POLITECNICO DI TORINO

Corso di Laurea
in Ingegneria Gestionale

Tesi di Laurea Magistrale

Progettazione modello dati e costruzione dashboard per l'ottimizzazione della scelta fornitori attraverso l'analisi di indicatori di performance

Relatore
Prof.ssa Tania Cerquitelli
Tutor Aziendale
Dott. Roberto Macrì

Candidato
Alessia Cantini

Anno Accademico 2022-2023

A babbo e mamma

Sommario

Il processo di selezione dei fornitori influenza in modo rilevante la vita di un'azienda, le spese effettuate durante il processo di approvvigionamento contribuiscono per una buona parte al costo del prodotto finale. Per questo motivo si rende necessario un processo efficace in grado di rendere minimi tempi e costi. Questo progetto nasce all'interno dell'azienda Mediamente Consulting s.r.l. e si rivolge a un'azienda cliente nel settore Manufacturing; l'obiettivo è di minimizzare costi e apportare valore aggiunto, leggendo ed interpretando i dati relativi al processo di approvvigionamento in modo da individuare i fornitori più in linea con le esigenze aziendali. Dopo una prima fase di esplorazione e analisi del contesto, si sono evidenziate quelle che erano le richieste del cliente e deciso il piano da attuare. Sono stati creati un gruppo di indicatori volti a descrivere le capacità dei fornitori in termini di tempestività, flessibilità, puntualità, prezzo e performance. I fornitori sono successivamente divisi in gruppi tramite tecniche di Data Mining. Negli step finali, il tutto si è concentrato sulla visualizzazione grafica delle informazioni ricavate in quanto le informazioni raccolte non hanno valore se non le si riescono ad interpretare. Sono state create dashboard che riassumono le analisi effettuate, dando così vita ad uno strumento interattivo, navigabile, capace di declinare i fornitori in tutte le loro dimensioni di analisi. Gli strumenti utilizzati per la buona riuscita del progetto sono stati SQL Server Management Studio, il linguaggio di programmazione Python e il tool di Data Visualization IBM Cognos Analytics.

Indice

Introduzione 5

1 Decision Support System 7
1.1 Vendor Rating . . . 9
1.2 Utilizzo dei Big Data . . . 11
1.2.1 Data Warehouse e Data Mart . . . 11
1.2.2 Normalizzazione tabelle . . . 12
1.3 Introduzione al caso studio: Litokol s.p.a. . . . 15
1.3.1 Analisi requisiti . . . 16

2 Data Science 18
2.1 Processo di estrazione della conoscenza . . . 19
2.1.1 Data Exploration . . . 20
2.1.2 Data Preprocessing . . . 20
2.1.3 Data mining . . . 21
2.1.4 Clustering . . . 22
2.2 Business Intelligence . . . 24
2.2.1 Utilizzo della BI per apportare valore aggiunto . . . 25

3 Caso studio: Litokol s.p.a. 27
3.1 Obiettivo di business . . . 27
3.2 Esplorazione dati . . . 28
3.3 Pre-processing . . . 36
3.4 Definizione KPI . . . 38
3.4.1 Indice di performance (IQ) . . . 41
3.4.2 Indice di puntualità (IPu) . . . 41
3.4.3 Indice di tempestività (IT) . . . 43
3.4.4 Indice di flessibilità (IF) . . . 44
3.4.5 Indice di prezzo (IPr) . . . 45
3.4.6 Indici di rilevanza . . . 46
3.4.7 Tabella riassuntiva KPI . . . 47

3.5 Preparazione dati per clustering 48
3.6 Clustering fornitori 51
3.6.1 Validazione clustering 57

4 Modello e integrazione dati 62
4.1 Data Integration 62
4.2 Creazione dello Snowflake Schema 64
4.2.1 Framework Manager (FM) 66

5 Data Visualization 68
5.1 IBM Cognos 69
5.2 Dashboard Globale 70
5.2.1 Navigazione interattiva della dashboard 74
5.3 Dashboard per Cluster 75
5.4 Dashboard di Dettaglio 77
5.4.1 Navigazione interattiva della dashboard 79

6 Conclusioni 82
6.1 Implementazione e sviluppi futuri 82

Acronimi 84

Elenco delle figure 85

Elenco delle tabelle 87

A Creazione tabella per tipologia pagamento 88

B Creazione tabella riasuntiva KPI 90

C Codice Python per cluster 91

Riferimenti bibliografici 99

Introduzione

Questo elaborato ha lo scopo di descrivere le attività che sono state svolte durante il tirocinio presso l'azienda Mediamente Consulting con sede in Corso Francesco Ferrucci 112 a Torino. Tante sono le attività che questa azienda svolge per soddisfare le richieste di una clientela proveniente da aree di mercato diverse, ognuna con le proprie esigenze e caratteristiche. Il loro punto di forza è la presenza di un forte know-how interno e la suddivisione in unità tematiche con competenze diverse; le aree in cui si divide Mediamente consulting sono: Corporate Performance Management, Advanced Analytics, Business Intelligence, Data Integration e Management e Infrastruttura Tecnologica. Io ho lavorato a stretto contatto con coloro facenti parte della business unit che si occupa di Advanced Analytics e Business Intelligence. Lo scopo del tirocinio e di stesura della tesi è stato quello di partecipare ad un progetto al fine di soddisfare una esigenza di una delle aziende clienti di Mediamente. Le attività di formazione eseguite sono infatti state finalizzate allo studio degli strumenti con i quali lavorano ogni giorno i dipendenti di Mediamente e che sono stati successivamente utilizzati anche da me per lo svolgimento delle mie attività. Essendo una società di consulenza informatica, la sua mission è quella di aiutare le aziende nella gestione dei dati e sviluppo del reparto IT, per questo motivo in questa tesi parleremo di un'azienda italiana inserita nel campo manufacturing.

Il progetto che ho seguito insieme al team è nato a seguito della chiusura di un precedente progetto sempre con la stessa azienda, in quel caso è stata costruita una dashboard per controllare i dati relativi alle vendite con un focus sul mercato italiano e uno su quello estero, andando a osservare nel dettaglio i clienti e i prodotti che essi acquistavano. Dopo la conclusione di questo progetto è stato richiesto anche qualcosa che, separatamente, analizzasse i dati degli acquisti. L'obiettivo di questa tesi è quindi di analizzare, rendere leggibili e più velocemente comprensibili, i dati relativi al processo di approvvigionamento. Dopo un primo periodo di formazione si è iniziato ad analizzare i dati provenienti dai database dell'azienda, si è creato un insieme di KPI efficaci a misurare gli andamenti real time e nel tempo delle performance dei fornitori. Grazie a questi indicatori è stato possibile suddividere i fornitori in gruppi in modo unire coloro con caratteristiche e

dividerli da coloro che invece hanno performance diverse. In questo modo si vuole aiutare l'azienda a prendere decisioni riguardanti la scelta di fornitori che, in base a tempistiche e prezzi, siano migliori per rifornire un determinato prodotto in un determinato periodo. Per fare questo si sono utilizzate tecniche di data mining, in particolare parleremo dell'utilizzo degli algoritmi di clustering. Alla fine del processo è stata creata una dashboard di visualizzazione dati per dare in mano alla unità di decision making dell'azienda un potente strumento in grado di mostrare l'andamento dei dati con un insieme di grafici e tabelle. Il risultato finale è di poter selezionare i fornitori più adeguati in base a puntualità, flessibilità, tempestività, prezzo e performance e capire se le risorse economiche dell'azienda sono destinate a fornitori più o meno performanti.

La tesi si divide in 6 capitoli: il primo capitolo racchiude la teoria legata al Decision Support System dell'azienda e l'importanza che con il passare degli anni viene data alla raccolta di dati in tutte le aree aziendali; in questo capitolo verrà inoltre introdotta l'azienda a cui faremo riferimento e le richieste che hanno portato all'avviamento del progetto. Il secondo capitolo spiega le tecniche utilizzate per estrapolare le informazioni dai dati raccolti. Il terzo capitolo invece racchiude tutte le attività svolte nella prima parte di stesura della tesi che hanno portato alla creazione del portafoglio di indicatori selezionati per le misurazioni delle performance dei fornitori e la descrizione e la motivazione sugli algoritmi di clustering utilizzati. Il quarto e quinto capito si soffermano invece sulla seconda parte di attività svolte che comprendono la creazione del modello dati utilizzato, proveniente principalmente dal Data Warehouse dell'azienda, e la descrizione delle dashboard create grazie agli strumenti di Business Intelligence che il mercato propone per visualizzare i dati. Per finire vi è il capitolo relativo alle conclusioni e all'utilizzo futuro che potrà fare l'azienda con i risultati ottenuti dal lavoro.

Capitolo 1

Decision Support System

In questo capitolo verranno descritte le componenti del sistema di supporto alla decisioni o Decision Support System (DSS). Le aziende stanno modificando il modo in cui prendono le decisioni, per emergere nel mercato è necessario conoscere bene le performance dei processi operativi, tenerli sotto controllo e modificarli nel caso stiano andando in una direzione sbagliata. Per riuscire a fare questo si entra nel mondo *data driven*, un contensto in cui tutte le decisioni sono prese sulla base di dati raccolti nel tempo. In particolare, nei capitoli a seguire, vedremo come questo processo viene applicato per supportare le decisioni dell'azienda Litokol s.p.a. (meglio descritta nel capitolo 1.3), analizzando nel dettaglio le sue esigenze di acquisto giornaliero di materiale per le attività operative. Parleremo dell'importanza che al giorno d'oggi assumono i dati all'interno dell'organizzazione e come essi possono essere storicizzati per essere facilmente accessibili ed analizzati. Il DSS nasce per indicare tutta una serie di tecniche e metodologie informatiche volte ad analizzare i dati per scopi decisionali. E' importante che le aziende prendano decisioni sulla base dei dati raccolti giorno per giorno in modo "intelligente". Il termine viene usato per la prima volta negli anni '70 e indicava un sistema informatico a supporto del processo decisionale che negli anni a venire ha assunto significati sempre più ampi. Le componenti di questo sistema a supporto delle decisioni, come indica la figura 1.1, sono il data warehouse, gli strumenti di back-end, per il trattamento e la preparazione del dato, e gli strumenti di data mining e business analitycs per l'analisi e la visualizzazione dei dati. Nel seguito di questo elaborato verranno ripresi tutti questi concetti perché serviranno al raggiungimento dell'obiettivo finale.

Il processo di supporto alle decisioni assiste tutti i livelli della vita aziendale nelle decisioni che ogni giorno devono essere prese all'interno dell'organizzazione, che siano esse scelte strategiche, operative o di qualsiasi altro genere. Non esiste al giorno d'oggi una decisione che non sia presa sulla base dell'osservazioni di dati, il processo di raccolta dei dati è infatti diventato di fondamentale importanza. Ogni livello aziendale necessiterà di analizzare dati provenienti da fonti diversi in grado

di descrivere il processo in tutte le sue parti; perciò, un buon sistema DSS deve essere in grado di suddividere e elaborare dati in modo funzionale all'utilizzo che l'utente finale ne dovrà fare. Per fare un esempio, il top management lavorerà con dati aggregati che possono mostrare l'andamento generale delle attività o una previsione di ciò che accadrà in futuro; mentre livelli più bassi dell'organigramma lavoreranno con dati di dettaglio e specifici del settore a cui si fa riferimento.

Il Decision Support System è diventato un punto chiave per la competitività aziendale. Il mondo sta diventando sempre più data-driven[1], i dati sono analizzati, studiati e confrontati con quelli passati. Un singolo dato non dà alcuna informazione ma più dati, se trattati con le tecniche giuste, possono consentire di ottenere una miriade di informazioni, il trucco è quello di osservare come essi cambiano in base all'orizzonte temporale, in base alla distribuzione geografica e così via. Il DSS ha integrato negli ultimi anni strumenti di business intelligence e data analytics per consentire di fare analisi sempre più complesse e leggere i dati in modo veloce. Grazie a questo sistema è ora possibile creare file di report in aggiornamento real time che scaturiscono dall'analisi di grandi quantità di dati e questo avviene ad una velocità sempre crescente, che un essere umano non potrebbe replicare.

Come si è detto il DSS ha il compito di accompagnare tutte le sfaccettature della vita aziendale: in particolare ha inizio con l'individuazione di un problema, segue dopo la fase di implementazione delle diverse soluzioni alternative per la sua risoluzione e infine vi è la scelta di quella più adeguata a migliorare la vita aziendale.

In questo elaborato mi soffermerò nell'analisi della prima parte della supply chain management di Litokol s.p.a.; cioè quella che ha in compito di controllare il processo di approvvigionamento di materiale e quindi prendere decisioni sulla selezione del portafoglio di fornitori adeguato. Questo processo prende il nome di Vendor Rating ed è utilizzato dai buyer per capire se un fornitore è adatto a rifornire un determinato prodotto all'azienda. I fornitori vengono descritti e valutati attraverso indicatori di vario tipo che uniti formato un unico valore descrittivo del fornitore e delle sue caratteristiche in modo da creare una classifica finale e poter scegliere i venditori con Vendor Rating più alto.

In figura 1.1 viene mostrato il processo di gestione del dato. Come prima cosa i dati vengono prelevati da una o più sorgenti esterne. Una volta trasformati e resi consistenti gli uni con gli altri vengono caricati all'interno del Data Warehouse e successivamente suddivisi in più piccoli Data Mart tematici per poi essere analizzati attraverso operazioni di business intelligence, ad esempio, creazione di report o visualizzazione tramite grafici e istogrammi (per approfondimento si veda il capitolo 1.2.1).

[1] Approccio scientifico basato sui dati

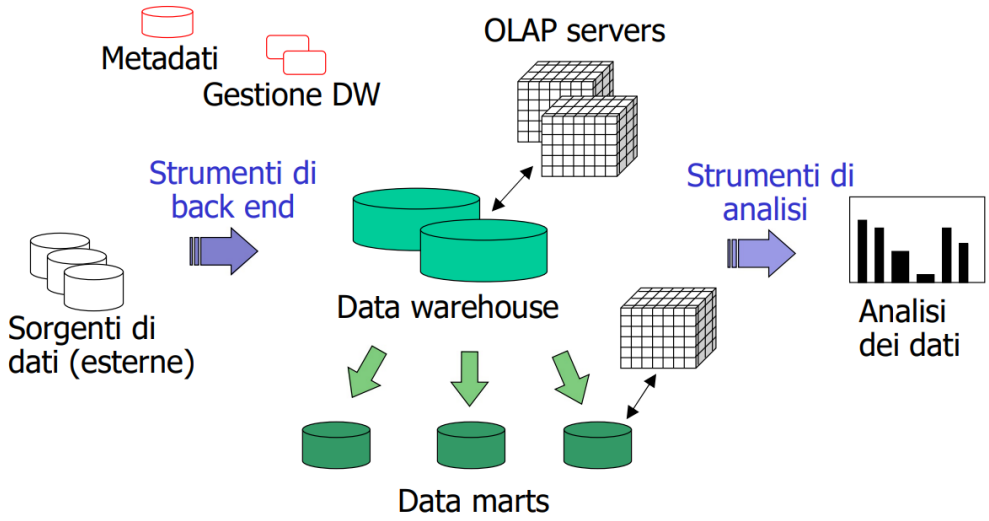


Figura 1.1: Struttura Data Warehouse. Fonte: slide corso Business Intelligence

1.1 Vendor Rating

Il vendor rating è uno strumento di valutazione della performance della supply chain che si basa essenzialmente sulla valutazione dei fornitori per indicare quelli che possono garantire un servizio idoneo alle esigenze dell'azienda. Essere in grado di ottenere dal mercato un prodotto della giusta qualità e/o del giusto prezzo, nella giusta quantità e nel giusto tempo è una caratteristica di elevata importanza. Emerge infatti che il costo delle materie prime contribuisce per le aziende per circa il 70% del costo del prodotto finale. Un buon processo di vendor rating può quindi contribuire all'accrescimento del margine di profitto aziendale ed essere in grado di:

1. scovare e mantenere i fornitori migliori;
2. quantificare una molteplicità di aspetti tradizionalmente gestiti in modo slegato;
3. rendere trasparente il rapporto con i fornitori;
4. monitorare l'evoluzione del fornitore.

Questo processo ha tre effetti diretti: finanziari, in quanto gli acquisti rappresentano una parte significativa della spesa di produzione totale; operativi, in quanto cattive prestazioni al lato acquisti possono portare a fermi della catena di produzione; e competitivi, in quanto l'efficacia del processo di approvvigionamento può dare un buon vantaggio nel segmento di mercato in cui opera l'azienda.

Un ruolo importante nella elaborazione del vendor rating è stato quello di Dickson (1966) nel suo lavoro "An analysis of vendor selection systems and decisions" nel quale definisce 23 criteri di valutazione dei fornitori, tra questi ci sono sia criteri qualitativi che quantitativi. A questi criteri è stato successivamente assegnato un punteggio e infine, sono stati ordinati e classificati in base al loro livello di importanza. Dal suo studio è emerso che il criterio per eccellenza debba essere la qualità, seguita dal delivery, le performance storiche, le garanzie e le politiche di reclamo, la capacità produttiva, il prezzo e così via... L'ultima posizione è invece assegnata agli accordi reciproci e alle future forniture. Lo studio di Dickson risale al 1996, negli anni alcune esigenze delle aziende sono cambiate, ma i suoi 23 criteri rimangono sempre una buona base di partenze per chi si appresta a implementare un sistema di valutazione fornitori.

Tre sono i metodi della letteratura per la valutazione dei fornitori: il metodo della media ponderata semplice, il metodo categorico e il cost-ratio plan. Il primo consiste nella individuazione di una serie di criteri che saranno usati per determinare un criterio unico finale; il secondo consiste nel categorizzare ogni criterio in base a classi di bontà del servizio offerto (esempio: "buono", "neutrale" e "insoddisfacente") dopodiché viene calcolato un overall finale sulla base del numero di criteri per ogni classe; mentre l'ultimo consiste nella individuazione del costo sostenuto per ogni acquisto sommando poi il costo dei servizi aggiuntivi come la qualità, il delivery e il tipo di servizio offerto. Per questa tesi è stato scelto di utilizzare il primo criterio, in quanto è più preciso del metodo categorico ma senza la rigidità del cost-ratio.

Il primo step del vendor rating è quello di stabilire i criteri con i quali i fornitori verranno valutati. Ogni fornitore è valutato utilizzando tutti i criteri selezionati, e il criterio generale viene calcolato come somma pesata di tutti i criteri singoli. I pesi assegnati ai criteri sono dati in base all'importanza che la singola azienda dà ad ognuno di essi a seconda della loro rilevanza nella attività operativa (la somma dei pesi fra loro sarà pari a 100). Questi pesi non possono essere calcolati sulla base di ricerche, ma devono essere forniti direttamente dal reparto che si occupa della valutazione fornitori, in quanto essi sono diversi a seconda dell'azienda e del modo in cui vengono utilizzati.

Per scegliere gli indicatori più adeguati a descrivere le performance dei fornitori della azienda siamo partiti dall'analisi dei dati che avevamo a disposizione. I fornitori verranno valutati secondo tre diversi punti di vista: tempistiche, costi e tipologia del servizio offerto; osservando i dati raccolti negli ultimi cinque anni (periodo di osservazione) degli acquisti effettuati.

Alla fine del processo l'azienda sarà quindi in grado di dividere i fornitori in gruppi più o meno performanti e decidere azioni correttive da attuare per rendere più efficiente il rapporto di fornitura con uno specifico fornitore.

1.2 Utilizzo dei Big Data

Chi prende le decisioni in azienda ha bisogno di un'enorme quantità di informazioni. Queste informazioni al giorno d'oggi derivano dai dati. La sfida più importante che i decision maker si trovano ad affrontare è appunto quella di trasformare i dati in informazioni utili a sostegno delle decisioni. Il volume dei dati che, dagli ultimi anni, possiamo immagazzinare è cresciuto enormemente e per questo si è iniziato a parlare di BIG DATA. Le aziende di oggi, generano continuamente dati che, per essere utilizzati, devono essere storicizzati utilizzando infrastrutture hardware specializzate; queste strutture prendono il nome di Data Wharehouse. Per big data si fa riferimento a tutti quei dati caratterizzati da ("Flyip", 2022):

1. Volume: i dati raccolti crescono molto rapidamente e raggiungono grandi volumi;

2. Velocità: fa riferimento alla raccolta dei dati nel tempo. L'elevata velocità di generazione porta necessariamente ad una elevata velocità di processamento del dato;

3. Varietà: fa riferimento alla natura dei dati; i quali possono essere strutturati, semi-strutturati e non strutturati;

4. Veridicità: fa riferimento alla qualità del dato. Quando si raccoglie un dato è necessario anche controllare che esso sia in linea e conforme al dominio di appartenenza. Raccogliere dati senza guardarli porta solo costi e non benefici;

5. Valore: fa riferimento alla capacità di trasformare l'output dell'elaborazione in azione che porta valore aggiunto per chi la esegue.

1.2.1 Data Warehouse e Data Mart

Tra il sistema gestionale che utilizza l'organizzazione per tenere traccia delle sue attività e il processo di decision making sta il Data Warehouse. Un Data Warehouse può essere centralizzato o decentralizzato; centralizzato significa che vi è un singolo database fisico che contiene tutti i dati per il supporto decisionale ed è quindi separato dai database operazionali dell'azienda. Essendo la quantità di dati presente nel data warehouse elevata, viene spesso alleggerito il carico di lavoro dividendo i dati in aree tematiche che prendono il nome di Data Mart. Un Data Mart è quindi un database tematico che contiene solo una parte dei dati raccolti e che fanno riferimento ad una singola area aziendale. In questa tesi verranno analizzati solo i dati relativi agli acquisti dell'azienda e perciò sarà creato un Data Mart con all'interno dati relativi agli ordini, ai fornitori e ai prodotti acquistati. Il Data Mart segue la tecnologia di database relazionali, è perciò organizzato in

tabelle in cui ogni riga rappresenta un record ed è caratterizzata da una serie di attributi organizzati in colonne.

I dati all'interno del Data Warehouse devono rispettare determinate caratteristiche:

1. Orientamento al soggetto: cioè organizzati per aree tematiche. I dati si riferiscono a soggetti rilevanti (prodotti, clienti, fornitori, date) e tutti quelli che non rientrano in questi soggetti vengono eliminati dalle analisi. Per prendere la decisione su quali dati usare e quali no, si deve tener conto di chi utilizzerà quei dati a valle dell'attività di analisi;

2. Integrati: cioè i dati devono essere omogenei e consistenti, eliminando i problemi di incompatibilità;

3. Riutilizzabili: cioè i dati devono essere conservati in modo accurato in modo da poterli utilizzare in qualsiasi momento;

4. Non volatili: cioè i dati devono essere conservati permanentemente, non possono essere modificati così da mantenere la loro integrità nel tempo.

Per la progettazione logica del Data Mart viene creato uno schema a stella. Il nome deriva dalla posizione che assumono le diverse tabelle che compongono lo schema. Al centro si trova la tabella dei fatti, cioè la tabella in cui vengono storicizzate tutte le transazioni e, collegate a questa, ci sono le tabelle delle dimensioni, in cui vengono descritti più nel dettaglio alcuni degli attributi presenti. La particolarità di questo schema è che la tabella dei fatti avrà dimensioni molto grandi, mentre le tabelle delle dimensioni avranno dimensioni più piccole. Questa ultime tabelle hanno un solo collegamento con la tabella dei fatti grazie all'utilizzo di chiavi primarie e chiavi esterne. Le tabelle delle dimensioni hanno una chiave primaria che va a identificare univocamente i record di quella tabella, mentre la tabella dei fatti ha una chiave primaria formata dall'unione di tutte le chiavi esterne che richiamano le tabelle dimensioni. L'obiettivo dell'utilizzo di questo schema è quello di ottimizzare la lettura e la comprensione dei dati, evitando ripetizioni delle informazioni e problemi di aggiornamento. Spesso è necessario che anche le tabelle delle dimensioni abbiamo delle tabelle a loro collegate e, in questo caso, si parla di snowflake schema.

1.2.2 Normalizzazione tabelle

La normalizzazione del database, nel linguaggio informatico, prevede la creazione di tabelle e relazioni al fine di:

1. ridurre la ridondanza: ad esempio, se all'interno della tabella relativa agli acquisti sono presenti più record associati allo stesso fornitore non viene salvato ogni volta il nome del fornitore ma semplicemente un numero, o chiave, che si riferisce a quel fornitore nella tabella di anagrafica fornitori;

2. ridurre la dipendenza incoerente: ad esempio, all'interno della tabella relativa agli acquisti non ha senso inserire le informazioni relative all'indirizzo email del fornitore, questo tipo di dati deve essere inserito nella tabella relativa all'anagrafica fornitori.

Perciò, per rendere un database normalizzato, è necessario avere più tabelle delle dimensioni in cui ogni record è identificato dalla sua chiave che viene poi richiamata nella tabella dei fatti. Ad esempio, nella nostra tabella dei fatti ad ogni acquisto è associato un fornitore; nella tabella dei fatti non ci sono tutti gli attributi relativi a quel fornitore ma solamente la chiave esterna della tabella da cui andare ad estrapolare le informazioni. Per far comunicare la tabella dei fatti con la tabella dei fornitori basterà quindi andare ad effettuare una inner join. La normalizzazione ha anche una conseguenza indiretta che consente di ritrovare più velocemente le informazioni relative ad un oggetto in quanto saranno presenti nella tabella dell'oggetto e che ha tutti gli attributi necessari a descriverlo. Per rendere un database normalizzato ci sono tre regole che corrispondono a tre livelli di normalizzazione:

1. Prima forma normale: consiste nella creazione della tabella fornitori collegata alla tabella degli acquisti. Se viene aggiunto un nuovo fornitore questo verrà inserito nella relativa tabella.

2. Seconda forma normale: consiste nel creare una tabella separata anche per quei valori richiesti da record provenienti da tabelle diverse. Ad esempio, il codice prodotto sarà collegato sia alla tabella degli acquisti, sia al fornitore che lo produce, il codice prodotto è quindi inserito in una tabella separata con la relativa chiave primaria;

3. Terza forma normale: creare tabelle separate per raccogliere informazioni aggiuntive rispetto a quelle presenti nelle altre tabelle. Ad esempio, nella tabella dei fornitori inserisco tutti coloro che hanno un rapporto di fornitura con l'azienda, anche se le relative chiavi primarie non vengono mai richiamate nella tabella di acquisto materie prime. In questo modo posso ricavare l'elenco di tutti i fornitori.

