ALMA MATER STUDIORUM – UNIVERSITA' DI BOLOGNA

SCUOLA DI ECONOMIA E MANAGEMENT

Corso di Laurea Magistrale in Economia e Politica Economica

TECNICHE DI MACHINE LEARNING PER LA CLASSIFICAZIONE DEI PRODOTTI OGGETTO DI COMMERCIO ESTERO TRA DIFFERENZIABILI E COMMODITY.

Analisi dei dati di panel

Presentata da:

Lucrezia Fenudi

950808

Relatore:

Prof.ssa Maria Elena Bontempi

APPELLO IV

ANNO ACCADEMICO 2020 / 2021

Sommario

Sommario figure

Sommario tabelle

Capitolo 1

1. Introduzione

È noto che il commercio internazionale possa apportare vantaggi ai Paesi coinvolti nello scambio. Come comunemente avviene nei modelli economici, se tentassimo di semplificare la realtà tralasciassimo eventuali misure protezionistiche dei mercati nazionali (come dazi o standard), potremmo individuare tali vantaggi sia dal lato delle imprese che da quello dei consumatori: le prime potrebbero agire in un mercato non segmentato e operare all'estero alle medesime condizioni nazionali, mentre per i consumatori sarebbe disponibile una maggiore varietà di beni ad un minor costo, grazie all'aumento del grado di concorrenza.

Nella letteratura economica si è a lungo dibattuto sulle ragioni del commercio internazionale: a partire dalle teorie classiche, come quella di A. Smith sui vantaggi assoluti o di D. Ricardo sui vantaggi relativi nella produzione, per passare poi alla teoria sulle differenze nella dotazione dei fattori produttivi del modello di Heckscher-Ohlin. Le nuove teorie, a partire da P. Krugman, focalizzano l'attenzione non più sul vantaggio nell'utilizzo dei fattori ma piuttosto sulle economie di scala che possono svilupparsi; il propulsore agli scambi è la struttura delle preferenze dei consumatori, il cui benessere è in qualche modo proporzionale alla varietà di beni disponibile – *amore per la varietà* – e dipende da quanto la sua scelta sia simile alla sua *varietà ideale.*

Nel quadro delle nuove teorie si definisce quindi la differenziazione di prodotto, una strategia che consente ad un'impresa di offrire un bene che sia, realmente o intrinsecamente, differente da quello delle imprese concorrenti. Tale strategia è particolarmente rilevante in fase di inserimento nei mercati esteri, poiché è l'unicità del prodotto a consentire che l'acquirente sia disposto a pagare un *premium price,* in ragione del fatto che il bene meglio soddisfa le sue esigenze. La differenziazione è quindi strategicamente rilevante, in quanto attraverso questo differenziale di prezzo l'impresa può raggiungere una situazione definita di *above-average performance.* Al contrario, le commodity sono beni altamente standardizzati, per cui nessuno è disposto a riconoscere un prezzo più elevato di quello di mercato.

Data la rilevanza strategica della differenziazione, per l'impresa risulta di notevole importanza indagare sulla natura del prodotto che si accinge a offrire sul mercato estero. Ci si chiede, quindi, in che modo sia possibile distinguere un prodotto differenziabile da una commodity. Una prima analisi potrebbe volgersi alla dispersione del prezzo; nel mercato dei beni differenziati, infatti, è naturale che i prodotti appartenenti alla stessa categoria possano essere

venduti a prezzi molto diversi, in ragione della differenza qualitativa reale o percepita dai consumatori. Quindi, se nello stesso periodo e sullo stesso mercato sono importati beni appartenenti allo stesso codice doganale, ma con prezzi diversi, questa può essere considerata una indicazione di differenziabilità del bene. Tuttavia, considerare tale dispersione come unico indicatore potrebbe essere fuorviante: in un così complesso sistema di mercato, si rilevano delle inefficienze che alterano il funzionamento della *legge del prezzo unico* dei beni omogenei. In circostanze di asimmetria informativa, costi di ricerca o elevata concentrazione di mercato, la dispersione di prezzo si configura come sintomo di fallimento di mercato e non solo come segnale di differenza qualitativa.

Il presente lavoro è stato sviluppato durante il tirocinio presso StudiaBo srl e l'obiettivo è stato quello di costruire un algoritmo di Machine Learning, all'interno del quale inserire variabili ulteriori alla dispersione di mercato, per la classificazione dei prodotti oggetto di commercio estero nelle due classi sopra citate: differenziabili e commodity. Non esiste una classificazione universale di riferimento, pertanto quella prodotta attraverso questo studio verrà confrontata con quella proprietaria StudiaBo; l'intento è quindi quello di produrre un algoritmo che possa funzionare in maniera automatizzata.

La scelta delle variabili a priori è il frutto della rassegna degli studi empirici disponibili sull'argomento e di ragionamenti sviluppati in merito: ulteriori analisi riguardano, ad esempio, la correlazione tra i prezzi tra punti diversi della distribuzione, o ancora il calcolo di indici di concentrazione di mercato o di quote di tipologia commercio.

La struttura è la seguente: il Capitolo [2] passa in rassegna i riferimenti teorici sui quali si è costruita l'analisi; il Capitolo [3] contiene la descrizione delle banche dati impiegate, Ulisse e Comext, sul commercio estero; al Capitolo [4] è invece raccontata la metodologia per la definizione delle variabili inserite successivamente negli algoritmi di Machine Learning, descritti e spiegati nel corso del Capitolo [5], alla fine del quale sono inoltre riportati i risultati ottenuti. Al Capitolo [6] sono infine riassunte le conclusioni. Il lavoro è corredato con Appendice [1], Appendice [2], Appendice [3] e Appendice [4], che riportano grafici e tabelle ulteriori al corpo centrale del testo.

Capitolo 2

2. Fondamenti teorici

2.1. Le teorie sul commercio internazionale

Nel corso dei seguenti paragrafi verranno esposte le principali teorie che studiano le determinanti e le cause dei flussi di commercio internazionale.
Tralasciando alcuni tipi di protezionismo talvolta presenti, come dazi o standard[1], i vantaggi del commercio internazionale si possono individuare sia nel lato dell'offerta che in quello della domanda. Le imprese possono infatti agire in un mercato non segmentato, operando all'estero alle stesse condizioni nazionali, mentre i consumatori possono accedere ad una maggiore varietà di beni e, grazie al maggior grado di concorrenza, ad un minor costo.

2.1.1. Uno sguardo d'insieme: commercio inter-settoriale e intra-settoriale

Le varie teorie sul commercio internazionale mirano a giustificare i flussi commerciali tra paesi individuando differenti propulsori per lo scambio. Tra le ragioni che spiegano il commercio si possono elencare:

- Differenti dotazioni di risorse e/o tecnologiche tra Paesi.
- Volontà di conseguire economie di scala, o rendimenti crescenti, nella produzione.
- Possibilità di accedere ad una maggiore varietà di prodotti.

La letteratura di riferimento è quella che analizza il commercio di tipo inter-industriale e intra-industriale. In linea generale, è possibile dividere tali teorie in due principali filoni:

1. La teoria tradizionale sostiene il commercio internazionale di tipo inter-settoriale, ossia che questo avvenga tra settori differenti in quanto i paesi possiedono diverse dotazioni fattoriali e si basa sull'analisi dei vantaggi comparati. I fattori di produzione sono capitale, lavoro e tecnologia.

[1] Ambientali o fiscali.

2. Le Nuove Teorie sul Commercio Internazionale (NTCI) indagano il commercio intra-settoriale, superando la questione delle dotazioni fattoriali e analizzando il commercio tra paesi all'interno dello stesso settore industriale, introducendo così il concetto di differenziazione di prodotto negli scambi. Il commercio intra-settoriale tra le moderne economie industrializzate non era previsto dalle teorie tradizionali, trattandosi di economie con simile dotazione fattoriale, tecnologica e con consumatori dalle preferenze simili, mentre ricopre invece una parte sostanziale del commercio internazionale.

L'importanza relativa di commercio intra e inter settoriale dipende da quanto sono simili i paesi. Più i paesi sono diversi – in termini di dotazioni – più il loro commercio è di tipo inter-settoriale. Al contrario, più sono simili, più il loro commercio è di tipo intra settoriale: Lancaster (1979) evidenzia come economie simili possano generare un volume di commercio reciproco persino maggiore rispetto a quello tra economie differenti.

2.1.2. La teoria tradizionale sul commercio internazionale: Ricardo e Hecksher-Ohlin

Tra le teorie appartenenti al primo filone si possono annoverare quella di Ricardo[2] e quella di Hecksher-Ohlin[3]. Ricardo è stato il primo economista a distinguere tra commercio nazionale ed internazionale, volendo mostrare che seguissero differenti regole. Le formulazioni classiche e neoclassiche della teoria del vantaggio comparato si differenziano per gli strumenti utilizzati, ma condividono la stessa logica per cui le forze di mercato spingono i fattori di produzione verso il loro migliore utilizzo nell'economia. Il libero commercio internazionale andrebbe a vantaggio di tutti i paesi partecipanti perché incrementerebbe la produzione complessiva e il consumo grazie alla specializzazione basata, appunto, sul vantaggio comparato.

La teoria dei vantaggi comparati - o modello ricardiano - assume che la specializzazione produttiva di un Paese avvenga per il bene sul quale possiede un vantaggio comparato. Ciò implica che un Paese sia più efficiente di un altro e che ci siano differenze per quanto riguarda il costo del lavoro: per quel bene, il costo opportunità in termini di altri beni è minore rispetto agli altri paesi. Il vantaggio comparato comporta l'esistenza di scambi mutualmente

[2] Ricardo, D. "*On the Principles of Political Economy and Taxation"* (1817)
[3] E. Hecksher, B. Ohlin (1919-1924)

vantaggiosi tra due paesi, anche in presenza di svantaggio assoluto di uno nei confronti dell'altro. Questa teoria assume l'immobilità di lavoro e capitale tra le nazioni, per «l'insicurezza immaginaria o reale del capitale, quando non è sotto l'immediato controllo del suo proprietario, così come la naturale riluttanza che ogni uomo ha a lasciare il suo paese natale e i suoi legami, e ad affidarsi con tutte le sue abitudini fisse, a uno strano governo e a nuove leggi[4]».

Un avanzamento nella teoria tradizionale si è avuto con lo sviluppo della teoria di Heckscher-Ohlin delle "dotazioni fattoriali". È di base costruita sulla teoria ricardiana, tuttavia assume che le differenze tra Paesi risiedano nelle dotazioni dei fattori produttivi - lavoro, capitale e terra - e non nella dotazione tecnologica, ed esamina gli effetti del commercio internazionale sulle remunerazioni di tali fattori. Le diverse dotazioni comportano differenze nelle strutture produttive e, di conseguenza, nel commercio. La teoria afferma che venga esportato il bene nella cui produzione viene utilizzato più intensivamente il fattore che nel paese è relativamente abbondante e poco costoso, mentre viene importato il bene che utilizza intensivamente il fattore relativamente scarso e costoso. Quindi, le nazioni con un rapporto capitale/lavoro elevato esporteranno beni ad alta intensità di capitale, e viceversa.

Di conseguenza il commercio dovrebbe tipicamente essere tra nazioni complementari, e cioè le nazioni con elevata dotazione di capitale dovrebbero commerciare con quelle ad alta dotazione di lavoro.

2.1.3. La Nuova Teoria sul Commercio Internazionale: Krugman

Gli economisti non neoclassici confutano i presupposti delle teorie del vantaggio comparato, sostenendo che si basino su presupposti né teoricamente né empiricamente validi, quali:

- Non mobilità dei fattori produttivi a livello internazionale, centrale nella teoria ricardiana
- Assenza di esternalità
- Mobilità delle risorse produttive tra settori industriali

Secondo la teoria tradizionale, infatti, l'intensificarsi delle relazioni commerciali avrebbe dovuto portare ad una maggiore specializzazione produttiva e dar luogo ad un commercio

[4] D. Ricardo, *On the Principles of Political Economy and Taxation* (1817)

inter-settoriale. I flussi commerciali, tuttavia, si sono sviluppati soprattutto tra paesi industrializzati e hanno avuto come oggetto beni sostanzialmente simili, in questo frangente viene quindi definito il cosiddetto commercio intra-settoriale[5]. Questo fenomeno fa quindi nascere il filone delle NTCI che considerano modelli di mercato caratterizzati dalla cosiddetta concorrenza monopolistica, la cui prima formulazione è dovuta a Chamberlin (1933)[6] nella sua "Teoria della concorrenza monopolistica" e ripresa, tra gli altri, da Krugman (1979). Queste teorie, quindi, considerano l'esistenza di mercati non perfettamente concorrenziali, come la concorrenza monopolistica, in cui un elevato numero di produttori relativamente piccolo vendono prodotti differenziati e sono price-setters. Questa struttura di mercato è un caso particolare di oligopolio in cui sono presenti economie di scala, ciascuna impresa è in grado di differenziare i propri prodotti e considera come dati i prezzi praticati dalle imprese concorrenti. Queste assunzioni implicano l'assenza degli atteggiamenti che emergono nel contesto generale di un oligopolio: comportamento collusivo e comportamento strategico. Questa forma di mercato è la più importante forma competitiva nell'analisi delle moderne economie altamente tecnologiche, dal momento in cui l'assunzione di perfetta competizione crolla in presenza di una diversa struttura di preferenze e infinite variazioni nella specificazione dei prodotti.

Tale modelli possono essere impiegati per mostrare come il commercio internazionale conduca a:

- Prezzo medio inferiore, grazie alle economie di scala.
- Disponibilità di una maggiore varietà di prodotti, grazie alla differenziazione.
- Commercio intra-settoriale.

Nel tentativo di spiegare il commercio intra-settoriale tra paesi con dotazioni simili, le nuove teorie analizzano quindi il tipo di struttura di mercato e considerano inoltre alcune variabili microeconomiche, come le preferenze dei consumatori. Il ruolo delle preferenze è centrale per il tema: il principale contributo di Krugman[7] si definisce nell'affermare che le preferenze individuali siano profondamente differenti anche all'interno di uno stesso prodotto. Entra quindi in gioco la varietà di prodotto – appariscente o intrinseca che sia.

[5] Il commercio intra-settoriale è difficile da misurare, in quanto concerne la definizione di prodotti e settori industriali. I principali indici sono Balassa (1965) sui vantaggi relativi comparati, Grubel e Llyod (1971) sul commercio intra-settoriale di un prodotto.

[6] Chamberlin, E.H, "*Theory of Monopolistic Competition" The Economic Journal,* December 1933

[7] Krugman, P.R. "*Increasing Returns, Monopolistic competition and International Trade" American Economic Review,* December 1980, 70, 950-9.

È inoltre riconosciuto il ruolo che le economie di scala giocano in alternativa al tema delle dotazioni fattoriali e tecnologiche nel spiegare la specializzazione internazionale e il commercio. Balassa (1967) e Kravis (1971) sostengono che le economie di scala abbiano avuto un ruolo cruciale durante la crescita del dopoguerra per il commercio tra le nazioni industrializzate. Vengono considerate economie di scala interne, ove il costo unitario dipende dalle dimensioni della singola impresa e non necessariamente dal settore nel suo complesso.
Uno dei risultati dell'analisi è che il grado di sviluppo dei paesi determina il tipo di commercio: più i paesi raggiungono lo stesso grado di sviluppo, più è possibile che si instauri un commercio intra-settoriale. Date dotazioni simili, infatti, si verificano specializzazioni industriali di tipo nazionale.
Alla luce dell'importanza che il commercio intra settoriale svolge all'interno del commercio mondiale, questo lavoro focalizza l'attenzione sullo studio della tipologia dei beni oggetto, indagando a quale essi appartengano.

2.2. La differenziazione di prodotto.

Nella trattazione delle teorie del commercio internazionale è stato introdotto il concetto di differenziazione di prodotto, senza tuttavia darne una definizione.
La differenziazione di prodotto è una strategia di marketing messa in atto dalle imprese per far sì che il prodotto sia, in maniera effettiva o percepita, differente da quello offerto dalle imprese concorrenti. Le caratteristiche di un prodotto, incluso il prezzo, possono essere assimilate agli elementi di un vettore, tali che possano essere confrontate. Per questo la differenziazione è definita anche come condizione di marketplace[8]; nel posizionamento nel mercato, i prodotti non sono percepiti allo stesso modo nel vettore delle diverse caratteristiche, incluso il prezzo.

2.1.1. Differenziazione come strategia competitiva: above-average performance

La competizione è il core del successo o del fallimento di un'impresa e consiste nella ricerca di un buon posizionamento nel settore, ossia una posizione profittevole e sostenibile rispetto alle forze che determinano la competizione stessa. La scelta della strategia competitiva

[8] La più semplice rappresentazione della differenziazione di prodotto è quella fornita dal modello spaziale di Hotelling, nel quale i venditori sono posizionati in vari punti lungo un segmento di una retta e i consumatori sono distribuiti lungo lo stesso, ognuno dei quali acquista in base al costo dell'offerta (dato dal prezzo più il costo di trasporto).

è quindi cruciale per l'attività imprenditoriale, a prescindere dalla profittabilità del settore in cui si opera: anche se non tutti i settori industriali offrono pari opportunità di profittabilità sostenibile, anche in quelli più attrattivi è necessario ottenere un buon posizionamento.

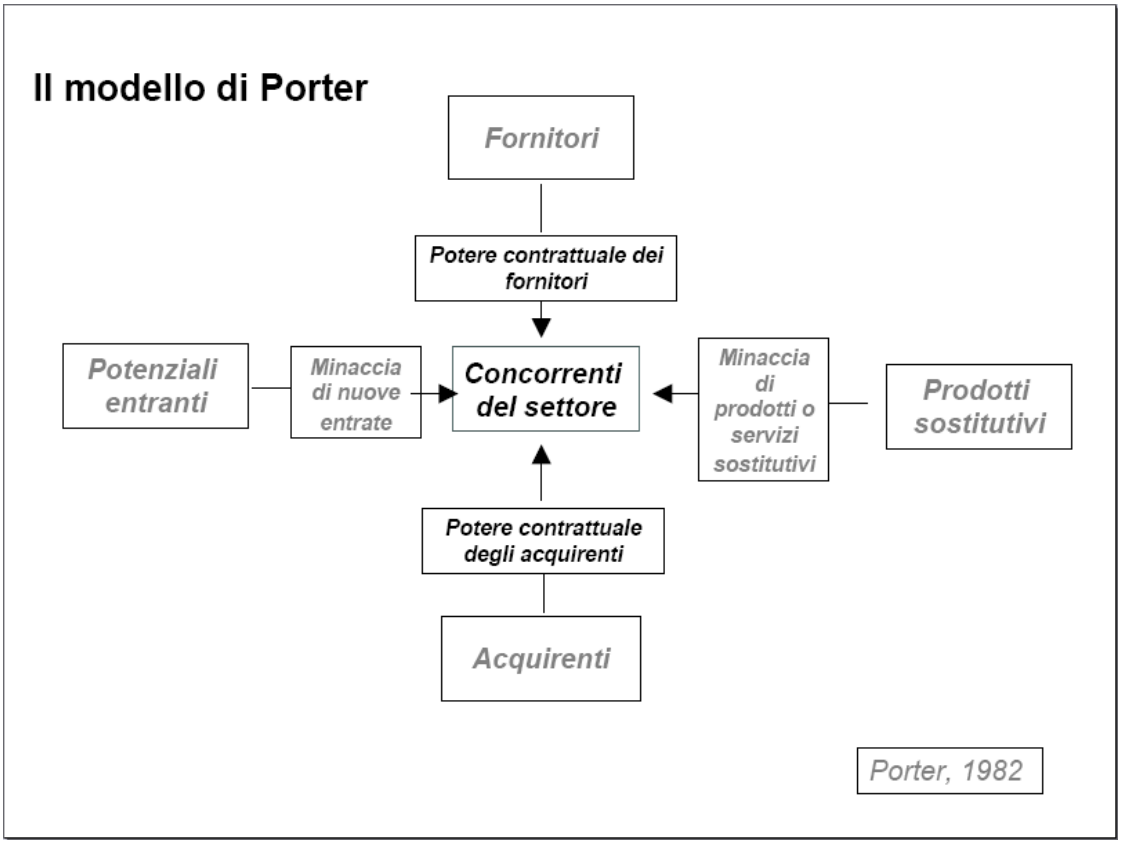

Figura 1: Le cinque forze competitive di Porter[9]

Nel modello di Porter vengono descritte le cinque forze che determinano la profittabilità di un'impresa, che non è pertanto funzione del solo prodotto, ma della struttura industriale nel suo complesso. Queste sono in grado di influenzare prezzi, costi e investimenti necessari.
Una questione fondamentale per le strategie competitive è, appunto, il posizionamento relativo all'interno dell'industria: il posizionamento determina se la performance di un'impresa è al di sotto o al di sopra della media di settore. Un requisito fondamentale per ottenere una *above-average performance* nel lungo periodo è quella di avere un vantaggio competitivo sostenibile: uno dei modi per ottenerlo è la differenziazione, che deriva dall'abilità dell'impresa di far fronte alle cinque forze in maniera migliore rispetto alla concorrenza.

[9] https://commons.wikimedia.org/w/index.php?curid=9402891

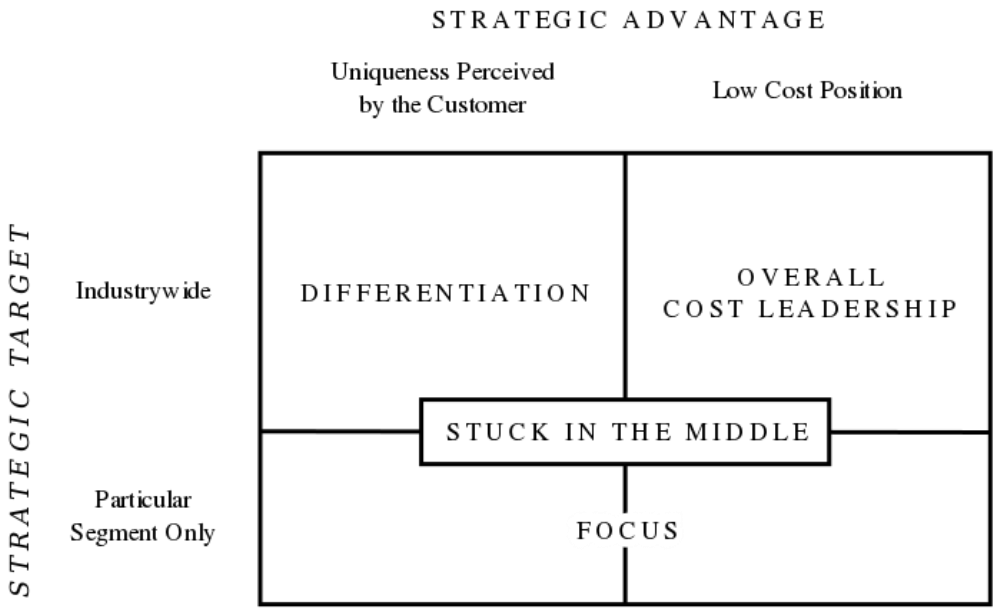

Figura 2: Le tre strategie di base di Porter[10]

«La seconda strategia di base è la differenziazione. In questo caso un'azienda mira a essere unica nel proprio settore industriale in rapporto ad alcune variabili ritenute molto importanti dai clienti. Essa sceglie una o più caratteristiche che sono percepite come importanti da molti clienti in un settore, e si mette nelle condizioni di soddisfare quei bisogni in modo ineguagliabile (come per esempio la Ferrari, grandi firme di moda). Tale unicità viene compensata con prezzi superiori alla media»[11].

