UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E GESTIONE DELLE IMPRESE

TESI DI LAUREA TRIENNALE

Cluster Analysis per la segmentazione della clientela utilizzando il Software SAS® ENTERPRISE MINER™

Relatore: Prof.ssa DULLI SUSI

Laureando: ZARAMELLA LUCA

Matricola: 484524

ANNO ACCADEMICO 2005-2006

Alla mia Famiglia

INDICE

Introduzione:

Il progetto realizzato e descritto in questa tesi punta a segmentare un insieme di dati mediante Cluster Analysis in modo da poter offrire un servizio differenziato alla clientela.
L'azienda che mi ha permesso di svolgere quest'esperienza di stage non aveva a disposizione un software per poter effettuare l'analisi ma aveva una serie di dati di loro potenziali clienti in forma cartacea. L'attività di stage si è quindi svolta in due parti, una di inserimento dati nel database aziendale e una di analisi utilizzando Enterprise Miner™ di SAS® disponibile in facoltà.
Nella tesi, oltre ad essere descritta tutta la parte di analisi dei dati mediante statistiche, grafici e risultati finali, è presente una parte teorica, che mira a spiegare la metodologia applicata e gli strumenti usati e a far capire l'importanza che assume il cliente nelle attuali realtà aziendali.

1. L’importanza del cliente

L’azienda è un sistema aperto e dinamico, fortemente influenzato dal mercato esterno e ricco di variabili. Nella realtà attuale questo mercato è incerto, difficile de prevedere e competitivo; ne consegue che la sopravvivenza delle aziende è garantita dalla ricerca di un vantaggio competitivo strategico in termini di informazioni che permettono di attuare scelte strategiche da effettuarsi in termini brevi. Per avere successo in un ambiente di questo tipo è necessario soddisfare i bisogni dei clienti, per questo motivo è fondamentale raccogliere dati relativi alla clientela effettiva o potenziale per poi estrarre da questi l’informazione utile che permetterà di individuare i soggetti più profittevoli per l’azienda. Per poter ricavare delle informazioni utili è però fondamentale inserire i dati in un database (database marketing) al quale possiamo applicare tecniche per segmentare la clientela ed indirizzare campagne di marketing diverse a segmenti diversi.

I dati inseriti nel database marketing devono essere integrati, riutilizzabili e non volatili. Nella customer table si trovano tutte le informazioni relative al cliente e questa è caratterizzata da un solo record per cliente che sarà identificato univocamente da una chiave.
L’impresa deve perciò cercare di mettere al centro di tutte le sue attività il cliente con l’obiettivo di capirne i bisogni e le richieste, cercando di offrirgli prodotti e servizi che lo soddisfino e che siano più attrattivi rispetto a quelli della concorrenza.
L’azienda che opera a scopo di business deve comunque servire solo determinati gruppi di clienti: quelli che possono creare profitto e quelli che con più probabilità possono accettare l’offerta, lasciando perdere quelli meno profittevoli.

1.1 L’orientamento al cliente

Per prima cosa è importante dire che un cliente acquista un determinato prodotto o servizio per soddisfare un determinato bisogno. Le aziende vendono i benefici di un determinato prodotto o servizio, quindi devono capire le aspettative del cliente e in basa a queste cercare di offrire dei prodotti o servizi atti a soddisfare queste necessità.

Ci sono però bisogni che non sono espressi dai consumatori, perché non sempre il consumatore è in grado di dire quali prodotti vuole o desidera. Tuttavia è in grado di esprimere che problemi ha incontrato con quelli che già esistono; compete all’azienda l’abilità di capirli, o di utilizzare l’ingegno, per sviluppare un prodotto o servizio innovativo e cercare di indirizzare il cliente verso questo.

Questo significa cercare di entrare in un mercato dove non c’è concorrenza, entrarvi per primi ed avere un vantaggio sui possibili entranti. Per mantenere tale vantaggio è necessario comunque lavorare per fidelizzare la clientela perché un cliente ritorna e riacquista un prodotto o servizio solo se soddisfatto del primo acquisto.

Un cliente di valore non è un cliente statico ma è un cliente che cambia nel tempo. L’azienda deve capire i comportamenti del cliente per cercare di prevedere i principali eventi che caratterizzano il suo ciclo di vita. La conoscenza del cliente, che è rappresentata dai dettagli delle interazioni con l’azienda, può essere raggruppata in alcune fasi che costituiscono il suo ciclo di vita (*), utile a comprendere meglio il suo comportamento. La metodologia che supporta con metodi quantitativi la gestione del ciclo di vita del cliente, è quella del Data Mining, che può essere applicato ad ogni fase di questo ciclo.

---

* : Berry & Linoff – Data Mining

1.2 Ciclo di vita del cliente

All'inizio, i clienti che si trovano nel mercato target, dimostrano un certo interesse, diventano poi clienti effettivi che, a seconda del loro comportamento, possono essere considerati dall'azienda clienti di grande valore, alto potenziale o di scarso valore. Poi ad un certo punto cessano di essere clienti.

Prospect: Sono tutti i potenziali clienti a cui si indirizza una campagna. Un evento importate per cercare di acquisire prospect è la campagna di acquisizione cioè un iniziativa commerciale diretta al mercato target con lo scopo di far conoscere un prodotto o un servizio ai potenziali clienti. È indispensabile mirare la campagna di marketing verso un segmento specifico in cui la clientela sia più sensibile.

Il mercato target è composto da clienti potenziali, quindi sconosciuti e perciò non esistono dati a disposizione. Però è possibile applicare tecniche di Data Mining ad analoghe campagne precedenti per cercare di avere un profilo di prospect che hanno risposto positivamente alle precedenti campagne. L'utilizzo del Data Mining in questo caso consente di costruire un modello previsivo che in base all'esperienza accumulata permetta di identificare i potenziali clienti. Ì record sono classificati in base a un comportamento futuro prevedibile o a un valore futuro stimato. I dati storici servono per costruire un modello che spieghi il comportamento osservato.

In alternativa può essere utile ricorrere all'esplorazione dei dati e al Data Mining non supervisionato e tracciare un profilo del potenziale cliente, oppure è possibile individuare i prospect che hanno un profilo analogo a quello dei clienti migliori che già l'azienda possiede.

Spesso è possibile utilizzare il Data Mining per individuare i migliori segmenti di clientela per poi poter indirizzare le campagne verso quei profili.

***Responder**:* Sono tutti coloro che si sono dimostrati interessati o hanno deciso di aderire alla promozione. Il processo mediante il quale si trasformano i clienti prospect in clienti responder dipende dal settore in cui opera l'azienda.
Quando un potenziale cliente, del quale non si hanno dati, accoglie l'offerta della campagna diventa un responder e quindi il cliente diventa individuabile, con nome e cognome, e può essere raggiunto o contattato nel tempo. In questa fase il cliente non ha ancora effettuato nessun acquisto ma i responder sono quelli che con maggiore probabilità lo faranno.
Il Data Mining viene utilizzato in questa fase per capire quali prospect diventeranno responder e tra i responder chi diventerà cliente effettivo. Lo strumento più usato è quello dei modelli previsionali.

Clienti effettivi: Coloro che utilizzano il prodotto o il servizio. Chi entra per la prima volta in questa categoria è un nuovo cliente e il suo comportamento iniziale è utile per fare previsioni sul suo comportamento futuro.
I responder diventano clienti effettivi quando il cliente ha effettuato il primo acquisto, cioè quando si istaura un rapporto di natura economica tra l'azienda e il cliente.

Il comportamento iniziale è un elemento fondamentale per poter prevedere il comportamento futuro; la tipologia di comportamento si evidenzia fin dai primi acquisti.

Eventi che caratterizzano il ciclo di vita del cliente effettivo:

1. L'azienda deve stimolare il cliente a fare un uso abituale del prodotto o servizio perché una delle principali fonte di entrate è proprio l'utilizzo abituale.
2. L'azienda deve promuovere campagne di cross-selling mirate a stimolare i clienti all'acquisto di altri prodotti o servizi cercando di allargare la gamma aziendale.

3. L’azienda deve promuovere campagne di up-selling mirate a stimolare i clienti all’acquisto di altri prodotti o servizi (di costo maggiore) cercando di acquisire una nicchia di mercato.

Gli eventi legati all’utilizzo forniscono informazioni utili per applicare tecniche di Data Mining per individuare i pattern di comportamento del cliente.

I clienti effettivi sono una fonte di dati molto importante e applicando le tecniche di Data Mining a questi dati è possibile prevedere quanti clienti risponderanno alle campagne di up-selling e cross-selling e con quale probabilità (scoring system). Per poter utilizzare questi dati però, va effettuata l’estrazione di determinate caratteristiche, perché i dati che ci forniscono il comportamento del cliente effettivo sono molto dettagliati e non tutte le informazioni servono allo scopo-obiettivo.

Ex clienti: Sono tutti i clienti che non usano più il prodotto o che hanno abbandonato il servizio, volontariamente o meno.
Ad un certo punto i clienti cessano di essere tali. Occorre distinguere gli abbandoni in:
- volontari. Tutti quei clienti effettivi che decidono di non acquistare più quel prodotto per vari motivi (il cliente si è trasferito in un’altra area geografica non servita dall’azienda, ha cambiato abitudini di vita, è passato alla concorrenza oppure non ha più la necessità di usare il prodotto o servizio).
- forzati. Questi si riferiscono a clienti che cessano di essere buoni clienti (per esempio perché non saldano i debiti)

Importante è usare il Data Mining per distinguere tra abbandoni volontari e non, per prevedere quali clienti appartengono ad una o all’altra tipologia, in modo da risparmiare le spese di una campagna fedeltà mirata magari a chi non ha intenzione di rimanere cliente dell’azienda.

Tuttavia, quando un cliente effettivo diventa ex cliente, non tutto è ancora perduto, per riconquistarne la fiducia vengono lanciate le campagne di recupero (Churn Analysis).

1.3 Segmentazione della clientela

I clienti di una azienda, siano essi effettivi o potenziali, non sono tutti uguali anzi, sono tutti diversi tra loro, hanno diversi bisogni, diverse aspettative e diverse preferenze. Una segmentazione efficace porta a rappresentare la varietà di uno specifico mercato, in cui opera l'azienda, in piccoli insiemi di clienti omogenei al loro interno ed eterogenei tra di loro.

1.3.1 Criteri per segmentare un mercato

Criterio SOCIO-DEMOGRAFICO ➔ uso le variabili anagrafiche (età, professione, classe di reddito, luogo di residenza)

Variabili COMPORTAMENTALI ➔ esprimono un aspetto del comportamento di acquisto. Rapporto del consumatore con l'acquisto del prodotto (non consumatore, consumatore potenziale ma non effettivo, ex consumatore, consumatore effettivo (al primo acquisto o stabile)).

BENEFIT SEGMENTATION ➔ segmentazione basata sui benefici attesi che il consumatore si aspetta. Collegata alle variabili comportamentali. Segmentazione su base strategica.

SEGMENTAZIONE PER STILI DI VITA ➔ raggruppa i consumatori a seconda degli stili di vita. Lascia da parte il comportamento del consumatore. Hanno un collegamento con le variabili socio-demografiche.

Gli ultimi due vengono definiti criteri di segmentazione a posteriori perché nascono da una ricerca fatta direttamente sui consumatori.
In determinati mercati si usano più criteri di segmentazione. In questo caso i segmenti diventano molto più circoscritti e la possibilità di arrivare a segmenti che contengono un solo cliente è elevata, (micronizzazione del mercato) e il prodotto o servizio offerto è estremamente personalizzato. Nel caso opposto, cioè nel caso di segmenti molto grandi, potremmo arrivare ad una macronizzazione del mercato

1.3.2 Requisiti per una buona segmentazione

I segmenti devono avere una serie di requisiti che li rendano utili per la strategia di marketing. Per questo motivo il fatto di applicare più criteri di segmentazione potrebbe non portare dei benefici ma bensì delle complicazioni.
I principali requisiti sono:

1. I segmenti devono essere omogenei al loro interno ed eterogenei tra di loro. Perciò un cliente, o potenziale cliente, lo posso attribuire ad un solo segmento. L'utilizzo di più criteri porta si al verificarsi dell'omogeneità interna ma può violare il requisito dell'eterogeneità tra i segmenti.
2. I segmenti devono essere misurabili altrimenti non si riesce a quantificarli e l'azienda non riuscirebbe a posizionarsi.
3. Ogni segmento deve essere significativo dal punto di vista economico. Un segmento piccolo può non risultare conveniente, soprattutto se composto da piccoli clienti.
4. I segmenti devono essere accessibili rispetto alle azioni che permettono all'impresa di posizionarsi, rispetto alle risorse e rispetto alle competenze dell'impresa. Un segmento sovraffollato è limitamente accessibile.

5. I segmenti devono essere tendenzialmente stabili nel tempo. Fattori di cambiamento possono essere legati alle strategie dell'impresa oppure all'ambiente esterno che l'impresa non riesce a dominare ma con il quale può integrarsi.

Va comunque detto che possono esistere motivazioni strategiche che permettono al segmento di non avere determinati requisiti o di non avere requisiti.
Importante infine è monitorare nel tempo la segmentazione perché cambiamenti che interessano il mercato o l'impresa stessa potrebbero comportare cambiamenti più o meno importanti nei segmenti e nei loro requisiti. Per esempio un'innovazione tecnologica potrebbe rendere profittevole un segmento che prima non lo era.

1.3.3 Strategie di segmentazione

Una volta segmentato il mercato bisogna scegliere la strategia da adottare:

- Adattamento ➔ Mi adatto e decido il numero di segmenti in cui operare.
- Modifica ➔ Cerco di modificare la situazione per renderla più favorevole alle mie caratteristiche. È rischiosa perché modifico il comportamento del consumatore. La strategia di modifica del mercato segmentato può avvenire in diversi modi:

1. Aumentando il livello di segmentazione, ovvero creando un nuovo segmento in aggiunta a quelli che già esistono. I clienti, o potenziali, che fanno parte di questo segmento provengono da segmenti che già esistevano, perciò viene modificata la loro funzione di domanda.

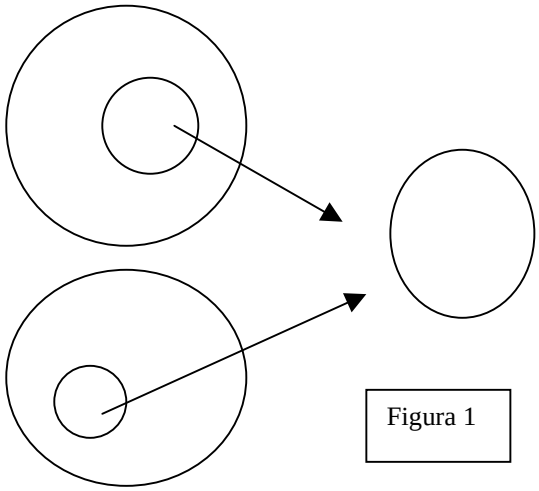


Figura 1

2. Si crea un nuovo segmento eliminandone uno che già esisteva. Non viene modificato il livello di segmentazione.

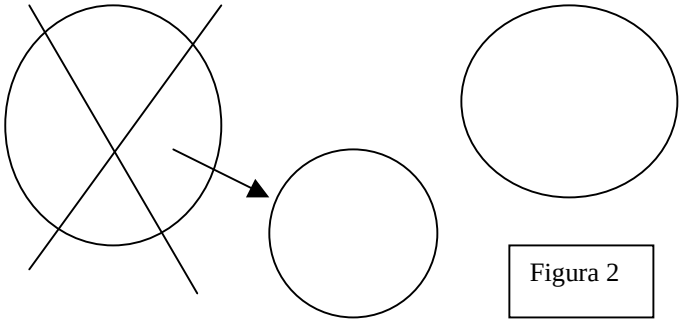


Figura 2

3. Si crea un nuovo segmento dalla fusione di più segmenti. In questo caso ho una riduzione del livello di segmentazione.

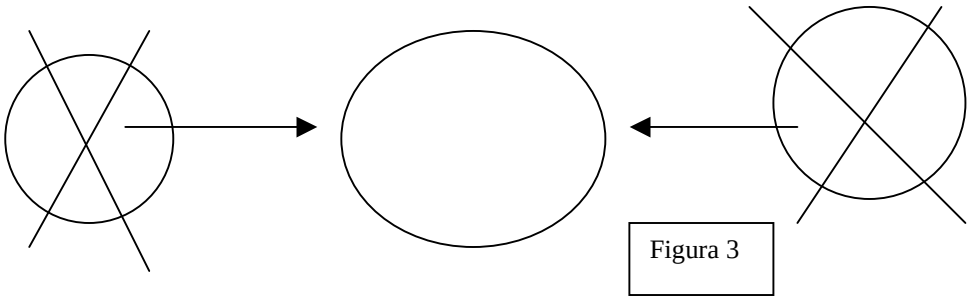


Figura 3

4. Si può frammentare un segmento in più segmenti. I clienti o potenziali si distribuiscono tra i nuovi segmenti.

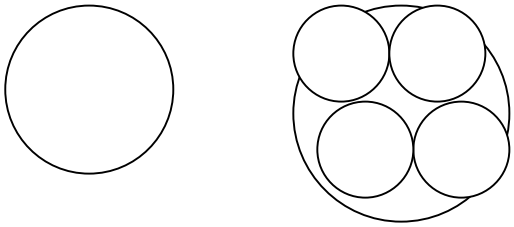


Figura 4

5. Si può far passare una parte dei clienti, o potenziali, da un segmento all'altro senza avere una variazione del livello di segmentazione.

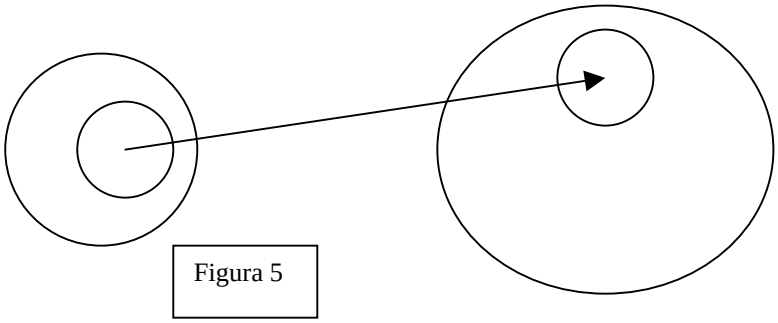


Figura 5

6. Si può spostare un intero segmento in un altro e il livello di segmentazione diminuisce.

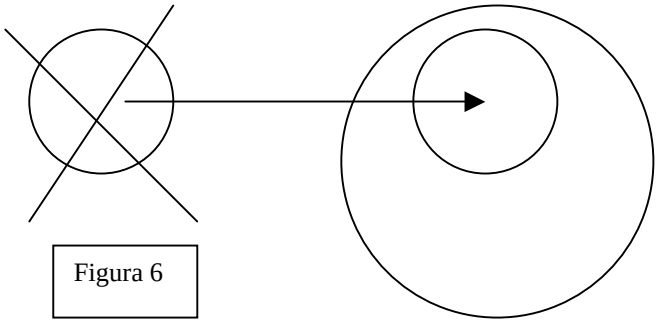


Figura 6

Una volta identificati i segmenti che formano il mercato obiettivo e scelta la strategia di fondo (adattamento o modifica), l'azienda dovrà scegliere una *copertura*, cioè dovrà decidere in quali e in quanti segmenti operare.

Se l'azienda decide di adattarsi al mercato allora dovrà decidere se coprire l'intero mercato con offerte di marketing differenziate per ogni segmento oppure focalizzare la propria attenzione in un determinato segmento.

Se l'azienda decide invece di creare un nuovo segmento per poi focalizzare la propria attenzione su esso (dunque decide di modificare il mercato) allora dovrà adottare una serie di strategie che le permettano di farlo.

Il *posizionamento* è la fase ultima del processo di segmentazione del mercato. Posizionare un prodotto o un servizio significa sviluppare un determinato marketing mix per un determinato segmento, perciò significa definire il prodotto o il servizio, definire una strategia di comunicazione, definire una strategia di distribuzione e definire il prezzo per ogni segmento che si intende coprire. Molta importanza ha la strategia di comunicazione perché attraverso un'appropriata comunicazione l'azienda comunica al mercato il posizionamento che vuole ottenere.

Lo sviluppo di una strategia di marketing che si rivolge ad un determinato segmento, dunque ad un determinato target di consumatori, richiede la raccolta e l'analisi dei dati relativi ai potenziali clienti ma anche all'ambiente esterno, per poter identificare le opportunità e le minacce.

Le ricerche di mercato consentono di raccogliere dati relativi ai consumatori, ai concorrenti, alle condizioni economiche globali e così via, e permettono di individuare nuovi mercati per prodotti o servizi già esistenti, nuovi segmenti e di anticipare le mosse dei concorrenti.

Le ricerche di mercato servono per capire come segmentare il mercato e quali opportunità o minacce esistono all'interno di ogni segmento. Molte altre sono le finalità però: previsione delle vendite, determinazione del prezzo e così via….

Per arrivare a questo bisogna raccogliere i dati utili alla ricerca di mercato. I dati possono essere primari, cioè informazioni che devono essere raccolte per la prima volta, o secondari ovvero informazioni già esistenti (*). L'analisi dei dati secondari ha la precedenza su quella dei dati primari solo ed esclusivamente per un motivo economico.

I dati secondari possono essere dati interni all'azienda (vecchi piani di marketing, contabilità generale….) oppure esterni (raccolti da terzi e resi pubblici). Con la diffusione di internet sono molto facili da reperire e quindi molto usati dalle aziende.

A cosa servono tutti questi dati?

1. Per stabilire chi saranno i potenziali clienti, cioè tutte quelle persone o organizzazioni che hanno bisogno del prodotto o del servizio e che possono acquistarlo.
2. Per cercare di prevedere le condizioni di mercato che si potranno verificare

---

Figure 1,2,3,4,5,6 : Lucidi del corso Economia e Gestione Imprese II

* : Lucidi del corso Analisi di mercato I

2. Metodologia e strumenti

2.1 Data Mining

Il Data Mining è un processo analitico di scoperta di relazioni, di pattern e di informazioni precedentemente sconosciute e potenzialmente utili presenti all'interno di grandi database.

Un pattern indica una struttura, un modello, una rappresentazione sintetica dei dati.

Il risultato di tale processo è una quantità molto piccola ma molto preziosa di informazione che non deve esaurirsi in una volta sola perché il processo è costoso. Per questo motivo l'informazione creata viene messa in circolo con tutti gli altri dati per poter essere utilizzata più volte, anche assieme ai dati stessi per creare altra informazione.

L'informazione ottenuta può essere tramutabile in azioni commerciali allo scopo di ottenere un vantaggio di business e aumentare la profittabilità. Infatti la diversificazione e la globalizzazione hanno aumentato il livello di competitività e ottenere un vantaggio sulla concorrenza è diventato fondamentale per la sopravvivenza di un'azienda.

