UNIVERSITA' POLITECNICA DELLE MARCHE

FACOLTA' DI **INGEGNERIA**

Corso di Laurea magistrale in **Ingegneria Gestionale**

Analisi del mercato dei pagamenti digitali tramite strumenti di Data Analytics

Analysis of the digital payment market through Data Analytics

Relatore: Chiar.mo
Prof. **Filippo Emanuele Ciarapica**

Tesi di Laurea di:
Caterina Vardè

Correlatore:
Ing. **Alberto Danese**

A.A. **2021 / 2022**

“Sto ancora imparando …”
Michelangelo B. a 87 anni

Indice

ELENCO DELLE TABELLE

ELENCO DELLE FIGURE

ELENCO DEGLI SCRIPT

INTRODUZIONE

I sistemi di pagamento esistono da secoli e hanno sempre avuto un importante ruolo all'interno delle società, indipendentemente dal tipo di trasferimento desiderato. L'evoluzione dei sistemi di pagamento negli anni, è stato considerevole: si è passati dal semplice e fisico baratto al complesso e astratto mondo delle innovazioni tecnologiche. L'avvento tecnologico e la globalizzazione hanno permesso l'evoluzione dei sistemi di pagamento.

Dagli anni Novanta la diffusione di internet e la rapida adozione dei dispositivi mobili ha fatto sì che anche i pagamenti si spostassero da un luogo fisico a uno digitale.

Nexi Group è la PayTech Europea che offre servizi e infrastrutture per il pagamento digitale per banche, aziende e pubblica amministrazione, compresi l'*issuing* (emissione di carte di pagamento) e l'*acquiring* (la fornitura dei servizi di accettazione dei pagamenti).

Nexi Group nasce dall'unione di Nexi, Nets e SIA tre dei maggiori player europei nel mercato dei pagamenti; gestisce circa 170 milioni di carte di pagamento e circa 2,2 milioni di punti vendita convenzionati.

All'interno del mondo dell'acquiring, che include sia la fornitura di POS fisici su cui i clienti appoggiano o strisciano la carta, sia lo sviluppo di payment gateway implementati sui siti di e-commerce, ***LAKA (Large Account Key Account)*** è una definizione interna a Nexi Group per individuare i merchant che hanno un transato superiore ai 24 milioni di euro. I LAKA oltre ad essere gli esercenti caratterizzati da grandi volumi di transato sono anche i merchant che hanno una **particolare visibilità** / **ruolo** a livello nazionale, rispetto agli SME (Small Medium Enterprice).

Il progetto di tesi, che è stato sviluppato analizza il mercato dei pagamenti digitali tramite le transazioni di Nexi con l'obiettivo di:

1. **Classificare i Laka,** sfruttando il processo informatico denominato LMS (**L**aka **M**onitoring **S**ystem), già presente in Nexi:
2. **Ricercare della similarità tra due Laka**, andando a creare degli indici di similarità tra coppie di Laka in base alle abitudini di acquisto dei titolari delle carte;

3. Applicare una cluster analysis, tramite un approccio gerarchico sugli indici di similarità, in modo da creare gruppi ”affini” di Laka su cui intraprendere azioni commerciali e marketing.

Capitolo 1

PAGAMENTI DIGITALI: NEXI

1.1 **Moneta elettronica**

L'antenata delle carte di pagamento che conosciamo oggi è riconducibile ai primi anni del 1900, alcune ditte Americane permettevano ai clienti più facoltosi di aprire delle linee di credito consegnando loro delle carte solitamente di cartone. L'anno che segna la nascita vera e propria della carta di credito è il 1950 quando **Diners Club Inc**. introdusse la prima carta di credito che poteva essere usata per acquistare una varietà di prodotti e servizi. [1]. Diners Club Inc. concedeva al titolare della carta fino a 60 giorni di credito per perfezionare il pagamento dei beni acquistati.

La prima banca ad implementare il sistema della carta di credito fu la Franklin National Bank di New York, che nel 1951 infatti introdusse le **"**Charge-It Cards **".** In seguito, anche altre banche capirono il potenziale

e così nel 1958 l'American Express mise in circolazione la sua prima carta "Don't leave home without it".

La tecnologia fu fondamentale per l'evoluzione delle carte di pagamento. Nel 1967 fu inventato l'ATM (Automated Teller Machine), sistema che permetteva di prelevare a tutte le ore del giorno dal proprio conto bancario denaro contante in modo automatico. Per prelevare negli ATM, serviva inizialmente un cheque monouso. Nel 1979, con i progressi nel campo dell'elettronica le carte di credito vengono dotate di una banda magnetica. All'esercente ora bastava passare la banda magnetica nel terminale per procedere al perfezionamento dell'acquisto. Dopo il 1980, negli Stati Uniti, vi fu una significativa crescita nell'utilizzo di carte di pagamento e nel 1983 il 43% delle famiglie possedeva una carta di credito mentre erano il 62% nel 1992[1].

In **Italia** la prima carta di credito fece la sua comparsa nel 1958 ed apparteneva al gruppo Diners Club. Alla fine degli anni Sessanta anche Bankamericard, American Express entrarono nel mercato italiano. Inizialmente, le carte di credito venivano usate principalmente da stranieri, che venivano in Italia per motivi turistici o di business, e da italiani appartenenti ad un'alta classe sociale.

Fino alla metà degli anni Ottanta le carte di credito presenti sul mercato italiano appartenevano ad associazioni e banche statunitensi.

È solo dal 1986, anno di nascita dei **Servizi Interbancari** (SI) in Italia, che il sistema interbancario italiano entra a far parte del mercato delle carte di credito. Servizi Interbancari diventò poi CartaSI S.p.a., oggi **Nexi Group** raggiungendo in poco tempo i vertici del mercato interno principalmente grazie al numero di banche aderenti ed alla diversità delle carte e dei servizi offerti, in grado di soddisfare tutte le esigenze; a distanza di oltre 30 anni Nexi Group è tuttora l'emittente di carte di pagamento principale in Italia, con un costante aumento del numero dei titolari.

1.2 **Evoluzione dei pagamenti digitali**

I pagamenti digitali sono sempre più diffusi in Italia, e vengono visti come un processo di pagamento semplice e sicuro. A fine 2021 hanno raggiunto i 327 milioni di euro con una crescita del 22% rispetto al 2020 [2].

L'effetto combinato della pandemia e degli incentivi ai consumatori quali il cashback ha accelerato il cambiamento di abitudini da parte degli

italiani. Gli italiani hanno fatto sempre più ricorso nel 2021 ai pagamenti digitali con preferenza per le carte contactless, con la paura di contatto fisico nella pandemia ha giocato a favore dei pagamenti con carte contactless negli acquisti in negozio (126,5 miliardi di euro il valore delle transazioni nel 2021). La crescita in assoluto più alta, è stata però registrata dai pagamenti con smartphone e i wearable che hanno superato infatti i 7 miliardi di euro, raddoppiando il loro valore rispetto al 2020 (+106%). Un successo da ricondursi alla combinazione di una serie di fattori: semplicità, velocità e utilità, percepita nell'uso quotidiano che portano gli utilizzatori di questi strumenti a preferirli rispetto ad altri metodi, poiché il mobile wallet va a sostituire il portafoglio.

Tra i player italiani che hanno beneficiato di questo trend troviamo Nexi Group che ha registrato in Italia nel 2021 un aumento del 122% delle transazioni con smartphone in negozio: la PayTech, nel 2021 ha gestito circa il 70% del totale transazioni mobile in store in Italia, pari a 7 miliardi di euro. L'86% di questo importo, in crescita rispetto all'84% del 2020, è stato generato dagli acquisti effettuati con app che prevedono la virtualizzazione della carta nello smartphone, come Google Pay, Samsung Pay e Apple Pay. Ne è ulteriore conferma la crescita del 50% delle carte

Nexi registrate su queste app di pagamento mobile. Il 2021 si è caratterizzato anche per un incremento importante delle soluzioni di incasso digitale più innovative: il 54% dei merchant clienti delle banche partner di Nexi ha attivato PayByLink, il servizio che consente agli esercenti di usare la posta elettronica o i canali social per inviare ai propri clienti un link di pagamento; il 32% si è dotato di un Mobile Pos, device che permette l'abbinamento immediato Pos-Smartphone. È poi cresciuta dell'85% l'adozione di Easy Delivery e Easy Calendar, soluzioni che consentono di ricevere ordini e pagamenti via social e di creare una vetrina digitale per gli appuntamenti pagando a distanza o sul posto. Mentre il numero di transazioni effettuate con carte Nexi è aumentato del 28% con un'accelerazione più marcata nell'ultimo trimestre dell'anno (+30%) [2].

I pagamenti digitali sono in crescita anche nel 2022 dato riportato nella 20^0 edizione dell'Osservatorio Carte di Credito e Digital Payment curato da Assofin, Ipsos e Nomisma, infatti nel primo semestre 2022 i pagamenti digitali continuano la loro crescita registrando un balzo del 24%.

Capitolo 2
LO STATO DELL'ARTE NEL MONDO DELLA DATA SCIENCE

L'obiettivo di questo capitolo è quello di fornire una panoramica teorica delle tecniche e degli algoritmi che sono stati utilizzati durante lo svolgimento del progetto, oltre a definire il concetto di Data Science.
Il termine Data Science viene spesso utilizzato come sinonimo di Intelligenza artificiale, si tratta però di due discipline distinte anche se interconnesse e spesso vi è una certa confusione anche nell'utilizzo di termini quali Machine Learning e Deep Learning. Talvolta usati impropriamente. Di seguito sono riportate le relative definizioni [3] [4]:

- *Intelligenza Artificiale*: si tratta di un termine generico che indica il campo dell'informatica dedicato alla creazione di sistemi che sono in grado di risolvere problemi e di riprodurre attività proprie dell'intelligenza umana. Spesso il termine intelligenza artificiale è abbreviato in AI.

- *Machine Learning*: l'apprendimento automatico è una applicazione dell'intelligenza artificiale che consente alle macchine di imparare dai dati, senza che queste siano programmate in maniera esplicita. L'algoritmo riceve una serie di dati ed è capace di apprendere, modificando e migliorando le predizioni mano a mano che riceve più informazioni su ciò che sta elaborando. Per questo motivo, gli algoritmi di Machine Learning tenteranno di minimizzare gli errori e massimizzare la probabilità che le loro previsioni siano vere. Spesso il termine Machine Learning è abbreviato in ML.
- *Deep Learning*: l'apprendimento profondo è un sottogruppo del Machine Learning. È basato su modelli e algoritmi computazionali che imitano l'architettura delle reti neurali biologiche nel cervello umano. Tali algoritmi costruiscono e utilizzano le cosiddette reti neurali artificiali, note anche con l'abbreviazione ANN. Spesso il termine Deep Learning è abbreviato in DL.
    - Data Science, è, un campo interdisciplinare che utilizza metodi, processi, algoritmi e sistemi scientifici per estrarre

conoscenza e informazioni da insiemi di dati che possono essere strutturati o non strutturati [5]. L'obiettivo di questa disciplina è sviluppare strategie e modelli per l'analisi dei dati con il fine di ottenere nuove informazioni che verranno poi sfruttate in altri ambiti.

La figura 2.1 mostra graficamente la gerarchia tra i concetti appena definiti.

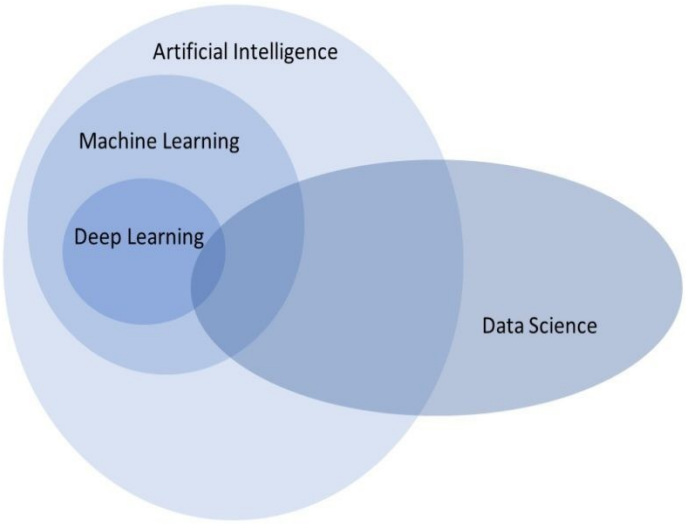


Figura 2.1 Schema sulla Data Science

La Data Science è dunque una disciplina trasversale a cui le aziende fanno riferimento per ottenere nuove informazioni sui propri clienti o sul mercato. Oggi i dati sono diventati uno degli aspetti più importanti

per le aziende, infatti più dati si raccolgono dall'attività di un cliente e più informazioni è possibile ottenere per migliorare il proprio prodotto o per ottenere vantaggi competitivi.

In questo lavoro di tesi abbiamo combinato le tecniche del Machine Learning che sono alla base per eseguire operazioni di pulizia dei dati, classificazione e previsione ed i processi di estrapolazione delle informazioni dai dati della Data science.

Nell'ambito del Machine Learning esistono diverse categorie di apprendimento, nei prossimi capitoli si descriveranno l'apprendimento supervisionato e l'apprendimento non supervisionato.

2.1 **Apprendimento supervisionato**

L'apprendimento supervisionato è una tecnica di apprendimento automatico che mira a istruire un sistema informatico in modo da consentirgli di elaborare automaticamente previsioni sulla base di una serie di esempi.

Nell'apprendimento supervisionato l'insieme di dati da utilizzare viene diviso in tre parti: training set, validation set e test set, come mostrato nella figura 2.2. Innanzitutto viene fornito all'algoritmo un

dataset di esempio, cioè il training set, nel quale i dati sono etichettati. Questo significa che gli esempi sono composti da una coppia di dati contenenti il dato originale e il risultato atteso, cioè la variabile di risposta. Il compito dell'algoritmo di Machine Learning è trovare la funzione che modelli la relazione tra i due, in modo da essere in grado di fare previsioni su nuovi dati in cui l'esito non è noto. Nel caso in cui la variabile di risposta sia di tipo categorica si parla di classificazione, invece nel caso di variabili di risposta continue si parla di regressione. Esistono diversi algoritmi di apprendimento automatico per creare funzioni che eseguano classificazioni o regressioni, ad esempio Support Vector Machine, Naive Bayes, Logistic Regression, Decision Trees, K-Nearest Neighbors e Random Forest. In seguito, il modello generato durante la fase di training viene validato sul validation set. Questo permette di selezionare il modello più performante, per esempio in termini di accuratezza. L'ultimo passaggio consiste nel valutare la performance del modello selezionato su un nuovo set di dati, il test set [8]. La suddivisione del dataset originale appena descritta è necessaria per garantire di non incorrere nel cosiddetto *overfitting*, con tale termine si indica la situazione in cui il modello generato risulta essere troppo

sensibile alle caratteristiche particolari del training set piuttosto che alle caratteristiche generali del problema che si sta analizzando.

