Client Segmentation e Market Basket Analysis:

Costruzione di un modello predittivo applicato al sistema portuale Import/Export

Mattia Biggi
mat: 412130

Relatore
Prof.ssa Anna Monreale
Tutor Aziendale
Dott. Gianluca Schiavina

UNIVERSITÀ DI PISA

Corso di Laurea Magistrale in Informatica
Università degli Studi di Pisa

Anno Accademico

2015/2016

Indice

Elenco delle figure

Prefazione

Il lavoro svolto consiste nell'utilizzo di strategie di Data Mining, attualmente impiegate in contesti come mobilità, telecomunicazione e marketing al mondo dell'Import/Export. Questo contesto è costituito da tutte le procedure doganali e fiscali necessarie per esportare ed importare merci, in vigore secondo gli standard italiani e comunitari.
Emerge in questo contesto la figura dello **Spedizioniere Doganale**, figura professionale nel campo del trasporto merci (impersonata sia un singolo individuo che una organizzazione), definita anche *Architetto del Trasporto*. Il suo compito principale è: *"evadere la necessità di un cliente di trasferire i materiali acquistati o venduti, spesso da e verso l'estero, aiutandolo nell'evasione di tutte le pratiche doganali e fiscali necessarie, nonché nel reperimento dei mezzi di trasporto più idonei* [14]".
Nel caso specifico si è presa in considerazione un'azienda di grandi dimensioni che opera in questo campo. Il traffico generato è molto elevato e tutte le operazioni effettuate sono fatturate al cliente con un determinato prezzo ed un identificativo univoco. Questo aspetto è stato associato al mondo del marketing ed al modello Market Basket, considerando tutte le voci fatturate per conto di un cliente come l'effettiva lista dei suoi acquisti. Lo scopo è quello di costruire un modello in grado di prevedere il traffico futuro di un cliente, in base al comportamento assunto in un particolare periodo di tempo arbitrario.

Capitolo 1

Introduzione

La rapida evoluzione delle tecnologie informatiche, in particolare il costo sempre più economico della memoria fisica per l'archiviazione dati e l'aumentare della potenza di calcolo, procedono di pari passo con l'espansione della rete e la digitalizzazione delle più comuni procedure.

Una delle conseguenze dirette è la continua espansione delle infrastrutture informatiche ed un crescente afflusso di clienti altamente dinamici, i quali lasciano numerose tracce delle loro abitudini e preferenze. Queste informazioni sono sparse ovunque in qualsiasi banca dati di qualsiasi azienda che opera nel settore. La disponibilità di questi dati ha favorito negli ultimi decenni lo studio di tecniche di **Data Mining**, che stanno acquisendo sempre più importanza e sono usate spesso per estrarre informazioni aggiuntive su cui basare opprtune strategie commerciali [6]. Il fenomeno descritto ha interessato particolarmente il mondo dell'Import/Export. Le pratiche fiscali e doganali stanno tutt'ora evolvendo verso la completa digitalizzazione e le aziende che operano nel settore, devono adeguare rapidamente i loro strumenti. Le case di spedizionieri [14], hanno digitalizzato rapidamente tutte le procedure fiscali e doganali, e tutta la documentazione inerente lo sdoganamento delle merci. Oltre ad una maggiore efficienza in termini di organizzazione e soprattutto tempistiche di esecuzione, le banche dati delle aziende hanno iniziato a popolarsi di informazioni riguardanti il traffico effettuato dalla propria clientela. Queste informazioni contengono dati statistici sul fatturato al cliente e sui servizi di cui usufruisce, ma non solo, infatti la presenza di uno storico delle fatture emesse comprensivo di data, tipo di servizio e prezzo, permette l'applicazione di numerose tecniche di Data Mining, al fine di classificare i clienti in base alle loro abitudini o preferenze oppu-

re prevedere il loro comportamento basandosi sui loro movimenti trascorsi. In sostanza questa evoluzione ha reso il contesto Import/Export, analogo al contesto del Marketing nel quale ogni cliente lascia una traccia evidente delle sue abitudini e preferenze commerciali.
L'obiettivo del progetto è quello di analizzare il comportamento della clientela dell'azienda Ser.Nav s.r.l. (Agenzia Marittima che opera nei porti di La Spezia, Genova e Livorno, Capitolo 3), applicando determinate tecniche di Data Mining, allo scopo di individuare similitudini tra i clienti e sfruttare sia il comportamento individuale che quello collettivo, per proporre possibili previsioni sui futuri servizi richiesti, permettendo all'azienda una maggiore precisione nella costruzione di un piano tariffario personalizzato ed adeguato alle esigenze del cliente.

Struttura della Tesi

Il documento è strutturato in tre parti principali. La prima parte è composta dai Capitoli 2 e 3 e consiste in una fase introduttiva dove è descritto il contesto sia dal punto di vista accademico che aziendale. Nel Capitolo 2 si discutono le principali metodologie e tecniche presenti in letteratura, con particolare approfondimento di quelle che sono state rilevanti per la realizzazione dell'intero progetto. Il Capitolo 3 descrive la realtà aziendale di riferimento, esponendo nel dettaglio le necessità aziendali, la natura delle informazioni inerenti il traffico della clientela e le conseguenti procedure di elaborazione dati attuate per l'estrazione delle informazioni.
La seconda parte riguarda la descrizine del processo analitico costruito e comprende i Capitoli 4 e 5. Al loro interno si espone la soluzione proposta per soddisfare le necessità aziendali, mostrando nel dettaglio il processo di Data Mining messo in atto, il formato dell'input e le relative strutture di cui è stata necessaria la creazione, per concludere con il processo di validazione adottato per verificarne l'efficienza.
L'ultima parte comprende i capitoli 6 e 7, in essa è mostrato l'utilizzo ed i risultati ottenuti, per concludere proponendo alcune possibili estensioni del progetto. Nel Capitolo 6, sono mostrate le interfacce software costruite per facilitare l'utilizzo e la manutenzione del progetto, il tutto seguito dai risultati ottenuti nel processo di validazione. Infine il Capitolo 7 propone le strategie di Data Mining che possono essere introdotte e le possibili estensioni di quelle attuate.

Capitolo 2

Stato dell'Arte

In questo capitolo verrà trattata la letteratura relativa ai contenuti sviluppati. Saranno illustrate le principali tecniche di Data Mining in uso per affrontare sia **Customer Segmentation** che **Predictive Analytics**, con particolare riferimento al concetto di Clustering ed all'utilizzo di Regole d'Associazione, in quanto sono alla base della strategia utilizzata in questo progetto.
In seguito saranno illustrati i principali contesti applicativi: *Mobilità*, *Telecomunicazioni* e *Decision Making*.

2.1 Data Mining

Il Data Mining consiste in tutte quelle che sono le metodologie per l'estrazione di informazioni o comportamenti particolari avendo a disposizione una grande quantità di dati e la successiva interpretazione dei risultati.

Tale processo si affianca all'analisi statistica, ma differentemente da essa, non estrae informazioni generali, lo scopo è la ricerca delle correlazioni tra le categorie o tra gli individui che costituiscono la base dati analizzata [7].
Il Data Mining ha una duplice valenza:

- Utilizzo di tecniche analitiche per l'estrazione di informazioni implicite da dati strutturati, in modo da poter essere elaborate ed utilizzate nel modo più efficiente per operazioni di classificazione, predizione, segmentazione.

- Ricerca di pattern significativi o comportamenti ricorrenti per mezzo di procedure automatiche o semi-automatiche sempre su grandi quantità di dati.

Il Data Mining possiede un ruolo tanto importante quanto nuovo nell'ambito del marketing. E' di fondamentale importanza tenere conto che tali tecniche non ambiscono a superare la valutazione degli esperti che operano nel settore, ma ad agevolare il loro lavoro. Gli algoritmi che verranno illustrati hanno molto potenziale ma non possono funzionare al pieno delle loro capacità senza una figura, esperta del dominio, in loro supporto soprattutto per quanto riguarda la validazione dei risultati.
Tuttavia questa attività è utilizzata non solo in ambito commerciale, ma anche nella ricerca scientifica, nell'ottimizzazione di siti web, nella rilevazione di comportamenti fraudolenti [8].

2.2 Customer Analytics

Il processo in considerazione prende il nome di Customer Analytics. Consiste nell'utilizzare i dati relativi al **comportamento dei clienti** per ricavare delle decisioni di marketing attraverso i processi di **Market Segmentaton** e **Predictive Analytics**. Queste informazioni sono utilizzate per la gestione di offerte e delle relazioni con i clienti [9].

2.2.1 Customer Behavior

Consiste nello studio di individui, gruppi o organizzazioni e dei processi che essi usano. Lo scopo è quello di selezionare prodotti, servizi, idee, nel tentativo di capire il loro processo decisionale. Questa conoscenza permette di raffinare le strategie e le politiche di marketing per soddisfare al meglio i bisogni dei compratori. Tale studio è totalmente incentrato sul comportamento del cliente il quale svolge tutte le funzioni dalla ricerca, alla scelta e all'acquisto del prodotto. Per questi motivi si tratta di un campo molto eterogeneo e di difficile analisi, sia per quanto riguarda la conoscenza del dominio che la conoscenza delle tecnologie e metodologie informatiche e economiche necessarie.

2.2.2 Market Segmentation

Market Segmentation è una strategia di marketing che consiste nella suddivisione di un ampio insieme di individui, in sottoinsiemi legati da comuni proprietà. Questa strategia comprende anche tutto il processo di individuazione e valutazione delle caratteristiche condivise.

Le numerose strategie di segmentazione sono usate generalmente per identificare e categorizzare i clienti, creando profili confrontabili e permettendo l'attuazione di procedure di marketing per meglio assecondare le esigenze dei clienti.

Le proprietà comuni su cui basare il processo di Market Segmentation sono di molteplice natura:

- **Geografica**
  Suddivisione effettuata secondo criteri geografici come stato di appartenenza, regione, città, codice postale. Lo scopo è quello di categorizzare i clienti in base al territorio, per tutte quelle tipologie di business che sono influenzate da fattori come le condizioni climatiche, la presenza di grandi centri abitati o prossimità rispetto ai punti di vendita.

- **Demografica**
  Questa suddivisione si basa su elementi demografici come età, sesso, religione, occupazione, educazione.

- **Comportamentale**
  Ritenuta una delle migliori basi di partenza per la suddivisione dei clienti, la strategia comportamentale genera gruppi in base alle abitudini commerciali, dal tipo di servizi utilizzati, alla tipo di beni acquistati.

- **Psicografica**
  Si prendono in considerazione le opinioni del cliente ed i suoi interessi e soprattutto come essi vengono influenzati da fattori esterni, questa analisi è importante perché identifica lo stile di vita dell'individuo.

- **Occasionale**
  Suddivisione basata su fattori occasionali, considerando circostanze particolari si cerca di raggruppare i clienti che possono essere soggetti ad eventi casuali, si ottengono informazioni riguardo ai bisogni del cliente ed alle sue reazioni. Questo approccio non è assoluto, ogni cliente può far parte di differenti insiemi in base alle circostanze a cui è soggetto.

- **Culturale**
  Classificazione dei clienti in base alle origini culturali, misura molto influente nella vita del singolo individuo. La suddivisione basata sulle origini culturali permette di creare strategie commerciali che rispettano esattamente le esigenze di particolari popolazioni.

- **Multi-Attributo**
  Sono possibili anche approcci ibridi, come la combinazione di dati geografici e culturali, oppure combinazioni psicografiche e demografiche. Ogni dominio è generalmente influenzato da molteplici fattori e di conseguenza la classificazione dei clienti trae maggiore beneficio da questo tipo di strategie.

2.2.3 Predictive Analytics

L'analisi predittiva comprende una varietà di strategie sia statistiche che tecniche come la modellazione predittiva, machine learning e data mining che analizzano i fatti attuali e storici per fare previsioni su eventi futuri.

I modelli predittivi sfruttano i pattern individuati dalle transazioni passate e catturano relazioni tra molteplici attributi, l'analisi predittiva fornisce un indice di predizione, espresso generalmente come probabilità, per ogni individuo del dominio. Lo scopo è quello di determinare, informare o influenzare i processi decisionali, permettendo di contenere rischi o individuare potenziali nascosti sia positivi che negativi.

2.2.4 Customer Relationship Management

Customer relationship management (CRM) consiste nella gestione delle interazioni tra le compagnie e la loro clientela, considerando sia i clienti posseduti che i potenziali clienti futuri. Esistono tre tipi di CRM:

- **Operativo**: soluzioni metodologiche e tecnologiche per automatizzare i processi di business che prevedono il contatto diretto con il cliente.

- **Analitico**: procedure e strumenti per migliorare la conoscenza del cliente attraverso l'analisi e la revisione dei dati riguardanti il suo comportamento.

- **Collaborativo**: metodologie e tecnologie integrate con gli strumenti di comunicazione per gestire il contatto con il cliente.

2.3 Tecniche di Segmentazione

Con Customer Segmentation si intende il processo intento a dividere i clienti in distinte, significative ed omogenee categorie basate sia su una singola proprietà che sulla combinazione di molteplici. Nell'ambito del marketing lo scopo è quello di differenziare i clienti permettendo di applicare le migliori strategie ed i modelli più idonei.

Come principale tecnica di segmentazione, vedremo nel dettaglio il **Clustering**. In seguito saranno descritti i principali algoritmi spiegando sia i punti di forza che i punti deboli, per poi trattare il processo di validazione.

2.3.1 Clustering

L'obiettivo della clusterizzazione è di organizzare gli oggetti esaminati in gruppi, i quali condividono proprietà simili. Il Clustering si può considerare uno dei più importanti **metodi di apprendimento non supervisionato** [13]. Come ogni metodo appartenente a questa categoria, non fa uso di identificatori determinati a priori per intuire la possibile struttura dei dati. Un cluster può essere definito come una collezione di oggetti simili tra loro e dissimili da elementi negli altri cluster.

Esistono varie classificazioni delle tecniche di clustering, una prima categorizzazione dipende dalla possibilità che un elemento possa o meno essere assegnato a più cluster [7]:

- **Clustering Esclusivo**: Ogni elemento può appartenere solamente ad un cluster, ossia le intersezioni tra i clusters sono sempre insiemi vuoti, questa procedura prende anche il nome di *Hard Clustering*.

- **Clustering Inclusivo**: Ogni elemento può appartenere a più cluster contemporaneamente, con un indice che decreta il grado di appartenenza ad ogni cluster, procedura che prende il nome di *Soft* o *Fuzzy Clustering*.

Un'altra tecnica consiste nella tipologia di suddivisione dello spazio:

- **Clustering Partizionale**: Si utilizza il concetto di **distanza** tra gli elementi, i quali appartengono ad un particolare gruppo in base alla loro relazione con un punto significativo del dataset.

- **Clustering Gerarchico**: Si costruisce una gerarchia di partizioni, costruita sia per aggregazione che per divisione, mediante una rappresentazione ad albero che prende il nome di *Dendogramma*.

Esistono altre suddivisioni per quanto rigurada il Clustering Partizionale, più dettagliate, le quali si differenziano per la valutazione della distanza tra gli elementi e la relativa creazione dei cluster:

- *Well-Separated Clusters*: Ogni punto all'interno del cluster è più simile agli elementi del suo gruppo rispetto a qualsiasi altro elemento che non appartiene al cluster.

- *Center-Based Clusters*: Mantiene le proprietà dei cluster Well-Separeted, ma la similitudine è decretata rispetto ad un punto rappresentativo del cluster, spesso si considera il **Centroide** oppure il **Medoide**.

- *Contiguous Clusters*: La composizione del cluster è basata sul concetto di *contiguità*, ossia una volta stabilita la distanza tra gli elementi, si considerano come cluster tutti gli insiemi contigui che risultano separati tra di loro.

- *Density Based*: Si identifica come cluster ogni regione molto densa di elementi mappati secondo una particolare funzione di distanza. In questa tipologia di clustreing esiste il concetto di *Noise* che, come vedremo nel dettaglio in seguito, include la possibilità che non tutti gli elementi siano assegnati ad un cluster.

- *Objective Function*: Si cercano cluster che massimizzano o minimizzano una particolare funzione obiettivo, si enumerano tutti i modi possibili per clusterizzare i punti e si valuta la qualità della soluzione per ogni risultato rispetto ad una funzione obiettivo, questa procedura viene affrontata spesso euristicamente in quanto risulta NP-Hard.

In seguito sono descritte le principali strategie di Clustering ed algoritmi utilizzati.

Clustering Basato su Centroidi

Il clustering basato su centroidi è di tipo partizionale e ogni cluster è rappresentato da un prototipo chiamato *centroide* che tipicamente è la media tra

le distanze dei punti del cluster. Uno dei più famosi algoritmi di clustering appartenenti a questa categoria è il K-Means che richiede di specificare il numero $\boldsymbol{K}$ di cluster che si vogliono ottenere (conferendo il nome K-Means). L'algoritmo iterativamente elegge i K centroidi del cluster, ed ogni elemento viene associato al centroide più vicino. L'algoritmo è il seguente:

Algorithm 1 K-Means

```
1: function K-Means(clusters K)
2:     Elezione K Centroidi
3:     repeat
4:         Assegnamento di ogni elemento al punto K piú vicino
5:         Ricalcolo dei K Centroidi
6:     until I Centroidi non variano
```

Inizialmente i centroidi vengono scelti randomicamente, mentre nelle iterazioni successive dell'algoritmo essi consistono tipicamente nella media tra le distanze dei punti del cluster. Esistono differenti metodologie per calcolare tale distanza: *Distanza Euclidea*, *Cosine Similarity*, *Correlazione*. L'algoritmo converge per le misure di similitudine elencate, tale convergenza si manifesta principalmente nelle prime iterazioni, seguite da una fase di assestamento, infatti spesso la condizione di stop viene rilassata, ammettendo una soglia minima di cambiamento tra i centroidi.
La scelta dei centroidi è una fase molto sensibile, infatti vengono applicate le seguenti tecniche per risolvere, anche se non completamente, il problema:

- Si eseguono molteplici esecuzioni, stimando i centroidi in modi differenti oppure semplicemente randomicamente, in seguito si valuta la qualità del risultato ottenuto, per mezzo degli strumenti di validazione che saranno descritti in seguito.

- Si utilizza la procedura di *Clustering Gerarchico* (spiegata nella sezione successiva), si eseguono K suddivisioni e si calcolano i centroidi dei cluster ottenuti, questi saranno i punti di partenza per l'algoritmo K-Means.

- Si stima un numero di centroidi $N > K$, e vengono considerati solamente i K migliori.

- Tecniche di postprocessing, come eliminazione di piccoli clusters, unione di cluster molto simili tra di loro e suddivisione di cluster troppo grandi.

