# ***CAPITOLO 3: DATA MINING PER IL GEO-POSITIONING***

Negli ultimi decenni, lo sviluppo di informazioni e comunicazioni delle tecnologie hanno danno nuova vitalità al marketing aziendale. I dati da immagazzinare ed analizzare stanno aumentando a un ritmo molto rapido, probabilmente 1000 volte rispetto a cinque anni fa. Tuttavia, i dati e gli utili aziendali non lo sono direttamente proporzionale.

La tecnologia Data mining nel marketing è un'applicazione relativamente universale. Queste applicazioni sono riferite a una Boundary Science, una varietà di teorie scientifiche basate principalmente sulle discipline di base dell’Information Technology, del Marketing e dello studio dei metodi Statistici che sta alla base di ogni possibile algoritmo. Inoltre, il data mining fa riferimento anche a discipline letterarie e comportamentali per valutare meglio le caratteristiche di un cliente, come la psicologia e la sociologia [20].

In generale, attraverso l’estrazione, il trattamento e lo smaltimento di una grande quantità di informazioni per identificare l’interesse, le preferenze e i comportamenti di specifici gruppi o dei singoli consumatori, le abitudini di consumo, ma soprattutto, la domanda, orientando le vendite per un marketing dal contenuto specifico.

Poiché l'automazione è popolare in tutto il settore, le imprese che gestiscono i processi devono avere molti dati operativi. I dati non sono raccolti allo scopo di analisi, ma provengono da operazioni commerciali. L'analisi di questi dati conferisce al decision-maker il valore reale delle informazioni, al fine di ottenere profitti.

Le informazioni commerciali provengono dal mercato attraverso vari canali come, ad esempio, il processo di acquisto tramite credito carta dove possiamo raccogliere i dati di consumo del cliente, come ora, luogo, beni o servizi interessanti interessati, prezzi voluti e il livello di capacità di ricezione. Inoltre, le imprese possono anche acquistare una varietà di informazioni sui clienti da altri società di consulenza.

Il marketing basato sul data mining di solito può creare sulle vendite delle promozioni specifiche per il cliente secondo i suoi precedenti record di acquisto.  Le più comuni applicazioni nel settore bancario, assicurativo, sistema di traffico, vendita al dettaglio e in campo commerciale.

Come già descritto nello Stato dell’arte, le tecnologie e le analisi del marketing sono basate sull’analisi del mercato, come la predizione, la segmentazione e la classificazione del cliente, il profiling e il cross-selling.

***Immagine che contiene testo

Descrizione generata con affidabilità molto elevata***

Figura 1: Applicazione Del Data Mining Per Il Marketing

Esse possono essere utilizzate anche per operazioni di valutazione del credito e frode.

Il processo di base del data mining nel marketing mostra come segue:

* *Preparare i dati primitivi*: Include informazioni di carattere individuale (età, sesso, hobby, background, professione, indirizzo, codice postale e reddito), la precedente esperienza di acquisto e la relazione all'interno dei clienti. La preelaborazione dei dati primitivi è molto importante per selezionare i potenziali clienti.
* *Stabilire un determinato modello*: Questo modello può utilizzare molto tecnologie tradizionali di data mining e molte tecnologie da altri argomenti correlati. Tuttavia, il problema che tali tecnologie dovrebbe risolvere è quello di individuare il mercato migliore o accettabile, all'interno fonte di dati limitata, tempo limitato e spese limitate.

In definitiva, utilizzare questo modello per selezionare i clienti e decidere il piano di marketing.

Nel nostro progetto andremo a idealizzare una possibile predizione dei dati ISTAT italiani del 2018 tramite la regressione lineare, in quanto adattabili solo fino al 2017 dal sito omonimo [25]. Inoltre, si svilupperà una classificazione CART per capire lee future aspettative dei negozi attualmente aperti, e per trovare una possibile ideale locazione per aprire un nuovo store.

## ***3.1*** ***R – PREDICTION: LINEAR REGRESSION PER I DATI ISTAT 2018***

Per una analisi completa e affidabile, è molto importante l’integrità e la completezza di dati. Dopo una ricerca in vari siti dedicati, si è scelto di analizzare e prevedere i dati dell’Istituto Nazionale di Statistica, comunamente denominato ISTAT.