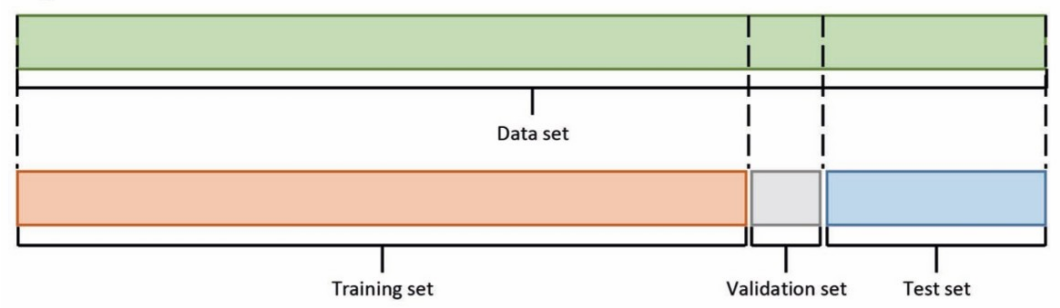


Figura 2.2 Divisione del dataset in training set, validation set e test set

2.2 **Apprendimento non supervisionato**

A differenza dell'apprendimento supervisionato, gli approcci non supervisionati sono utilizzati quando le etichette delle variabili in input non sono note a priori, di conseguenza tali metodi tentano di riconoscere relazioni e pattern solo dalle caratteristiche dei dati e senza utilizzare una categorizzazione come visto per gli algoritmi di Supervised Learning. L'apprendimento non supervisionato ha alcuni obiettivi: comprendere la distribuzione dei dati, raggrupparli sulla base di caratteristiche simili o ridurre le loro dimensioni in modo da costruirne una versione più sintetica ed efficace. Utilizzando i metodi

di Unsupervised Learning può non essere semplice misurare quanto l'algoritmo sia accurato, poiché le prestazioni sono spesso soggettive e specifiche al dominio del problema. In questa tesi ci concentreremo in particolare su algoritmi di tipo non supervisionato, e nello specifico su algoritmi di clustering. Vediamo quindi nel dettaglio questa famiglia di algoritmi.

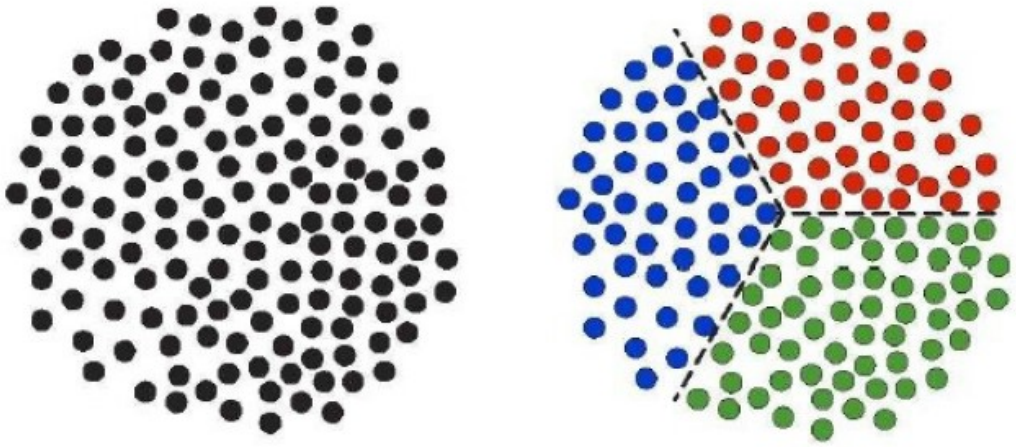
2.2.1 *Cluster Analisys*

La cluster analysis è un metodo statistico per ricavare da una popolazione di dati una struttura a gruppi. La cluster analysis è stata proposta e sperimentata a partire dagli anni 60', nell'affermazione riportata da Tyron e Bailey nel loro lavoro del 1970 [19]:

"Understanding our world requires conceptualizing the similarities and differences between the entities that compose it".

Le tecniche di cluster sono state applicate ad un'ampia varietà di problemi di ricerca: nel campo della medicina, della psichiatria, in archeologia, nel marketing e nell'industria. In generale ogni qualvolta si deve classificare una grande mole di informazioni in gruppi espressivi e trattabili, la Cluster Analysis si rivela un ottimo strumento. Perciò con il

termine *cluster analysis*, o analisi dei gruppi o delle classi, si intendono le procedure che permettono di individuare, all'interno di un insieme di oggetti di qualsiasi natura, alcuni sottoinsiemi, i clusters appunto, mutuamente esclusivi e tendenzialmente omogenei al loro interno.



Insieme di dati Dati riorganizzati in gruppi

Figura 2.3 Rappresentazione dei dati e dei gruppi ottenuti con la cluster analysis

Una volta terminato il procedimento, i cluster finali dovrebbero esibire un'alta omogeneità interna (intra-cluster) ed un'alta eterogeneità esterna (inter-cluster). In altre parole in una classificazione avvenuta con successo, gli oggetti all'interno dei cluster saranno "molto simili e affini tra loro e al tempo stesso molto differenti e distanti dagli altri cluster.

Un processo di Clustering si articola in alcune fasi fondamentali e richiede una serie di scelte, tra le quali quella di uno specifico algoritmo di raggruppamento. É necessario considerare diversi aspetti:

- la *scelta delle caratteristiche* in base alle quali raggruppare gli oggetti;
- la scelta di una adeguata *misura della dis/somiglianza* esistente fra gli oggetti;
- la scelta dell'*algoritmo di raggruppamento*;
- la *validazione* dei risultati ottenuti.

Tali scelte possono condizionare notevolmente la soluzione di Clustering ottenuta, che potrebbe essere di difficile interpretazione o ancor peggio potrebbe non riflettere la naturale struttura dei dati.

Nei seguenti paragrafi saranno illustrate le alternative sulla base delle quali scegliere una strategia che sia coerente con l'obiettivo dell'analisi.

2.2.2 *Scelta delle caratteristiche*

La quantità e la varietà di dati ad oggi disponibile è aumentata esponenzialmente nel tempo. A prescindere dai volumi in gioco, la

struttura per effettuare una cluster analysis è nella maggior parte dei casi una semplice matrice di dati così rappresentata formalmente:

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

Dove xij rappresenta l'attributo/valore della j-esima caratteristica osservato sull'i-esima unità. A partire da una struttura dati di questo tipo, il primo passo di una *Cluster Analysis* è scegliere se si intende raggruppare le righe o le colonne della matrice dati e rispetto a quali caratteristiche ricercare il raggruppamento. La scelta delle variabili, rispetto alle quali raggruppare le unità, è naturalmente legata agli obiettivi dell'analisi. Dalla matrice di partenza devono infatti essere selezionate quelle variabili che si ritengono significative per l'identificazione dei *cluster*. La riduzione della dimensionalità può essere ottenuta sia attraverso tecniche di trasformazione delle variabili, sia attraverso metodi di selezione delle variabili, la scelta tra questi due approcci per la selezione delle variabili è determinante per l'efficacia di un'applicazione di *Clustering*. Un "ottimale" selezione delle variabili

può ridurre notevolmente il carico di lavoro e semplificare le successive fasi dell'analisi.

2.2.3 *Scelta di un criterio di valutazione della somiglianza*

L'obiettivo della *Cluster Analysis* come detto è quello di suddividere un insieme, eterogeneo al suo interno, in un certo numero di gruppi secondo il livello di dis/somiglianza tra le unità che lo compongono.
Ovviamente, la scelta di una misura di dis/somiglianza influisce direttamente sulla formazione dei cluster. Per poter applicare molti degli algoritmi di Clustering, è possibile trasformare la matrice dei dati originaria in una matrice del tipo simmetrica n x n, dove il generico elemento ij esprime la prossimità tra l'unità i e l'unità j. Molti metodi di analisi dei dati in statistica si basano sul calcolo della similarità o dissimilarità tra unità.
DEFINIZIONE: Un indice di similarità è un'applicazione s su un insieme E nel campo dei numeri reali non negativi nello spazio ExE, tale che:

a) s(Xi , Xj)=s(Xj , Xi) per ogni (i,j) ExE (simmetria)
b) s(Xi , Xi)=s(Xj , Xj)=Max per ogni i e j diversi tra loro (similarità massima)

Analogamente, un indice di dissimilarità è un indice simmetrico che assume valore zero quando le due unità coincidono.

a) d(Xi , Xj)=d(Xj , Xi) per ogni (i,j) ExE (simmetria)
b) d(Xi , Xi)=0 e d(Xi , Xj)=0 per i = j (distanza minima)

Inoltre vale la disuguaglianza triangolare:

d(Xi , Xj) ≤ d(Xi , Xk) + d(Xk , Xj) per ogni i, j, k

Nel caso di caratteri quantitativi possono essere utilizzati vari tipi di indici di distanza [21] di cui i più utilizzati sono:

- *Distanza euclidea*: corrisponde al concetto geometrico di distanza nello spazio, in particolare è una misura della lunghezza del segmento avente per estremi i due punti $2^{d_{iJ}} = d_{iJ} = \sqrt{\sum_{S=1}^{P} \left(x_{i_S} - x_{JS}\right)^2}$
- *Quadrato della distanza euclidea:* qualora si voglia dare un peso progressivamente maggiore agli oggetti che stanno oltre una certa distanza;
- *Distanza di Manhattan:* è semplicemente la differenza media fra le dimensioni $1^{d_1 k} = \sum |x_{hv} - x_{kv}| w_v$ consigliata in generale quando le variabili di classificazione sono su scala ordinale;

- *Distanza di Chebychev:* può essere appropriata nei casi in cui si voglia definire due oggetti come "differenti" se essi sono diversi in ciascuna delle dimensioni: $c_{=\sum \max|x_{hv}-x_{kv}|}^{dhk}$
- Distanza e indice di Jaccard: sono due metriche utilizzate in statistica per confrontare la somiglianza/diversità tra campioni quest'ultima è di particolare interesse poiché la distanza che ho utilizzato per questo lavoro di tesi;
    - L'*indice di Jaccard* è il rapporto tra la grandezza dell'intersezione degli insiemi considerati e l'unione degli insiemi: $J(A,B) = \left|\frac{A \cap B}{AUB}\right|$
    - La *distanza di Jaccard* misura la dissomiglianza tra gli insiemi. Consiste semplicemente nel sottrarre l'indice Jaccard da 1. $Jd(A,B) = 1 - J(A,B)$

2.2.4 *Scelta di un algoritmo di raggruppamento*

Un algoritmo di Clustering nasce dalla combinazione della scelta di una misura di prossimità/distanza tra le unità da raggruppare e della scelta di una funzione criterio in base al quale effettuare il raggruppamento. La distinzione che normalmente viene proposta è fra:

- *metodi gerarchici* in cui viene costruita una gerarchia di partizioni annidate caratterizzate da un numero (de)crescente di gruppi;
- *metodi iterativi non gerarchici*. In cui un insieme di unità viene suddiviso in un pre-specificato numero di gruppi, ottenendo così un'unica partizione dei dati.

Per questo lavoro di tesi abbiamo usato un algoritmo di raggruppamento gerarchico dettagliato di seguito.

2.2.4.1 Hierarchical Clustering e Agglomerative Clustering

Sono due le modalità principali per fare cluster gerarchico.

Il **clustering agglomerativo**, scelto per questo lavoro di tesi, adotta un approccio detto bottom up, cioè dal basso verso l'alto, nel quale

si inizia inserendo ciascun elemento in un cluster diverso e successivamente si procede accorpando gradualmente coppie di cluster ampliando la soglia di similarità. L'esecuzione termina quando tutti gli oggetti sono accorpati in un unico cluster, nel quale tutti gli elementi sono considerati simili. Al contrario, il **clustering divisivo** adotta un approccio detto top down, cioè dall'alto verso il basso, nel quale tutti gli elementi inizialmente si trovano in un solo cluster che ricorsivamente viene suddiviso in cluster più piccoli.

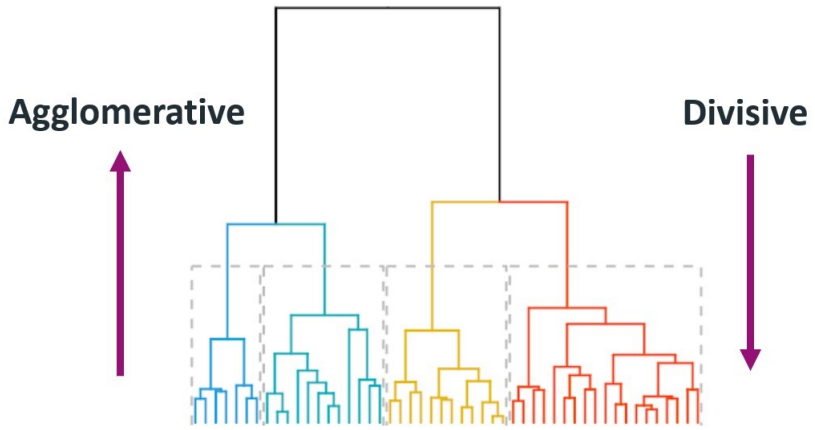


Figura 2.4 Agglomerative Clustering e Divisive Clustering

L'output del Hierarchical Clustering viene rappresentato da un dendrogramma. Il modo in cui si decide quali cluster devono essere

combinati, nel caso di Agglomerative Clustering, o quale cluster deve essere suddiviso, nel caso di Divisive Clustering, viene definito da una misura di dissimilarità tra i cluster. Questo si ottiene utilizzando una metrica appropriata, la quale quantifica la distanza tra coppie di elementi, oltre a un criterio di collegamento che ha il compito di definire la dissimilarità presente tra due cluster come funzione di distanze tra gli elementi nei cluster.

2.2.4.2 Metriche delle distanze

Metodo del legame medio, [Average-Linkage]: si tratta del valore medio aritmetico di tutte le distanze tra gli elementi;

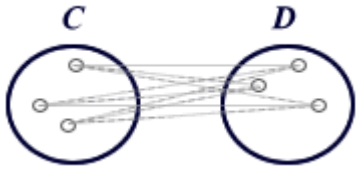


Figura 2.5 Average Linkage

Si uniscono i due gruppi che presentano la più piccola distanza così definita

L'adozione di questo algoritmo per la composizione dei gruppi semplifica notevolmente la composizione dell'albero costruito con l'algoritmo completo, mente rispetto a quello costruito sull'algoritmo singolo rappresenta una movimentazione e differenziazione. Essendo basato sulla media delle distanze, i risultati sono più attendibili e i gruppi risultano più omogenei e ben differenziati tra di loro.

Metodo del centroide, vanno determinati i vettori contenenti i valori medi delle *p* variabili in tutti gruppi (centroidi), e le distanze tra i gruppi viene assunta pari alla distanza tra i rispettivi centroidi. Se $\overline{X}_C$ e $\overline{X}_D$ sono i centroidi avremo:

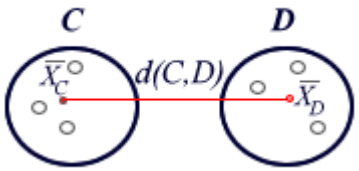


Figura 2.6 Centroide

Metodo di Ward differisce in parte dai precedenti, in quanto suggerisce di riunire, ad ogni tappa del processo, i due gruppi dalla cui fusione deriva il minimo incremento possibile della devianza "entro".

$$DEV_T = \sum_{s=1}^{p}\sum_{i=1}^{n}(x_{is} - \bar{x}_s)^2 = \sum_{i=1}^{n}\sum_{s=1}^{p}(x_{is} - \bar{x}_s)^2$$

dove $\bar{x}_s$ è la media della variabile s con riferimento all'intero collettivo. Data una partizione in g gruppi, tale devianza può essere scomposta in:

$$DEV_{IN} = \sum_{k=1}^{g}\sum_{s=1}^{p}\sum_{i=1}^{n_k}(x_{is} - \bar{x}_{s,k})^2$$

che è la devianza entro i gruppi riferita alle p variabili con riferimento al gruppo k, dove $\bar{x}_{s,k}$ è la media della variabile s con riferimento al gruppo k;

$$DEV_{OUT} = \sum_{s=1}^{p}\sum_{k=1}^{g}(\bar{x}_{s,k} - \bar{x}_s)^2 n_k$$

che è la devianza tra i gruppi. Come noto

$$DEV_T = DEV_{IN} + DEV_{OUT}$$

Nel passare da k+1 a k gruppi (aggregazione) DEV_{IN} aumenta, mentre ovviamente DEV_{OUT} diminuisce. Ad ogni passo metodo di Ward si aggregano tra loro quei gruppi per cui vi è il minor incremento della devianza entro i gruppi.