La strategia di differenziazione, se vincente, genera un vantaggio competitivo per la compagnia che la mette in atto, dal momento in cui al prodotto viene riconosciuto un *premium price*. Il concetto di differenziazione è un concetto che ruota, quindi, attorno a quello di qualità che il consumatore attribuisce ad un determinato prodotto; per questo motivo, nel corso di questo lavoro, i prodotti differenziati verranno definiti anche come prodotti "*quality"*.

2.2.2. Differenziazione verticale e orizzontale

In una struttura di mercato in cui sono disponibili prodotti differenziati il beneficio dei consumatori non è valutato non solo in termini di prezzi, ma anche di varietà di prodotto.

10 https://upload.wikimedia.org/wikipedia/commons/0/0b/Michael_Porter%27s_Three_Generic_Strategies.svg

11 Porter, M.E. *"Competitive Advantage: Creating and Sustaining Superior Performance", The Free Press* (1985)

Nella precedente sezione sono state prese in esame le NTCI; secondo Lancaster[12], la differenziazione è di due tipi:

- Orizzontale, i prodotti hanno lo stesso livello qualitativo ma caratteristiche – reali o presunte – che determinano un diverso ordinamento delle preferenze da parte dei consumatori. I prodotti sono collocati in uno spazio delle caratteristiche secondo preferenze soggettive e non è possibile classificarli in base a criteri oggettivi. Quanto più si distanziano, maggiore è la differenziazione e quindi minore la sostituibilità.
- Verticale, i prodotti differiscono per livello qualitativo. Un bene è classificato in base a tutto il vettore delle caratteristiche secondo criteri oggettivi, la scelta è quindi determinata dal livello del reddito del consumatore e non dalle preferenze soggettive.

Nella definizione del commercio intra-industriale vengono catturate categorie differenti:

- Commercio orizzontale di prodotti simili ma con varietà differenti, si riferisce al commercio simultaneo di beni classificati nello stesso settore e allo stesso stadio di produzione (ad esempio, la Corea del Sud che importa ed esporta telefoni allo stadio finale del processo produttivo). Dal momento in cui sono prodotti con tecnologia simile e incorporano funzioni simili, sono classificati nello stesso settore, anche se appaiono differenti e le diverse marche soddisfano i desideri dei diversi consumatori in modo differente.
- Commercio verticale di prodotti distinguibili sia per qualità che per prezzo, si riferisce al commercio simultaneo di beni nello stesso settore ma in diversi stadi del processo produttivo, tipico della frammentazione della produzione in cui ogni stadio avviene in locazioni differenti (ad esempio, la Cina importa componenti di computer e li assembla per poi esportarli).

2.2.3. Le preferenze individuali: *Love for variety* e *Ideal variety*

«Le economie coinvolte negli scambi possiedono una struttura industriale caratterizzata da gruppi di prodotti differenziati: i ‘gruppi’ consistono in una classe di prodotto

[12] Lancaster, K.J. *“Intra-Industry Trade under Perfect Monopolistic Competition.” Journal of International Economics, 10, 1980, 151-75.*

in cui tutti i prodotti possiedono le stesse caratteristiche e i diversi prodotti all'interno del gruppo hanno tali caratteristiche in diversa proporzione.»[13].

Così Lancaster definisce la differenziazione di prodotto, e ciascuna specificazione di un bene dipende dalle differenti proporzioni delle caratteristiche. Le specificazioni sono potenzialmente infinite all'interno di ciascun gruppo.
In riferimento a tale tipologia di beni, nella teoria economica esistono due approcci per rappresentare le preferenze del consumatore in relazione alle diverse specificazioni:

- *Ideal variety,* si assume che ogni consumatore abbia, per ciascun prodotto, una propria varietà ideale e cioè una specifica proporzione di caratteristiche. L'individuo può scegliere tra tutte le specificazioni disponibili all'interno del gruppo ma scegliendo un solo prodotto, dal momento che non sono combinabili; nell'effettuare le proprie scelte, dati i prezzi e il reddito, selezionerà la varietà più prossima alla sua ideale.
- *Love for variety,* si assume che le preferenze di ogni consumatore siano definite sull'insieme di tutte le varietà disponibili di un determinato prodotto (e quindi all'interno del gruppo) e che il consumatore, dal momento in cui possiede una ideal variety che può non essere presente sul mercato, nel massimizzare la propria soddisfazione è portato a domandare quante più varietà possibili nell'ottica, appunto, dell'amore per la varietà.
-

2.3. Le commodity

Nel quadro dell'analisi è necessario definire anche la seconda categoria di beni che saranno oggetto di studio: i beni commodity, o indifferenziati, sono beni omogenei caratterizzati da quasi perfetta – o perfetta – sostituibilità. Nell'ambito della differenziazione definita come marketplace, la classe delle commodity è caratterizzata dall'uguaglianza di tutte le alternative e della percezione di esse nel vettore di prezzi e caratteristiche, sia fisiche che non.
È definita come tipologia di bene fungibile, ossia avente caratteristiche identiche che prescindono dal produttore. Questi prodotti sono caratterizzati da elevata standardizzazione e possono essere prodotti di base non lavorati, merci finite o altre tipologie di prodotti/servizi la

[13] Lancaster, K.J. *"Intra-Industry Trade under Perfect Monopolistic Competition." Journal of International Economics,* 10, 1980, 151-75.

cui tecnologia diventa pubblica in seguito alla scadenza del brevetto o altre questioni normative; questo processo è definito come *commoditization*.

2.3.2. La *legge del prezzo unico*

Come detto, i beni omogenei sono prodotti altamente standardizzati, come ad esempio le materie prime energetiche, i metalli naturali o beni di consumo come zucchero e caffè. Secondo la legge del prezzo unico, in condizioni di concorrenza, assenza di costi di trasporto e altre barriere doganali, beni identici *devono* avere lo stesso prezzo in qualsiasi mercato. L'elevata standardizzazione di questi beni ne consente un'agile negoziazione sui mercati internazionali, principalmente come attività sottostante di particolari strumenti derivati, i futures, nei seguenti mercati: New York Mercantile Exchange (NYMEX), Chicaco Board of Trade (CBOT), Intercontinental Exchange (ICE), Chicago Mercantile Exchange (CME), London Metal Exchange (LME) e New York Board of Trade (NYBOT). Il loro prezzo, quotato in borsa, è determinato dal mercato e coincide con quello di equilibrio. In questo caso, la legge del prezzo unico è certamente valida e assicurata dalle eventuali opportunità di arbitraggio, il cui meccanismo ha come risultato l'eliminazione delle differenze di prezzo.

2.4. La differenziazione sui mercati esteri

2.4.1. Vantaggio competitivo e premium price

Come anticipato, per le imprese è fondamentale individuare la strategia competitiva che ne consenta la sopravvivenza sul mercato. Nella fase di inserimento nei mercati è quindi di notevole rilevanza indagare sulla natura di un prodotto, così da attuare la strategia più consona. Il potenziale vantaggio competitivo ottenuto da chi attua la differenziazione è infatti riconosciuto dal mercato di destinazione sotto forma di disponibilità a pagare un *premium price*: a differenza delle commodity, quindi, la differenza qualitativa tra beni simili si riflette sul prezzo degli stessi e viene perciò meno la legge del prezzo unico. Nel caso di beni differenziabili, l'unicità del prodotto fa sì che l'acquirente sia disposto a pagare un prezzo maggiorato, in ragione del fatto che il bene meglio soddisfa le sue esigenze (ossia, somiglia di più alla *ideal variety.* La differenziazione si configura quindi come strategicamente rilevante

nell'approccio ai mercati esteri, in quanto attraverso questo differenziale di prezzo l'impresa può raggiungere la situazione definita come *above-average performance.*

2.4.2. La dispersione di prezzo

Data questa premessa, quindi, è naturale che nel mercato dei beni differenziati prodotti appartenenti alla stessa categoria vengano venduti a prezzi molto diversi. Tale differenza può essere espressa dalla dispersione di prezzo, che cattura quindi differenze qualitative misurabili e non. Di conseguenza, la presenza sul mercato di dispersione dei prezzi può essere considerata un segnale di differenziabilità del bene.

2.4.3. I fallimenti di mercato

Si potrebbe erroneamente pensare di considerare la dispersione come unico indicatore di differenziabilità; questo potrebbe essere fuorviante, in quanto all'interno di un così complesso sistema di mercato si rilevano delle inefficienze che alterano il funzionamento della legge del prezzo unico.

In determinati casi la dispersione si configura come sintomo di fallimenti di mercato, quali:

1. **Asimmetrie informative**, ossia una situazione in cui le parti dispongono di informazioni differenti. Infatti, è possibile che diversi acquirenti abbiano informazioni diverse riguardo il prezzo di un bene e siano quindi portati a pagare prezzi diversi. Se i consumatori concentrano la ricerca su pochi offerenti, o siti web, acquisiscono diversi livelli di informazione.
2. **Costi di ricerca**, che possono essere molto significativi nel processo di ricerca del prezzo più basso. Il desiderio di un acquirente di conoscere tutti i prezzi praticati sul mercato prima di attuare la propria strategia di acquisto è utopico e richiederebbe un notevole dispendio di risorse e tempo. Baye et al.[14] mostrano che anche in presenza di piattaforme digitali di comparazione dei prezzi, sussistono situazioni in cui il prezzo praticato può differire per lo stesso prodotto perché i consumatori sono sensibili al

[14] Baye et al., *Price Dispersion in the Small and in the Large: Evidence from an Internet Price Comparison Site* (2001)

tempo impiegato nella ricerca. Zhen et al.[15] evidenziano che i costi possono essere particolarmente elevati nei giorni feriali, durante i quali la dispersione è di conseguenza più alta. In maniera quasi paradossale, inoltre, la concorrenza tra piattaforme di ricerca dei prezzi può incrementare la dispersione perché aumentano i costi di ricerca e comparazione per il buyer. Non in ultimo, il buyer può incorrere nel fenomeno di cosiddetto "platform lock-in", per il quale si trovano chiusi in una sola piattaforma per questioni legate a pubblicità, sostegni finanziari, servizi aggiuntivi etc.

3. **Concentrazione del mercato,** all'aumentare della quale aumenta anche la dispersione tra i prezzi: l'aumento del numero di venditori e la maggiore concorrenza riducono la capacità dei venditori di monopolizzare il prezzo di mercato, facendo sì che i venditori offrano prezzi più simili al costo marginale. La dimensione della dispersione dei prezzi è quindi negativamente correlata al numero di venditori.

[15] Zen et al., *'Law of One Price' in the Internet Era: Search Cost, Platform Competition and Customer Lock-in*

Capitolo 3

3. Descrizione dei dati

Per il presente lavoro sono state utilizzate due diverse banche dati disponibili in StudiaBo, la cui descrizione è di seguito riportata.

3.1. Banche dati e codici doganali

Ai fini dell'analisi vengono utilizzati i dati disponibili nel tool Analytics di ExportPlanning. Questo tool sui mercati esteri contiene diverse tipologie di Datamart[16], quelli utilizzati sono:

1) Datamart Ulisse di Commercio internazionale Annuale, contenente i dati annuali storici dei flussi di commercio estero con scomposizione per fascia di prezzo elaborati da StudiaBo. Contiene dati a partire dal 1995.
2) Datamart Congiuntura paesi UE di Commercio internazionale Trimestrale, contenente i dati trimestrali sui flussi di commercio estero dichiarati dalle imprese appartenenti ai paesi dell'Unione Europea. Contiene dati a partire dal 2000.

L'analisi verrà condotta sulle due banche dati in modo da poter effettuare dei confronti

La Classificazione Prodotti Ulisse è stata sviluppata a diversi livelli di aggregazione:

- **UL20** è il livello più aggregato, corrispondente al concetto di sistema;
- **UL200** è un livello di aggregazione intermedio, corrispondente al concetto di Industria/Settore;
- **UL3000** è il livello più disaggregato, corrispondente al concetto di prodotto omogeneo.

Il livello **UL20** è composto dalle seguenti voci:

A1: Materie prime naturali

A2: Materie prime industriali

B1: Beni alimentari intermedi e finali non confezionati

B2: Beni intermedi in materie tessili e pelli

[16] Con Datamart si indica un archivio di dati omogenei.

B3: Beni intermedi in carta e in legno

B4: Beni intermedi in metallo

B5: Beni intermedi chimici

B6: Beni intermedi in minerali non metalliferi

C1: Beni e prodotti per le costruzioni

D1: Componenti elettroniche

D2: Componenti meccaniche ed ottiche

D3: Componenti per i mezzi di trasporto

D4: Elettrotecnica

E0: Alimentari confezionati e bevande

E1: Prodotti finiti di largo consumo

E2: Prodotti finiti per la persona

E3: Prodotti finiti per la casa

E4: Prodotti e strumenti per la salute

F1: Strumenti e attrezzature per ICT e servizi

F2: Strumenti e attrezzature per l'industria

F3: Mezzi di trasporto e per l'agricoltura

F4: Macchine e impianti per i processi industriali

F5: Impiantistica industriale

G1: Armi e munizioni

Z9: Dati confidenziali

Le elaborazioni e le analisi verranno effettuate sul livello UL3000, con i corrispondenti codici HS a 6 cifre, mentre i risultati finali verranno riportati a livello di Sistema, la cui classificazione è stata appena elencata. Per comodità espositiva i sistemi verranno definiti Settori Industriali.

3.1.1. La banca dati Ulisse

I dati di esportazioni e di importazioni dichiarati dai diversi paesi mondiali rappresentano un database ricco di contenuto informativo, grazie ai quali è possibile avere informazioni sulla dinamica dei diversi mercati ad un elevato livello di dettaglio conoscendo il prezzo a cui viene venduto un dato bene su uno specifico mercato e soprattutto se quel mercato accetta o meno di pagare un *premium price* per la maggior qualità offerta da un venditore. Tuttavia, spesso queste informazioni non sono immediatamente fruibili da una semplice lettura dei dati, poiché le dichiarazioni dei diversi paesi possono contenere dati che non corrispondono ai flussi effettivi di commercio con l'estero per alcuni principali motivi:

- Procedure di stima. I dati possono non essere raccolti nel momento in cui le merci attraversano le frontiere doganali, ma essere frutto di una stima, come nel caso del commercio tra paesi UE i cui valori sono stimati dagli istituti di statistica sulla base delle dichiarazioni Iva effettuate dalle varie imprese.
- Classificazione merceologica. Non tutti i paesi usano la stessa classificazione, quindi i flussi relativi a prodotti simili possono essere attribuiti a codici merceologici diversi. Questa problematica è stata in parte ovviata dal lavoro svolto dal World Custom Organization che ha sviluppato la classificazione Harmonized System, usata ormai da quasi 150 paesi.
- Confidenzialità. Talvolta i flussi di commercio estero per paese e prodotto sono ritenuti confidenziali e sono quindi segretati, questo incide maggiormente sui paesi più piccoli.
- Ritardi. I paesi con una struttura amministrativa meno consolidata tendono a produrre le statistiche sul commercio con ritardo. L'Onu è impegnato nel finanziamento di progetti che hanno l'obiettivo di migliorare le procedure di raccolta, elaborazione e pubblicazione dei dati di commercio estero.

Di conseguenza, per poter estrarre informazioni significative ed affidabili dai dati di base è necessario utilizzare degli strumenti metodologici che distinguano la misura del fenomeno dal rumore statistico, sfruttando sia i vantaggi insiti nei dati (elevata numerosità, doppia dichiarazione, dati riferiti alla popolazione), sia le metodologie di data mining che riguardano:

- **Outlier**, per l'individuazione di errori di misura e sostituzione con valori più coerenti
- **Dati mancanti**, il cui valore è inserito con una stima coerente con le informazioni nella banca dati

- **Normalizzazione**, elaborazione che consente di ottenere per ogni variabile una misurazione univoca e quindi confrontabile
- **Congiuntura**, consente di produrre una pre-stima di un dato annuale, a partire dalle informazioni congiunturali
- **Fasce di qualità/prezzo**, consente di valutare se il bene può essere oggetto di offerta differenziata. StudiaBo attribuisce una specifica fascia qualitativa o di prezzo per ciascun flusso.
- **Quantità a prezzi costanti**, dato che le rilevazioni riguardano valori e unità di misura fisica, dal loro rapporto si ottiene il valore medio unitario, spesso identificato con il prezzo. Tale misura incorpora le variazioni qualitative a cui il prodotto viene commercializzato. StudiaBo ha costruito la variabile quantità a prezzi costanti in modo che includa l'effetto qualità: gli indicatori di prezzo che derivano dall'uso di questa misura possono essere considerati più attendibili rispetto ai VMU, perché non inficiati da variazioni qualitative del prodotto.

Con tali metodologie si è prodotta una banca dati altamente informativa sulle caratteristiche del fenomeno Commercio Estero.

La banca dati Ulisse è stata costruita a partire dalle informazioni disponibili da diverse fonti di analisi economica, in primis la Divisione Statistica delle Nazioni Unite, che aggiorna il database Comtrade. UN Comtrade riunisce le dichiarazioni annuali di più di 170 paesi verso i propri partner commerciali, il cui livello di dettaglio a livello di codice prodotto in classificazione merceologica è HS a 6 cifre (Harmonized System è, quindi, la classificazione doganale di riferimento). Stima di riuscire a esplicitare più del 95% del commercio mondiale, riportando il valore in dollari e la quantità commercializzata, espressa in kg e/o in unità di misura supplementare. Oltre a questo database, la banca dati Ulisse è periodicamente aggiornata con informazioni congiunturali provenienti da altre fonti:

- Banca dati Comext sul commercio con l'estero mensile dei paesi UE ed Efta, prodotta da Eurostat
- Banca dati Comtrade a livello mensile
- Banca dati UsaTrade sul commercio con l'estero mensile degli Stati Uniti, prodotta dall'U.S. Censis Bureau con una classificazione HS a 10 digits

3.1.2. La banca dati Comext

I dati inerenti export e import dichiarati dai paesi UE rappresentano un database altamente informativo, grazie anche al frequente aggiornamento e il minuzioso dettaglio merceologico. Anche in questo caso, ma per motivi differenti, i dati sono stati elaborati attraverso una specifica metodologia per sfruttarne il contenuto informativo. I principali motivi per cui i dati non sono immediatamente fruibili sono:

- **Procedura di stima,** con il Mercato Unico Europeo e la rimozione della dogana i dati sul commercio vengono raccolti dagli istituti di statistica attraverso la modalità Intrastat basata sulle dichiarazioni Iva. Generalmente, queste procedure di stima sono tuttavia di elevata affidabilità
- **Confidenzialità,** a causa di cui alcuni flussi sono segretati per salvaguardare le strategie competitive dei diversi paesi. La quota di flusso segretato può arrivare a superare il 5% dell'export.
- **Rotture delle serie storiche,** la Nomenclatura Combinata a 8 digits (NC8) è sottoposta a revisione ogni anno e questi frequenti cambiamenti rendono difficile la lettura dei dati storici a causa di possibili rotture temporali nel passaggio da una revisione all'altra. Questo è il più grande ostacolo alla fruizione efficace dei dati di commercio di fonte Eurostat.

Nuovamente, per poter estrarre informazioni significative ed affidabili dai dati di base è necessario utilizzare degli strumenti metodologici che distinguano la misura del fenomeno dal rumore statistico, sfruttando i vantaggi insiti nei dati (elevata numerosità, dati riferiti alla popolazione), che consentono a loro volta di utilizzare efficacemente alcune metodologie di data mining, quali:

- **Costruzione delle serie storiche,** attraverso uno strumento informativo costruito da StudiaBo che è in grado di ricostruire l'evoluzione del commercio con l'estero dei diversi prodotti.
- **Armonizzazione,** per l'eventuale rettifica dei dati più recenti che sono passibili di modifiche e revisioni in quanto non ancora certificati da Eurostat e Stati Membri. Le imprese europee, infatti, forniscono informazioni con tempistiche diverse, a seconda della loro dimensione.

- **Outlier,** i dati riguardanti quantità fisiche sono trattati al fine di individuare eventuali errori di misura ed essere quindi rimpiazzati con valori più coerenti al resto delle osservazioni.
- **Pre stime,** StudiaBo si avvale di modelli econometrici ARMA per effettuare pre stime del trimestre in corso.

Tali strumenti consentono di produrre una banca dati altamente informativa del fenomeno Commercio Estero UE.

La banca dati Congiuntura paesi UE contiene quindi i dati che Eurostat Comext riporta per i paesi UE e i 4 paesi Efta[17] sulle dichiarazioni di commercio estero mensili, ossia i flussi verso gli oltre 150 paesi mondiali della classificazione Ulisse. I membri UE utilizzano una classificazione più dettagliata rispetto a quella armonizzata: la Nomenclatura Combinata a 8 digits (CN8), che consiste in una specifica del Harmonized System e ha validità annuale. Per ogni flusso mensile, Comext riporta il valore in euro e la quantità commercializzata espressa in kg e/o unità di misura supplementare. I dati vengono aggiornati mensilmente a scadenze fisse, con un ritardo di 6-10 settimane rispetto al mese concluso. I flussi dei paesi Efta sono invece dichiarati in classificazione HS a 6 digits.

3.2. La classificazione StudiaBo

StudiaBo dispone di una classificazione proprietaria dei prodotti, effettuata manualmente sfruttando le conoscenze del mercato, per i prodotti più noti, e basandosi su ipotesi per i prodotti meno noti. Non esiste una classificazione generale di riferimento, pertanto si prenderà come confronto quella StudiaBo.

[17] Islanda, Liechtenstein, Norvegia e Svizzera.

Capitolo 4

4. Definizione features

Nel seguente capitolo è descritta la metodologia utilizzata sia per la creazione delle strutture dati che per la definizione delle features[18], per le quali le strutture sono base necessaria. Questo consiste nell'effettuare una rielaborazione delle banche dati esistenti così da avere a disposizione quanto necessario al calcolo delle variabili. Inoltre, la rielaborazione consente di renderle omogenee nel contenuto: in questo modo le features saranno confrontabili.

Ai fini dell'analisi viene utilizzato il linguaggio Python.

La scelta delle features da definire è frutto di ragionamenti basati sullo studio della letteratura di riferimento. È infatti possibile classificarle in due categorie:

- Features di prezzo: dispersione interquartilica e correlazione tra i prezzi del primo e del terzo quartile.
- Features di mercato: indice di concentrazione di mercato e indice di commercio intra-settoriale.

Nel corso del capitolo ne verrà spiegata la ratio economica.

4.1. Strutture dati

Il primo passo prima di procedere alla definizione delle features e, in seguito, alla classificazione tramite algoritmi di machine learning, è quello di rendere le banche dati omogenee.

A partire dalle banche dati Ulisse e Comext, descritte nel precedente capitolo, vengono definite due strutture dati differenti: la prima verrà utilizzata per calcolare i quantili e ottenere quindi le features di dispersione e correlazione, la seconda sarà invece strumentale alla definizione delle features di concentrazione e commercio intra-settoriale.