I dati effettivi hanno un orizzonte che parte dal 2004 al 2017. Essendo nel 2019, I dati forniti non si possono considerare completi. Perciò, si è deciso di affrontare il problema, scegliendo 5 indicatori tra tutti quelli proposti a livello regionale, con un orizzonte temporale di dieci anni, considerando i dati dal 2007 al 2017 per fare una previsione riferita al 2018.

La ricerca per decidere gli indicatori adatti all’analisi, si è svolta seguendo una procedura matriciale, facendo un mapping diviso in aree geografiche e serie temporale coperta, inserendo il tutto in una tabella esplicativa mostrata come segue:

Tabella 1: Matrice Dati ISTAT

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TABELLA DATI ISTAT | INDICATORI / TITOLI | COPERTURA | | | | | TEMPO | |
| **Italia** | **zone** | **regione** | **provincia** | **comune** | **dati fino al 2017** | **dati fino al 2018** |
| FAMIGLIE | Incidenza della povertà assoluta |  |  |  |  |  |  |  |
| Reddito familiare netto medio  (esclusi i fitti imputati) |  |  |  |  |  |  |  |
| Spesa media mensile familiare per beni  e servizi non alimentari |  |  |  |  |  |  |  |
| Spesa media mensile familiare totale |  |  |  |  |  |  |  |
| Trasporti | Autobus circolanti |  |  |  |  |  |  |  |
| Autovetture |  |  |  |  |  |  |  |
| Rete autostradale |  |  |  |  |  |  |  |
| Rete ferroviaria in esercizio |  |  |  |  |  |  |  |
| Trasporto di merci su strada |  |  |  |  |  |  |  |
| Lavoro | Tasso di disoccupazione |  |  |  |  |  |  |  |
| Tasso di disoccupazione giovanile |  |  |  |  |  |  |  |
| Macro-Economia | Consumi finali interni |  |  |  |  |  |  |  |
| Investimenti fissi lordi |  |  |  |  |  |  |  |
| Pil pro capite |  |  |  |  |  |  |  |
| Territorio | Densità della popolazione dei comuni |  |  |  |  |  |  |  |
| Densità della popolazione dei comuni con superficie da 1.001 a 2.000 |  |  |  |  |  |  |  |
| Densità della popolazione dei comuni con superficie da 2.001 a 6.000 |  |  |  |  |  |  |  |
| Densità della popolazione dei comuni con superficie da 6.001 a 25.000 |  |  |  |  |  |  |  |
| Densità della popolazione dei comuni con superficie fino a 1.000 |  |  |  |  |  |  |  |
| Densità popolazione comuni con superficie superiore ai 25.000 |  |  |  |  |  |  |  |
| Permessi di costruire - abitazioni in nuovi fabbricati residenziali |  |  |  |  |  |  |  |
| Permessi di costruire - m2 utili abitabili in nuovi fabbricati residenziali |  |  |  |  |  |  |  |
| Popolazione residente media |  |  |  |  |  |  |  |
| Turismo | Totale arrivi |  |  |  |  |  |  |  |

Si è deciso di scegliere uno o massimo due indicatori per tipo a livello regionale, e di scegliere la popolazione residente e il turismo a livello comune.

Per eseguire la predizione, si è utilizzato un linguaggio dedicato denominato “R” è un linguaggio e un ambiente per il calcolo statistico e la grafica. È un progetto GNU e fornisce un'ampia varietà di modelli statistici (modellazione lineare e non lineare, test statistici classici, analisi di serie temporali, classificazione, clustering, ...) e tecniche grafiche ed è altamente estensibile. Il linguaggio R fornisce un'opzione Open Source per la partecipazione a tale attività, con il supplemento di R. Uno dei punti di forza di R è la facilità con cui è possibile produrre trame di qualità di pubblicazione ben progettate, compresi simboli matematici e formule dove necessario. È stata prestata grande attenzione alle impostazioni predefinite per le scelte di progettazione minori nella grafica, ma l'utente mantiene il controllo completo [27].

R-studio è un ambiente di sviluppo integrato per R, con una console, un editor di evidenziazione della sintassi che supporta l'esecuzione diretta del codice e strumenti per il tracciamento, la cronologia, il debug e la gestione dello spazio di lavoro [26].

Prima dell’implementazione del codice, è stato necessario creare dei fogli di lavoro univoci per ogni indicatore, favorendo una predizione efficace dei dati, data la diversità di origine degli stessi.