- Si utilizza l'algoritmo *Bisecting K-Means*, esso consiste in un approccio gerarchico attraverso il quale partendo da un unico cluster, si suddivide tramite algoritmo **2-Means** un numero arbitrario di volte, si prende l'iterazione che ha prodotto i migliori cluster e si applica ricorsivamente l'algoritmo, a ciascun cluster scelto fino a che non si ottengono i **K** cluster desiderati.

La complessità dell'algoritmo è $O(n*K*I*d)$ dove n è il numero di punti, K il numero di cluster, I il numero di iterazioni e d il numero di attributi su cui si basa la funzione per il calcolo della distanza utilizzata.

In conclusione l'algoritmo K-Means presenta difficoltà nella gestione di dati la cui presenza di outliers è troppo elevata, infatti sono spesso eseguite procedure di *Preprocessing* per attenuare la problematica, inoltre, come detto in precedenza la scelta dei centroidi è spesso problematica, soprattutto quando si ha a che fare con dati ad elevata densità. Tuttavia K-Means risulta uno degli algoritmi più utilizzati soprattutto per quanto riguarda il problema della Customer Segmentation affrontato in questo progetto.

Clustering Gerarchico

Il clustering gerarchico produce un set di cluster annidati tra loro, organizzati come un albero gerarchico. Si può visualizzare come un Dendogramma e grazie ad esso non è necessario stabilire a priori il numero **N** di cluster da ricavare, poiché esso può essere scelto in seguito prendendo il livello N-esimo del dendogramma [15]. Questa tecnica si suddivide in due approcci:

- **Agglomerativo**: Il processo inizia considerando ogni punto come un cluster, ad ogni step si unificano i punti secondo una particolare funzione di similitudine arbitraria, fino ad ottenere un cluster unico ed il relativo dendogramma.
  Questo approccio si basa sullo sviluppo di una *Matrice di Prossimità* tra i cluser e risulta di fondamentale importanza la funzione per il calcolo della similitudine tra due cluster, esistono differenti varianti:

  - *Single Linkage*, distanza minima tra tutti i loro punti.

  - *Complete Linkage*, distanza massima tra tutti i loro punti.
  - *Group Average*, media della distanza a coppie tra tutti i loro punti.
  - *Centroid-Distance*, distanza tra i centroidi dei cluster.
  - *Ward's Method*, distanza basata sull'errore quadratico medio dato dall'unione dei due cluster [7].

- **Divisivo**: Caso complementare in cui si parte da un unico cluster e si suddivide ad ogni iterazione, fino ad ottenere un numero di cluster pari al numero di punti che costituiscono la base dati.

Entrambi gli approcci possono essere interrotti al raggiungimento di un numero di cluster N desiderato. Lo spazio richiesto è di $O(N^2)$ a causa della matrice di prossimità, mentre la complessità è nell'ordine di $O(N^3)$, dovuto dal fatto che ci sono N step ad ogni aggiornamento della matrice di prossimità che deve essere costantemente aggiornata. Come in K-Means, la presenza di outliers condiziona negativamente questo approccio, inoltre presenta difficoltà nella gestione di cluster le cui dimensioni sono molto variabili, per esempio tende a rompere cluster molto grandi a prescindere dalla loro validità.

Density-Based Clustering

Il density-based clustering si basa sul concetto di **Densità**. L'idea di base è trovare clusters definiti implicitamente da regioni ad alta densità separate da regioni a bassa densità. Uno degli algoritmi più famosi di questa categoria è il DBSCAN che usa due parametri per identificare aree dense: un raggio ε, che serve a identificare un'area attorno ad un determinato punto, e un numero minimo di punti *MinPts* che devono essere presenti all'interno del raggio ε. Ogni punto viene etichettato secondo 3 differenti categorie:

- **Core Point**: tutti i punti che superano la soglia *MinPts* all'interno del raggio ε.

- **Border Point**: tutti i punti che non superano la soglia *MinPts* ma nel loro raggio ε hanno almeno un Core Point.

- **Noise Point**: tutti i punti che non sono Border o Core Point.

L'algoritmo parte da un punto casuale. Sono calcolati tutti i punti compresi nel raggio ε e se contiene un numero *MinPts* di punti, viene creato un nuovo cluster altrimenti viene etichettato come Noise-Point. Il punto potrebbe essere successivamente ritrovato in quanto incluso nel raggio ε di un vicino e di conseguenza essere inserito in un cluster.
Se un punto è associato ad un cluster, sono inseriti in esso anche i punti presenti all'interno del suo raggio ε, e di conseguenza anche i loro vicini all'interno sempre del raggio stabilito. Questo processo continua fino a quando non sono stati inseriti tutti vicini ottenuti, ogni punto a cui è associato un cluster viene marcato come visitato e l'algoritmo prosegue eseguendo la stessa procedura per un punto successivo che non è ancora stato visitato.

L'algoritmo ha complessità $O(n^2)$ che tuttavia può essere ridotta a $O(n \log n)$ tramite utilizzo di strutture indicizzate per l'interrogazione del vicinato.
Il punto di forza di questo approccio è dato dalla buona gestione di outliers e dalla conseguente capacità di riuscire a gestire cluster di forme e dimensioni molto differenti. Tuttavia risulta inefficiente quando si ha a che fare con dati che son caratterizzati da densità troppo variabili.

Validazione

Il processo di validazione compone la fase più importante di tutto il processo di clustering, esso permette di verificare la qualità della clusterizzazione ottenuta. Esistono differenti tipi di validazione:

- **SSE Sum of Squared Errors**: Si tratta di una **misura interna** ossia che si basa solamente sull'insieme di dati clusterizzati, per ogni punto l'errore consiste nella distanza dai cluster ottenuti:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2$$

  Dove k è il numero di clusters, C il set degli oggetti all'interno del cluster ed m il centroide del cluster. Questo indice, se applicato al singolo cluster, misura la sua **coesione** ossia come sono strettamente correlati gli elementi all'interno del cluster. Altrimenti se calcolato tra cluster differenti, fornisce una stima della **separazione**, ossia come i cluster sono ben separati tra di loro. Queste caratteristiche permettono di eseguire una stima della soluzione ottenuta durante ogni momento

della procedura di clustering, ad esempio in K-Means non solo permette di valutare il parametro K, ma consente di interrompere l'algoritmo prima della sua naturale condizione di stop, se SSE rispetta opportune soglie.

- **Silhouette Coefficient**: Questo indice rientra nelle **misure interne**, infatti combina entrambi i concetti di coesione e separazione. Per ogni individuo si calcola la distanza media all'interno del proprio cluster D_{int} e la distanza media minima tra tutti gli altri cluster D_{ext}. Il coefficiente di Silhouette S è calcolato secondo la formula:

$$S = \begin{cases} 1 - \frac{D_{int}}{D_{ext}} & \text{if } D_{int} < D_{ext}, \\ \frac{D_{ext}}{D_{int}} - 1 & \text{if } D_{int} > D_{ext}, \\ 0 & \text{if } D_{int} = D_{ext} \end{cases}$$

  Tipicamente $-1 \leq S \leq 1$, dove 1 è il valore ottimo.

- **Entropia e Purità**: In questo caro rientriamo nelle **misure esterne**, ossia indici che si basano su fattori esterni alla clusterizzazione ottenuta. Infatti questo prevede che gli elementi siano caratterizzati da proprietà che li classificano distintamente. Per ogni cluster viene calcolata la distribuzione dei suoi elementi, rispetto alle classi a cui appartengono. Per il cluster j si calcola p_{ij} ossia la probabilità che un elemento del cluster j appartenga alla classe i nel modo seguente: $p_{ij} = \frac{m_{ij}}{m_j}$ dove m_j è il numero di elementi nel cluster j e m_{ij} è il numero di elementi della classe i all'interno del cluster j.
  Utilizzando la distribuzione della classe, con L il numero di classi, l'entropia e di ogni cluster j è calcolata utilizzando la formula:

$$e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$$

  L'entropia totale è calcolata come la somma delle entropie di tutti i cluster, pesata per le dimensioni di ognuno di essi. Considerando m il numero totale di elementi, m_j la dimensione del cluster j e K il numero di cluster, otteniamo la seguente formula:

$$e = \sum_{i=1}^{K} \frac{m_i}{m_j} e_j$$

Per quanto riguarda l'indice di purità di un cluster j, si definisce con la formula $purity_j = \max p_{ij}$. La purità totale è la media pesata rispetto alle dimensioni dei singoli cluster e si ottiene con la formula:

$$purity = \sum_{i=1}^{K} \frac{m_i}{m_j} purity_j$$

2.4 Predizione e Classificazione

In Data Mining esistono due forme di analisi che possono essere utilizzate per estrarre modelli e descrivere importanti classi o prevedere futuri pattern nei dati. Queste due forme sono la Predizione e la Classificazione.

2.4.1 Classifciazione

Con Classificazione si intende il processo che data una collezione di record, denominata **Training Set**, cerca di costruire un **modello** in grado di attribuire una caratteristica, denominata attributo **Classe**, basandosi sulla combinazione delle altre proprietà che caratterizzano il singolo individuo della popolazione. Una volta ottenuto il modello questo può essere usato per predire la classe di nuove istanze di record per cui la classe è sconosciuta. Dopo la costruzione del modello la fase più importante è la sua validazione, che avviene applicando il modello ad una partizione dei dati disposizione, denominata **Test Set**. Questo processo è fondamentale per determinare la qualità del modello. In seguito sono riportati alcuni dei principali metodi presenti in letteratura per affrontare il processo di classificazione:

- **Decision Trees** [7]: Descrive una struttura ad albero dove i nodi foglia rappresentano le classificazioni e le ramificazioni l'insieme degli attributi che le determinano. Di conseguenza ogni nodo interno risulta essere una macro-classe costituita dall'unione delle classi associate ai suoi nodi figli. Il meccanismo si basa sul concetto di **Split**, ossia come ogni attributo viene suddiviso in base ai valori che può assumere. il tipo di attributo *Nominale*, *Ordinale* o *Continuo* influisce molto sulle

tipologie di split, il quale può consistere in una suddivisione binaria oppure multipla.
Per quanto riguarda gli attributi nominali e ordinali, questi split avvengono semplicemente dividendo, in molteplici sottoinsiemi, tutti i valori assunti dall'attributo. Nel caso di attributi continui è opportuno effettuare procedure di discretizzazione per renderlo categorico oppure suddividerlo in sottoinsiemi espressi con disequazioni.
I differenti algoritmi esistenti in letteratura si differenziano in base alla strategia impiegata, ad ogni nodo, per la valutazione dello Split, esistono infatti differenti indici:

- **Gini Index** è usato in CART, SLIQ, SPRINT [16]. Considerando $p(j|t)$ la frequenza relativa della classe j al nodo t questo è definito come segue:

  $$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

  Nel caso il nodo p sia suddiviso in k partizioni allora la qualità dello split si calcola nel seguente modo, con n_i il numero di record al figlio i ed n il numero di record al nodo p:

  $$GINI_{p-split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

- **Gain Index** è usato in ID3, C4.5 [16] e si basa sul concetto di entropia, indice relativo alla omogeneità del nodo. Indicando con $p(j|t)$ la relativa frequenza della classe j al nodo t questo indice è definito come segue:

  $$Entropy(t) = -\sum_j p(j|t) \log p(j|t)$$

  Il **Gain**, misura la riduzione dell'entropia ottenuta eseguendo un particolare split su un nodo p:

  $$GAIN_{p-split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

- **Classification Error** misura l'errore di classificazione commesso al nodo t. Indicando con $p(j|t)$ la relativa frequenza della classe j al nodo t questo indice è definito come segue:

  $$ERROR(t) = 1 - max_j P(j|t)$$

- **Classificatori basati su istanze**: Consiste in una famiglia di algoritmi i quali, anzi che eseguire generalizzazioni esplicite, confrontano nuove istanze direttamente con i record analizzati e opportunamente memorizzati dal training set. Degna di nota è la procedura **Nearest-Neighbor** che utilizza una particolare ed arbitraria metrica per il calcolo della distanza ed un parametro k rappresentante il numero minimo di vicini da estrarre [7]. Per ogni record che deve essere classificato, si calcola la distanza dal training set identificando i k record ritenuti più vicini e si usano i valori assunti dai loro attributi per classificare il record in esame.

- **Classificatori Byesiani**: Consiste in un framework probabilistico per risolvere il problema della classificazione. Si basa fortemente sul concetto di **Probabilità Condizionata**:

  $$P(C|A) = \frac{P(A,C)}{P(A)} \qquad P(A|C) = \frac{P(A,C)}{P(C)}$$

  Ne segue il **Teorema di Bayes** [17]:

  $$P(C|A) = \frac{P(A|C)P(C)}{P(A)}$$

  Si considerano gli attributi e la classe come variabili casuali. Dato un record con attributi $(A_1, A_2, ..., A_n)$, l'obiettivo è quello di prevedere la classe C, ossia vogliamo trovare il valore di C che massimizza la probabilità $P(C|A_1, A_2, ..., A_n)$. Grazie al teorema di Bayes, si ottiene un problema di ottimizzazione equivalente che consiste nel trovare C che massimizza: $P(A|C) = \frac{P(A,C)}{P(C)}$. Esistono differenti modi per la stima di tale probabilità basandosi sui dati, come distribuzione normale, stima di densità, m-estimate, Laplace [7].

- **Support Vector Machine (SVM)** : la classificazione viene eseguita trovando l'iperpiano che massimizza il margine tra due classi. I vettori (possibili attributi della classe) che definiscono l'iperpiano, sono definiti **vettori di supporto**. Il vantaggio di questo metodo consiste nel fatto che se i dati sono linearmente separabili, allora esiste un minimo globale unico. Una SVM ideale dovrebbe produrre un iperpiano che separa completamente i vettori di due classi non sovrapposte. In genere la completa separazione non è sempre possibile, spesso si arriva ad ottenere un modello con troppi possibili casi che comporta una classificazione non corretta [18].

Validazione

Questo processo è di fondamentale importanza, in quanto permette di valutare le prestazioni del modello costruito e di poterlo confrontare con altre possibili modellazioni. Le misure di valutazione si basano sul Test-Set, partizione dei dati su cui applicare il modello predittivo. L'applicazione del modello sul Test-Set produce la **Matrice di Confusione**, ossia matrice indicante l'incidenza tra le classi predette ed il loro valore reale dei record nel Test-Set. Si possono quindi determinare le seguenti tipologie di previsione:

- **True Positive** : Predizioni Positive Corrette
- **False Positive** : Predizioni Negative Corrette
- **True Negative** : Predizioni Positive Errate
- **False Negative** : Predizioni Negative Errate

Queste possono essere applicate a qualsiasi tipologia di attributo, non solamente alle classi binarie. Le metriche più utilizzate sono:

- Accuracy: $\frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$
- Precision: $\frac{TruePositive}{TruePositive+FalsePositive}$
- Recall: $\frac{TruePositive}{TruePositive+TrueNegative}$
- F-Measure: $\frac{2\times TruePositive}{TruePositive+TrueNegative+FalsePositive}$

2.4.2 Predizione

Con il concetto di Predizione, si intende un meccanismo attraverso il quale, differentemente dalla Classificazione, non si cerca di assegnare una classe particolare ad un individuo, ma si possono effettuare previsioni sui singoli valori degli attributi, e su quello che sarà il comportamento futuro dell'individuo. I metodi utilizzati per queste procedure infatti sono analoghi, la differenza risiede nello scopo che si vuole raggiungere: prevedere una tipologia, un attributo, una classe dell'individuo consiste in una classificazione, mentre prevedere i valori specifici o in range degli attributi, consiste in una previsione [19].

Mining Regole di Associazione

La base di partenza di un algoritmo per l'estrazione di regole associative è costituita da un insieme di **Transazioni**. Ogni transazione consiste in un insieme di item. Estrarre le *Regole di Associazione* consiste nel prevedere l'occorrenza di un item in base all'occorrenza di altri item compresi anch'essi nelle transazioni a disposizione.
Risulta importante definire alcuni concetti alla base di questa tecnica:

- **Itemset**: Collezione di uno o più elementi generalmente definito per mezzo del parametro k, indicativo della sua dimensione nella forma $k\text{-}Itemset$.
- **Supporto Itemset**: Dato un itemset I, il supporto è la frazione delle transazioni che contengono I e si indica con $supp(I)$.
- **Itemset Frequente**: Tutti gli itemset che superano un'arbitraria soglia minima di supporto.

Una **Regola di Associazione** è un'implicazione espressa nella forma:

$$X \rightarrow Y \text{ con } X, Yitemset$$

dove X prende il nome di *Premessa* ed Y *Conseguenza* della regola. Oltre al supporto, visto precedentemente, esiste un'altra forma di validazione della regole che tiene conto sia della premessa che della conseguenza: la **Confidenza**. Essa indica quanto spesso una particolare regola è verificata, consiste nella proporzione tra il numero delle transazioni che contengono l'intera regole e le transazioni che contengono la premessa:

$$conf(X \rightarrow Y) = supp(X \cup Y)/supp(X)$$

Formalmente il supporto $supp(X \cup Y)$ può essere riscritto come la probabilità congiunta $P(E_X \cap E_Y)$, dove E_X e E_Y sono tutte le transazioni che contengono X o Y rispettivamente. Quindi possiamo esprimete la confidenza come la probabilità condizionata $P(E_Y|E_X)$.

Dato un set di transazioni, l'obiettivo consiste nell'estrazione di tutte le regole che rispettano le soglie arbitrarie di supporto e confidenza. La loro estrazione non può essere eseguita con un approccio Brute-Force, a causa dell'elevato numero di regole che possono essere generate. Per ridurre il numero di possibili regole, si sfrutta il **Principio Apriori**.