[18] Nell'ambito del machine learning, le features sono le variabili.

4.1.1. Banca dati Ulisse

Il contenuto originario delle banche dati è descritto nel precedente capitolo.
La banca dati Ulisse è stata ripulita dai valori del sistema Z9[19], inutili ai fini dell'analisi.
I dati sono stati raggruppati per anno, paese esportatore e paese importatore, sommando sia rispetto ai valori monetari delle esportazioni e sia alle quantità.
La prima colonna definita è PERIOD, che contiene l'anno a cui sono riferiti i valori delle altre colonne.
Sono poi stati selezionate dalla colonna dei valori V solo le esportazioni, in modo da evitare le ripetizioni con le importazioni mirror[20].
A questo punto è possibile definire la colonna dei prezzi, PK, i cui valori sono ottenuti rapportando il valore monetario del flusso e la quantità; questo corrisponde al Valore Medio Unitario (VMU), espresso in valuta USD. Una ulteriore colonna ITEM è costruita in modo da contenere i nomi dei due paesi coinvolti nello scambio: il primo corrisponde al paese esportatore, il secondo al paese importatore.
La prima struttura dati è quindi composta dalle quattro colonne PERIOD, ITEM, V e PK.

Per la seconda struttura dati, i valori vengono raggruppati per anno, paese esportatore e paese importatore e sommati rispetto ai valori monetari delle esportazioni e delle importazioni. Vengono definite le colonne PERIOD, PAE, VX e VM che corrispondono rispettivamente all'anno, al paese esportatore/importatore, al valore delle esportazioni e al valore delle importazioni. Non è necessario ricavare i prezzi.
La seconda struttura dati è quindi composta dalle quattro colonne PERIOD, PAE, VX e VM.

4.1.2. Banca dati Comext

Si vogliono ottenere le stesse strutture dati a partire dalla banca dati Comext, che contiene informazioni differenti poiché non sono espressi in termini di paese esportatore e paese importatore, ma di paese dichiarante e paese partner. Perciò, il dichiarante si configura rispettivamente come esportatore nel caso delle esportazioni e come importatore nel caso delle importazioni.

[19] Il contenuto è descritto nel Capitolo 3
[20] Nella colonna V le esportazioni dal Paese A al Paese B corrispondono alle importazioni del Paese B dal Paese A: sono i dati mirror, si vogliono pertanto evitare ripetizioni.

Essendo i dati trimestrali, devono prima essere aggregati per anno, sommando quindi i valori dei quattro trimestri.
Viene quindi nuovamente ottenuta la colonna dei prezzi, che sono tuttavia espressi in valuta Euro, così come i valori monetari. Perciò, vengono convertiti in valuta USD al tasso di cambio medio pari a 1.15[21]. Ancora, la colonna ITEM contiene prima il paese esportatore seguito dal paese importatore. Come in precedenza, la prima struttura dati è quindi composta dalle quattro colonne PERIOD, ITEM, V e PK.

Per la seconda struttura dati vengono selezionati rispettivamente i valori delle esportazioni e delle importazioni per ottenere le colonne VX e VM, in termini di valore e non di prezzo. La colonna PAE contiene l'elenco univoco dei paesi a cui sono associati i rispettivi valori esportati e importati. I dati vengono raggruppati per periodo e paese, sommando i valori di export e import. Come in precedenza, la seconda struttura dati è quindi composta dalle quattro colonne PERIOD, PAE, VX e VM.

A partire da due banche dati differenti si ottengono, quindi, due strutture dati omogenee. È a questo punto possibile procedere con la definizione delle features. Di seguito verrà brevemente esposta la ratio economica alla base della scelta delle variabili congiuntamente alla metodologia utilizzata per definirle.

4.2. Indice di dispersione dei prezzi

Nella fase di inserimento nei mercati esteri, è di notevole rilevanza indagare sulla natura di un prodotto, cercando di individuare le i driver che consentono di distinguere un prodotto differenziabile da una commodity.
Come anticipato nel Capitolo 2, infatti, la differenziazione è una strategia che consente di offrire un bene che, per le sue caratteristiche, può essere venduto ad un premium price. Nel mercato dei beni differenziati è naturale che prodotti appartenenti alla stessa categoria vengano venduti a prezzi molto differenti, presentando quindi dispersione dei prezzi. Quindi, se nello stesso mercato e nello stesso periodo sono importati beni merceologicamente omogenei, ma con prezzi diversi, la dispersione di prezzo può essere considerata un indicatore di differenziabilità. Ad esempio, si prendano in considerazione i seguenti grafici, che riportano il prezzo del primo, secondo e terzo quartile di alcuni prodotti[22]:

[21] Sono presenti dati precedenti all'entrata in vigore dell'Euro, per cui la conversione è approssimativa.
[22] L'elenco dei prodotti presi in esame e gli ulteriori grafici sono riportati in Appendice 1

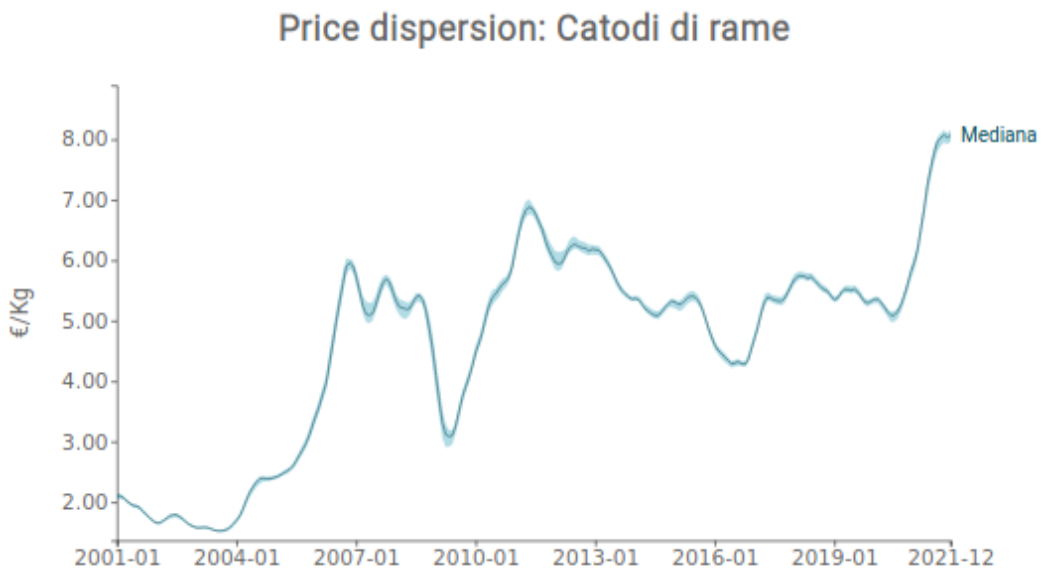

Figura 3. HS740311 - Fonte: Elaborazioni ExportPlanning

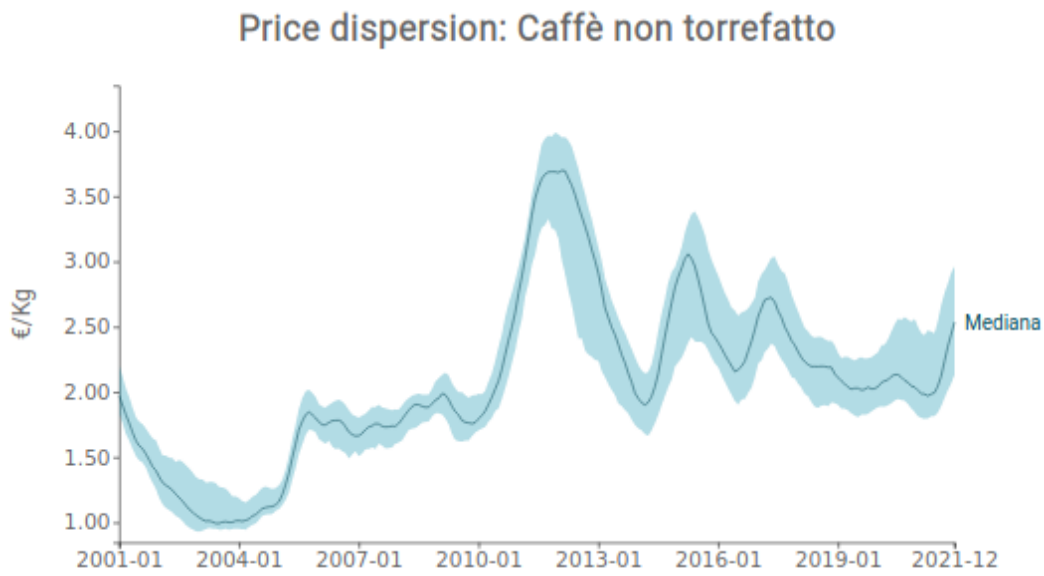

Figura 4. HS090111 - Fonte: Elaborazioni ExportPlanning

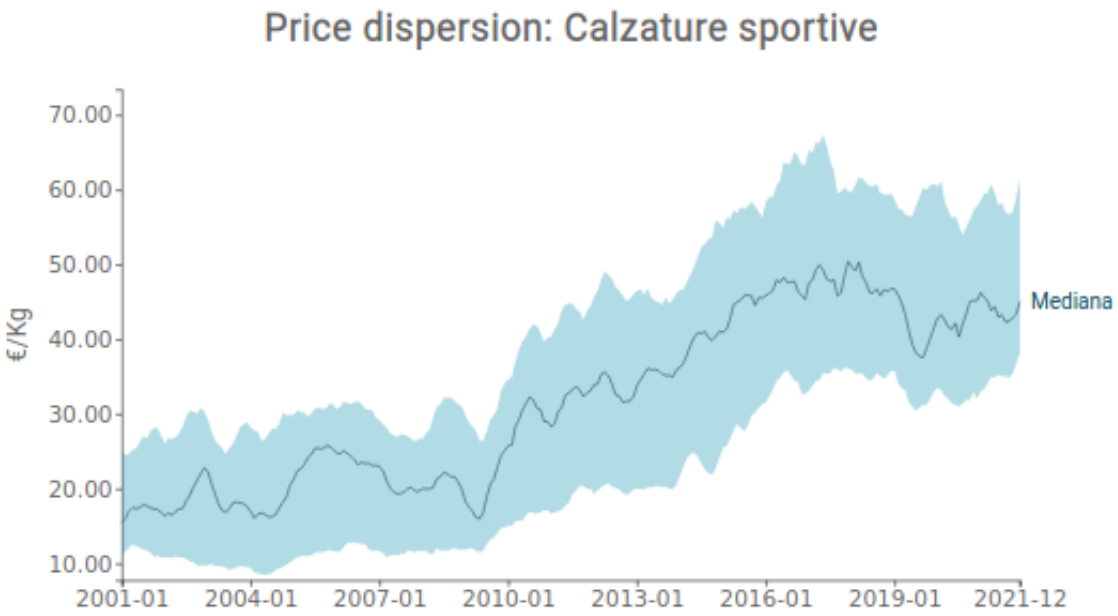

Figura 5. HS640219 - Fonte: Elaborazioni ExportPlanning

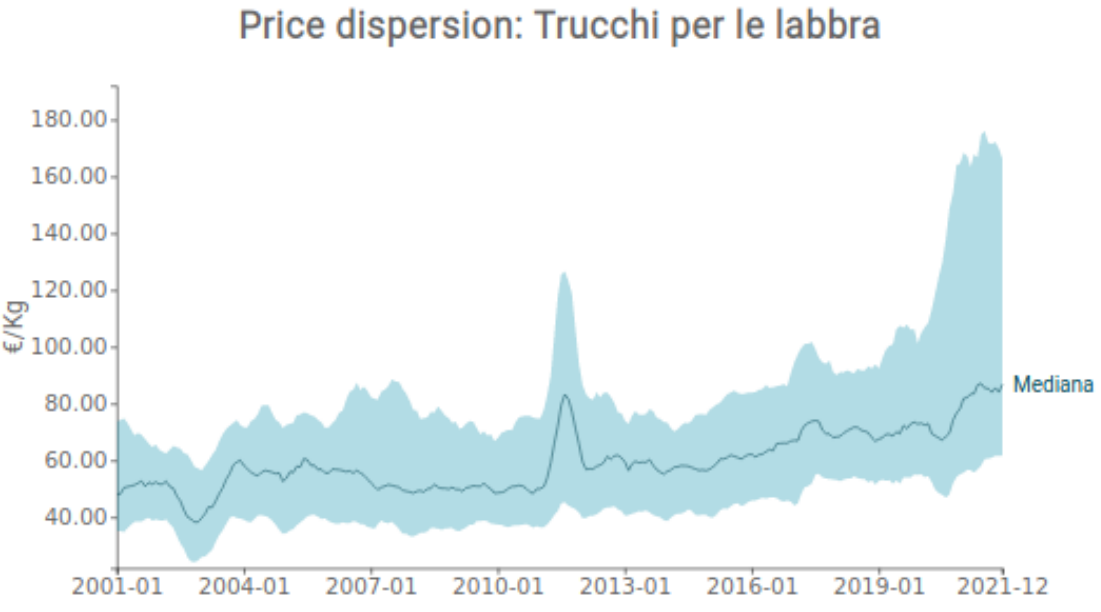

Figura 6. HS330410 - Fonte: Elaborazioni ExportPlanning

L'area colorata corrisponde alla dispersione tra i prezzi del I e del III quartile (rappresentati rispettivamente dal perimetro inferiore e superiore dell'area stessa). La linea più scura rappresenta invece i prezzi del II quartile, corrispondenti alla mediana.
Come mostrato in *Figura 3,* i catodi di rame sono chiaramente non differenziabili in quanto presentano una dispersione di prezzo praticamente nulla.
Passando alla *Figura 4* invece, si riscontra una dispersione di prezzo leggermente più elevata, pur se contenuta. Sappiamo che il caffè è certamente una commodity, dal momento in cui è quotato alla NYMEX; se non fossimo a disposizione di questa informazione, sarebbe necessario basarsi solo sugli indicatori di differenziabilità per discriminare la natura del prodotto.
La *Figura 5* e la *Figura 6* mostrano due prodotti con una elevata dispersione di prezzo: questo è un primo e chiaro segnale di differenziabilità, tuttavia non incontrovertibile: sarà necessaria l'analisi di altri indicatori. Dal momento in cui la dispersione è molto più elevata nel caso di beni differenziabili, si potrebbe allora pensare di poterla considerare come unico indicatore di differenziabilità; questo potrebbe essere fuorviante, in quanto a cause delle inefficienze di mercato si rende necessaria l'analisi di altri fattori discriminanti.
Dal punto di vista applicativo la variabile di dispersione dei prezzi, denominata range, è ottenuta a partire dalla prima struttura dati di entrambi i database. Dal momento in cui le strutture dati sono state rese omogenee, la metodologia è identica sia con i dati Ulisse che con i dati Comext.
Per il calcolo di questa feature viene considerato il periodo dal 2017 al 2019[23] incluso.
In primis, si calcolano i quantili: è perciò necessario ottenere il peso percentuale relativo di ciascun flusso, rapportandolo al valore totale del commercio. Viene poi ottenuta la distribuzione dei flussi, a partire dalla quale si possono ricavare i quartili.
La dispersione è infine calcolata come scarto interquartilico: contiene il 50% dei valori della distribuzione ed esprime quanto i prezzi si allontanino dal loro valore mediano. Il range è espresso in termini percentuali.

4.3. Indice di correlazione

Come accennato in precedenza, l'indice di dispersione non è sufficiente come unica valutazione per la natura di un prodotto.

[23] Sono disponibili dati più recenti, non utilizzati in quanto riferiti al periodo della pandemia da Covid-19.

Un'ulteriore analisi può riguardare la misura dell'intensità della correlazione tra i prezzi di fascia alta e quelli di fascia bassa, per un dato prodotto. Si tratta quindi di verificare se, e con quale forza, i movimenti dei prezzi del primo e del terzo quartile siano tra loro legati.
L'ipotesi di partenza è che se a causa di inefficienze di mercato si verifichi dispersione di prezzo anche sui mercati delle commodity, i prezzi di queste ultime siano comunque legati da un'alta correlazione. In caso il bene sia differenziabile, invece, ad un aumento dei prezzi del terzo quartile non necessariamente ne corrisponderebbe quello dei prezzi del primo, in quanto le differenze di prezzo sono legate alla natura qualitativa del bene piuttosto che ad una inefficienza del mercato.
Analizzando nuovamente i grafici in *Figura 3* e in *Figura 4,* a giudicare solo dalla dispersione si potrebbe infatti pensare che il caffè sia un prodotto differenziabile; tuttavia le oscillazioni delle diverse fasce di prezzo sono perfettamente concordi, segnalando un indice di correlazione molto elevato. Nei casi in *Figura 5* e *Figura 6,* invece, in alcuni periodi all'aumento (o diminuzione) dei prezzi di fascia alta non corrisponde l'aumento (o diminuzione) dei prezzi di fascia bassa.
Quindi, il fatto che per un determinato prodotto sia presente dispersione di prezzo, non è indicazione incontrovertibile di differenziabilità, in quanto tale dispersione può essere sintomo di inefficienze di mercato che non possono e non devono essere ignorate in fase di valutazione, per scongiurare conclusioni errate.
La variabile di correlazione mira a catturare l'intensità del legame tra i prezzi nel primo quartile e quelli nel terzo quartile. Anch'essa viene quindi ottenuta a partire dalla prima struttura dati, con gli stessi quantili e lo stesso span temporale utilizzati per la precedente feature.
Combinando queste prime informazioni grafiche disponibili sui beni, i primi due risultano non differenziabili: nessun acquirente sarà disposto a pagare un prezzo più elevato di quello quotato. Lo stesso non avviene nel caso delle calzature e dei trucchi, che possono essere considerati differenziabili, almeno in prima analisi.

4.4. Indice di concentrazione di mercato

Le prime due variabili descritte hanno focus sui prezzi. Per un modello più approfondito sono state definite altre due variabili, basate questa volta sui volumi scambiati.

La prima è una misura della concentrazione di mercato: l'ipotesi a priori è che i beni differenziabili siano caratterizzati da bassa concentrazione, mentre i beni omogenei da elevata concentrazione. Al contrario, l'export di prodotti qualitativamente differenziabili sarebbe meno concentrato, per il fatto che tali beni non richiedono specifiche condizioni territoriali, climatiche o di altra natura per la loro produzione e potrebbero potenzialmente essere prodotti da qualsiasi Paese.

La variabile di concentrazione di mercato è calcolata con l'indice di Herfindal[24] (HHI), usato indicatore del grado di competizione nel mercato di ciascun prodotto. Per il calcolo di questa feature vengono utilizzati i dati del 2019[25]. È così calcolato:

$$H = \sum_{i=1}^{N} s_i^2$$

L'indice è ottenuto dalla somma del quadrato del market share percentuale (s_i), ed è stato calcolato sia per le prime 3 (HH3) che per le prime 5 (HH5) nazioni nel commercio mondiale, per cogliere eventuali differenze nella concentrazione tra i paesi top sul mercato. Il market share è dato dal rapporto tra le esportazioni di un paese e le esportazioni totali. Il valore dell'indice indica quindi la quota di mercato detenuta dai primi tre o dai primi cinque paesi nell'export di un determinato prodotto. Il valore dell'indice di Herfindahl varia da 0 a 1. Quando l'indice si avvicina al valore zero (0) indica una concentrazione minima del potere di mercato. Viceversa, quando l'indice si avvicina al valore uno (1) indica una concentrazione massima del potere di mercato. L'indice di Herfindahl considera tutte le imprese operanti sul mercato e, pertanto, l'indicatore di concentrazione non è influenzato dal numero delle imprese. Ciò consente di calcolare in modo univoco un indicatore della concentrazione del potere di merca to indipendentemente dal numero delle imprese che vi operano.

4.5. Indice di commercio intra settoriale

Come anticipato nel Capitolo 2, le teorie tradizionali sul commercio internazionale ipotizzano un flusso di scambi di tipo inter settoriale e, quindi, tra beni merceologicamente differenti ed assimilabili alle commodities. Le nuove teorie parlano invece di commercio intra

[24] Kelly, W.A. Jr, *A Generalized Interpretation of the Herfindal Index* (1981), Southern Economic Journal , Jul., 1981, Vol. 48, No. 1 (Jul., 1981), pp. 50-57

[25] Per le features annuali sono stati selezionati i dati dell'ultimo anno pre-Covid19.

settoriale, perciò tra beni appartenenti allo stesso settore – o, nel nostro caso, stesso codice doganale. Una ulteriore ed ultima feature mira, quindi, a quantificare la quota di commercio intra settoriale per ogni codice prodotto. Ciò equivale a valutare la rilevanza della differenziazione verticale negli scambi internazionali. A questo proposito, nell'analisi empirica viene utilizzato l'indice di Grubel-Lloyd[26], L'indice è così calcolato:

$$GL_i = \frac{(X_i + M_i) - |X_i - M_i|}{X_i + M_i} = 1 - \frac{|X_i - M_i|}{X_i + M_i} \quad ; \; 0 \leq GL_i \leq 1$$

Per il calcolo di questa feature vengono utilizzati i dati del 2019. Nella formula, X_i sono le esportazioni e M_i sono le importazioni. Il valore dell'indice compreso nell'intervallo tra 0 e 1, si è quindi reso necessario definire una soglia oltre la quale un flusso di commercio viene categorizzato come intra-trade: tale soglia è stata fissata a 0.7[27].
Inoltre, poiché è necessario ottenere un unico indice per ciascun codice prodotto (HS), si è ricavato quello aggregato per tutte le nazioni. Perciò, è stata assegnata ad ogni Paese una variabile dummy che assuma valore pari a:

- **1**, se l'indice individuale supera la soglia 0.7
- **0**, se l'indice individuale non supera la soglia 0.7

A questo punto, l'indice generale viene calcolato rapportando:

- Al numeratore, la media del valore dei flussi corrispondenti ai soli paesi il cui indice individuale supera la soglia
- Al denominatore, la media del valore dei flussi di tutti i paesi

L'indice generale è poi trasformato in valore percentuale ed esprime la quota di commercio intra-settoriale del prodotto.
Il dataset finale contiene per riga i codici doganali dei prodotti a cui corrispondono, per colonna, i valori delle diverse features. Le colonne sono, quindi: HS, Range, Correlazione,

[26] Grubel, H. P. Lloyd, P. "*The Empirical Measurement of Intra-Industry Trade*", Economic Record , vol. 47, n°4, 1971, p. 494-517
[27] La soglia è stata definita in maniera arbitraria, ritenendola coerente al fatto che GL_i=1 indica che il commercio del bene in questione è totalmente di tipo interindustriale (cioè X_i=M_i).

HHI3, HHI5 e GL, sia per i dati Ulisse che per i dati Comext. A partire da questo dataset è possibile implementare le procedure di machine learning, che verranno descritte nel prossimo capitolo.

Le variabili appena considerate rappresentano una forte semplificazione di un fenomeno complesso quale il commercio mondiale; molti altri fattori dovrebbero essere considerati nel tentativo di spiegare la differenza qualitativa tra i beni scambiati: forma di mercato, grado di sostituibilità tra i beni, competitività, strategie di prezzo o altri fenomeni congiunturali e strutturali. Inoltre, e non di minor rilevanza, si sottolinea il fatto che i dati utilizzati ai fini dell'analisi sono dati aggregati e i prezzi sono calcolati utilizzando una proxy (i VMU). Tale approccio non considera l'eterogeneità in essere tra le imprese, dal punto di vista dimensionale, di posizionamento, di redditività, di politiche di pricing ecc.