A livello Regionale si è deciso di scegliere:

* Settore Trasporti: Rete ferroviaria in esercizio.
* Settore Famiglia: Spesa media mensile familiare per beni e servizi non alimentari, reddito medio.
* Settore Macro-Economia: Pil Pro Capite.
* Settore Lavoro: Tasso di disoccupazione.

Mentre, a livello comunale, si è evidenziato la popolazione residente e il turismo.

La suddivisione, onde evitare meccanismi complicati e macchinosi ed eventuali errori di copiatura, di svolgerla tramite Talend Open Studio [28], software di ETL già ampiamente discusso nel capitolo precedente.

Il job Prende I dati completi precedentemente caricati nella Staging Area e tramite l’utilizzo di Query, Ogni tabella viene interrogata in modo da estrarre nel foglio Excel dedicato, solo i dati regionali inerenti alla nostra analisi.

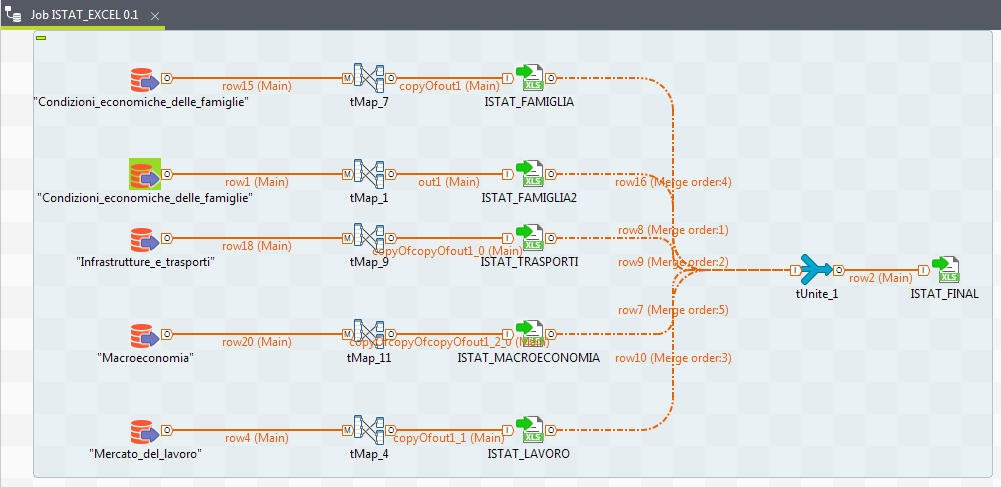


Figura 2: Job ISTAT Excel

Per esempio, per estrarre solo i dati relativi al reddito medio delle famiglie, si è considerato la tabella dedicata ai dati delle condizioni economiche delle famiglie dell’Istituto Nazionale di Statistica Italiano e si è estratto Il reddito Medio, evidenziandolo con la clausola SQL IN.. La stessa cosa è stata fatta per escludere i dati nazionali o relativi alle zone di appartenenza escludendo gli altri campi (Centro Nord, Nord-Est, …) con un NOT IN.

Naturalmente, le stesse operazioni sono state svolte per gli altri indicatori.

La Query usata nel caso del reddito è la seguente:

"***SELECT*** *OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.Settore,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.Indicatore,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.\"Modalitа\",*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.\"Unitа di misura\",*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.Territorio,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.Fonte,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2004,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2005,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2006,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2007,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2008,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2009,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2010,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2011,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2012,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2013,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2014,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2015,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2016,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2017,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.ANNO\_2018,*

*OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie.Note*

***FROM*** *OPEN\_DATA\_ITALY.Condizioni\_economiche\_delle\_famiglie*

***WHERE*** *Indicatore* ***in*** *('Reddito familiare netto medio (esclusi i fitti imputati)')*

***and*** *Territorio* ***Not in*** *('Nord-ovest',*

*'Bolzano/Bozen',*

*'Trento',*

*'Nord-est','Nord',*

*'Centro',*

*'Centro-Nord',*

*'Mezzogiorno',*

*'Italia')"*

Creando i File Excel univoci per ogni indicatore è ora possibile svolgere la Regressione Lineare.

## ***3.2 CLUSTERING: K-MEANS***

## ***3.3 CLASSIFICAZIONE: CART***