Queste regole sono indipendenti fra loro e non sono sempre applicabili, per capire quale usare per facilitare e rendere più flessibile l'analisi si deve tener conto del contesto di riferimento.

Table 1: Tabella Acquisti

NUM_ORDINE	DATA_ORDINE	KEY_FORNITORE	KEY_PRODOTTO
FO22R0100262	2022-03-22	2409	19799
FO22R0100511	2022-06-08	2504	7264
FO22R0100542	2022-06-15	5336	3645
FO22R0100644	2022-07-12	2433	16119
...	...	...	...

Table 2: Tabella prodotti

KEY_PRODOTTO	COD_PRODOTTO	DES_PRODOTTO	KEY_CATPRO08
3645	ATR453	INCHIOSTRO NERO A BASE ACQUA	3
7264	MPC0865.1	ORGAL ROOFLEX S6	15
16119	MPS0000	SABBIA SILICEA MEDIA	17
19799	SCHBLT010	SECCHIO BILTEMA LA315 5KG	12
...	...	...	...

Table 3: Tabella Fornitori

KEY_FORNITORE	COD_FORNITORE	DES_FORNITORE	INDIRIZZO
2409	2034780	Nome1	VIA —-
2433	0003457	Nome2	VIA —-
2504	2072185	Nome3	VIA —-
5336	2082275	Nome 4	VIA —-
...	...	...	...

Figura 1.2: Tabella Acquisti, Prodotti e Fornitori. Fonte: Elaborazione personale.

Le tabelle in figura 1.2 sono un esempio di tabelle normalizzate presenti nel database di Litokol, per estrarre le informazioni da queste tabelle è stato utilizzato SQL Server Management Studio[2]. La tabella Acquisti mostra tutti gli acquisti effettuati con il relativo numero ordine, la data di effettuazione ordine, il fornitore a cui è stato ordinato il materiale e il prodotto ordinato, questi ultimi richiamati tramite la chiave esterna relativa al record delle rispettive tabelle di anagrafica. La tabella Prodotti è invece formata dalla chiave primaria relativa al prodotto, il suo codice, la descrizione testuale e la categoria di appartenenza del prodotto, anche questa richiamata tramite l'utilizzo di chiavi esterne. La tabella Fornitori ha invece la chiave primaria che identifica il singolo fornitore, il codice del fornitore, il nome del fornitore e l'indirizzo della sede di appartenenza del fornitore (il nome e l'indirizzo non possono essere visualizzati chiaramente perché dati sensibili). Le tabelle riportate sono solo una parte di quelle presenti nel database, le quali hanno

[2]Applicazione software rilasciata da Microsoft che permette di leggere e scrivere tabelle con l'utilizzo del linguaggio SQL

molti più attributi per ciascun record[3].

1.3 Introduzione al caso studio: Litokol s.p.a.

Questo elaborato si è svolto in collaborazione con l'azienda di consulenza Mediamente Consulting. Come azienda di consulenza il suo obiettivo è quello di gestire e valorizzare i dati dei suoi clienti al fine di portare maggiore conoscenza riguardo all'andamento del loro business nel mercato e in confronto con i competitor. Per questa tesi ci soffermeremo su Litokol s.p.a., un'azienda italiana nel campo manifatturiero che produce, realizza e commercializza prodotti per l'edilizia sia in in modalità B2B[4] che B2C[5]. Litokol nasce nel 1968 a Rubiera, città dell'Emilia Romagna, centro del comprensorio ceramico. In quegli anni il mercato degli elementi e materiali di costruzioni si basavano sulle conoscenze tradizionali ed era restio a cambiare il suo punto di vista. Con idee innovative Litokol è invece riuscita ad affermarsi in campo nazionale ed internazionale nel mercato dell'industria ceramica. Già a partire dagli anni '70 Litokol dà avvio a rapporti internazionali che porteranno l'azienda ad aprire sedi produttive in diversi stati come l'Ucraina, la Russia, l'Armenia e la Cina. Negli anni ottiene diversi brevetti per le invenzioni di materiali sempre più performanti alle esigenze del mercato e ampia la gamma di prodotti offerti quali: adesivi cementizi, adesivi in dispersione e reattivi, sigillanti, sigillanti siliconici, prodotti per il trattamento delle superfici e per la impermeabilizzazione e tanti altri. La mission di questa azienda è di realizzare prodotti innovativi e altamente performanti concentrandosi sempre su nuovi prodotti, nuove tecnologie e nuovi mercati.

Negli ultimi anni ha deciso di concentrare le sue energie allo studio e analisi dei dati provenienti dal gestionale di cui fa uso in modo da estrapolare conoscenza per supportare il processo decisionale. In questo elaborato parleremo ed analizzeremo i dati relativi agli acquisti di tutti i prodotti utilizzati nei suoi processi oppure di prodotti commercializzati. Lo scopo finale sarà quello di realizzare una dashboard[6] che possa portare valore aggiunto e vantaggio competitivo all'azienda sul mercato dando la possibilità di visualizzare in maniera grafica le performance dei

[3]Riga di una tabella che identifica un oggetto con tutti i suoi attributi.

[4]Scambi commerciali che avvengo fra due aziende.

[5]Scambi commerciali che avvengono tra azienda e consumatore.

[6]Le dashboard di business intelligence sono soluzioni di visualizzazione dei dati e gestione delle informazioni, usate per analizzare i dati. I creatori del contenuto possono usare elementi interattivi, come filtri e azioni, per combinare grafici, diagrammi e report su un singolo schermo per panoramiche istantanee ("Tableau", 2022).

fornitori tenendo in considerazione tutti gli acquisti effettuati negli ultimi anni, le tempistiche e le quantità ordinate.

1.3.1 Analisi requisiti

La prima parte dello sviluppo di un progetto, e fondamentale per la sua buona riuscita, è quella dell'analisi dei requisiti. É un processo che ha come obiettivo quello di definire le aspettative degli stakeholder. Ogni cliente ha le sue specifiche richieste che devono essere ascoltate e analizzate nel dettaglio per fare in modo che l'output finale sia coerente e rispetti tutte le necessità richieste. Ogni progetto è unico, solo in questo modo è possibile assicurare un servizio customizzato e di buona qualità. L'analisi dei requisiti precede quindi la parte di sviluppo della soluzione e accompagna la fase di scelta degli strumenti e delle metriche da utilizzare durante tutto il progetto.
I requisiti selezionati sono quelli relativi al progetto di realizzazione della dashboard direzionale per il controllo degli acquisti e sono quindi inerenti a ciò di cui l'azienda ha bisogno di visualizzare nell'output finale; derivano da riunioni con il responsabile del dipartimento IT di Litokol S.p.A.:

- Acquistare materiale di buona qualità e conforme alle specifiche richieste, sia riguardo alle materie prime utilizzate, sia riguardo la spedizione;
- spedizione con tempi brevi e flessibilità nelle richieste di consegna;
- mantenimento degli accordi contrattuali;
- costi bassi (avere sconti per ordini di grandi quantità e ripetitivi);
- costo di ordinazione basso, in termini di tempo impiegato e risorse utilizzate;
- facilità di comunicazione con il fornitore dal momento dell'acquisto al momento della consegna;
- dilazione di pagamento.

La fase di analisi dei requisiti è proseguita da diverse comunicazioni con il responsabile del dipartimento IT prima citato per definire quei requisiti che potevano essere misurati grazie ai dati presenti nel gestionale utilizzato dall'azienda. Da questa comunicazione è emerso che:

- non è presente un vero e proprio accordo per assistenza post-vendita o assicurazioni di qualità con i fornitori;

- non è presente la data di entrata merce in magazzino, le date importanti da tenere in considerazione sono quella di ordine[7], la data di consegna richiesta[8] e la data di consegna prevista[9];
- le fatture sono registrate con il codice articolo, perciò, il prezzo unitario è scorporato dal resto;
- la maggior parte delle spedizioni è Franco Destino[10];
- la gestione dei resi non è di interesse in questo momento;
- non è necessario avere sulla dashboard indicazioni riguardanti la comunicazione con il fornitore;

A questo punto è iniziata la fase di analisi dei dati presenti nel Data Warehouse (DWH) di Litokol, con la finalità di definire e calcolare i Key Performance Indicator (KPI) più adatti alla riuscita del progetto.

[7]Data in cui viene fatto l'ordine, in cui ci si accorda quando la fornitura arriverà a destinazione.

[8]Data in cui Litokol richiede che venga consegnato il materiale.

[9]Data in cui il fornitore può consegnare il materiale. Può essere maggiore, minore o uguale la data richiesta.

[10]I rischi e le spese di spedizione sono a carico del fornitore fino a destinazione.

Capitolo 2

Data Science

I dati raccolti hanno come obiettivo quello di sostenere chi deve prendere le decisioni in azienda con informazioni dettagliate e complete. La Data Science è la scienza che va in aiuto all'essere umano per analizzare grossi volumi di dati al fine di estrarre conoscenza che altrimenti rimarrebbe nascosta tra le celle del database. Questa conoscenza viene poi utilizzata per: prendere decisioni, prevedere il futuro e comprendere il passato e il presente. L'enorme quantità di dati che al giorno d'oggi abbiamo a disposizione non garantirebbe in altro modo ad un essere umano di analizzarli in tempi ragionevoli, la data science invece utilizza un insieme di discipline volte a estrapolare queste informazioni:

1. Matematica e statistica: utilizzo di equazioni e formule per svolgere le analisi e valutare i risultati degli algoritmi utilizzati;

2. Programmazione: utilizzo di algoritmi e path di attività per estrarre conoscenza dai dati;

3. Conoscenza del dominio: conoscenza del settore dell'attività in cui viene svolta l'analisi.

Le tecniche necessarie dipendono dai dati che si hanno a disposizione. Non esiste la possibilità di conoscere a priori l'algoritmo giusto da utilizzate. Il data scietist è colui che ha le conoscenze e l'esperienza per utilizzare le varie tecniche di data science e riconoscere quella più adeguata al caso in analisi.

2.1 Processo di estrazione della conoscenza

In questa tesi verrà applicato un processo di estrazione della conoscenza di tipo data driven[1]. Nella figura 2.1 viene riportato graficamente il KDD Process (Knowledge Discovery in Databases), processo con il quale la Data Science riesce a estrarre conoscenza e valore dai dati.

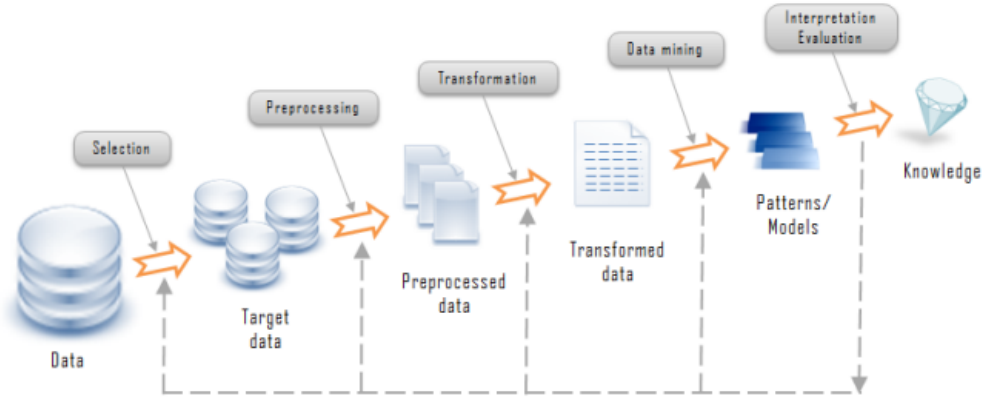


Figura 2.1: Processo KDD. Fonte: slide corso Business Intelligence

Il primo step di questo processo parte dallo studio e l'approfondimento del dominio di riferimento del progetto che deve essere eseguito, dalla conoscenza già nota a priori e dalla definizione dell'obiettivo finale. Dopodiché, da una collezione di dati più o meno grande, viene selezionato un dataset (o target data), cioè una porzione dai dati più piccola per rendere più semplice e veloce la focalizzazione sui dati utili. Una volta che si ha la porzione di dati si può procedere con la fase di preprocessing, la quale si compone di una fase esplorativa, dove i dati vengono analizzati nel dettaglio; e una fase di data cleaning, in cui invece i dati vengono puliti da eventuale rumore presente. In questa fase vengono decise delle strategie di gestione di eventuali dati mancanti o errati. Dopodiché si passa alla fase di trasformazione dei dati in cui è possibile ridurre le dimensioni di analisi per ridurre il numero di variabili da sottoporre all'analisi. A seguire, vi è la fase di scelta e utilizzo degli algoritmi di data mining che più si adeguano ai dati a disposizioni; possono essere tecniche di classificazione, di clustrering o di estrazione di regole di associazione. Una volta applicato il modello arriviamo all'estrazione della vera e propria conoscenza. Una volta applicati questi metodi si avranno a disposizione conoscenze che non erano note a priori. In fine, vi è l'incorporazione della conoscenza estratta nei processi e il report delle informazioni ricavate dal

[1]Le analisi data driven sono quelle che estraggono conoscenza direttamente dai dati

KDD con la presentazione dei risultati ottenuti all'esperto di dominio, il quale deciderà se utilizzare tale conoscenza nell'operatività aziendale.

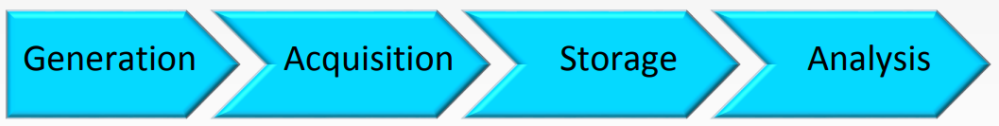


Figura 2.2: KDD Fasi. Fonte: slide corso Business Intelligence

2.1.1 Data Exploration

La fase iniziale di un processo di data science consiste nell'acquisire conoscenza sui dati da analizzare. Questa fase prende il nome di data exploration. L'esplorazione dei dati consiste nel caratterizzare i dati in termini di struttura, distribuzione e analisi delle relazioni, l'obiettivo finale è quello di capire l'utilizzo che può essere fatto con i dati a disposizione. Maggiore è l'attenzione che dedichiamo a questa fase più sarà semplice effettuare le attività successive. Ogni variabile va caratterizzata cioè si deve conoscere la sua distribuzione e le statistiche di base come valore massimo, minimo, valore medio, deviazione standard e così via. Le analisi possono essere monovariate, cioè focalizzate su un attributo per volta, o analisi multivariate, cioè che prendono in considerazione più attributi diversi contemporaneamente. La data exploration si compone di:

1. Outlier detection: individuazione ed eliminazione di tutti i valori errati che portano rumore all'interno del dataset;
2. Correlation analysis: descrizione delle relazioni principali tra attributi;
3. Feature Engineering: trasformare i dati in un formato più utile alle analisi.

Alla fine del processo è possibile individuare trend o eventi nascosti che prima non era possibile conoscere. Il risultato di questa fase può essere mostrato attraverso l'utilizzo di strumenti di visualizzazione.

In questo elaborato lavoreremo con dati esterni all'azienda di consulenza che provengono da un'altra organizzazione; la fase di esplorazione ha quindi un valore molto alto per conoscere quali dati sono a nostra disposizione, in che numero, in che modo e con quale frequenza questi sono registrati. Quando questa conoscenza è difficile da ottenere si comunica con l'esperto di dominio dell'azienda cliente.

2.1.2 Data Preprocessing

Una volta selezionati i dati che saranno utilizzati per effettuare le analisi e per raggiungere gli obiettivi preposti è necessario prepararli in quanto, all'interno del

database, ci possono essere dati errati o mancanti che possono portare ad analisi non corrette. É quindi compito dell'analista preparare i dati e renderli pronti per l'utilizzo al fine di preservarne la fiducia nella conoscenza estratta da essi. Questa fase prende il nome di data preprocessing. Le attività principali del data preprocessing sono:

- data cleaning: cioè analizzare la qualità dei nostri dati per capire se il contenuto informativo è corretto o se presenta alcuni errori ad esempio rimuovere duplicati, riempire dati mancanti e eliminare outlier. L'obiettivo è quello di diminuire il rumore dei dati e risolvere le inconsistenze. Per fare questo sono necessarie conoscenze di tecniche statistiche e del dominio delle variabili;
- data integration: si rende necessaria quando si hanno dati provenienti da sorgenti diverse che raccolgono dati con natura e frequenza diversa. Le fonti potrebbero ad esempio essere da database, fogli di calcolo o documenti di testo. Lo scopo è quello di creare un unico luogo di dati;
- data transformation: questo serve a trasformare i dati in un formato più utile alle tecniche di data mining, ad esempio, la normalizzazione dei dati o la creazione di variabili dummy;
- data reduction: attività che ha lo scopo di ridurre il set di dati, selezionando solo quelli utili ai processi di data mining. Ad esempio, possono essere fatte attività di feature selection o feature extraction[2];
- data discretization: la discretizzazione ha lo scopo di trasformare dati continui in dati categorici. Anche in questo caso si rende utile quando le tecniche di data mining utilizzate lo richiedono.

2.1.3 Data mining

Il data mining è quella parte di data science che si occupa di analizzare dati con l'utilizzo di vari algoritmi alla scoperta di schemi o modelli per trovare le relazioni presenti fra i dati. Ciò significa prevalentemente estrarre conoscenza, precedentemente non nota, dai dati. La data mining consente di analizzare dati quantitativi, qualitativi, testuali, immagini e suoni su più variabili a disposizione. L'estrazione della conoscenza può avvenire tramite (Mikut & Reischl, 2011):

1. Apprendimento supervisionato: dove la variabile di output è conosciuta; ogni record utilizzato in input ha il relativo output che viene utilizzato per insegnare all'algoritmo le regole del modello che applicherà successivamente a quegli

[2]tecniche di riduzione delle dimensioni di un record per eliminare la ridondanza e selezionare solo i campi utili alle tecniche di machine learning

input che non hanno il relativo output. Il dataset viene quindi suddiviso in due parti: il dataset di training e quello di test. I metodi che utilizzano queste tecniche sono:

   (a) Classificazione: consiste nella predizione di classe. Ogni record appartiene ad una e una sola classe.

   (b) Classificazione fuzzy: consiste nella predizione di classe. Ogni record può appartenere ad ogni classe con un livello di appartenenza ad ognuna che varia tra 0 e 1.

   (c) Regressione: consiste nella predizione di un valore di output con l'utilizzo di serie temporali.

2. Apprendimento non supervisionato: dove la variabile di output non è conosciuta per nessuno degli input. Ci pensa quindi l'algoritmo a cercare un senso tra i dati. Tra questi metodi abbiamo:

   (a) Clustering: ricerca di gruppi di record simili tra loro utilizzando algoritmi nitidi o fuzzy.

   (b) Regole di associazione: trova gruppi di item che ricorrano frequentemente insieme negli esempi.

3. Apprendimento semi-supervisionato: dove la variabile di output è conosciuta solo per alcuni esempi.

Vedremo in seguito l'utilizzo di algoritmi di clustering applicati a questa tesi per il raggruppamento di fornitori con caratteristiche simili tra loro.

2.1.4 Clustering

Il clustering, come accennato in precedenza, è una tecnica non supervisionata che non necessita dell'etichetta di classe. L'utilizzo di questa tecnica consente di identificare gruppi di dati che hanno caratteristiche simili al fine di ottenere un partizionamento di oggetti che sono simili se appartenenti allo stesso gruppo e dissimili se appartenenti a gruppi diversi.

Le tecniche della cluster analysis si basano sul concetto di somiglianza e distanza, non funzionerebbero con unità di misura dipendenti dai dati, perciò, prima di un qualsiasi algoritmo di clustering è necessario preparare i dati attraverso normalizzazioni o standardizzazioni. La similarità, e perciò anche la dissimilarità, è calcolata secondo gradi di distanza, il grado di correlazione e il grado di associazione.

I metodi di clustering presenti in letteratura si dividono in diversi tipi a seconda della metodologia con cui dividono i record in gruppi:

- Cluster Gerarchico: crea un set di cluster nidificati sotto forma di albero gerarchico.
  - Cluster agglomerativo: all'inizio del processo ogni punto rappresenta un clusetr distinto, man mano che si procede con l'algoritmo questi vengono raggruppati secondo la similarità con gli altri cluster.
  - Cluster divisionale: all'inizio del processo tutti i punti appartengono ad un unico cluster, successivamente vengono disaggregati per formare più cluster.
- Cluster non gerarchico: detto anche clustering partizionale, poiché crea delle partizioni (gruppi) di dati massimizzando la similarità di punti all'interno di uno stesso cluster e la dissimilarità di punti appartenenti a cluster diversi.
  - K-means: è un clustering partizionale nel quale ogni cluster è associato al suo centroide e ogni punto è associato al cluster più vicino al suo centroide. Il centroide è il baricentro di tutti i punti appartenenti al cluster, non è quindi un vero e proprio punto presente nel dataset. Il numero di cluster deve essere deciso a priori.
  - K-medoid:deriva dal k-means, perciò hanno le stesse caratteristiche. La differenza sostanziale è che questo algoritmo usa come centroide un punto presente all'interno del dataset. Il K-medoid è più robusto al rumore e agli outliers.
  - Fuzzy clustering: tipo di cluster in cui ogni punto può appartenere a più di un cluster con una percentuale di appartenenza.
- Clustering density-based: algoritmi di clustering basato sul concetto di densità.
  - DBSCAN: raggruppa punti ad alto livello di intensità l'uno con l'altro. Se la distanza fra due punti è minore di epsilon (valore dato in input) questi vengono inseriti in un unico cluster.
    Parametri:
    • eps: due punti sono considerati vicini se la distanza fra loro è minore di epsilon.
    • min samples: il minimo numero di vicini che un dato punto deve avere per essere considerato all'interno del cluster.
    • metrica: è la metrica usata per calcolare la distanza tra i punti.
    Tipologie di punti:
    • Core Point: punto che ha almeno il numero minimo di punti selezionato ad una distanza minore di epsilon.

- Border Point: un punto che non ha il minimo numero di vicini selezionato ma è vicino ad un core point.
- Noise Point: punti che non sono vicini abbastanza ad un core point.

Gli algoritmi di clustering basati su densità possono essere usati anche a scopo di outlier detection[3]. L'output del cluster sarà quindi la suddivisione del dataset in gruppi: uno o più gruppi che corrisponderanno ai dati puliti e di buona qualità e uno che invece sarà formato da quei record aventi caratteristiche fuori dalla norma. I dati identificati come outlier potranno essere analizzati singolarmente per capire se sono effettivamente rumore oppure no. Si parla in questo caso di outlier detenction multivariata, dato un database con n attributi diversi, andiamo a vedere se ci sono record che presentano una distribuzione diversa rispetto agli altri.

2.2 Business Intelligence

La Business Intelligence (BI) è una disciplina che abbraccia tutti i settori di un'organizzazione combinando business analytics, data mining, visualizzazione dati, strumenti e infrastrutture per i dati e per permettere alle organizzazioni di prendere decisioni basate sui dati. Essa si pone in modo trasversale all'interno delle attività aziendale per innovare, supportare, controllare, monitorare e ridurre le inefficienze. Alla base della business intelligence vi è un processo di estrazione e gestione dei dati che si articola in 4 livelli:

1. Livello delle sorgenti;
2. Livello di alimentazione: estrazione, trasformazione e caricamento (ETL);
3. Livello del Data warehouse;
4. Livello di Analisi.

La BI racchiude tutto l'insieme di metodologie, strumenti, tecniche per la raccolta e l'analisi dei dati. Più i dati sono presenti in grandi quantità e più c'è bisogno di un sistema che aiuti il management a prendere le decisioni sulla base delle informazioni che nascondono. La BI deve presentare le seguenti caratteristiche: essere *veloce*, riuscire a leggere ed elaborare i dati in modo quasi istantaneo anche quando questi sono presenti in grandi quantità; *facile*, non serve analizzare i dati se l'utente finale non riesce ad interpretarli, la BI deve quindi utilizzare un linguaggio facile da comprendere e immediato nell'interpretazione; *aggregante*, i

[3]Individuazione di elementi rumorosi all'interno del dataset

dati raccolti devono provenire da tutte le fonti a disposizione ed è necessario saper trattare anche dati di tipologie diverse (ad esempio numerici, di testo e così via); *storico*, tenere in considerazione dati da un arco temporale adeguato alle informazioni che si vogliono estrapolare (D'Ovidio & Villari, 2009).

2.2.1 Utilizzo della BI per apportare valore aggiunto

Lo scopo principale del BI è quello di rispondere alle domande che le aziende si pongono ogni giorno al fine di rispondere velocemente ai cambiamenti del mercato modificando strategie e decisioni. Ecco perché i dati all'interno di un'azienda ricoprono un ruolo molto importante, essi devono supportare il processo decisionale e per fare questo devono essere scelti con cura, trasformati e analizzati. L'obiettivo degli analisti è quello di estrarre valore dai dati utilizzando tecniche diverse a seconda delle tipologie di dati che hanno a disposizione. L'analisi aziendale è un processo ciclico che comprende fasi di accesso ai dati, identificazione, esplorazione e condivisione delle informazioni. Non si avrà quindi una fine del processo di BI ma questo accompagnerà l'azienda in tutti i suoi giorni di vita.

La BI è il processo di estrazione dei dati da un database OLAP, tecnologia di database ottimizzata per l'esecuzione di query e la creazione di report. La caratteristica principale che distingue questo tipo di database è l'organizzazione dei dati attraverso cubi anziché tabelle; in questo modo è possibile archiviare e analizzare una quantità più alta di dati in modo facile e veloce. I dati vengono così organizzati secondo vari livelli di dettaglio (ad esempio per la dimensione tempo è possibile raggruppare i dati per anno, per trimestre, per mese e così via).

Il contesto decisionale di tutte le aziende si basa sulla BI. Le aziende modificano strategie e metodologie sulla base delle conoscenze che ricavano dai dati perché permettono di inferire conoscenza che generalmente non è nota a priori. Per riuscire a eseguire le giuste analisi è necessario seguire step predefiniti che comprendono:

1. Tecniche di preparazione dei dati: fase preliminare in cui vengono analizzati i dati a disposizione in modo da dare forma all'obiettivo di business da raggiungere e applicare le giuste tecniche per preparare il dato alle analisi successive. Spesso più del 60% del progetto è dedicato a questa fase.