Capitolo 5

5. Algoritmi di Machine Learning e risultati

Il Machine Learning, o apprendimento automatico, nasce dall'idea che i computer possano imparare ad eseguire delle funzioni specifiche, senza programmazione, grazie al riconoscimento di schemi nei dati attraverso algoritmi che imparano in modo iterativo. È una branca dell'intelligenza artificiale che raccoglie metodi di statistica computazionale, riconoscimento di pattern, reti neurali artificiali, filtraggio adattivo e data mining; utilizza metodi statistici per migliorare la performance di un algoritmo nell'identificazione dei pattern nei dati. È fondamentale per il buon funzionamento di tali processi che i dati vengano preparati in modo adeguato, così come è fondamentale la fase di feature creation (ossia la definizione delle variabili).
L'obiettivo del lavoro è quello di sviluppare un algoritmo di Machine Learning che produca la distinzione dei prodotti oggetto di commercio mondiale tra differenziabili e commodity. Gli algoritmi sono i "motori" del Machine Learning e ne esistono due tipologie differenti, la cui differenza consiste nel modo in cui l'algoritmo apprende dai dati per fare le previsioni.
StudiaBo è già in possesso di una classificazione proprietaria dei prodotti, pertanto il presente lavoro ha come scopo quello di disporre di un procedimento automatizzato che, sulla base delle informazioni a disposizione, produca una nuova classificazione di tutti i codici HS. La classificazione esistente verrà inoltre utilizzata come metro di giudizio dei risultati, almeno per i prodotti di cui si ha l'assoluta certezza riguardo la categoria di appartenenza.

5.1. Analisi preliminare: matrici di correlazione

Prima di passare all'analisi vera e propria, è prassi svolgere una valutazione preliminare delle variabili che si di includere nel modello. Vengono di seguito riportate le matrici di correlazione tra le variabili, calcolate a partire dai dataset finali per entrambe le banche dati. Lo scopo di quest'analisi è verificare se la forza e la direzione della relazione tra le variabili sia coerente alle ipotesi sviluppate in precedenza.

	RANGE	CORR	HER3	GRUBEL
RANGE	1	-0.33	-0.20	0.24
CORR	-0.33	1	0.04	-0.05
HER3	-0.20	0.04	1	-0.48
GRUBEL	0.24	-0.05	-0.48	1

Tabella 1: Matrice di correlazione (Dati Ulisse)

	RANGE	CORR	HER3	GRUBEL
RANGE	1	-0.28	-0.13	0.11
CORR	-0.28	1	0.04	-0.02
HER3	-0.13	0.04	1	-0.36
GRUBEL	0.11	-0.02	-0.36	1

Tabella 2: Matrice di correlazione (Dati Comext)

I valori delle correlazioni tra le features sono abbastanza coerenti alle aspettative, quantomeno nel verso, un po' meno nell'intensità. Più nel dettaglio:

- Tra range e correlazione è negativa e pari a -0.28, confermando la relazione inversa tra dispersione del prezzo e correlazione dei prezzi della distribuzione. Considerando il valore medio delle correlazioni, questa è abbastanza elevata.
- Tra range e indice di Herfindal è negativa e pari a -0.13, anche in questo caso confermando le aspettative sulla relazione tra concentrazione di mercato e dispersione del prezzo. Tuttavia, la forza della correlazione è abbastanza ridotta.
- Tra range e indice di Grubel è positiva, sempre coerentemente alle aspettative, ma ugualmente debole.
- Tra gli indici di Herfindal e Grubel è riscontrabile la correlazione più forte, pari a -0.36. Anch'essa conferma l'ipotesi di relazione inversa tra concentrazione di mercato e quota di commercio intra-settoriale.

- Tra correlazione e Herfindal è positiva ma praticamente nulla, mentre tra correlazione e Grubel è negativa ma ugualmente irrisoria.

5.2. Modelli di apprendimento non supervisionato

Come anticipato, esistono due tipologie di algoritmi di Machine Learning differenti, la cui differenza consiste nel modo in cui apprendono dai dati per poi fare le previsioni.
Il primo approccio è quello del Machine Learning non supervisionato: in questa tecnica di apprendimento automatico, il processo elabora le informazioni del database fornito come input in modo indipendente, identificando processi e schemi complessi senza la guida costante dello sviluppatore. La formazione è basata su dati non etichettati e non strutturati: cioè, le classi non sono note a priori ma devono essere apprese automaticamente, confrontando i dati alla ricerca di similarità o differenze; i dati vengono quindi riclassificati ed organizzati sulla base di caratteristiche comuni, così da poter effettuare ragionamenti e previsioni sugli input successivi.

5.2.1. Introduzione al *clustering*

Uno degli algoritmi a disposizione è il *clustering*, implementato in Python con la libreria Scikit-learn, e in particolare il modulo k-Means. Questo algoritmo calcola centroidi[28] e ripete il procedimento in modo iterativo finché non trova il centroide ottimale. Inizialmente, vengono create k partizioni (*cluster*) e vengono assegnati i punti iniziali a ciascun cluster in modo casuale; quindi, calcola il centroide di ogni gruppo e costruisce in seguito una nuova partizione, associando ogni punto in ingresso al gruppo il cui centroide è più vicino ad esso. Vengono infine ricalcolati i centroidi per i nuovi gruppi e così via, finché l'algoritmo non converge.
Con questo metodo, i punti corrispondenti ai dati vengono assegnati ai cluster in modo tale che la somma dei quadrati delle distanze tra i punti e il centroide sia il più piccola possibile.
Il numero di cluster che devono essere generati fa parte dei parametri di input: in questo studio, il numero di cluster (k) è pari a 2. Ad un cluster, saranno assegnati i codici HS6 di tipo commodity e all'altro cluster quelli di tipo differenziabile. È importante sottolineare che, una volta che il procedimento è terminato e i cluster sono stati definiti, è necessario etichettare

[28] In uno spazio euclideo n-dimensionale, il centroide è la posizione media di tutti i punti in tutte le direzioni coordinate.

manualmente i due cluster: il compito della macchina è infatti quello di elaborare una grande mole di dati e dividerli secondo l'algoritmo scelto, ma in ultimo è necessario applicare i ragionamenti affinchè il potenziale informativo dell'algoritmo sia massimizzato.
Per una più accurata analisi, è stato effettuato un tentativo ponendo un numero di cluster pari a 4, al fine di tentare di isolare eventuali outlier o prodotti cosiddetti "intermedi", tuttavia il risultato è stato quello di ottenere 4 cluster di numerosità più o meno simile e contenenti prodotti da tutte le industrie; dal momento in cui non è stato possibile etichettare alcun cluster come outlier, si è preferito tornare alla clusterizzazione semplificata.

5.2.2. Silhouette score e R^2

Il primo è quello di effettuare una selezione preliminare del modello da utilizzare, calcolando il *silhouette score:* questo score è un metodo della libreria Scikit-learn che consente di valutare la performance dell'algoritmo di machine learning non supervisionato. Poiché l'obiettivo dell'algoritmo è quello di creare cluster con osservazioni simili, la valutazione della performance è effettuata con una misura di similarità o dissimilarità tra i cluster. Può assumere valori nell'intervallo [-1, 1], dove *ss* = -1 indica un clustering errato; *ss = 0* indica cluster sovrapposti e *ss* = 1 indica clustering denso (cioè, in cui i cluster sono ben separati tra loro e le osservazioni all'interno di ciascun cluster sono vicine).
Questa misura è stata ottenuta per le principali combinazioni di features a disposizione, eccetto per le combinazioni che escludessero le features *range* o *correlazione,* considerate punti cardine dell'analisi in quanto basate sui prezzi.
Le varie combinazioni sono ottenute creando, a partire dal dataset finale contenente i valori di tutte le variabili, per entrambe le banche dati e per tutti i codici HS, tanti dataset in cui sono state volta per volta estratte solo le variabili di interesse.
Lo score è calcolato sulle variabili considerate congiuntamente.

Come accennato in precedenza, l'algoritmo fornisce come output i due cluster: le etichetta *price* o *quality* devono essere assegnate manualmente a ciascun cluster. Per effettuare questa assegnazione viene calcolato il valore medio per ciascuna variabile nei due cluster e viene applicata una funzione di assegnazione alle categorie che lavora nella seguente maniera:

- Se il valore medio del range del primo cluster è maggiore del valore medio del secondo cluster, allora il primo cluster contiene prodotti *quality* e il secondo prodotti *price,* e viceversa.

- Se il valore medio della correlazione del primo cluster è minore del valore medio del secondo cluster, allora il primo cluster contiene prodotti *quality* e il secondo prodotti *price,* e viceversa.
- Se il valore medio dell'indice di Herfindal del primo cluster è minore del valore medio del secondo cluster, allora il primo cluster contiene prodotti *quality* e il secondo prodotti *price,* e viceversa.
- Se il valore medio dell'indice di Grubel del primo cluster è maggiore del valore medio del secondo cluster, allora il primo cluster contiene prodotti *quality* e il secondo prodotti *price,* e viceversa.

Il clustering avviene quindi sulla base delle variabili considerate contemporaneamente, ma la fase di labelling può essere fatta considerando una sola variabile per volta, e non i loro valori in modo congiunto: è quindi necessario verificare che l'etichetta finale sia effettivamente coerente con i valori delle feature. Questo consente inoltre di ottenere un indice di adattamento per ogni variabile, denominato R^2: questa misura calcola la percentuale di codici che con il clustering vengono assegnati identicamente alla classificazione StudiaBo esistente. È importante sottolineare che tale misura non può essere valutata asetticamente, in quanto è calcolata sulla base della classificazione StudiaBo esistente: è perciò certamente attendibile per determinati settori industriali, che possiedono caratteristiche notoriamente price o quality, mentre per prodotti più "di nicchia", sui quali non si hanno abbastanza conoscenze del mercato, un valore basso non corrisponde necessariamente ad una errata previsione. Perciò, è direttamente interpretabile dal momento in cui si assume la vecchia classificazione come corretta nella sua totalità.

Di seguito, vengono riportate e analizzate le tabelle riassuntive per tutti i modelli considerati, sia per i dati Ulisse che per i dati Comext. Le tabelle riportano, per i due cluster, il numero di prodotti, il valore medio delle variabili e l'etichetta assegnata al cluster. Ad ogni modello è associata la rispettiva misura di performance e l'indice di adattamento (R^2) delle features.

a) RANGE, CORRELAZIONE, HERFINDAL, GRUBEL INDEX (Ulisse)

CLUSTER	N PROD	RANGE	CORR	HER 3	GI	TYPE
0	2584	69.6	0.61	46.7	33.7	**Quality**
1	2781	36.8	0.78	65.2	13.8	**Price**

Tabella 3: Specificazione 1 (Ulisse)

Silhouette score:

0.42

Indice di adattamento:

R^2range: 0.598

R^2corr: 0.598

R^2HER: 0.598

R^2GI: 0.598

Il primo modello include tutte le variabili; delle due misure di concentrazione di mercato si è scelto di includere solo quello calcolato sulle TOP 3, in quanto l'inclusione di quella sulle TOP 5 è parsa ridondante. Il modello in questione genera due cluster di numerosità simile e la denominazione dei due cluster è coerente ai valori medi delle features:

- Al cluster 0 sono stati assegnati 2584 codici HS6, caratterizzati da un'alta dispersione, correlazione e indice di Herfindal relativamente minori e indice di Grubel più elevato: l'etichetta *Quality* è pertanto coerente a questi risultati.
- Al cluster 1 sono stati assegnati 2781 codici HS6, caratterizzati invece da minore dispersione, correlazione e indice di Herfindal più elevati e indice di Grubel inferiore: è stato coerentemente etichettato come *Price.*

Il valore del silhouette score è 0.42, un valore intermedio che consente di considerare questa specificazione come abbastanza buona. I valori dei diversi R^2 sono identici, mostrando come le diverse variabili funzionino ugualmente bene nella classificazione.

b) RANGE, CORRELAZIONE, GRUBEL INDEX (Ulisse)

CLUSTER	N PRODOTTI	RANGE	CORR	GI	TYPE
0	3060	36	0.82	17	**Price**
1	2305	75	0.52	32	**Quality**

Tabella 4: Specificazione 2 (Ulisse)

Silhouette score:

0.45

Indice di adattamento:

R^2range: 0.595

R^2corr: 0.595

R^2GI: 0.595

Il secondo modello non include l'indice di Herfindal, la prima differenza con il precedente modello è riscontrabile nella numerosità dei cluster, infatti il primo contiene circa 700 prodotti in più del secondo:

- Al cluster 0 sono stati assegnati 3060 codici HS6, con bassa dispersione, elevata correlazione e indice di Grubel relativamente minore: i prodotti sono *Price.*
- Al cluster 1 sono stati assegnati 2305 codici HS6, aventi invece da elevata dispersione, correlazione media e indice di Grubel circa doppio al primo: i prodotti sono *Quality.*

Il valore del silhouette score è 0.45, mostrando quindi un leggero miglioramento rispetto alla prima specificazione. I valori dei diversi R^2 restano invece invariati.

c) RANGE, CORRELAZIONE (Ulisse)

CLUSTER	N PRODOTTI	RANGE	CORR	TYPE
0	2045	77.8	0.46	**Quality**
1	3320	37	0.84	**Price**

Tabella 5: Specificazione 3 (Ulisse)

Silhouette score:

0.29

Indice di adattamento:

R^2corr: 0.558

R^2GI: 0.558

La terza specificazione include solo le variabili di prezzo, che sembrano comunque funzionare abbastanza bene, tuttavia la performance del modello diminuisce notevolmente in seguito all'esclusione di una terza variabile e per questo motivo questa non sarà considerata tra le scelte.

d) CORRELAZIONE, GRUBEL INDEX (Ulisse)

CLUSTER	N PRODOTTI	CORR	GI	TYPE
0	3270	0.8	15	**Price**
1	2095	0.53	37	**Quality**

Tabella 6: Specificazione 4 (Ulisse)

Silhouette score:

0.03

Indice di adattamento:

R^2corr: 0.598

R^2GI: 0.598

Questa specificazione può essere esclusa in quanto lo score è prossimo al valore nullo, per cui il processo di clusterizzazione non è robusto.

Per completezza, si riporta un'ultima specificazione che giustifica la scelta iniziale di non considerare le specificazioni che escludessero le variabili di prezzo:

e) HERFINDAL INDEX, GRUBEL INDEX (Ulisse)

CLUSTER	N PRODOTTI	CORR	GI	TYPE
0	3270	0.8	15	**Price**
1	2095	0.53	37	**Quality**

Tabella 7: Specificazione 5 (Ulisse)

Silhouette score:

0.022

Indice di adattamento:

R^2corr: 0.558

R^2GI: 0.558

Come si può notare, nonostante l'accuratezza rispetto alla vecchia specificazione rimanga abbastanza buona, il modello non è per niente performante: l'algoritmo, includendo solo queste variabili, non riesce nell'intento di generare cluster densi e separati. Questo risultato consente di escludere, nelle successive analisi, le specificazioni che non includano le variabili di prezzo, considerate già a priori estremamente rilevanti e significative.

Di seguito si riportano i risultati delle medesime specificazioni applicate ai dati COMEXT:

a) RANGE, CORRELAZIONE, HERFINDAL, GRUBEL INDEX (Comext)

CLUSTER	N PROD	RANGE	CORR	HER 3	GI	TYPE
0	2695	81.4	0.51	63.1	49.7	**Quality**
1	2670	45.6	0.66	83.3	18.9	**Price**

Tabella 8: Specificazione 1 (Comext)

Silhouette score:

0.49

Indice di adattamento:

R^2range: 0.629

R^2corr: 0.629

R^2HER: 0.371

R^2GI: 0.629

b) RANGE, CORRELAZIONE, GRUBEL INDEX (Comext)

CLUSTER	N PRODOTTI	RANGE	CORR	GI	TYPE
0	1818	97.12	0.22	39.36	**Quality**
1	3547	46.40	0.77	31.80	**Price**

Tabella 9: Specificazione 2 (Comext)

Silhouette score:

0.53

Indice di adattamento:

R^2range: 0.553

R^2corr: 0.553

R^2GI: 0.553

c) RANGE, CORRELAZIONE (Comext)

CLUSTER	N PRODOTTI	RANGE	CORR	TYPE
0	3624	47.81	0.77	**Price**
1	1741	96.42	0.20	**Quality**

Tabella 10: Specificazione 3 (Comext)

Silhouette score:

0.70

Indice di adattamento:

R^2corr: 0.542

R^2GI: 0.542

d) CORRELAZIONE, GRUBEL INDEX (Comext)

CLUSTER	N PRODOTTI	CORR	GI	TYPE
0	2284	0.56	58.21	**Quality**
1	3081	0.61	16.69	**Price**

Tabella 11: Specificazione 4 (Comext)

Silhouette score:

0.61

Indice di adattamento:

R^2corr: 0.599

R^2GI: 0.599

Nel complesso, l'algoritmo con i dati COMEXT mostra un silhouette score più elevato in tutte le specificazioni, mentre gli indici di adattamento sono pressochè identici. Nelle specificazioni con i dati Ulisse, il valore massimo del silhouette score è quello relativo al modello b), pari a 0.45; lo stesso modello con i dati Comext ha invece uno score pari a 0.53, mostrando quindi una clusterizzazione più efficace. Il valore più elevato con i dati Comext è invece associato al modello d), con un valore pari a 0.70; lo stesso modello con i dati Ulisse presentava un valore praticamente nullo.
Infine, si è scelto di utilizzare la stessa specificazione per entrambe le banche dati, inserendo nell'algoritmo i modelli riportati al punto b) di entrambe le banche dati.

5.2.3 Risultati clustering per settore industriale

Una volta scelta la specificazione, sulla base alle valutazioni appena descritte, si passa alla seconda fase del clustering. In questa fase viene effettuato un raggruppamento dei codici per codice industriale, al fine di effettuare un confronto settore per settore con la vecchia classificazione. Il metro di valutazione è ancora l'indice di adattamento, calcolato in questo caso per specifico settore. Questa fase è fondamentale per verificare se, almeno nei settori noti a priori, l'algoritmo effettui una corretta assegnazione dei prodotti.

Codice settore	R2 Ulisse	R2 Comext	Descrizione settore
A1	0,71	0,67	Materie prime naturali
A2	0,71	0,73	Materie prime industriali
B1	0,75	0,81	Beni alimentari intermedi e finali non confezionat
B2	0,59	0,55	Beni intermedi in materie tessili e pelli
B3	0,62	0,62	Beni intermedi in carta e in legno
B4	0,6	0,58	Beni intermedi in metallo
B5	0,51	0,59	Beni intermedi chimici
B6	0,44	0,47	Beni intermedi in minerali non metalliferi
C1	0,47	0,52	Beni e prodotti per le costruzioni
D1	0,68	0,6	Componenti elettroniche
D2	0,74	0,46	Componenti meccaniche ed ottiche
D3	0,47	0,27	Componenti per i mezzi di trasporto
D4	0,76	0,46	Elettrotecnica
E0	0,4	0,27	Alimentari confezionati e bevande
E1	0,71	0,43	Prodotti finiti di largo consumo
E2	0,47	0,44	Prodotti finiti per la persona
E3	0,34	0,41	Prodotti finiti per la casa
E4	0,52	0,43	Prodotti e strumenti per la salute
F1	0,62	0,32	Strumenti e attrezzature per ICT e servizi
F2	0,73	0,43	Strumenti e attrezzature per l'industria
F3	0,26	0,25	Mezzi di trasporto e per l'agricoltura
F4	0,48	0,31	Macchine e impianti per i processi industriali
F5	0,72	0,36	Impiantistica industriale
G1	0,65	0,45	Armi e munizioni
Media	0,58	0,48	Overall
Mediana	0,61	0,46	Overall

Tabella 12: Clustering per settore industriale

Come mostrato in tabella, alcuni settori mostrano un adattamento alla vecchia classificazione abbastanza elevato: si prenda ad esempio il settore A2 o il settore B1. Altri, invece, decisamente basso: ad esempio, il settore F3. Questi risultati aggregati sono utili per valutare l'algoritmo: in prima analisi, sembrerebbe funzionare discretamente per i settori con caratteristiche ben definite (come, appunto, il settore B2 che contiene prodotti notoriamente price, così come il settore A2). D'altro canto, funziona meno bene per quei settori che contengono prodotti meno noti e di cui, quindi, è più difficile valutarne la natura. Inoltre, questa valutazione è basata sull'indice R^2, che ricordiamo essere valido solo se si assume completamente corretta la classificazione esistente.

5.3. Modelli di apprendimento supervisionato

Il secondo approccio utilizzato è quello del Machine Learning supervisionato: questo algoritmo agisce sotto la guida del programmatore, che "insegna" all'algoritmo i risultati da generare, facendo apprendere l'algoritmo da un set di dati già etichettato. Il nome supervisionato deriva, appunto, dal fatto che la base dati di esempio è scelta dal supervisor: è il cosiddetto *training set.* Il training set contiene n esempi, ciascuno dei quali consiste in un vettore di x_i caratteristiche (le features) e una etichetta y_i (label). Il training set è utilizzato per applicare le conoscenze acquisite al set di dati non ancora classificati, fornendo così le previsioni (o modello di classificazione).
Nel nostro caso, il training set consiste nel fornire all'algoritmo una serie di prodotti che a priori sono noti come quality o price.

5.3.1 Introduzione alla *classificazione*

La classificazione è una categoria del Machine Learning supervisionato che consiste nell'identificare a quale categoria appartiene un elemento in input, sulla base di un modello di classificazione ottenuto in apprendimento automatico. Ogni oggetto osservato è chiamato *istanza*, mentre l'algoritmo che implementa la classificazione è chiamato *classificatore:* quest'ultimo esamina l'istanza e la etichetta con una classe.
Dato che nel nostro caso le classi sono due (quality/price), si parla pertanto di classificazione binomiale o binaria.
L'algoritmo viene implementato in Python con la libreria di apprendimento automatico Scikit-learn, che contiene algoritmi di classificazione, oltre a quelli di clustering. Operativamente, è necessario selezionare dei prodotti, o classi di prodotti, da inserire come dati di input etichettati: più precisamente, vengono inseriti i prodotti appartenenti a due settori industriali di cui, a priori, si conosce la tipologia di appartenenza. I settori selezionati:

- B1 Beni alimentari intermedi e non confezionati, con label *Price*
- E2 Prodotti finiti per la persona, con label *Quality*

5.3.2. Accuracy Score e R^2

Un metodo implementabile con libreria Scikit-learn per la valutazione dei modelli è il *train test split,* un'utile funzione che consente di separare il training set selezionato (settori B1

e E2) in due subset: un train subset e un test subset. In particolare, il default prevede che il train set equivalga al 25% del totale, così da valutare il modello sul restante 75% dei dati. È una tecnica di valutazione della performance dell'algoritmo; la specificazione è la medesima scelta per il clustering che include le variabili Range, Correlazione e Grubel Index.
I modelli utilizzati sono:

a) *Logistic regression,* è un modello lineare per la classificazione che modella le probabilità che descrivono i possibili outcome utilizzando la funzione logistica. È usata per predire la probabilità di una variabile dipendente categorica come P(Y=1) in funzione di X.
b) *K-Nearest Neighbors,* è un modello non parametrico di classificazione che assegna ad ogni punto una classe. Computa la similarità tra l'input e ogni istanza del training set.
c) *Linear Discriminant Analysis,* sviluppa un modello di probabilità per classe basato sulla distribuzione delle osservazioni per ogni input. Ogni nuovo esempio è poi classificato calcolando la probabilità condizionata di appartenere a ciascuna classe.
d) *Gaussian Naive Bayes,* il più semplice e veloce, basato sul teorema della probabilità di Bayes.