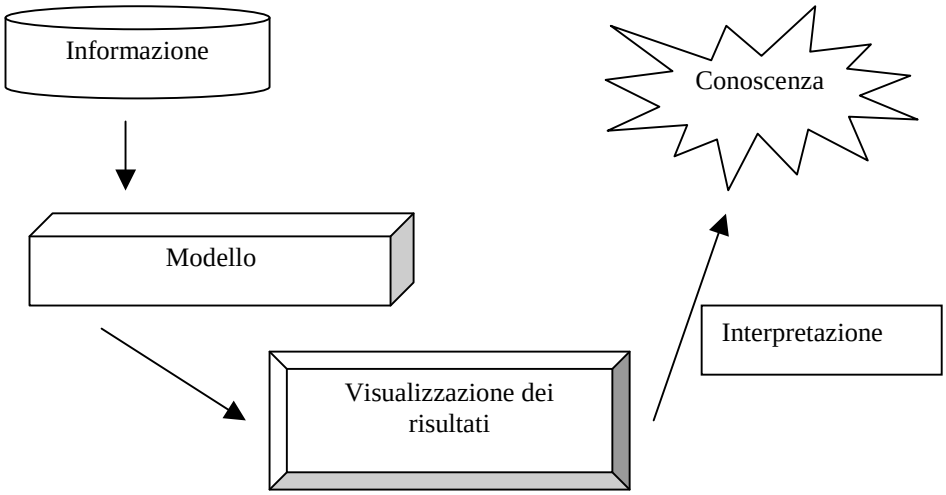


Figura: Lucidi corso Sistemi informativi Aziendali

Il Data Mining serve per :

1. **Classificare:** si assegna una nuova variabile ad ogni cliente che identifica l'appartenenza ad una determinata classe. Si applica un modello a dei dati grezzi che verranno poi classificati. Si può così dividere la propria clientela e dedicare gli sforzi di una campagna di Marketing ad una determinata classe. Per esempio si possono dividere i clienti secondo il reddito (basso, medio, alto). Esiste comunque un numero di classi già note e l'obiettivo è quello di inserire ogni record (cliente) in una determinata classe.
2. **Stimare:** mentre la classificazione usa valori discreti, la stima usa valori continui. In base ad un determinato input, usiamo la stima per individuare il valore di una variabile continua sconosciuta (il reddito per esempio).

Spesso stima e classificazione vengono utilizzate insieme, per esempio posso fare prima una stima che mi calcola un punteggio o una probabilità da assegnare ad ogni cliente e successivamente posso classificare i clienti in base alla variabile stimata.

3. **Fare previsioni:** può essere considerata una classificazione o una stima perché i clienti sono classificati in base ad un

comportamento futuro prevedibile o stimato. Molta importanza hanno i dati storici perché servono per costruire un modello che spieghi il comportamento futuro in base a quello passato.

4. **Raggruppare per affinità o regole di associazione:** l'obiettivo è di stabilire quali oggetti (in genere prodotti) possono abbinarsi. Si può utilizzare il raggruppamento per affinità per pianificare la produzione dei prodotti sugli scafali o nei cataloghi in modo che gli articoli, che vengono acquistati insieme, si trovino il più possibile vicini.
5. **Clustering:** significa segmentare un gruppo di clienti eterogenei in gruppi omogenei. È diverso dalla classificazione perché non si usano classi predefinite, infatti i record sono raggruppati in base alle analogie, è il ricercatore che deve stabilire il significato da dare ad ogni cluster.
6. **Descrizione e visualizzazione:** una descrizione efficace di uno specifico comportamento indica da dove partire per cercare una spiegazione. La visualizzazione dei dati è una forma molto efficace di Data Mining descrittivo, è molto più immediato ricavare utili informazioni da dati visivi.

Le prime tre tecniche sono esempi di *Data Mining diretto*: lo scopo è di usare i dati disponibili per creare un modello che dia come output una specifica variabili obiettivo.
Le altre tre tecniche sono esempi di *Data Mining indiretto*: non ho più una variabile target ma l'obiettivo è quello di stabilire una precisa relazione tra tutte le variabili.

Il Data Mining è molto usato nel settore marketing vista la presenza di grosse quantità di dati da elaborare per ricavarne informazioni utili. Questi dati sono tutti raccolti in un database marketing e si riferiscono a tutti i potenziali clienti (prospect) di una campagna di mercato. Questi dati possono descrivere il comportamento del cliente

già acquisito o possono contenere una serie di informazioni grezze di tipo demografico sui possibili clienti.
Il Data Mining permette all'azienda di ridurre le spese non contattando la clientela che difficilmente risponderà all'offerta.

2.1.1 Fasi del processo di Data Mining

1. Determinazione del problema di business: Il primo passo del processo di Data Mining consiste nel definire l'obiettivo dell'analisi.
2. Selezione ed organizzazione dei dati: Una volta determinato l'obiettivo di business bisogna raccogliere e selezionare i dati necessari per l'analisi.
3. Analisi esplorativa dei dati: L'analisi statistica vera e propria inizia con l'analisi preliminare dei dati che consiste in una prima valutazione delle variabili statistiche per una eventuale eliminazione o trasformazione.

   Tramite un'adeguata analisi delle variabili statistiche è possibile individuare la presenza di valori anomali che non vanno necessariamente eliminati perché possono contenere informazioni utili al raggiungimento degli obiettivi del Data Mining.

   L'analisi esplorativa può essere utile anche per individuare quali variabili sono tra loro correlate in modo da eliminarne una, visto che le informazioni contenute sono equivalenti.

4. Data Mining: Per prima cosa bisogna scegliere la tecnica da usare che dipende dall'obiettivo dell'analisi (punto 1) e dai dati disponibili. Utilizzo poi la tecnica decisa per arrivare all'informazione. Spesso non si usa una tecnica sola per poi valutarle e scegliere quella migliore.

5. Interpretazione dei modelli identificati: Analisi e verifica dei risultati con possibile retroazione ai punti precedenti per ulteriori iterazioni al fine di migliorare l'efficacia dei modelli trovati.
6. Consolidamento della conoscenza scoperta: Integrazione della conoscenza e valutazione del sistema mettendo a confronto i risultati con l'effettivo andamento della realtà e produzione della documentazione agli utenti finali o a terze parti interessate.

2.1.2 Tecniche di Data Mining e applicazioni

Tecniche

- Tecniche di visualizzazione dei dati: mediante grafici multidimensionali è possibile identificare relazioni complesse per scoprire l'informazione nascosta.
- Clustering: tecniche di classificazione che divide una popolazione in sottogruppi. Le tecniche di clustering e le reti neurali non supervisionate consentono di effettuare operazioni di segmentazione della clientela.
- Reti neurali: per risolvere problemi di classificazione e di previsione. Utilizzate in un ambiente dinamico, dove i dati cambiano sempre, vista la loro capacità di generalizzare la conoscenza appresa.
- Alberi decisionali: rappresentazioni grafiche costruite suddividendo i dati in sottogruppi omogenei. La rappresentazione avviene in modo gerarchico ad albero.

Le reti neurali supervisionate e gli alberi di classificazione, fanno uso della conoscenza acquisita per classificare nuovi oggetti o prevedere nuovi eventi.

- Individuazione di associazioni: tecniche esplorative per misurare l'affinità dei prodotti. Individuano gruppi di prodotti legati da analoghe abitudini di acquisto.

Applicazioni

- Scoring system: È un particolare approccio di analisi incentrato sull'assegnazione ai singoli clienti (prospect) della probabilità di adesione ad una campagna commerciale. La finalità è quella di classificare i clienti o gli eventuali prospect in modo tale da attuare azioni di marketing diversificate a seconda dei target individuati.
  L'obiettivo è quello di costruire un modello predittivo in modo da individuare una relazione tra una serie di variabili comportamentali e una variabile obiettivo che rappresenta l'oggetto di indagine. Il modello produce come risultato un punteggio (score) che indica la probabilità di risposta positiva alla campagna (il cliente aderisce o non aderisce alla campagna promozionale).
- Segmentazione della clientela: applicazione di tecniche di clustering per individuare gruppi omogenei calcolati secondo variabili comportamentali o socio-demografiche. L'individuazione delle diverse tipologie permette di effettuare campagne di marketing mirate.
- Market basket analysis: applicazione di tecniche di associazioni a dati di vendita per individuare quali prodotti vengono acquistati insieme. Utile per la disposizione dei prodotti sugli scafali.

2.2 I dati

Il Data Mining che lavora sul cliente, richiede che a ogni riga (record) corrisponda un singolo cliente, che viene inteso come l'unità di

azione e che può fornire utili informazioni per comprendere meglio i pattern.

I dati sono quindi strutturati in una serie di righe e colonne

Righe

Spesso i clienti non sono identificabili. Gli utenti di un sito web per esempio, non hanno nome e cognome ma solo un indirizzo IP che corrisponde ad un cliente. Dalla stessa macchina però può collegarsi un altro utente diverso dal primo oppure può succedere che il primo utente utilizzi un altro computer.

Non sempre per tutti i clienti sono disponibili tutti i dati, quindi i dati effettivi utilizzabili sono quelli che si riferiscono ad un sottogruppo di clienti con valori validi solo per alcuni campi, per esempio se devo svolgere una campagna via e-mail, le righe che prenderò in esame saranno solo quelle che identificano clienti con indirizzi di posta elettronica validi.

Colonne

Le colonne, ovvero le variabili statistiche, rappresentano i dati relativi a ciascun record. Il range della colonna è l'insieme dei valori che quella colonna può assumere.

Colonne unarie:

È possibile che una colonna abbia un unico valore (colonne unarie). Non offrono nessuna informazione utile che ci permetta di distinguere un record dall'altro perciò una variabile di questo tipo viene ignorata dall'analisi di Data Mining.

La formazione di colonne unarie può essere il risultato di un'analisi mirata ad uno specifico sottogruppo di clienti, il campo che definisce questo sottogruppo presenta lo stesso valore per tutti i clienti. Al momento di costruire il modello questo campo verrà però ignorato.

Colonne quasi unarie:

In questo tipo di colonne quasi tutti i record presentano lo stesso valore per quella specifica variabile, quindi i valori diversi sono molto pochi.
Questo tipo di variabili statistiche possono essere ignorate solo quando una parte trascurabile dei record assume valori diversi, cioè quando un gruppo di clienti non significativo si differenzia dalla massa. Normalmente una colonna che ha il 95-99% dei valori identici può essere ignorata, ma prima di prendere una decisione così importante bisogna capire perché certi valori di quella colonna sono così rari. Per esempio il numero di persone che ogni giorno decidono di abbonarsi ad una rivista può essere basso, ma con il passare del tempo questo numero potrebbero aumentare; bisogna quindi sommare gli abbonamenti per un lungo periodo di tempo.

Colonne con valori unici:
Una colonna può anche assumere valori diversi per ogni record e quindi venere utilizzata per identificare esattamente le righe (nominativi dei clienti, indirizzi, numeri telefonici). Anche queste colonne sono povere di informazione utile perché identificano in modo univoco ogni cliente e quindi non hanno alcun valore previsionale.
Queste colonne però possono essere utilizzate per calcolare le variabili derivate, per esempio i numeri di telefono e gli indirizzi mi possono dare informazioni sulla distribuzione geografica. Una volta estratte le informazioni utili le colonne originarie vanno ignorate.

Quando una colonna è troppo correlata alla colonna target può significare che questa è un equivalente. È possibile che delle campagne esplorative siano state focalizzate su un determinato segmento, per esempio clienti con età inferiore a 40 anni e con figli. Tutti i responder avranno queste caratteristiche, ma l'età e il fatto di

avere figli sono variabili che non possono essere prese in considerazione per ulteriori analisi. Per fare del Data Mining è importante sapere chi ha accettato l'offerta, ma importante è anche sapere chi è stato contattato.

2.2.1 Ruolo e proprietà delle variabili statistiche

Ruolo

- Colonne di input ➔ impiegate come input nel modello.
- Colonne target ➔ usate solo nei modelli previsionali, rappresentano le informazioni interessanti (propensione all'acquisto di un determinato prodotto).
  Per i modelli descrittivi le colonne target non servono.
- Colonne ignorate ➔ quelle che non vengono utilizzate. Queste hanno un ruolo importante nel clustering, non vengono utilizzate per la costruzione dei cluster ma la loro distribuzione all'interno di essi può dare dettagli importanti o interessanti sui clienti.

I tre ruoli elencati sopra sono fondamentali ma esistono anche ruoli avanzati. Di seguito sono elencati alcuni esempi da SAS Enterprise Miner:

- Colonne di identificazione ➔ identificano univocamente i record, in genere vengono tralasciate per scopi di analisi.
- Colonne di peso ➔ stabiliscono il peso da assegnare ad ogni record, per esempio per creare un campione pesato (un cliente può valere di più di un altro cliente).
- Colonne di costo ➔ specifica il costo associato ad una riga. Posso attribuire così ad ogni cliente un costo.

Proprietà

L'ordine è la principale proprietà di misura. Sono dette *categoriche*, o nominali, quelle colonne alle quali non si può attribuire un ordine. Per esempio i colori non seguono nessun ordine, non si può dire che il rosso viene prima del verde.

Le colonne ordinate si distinguono invece in:

- *Ranghi* ➔ hanno un ordine ma non consentono calcoli aritmetici (basso, alto, medio).
- *Intervalli* ➔ hanno un ordine, consentono la sottrazione ma non necessariamente la somma (per esempio per una data, ha senso chiedere quanti giorni intercorrono tra questa e un'altra, ma non ha senso raddoppiarla).
- *Valori numerici* ➔ hanno un ordine e consentono qualsiasi calcolo aritmetico.

I dati possono essere stringhe o numeri ma questo non significa che il primo tipo è categorico e il secondo no. Possiamo trovare codici che contengono cifre ma che non possono essere ordinati, mentre possiamo trovare codici che contengono stringe e che hanno un determinato ordine.

Le *categorie* sono un insieme di valori, in genere delle etichette che non hanno un ordinamento proprio (gli stati). Certe volte le categorie sono rappresentate da numeri che non seguono un ordine preciso. In questo caso bisogna ridefinire il tipo in modo che i numeri rappresentino delle categorie, in quanto gli strumenti di Data Mining trattano tutti i numeri come valori numerici.

I *ranghi* sono simili alle categorie, con la sola differenza che hanno un ordine. Spesso sono il risultato di una suddivisione in fasce (**binning**). Se la mia popolazione ha una distribuzione non omogenea di valori, con il binning posso andare alla ricerca di bin (fasce) che contengono lo stesso numero di valori, l'ampiezza può

variare ma il numero dei valori è più o meno sempre lo stesso. Questo può essere un modo per gestire i valori anomali.
Posso avere anche dei bin di ampiezza uguale. Questo non elimina il problema dei valori anomali anzi, potrebbe peggiorare la situazione.
Per *intervallo* si intende la differenza tra due valori. Per esempio le date e gli orari, quanti giorni mancano al 17 agosto se oggi siamo il 2 agosto? 15! Non ha senso fare altre operazioni.
Tra i *valori numerici* troviamo tutti quei dati che consentono qualsiasi tipo di operazioni (somma totale di pezzi acquistati).

2.2.2 Variabili derivate

Le variabili derivate sono colonne non presenti in origine ma ricavate da altre variabili.
Spesso esistono più variabili derivate che hanno lo stesso contenuto informativo e che identificano lo stesso pattern. Se ho una variabile derivata che indica il numero totale di chiamate urbane e interurbane, una che indica il tempo trascorso in chiamate urbane e una che indica il tempo trascorso in chiamate interurbane e tutte e tre sono uguali a zero, allora l'informazione che ottengo è sempre la stessa, ovvero non è stata effettuata nessuna chiamata.

Outlier: (Valori rari)

Cosa fare?

- *Non fare nulla:* alcuni algoritmi (alberi decisionali) non sono sensibili alla presenza degli outlier perché usano il rango delle variabili numeriche. Altri algoritmi (reti neurali) sono molto

sensibili e basta la presenza di pochi valori rari per comprometterne il funzionamento.

- *Filtrare le righe che li contengono:* potrebbe portare ad una distorsione nei dati, ma è vero anche che potrebbe essere una buona idea per non considerare gli acquisti dei non clienti: se in un supermercato ignoro gli acquisti che si discostano di molto dalla media significa che prendo in considerazione solo clienti abituali, cioè che fanno acquisti in media.
- *Ignorare le colonne:* soluzione estrema. La colonna può essere sostituita da informazioni relative alla colonna.
- *Sostituire gli outlier:* il più comune. Può essere "null" per algoritmi che gestiscono bene questo valore, ma può essere anche zero, la media, il massimo o minimo…..
- *Binning:* usando intervalli di peso uguale. Per esempio posso dividere i salari in alto, basso e medio.

Combinazioni di colonne:

Posso calcolare il contenuto di una variabile derivata prendendo i dati già presenti in una riga e calcolando nuovi valori:

- Indice di obesità = statura^2/peso
- N° pagine web visitate/n° acquisti

L'obbiettivo è comunque quello di trovare nuove informazioni utili da poter integrare con i dati già a disposizione.

Sommarizzazioni:

Variabili derivate che sono il risultato di una ricerca di informazione sulle dimensioni di un record.

Se voglio calcolare la redditività media per ogni zona identificata dal CAP devo aggregare i dati con lo stesso CAP e poi aggiungo la nuova informazione, la redditività media, usando come chiave il CAP.

Estrarre elementi utili da singole colonne:

Dalle date posso estrarre molte informazioni utili che possono spiegare il comportamento del cliente:

- giorni della settimana
- giorno del mese
- mese dell'anno
- giorno dell'anno
- trimestre dell'anno
- festività, non festività
- lavorativo, non lavorativo

Tutte queste possono essere delle nuove colonne che si aggiungono ai dati già disponibili. Queste nuove variabili sono utilissime per le sommarizzazioni, per esempio posso sapere il comportamento dei clienti nei vari giorni lavorativi o come si comportano nei giorni di festa e così via….

Anche gli orari possono essere utili per ricavare alcune informazioni aggiuntive, posso dividere la giornata in mattino, pomeriggio, sera, notte e vedere l'andamento delle vendite durante questi orari.

Sono comunque molti i tipi di dati dai quali posso estrarre nuove informazioni:

- numeri telefonici ➔ mi danno informazioni sulla distribuzione geografica, il prefisso indica la provincia.
- Indirizzi ➔ dal numero civico posso capire se il residente abita in una casa singola.
- ……..

Estrarre queste informazioni è molto importante ai fini del Data Mining.

Serie temporali:

Le serie temporali rappresentano i dati che si ripresentano più volte a precisi intervalli di tempo.
Per poter utilizzare questi dati in modo migliore è necessario normalizzarli all'ultima data disponibile. Se l'oggetto di studio è l'abbandono, avrò numerosi clienti che lasciano in momenti diversi. Per poter costruire un modello che descriva questi clienti è necessario riallineare i dati rispetto alla data di abbandono, prendendo in considerazione il mese finale di ciascun cliente eliminando però la stagionalità ed altre informazioni che sono comunque recuperabili tramite l'aggiunta di variabili derivate.
Un esempio utile di serie temporali sono i dati relativi all'uso di telefoni cellulari, sono serie temporali perché i dati vengono raccolti e analizzati mensilmente. Possiamo distinguere anche diversi tipi di clienti:

- clienti stabili: il loro profilo è ogni mese lo stesso;
- clienti in crescita: l'uso del telefono cresce in maniera costante;
- ricevitore: il loro profilo presenta solo chiamate ricevute;
- mittenti: il loro profilo presenta solo chiamate in uscita.

Riepilogo

I dati devono avere il seguente formato per essere utilizzabili ai fini del Data Mining:

1. tutti i dati devono stare in una sola tabella o vista;
2. ogni riga deve corrispondere a un'istanza di dato rilevante a fini pratici;
3. le colonne unarie vanno ignorate;
4. le colonne con valori unici vanno ignorate, anche se possono essere utili per calcolare colonne derivate;

5. per i modelli previsionali, una volta identificata la variabile target, tutte le colonne equivalenti vanno ignorate.

2.2.3 Come si presentano i dati e da dove provengono

Nel mondo reale i dati non sono mai pronti per essere utilizzati dal Data Mining, quindi una volta raccolti bisogna trasformarli nel formato adatto agli algoritmi che si vogliono usare.
I dati per il Data Mining devono essere importati da altri sistemi (possono essere immagazzinati in database relazionali, log file, ecc…) e tutti o quasi i sistemi operazionali possono esportare i dati.

Sistemi operazionali:
I sistemi operazionali sono tutti quei sistemi usati per far funzionare l'azienda (*):

- bancomat
- web server e database per e-commerce
- sistemi di fatturazione
- …….

I sistemi operazionali sono una fonte ricchissima di dati, dati che vengono raccolti direttamente dal punto di contatto con il cliente. Non tutti i S.O. però sono in grado di raccogliere i dati e quindi l'azienda dovrà rivolgersi ad altre fonti (sondaggi, profili di mercato e intuizioni) per avere a disposizione qualcosa su cui fare del Data Mining, aumentando la spesa e ottenendo dati incompleti.

I dati memorizzati sui sistemi operazionali non sono immediatamente accessibili, perché ci sono attività (per esempio la fatturazione) che hanno la precedenza sulle attività di business intelligence.
I dati inoltre sono sempre sporchi.
Tutti i dati che un'azienda possiede sono immagazzinati in un datawarehouse e si trovano tutti in un solo posto, pronti per essere utilizzati.
I datawarehouse sono database relazionali che presentano centinaia di tabelle descritte da migliaia di campi, I dati vengono inseriti nel sistema, puliti e verificati. È possibile fare del Data Mining anche senza i datawarehouse, anche se quest'ultimi sono una fonte utilissima di dati.

---

* : Berry & Linoff, Data Mining

Spesso i clienti sono invitati a fornire informazioni personali su di loro, questo accade per i sondaggi e le inchieste. Tutti questi dati però devono essere trattati con molta cautela perché:

- La gente se può non risponde alle domande. C'è però un gruppo ristretto di persone che lo fanno, questi sono la minoranza e quindi non rappresentano tutta la popolazione.
- Le risposte possono essere non del tutto corrette, o per errori di battitura o per volontà delle persone stesse.
- Le inchieste condotte nel passato potrebbero non essere confrontabili con quelle più recenti, perché la popolazione di riferimento potrebbe cambiare con il passare del tempo.
- I dati raccolti spesso sono incompleti, perciò non sono utilizzabili come input per i modelli.

Nonostante tutto però i sondaggi e le inchieste sono molto utili per avere maggiori informazioni sui clienti. I risultati possono servire per trovare un nuovo approccio alla commercializzazione di un prodotto, o per ricavare un nuovo tema per una campagna pubblicitaria e così via…..

Quanti dati?

Più dati ci sono e meglio è……

La quantità di dati disponibili dipende dal rapporto tra azienda-cliente.

I prospect offrono la minore disponibilità e spesso il loro elenco viene acquistato da terzi. Le campagne pubblicitarie vengono mirate secondo una divisione demografica e comunque non si sa nulla di chi riceve il messaggio pubblicitario finché il prospect non decide di rispondere alla campagna. Spesso quando un cliente potenziale chiede informazioni relative ad un prodotto o ad un servizio, lascia delle tracce che sono informazioni utili e importanti ma che spesso sono dati imprecisi e incompleti. I dati generati dal comportamento dei clienti effettivi contengono informazioni più precise e informazioni riguardanti i segmenti a cui appartengono.