2. Tecniche di esplorazione dei dati: i dati devono essere interpretati correttamente. Si usano le tecniche statistiche per prendere consapevolezza delle informazioni. Questa analisi dà la possibilità di caratterizzare il dataset. Da questa fase possono venir fuori informazioni che all'azienda non erano note in precedenza.

3. Algoritmi di estrazione della conoscenza: da qui viene estratto valore dai dati. usando tecniche diverse raggruppate in 4 macrocategorie:

(a) algoritmi di classificazione: predire il trend di uno o più attributi nel futuro

(b) algoritmi di clustering e indici di qualità: tecniche non supervisionate che vanno a identificare dei sottogruppi di dati. Le tecniche non supervisionate si basano solo sui dati a disposizione. Il ruolo del data scientist[4] sarà poi quello di andare a caratterizzare ogni gruppo e cercare di interpretarli correttamente all'interno del contesto applicativo.

(c) algoritmi di estrazione di regole di associazione

(d) Analisi di serie temporali e regressione

[4]Colui che ha la capacità di gestire i Big Data e trarne informazioni rilevanti

Capitolo 3

Caso studio: Litokol s.p.a.

Questo capitolo descrive i passaggi fondamentali che sono stati eseguiti durante i mesi di elaborazione della tesi: in particolare verrà trattata la fase di esplorazione dei dati, la fase di scelta e calcolo dei KPI più efficaci a descrivere le performance dei fornitori e infine la fase di utilizzo di tecniche di data mining per l'estrazione della conoscenza.

3.1 Obiettivo di business

Un progetto nasce là dove esista una necessità da soddisfare o una mancanza da colmare; ogni progetto ha quindi un determinato scopo e uno specifico risultato da realizzare che viene raggiunto grazie allo studio dello stato dell'arte fino a quel momento conosciuto e al successivo sviluppo di nuove tecniche e soluzioni che migliorano ciò che fino a quel momento è utilizzato. L'obiettivo deve essere:

- specifico: deve essere ben definito a chi si rivolge, dove va raggiunto e come raggiungerlo;
- misurabile: una volta raggiunto l'obiettivo deve essere compreso;
- accessibile: deve essere realistico;
- rilevante: deve apportare valore aggiunto alle attività svolte;
- finito: deve avere una data di fine in cui tutti gli obiettivi devono essere raggiunti.

L'obiettivo che ci siamo posti di raggiungere con questa tesi è quello di analizzare i dati relativi agli acquisti dell'azienda al fine di ricavare informazioni utili per attuare strategie di approvvigionamento personalizzate per gruppi di fornitori. In

particolare ci soffermeremo ad analizzare le tempistiche dalla richiesta di materiale fino alla sua consegna e l'andamento dei costi dei singoli prodotti nel tempo. L'obiettivo è quello di realizzare uno strumento in grado di assistere il decision maker[1] dell'azienda a prendere decisioni. Con la nostra dashboard si vuole ottenere informazioni riguardanti i fornitori per capire se coloro che sono maggiormente rilevanti per quantità acquistata e importo speso sono anche quelli che hanno livelli di performance più alti, oppure se sono presenti fornitori non molto utilizzati ma che invece sono in grado di approvvigionare materiale con tempi e costi più bassi. Al fine di riuscire a selezionare il portafoglio di fornitori con coloro che hanno le giuste caratteristiche, ovvero quelle richieste dal business di Litokol.

3.2 Esplorazione dati

Per la fase di esplorazione dati si è utilizzato inizialmente SQL Server Management Studio per prendere conoscenza dei dati tramite query e, successivamente, Cognos di IBM[2] per visualizzare la conoscenza estratta. Sono stati utilizzati questi due tool perché il primo ha permesso di conoscere approfonditamente i dati in modo veloce ma con un limite, quello di poter ottenere come output solo tabelle che spesso non sono facili da interpretare. Il secondo ha invece permesso di trasformare quelle tabelle in grafici, istogrammi, diagrammi a torta e così via, vedendo così ogni sfaccettatura dei dati. Grazie a SQL Server Management studio è stato anche possibile fare una selezione delle informazioni che si sarebbero volute visualizzare nelle dashboard e quelle che invece non era necessario riportare. Le immagini che seguono sono screen derivanti dal tool di visualizzazione utilizzato.

La figura 3.1 mostra la distribuzione dei dati in data 31/01/2023; la specifica della data è importante in quanto i dati si modificheranno nel tempo a seconda degli acquisti che vengono effettuati ogni giorno. Nel database sono presenti dati con orizzonte temporale dal 2012 fino al presente, il primo grafico a barre mostra la somma degli importi pagati per gli acquisti di ogni anno, mentre quello sottostante mostra il numero di acquisti effettuati. Si vede che i dati dell'anno 2012 sono molto scarsi, questo ci fa capire che in quell'anno non venivano registrati tutti gli acquisti all'interno del database, sarà quindi un anno che non prenderemo in considerazione per le considerazioni future. Dopo il 2013, l'importo pagato per l'acquisto della merce, è sceso per poi stabilizzarsi intorno agli anni 2017-2019 ad un valore oltre la decina di milioni. Negli anni successivi è cresciuto nuovamente, fino a raggiungere il picco di massimo nel 2021. Per finire, l'anno 2023 ha un valore molto basso perché il template è stato creato alla fine del mese di gennaio, mancano perciò

[1]Figura chiamata a prendere le decisioni all'interno dell'azienda.

[2]tool di data visualization

tutti gli altri mesi. Il numero di ordini invece oscilla con un andamento positivo verso l'alto. Al contrario di ciò che si penserebbe, non segue lo stesso andamento dell'importo speso.

In totale ci sono 487 diversi fornitori all'interno del database; solo alcuni di questi sono utilizzati frequentemente, altri si riferiscono a forniture avvenute in anni passati e successivamente non più utilizzati e altri ancora non sono fornitori di materie prime utili al core business ma di prodotti di vario genere. Sono presenti 7230 diversi prodotti nel database, suddivisi in 31 categorie[3]. Le categorie più rilevanti (cioè quelle che hanno un impatto di costo più elevato per l'azienda) si possono vedere nel grafico a torta mostrato in figura 3.1 in cui le aree degli spicchi sono date dalla percentuale di budget speso per acquistare quella determinata categoria. Ci sono due categorie che occupano il 50% della spesa per il materiale acquistato; queste categorie sono le *materie prime e additivi* (in viola scuro nel grafico) e gli *imballaggi* (di colore viola chiaro). A seguire troviamo le *materie prime e resine liquide*, le *materie prime inerti* e i *profili*. Una di queste categorie non ha una specifica definizione ma viene sostituita da un valore nullo, in questo caso si sono prese in considerazione diverse soluzioni:

- eliminare gli ordini che hanno prodotti appartenenti a questa categoria;
- assegnare a questi prodotti il nome della categoria più frequente;
- considerare questa categoria come nuova categoria chiamata "Altro" a sé stante dalle altre.

Dato che il numero di record con questa caratteristica è molto rilevante, la prima opzione non si è ritenuta adeguata in quanto avrebbe fatto perdere molte informazioni. Guardando più nel dettaglio i prodotti appartenenti a questa categoria si è constatato che siano molto diversi tra loro, perciò, non era giusto assegnarli il nome della categoria più frequente, perciò si è scelta l'ultima opzione.

Come ultima informazione, dalla figura 3.1 si può vedere la tabella che mostra i codici fornitori e il relativo importo di acquisto e la quantità di materiale associata, la tabella è ordinata per importo decrescente; si nota che i due valori non hanno lo stesso andamento, ci sono fornitori che hanno un importo elevato ma una quantità acquista non alta, questo vuol dire che i prodotti da essi acquistati hanno un costo elevato e sarà quindi importante capire se le performance di acquisto di questi fornitori sono elevate e il linea con le richieste di Litokol s.p.a..

La tabella dei fatti presa in esame, come si è detto, è la tabella relativa agli acquisti. Questa tabella ha un numero molto elevato di attributi. Non essendo mai stata utilizzata prima per altri progetti, è stato utile soffermarci sull'analisi di ogni attributo per capire le informazioni utili che si potevano ricavare.

[3]Sono famiglie di prodotto, anche chiamate 'serie'.

Figura 3.1: Esplorazione dati al 31/01/2023.
Fonte: elaborazione personale.
Tool: IBM Cognos.

Importo ordinato, spedito e fatturato:
La figura 3.2 mostra un grafico a linee in cui è riportato per ogni mese dell'anno 2022 la somma dell'importo ordinato, spedito e fatturato. Si vede che per ogni mese l'importo ordinato è sempre molto superiore all'importo spedito e fatturato; questi ultimi hanno invece un andamento pressoché identico. Questo vuol dire che al momento dell'ordinazione viene valorizzato solo il campo dell'ordinato, nel momento in cui viene spedito il materiale è invece valorizzato il campo dello spedito solo del valore della merce effettivamente spedita che poi arriverà al valore totale al momento dell'evasione dell'ordine e questo può avvenire a distanza di molti mesi dalla data di ordinazione (analoga considerazione per il campo fatturato).

Per vedere quali ordini sono aperti (non ancora evasi totalmente) e quali chiusi vi è un campo nella tabella dei fatti che richiama, grazie all'utilizzo di una chiave esterna, alla tabella DW_ST_RIGA. Quando questo campo è valorizzato a stato 'PO' significa che l'ordine è aperto. Negli anni 2021, 2022 e 2023 ci sono rispettivamente 1, 36 e 143 ordini non ancora evasi. Ciò è in linea con le tempistiche di spedizione. In generale gli ordini sono evasi nel giro di qualche mese, non ci sono ordini aperti più vecchi del 2021.

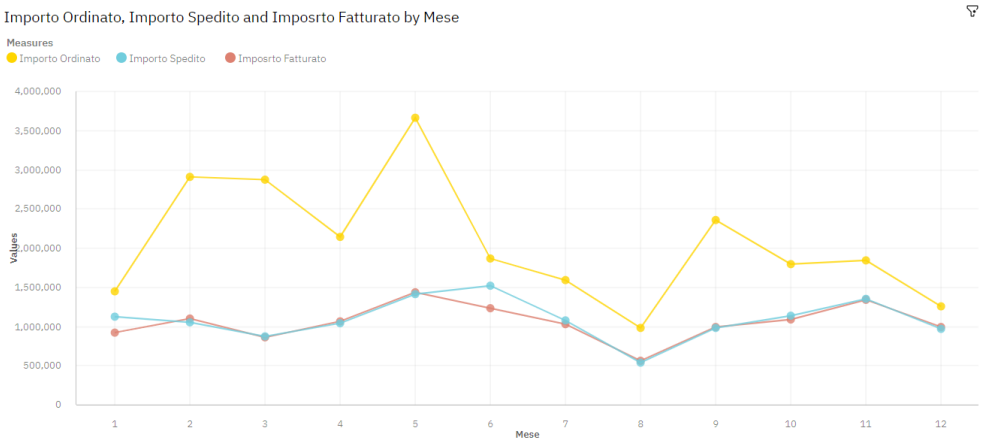


Figura 3.2: Importo Ordinato, Spedito, Fatturato nel 2022.
Fonte: elaborazione personale.
Tool: IBM Cognos.

DW_TP_PAG:

Nel database di partenza ci sono 106 tipologie diverse di metodo di pagamento che differiscono tra loro per metodologia e per dilazione di pagamento. Le metodologie presenti sono: bonifico bancario, ricevuta bancaria, rimessa diretta e SEPA[4]; i giorni di pagamento erano invece molto distribuiti e andavano dai 10 giorni fino ai 150 giorni. Si è quindi deciso di dividere queste metodologie in 4 classi diverse che chiameremo: A, B, C e D.

- Classe A: nella classe A sono stati inseriti tutti i pagamenti che permettono una dilazione sopra i tre mesi (>= 90 giorni);
- Classe B: nella Classe B sono stati inseriti tutti i pagamenti che permettono una dilazione di pagamento superiore a un mese e inferiori ai tre mesi (> 30 e <90), più quelle di un mese ma che comprendono sconti perché considerato un fattore positivo per l'azienda;
- Classe C: nella Classe C sono stati inseriti tutti i pagamenti che permettono una dilazione pari a un mese. (=30 giorni);
- Classe D: nella Classe D sono stati inseriti i pagamenti anticipati, quelli immediati, quelli inferiori ad une mese (<30 giorni), i bonifici a vista e tutte le tipologie che non avevano al loro interno un'informazione sui giorni di dilazione;

[4]Single Euro Payments Area

La Distinzione in classi è stata fatta creando una vista su SQL Server Management Studio grazie alla query in appendice A.

Data ordine, data consegna richiesta e data consegna prevista:
Ogni record della tabella dei fatti è associato a tre date:

- data Ordine: è la data in cui viene effettuato l'ordine e in cui si decide la quantità e l'importo da pagare;
- data consegna richiesta: è la data in cui Litokol richiede che venga consegnata la fornitura in modo da assicurare il servizio ai suoi clienti e in modo da non bloccare la sua attività operativa per mancanza di materiale in stock;
- data consegna prevista: è la data in cui il fornitore si impegna a consegnare la merce al momento dell'ordinazione, può quindi essere uguale, antecedente o successiva alla data richiesta.

Salta subito all'occhio che non è presente la data di entrata merce a magazzino, parlando con l'esperto di dominio (il responsabile del reparto IT di Litokol) è fuoriuscito che questa data veniva registrata con un altro gestionale e non era interessante per Litokol. L'azienda si basa infatti sulla data prevista e richiesta per determinare i ritardi.

Fornitori:
Come accennato in precedenza, 487 fornitori riforniscono Litokol dal 2012, il numero diminuisce a 336 considerando gli acquisti dal 2018, fino ad arrivare a soli 191 nell'ultimo anno (2022); questo ci dice che negli anni i fornitori cambiano, possono essere cessati contratti di fornitura di vecchia data per iniziare nuovi rapporti con fornitori ritenuti più adeguati alle richieste. Vediamo ora la distribuzione dei fornitori in base all'importo ad essi pagato negli ultimi 5 anni (dal 2018 ad oggi) in figura 3.3. Nel grafico di Pareto sono stati oscurati i nomi dei fornitori che si trovano sull'asse delle ascisse perché dati sensibili che non è possibile mostrare. Il grafico mostra che l'80% del budget a disposizione per gli acquisti viene affidato ai primi 40 fornitori per somma di importo. Questo vuol dire che il portafoglio di fornitori rilevante non è molto alto. Questi fornitori saranno quindi quelli da considerare più critici a livello di analisi delle performance.

La distribuzione geografica dei fornitori si basa sulle informazioni presenti in tre campi: Nazione, regione e provincia. Come si vede in figura 3.3, nel grafico a bolle sulla sinistra, la nazione da cui proviene la maggior parte dei fornitori è l'Italia, seguita dalla Germania; per i restanti, il valore della nazione è nullo. Andando ad analizzare nel dettaglio i fornitori italiani, nel grafico a bolle di destra si vede la suddivisione tra le regioni; le regioni con una presenza di fornitori maggiore sono

Lombardia, Emilia e Marche.

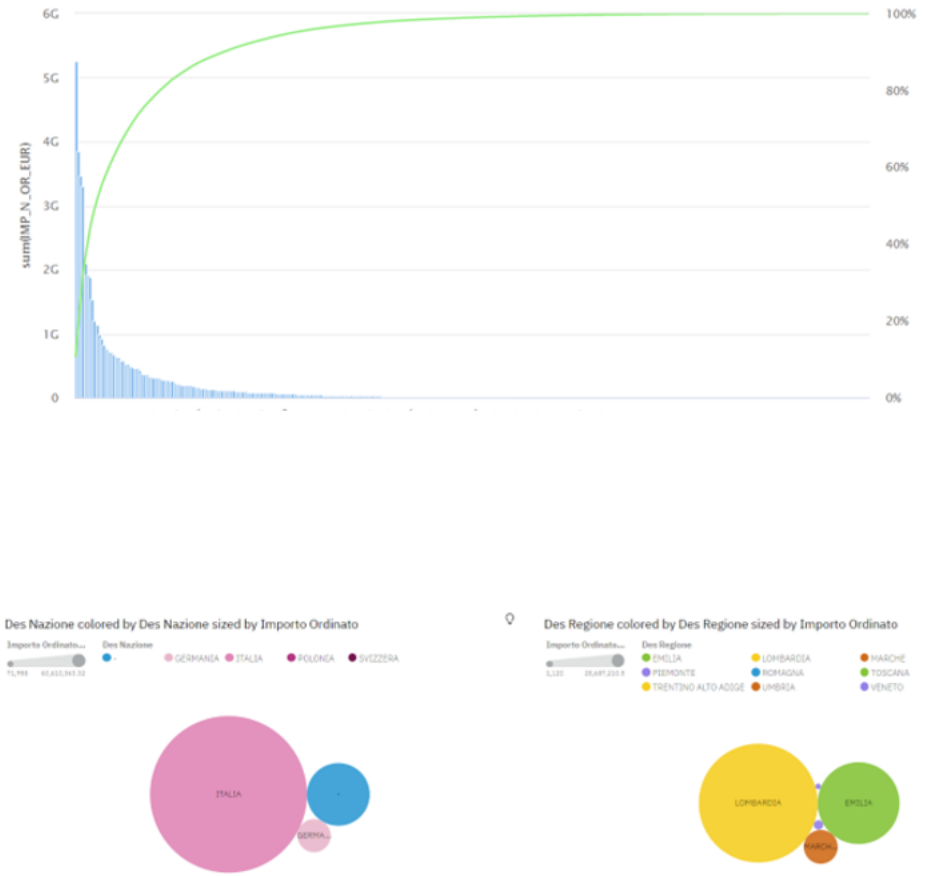


Figura 3.3: Distribuzione Fornitori.
Fonte: elaborazione personale.
Tool: IBM Cognos e Rapid Miner.

Prodotti:
Nel database sono presenti 7230 prodotti diversi dal 2012, che si riducono a 4350 considerando gli anni dal 2018, fino ad arrivare a 1650 prodotti diversi acquistati nell'ultimo anno (2022). In figura 3.4 è rappresentato il diagramma di Pareto dei prodotti, esso mostra che l'80% dei prodotti dal punto di vista dell'importo speso è dovuto a 200 prodotti. Il grafico a linee sottostante mostra invece l'andamento della spesa per i prodotti appartenenti alle categorie più rilevanti trovate in precedenza. Si nota che la serie *Materie Prime e Additivi* è l'unica che ha avuto un picco considerevole nell'anno 2021 arrivando a costare per Litokol più di 15M di euro. Le altre serie hanno invece un andamento pressoché costante nel tempo.

Un'altra informazione ricavata dal database a disposizione è la presenza di pochi fornitori che riforniscono tanti prodotti diversi e tanti fornitori che riforniscono un solo prodotto. Per quanto riguarda i prodotti si ha invece che uno stesso prodotto difficilmente è rifornito da fornitori diversi; ci sono quindi fornitori specifici per determinati prodotti. Da qui deriva l'obiettivo che ci siamo posti per questo progetto: non è possibile scegliere quale fornitore sia più adeguato a rifornire un determinato prodotto ma si rende necessario capire se un fornitore associato ad uno specifico prodotto rispetta le esigenze e le richieste di tempistiche e qualità che caratterizzano il business di Litokol. Se così non fosse, sarebbe quindi necessario valutare se è presente un altro fornitore nel portafoglio di quelli già presenti che abbia buone performance e possa rifornire quel prodotto.

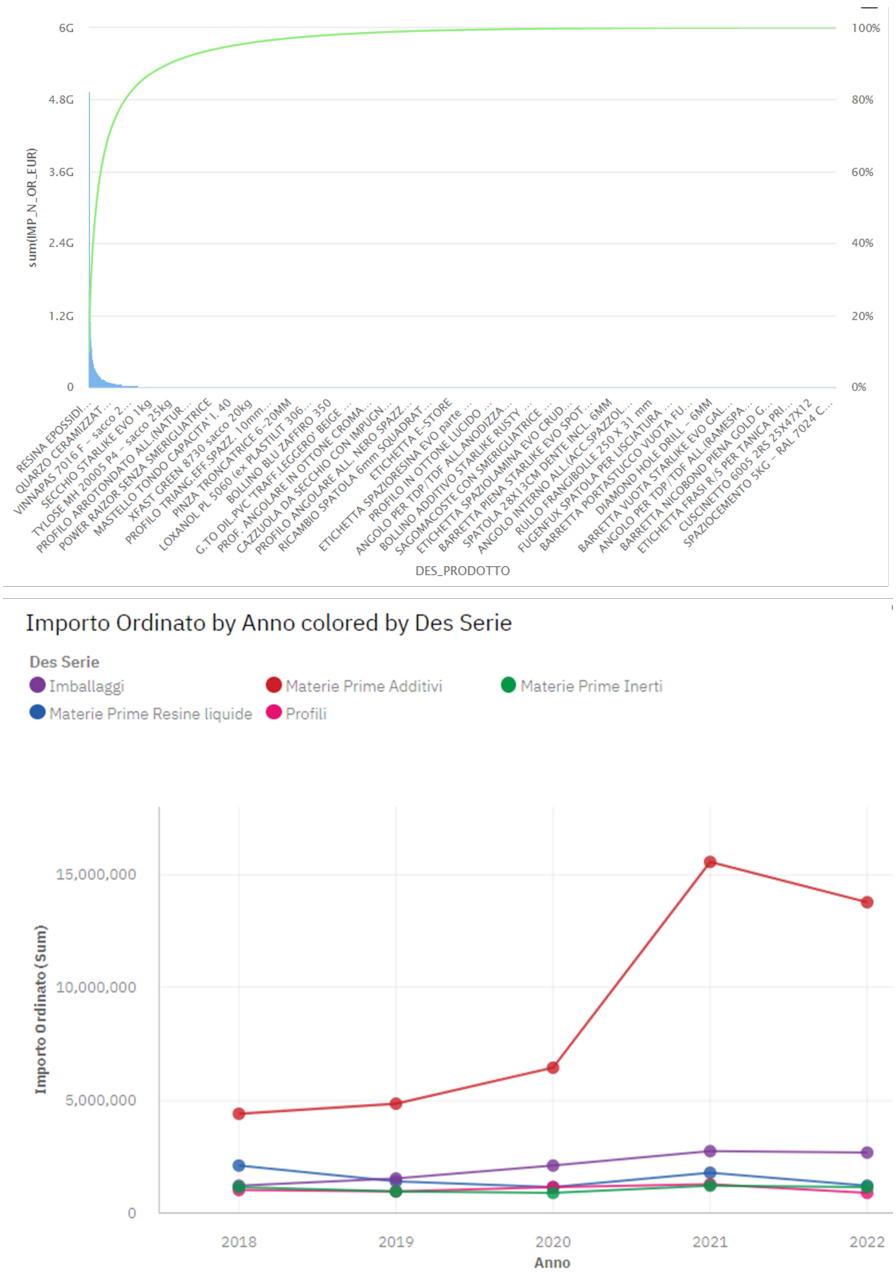


Figura 3.4: Distribuzione Prodotti.
Fonte: elaborazione personale.
Tool: IBM Cognos e Rapid Miner.

3.3 Pre-processing

La prima attività di data cleaning effettuata sul database di analisi è stata quella di eliminare tutti i record relativi ad acquisti avvenuti prima del 2018, i motivi di questa scelta sono due:

- i dati raccolti in anni precedenti al 2018 sono molto sporchi, hanno valori nulli o incongruenti con il dominio;
- per l'analisi è stato giudicato efficace un periodo temporale di cinque anni (dal 2018 al 2022).

Per eliminare questi dati è stato creato un filtro sul tool di IBM Cognos, Framework Manager (si veda capirotlo 4.2.1 in cui viene descritto lo strumento). Una volta presi in analisi questi dati si è visto che alcuni presentavano anch'essi dei valori che portavano rumore all'interno del dataset, per questo motivo sono state effettuate le seguenti azioni di data cleaning:

- Alcuni record avevano valorizzata la data di consegna richiesta o la data di consegna prevista all'anno 1900; parlando con l'esperto di dominio si è capito che questa data era presente qualora il dato non fosse stato disponibile. Si è deciso quindi di eliminare tali record dal dataset in quanto non associabili ad un anno in analisi.
- alcuni record avevano l'importo pagato per l'acquisto e la quantità acquistata minore o uguale a zero. In questo caso, sempre parlando con l'esperto di dominio, si è compreso che simboleggiavano materiali non conformi e quindi restituiti oppure omaggi del fornitore. Essendo questi casi molto rari all'interno del dataset non è stato possibile fare analisi più approfondite, perciò sono stati eliminati.
- Alcuni record avevano un valore di importo ordinato e importo fatturato diverso da quello spedito; in caso di ordini chiusi in data antecedente a quella corrente questi tre importi devono coincidere. In questo caso, parlando con l'esperto di dominio siamo rientrati in una particolarità che non era spiegata dalle logiche utilizzate internamente per storicizzare i dati, perciò tali dati erano valorizzati in modo errato. In questo caso, non compromettendo le analisi da effettuare successivamente sono stati lasciati all'interno del dataset.

Andando avanti con le analisi è stato notato la presenza di un pattern frequente tra i dati: alcuni prodotti erano presenti più di una volta all'interno di uno stesso numero ordine e avevano uguali attributi ad eccezione dell'importo pagato. Questo fatto portava ad avere un costo unitario molto più basso rispetto alla media per quel prodotto. Si è perciò arrivati alla conclusione che quei record erano valorizzati

con dati sporchi probabilmente a causa di successive integrazioni o aggiustamenti sui dati che lasciavano il record invariato ma con un valore di importo aggiuntivo a quello presente in precedente nell'ordine. Per modificarli e renderli consistenti e utili per le analisi è stato necessario raggruppare quei record sommando l'importo ordinato. Vediamone un esempio relativo al prodotto con codice "EAC.102.80": la tabella 3.1 mostra i dati come vengono prelevati direttamente dalla sorgente; mentre la tabella 3.2 mostra i dati dopo il raggruppamento.