Principio Apriori

Questo principio si basa sulla proprietà *anti-monotona* del supporto, che ci permette di stabilire con certezza che se un itemset non risulta frequente, allora nemmeno tutti gli itemset che lo contengono risulteranno frequenti. Tale proprietà è cosi formalizzata, con X e Y itemset:

$$\forall X, Y : (X \subseteq Y) \Rightarrow supp(X) \geq supp(Y)$$

Questa proprietà è alla base dell'algoritmo *Apriori*, procedura che partendo da tutti i possibili item con cardinalità 1, costruisce tutti i gli itemset di dimensione $n + 1$ con n la dimensione dell'itemset di partenza e ad ogni iterazione verifica se l'itemset generato è frequente o meno. La proprietà anti-monotona permette di escludere itemset non frequenti e di conseguenza tutti possibili itemset derivanti da essi. Gli step da cui è costituita la procedura sono i seguenti:

Algorithm 2 Apriori

```
1: function APRIORI(set transazioni T, minSupport)
2:     k = 1
3:     Generazione itemset con cardinalitá 1
4:     repeat
5:         Generazione itemset di cardinalitá k + 1
6:         Eliminazione itemset contenenti non frequenti
7:         Calcolo support itemset generati
8:         Eliminazione itemset non frequenti
9:     until Non sono generati ulteriori itemset frequenti
```

Segue l'algoritmo implementato nel dettaglio:

Algorithm 3 Apriori

```
1: function APRIORI(T, ϵ)
2:     L_1 ← {large 1 − itemsets}
3:     k = 2
4:     while L_{k−1} ≠ ∅ do
5:         C_k ← Generate(L_{k−1})
6:         for transaction t ∈ T do
7:             C_t ← Subset(C_{k1}, t)
8:             for candidates c ∈ C_t do
9:                 count[c] ← count[c] + 1
10:        L_k ← {c ∈ C_k | count[c] ≥ ϵ}
11:        k ← k + 1
12:    return ⋃_k L_i
```

Generazione Regole

Al termine di questa procedura, otteniamo tutti gli itemset che hanno superato la soglia supporto. Bisogna procedere con l'estrazione delle regole di associazione dagli itemset ottenuti. Le regole generate saranno valutate in base alla loro Confidenza (soglia arbitraria), e quest'ultima generalmente non gode della proprietà *anti-monotona*, ma la confidenza delle regole generate dal solito itemset possiede la seguente proprietà, indicando con $Conf(X \Rightarrow Y)$ la confidenza della regola $X \Rightarrow Y$:

$$Conf(ABC \Rightarrow D) \geq Conf(AB \Rightarrow CD) \geq Conf(A \Rightarrow BCD)$$

Si evince che la confidenza è *anti-monotona* rispetto al numero di item che compongono la premessa della regola.
Si procede quindi generando le regole che possiedono solo un item nella conseguenza, vengono eliminate tutte le regole che non superano la soglia minima di confidenza. Sulla base delle regole rimaste, si procede generando e valutando le regole con un item addizionale nella conseguenza, procedendo fino a che non sono state generate tutte le possibili regole.

Validazione Regole

Le regole estratte sono sottoposte ad un'ulteriore fase di postprocessing, in quanto la confidenza può alle volte essere fuorviante come indice di validità

per una regola. Quato aspetto emerge per itemset che fanno parte della premessa di una regola, caratterizzati da alto supporto. Un itemset molto frequente tende ad alzare l'indice di confidenza delle regole di cui esso costituisce la premessa, indipendentemente dal fatto che la regola sia contestualmente valida. Per ottenere una validazione più affidabile, viene fatto affidamento ad un indice che tiene conto dell'indipendenza statistica tra premessa e conseguenza della regola, il **Lift**:

$$Lift = \frac{P(E_Y|E_X)}{P(E_Y)}$$

Dove E_X e E_Y sono tutte le transazioni che contengono X o Y rispettivamente.
Per la similarità tra regole si utilizza come misura il **coefficiente di Jaccard**, dati due set A e B questo indice J è dato dal rapporto tra la dimensione dell'intersezione con la dimensione dell'unione dei due set:

$$J(A,B) = \frac{|(A \cap B)|}{|A \cup B|}$$

Esistono molteplici forme di validazione per le regole di associazione [7], sono state riportate solamente le misure interessate dal progetto svolto.

Pattern Sequenziali

L'estrazione di pattern sequenziali rientra nella categoria "Structured Mining" [10], ossia l'estrazione di informazioni da **dati semi-strutturati**. I pattern ottenuti consistono in sequenze di elementi i quali si susseguono generalmente in sequenza temporale, ma anche geografica [7]. I dati semi strutturati sono necessari e risulta fondamentale la fase di **preprocessing dei dati**[1], attraverso la quale si crea una base dati costituita da **sequenze di item**.
Analogamente all'estrazione delle regole di associazione, anche in questo contesto è prevista la dichiarazione di una **soglia minima di frequenza**, che il pattern dovrà rispettare per essere preso in considerazione. Formalmente una sequenza è definita nel seguente modo:

$$s = \langle e_1, e_2, ..., e_j \rangle$$

Dove ogni e_j sono dette *Transazioni*. Ogni transazione consiste in una collezione di item o eventi:

[1]Secondo step del processo KDD, esposto nel paragrafo 5.1

$$e_i = \{i_1, i_2, i_n\}$$

Ad ogni item è attributio uno specifico istante temporale oppure una specifica posizione. La lunghezza di una sequenza è data dal numero k di transazioni che contiene e prende il nome di $k-$sequenza.
Le precedenti definizioni sono necessarie per introdurre il concetto di sottosequenza, una sequenza $\langle a_1, a_2, ..., a_n \rangle$ si definisce **sottosequenza** di $\langle b_1, b_2, ..., b_m \rangle$ con $m \geq n$, se esiste una sequenza di indici $i_1 < i_2 < ... < i_n$ tali che:

$$a_1 \subseteq b_{i1}, a_2 \subseteq b_{i2}, ..., a_n \subseteq b_{in}$$

Questa definizione ci permette di raffinare il concetto di frequenza in **supporto**. Il supporto di una sottosequenza α è definito dal rapporto tra le sequenze che contengono α ed il totale. Come per le regole di associazione, si sfrutta il principio **Apriori**, secondo il quale, se una sequenza soddisfa la soglia minima id supporto, allora tale soglia è soddisfatta anche tutte le sue sottosequenze. Questo approccio vede come protagonista l'algoritmo Generalized Sequential Patterns, il quale si basa sulla generazione di sequenze frequenti a partire dai singoli item che soddisfano la soglia minima di supporto (analogamente all'algoritmo Apriori esposto nella Sezione 2.2.2) [7].
La particolarità dei Sequential Pattens consiste nella loro relazione con gli intervalli temporali che intercorrono tra le transazioni che costituiscono le sequenze e tra le sequenze stesse. Questo aspetto permette di raffinare la ricerca delle sequenze frequenti, in base all'arco di tempo che esse ricoprono ed in base all'intervallo temporale tra gli item che costituiscono la transazione. Emergono differenti approcci basati sul tipo di vincolo temporale, come l'estrazione di sequenze immediatamente consecutive, oppure non troppo dilazionate nel tempo.

2.5 Contesti Applicativi

Il concetto di Customer Analytics vede al giorno d'oggi la sua maggiore applicazione nei contesti caratterizzati da un'elevata dinamicità dell'individuo analizzato, sia in termini di quantità di traffico che di eterogeneità delle sue azioni. I contesti che traggono i maggiori benefici da questo tipo di approccio sono infatti la mobilità, il marketing e le telecomunicazioni.

2.5.1 Mobilità

L'Individuo analizzato può essere sia una persona fisica che un mezzo di trasporto specifico, l'analisi è effettuata sull'insieme degli spostamenti effettuati.

Mobility Profiling

Il profiling basato sulla mobilità di un individuo (persona fisica o mezzo di trasporto), consente non solo di individuare i luoghi di interesse principali, ma anche di capire il comportamento del singolo rispetto ad una popolazione di individui. Questo accorgimento permette di poter ricavare previsioni sui possibili spostamenti di un individuo, avendo a disposizione i dati riguardo alla sua mobilità in tempo reale. Questi dati sono accessibili per mezzo di telefoni, oppure sensori GPS installati ormai su tutti i veicoli o mezzi di trasporto.
Questo approccio è stato analizzato e messo in pratica in [1]. L'idea si basa sul fatto che tipicamente ogni utente visita sistematicamente con regolarità un determinato set di destinazioni. Per ogni utente si creano dei veri e propri profili basati sulle informazioni di mobilità accumulate e queste informazioni personali sono usate, sia personalmente che collettivamente, per prevedere gli spostamenti dell'utente. La condivisione dei profili permette di determinare i futuri spostamenti, basandosi sul fatto che un percorso sistematico di un individuo è molto probabile che sia sistematico anche per altri utenti. Nel caso non sia consentita la condivisione dei dati, l'utente può contare sulle sue informazioni personali per delle previsioni comunque accurate. Nel lavoro svolto si utilizzano le strategie di data mining citate precedentemente per creare i *"Mobility Profiles"* costituiti dalle traiettorie sistematiche di un individuo, infatti esse sono raggruppate tramite Density Based Clustering, con opportune funzioni di distanza. Ogni utente possiede un profilo di mobilità ed il calcolo della similitudine tra la rotta intrapresa ed i profili degli altri utenti, permette questo tipo di previsioni.

Sempre nel campo della predizione di traiettorie, è degno di nota l'approccio adottato in [2] nel quale si sfrutta il principio Apriori, ottenendo un set di traiettorie frequenti, e ricavando determinate regole di associazione che soddisfano criteri di validità (supporto, confidenza, lift..) prestabiliti. Tra tute le regole ottenute, si impiegano strategie di comparazione per filtrare quelle più simili al percorso intrapreso e le loro conseguenze permettono di determinare potenziali traiettorie future.

2.5.2 Marketing

Nel campo del marketing, l'intensa competizione e l'aumento delle possibili scelte da parte della clientela, scaturiscono sempre nuove pressioni e di conseguenza nuove esigenze sui processi decisionali applicati [3]. Queste esigenze sono soddisfatte in modo sempre più efficiente dalle procedure di data mining esposte. Infatti il riconoscimento di pattern è una componente molto importante per la conoscenza del cliente, permettendo di migliorare i rapporti tra esso e l'azienda. In [4] vediamo esattamente come applicare le regole di associazione per adempire al processo di *Customer Profiling*. I profili costruiti sono costituiti da due aspetti: descrittivo e comportamentale. Il primo si basa su nozioni come genere, età, stato sociale, mentre il secondo sulle informazioni estratte dallo storico delle sue transazioni. L'analisi comportamentale prevede una fase di estrazione delle regole d'associazione relative al cliente. Queste regole sono raggruppate per similarità e filtrate secondo template ritenuti significativi o irrilevanti, infine si procede rimuovendo le regole ridondanti ossia quelle che possono essere dedotte da altere regole presenti nell'insieme.

2.5.3 Telecomunicazioni

Il campo delle telecomunicazioni è caratterizzato da un elevato traffico, che è costituito da un insieme altamente eterogeneo di informazioni: dati riguardanti la mobilità, dati sulle abitudini di acquisto, dati sull'utilizzo dei servizi offerti, profili non comportamentali dei clienti, ossia età, genere, data e luogo di nascita. Tutte queste informazioni sono in possesso delle compagnie in quanto ogni utente è generalmente vincolato da contratto. La magnitudine di queste informazioni non può permettere analisi efficienti senza l'ausilio di mezzi informatici, infatti le procedure di clustering e classificazione vedono in questo settore il più alto impiego. In [5] si creano nuovi classi che determinano il profilo comportamentale del cliente, per mezzo di differenti algoritmi di clustering. Questi profili sono confrontati con i dati descrittivi del cliente per poi applicare la metodologia di apprendimento supervisionato delle Support Vector Machine [18], unificando i due concetti, descrittivo e comportamentale, al fine di ottenere una Customer Segmentation adatta al contesto.

Capitolo 3

L'Azienda Ser.Nav s.r.l.

Questo Capitolo tratta le informazioni principali sull'azienda ed il contesto in cui è stato svolto l'intero progetto di tesi. Infine è descritta l'interazione e l'elaborazione delle informazioni descrivendo sia la base dati da cui sono state estratte che il modello relazionale costruito.

3.1 L'Azienda

Ser.Nav s.r.l. è agenzia marittima e spedizioniere doganale, facente parte di un gruppo che copre tutte le operatività attinenti le spedizioni, fanno infatti parte del gruppo: centri assistenza doganali (CAD), società di trasporto, di magazzinaggio e di logistica e distribuzione. Essa si occupa del traffico Import/Export riguardante i porti di Genova, La Spezia e Livorno e nella sua banca dati sono presenti oltre 600 clienti, tra cui molte delle più grosse case di spedizione a livello mondiale. Essa possiede oltre 150 dipendenti suddivisi nelle sedi di Genova, Livorno e La Spezia (sede principale). Nell'ultimo decennio l'azienda ha investito nella formazione e nello sviluppo di un ufficio informatico interno, il quale si occupa della creazione, manutenzione ed assistenza di software gestionale per l'agevolazione della produttività.
Ogni giorno vengono evase pratiche doganali relative all'importazione ed esportazione di centinaia di Container, il tutto tramite software auto-prodotto e mantenuto interamente. Il team informatico è in continuo sviluppo e costante monitoraggio di tutte le attività svolte.
La tipologia di cliente è molto eterogenea, varia dal privato alle grandi multinazionali dei trasporti che hanno ritmi e carichi di lavoro molto differenti.

3.2 Necessità Aziendali e Definizione del Problema

L'Azienda Ser.Nav s.r.l. ha la necessità di identificare per ogni cliente delle possibili offerte in termini di servizi. Attualmente, quest'attività è realizzata tramite statistiche validate dall'esperienza dei diretti responsabili della sezione marketing.

L'obiettivo di questa tesi è di definire un processo analitico che permette di identificare delle tariffe adeguate al cliente basandosi sulle sue preferenze e abitudini. L'idea è quella di sfruttare modelli di data mining per calibrare al meglio le offerte, diminuendo e aumentando i prezzi dei servizi secondo opportune previsioni basate su un'attenta analisi del cliente.

Per affrontare questo problema abbiamo deciso di applicare metodologie di **Market Basket Analysis** [20], considerando le operazioni fatturate a ciascun cliente come se fossero dei beni acquistati, e di conseguenza l'intera fattura diventa lo "scontrino" del cliente.

3.3 Natura ed Estrazione delle Informazioni

Il traffico relativo alla fatturazione è gestito in gran parte dal sistema informativo aziendale AS400 [21]. Oltre alle sue consuete funzionalità, permette analisi statistiche limitate e non consente l'accesso diretto alle informazioni inerenti il traffico di ciascun cliente. L'interazione con la base dati contenente le informazioni interessate è molto delicata, in quanto la sua consultazione massiva causa un rallentamento delle prestazioni del server dovuta ad un'eccessiva occupazione. Pertanto è stata necessaria una lunga e laboriosa fase di estrazione dati, non solo a causa dell'impiego fisico delle risorse, ma anche per la struttura del database in esame e dei record al suo interno: si tratta di un database relazionale denominato DB2 [23], modificato appositamente in base all'operatività aziendale, all'interno del quale le tabelle sono nominate con brevi sigle e le colonne di ogni tabella sono distinte per mezzo di codici numerici, rendendo obbligatorio l'aiuto sia del reparto fatturazione che del reparto tecnico addetto alla gestione della base dati. Tale database è suddiviso in 4 sezioni, una per ogni società appartenente al gruppo Ser.Nav: Marittima Servizi, I.C.S Livorno, I.C.S. Genova ed ovviamente Ser.Nav che svolge ruolo coordinante. Ogni sezione è caratterizzata da un codice libreria e l'accesso alle singole tabelle è effettuato specificando prima la libreria poi

la tabella, ad esempio l'accesso alla tabella FTSTTE (*Testata Fatture)* della società *Ser.Nav* avviene nel seguente modo:

Liberia: ADEFINTESE → *ADEFINTESE.FTSTTE*

Il solito accesso per la società *I.C.S. Genova* risulta:

Liberia: ADEFINTEIC → *ADEFINTEIC.FTSTTE*

In parallelo al processo di estrazione dati, è stato progettato un modello relazionale e costruita una base dati dedicata dove inserire le informazioni ottenute. In seguito sono mostrate le principali fasi del processo.

3.3.1 Estrazione Clienti

Il cliente è caratterizzato dal nome e dalla Partita IVA come identificatore univoco, tuttavia all'interno di AS400 essa non è univoca: sono inserite le singole filiali, identificate univocamente da un codice AS400, da un campo descrittivo relativo alla provincia di appartenenza e dalla Partita IVA.
Sono stati estratti tutti i codici clienti per un totale di 1299 voci, e raggruppati in base alla Partita IVA ottenendo 759 clienti distinti. Inoltre è stato necessario procedere manualmente alla raffinazione dei record ottenuti, sotto la supervisione di addetti alla fatturazione, poiché alcuni clienti presentavano molteplici Partite IVA, causa di cambi gestione o filiali estere. In totale sono presenti 631 clienti. Tuttavia è stato necessario creare due tabelle per la gestione dei clienti:

- **CLIENTI_PIVA:** Tabella contenente il cliente ad un livello più generale possibile, ossia caratterizzato dalla descrizione e dalla Partita IVA; inoltre è presente il flag *APPARTIENE_MODELLO* il quale specifica la rilevanza del cliente rispetto al traffico presente, per contraddistinguere i clienti obsoleti o irrilevanti in termini di fatturato e/o numero di transazioni.

- **CLIENTI:** Tabella nella quale son state inserite tutte le filiali relative ad ogni cliente, con riferimento tramite chiave esterna alla tabella precedente. Queste informazioni sono di fondamentale importanza, poiché (come verrà esposto nella procedura di estrazione del traffico) ogni transazione presente in fattura fa riferimento al codice AS400 della filiale, non alla Partita IVA. Quindi è stato necessario campionare questi dati per essere in grado durante l'estrazione dei movimenti, di individuare il cliente responsabile.

3.3.2 Campionamento Operazioni

In totale sono presenti 51 tipi di voci fatturabili al cliente. Per determinarle sono stati necessari confronti con i responsabili della fatturazione, per comprendere la struttura delle tariffe associate al cliente. All'interno di AS400 le voci di fatturazione, che identificano una determinata produttività, possono essere abbinate ad uno o più codici, per motivi operativi interni. Ogni codice ha un costo associato che varia, per questioni commerciali, a seconda della società e del porto in cui l'operazione viene svolta e fatturata, ad esempio:

Operazione	**Società**	**Porto**	**Codice**	**Prezzo**
Operazione Definitiva Container	Ser.Nav	SP	ID4, I44	45€
Operazione Definitiva Container	Ser.Nav	GE	ID5, I45	45€
Operazione Definitiva Container	Ser.Nav	LI	ID1, 114	48€
Operazione Definitiva Container	Marittima Servizi	SP	I44	45€
Operazione Definitiva Container	Marittima Servizi	GE	I45	45€
Operazione Definitiva Container	Marittima Servizi	LI	113, 114	48€
Operazione Definitiva Container	ICS Genova	SP	I46	46€
Operazione Definitiva Container	ICS Genova	GE	I47	46€
Operazione Definitiva Container	ICS Genova	LI	115	46€
Operazione Definitiva Container	ICS Livorno	SP	I48	46€
Operazione Definitiva Container	ICS Livorno	GE	I49	46€
Operazione Definitiva Container	ICS Livorno	LI	H71, 118	46€
...	...	...	...	...

Ne consegue che per ognuna delle 51 operazioni catalogate, sono associati minimo 12 codici. Questi codici non sono univoci, quindi per identificare un'operazione durante l'estrazione è necessario sapere oltre ad esso, la società ed il porto di riferimento come è spiegato nella seguente sezione. Le tabelle in esame sono:

- **OPERAZIONI:** Questa tabella conteniene le voci di fatturazione principali, ossia quelle che compaiono nella tariffa del cliente. Sono presenti campi descrittivi come la tipologia (se si tratta di IMPORT o EXPORT) e la sottotipologia (Voci Accessorie, Servizi Aggiuntivi, Deposito IVA, Pratica Merci Pericolose ...). Infine sono presenti altri parametri descrittivi, utilizzati solamente per la creazione delle tariffe proposte al cliente. Tra questi parametri è presente anche il tipo

di importo associato all'operazione: il parametro binario *PASSIVO*, specifica se la voce è un importo passivo, cioè un costo sostenuto dall'azienda per fornire una determinata prestazione, a fronte del quale è stato realizzato un relativo importo da fatturare al cliente.