Importante infine registrare i gruppi esposti alle diverse campagne per poter dividere chi ha effettivamente risposto da chi non ha risposto.

Spesso però i dati a disposizione per costruire il modello sono troppi, per esempio quando si usa solo la visualizzazione, perciò è fondamentale usare la campionatura per ridurre il numero dei dati.

Se invece usiamo le reti neurali c'è un limite inferiore di dati da soddisfare.

Se trattiamo dati relativi ai prospect allora il numero di righe potrebbe essere molto elevato visto che i potenziali clienti sono tanti, mentre le

colonne potrebbero non essere molte visto che poche sono le informazioni che abbiamo su di loro. Se invece trattiamo dati sulla clientela esistente il numero di righe (di clienti) diminuiscono, perché non riusciremo mai a raggiungere tutti i potenziali clienti vista la loro normale diversità, ma le colonne aumenteranno perché abbiamo più dati a disposizione visto che ora conosciamo la nostra clientela.

Campionatura

La campionatura è il processo mediante il quale partendo dal set di dati originario arrivo ad un insieme di dati più ristretto. Nel caso dei modelli previsionali il campione deve essere rappresentativo, cioè deve avere caratteristiche simili rispetto alla popolazione di riferimento. Per verificare se lo è effettivamente basta controllare la distribuzione del campione e metterla a confronto con quella della popolazione, se le due distribuzioni sono simili allora il campione è rappresentativo.

Se devo trattare dati categorici allora un campione rappresentativo è formato dai dati più comuni nella popolazione, mentre se tratto dati numerici, la media e la deviazione standard devono essere simili. Se questo non accade allora bisogna generare un altro campione.

Se nella popolazione ho dei dati rari allora devo applicare la *sovracampionatura.*

La sovracampionatura è il processo mediante il quale costruisco un insieme con una quantità maggiore di dati rari rispetto a quelli comuni.

Dati sporchi

I dati sono sempre sporchi!!!!!! (*Berry & Linoff, Data Mining*)

Dati mancanti:

I dati mancanti esistono per quattro motivi:

1. valori vuoti: forniscono informazioni rilevanti. Se un cliente non mette il suo numero telefonico può voler dire che non vuole essere disturbato.
2. valori inesistenti: un cliente può non avere un indirizzo e-mail, in questo caso se devo costruire un modello per avere una previsione sull'utilizzo futuro della posta elettronica non posso considerare questi clienti.
3. dati incompleti.
4. dati non raccolti.

Come posso trattarli?

- Non fare nulla: se ce ne sono pochi il modello potrebbe non risentirne.
- Eliminare le righe che li contengono: potrei avere una distorsione nei dati cioè potrei eliminare i clienti che rappresentano molto bene la popolazione di riferimento.
- Ignorare le colonne: applicabile se gli spazi vuoti sono pochi o in caso sostituirli con dei flag per indicare se i dati erano presenti o no.
- Prevedere nuovi valori: posso sostituire i dati mancanti con il valore medio o con quello più frequente.
- Costruire un altro modello: posso segmentare i clienti in base ai dati disponibili.

Valori errati:

Sono tutti quei valori che non risultano validi per la colonna o che sono per qualche ragione incomprensibili.

Si hanno dati errati perché questi vengono raccolti in modo non corretto, o perché vengono immessi in modo sbagliato (errori di inserimento).

2.3 Cluster Analysis

La Cluster Analysis è una tecnica utilizzata per segmentare la clientela a partire dai dati anagrafici ma anche dalle informazioni sull'utilizzo del prodotto o dalle risposte di questionari e permette di arrivare all'individuazione dei segmenti.

I segmenti vengono anche chiamati cluster o gruppi; in pratica un insieme di clienti viene diviso in sottogruppi omogenei al loro interno e diversi tra di loro, rispetto a delle variabili date in input.

Ogni cliente i-esimo ha p variabili (x_{i1},….., x_{ip}) che possono essere qualitative e quantitative. Per stabilire la diversità tra i gruppi vengono calcolate delle distanze.

2.3.1 Distanza tra due elementi

Per calcolare la distanza tra due elementi h e k, d(h,k) è necessario calcolare la distanza per ciascuna delle p variabili considerate ovvero $d_j(x_{hj} , x_{kj})$ con j=1,….,p.

Nel caso di variabili quantitative:

$d_j(x_{hj},x_{kj}) = (x_{hj} - x_{kj})^2$ è il quadrato della distanza euclidea. [1]

Nel caso di variabili qualitative:

$d_j(x_{hj},x_{kj}) = 1 – I(x_{hj}= x_{kj})$ [2]

dove $I(x_{hj}= x_{kj})$ vale 1 se le due variabili sono uguali e 0 altrimenti. Se le variabili sono qualitative ordinali allora viene assegnato loro un punteggio 1,2,…m e vengono considerate come variabili quantitative.

Una volta scelta la funzione che calcola la distanza per ogni variabile per calcolare la distanza tra due elementi posso sommare le p funzioni:

$$d_{hk} = \sum_{j=1…p} (X_{hj} – X_{kj})^2 \quad [3]$$

se le variabili utilizzate sono in parte quantitative, in parte qualitative e in parte qualitative ordinali la distanza d_{hk} va calcolata separatamente per i tre tipi di variabili e poi combinate insieme con la possibilità di scegliere tre pesi per dare valore diverso a variabili di tipo diverso.

$$d_{hk} = [w_1 d_1 + w_2 d_2 + w_3 d_3] / (w_1+ w_2+ w_3) \quad [4]$$

dove d_1 è la distanza complessiva tra due elementi per le variabili quantitative;
d_2 è la distanza complessiva tra due elementi per le variabili qualitative;
d_3 è la distanza complessiva tra due elementi per le variabili qualitative ordinali.
Le distanze vengono poi messe in una matrice di dissimilarità D_{nxn}, con diagonale nulla ed elementi positivi. Questa matrice è la base di partenza per molti algoritmi di classificazione.

---

[1] [2][3][4] : Azzalini & Scarpa, Analisi dei dati e Data Mining

2.3.3 Algoritmi di classificazione

Esistono due tipi di algoritmi di classificazione:

1. **Gerarchici:** ogni gruppo fa parte di un gruppo più grande che è a sua volta contenuto in un gruppo ancora più grande, fino ad ottenere un gruppo unico che contiene tutta la popolazione.
2. **Non gerarchici:** negli algoritmi non gerarchici bisogna avere il numero dei cluster che si vuole ottenere. L'algoritmo, in base a questo numero, cerca di ottenere la migliore classificazione possibile.

Gli algoritmi gerarchici si dividono in:

- *Scissori:* parte da un cluster contenente tutta la popolazione a va a scindere fino ad ottenere cluster con un solo elemento.

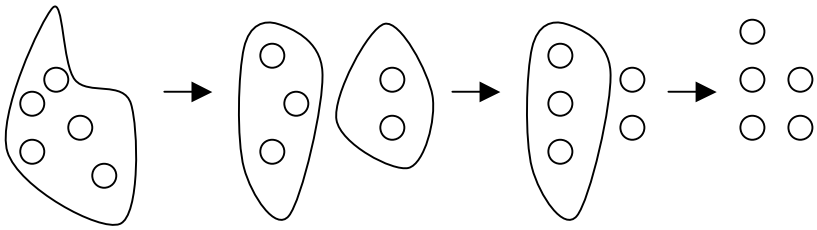


Figura: lucidi corso Sistemi Informativi Aziendali

- *Aggregativi:* il procedimento è inverso, parta da una situazione con cluster che contengono un solo elemento e li aggrega fino ad arrivare ad un unico cluster contenente tutti gli elementi.

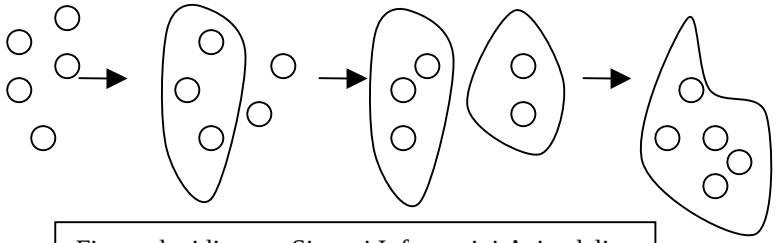


Figura: lucidi corso Sistemi Informativi Aziendali

Gli algoritmi non gerarchici invece si dividono in:

- *A partizioni:* ogni elemento va dentro ad un cluster in modo univoco.
- *A classi sovrapposte:* ogni elemento va dentro ad un cluster con un certo grado di appartenenza.

Gli algoritmi gerarchici, e non, offrono vantaggi e svantaggi:

- i primi non usano il numero dei gruppi a priori ma sono onerosi dal punto di vista computazionale perché vanno a calcolare la matrice delle distanze ad ogni iterazione e possono essere poco efficienti se ho a disposizione tanti dati e sono fortemente influenzati dalla presenza di outliers;

- gli algoritmi del secondo tipo invece risultano più efficienti in termini computazionali, se si hanno a disposizione tanti dati, sono meno influenzati da valori anomali e permettono che un'unità statistica possa cambiare il proprio cluster durante il processo iterativo.

Algoritmi gerarchici

Gli algoritmi gerarchici producono una struttura gerarchica dei dati, cioè una struttura ad albero binario.

Il problema adesso sta nel calcolare la distanza tra due gruppi. Nella matrice di dissimilarità abbiamo le distanze tra gli elementi che per gli algoritmi gerarchici agglomerativi corrispondono ai gruppi iniziali. Perciò, come primo passo, posso aggregare gli elementi più simili formando gruppi di due elementi. Per calcolare poi la distanza fra gruppi di più elementi le misure più usate sono:

Del legame singolo ➔ calcolo le distanze tra tutti gli elementi del primo gruppo e tutti gli elementi del secondo gruppo. La distanza tra i due gruppi è quella minima.

Del legame completo ➔ calcolo le distanze tra tutti gli elementi del primo gruppo e tutti gli elementi del secondo gruppo. La distanza tra i due gruppi è quella massima.

Del legame medio ➔ calcolo le distanze tra tutti gli elementi del primo gruppo e tutti gli elementi del secondo gruppo. La distanza tra i due gruppi è la media ponderata delle distanze. La media va ponderata perché i due cluster possono contenere un numero di elementi non uguale.

Metodo di Ward ➔ Tecnica gerarchica agglomerativa, usata per assegnare la numerosità dei cluster. Ad ogni passaggio agglomerativo i due cluster che si fondono sono quelli con il più

piccolo incremento nella somma totale del quadrato della distanza all'interno del cluster.

I raggruppamenti cambiano con la misura adottata.

Algoritmi non gerarchici

Il metodo più noto e anche il più usato è quello delle K-medie.
L'obiettivo è di individuare dei punti di aggregazione, centroidi, attorno ai quali costruire i gruppi; le osservazioni vanno dentro al gruppo che contiene il centroide più vicino.
I centroidi non sono fissi ma variano al procedere dell'algoritmo.
Procedimento: una volta suddivisa la popolazione in k gruppi, k è il numero di cluster che si vogliono ottenere ed è noto a priori, la dissimilarità totale può essere intesa come la dissimilarità complessiva all'interno dei gruppi, D_{entro}, più la dissimilarità complessiva tra i gruppi, D_{tra}. Il nostro obiettivo è quello di scegliere i gruppi in modo da ottenere una situazione di omogeneità all'interno e di eterogeneità tra di loro, quindi l'obiettivo è minimizzare D_{entro} oppure massimizzare D_{tra}. L'algoritmo esamina tutte i gruppi possibili è sceglie la situazione migliore.
L'algoritmo delle K-medie utilizza la distanza euclidea ed è applicato solo a variabili quantitative continue.

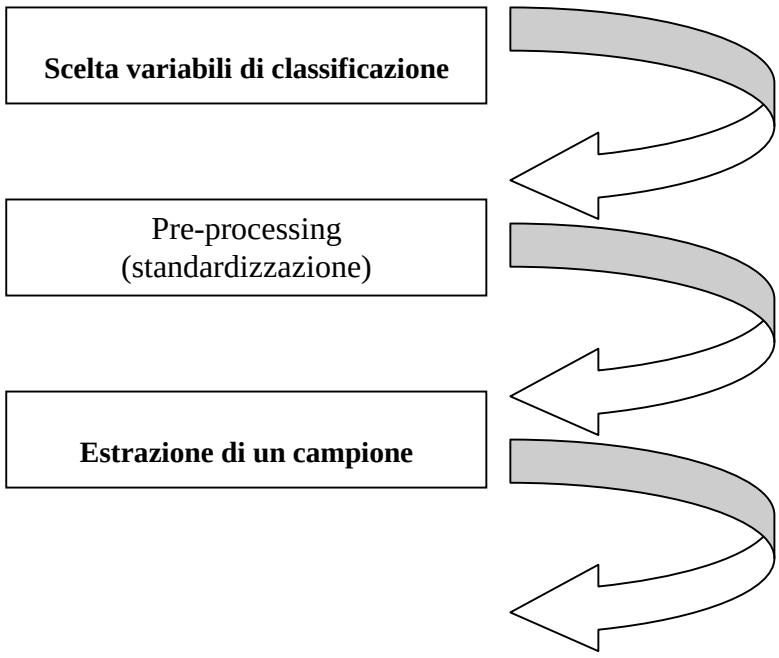
$$D_{entro} = 2 \sum_{k=1..K} \sum_{G(i)=1..k} || x_i - m_k ||^2 \qquad [5]$$

Dove m_k è il vettore formato dalle medie aritmetiche di ciascuna variabile.
Se invece si vuole utilizzare l'algoritmo con variabili qualitative allora bisogna scegliere un'altra forma di dissimilarità, introducendo il concetto di menoide. (*)

---

[5] (*): Azzalini & Scarpa, Analisi dei dati e Data Mining

2.3.3 Fasi del processo di classificazione



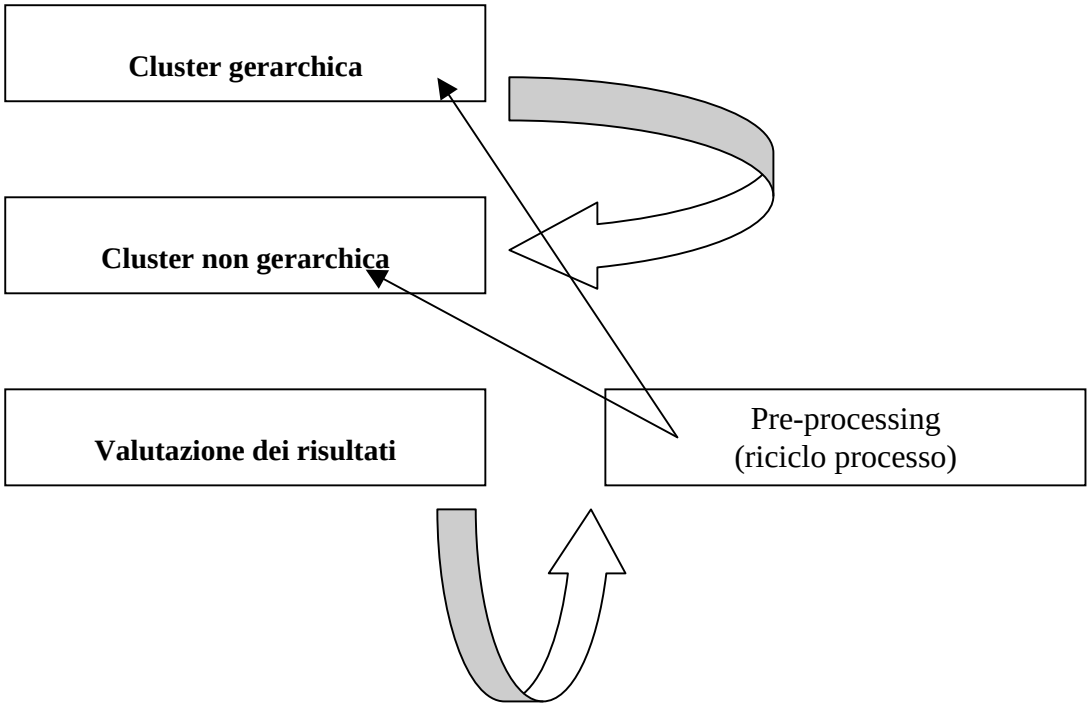


Figura: lucidi lezione Sistemi Informativi Aziendali

Obiettivi:

- Individuare cluster non noti a priori, significa che a priori non so le caratteristiche dei cluster;
- Offrire un servizio differenziato alla clientela. Maggiore è la diversità fra i cluster maggiore è la differenziazione che posso ottenere;
- Effettuare azioni di micro-marketing mirate.

Scelta delle variabili di classificazione e standardizzazione:

Dato un insieme di unità statistiche é relative variabili e necessario decidere quali variabili utilizzare nel processo o eventuali trasformazioni.

Le variabili scelte devono essere sottoposte ad una standardizzazione per renderle indipendenti dalla scala di misura.

Estrazione di un campione:

Solo nel caso di una grossa quantità di dati disponibili.

Cluster gerarchica:

Si effettua una cluster gerarchica per avere il numero dei cluster ottimale.

Cluster non gerarchica:

Una volta ottenuto il numero dei cluster posso usarlo per effettuare la cluster non gerarchica il cui risultato sarà quello finale.

Valutazione dei risultati:

Per valutare la bontà della classificazione ottenuta si può usare il valore di certi indicatori:

1. Frequency ➔ numero di unità statistiche appartenenti a ciascun cluster. I vari gruppi non devono contenere un numero di elementi molto diverso, non ci devono essere cluster troppo vuoti oppure cluster troppo numerosi nei quali si concentri la maggior parte delle osservazioni (elephant cluster).
2. Massima distanza dal seme all'osservazione ➔ indica la distanza massima tra il centroide di ciascun cluster e la relativa osservazione maggiormente distante. Valori piccoli di questo indicatore stanno ad indicare una buona classificazione.

3. Distanza tra cluster ➔ indica la distanza tra i centroidi dei cluster individuati. Valori alti indicano che c'è una netta separazione tra i gruppi.
4. R-Squared ➔ indica la quota di variabilità spiegata a livello totale e relativamente a ciascuna delle variabili.

2.4 SAS® e metodologia SEMMA

Il Data Mining è un processo e di conseguenza, la soluzione di un problema di Data Mining necessita di una soluzione software integrata che racchiuda tutti i passi di tale processo.

Il software SAS® Enterprise Miner™ racchiude tutti i passi per un progetto di Data Mining. Infatti, partendo dal campionamento dei dati

permette di effettuare analisi e modelli di vario tipo che portano all'informazione di business ricercata.
Enterprise Miner utilizza una singola interfaccia utente grafica (GUI) di tipo point-and-click che permette di arrivare all'informazione partendo da una serie di dati. L'utilizzo avviene mediante barre degli strumenti, menu, finestre e schede che forniscono tutti gli strumenti di Data Mining.
L'interfaccia è progettata in modo da poter essere usata dagli analisti aziendali con nozioni minime di statistica, e dagli esperti statistici con la possibilità per questi di esplorare i dettagli.

La metodologia SEMMA

SEMMA sta per Sample, Explore, Modify, Model e Assess e rappresenta la struttura generale per organizzare un progetto di Data Mining. In altre parole sono le fasi che permettono di arrivare all'informazione, partendo da un insieme di dati strutturati. Non tutti i progetti di Data Mining devono seguire ogni passo della metodologia SEMMA che offre agli utenti un metodo strutturato per concettualizzare, creare e valutare i progetti di Data Mining.
La rappresentazione visiva di questa struttura è un PFD che illustra graficamente i passi di ogni singolo progetto.

Nodi di campionamento

Quando si ha a disposizione una grossa quantità di dati è utile utilizzare un campione di questi per il processo di Data Mining in modo da ridurre il tempo necessario per l'elaborazione.
Input data source: consente di accedere e interrogare data set SAS per utilizzarli in progetti di Data Mining. Dopo aver inserito il nome del data set inizia la pre-elaborazione dell'origine dei dati, cioè la generazione automatica di metadati, come i ruoli delle variabili in

input e le statistiche di riepilogo. Per default utilizza un campione di 2000 osservazioni per determinare i metadati che non va confuso con il campione costruito ai fini dell'analisi.

È possibile modificare il ruolo del data set, per default è utilizzato come data set grezzo, ma è possibile fare in modo che il ruolo dei dati sia di addestramento, validazione o test.

Consente di ridefinire i ruoli, la scala di misura, i formati e le etichette delle variabili, che per default sono determinate dal software durante la pre-elaborazione dei dati, basandosi sulle informazioni ricavate dal campione dei metadati.

Per le variabili continue permette di visualizzare una tabella contenente le statistiche di riepilogo (valore minimo, massimo, media, deviazione standard…..).

È possibile utilizzare più nodi input data source in un singolo progetto.

Nodo Sampling: consente di estrarre un campione dal data set di input. Ovviamente il campionamento è consigliato solo se si ha a disposizione una grossa quantità di dati perché l'unica utilità è quella di ridurre il tempo di elaborazione.

Il nodo deve essere preceduto dal nodo input data source e può essere seguito da nodi di esplorazione e stima dei modelli e restituisce come output un data set contenente le osservazioni del campione.

Il campionamento può essere: casuale, ogni n osservazioni, stratificato, sulle prime osservazioni di un data set di input. Per qualsiasi tipo di campionamento è possibile specificare il numero di osservazioni o la percentuale della popolazione da selezionare.

Nodo Data partition: dopo il campionamento i dati possono essere partizionati prima di iniziare la costruzione dei modelli.

Il nodo data partition consente di partizionare i dati del data set iniziale o del campione in data set singoli utilizzati per:

- Costruire il modello (training);
- Valutare la bontà del modello costruito con i dati dell'insieme training (validation);
- Una valutazione finale del modello scelto (testing). Questo è opzionale.

Per partizionare le osservazioni viene eseguito un campionamento casuale stratificato o semplice, specificando la quota delle osservazioni da assegnare ad ogni partizione.

Nodi di esplorazione

Utili per esplorare e modificare i propri dati. Si studia una variabile alla volta, si può così notare la presenza di valori mancanti o anomali ed eliminare la variabile.

Nodo distribution explorer: consente di esplorare grandi quantità di dati in ambiente grafico. È possibile generare istogrammi multidimensionali per variabili categoriche o continue.

Nodo insight: consente di esplorare e analizzare in modo interattivo i dati attraverso molteplici grafici e analisi statistiche.

Nodo associations: consente di effettuare una association discovery, cioè permette di identificare gli elementi che si presentano contemporaneamente in un dato evento. Le associazioni vengono scoperte in base a conteggi del numero di volte in cui gli elementi si presentano da soli o in combinazione in un database.

Nodo variable selection: se le variabili in input sono in quantità molto elevata questo nodo può aiutare a ridurre il numero di variabili in input eliminando quelle che sono collegate alla variabile target.

Il nodo può identificare le variabili che risultano utili per stimare le variabili target, ricerca le variabili con una relazione gerarchica (provincia e CAP) e tiene solo la variabile che presenta i maggiori

dettagli eliminando l'altra, o viceversa, elimina le variabili con una alta percentuale di valori missing (per default 50%), elimina le variabili categoriche che hanno una singola modalità o che hanno un numero di modalità maggiore di n.