Una volta effettuata la divisione dei codici HS tra train e test set in modo randomico, viene stimata l'accuratezza dei vari modelli attraverso l'*Accuracy score*. Tale score quantifica la qualità delle previsioni, valutando l'accuratezza del subset: ossia, quanto le previsioni effettuate sul subset test corrispondono al train set originale.
Di seguito si riportano i risultati per i dati Ulisse:

a) Accuracy of Logistic regression classifier on training set: 0.64
   Accuracy of Logistic regression classifier on test set: 0.67

b) Accuracy of K-NN classifier on training set: 0.77
   Accuracy of K-NN classifier on test set: 0.60

c) Accuracy of LDA classifier on training set: 0.64
   Accuracy of LDA classifier on test set: 0.66

d) Accuracy of GNB classifier on training set: 0.64
   Accuracy of GNB classifier on test set: 0.65

L'accuratezza è così calcolata[29]:

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

La valutazione dell'algoritmo di machine learning è una parte essenziale. Un grande problema in questa fase sorge nel momento in cui si fronteggia con dei training set sbilanciati: ossia, quando l'algoritmo deve effettuare una classificazione binaria ma nel training set una delle due classi è preponderante (ad esempio, 70% dei prodotti sono Price e solo 30% sono Quality). Si ovvia questo problema nel momento in cui vengono inseriti nel training set due settori industriali, di cui uno 100% Price (B1) e l'altro 100% Quality (E2).[30] Inoltre, la dimensione dei due set è simile: la classe B1 contiene 550 codici HS e la classe E2 ne contiene 484.
Quindi, un training set bilanciato permette di ovviare al problema dell'Accuracy paradox, che occorre quando un'elevata misura dell'accuratezza semplicemente riflette la distribuzione della classe sottostante.
Si può quindi passare all'analisi dei valori dell'accuratezza, dove:

- Accuracy on training set indica l'accuratezza del modello sugli esempi su cui è costruito
- Accuracy on test set indica l'accuratezza di un modello su input non conosciuti

Ciò che accade nel caso del modello al punto b) è un problema di *overfitting*, ed emerge dalla differenza tra l'accuratezza nel training set e quella nel test set. Ciò avviene quando il modello apprende dei meccanismi adatti al training set, ma che in fase di generalizzazione non funzionano al di fuori dello stesso. Una misura di overfitting è fornita dalla matrice di confusione[31]. Tale matrice descrive la performance del modello in una classificazione binaria.
Si riportano adesso i risultati per i dati Comext:

a) Accuracy of Logistic regression classifier on training set: 0.77
   Accuracy of Logistic regression classifier on test set: 0.80

b) Accuracy of K-NN classifier on training set: 0.82
   Accuracy of K-NN classifier on test set: 0.78

[29] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
[30] Secondo la classificazione StudiaBo.
[31] Appendice 2

c) Accuracy of LDA classifier on training set: 0.75
   Accuracy of LDA classifier on test set: 0.79
d) Accuracy of GNB classifier on training set: 0.75
   Accuracy of GNB classifier on test set: 0.79

Anche nel caso dell'algoritmo supervisionato, l'indice di accuratezza migliora con l'utilizzo della banca dati Comext, in particolare per il modello K-NN.
Come accennato in precedenza, l'Accuracy score dipende dalla distribuzione del test set e training set. I settori selezionati per il training set nell'algoritmo supervisionato sono stati selezionati valutando l'adattamento per settore industriale dai risultati dell'algoritmo non supervisionato: questi due settori avevano un indice R^2 tra i più elevati.
Per provare a migliorare l'accuratezza del modello, a partire dai risultati del clustering vengono selezionati i settori in oggetto e poi solo i codici accuratamente previsti rispetto alla vecchia classificazione. Vengono nuovamente implementati i modelli di classificazione, sulla base del nuovo training set che contiene solo i codici accuratamente previsti: tali codici sono quindi classificati come Price/Quality da entrambe le classificazioni. Questo intervento di supervisione ha l'intento di diminuire la variabilità interna ai settori del training set: selezionando solo i codici accuratamente previsti, aumenta la probabilità che questi siano etichettati in modo corretto e si massimizza il potenziale informativo dei dati forniti.
I nuovi risultati dell'Accuracy score per i dati Ulisse sono i seguenti:

a) Accuracy of Logistic regression classifier on training set: 0.99

   Accuracy of Logistic regression classifier on test set: 1.00

b) Accuracy of K-NN classifier on training set: 0.99

   Accuracy of K-NN classifier on test set: 0.99

c) Accuracy of LDA classifier on training set: 0.98

   Accuracy of LDA classifier on test set: 0.98

d) Accuracy of GNB classifier on training set: 0.96

   Accuracy of GNB classifier on test set: 0.97

Mentre per i dati Comext:

a) Accuracy of Logistic regression classifier on training set: 0.99

   Accuracy of Logistic regression classifier on test set: 0.99

b) Accuracy of K-NN classifier on training set: 0.99

   Accuracy of K-NN classifier on test set: 0.99

c) Accuracy of LDA classifier on training set: 0.98

   Accuracy of LDA classifier on test set: 0.98

d) Accuracy of GNB classifier on training set: 0.98

   Accuracy of GNB classifier on test set: 0.97

È evidente come l'indice di accuratezza sia notevolmente migliorato per tutti i modelli e in entrambe le banche dati; per la seconda fase della classificazione viene quindi selezionato il modello K-NN, che presenta la medesima accuratezza tra train e test set e tra banche dati.

5.3.3. Risultati classificazione per settore industriale

Di seguito si riportano i risultati per settore industriale ottenuti con il modello K-NN appena descritto.

Codice settore	R2 Ulisse	R^2 Comext	Descrizione settore
A1	0.77	0.76	Materie prime naturali
A2	0.76	0.78	Materie prime industriali
B2	0.72	0.71	Beni intermedi in materie tessili e pelli
B3	0.72	0.54	Beni intermedi in carta e in legno
B4	0.62	0.61	Beni intermedi in metallo
B5	0.57	0.56	Beni intermedi chimici
B6	0.48	0.45	Beni intermedi in minerali non metalliferi
C1	0.45	0.45	Beni e prodotti per le costruzioni
D1	0.57	0.60	Componenti elettroniche
D2	0.58	0.62	Componenti meccaniche ed ottiche
D3	0.35	0.36	Componenti per i mezzi di trasporto
D4	0.66	0.67	Elettrotecnica
E0	0.28	0.28	Alimentari confezionati e bevande
E1	0.58	0.61	Prodotti finiti di largo consumo
E3	0.26	0.25	Prodotti finiti per la casa
E4	0.36	0.35	Prodotti e strumenti per la salute
F1	0.45	0.42	Strumenti e attrezzature per ICT e servizi
F2	0.64	0.63	Strumenti e attrezzature per l'industria
F3	0.18	0.19	Mezzi di trasporto e per l'agricoltura
F4	0.29	0.32	Macchine e impianti per i processi industriali
F5	0.56	0.60	Impiantistica industriale
G1	0.55	0.55	Armi e munizioni
Media	0.52	0.51	Overall
Mediana	0.57	0.56	Overall

Tabella 13: Classificazione per settore industriale

In media, il valore dell'indice di adattamento è inferiore rispetto al precedente risultato di clustering. Ancora una volta, si evidenzia il problema di non avere una classificazione unica di riferimento; infatti, nonostante l'Accuracy score abbia un valore molto elevato sul test set, in seguito alla selezione dei dati, il confronto con la vecchia classificazione sembra non riportare risultati troppo soddisfacenti.

Per alcuni settori il valore dell'indice è rimasto invariato, per altri è peggiorato, come nel caso del settore F3: Mezzi di trasporto e per l'agricoltura.[32] La classificazione StudiaBo classifica tutti i prodotti di questo settore come quality, mentre la nuova classificazione ne assegna a

[32] L'elenco completo dei prodotti è riportato in Appendice

questa classe solo il 20%. Un'analisi più dettagliata mostra un valore medio delle features pari a 41.3 per la dispersione dei prezzi, 0.76 per la correlazione e 23.5 per l'indice di Grubel. Tali valori corrispondono, secondo le ipotesi, a prodotti appartenenti alla classe quality: tuttavia, tale settore comprende prodotti come autoveicoli per il trasporto di persone o, ancora, motocicli, la cui conoscenza del mercato consente di affermare che tali prodotti siano certamente differenziabili. Basti pensare al commercio europeo di automobili, come chiaro esempio della differenziazione qualitativa (ancora, reale o percepita) che consente ad alcuni marchi di vendere i propri prodotti ad un prezzo molto elevato.

Per quanto riguarda il resto dei prodotti appartenenti a questo settore, non si ha una conoscenza approfondita del mercato: è possibile che tali prodotti non siano propriamente price, ma che le preferenze dei consumatori siano piuttosto omogenee, e trattandosi di mezzi funzionali all'attività lavorativa o al trasporto delle persone rilevino in maniera "indiretta" nell'utilità individuale. È perciò possibile considerare tali prodotti come beni *intermedi.*

5.4. Risultati per codici selezionati

HS6	Descrizione	RANGE	CORR	GI	StudiaBo	Previsione UL	Previsione CO	Settore
HS090111	Caffé non torrefatto	23.35	0.99	5.97	price	price	price	A1
HS120190	Baccelli di soia	5.37	0.99	0.43	price	price	price	A1
HS260111	Minerali di ferro	10.92	0.99	1.67	price	price	price	A1
HS270400	Coke e semi-coke di carbon fossile	19.85	0.99	4.94	price	price	price	A2
HS270900	Oli di petrolio	9.23	1.00	3.48	price	price	price	A1
HS280410	Idrogeno	3.17	0.99	6.91	price	price	price	B5
HS330410	Trucchi per le labbra	54.43	0.92	55.09	quality	price	price	E2
HS330510	Shampoo	66.06	0.83	51.48	quality	quality	price	E1
HS400110	Lattice di gomma naturale	14.90	0.97	5.89	price	price	price	A1
HS530911	Tessuti di lino	49.83	0.81	15.73	price	price	price	B2
HS610120	Cappotti	53.94	0.45	25.26	quality	quality	price	E2
HS611211	Tute sportive	56.81	0.77	31.55	quality	quality	price	E2
HS640219	Calzature per lo sport	38.54	0.99	33.52	quality	price	price	E2
HS650699	Cappelli	43.80	-0.39	14.94	quality	quality	price	E2
HS721810	Acciai inossidabili	51.81	0.55	48.21	price	quality	quality	A2
HS740311	Rame raffinato	2.91	0.98	8.75	price	price	price	A2
HS750210	Nichel	5.00	1.00	15.02	price	price	price	A2
HS780199	Piombo greggio	21.10	0.97	16.76	price	price	price	A2
HS870120	Trattori stradali per semirimorchi	15.84	0.48	13.02	quality	price	price	F3
HS940171	Mobili per sedersi	69.01	0.60	20.98	quality	quality	price	F1
HS940510	Lampadari	69.44	0.44	29.72	quality	quality	price	E3
HS970110	Quadri, pitture e disegni	0.43	1.00	41.66	price	price	quality	E2

Tabella 14: Confronto risultati per alcuni codici HS

La precedente tabella riporta i risultati della previsione effettuata, attraverso l'algoritmo di clustering, per la selezione di codici utilizzata per l'analisi preliminare.
Come abbiamo visto, con questo algoritmo gli indici di adattamento per settore industriale presentano una media vicina al 60% per i dati Ulisse e al 50% per i dati Comext. Si prende adesso in esame il confronto tra la nuova previsione e la classificazione Studiabo, per i codici HS6 su cui si è effettuata l'analisi preliminare. Si possono effettuare alcune considerazioni:

- I codici appartengono ai settori industriali che mostrano i più elevati indici di adattamento
- Dei 22 codici nel campione, la classificazione StudiaBo ne inserisce 13 nel gruppo price e 9 nel gruppo quality; su questo campione la classificazione sembra funzionare bene, mostrando un indice di adattamento dell'86.4%.

Sono tre i prodotti che vengono classificati diversamente:

1) Calzature sportive vengono classificate come *price.* Analizzado i risultati delle variabili, si evidenzia una dispersione molto elevata ma anche una correlazione praticamente perfetta, pari a 0.99. Inoltre, anche l'indice di commercio intra-settoriale presenta un valore abbastanza elevato. L'analisi di questi risultati, congiuntamente alla conoscenza del mercato, porta a considerare errata la classificazione di questo prodotto, che può quindi considerarsi *quality.*
2) Acciaio inossidabile, classificato come *quality.* Anche in questo caso, la classificazione è da considerarsi errata, in quanto è un bene quotato, quindi una commodity.
3) Trattori stradali per semirimorchi, classificati come price. Presentano bassa dispersione e indice di Grubel, con una correlazione media tra i prezzi di primo e terzo quartile; nonostante sia un bene differenziabile per caratteristiche o qualità, si potrebbe considerare tale bene come *intermedio* e cioè per ogni tipologia esiste un prezzo di riferimento oltre cui nessun consumatore è disposto ad acquistare il bene.

5.5. Risultati per campioni casuali

Si riportano di seguito i risultati derivanti dal clustering effettuato su due campioni casuali di numerosità n=30. Il campionamento viene effettuato con funzione del pacchetto random dalla lista dei 5365 codici HS. Vengono omessi dalla tabella i valori delle features per renderla più facilmente leggibile.

HS17	Descrizione prodotto	StudiaBo	Previsioni U	Previsioni C	Codice settore
HS010129	Cavalli, vivi	price	price	price	A1
HS021011	Prosciutti, spalle e loro pezzi, di suidi	quality	price	price	E0
HS090220	The verde	price	price	price	A1
HS121293	Canne da zucchero	price	price	quality	A1
HS140420	Linters di cotone	price	quality	price	B1
HS200979	Succhi di mela	quality	price	price	E0
HS210420	Preparazioni alimentari omogeneizzate	quality	quality	price	E0
HS252330	Cementi alluminosi	price	price	price	C1
HS271019	Oli medi	price	price	price	A2
HS290511	Metanolo	price	price	price	A2
HS290721	Resorcinolo	price	price	price	A2
HS293980	Alcaloidi	price	price	quality	B5
HS300215	Prodotti immunologici	quality	price	price	E4
HS300220	Vaccini per la medicina umana	quality	price	price	E4
HS310490	Sali di potassio	price	price	quality	B5
HS401695	Materassi pneumatici	price	quality	quality	B5
HS442010	Statuette e oggetti ornamentali di legno	quality	price	price	E3
HS481151	Carta e cartone colorati	price	price	quality	B3
HS620453	Gonne e gonne-pantaloni	quality	price	price	E2
HS710812	Oro	price	price	price	A2
HS730451	Tubi e profilati cav	price	price	price	B4
HS750711	Tubi di nichel	price	price	price	A2
HS840490	Apparecchi per caldaie	quality	quality	price	D2
HS842123	Apparecchi per filtrare gli oli	quality	quality	price	D3
HS851310	Lampade elettriche portatili	quality	price	price	E3
HS852321	Schede con pista magnetica	price	quality	quality	D1
HS852849	Monitor con tubo catodico	quality	price	quality	E3
HS903033	Strumenti la misura della tensione elettrica	quality	quality	quality	F2
HS911320	Cinturini per orologi	quality	quality	price	E2
HS940550	Apparecchi per l'illuminazione	quality	price	quality	E3

Tabella 15: Risultati campione casuale 1

Per questo primo campione, i codici correttamente previsti 18/30 (60%) con la banca dati Ulisse e 13/30 con la banca dati Comext (43%). Il campione, secondo la classificazione StudiaBo, è composto da 14 prodotti quality e 16 prodotti price. Il clustering classifica come prodotti quality solo 8 e 9 prodotti (rispettivamente con Ulisse e Comext), di cui coerenti alla vecchia classificazione solo 5 nel caso dei dati Ulisse e 2 dei dati Comext. Il clustering nel campione presenta quindi una distorsione nella classificazione verso la classe dei prodotti price, che risultano essere la classe a cui vengono assegnati il maggior numero di prodotti e

anche la classe che presenta la percentuale più elevata di codici corrispondenti tra le due classificazioni.

HS6	Descrizione prodotto	StudiaBo	Previsione U	Previsione C	Codice settore
HS010613	Cammelli, vivi	price	quality	price	A1
HS020621	Lingue di bovini	price	quality	price	B1
HS100860	Triticale	price	quality	price	A1
HS250900	Creta	price	quality	price	A1
HS262011	Metalline di galvanizzazione	price	price	price	A1
HS330520	Preparazioni permanenti per capelli	quality	quality	price	E2
HS340540	Paste abrasive	quality	price	price	E1
HS382311	Acido stearico industriale	price	price	price	A2
HS391729	Tubi di materie plastiche	price	quality	price	B5
HS440799	Legnotagliato	price	price	quality	B3
HS480640	Carta pergamina	price	price	price	B3
HS520839	Tessuti di cotone	price	quality	price	B2
HS521159	Tessuti prevalentememte di cotone	price	quality	price	B2
HS551012	Filati	price	price	price	B2
HS551622	Tessuti prevalentemente di fibre artificiali	price	price	quality	B2
HS610190	Cappotti di materie tessili	quality	price	price	E2
HS611130	Indumenti per bambini piccoli	quality	price	quality	E2
HS611595	Calze, calzettoni, calzini	quality	price	quality	E2
HS620111	Cappotti di lana o di peli fini	quality	quality	price	E2
HS710210	Diamanti non scelti	price	price	price	A1
HS721119	Prodotti piatti di ferro	price	price	price	A2
HS721913	Prodotti piatti di acciai inossidabili	price	price	price	A2
HS730422	Aste di perforazione	price	quality	quality	B4
HS750620	Lamiere di nichel	price	price	price	A2
HS830590	Graffette	quality	price	price	B4
HS847681	Macchine automatiche per la vendita	quality	quality	price	F1
HS851529	Apparecchi per la saldatura	quality	quality	price	F2
HS871160	Motocicli	quality	quality	quality	F3
HS880330	Parti di aeroplani e di elicotteri	quality	quality	quality	D3
HS950621	Tavole a vela per la pratica di sport nautici	quality	price	price	E2

Tabella 16: Risultati campione casuale 2

L'indice di adattamento nel caso del secondo campione è pari a 16/20 con i dati Ulisse e 15/20 con i dati Comext. Di questi, sono di tipo price 10 per i dati Ulisse e 11 per i dati Comext. Secondo la classificazione StudiaBo, il campione è composto da 12 prodotti quality e 18 prodotti price; il rapporto è invece rispettivamente 14-16 per i dati Ulisse e 7-23 per i dati Comext. La correttezza nella classificazione sembra avvenire più frequentemente nel caso dei prodotti price, rilevando quindi la difficoltà dell'algoritmo nell'individuazione dei prodotti quality.

Infine, analizzando i risultati del clustering sul totale degli oltre 5000 codici HS si rileva che:

- La classificazione StudiaBo assegna 2215 prodotti alla tipologia quality e 2773 prodotti alla tipologia price

- La previsione assegna 2108 prodotti alla tipologia quality e 2880 prodotti alla tipologia price

Di questi, sono identicamente assegnati 1184 codici quality (56%) e 1829 codici price (64%), confermando quindi la maggiore difficoltà nella classificazione dei prodotti di tipo quality. In generale, la classe prevalente è quella dei prodotti price sia nella classificazione StudiaBo che nella previsione derivante dagli algoritmi di Machine Learning.

In conclusione, è possibile affermare che gli algoritmi costruiti hanno una buona capacità previsionale per i prodotti appartenenti a settori industriali con caratteristiche chiaramente definibili a priori, mentre riscontra difficoltà nell'assegnazione dei codici appartenenti a settori con caratteristiche intermedie, tali per cui le features individuate non consentono all'algoritmo di riconoscere dei pattern comuni a tutti i prodotti. Per risolvere tale problematica, sarebbe necessario uno studio più approfondito che consenta l'individuazione di tali settori "intermedi", così da studiarne le caratteristiche e sviluppare delle features ad hoc per migliorare l'efficienza degli algoritmi di previsione.

5.6. Metodologie a confronto

La seguente tabella riassume i risultati confrontando le due metodologie, supervisionata e non supervisionata, e le banche dati. Nelle colonne della classificazione sono presenti due righe vuote: sono i settori selezionati come training set.
I risultati sono gli stessi descritti ai paragrafi 5.2.3. e 5.3.3., vengono di seguito nuovamente riportate per visualizzare le due metodologie e i loro risultati nella stessa tabella.

Codice settore	Clustering U	Clustering C	Classificazione U	Classificazione C	Descrizione settore
A1	0.71	0.67	0.77	0.76	Materie prime naturali
A2	0.71	0.73	0.76	0.78	Materie prime industriali
B1	0.75	0.81			Beni alimentari intermedi e finali non confezionati
B2	0.59	0.55	0.72	0.71	Beni intermedi in materie tessili e pelli
B3	0.62	0.62	0.72	0.54	Beni intermedi in carta e in legno
B4	0.6	0.58	0.62	0.61	Beni intermedi in metallo
B5	0.51	0.59	0.57	0.56	Beni intermedi chimici
B6	0.44	0.47	0.48	0.45	Beni intermedi in minerali non metalliferi
C1	0.47	0.52	0.45	0.45	Beni e prodotti per le costruzioni
D1	0.68	0.6	0.57	0.60	Componenti elettroniche
D2	0.74	0.46	0.58	0.62	Componenti meccaniche ed ottiche
D3	0.47	0.27	0.35	0.36	Componenti per i mezzi di trasporto
D4	0.76	0.46	0.66	0.67	Elettrotecnica
E0	0.4	0.27	0.28	0.28	Alimentari confezionati e bevande
E1	0.71	0.43	0.58	0.61	Prodotti finiti di largo consumo
E2	0.47	0.44			Prodotti finiti per la persona
E3	0.34	0.41	0.26	0.25	Prodotti finiti per la casa
E4	0.52	0.43	0.36	0.35	Prodotti e strumenti per la salute
F1	0.62	0.32	0.45	0.42	Strumenti e attrezzature per ICT e servizi
F2	0.73	0.43	0.64	0.63	Strumenti e attrezzature per l'industria
F3	0.26	0.25	0.18	0.19	Mezzi di trasporto e per l'agricoltura
F4	0.48	0.31	0.29	0.32	Macchine e impianti per i processi industriali
F5	0.72	0.36	0.56	0.60	Impiantistica industriale
G1	0.65	0.45	0.55	0.55	Armi e munizioni

Tabella 17: Metodologie a confronto

Capitolo 6

6. Conclusioni

Il tema centrale del lavoro è come nell'ambito del commercio internazionale sia di notevole rilievo la differenziazione qualitativa dei prodotti. Nell'operare in un mercato, l'impresa deve scegliere la strategia da attuare per ottenere un vantaggio competitivo; la differenziazione è tra le strategie che consentono di conseguire una performance al di sopra della media. Questo accade grazie al fatto che la maggiore varietà di prodotti meglio soddisfa le esigenze dei consumatori, la cui disponibilità a pagare è crescente rispetto al benessere derivante dall'acquisto del prodotto. Quindi, aumenta man mano che il prodotto si avvicina alla varietà ideale. Il premium price non si configura, invece, nei mercati in cui vengono scambiati beni omogenei, che devono quindi essere venduti al prezzo di mercato.