- **CODICI_OPERAZIONI:** Tabella nella quale sono stati inseriti tutti i codici AS400 con riferimento tramite chiave esterna al porto, alla società ed all'operazione. Questo campionamento è necessario in quanto una stessa operazione può avere un prezzo differente in base al porto ed alla società che emette fattura.

3.3.3 Estrazione Traffico

Per questioni di architettura del sistema AS400 interne, per ogni fattura emessa il sistema AS400 popola due tabelle parallele. La prima, denominata FTSTTE , contiene la testata della fattura, dentro la quale è specificata la posizione (codice identificativo unico della fattura) ed i dettagli principali, come il totale di fattura, data, codice cliente, codice società, sigla del porto, valuta e tasso di cambio. Il nostro interesse ricade nella seconda tabella popolata, denominata FTSTD01, la quale comprende il dettaglio di tutte le fatture emesse. Al suo interno troviamo un record per ogni singola voce fatturata. Ogni record possiede, la data ed il riferimento alla posizione della testata ed in aggiunta troviamo il codice dell'operazione ed il prezzo, con relativa valuta e tasso di cambio. L'estrazione del traffico consiste in una ricerca parallela all'interno di queste due tabelle, al fine di estrarre il traffico di ogni cliente. La tabella creata prende il nome di **TRAFFICO_OPERAZIONI** ed è strutturata nel modo seguente:

- *ID_CODICE_OPERAZIONE:* Consiste nel riferimento al codice AS400. L'individuazione di questo riferimento non è immediata, poiché per come sono state campionate le operazioni in base al codice, al porto ed alla società, sono necessarie tutte queste informazioni per ottenere il record corretto da associare. Di conseguenza è necessaria una query nella tabella contenente le testate, andando a ricercare la posizione della fattura, ottenendo sia la società che il porto a cui la voce dettaglio si riferisce. Ottenute queste informazioni è possibile recuperare l'identificativo dell'operazione corretto ed associarlo al record.

- *ADDEBITO:* Prezzo della singola operazione convertito in euro, in base alla valuta specificata.
- *DATA* : Data di emissione della fattura.
- *ID_CLIENTE:* Riferimento al cliente intestatario della fattura.

Il popolamento della base dati è eseguito per ogni cliente. Iterativamente si ottengono tutti i codici delle filiali associate, e per ogni società si effettua una query nella tabella contenente i dettagli delle fatture, nel range di date in cui si vuole estrarre il traffico del cliente. I record ottenuti sono elaborati come spiegato ed inseriti nella base dati.

3.4 La Base di Dati

E' stato costruito un apposito Database per la gestione delle informazioni da elaborare e la costruzione del modello. La struttura del database è la seguente:

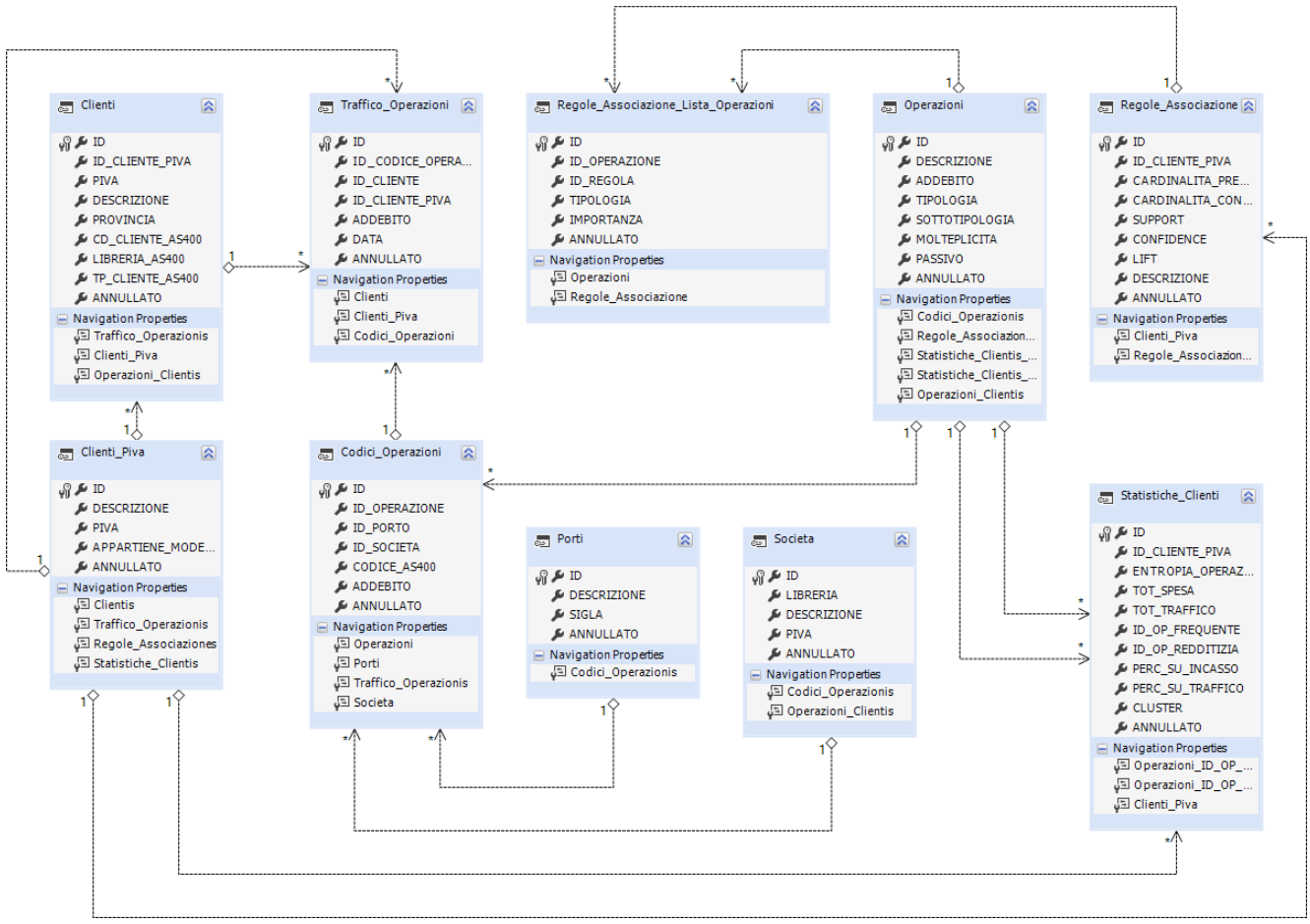


Figura 3.1: *Diagramma Relazionale relativo alla Base Dati costruita*

Degne di nota sono le tabelle:

- *REGOLE_ASSOCIAZIONE*: Tabella nella quale sono archiviate le regole di associazione estratte. Per ogni regola è presente il riferimento al cliente, le sue proprietà relative alla validazione: Supporto, Confidenza

e Lift; le sue proprietà statistiche, ossia quanto incide sul guadagno e sul traffico dell'azienda.

- *REGOLE_ASSOCIAZIONE_LISTA_OPERAZIONI*: Questa tabella è popolata con le singole operazioni che costituiscono la regola ed il loro indice di importanza che verrà spiegato in dettaglio durante la descrizione della procedura di analisi dei dati (Sezione 5.2.3). Sinteticamente esso rappresenta una combinazione tra la frequenza ed il costo dell'operazione nell'intervallo temporale analizzato.
  Per ogni operazione è specificato se fa parte della premessa o della conseguenza della regola. Ogni record di questo tipo possiede il riferimento (chiave esterna) ad un record *REGOLE_ASSOCIAZIONE*, ossia alla regola principale:

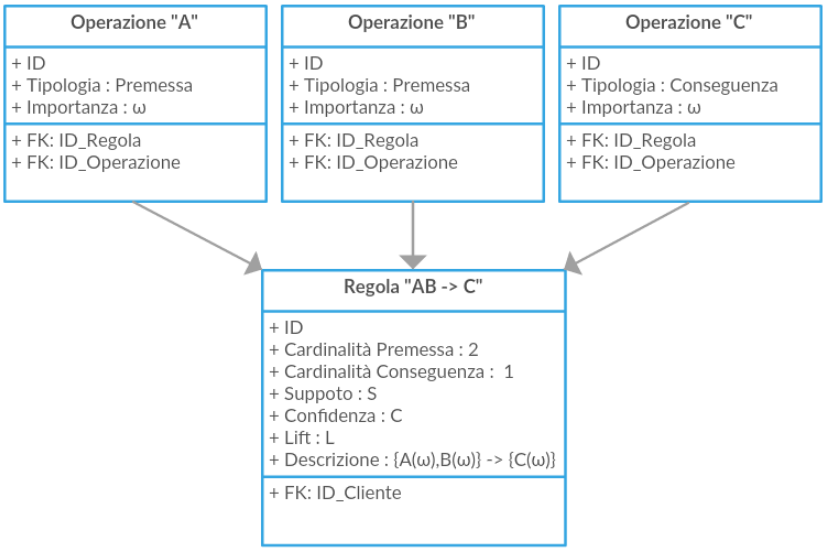


Figura 3.2: *Struttura Regola Associazione all'interno della base dati*

- *STATISTICHE_CLIENTI*: Questa tabella completa il modello in quanto contiene, per ogni cliente, le informazioni che compongono il profilo statistico del cliente ossia totale fatturato, totale traffico, la percentuale di fatturato e traffico rispetto al totale ricavato dall'azienda nel periodo ricoperto dal modello ed un indice relativo all'entropia delle

operazioni[1], per cogliere anche l'omogeneità delle tipologie di servizi di cui il cliente usufruisce.

L'insieme di queste tabelle costituisce il modello che verrà spiegato dettagliatamente nei capitoli seguenti.

[1]Entropia calcolata per mezzo della formula di *Shannon* [22]

Capitolo 4

Processo Analitico

In questo capitolo viene spiegato l'intero processo analitico pensato e realizzato per adempire alle necessità aziendali. Inizialmente verrà esposta la struttura generale, per poi spiegare nel dettaglio tutte le fasi che lo costituiscono. Si ricorda che le necessità aziendali richiedono uno strumento in grado di analizzare il comportamento dei clienti e generare previsioni che tengano conto, non solo della tipologia di servizio, ma anche del suo impatto sul traffico e sul fatturato.

4.1 Struttura del Processo

Il processo è suddiviso in due fasi principali, la prima consiste nella costruzione di un modello analitico, il quale descrive il comportamento di ogni cliente assunto in un range arbitrario di date, mentre la seconda è la fase predittiva, la quale specificando, anche in questo caso, un intervallo temporale, analizza il comportamento del cliente e sfrutta le informazioni contenute nel modello per generare previsioni sul traffico futuro del cliente.

4.1.1 Struttura del Modello

Come specificato nella fase introduttiva del capitolo, il modello deve essere costruito sulla base di un intervallo temporale abbastanza ampio da poter catturare il comportamento del cliente. Le fasi che costituiscono la sua creazione sono spiegate generalmente in seguito, per poi essere trattate nel dettaglio nelle sezioni successive:

- **Analisi Comportamentale**: Si estraggono le regole di associazione relative al range temporale scelto. Risulta necessario specificare una soglia S di supporto, per la generazione di itemset frequenti ed una soglia C di confidenza per l'estrazione delle regole. Infine tali regole saranno classificate in base al *lift* (un indice di importanza, Sezione 2.4.2), e dato che esse potranno essere eccessivamente numerose, bisogna specificare un valore K_{top} che limiterà il processo di estrazione.

- **Analisi Statistica**: I clienti sono analizzati sulla base del traffico effettuato nel range di date prestabilito. Sono creati e salvati i profili statistici ricavati dalle informazioni sul traffico e dal fatturato relativo al periodo.

- **Segmentazione**: I profili statistici sono raggruppati tramite procedure di clustering, opportunamente validate. Creando per ogni cliente un insieme di individui simili, sulla base delle informazioni statistiche che li caratterizzano.

Analisi Comportamentale

Il comportamento del cliente viene catturato e salvato per mezzo delle regole di associazione. Le regole sono estratte tramite l'algoritmo Apriori. Si procede partizionando temporalmente l'intervallo di tempo che ricopre il modello. Ogni partizione contiene una porzione del traffico effettuato dal cliente. Il traffico è costituito da una lista di operazioni con relativa data e prezzo fatturato.
Per poter applicare l'algoritmo, ogni operazione viene analizzata calcolando la sua influenza in percentuale sia sul traffico che sul fatturato, **sempre relativa alla singola partizione**. Nel dettaglio, il modello ricopre una porzione del dataset D_M, che viene partizionata in intervalli:

$$D_M = \langle Int_1, Int_2, \ldots, Int_n \rangle$$

Ogni intervallo si presenta come una lista con operazioni che possono ripetersi e che ogni volta possono avere prezzi differenti:

$$\{\langle Op_1, 450€\rangle\langle Op_2, 50€\rangle\langle Op_1, 200€\rangle\langle Op_3, 45€\rangle\langle Op_3, 350€\rangle\langle Op_4, 180€\rangle\}$$

Per ogni operazione Op_i si calcola la sua frequenza ed il suo fatturato, relativi entrambi all'intervallo. Si ottengono due vettori bidimensionali,

dove Op_i rappresenta l'$i-$esima operazione e F_i, G_i rispettivamente la sua frequenza ed il suo guadagno:

$$V_{frequenze} = \langle Op_i, F_i \rangle \qquad V_{guadagni} = \langle Op_i, G_i \rangle$$

Questi vettori devono essere combinati per mezzo di una apposita funzione che unifichi entrambi i vettori senza perdere significato (Sezione 5.2.3), ottenendo un unico vettore dei pesi ω:

$$V_\omega = \langle Op_i, \omega_i \rangle$$

Si calcola il vettore V_ω[1] per ogni intervallo Int_n, ottenendo il un nuovo dataset $\langle D_{Int_n} \rangle$. Effettuando questa operazione per ogni cliente, otteniamo una nuova base dati, predisposta per l'applicazione dell'algoritmo Apriori. Denotando con C_i l' $i-$esimo cliente, il dataset presenta questa struttura:

$$D_M = \{\langle C_1, D_{Int_n} \rangle \langle C_2, D_{Int_n} \rangle \dots \langle C_i, D_{Int_n} \rangle\}$$

L'output dell'algoritmo consiste in una collezione di itemset, con cardinalità ≥ 2. Ogni itemset è un vettore di coppie $\langle Op, \omega \rangle$ ed è caratterizzato dalla sua frequenza (supporto) all'interno del dataset. Successivamente è necessario estrarre le regole di associazione dagli itemset ottenuti. In questa fase è di fondamentale utilizzo la *proprietà anti-monotona della confidenza* (Sezione 2.4.2), la soglia minima di confidenza è un parametro arbitrario. In questo modo sono generate tutte le possibili regole di associazione e viene calcolato anche il $Lift$ di ognuna di esse, parametro secondo il quale saranno filtrate durante l'archiviazione. Il numero di regole che vengono estratte è variabile, dato che esse derivano dal traffico svolto dal cliente, si possono ottenere dalle poche decine alle migliaia di regole. Queste regole descrivono il comportamento del cliente, per questo motivo maggiore è il numero di regole, maggiore è la possibilità di ottenere una previsione, **ma non ci sono garanzie riguardo alla sua validità**. Di conseguenza non sono archiviare tutte le regole ottenute, ma solamente un numero limitato K_{top}, parametro del modello. Le regole vengono raggruppate per cardinalità, successivamente ogni set ottenuto è ordinato in base agli indici di valutazione della regola nel seguente ordine $lift$, $confidenza$, $supporto$. Dopo l'ordinamento si procede con l'archiviazione nella base dati delle K_{top} regole di ogni set.

[1]Risulta opportuno specificare che la funzione per il calcolo di ω è parametrica, la costruzione del modello è predisposta per utilizzare anche funzioni differenti.

Analisi Statistica

L'analisi statistica del traffico svolto dal cliente procede analizzando tutte le transazioni effettuate da esso per determinare e salvare un profilo che sarà alla base del processo di *Segmentazione*. Inoltre questo profilo fornirà informazioni comparative sul comportamento del cliente. Il profilo ottenuto è costituito dalle seguenti informazioni:

1) **Totale Traffico e Fatturato**: Rispettivi totali del numero di operazioni fatturate ed ammontare della spesa effettuata.

2) **Percentuale Traffico e Fatturato**: Valori relativi in percentuale al traffico e fatturato complessivo dell'azienda.

3) **Entropia Operazioni**: Indice rappresentante l'eterogeneità delle operazioni svolte, calcolato per mezzo della formula di *Shannon* [22].

4) **Operazioni Rilevanti**: Identificativi dell'operazione più frequente, quella più redditizia e fatturato medio per operazione. Informazioni non utilizzate per la seguente procedura di clusterizzazione, ma sono riportati a scopo comparativo.

Il processo costruisce e archivia un profilo per ogni cliente che presenta almeno una transazione all'interno dell'intervallo temporale ricoperto dal modello.

Segmentazione

Basandosi sui loro profili statistici, i clienti sono raggruppati al fine di stabilire insiemi correlati da caratteristiche comuni. Questo processo è svolto tramite un *Centroid Based Clustering* per mezzo dell'algoritmo K-Means (Sezione 2.3.1). Tale processo deve tener conto di molteplici attributi presenti nel profilo del cliente, infatti si attua una clusterizzazione n-dimensionale. Viene creata una matrice $M_{cluster}$ di dimensione $n \times C$, dove n è il numero degli attributi e C il numero dei clienti ossia la popolazione di partenza. In base a come sono stati costruiti i profili statistici, la clusterizzazione viene effettuata su 3 attributi: **Percentuale Fatturato** S, **Percentuale Traffico** T e **Entropia** E, creando i vettori:

$$V_c = \{S_c, T_c, E_c\} \qquad \forall \text{ cliente } C$$

L'insieme di questi vettori costituisce la matrice $M_{cluster}$, input dell'algoritmo di clustering. L'algoritmo richiede come unico parametro il numero di cluster k. Per valutare il miglior valore di k, l'algoritmo viene eseguito $\forall k$ $s.t.$ $2 < k \leq \frac{C}{3}$, ricordando che C è la cardinalità della popolazione. Per ogni cluster ottenuto si calcola sia la misura interna **SSE** (Sum of Squared Errors) che il coefficiente di **Slhouette** (Sezione 2.3.1), in seguito si analizza la distribuzione di entrambi gli indici, scegliendo il valore di k più idoneo. Al termine del raggruppamento, all'interno di ogni profilo statistico, si inserisce il riferimento al cluster di appartenenza.