Nodi di modifica

Consentono tutte le modifiche sulle variabili, compresa la modifica degli attributi dei data set (come i nomi e i ruoli), e di creare nuove variabili utili all'analisi; permettono inoltre di applicare filtri per escludere valori anomali.

Nodo data set attributes: consente di modificare gli attributi dei data set come i nomi, le descrizioni e il ruolo.

Nodo trasform variables: consente di creare nuove variabili che sono la trasformazione di variabili già esistenti. Le trasformazioni possono essere utili per stabilizzare varianze, eliminare la non linearità, correggere la non normalità. Le possibili trasformazioni possono essere logaritmica, radice quadrata, inversa….. con la possibilità di inserire una propria formula.

Il nodo trasform variables consente di raggruppare una variabile continua in una variabile di tipo ordinale (trasformazioni binning) attraverso vari metodi:

- Trasformazione quantile che suddivide la distribuzione di frequenza di una variabile in n classi contenenti lo stesso numero di osservazioni.
- Trasformazione bucket che divide la variabile in n classi di uguale ampiezza, in base alla differenza tra il valore minimo e massimo. Le classi però contengono un numero di osservazioni non uguale.

Nodo filter outliers: permette di applicare un filtro sui dati in modo da escludere osservazioni indesiderate, come i valori rari per le

variabili categoriche, i valori estremi per le variabili continue, le osservazioni con valori missing.

Nodo replacement: la tabella in input può contenere osservazioni che hanno valori missing per una o più variabili. Per default Enterprise Miner non utilizzerà quell'osservazione, anche se scartare osservazioni incomplete può voler dire eliminare informazioni utili. Con il nodo replacement si possono sostituire valori missing per variabili intervallo o categoriche in modo da poter utilizzare tutte le osservazioni.

Nodo clustering: esegue un raggruppamento delle osservazioni.

Dopo aver effettuato il clustering, le caratteristiche dei gruppi possono essere analizzate graficamente e si possono così vedere le differenze tra i vari cluster.

L'identificativo del cluster di appartenenza di ogni osservazione può essere passato ad altri nodi per essere utilizzato come variabile di input, di identificazione, di gruppo o di target.

Nodi di modellazione

Costruiscono i seguenti modelli: regressione lineare e logistica, alberi decisionali, reti neurali.

Nodo regression: consente di stimare modelli di regressione lineare o logistica. La prima permette di stimare il valore di una variabile continua (target) come una funzione lineare di una o più variabili date in input; la seconda invece prevede la probabilità che una variabile target binaria o ordinale assuma un determinato valore.

Nodo tree: consente di creare alberi decisionali per classificare o stimare la decisione più appropriata quando si specificano più decisioni alternative.

La classificazione avviene in funzione di una variabile target binaria o nominale; in caso di una variabile target continua il risultato è allora la stima.

Nodo neural network: una rete neurale è un'applicazione che cerca di simulare la neurofisiologia del cervello umano e impara a identificare le relazioni nei dati a partire da un campione rappresentativo. Possono così produrre previsioni e stime.

Nodo ensemble: per combinare i risultati dei nodi sopra elencati e creare un singolo modello integrato.

I nodi sopra elencati dispongono di una funzione chiamata ***Model Manager*** in cui è possibile memorizzare e accedere ai modelli. In questo modo è possibile confrontare modelli diversi.

Nodi di valutazione

Nel caso di utilizzo di più modelli permette di scegliere quello migliore. Le statistiche da valutazione del modello sono calcolate nei nodi di modellazione nei quali si possono confrontare modelli creati con nodi uguali; nei nodi di valutazione è possibili invece confrontare modelli fatti con nodi diversi.

Nodo assessment: i criteri di valutazione sono i ricavi previsti e reali ottenuti dai risultati dei modelli, indipendenti dalla dimensione del campione e dal tipo di algoritmo usato, quindi sono confrontabili.

Input del nodo assessment:

- Scored data set: sono una serie di probabilità a posteriori per ogni valore di una variabile target binaria, nominale o ordinale; se la variabile è continua allora è una stima. I nodi di modellazione producono come output uno scored data set.
- Matrice di decisione: è una tabella dei profitti e dei costi attesi per ogni possibile decisione associata alla variabile target.

Nodo reporter: permette di generare report a partire dal processo di Data Mining. Il nodo mette insieme i risultati dell'analisi in un report in formato HTML e che può essere visto attraverso un Web Browser. Il report contiene il titolo, l'identificativo dell'utente, la data di creazione e i commenti, l'immagine del diagramma di flusso, i parametri definiti nei nodi e i risultati.

Nodi di utility

Nodo group processing: consente di definire variabili di raggruppamento, per ottenere analisi separate per ogni modalità della variabile di raggruppamento.

È possibile inserire un solo nodo group processing per ogni diagramma di flusso del processo.

3. Progetto

L'azienda: *

EGQ Consulenza d'azienda Srl è una società che opera nel settore della consulenza di direzione a partire dal 1991. La società opera a livello nazionale sia da sola, sia attraverso un network di società collegate; l'area di intervento principale è comunque costituita dal Triveneto, dove la società ha la sede e le proprie unità operative.

EGQ svolge la propria attività di consulenza di direzione e formazione in aziende appartenenti a diversi settori merceologici

(agroalimentare, arredamento, biomedicale, chimico, edile, elettrico, elettronico, installazioni elettriche, legno, metalmeccanico, segnaletica stradale, imballaggio e logistica, servizi in generale fra cui il servizio turistico alberghiero), offrendo la pluriennale esperienza della propria struttura per la progettazione, lo sviluppo e l'implementazione di programmi di consulenza e formazione.

EGQ Consulenza d'azienda si rivolge sia all'industria sia ad aziende operanti nei servizi privati e pubblici (Pubblica Amministrazione) tramite collaboratori di provata esperienza professionale per la progettazione, lo sviluppo e l'implementazione di programmi di consulenza e formazione.

Determinazione del problema di business:

L'obiettivo del progetto è la segmentazione dei potenziali clienti per poter offrire servizi diversi a gruppi di clienti diversi.

Per potenziali clienti intendo tutte le aziende che sono state già contattate e che potrebbero essere interessate ma che non sono ancora clienti effettivi.

---

* : www.egq.it

Selezione ed organizzazione dei dati:

L'azienda aveva una grossa quantità di dati disponibili su carta, i dati sono stati quindi inseriti in un database. Questa è una fase fondamentale del processo di Data Mining.

Ogni potenziale cliente è descritto da queste variabili demografiche:

- Partita IVA che funge da chiave, cioè identifica univocamente un record (un potenziale cliente) e che, se non è disponibile viene sostituita con un codice progressivo.
- Società, il nome dell'azienda.
- Indirizzo

- Città
- CAP
- Provincia
- Nazione, uguale per tutti i dati inseriti.
- Telefono
- Telefono secondario
- Fax
- E-mail
- Sito web

Ad ogni record è associata una scheda clienti che contiene i seguenti dati:

1. dimensioni per fatturato che può assumere come valore
    - micro (da 0 a 2 mil €)
    - piccola (da 2 a 10 mil €)
    - media (da 10 a 50 mil €)
    - grande (oltre 50 mil €)

2. dimensioni per dipendenti
    - micro (fino a 10 dipendenti)
    - piccola (da 10 a 50 dipendenti)
    - media (da 50 a 250 dipendenti)
    - grande (oltre 250 dipendenti)
3. settore merceologico
    - caccia – pesca (fino a confezione all'ingrosso, non distribuzione finale)
    - agricoltura – floricoltura – allevamento (fino a confezione all'ingrosso, non distribuzione finale)

- industria estrattiva (minerale, idrico/petrolifero, gassoso, fino a fornitura all'ingrosso, non distribuzione finale)
- produzione di materie prime di base e semilavorati (da trasformare in componenti o prodotti finiti)
- produzione/trasformazione di prodotti secondari (recupero, rigenerazione, anche quando il prodotto non va venduto, ma smaltito a costi o condizioni meno severe)
- trattamento superficiale per terzi (galvanica, verniciatura, serigrafia, lavorazioni superficiali varie senza aggiunta di componenti)
- produzione di componentistica (per tutti i settori e tecnologie, oggetti o componenti con una funzione)
- prodotti di largo consumo (alimentare, pulizie, cosmesi, chimico, farmaceutico utilizzabili dal cliente finale)
- prodotti finali per l'informazione (cartacei, SW, magnetici, anche pubblicitari o artistici)
- costruzioni, installazioni, restauri, cantieristica, grande engineering anche pluriennale
- arredamento in generale (casa, ufficio, illuminazione, serramenti, anche ornamentale)
- abbigliamento in generale (maglieria, calzature, abiti, bigiotteria, sanitario)
- oreficeria, gioielleria, vetro, ceramica, strumentazione musicale, sportiva, tempo libero, giocattoli
- macchine, attrezzature meccaniche, elettromeccaniche, elettroniche per la casa e l'ufficio dell'utente finale o dell'azienda
- produzione di mezzi di trasporto pubblico e privato (aria, terra, acqua)

- commercio di prodotti/servizi materiali (intermediari e agenzie comprese)
- pulizie, manutenzioni, riparazioni, noleggi, vigilanza, lavanderia, disinfestazioni e servizi materiali (ma non trattamenti superficiali di completamento prodotti)
- trasporti in genere (aria, terra, acqua)
- attività immobiliari in genere
- alberghi, ristorazione, turismo, sport, tempo libero, spettacolo
- attività sanitarie, benessere, bellezza, assistenza privata, organizzazioni no profit, organizzazioni confessionali e politiche
- servizi pubblici istituzionali
- postelegrafonici, elaborazione e gestione dati, trattamento dell'informazione in conto terzi
- prestazioni intellettuali, consulenze, produzione SW, ricerca, marketing, istruzione, pubblicità (non mezzi per)
- credito, assicurazioni, leasing, intermediazioni e servizi finanziari, gestione fondi e pensioni

4. tipologia di prodotto
    - terzista o controlavorista (prodotto e mercato di terzi)
    - terzista con ingeneria (vende servizio, ma prodotto e mercato di terzi)
    - prodotto proprio ma rete di vendita di terzi
    - prodotto e rete di vendita propri
    - prodotto di terzi e rete di vendita propri (commerciante – grossista)
5. mercato di sbocco
    - utente finale
    - negozio al minuto

- grossista
- grande distribuzione organizzata
- cercato dal cliente

6. proprietà
    - parte di gruppo nazionale
    - parte di gruppo estero
    - prima generazione padronale
    - prima generazione organizzata
    - prima generazione con figli operativi
    - successiva generazione senza management
    - successiva generazione con management
    - pubblic company
7. propensione all'export
    - solo Italia (obbligo per natura prodotto/servizio)
    - solo Italia con potenzialità di export
    - export tramite terzi
    - export in minima parte senza propria organizzazione
    - export con propria organizzazione

Completato l'inserimento i dati sono stati esportati in un file excel e poi trasformati in un data set SAS per poterli utilizzare con Enterprise Miner.

I dati sono strutturati, un potenziale cliente per riga descritto da una serie di variabili statistiche tutte di tipo qualitativo.

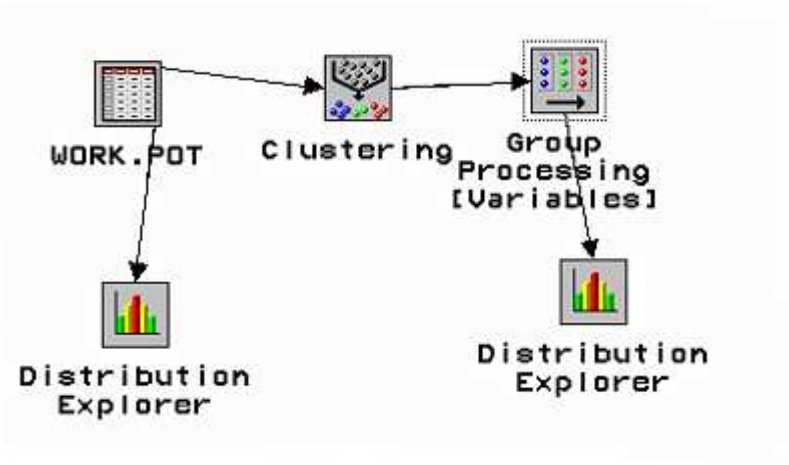


Figura: progetto creato con SAS® Enterprise Miner™

Analisi esplorativa dei dati:

In questa fase entra in gioco Enterprise Miner™ di SAS®, infatti dopo aver trasformato il file excel in un data set ho utilizzato il nodo input data source per poter accedere e interrogare il data set. Dopo aver inserito il nome del data set, premendo il tasto invio inizia la pre-elaborazione dell'origine dei dati, cioè la generazione automatica dei metadati, come i ruoli delle variabili in input e le statistiche di riepilogo.

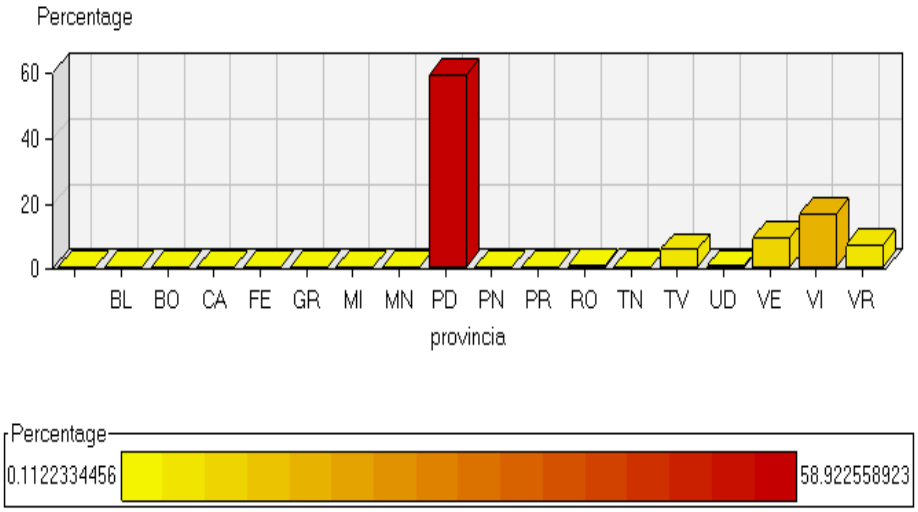
In questo nodo è possibile, tramite la scheda variables, ridefinire il ruolo delle variabili.

Essendo il numero di dati molto piccolo, 918, non è necessario estrarre un campione dei dati.

Collegando il nodo input data source al nodo distribution explorer è possibile esplorare i dati per poter individuare la presenza di valori anomali.

L'esplorazione è stata fatta per una variabile alle volta utilizzando degli istogrammi.

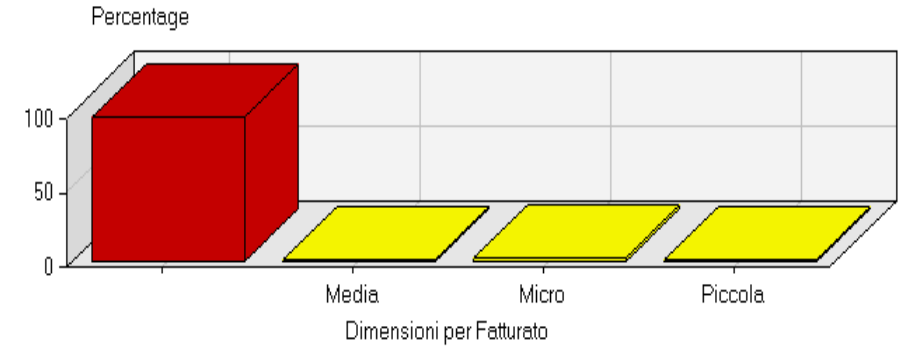
Variabile provincia:

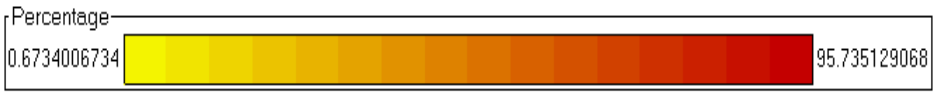


La maggior parte dei potenziali clienti dell'azienda hanno la sede a Padova, circa il 60%, mentre una bassa percentuale di essi hanno sede a Treviso, Venezia, Vicenza e Verona. La percentuale è quasi nulla per le altre zone. È una situazione normale visto che l'EGQ opera maggiormente nel Veneto ed in particolar modo nel padovano avendo la sede in provincia di Padova.

La variabile provincia verrà comunque usata per la determinazione dei cluster perché nonostante la percentuale di PD sia molto alta ci sono altre quattro zone (TV, VE, VI, VR) con percentuale molto più bassa ma comunque valida.

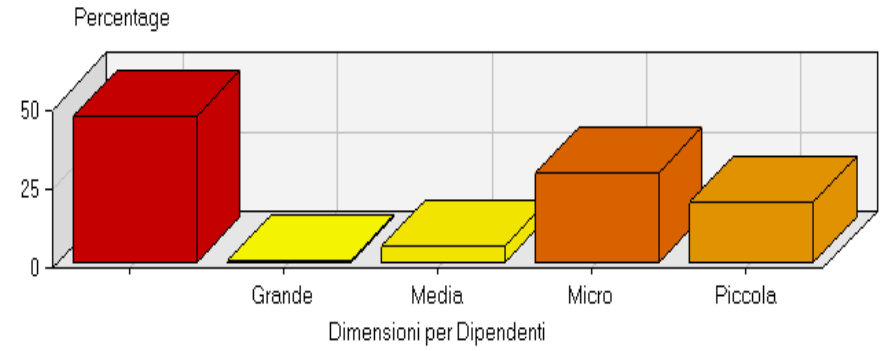
Variabile dimensioni per fatturato:

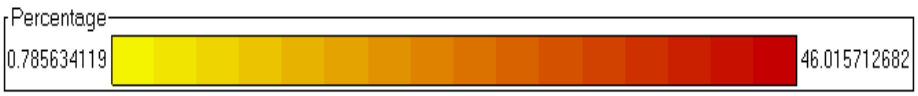




La variabile dimensione per fatturato non fornisce nessuna informazione utile, infatti più del 95% dei record non contengono nessun valore. Questo accade perché i potenziali clienti sono restii a dare un'informazione di questo tipo, la danno solo nel momento in cui diventano clienti effettivi. Ovviamente questa variabile sarà ignorata in sede di analisi.

Variabile dimensione per dipendenti:

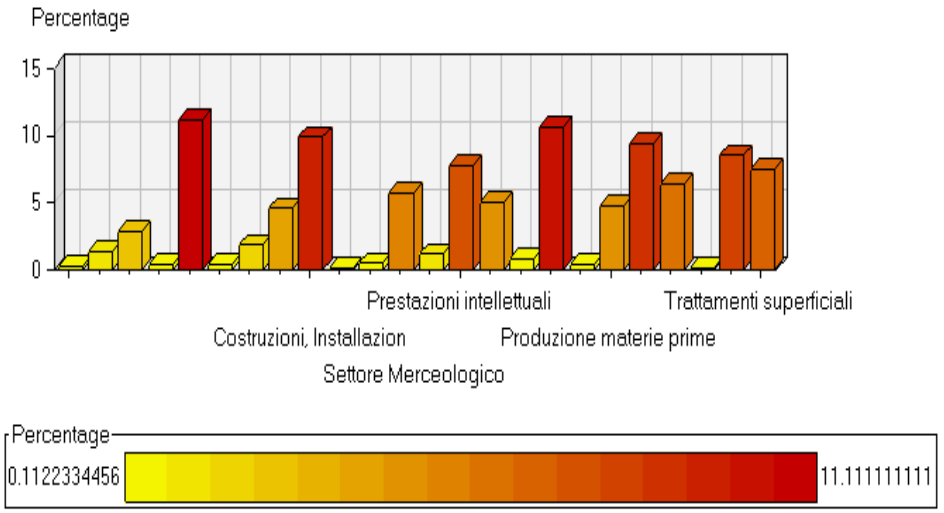




Anche in questa variabile c'è la presenza di valori non dati, circa il 46% dei record non hanno un valore per il numero di dipendenti. Però in questo caso la differenza sta nel fatto che più della metà

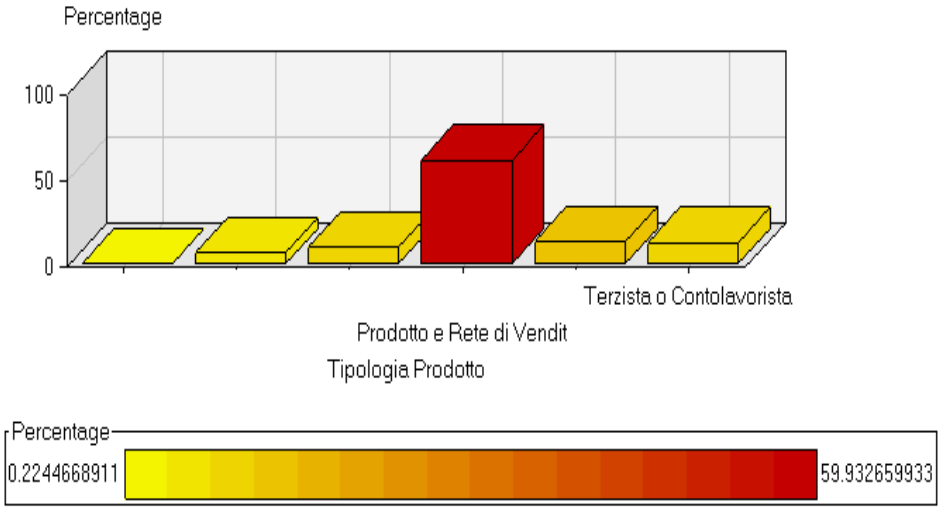
hanno un valore quindi la variabile verrà utilizzata per trovare i cluster.

Variabile settore merceologico:



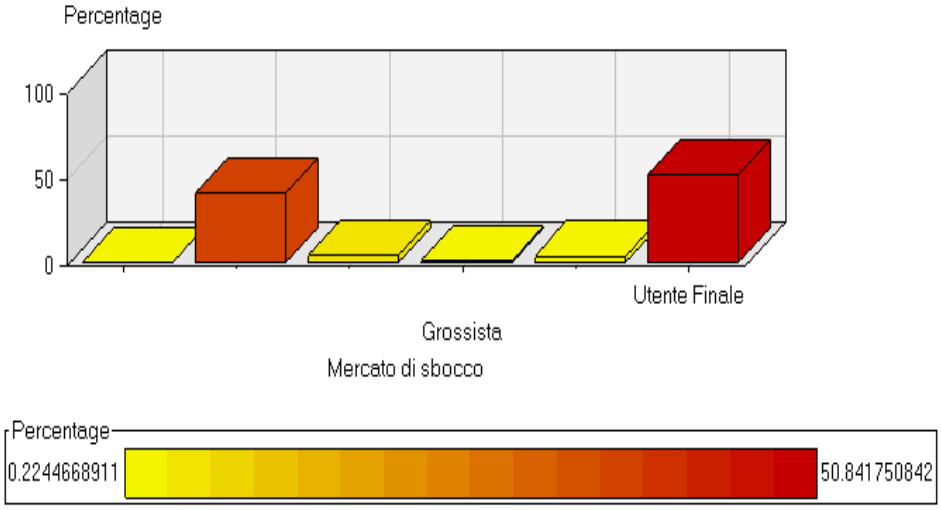
Questa variabile si distribuisce abbastanza tra le osservazioni, per questo motivo il contenuto informativo è molto elevato. L'utilizzo di questa variabile per la determinazione dei cluster è molto importante.