2.2.4.3 Iterative clustering e K-Means

K-Means è un algoritmo di apprendimento non supervisionato il cui scopo è quello di trovare un numero fisso di cluster all'interno di un dataset. Come descritto precedentemente, i cluster sono i gruppi nei quali vengono divisi i dati in base alla loro somiglianza, in K-Means il numero dei cluster viene scelti a priori, prima dell'esecuzione dell'algoritmo stesso. Ogni cluster raggruppa un particolare insieme di dati, che vengono chiamati data points. Per ogni cluster viene definito un centroide, cioè un punto che si trova al centro di un determinato cluster. K-Means è un algoritmo iterativo, questo significa che ripete le seguenti fasi:

- *Inizializzazione*: nella prima fase si definiscono i parametri di input per eseguire l'algoritmo. In particolare si sceglie il

dataset e il numero di centroidi iniziali da utilizzare che vengono disposti casualmente. Scegliendo il numero di centroidi si scelgono i cluster di cui il dataset sarà composto, cioè i raggruppamenti che si vogliono eseguire e successivamente visualizzare.

- *Assegnazione del cluster*: nella seconda fase l'algoritmo analizza ogni data point e lo assegna al centroide più vicino, quindi ad un determinato cluster. Per eseguire questa operazione viene calcolata la distanza euclidea tra ogni data point e ogni centroide.
- *Aggiornamento della posizione del cenroide:* dopo la seconda fase è probabile che si siano formati nuovi cluster, poiché a quelli precedenti si sono assegnati o tolti alcuni data points. Di conseguenza viene ricalcolato il punto esatto del centroide, modificandone la posizione. Tale posizione è la media di tutti i data points che sono stati assegnati al nuovo cluster.

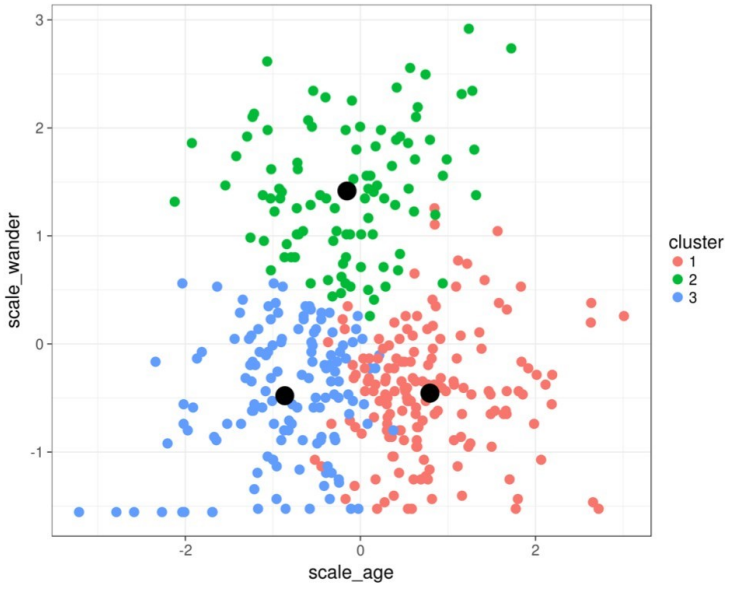


Figura 2.7 Clustering con l'algoritmo K-Means

L'algoritmo K-Means ripeterà la seconda e la terza fase finché i centroidi non si modificano, cioè finché non si raggiungerà un punto di convergenza tale per cui non sono più necessarie modifiche ai cluster. In questo caso viene raggiunta la condizione di stop, ad esempio quando non sono più presenti data points che cambiano cluster, quando la somma delle distanze è ridotta al minimo, oppure quando si raggiunge un numero massimo di iterazioni eseguite. K-Means è un algoritmo che ha tempi di esecuzione rapidi, adatto nel caso in cui sia noto il numero di

cluster a priori ed è possibile ottenere gruppi distinti dal dataset [13].

2.2.4.4 DBSCAN

L'algoritmo DBSCAN, acronimo di Density-Based Spatial Clustering of Applications with Noise è un metodo di clustering molto utilizzato e basato sulla densità, poiché il suo scopo è quello di collegare insiemi di punti che abbiano una densità sufficientemente alta. Inoltre è in grado individuare cluster che presentano forme arbitrarie. Al contrario delle tecniche K-Means e Agglomerative Clustering viste nei capitoli precedenti, l'algoritmo DBSCAN non necessita come input iniziale del numero dei cluster. I due input di cui ha bisogno sono il parametro eps (ε), e il parametro minpoints. Eps definisce il raggio della distanza, mentre minpoints è una soglia che indica il numero di vicini di ciascuna osservazione. Dato un insieme di punti che si vogliono raggruppare in cluster, innanzitutto essi sono classificati dall'algoritmo DBSCAN in una delle seguenti tre categorie:

- Core point: un campione p che ha almeno un certo numero minpoints di punti vicini, entro un intervallo determinato dal

parametro eps. Tali punti si dicono direttamente raggiungibili dal punto p.

- Border point: un campione che ha un numero inferiore di minpoints come vicini, comunque collegati (direttamente o indirettamente) a un core point.
- Outlier: un campione che ha un numero inferiore di minpoints come vicini e non è collegato ad alcun core point.

Gli altri concetti principali su cui si basa l'algoritmo DBSCAN sono descritti di seguito:

- Direttamente raggiungibile in densità: dati due punti q e p, il punto q è detto direttamente raggiungibile da p se non sono lontani più di una certa distanza eps. Come descritto in precedenza, il parametro eps è impostato dall'utente e indica l'area entro cui cercare punti vicini.
- Raggiungibile in densità: dati due punti q e p, q si dice raggiungibile in densità p se esiste una sequenza p1...pn di punti con p1=p e pn=q nella quale ognuno di essi è direttamente raggiungibile dal suo predecessore.

- Densamente connesso: dati due punti q e p, si dicono connessi in densità se esiste un punto o tale che sia o-p che o-q siano raggiungibili in densità.

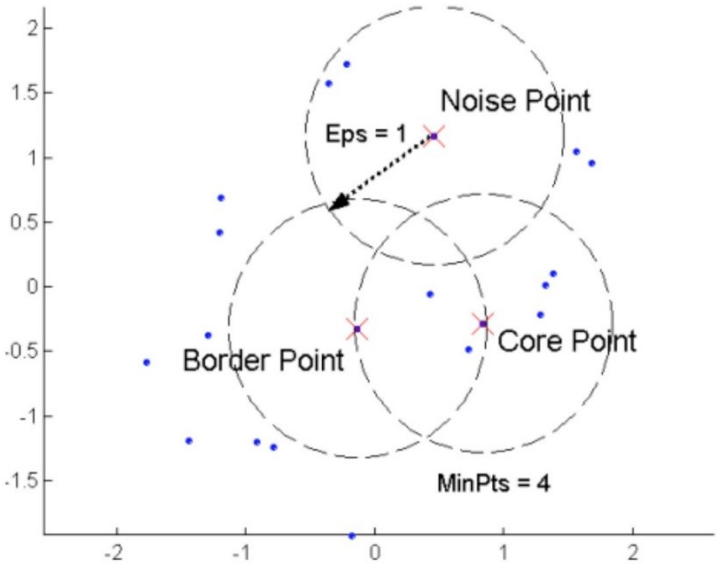


Figura 2.8 Clustering con l'algoritmo DBSCAN

Di conseguenza un cluster è un insieme in cui tutti i punti al suo interno sono connessi in densità. Inoltre, se un punto è connesso in densità ad un altro punto del cluster allora anch'esso fa parte del cluster. DBSCAN inizia la sua esecuzione selezionando casualmente un punto p non ancora visitato e ne calcola il suo vicinato in base al parametro

eps. Se tale vicinato contiene un numero sufficiente di punti in base al parametro minpoints, viene creato un nuovo cluster. Se ciò non avviene il punto viene etichettato come outlier e successivamente potrebbe essere ritrovato in un altro vicinato sufficientemente grande riconducibile ad un punto differente, entrando così a fare parte di un diverso cluster. Se un punto viene associato ad un cluster di conseguenza anche tutti i punti del suo vicinato, calcolato in base ad eps, entreranno a fare parte del medesimo cluster, lo stesso avviene anche per i loro vicinati. Tale processo continua finché il cluster è completato e sono stati visitati tutti gli altri punti [13].

2.2.5 *Validazione e interpretazione dei risultati*

La Cluster Analysis, facendo parte delle tecniche di analisi multidimensionale di tipo esplorativo, non necessita di alcun tipo di assunzione a priori, l'obiettivo è quindi individuare la naturale classificazione dei dati. Il problema è sostanzialmente quello di decidere se la soluzione ottenuta riflette la struttura naturale dei dati o è stata indotta dall'algoritmo di Clustering scelto. D'altra parte l'assenza di gruppi noti a priori in un processo di Clustering ha reso difficile trovare un indicatore

adeguato per valutare se la soluzione ottenuta, il numero di cluster, la loro forma è ammissibile o meno. L'insieme delle tecniche che mirano ad una valutazione quantitativa e oggettiva dei risultati di un processo di Clustering va sotto il nome di *Cluster validity methods* [22]. In tal senso un risultato è valido se rappresenta la migliore approssimazione possibile della realtà. Nell'ambito del Clustering la validazione di una partizione o di una gerarchia di partizioni annidate può essere effettuata tenendo conto dei seguenti criteri:

- *oggettività*, per cui i ricercatori che lavorano indipendentemente sullo stesso insieme di dati devono giungere agli stessi risultati;
- *stabilità* dei risultati del Clustering, operando su dati equivalenti;
- *capacità* predittiva delle variabili su un nuovo insieme di dati.

Jain e Dubes ritengono valida una soluzione di Clustering nella misura in cui la partizione o la gerarchia individuata fornisce una vera informazione sui dati o, in altri termini quanto tale soluzione sia in grado di riflettere le caratteristiche intrinseche dei dati [23].

In letteratura sono tradizionalmente individuati tre approcci per investigare la validità del Clustering:

- *criteri di validazione esterni*, che misurano quanto i cluster individuati corrispondono a etichette di classe fornite esternamente;
- *criteri di validazione interni*, che misurano quanto una soluzione di Clustering si adatta bene ai dati, quando i dati sono la sola informazione disponibile;
- *criteri di validazione relativi*, che misurano la bontà di una soluzione di Clustering confrontandola con i risultati ottenuti da altri algoritmi di Clustering o dallo stesso algoritmo, ma usando differenti valori dei parametri.

Sulla base di questi tre criteri, che definiscono delle strategie generali, sono stati definiti una serie di indici di validità.

Per il calcolo degli indici di *validazione esterna* sono necessarie le informazioni circa le etichette di classe degli oggetti su cui si esegue il Clustering.

Gli indici di *validazione interna* si basano sui concetti di coesione (Cluster Cohesion) e di separazione (Cluster Separation). Questi indicatori, infatti, misurano quantitativamente quanto la partizione ottenuta risponde

all'obiettivo del Clustering, individuazione di gruppi coesi e ben separati tra loro.

L'indice di *validazione relativo* è una misura di validazione supervisionata o non supervisionata, usata per eseguire un confronto. Pertanto gli indici relativi non sono realmente un gruppo separato di misure di validazione, ma rappresentano uno specifico uso degli indici interni ed esterni.

Di seguito sono indicati i principali indici di validazione interna, basati sui concetti di coesione e separazione dei cluster.

Indice Silhouette [24]: Kaufman and Rousseeuw (1990) hanno proposto il Silhouette Coefficient è un indice che misura la coesione e la separazione di un cluster. Dato un punto *i* appartenente al cluster C, sia a(i) la distanza media del punto i da tutti i punti appartenenti al cluster C. Si può quindi interpretare a(i) come una misura di quanto il punto i sia dissimile dal suo cluster. Sia b(i) la più piccola distanza media del punto i da ogni altro cluster di cui i non è membro. Il cluster cui corrisponde il minimo di b(i) è chiamato *neighbouring cluster*, poiché rappresenta dopo C il miglior cluster cui assegnare i. Un valore grande di b(i) implica che i sarebbe mal assegnato al suo neighbouring cluster. L'indice Silhouette per ciascun punto del dataset è ottenuto come segue: $s(i) = \frac{b(i)-a(i)}{\max[a(i),b(i)]}$

Indice Dunn [25]: valuta la qualità di una soluzione di Clustering in termini di rapporto tra la separazione e la coesione dei gruppi. Nello specifico questo indice considera la minima distanza a coppie tra i punti appartenenti a differenti cluster come misura della separazione tra i cluster e il massimo diametro tra tutti i cluster come misura della coesione, dove il diametro di un cluster è definito come la distanza massima che separa due punti distinti appartenenti allo stesso cluster e può essere considerato una misura della dispersione dei diversi cluster.

Indice Davies Bouldin [26]: Per ogni cluster C, sono calcolate le similarità tra C e tutti gli altri cluster, e il valore più alto è assegnato a C come sua similarità. L'indice Davies Bouldin è ottenuto come media delle similarità tra ogni cluster e il rispettivo cluster più simile ad esso. Si desidera che i cluster siano il meno possibile simili l'uno con l'altro, pertanto più piccolo sarà il valore dell'indice migliore sarà la configurazione ottenuta.

Capitolo 3
TECNOLOGIE UTILIZZATE

In questo capitolo sono riportate le principali nozioni teoriche relative alle tecnologie utilizzate durante il progetto di tesi. L'obiettivo è spiegare le caratteristiche intrinseche di ciascuna tecnologia adottata e come le stesse sono state integrate per la realizzazione del progetto finale.

3.1 **Amazon Web Services**

Amazon Web Services ambiente cloud scelto da Nexi Group nel 2019 per attività analitiche. Il cloud computing è un modello per abilitare, tramite la rete, l'accesso diffuso ad un insieme condiviso e configurabile di risorse di elaborazione (esempio: reti, server, memoria, applicazioni e servizi) che possono essere acquisite e rilasciate rapidamente e con minimo sforzo di gestione o di interazione con il fornitore di servizi [27]. Questo rappresenta un grande vantaggio per le aziende per insieme di motivi, tra cui un minore impegno sulla gestione dell'infrastruttura

tecnologica risparmiando risorse che potranno invece essere dedicate al loro business. Il Cloud Computing può essere suddiviso in tre grandi categorie, dipendentemente dallo scopo per cui viene creato. Abbiamo quindi:

- **Infrastructure as a Service (Iaas)**, dove l'infrastruttura hardware, la rete, lo storage vengono resi disponibili come servizi;
- **Platform as a Service (PaaS)**, dove è la piattaforma applicativa, il sistema operativo, a essere fruibile come servizio con la possibilità di sviluppare soluzioni software;
- **Software as a Service (SaaS)**, dove l'applicazione diventa un servizio fruibile su richiesta.

Tipicamente vengono identificate quattro categorie di Cloud, tali categorie sono:

- Public Cloud
- Private Cloud
- Community Cloud
- Hybrid Cloud

In Nexi Group, per le attività analitiche, viene utilizzato il Public Cloud AWS, naturalmente con account dedicati e la massima attenzione alle tematiche di compliance e sicurezza.

3.2 **Python**

Python è un linguaggio di programmazione orientato agli oggetti, alquanto intuitivo e di facile utilizzo. Tale caratteristica, di fatti, lo rende uno dei linguaggi di programmazione maggiormente utilizzati dai programmatori o analisti. Python è adatto allo sviluppo di applicazioni distribuite e all'elaborazione di script utili per l'analisi di dati. Inoltre, è multipiattaforma, ovvero lo si esegue sui principali sistemi operativi senza alcun problema. Tuttavia, il principale vantaggio associato al linguaggio Python è relativo alla vasta disponibilità di librerie e framework, open-source, che forniscono un supporto efficace in diversi ambiti, dal machine learning, allo sviluppo di applicazioni web. In particolare, tali librerie (descritte nei prossimi capitoli) sono risultate valide nello sviluppo del lavoro di tesi, in quanto hanno permesso lo svolgimento di attività come, operazioni di scrittura e lettura di file, manipolazione dei dati,

elaborazione di report, visualizzazione grafica dei dati, l'applicazione di algoritmi di apprendimento automatico etc. In particolare è stata utilizzata la versione 3 di Python. L'IDE (acronimo di Integrated Development Environment) utilizzato per la scrittura e il testing del codice Python è Spyder.