Tabella 3.1: Dati relativi al prodotto EAC.102.80 (prima)

NUM_ORDINE	DATA_ORDINE	COD_FORNITORE	QUANTITÁ	IMPORTO	Cu
FO21R0100014	07/01/2021	0004169	10	40,37	4,037
FO21R0100729	21/07/2021	0004169	3	12,13	4,043
FO21R0100854	20/09/2021	0004169	5	20,76	4,152
FO21R0100988	03/11/2021	0004169	2	8,08	4,040
FO22R0100219	09/03/2022	0004169	5	21,79	4,358
FO22R0100387	03/05/2022	0004169	5	21,21	4,242
FO22R0100421	13/05/2022	0004169	10	42,90	4,290
FO18R0100300	04/04/2018	2025864	15	58,49	3,899
FO19R0100385	29/04/2019	2025864	7	27,04	3.863
FO19R0100470	20/05/2019	2025864	10	38,06	3,806
FO19R0100662	05/07/2019	2025864	10	37,55	3,755
FO20R0100356	24/04/2020	2025864	10	38,99	3,899
FO20R0100468	26/05/2020	2025864	7	27,01	3,859
FO20R0100468	26/05/2020	2025864	10	38,59	3,859
FO20R0100584	24/06/2020	2025864	10	38,58	3,858
FO20R0100921	21/09/2020	2025864	10	15,49	1,549
FO20R0100921	21/09/2020	2025864	10	23,08	2,308

Tabella 3.2: Dati relativi al prodotto EAC.102.80 (dopo)

NUM_ORDINE	DATA_ORDINE	COD_FORNITORE	QUANTITÁ	IMPORTO	Cu
FO21R0100014	07/01/2021	0004169	10	40,37	4,037
FO21R0100729	21/07/2021	0004169	3	12,13	4,043
FO21R0100854	20/09/2021	0004169	5	20,76	4,152
FO21R0100988	03/11/2021	0004169	2	8,08	4,040
FO22R0100219	09/03/2022	0004169	5	21,79	4,358
FO22R0100387	03/05/2022	0004169	5	21,21	4,242
FO22R0100421	13/05/2022	0004169	10	42,90	4,290
FO18R0100300	04/04/2018	2025864	15	58,49	3,899
FO19R0100385	29/04/2019	2025864	7	27,04	3.863
FO19R0100470	20/05/2019	2025864	10	38,06	3,806
FO19R0100662	05/07/2019	2025864	10	37,55	3,755
FO20R0100356	24/04/2020	2025864	10	38,99	3,899
FO20R0100468	26/05/2020	2025864	7	27,01	3,859
FO20R0100468	26/05/2020	2025864	10	38,59	3,859
FO20R0100584	24/06/2020	2025864	10	38,58	3,858
FO20R0100921	21/09/2020	2025864	10	38,57	3,857

La tabella 3.2 è stata ricavata tramite la realizzazione di una query in SQL Server Management Studio.

3.4 Definizione KPI

Ai fini della valutazione delle performance dei fornitoti si fa riferimento al sistema del Vendor Rating. Lo standard che definisce le linee guida del Vendor Rating System è IS:12040-2001; nel quale vengono stipulati i 5 fattori chiave che determinano una buona analisi dei fornitori. I cinque fattori sono: Qualità, prezzo, trasporto, servizio e sistema. La norma assegna a ciascun criterio un peso relativo alla sua importanza per la prosecuzione degli accordi di fornitura:

- Qualità: 35%
- Trasporto: 20%
- Prezzo: 10%
- Servizio: 5%
- Sistema: 30%

Per valutare queste caratteristiche devono essere create delle misure oggettive delle prestazioni dei fornitori, queste misure prendo il nome di Key Performance Indicator (KPI). I KPI sono indici che misurano l'efficacia con cui un'azienda sta raggiungendo gli obiettivi prefissati; sono valori misurabili attraverso formule costruite per l'identificazione di ogni sfaccettatura del processo in esame e per rispondere alle domande che l'azienda si pone ogni giorno. I dati a nostra disposizione non permettono di ricavare tutti e 5 i fattori. In particolar modo, parlando con l'esperto di dominio, si è deciso di non considerare la qualità nella nostra analisi, in quanto i dati relativi ai resi o ai controlli di qualità non sono a nostra disposizione. Per quanto riguarda il servizio e il sistema offerto questi sono indici qualitativi che non è possibile misurare attraverso i dati. In questo elaborato ci soffermeremo solo su ciò che è quantitativo.
Per misurare le performance di trasporto e prezzo di ciascun fornitore sono stati selezionati 5 KPI, i quali andranno poi a formare un unico indicatore, complessivo del rapporto di fornitura.

Non tutte le misure che è possibile ricavare dai dati sono necessarie, ogni azienda ha esigenze e richieste specifiche che necessitano di indicatori che non saranno mai identici a quelli usati da altre aziende. Per misurare l'efficacia ed efficienza dei fornitori utilizzeremo quindi un set di misurazione create per lo specifico caso in analisi. Il SVR creato si compone di diverse parti: puntualità, tempestività, flessibilità e disponibilità. Per ogni indicatore viene definito uno o più indici quantitativi e il valore finale del SVR sarà la media pesata di tutti questi indici. I pesi che entrano nella formula vengono dati dall'azienda in base a quanto è ritenuto importante quello specifico indicatore.

Gli indicatori selezionati sono calcolati per singolo fornitore, quindi ognuno avrà esattamente sei indicatori ad esso associati:

- Indice di Performance: quantità di materiale ordinato che è stato consegnato in tempo rispetto alla quantità di materiale ordinato consegnato in ritardo.
- Indice di Puntualità: numero di consegne effettuate in ritardo rispetto a quelle consegnate in tempo.
- Indice di Tempestività: media dei giorni in cui viene consegnata la fornitura.
- Indice di Tempo: media ponderata dei tre indicatori precedenti.
- Indice di Flessibilità: misura la capacità di consegnare la merce esattamente alla data richiesta.
- Indice di Prezzo: misura quanto il prezzo richiesto per uno specifico prodotto dista dal prezzo più basso pagato per quel prodotto in precedenza.
- Indice di rilevanza al prezzo: misura la percentuale di importo destinato agli acquisti per un determinato fornitore rispetto al totale di tutti gli acquisti.
- Indice di rilevanza alla quantità: misura la percentuale di quantità acquistata da un determinato fornitore rispetto al totale di tutti gli acquisti.

La figura 3.5 mostra come i primi sei indicatori elencati si collegano fra loro; gli indicatori di rilevanza non saranno usati per il calcolo del punteggio totale delle prestazioni del fornitore ma verranno utilizzati al momento della creazione della dashboard finale.

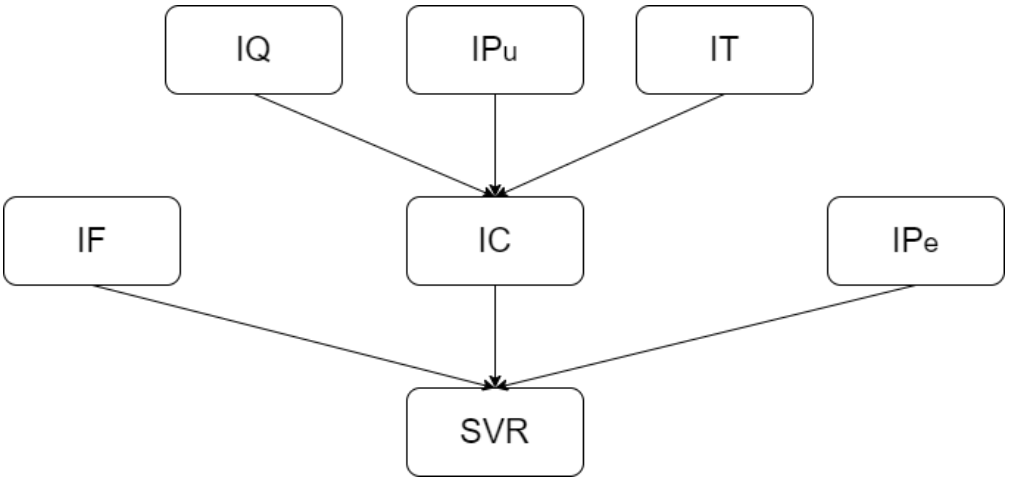


Figura 3.5: Albero dei KPI. Fonte: elaborazione personale.

Il vednor rating finale sarà quindi dato dalla seguente formula:

$$SVR = -w_{IF} * IF + w_{IC} * IC + w_{IPr} * IPr \qquad \forall fornitore \qquad (3.1)$$

Dove:
w_{IF}, w_{IC} e w_{IPr} sono i pesi relativi ai sotto-indicatori IF, IC, IPr.
IC: indice di consegna, assegna un punteggio al fornitore sulla base delle tipologie di consegne.
IF: indice di flessibilità, assegna un punteggio al fornitore sulla base della sua capacità di rispettare le richieste di Litokol.
IPr: indice di prezzo, assegna un punteggio al fornitore sulla base del prezzo effettuato.

In particolare l'indice di consegna (IC) è formato da tre tipi di indicatori diversi: l'indice di performance (IQ), l'indice di tempestività (IT) e l'indice di Puntualità (IP_U) nel seguente modo:

$$IC = w_{IQ} * IQ - w_{IT} * IT + w_{IPu} * IPu \qquad \forall fornitore \qquad (3.2)$$

Dove:
w_{IQ}, w_{IT} e w_{IP_U} sono i pesi relativi ai sotto-indicatori IQ, IT e IP_U.

Chiameremo IF, IC e IPr indicatore di primo livello; mentre IQ, IT e IPu indicatori di secondo livello. L'indicatore di tempestività e di flessibilità contribuiscono negativamente alla formazione dell'indicatore finale aggregato in quanto un aumento di questi indici dimostra una cattiva performance del fornitore al contrario degli altri indici che lo compongono, questo è il motivo per cui nella formula è presente un meno in corrispondenza di questi indicatori.

Per assegnare dei valori ai pesi di ogni indicatore è stato somministrato un questionario al responsabile del reparto acquisti di Litokol s.p.a., in quanto ogni indicatore avrà una sua specifica importanza che dipende dalle caratteristiche dell'azienda e del servizio che Litokol vuole fornire ai suoi clienti. I pesi individuati per gli indicatori di secondo livello sono:

- Indice di Performance: 30%
- Indice di puntualità: 30%
- Indice di tempestività: 40%

I pesi per gli indicatori di primo livello sono:

- Indice di tempo: 15%
- Indice di flessibilità: 40%
- Indice di prezzo: 45%

3.4.1 Indice di performance (IQ)

L'indice di performance misura la quantità di merce consegnata secondo le tempistiche indicate in fase di ordine consegnato in tempo[5] e la quantità di merce consegnata in ritardo. É stato necessario fare a priori, una valutazione sull'eventualità di considerare anche un anticipo[6] un fattore negativo per l'azienda in quanto crea disagi in termini di stoccaggio del materiale e aumenta i costi di magazzino. Nonostante ciò, una fornitura in anticipo garantisce anche un prosecuzione indisturbata dell'attività operativa e quindi una maggior qualità del servizio offerto al cliente finale. Si è quindi concluso che la merce consegnata in anticipo dovesse essere considerata come la merce arrivata in tempo.
La formula utilizzata è:

$$Dr = (\frac{Q_a}{Q_c}) - (\frac{Q_b}{Q_c}) \qquad \forall fornitore, riga \tag{3.3}$$

dove:
Q_a: quantità di merce consegnata in tempo o anticipo;
Q_b: quantità di merce consegnata in ritardo;
Q_c: quantità di merce totale consegnata;

L'indicatore può variare da -1 a 1; è negativo quando la quantità consegnata in ritardo è maggiore di quella consegnata in anticipo, positivo altrimenti, nullo quando le due quantità si equivalgono.

Nella tabella 3.3 viene mostrato un esempio riportato da SQL Server Management studio. Per ogni fornitore viene ricavato uno e un solo indicatore relativo alle performance di tutti gli acquisti.

3.4.2 Indice di puntualità (IPu)

L'indice di puntualità monitora il numero di consegne effettuate con un ritardo, rispetto al numero di consegne totali. Come in precedenza, sono state considerate

[5]Merce con data di consegna richiesta uguale a prevista

[6]merce consegnata prima della data richiesta da Litokol

Tabella 3.3: Indice di Performance

COD FORNITORE	Q_C	$\%Q_a$	$\%Q_b$	**KPI**
0004903	4930	0,188640	0,811359	-0,622719
2071511	8	0,50	0,50	0,00
2038490	7475	0,732441	0,267558	0,46883
2043245	92866	0,763637	0,236362	0,527275
...	...	...	...	...

consegne in ritardo quelle che avevano una data di consegna prevista maggiore della data di consegna richiesta. Le consegne in ritardo sono considerate negative perché portano a vari problemi di produzione e influiscono quindi sul servizio offerto al cliente finale.
La formula utilizzata è:

$$IPu = 1 - (\frac{N_r}{N_{tot}}) \qquad \forall fornitore \tag{3.4}$$

dove:
N_r: numero di consegne in ritardo;
N_{tot}: numero di consegne totali

Il valore dell'indicatore varia da 0 a 1. Il valore sarà 0 quando tutti gli ordini arrivano con ritardo, mentre 1 quando tutti gli ordini arrivano in tempo. Nella tabella 3.4 viene mostrato un esempio riportato da SQL Server Management studio. Per ogni fornitore viene ricavato uno e un solo indicatore che indica il livello di affidabilità alla puntualità.

Tabella 3.4: Indice di Puntualità

COD FORNITORE	N_r	N_{tot}	**KPI**
0004903	1	2	0,5
2071511	1	5	0,
2038490	1	5	0,8
2043245	3	22	0,864
...	...	...	...

3.4.3 Indice di tempestività (IT)

L'indice di tempestività misura la velocità in giorni con cui il fornitore è in grado di consegnare la fornitura richiesta. É stato necessario ponderare l'indicatore in base alla quantità richiesta per ogni ordine in quanto esso ha una dipendenza diretta con la velocità di produzione e consegna. Ogni fornitore può inoltre avere da uno a più indicatori di tempestività in base al numero di prodotti che esso fornisce all'azienda.
La formula utilizzata è:

$$IT = \sum_{i=1}^{N}((Data_cons_prev - Data_ordine) * \frac{Q_i}{Q_{tot}}) \qquad \forall fornitore, prodotto \tag{3.5}$$

dove:
Data_cons_perv: data in cui il fornitore ha previsto di consegnare la merce al momento dell'ordine;
Data_ordine: data in cui è stato effettuato l'ordine;
Q_i: quantità ordinata per quello specifico ordine;
Q_{tot}: quantità ordinata in totale.
N: numero di ordini

L'indicatore restituisce il numero di giorni medi che un fornitore utilizza per consegnare il materiale richiesto perciò più l'indicatore è alto più è negativo per l'azienda. La tabella 3.5 mostra un esempio riportato da SQL Server Management Studio; ogni fornitore può apparire in una o più record a seconda di quanti prodotti è chiamato a rifornire l'azienda.

Tabella 3.5: Indice di Tempestività

COD FORNITORE	COD PRODOTTO	**KPI**
0004903	PLTEPEQ001	3,434
2071511	ARG.MNT.01	6,999
2071511	ARG.RCM.01	3,0
2038490	MPC0266	4.980
2038490	MPC0714	5,60
2043245	CRTL1/6	6,294
...	...	...

Per ricavare un solo indicatore per ogni fornitore è stata eseguita una media aritmetica nei casi di presenza di più prodotti per un singolo fornitore.

L'Indice di Performance, l'Indice di Puntualità e l'Indice di Tempestività saranno sommati e pesati secondo la loro relativa importanza per formare l'indicatore di consegna totale (IC).

3.4.4 Indice di flessibilità (IF)

L'indice di flessibilità misura la capacità del fornitore di consegnare il materiale esattamente il giorno in cui viene richiesto dall'azienda (quindi quando la data prevista è uguale alla data richiesta). Anche questo indicatore, come quello precedente, è ponderato in base alla quantità ordinata per ogni ordine e ogni fornitore può avere da uno a più indici di flessibilità in base al numero di prodotti che vengono acquistati.
La formula utilizzata è:

$$IF = \sum_{i=1}^{N}((Data_cons_prev - Data_cons_rich) * \frac{Q_i}{Q_{tot}}) \qquad \forall fornitore, prodotto \tag{3.6}$$

dove:
Data_cons_perv: data in cui il fornitore ha previsto di consegnare la merce al momento dell'ordine;
Data_cons_rich: data in cui Litokol desidera ricevere l'ordine;
Q_i: quantità ordinata per quello specifico ordine;
Q_{tot}: quantità ordinata in totale.
N: numero di ordini

L'indicatore restituisce il numero medio di giorni tra la data richiesta e la data prevista. Assumere valori negativi quando si verificano anticipi, positivi quando si verificano ritardi e nulli quando le due date si equivalgono. La tabella 3.6 riporta un esempio da SQL Server Management Studio.

Tabella 3.6: Indice di Flessibilità

COD FORNITORE	COD PRODOTTO	**KPI**
0004903	PLTEPEQ001	0,81
2071511	ARG.MNT.01	0,0
2071511	ARG.RCM.01	0,0
2038490	MPC0714	0,0
2038490	MPC0266	3,23
2043245	CRTL1/6	-5,647
...	...	...

Per ricavare un solo indicatore per ogni fornitore è stata eseguita una media aritmetica nei casi di presenza di più prodotti per un singolo fornitore.

3.4.5 Indice di prezzo (IPr)

Non è stato facile trovare un indicatore di prezzo che potesse rappresentare la performance di un fornitore. Analizzando i dati si è constatato che i prezzi unitari di un prodotto hanno una variazione molto alta, questa variazione è dovuta a:

- Anno di realizzazione dell'ordine;
- quantità ordinata;
- distanza tra la data dell'ordine e la data di consegna richiesta.

Ogni fornitore ha i suoi criteri di assegnazione del prezzo e, con numeri così alti di dati, non è possibile tenerli tutti in considerazione. Non rimane che misurare quanto un fornitore sia capace di mantenere un prezzo basso nonostante queste variazioni. Per misurarlo si è deciso di creare l'indice di prezzo come percentuale di distanza tra il costo unitario più basso sostenuto per l'acquisto di quel prodotto e il costo unitario attuato per quello specifico ordine.
La formula utilizzata è:

$$IPr = \frac{P_L}{P_Q} \qquad \forall fornitore, ordine \tag{3.7}$$

dove:
P_L: prezzo unitario più basso effettuato per quel prodotto da tutti i fornitori;
P_Q: prezzo unitario dell'ordine.

Per ricavare un unico indicatore che mostrasse la performance totale del fornitore si è fatta la media dei suoi IP:

$$\overline{IP} = \sum_{i=1}^{N}(\frac{IP_i}{N}) \qquad \forall fornitore, prodotto \tag{3.8}$$

dove: N: numero di ordini fatti al fornitore per il prodotto j-esimo.

L'indicatore può assumere valore da 0 a 1; tende a 0 quando il prezzo aumenta rispetto al più basso sostenuto, mentre tende a 1 quando il prezzo si avvicina al più basso sostenuto. Nella tabella 3.7 sono stati riportati degli esempi da SQL Server Management Studio.

Tabella 3.7: Indice di Prezzo

COD FORNITORE	COD PRODOTTO	**KPI**
0004903	PLTEPEQ001	0,488
2071511	ARG.MNT.01	0,013
2071511	ARG.RCM.01	0,011
2038490	MPC0714	1
2038490	MPC0266	0,730
2043245	CRTL1/6	0,613
...	...	...

3.4.6 Indici di rilevanza

In fine è stato necessario conoscere quali e quanti fornitori sono determinanti per l'attività operativa di Litokol e quanto l'azienda dipenda da questi. Per ottenere questi dati sono stati creati altri due KPI che, a differenza di quelli descritti in precedenza, non andranno a formare l'indicatore generale del fornitore. La dipendenza da un fornitore si misura dalla quantità di materiale ordinata da esso rispetto alla quantità ordinata totale e dall'importo pagato al fornitore rispetto all'importo totale speso per gli acquisti.
Le due formule utilizzate sono:

$$R_imp_i = \frac{\sum_{j=1}^{P}(imp_n_or_{ij})}{\sum_{i=1}^{F}(imp_n_or_i)} \qquad \forall fornitore \tag{3.9}$$

Dove:
P: numero di prodotti acquistati dal fornitore i-esimo;
F: numero di fornitori;
imp_n_or: importo speso.

E

$$R_qta_i = \frac{\sum_{j=1}^{P}(qta_{ij})}{\sum_{i=1}^{F}(qta_i)} \qquad \forall fornitore \tag{3.10}$$

Dove:
P: numero di prodotti acquistati dal fornitore i-esimo;
F: numero di fornitori;
qta: quantità ordinata.

La tabella 3.8 mostra un esempio tratto da SQL Server Management Studio in cui ogni fornitore è affiancato dalla sua percentuale di rilevanza in termini di importo pagato e quantità ordinata.

Tabella 3.8: Rilevanza in percentuale

COD FORNITORE	**% Imp**	**% Qta**
2081695	15.366800	5.146500
2005196	7.738100	1.272500
0094411	6.603600	1.222900
0013599	4.630900	3.486200
...	...	...

Gli indicatori variano da 0 a 100; il valore sarà vicino a zero se il fornitore non viene utilizzato spesso durante il processo di approvvigionamento; mentre assumerà valori più alti se l'importo e la quantità acquistata sono rilevanti.

3.4.7 Tabella riassuntiva KPI

In questa sezione è mostrata una tabella che riassume le caratteristiche di ogni indicatore. In particolare, avremo il nome dato al KPI, la sua descrizione, il range di variazione dei valori che esso può assumere, la sua unità di misura e per finire la descrizione del suo andamento se positivo o negativo al crescere del valore.

Indice	Descrizione	Range	Unità Misura	Andamento
Indice Performance	quantità di materiale ordinato a ciascun fornitore che è stato consegnato in tempo rispetto alla quantità di materiale ordinato consegnato in ritardo	[-1,+1]	Percentuale	Se aumenta è positivo
Indice di puntualità	numero di consegne effettuate in tempo rispetto al totale	[0,+1]	Percentuale	Se aumenta è positivo
Indice di tempestività	media dei giorni in cui un fornitore è in grado di consegnare la fornitura	[0,+infinito]	Giorni	Se aumenta è negativo
Indice di flessibilità	capacità del fornitore di consegnare la merce esattamente alla data richiesta	[-infinito,+infinito]	Giorni	Se aumenta è negativo
Indice di Prezzo	misura quanto il prezzo richiesto dal fornitore per uno specifico prodotto dista dal prezzo più basso pagato per quel prodotto in precedenza	[0,+1]	Percentuale	Se aumenta è positivo
Indice Rilevanza Importo	Quantità di importo pagato ad un fornitore rispetto al totale	[0,+1]	Percentuale	-
Indice Rilevanza Quantità	Quantità acquistata da un fornitore rispetto al totale	[0,+1]	Percentuale	-

Figura 3.6: Riassunto KPI. Fonte: elaborazione personale.

3.5 Preparazione dati per clustering

La Cluster Analysis permette alle organizzazioni di comprendere meglio i propri fornitori identificando i venditori con caratteristiche simili. Quando si ha un numero elevato di fornitori è importante attuare strategie di approvvigionamento diverse a seconda delle prestazioni di ognuno. Avere una strategia per ogni fornitore sarebbe però impossibile e troppo dispendioso, ecco perché si rende necessaria la suddivisione dei fornitori in gruppi di dimensioni più o meno grande che separano fornitori con caratteristiche simili dagli altri. Alla fine del processo sarà necessario studiare una strategia di approvvigionamento per ogni cluster con un notevole risparmio di tempo e costi. Sarà inoltre possibile selezionare il fornitore sulla base del cluster di appartenenza capace di rispondere alle esigenze dell'organizzazione.

Prima di poter applicare gli algoritmi di cluster è necessario preparare dati in modo tale che i risultati ottenuti siano corretti, puliti e soprattutto affidabili. Per le analisi successive considereremo i 5 indicatori di performance, puntualità, tempestività, flessibilità e prezzo.

Per questo paragrafo e il successivo (3.6) le analisi effettuate sono eseguite utilizzando l'ambiente di elaborazione Jupyter Notebook, applicazione web open source basata sul modello client server che permette la creazione e l'esecuzione di blocchi di codice Python.

Normalizzazione:
Questo paragrafo fa riferimento alla normalizzazione statistica dei dati. La normalizzazione è una tecnica usata per la preparazione dei dati prima di applicare tecniche di Machine Learning. Esistono vari tipi di normalizzazioni, sono utili quando i dati a disposizione variano su scale diverse tra loro per portarli ad una scala comune. Useremo come input per gli algoritmi di machine learning i KPI calcolati nel paragrafo precedente. Alcuni di questi indici variano tra 0 e 1, altri possono assumente valori negativi e altri ancora possono assumere valori molto alti. La scala diversa tra i dati potrebbe creare problemi nel momento in cui si vuole calcolare il KPI aggregato oppure per fare analisi di Avanced Analytics. In figura 3.7 viene mostrata la variazione degli indici calcolati.

I possibili metodi di normalizzazione sono:

1. Zscore: converte tutti i valori in un punteggio relativo alla normale standard. La formula utilizzata è:

$$v' = \frac{v - mean}{stand_dev} \tag{3.11}$$

   Questo tipo di normalizzazione è in grado di individuare gli outlier, ma può produrre dati con scale diverse.

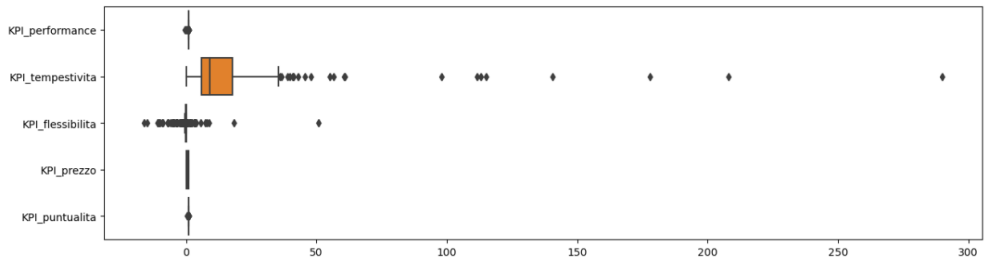


Figura 3.7: Boxplot per la dispersione degli indici.
Fonte: elaborazione personale.
Tool: Jupyter Notebook.

2. MinMax: rende tutti i valori compatibili con una scala di intervallo [0,1] usando il valore minimo e massimo che assumono i valori considerati. La formula è:
$$v' = \frac{v - min}{max - min} \tag{3.12}$$
Questo tipo di normalizzazione produce dati aventi uguale scala, ma non è in grado di individuare gli outlier.