Variabile tipologia prodotto:



La situazione per questa variabile e più o meno simile alla dimensione per fatturato perché ho un valore che si differenzia dalla massa. In questo caso però il contenuto informativo c'è perché non ho valori vuoti, tranne lo 0,22% che è irrilevante. Questa variabile sarà quindi utilizzata per determinare i cluster.

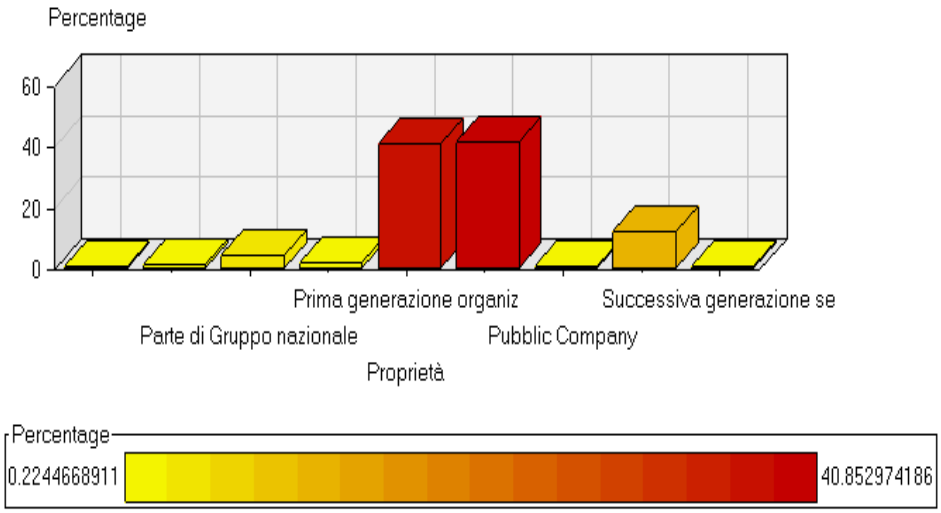
Variabile mercato di sbocco:



Lo 0,22% non ha un valore, però un certo contenuto informativo c'è e tende a spaccare in due la popolazione dato che il 50% assume come valore "utente finale" e circa il 47% assume come valore

“cercato dal cliente”. Il contenuto informativo non è molto elevato però la variabile sarà utilizzata per determinare i cluster.

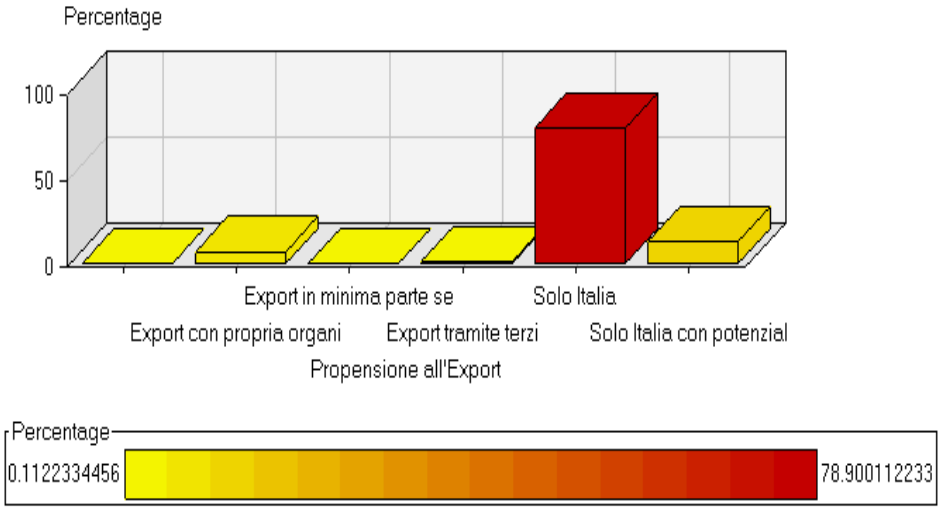
Variabile proprietà:



Abbiamo una situazione più o meno analoga a quella precedente, con la differenza che questa variabile non spacca la popolazione in due ma si differenzia in modo maggiore con due picchi che si riferiscono ai valori “prima generazione organizzata” e “prima generazione padronale” e con percentuali non del tutto trascurabili per i valori “successiva generazione con management”, circa 19%, e “parte di gruppo nazionale”, circa il 5%.

Lo 0,22% delle osservazioni non ha valore.

Variabile propensione all'export:



Il 79% delle osservazioni assumono come valore "solo Italia" per questa variabile. Questo potrebbe portare alla decisione di escluderla dall'analisi dei cluster, ma la presenza del 10% per il valore "solo Italia con potenzialità di export" e del 6% circa per il valore "export con propria organizzazione" porta a decidere di tenere anche questa variabile per la determinazione dei cluster anche se le percentuali sono molto più basse rispetto al picco.

Le altre variabili statistiche non sono state esplorate perché assumono valore diverso per ogni osservazione. Non offrono dunque informazioni utili, infatti già nel nodo input data source non era stato dato loro il ruolo di input.

Una volta esplorate tutte le variabili il fatto di avere una percentuale di almeno 0,11% di valori vuoti in tutte le variabili mi ha fatto pensare di avere un record completamente vuoto. Quindi sono andato a controllare il file excel ed infatti un record era del tutto senza valori, l'ho cancellato ed ho modificato anche il data set. Infine ho iniziato con la Cluster Analysis.

Cluster Analysis

L'analisi esplorativa dei dati mi ha permesso di capire quali variabili statistiche possono essere utilizzate per determinare i cluster.

Il nodo clustering mi permette di fare la Cluster Analysis collegandolo con il nodo input data source.

L'interfaccia del nodo clustering comprende la seguente scheda:

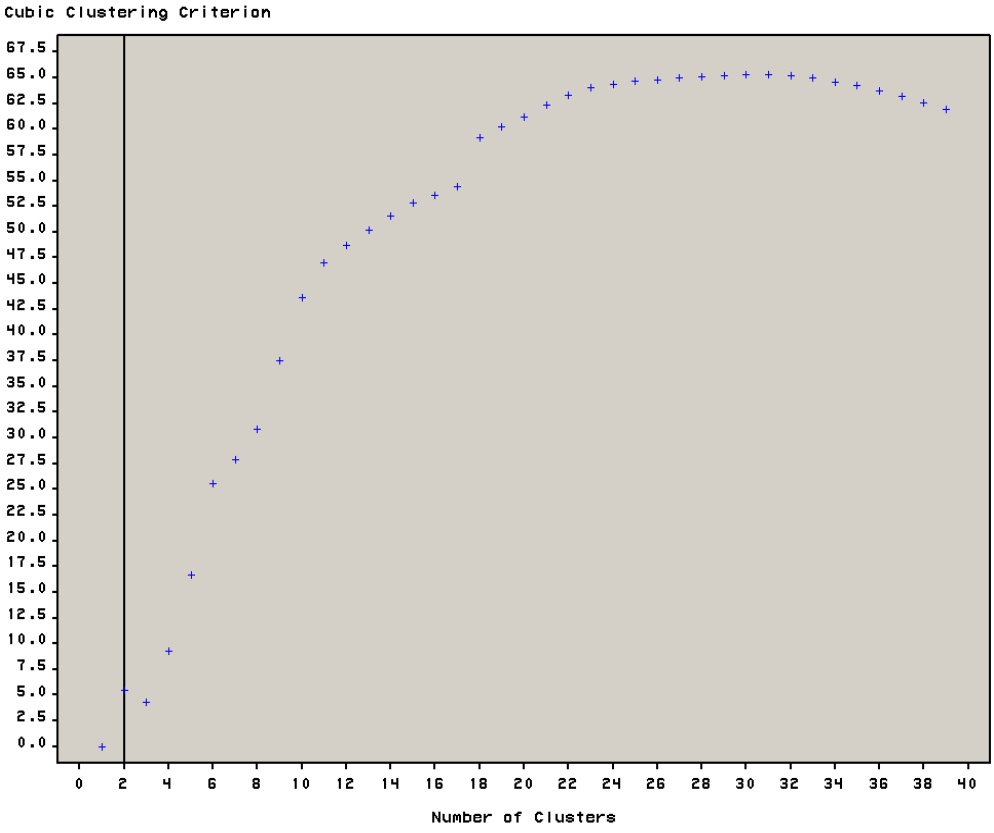
Cluster ➔ serve per impostare il numero massimo di cluster o per impostare la selezione automatica del numero ottimale (number of cluster) e per definire l'identificativo del segmento cioè il nome della variabile, la relativa etichetta e il ruolo nel data set.

Come detto dopo l'analisi esplorativa dei dati le variabili usate per la segmentazione sono la provincia, la dimensione per dipendenti, il settore merceologico, la tipologia di prodotto, il mercato di sbocco, la proprietà e la propensione all'export.

Name	Status	Model Role	Measurement	Type	Format	Label
SOCIET_	don't use	id	nominal	char	$68.	società
GRUPPO	don't use	rejected	binary	char	$13.	gruppo
INDIRIZZO	don't use	rejected	nominal	char	$46.	indirizzo
CAP	don't use	rejected	nominal	char	$5.	CAP
CITT_	don't use	rejected	nominal	char	$34.	città
PROVINCIA	use	input	nominal	char	$2.	provincia
TELEFONO	don't use	rejected	nominal	char	$11.	telefono
TELEFONO_SECONDARIO	don't use	rejected	nominal	char	$10.	telefono second
FAX	don't use	rejected	nominal	char	$11.	fax
E_MAIL	don't use	rejected	nominal	char	$37.	e-mail
SITO_WEB	don't use	rejected	nominal	char	$52.	sito web
PARTITA_IVA_CODICE_FISCALE	don't use	rejected	nominal	char	$16.	partita IVA/cod
DIMENSIONI_PER_FATTURATO	don't use	input	nominal	char	$8.	Dimensioni per
DIMENSIONI_PER_DIPENDENTI	use	input	nominal	char	$8.	Dimensioni per
SETTORE_MERCEOLOGICO	use	input	nominal	char	$127.	Settore Merceol
TIPOLOGIA_PRODOTTO	use	input	nominal	char	$44.	Tipologia Prodo
MERCATO_DI_SBOCCO	use	input	nominal	char	$33.	Mercato di sboc
PROPRIET_	use	input	nominal	char	$39.	Proprietà
PROPENSIONE_ALL_EXPORT	use	input	nominal	char	$51.	Propensione all

Da notare che la variabile dimensioni per fatturato, anche se è in input, non è stata usata per l'individuazione dei cluster (dont'use) per i motivi elencati in precedenza.

La prima parte dell'analisi consiste nella determinazione dei cluster in modo automatico.
Una volta sottomesso il processo il risultato è stato il CCC plot (Cubic Clustering Criterion), che è un indicatore della bontà del modello, tanto più alto è il valore di CCC, tanto più i gruppi sono omogenei al loro interno e distinti tra loro.
L'obiettivo, in questa prima parte è scegliere il valore per il quale CCC si impenna, per cui con un solo gruppo in più, la bontà aumenta di molto.

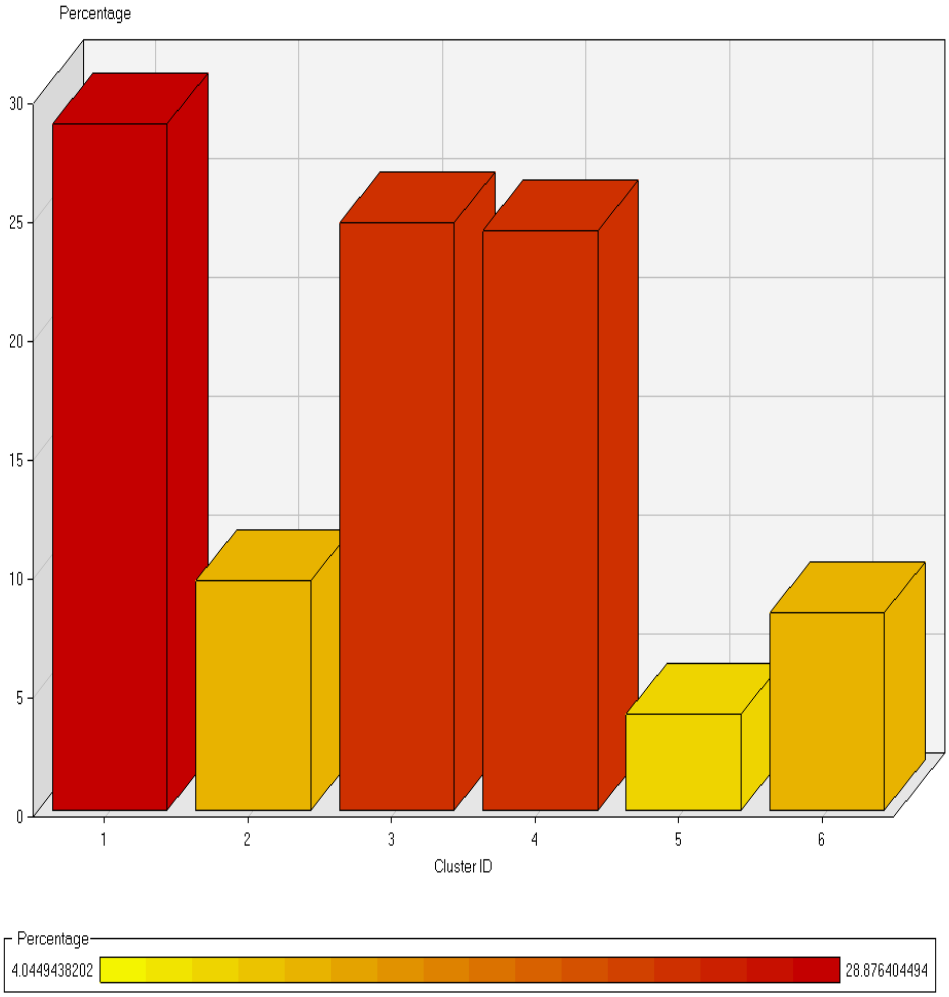


Il passaggio tra 4 cluster e 6 cluster è quello che fa aumentare di più la bontà del modello per cui il numero di gruppi ottimale è 6.
Riefettuo l'analisi impostando il numero dei cluster a 6.

Il nodo clustering mi ha dato come risultato le statistiche per ciascun cluster e 6 data set contenenti i clienti appartenenti a ciascun cluster con tutte le variabili statistiche iniziali, con la possibilità di esportarli in un file excel.
Per vedere se è stata fatta una buona segmentazione bisogna controllare se vi è la presenza di un elephant cluster:

CLUSTER	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	P
1	257	0.2237709137	2.5943121708	2	1.5527041512	***
2	86	0.2050342988	2.3589123518	1	1.5527041512	***
3	220	0.1918449893	2.5467018687	4	1.3843838773	***
4	217	0.1880755637	2.4171169955	3	1.3843838773	0
5	36	0.2524938601	2.5696091447	4	1.7441032107	0
6	74	0.2338381449	2.5499713876	1	1.8348707467	0

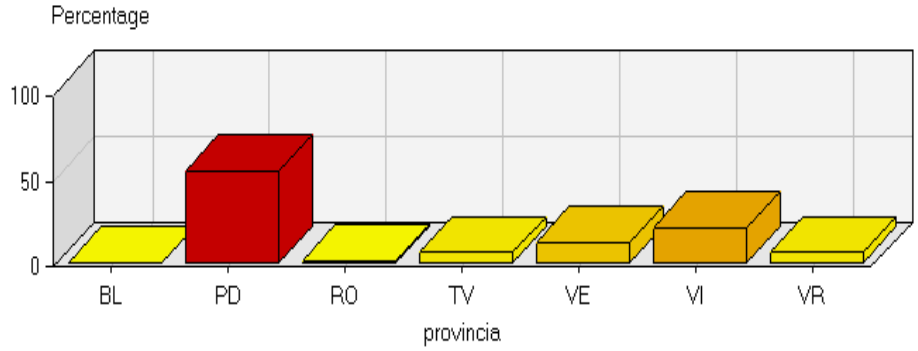
Guardando la frequenza all'interno di ogni cluster non sembra esserci un elephant cluster, ma tre gruppi molto popolati e altri tre molto poco popolati, quindi ritengo che la Cluster Analysis ha dato un risultato non ottimo ma abbastanza buono.

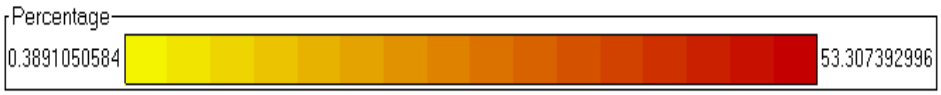


Adesso che i gruppi sono formati, l'ultimo passaggio è quello dell'analisi dei cluster utilizzando il nodo group processing e collegandolo al nodo clustering.

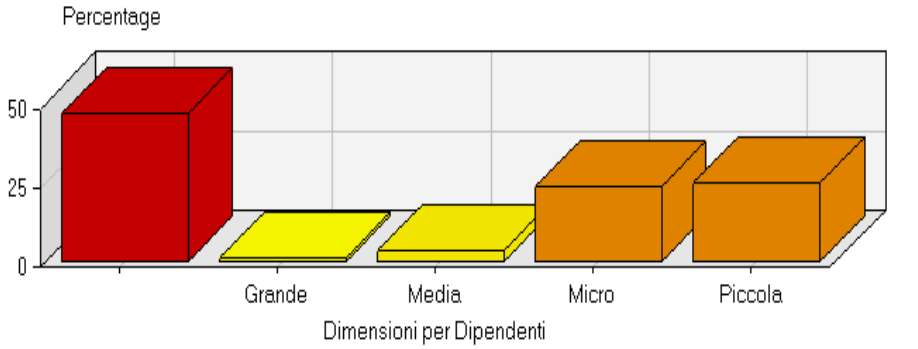
Il nodo group processing consente di definire variabili di raggruppamento e mi permette di esplorare, sempre con il nodo distributio explorer, una variabile alla volta per ogni cluster.

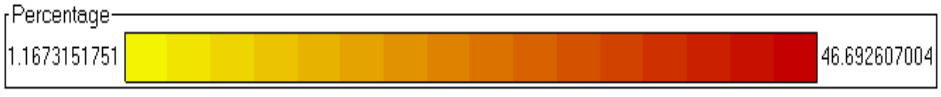
Cluster 1



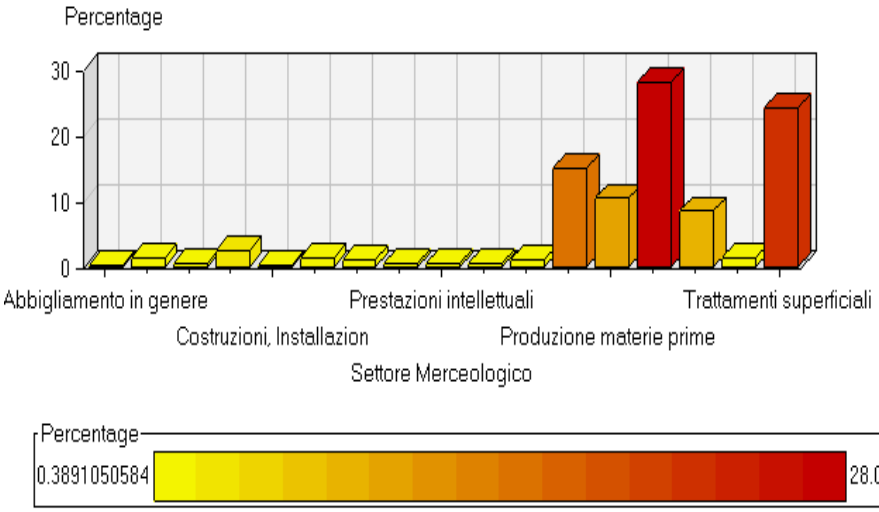


BL ➔ %0.39
PD ➔ %53.31
RO ➔ %1.17
TV ➔ %6.23
VE ➔ %12.06
VI ➔ %20.62
VR ➔ %6.23

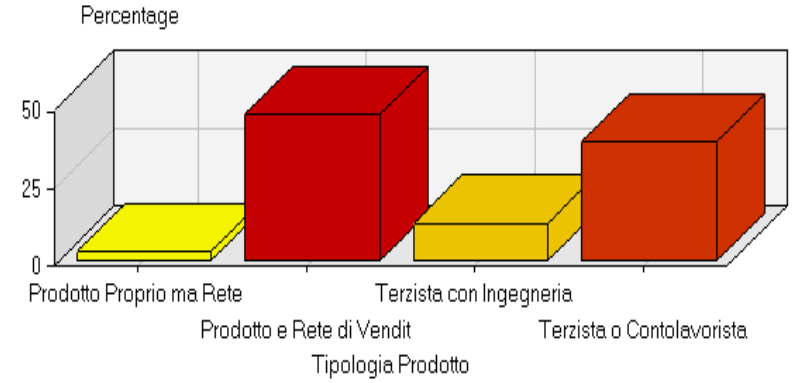


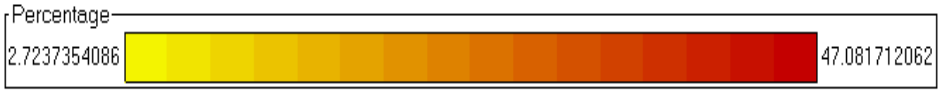


Grande ➔ %2.19
Media ➔ %6.57
Micro ➔ %44.53
Piccola ➔ %46.72

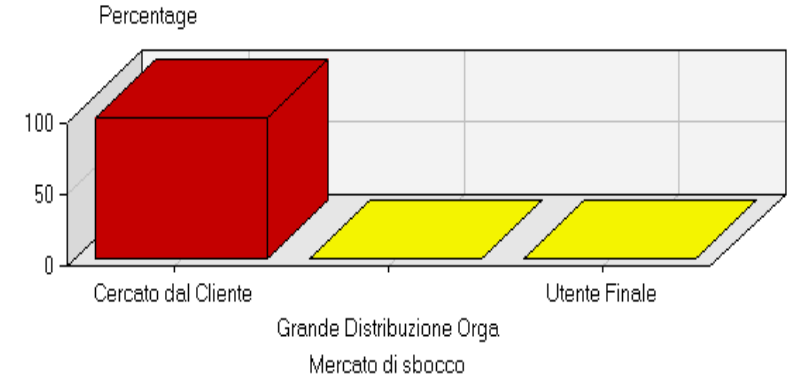


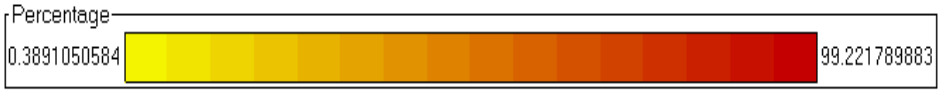
Abbigliamento in genere ➔ %0.39
Agricoltura, Floricoltura, Allevamento ➔ %1.56
Arredamento in generale ➔ %0.78
Attività sanitarie, Benessere, Bellezza, Assistenza privata, Organizzazioni no profit, Organizzazioni confessionali e politiche ➔ %2.72
Costruzioni, Installazioni, Restauri, Cantieristica, Grande Engineering anche pluriennale ➔ %0.39
Industria estrattiva ➔ %1.56
Macchine, Attrezzature meccaniche, elettromeccaniche, elettroniche per la casa e l'ufficio dell'utente finale e dell'azienda ➔ %1.17
Oreficeria, Gioielleria, Vetro, Ceramica, Gadget, Strumentazione musicale, sportiva, tempo libero, giocattoli ➔ %0.78
Prestazioni intellettuali, Consulenze, Produzione SW, Ricerca, Marketing, Istruzione, Pubblicità ➔ %0.78
Prodotti di largo consumo ➔ %0.78
Prodotti finali per l'Informazione ➔ %1.17
Produzione di Componentistica ➔ %15.18
Produzione materie prime di base e semilavorati ➔ %10.51
Produzione/Trasformazione di prodotti secondari ➔ %28.02
Pulizie, Manutenzioni, Riparazioni, Noleggi, Vigilanza, Lavanderia, Disinfestazioni e servizi materiali vari ➔ %8.56
Trasporti in genere ➔ %1.56
Trattamenti superficiali per terzi ➔ %24.12