4.1.2 Fase Predittiva

La fase predittiva consiste nell'analizzare il traffico di un cliente particolare, dato un intervallo temporale specifico e proporre delle previsioni sul suo futuro comportamento, sfruttando le informazioni contenute nel modello: si ottengono dal modello tutti i clienti appartenenti allo stesso cluster del cliente selezionato. Sempre sfruttando il modello, si ottengono le regole di associazione di questo sottoinsieme di clienti; successivamente si analizza il comportamento del cliente in esame, solamente riguardo al range di date specificato. Tale analisi è analoga a quanto svolto nella fase di elaborazione del dataset per l'algoritmo Apriori durante l'analisi comportamentale (Sezione 4.1.1), ottenendo il vettore $V_\omega = \langle Op_i, \omega_i \rangle$. Sono generati tutti i possibili itemset costituiti dagli elementi presenti nel vettore. La previsione consiste nel proporre all'utente le **conseguenze** delle regole le quali **premesse** sono simili agli itemset generati. Tale similitudine è calcolata con il coefficiente di Jaccard (Sezione 2.4.2), secondo una soglia di tolleranza arbitraria. Segue l'algoritmo e tutte le sue procedure:

Algorithm 4 $Predict_{Op}$

1: **function** $Predict_{Op}(C, T, J)$
2: $Profilo\ P_C = CalcolaProfilo(C, T)$
3: $Lista\ L_{Vicini} = Modello \rightarrow GetVicini(C)$
4: $Lista\ Vicini_{regole} = Modello \rightarrow GetRegole(L_{Vicini})$
5: $Lista\ V_\omega = CreaVettore_\omega(C, T)$
6: $Lista\ Itemset_\omega = CalcolaItemset(V_\omega)$
7: $Lista\ Predizioni_C = CalcolaPredizioni(Itemset_\omega, Vicini_{regole}, J)$
8: **return** $Predizioni_C$

(1) ***Input(Cliente** C, **Timespan** T, **double** J)*: C è il cliente di cui si vuole prevedere il traffico e la relativa importanza delle operazioni; T è l'intervallo temporale in cui si vuole analizzare il comportamento del cliente C, J è la soglia che devono rispettare due itemset per essere ritenuti simili, espressa per mezzo del *coefficiente di Jaccard*.

(2) ***CalcolaProfilo(Cliente** C, **Timespan** T)*: Calcolo del profilo statistico del cliente indicato relativo all'intervallo T.

(3) ***Model** → **GetVicini(Cliente** C)*: Restituisce, direttamente dal modello, la lista dei clienti appartenenti al solito segmento del cliente C.

(4) ***Model.GetRegole(lista** L_{Vicini})*: Restituisce, direttamente dal modello, tutte le regole di associazione appartenenti ai vicini del cliente C.

(5) ***CreaVettore**$_\omega$**(Cliente** C, **Timespan** T)*: Restituisce il vettore bidimensionale V_ω, ossia la struttura contenente tutte le operazioni ed il loro indice di importanza ω, calcolate nell'intervallo T.

(6) ***CalcolaItemset(lista** V_ω)*: Dato il vettore in input, crea tutte le possibili combinazioni che si possono ottenere con le operazioni caratterizzate da $\omega \neq 0$.

(7) ***CalcolaPredizioni(lista** I, **lista** R, **double** J)*: Viene computato il coefficiente di Jaccard tra tutti gli elementi di I e le **premesse** di R. Si selezionano le **conseguenze** delle regole in R, il cui coefficiente supera la soglia J. Questo insieme è processato estraendo tutte le operazioni distinte ed il peso ω associato. Nel caso di occorrenze multiple della stessa operazione, si prende l'operazione appartenente alla regola che massimizza il lift.

(8) ***Output***: Si restituiscono le operazioni predette e la relativa importanza.

Le previsioni che si propongono al cliente consistono in una lista di coppie $\langle (Op_i, \omega), C_i \rangle$ contente l'operazione, specificando il relativo impatto sul fatturato e sul traffico ed anche il riferimento al cliente C_i da cui è stata generata.

4.2 Validazione del Modello

Il processo di validazione si applica ad un intervallo temporale che corrisponde ad una porzione del dataset che non è ricoperta dal modello predittivo esposto in precedenza. Per ogni cliente, iterativamente si applica la procedura per il calcolo delle previsioni, basata su un **intervallo antecedente** al periodo specificato. In questo modo si ottiene un numero P di previsioni:

$$P = \sum_{c \in C} P_c$$

dove C è l'insieme dei clienti e P_c le previsioni del cliente c ottenute.
Successivamente, si analizza il periodo in input per cogliere effettivamente il comportamento di ogni cliente. Per ogni coppia (operazione, importanza) si verifica se essa è stata effettivamente svolta con la stessa importanza predetta, ottenendo l'insieme T:

$$T = \sum_{c \in C} T_c$$

dove C è l'insieme dei clienti e T_c le previsioni corrette del cliente c ottenute.
Ricordando quanto detto sulle metodologie di validazione (Sezione 2.3.1), T_c costituisce l'insieme dei *True Positive* relativi al cliente c.
L'insieme dei *True Negative* N si ottiene dal complemento rispetto a P:

$$N = P \setminus T.$$

Segue il solito ragionamento applicato al cliente c:

$$N_c = P_c \setminus T_c.$$

Grazie a queste informazioni possiamo calcolare l'indice di **Precisione** per ogni cliente ed anche la precisione globale $Precisione_{TOT}$ e la precisione media $Precisione_{AVG}$ del modello:

$$Precisione_c = \frac{T_c}{T_c + N_c}$$

$$Precisione_{TOT} = {}^{T}/_{P}$$

$$Precisione_{AVG} = \frac{\sum_{c \in C} P_c}{C}$$

Il procedimento di validazione si riassume con il seguente algoritmo:

Algorithm 5 *Validation*

1: **function** $Validation(T,\ J,\ k)$
2: $lista\ C = getClientiDistinti(T)$
3: $timespan\ T' = T.getIntervalloPrec(T.days \cdot k)$
4: $int\ P = 0$
5: $int\ T = 0$
6: $double\ Precision_{AVG} = 0.0$
7: **for all** $c\ in\ C$ **do**
8: $lista\ P_c = Predict_{Op}(c, T', J)$
9: $P \leftarrow P + |P_c|$
10: $lista\ V_\omega = CreaVettore_\omega(c, T)$
11: $lista\ T_c = P_c.intersect(V_\omega)$
12: $double\ Precision_c = {}^{T_c}/_{P_c}$
13: $Precision_{AVG} \leftarrow Precision_{AVG} + Precision_c$
14: $T \leftarrow T + |T_c|$
15: $double\ Precision_{TOT} = {}^{T}/_{P}$
16: $Precision_{AVG} \leftarrow Precision_{AVG}/|C|$
17: **return** $\langle Precision_{TOT}, Precision_{AVG} \rangle$

La funzione *getClientiDistinti(T)* restituisce tutti i clienti che possiedono almeno una transazione all'interno del *timespan* T, mentre la funzione *getIntervalloPrec(n)* applicata ad un intervallo di date T, restituisce l'intervallo di ampiezza n antecedente a T.

La validazione dei risultati dovrà essere affidata anche ad esperti del dominio: per verificare l'affidabilità delle regole estratte e delle previsioni ottenute, si effettuano dei prelevamenti a campione ed il personale addetto al marketing e fatturazione dovrà verificare la veridicità di tale comportamento da parte del cliente in esame.

Capitolo 5

Processamento dei Dati

La realizzazione del modello e soprattutto delle strutture ad esso necessarie, ha necessitato di una laboriosa fase di analisi dei dati. Considerando più generalmente le varie fasi affrontate per la realizzazione del processo analitico, esse possono essere raccolte in quello che viene definito il processo di **Knowledge Discovery and Data Mining** o più brevemente **KDD**.

5.1 Step del Processo KDD

Il processo KDD (Figura 5.1) consiste in una procedura non banale suddivisa in step, la quale partendo dai dati, nella loro forma più grezza, li analizza ed elabora per renderli idonei a determinate procedure di Data Mining, allo scopo di identificare e validare pattern validi, comprensivi ed utili. Il processo è composto da 4 step principali ed essendo un processo **iterativo**, ad ogni fase, in base ai risultati ottenuti, è spesso necessario tornare ad uno step precedente:

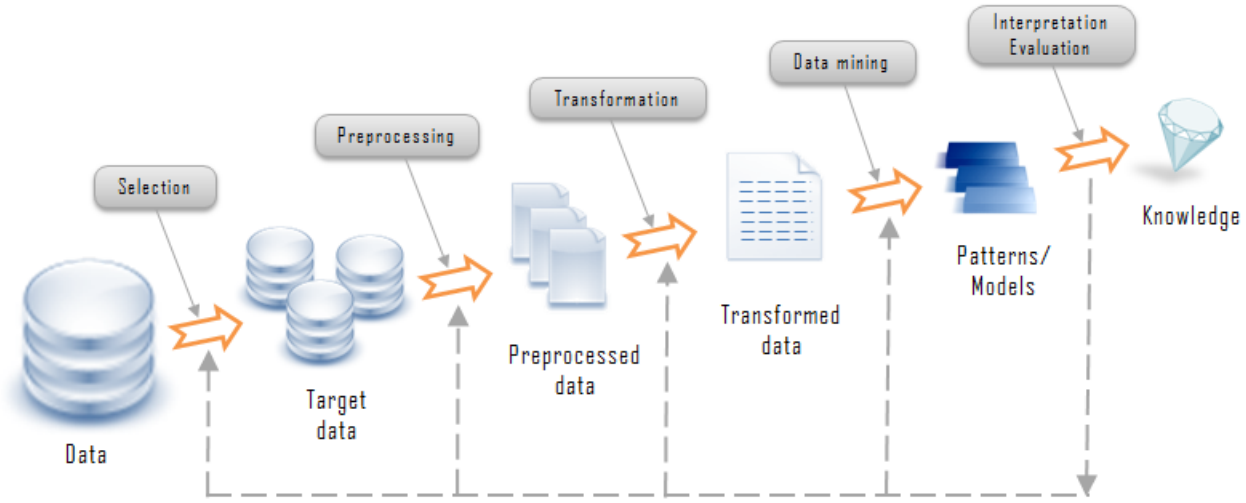


Figura 5.1: *Processo Knowledge Discovery and Data Mining [25]*

- **Consolidazione dei Dati**: A partire dai dati si determina una lista preliminare degli attributi necessari per l'analisi e della loro ubicazione, per poi essere consolidati assieme in una base dati dedicata.
  In seguito si effettua una procedura di *"pulizia"*: i dati sono analizzati una prima volta per testarne la validità e rimuovere dati incoerenti, ossia se sono presenti errori all'interno delle informazioni (dati mancanti, frammentati oppure formato incompatibile). Si conclude questa fase con una prima valutazione della quantità di dati ottenuta, qualora le dimensioni risulterebbero proibitive, si procede con analisi più selettive.

- **Selezione e Preprocessamento**: Si estraggono dei campioni di dati per valutare la ridondanza delle informazioni e cercare possibili combinazioni tra gli attributi ottenuti, principalmente per mezzo di funzioni di aggregazione. Si valutano in seguito possibili semplificazioni o specializzazioni sui range ricoperti dagli attributi. Si valutano anche possibili normalizzazioni dei dati, trasformazione di attributi continui in ordinali, intervalli o percentuali.

- **Data Mining**: L'obiettivo di questa fase è l'individuazione di pattern e modelli.

- **Interpretazione e Valutazione**: L'ultimo step è il più delicato in quanto richiede di valutare la qualità delle informazioni ottenute. Si possono usare strumenti statistici per valutare i risultati. In caso di predizione o classificazione è necessario eseguire dei test su apposite partizioni di dati per verificare, per mezzo di molteplici indici, la veridicità delle predizioni. Risulta molto importante anche la validazione da parte di esperti del dominio; infatti, la combinazione tra procedure di data mining e la conoscenza del contesto, permette di combinare strumenti informatici e statistici con conoscenze accurate, consentendo al complessivo processo KDD di raggiungere la sua massima efficienza.

Come accennato precedentemente, il processo è iterativo. Ad ogni step può emergere la necessità di eseguire dei passi indietro: durante il preprocessamento può emergere la necessità di attributi addizionali per eseguire combinazioni più adeguate e di conseguenza sarà necessaria un'altra fase di consolidazione; applicando procedure di data mining spesso sono necessarie elaborazioni come l'adozione di scale logaritmiche e conseguente elaborazione dell'input. Infine spesso le procedure di validazione non soddisfano le soglie minime, rendendo necessarie modifiche alle strategie di data mining utilizzate, come la revisione dei parametri del modello la sua eventuale modifica.

5.2 Analisi ed elaborazione dati

La prima fase del processo KDD descritta in precedenza, corrisponde a quanto svolto nella Sezione 3.3, ossia all'estrazione delle informazioni e alla costruzione di una base dati relazionale in grado di ospitare sia le informazioni che compongono il modello, che il traffico complessivo e le relative informazioni descrittive di tutti i clienti e dei loro piani tariffari.
La fase di preprocessamento dei dati si è rivelata più impegnativa, in quanto sono emerse problematiche come presenza di outliers tra i clienti e tra le date di fatturazione. In seguito verranno mostrate le procedure applicate per questa particolare fase.

5.2.1 Analisi del Traffico Clienti

Sono presenti in totale 631 clienti distinti, per un totale di 1299 filali. Analizzando la distribuzione delle transazioni per ogni cliente, è emerso che esistono

molti clienti obsoleti (traffico antecedente al 01/01/2014) e molti occasionali, i quali possiedono basso volume di traffico nell'arco dell'intero dataset.

Numero Clienti	**Numero transazioni**	**Inclusi nel modello**
311	0 (clienti obsoleti)	NO
122	da 1 a 99	NO
17	da 100 a 199	NO
60	da 200 a 999	SI
66	da 1001 a 4999	SI
18	da 5000 a 9999	SI
10	da 10000 a 19999	SI
12	da 20000 a 49999	SI
2	da 50000 a 99999	SI
6	oltre 100000	SI

Questa caratteristica è stata esposta e conseguentemente discussa con esperti del dominio, introducendo un parametro per contraddistinguere i clienti che non sono rilevanti a causa del numero limitato di operazioni effettuate. Sono stati esclusi i clienti con meno di 200 transazioni nell'arco delle date ricoperte dal dataset. Il tutto avviene durante la creazione dei profili statistici: il profilo statistico e le regole di associazione sono ugualmente estratti ma sono esclusi dalle procedure di clustering e conseguentemente di predizione (Sezione 3.3.1).
Il numero di clienti effettivamente inseriti nel modello è di 174.

5.2.2 Analisi delle Date Fatturazione

Procedura per la determinazione dell'intervallo temporale necessario all'input dell'algoritmo Apriori. In primo luogo è stato estratto il totale delle transazioni presenti ed è stato calcolato il numero di operazioni effettuate dall'insieme di tutti i clienti, raggruppate per data di fatturazione. Successivamente, sono state campionate per ogni data, il numero di transazioni associate, ottenendo il grafico seguente:

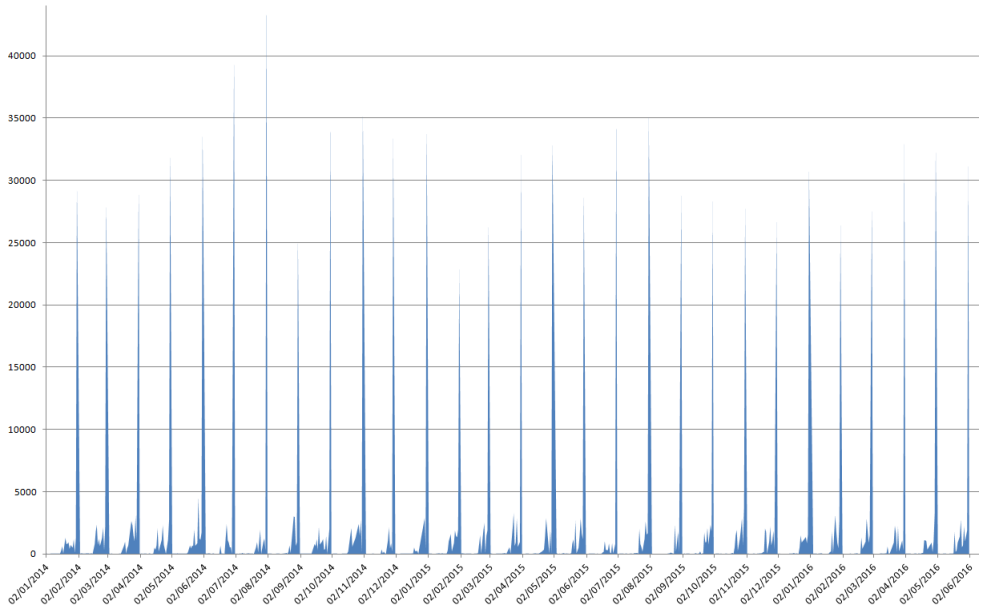


Figura 5.2: *Plotting traffico totale rispetto alle date*

Si evidenzia un comportamento particolare nei dati, ovvero esiste una periodicità nelle date in cui si effettuano le fatturazioni. Per comprendere il perché di questo comportamento è stato necessario l'ausilio degli addetti alla fatturazione. Tale comportamento è dovuto dal fatto che l'azienda tende a fatturare periodicamente con cadenze abbastanza regolari in base al traffico svolto dal cliente. Per ogni cliente si accumula il traffico effettuato in un periodo che va dai 7 ai 21 giorni, esistono tuttavia delle eccezioni, ossia date nelle quali sono fatturati pochissimi servizi. Anche questo comportamento ha trovato spiegazione: alcuni servizi offerti dall'Azienda, consistono nell'anticipo di tasse doganali (Dazio, IVA) e conseguente rimborso da parte del cliente interessato. Trattandosi di un costo passivo, questi rimborsi sono fatturati istantaneamente. Sono state estratte le date con un numero inferiore a 10 operazioni, ottenendo le seguenti informazioni:

- Sono presenti 671 date distinte
- Le date con un traffico < 10 sono 109 ossia $\sim 16.2\%$ delle date, per un totale traffico di 637 operazioni, meno dello 0.0005%.