3. Decimal Scaling: anche questo tipo di normalizzazione rende i valori compatibili con una scala di intervallo [0,1] ma per farlo utilizza una divisione con una potenza di 10. La formula utilizzata è:
$$v' = \frac{v}{10^j} \tag{3.13}$$
dove j è il numero di cifre che sono presenti nel valore più grande tra tutti quelli presi in considerazione.

Per la nostra analisi è stato necessario normalizzare gli indicatori prima di eseguire gli algoritmi di clustering. É stata scelta la normalizzazione secondo zscore, in quanto ritenuta più adeguata a rappresentare ogni sfaccettatura dei dati evidenziando meglio i valori anomali che si discostano dalla massa.
Dopo la standardizzazione zscore, la distribuzione dei dati si presenta così (fig. 3.8):

Dalla figura emerge che tutti gli indici sono ora centrati al valore zero e che gli indici di prezzo e tempestività hanno una variabilità più alta rispetto agli altri. In generale, sono presenti in tutti gli indici, a differenza di quello di prezzo, dei valori anomali rappresentati dai punti che si estendono fuori della scatola centrale. La mancanza di una vera e propria scatola centrale per flessibilità e puntualità indica che la maggior parte dei fornitori è caratterizzato dallo stesso valore dell'indicatore, questo fa sì che coloro che invece hanno un valore diverso siano considerati anomali.

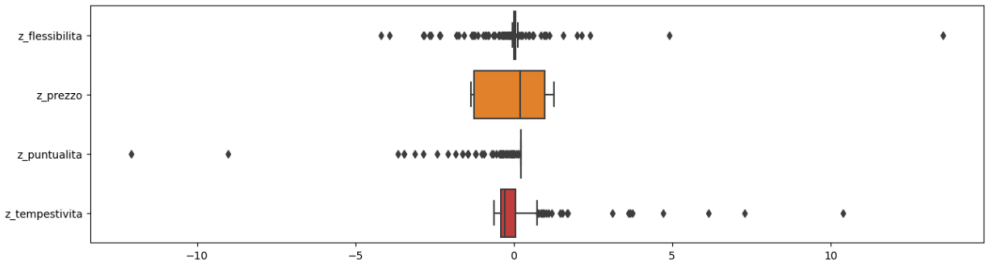


Figura 3.8: Indici standardizzati.
Fonte: elaborazione personale
Tool: Jupyter Notebook.

Correlazione:
Prima di utilizzare gli indici come input per il clustering è necessario controllare un'altra caratteristica importante. Gli indicatori non devono avere un alto grado di correlazione gli uni con gli altri. Per capire se i KPI selezionati sono dipendenti si è calcolata la matrice di correlazione e i risultati sono mostrati in figura 3.9.

Figura 3.9: Matrice di correlazione.
Fonte: elaborazione personale.
Tool: Jupyter Notebook.

Come si vede dalla figura, gli indici sono indipendenti tra loro. Il valore più alto è quello tra l'indice di performance e puntualità. Questo valore non ci sorprende in quanto l'indice di performance misura la quantità di materiale consegnato in ritardo rispetto a quello consegnato in tempo, mentre l'indice di puntualità misura il numero di consegne effettuate in ritardo. Quando il numero di consegne in ritardo cresce crescerà anche la quantità del materiale. Dopo questa analisi si è quindi

deciso di eliminare dalle analisi successive di clustering l'indice di performance, mantenendo solo quello di puntualità.

3.6 Clustering fornitori

Questa sezione descrive tutti i passaggi, le considerazioni e le scelte che sono state prese durante la stesura dell'elaborato, al fine di ottenere una conoscenza dai dati che possa portare valore aggiunto al cliente finale. É stata effettuata un'analisi multivariata utilizzando come input per gli algoritmi i KPI descritti in precedenza. Non è possibile sapere a priori quale algoritmo di cluster tra quelli presenti in letteratura performa meglio con il dataset che abbiamo a disposizione, perciò, ne è stato applicato più di uno.

Per le elaborazioni dei dati che seguono e per l'utilizzo degli algoritmi è stato usato il linguaggio di programmazione Python (il codice si trova in allegato C). Per consentire di utilizzare i dati del database di Litokol è stato necessario innanzitutto stabilire la connessione tra il computer locale, in cui è stato scritto il codice, e il database; utilizzando come informazioni il driver, l'indirizzo IP, il nome, l'username e la passward di accesso al database. Una volta stabilita la connessione è stato possibile leggere i dati grazie ad una query di estrazione delle informazioni. A questo punto tutti i dati necessari erano disponibili alle elaborazioni successive.

Di seguito sono elencati i passaggi eseguiti per arrivare al risultato finale; i passaggi relativi all'applicazione degli algoritmi di cluster e alla validazione verranno poi descritti in modo più dettagliato nel seguito del capitolo:

1. connessione al database di Litokol;
2. lettura della tabella Tab_KPI;
3. normalizzazione degli indici tramite metodo zscore;
4. matrice di correlazione;
5. KNN e applicazione algoritmo DBSCAN sull'interno dataset;
6. creazione dei gruppi *Fornitori0* e *Fornitori Particolari*;
7. *Fornitori0*
   (a) KNN e applicazione algoritmo DBSCAN;
   (b) Elbow Method e applicazione algoritmo Kmeans;
   (c) applicazione cluster gerarchico;

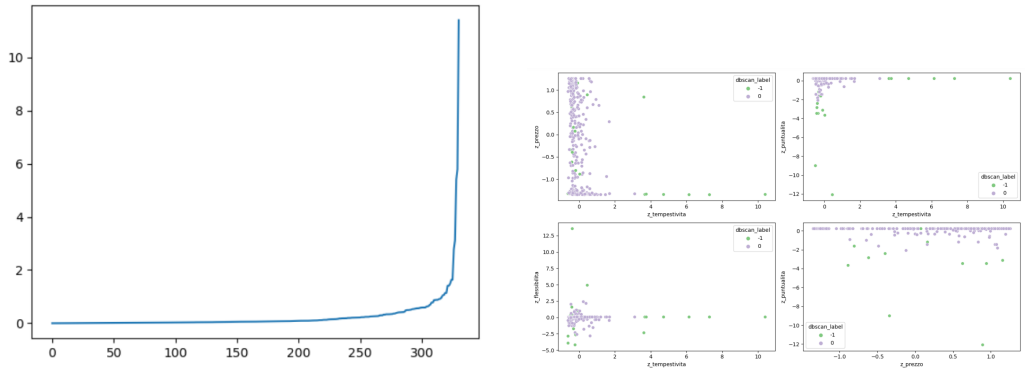
8. *Fornitori Particolari*

   (a) KNN e applicazione algoritmo DBSCAN ;

9. Validazione delle analisi tramite SVD e Silhouette.

I primi 4 punti dell'elenco sono stati descritti in precedenza, nel seguito del capitolo verrà approfondito l'esecuzione degli algoritmi selezionati per le analisi.

5-6. DBSCAN:
Input: intero dataset.
Il primo passo è stato quello di trovare il valore di epsilon, si è quindi fatto uso del grafico K-NN in cui k rappresenta il numero di dimensioni del dataset +1, nel nostro caso 5 perché i KPI che elabora l'algoritmo sono quattro. Il valore letto sull'asse delle ordinate, che corrisponde al punto di gomito della curva, è il valore di epsilon cercato (fig. 3.10a).



(a) *Grafico K-NN* (b) *Cluster*

Figura 3.10: Applicazione DBSCAN.
Fonte: elaborazione personale.
Tool: Jupyter Notebook.

Il valore di epsilon scelto è 1.5, mentre il valore di min_sample è 5. Si vanno così a creare due cluster: il cluster 0, con il maggior numero di fornitori, rappresenta il gruppo di fornitori con caratteristiche simili tra loro; e il cluster -1 è quello che il dbscan forma con tutti i dati considerati rumorosi e raccoglie tutti i fornitori che si discostano dagli altri perché caratterizzati da KPI fuori dalla media. In particolare, il cluster 0 è formato da 311 fornitori e il cluster -1 da 20 fornitori. La distribuzione dei punti all'interno dei cluster si può visualizzare in figura 3.10b in cui i punti grigi sono quelli appartenenti al cluster 0 e i punti verdi quelli appartenenti al cluster -1. Nonostante l'algoritmo consideri i fornitori appartenenti al cluster -1 degli outlier, essi non sono stati eliminati dal dataset ma saranno analizzati successivamente per

capire il motivo di tale suddivisione. Tali fornitori non sono formati da dati errati, ma semplicemente da dati fuori dalla media.

Il dataset a nostra disposizione è formato quindi da alcuni fornitori che si discostano notevolmente dalla massa. Per questo si è deciso di separare questo gruppo sotto il nome di *Fornitori particolari*. Questa distinzione si è resa necessaria per poter andare ad analizzare successivamente con altri algoritmi di clustering i fornitori che fanno parte del cluster 0, che chiameremo *Fornitori0*, per capire se sono presenti particolarità rilevanti anche in quel gruppo. I *fornitori particolari* sono analizzati separatamente.

7a. DBSCAN:
Input: tabella con il gruppo *Fornitori0*
Rifacendo gli stessi passaggi utilizzando l'algoritmo DBSCAN otteniamo comunque un unico cluster e il cluster dei dati rumorosi. Non essendo un risultato rilevante per il proseguimento delle analisi si è deciso di applicare algoritmi diversi in quanto per i dati a disposizione il DBSCAN non è l'algoritmo giusto.

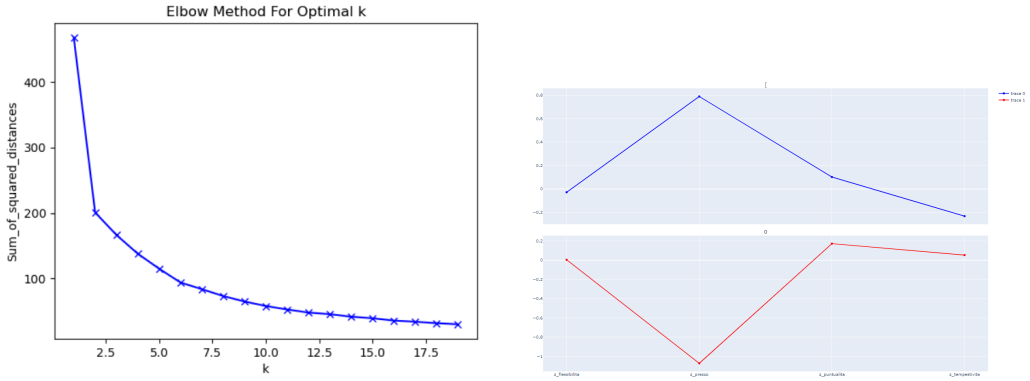
7b. KMeans:
Input: tabella con il gruppo *Fornitori0*
L'algoritmo Kmeans necessita in input del valore di k, cioè del numero di cluster che si vogliono ottenere come output. Per conoscere il numero più adeguato da inserire come k è stato utilizzato l'Elbow Method. L'Elbow Method, anche chiamato *Metodo del Gomito*, è un grafico a linea sul piano cartesiano in cui sull'asse delle ascisse si ha il numero di cluster e sull'asse delle ordinate l'indice Sum of Square Error (SSE). La curva spiega la percentuale di varianza calcolata attraverso il calcolo del SSE:

$$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.14}$$

Dove:
n: numero di cluster

Viene scelto il numero di cluster corrispondente al cambio di pendenza che esegue la curva da un certo punto in poi. L'andamento decrescente della curva viene spiegato dall'aumento dei cluster: maggiore sarà il numero di cluster, minore sarà l'errore, cioè la distanza media, tra ogni punto all'interno del cluster. In altre parole, se vengono formati tanti cluster allora ogni singolo punto sarà spiegato maggiormente dal cluster di appartenenza. Non è possibile però aumentare il numero di cluster a piacere in quanto da un certo punto in poi l'aggiunta di un cluster non aggiunge valore alle analisi ma, al contrario, può portare al caso di *overfitting*, cioè spiega in modo eccessivo i dati. Di seguito viene riportato il grafico dell'Elbow Method (fig. 3.11a), per il set di dati preso in analisi il numero di cluster più

adeguato è 2 in quanto la curva fa un ginocchio in quel punto.



(a) *Elbow Method* (b) *Cluster*

Figura 3.11: Applicazione Kmeans.
Fonte: elaborazione personale.
Tool: Jupyter Notebook.

É stato così applicato l'algoritmo KMeans con K=2 e i cluster ottenuti sono quelli rappresentati in figura 3.11b in cui vi sono due piani cartesiani che hanno in ascissa le dimensioni utilizzate per la suddivisione in gruppi e in ordinate il valore medio che ogni cluster assume per quella dimensione. Il cluster rappresentato con la linea blu è formato da 184 fornitori mentre quello rosso ne ha al suo interno 127. Analizzando i due grafici, i cluster formati si differenziano specialmente dal prezzo che essi richiedono per il materiale acquistato. Il gruppo di fornitori blu attua prevalentemente un prezzo molto buono, vicino al minimo mai richiesto, mentre il gruppo di fornitori rosso è caratterizzato da prezzi più alti della media. Un'altra distinzione riguarda l'indice di tempestività, il cluster blu ha in media un indice più basso, mentre il cluster rosso ha in media un indice più alto. Per gli altri KPI invece non si riscontrano sostanziali differenze tra i due cluster. Questo risultato non ci sorprende in quanto dalla figura 3.8 era chiaro che gli indici con una variabilità più alta erano proprio quello di prezzo e di tempestività.

7c. Clustering gerarchico
Input: tabella con il gruppo *Fornitori0*
Dopo l'utilizzo del Kmeans si è applicato l'algoritmo di clustering gerarchico per vedere se portava a risultati diversi. A differenza del Kmeans non richiede un numero di cluster da definire a priori e analizza tutte le possibili combinazioni di oggetti. L'output di questo algoritmo è il dendrogramma, cioè un grafico ad albero in cui sull'asse delle ordinate c'è il valore della distanza che assumono i cluster creati mentre sull'asse delle ascisse avremo tutti i fornitori presenti nel nostro

database di *Fornitori0*. Vediamo in figura 3.12 il dendrogramma creato in output da questo algoritmo.

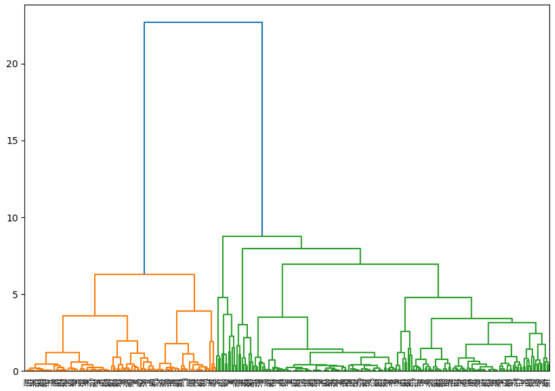


Figura 3.12: Dendrogramma.
Fonte: elaborazione personale.
Tool: Jupyter Notebook.

Alla base del dendrogramma si ha un numero di cluster pari al numero di fornitori, in quanto ogni fornitore rappresenta un cluster, nei passi successivi vengono raggruppati i fornitori più simili a due a due, fino ad ottenere un unico cluster alla fine del processo. Per misurare la distanza fra due cluster si possono utilizzare diverse misure, e ogni metodo produce una diversa partizione degli oggetti:

- single Linkage: si uniscono i cluster con distanza minima, avente distanza calcolata come la più piccola distanza fra due oggetti appartenenti a cluster diversi.

- Complete Linkage: si uniscono i cluster con distanza minima, avente distanza calcolata come la più grande distanza tra due oggetti presenti in due cluster diversi.

- Average Linkage: si uniscono i cluster con distanza minima, avente distanza calcolata come la media delle distanze fra tutti i punti presenti in due cluster diversi.

- Centroid Linkage: si uniscono i cluster con distanza minima, avente distanza calcolata come quadrato della distanza euclidea tra i centroidi di cluster diversi.

- Ward Linkage: vengono uniti due cluster la cui unione provoca il minimo incremento della devianza.

Durante l'utilizzo di questo algoritmo sono state provate diverse tecniche di linkage e il metodo che portava ad un risultato migliore è quello del Ward Linkage utilizzando come misurazione la distanza euclidea. Il dendrogramma finale è riportato in figura 3.12. A partire dal basso della figura si vede che i primi cluster creati sono molto vicini fra loro, infatti le unioni dei gruppi si formano molto in basso. Quando il cluster verde e il cluster rosso sono creati, i due si riuniscono a una distanza molto alta di circa 22. Il punto giusto in cui *tagliare* il dendrogramma per formare k cluster distinti è quello in cui il passaggio da k gruppi a k+1 gruppi si forma ad una distanza elevata. Guardando il dendrogramma formato è giusto quindi tagliare ad un'altezza di 10 perché prima di quella i cluster formati sono molto vicini fra loro. A dimostrazione del fatto che, anche con questo algoritmo, la partizione migliore del set di dati è in due grandi gruppi al cui interno sono presenti fornitori simili tra loro e molto diversi da quelli presenti nell'altro gruppo.

8a. Fornitori Particolari:
Per finire, siamo andati a vedere più nel dettaglio i fornitori che nel primo step di clustering si erano ritenuti particolari perché si discostavano in maniera significativa dagli altri. Per capire perché questi fornitori non avevano performance simili agli altri è stato applicato nuovamente l'algoritmo dbscan. In questo modo è stato possibile osservare diversi livelli di densità tra i fornitori. Una volta osservati i passaggi richiesti dall'algoritmo e descritti in precedenza, si è scelto un epsilon=4 e un numero minimo di punti=2. Le partizioni ottenute sono 3 e si possono vedere in figura 3.13.

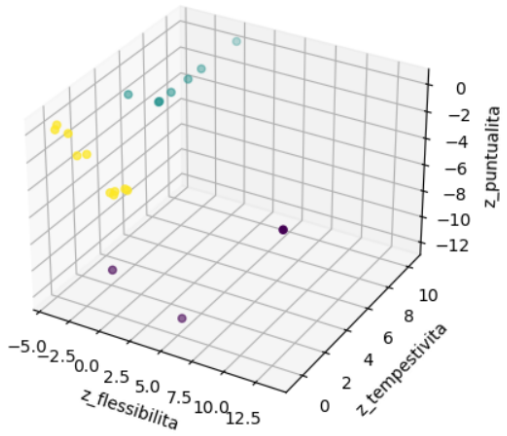


Figura 3.13: Cluster Fornitori Particolari.
Fonte: elaborazione personale.
Tool: Jupyter Notebook.

Questi cluster non si distinguano in base al prezzo come invece avveniva per i Fornitori0 ma si differenziano per il valore assunto dagli altri indici. Il cluster giallo è caratterizzato da flessibilità e tempestività medio-bassa; il cluster blu è caratterizzato da alti valori di tempestività e alti prezzi. Per finire ci sono tre fornitori considerati fuori da ogni cluster, dato che sono di numero basso, si è deciso di assegnarli al cluster loro più vicino e cioè al cluster giallo in quanto separarli dagli altri non aggiunge valore.

3.6.1 Validazione clustering

Una volta finita la fase di applicazione degli algoritmi si passa a quella della visualizzazione e della validazione dei cluster ottenuti. Per poter visualizzare punti descritti da più attributi vengono usate tecniche di riduzione delle dimensioni dei dati che prendono il nome di Principal Component Analysis. La tecnica più usata è la Singular Value Decomposition (SVD), in questo modo è possibile trasformare dati a N-dimensioni in dati a K-dimensioni, con K scelto dall'analista. Per la nostra analisi è stato scelto un K=2 in modo da plottare i dati attraverso un piano cartesiano.
Altre misure che vengono usate per la validazione dei cluster sono:

- Indici Esterni: usati per valutare se le etichette di cluster corrispondono alle etichette date dall'esterno, ad esempio l'indice di coesione o separazione.
- Indici Interni: usati per misurare la bontà di cluster ottenuti non considerando informazioni provenienti dall'esterno, ad esempio l'indice di Silhouette.
- Indici Relativi: usati per comparare due tecniche di cluster diverse tramite l'utilizzo degli indici nominati sopra.

In seguito verrà calcolato il valore di Silhouette per i gruppi ottenuti dai diversi algoritmi testati. Questo indice può assumere valore tra -1 e +1. Un valore vicino a +1 significa che i punti di un cluster sono ben diversi dai punti di un altro cluster, un valore vicino a -1 invece dice che i punti dei cluster potrebbero essere stati assegnati al gruppo sbagliato.

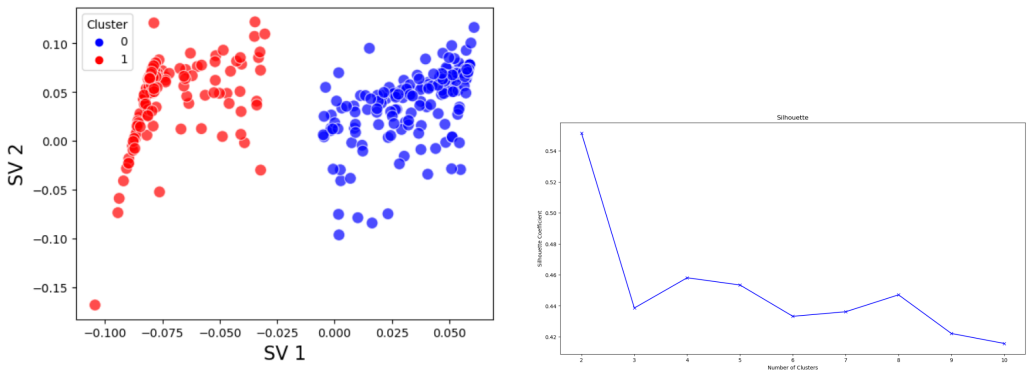
KMEANS:
Vediamo come si presenta il risultato ottenuto con l'algoritmo Kmeans. Per prima cosa mostriamo i dati all'interno di un piano cartesiano attraverso la tecnica SVD. Si è scelto di mostrare i due gruppi tramite un piano a due dimensioni arrivando così ad ottenere la rappresentazione dei cluster come in figura 3.14a. I gruppi formati si presentano ben distinti l'uno con l'altro, fattore positivo da cui si può concludere che i fornitori appartenenti a diversi cluster abbiano caratteristiche diverse fra loro.

Si è successivamente scelto di calcolato l'indice di Silhouette essendo il più completo tra quelli proposti e in grado ti tenere in considerazione sia il grado di coesione che di separazione tra cluster.
L'indice di Silhouette viene calcolato come segue:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3.15}$$

Dove:
a(i): la media di dissimilarità del punto i con tutti gli altri punti dello stesso cluster.
b(i): il minimo valore tra le medie di dissimilarità che il punto i ha in relazione agli altri cluster.

In figura 3.14b viene mostrato come la Silhouette cambia in relazione al numero di cluster formati che si leggono nelle ascisse. Con un numero di partizioni uguale a 2 otteniamo un valore di Silhouette pari a 0.55, valore che si abbassa notevolmente aumentando il numero di cluster, ciò a dimostrazione del fatto che 2 sono i gruppi più rilevanti all'interno del dataset di fornitori.



(a) *Singular Value Decomposition* (b) *Silhouette*

Figura 3.14: Validazione cluster.
Fonte: elaborazione personale.
Tool:Jupyter Notebook.

Cluster Gerarchico:
Vediamo invece come si presenta il cluster gerarchico con la tecnica SVD in figura 3.15. I cluster formati non sono ben distinti come avveniva con l'algoritmo Kmeans. Inoltre, l'indice di Silhouette vale 0.53, quindi leggermente più basso di quello calcolato con l'altro algoritmo.

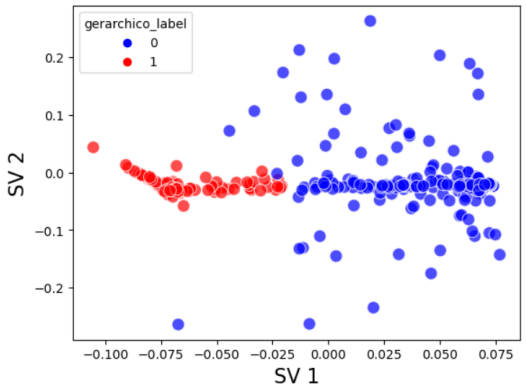


Figura 3.15: Singular Value Decomposition per cluster gerarchico.
Fonte: elaborazione personale.
Tool: Jupyter Notebook.

Conclusioni:
A questo punto è possibile scegliere gli algoritmi più utili all'analisi. Siamo partiti da un dataset di circa 330 fornitori caratterizzati da indici di tempestività, flessibilità, puntualità e prezzo. Dopo una prima scrematura iniziale in cui sono stati divisi i fornitori che avevano indici pressoché simili da quelli che invece avevano performance diverse dalla massa grazie all'algoritmo basato sul concetto di densità, è stato possibile trovare i gruppi di *Fornitori particolari* e *Fornitori0*. Sono stati poi utilizzati due tipi di algoritmi, uno partizionale (Kmeans) e uno gerarchico, per il gruppo di *Fornitori0*. Il processo di validazione ha evidenziato un algoritmo più performante di un altro. L'algoritmo Kmeans, oltre ad avere un indice di Silhouette più alto, ha anche portato alla creazione di gruppi di fornitori molto più distinti l'uno dall'altro. Successivamente anche dal gruppo di *Fornitori particolari* sono stati evidenziati due gruppi grazie all'utilizzo dell'algoritmo dbscan.