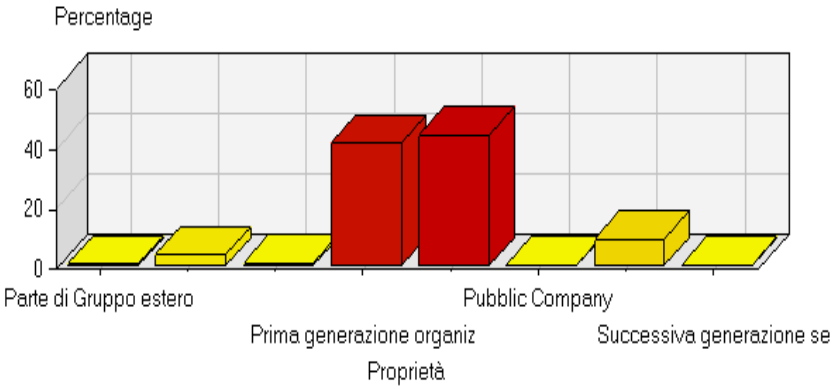


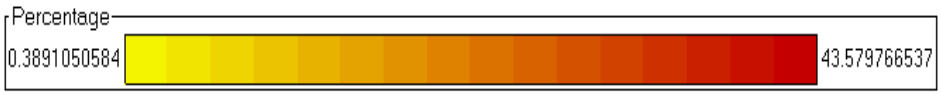
Prodotto Proprio ma Rete di Vendita di Terzi ➔ %2.72
Prodotto e Rete di Vendita Propri ➔ %47.08
Terzista con Ingegneria ➔ %12.06
Terzista o Contolavorista ➔ %38.13



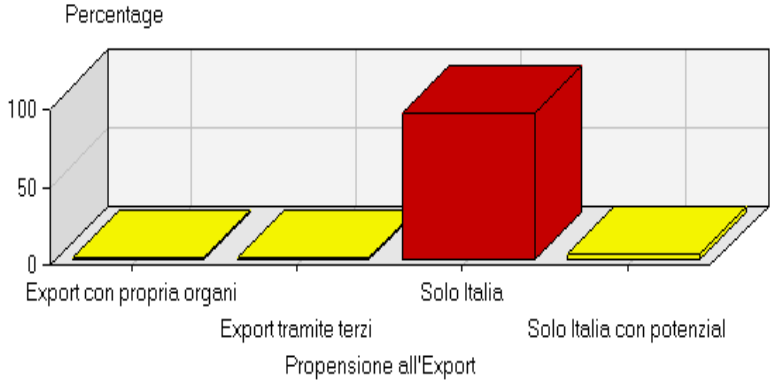


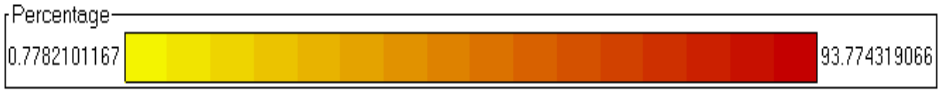
Cercato dal Cliente ➔ %99.22
Grande Distribuzione Organizzata ➔ %0.39
Utente Finale ➔ %0.39





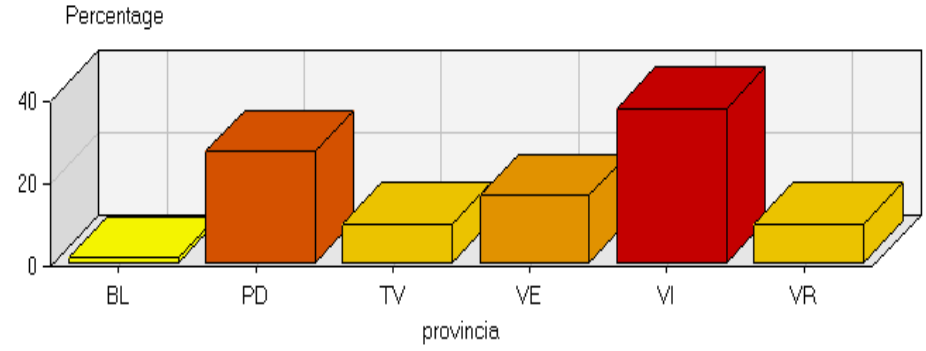
Parte di Gruppo estero ➔ %0.78
Parte di Gruppo nazionale ➔ %3.89
Prima generazione con figli operativi ➔ %1.17
Prima generazione organizzato ➔ %40.86
Prima generazione-padronale ➔ %43.58
Pubblic Company ➔ %0.39
Successiva generazione con management ➔ %8.95
Successiva generazione senza management ➔ %0.39

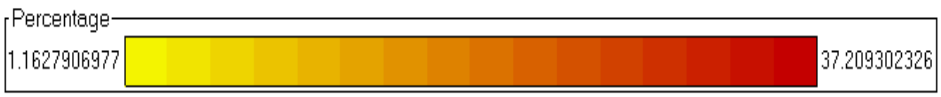




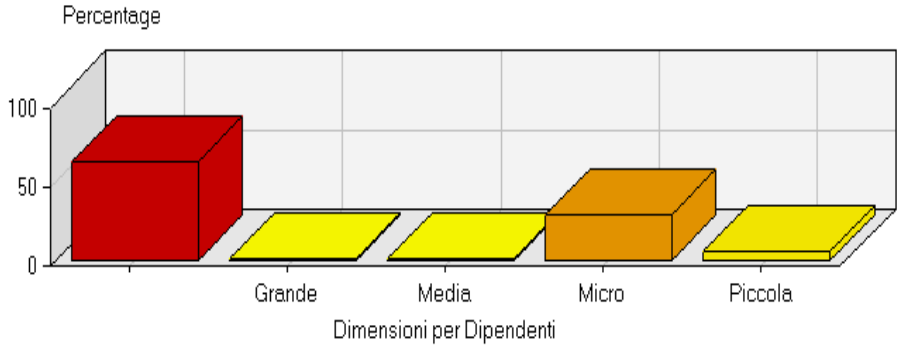
Export con propria organizzazione ➔ %0.78
Export tramite terzi ➔ %1.56
Solo Italia ➔ %93.77
Solo Italia con potenzialità di export ➔ %3.89

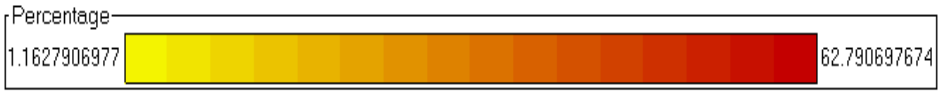
Cluster 2



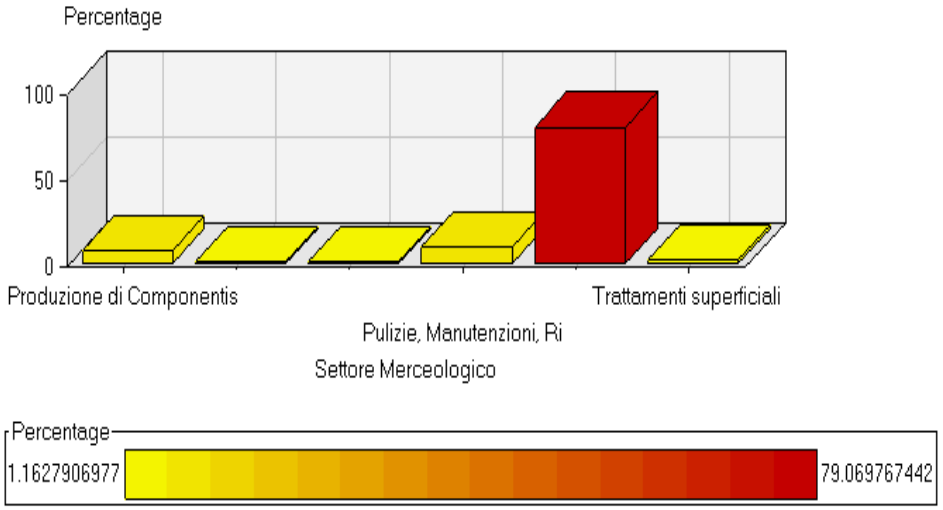


BL ➔ %1.16
PD ➔ %26.74
TV ➔ %9.30
VE ➔ %16.28
VI ➔ %37.21
VR ➔ %9.30

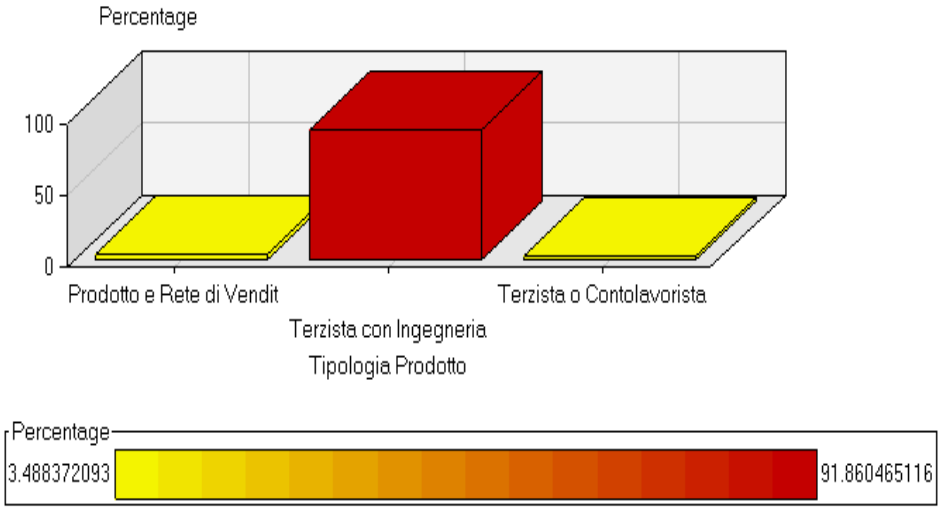




Grande ➔ %3.13
Media ➔ %3.13
Micro ➔ %78.13
Piccola ➔ %15.63

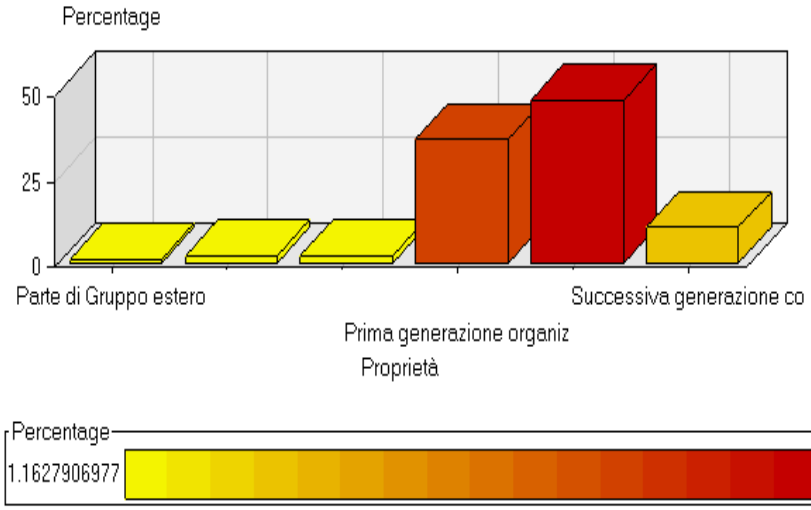


Produzione di Componentistica ➔ %6.98
Produzione materie prime di base e semilavorati ➔ %1.16
Produzione/Trasformazione di prodotti secondari ➔ %1.16
Pulizie, Manutenzioni, Riparazioni, Noleggi, Vigilanza, Lavanderia, Disinfestazioni e servizi materiali vari ➔ %9.30
Trasporti in genere ➔ %79.07
Trattamenti superficiali per terzi ➔ %2.33

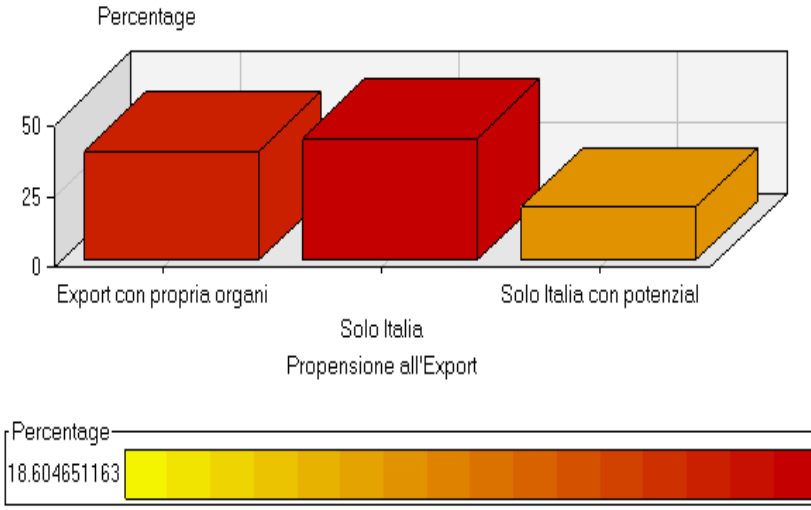


Prodotto e Rete di Vendita Propri ➔ %4.65
Terzista con Ingegneria ➔ %91.86
Terzista o Contolavorista ➔ %3.49

MERCATO_DI_SBOCCO ➔ Cercato dal Cliente ➔ %100.0

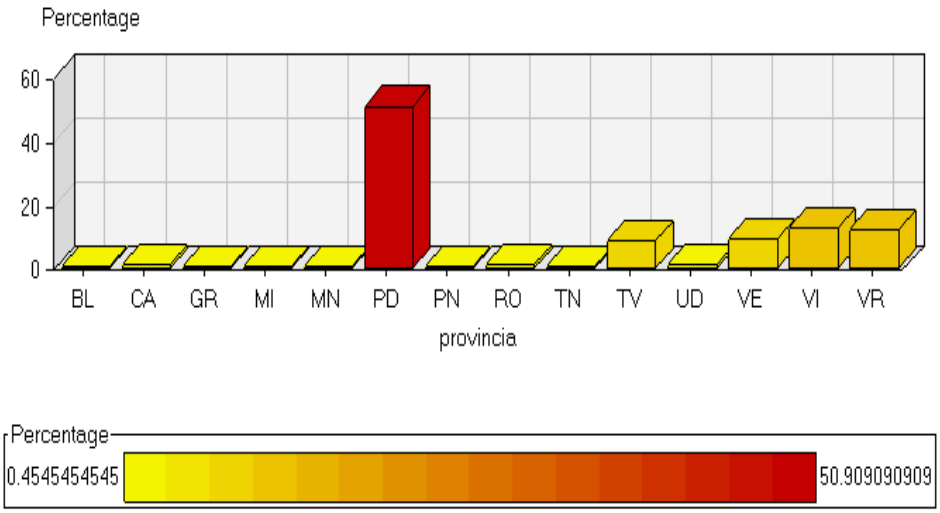


Parte di Gruppo estero ➔ %1.16
Parte di Gruppo nazionale ➔ %2.33
Prima generazione con figli operativi ➔ %2.33
Prima generazione organizzato ➔ %36.05
Prima generazione-padronale ➔ %47.67
Successiva generazione con management ➔ %10.47

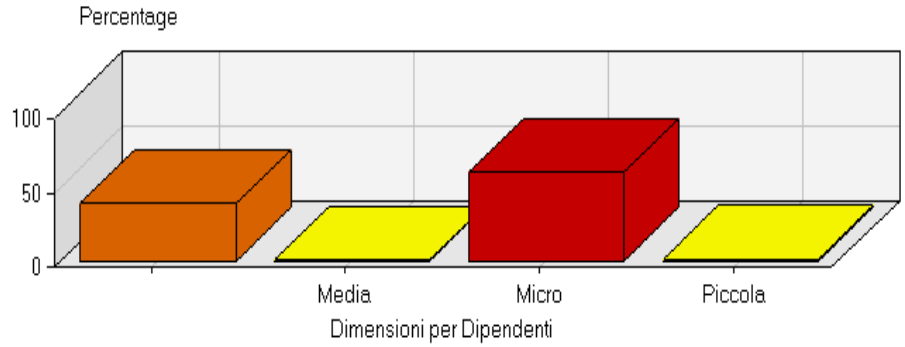


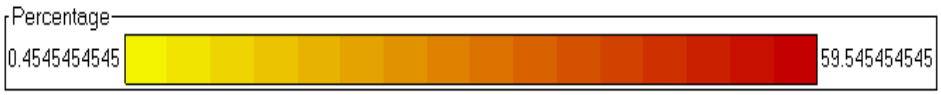
Export con propria organizzazione ➔ %38.37
Solo Italia ➔ %43.02
Solo Italia con potenzialità di export ➔ %18.60

Cluster 3

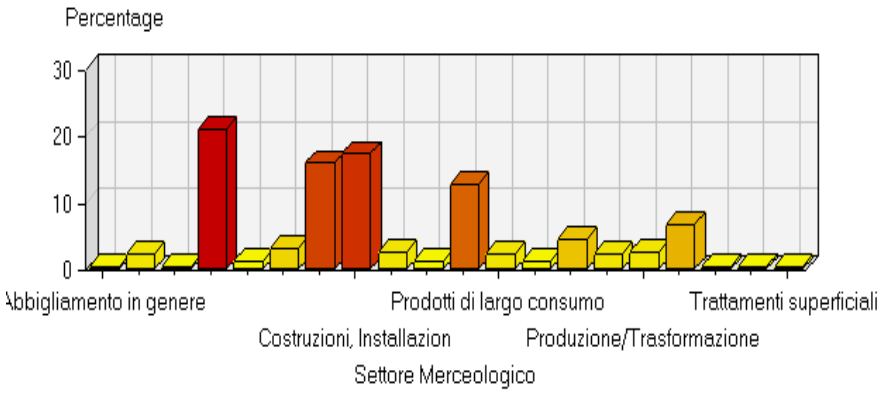


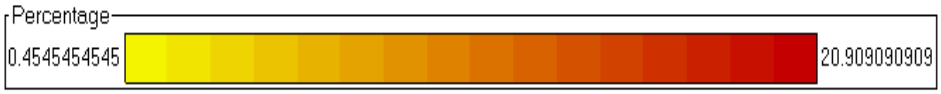
BL ➔ %0.45
CA ➔ %1.36
GR ➔ %0.45
MI ➔ %0.45
MN ➔ %0.45
PD ➔ %50.91
PN ➔ %0.45
RO ➔ %1.36
TN ➔ %0.45
TV ➔ %8.64
UD ➔ %0.91
VE ➔ %9.09
VI ➔ %12.73
VR ➔ %12.27





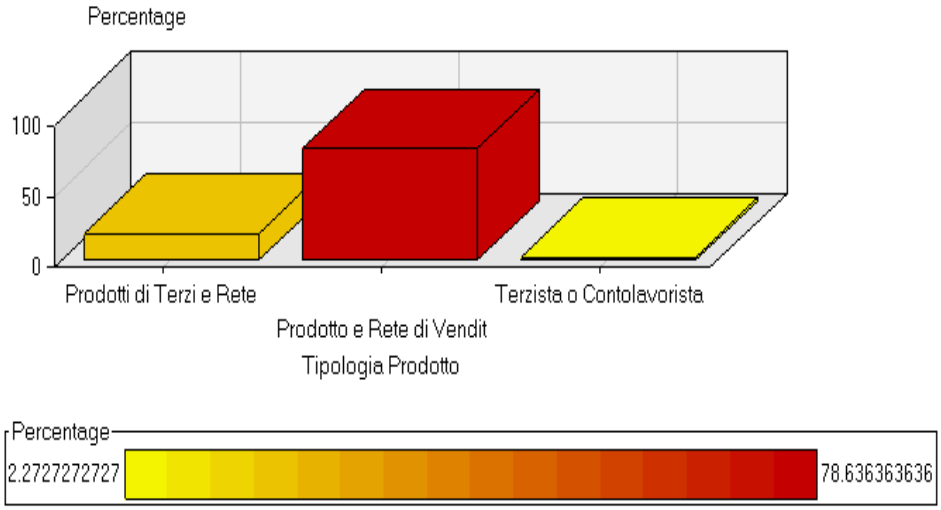
Media ➔ %0.74
Micro ➔ %97.04
Piccola ➔ %2.22



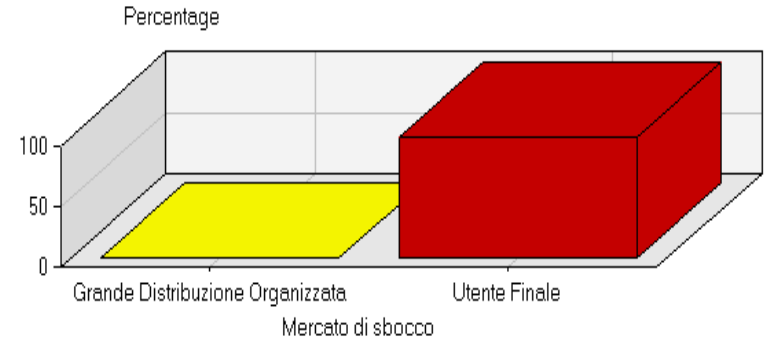


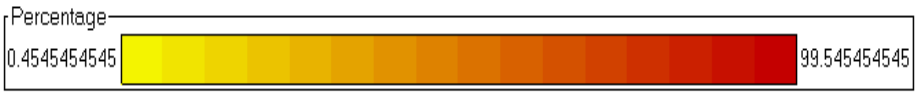
Abbigliamento in genere ➔ %0.45
Agricoltura, Floricoltura, Allevamento ➔ %2.27
Alberghi, Ristorazione, Turismo, Sport, Tempo libero, Spettacolo ➔ %0.45
Arredamento in generale ➔ %20.91
Attività Immobiliai in genere ➔ %1.36
Attività sanitarie, Benessere, Bellezza, Assistenza privata, Organizzazioni no profit, Organizzazioni confessionali e politiche ➔ %3.18
Commercio di prodotti/servizi materiali ➔ %15.91
Costruzioni, Installazioni, Restauri, Cantieristica, Grande Engineering anche pluriennale ➔ %17.27