Spyder è una ambiente di sviluppo integrato multipiattaforma open source e facilita l'installazione e la configurazione delle varie librerie Python usate, tra le quali la libreria numpy, pandas, matplotlib, plotly, scikit-learn.

3.2.1 Pandas

Pandas è una libreria open source realizzata per il linguaggio di programmazione Python. È stata ideata dall'informatico americano Wes McKinney e la prima versione è stata rilasciata nel 2008. Si tratta di una libreria software specializzata nella gestione e nell'analisi dei dati. In particolare, offre operazioni e strutture dati per la manipolazione di tabelle numeriche e serie. Pandas è utilizzato principalmente per l'analisi dei dati e permette di importare i dataset da vari formati di file come CSV, JSON, SQL e Microsoft Excel... Durante la realizzazione del progetto, la libreria

Pandas è stata utilizzata per implementare per la lettura del file csv dove sono presenti i dati estratti con la query e l'elaborazione del Data Frame.

3.2.2 NumPy

NumPy è stata ideata dalla data scientist americano Travis Oliphant e la prima versione è stata rilasciata nel 1995. La libreria numpy permette una facile gestione di diverse strutture dati quali array multidimensionali e matrici, oltre ad una estesa collezione di funzioni matematiche di alto livello che permettono di gestire e manipolare in modo efficiente tali strutture dati, grazie alle funzioni matematiche di alto livello messe a disposizione. Durante la realizzazione del progetto, la libreria Numpy è stata utilizzata insieme al linguaggio Python per implementare gli script relativi alla gestione e all'elaborazione dei dati.

3.2.3 SciPy

SciPy è una libreria open source ideata da Travis Oliphant, Eric Jones, e Pearu Peterson, la prima versione di SciPy è stata rilasciata nel 2001. Si tratta di una libreria composta da un insieme di algoritmi e

strumenti matematici dedicati al linguaggio di programmazione Python. In particolare include moduli dedicati all'ottimizzazione, all'integrazione, al l'algebra lineare, funzioni speciali, funzioni per l'elaborazione di segnali e di immagini, risolutori e altri strumenti utilizzati nell'ambito della scienza e dell'ingegneria. La struttura dati di base utilizzata da SciPy è un array multidimensionale fornito dalla libreria NumPy. Durante la realizzazione del progetto, la libreria Numpy si è utilizzata insieme al linguaggio Python per implementare gli script relativi alla gestione e all'elaborazione dei dati, l'analisi della correlazione per la creazione dei cluster e del dendrogramma.

3.2.4 Scikit-learn

Scikit-learn è una libreria open source ideata dalla data scientist David Cournapeau e la prima versione è stata rilasciata nel 2007. Si tratta di una libreria software dedicata all'apprendimento automatico, include dozzine di algoritmi e modelli di classificazione, regressione, clustering, etc... Scikit-learn è diventata una delle librerie più utilizzate nell'ambito dell'apprendimento automatico, sia supervisionato che non supervisionato, grazie alla vasta gamma di strumenti che offre, ma anche

grazie alla sua API ben documentata, facile da usare e versatile. Scikit-learn è progettato per operare insieme alle librerie NumPy e SciPy, inoltre si integra bene con molte altre librerie Python, come Matplotlib e Pandas. Durante la realizzazione del progetto, la libreria Scikit-learn è stata utilizzata per implementare gli algoritmi di Machine Learning di Agglomerative Clustering.

3.2.5 Matplotlib

Matplotlib è una libreria open source ideata dal neurobiologo americano John Hunter e la prima versione è stata rilasciata nel 2003.

Si tratta di una libreria software di plotting che utilizza anche la libreria matematica NumPy per la creazione di grafici di tipologie diverse. Ad esempio grafici a linee, istogrammi e scatter plot. Matplotlib rende disponibili funzioni dedicate per l'inserimento di grafici all'interno delle applicazioni Python. Durante la realizzazione del progetto, la libreria Matplotlib è stata utilizzata insieme al linguaggio Python per implementare gli script relativi all'analisi, alla correlazione e al clustering dei dati, in particolare per la creazione di alcuni grafici.

3.3 **BitBucket**

Per favorire la collaborazione di tutto il team di sviluppo, in Nexi si utilizza il repository Cloud BitBucket, uno strumento di gestione della versione Git basato sul web, sviluppato da Atlassian. BitBucket, come altri sistemi di controllo di versioni, consente di creare repository in cui gli sviluppatori possono caricare il proprio codice o modificare file esistenti, promuovendo una gestione collaborativa del codice.

BitBucket consente di vedere chi ha fatto le modifiche e quando, risolve gli eventuali conflitti tra le versioni differenti del codice sviluppare in parallelo e permette di tornare alle versioni precedenti in caso di necessità. Il suo funzionamento si basa sui comandi *Git,* gestibili in maniera semplice e rapida anche con client GUI come Sourcetree.

Capitolo 4
SEGMENTAZIONE DEI GRANDI ESERCENTI ATTRAVERSO LA TECNICA DELL'AGGLOMERATIVE CLUSTERING

In questo capitolo viene descritta, in modo dettagliato, l'architettura del progetto di tesi, riportando i passaggi sequenziali che sono stati compiuti. L'obiettivo è quello di mostrare le scelte che sono state attuate e le metodologie applicate in funzione dell'obiettivo proposto dalla tesi, ossia la realizzazione di un sistema di raccomandazione, in grado di fornire all'ufficio di marketing uno strumento utile a definire una campagna commerciale.

Nei successivi paragrafi viene descritto l'approccio metodologico adottato, come riportato nella (fig. 4.1), che partendo da un dataset di Laka, ha permesso la clusterizzazione del comportamento dei possessori delle carte di credito in relazione con i Laka.

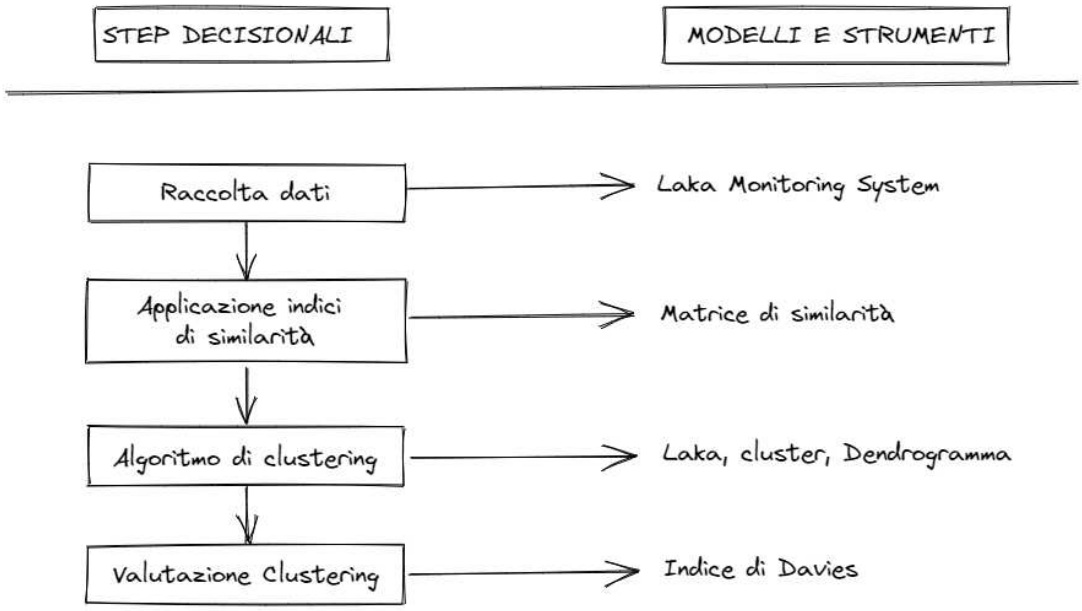


Figura 4.1 Processo di clustering

4.1 **Presentazione del caso applicativo**

NEXI Group fornisce un dataset come risultato di un suo processo interno denominato LMS (**L**aka **M**onitoring **S**ystem), che attraverso algoritmi di String Matching classifica i Merchant[1] attraverso l'analisi delle loro transazioni su carte NEXI. Un estratto delle informazioni prodotte è rappresentato nella Tabella 4.1 riportata di seguito.

[1] Esercenti

id_mov	id_car	dt_mov	...	fl_cash	fl_mov_internet	fl_domestic	te_insegna_ese	nm_nome_laka	va_spe_eur	anno	mese	giorno
15923277200002022-06-19-11.33.00.076132	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	54.9	2022	6	19
14623247300002022-06-19-11.33.00.024539	xxx	18/06/2022	...	N	N	S	ALI'	ALI' (SUPERMERCATO)	25.56	2022	6	19
12348178500002022-06-19-11.33.00.057887	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	95.72	2022	6	19
16090280400002022-06-19-11.33.00.059811	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	70.85	2022	6	19
16785153400002022-06-19-11.33.00.069639	xxx	18/06/2022	...	N	N	S	TIGOTA'	TIGOTA	72.25	2022	6	19
00988016600002022-06-19-20.33.00.004581	xxx	18/06/2022	...	N	N	S	ALLIANZ DIRECT SPA	ALLIANZ DIRECT	46.0	2022	6	19
09679281700002022-06-19-11.33.00.046638	xxx	18/06/2022	...	N	N	S	SUPERMERCATO FAMILA	FAMILA	35.38	2022	6	19
17516964900002022-06-19-11.33.00.033699	xxx	18/06/2022	...	N	N	S	PRIX QUALITY	PRIX (SUPERMERCATI)	10.01	2022	6	19
108260238000002022-06-19-17.06.00.109888	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	15.31	2022	6	19
022878168000002022-06-19-11.33.00.088559	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	65.45	2022	6	19
039027896000002022-06-19-11.33.00.081473	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	120.84	2022	6	19
057934693000002022-06-19-11.33.00.091598	xxx	18/06/2022	...	N	N	S	TIGOTA'	TIGOTA	52.29	2022	6	19
190478333000002022-06-19-11.33.00.049605	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	64.83	2022	6	19
102754000000002022-06-19-20.33.00.000695	xxx	17/06/2022	...	N	N	S	INARCASSACARD A SALDO	CASSE PREVIDENZA / ASSICURAZIONE	1564.0	2022	6	19
062656848000002022-06-19-20.33.00.000663	xxx	17/06/2022	...	N	N	S	INARCASSACARD A SALDO	CASSE PREVIDENZA / ASSICURAZIONE	1564.0	2022	6	19
008491931000002022-06-19-20.33.00.000682	xxx	17/06/2022	...	N	N	S	INARCASSACARD A SALDO	CASSE PREVIDENZA / ASSICURAZIONE	1564.0	2022	6	19
144452645000002022-06-19-20.33.00.000712	xxx	17/06/2022	...	N	N	S	INARCASSACARD A SALDO	CASSE PREVIDENZA / ASSICURAZIONE	1564.0	2022	6	19
123175974000002022-06-19-20.33.00.000694	xxx	17/06/2022	...	N	N	S	INARCASSACARD A SALDO	CASSE PREVIDENZA / ASSICURAZIONE	1564.0	2022	6	19
119394872000002022-06-19-20.33.00.000685	xxx	17/06/2022	...	N	N	S	INARCASSACARD A SALDO	CASSE PREVIDENZA / ASSICURAZIONE	1564.0	2022	6	19
024037185000002022-06-19-20.33.00.000704	xxx	17/06/2022	...	N	N	S	INARCASSACARD A SALDO	CASSE PREVIDENZA / ASSICURAZIONE	1564.0	2022	6	19
035519478000002022-06-19-11.33.00.020317	xxx	18/06/2022	...	N	N	S	SUPERMERCATO DOK	DOK (SUPERMERCATO)	149.67	2022	6	19
137442160000012022-06-19-20.33.00.000713	xxx	17/06/2022	...	N	N	S	INARCASSACARD A SALDO	CASSE PREVIDENZA / ASSICURAZIONE	1564.0	2022	6	19
187563623000002022-06-19-11.31.00.052101	xxx	17/06/2022	...	N	N	S	IPERCOOP CENTRONOVA	COOP ITALIA	165.36	2022	6	19
199880098000002022-06-19-11.33.00.087620	xxx	18/06/2022	...	N	N	S	UNICOOP-FIRENZE	COOP ITALIA	235.98	2022	6	19
026768846000002022-06-19-11.33.00.022056	xxx	18/06/2022	...	N	N	S	SUPERMERCATO IPERLANDO	IPERLANDO	145.91	2022	6	19
173566304000002022-06-19-11.33.00.019680	xxx	18/06/2022	...	N	N	S	CONAD SUPERSTORE	CONAD	99.67	2022	6	19
158231985000002022-06-19-11.33.00.028579	xxx	18/06/2022	...	N	N	S	SUPERMERCATO DOK	DOK (SUPERMERCATO)	38.4	2022	6	19
145223761000002022-06-19-11.33.00.078921	xxx	18/06/2022	...	N	N	S	KING SPORT & STILE	KING (SPORT & STYLE)	76.9	2022	6	19
076224641000002022-06-19-11.33.00.063809	xxx	18/06/2022	...	N	N	S	ALI'	ALI' (SUPERMERCATO)	26.59	2022	6	19
133129566000002022-06-19-11.33.00.045705	xxx	18/06/2022	...	N	N	S	HAPPY CASA STORE	HAPPY CASA	4.99	2022	6	19

Tabella 4.1 Estratto dataset generato dal processo LMS (n.b. per motivi di spazio si riporta un sub-set dati)

Il dataset completo per il periodo gennaio/ottobre 2022 conteneva più di 650.000.000 di movimenti di carte di credito, per questo lavoro di tesi vengono prese in esame solamente le variabili e i casi ritenuti utili al nostro scopo, perciò il dataset completo è stato filtrato estraendo solamente i grandi esercenti (Laka) ed escludendo la grande distribuzione alimentare, ottenendo così un nuovo dataset di circa 195.000 movimenti distribuiti su 477 Laka.

4.2 **Obiettivo dell'analisi**

Lo scopo di questa analisi è estrarre valore dai dati "grezzi" in termini di conoscenza e informazioni con l'obiettivo di fornire al reparto marketing di NEXI group uno strumento utile nella definizione di campagne di marketing basate sulla comprensione dei Laka relazionati dagli acquisti effettuati da ogni cliente. La tecnica utilizzata è il clustering gerarchico che appartiene agli algoritmi di apprendimento automatico non supervisionati.

Per questo lavoro di tesi si è scelto di utilizzare la libreria SciPy di Python per raggruppare i dati in cluster, utilizzando:

- Il metodo *linkage*: per creare i cluster
- Il metodo *fcluster*: per predire le etichette

4.3 **Processo di Cluster Analysis**

Il caso applicativo è stato trattato secondo le seguenti fasi:

- Processo di raccolta dei dati
- Definizione degli indici di Similarità e matrice di Jaccard

- Sviluppo algoritmo di Clustering
- Valutazione dei risultati ottenuti

4.3.1 *Processo di raccolta dei dati*

All'interno di NEXI Group esiste un processo informatico denominato LMS (**L**aka **M**onitoring **S**ystem), come riportato nella (fig. 4.2), che attraverso algoritmi di String Matching si occupa di classificare i merchant attraverso l'analisi delle loro transazioni su carte NEXI: un merchant alto-transante viene classificato come LAKA, mentre tutto il rimanente viene classifico come SME.

I merchant nel processo di classificazione vengono anche suddivisi in Local (nazionali) o Global (internazionali) ed E-commerce o fisico.