Nel corso del lavoro si è cercato di individuare i driver che consentono di classificare i prodotti nelle categorie di bene differenziabile o bene omogeneo: si è evidenziata in primis la dispersione di prezzo, che è certamente un segnale della differenza qualitativa tra i beni, ma si è anche visto come questa possa essere sintomo di inefficienze di mercato quali asimmetrie informative o elevati costi di ricerca. Quindi, per identificare l'origine della dispersione si è integrata l'analisi della correlazione tra prezzi di fascia bassa e di fascia alta: anche in presenza di un range di prezzi ampio, se tale correlazione è elevata significa il bene è omogeneo, in quanto le dinamiche di prezzi bassi e alti sono molto simili perché influenzate allo stesso modo dalle variabili da cui il prezzo dipende. Viceversa, tali dinamiche sono più indipendenti nel il caso dei beni differenziabili: prezzi elevati riflettono qualità più elevata o specificazioni di prodotto differenti da quelle associate ai prezzi più bassi.

La questione della differenziabilità è stata messa in luce dalla letteratura sul commercio intra settoriale, che prova a spiegare l'esistenza del commercio tra paesi con dotazioni simili, che le teorie tradizionali sul commercio internazionale non avevano previsto e non riuscivano, quindi, a spiegare. Una misura di tale fenomeno è stata fornita dal lavoro di Grubel-Lloyd e ripresa nel presente studio.

Un quarto e ultimo fattore preso in considerazione è quello della concentrazione di mercato, ipotizzando che i beni omogenei siano caratterizzati dalla struttura di mercato dell'oligopolio con prodotti omogenei, in cui quindi il grado di concentrazione è elevato e le imprese competono su prezzi e quantità (ossia, poche nazioni si spartiscono il commercio di dati beni, per la cui produzione sono necessari determinati fattori ambientali o esistono barriere

all'ingresso per gli elevati costi, per cui l'ingresso non è profittevole). Il mercato dei beni differenziabili è invece caratterizzato dalla struttura tipica della concorrenza monopolistica, in cui agiscono un elevato numero di imprese con basso potere di mercato e in cui non sussistono le dinamiche strategiche tipiche dell'oligopolio: si ipotizza quindi una concentrazione di mercato inferiore. Questo non è in contraddizione con la presenza di economie di scala, per le quali un paese potrebbe specializzarsi nella produzione di una specifica varietà concentrando la produzione in un unico stabilimento. Si prenda in analisi, ad esempio, il mercato europeo delle automobili, che sono un prodotto notoriamente differenziabile: l'Italia è specializzata nella produzione di utilitarie e la Germania nelle auto sportive. È quindi vero che esiste una specializzazione, e quindi concentrazione, nazionale, ma solo di una varietà di prodotto. Per la misura di questo fenomeno è stato calcolato l'indice di Herfindal-Hirschmann.

Dopo aver individuato le variabili potenzialmente utili nella distinzione dei prodotti, sono stati costruiti dei modelli di Machine Learning per produrre la classificazione. L'analisi qualitativa sui modelli ha evidenziato come la migliore combinazione fosse quella che esclude la misura di concentrazione di mercato, includendo quindi le variabili di dispersione e correlazione dei prezzi e indice di commercio intra settoriale. Si è infine visto, tra i risultati, come questi algoritmi funzionino nel caso di prodotti con caratteristiche ben definite e la cui classe è nota a priori (come mostra la Tabella 14).

Per quanto riguarda i prodotti appartenenti a quelli che possono essere definiti come settori intermedi (ossia, settori che possiedono caratteristiche miste, tipiche sia dei prodotti omogenei che di quelli differenziabili), la valutazione ottenuta dal confronto con la classificazione esistente in StudiaBo ha una duplice interpretazione: da un lato, un valore basso dell'indice di adattamento può significare la difficoltà dell'algoritmo nel cogliere le lievi differenze tra i prodotti intermedi, restituendo una classificazione errata. D'altra parte, non esistendo una classificazione generale di riferimento, è possibile che la classificazione StudiaBo non sia corretta nella sua totalità, soprattutto in riferimento ai prodotti di cui si ha meno conoscenza del mercato di riferimento, pertanto un basso indice di adattamento non è necessariamente il risultato di un'errata classificazione.

Il risultato della classificazione dipende strettamente dalle variabili inserite nel modello di apprendimento automatico, pertanto uno sviluppo del lavoro potrebbe essere sicuramente rivalutare o rifinire tali variabili.

Si possono inoltre evidenziare alcune criticità nella metodologia:

1) Codici doganali.

   Come abbiamo visto, il commercio intra settoriale nasce nel momento in cui due nazioni importano ed esportano simultaneamente beni o servizi simili. La somiglianza tra beni è identificata attraverso i codici doganali: un primo problema potrebbe quindi sorgere nel momento in cui si sceglie il livello di dettaglio con cui operare. In questo frangente, è stata utilizzato il sistema armonizzato a 6 cifre; la classificazione potrebbe essere fatta con un livello di dettaglio ancora superiore, utilizzando il sistema ad 8 cifre. Tale dettaglio nella divisione dei prodotti andrebbe ad eliminare parzialmente la dispersione di prezzo imputabile all'aggregazione di prodotti di tipologie differenti, consentendo di isolare quella dovuta ad una differenziazione qualitativa o ad eventuali inefficienze di mercato.

2) Indice di Grubel.

   Nel 1975 Grubel e Lloyd forniscono una misura del commercio intra settoriale, calcolato con esportazioni e importazioni nazionali di ciascun prodotto. Un altro problema è pertanto relativo alla costruzione di tale indice, al fine del calcolo della variabile da inserire nel modello: infatti, quello costruito da Grubel e Lloyd è stato il punto di partenza per ottenere l'indice relativo al prodotto, per ogni paese presente nel dataset. Si è tuttavia reso necessario il calcolo di un indice aggregato, pertanto l'indice finale differisce da quello originale e il potenziale informativo è senza dubbio differente. Innanzitutto, sono stati selezionati i soli paesi il cui indice supera la soglia di 0.70, l'indice aggregato è poi ottenuto rapportando la media delle esportazioni e delle importazioni dei soli paesi che superano la soglia alla media delle esportazioni e delle importazioni totali. Questo valore può quindi subire delle distorsioni, in quanto include i valori di paesi in tutti gli stadi di sviluppo e la componente intra settoriale può essere inficiata dall'aggregazione dei dati. Inoltre, dal momento in cui è il risultato di una aggregazione tra nazioni, è possibile che i prodotti technology-intensive risentano della struttura della domanda delle economie in via di sviluppo. Una selezione dei paesi sviluppati potrebbe isolare una struttura più omogenea delle preferenze dei consumatori (ad esempio, i paesi OECD).

3) Indice di Herfindal.

   Ne è stata utilizzata una versione semplificata, ossia la sommatoria delle quote di mercato delle prime n nazioni. La definizione dell'indicatore potrebbe essere affinata,

usando ad esempio quella inserita tra gli indicatori UNCTAD[33], che utilizza la versione normalizzata per misurare il grado di concentrazione di mercato per nazione d'origine. Il rapporto UNCTAD sulla concentrazione dell'export mostra inoltre come per determinate commodity[34] sia effettivamente vero che l'export è concentrato su poche nazioni, mentre per altre[35] l'export è meno concentrato. Anche la struttura del mercato petrolifero, che ha sperimentato cambiamenti significativi negli anni, come il phasing out in alcune nazioni, ha comunque mantenuto un indice di concentrazione stabile intorno a 0.15 per l'introduzione di nuovi produttori, lo sviluppo di nuove tecnologie e presenza di conflitti e instabilità mostrando di essere un mercato dinamico, in cui è presente una facile compensazione tra produttori. La variabile non è risultata utile nelle per la specificazione del modello.

4) Assenza di microdati a livello di impresa.
   Queste metodologie di stima presumono l'impiego di dati su alcune variabili macroeconomiche e sui dati di commercio internazionale disaggregati a livello merceologico e geografico. L'utilizzo di tali dati aggregati e, nello specifico, dei valori medi unitari come proxy dei prezzi può generare alcuni problemi interpretativi. Tali approcci, infatti, non tengo conto dell'eterogeneità individuale tra le imprese che esportano. La disponibilità di informazioni dettagliate sulle singole imprese consentirebbe una più accurata valutazione. Tuttavia, l'analisi a livello settoriale richiederebbe una quantità e un livello di dettaglio delle informazioni difficilmente disponibile.

In uno studio dell'OECD (2002) che riassume la crescente importanza del commercio intra settoriale, ne vengono evidenziate le seguenti caratteristiche:

- Questa tipologia di commercio ha sperimentato una crescita significativa a partire dagli anni 80 in molte nazioni OECD
- È particolarmente elevato per alcuni prodotti manufatti basati sulla differenziazione (o frammentazione produttiva) come prodotti chimici, trasporti, componenti elettroniche
- È particolarmente elevato nelle economie cosiddette "supertrading", che importano ed esportano per più della metà del PIL
- È connesso con i flussi di IDE, soprattutto per le economie in transizione

[33] https://unctadstat.unctad.org/EN/IndicatorsExplained.html
[34] Seta (per l'80% in Cina), altre fibre tessili, sughero, uranio, gas da carbone e ceramica
[35] Energia elettrica, legno, frutta e bevande analcoliche

In definitiva, è una quota importante del commercio nel mondo globalizzato, ma lo è prevalentemente per le nazioni sviluppate: si potrebbe quindi verificare se e in che modo la classificazione cambierebbe se l'algoritmo fosse applicato non al commercio mondiale, ma solo a quello relativo ai paesi sviluppati, in cui il commercio intra settoriale è appunto più sviluppato. Si aderirebbe in qualche modo all'ipotesi di dotazioni simili, nonché gusti (e reddito) pressoché omogenei tra i consumatori.

Si può concludere che la difficoltà nella valutazione della classificazione ottenuta dipende sicuramente dalla mancanza della classificazione generale di riferimento e inoltre, più che una distinzione binaria tra "commodity" o "differenziabili", si potrebbe pensare ad una scala a più gradi di "commoditization" che dipendono dai fattori presi in esame, a partire dalle preferenze dei consumatori.

APPENDICE 1: Grafici di dispersione

HS17	DESHS17_IT	StudiaBo
HS090111	Caffé non torrefatto	price
HS120190	Baccelli di soia	price
HS260111	Minerali di ferro	price
HS270400	Coke e semi-coke di carbon fossile	price
HS270900	Oli di petrolio	price
HS280410	Idrogeno	price
HS330410	Trucchi per le labbra	quality
HS330510	Shampoo	quality
HS400110	Lattice di gomma naturale	price
HS530911	Tessuti di lino	price
HS610120	Cappotti	quality
HS611211	Tute sportive	quality
HS640219	Calzature per lo sport	quality
HS650699	Cappelli	quality
HS721810	Acciai inossidabili	price
HS740311	Rame raffinato	price
HS750210	Nichel	price
HS780199	Piombo greggio	price
HS870120	Trattori stradali per semirimorchi	quality
HS940171	Mobili per sedersi	quality
HS940510	Lampadari	quality
HS970110	Quadri, pitture e disegni	price