Macchine, Attrezzature meccaniche, elettromeccaniche, elettroniche per la casa e l'ufficio dell'utente finale e dell'azienda ➔ %2.73
Oreficeria, Gioielleria, Vetro, Ceramica, Gadget, Strumentazione musicale, sportiva, tempo libero,
giocattoli ➔ %1.36
Prestazioni intellettuali, Consulenze, Produzione SW, Ricerca, Marketing, Istruzione, Pubblicità ➔ %12.73
Prodotti di largo consumo ➔ %2.27
Prodotti finali per l'Informazione ➔ %1.36
Produzione di Componentistica ➔ %4.55
Produzione materie prime di base e semilavorati ➔ %2.27
Produzione/Trasformazione di prodotti secondari ➔ %2.73
Pulizie, Manutenzioni, Riparazioni, Noleggi, Vigilanza, Lavanderia, Disinfestazioni e servizi materiali vari ➔ %6.82
Servizi Pubblici istituzionali ➔ %0.45
Trasporti in genere ➔ %0.45
Trattamenti superficiali per terzi ➔ %0.45

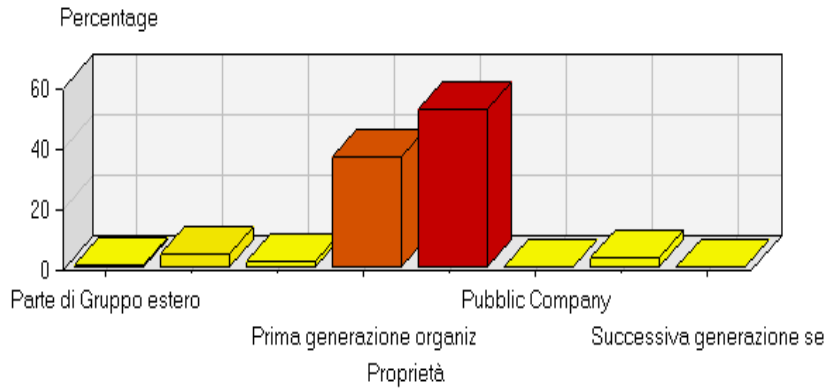


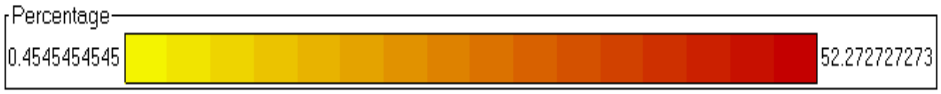
Prodotti di Terzi e Rete di Vendita Propri ➔ %19.09
Prodotto e Rete di Vendita Propri ➔ %78.64
Terzista o Contolavorista ➔ %2.27



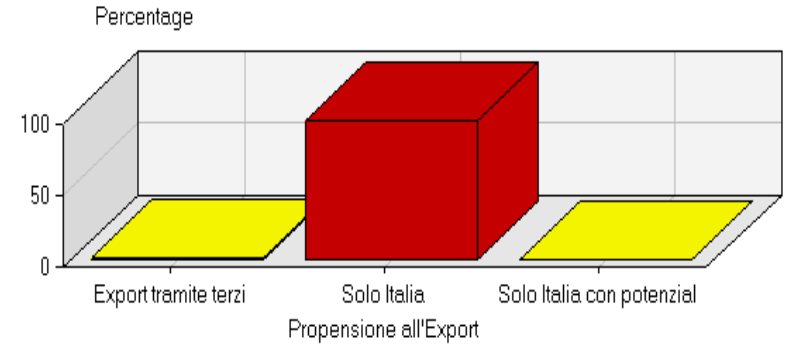


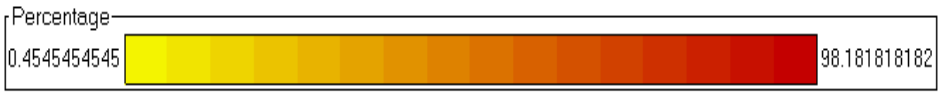
Grande Distribuzione Organizzata ➔ %0.45
Utente Finale ➔ %99.55





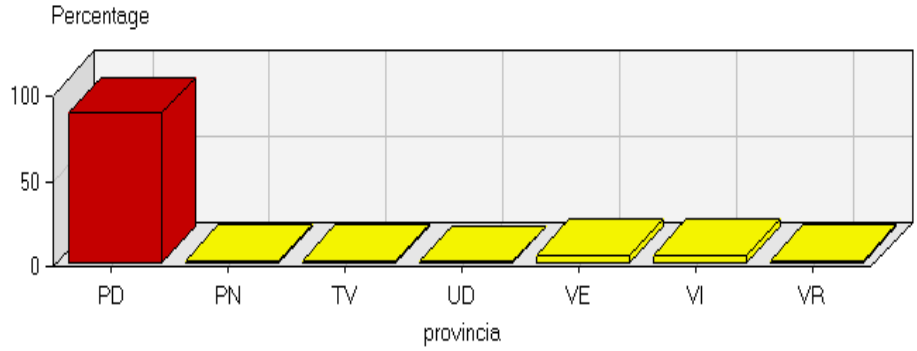
Parte di Gruppo estero ➔ %0.91
Parte di Gruppo nazionale ➔ %4.55
Prima generazione con figli operativi ➔ %1.82
Prima generazione organizzato ➔ %36.36
Prima generazione-padronale ➔ %52.27
Pubblic Company ➔ %0.45
Successiva generazione con management ➔ %3.18
Successiva generazione senza management ➔ %0.45

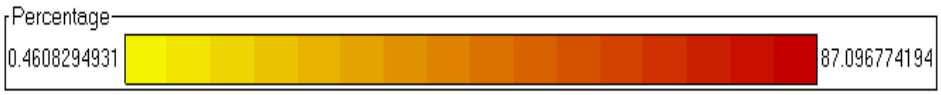




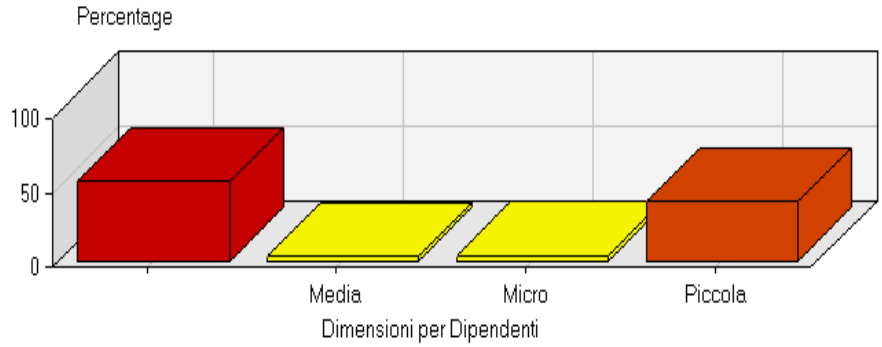
Export tramite terzi ➔ %1.36
Solo Italia ➔ %98.18
Solo Italia con potenzialità di export ➔ %0.45

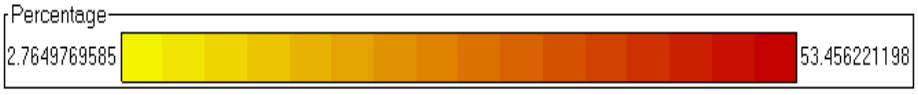
Cluster 4



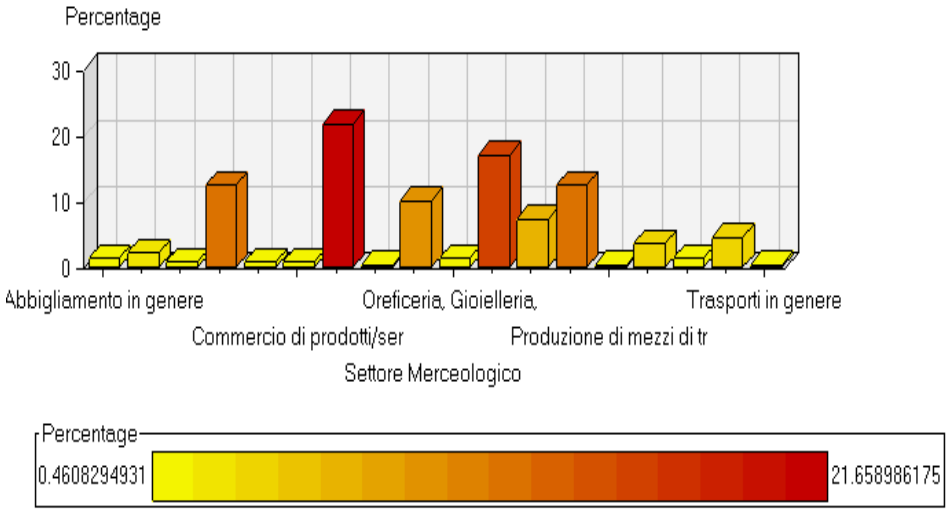


PD ➔ %87.10
PN ➔ %0.92
TV ➔ %1.38
UD ➔ %0.46
VE ➔ %4.61
VI ➔ %4.61
VR ➔ %0.92

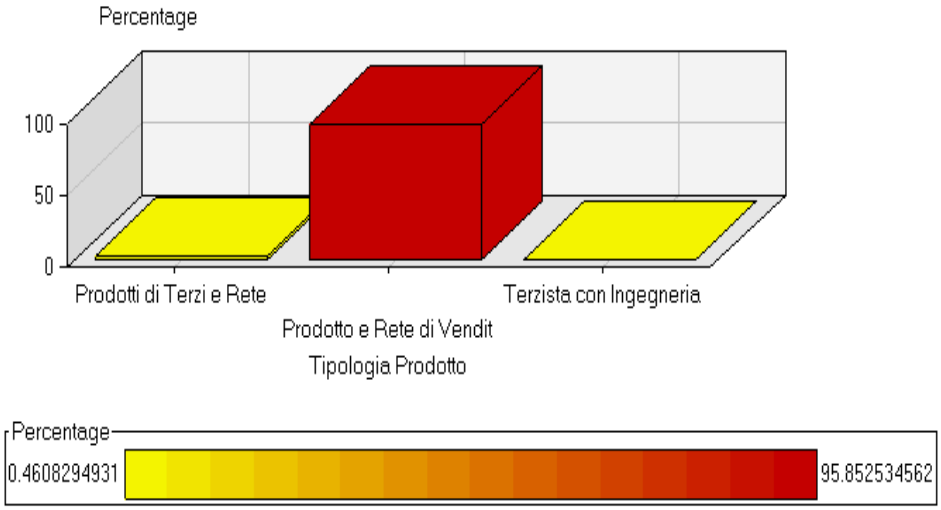




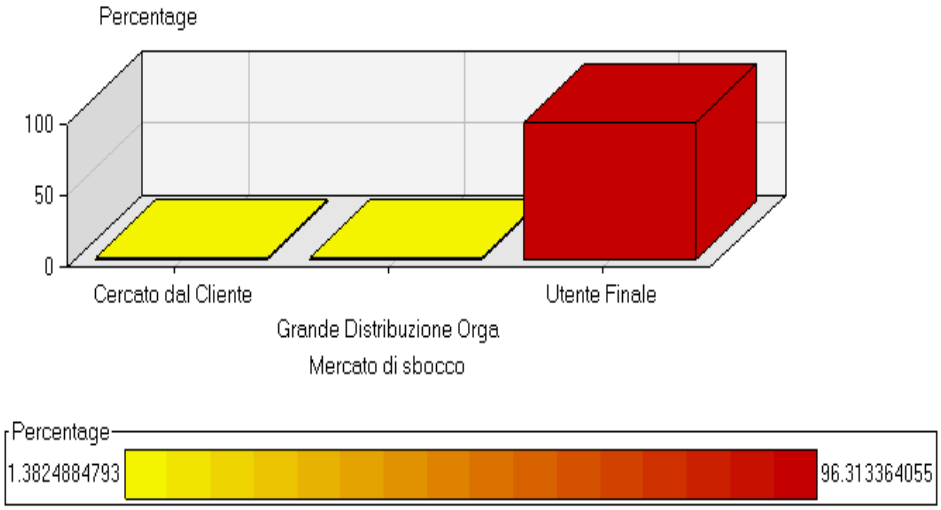
Media ➔ %5.94
Micro ➔ %7.92
Piccola ➔ %86.14



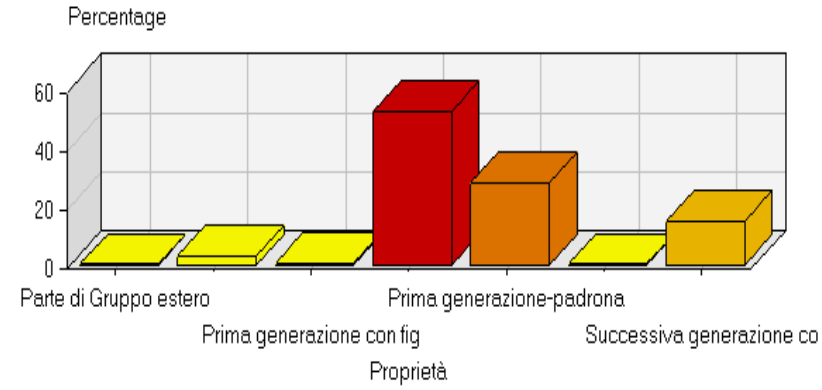
Abbigliamento in genere ➔ %1.38
Agricoltura, Floricoltura, Allevamento ➔ %2.30
Alberghi, Ristorazione, Turismo, Sport, Tempo libero, Spettacolo ➔ %0.92
Arredamento in generale ➔ %12.44
Attività sanitarie, Benessere, Bellezza, Assistenza privata, Organizzazioni no profit, Organizzazioni confessionali e politiche ➔ %0.92
Commercio di prodotti/servizi materiali ➔ %0.92
Costruzioni, Installazioni, Restauri, Cantieristica, Grande Engineering anche pluriennale ➔ %21.66
Credito, Assicurazioni, Leasing, Intermediazioni e Servizi Finanziari, Gestione Fondi e Pensioni ➔ %0.46
Macchine, Attrezzature meccaniche, elettromeccaniche, elettroniche per la casa e l'ufficio dell'utente finale e dell'azienda ➔ %10.14
Oreficeria, Gioielleria, Vetro, Ceramica, Gadget, Strumentazione musicale, sportiva, tempo libero, giocattoli ➔ %1.38
Prestazioni intellettuali, Consulenze, Produzione SW, Ricerca, Marketing, Istruzione, Pubblicità ➔ %17.05
Prodotti di largo consumo ➔ %7.37
Produzione di Componentistica ➔ %12.44
Produzione di mezzi di trasporto pubblico e privato ➔ %0.46
Produzione materie prime di base e semilavorati ➔ %3.69
Produzione/Trasformazione di prodotti secondari ➔ %1.38
Pulizie, Manutenzioni, Riparazioni, Noleggi, Vigilanza, Lavanderia, Disinfestazioni e servizi materiali vari ➔ %4.61
Trasporti in genere ➔ %0.46

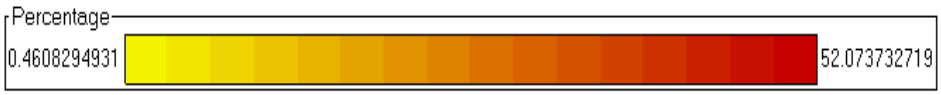


Prodotti di Terzi e Rete di Vendita Propri ➔ %3.69
Prodotto e Rete di Vendita Propri ➔ %95.85
Terzista con Ingegneria ➔ %0.46

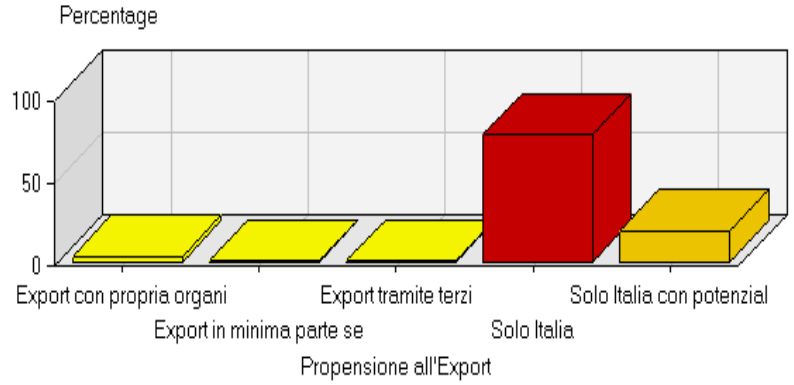


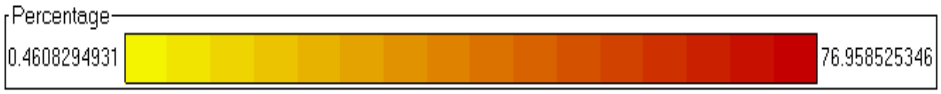
Cercato dal Cliente ➔ %2.30
Grande Distribuzione Organizzata ➔ %1.38
Utente Finale ➔ %96.31





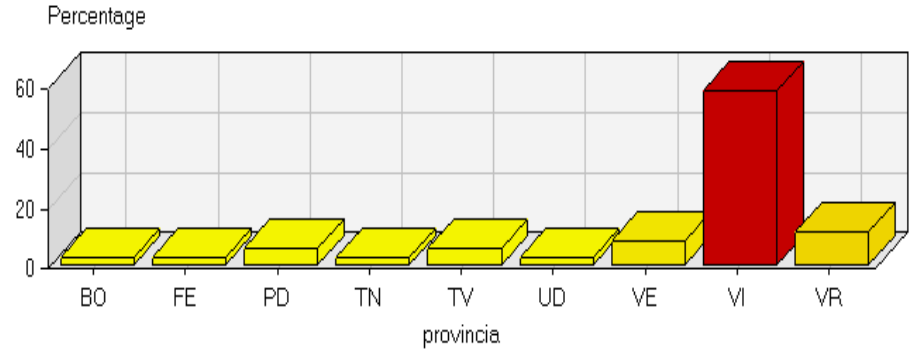
Parte di Gruppo estero ➔ %0.46
Parte di Gruppo nazionale ➔ %3.23
Prima generazione con figli operativi ➔ %0.92
Prima generazione organizzato ➔ %52.07
Prima generazione-padronale ➔ %28.11
Pubblic Company ➔ %0.46
Successiva generazione con management ➔ %14.75

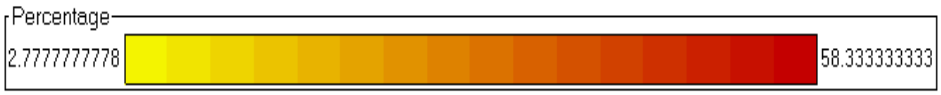




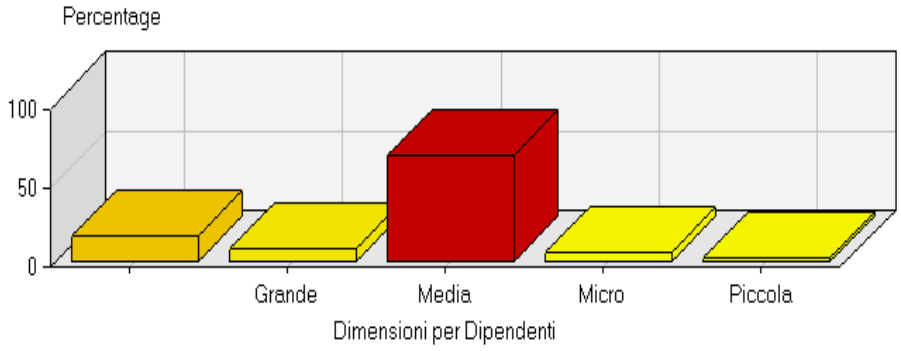
Export con propria organizzazione ➔ %3.23
Export in minima parte senza propria organizzazione ➔ %0.46
Export tramite terzi ➔ %0.46
Solo Italia ➔ %76.96
Solo Italia con potenzialità di export ➔ %18.89

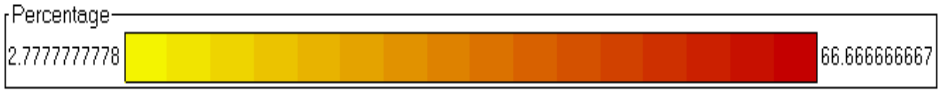
Cluster 5



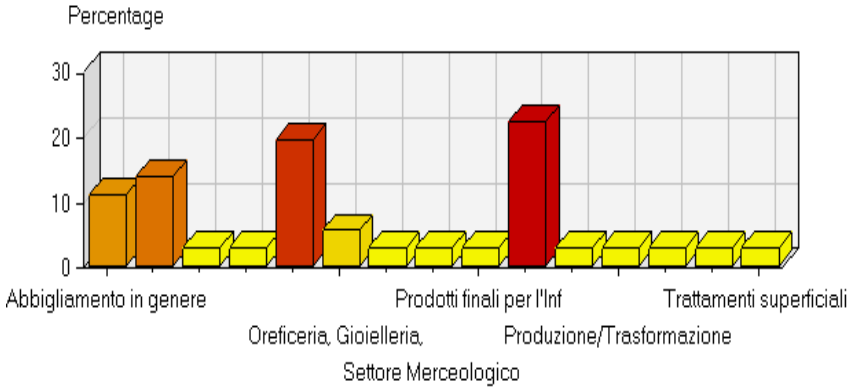


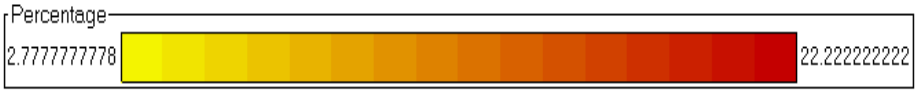
BO ➔ %2.78
FE ➔ %2.78
PD ➔ %5.56
TN ➔ %2.78
TV ➔ %5.56
UD ➔ %2.78
VE ➔ %8.33
VI ➔ %58.33
VR ➔ %11.11