Per questo lavoro di tesi prendiamo in considerazioni solo merchant Local.

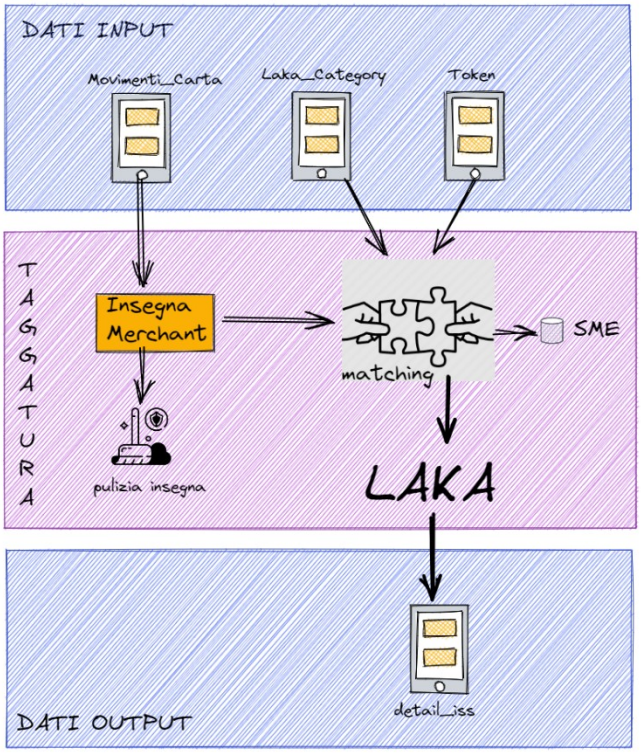


Figura 4.2 Laka Monitoring System

Il processo LMS accede a un RDS (Repository Data Store) su cloud AWS, e si compone nelle seguenti tre fasi successive:

1° Fase: acquisizione dati - L'acquisizione dei dati avviene attraverso l'interrogazione delle tabelle movimenti, anagrafica e token.

La tabella *movimenti* contiene i dati dei movimenti effettuati; in questa tabella sono presenti in tabella molte informazioni utili alla

classificazione delle transazioni e anche alla profilazione dei Merchant.

Tabella *anagrafica* in cui è presente l'anagrafica di tutti i Laka già classificati dal processo LMS.

Tabella *token* contiene il dizionario il dizionario utilizzato dall'algoritmo di String Matching per la classificazione dei Laka.

2° Fase: taggatura - Il processo di taggatura del Laka, riportato nella figura 4.3, avviene tramite la lettura dell'insegna del negozio relativa al movimento, la quale viene tokenizzata e processata. Alla fine della fase di processing (ossia di pulizia) viene verificato che sia presente nei token della tabella. Il token può essere Monogramma o Bigramma.

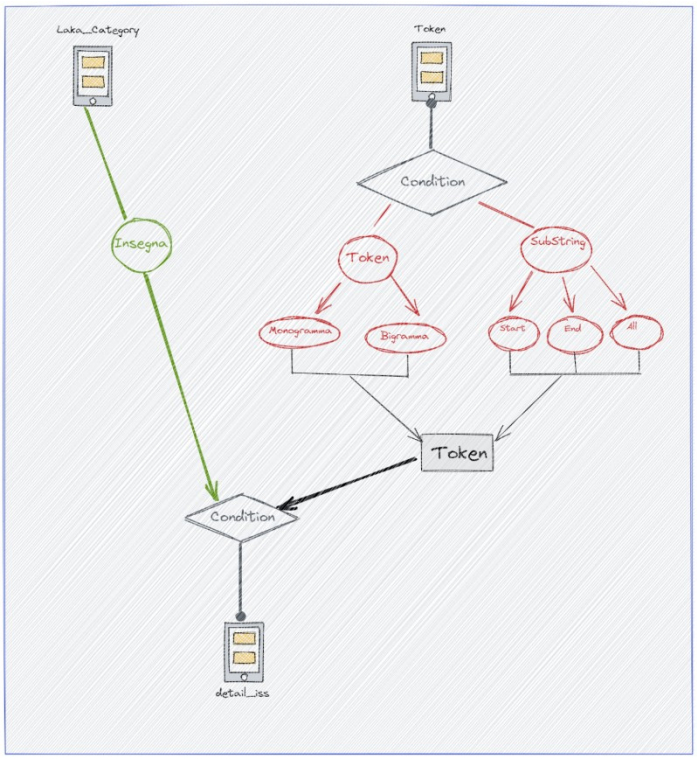


Figura 4.3 Processo di taggatura Laka

3° Fase: esportazione dei dati - Il risultato della fase di taggatura viene infine inserito nella tabella *detail*, presente sul repository AWS, le informazioni in essa contenute permettono di attribuire ad ogni insegna del merchant l'appartenenza ai Laka classificati. Un estratto dei dati è riportato nella tabella 4.2 riportata di seguito.

id_mov	id_car	dt_mov	...	fl_cash	fl_mov_internet	fl_domestic	te_insegna_ese	nm_nome_laka	va_spe_eur	anno	mese	giorno
159232772000002022-06-19-11.33.00.076132	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	54.9	2022	6	19
146232473000002022-06-19-11.33.00.024539	xxx	18/06/2022	...	N	N	S	ALI'	ALI' (SUPERMERCATO)	25.56	2022	6	19
123481785000002022-06-19-11.33.00.057887	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	95.72	2022	6	19
160902804000002022-06-19-11.33.00.059811	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	70.85	2022	6	19
167851534000002022-06-19-11.33.00.069639	xxx	18/06/2022	...	N	N	S	TIGOTA'	TIGOTA	72.25	2022	6	19
009880166000002022-06-19-20.33.00.004581	xxx	18/06/2022	...	N	N	S	ALLIANZ DIRECT SPA	ALLIANZ DIRECT	46.0	2022	6	19
096792817000002022-06-19-11.33.00.046638	xxx	18/06/2022	...	N	N	S	SUPERMERCATO FAMILA	FAMILA	35.38	2022	6	19
175169649000002022-06-19-11.33.00.033699	xxx	18/06/2022	...	N	N	S	PRIX QUALITY	PRIX (SUPERMERCATI)	10.01	2022	6	19
108260238000002022-06-19-17.06.00.109888	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	15.31	2022	6	19
022878168000002022-06-19-11.33.00.088559	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	65.45	2022	6	19
039027896000002022-06-19-11.33.00.081473	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	120.84	2022	6	19
057934693000002022-06-19-11.33.00.091598	xxx	18/06/2022	...	N	N	S	TIGOTA'	TIGOTA	52.29	2022	6	19
190478333000002022-06-19-11.33.00.049605	xxx	18/06/2022	...	N	N	S	CONAD	CONAD	64.83	2022	6	19
102754000000002022-06-19-20.33.00.000695	xxx	17/06/2022	...	N	N	S	INARCASSACARD A SALDO	CASSE PREVIDENZA / ASSICURAZIONE	1564.0	2022	6	19

Tabella 4.2 Estratto dati della tabella detail

4.3.2 *Definizione degli indici di similarità e matrice di Jaccard*

A partire dai dati raccolti con il progetto LMS che identifica i Laka, si possono applicare vari coefficienti di similarità. Il loro scopo è quello di definire il grado di correlazione tra due Laka generici L1, L2 in base al volume del transato. L'indice usato è Jaccard Similarity Index (*J*). La correlazione tra il Laka 1 e il Laka 2 relativamente alla parte i è definita nel modo seguente:

$$S_{l1,l2} = \frac{\sum_{i=1}^{N} X_{i,l1,l2}}{\sum_{i=1}^{N} (X_{i,l1,l2} + Y_{i,l1,l2})} \quad 0 \leq S_{l1,l2} \leq 1$$

$$X_{i,l1,l2} = \begin{cases} 1, se\ i\ transa\ su\ entrambio\ i\ laka \\ 0, altrimenti \end{cases}$$

$$Y_{i,l1,l2} = \begin{cases} 1, se\ i\ transa\ su\ laka\ 1\ o\ laka\ 2 \\ 0, altrimenti \end{cases}$$

per un valore dell'indice di Jaccard pari a 1 la correlazione è massima (casi molto rari), per un valore pari a 0 non esiste correlazione.

Partendo dall'output del processo LMS, come descritto nel paragrafo precedente, ed escludendo la categoria merceologica alimentari e considerando solo il transato nel periodo che va dal 1gennaio 2022 al 30 settembre 2022, viene calcolato il coefficiente di Jaccard attraverso lo script SQL (script 4.1) riportato di seguito.

```
with
LAKA_COUNT as (
    select nm_nome_laka, count(distinct id_car) count_laka
    from laka_db.detail
        inner join laka_db.anagrafica c using(nm_nome_laka)
    where c.nm_category = 'LOCAL' and anno = '2022' and nm_nome_laka not like 'SME%'
and  c.te_ctg_2 <> 'DISTRIBUZIONE MODERNA'
    group by 1
    order by 1
),
LAKA_CAR as (
    select c.te_ctg_2, nm_nome_laka, id_car
    from laka_db.detail
        inner join laka_db.anagrafica c using(nm_nome_laka)
    where c.nm_category = 'LOCAL' and anno = '2022' and nm_nome_laka not like 'SME%'
and  c.te_ctg_2 <> 'DISTRIBUZIONE MODERNA'
    group by 1,2,3
    order by 1,2,3
),
LAKA_COUPLE_CAR as(
    select A.te_ctg_2 ctg_1, A.nm_nome_laka laka_1, B.te_ctg_2 ctg_2, B.nm_nome_laka
laka_2, count(*) count_laka_12
    from LAKA_CAR A
        inner join LAKA_CAR B using(id_car)
    where A.nm_nome_laka <> B.nm_nome_laka and
          A.te_ctg_2 <> B.te_ctg_2
    group by 1,2,3,4
    order by 1,2,3,4
)
select ctg_1, ctg_2,laka_1, laka_2, A.count_laka count_laka_1, B.count_laka
count_laka_2, count_laka_12,  1.0 * count_laka_12 /(A.count_laka + B.count_laka)
similarity
from LAKA_COUPLE_CAR C
    inner join LAKA_COUNT A on A.nm_nome_laka = C.laka_1
    inner join LAKA_COUNT B on B.nm_nome_laka = C.laka_2
order by 6 DESC
```

Script 4.1 SQL per la generazione dataset similarità dei Lapa applicando Jaccard

dove il risultato dell'estrazione è riportato di seguito:

ctg_1	ctg_2	laka_1	laka_2	count_laka_1	count_laka_2	count_laka_12	similarity
PEDAGGI	CARBURANTI E STAZIONI DI SERVIZIO	AUTOSTRADE	ENI (AGIP)	845745	1048184	318588	16,82%
RISTORANTI	CARBURANTI E STAZIONI DI SERVIZIO	AUTOGRILL	ENI (AGIP)	506418	1048184	228832	14,72%
RISTORANTI	PEDAGGI	AUTOGRILL	AUTOSTRADE	506418	845745	198666	14,69%
CARBURANTI E STAZIONI DI SERVIZIO	RISTORANTI	NUOVA SIDAP	AUTOGRILL	191020	506418	89862	12,88%
CARBURANTI E STAZIONI DI SERVIZIO	PEDAGGI	TAMOIL	AUTOSTRADE	499989	845745	171131	12,72%
SERVIZI VARI	TELEFONIA	LINKEM	TISCALI	11664	17089	3570	12,42%
PROFUMERIE	VESTIARIO	KIKO	TEZENIS	171308	347922	64264	12,38%
CARBURANTI E STAZIONI DI SERVIZIO	PEDAGGI	IP (GRUPPO API)	AUTOSTRADE	422393	845745	147075	11,60%
ATTIVITA RICREATIVE	TRASPORTI PERSONE	TICKETONE (EVENTIM)	TRENITALIA	323895	597360	101503	11,02%
CARBURANTI E STAZIONI DI SERVIZIO	RISTORANTI	TAMOIL	AUTOGRILL	499989	506418	109826	10,91%
CARBURANTI E STAZIONI DI SERVIZIO	RISTORANTI	IP (GRUPPO API)	AUTOGRILL	422393	506418	100676	10,84%
CASALINGHI	VESTIARIO	ACQUA E SAPONE	OVIESSE (OVS)	388448	542408	98365	10,57%
ENTI PUBBLICI	SERVIZI VARI	LEXTEL	VISURA,IT	3341	17029	2143	10,52%
VESTIARIO	CARBURANTI E STAZIONI DI SERVIZIO	OVIESSE (OVS)	ENI (AGIP)	542408	1048184	156341	9,83%
TRASPORTI PERSONE	PEDAGGI	TRENITALIA	AUTOSTRADE	597360	845745	141067	9,78%
VESTIARIO	PROFUMERIE	CALZEDONIA	KIKO	297943	171308	45558	9,71%
CASALINGHI	VESTIARIO	ACQUA E SAPONE	TEZENIS	388448	347922	71166	9,66%
CASALINGHI	VESTIARIO	RISPARMIO CASA	GLOBO	234248	210675	42703	9,60%
VESTIARIO	CASALINGHI	TEZENIS	KASANOVA	347922	237307	54473	9,31%
RISTORANTI	TRASPORTI PERSONE	AUTOGRILL	ATM - MILANO	506418	187019	64035	9,23%
VESTIARIO	PROFUMERIE	ZARA	KIKO	545786	171308	65689	9,16%
PROFUMERIE	VESTIARIO	KIKO	STRADIVARIUS	171308	105593	24777	8,95%
TRASPORTI PERSONE	RISTORANTI	TRENITALIA	AUTOGRILL	597360	506418	98358	8,91%
CASALINGHI	ALTRI PRODOTTI RETAIL	RISPARMIO CASA	MAURY'S	234248	141191	33415	8,90%
VESTIARIO	CASALINGHI	OVIESSE (OVS)	KASANOVA	542408	237307	68710	8,81%
TRASPORTI PERSONE	VESTIARIO	TRENITALIA	ZARA	597360	545786	99757	8,73%
...	...	...	...	...	...	...	...

