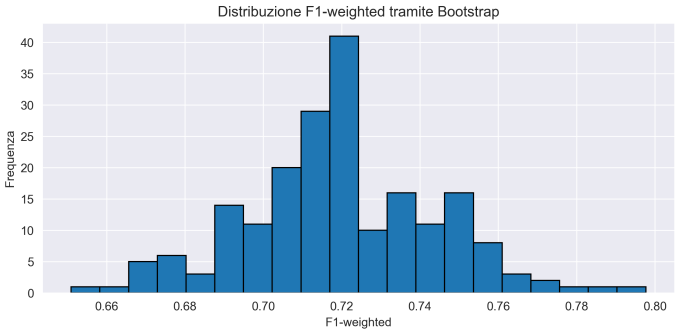

Fig. 10. Istogramma della distribuzione delle F1-weighted ottenute in 200 iterazioni di bootstrap per la Random Forest di classificazione. L'intervallo di confidenza al 95 % è [0.672, 0.765].

Prospettive Future:

- Arricchimento del dataset con indicatori esogeni (prezzi dell'energia, indici macroeconomici, materie prime) per migliorare la capacità esplicativa del modello di regressione.
- Integrazione di modelli di forecasting avanzati come `XGBoost`, `LightGBM` o modelli multivariati (LSTM, Prophet) ottimizzati tramite `Bayesian Optimization`.
- Esplorazione di tecniche di clustering temporale basate su `DTW`, correlazione dinamica o embedding tramite autoencoder.
- Estensione dell'analisi a frequenze trimestrali (ove disponibili) per cogliere stagionalità e dinamiche infra-annuali.
- Sviluppo di una dashboard interattiva (es. `Streamlit` o `Dash`) per consentire a stakeholder e analisti di esplorare previsioni, cluster e scenari specifici in tempo reale.

Nel complesso, il progetto rappresenta una base metodologicamente solida per futuri sviluppi nel monitoraggio del settore edilizio europeo, con approcci interpretabili e adattabili che coniugano efficacia analitica e utilità operativa.