Tabella 18: Selezione codici HS

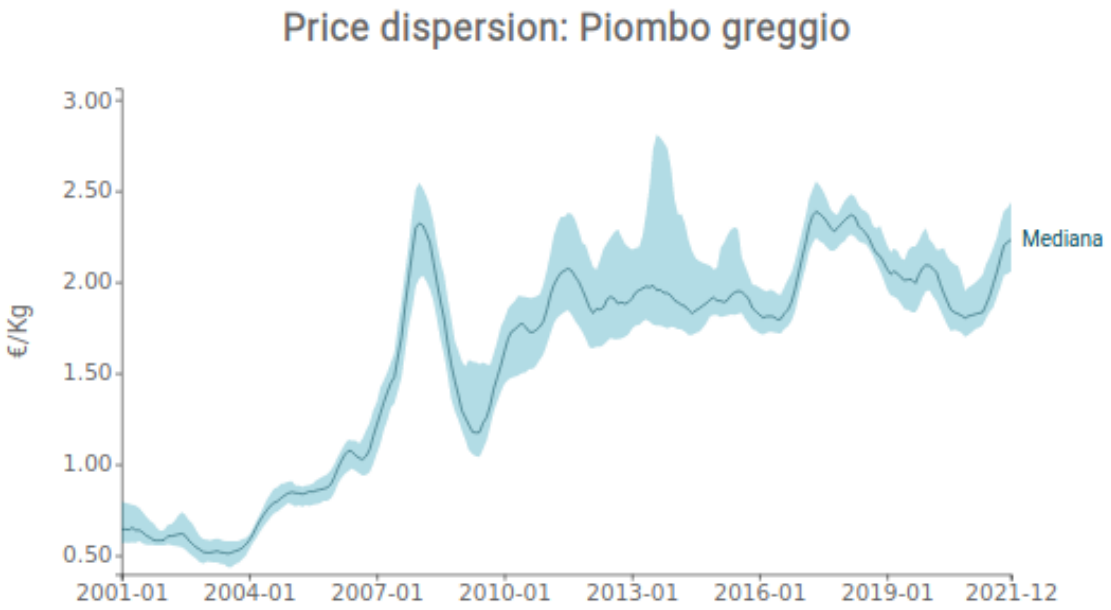

Figura 7. HS780199 - Fonte: Elaborazioni ExportPlanning

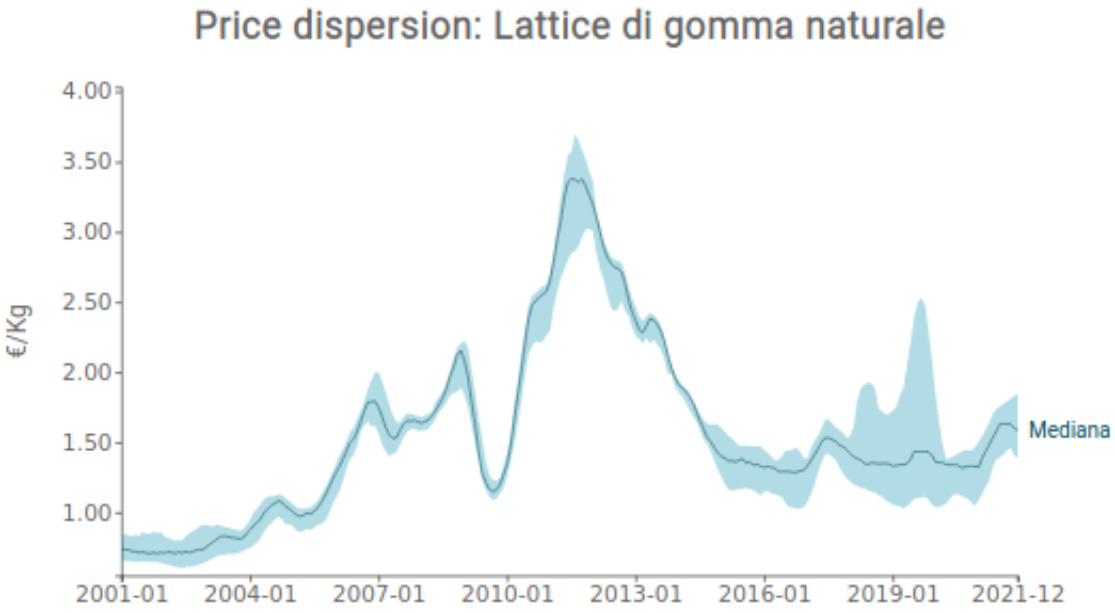

Figura 8. HS400110 - Fonte: Elaborazioni ExportPlanning

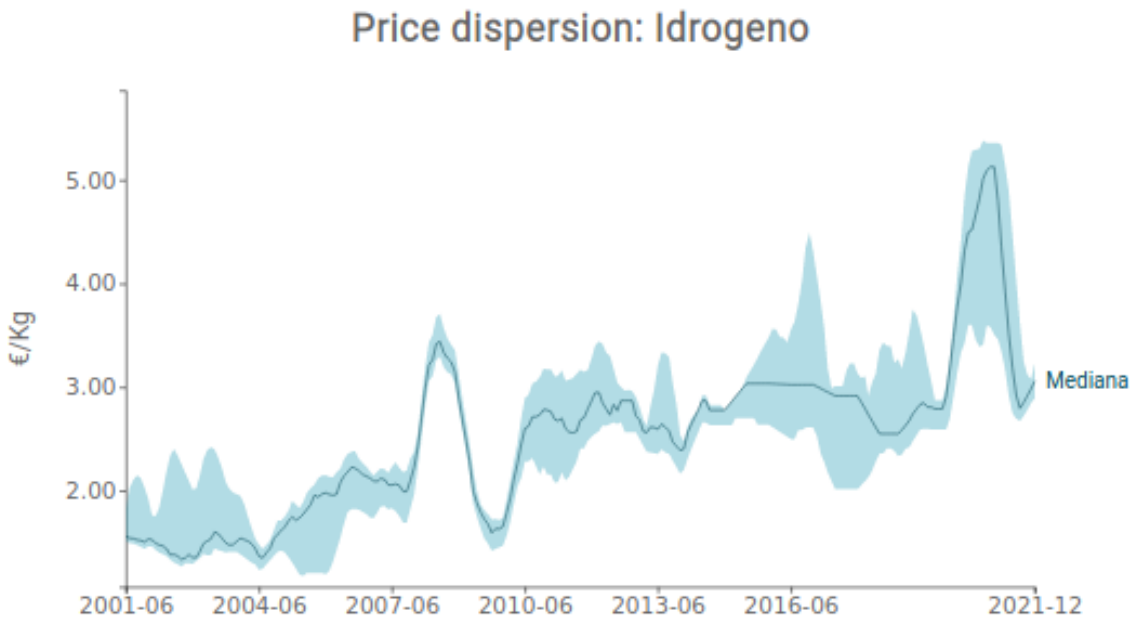

Figura 9. HS280410 - Fonte: Elaborazioni ExportPlanning

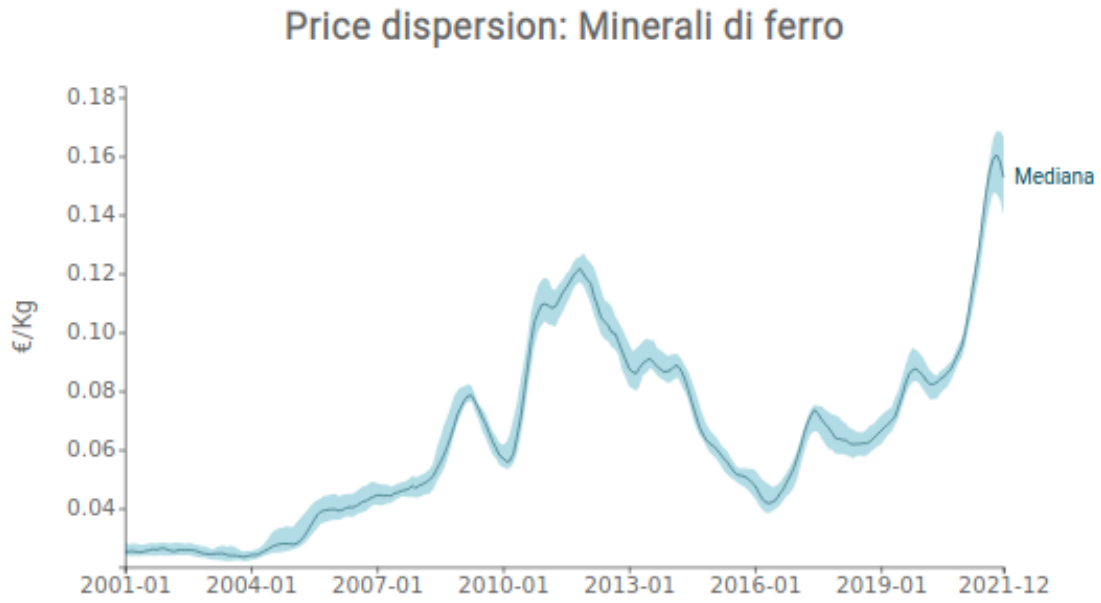

Figura 10. HS260111 - Fonte: Elaborazioni ExportPlanning

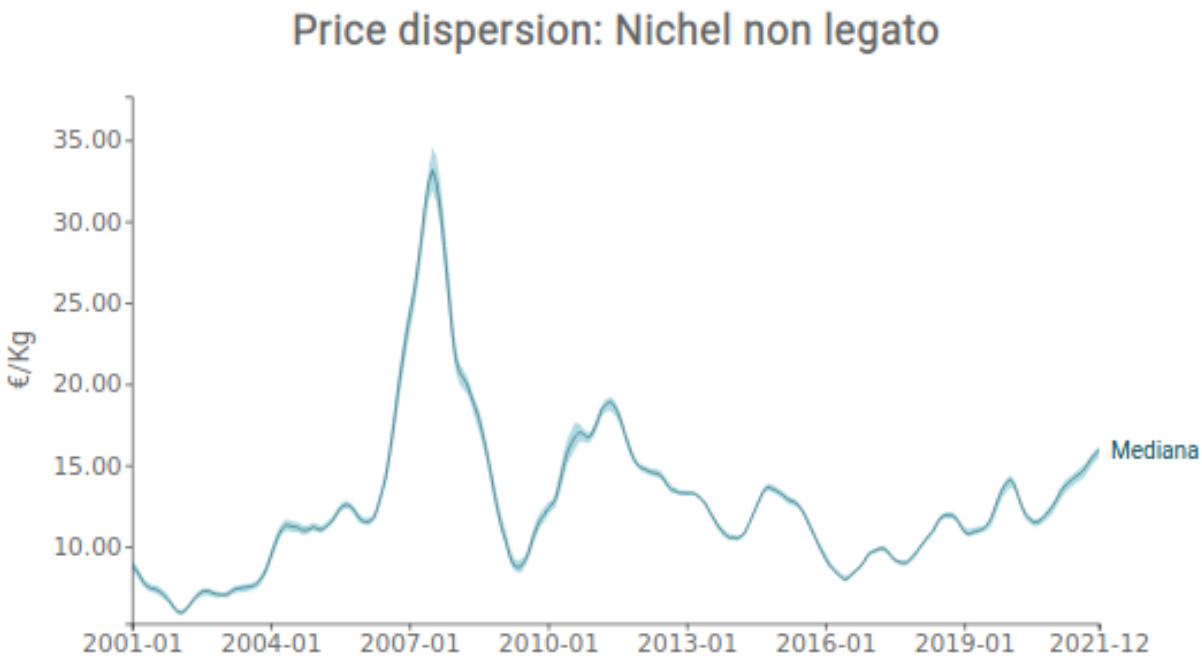

Figura 11. HS750210 - Fonte: Elaborazioni ExportPlanning

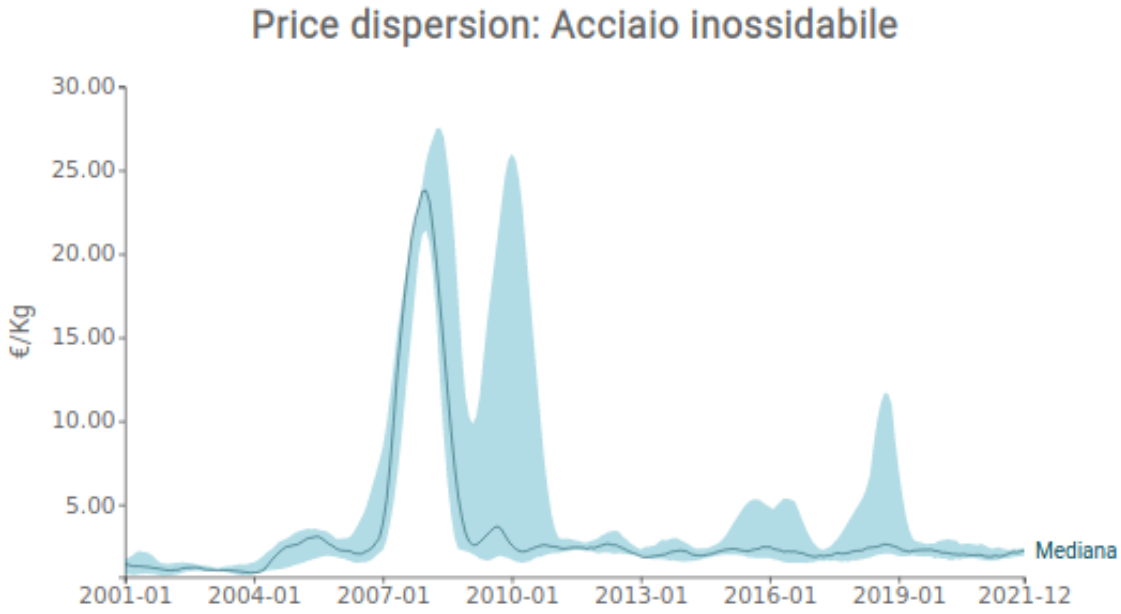

Figura 12. HS721810 - Fonte: Elaborazioni ExportPlanning

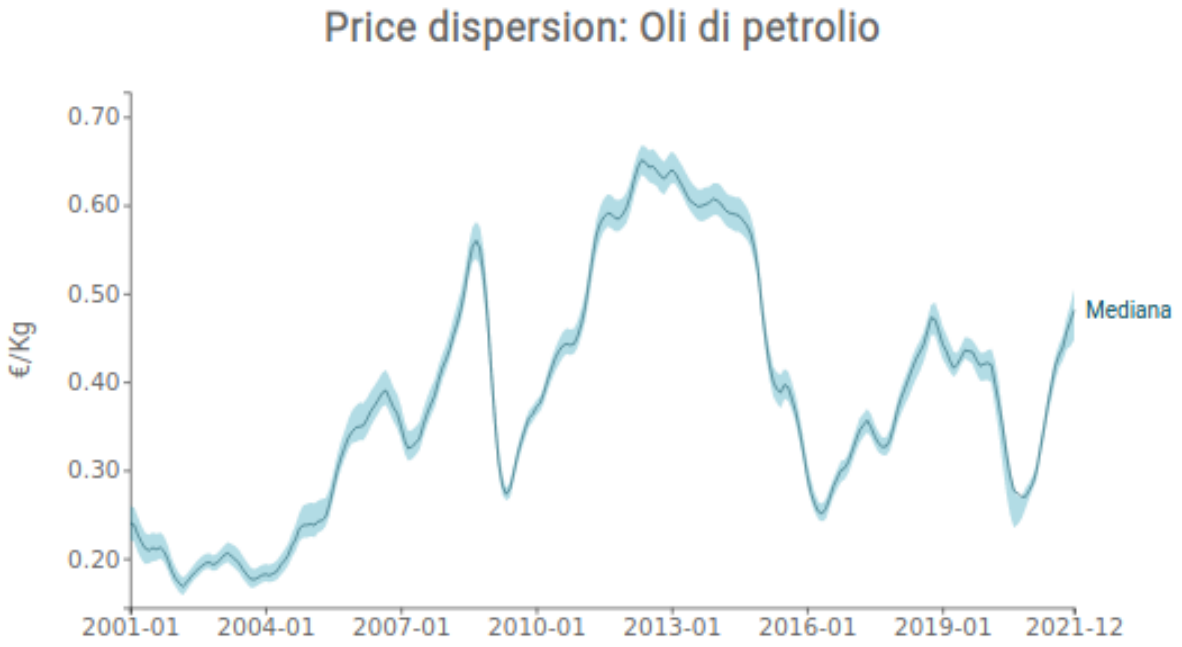

Figura 13. HS270900 - Fonte: Elaborazioni ExportPlanning

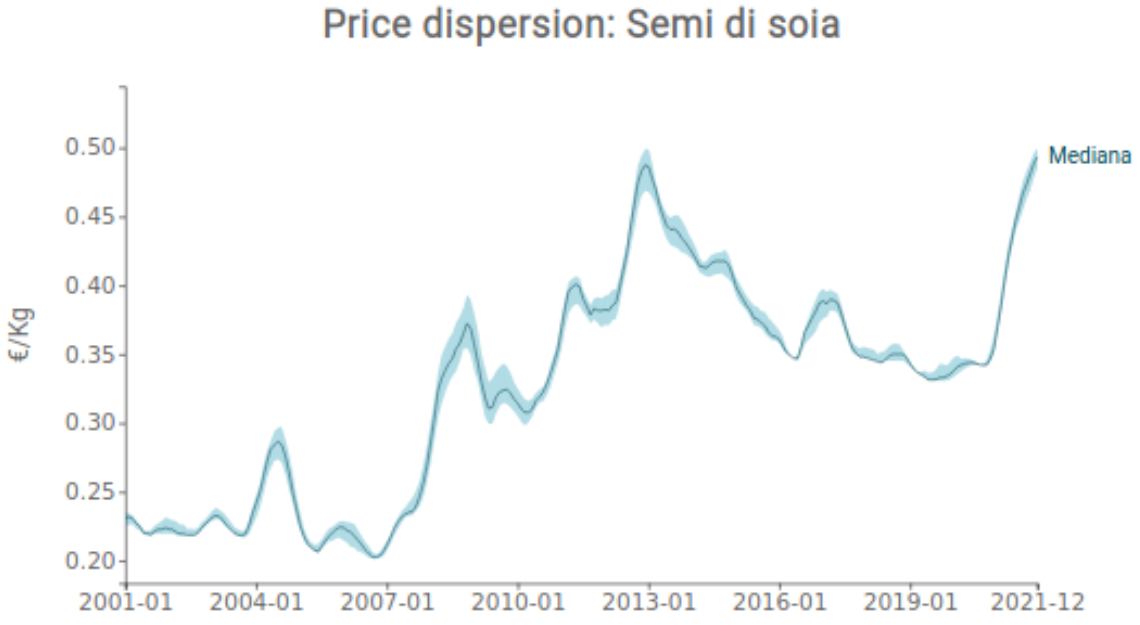

Figura 14. HS120190 - Fonte: Elaborazioni ExportPlanning

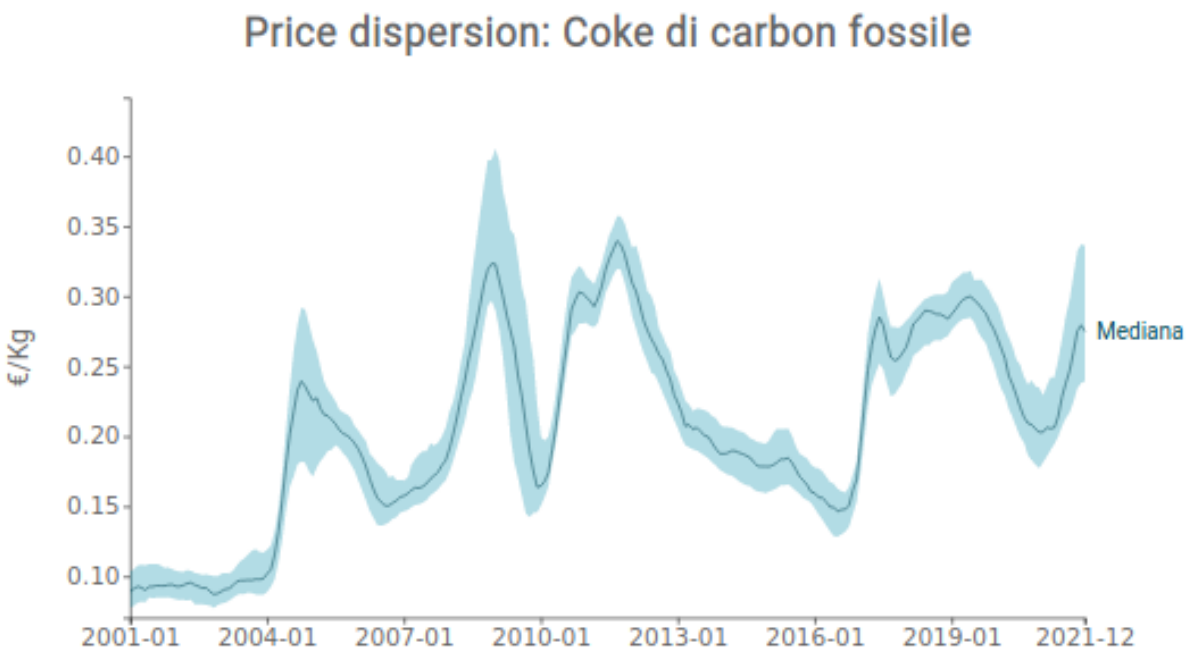

Figura 15. HS270400 - Fonte: Elaborazioni ExportPlanning

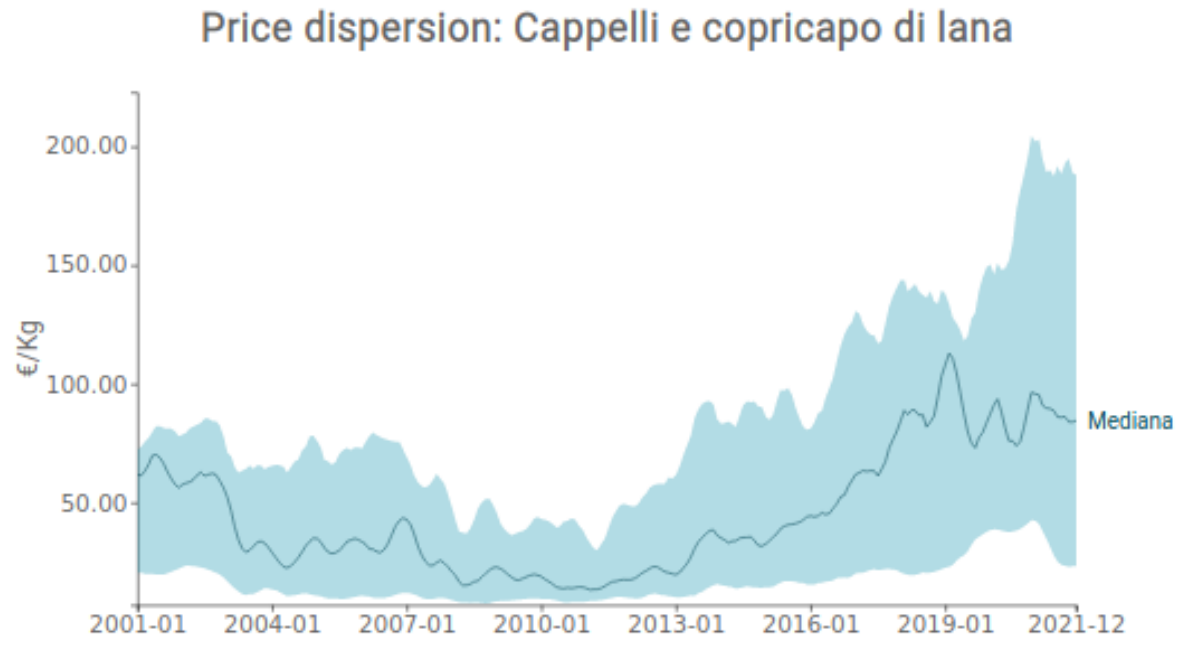

Figura 16. HS650699 - Fonte: Elaborazioni ExportPlanning

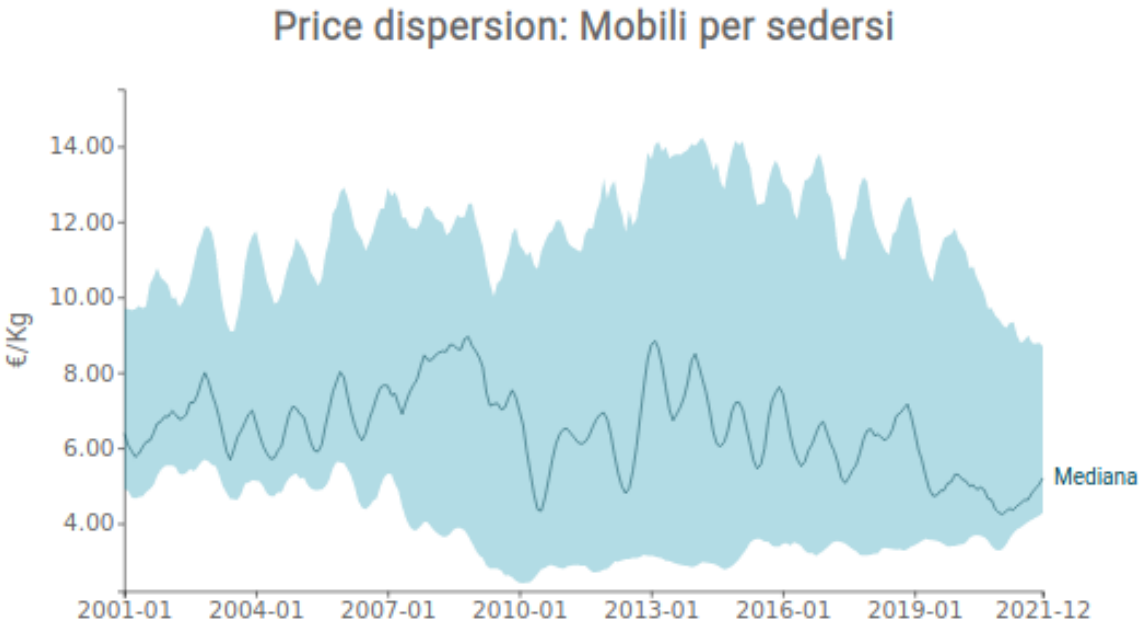

Figura 17. HS940171 - Fonte: Elaborazioni ExportPlanning

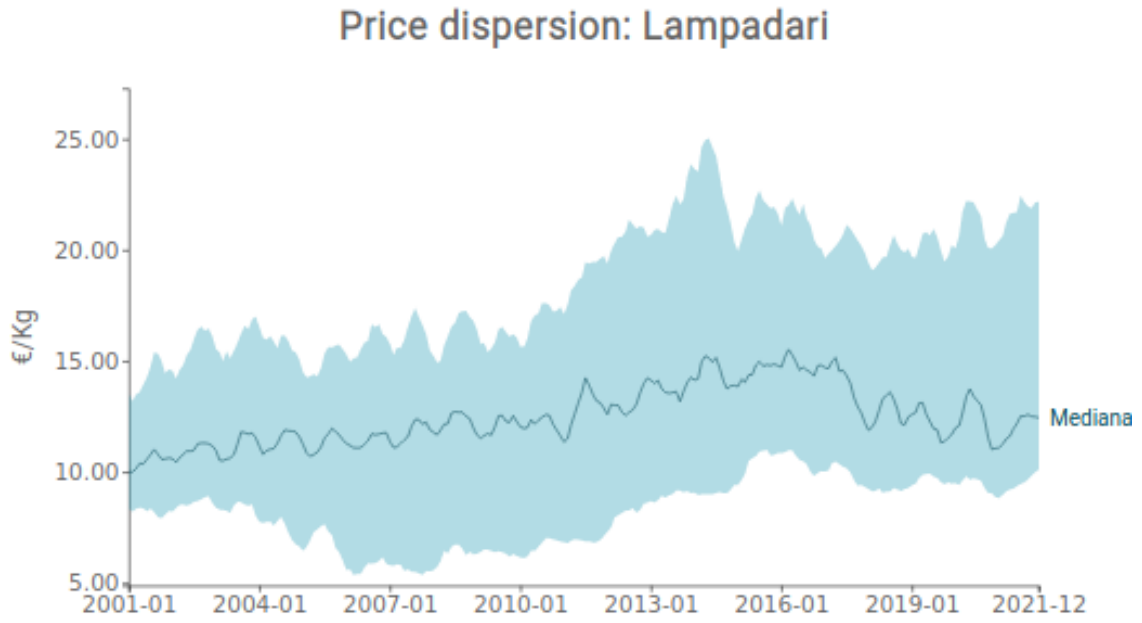

Figura 18. HS940510 - Fonte: Elaborazioni ExportPlanning

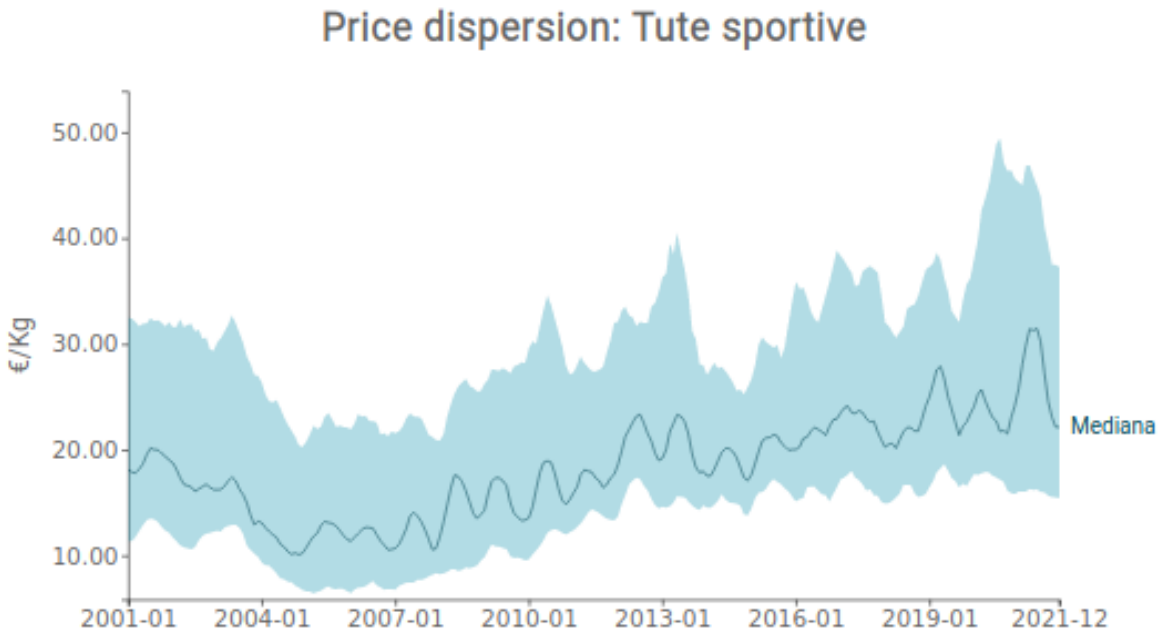

Figura 19. HS611211 - Fonte: Elaborazioni ExportPlanning

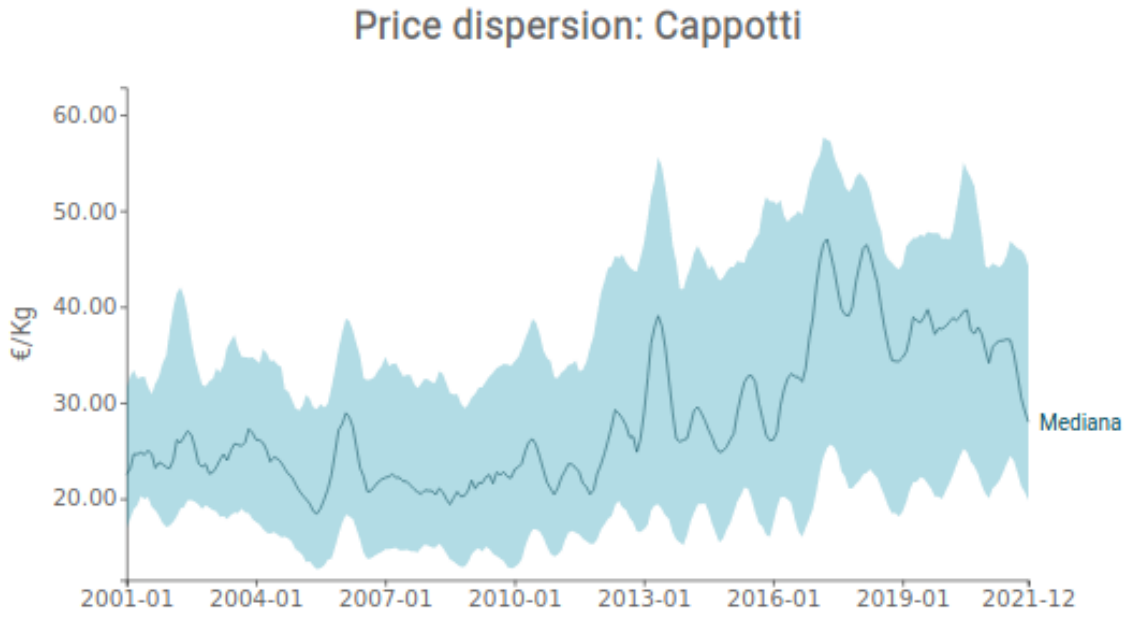

Figura 20. HS610120 - Fonte: Elaborazioni ExportPlanning

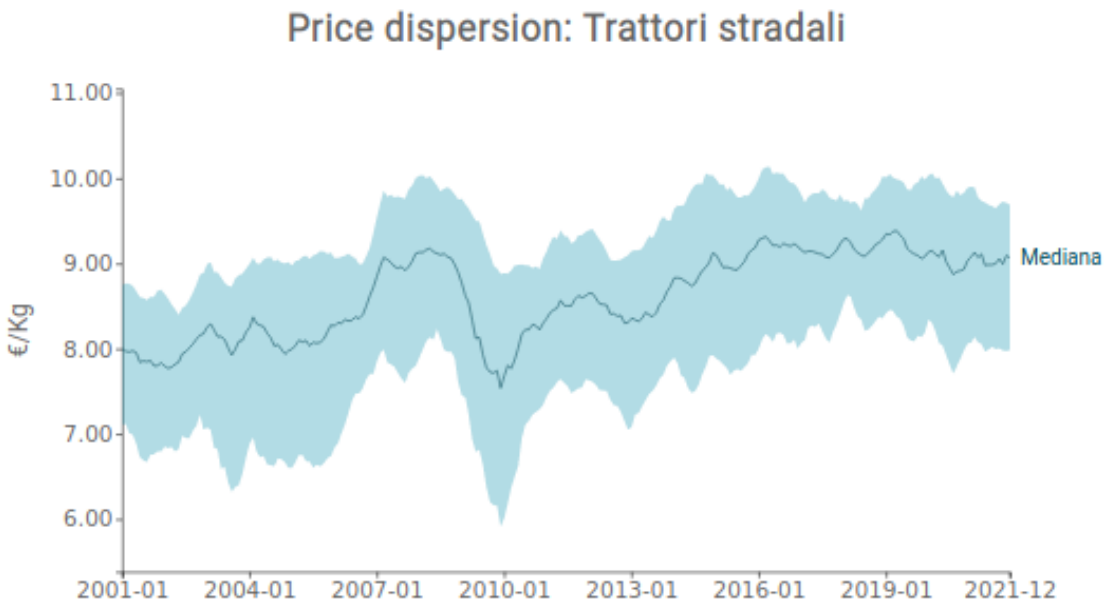

Figura 21. HS870120 - Fonte: Elaborazioni ExportPlanning

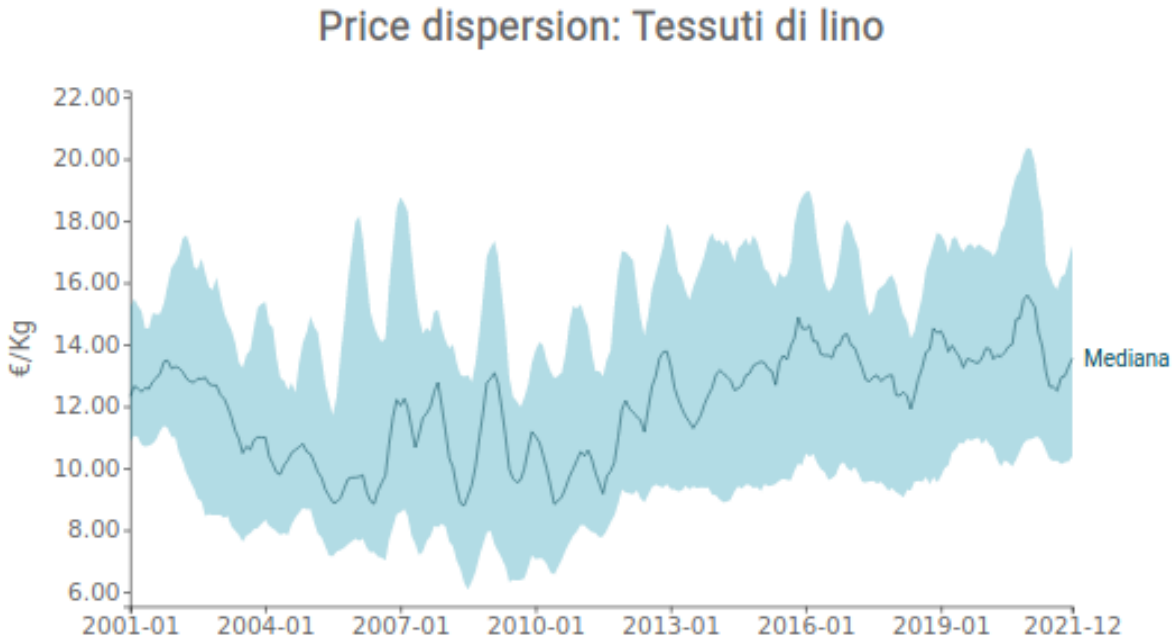

Figura 22. HS530911 - Fonte: Elaborazioni ExportPlanning

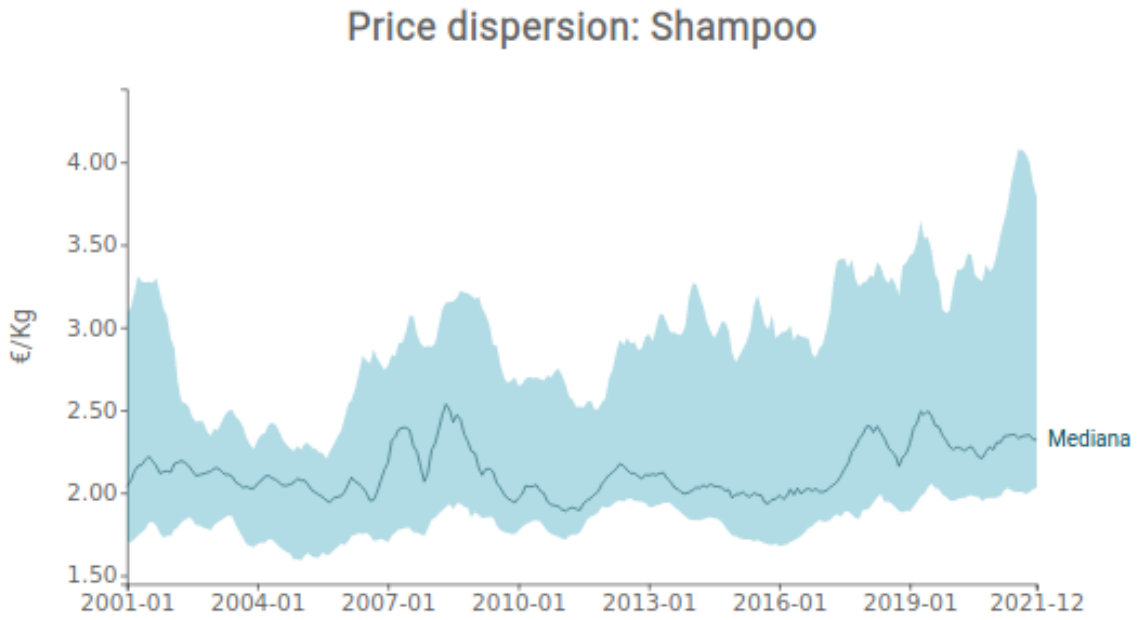

Figura 23. HS330510 - Fonte: Elaborazioni ExportPlanning

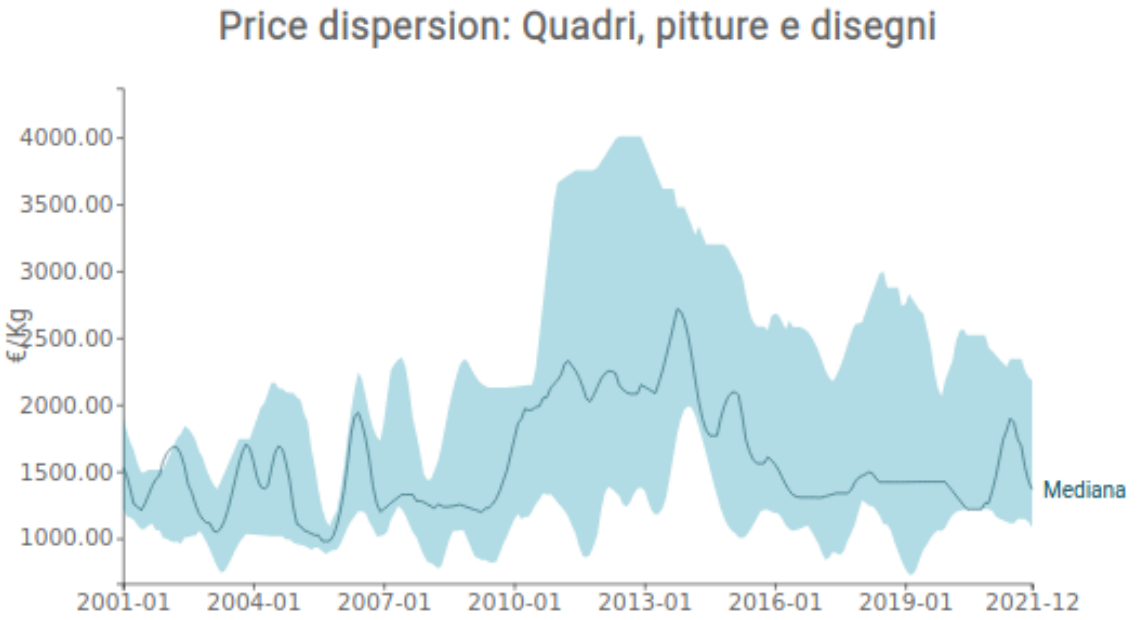

Figura 24. HS970110 - Fonte: Elaborazioni ExportPlanning

APPENDICE 2: Matrice di confusione

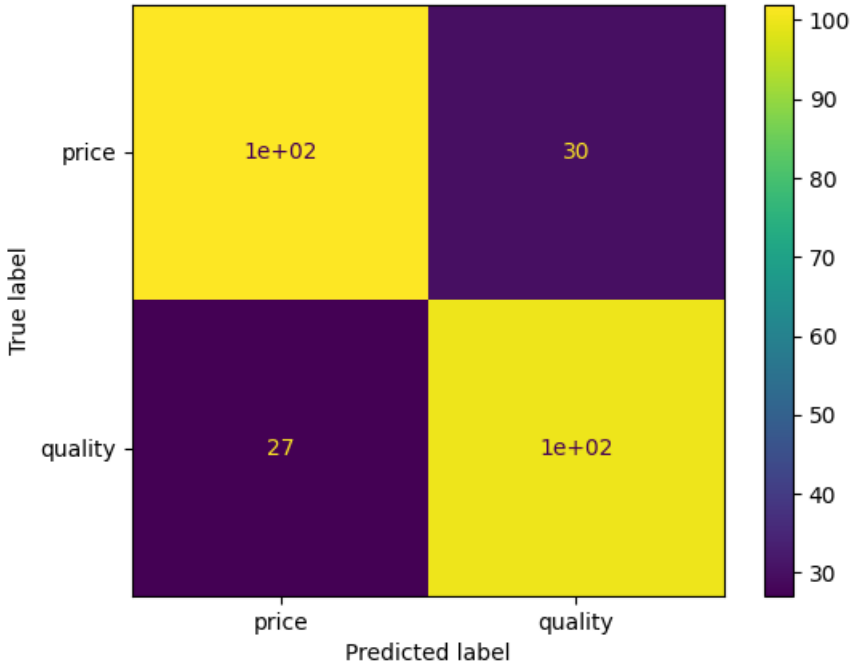

Figura 25: Matrice di confusione

Sull'asse orizzontale sono riportate le categorie previste dal modello, sul quello verticale le categorie reali del training set. Quindi, ciascun quadrante indica il numero di codici accuratamente assegnati (in giallo) o erratamente assegnati (in viola). Sul totale dei codici prodotto, 57 sono stati classificati in modo errato, di cui 27 classificati erratamente come price e 30 come quality. Pertanto, non si evidenzia una particolare distorsione di errore verso una delle due categorie.

APPENDICE 3: Risultati classificazione

Codice settore	R^2 Ulisse	R^2 comext	Descrizione settore
A1	0.64	0.64	Materie prime naturali
A2	0.69	0.69	Materie prime industriali
B2	0.53	0.53	Beni intermedi in materie tessili e pelli
B3	0.54	0.54	Beni intermedi in carta e legno
B4	0.51	0.51	Beni intermedi in metallo
B5	0.5	0.5	Beni intermedi chimici
B6	0.42	0.42	Beni intermedi in minerali non metalliferi
C1	0.45	0.45	Beni e prodotti per le costruzioni
D1	0.51	0.51	Componenti elettroniche
D2	0.75	0.75	Componenti meccaniche ed ottiche
D3	0.54	0.54	Componenti per i mezzi di trasporto
D4	0.71	0.71	Elettrotecnica
E0	0.49	0.49	Alimentari confezionati e bevande
E1	0.58	0.58	Prodotti finiti di largo consumo
E3	0.46	0.46	Prodotti finiti per la casa
E4	0.57	0.57	Prodotti e strumenti per la salute
F1	0.65	0.65	Strumenti e attrezzature per ICT e servizi
F2	0.65	0.65	Strumenti e attrezzature per l'industria
F3	0.39	0.39	Mezzi di trasporto per l'agricoltura
F4	0.5	0.5	Macchine e impianti per i processi industriali
F5	0.5	0.5	Impiantistica industriale
G1	0.55	0.55	Armi e munizioni
Media	0.55	0.5	Overall
Mediana	0.54	0.55	Overall

Tabella 19: Classificazione per settore industriale

APPENDICE 4: Prodotti settore F3

HS6	Descrizione prodotto	RANGE	CORR	GI	StudiaBo	Previsione
HS842441	Irroratrici nebulizzatrici portatili	43.22	0.76	23.60	quality	price
HS842449	Irroratrici nebulizzatrici	48.26	0.87	4.74	quality	price
HS842482	Apparecchi meccanici per l'agricoltura	40.68	0.89	22.02	quality	price
HS842710	Carrelli semoventi a motore elettrico	22.98	0.84	26.51	quality	price
HS842720	Carrelli semoventi (non elettrici)	42.64	0.87	45.11	quality	price
HS842911	Apripista su cingoli	22.84	0.90	29.68	quality	price
HS842919	Apripista su ruote	42.25	0.91	22.16	quality	price
HS842920	Livellatrici	36.02	0.89	31.30	quality	price
HS842930	Ruspe spianatrici	25.39	0.77	7.20	quality	price
HS842940	Compattatori e rulli compressori	41.40	0.48	10.05	quality	price
HS842951	Caricatori e caricatrici-spalatrici	33.53	0.85	38.75	quality	price
HS842959	Pale meccaniche, escavatori	44.40	0.82	20.37	quality	price
HS843031	Macchine per perforare trafori e gallerie	68.04	0.67	4.97	quality	price
HS843049	Macchine per l'estrazione della terra	54.82	0.65	21.84	quality	price
HS843069	Macchine per comprimere il terreno	29.21	0.76	10.39	quality	price
HS843221	Aratri per l'agricoltura, l'orticoltura	32.52	0.97	20.00	quality	price
HS843229	Polverizzatori per l'agricoltura	43.11	0.92	23.06	quality	price
HS843239	Seminatrici, piantatrici e trapiantatrici	29.70	0.58	24.17	quality	price
HS843242	Spanditori di letame	43.61	0.73	27.86	quality	price
HS843280	Distributori di concimi	43.28	0.86	21.11	quality	price
HS843330	Falciatrici	42.86	0.65	26.99	quality	price
HS843340	Macchine ed apparecchi da fienagione	28.96	0.91	23.32	quality	price
HS843351	Presse da paglia o da foraggio	29.09	0.97	39.97	quality	price
HS843352	Mietitrici-trebbiatrici	41.10	0.93	12.10	quality	price
HS843359	Macchine per la raccolta di radici o tuberi	24.36	0.86	15.37	quality	price
HS843410	Macchine per pulire uova, frutta	53.53	0.71	27.27	quality	price
HS843610	Mungitrici	41.31	0.82	22.32	quality	price
HS843629	Incubatrici ed allevatrici per l'avicoltura	51.31	0.84	15.17	quality	price
HS843680	Macchine per l'avicoltura	41.15	0.56	9.39	quality	price
HS847432	Betoniere per cemento	64.01	0.67	14.46	quality	price
HS847439	Macchine per mescolare	52.58	0.57	14.04	quality	price
HS860110	Macchine per l'edilizia	23.02	0.89	25.16	quality	price
HS860120	Locomotive e locotrattori	28.97	0.92	7.00	quality	price
HS860290	Locomotive diesel-elettriche	30.45	0.97	12.72	quality	price
HS860390	Automotrici a presa di corrente	40.51	0.80	17.50	quality	price
HS860400	Automotrici ed elettromotrici	34.59	0.75	10.20	quality	price