Tabella 4.3 Estratto del dataset delle similarità dei Laka applicando Jaccard

La fase successiva è la costruzione della matrice di similarità, di seguito un estratto:

	ACQUA E SAPONE	ATM - MILANO	AUTOGRILL	AUTOSTRADE	CALZEDONIA	ENI (AGIP)	GLOBO	IP (GRUPPO API)	KASANOVA	KIKO	MONDADORI	LINKEM	MAURY'S	NUOVA SIDAP	OVIESSE (OVS)	RISPARMIO CASA	TRENITALIA	TAMOIL	TEZENIS	TICKETONE (EVENTIM)	...
ACQUA E SAPONE	1,000																				...
ATM - MILANO	0,029	1,000																			...
AUTOGRILL	0,063	0,092	1,000																		...
AUTOSTRADE	0,061	0,061	0,147	1,000																	...
CALZEDONIA	0,073	0,039	0,060	0,053	1,000																...
ENI (AGIP)	0,083	0,045	0,147	0,168	0,059	1,000															...
GLOBO	0,082	0,010	0,038	0,041	0,000	0,051	1,000														...
IP (GRUPPO API)	0,066	0,040	0,108	0,116	0,046	0,000	0,055	1,000													...
KASANOVA	0,000	0,039	0,058	0,048	0,068	0,056	0,047	0,047	1,000												...
KIKO	0,066	0,039	0,047	0,038	0,097	0,040	0,046	0,036	0,071	1,000											...
MONDADORI	0,038	0,000	0,089	0,098	0,053	0,077	0,022	0,053	0,037	0,035	1,000										...
LINKEM	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,001	1,000									...
MAURY'S	0,066	0,008	0,027	0,030	0,033	0,039	0,076	0,042	0,036	0,035	0,000	0,003	1,000								...
NUOVA SIDAP	0,032	0,050	0,129	0,078	0,031	0,000	0,024	0,000	0,032	0,024	0,001	0,002	0,025	1,000							...
OVIESSE (OVS)	0,106	0,039	0,080	0,082	0,000	0,098	0,000	0,071	0,088	0,080	0,000	0,002	0,053	0,038	1,000						...
RISPARMIO CASA	0,000	0,015	0,046	0,048	0,047	0,060	0,096	0,060	0,000	0,048	0,000	0,002	0,089	0,031	0,078	1,000					...
TRENITALIA	0,038	0,000	0,089	0,098	0,053	0,077	0,022	0,053	0,037	0,091	0,090	0,004	0,016	0,039	0,056	0,024	1,000				...
TAMOIL	0,064	0,046	0,109	0,127	0,046	0,000	0,047	0,000	0,049	0,037	0,000	0,002	0,037	0,000	0,076	0,058	0,025	1,000			...
TEZENIS	0,097	0,042	0,069	0,064	0,000	0,072	0,000	0,056	0,093	0,124	0,000	0,002	0,046	0,033	0,000	0,067	0,000	0,062	1,000		...
TICKETONE (EVENTIM)	0,027	0,048	0,051	0,057	0,047	0,043	0,018	0,032	0,026	0,028	0,001	0,004	0,013	0,025	0,037	0,018	0,110	0,032	0,040	1,000	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Tabella 4.4 Estratto della matrice di correlazione sulle similarità di Jaccard

La matrice delle correlazioni sulle similarità di Jaccard (tab. 4.4) rappresenta uno strumento fondamentale nell'analisi perché in essa vengono evidenziate le relazioni tra le varie coppie di variabili, dove il valore 1 indica correlazione massima e il valore 0 correlazione nulla. Nel nostro caso dalla **matrice di correlazione** si evince la presenza di Laka con una discreta correlazione, tralasciando le correlazioni ovvie, tra i maggiormente correlati indichiamo:

- Nova Sidap con Autogrill (c=0.129)
- Tezenis con Kiko (c=0.124)
- TicketOne con Trenitalia (c=0,110)
- Oviesse con Acqua e Sapone (c=0.106)
- Kiko con Calzedonia (c=0.097)

4.3.3 *Sviluppo algoritmo di Clustering*

La matrice delle distanze di Jaccard appena ottenuta può essere utilizzata per identificare gruppi di Laka simili, questo grazie alla disponibilità di una grande varietà di algoritmi euristici, i quali ci

permettono, attraverso un approccio gerarchico agglomerativo, di generare i cluster definitivi.

Inizialmente ogni Laka forma un cluster, successivamente cluster distinti vengono agglomerati insieme nell'ordine dettato dall'algoritmo scelto, e ad ogni passaggio vengono aggiornati gli indici delle distanze dei nuovi gruppi. Il tutto viene ripetuto fino ad ottenere un unico cluster che contiene tutti gli altri. Come riportato nel capitolo 4, ci sono vari algoritmi euristici che permettono di raggiungere questo scopo, in questo lavoro di tesi, partendo dalla Matrice di Similarità (tab. 4.4) sono stati applicati alcuni di questi algoritmi attraverso lo sviluppo di uno script nel linguaggio Python, come riportato di seguito:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
from scipy.spatial.distance import pdist
from sklearn import metrics
...

D = pdist(corrLaka, metric = 'cityblock')
H = linkage(D, method = '…',metric='jaccard' )
plot_with_labels(H, 12)
```

Script 4.2 Python usato per la clusterizzazione (linkge)

Per ogni algoritmo utilizzato è stato valutato e contestualizzato il risultato. Algoritmi euristici utilizzati per questo lavoro di sono:

- Single linkage
- Complete linkage
- Ward linkage

Il risultato del processo di clustering non dipende solo dalla regola adottata per il raggruppamento, ma dipende anche dalla scelta del valore di soglia che taglia il dendrogramma. Esso è il valore minimo dell'indice al quale si accetta la formazione di cluster, dopo questo coefficiente si bloccano le fusioni e si ottiene la formazione finale delle celle. Per questo lavoro di tesi si è scelto una **threshold** pari a 12.