- Le operazioni presenti in questo insieme, sono principalmente ad addebito passivo, eccetto alcuni casi isolati che possono essere trascurati in quanto ricoprono una percentuale irrilevante sul totale:

Operazione	**Frequenza**	**Tipologia**
Diritti Doganali su bolla IMZ	515	Passivo
Diritti Doganali bolle interne	48	Passivo
Rimborso dazio	18	Passivo
Terminal Gate e Security Fee VTE Import	11	Passivo
Terminal Gate e Security Fee VTE Export	9	Passivo
Spese addizionali Import per contenitore Genova S.Benigno	9	Passivo
Emissione temporanea esportazione	8	Passivo
Spese bancarie	7	Attivo
Port Fees per contenitore	6	Attivo
Svincolo presso la compagnia	6	Attivo

Ricapitolando, per queste operazioni è l'Azienda che, a causa degli iter doganali, è costretta ad anticipare le tasse per velocizzare le procedure di sdoganamento merci, per poi ricevere rimborso da parte del cliente. A causa di questo comportamento, non si possono considerare le singole date di fatturazione, ma si effettua il calcolo della media, in giorni, tra le date di fatturazione di ogni cliente, ottenendo un valore I_{Gap} il quale determinerà l'ampiezza degli intervalli temporali per la costruzione delle strutture necessarie all'algoritmo Apriori (Sezione 4.1.1). L'insieme delle operazioni nell'intervallo andrà a costituire lo "scontrino" del cliente come previsto dal modello Market Basket.

5.2.3 Analisi Operazoni

In seguito è stato analizzato l'insieme delle transazioni allo scopo di capire quali sono le tipologie più frequenti e quelle più redditizie e determinare la funzione per il calcolo di ω, indice rappresentante l' *"importanza"* dell'operazione, combinando sia l'impatto sul traffico che sul fatturato (Sezione 4.1.1). Sono presenti 51 tipologie di operazioni distinte. Il traffico totale estratto ammonta a 1.357.433 transazioni per un totale fatturato pari a 234.821.386,00€. Segue la lista delle voci fatturazione e la loro influenza sul traffico totale:

Operazione	**Totale Fatturato**	**% Totale Fatturato**	**Frequenza**	**% Frequenza**	ω
Diritti Doganali su Bolla IMZ	122575359	84,854	10570	0,85	42,852
Operazione Definitiva Container	4428859	3,066	80374	6,44	4,753
Operazione Doganale Container	3051125	2,112	130165	10,44	6,276
Diritti Doganali Bolle Interne	2656072	1,839	121047	9,7	5,7695
Messa a Disposizione per Verifica Radiometrica	2649323	1,834	11665	0,94	1,387
Diritti Doganali Bolle Esterne	1005708	0,696	66585	5,34	3,018
Emissioni in T1	879285	0,609	21103	1,69	1,1495
Rilascio Certificato Radiometrico	870370	0,603	7873	0,63	0,6165
Scarico Bolla	852753	0,59	66315	5,32	2,955
Memorandum Entrata e Ordinativo Imbarco	649847	0,45	105909	8,49	4,47
Diritti Doganali su Bolla T1	447024	0,309	29782	2,39	1,3495
Nostra Assistenza Verifiche Doganali Scanner o Radiometriche	373991	0,259	12964	1,04	0,6495
Costi Introduzione Estrazione Deposito IVA	359509	0,249	2644	0,21	0,2295
Spese Export Control System	342141	0,237	110744	8,88	4,5585
Svincolo Presso la Compagnia	290365	0,201	54949	4,41	2,3055
Formalita' Sanita' Marittima (escluso diritti	290182	0,201	11127	0,89	0,5455
Fuori Orario Doganale	289797	0,201	24713	1,98	1,0905
Port Fees Export	281787	0,195	61924	4,96	2,5775
Terminal Gate & Security Fee VTE Export	255709	0,177	51341	4,12	2,1485
Terminal Gate & Security Fee VTE Import	253010	0,175	30891	2,48	1,3275
Uscita Contenitore	249937	0,173	22605	1,81	0,9915
Port Fees per Contenitore	222703	0,154	51922	4,16	2,157
Introduzione Merci Pericolose in Parco IMO Genova	188637	0,131	8808	0,71	0,4205
Spese Addizionali Export per Container Genova S.Benigno	145880	0,101	53402	4,28	2,1905
Emissione Certificati Circ. EUR1 Standard	81883	0,057	13786	1,11	0,5835
Spese Addizionali Import per Container Genova S.Benigno	78383	0,054	23324	1,87	0,962
Singoli Successivi Import	78143	0,054	6341	0,51	0,282
Fidejussione per Deposito IVA	72797	0,05	1691	0,14	0,095
Uffici Divieti o Scarico Licenza	71302	0,049	368	0,03	0,0395
Costo Chimico del Porto La Spezia	57027	0,039	835	0,07	0,0545
Emissione Temporanea Esportazione	48813	0,034	1499	0,12	0,077
Spese Bancarie	47650	0,033	24584	1,97	1,0015
Assistenza Verifiche Doganali e Scanner	44736	0,031	1546	0,12	0,0755
Controllo Documentale Export	35489	0,025	1982	0,16	0,0925
Controllo Documentale Import	29402	0,02	1679	0,13	0,075
Deposito Garanzia e presa in carico Bolla Doganale	29095	0,02	1366	0,11	0,065
Rimborso Dazio	25724	0,018	970	0,08	0,049
Diritto Esecuzione Mandato	21638	0,015	4427	0,35	0,1825
Singoli Successivi Export	19870	0,014	673	0,05	0,032
Emissione Certificati Circ. ART1 Standard	17325	0,012	3277	0,26	0,136
Allibramento T1	17102	0,012	2137	0,17	0,091
Utilizzo ns Garanzia	12958	0,009	624	0,05	0,0295
Emissione Bonifico Bancario	10694	0,007	2366	0,19	0,0985
Nostra Assistenza Formalita' Introduzione Estrazione Deposito	8099	0,006	368	0,03	0,018
Scarico DAA	8059	0,006	941	0,08	0,043
Nostra Assistenza Uscita Merci IMO	7572	0,005	291	0,02	0,0125
Proroga Delivery Order	6688	0,005	1203	0,1	0,0525
Plico Posta Nave	6385	0,004	878	0,07	0,037
Scarico Licenze/Agrex	3948	0,003	463	0,04	0,0215
Emissione Certificati Circ. EUR1 Mexico	3506	0,002	324	0,03	0,016
Emissione P/C	176	0,0001	10	0,01	0,00505

Tali dati mostrano come alcune tipologie di operazioni siano molto influenti sia a livello di traffico che di ammontare fatturato. Analizzando tali dati con i responsabili marketing è stato valutato che entrambi gli attributi, traffico e fatturato, sono equamente importanti. Si è pensato di utilizzare, come misura unica ω, la media aritmetica tra le percentuali relative al traffico ed al fatturato, creando un indice unico, ma soprattutto normalizzato.

Si ricorda che per ogni operazione sono creati due vettori bidimensionali, dove Op_i rappresenta l'$i-$esima operazione e F_i, G_i rispettivamente la sua frequenza ed il suo guadagno relativi:

$$V_{frequenze} = \langle Op_i, F_i \rangle \qquad V_{guadagni} = \langle Op_i, G_i \rangle$$

Ottenendo il vettore:

$$V_\omega = \langle Op_i, \omega_i \rangle \qquad \text{con} \quad \omega = \frac{F_i + G_i}{2}$$

Analizzando questo nuovo indice si è evidenziato che, fatta eccezione per l'operazione *"Diritti doganali su bolla IMZ"* (che ricopre oltre l'80% del fatturato), le rimanenti operazioni hanno un indice di importanza in un range $0 < \omega \leq 6$. Questa analisi è stata effettuata a livello globale e per mezzo della consulenza di esperti nel dominio, sono state create quattro categorie: *Nulla, Bassa, Media, Alta.*

- **Nulla**: ω nullo, l'operazione non è stata mai svolta
- **Bassa**: $0 < \omega \leq 1\%$
- **Media**: $1\% < \omega < 4\%$
- **Alta**: $\omega \geq 4\%$

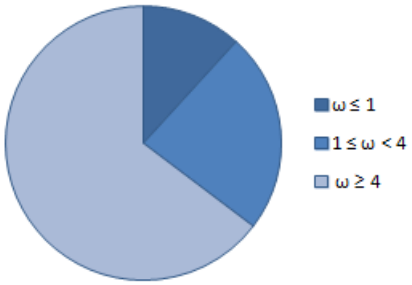


Figura 5.3: *Distribuzione delle categorie sulle operazioni globali*

Inoltre, sempre in concordanza con esperti del dominio, dato che il tipo e l'ammontare di traffico svolto varia da cliente a cliente, l'indice ω non è stato determinato globalmente per ogni operazione, ma esso è stato valutato singolarmente per ogni cliente: per ogni tipologia di operazione svolta, si misura la sua incidenza sul traffico e sul fatturato relativo al periodo indicato. Questo aspetto è molto rilevante perché non è detto che una determinata operazione risulti importante allo stesso modo per tutti i clienti.

La funzione utilizzata per il calcolo di ω è stata predisposta per essere facilmente sostituibile all'interno del modello, ed il software è stato predisposto per l'utilizzo di funzioni differenti, tuttavia l'affidabilità delle funzioni utilizzate per svolgere questo compito, rimane a discrezione dell'utilizzatore, dato che si tratta di un dato molto sensibile, per il quale spesso è necessario eseguire procedure di analisi dei dati e validazione per capirne l'efficienza.

Capitolo 6

Utilizzo e Risultati Ottenuti

Il software è predisposto per l'integrazione con i sistemi attualmente in vigore nel sistema aziendale. Tali software prevedono l'autenticazione obbligatoria da parte dell'utente il quale può assumere differenti ruoli: **Admin**, **Utente Fatturazione**, **Utente Operativito** e **Utente Semplice**.

6.1 Utilizzo

L'accesso al software è predisposto solamente per **Admin** e **Utente Fatturazione** e le rispettive funzionalità sono mostrate nel diagramma in Figura 6.1:

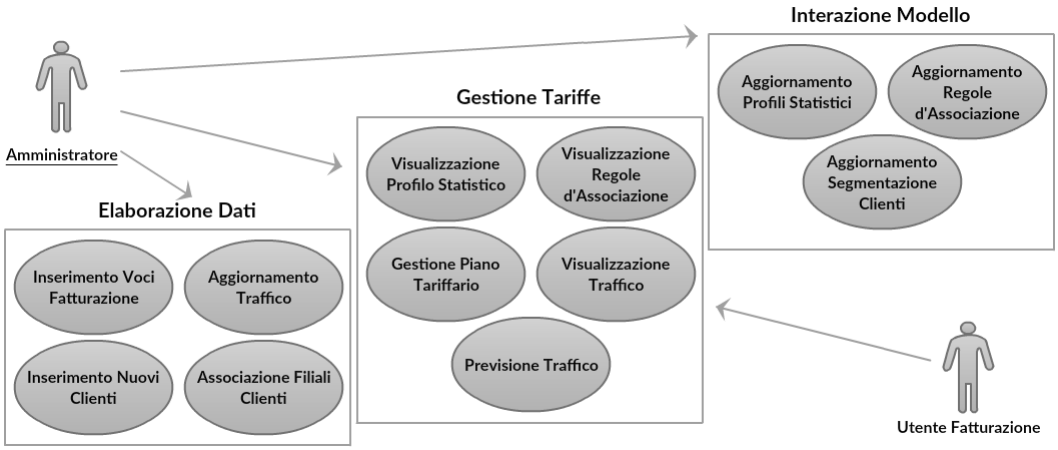


Figura 6.1: *Diagramma Casi d'Uso relativo agli utenti esistenti*

6.1.1 Gestione Tariffe

L'*utente fatturazione* potrà selezionare un cliente e consultare sia le informazioni statistiche, che il comportamento del cliente:

- Visualizzare i clienti presenti nel sistema, ossia quelli inclusi nella costruzione del modello. Inoltre esso potrà selezionare un qualsiasi cliente per visualizzare il profilo statistico globale e tutte le regole di associazione estratte dalle sue transazioni, con relativi indici di validazione: *Frequenza*, *Probabilità* e *Validità* (ossia supporto, confidenza e lift), inoltre è specificata anche l'importanza della regola[1] (Figura 6.2).

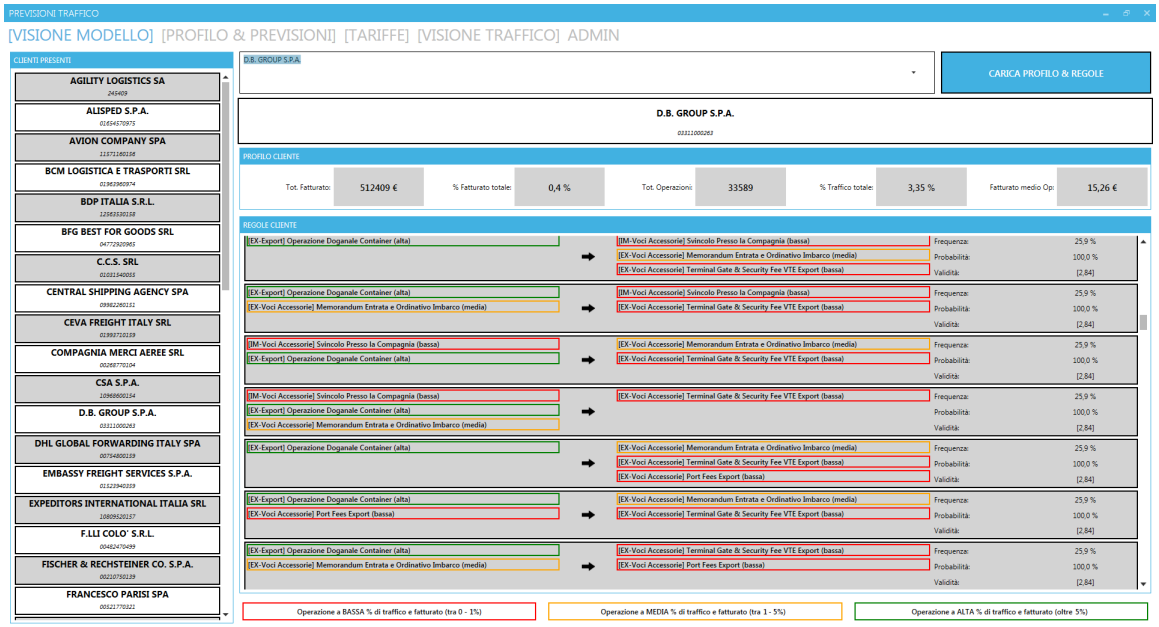


Figura 6.2: *Interfaccia utilizzo sezione per la visione profilo e regole cliente*

- Visionare, modificare ed inserire piani tariffari personalizzati. L'utente è in grado di inserire la tariffa associata ad ogni operazione, includere ed escludere opportuni servizi non richiesti e generare il preventivo, specificando l'azienda che deve comparire nell'intestazione (Figura 6.3).

- Visionare il traffico di un determinato cliente, specificando il periodo. L'utente è in grado di esaminare anche le singole transazioni effettuate

[1] Indice ω stimato con la funzione esposta nel Capitolo 5.2.3

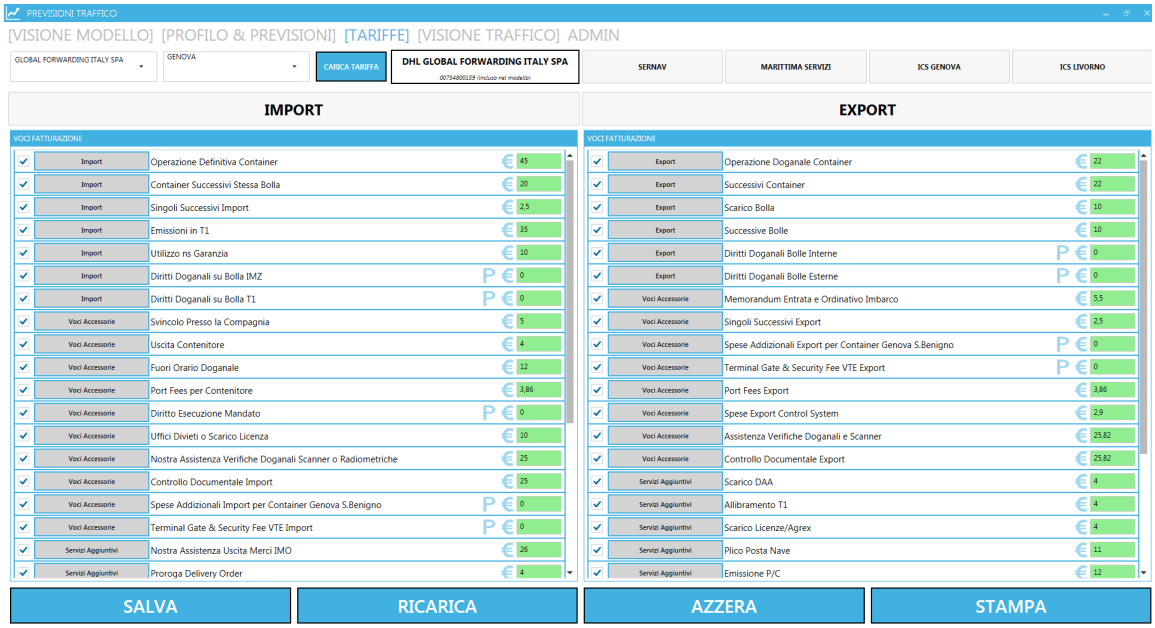


Figura 6.3: *Interfaccia utilizzo sezione gestione tariffe*

nel range di date desiderato. Potrà anche visualizzare il costo e la provincia relativa alla filiale a cui è stata fatturata tale operazione (Figura 6.4).

- Visionare il profilo statistico, sia relativo a un periodo che globale, e la tipologia di traffico svolta. Si visualizzano i grafici indicanti le operazioni più redditizie e quelle più frequenti svolte dal cliente nell'arco di tempo indicato (Figura 6.5).

- Calcolare le previsioni sul traffico, mostrando sia le operazioni ottenute in base al proprio comportamento, che quelle generate dai clienti simili. Per ognuna di esse è specificato il cliente da cui è stata estratta la previsione e sono mostrati tutti i clienti simili ottenuti (Figura 6.5).