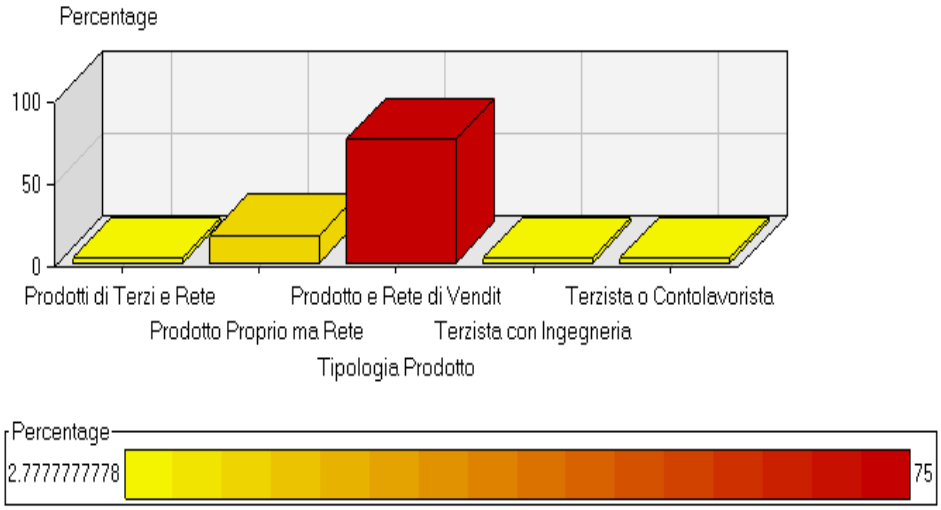


Grande ➔ %10.00
Media ➔ %80.00
Micro ➔ %6.67
Piccola ➔ %3.33

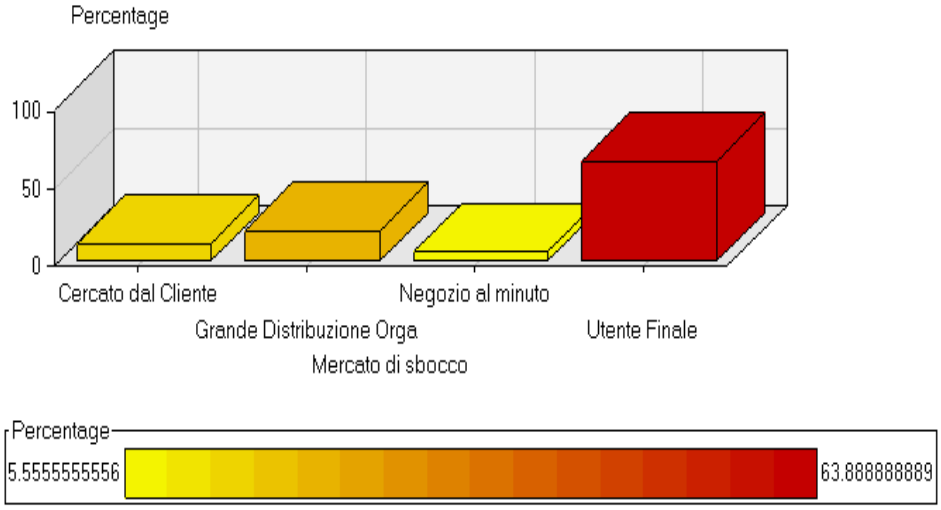




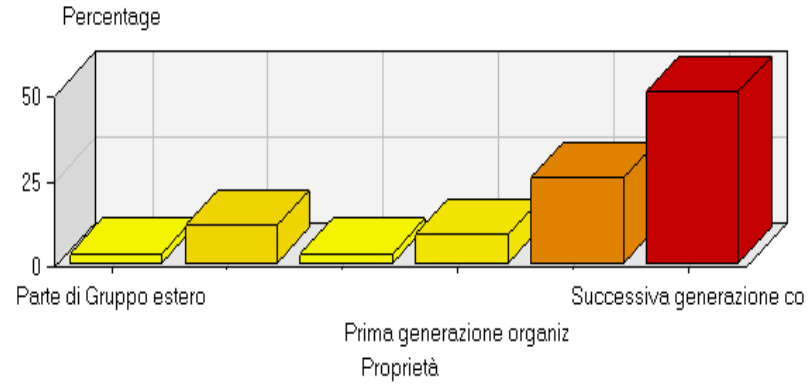
Abbigliamento in genere ➔ %11.11
Arredamento in generale ➔ %13.89
Commercio di prodotti/servizi materiali ➔ %2.78
Costruzioni, Installazioni, Restauri, Cantieristica, Grande Engineering anche pluriennale ➔ %2.78
Macchine, Attrezzature meccaniche, elettromeccaniche, elettroniche per la casa e l'ufficio dell'utente finale e dell'azienda ➔ %19.44
Oreficeria, Gioielleria, Vetro, Ceramica, Gadget, Strumentazione musicale, sportiva, tempo libero,
giocattoli ➔ %5.56
Prestazioni intellettuali, Consulenze, Produzione SW, Ricerca, Marketing, Istruzione, Pubblicità ➔ %2.78
Prodotti di largo consumo ➔ %2.78
Prodotti finali per l'Informazione ➔ %2.78
Produzione di Componentistica ➔ %22.22
Produzione materie prime di base e semilavorati ➔ %2.78
Produzione/Trasformazione di prodotti secondari ➔ %2.78
Pulizie, Manutenzioni, Riparazioni, Noleggi, Vigilanza, Lavanderia, Disinfestazioni e servizi materiali vari ➔ %2.78
Trasporti in genere ➔ %2.78
Trattamenti superficiali per terzi ➔ %2.78

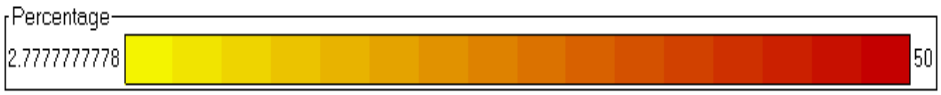


Prodotti di Terzi e Rete di Vendita Propri ➔ %2.78
Prodotto Proprio ma Rete di Vendita di Terzi ➔ %16.67
Prodotto e Rete di Vendita Propri ➔ %75.00
Terzista con Ingegneria ➔ %2.78
Terzista o Contolavorista ➔ %2.78

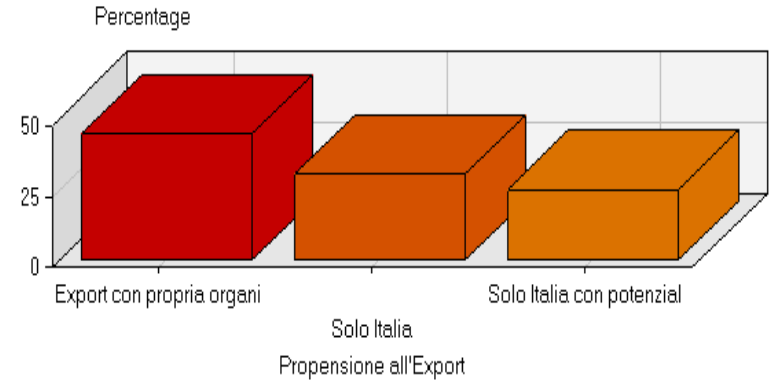


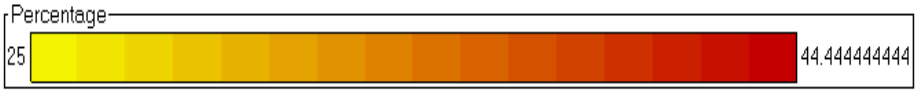
Cercato dal Cliente ➔ %11.11
Grande Distribuzione Organizzata ➔ %19.44
Negozio al minuto ➔ %5.56
Utente Finale ➔ %63.89





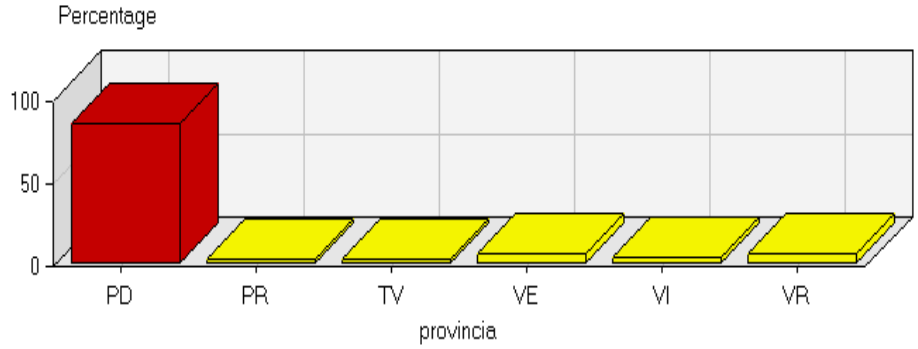
Parte di Gruppo estero ➔ %2.78
Parte di Gruppo nazionale ➔ %11.11
Prima generazione con figli operativi ➔ %2.78
Prima generazione organizzato ➔ %8.33
Prima generazione-padronale ➔ %25.00
Successiva generazione con management ➔ %50.00

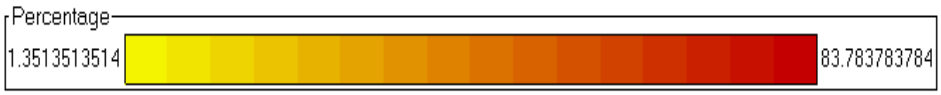




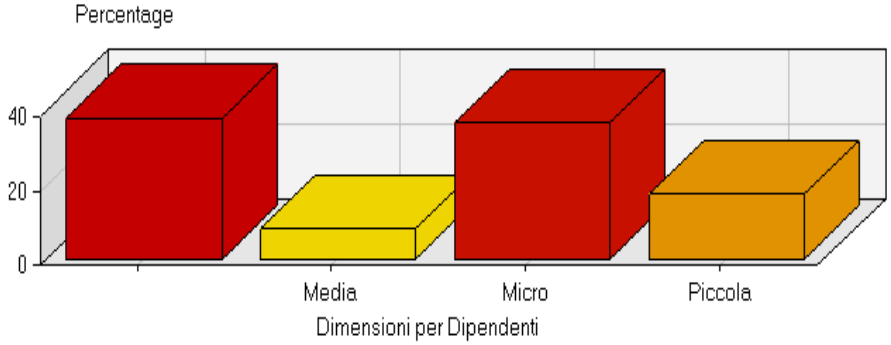
Export con propria organizzazione ➔ %44.44
Solo Italia ➔ %30.56
Solo Italia con potenzialità di export ➔ %25.00

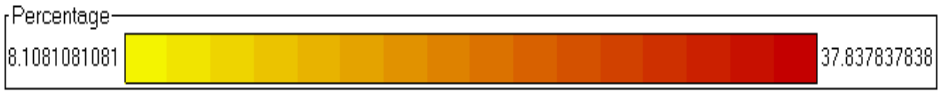
Cluster 6



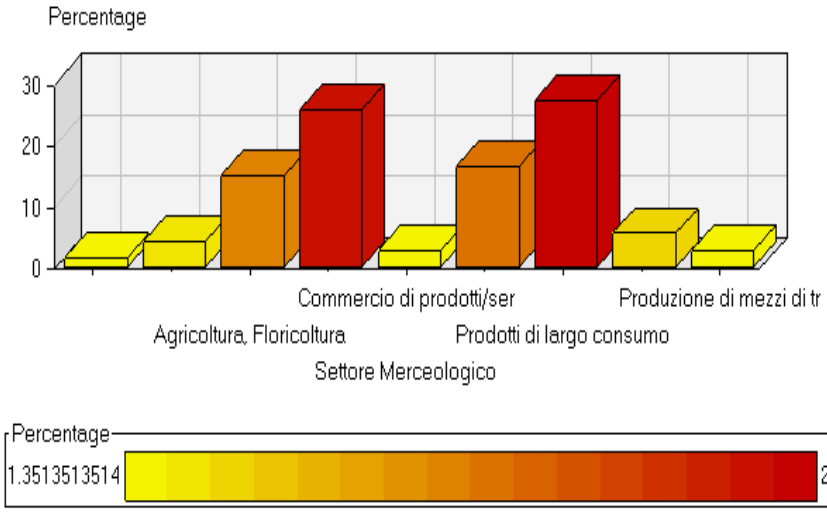


PD ➔ %83.78
PR ➔ %1.35
TV ➔ %1.35
VE ➔ %5.41
VI ➔ %2.70
VR ➔ %5.41

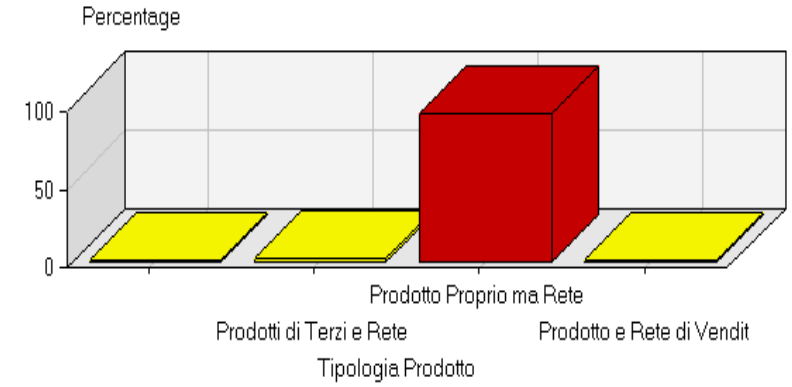


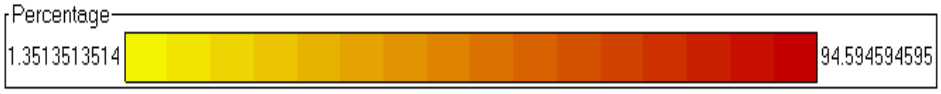


Media ➔ %13.04
Micro ➔ %58.70
Piccola ➔ %28.26

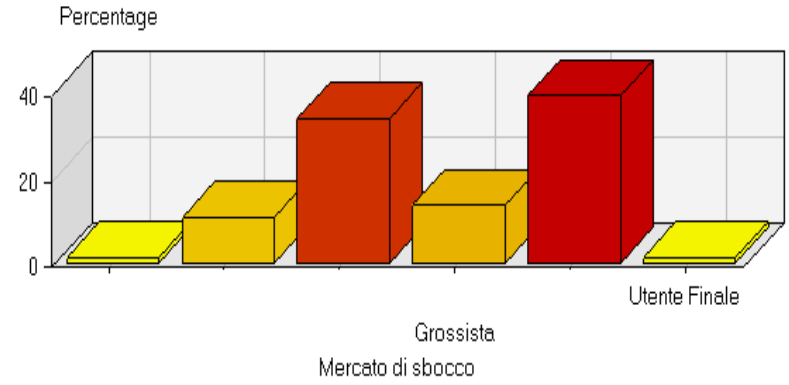


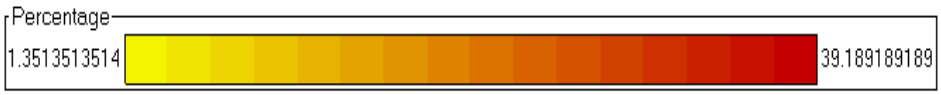
Abbigliamento in genere ➔ %4.11
Agricoltura, Floricoltura, Allevamento ➔ %15.07
Arredamento in generale ➔ %26.03
Commercio di prodotti/servizi materiali ➔ %2.74
Macchine, Attrezzature meccaniche, elettromeccaniche, elettroniche per la casa e l'ufficio dell'utente finale e dell'azienda ➔ %16.44
Prodotti di largo consumo ➔ %27.40
Produzione di Componentistica ➔ %5.48
Produzione di mezzi di trasporto pubblico e privato ➔ %2.74



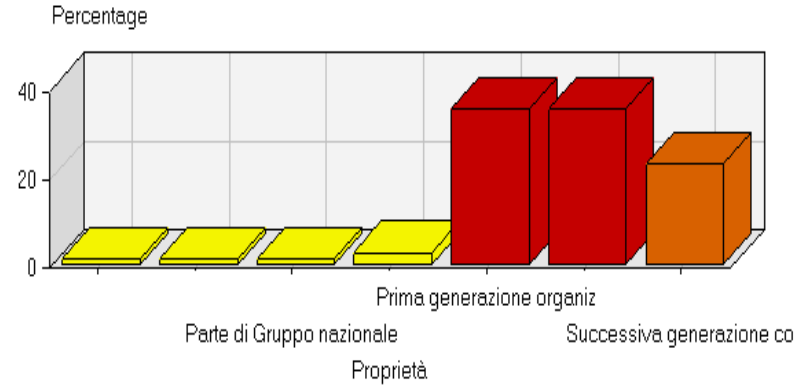


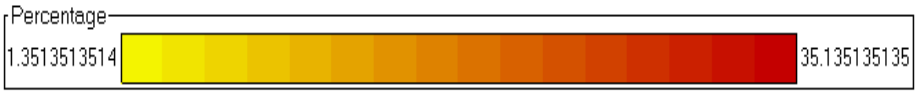
Prodotti di Terzi e Rete di Vendita Propri ➔ %2.74
Prodotto Proprio ma Rete di Vendita di Terzi ➔ %95.89
Prodotto e Rete di Vendita Propri ➔ %1.37



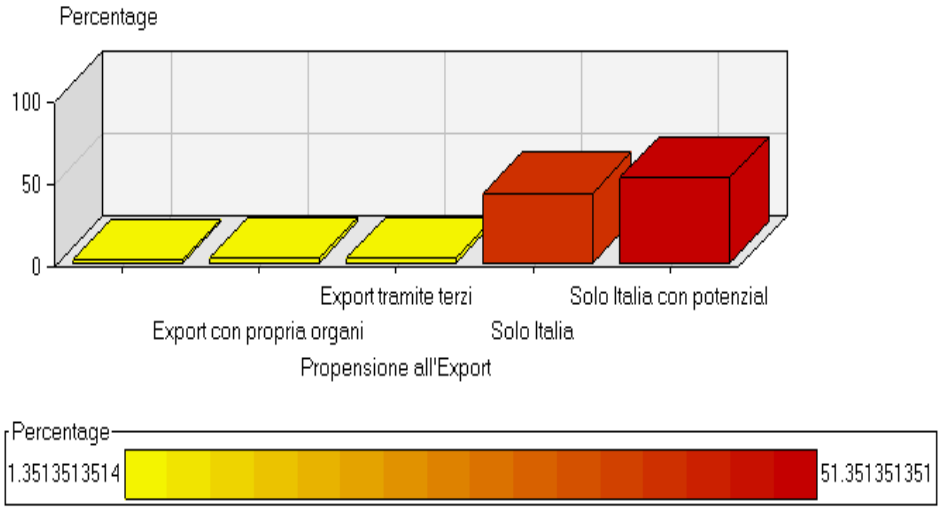


Cercato dal Cliente ➔ %10.96
Grande Distribuzione Organizzata ➔ %34.25
Grossista ➔ %13.70
Negozio al minuto ➔ %39.73
Utente Finale ➔ %1.37





Parte di Gruppo estero ➔ %1.37
Parte di Gruppo nazionale ➔ %1.37
Prima generazione con figli operativi ➔ %2.74
Prima generazione organizzato ➔ %35.62
Prima generazione-padronale ➔ %35.62
Successiva generazione con management ➔ %23.29



Export con propria organizzazione ➔ %2.74
Export tramite terzi ➔ %2.74
Solo Italia ➔ %42.47
Solo Italia con potenzialità di export ➔ %52.05

Una volta esplorate tutte le variabili statistiche di interesse per tutti e sei i cluster bisogna dare una descrizione generale per ogni gruppo.

Cluster 1: **terzisti o controlavoristi e lavanderie ➔** sono il gruppo più numeroso, ne fanno parte i terzisti o controlavoristi e le lavanderie, che si caratterizzano dal fatto di operare solo in Italia e che per il loro lavoro sono cercati dal cliente, sono distribuiti in gran parte nel padovano ma coprono anche città come Vicenza e Venezia. La aziende contenute in questo cluster non sono di grandi dimensioni, al massimo 50 dipendenti, con una gestione padronale o organizzata.

Cluster 2: **autotrasportatori ➔** le aziende all'interno di questo cluster si distribuiscono quasi in modo uniforme in tutte le province del veneto, con esclusione di Rovigo e Belluno, riguardano il settore trasporti in generale, quindi sono autotrasportatori che operano in Italia ma anche all'estero con propria organizzazione, sono cercati dal cliente perché offrono un servizio ma il prodotto e la rete di vendita sono del cliente.

Cluster 3: **industria classica (micro), scuole non di PD e commercianti ➔** questo cluster contiene aziende che operano in vari settori, dalle imprese edili alle scuole passando per quelle che producono arredamento in generale e commercianti, sono comunque tutte aziende micro (fino a 10) per quanto riguarda il numero di dipendenti, vendono un prodotto proprio ed hanno una propria rete di vendita, dunque si rivolgono direttamente all'utente finale operando solo in Italia.

Cluster 4: **industria classica (piccola) e scuole padovane ➔** in questo cluster sono contenute aziende padovane con un numero di dipendenti che varia da 10 a 50, che operano nel settore dell'industria classica. Offrono un prodotto proprio e hanno una rete di vendita propria, dunque si rivolgono all'utente finale. Operano in gran parte in Italia con una piccola percentuale di aziende con potenzialità di export. Le scuole operano solo in Italia.

Cluster 5: **imprese medie ➔** questo cluster si differenzia dagli altri in particolar modo perché contiene aziende che operano nella provincia di Vicenza e che sono di media dimensione, dal punto di vista del numero di dipendenti, e operanti nell'industria classica offrendo al mercato un prodotto proprio e gestendo una propria rete di vendita. Si rivolgono all'utente finale e sono gestite da manager o in modo padronale. Si caratterizzano dal fatto che operano in Italia e all'estero con propria organizzazione.

Cluster 6: **industria classica e aziende agricole ➔**questo cluster contiene aziende del padovano non di grandi dimensioni, al massimo 50 dipendenti, che operano nell'industria classica con produzione di componenti, di macchine, attrezzature meccaniche, elettromeccaniche, elettroniche e di prodotti alimentari, ma contiene anche aziende agricole. Non gestiscono direttamente la vendita del proprio prodotto ma si affidano alla grande distribuzione organizzata o negozi al minuto. Operano in Italia ma con potenzialità di export, magari tramite terzi.

Conclusioni:

Il risultato finale è stata l'individuazione di sei clusters ognuno dei quali contiene aziende con un comportamento simile. Questo comportamento è stato determinato da me sulla base delle statistiche risultanti dall'elaborazione dei dati effettuata con Enterprise Miner™.

L'azienda potrà usufruire di questi clusters per capire quali potenziali clienti contattare in base al servizio da proporre in modo da avere una risposta positiva.

L'esperienza di stage mi ha permesso di utilizzare Enterprise Miner™ e di capirne meglio il funzionamento utilizzando dati reali.

BIBLIOGAFIA:

- **Marketing Management** – R. Winer: Utilizzato per descrivere l'orientamento al cliente.

- **Concetti e strumenti di marketing** - Roberto Grandinetti: utile per la parte che tratta della segmentazione della clientela, dei criteri e dei requisiti e delle strategie di segmentazione.

- **Metodi di Data Mining per il customer relationship management** – Nicola Del Ciello, Susi Dulli, Alberto Saccardi: Utilizzato per descrivere le tecniche di Data Mining e le applicazioni, per il paragrafo della Cluster Analysis e per la metodologia SEMMA.

- **Data Mining** – Berry, Linoff: Utilizzato per la parte che descrive il ciclo di vita del cliente, il Data Mining e tutto il paragrafo dei dati.

- **Analisi dei dati e Data Mining** – A. Azzalini, B. Scarpa: Utilizzato per la parte che descrive la Cluster Analysis, le distanze tra due elementi e gli algoritmi di classificazione.

- **Tecniche di Data Mining con il software SAS® Enterprise Miner™** – Scritto da Sabina Silani e Marina Tarantino: Utilizzato per la parte di SAS e per la realizzazione del progetto.

- Appunti e lucidi dei corsi **Sistemi Informativi Aziendali** per le fasi del processo di classificazione e per quello di Data Mining ed **Economia e Gestione delle Imprese II** per la segmentazione del mercato e rispettive immagini.