```
...
Clusters = fcluster(H,t =12, criterion='distance')
...
```

Script 4.3 Python usato per la predizione delle etichette (fcluster)

Nei paragrafi successivi analizziamo i risultati ottenuti per ogni algoritmo selezionato.

4.3.4 *Valutazione dei risultati ottenuti*

Single Linkage - I risultati ottenuti dall'elaborazione del seguente algoritmo scritto in Python al quale viene fornito in input la matrice di similarità calcolata tramite il Jaccard Similarity Index (tab. 4.5) sono i seguenti:

Cluster	Numero Laka
cluster n. 1	2
cluster n. 2	2
cluster n. 3	4
cluster n. 4	1
cluster n. 5	2
cluster n. 6	1
cluster n. 7	2
cluster n. 8	2
cluster n. 9	2
cluster n. 10	444
cluster n. 11	1
cluster n. 12	1
cluster n. 13	1
cluster n. 14	1
cluster n. 15	1
cluster n. 16	1
cluster n. 17	1
cluster n. 18	1
cluster n. 19	1
cluster n. 20	1
cluster n. 21	1
cluster n. 22	1
cluster n. 23	1
cluster n. 24	1
cluster n. 25	1

Tabella 4.5 Attribuzione dei cluster per single linkage

La tabella riassume le aggregazioni effettuate dall' algoritmo, di seguito il dendrogramma:

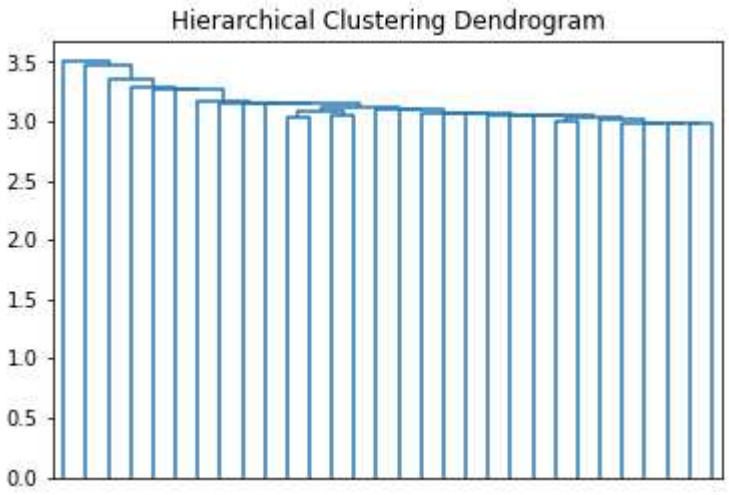


Figura 4.4 Dendrogramma Single Linkage

Nell'asse verticale la gerarchia di aggregazione, mostrano nell'ordine le varie fusioni ed il rispettivo valore alle quali avvengono, riducendo sempre più il coefficiente di similarità. Come si può notare dall'estratto della tabella e dal dendrogramma questo algoritmo con una **threshold** pari a **3** genera un totale di **25 cluster** con un coefficiente **Davies** pari a **1.4243**, il limite del metodo del legame singolo è che ha la tendenza a concatenare quasi tutti i casi in un unico grande gruppo e si mantengono separati solo piccoli gruppi o casi isolati.

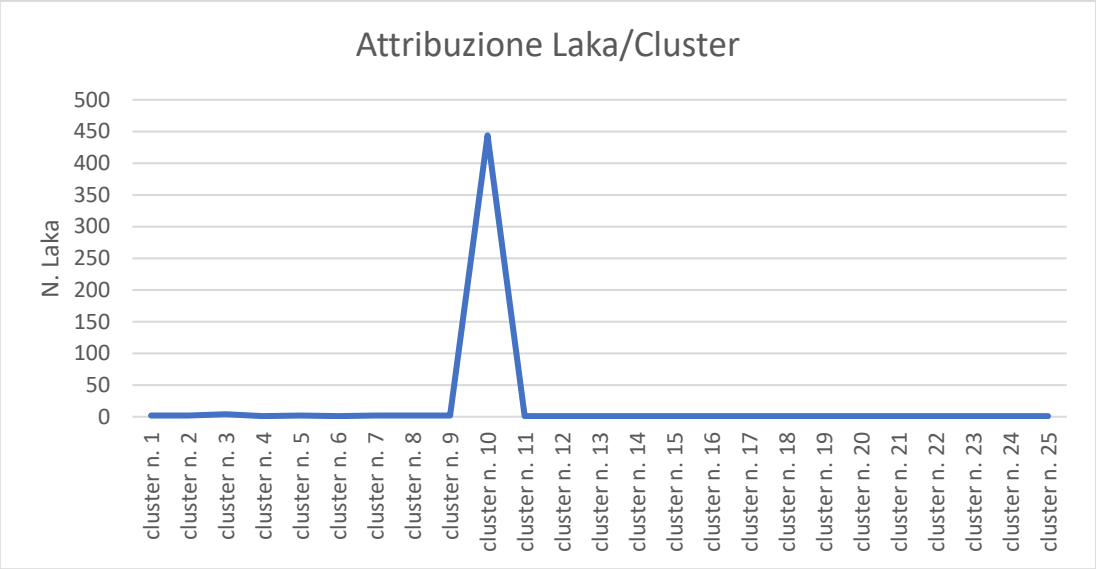


Figura 4.5 Distribuzione Laka ai cluster: Single Linkage

Come possibile notare la distribuzione della clusterizzazione con l'algoritmo ***Single*** è fortemente sbilanciata verso il cluster n. 10.

Complete Linkage - I risultati ottenuti dall'elaborazione del seguente algoritmo scritto in Python al quale viene fornito in input la matrice di similarità calcolata tramite il Jaccard Similarity Index (tab. 4.6) sono i seguenti:

Cluster	nLaka
cluster n. 1	250
cluster n. 2	27
cluster n. 3	57
cluster n. 4	8
cluster n. 5	22
cluster n. 6	22
cluster n. 7	5
cluster n. 8	7
cluster n. 9	3
cluster n. 10	6
cluster n. 11	21
cluster n. 12	2
cluster n. 13	4
cluster n. 14	2
cluster n. 15	3
cluster n. 16	6
cluster n. 17	5
cluster n. 18	4
cluster n. 19	4
cluster n. 20	4
cluster n. 21	9
cluster n. 22	2
cluster n. 23	2
cluster n. 24	2

Tabella 4.6 Attribuzione dei cluster per complete linkage

La tabella riassume le aggregazioni effettuate dall' algoritmo, che si possono rappresentare in un dendrogramma:

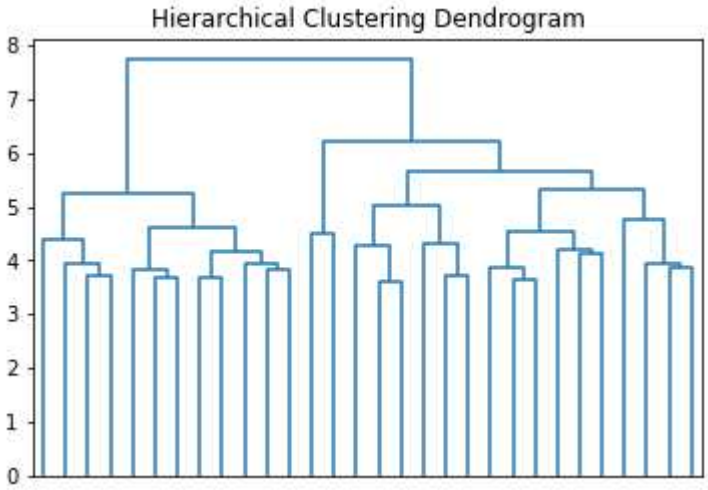


Figura 4.6 Dendrogramma Complete Linkage

Come si può notare dall'estratto della tabella e dal dendrogramma anche questo algoritmo con una **threshold** pari a **3.8** genera un totale di **24 cluster** e un coefficiente **Davies** pari a **6.4248**.

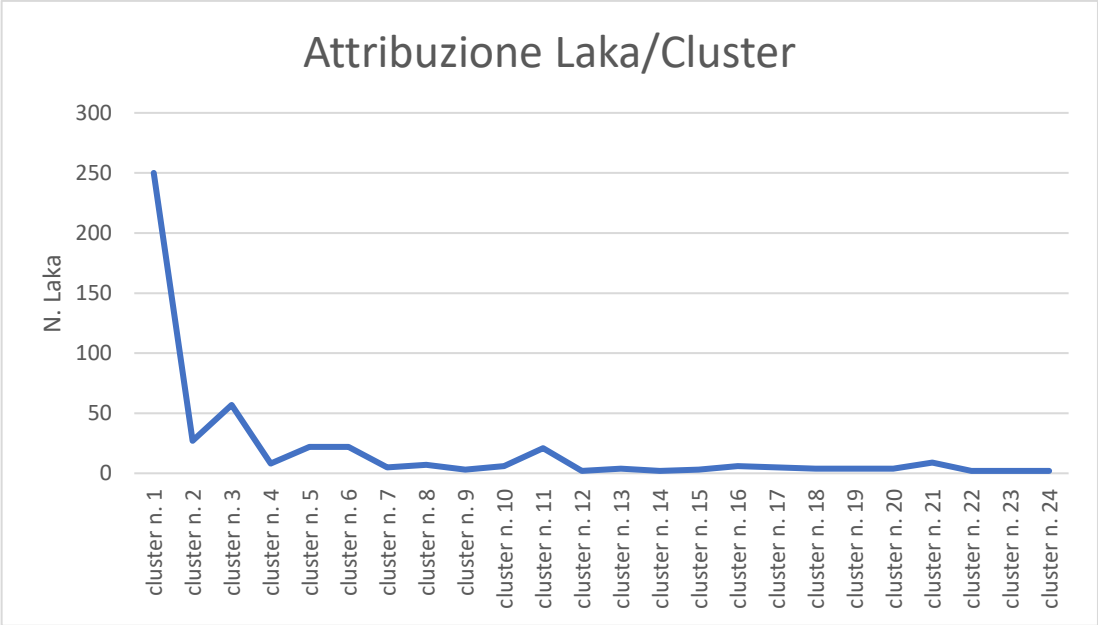


Figura 4.7 Distribuzione Laka ai cluster: Complete Linkage

Anche in questo caso è possibile notare la distribuzione della clusterizzazione che l'algoritmo ***Complete*** determina è fortemente sbilanciata verso il cluster n. 1con 250 Laka attribuiti.

Ward Linkage - I risultati ottenuti dall'elaborazione del seguente algoritmo scritto in Python al quale viene fornito in input la matrice di similarità calcolata tramite il Jaccard Similarity Index (tab. 4.7) sono i seguenti:

Cluster	nLaka
cluster n. 1	16
cluster n. 2	14
cluster n. 3	10
cluster n. 4	6
cluster n. 5	14
cluster n. 6	9
cluster n. 7	24
cluster n. 8	6
cluster n. 9	21
cluster n. 10	6
cluster n. 11	4
cluster n. 12	3
cluster n. 13	6
cluster n. 14	7
cluster n. 15	64
cluster n. 16	64
cluster n. 17	29
cluster n. 18	21
cluster n. 19	60
cluster n. 20	18
cluster n. 21	30
cluster n. 22	12
cluster n. 23	33

Tabella 4.7 Attribuzione dei cluster per ward linkage

La tabella riassume le aggregazioni effettuate dall' algoritmo, che si possono rappresentare in un dendrogramma:

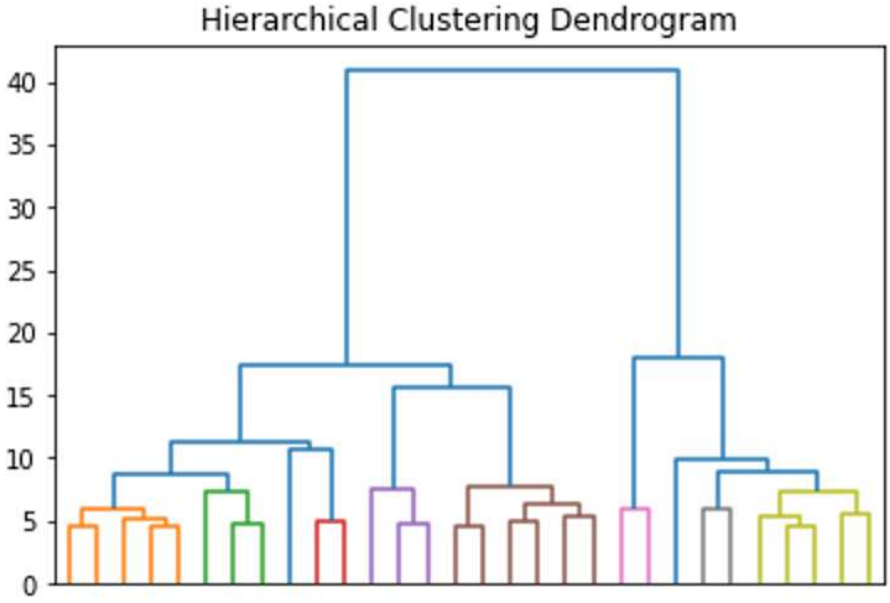


Figura 4.8 Dendrogramma Ward Linkage

Come si può notare dall'estratto della tabella e dal dendrogramma anche questo algoritmo con una **threshold** pari a **5** genera un totale di **23 cluster** e un coefficiente **Davies** pari a **7.4926**.

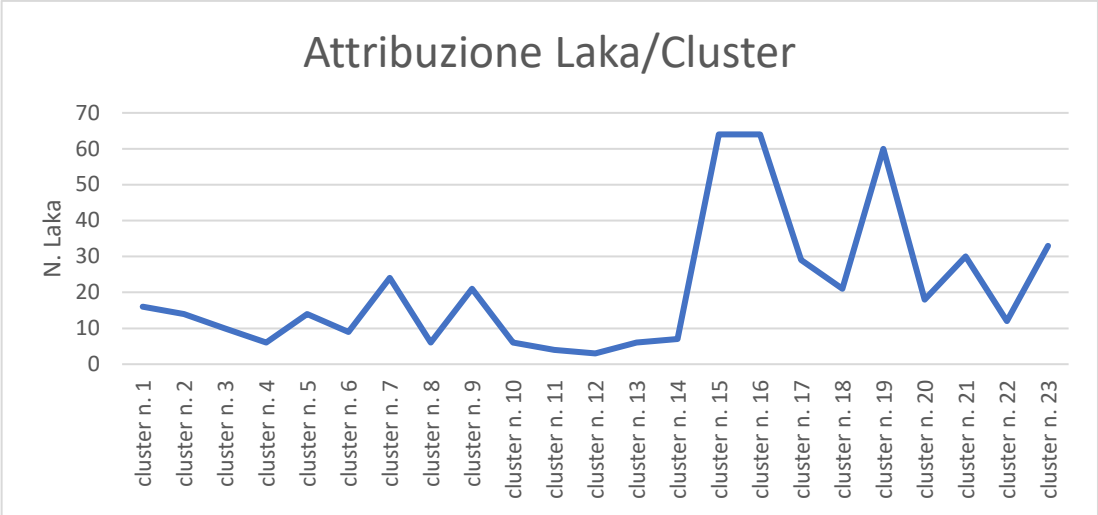


Figura 4.9 Distribuzione Laka ai cluster: Ward Linkage

L'algoritmo Ward restituisce la migliore distribuzione dei cluster, come si può notare del grafico di figura 4.9.

A questo punto possiamo valutare quale sia l'algoritmo che meglio definisce i cluster in base allo Score restituito dall'indice di Davis:

Algoritmo	Threshold	Cluster	Score (davies)
Single	3	25	1,4283
Complete	3,8	24	6,4248
Ward	5	23	7,4926

Tabella 4.8 confronto tra algoritmi e parametri selezionati

Come possiamo notare l'algoritmo Ward restituisce il miglior score pari a 7,4926, partizionando i Laka in 23 cluster, partendo da questo risultato per meglio visualizzare la distribuzione dei Laka in base alla loro clusterizzazione li aggreghiamo in base alla loro categoria merceologica, tabella 4.9.

Categoria merciologica	Media Similarity	Num. Laka	Num. Cluster
ACCESSORI CASA	0,0028	4	2
AGENZIE VIAGGIO	0,0033	3	3
ALBERGHI	0,0024	3	2
ALTRI PRODOTTI RETAIL	0,0030	4	2
ANIMALI E ACCESSORI	0,0036	4	4
ARREDAMENTO	0,0030	4	4
ASSICURAZIONI E FINANZIARIE	0,0025	21	7
ASSOCIAZIONI E NO PROFIT	0,0008	5	1
ATTIVITA RICREATIVE	0,0034	28	8
CALZATURE	0,0050	12	5
CARBURANTI E STAZIONI DI SERVIZIO	0,0059	26	6
CARTOLIBRERIE	0,0066	10	5
CASALINGHI	0,0060	10	6
DIRECT MARKETING E VPC	0,0025	4	4
ELETTRODOMESTICI	0,0035	8	6
ENTI PUBBLICI	0,0030	16	7
GIOCATTOLI	0,0066	2	2
GIOIELLERIE	0,0032	12	6
GRANDE DISTRIBUZIONE NON ALIMENTARE	0,0072	19	8
INFORMATICA	0,0059	8	5
LINEE AEREE	0,0025	5	3
NON DEFINITA	0,0122	20	1
PEDAGGI	0,0318	2	2
PIANTE E FIORI	0,0093	2	2
PRODOTTI SANITARI	0,0049	14	8
PRODOTTI VARI	0,0018	4	1
PROFUMERIE	0,0046	15	9
RISTORANTI	0,0048	38	8
SCUOLE	0,0023	7	2
SERVIZI FOTOGRAFIA E RIPRODUZIONE	0,0008	5	1
SERVIZI MEDICO-SANITARI	0,0024	7	4
SERVIZI NOLEGGIO	0,0021	8	5
SERVIZI VARI	0,0023	28	6
SERVIZI VIA CAVO	0,0149	12	2
SPECIALISTI E DISTRIBUZIONE TRADIZIONALE	0,0032	8	4
TELEFONIA	0,0101	8	4
TRASPORTI COSE	0,0006	23	2
TRASPORTI PERSONE	0,0059	22	9
VENDITA E SERVIZI VEICOLI	0,0035	10	6
VESTIARIO	0,0008	36	3

Tabella 4.9 distribuzione delle categorie merceologiche per laka d'appartenenza

La tabella per ogni categoria merceologica riporta:

1. La media delle similarità di tutti i Laka appartenenti alla categoria merceologica.
2. Il numero di Laka appartenenti alla categoria merceologica.
3. Il numero di cluster che contengono almeno un Laka appartenente alla categoria merceologica.

Di seguito il grafico ad albero della distribuzione delle categorie merceologiche sulla clusterizzazione dei Laka.

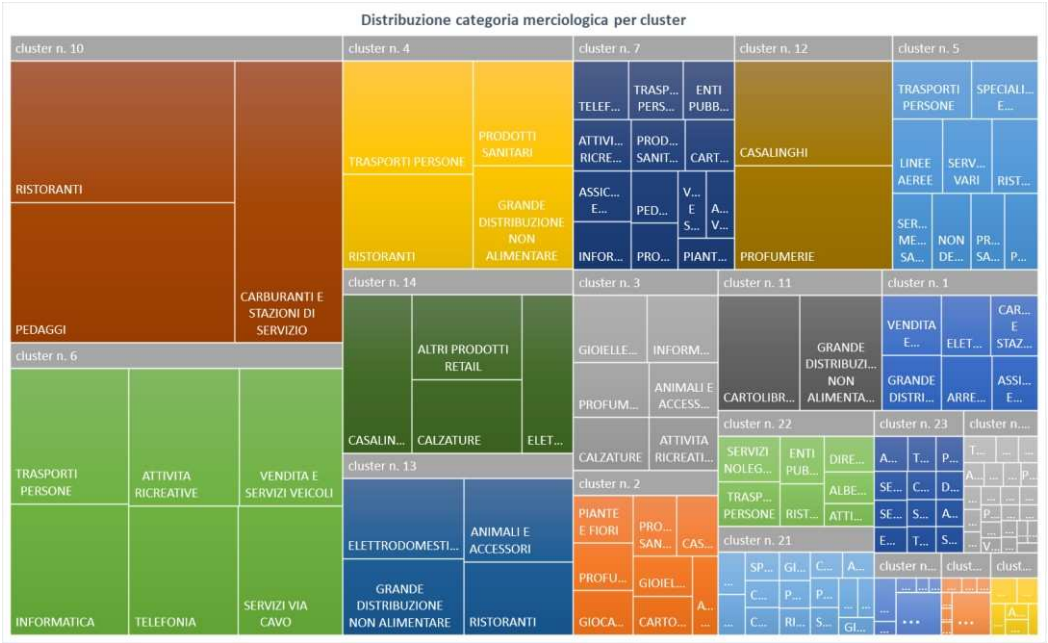


Figura 4.10 distribuzione delle categorie merceologiche per laka d'appartenenza

I risultati ottenuti ci permettono di ipotizzare azioni di marketing mirate che possono sfruttare la sinergia non scontata tra esercenti appartenenti spesso a categorie merceologiche molto diverse tra loro, ma che grazie a questa analisi, risultano avere una clientela discretamente sovrapposta.

Osservando la matrice di similarità (tab. 4.4) notiamo la presenza di relazioni di similarità significative tra diversi Laka, le relazioni generano delle dipendenze logiche rappresentabili attraverso un grafo non orientato e pesato, dove i nodi corrispondono ai Laka, gli archi alle relazioni di similarità e infine i pesi degli archi alla percentuale di similarità che lega i due nodi, un esempio è riportato nella Figura 4.11.

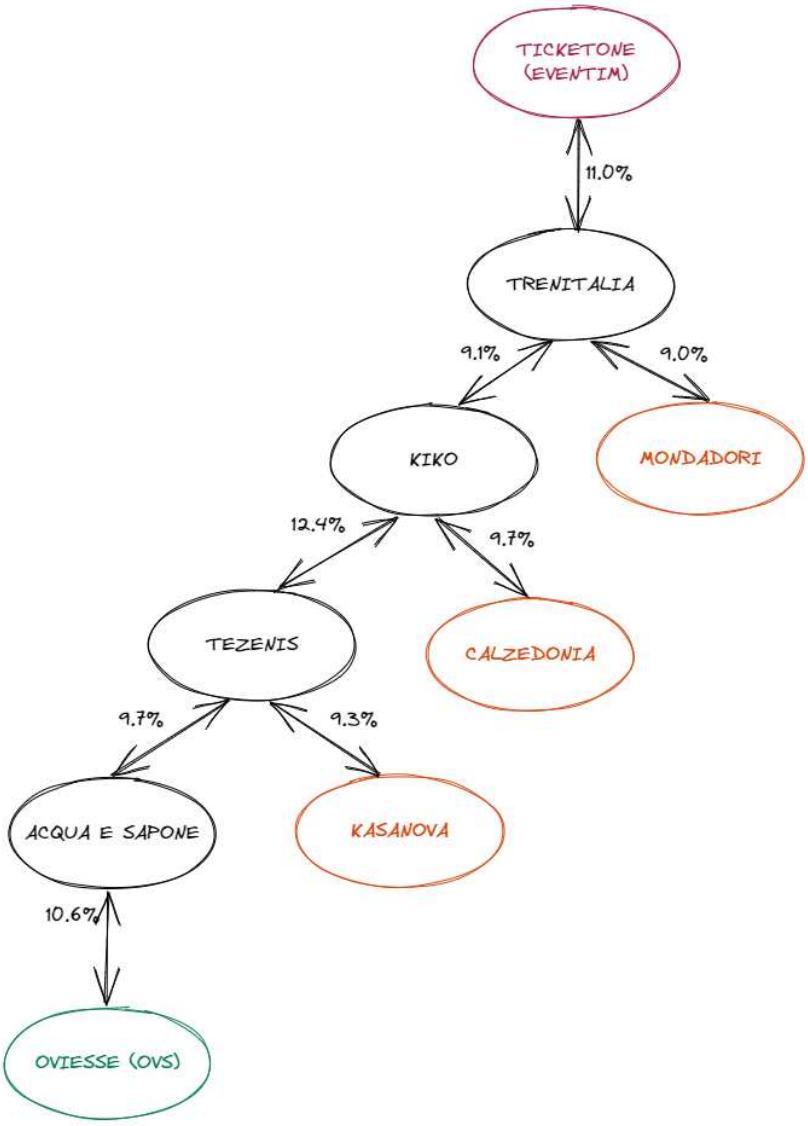


Figura 4.11 Grafo non orientato pesato delle similarità fra un subset di Laka.

Il grafo (fig. 4.11) rappresenta le relazioni tra un subset dei Laka analizzati.

La comprensione del mercato e delle abitudini dei consumatori che derivano da questa analisi permette ad esempio di giustificare diverse

forme di incentivazione prevedibili (es. suggerire o incentivare l'acquisto di un biglietto *Trenitalia* a fronte dell'acquisto di un biglietto *TicketOne* per un concerto di Maneskin a Milano), ma anche meno scontate, ad esempio studiando l'iterazione tra alcuni brand del mondo dell'abbigliamento e negozi specialistici della grande distribuzione non alimentare.

Più in generale individuando le relazioni fra le categorie merceologiche di appartenenza dei Laka e la loro clusterizzazione (tab. 4.10), è possibile definire il rispettivo grafo delle dipendenze logiche (fig. 4.12).

Laka	Categoria Merceologica	Cluster
TICKETONE (EVENTIM)	ATTIVITA RICREATIVE	Cluster n. 6
TRENITALIA	TRASPORTI PERSONE	
CALZEDONIA	VESTIARIO	Cluster n. 8
TEZENIS	VESTIARIO	
OVIESSE (OVS)	VESTIARIO	
MONDADORI	CARTOLIBRERIE	Cluster n. 11
KIKO	PROFUMERIE	Cluster n. 12
KASANOVA	CASALINGHI	
ACQUA E SAPONE	CASALINGHI	
…	…	…

Tabella 4.10 Categorie merceologiche NEXI, Cluster per Laka

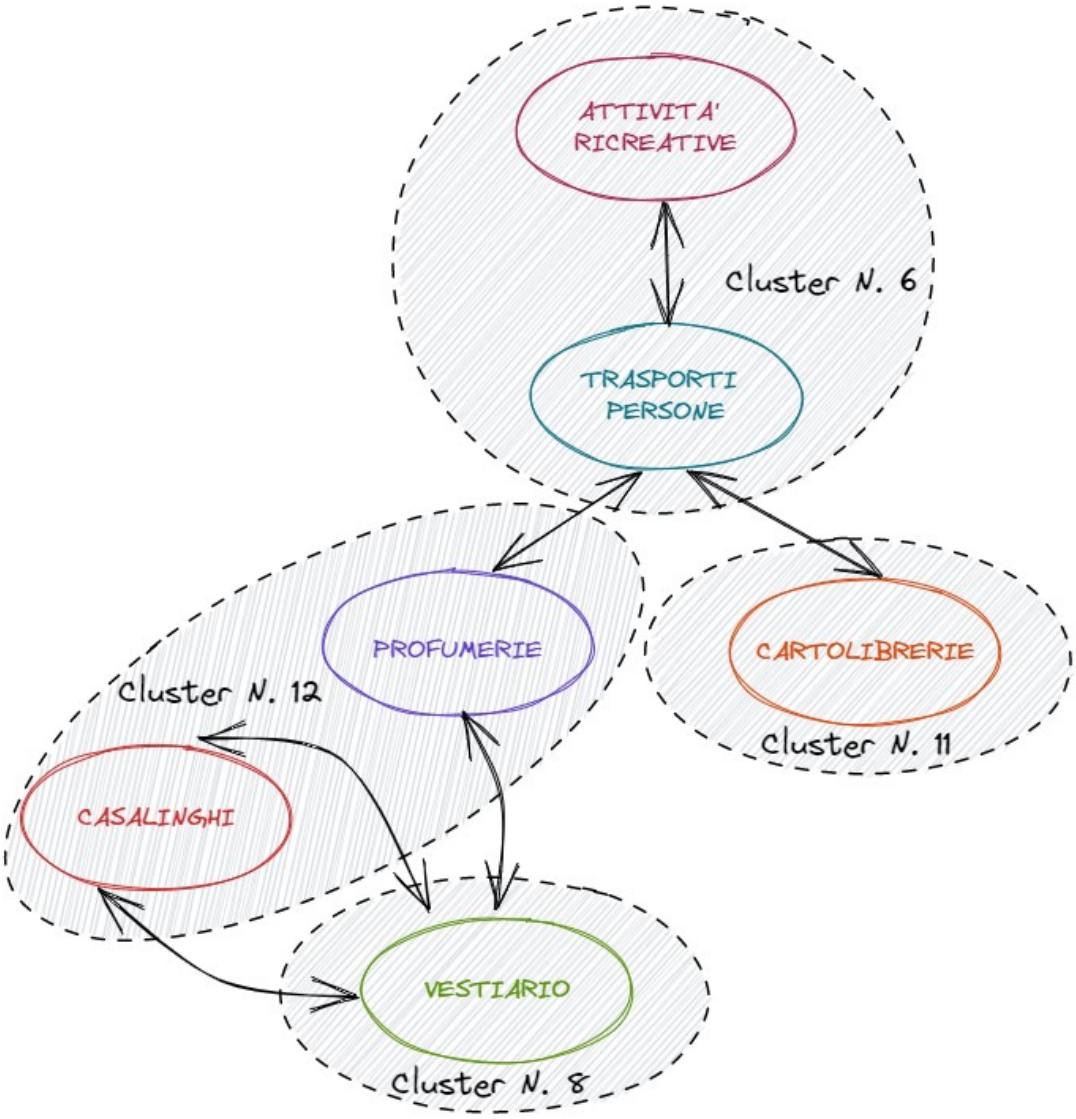


Figura 4.12 Grafo delle dipendenze fra le categorie merceologiche e i cluster

e in modo del tutto analogo sarà possibile definire delle azioni di marketing rivolte non ai singoli Laka ma bensì a categorie merceologiche di Laka o in modo ancora più aggregato a cluster di Laka.

Capitolo 5
SVILUPPI FUTURI: RECOMMENDER SYSTEM

La cluster analysis realizzata rappresenta un punto di partenza per fornire una lista di Laka da raccomandare ad ogni utente sulla base degli acquisti passati. Nello specifico verrà implementato un sistema di raccomandazioni appartenente alla categoria degli approcci collaborativi used based che prende in considerazione la similarità dei comportamenti di acquisto degli utenti per predire/suggerire gli acquisti futuri. Da questo principio costruiremo un modello più adatto ai fini di un'estrazione e visualizzazione di una *Reccommendation list.*

I Recommender System sono strumenti software progettati come una particolare forma di filtraggio intelligente delle informazioni con l'obiettivo di suggerire item agli utenti. Essi sono utilizzati per generare una lista ordinata di proposte realizzata su misura tramite la scoperta di similarità tra utenti e/o elementi a catalogo. Al giorno d'oggi si fa ampio utilizzo di questi sistemi; un esempio è quando cerchiamo un prodotto su Amazon, dove oltre alle nostre ricerche si presentano quei "prodotti

consigliati", acquistati da utenti che hanno gusti simili ai nostri, o prodotti correlati a prodotti precedentemente acquistati. U altro esempio "classico" è attraverso Spotify che crea delle playlist "ad-hoc" per l'ascoltatore, in funzione dei brani ascoltati.

Gli obiettivi di un Recommender System possono essere molteplici e variano in funzione dell'ambito in cui è utilizzato. I più importanti sono incrementare:

- *Le vendite dei prodotti*
- *La soddisfazione dell'utente.*
- *La Fedeltà dell'utente*
- *La comprensione dei bisogni degli utenti*

Per i Recommender System, come per tutti i problemi di Machine Learning, le tecniche e i modelli che si usano dipendono fortemente dalla quantità e dalla qualità dei dati in possesso.