6.1.2 Elaborazione Dati

Le procedure di elaborazione dati sono eseguibili solamente da utente autenticato come *amministratore*. Esse consistono nella gestione della clientela, delle

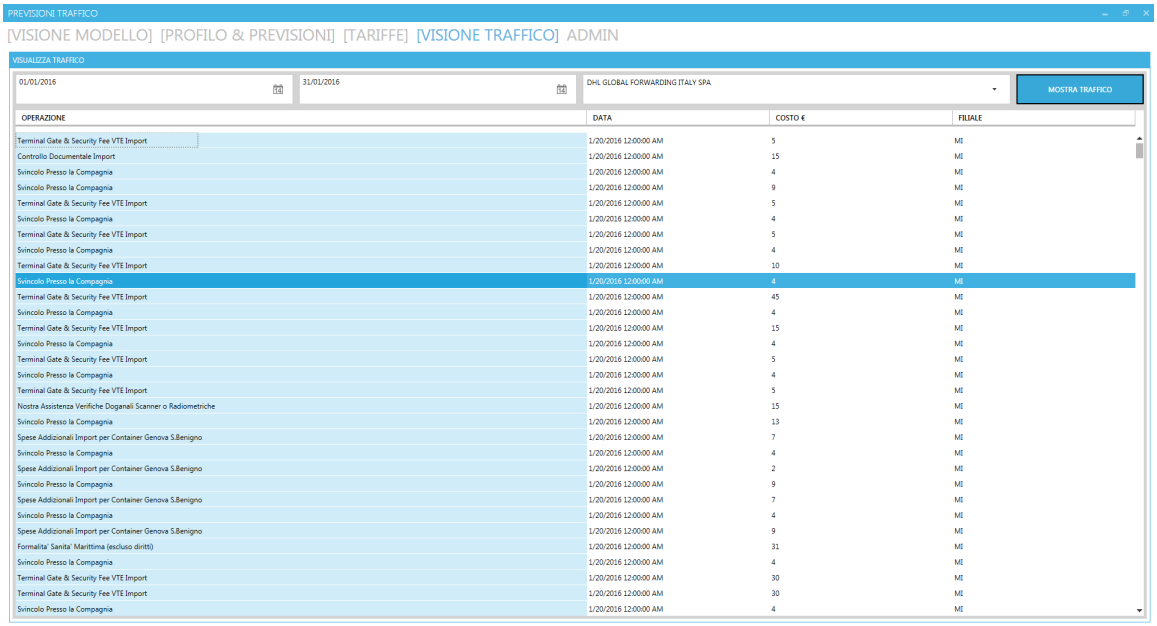


Figura 6.4: *Interfaccia visualizzazione traffico relativo al cliente*

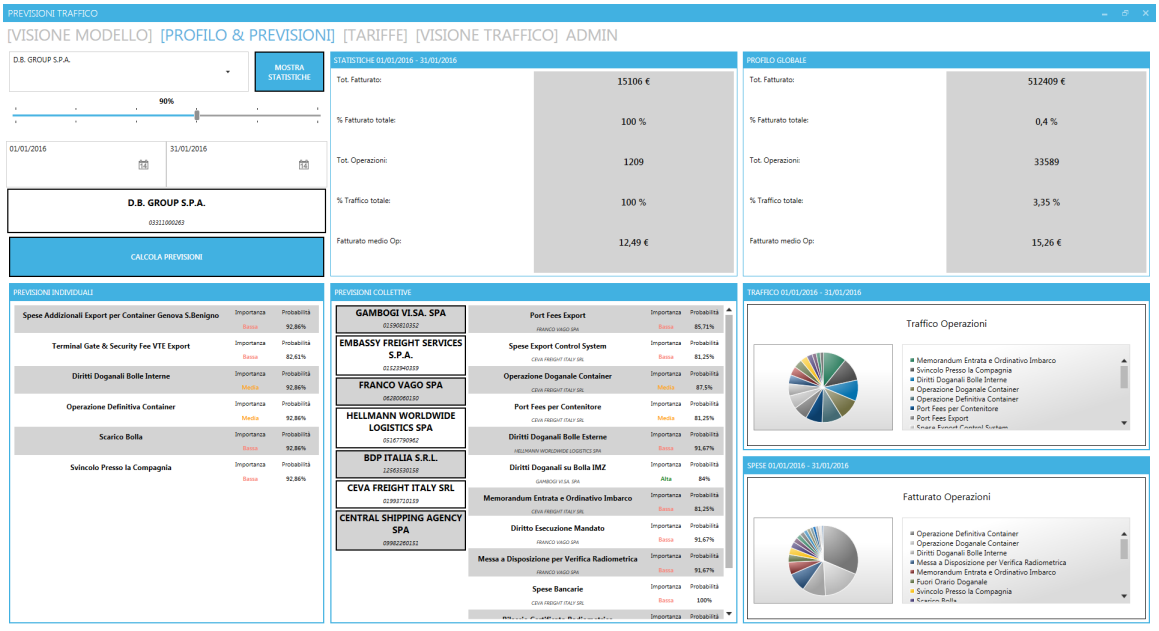


Figura 6.5: *Interfaccia utilizzo sezione per la visione statistiche e calcolo previsioni*

tipologie di voci fatturabili al cliente e nell'interazione con la base dati AS400 per l'aggiornamento del traffico complessivo (Figura 6.6). Nel dettaglio:

- Inserire un nuovo cliente, ed associare ad esso le sue filiali. In caso di nuove filiali, possono direttamente essere associate a clienti già presenti. Inserendo un nuovo cliente, automaticamente si genera un piano tariffario base.
- Inserire nuove voci di fatturazione ed i relativi codici AS400 associati, società e porto di imbarco/sbarco. In questo modo è creata una voce fatturazione che potrà essere rilevata dalla procedura di estrazione del traffico.
- Aggiornare i dati relativi al traffico. Impostando il range di date, si avvia la procedura di interazione con la base dati AS400 (Sezione 3.3), incrementando il volume di traffico presente nella base dati e la conseguente possibilità di effettuare previsioni basate sul comportamento del cliente nel periodo inserito.

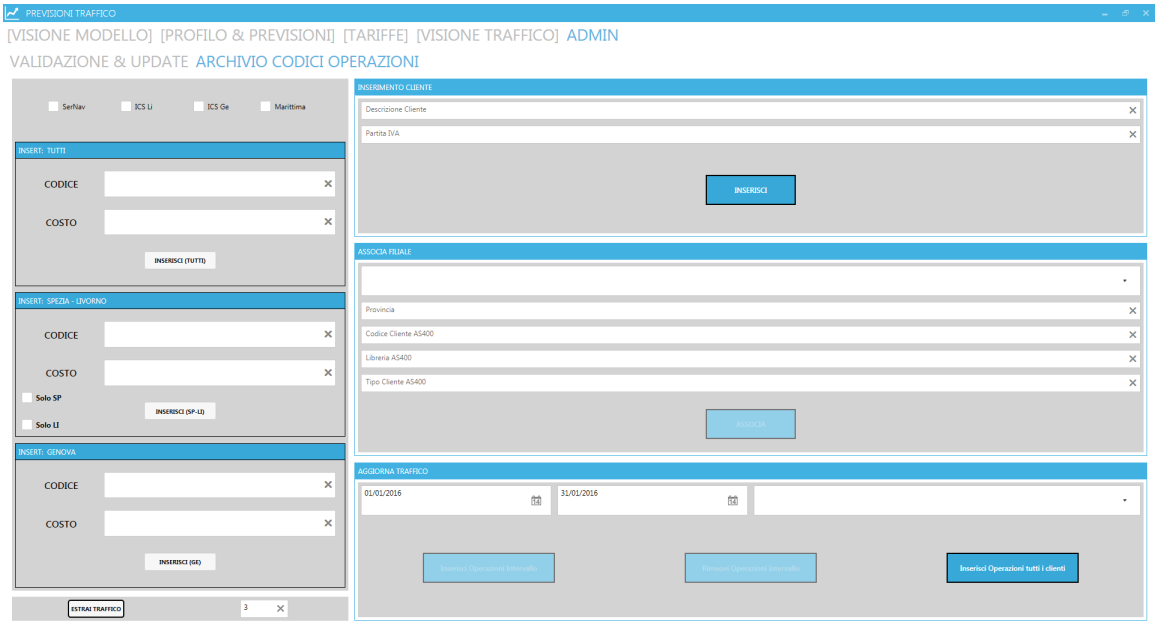


Figura 6.6: *Interfaccia inserimento clienti, associazione codici ed aggiornamento traffico*

6.1.3 Interazione Modello

L'amministratore, oltre che modificare il modello, sarà in grado di validare la sua efficienza su range di date arbitrari. Il processo di validazione viene eseguito per tutti i clienti e per ciascuno di essi si riporta la precisione media e quella globale ottenuta ed infine la precisione complessiva.
Il modello può essere modificato o aggiornato: specificando un cliente in particolare si possono aggiornare, o inserire se non presente, sia le sue regole di associazione che il suo profilo statistico, altrimenti è possibile eseguire questa operazione globalmente. In entrambi i casi è necessario specificare l'intervallo temporale e gli altri parametri input necessari al modello (Figura 6.7).

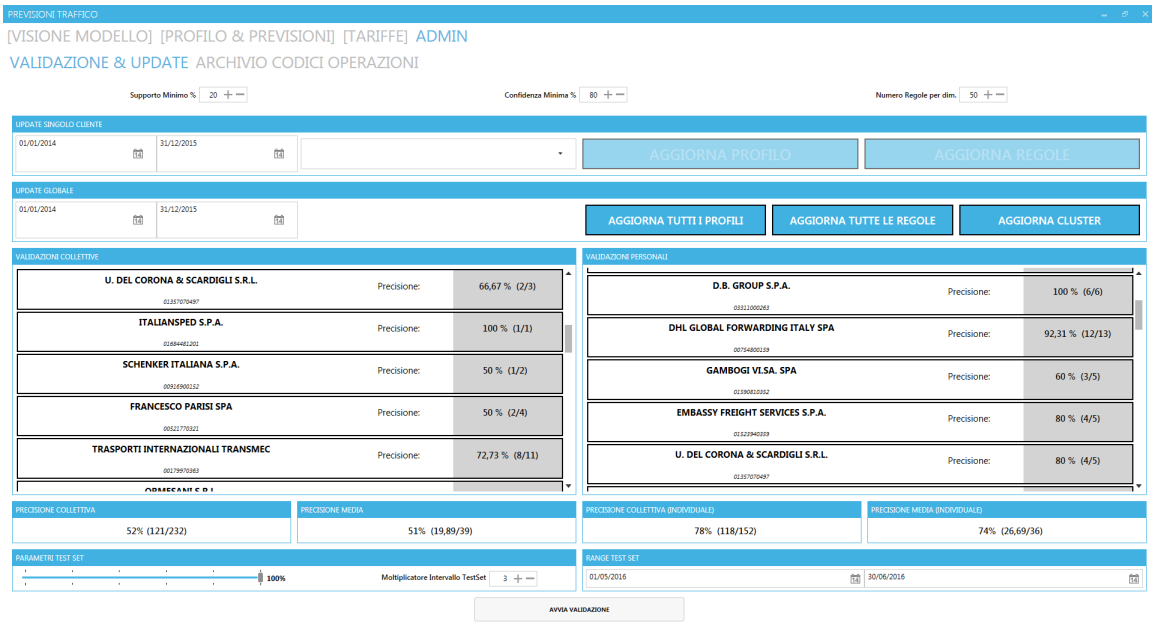


Figura 6.7: *Interfaccia utilizzo sezione update e validazione modello*

Inoltre è possibile aggiornare la segmentazione dei clienti, avviando la procedura di clustering. Questa procedura esegue l'algoritmo K-Means al variare del parametro k e al termine visualizza l'andamento degli indici **SSE** e **Silhouette** (Sezione 2.3.1). Attraverso un'opportuna interfaccia, l'amministratore potrà valutare e selezionare il valore di k ritenuto più idoneo (Figura 6.8).

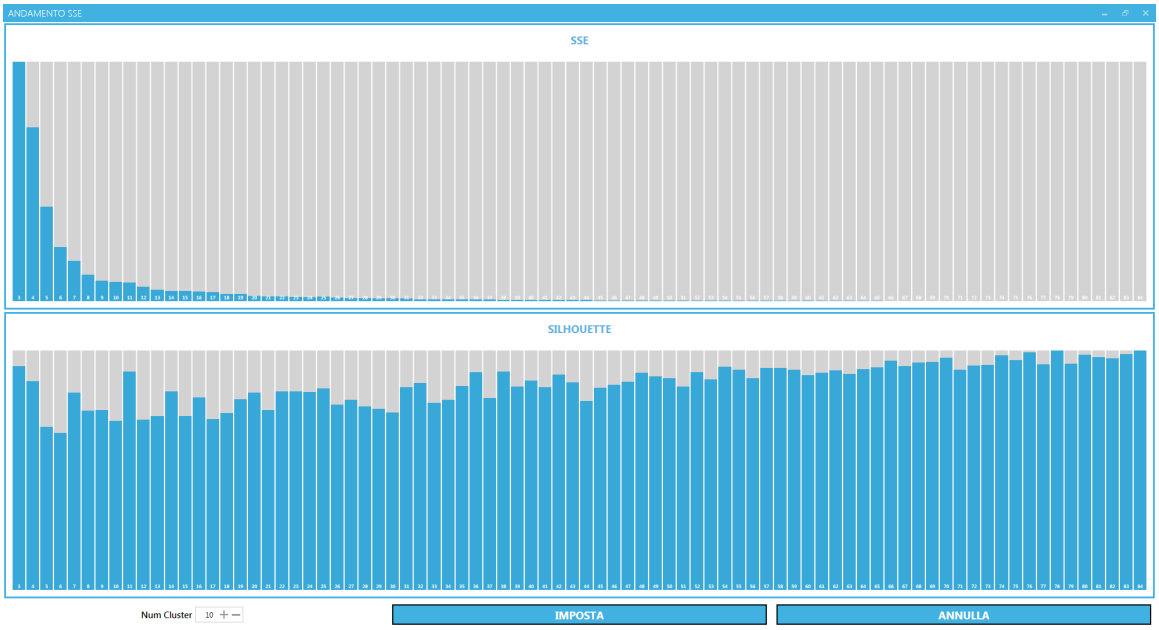


Figura 6.8: *Interfaccia valutazione del parametro k relativo all'algoritmo K-Means*

6.2 Risultati Ottenuti

Per l'interpretazione dei risultati è opportuna la conoscenza approfondita del dominio di appartenenza. Tuttavia in seguito saranno riportati i test di validazione effettuati e le relative percentuali di precisione ottenute: $Precisione_{TOT}$ e $Precisione_{AVG}$ (Sezione 4.2). Il modello utilizzato ricopre le date che vanno dal 01/01/2014 al 31/12/2015, le regole di associazione sono state estratte con *Supporto*= 20%, *Confidenza*= 80% e $K_{top} = 10$ ed il valore k relativo alla procedura di segmentazione (algoritmo *K-Means*) pari a 11. Per quanto riguarda la fase predittiva, si è utilizzato un coefficiente $J = 100\%$ (Sezione 4.1.2). La procedura di validazione ricopre complessivamente l'intervallo dal 01/01/2016 al 30/06/2016 (non ricoperto dal modello), questa procedura prende un intervallo temporale Int_{test}, calcola le previsioni relative al periodo Int_{valid} antecedente ad esso e verifica se esse risultano effettivamente presenti all'interno di Int_{test}. L'ampiezza di Int_{valid} è parametrica, il parametro specificato consiste in un moltiplicatore z e si calcola:

$$Int_{valid} = z \times Int_{test}$$

Le validazioni seguenti sono state effettuate con Int_{test} da 30 a 90 giorni e z da 1 a 5. Nel dettaglio sono stati scelti combinazioni di Int_{test} e z tali che non si andasse ad interagire con traffico antecedente al 01/01/2016, escludendo completamente la porzione di dataset utilizzata per la creazione del modello.

Validation Set		**Test Set**	
Date	Num. Mesi	Date	Num. Mesi
{01/06/2016} {30/06/2016}	1	{01/05/2016} {31/05/2016}	1
{01/06/2016} {30/06/2016}	1	{01/04/2016} {31/05/2016}	2
{01/06/2016} {30/06/2016}	1	{01/03/2016} {31/05/2016}	3
{01/06/2016} {30/06/2016}	1	{01/02/2016} {31/05/2016}	4
{01/06/2016} {30/06/2016}	1	{01/01/2016} {31/05/2016}	5
{01/05/2016} {30/06/2016}	2	{01/03/2016} {30/04/2016}	2
{01/05/2016} {30/06/2016}	2	{01/01/2016} {30/04/2016}	4
{01/04/2016} {30/06/2016}	3	{01/01/2016} {31/03/2016}	3

Oltre alla precisione, per ogni validazione, è riportato l'andamento della precisione relativa ai singoli clienti, sia per le previsioni personali (ottenute dal proprio comportamento), che per le previsioni collettive (ottenute dal comportamento dei clienti ritenuti simili):

- **Validation Set 1 mese: {01/06/2016} {30/06/2016}**
  Test Set 1 mese: {01/05/2016} {31/05/2016}
  (*Figura* 6.9)

Precisione P	Collettiva	Personale
P_{TOT}	46.58%	82.57%
P_{AVG}	44.49%	79.87%

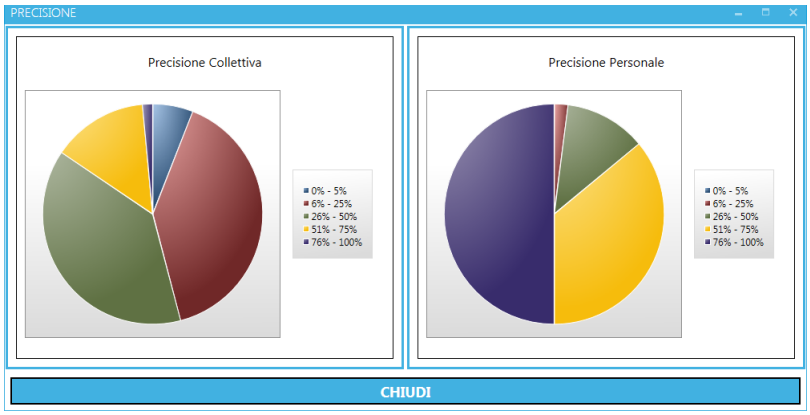


Figura 6.9: *Andamento Precisione [01/06-30/06 (2016)] - [01/05-31/05 (2016)]*

- **Validation Set 1 mese: {01/06/2016} {30/06/2016}**
  Test Set 2 mesi: {01/04/2016} {31/05/2016}
  (*Figura* 6.10)

Precisione P	Collettiva	Personale
P_{TOT}	48.21%	76.24%
P_{AVG}	47.11%	74.5%

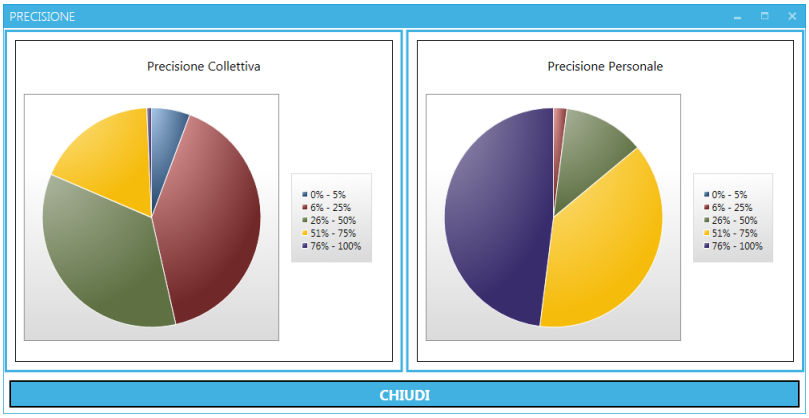


Figura 6.10: *Andamento Precisione [01/06-30/06 (2016)] - [01/04-31/05 (2016)]*

- **Validation Set 1 mese: {01/06/2016} {30/06/2016}**
  Test Set 3 mesi: {01/03/2016} {31/05/2016}
  (*Figura* 6.11)