HS860500	Veicoli per manutenzione strade	45.05	0.87	22.00	quality	price
HS860610	Vetture per viaggiatori	16.22	0.91	4.09	quality	price
HS860692	Carrozze e vagoni di treni	10.32	0.95	0.98	quality	price
HS860699	Carri per il trasporto di merci	38.93	0.83	6.07	quality	price
HS870130	Trattori stradali per semirimorchi	15.84	0.48	13.02	quality	price
HS870191	Trattori a cingoli	36.60	0.96	43.80	quality	price
HS870192	Trattori, motore <= 18 kW	21.26	0.84	15.40	quality	price
HS870193	Trattori, motore > 18 kW	19.92	0.94	12.99	quality	price
HS870194	Trattori, 37 kW< motore <= 75 kW	21.14	0.93	24.63	quality	price
HS870195	Trattori, 75 kW<motore <= 130 kW	20.46	0.97	27.42	quality	price
HS870210	Trattori, motore> 130 kW	18.75	0.98	35.66	quality	price

HS870220	Autoveicoli >= 10 persone (diesel)	34.04	0.72	30.22	quality	price
HS870230	Autoveicoli >= 10 persone (ibride)	35.30	0.71	36.98	quality	price
HS870321	Autoveicoli < 10 persone per neve, golf	26.22	0.91	17.51	quality	price
HS870322	Autoveicoli da turismo < 10 persone	26.42	0.91	20.86	quality	price
HS870323	Autoveicoli cilindrata > 1.000 cm3 e <= 1.500 cr	18.43	0.95	40.32	quality	price
HS870324	Autoveicoli cilindrata > 1.500 cm3 e <= 3.000 cr	23.72	0.97	20.07	quality	price
HS870331	Autoveicoli cilindrata > 3.000 cm3	11.98	0.91	20.29	quality	price
HS870332	Autoveicoli cilindrata <= 1.500 cm3	19.07	0.72	42.60	quality	price
HS870333	Autoveicoli cilindrata > 1.500 cm3 <= 2.500 cm3	21.25	0.98	56.95	quality	price
HS870340	Autoveicoli cilindrata > 2.500 cm3	28.45	0.97	42.31	quality	price
HS870421	Autocarri a cassone ribaltabile	19.16	0.88	33.57	quality	price
HS870422	Autoveicoli per il trasporto di merci=< 5 T	26.27	0.93	27.12	quality	price
HS870423	Autoveicoli per il trasporto di merci> 5 T =< 20	21.19	0.68	19.64	quality	price
HS870431	Autoveicoli per il trasporto di merci > 20 T	20.46	0.61	21.41	quality	price
HS870520	Gru-automobili	6.76	0.87	17.69	quality	price
HS870530	Derricks automobili per il sondaggio	30.88	0.93	25.81	quality	price
HS870540	Autopompe antincendio	28.88	0.89	9.78	quality	price
HS870590	Autocarri betoniere	50.82	0.84	21.85	quality	price
HS870911	Autoveicoli per usi speciali	41.02	0.79	21.42	quality	price
HS871120	Motocicli cilindrata =< 50 cm3	54.83	0.47	8.92	quality	price
HS871130	Motocicli cilindrata > 50 cm3 <= 250 cm3	63.68	0.67	14.56	quality	price
HS871140	Motocicli> 250 cm3 <= 500 cm3	28.24	0.65	12.96	quality	price
HS871150	Motocicli cilindrata > 500 cm3 <= 800 cm3	13.27	0.93	21.84	quality	price
HS871160	Motocicli cilindrata 800 cm3	25.48	0.94	37.06	quality	price
HS871620	Roulotte	3.54	0.82	3.54	quality	price
HS871639	Rimorchi con cisterna	36.41	0.54	23.63	quality	price
HS871640	Rimorchi trasporto di merci	31.83	0.71	27.87	quality	price
HS880220	Elicotteri > 2 000 kg	45.52	0.98	13.29	quality	price
HS880230	Aeroplani =< 2 000 kg	36.66	0.87	18.74	quality	price
HS880240	Aeroplani > 2 000 kg, ma <= 15 000 kg	29.78	0.97	13.57	quality	price
HS880260	Aeroplani > 15 000 kg	28.10	0.97	12.88	quality	price
HS880510	Veicoli spaziali, incl. i satelliti	22.10	0.86	27.60	quality	price
HS880529	Simulatori di combattimento aereo	32.66	0.97	1.54	quality	price
HS890110	Apparecchi al suolo di allenamento al volo	19.31	0.82	39.22	quality	price
HS890130	Navi cisterna	55.40	0.92	4.14	quality	price
HS890190	Navi frigorifere (escl. navi cisterna)	35.57	0.98	8.96	quality	price
HS890200	Navi per il trasporto di merci	59.14	0.68	17.94	quality	price
HS890310	Natanti per la pesca	65.96	0.78	21.97	quality	price
HS890392	Barche a vela e panfili a vela	40.94	0.90	15.23	quality	price
HS890399	Barche e panfili da diporto o da sport	18.34	0.90	36.52	quality	price
HS890520	Draghe	43.15	0.91	13.70	quality	price
HS890590	Piattaforme di perforazione o di sfruttamento	42.23	0.76	28.26	quality	price

Tabella 20: Mezzi di trasporto e per l'agricoltura

BIBLIOGRAFIA

Baye et al., "*Price Dispersion in the Small and in the Large: Evidence from an Internet Price Comparison Site", The Journal of Industrial Economics*, Vol. 52, No. 4, Dec., 2004, pp. 463-496

Dixit, A. and Stiglitz, J. E. *"Monopolistic Competition and Optimum Product Diversity." American Economic Review*, June 1977, *67*(3), 297-308.

Grubel, H. P. Lloyd, P. *"The Empirical Measurement of Intra-Industry Trade", Economic Record*, vol. 47, n°4, 1971, p. 494-517

Kelly, W.A. Jr. *"A Generalized Interpretation of the Herfindal Index", Southern Economic Journal*, Vol. 48, No. 1 (Jul., 1981), pp. 50-57

Krugman, P. R. *"Scale Economies, Product Differentiation, and the Pattern of Trade." Journal of International Economics*, November 1979, *9*, 469-80.

_____. *"Increasing Returns, Monopolistic Competition, and International Trade." American Economic Review*, December 1980, *70*, 950-9.

Lancaster, K. *"Variety, Equity, and Efficiency", Columbia University Press*, 1979.

_____. "*Intra-Industry Trade under Perfect Monopolistic Competition.", Journal of International Economics*, 10, 1980, 151-75.

Porter, M.E. *"Competitive Advantage: Creating and Sustaining Superior Performance", The Free Press* (1985)

Schumacher, R. *"Deconstructing the Theory of Comparative Advantage." World Economic Review 2: 83-105*, 2013.

SITOGRAFIA

Vitali, G. *"L'integrazione commerciale europea e le nuove teorie sul commercio internazionale"* (2011)
http://www2.ceris.cnr.it/homedipendenti/vitali/dispense2011/dispensa%20nuove%20teorie%20commercio%20ue.pdf

Borin, A. and Lamieri, M. *"Misurare La Qualità Dei Beni Nel Commercio Internazionale"* (May 2007)
https://www.researchgate.net/publication/255996632_Misurare_La_Qualita_Dei_Beni_Nel_Commercio_Internazionale

Zhen, S. et al. *"'Law of One Price' in the Internet Era: Search Cost, Platform Competition and Customer Lock-in"*
https://www.aeaweb.org/conference/2020/preliminary/paper/hzYbDNG3

https://www.okpedia.it/indice_di_herfindahl

https://it.frwiki.wiki/wiki/Indice_Grubel-Lloyd

https://it.wikipedia.org/wiki/Apprendimento_automatico

https://www.analyticsvidhya.com/blog/2021/11/understanding-k-means-clustering-in-machine-learningwith-examples/

https://www.eage.it/machine-learning/classificazione-machine-learning

https://scikit-learn.org/stable/

https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234