Si sono quindi formati quattro gruppi, per poter etichettare questi cluster secondo le loro caratteristiche distintive si è usato il Radar Chart in figura 3.16. Questa tipologia di grafici consente di mostrare diverse dimensioni di analisi in un grafico bidimensionale distribuendole nel piano su diverse assi aventi la medesima origine. In questo caso le aree colorate rappresentano i cluster e gli assi rappresentano gli indici utilizzati per la loro realizzazione. Ogni area colorata rappresenta un diverso cluster e i punti corrispondono alla media del relativo indice che compone quel cluster. Tenendo in considerazione che gli indici di tempestività e flessibilità variano al contrario rispetto agli altri, quando avremo un indicatore alto significa che le performance relative sono basse ecco perché verrà scritto *negativo*.
Il gruppo zero, di area blu, ha le seguenti caratteristiche:

- Performance: Positivo

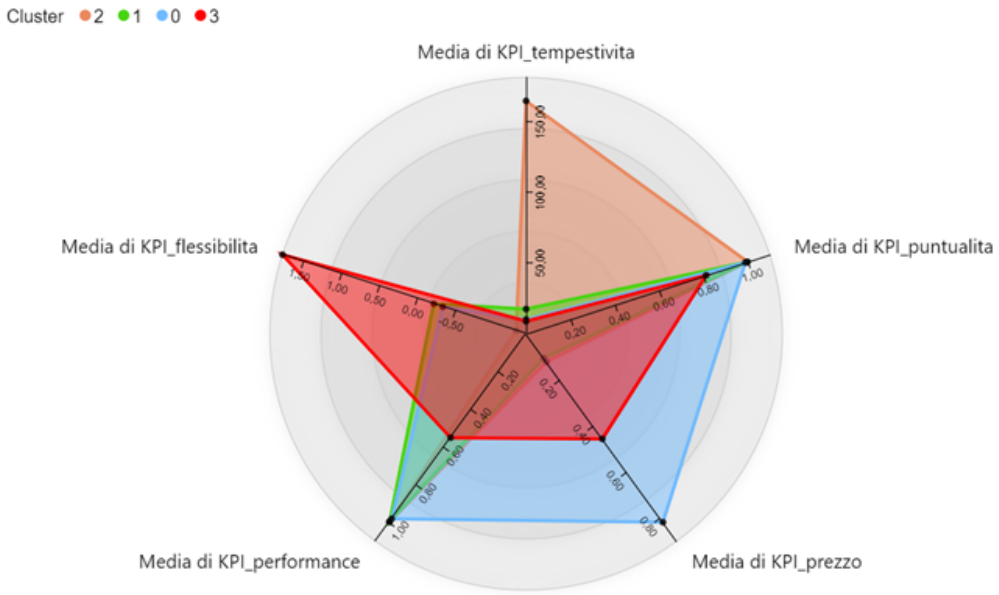


Figura 3.16: Radar Chart. Fonte: elaborazione personale

- Tempestività: Positivo
- Flessibilità: Medio
- Prezzo: Positivo
- Puntualità: Positivo

Il gruppo uno, di area verde, ha le seguenti caratteristiche:

- Performance: Positivo
- Tempestività: Positivo
- Flessibilità: Medio
- Prezzo: Negativo
- Puntualità: Medio

Il gruppo due, quello di area arancione, ha le seguenti caratteristiche:

- Performance: Positivo
- Tempestività: Negativo
- Flessibilità: Positivo

- Prezzo: Negativo
- Puntualità: Positivo

Il gruppo tre, quello di area rossa, ha le seguenti caratteristiche:

- Performance: Medio
- Tempestività: Positivo
- Flessibilità: Negativo
- Prezzo: Medio
- Puntualità: Negativo

Alla fine del processo abbiamo ottenuto quattro cluster descritti come in tabella 3.9.

Alla base di queste analisi vi è la necessità di capire quale fornitore sia più o meno adeguato nel momento in cui viene effettuato un ordine in base alle esigenze di Litokol. Ci possono essere periodi in cui è richiesto di spendere meno, oppure periodi in cui è di fondamentale importanza ricevere del materiale in tempi brevi. Grazie a queste divisioni è ora possibile andare ad individuare più velocemente il fornitore in grado di rispondere e soddisfare al meglio queste esigenze.

Un secondo obiettivo è quello di individuare gruppi di fornitori particolarmente performanti e gruppi di fornitori che invece non lo sono. In questo modo la fase di decision making sarà supportata da dati che prima non erano a disposizione; gli acquisti saranno indirizzati verso fornitori più veloci ed economici, andando mano a mano ad abbandonare quei fornitori che non si ritengono più adeguati a soddisfare le richieste dell'azienda.

A questo punto è possibile proseguire con l'inserimento della tabella relativa ai cluster formati su SQL Server Management Studio per poi portarla su Framework Manager per utilizzarla.

Tabella 3.9: Tabella riassuntiva cluster

Cluster	num_fornitori
0	184
1	127
2	7
3	13

Capitolo 4

Modello e integrazione dati

Prima di poter visualizzare i dati attraverso grafici e tabelle è necessario costruire il modello dati che relaziona tutte le tabelle contenenti i dati degli acquisti. Un modello di dati completo e ottimizzato evita la ridondanza dei dati, riduce i requisiti di archiviazione e consente un recupero efficiente e veloce delle informazioni.

4.1 Data Integration

In questo paragrafo verranno descritte le tabelle create che non erano presenti in precedenza nel database ma che si sono rese necessarie per le analisi svolte. Con il termine Data Integration si fa riferimento a quel processo in cui si uniscono dati provenienti da fonti diverse in un unico database. In questa tesi invece è stato necessario creare altre tabelle dalla rielaborazione di quelle esistenti in modo da ottenere dati aggregati pronti per essere utilizzati.
Le tabelle inserite nel database di Litokol sono quattro e verranno descritte nel dettaglio nel seguito del capitolo:

1. KPI_peso_TP_pag;
2. TAB_KPI;
3. KPI_rilevanza;
4. TAB_cluster.

1. KPI_peso_TP_pag
La prima tabella creata è stata quella che consentiva di dividere i metodi di pagamento sulla base della quantità di giorni di dilazione che permette ciascun fornitore. Il nome della tabella è preceduto dalla parola KPI per distinguere le nuove tabelle da quelle presenti in precedenza nel database. All'interno della tabella ogni fornitore è associato ad una classe secondo la suddivisione spiegata nel capitolo 3.2

sezione *DW_TP_PAG*. La tabella così creata ha cinque attributi (in grassetto la chiave primaria), il codice si trova in appendice A:

- **KEY_FORNITORE**
- COD_TP_PAG
- DES_TP_PAG
- Classe
- Peso classe

2. Tab_KPI

Dopodiché si è passati alla creazione delle tabelle per i KPI selezionati. Per iniziare si è creata una vista che leggesse i dati dalla tabella dei fatti per ottenere una tabella che avesse un record per numero ordine, calcolasse la quantità totale ordinata, eventuali giorni di ritardo e anticipo e descrivesse lo stato dell'ordine raggruppando le righe per numero ordine, fornitore, unità di misura utilizzata e data di consegna. Le successive creazioni di tabelle leggeranno da questa vista per ottenere i dati in modo più veloce.

Come già detto in precedenza, per l'applicazione degli algoritmi di clustering è stato necessario creare una tabella riassuntiva di tutti i KPI per singolo fornitore. Una frazione della tabella creata viene riportata di seguito (tabella 4.1), in cui la chiave primaria è la key fornitore, poi è presente il codice e i 5 KPI.

Tabella 4.1: Tabella riassuntiva KPI

KEY FORNITORE	COD FORNITORE	KPI_performance	KPI_tempestività	KPI_flessibilità	KPI_prezzo	KPI_puntualità
7	0004903	-0,55	3,33	0,78	0,39	0,5
52	2071511	1,00	11,33	-0,33	0,01	1,00
127	2038490	0,57	5,28	1,51	0,76	0,80
14	2043245	0,45	12,31	3,99	0,57	0,91
...	...	...	...	...	...	...

La tabella 4.1 è il risultato dell'unione di altre cinque tabelle (si veda il codice in appendice B, ognuna usata per il calcolo del singolo KPI e anch'esse aventi come chiave primaria la key fornitore. Tra queste tabelle è stata successivamente eseguita una left join per ottenere la tabella finale e poterla dare in input agli algoritmi di cluster.

3. KPI_rilevanza

É stata successivamente creata la tabella riassuntiva degli indici di rilevanza descritti nel capitolo 3.4.6. Ad ogni codice fornitore (chiave primaria della tabella) sono stati associati i due indici. Il primo descrive la percentuale di importo pagato

rispetto alla spesa totale per effettuare tutti gli acquisti; mentre il secondo descrive la percentuale di quantità acquistata da quel fornitore rispetto alla quantità di materia acquistata in totale.

4. TAB_Cluster
Passiamo ora all'ultima tabella creata, questa è stata creata in modo completamente diverso dalle altre, in quanto i dati inseriti derivano dall'esecuzione degli algoritmi di clustering. Una volta creata la tabella con i codici dei fornitori e il relativo cluster di appartenenza utilizzando l'editor Jupyter Notebook è stato creato un file **.txt**[1] esterno per poterlo caricare all'interno del database di Litokol. Usando SQL Server Management studio è stato possibile inserire la tabella esterna all'interno del database di Litokol tramite un passaggio guidato che richiedeva il percorso del file di origine, la tipologia di dati e la chiave primaria.

Si tenga in considerazione che i nomi date alle tabelle in questa fase verranno modificati in seguito per consentire a chi utilizza queste tabelle di comprendere meglio cosa esse contengono.

4.2 Creazione dello Snowflake Schema

In questa sezione varrà descritto il procedimento di creazione del modello dati che poi sarà utilizzato per leggere i dati con lo strumento di Data Visualization. A nostra disposizione vi era l'intero DWH di Litokol s.p.a. contenente tutti i dati provenienti dal loro gestionale relativi ad acquisti e vendite. Essendo la nostra analisi incentrata sullo studio e l'analisi dell'efficacia dei fornitori utilizzati durante la fase di acquisto dei prodotti per le attività operative, è stato necessario costruire un Data Mart tematico (si veda capitolo 1.2.1) con i dati relativi agli acquisti. Sono state quindi individuate le tabelle con le informazioni necessarie e, dove mancavano, ne sono state create delle nuove. Per rappresentare meglio i dati si è ritenuto più adeguato la creazione di un Snowflake Schema (si veda capitolo 1.2.1). La creazione dello schema è avvenuta facendo uso del tool Framework Manger di IBM Cognos (capitolo 4.2.1), con il quale è stato possibile leggere tabelle dal DWH e relazionarle tra loro.

Litokol organizza i suoi dati secondo tabelle normalizzate collegate tra loro tramite le chiavi primarie. Le tabelle in cui soffermeremo la nostra analisi e che andremo ad analizzare nel dettaglio sono:

- DW_ACQUISTI: la tabella dei fatti, la chiave primaria è data dall'unione delle chiavi primarie delle tabelle a lei collegate. Ogni record corrisponde ad

[1] File di testo.

un acquisto effettuato dall'azienda descritto da 85 campi.

- DW_FORNITORE: tabella dell'anagrafe fornitori, chiave primaria *KEY_FORNITORE*, in cui ogni record è relativo ad un fornitore descritto da 42 campi.
- DW_PRODOTTO: tabella dell'anagrafe prodotti, chiave primaria *KEY_PRODOTTO*, in cui ogni record è relativo ad un prodotto descritto da 45 campi.
- DW_ST_RIGA: tabella che raccoglie tutte le tipologie di stato dell'ordine, ad esempio se è concluso oppure no, chiave primaria *KEY_ST_RIGA*.
- DW_UM: tabella delle unità di misura utilizzate per valorizzare le quantità ordinate chiave primaria *KEY_UM*.
- DW_CATPRO08: tabella delle diverse categorie di prodotto presenti all'interno dell'azienda, chiave primaria *KEY_CATPRO08*.
- DW_ZONA_02 - 03 - 04: sono tre tabelle che descrivono rispettivamente la nazione, provincia e regione del fornitore di riferimento.
- KPI_PESO_TP_PAG: tabella che descrive la classe e il relativo peso della tipologia di pagamento associata.
- TAB_KPI: tabella che raccoglie, per ogni fornitore, il valore dei KPI selezionati, chiave primaria *KEY_FORNITORE*.
- TAB_cluster: tabella in cui si descrive ciascun fornitore secondo il cluster di appartenenza (si veda capitolo 4.1), chiave primaria *COD_FORNITORE*.

Le tabelle sono collegate tramite relazioni tra le chiavi primarie. In particolare abbiamo la tabella dei fatti *DW_ACQUISTI* collegata a *DW_FORNITORE*, *DW_PRODOTTO*, *DW_UM* e *DW_ST_RIGA*. La relazione creata è uno a molti; vicino alla tabella dei fatti si vede *1...n*, mentre vicino alle tabelle delle dimensioni vi è *1...1*. Ad ogni acquisto è associato un fornitore che è presente una sola volta nella tabella fornitori, mentre per ogni fornitore della tabella fornitori sono presenti uno o più righe della tabella degli acquisti; ragionamento analogo si può fare con le altre tabelle.

La tabella *DW_FORNITORE* è collegata con *DW_ZONA02*, *DW_ZONA03*, *DW_ZONA04* e *KPI_PESO_TP_PAG* con una relazione uno a molti, mentre con le tabelle *TAB_CLUSTER* e *TAB_KPI* con una relazione uno a uno in quanto per ogni fornitore è presente un solo cluster e un solo gruppo di KPI.

Per finire la tabella *DW_CATPRO08* è collegato con la tabella *DW_PRODOTTO* con una relazione uno a molti.

La figura 4.1 mostra come si presenta lo Snowflake Schema creato con le relazioni tra le tabelle.

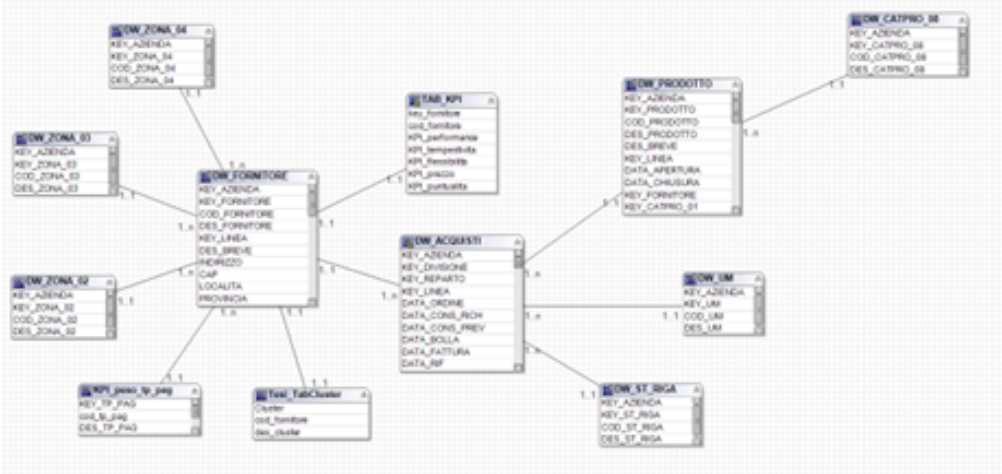


Figura 4.1: Snowflake Schema.
Fonte: elaborazione personale.
Tool: Framework Manager.

4.2.1 Framework Manager (FM)

Framework Manager (FM) è un tool di IBM Cognos grazie al quale è possibile realizzare il modello dati, fare le quadrature[2], analizzare i dati che sono stati inseriti nelle tabelle e visualizzare le join tra le tabelle. Al momento della creazione del modello si deve selezionare il tipo di sorgente da cui prendere i dati; nel mio caso i dati saranno ricavati direttamente dal DWH di Litokol. É importante considerare che su Framework Manager non si portano fisicamente i dati ma si creano le strutture per andare a interrogarli e a leggerli. FM tradurrà tutte le operazioni in interrogazioni SQL che estrapolano i dati dalle tabelle sorgenti. Per creare il modello si sono seguite le linee guida con cui Mediamente Consulting è solita lavorare. Il primo step è la creazione di due namespace[3]:

- Foundation Object View: strato "fisico" che include le tabelle così come sono nel DWH e nel quale si creano le relazioni (join).
- Consolidation Object View: una volta create le relazioni si richiamano le tabelle con il comando *select* di SQL server. In questo strato si devono rendere più comprensibili possibili i nomi di campi e tabelle che si leggeranno nella BI; verranno perciò cambiati i nomi delle tabelle e dei campi formattandoli come

[2]tecniche che hanno lo scopo di ottenere dati conformi tra il DWH e la BI

[3]Spazi di dati separati tra loro

si desidera; si possono inoltre fare dei campi calcolati (ad esempio concatenare tra loro due o più campi).

Nel caso studio lo strato di Foundation Object View conterrà l'insieme di tabelle selezionate per costruire lo snowflake schema. Il Consolidation Object View sarà invece formato da una rielaborazione di quelle tabelle (si veda figura 4.2), si avrà la tabella *Fornitore*, *Acquisti*, *Prodotto*, *Stato Ordine*, *KPI* e *Cluster*. La tabella Fornitore ha al suo interno, una selezione degli attributi presenti nella *DW_FORNITORE* e alcune informazioni aggiuntive derivante da altre tabelle, come la specifica della posizione geografica e dalla metodologia di pagamento a lui associata. Anche per la tabella Acquisti sono state fatte le stesse considerazione, per prima cosa sono stati selezionati solo gli attributi necessari e successivamente sono state aggiunte le informazioni relative all'unità di misura utilizzate e al dettaglio di anno e mese dell'acquisto. La tabella prodotto ha la possibilità di verificare anche la categoria merceologica di appartenenza. Infine, la tabella dei KPI racchiude tutti gli indicatori disponibili compresi quelli relativi alle performance, quelli di rilevanza e gli stessi indicatori ma standardizzati. All'interno di questa tabella è inoltre stato creato l'indice SVR grazie alla formazione di un campo calcolato secondo le equazioni 3.1 e 3.2.

Una volta creato il modello e riportato nella Consolidation Object View è possibile la creazione e pubblicazione del package. Questa sezione viene letta dai tool di BI per importare i dati.

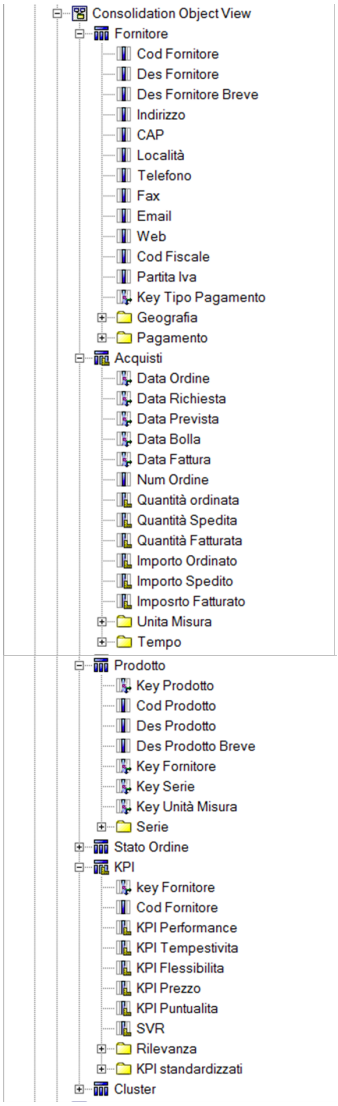


Figura 4.2: Consolidation Object View.
Fonte: elaborazione personale.
Tool: Framework Manager.

Capitolo 5

Data Visualization

L'ultimo capitolo di questa tesi descriverà i passaggi che sono stati realizzati per ottenere un insieme di dashboard che descrivono l'andamento delle performance dei fornitori. Con la Data Visualization le informazioni che provengono da strutture dati di grandi dimensioni si rendono di facile e immediata comprensione grazie all'utilizzo di visualizzazioni in grado di identificare i collegamenti e relazioni tra i dati. L'output di questo processo sarà una rappresentazione dei dati con forme e colori diversi tale per cui l'occhio e la mente umana possono comprenderli facilmente. Gli strumenti utilizzati sono pensati per coloro che utilizzeranno le dashboard per prendere decisioni riguardanti le strategie da attuare. La Business Intelligence è una disciplina che viene spesso presentata attraverso la visualizzazione dei dati poiché deve essere usato anche da coloro che non sono esperti in materia di analisi dati e l'utilizzo di colori, dimensioni e immagini sono tecniche utili a comprendere l'andamento positivo o negativo delle dimensioni in analisi e come esse si modificheranno nel tempo. Ogni progetto ha i suoi risultati e tali assumono valore solo al momento in cui essi sono compresi e valutati; grafici, tabelle, liste e dati di sommario sono gli strumenti fondamentali per riuscire a svolgere il lavoro di analista e acquisire conoscenza dai dati. I dati a noi affidati relativi al processo di approvvigionamento sono stati raccolti in dashboard interattive in cui è possibile navigare per raggiungere livelli di dettaglio sempre diversi.

L'importante non è visualizzare i dati ma dare uno strumento utile alle persone di business per prendere le decisioni. I punti su cui si basa la data visualization sono:

- Pensare all'utente: utilizzare un linguaggio comprensibile (non necessariamente tecnico) che tenga conto delle caratteristiche dell'utente.

- Evidenziare KPI: utilizzare pochi numeri che permettono di vedere l'andamento dell'azienda.

- Rappresentare l'organizzazione: ogni KPI deve avere un responsabile cui fare riferimento per chiedere approfondimenti o spiegazioni.
- Gestire i confronti: il numero assoluto di un indicatore spesso non ha valore, è importante confrontare il numero con il corrispondente del passato o con un valore target.
- Condividere il layout: rendere utile e comprensibile l'ambiente di analisi per l'utente finale. Non è detto che l'utilizzo corretto di un grafico per l'analista sia quello più funzionale per l'utilizzatore finale.
- Certificare i dati: l'informazione deve essere sicura, consistente. L'utente deve fidarsi di quello che sta guardando altrimenti non lo utilizzerà.

Vari sono gli strumenti che possono essere utilizzati per la visualizzazione dei dati che è possibile reperire dal mercato; l'approccio di vendita di una soluzione piuttosto che un'altra può essere *mirato* o *libero*. L'approccio mirato si ha quando il cliente chiede uno specifico strumento ed è perciò necessario conoscere tutti i punti di forza di quello strumento per poter ottenere la migliore rappresentazione. L'approccio *libero* invece si ha quando il cliente non ha una preferenza in particolare e da carta bianca dello strumento di utilizzo che ritiene maggiormente utile l'azienda di consulenza a cui si affida. In questo caso è quindi di fondamentale importanza capire quale strumento, tra quelli presenti nel mercato, sia più adeguato a rappresentare il contesto in analisi.

5.1 IBM Cognos

IBM Cognos è un tool che permette di rappresentare i dati con tabelle e grafici ma anche di fare delle operazioni matematiche per trasformarli, raggrupparli o segmentarli. Cognos è una piattaforma di Business Intelligence che supporta tutta l'attività analitica. Per questa tesi è stato scelto questo tool perché era già utilizzato dall'azienda cliente per la rappresentazione dei dati di vendita. É stato quindi richiesto di non cambiare lo strumento. IBM Cognos si compone di due parti principali:

- Team Content: è la parte pubblica, visibile ad ogni utilizzatore della piattaforma, in cui si inseriscono dashboard e report pronti per essere utilizzati;
- My Content: è la parte privata, in cui si trovano dashboard e report in elaborazione ed è visibile solo agli sviluppatori.

Le funzionalità di questo strumento sono molteplici, in particolare viene sfruttato per creare dashboard o report. La dashboard è utilizzata principalmente per

rispondere alle domande poste dall'organizzazione in modo visuale; può essere statica o dinamica e può essere implementata utilizzando filtri per dare risalto alle informazioni utili rispetto a tutte quelle presenti. Le sorgenti dati da cui è permesso leggere le informazioni sono varie: i packages di Framework Manager, il modello dati, file XLSX o CSV o dataset; i data source possono essere di singolo tipo o multiplo e questo è uno dei grandi cambiamenti che hanno portato la visualizzazione dati ad un livello superiore rispetto al passato. I report invece si differenziano dalle dashboard perché sono più mirati verso un preciso obiettivo e non sempre si basano su visualizzazioni grafiche ma vengono utilizzate liste o tabelle. Sono principalmente utilizzati in caso di reporting istituzionale, report pixel perfect e report stampabili e schedulabili.

Per raggiungere l'obiettivo di questa tesi faremo uso di tre dashboard diverse:

1. **Dashboard Globale:** in questa sezione verranno rappresentati i dati aggregati di ogni fornitore, per visualizzare una classica dei fornitori, da quelli più performanti a quelli meno performanti.

2. **Dashboard per Cluster:** in questa sezione si mostrano i risultati ottenuti dagli algoritmi di clustering.

3. **Dashboard di Dettaglio:** in questa sezione viene mostrata la balance scorecard di ogni fornitore, con il dettaglio di ogni singolo indicatore anno dopo anno.

La descrizione delle dashboard verrà ripresa in seguito più nel dettaglio. Sarà possibile navigare tra le diverse pagine attraverso semplici click del mouse; in particolare per arrivare alla pagina di dettaglio di uno specifico fornitore basterà selezionarlo dalla pagina principale. Questo tipo di procedimento prende il nome di *Drill Through* e permette di creare una pagina di destinazione in cui si aumenta il livello di dettaglio della informazione selezionata (in questo caso del fornitore).

Il primo step di creazione di una dashboard su IBM Cognos è la selezione della sorgente dati da cui vengono lette le informazioni. Per la nostra analisi è stato scelto il modello dati creato in precedenza su Framework Manager (si veda capitolo 4.2.1). I grafici selezionati sono stati scelti tenendo in considerazione il tipo di informazione e l'utilizzo che ne deve essere fatto.

5.2 Dashboard Globale

La dashboard Globale è quella che viene mostrata inizialmente, in cui si presenta l'andamento dei principali indicatori relativi al processo di approvvigionamento. La dashboard si presenta come in figura 5.1, alcuni dati sensibili sono stati oscurati.

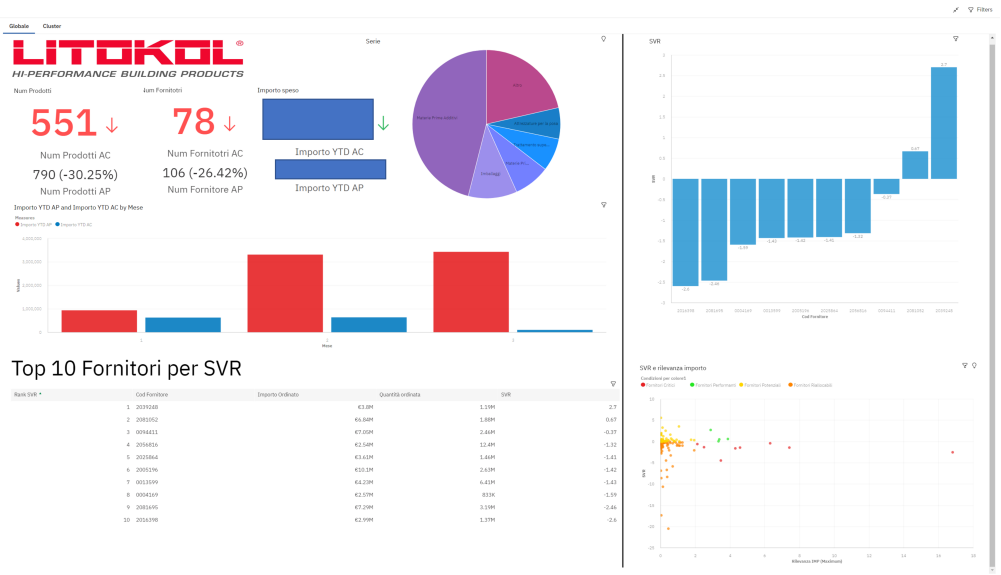


Figura 5.1: Dashboard Globale.
Fonte: elaborazione personale.
Tool: IBM Cognos.