Precisione P	**Collettiva**	**Personale**
P_{TOT}	48.81%	69.68%
P_{AVG}	47.91%	67.95%

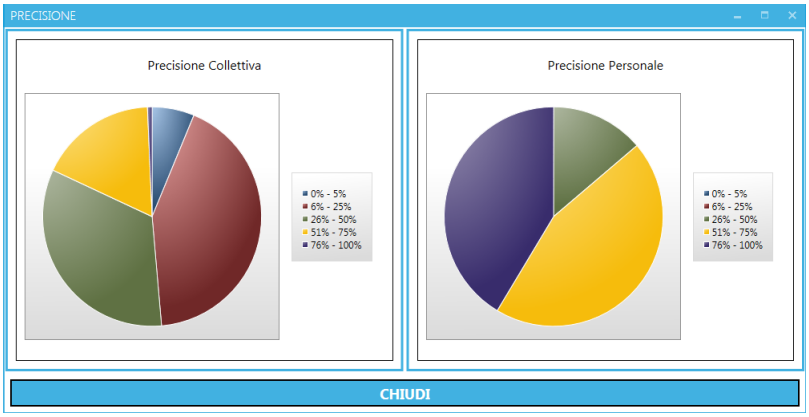


Figura 6.11: *Andamento Precisione [01/06-30/06 (2016)] - [01/03-31/05 (2016)]*

- **Validation Set 1 mese: {01/06/2016} {30/06/2016}**
  Test Set 4 mesi: {01/02/2016} {31/05/2016}
  (*Figura* 6.12)

Precisione P	**Collettiva**	**Personale**
P_{TOT}	49.72%	73.94%
P_{AVG}	48.79%	72.52%

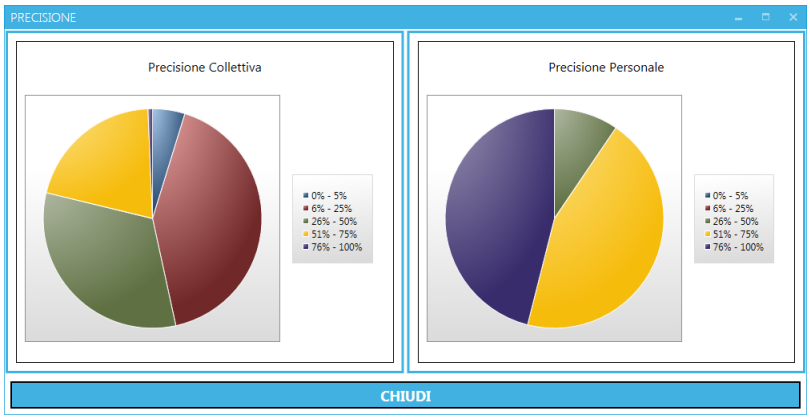


Figura 6.12: *Andamento Precisione [01/06-30/06 (2016)] - [01/02-31/05 (2016)]*

- **Validation Set 1 mesi: {01/06/2016} {30/06/2016}**
  Test Set 5 mesi: {01/01/2016} {31/05/2016}
  (*Figura* 6.13)

Precisione P	**Collettiva**	**Personale**
P_{TOT}	55.65%	72.71%
P_{AVG}	54.87%	70.63%

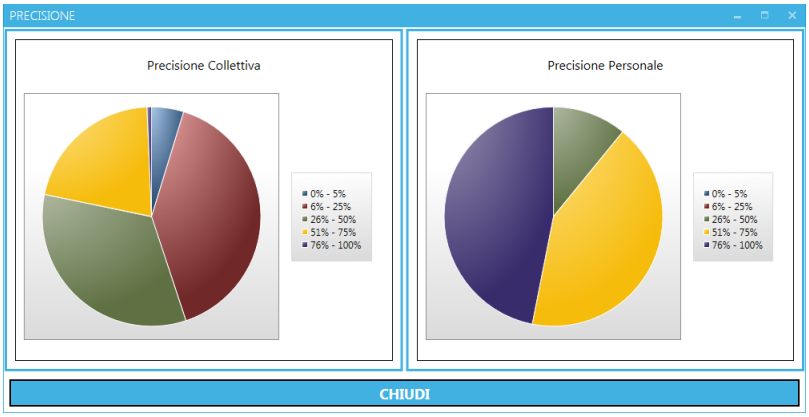


Figura 6.13: *Andamento Precisione [01/06-30/06 (2016)] - [01/01-31/05 (2016)]*

- **Validation Set 2 mesi: {01/05/2016} {30/06/2016}**
  Test Set 2 mesi: {01/03/2016} {30/04/2016}
  (*Figura* 6.14)

Precisione P	**Collettiva**	**Personale**
P_{TOT}	47.12%	69.97%
P_{AVG}	46.91%	68.65%

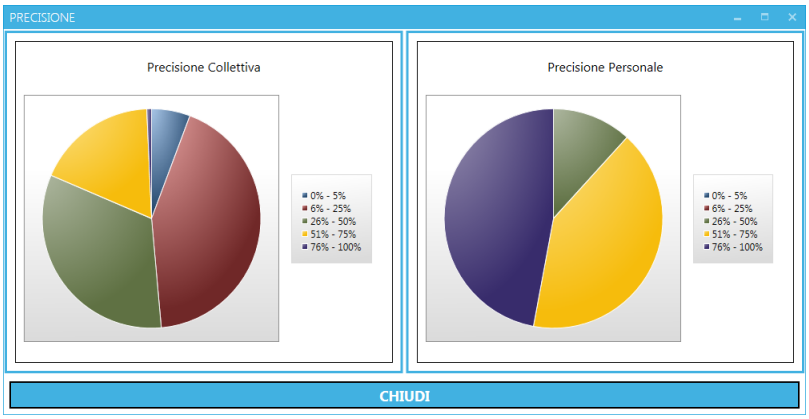


Figura 6.14: *Andamento Precisione [01/05-30/06 (2016)] - [01/03-31/05 (2016)]*

- **Validation Set 2 mesi: {01/05/2016} {30/06/2016}**
  Test Set 4 mesi: {01/01/2016} {30/04/2016}
  (*Figura* 6.15)

Precisione P	Collettiva	Personale
P_{TOT}	47.33%	72.56%
P_{AVG}	45.82%	70.79%

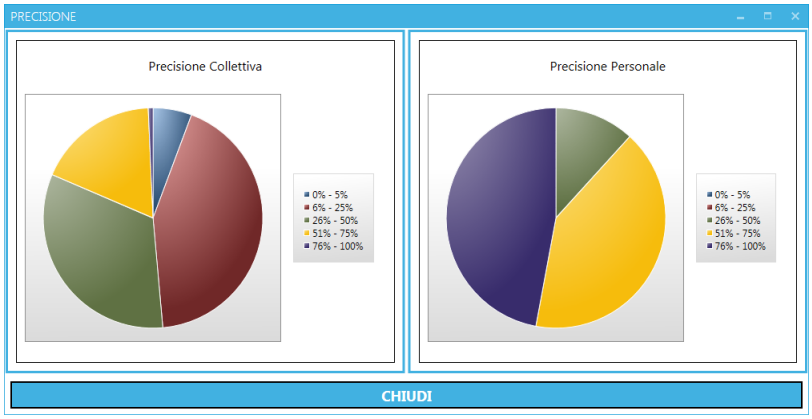
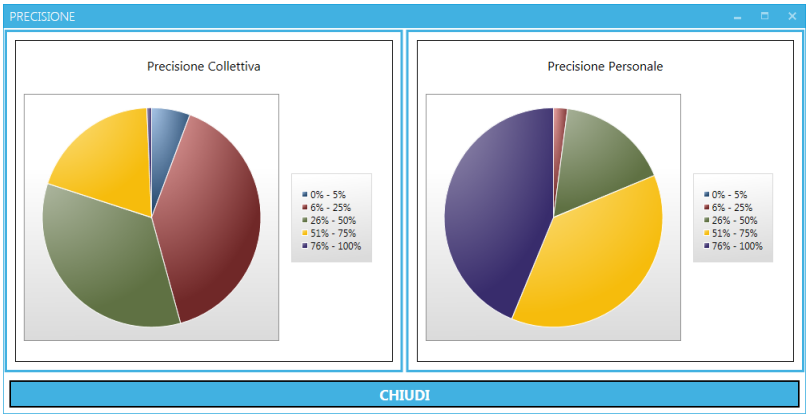


Figura 6.15: *Andamento Precisione [01/05-30/06 (2016)] - [01/01-31/05 (2016)]*

- **Validation Set 3 mesi: {01/04/2016} {30/06/2016}**
  Test Set 3 mesi: {01/01/2016} {31/03/2016}
  (*Figura* 6.16)

Precisione P	Collettiva	Personale
P_{TOT}	55.76%	69.28%
P_{AVG}	53.61%	68.45%



1

Figura 6.16: *Andamento Precisione [01/05-30/06 (2016)] - [01/01-31/05 (2016)]*

In totale, per ogni validazione, sono state generate una media di $\sim$ 1900 previsioni collettive. Per $\sim$ 150 clienti su 174 è stata generata almeno una previsione, di conseguenza per $\sim$ 24 clienti non sono state individuate previsioni collettive. Per quanto riguarda le previsioni personali, ne sono state generate in media $\sim$ 200 per ogni validazione, per $\sim$ 70 clienti. Si deduce che risulta più difficile ottenere una previsione generata dal comportamento personale del cliente, tuttavia analizzando la precisione complessiva, emerge che questo tipo di previsione risulta in media più affidabile, soprattutto per quanto riguarda intervalli di tempo più ristretti, contrariamente alle previsioni collettive, le quali riportano precisione maggiore rispetto ad intervalli temporali più ampi.

6.3 Ambiente di Sviluppo

Il software è stato sviluppato durante tirocinio universitario svolto presso l'azienda Ser.Nav s.r.l. Sono state impiegate le tecnologie messe a disposizione dall'azienda, la quale opera principalmente in ambiente di sviluppo Microsoft utilizzando il Framework .NET alla versione 4.5. Il software consiste in un'applicazione Windows Presentation Foundation, mentre il linguaggio di programmazione utilizzato è Visual Basic .NET. Queste decisioni sono state prese non per ragioni di efficienza, motivo per cui l'implementazione in lin-

guaggi come Java o C++ avrebbe dato risultati migliori a causa della grande quantità di librerie opensource disponibili. La ragione principale riguarda l'integrazione: il software è stato infatti predisposto per potersi integrare perfettamente agli strumenti attualmente in vigore in azienda ed essere inserito come funzionalità aggiuntiva.
Per quanto riguarda gli algoritmi utilizzati, è stata necessaria l'implementare dell'algoritmo Apriori, in quanto si necessita di una sua particolare versione in grado di gestire non solo l'item ma anche la sua importanza ed inoltre l'ambiente di sviluppo utilizzato non offre librerie opensource che implementato questo algoritmo e che si integrano efficientemente con il progetto.
La costruzione dei grafici statistici è eseguita per mezzo della libreria WPF Extended Toolkit [11], con licenza *Microsoft Public License (Ms-PL).*
L'algoritmo K-Means utilizzato è contenuto nella libreria Accord Framework .NET [12], con licenza: *GNU Lesser General Public License, version 2.1.*
Infine lo stile grafico dell'applicazione è stato realizzato grazie al toolkit open source Mahapps.Metro [24] con licenza *Microsoft Reciprocal License (MS-RL).*

Capitolo 7

Conclusioni

La tesi descrive l'attività di tirocinio svolto presso l'azienda Ser.Nav s.r.l., un'agenzia marittima e spedizioniere doganale. Ogni cliente dell'azienda ha piani tariffari a durata annuale, il rinnovo di queste tariffe consiste nella stipulazione di un piano tariffario in concordanza con le esigenze del cliente. La valutazione delle tariffe più idonee è effettuata dal personale addetto alla sezione marketing, con l'ausilio di software in grado di calcolare dati statistici sul traffico interrogando la base dati AS400, nella quale sono archiviate tutte le fatture emesse ed i relativi dettagli. L'interazione con questi strumenti tuttavia è molto delicata, poiché l'interrogazione di AS400 per statistiche troppo elaborate, può influire sulle sue prestazioni e di conseguenza rallentare l'operatività dell'azienda. Il progetto svolto offre un sistema di valutazione e previsione delle tariffe autonomo, il quale sfrutta un processo di data mining apposito per la previsione del traffico e l'individuazione di clienti simili tra loro. Il software creato è destinato proprio al personale dedito al marketing, il quale, data l'esperienza e la conoscenza del contesto, potrà usufruirne per calibrare al meglio i piani tariffari da applicare ai propri clienti.

7.1 Sviluppi Futuri

La base dati è in continua evoluzione, per il momento il traffico presente copre un intervallo di 30 mesi: dal 01/01/2014 al 30/06/2016. L'espansione di questa finestra temporale è in continua esecuzione, infatti è previsto

l'inserimento sia del traffico relativo all'anno 2013 che il traffico futuro che ricoprirà la seconda metà dell'anno 2016.

7.1.1 Inserimento Sequential Pattern Mining

L'ampliamento dell'intervallo temporale permette l'arricchimento del modello, per quanto riguarda la fase predittiva, con l'introduzione di *Sequential Pattern Mining*[1]. L'introduzione di questa strategia prevede un'opportuna fase di analisi dei dati e preprocessamento, per calibrare al meglio i parametri degli algoritmi necessari. Per ogni cliente saranno presenti non solo le principali regole di associazione, ma anche i pattern sequenziali più rilevanti, estratti secondo indici di validità e vincoli temporali opportunamente valutati.

7.1.2 Specializzazione del Clustering

Possono essere introdotti altri attributi su cui basare il processo di clustering, informazioni che non si basano interamente su dati statistici relativi al traffico svolto.
I dettagli presenti nel database di AS400, relativi alla fattura emessa, comprendono riferimenti alla tipologia di merci importate o esportate, inoltre sono presenti anche le informazioni riguardo al privato che ha usufruito della compagnia di trasporti affiliata per effettuare lo spostamento del materiale. L'estrazione di questi ulteriori dettagli rende possibile specificare una lista delle **tipologie di merci gestite** da ogni cliente, con relativi indici di rilevanza. Inoltre le informazioni sul privato comprendono anche dati riguardo la locazione delle merci, da dove provengono o dove sono destinate, rendendo possibile la creazione di un **profilo geografico** relativo all'operatività cliente.
Queste informazioni sono molto utili ed interessanti ma altrettanto delicate. La loro estrazione prevede L'impegno di risorse dedicate sia umane che fisiche ed un conseguente investimento da parte dell'azienda, per questo motivo sono state proposte come possibili sviluppi futuri.

[1] Tecnica predittiva esposta nella Sezione 2.4.2

Ringraziamenti

Ringrazio l'azienda Ser.Nav s.r.l. per avermi concesso l'opportunità di svolgere l'attività di tirocinio. In particolare ringrazio tutto l'ufficio informatico, composto da: Gianluca Schiavina, Simone Pagni, Stefano Colombo, Enrico Piras e Giacomo Galletto, i quali hanno sempre fornito il loro aiuto e soprattutto la loro migliore compagnia.
Ringrazio tutti i docenti del Corso di Laurea Magistrale Informatica, per aver contribuito alla mia formazione, in particolare ringrazio la Prof.ssa Anna Monreale per avermi costantemente seguito e aiutato sia durante lo svolgimento del progetto che nella stesura del documento di tesi.
Infine ringrazio tutti i miei amici e compagni di corso, il loro appoggio è stato fondamentale per superare le difficoltà incontrate durante il percorso di studio.

Bibliografia

[1] R. Trasarti, R. Guidotti, A. Monreale, F. Giannotti, "MyWay: Location prediction via mobility profiling". In: *Information System www.elsevier.com/locate/infosys*, 2014.

[2] Mikolaj Morzy, "Prediction of moving Object Location Based on Frequent trajectories". In: *The* 21^{st} *International Symposium on Computer and Information Sciences ISCIS*, 2006.

[3] Michael J. Shaw, Chandrasekar Subramaniam, Gek Woo Tan, Michael E. Welge, "Knowledge management and data mining for marketing". In: *Decision Support System 31 127-137*, 2001.

[4] Gediminas Adomavicius, Alexander Tuzhilin, "Using Data Mining Methods to Build Customer Profiles". In: *IT Professional Editorial Calendar*, 2001.

[5] S.M.H. Jansen, "Customer Segmentation and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behavior". In: *Master thesis in Operations Research at the University of Maastricht UM*, 2007.

[6] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Data Mining Techniques: For Marketing, Sales, and Customer Support". In: *John Wiley & Sons, Inc. New York, NY, USA*, 1997.

[7] Michael J. Berry, Gordon Linoff, In: *"Introduction to Data Mining"*, 2004.

[8] E.W.T. Ngaia, Yong Hub, Y.H. Wonga, Yijun Chenb, Xin Sunb, "The application of data mining techniques in financial fraud detection: A

classification framework and an academic review of literature". In: *Decision Support Systems, on quantitative methods for detection of financial fraud*, 2011.

[9] https://en.wikipedia.org/wiki/Customer_analytics, "Customer analytics" 2016.

[10] https://en.wikipedia.org/wiki/Structure_mining, "Structure mining" 2016.

[11] http://wpftoolkit.codeplex.com/, "WPF Extended Toolkit" 2016.

[12] http://accord-framework.net/, "Accord Framework .NET" 2016.

[13] https://en.wikipedia.org/wiki/Unsupervised_learning, "Unsupervised learning" 2016.

[14] https://en.wikipedia.org/wiki/Spedizioniere, "Spedizioniere" 2016.

[15] https://en.wikipedia.org/wiki/Hierarchical_clustering, "Hierarchical clustering" 2016.

[16] Wray Buntine, "Learning classification trees". In *W. Stat Comput (1992) 2: 63. doi:10.1007/BF01889584* 1992.

[17] https://en.wikipedia.org/wiki/Bayes%27_theorem, "Bayes' theorem" 2016.

[18] Dr. Saed Sayad, "Support Vector Machine - Classification (SVM)". In: *http://www.saedsayad.com/support_vector_machine.htm* 2016.

[19] Jiawei Han, Micheline Kamber, Jian Pei, In *Data Mining: Concepts and Techniques* 2000.

[20] https://en.wikipedia.org/wiki/Affinity_analysis, "Affinity analysis (Market Basket Analysis)" 2016.

[21] https://it.wikipedia.org/wiki/IBM_System_i, "Application System/400" 2016.

[22] https://en.wiktionary.org/wiki/Shannon_entropy, "Shannon Entropy" 2016.

[23] https://en.wikipedia.org/wiki/IBM_DB2, "IBM DB2" 2016.

[24] http://mahapps.com/, "mahapps.metro a UI toolkit for WPF" 2016.

[25] http://www.rithme.eu/?m=solutions&p=kdprocess&lang=en, "Knowledge Discovery and Data Mining" Rithme Consulting, 2016 .