É possibile dividere la pagina in sezioni tematiche: la sezione di sinistra mostra i dati relativi al monitoraggio dei fornitori secondo indicatori di business considerando due anni di analisi (quello corrente e quello precedente), mentre la sezione di destra racchiude le informazioni riguardanti l'indicatore globale e quello di rilevanza di importo descritti nel capitolo 3.4 calcolati sull'intero arco temporale considerato.

Come si può notare nell'angolo in alto a destra sono presenti due tab: il tab *Globale* (quello visualizzato) e il tab *Cluster* (che vedremo in seguito). Subito al di sotto dei tab è stata inserito il logo dell'azienda e tre indicatori aggregati:

- Num Prodotti: è il numero di prodotti che sono stati acquistati dall'inizio dell'anno (2023) fino al giorno corrente.
- Num Fornitori: è il numero di fornitori da cui sono stati acquistati i prodotti dall'inizio dell'anno (2023) fino al giorno corrente.
- Importo speso: calcola l'ammontare dell'importo speso per le materie prime dall'inizio dell'anno fino alla data corrente.

Tutti i valori sono comparati con il valore dell'indicatore nello stesso periodo dell'anno precedente, mostrando il numero e la percentuale di variazione. Per realizzare questi indicatori sono stati creati dei campi calcolati in Cognos. Questo

strumento si basa su un linguaggio di programmazione proprio di IBM; anche se, alcuni costrutti sono simili ad altri linguaggi, il modo di definire i calcoli e il modo di aggregarli e calcolarli è molto diverso. Il linguaggio si basa sulla creazione di funzioni indipendenti che utilizzano referenzialità e puntamenti sui campi e sulle strutture definite nella sorgente dati.

Di seguito viene riportato un esempio di calcolo utilizzato per visualizzare l'indicatore relativo all'importo tale speso nell'anno corrente e l'importo totale speso nell'anno precedente fino al mese desiderato. Sono due campi calcolati distinti, questo vuol dire che è possibile richiamarli anche singolarmente in altre visualizzazioni:

1. Importo speso YTD AC:
if([Consolidation Object View].[Acquisti].[Anno] = _year(current_date) and [Consolidation Object View].[Acquisti].[Mese] between 1 and _month(current_date))
then ([Consolidation Object View].[Acquisti].[Importo Ordinato])
else (0)

2. Importo speso YTD AP:
if([Consolidation Object View].[Acquisti].[Anno] = _year(_add_years(current_date, -1)) and [Consolidation Object View].[Acquisti].[Mese] between 1 and _month (current_date))
then ([Consolidation Object View].[Acquisti].[Importo Ordinato])
else (0)

In entrambi i casi vi è quindi un costrutto *if-then-else* che verifica inizialmente se la data a cui è associato un acquisto è dello stesso anno della data corrente e successivamente se il mese è compreso fra il primo mese dell'anno e il mese della data corrente. Se le condizioni risultano entrambe verificate allora il ramo then valorizzerà un nuovo campo con l'importo associato a quell'acquisto; altrimenti il ramo *else* valorizzerà il campo a zero. IN questo modo, andato a sommare tutti i campi otterremo solo il valore degli acquisti desiderato. Lo stesso ragionamento, ma modificando l'anno ad un valore pari all'anno corrente meno uno, è stato fatto per calcolare l'importo relativo all'anno precedente. Ogni campo è nominato con il percorso da cui deriva, ad esempio *[Consolidation Object View].[Acquisti].[Anno]* significa che il campo "anno" sta all'interno della tabella "Acquisti" nella sorgente dati di nome "Consolidation Object View".

Proseguendo con le visualizzazioni si vede un diagramma a torta a torta con le serie principali in cui vi è una fetta soprannominata "Altro" che racchiude tutte le altre serie più piccole e un diagramma a barre in cui viene mostrato quanto è stato speso al trascorrere dei mesi da gennaio fino al mese corrente con il ragionamento spiegato prima; le colonne rosse sono dati relativi all'anno precedente e le barre

blu sono quelle relative all'anno in corso.

Vi è poi nella parte inferiore una tabella in cui vengono mostrati i dieci fornitori più rilevanti per importo speso e vengono inoltre classificati sulla base del valore di SVR associato alle loro forniture. Questi fornitori vengono poi ripresi nella parte destra della dashboard in cui, grazie ad un istogramma vengono mostrati i fornitori e l'SVR associato in modo da osservare come la variazione dell'indicatore da un fornitore all'altro.

Per finire, in basso a destra, è stato inserito uno scatter plot per suddividere i fornitori in quattro gruppi secondo il valore di SVR e secondo la rilevanza di importo speso a ciascun fornitore considerando l'importo totale speso per gli acquisti (si veda l'indicatore Rilevanza importo al paragrafo 3.4.6). I quattro gruppi sono:

Tabella 5.1: Matrice fornitori

SVR	Fornitori Potenziali	Fornitori Performanti
	Fornitori Riallocabili	Fornitori Critici
	Rilevanza	

- Fornitori Performanti (in verde): hanno un alto valore di indice di rilevanza e un alto valore di SVR. Questi fornitori sono quelli in verde nel piano cartesiano e sono quelli che hanno ottime prestazioni e che vengono utilizzati maggiormente. Vista la significativa rilevanza di questi fornitori, un aumento seppur marginale del loro utilizzo potrebbe portare benefici per l'azienda.
- Fornitori Potenziali (in giallo): hanno bassa rilevanza ma un valore alto di SVR. Questi fornitori sono in giallo nel piano cartesiano. Avendo ottime prestazioni si potrebbe pensare di utilizzare questi fornitori per reperire materiali che provengono da fornitori meno performanti e aumentare così la percentuale di rilevanza.
- Fornitori Riallocabili (in arancione): hanno bassa rilevanza e basso SVR. Questi fornitori sono arancioni nel piano cartesiano. Dato la poca rilevanza è possibile pensare di acquistare i prodotti da altri fornitori con più alto SVR.
- Fornitori Critici (in rosso): hanno alta rilevanza ma basso SVR. Questi fornitori sono in rosso nel piano cartesiano. Visto l'impatto rilevante di questi

fornitori le basse performance possono sfociare in effetti negativi sulle prestazioni di servizio di Litokol. É essenziale attuare delle specifiche azioni di controllo e di gestione degli acquisti derivanti da questi fornitori dato che spesso non vi è la possibilità di sostituirli, data la rilevanza degli acquisti.

5.2.1 Navigazione interattiva della dashboard

Qualsiasi grafico di questa dashboard può essere utilizzato per aggiungere dei filtri e visualizzare i dati relativi a particolari fette di dati. Un esempio è riportato in figura 5.2 in cui sono stati eseguiti i seguenti passaggi:

1. selezione della categoria *Imballaggi* dal diagramma a torta;
2. aggiornamento automatico della dashboard per mostrare solo i dati relativi alla categoria di prodotto selezionata al punto 1;
3. selezione con il tasto destro del mouse del primo fornitore nella tabella in basso;
4. click con il maouse sul pulsante *Drill Through*;
5. indirizzamento verso la dashboard relativa al fornitore selezionato al punto 3.

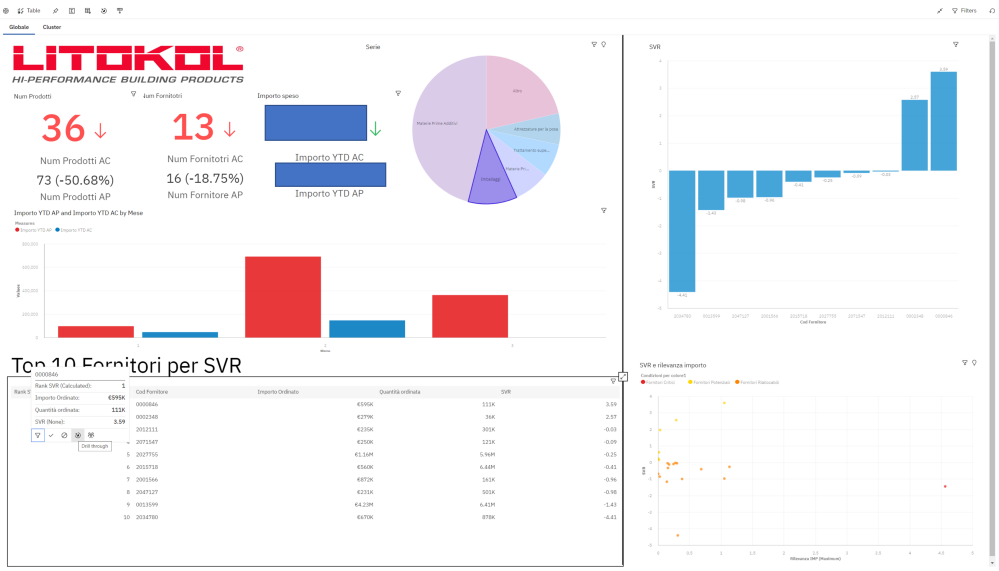


Figura 5.2: Dashboard Globale filtrata.
Fonte: elaborazione personale.
Tool: IBM Cognos.

In questa dashboard è stata quindi selezionata la serie *Imballaggi*. Ogni visualizzazione si modifica per mostrare solo i dati relativi agli imballaggi. In particolare, nella tabella in basso, sono adesso visualizzati i dieci fornitori con maggiore importo e maggiore SVR di questa serie, e lo stesso nello scatter plot alla sua destra. La finalità di questo è di confrontare i fornitori più simili per categorie di prodotto in modo da valutare se sia possibile approvvigionare il materiale da quelli più performanti.

Cliccato sul pulsante *Drill Through* è possibile navigare in un'altra dashboard che mostra i dati relativi ad un qualsiasi fornitore selezionato. Questa dashboard prende il nome di Dashboard di Dettaglio e verrà descritta in seguito.

5.3 Dashboard per Cluster

Questa dashboard mostra i dati relativi all'analisi di cluster effettuata (capitolo 3.6). É possibile visualizzarla cliccando sul tab relativo nella parte in alto a sinistra della pagina di Cognos. Questa dashboard si rende utile quando si vuole realizzare una ricerca del fornitore più adeguato tramite valore di singoli KPI. La figura 5.3 mostra come si presenta la pagina.

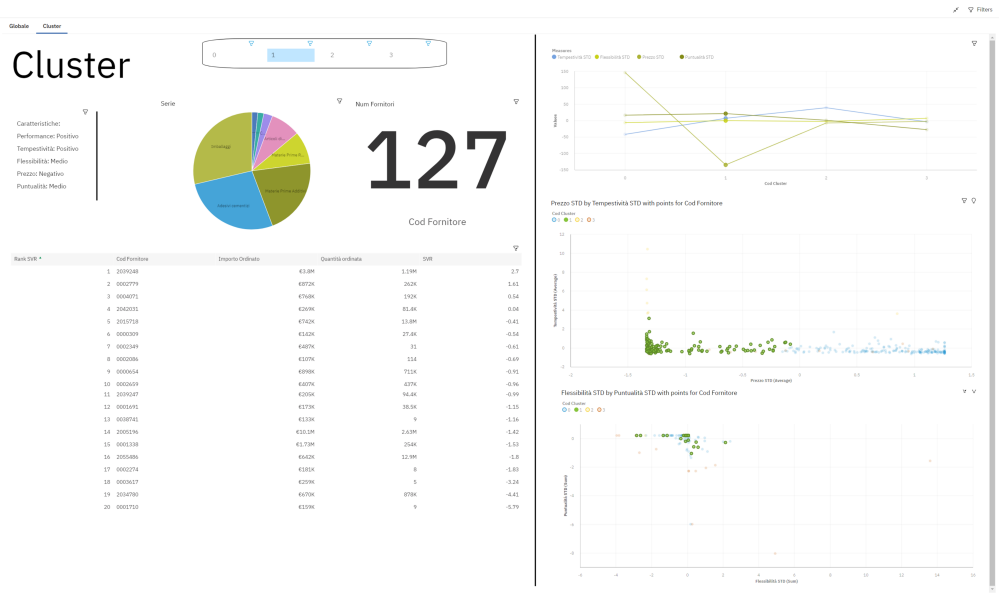


Figura 5.3: Dashboard Cluster.
Fonte: elaborazione personale.
Tool:IBM Cognos.

Anche in questa dashboard sono presenti due sezioni: la sezione di sinistra è quella relativa alla descrizione del cluster, mentre la sezione di destra mostra come

si posizionano i fornitori sulla base degli indicatori di prestazione.

In alto, accanto alla scritta Cluster, vi è un menù di selezione da cui è possibile selezionare il numero del cluster desiderato (in questo caso è stato selezionato il cluster 1). Successivamente, al di sotto sono elencate le caratteristiche del gruppo di fornitori in base agli indicatori di performance, tempestività, flessibilità, prezzo e puntualità seguendo una scala di bontà a tre valori: Positivo, Medio e Negativo. In questo modo l'azienda sarà in grado di visualizzare solo i fornitori che presentano determinate caratteristiche; ad esempio, se desidera un prodotto in tempi brevi andrà a selezionare il cluster con un indice di tempestività Positivo. Oppure, se fosse interessata ad un prodotto che consti poco e non ha tempi di consegna stretti, potrà selezionare il cluster con indice di prezzo positivo senza considerare il valore dell'indice di tempestività. Una volta scelto il cluster con gli indici adeguati alle richieste, è possibile vedere quali categorie di prodotto riforniscono i fornitori di quel gruppo e il numero di fornitori che lo compongono.

In basso vi è nuovamente la tabella presentata nella dashboard globale ma in questo caso mostra i dati relativi ai primi 20 fornitori secondo classifica SVR del cluster selezionato.

Nella parte destra della pagina sono presenti tre grafici: il primo mostra il grafico a linee che ha sull'asse delle ascisse i cluster e sull'asse delle ordinate la media dei KPI per ciascun gruppo. In questo caso i KPI sono quattro in quanto l'indicatore di performance non è stato usato dagli algoritmi per eseguire la suddivisione.

I due grafici in basso mostrano invece la distribuzione dei fornitori secondo i quattro indicatori. In figura 5.3 sono evidenziati solo i fornitori appartenenti al cluster 1 perché è quello selezionato dal menù iniziale.

Anche con questa dashboard è possibile filtrare ad un livello di dettaglio più approfondito i dati. Grazie al diagramma a torta è possibile selezionare una categoria di prodotti come si vede in figura 5.4.

Avendo selezionato la categoria *Imballaggi* si è modificato il numero di fornitori è sceso in quanto sette sono i fornitori che riforniscono la categoria Imballaggi e che fanno parte del cluster 1. La tabella sottostante mostra questi fornitori classificati seconso SVR e, anche in questo caso, è possibile selezionando un qualsiasi fornitore, eseguire un *Drill through* per navigare verso la dashboard di dettaglio relativa a quel fornitore.

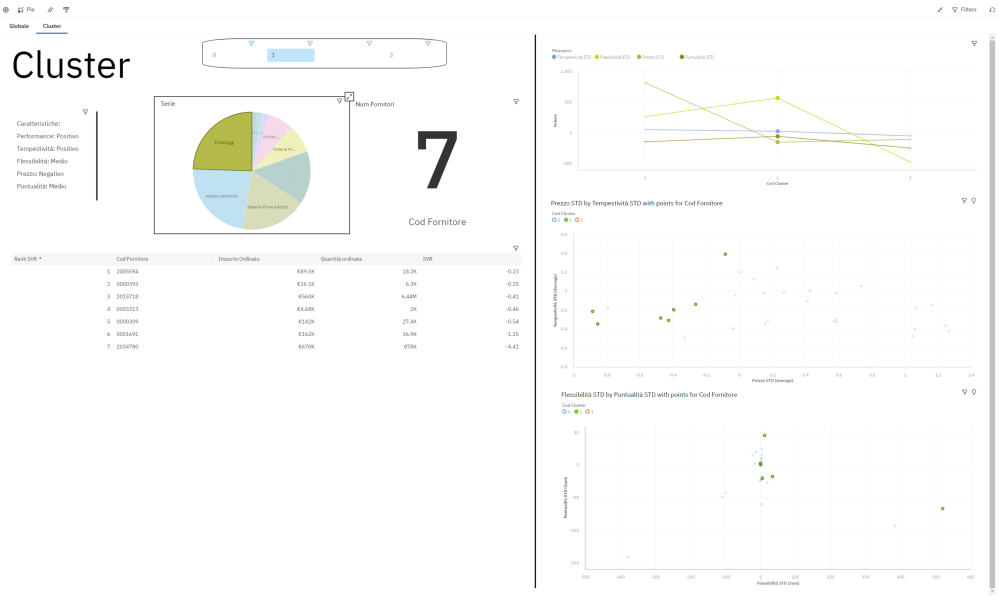


Figura 5.4: Dashboard Cluster filtrata.
Fonte: elaborazione personale.
Tool: IBM Cognos.

5.4 Dashboard di Dettaglio

Come ultima dashboard viene descritta quella con le informazioni di dettaglio di un singolo fornitore. Una volta compreso come un fornitore si posizione rispetto agli altri in termini di performance è necessario vedere quali sono le caratteristiche che lo influenzano maggiormente. Per questo è stata costruita la dashboard in figura 5.5. In questo modo ogni fornitore avrà la sua scheda di riferimento che racchiude tutte le informazioni importanti da conoscere al momento di una nuovo acquisto.

La figura mostra i dati relativi al fornitore di codice 0000846 e questo si vede dal filtro selezionato nella parte alta della pagina. La sezione evidenziata dai rettangoli rossi è infatti quella dedicata ai filtri, è possibile inserire delle condizioni per visualizzare solo parte dei dati. In particolare in questa sezione ci sono 3 filtri:

- Des serie: serve per selezionare una serie in particolare per mostrare solo i prodotti che ne fanno parte.
- Cod fornitore: per visualizzare solo il fornitore desiderato. Questo è l'unico filtro che sarà già selezionato al momento dell'apertura della dashboard.
- Cod prodotto: per selezionare un prodotto in particolare.

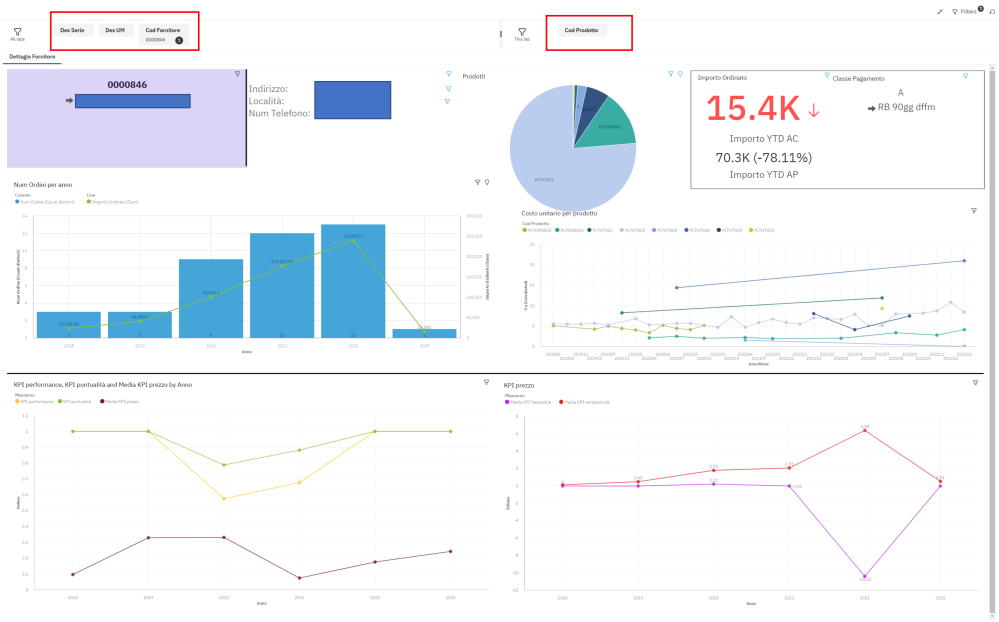


Figura 5.5: Dashboard di Dettaglio.
Fonte: elaborazione personale.
Tool: IBM Cognos.

I primi due stanno nel reparto *All tabs* perché sono quelli che prende in automatico al momento della selezione del fornitore dalla dashboard globale o da quella di cluster e, creando altri tab all'interno di questa dashboard li mantiene uguali per tutti; l'ultimo è inserito invece nella sezione *This tab* perché è un filtro non obbligatorio che può essere aggiunto successivamente alla apertura della dashboard e che rimane esclusivamente per questo tab.

La dashboard presenta in alto a sinistra i dati anagrafici del fornitore: il codice, il nome, l'indirizzo, la località in cui si trova la sede e il numero di telefono (questi sono dati sensibili che non è possibile mostrare perciò sono stati oscurati). Nel rettangolo di destra ci sono invece due indicatori, il primo mostra l'importo speso per il reperimento dei prodotti dall'inizio dell'anno alla data corrente confrontato con l'importo speso nello stesso periodo dell'anno precedente; il secondo indicatore invece mostra la classe e la tipologia di pagamento associata.

Proseguendo con la lettura della dashboard si ha un istogramma che mostra il numero di ordini effettuati nel corso degli anni (colonne blu) e la somma dell'importo associato (linea verde). Alla destra di questo istogramma si ha invece un diagramma a linee in cui viene rappresentato l'andamento nel tempo del costo unitario di ogni prodotto. Ogni linea colorata rappresenta un prodotto.

- per prima cosa si è creato un campo calcolato che permettesse di ricavare il

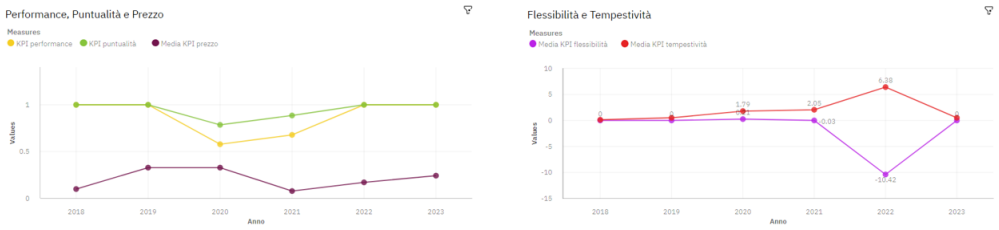
costo unitario del prodotto tramite la formula:

$$Cu = \frac{ImportoOrdinato}{QuantitàOrdinata} \tag{5.1}$$

- dopodiché si è aggiunto un campo calcolato per raggruppare non per singola data ma per anno e mese concatenando le due informazioni nel modo seguente:

  $[ConsolidationObjectView].[Acquisti].[Anno] * 100$
  $+ [ConsolidationObjectView].[Acquisti].[Mese]$

Passiamo ora alla descrizione dell'ultima parte, quella che si trova in basso. In questa sezione vi sono i grafici che mostrano l'andamento nel tempo degli indicatori di prestazione.



(a) *Performance, Puntualità e Prezzo* (b) *Tempestività e flessibilità*

Figura 5.6: Grafici a linee dei KPI.
Fonte: elaborazione personale.
Tool: IBM Cognos.

Il primo grafico mostra gli indicatori di Performance, di Puntualità e di prezzo, mentre il secondo mostra quelli di Tempestività e di Flessibilità. Questa separazione è dovuta alle scale in cui variano gli indicatori e al range di dati che possono assumere. Questi grafici permettono di fare delle valutazioni sull'andamento nel tempo, capire se ci sono stati dei peggioramenti importanti nelle prestazioni o dei sostanziali miglioramenti in modo da comprendere da cosa deriva il valore di SVR complessivo. Per riuscire a mostrare questi indicatori nel tempo è stato necessario creare dei campi calcolati su IBM Cognos utilizzando le stesse equazioni descritte al capitolo 3.4 ma aggregando per anno invece che per tutto il periodo di osservazione.

5.4.1 Navigazione interattiva della dashboard

Proseguiamo la navigazione delle informazione verso livelli di dettaglio sempre maggiori. Questa sezione è direttamente collegata al paragrafo 5.2.1 sovrastante.

Si pensi di voler acquistare un prodotto della categoria *Imballaggi* dal primo fornitore in termini di SVR. Una volta selezionato dalla Dashboard Globale, si atterra su quella di dettaglio per vedere il suo andamento nel corso del tempo. Il fornitore di codice 000846 si presenta come in figura 5.5; si può dire che le sue prestazioni siano state sempre positive ma che abbia avuto una piccola flessione nella puntualità e nella performance negli anni 2020 e 2021 che si è stabilizzata nuovamente al valore precedenti nel 2022. É possibile ricondurre questo andamento agli anni in cui l'Italia ha affrontato il periodo Covid, in cui il reperimento di materie prime era molto difficile. Ciò spiega sia il peggioramento delle tempistiche, sia l'incremento di prezzo che si può vedere dalla linea viola in figura 5.6a. La cosa fondamentale da osservare è che le curve hanno comunque un andamento crescente nel periodo successivo perciò è tornato ad avere buone prestazioni. Il secondo grafico in figura 5.6b, mostra invece che il fornitore nell'ultimo anno trascorso ha consegnato la merce in anticipo rispetto alla data richiesta da Litokol (indice di flessibilità negativo).

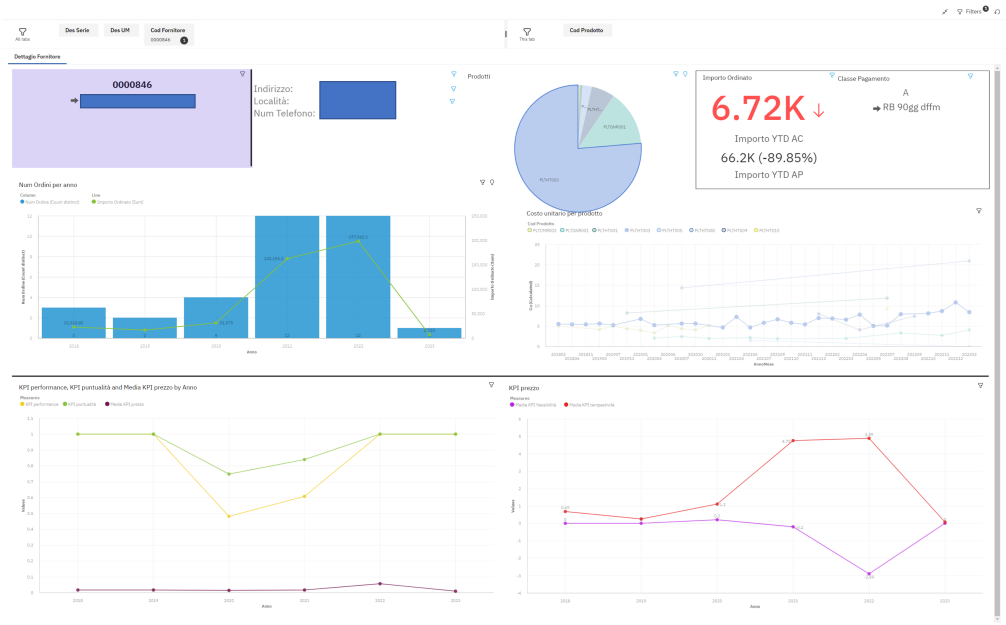


Figura 5.7: Dashboard di Dettaglio filtrata.
Fonte: elaborazione personale.
Tool: IBM Cognos.

Per finire, anche da questa dashboard può essere selezionato un qualsiasi prodotto dal diagramma a torta per visualizzare solo le informazioni relative ad esso, oppure è possibile selezionare solo uno degli anni presenti nell'istogramma. Nella figura 5.7 è stato selezionato il prodotto di codice PLTHT003 e così facendo tutte

le visualizzazioni della pagina si sono modificate automaticamente per visualizzare solo i dati relativi a quel prodotto.

Alla fine delle analisi sarà quindi più semplice capire se il fornitore è adeguato o meno a rifornire l'azienda dei suoi prodotti. Il fornitore 0000846 ha in generale buone prestazioni ed è riuscito a riprendersi velocemente dopo gli anni 2020 e 2021 nei quali ha avuto cali nella qualità del servizio. Questo può spingere Litokol ha incrementare il numero di prodotti acquistati da questo fornitore controllando che esso mantenga nuovamente negli anni gli stessi valori. Nell'ultimo anno ha consegnato la fornitura molto in anticipo rispetto la data prevista, perciò il rischio di rottura di stock dei prodotti che esso rifornisce è molto basso. Per questo può essere usato anche per prodotti di imballaggio provenienti da altri fornitori con indici di flessibilità e tempestività peggiori della media potendo richiedere sconti sulle grandi quantità e quindi ottenere costi unitari più bassi. Queste considerazioni si riflettano immediatamente sulla qualità del servizio che lo stesso Litokol dà ai suoi clienti. Un servizio rapito, puntuale e con bassi prezzi.

Capitolo 6

Conclusioni

Con questo elaborato si è realizzato uno strumento in grado di analizzare e leggere velocemente i dati relativi al processo di approvvigionamento di una delle aziende clienti di Mediamente Consulting facendo uso di vari strumenti in grado di gestire grandi quantità di dati organizzati in tabelle caratterizzati da grandi volumi di record e dimensioni di analisi. Il progetto è nato dalla necessità di prendere decisioni strategiche e operative per la vita dell'azienda basandosi sugli eventi avvenuti e avendo un quadro ampio e dettagliato del contesto per creare vantaggio competitivo rispetto alle altre aziende presenti sul mercato con lo scopo di arrivare prima dei competitor nella realizzazione di un servizio e con un prezzo modesto. Alla base di questa volontà vi è quindi la necessità di conoscere i fornitori e le loro prestazioni nel trascorrere del tempo. Dopo aver determinato il contesto di analisi e le esigenze dell'azienda cliente si è delineato il progetto da eseguire. Sono state eseguite tutte le fasi del sistema di supporto alle decisioni costruendo un processo fondato interamente dai dati e dalle misure che da essi si possono ricavare approfondendo vari temi come quello dell'estrazione e del caricamento, del data warehouse, delle tecniche di estrazione della conoscenza e di visualizzazione.

6.1 Implementazione e sviluppi futuri

Grazie alla creazione del modello dati utilizzato per la realizzazione di questo elaborato risulterà facile l'utilizzo di questo prodotto anche per sviluppi e modifiche future. Le dashboard realizzate saranno date in mano all'azienda cliente di Mediamente Consulting per la gestione del proprio business e potranno consultarle ogni qualvolta che deve essere presa una decisione di carattere operativo legata alla fornitura e applicazione di azioni correttive. Essendo strumenti direttamente collegati al database dell'azienda un'aggiunta o eliminazione di un fornitore o la crescita nel tempo degli acquisti effettuati risulteranno implementati real time

nella dashboard, perciò i dati visualizzati saranno costantemente aggiornati con i dati presenti fino alla data corrente.

In futuro sarà interessante rendere automatico anche la suddivisione in gruppi di fornitori tramite clustering che fino ad oggi è è soggetta a una classificazione statica. Così facendo avremo un raggruppamento che tenga conto delle modifiche di livelli di prestazioni dei fornitori nel corso del tempo.

Un altro possibile sviluppo sarà quello di integrare nell'indicatore complessivo (SVR) alcuni fattori qualitativi, come la qualità del servizio offerto in termini di: prontezza nelle emergenze, fornitura di documenti necessari in tempo e la capacità di evitare la ripetizione di errori, in modo da creare un indicatore capace di descrivere il fornitore da ogni punto di vista.

Acronimi

AC Anno Corrente

AP Anno Precedente

B2B Business to business

B2C Business to consumer

BI Business Intelligence

DSS Decision Support System

DWH Data Warehouse

ETL Extraction, Transformation, Load

FM Framework Manager

IT Information Technology

KPI Key Performance Indicator

SSE Sum of Square Error

SVR Service Vendor Rating

YTD Year To Date

Elenco delle figure

1.1 Struttura Data Warehouse. Fonte: slide corso Business Intelligence 9
1.2 Tabella Acquisti, Prodotti e Fornitori. Fonte: Elaborazione personale. 14

2.1 Processo KDD. Fonte: slide corso Business Intelligence 19
2.2 KDD Fasi. Fonte: slide corso Business Intelligence 20

3.1 Esplorazione dati al 31/01/2023. Fonte: elaborazione personale. Tool: IBM Cognos. 30
3.2 Importo Ordinato, Spedito, Fatturato nel 2022. Fonte: elaborazione personale. Tool: IBM Cognos. 31
3.3 Distribuzione Fornitori. Fonte: elaborazione personale. Tool: IBM Cognos e Rapid Miner. 33
3.4 Distribuzione Prodotti. Fonte: elaborazione personale. Tool: IBM Cognos e Rapid Miner. 35
3.5 Albero dei KPI. Fonte: elaborazione personale. 39
3.6 Riassunto KPI. Fonte: elaborazione personale. 47
3.7 Boxplot per la dispersione degli indici. Fonte: elaborazione personale. Tool: Jupyter Notebook. 49
3.8 Indici standardizzati. Fonte: elaborazione personale Tool: Jupyter Notebook. 50
3.9 Matrice di correlazione. Fonte: elaborazione personale. Tool: Jupyter Notebook. 50
3.10 Applicazione DBSCAN. Fonte: elaborazione personale. Tool: Jupyter Notebook. 52
3.11 Applicazione Kmeans. Fonte: elaborazione personale. Tool: Jupyter Notebook. 54
3.12 Dendrogramma. Fonte: elaborazione personale. Tool: Jupyter Notebook. 55
3.13 Cluster Fornitori Particolari. Fonte: elaborazione personale. Tool: Jupyter Notebook. 56

3.14 Validazione cluster. Fonte: elaborazione personale. Tool:Jupyter Notebook. 58
3.15 Singular Value Decomposition per cluster gerarchico. Fonte: elaborazione personale. Tool: Jupyter Notebook. 59
3.16 Radar Chart. Fonte: elaborazione personale 60

4.1 Snowflake Schema. Fonte: elaborazione personale. Tool: Framework Manager. 66
4.2 Consolidation Object View. Fonte: elaborazione personale. Tool: Framework Manager. 67

5.1 Dashboard Globale. Fonte: elaborazione personale. Tool: IBM Cognos. 71
5.2 Dashboard Globale filtrata. Fonte: elaborazione personale. Tool: IBM Cognos. 74
5.3 Dashboard Cluster. Fonte: elaborazione personale. Tool:IBM Cognos. 75
5.4 Dashboard Cluster filtrata. Fonte: elaborazione personale. Tool: IBM Cognos. 77
5.5 Dashboard di Dettaglio. Fonte: elaborazione personale. Tool: IBM Cognos. 78
5.6 Grafici a linee dei KPI. Fonte: elaborazione personale. Tool: IBM Cognos. 79
5.7 Dashboard di Dettaglio filtrata. Fonte: elaborazione personale. Tool: IBM Cognos. 80

Elenco delle tabelle

3.1 Dati relativi al prodotto EAC.102.80 (prima) 37
3.2 Dati relativi al prodotto EAC.102.80 (dopo) 37
3.3 Indice di Performance 42
3.4 Indice di Puntualità 42
3.5 Indice di Tempestività 43
3.6 Indice di Flessibilità 44
3.7 Indice di Prezzo 46
3.8 Rilevanza in percentuale 47
3.9 Tabella riassuntiva cluster 61

4.1 Tabella riassuntiva KPI 63

5.1 Matrice fornitori 73

Appendice A

Creazione tabella per tipologia pagamento

```
create table KPI_peso_tp_pag as

select distinct cod_tp_pag
, classe_tp_pag
, case classe_tp_pag
        when 'A' then 0.4
        when 'B' then 0.3
        when 'C' then 0.2
        when 'D' then 0.1
end peso_tp_pag
from(
select COD_TP_PAG
, case
        when COD_TP_PAG in ('8010', '2241','2140','2141','2151'
        ,'2311','2221','2220','2231','2331','2130','2131','2133'
        ,'6140','6131','6130','S023','S198','S080','S071','S022'
        ,'S192','S202')
        then 'A'
        when COD_TP_PAG in ('8009',     '2210','6313','2211','2103'
        ,'2106','2107','2109','2115','2102','2121', '2120','2125'
        ,'2127','2116','2117','2112','6313','6108','6103', '6271'
        , '6113','6114','6121','S057','S070','S024','S195','S024'
        , 'S201','S200','S199','S115','S197','S194','S016')
        then 'B'
        when COD_TP_PAG in ('6706','8001','9103','2110','2111'
```

```
        ,'8201','6110','6145','S053','S054','S193','S196','S015')
        then 'C'
        when COD_TP_PAG in ('0000','6308','0004','0093','6100'
        ,'6000','6999','6117','6117','6050','S108','S190','S111'
        ,'S109','S110','S191','8553','0014','8100','8102','0001'
        ,'2604','6098','6148','6104','8203','8550','8552','6101'
        ,'6300','S051','S001','S002','S003','S011')
        then 'D'
end classe_tp_pag
from DW_TP_PAG p, DW_FORNITORE f
where p.KEY_TP_PAG=f.KEY_TP_PAG
) tb
```

Appendice B

Creazione tabella riasuntiva KPI

```
create table TAB_KPI as

SELECT pe.key_fornitore, pe.cod_fornitore, pe.KPI KPI_performance,
t.KPI KPI_tempestivita, f.KPI KPI_flessibilita, pr.KPI KPI_prezzo
, case
when p.KPI is null then 1
else p.KPI
end KPI_puntualita
FROM KPI_performance pe
left JOIN KPI_puntualita p ON pe.key_fornitore = p.key_fornitore
left join KPI_tempestivita t on t.key_fornitore=pe.key_fornitore
left join KPI_flessibilita f on f.key_fornitore=pe.key_fornitore
left join KPI_prezzo pr on pr.key_fornitore=pe.key_fornitore
order by pe.cod_fornitore
```

Appendice C

Codice Python per cluster

```
!pip install pmdarima
!pip install pyodbc

import numpy as np # For data management
import pandas as pd # For data management
import pyodbc as po
import matplotlib.pyplot as plt
from pmdarima import auto_arima
import os

import matplotlib.pyplot as plt
from scipy.stats import zscore
import seaborn as sns
# For data visualization and specifically for pairplot()
import matplotlib.pyplot as plt # For data visualization

from sklearn.preprocessing import StandardScaler
#To transform the dataset
from sklearn.cluster import DBSCAN
# To instantiate and fit the model
from sklearn.metrics import pairwise_distances
# For Model evaluation
from sklearn.neighbors import NearestNeighbors
# For Hyperparameter Tuning

from sklearn.cluster import DBSCAN
from sklearn.datasets import make_blobs
```

```
from sklearn import metrics

from sklearn.cluster import KMeans

import plotly
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots

#connessione alla macchina litokol
connection_analytics = po.connect("Driver={SQL Server};"
                                  "Server=192.168.0.84;"
                                  "Database=DW_LITOKOL;"
                                  "uid=sa;pwd=%P4st!cc!n0$")
#lettura tabella tramite query
Tab_KPI = pd.read_sql_query("SELECT * FROM [dbo].[TAB_KPI]"
, connection_analytics)
KPI_df = pd.DataFrame(Tab_KPI)

#normalizzazione zscore
KPI_df['z_tempestivita']=zscore(KPI_df['KPI_tempestivita'])
KPI_df['z_performance']=zscore(KPI_df['KPI_performance'])
KPI_df['z_flessibilita']=zscore(KPI_df['KPI_flessibilita'])
KPI_df['z_prezzo']=zscore(KPI_df['KPI_prezzo'])
KPI_df['z_puntualita']=zscore(KPI_df['KPI_puntualita'])

#tabella con solo attributi normalizzati
zKPI_mx = KPI_df[["cod_fornitore", "z_tempestivita", "z_flessibilita"
,"z_prezzo","z_puntualita"]]

#MODELLO 1: DBSCAN

#find optimal epsilon
#analisi del ginocchio per selezionare un epsilon corretto.
#Approccio qualitativo
#selezionare valore epsilon corrispondente al punto di flesso
#n_neighbors = numero minimo di elementi in un cluster
#(parametro necessario per effettuare l'analisi che permette
#di definire l'epsilon)
from sklearn.neighbors import NearestNeighbors
neigh = NearestNeighbors(n_neighbors=5)
```

```
nbrs = neigh.fit(zKPI_mx.loc[:, zKPI_mx.columns.difference
(['cod_fornitore'])])
distances, indices = nbrs.kneighbors(zKPI_mx.loc[:, zKPI_mx.columns.
difference(['cod_fornitore'])])
distances = np.sort(distances, axis=0)
distances = distances[:,1]
plt.plot(distances) #visualizzazione del grafico KNN

#dbscan
#epsilon: distanza minima che devono avere due campioni per esser
#considerati vicini
#min samples: numero campioni minimo per definire un cluster.
#In generale a questo iper parametro va dato un
#valore >= numero dimensioni dataset + 1.
from sklearn.cluster import DBSCAN
from sklearn.datasets import make_blobs
from sklearn import metrics
db = DBSCAN(eps=1.5, min_samples=5).fit(zKPI_mx
.loc[:, zKPI_mx.columns.difference(['cod_fornitore'])])
clusters = db.labels_
print("Cluster created: " + str(list(set(clusters))))
#output dbscan
model_output_DBSCAN = zKPI_mx
#assegnamo la nuova label del cluster aggiungendo una nuova
#colonna al dataframe
model_output_DBSCAN['dbscan_label'] = clusters

#divido i due cluster in due dataframe diversi
Fornitori0 = model_output_DBSCAN
[model_output_DBSCAN['dbscan_label'] == 0 ]
FornitoriParticolari = model_output_DBSCAN[model_output_DBSCAN
['dbscan_label'] == -1 ]
Fornitori1 = Fornitori0.loc[:,
Fornitori0.columns.difference(['dbscan_label'])]

#MODELLO 2: KMeans
#optimal k for kmeans
#elbow method
from sklearn.cluster import KMeans
Sum_of_squared_distances = []
K = range(1,20)
```

```
for k in K:
    km = KMeans(n_clusters=k)
    km = km.fit(Fornitori0.loc[:, Fornitori0.columns.difference
    (['cod_fornitore','dbscan_label'])])
    Sum_of_squared_distances.append(km.inertia_)
plt.plot(K, Sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum_of_squared_distances')
plt.title('Elbow Method For Optimal k')
plt.show() #visualizzazione Elbow Method

A = 'z_prezzo'
B = 'z_flessibilita'
C = 'z_tempestivita'
D = 'z_puntualita'
X = Fornitori1[[A,B,C,D]]

# A list holds the silhouette coefficients for each k
silhouette_coefficients = []

# Notice you start at 2 clusters for silhouette coefficient
for k in range(2, 11):
  kmeans = KMeans(n_clusters=k)
  kmeans.fit(X)
  score = silhouette_score(X, kmeans.labels_)
  silhouette_coefficients.append(score)

#plt.style.use("fivethirtyeight")
plt.figure(figsize=(16,8))
plt.plot(range(2, 11), silhouette_coefficients, 'bx-')
plt.xticks(range(2, 11))
plt.xlabel("Number of Clusters")
plt.ylabel("Silhouette Coefficient")
plt.title('Silhouette')
plt.savefig('Silhouette US.jpeg')
plt.show() #visualizzazione grafico Silhouette

# Crea un'istanza di KMeans con 2 cluster
kmeans = KMeans(init="k-means++",n_clusters=2)

# Addestra il modello di K-means sul dataset
```

```
kmeans.fit(X)

# Assegna le etichette dei cluster a ogni oggetto del dataset
Fornitori1['Cluster'] = kmeans.predict(X)
Fornitori1.to_excel('Cluster US=4.xlsx')

# Crea un grafico a dispersione dei dati
plt.scatter(Fornitori1[A],Fornitori1[C], c=Fornitori1['Cluster']
, cmap='viridis')
plt.scatter(kmeans.cluster_centers_[:, 0],
kmeans.cluster_centers_[:, 1], s=200, c='red', marker='x',
label = 'Centroids')
plt.title('Clusters k-means')
plt.xlabel(A)
plt.ylabel(C)
#plt.savefig('Cluster k-means US=4')
plt.show()

#per il valore silhouette
labels = kmeans.labels_
print('Silhouette:',metrics.silhouette_score(X, labels,
metric='euclidean'))

#Cluster Visualization example on kmeans
#Subplots cycle
import plotly
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots

model_plot = Fornitori1.loc[:, Fornitori1.columns.difference
(['cod_fornitore'])]
fig = go.Figure()
fig = make_subplots(rows=model_plot['Cluster'].nunique(), cols=1
, subplot_titles= str(model_plot['Cluster'].unique()),
shared_xaxes=True, vertical_spacing=0.04)
color = ["blue","red","green",'violet','pink','yellow']
i = 0
for cluster_label in model_plot['Cluster'].unique():
    i +=1
    fig.add_trace(go.Scatter(x=model_plot.loc[model_plot['Cluster']
```

```
    == cluster_label ,model_plot.columns.difference(['Cluster']) ]
    .mean().index,y=model_plot.loc[model_plot['Cluster']
    == cluster_label ,model_plot.columns.difference(['Cluster']) ]
    .mean(),mode='lines+markers', line=dict(color=color[i-1]),
    visible=True), row = i, col = 1)
fig.update_xaxes(type='category')
fig.update_layout(height=1000, width=1500,
title_text="Cluster analysis")

fig.show()

var = 'cluster_'
plotly.offline.plot(fig, filename
='C:/Users/Mediamente/Desktop/'+var+'.html

#SVD
u,s,v = np.linalg.svd(Fornitori1.loc[:, Fornitori1.columns.difference
(['cod_fornitore','cluster'])], full_matrices=True)
labels= ['SV'+str(i) for i in range(1,3)]
svd_df = pd.DataFrame(u[:,0:2], index=Fornitori1["Cluster"].tolist()
, columns=labels)
svd_df=svd_df.reset_index()
svd_df.rename(columns={'index':'Cluster'}, inplace=True)
# specify colors for each continent
color_dict = dict({0:'Blue', 1: 'Red'})
# Scatter plot: SV1 and SV2
sns.scatterplot(x="SV1", y="SV2", hue="Cluster", palette=color_dict,
data=svd_df, s=100, alpha=0.7)
plt.xlabel('SV 1', fontsize=16)
plt.ylabel('SV 2', fontsize=16)

#MODELLO 2: DBSCAN

#find optimal epsilon
#analisi del ginocchio per selezionare un epsilon corretto.
#Approccio qualitativo
#selezionare valore epsilon corrispondente al punto di flesso
#n_neighbors = numero minimo di elementi in un cluster (parametro
necessario per effettuare l'analisi che permette di definire l'epsilon)
from sklearn.neighbors import NearestNeighbors
neigh = NearestNeighbors(n_neighbors=5)
```

```
nbrs = neigh.fit(FornitoriParticolari
.loc[:, FornitoriParticolari.columns.difference(['cod_fornitore'
,'dbscan_label'])])
distances, indices = nbrs.kneighbors(FornitoriParticolari
.loc[:, FornitoriParticolari.columns.difference(['cod_fornitore'
,'dbscan_label'])])
distances = np.sort(distances, axis=0)
distances = distances[:,1]
plt.plot(distances) #visualizzazione grafico KNN

#dbscan
#epsilon: distanza minima che devono avere due campioni per esser
#considerati vicini
#min samples: numero campioni minimo per definire un cluster. In
#generale a questo iper parametro va dato un valore >= numero
#dimensioni dataset + 1.
from sklearn.cluster import DBSCAN
from sklearn.datasets import make_blobs
from sklearn import metrics
db = DBSCAN(eps=4, min_samples=2).fit(FornitoriParticolari
.loc[:, FornitoriParticolari.columns.difference(['cod_fornitore'
,'dbscan_label'])])
clusters = db.labels_
print("Cluster created: "+str(list(set(clusters))))
#output dbscan
model_output_DBSCAN2 = FornitoriParticolari
#assegnamo la nuova label del cluster aggiungendo una nuova colonna
al dataframe
model_output_DBSCAN2['dbscan_label2'] = clusters
#conta elementi nei cluster creati
n_cluster=model_output_DBSCAN2[('dbscan_label2')]

#grafico 3D kmeans
from mpl_toolkits import mplot3d
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
fig = plt.figure()
ax = plt.axes(projection='3d')
ax.scatter3D(model_output_DBSCAN2["z_flessibilita"],
model_output_DBSCAN2["z_tempestivita"],model_output_DBSCAN2
```

```
["z_puntualita"], c=model_output_DBSCAN2["dbscan_label2"],
cmap='viridis')

ax.set_xlabel('z_flessibilita')
ax.set_ylabel('z_tempestivita')
ax.set_zlabel('z_puntualita')

plt.show() #visualizzazione grafico 3D

#creazione external tabele
Tab1 = Fornitori1.loc[:, Fornitori1.columns.difference
(['z_flessibilita','z_prezzo','z_puntualita','z_tempestivita'])]
Tab2 = FornitoriParticolari
.loc[:, FornitoriParticolari.columns.difference(['z_flessibilita'
,'z_prezzo'
,'z_puntualita','z_tempestivita','dbscan_label'])]
#Cambiamo numero al cluster
Tab2['dbscan_label2'] = Tab2['dbscan_label2'].replace
([0, 1, -1],[2, 3, 3])
Tab2.rename({'dbscan_label2': 'Cluster'}, axis=1, inplace=True)
#unione delle due tabelle
ExternalTable = pd.concat([Tab1, Tab2])
#creazione excel
ExternalTable.to_excel('ExternalTable.xlsx', sheet_name='A')
```

Riferimenti bibliografici

Arora, P., Varshney, S., et al. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, *78*, 507–512.

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*, *105*(9).

Chaudhuri, S., Dayal, U., & Ganti, V. (2001). Database technology for decision support systems. *Computer*, *34*(12), 48–55.

Dickson, G. W. (1966). An analysis of vendor selection systems and decisions. *Journal of purchasing*, *2*(1), 5–17.

D'Ovidio, F., & Villari, M. (2009). Soluzioni open source per la business intelligence geospaziale.

Jain, R., Singh, A. R., Yadav, H., & Mishra, P. (2014). Using data mining synergies for evaluating criteria at pre-qualification stage of supplier selection. *Journal of Intelligent Manufacturing*, *25*(1), 165–175.

Keen, P. G. (1980). Decision support systems: a research perspective. *Decision support systems: Issues and challenges: Proceedings of an international task force meeting*, 23–44.

Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(5), 431–443.

Syakur, M., Khotimah, B., Rochman, E., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP conference series: materials science and engineering*, *336*, 012017.

Trebuňa, P., Halčinová, J., Fil'o, M., & Markovič, J. (2014). The importance of normalization and standardization in the process of clustering. *2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 381–385.

Yahya, S., & Kingsman, B. (1999). Vendor rating for an entrepreneur development programme: a case study using the analytic hierarchy process method. *Journal of the Operational Research Society*, *50*(9), 916–930.

Yang, D., Ma, Z., & Buja, A. (2011). A sparse SVD method for high-dimensional data. *arXiv preprint arXiv:1112.2433*.

Sitografia

Cos'è la business intelligence? La tua guida alla BI e al perché è importante. (2022). https://www.tableau.com/it-it/learn/articles/business-intelligence#:~:text=Pi%C3%B9%20che%20indicare%20una%20%22cosa,obiettivo%20di%20migliorare%20le%20prestazioni.

Flyip. (2022). https://www.flyip.it/le-5v-dei-big-data-le-caratteristiche-di-una-massa-di-dati/#:~:text=Prima%20di%20entrare%20nel%20dettaglio,dati%20ricavati%20attraverso%20gli%20utenti

Normalizzare il componente dati. (2022). https://learn.microsoft.com/it-it/azure/machine-learning/component-reference/normalize-data

Tableau. (2022). https://www.tableau.com/it-it/learn/articles/business-intelligence/bi-dashboards#:~:text=Le%20dashboard%20di%20business%20intelligence,singolo%20schermo%20per%20panoramiche%20istantanee.