I recommender system possono essere classificabili in diverse categorie, in base al modo con cui vengono generati i suggerimenti.

Collaborative Filtering - L'idea alla base dell'approccio collaborative filtering considera che se due utenti A e B hanno

effettuato acquisti simili negli anni e di recente l'individuo A ha comprato in un merchant che B non ha ancora visto, allora è opportuno raccomandare quel merchant anche a B. In questo caso, per suggerire un prodotto o servizio da un più ampio set di alternative si attinge dalle preferenze degli utenti e non si tengono in considerazione la natura, le caratteristiche o i contenuti degli item raccomandati. Tanto più saranno le informazioni circa le preferenze degli utenti, tanto più accurati saranno i risultati. Possiamo dividere il filtraggio in 2 sotto categorie:

- User-based
- Item-based

Nell'approccio *user-based*, per fornire raccomandazioni si ricercano i k utenti più simili all'utente attivo, sulla base delle valutazioni fornite, e si predicono le valutazioni mancanti calcolando una media pesata delle valutazioni fornite dagli stessi. La similarità viene quindi calcolata tra le righe della matrice delle valutazioni per identificare gli utenti simili.

Nell'approccio *item-based*, invece, per predire la valutazione dell'utente attivo per un determinato articolo di interesse si

determina l'insieme S degli articoli maggiormente simili a tale articolo, sulla base delle valutazioni ricevute, e si predice la valutazione per tale articolo calcolando una media pesata delle valutazioni degli articoli simili.

Content-Based - Utilizza le descrizioni egli articoli per consigliare altri articoli simili, cercando di abbinare agli utenti elementi che sono simili a ciò che hanno apprezzato in passato.
Ad esempio, se a John piace il film di fantascienza futuristica Terminator, allora c'è un'alta probabilità che gli possa piacere un film di un genere simile, come Aliens.
I sistemi basati sul contenuto dipendono da due fonti di dati:

- La prima fonte è una descrizione di vari articoli in termini di attributi incentrati sul contenuto.
- La seconda fonte di dati è un profilo utente, che è generato dal feedback degli utenti su vari articoli.

Non sfruttando esplicitamente le valutazioni di altri utenti, nei sistemi basati sul contenuto gli altri utenti sono di poco interesse per la generazione di raccomandazioni.

Knowledge-Based - Sia i sistemi basati sui contenuti che quelli collaborativi richiedono una quantità significativa di dati sulle precedenti esperienze di acquisto e di valutazione. Inoltre, questi metodi non sono generalmente adatti a domini in cui il prodotto è altamente personalizzato.

Qui le raccomandazioni sono ottenute indipendentemente dalle valutazioni sui singoli utenti; esse sono il risultato di un'alta somiglianza tra le esigenze dei clienti e gli articoli oppure sono ottenute sulla base di regole di raccomandazione esplicite. La particolarità di un sistema basato sulla conoscenza è la sua alta interattività. Per tale motivo, questi sistemi sono definiti come sistemi che guidano un utente in modo personalizzato verso oggetti interessanti o utili in un ampio spazio di opzioni possibili o che restituiscono tali oggetti come output. Esistono due tipi fondamentali di sistemi di raccomandazione, questi sistemi sono:

- Sistemi di raccomandazione basati su vincoli: in questo caso, gli utenti specificano requisiti o vincoli

- Sistemi di raccomandazione basati su casi: si basano su un insieme di regole di raccomandazione esplicitamente definite

Entrambi gli approcci sono simili in termini di processo di raccomandazione: l'utente deve specificare i requisiti, e il sistema cerca di identificare una soluzione, se non si trova una soluzione l'utente deve cambiare i requisiti.

Ibrido - I tre approcci di raccomandazione discussi fino adesso sfruttano diverse fonti di informazione e seguono diversi paradigmi per dare suggerimenti agli utenti. Anche se producono risultati che sono personalizzati in base agli interessi presunti dei loro destinatari, essi operano con diversi gradi di successo in diversi campi di applicazione.

Il Collaborative Filtering sfrutta le valutazioni fornite agli articoli dalla comunità per suggerire le raccomandazioni; gli approcci Content-Based, invece, si basano sulle caratteristiche del prodotto e sulle descrizioni testuali, mentre gli algoritmi basati su Knowledge Based, ragionano su un modello di conoscenza

esplicito del dominio. Ciascuno di questi approcci ha i suoi pro e contro. I modelli degli utenti e le informazioni contestuali, i dati delle comunità e dei prodotti e i modelli di conoscenza costituiscono i potenziali tipi di input della raccomandazione. Tuttavia, nessuno degli approcci di base è in grado di sfruttarli tutti. Di conseguenza, è giustificata la costruzione di sistemi ibridi, che combinano i punti di forza di diversi algoritmi e modelli. La ricerca attuale sta andando in questa direzione.

CONCLUSIONI

Da quanto discusso in questa tesi si è compresa la sempre crescente importanza che hanno i pagamenti digitali nella società attuale, in particolare nel contesto di Nexi Group.

Grazie alla Data Science è stato possibile comprendere e analizzare i fenomeni reali tramite l'utilizzo dei dati, che adeguatamente manipolati e analizzati con una profonda conoscenza di dominio permettono molteplici possibilità di utilizzo.

Si è compresa l'importanza delle tecniche di intelligenza artificiale, nello specifico, grazie a tecniche avanzate di clustering, per realizzare un modello di apprendimento che permetta di scoprire relazioni non scontate tra diversi grandi esercenti (Laka) e quindi proporre soluzioni commerciali innovative per favorire una sempre maggiore adozione dei pagamenti digitali da parte della clientela

Questo lavoro di tesi è un esempio della direzione che si può osservare in tutte le aziende a forte connotazione tecnologica: l'uso dei dati e degli algoritmi non è più solo "a servizio" delle esigenze

operative o delle richieste del business, ma è a tutti gli effetti “al suo fianco”, in quanto rappresenta oggi la vera fonte di vantaggio competitivo per avere successo in tutti i mercati più competitivi.

BIBLIOGRAFIA

[1] EVANS, SCHMALENSEE, Innovation in payments, in Market Platform Dynamics, settembre, 2008, available on ssrn.com

[2] L. Incorvati, Addio al contante, nel 2021 balzo in avanti del 22% dei pagamenti digitali, *ilsole24ore*, 2022.

[https://www.ilsole24ore.com/art/addio-contante-2021-balzo-avanti-22percento-pagamenti-digitali-AElSlTMB?refresh_ce=1]

[3] Deepak Jakhar Ishmeet Kaur, "Artificial intelligence, machine learning and deep learning: definitions and differences". Clinical and Experimental Dermatology, 2019. [https://www.researchgate.net/publication/334014738_Artificial_intelligence _machine_learning_deep_learning_Definitions_and_differences]

[4] Yann LeCun, Yoshua Bengio e Geoffrey Hinton, "Deep learning". Nature, 2015. [https://www.nature.com/articles/nature14539]

[5] Vasant Dhar, "Data Science and Prediction". Communications of the ACM, 2013. [https://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction /fulltext]

[6] Chikio Hayashi, Noboru Ohsumi, Hans H. Bock, Keiji Yajima, Yutaka Tanaka e Y.Baba, "Data Science, Classification, and Related Methods". Springer, 1998. [https://www.springer.com/it/book/9784431702085]

[7] Tony Hey, Anthony J. G. Hey, Stewart Tansley e Kristin Michele Tolle, "The Fourth Paradigm: Data-intensive Scientific Discovery". Microsoft Research, 2009. [https://books.google.it/books?id=oGs_AQAAIAAJ&redir_esc=y]

[8] Sarah Guido e Andreas C. Mueller "Introduction to Machine Learning with Python: A Guide for Data Scientists". O'Reilly Media, 2016 [https://www.amazon.it/Introduction-Machine-Learning-Python-Sarah/dp /1449369413]

[9] Ornella Colpani, "Machine learning: the ability to predict applied to research and clinical practice". Giornale Italiano di Farmacoeconomia e Farmacoutilizzazione, 2019. [http://www.sefap.it/web/upload/GIFF2019-4_5_11.pdf]

[10] Daniel R. Schrider e Andrew D. Kern, "Supervised Machine Learning for Population Genetics: A New Paradigm". Trends in Genetics, 2018. [https://www.researchgate.net/publication/322397567_Supervised_ Machine _Learning_for_Population_Genetics_A_New_Paradigm]

[11] Xiangying Wang e Yixin Zhong, "Statistical learning theory and state of the art in SVM". The Second IEEE International Conference on Cognitive Informatics, 2003. [https://ieeexplore.ieee.org/abstract/document/1225953]

[12] Michael W. Berry, Azlinah Mohamed e Bee Wah Yap, "Supervised and Unsupervi- sed Learning for Data Science". Springer International Publishing, 2019. [https://books.google.it/books/about/Supervised_and_Unsupervised _Learning_for.html?id=0n6qzQEACAAJ&redir_esc=y]

[13] T. Soni Madhulatha, "An Overview on Clustering Methods". IOSR Journal of En- gineering, 2012.
[https://arxiv.org/abs/1205.1117]

[14] Christopher M. Bishop, "Pattern Recognition And Machine Learning". Springer Nature, 2006.
[https://www.amazon.it/Pattern-Recognition-Machine-Learning-Christopher /dp/0387310738]

[15] Sonish Sivarajkumar, "ReLU - Most popular Activation Function for Deep Neural Networks". Medium, 2019.
[https://medium.com/@sonish.sivarajkumar/relu-most-popular-activation-function-for-deep-neural-networks-10160af37dda]

[16] Yusuf Sani, Ahmed Mohamedou, Khalid Ali, Anahita Farjamfar, Mohamed Azman e Solahuddin Shamsuddin, "An Overview of Neural Networks Use in Anomaly In- trusion Detection Systems". IEEE, 2009.
[https://ieeexplore.ieee.org/document/5443289]

[17] Ian Goodfellow, Yoshua Bengio e Aaron Courville, "Deep Learning". MIT Press,2015.
https://books.google.it/books/about/Deep_Learning.html?id=Np9SDQAAQBAJ&redir_esc=y]

[18] Sebastian Ruder "An overview of gradient descent optimization algorithms". CornellUniversity, 2016.
[https://arxiv.org/abs/1609.04747]

[19] Tyron R.C., Bailey D.E. (1970) Cluster Analysis, McGraw-Hill.

[20] Tryon R.C. (1939) Cluster analysis: correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality, Edwards brother, Incorporated, lithoprinters and publishers.

[21] Hartigan, 1975

[22] Halkidi M., Batistakis Y., Vazirgiannis M. (2001) On clustering validation techniques, Journal of Intelligent Information Systems, 17(2-3), 107-145.

[23] Jain A., Dubes R., Algorithms for Clustering Data, Englewood Cliffs, NJ: Prentice-Hall, 1988.

[24] Rousseeuw P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics, 20, 53-65.

[25] Dunn† J.C. (1974) Well-separated clusters and optimal fuzzy partitions, Journal of cybernetics, 4(1), 95-104.

[26] Davies D.L., Bouldin D.W. (1979) A cluster separation measure, Pattern Analysis and Machine Intelligence, IEEE Transactions on, (2), 224-227.

[27] Definizione di cloud computing (2020) corso interno all'azienda, Alberto Danese

RINGRAZIAMENTI

A conclusione di questo elaborato vorrei ringraziare tutti coloro che mi hanno sostenuto durante questo percorso universitario.

Desidero innanzitutto ringraziare il Prof. Filippo Emanuele Ciarapica relatore di questa tesi, per il tempo che mi ha dedicato, la sua professionalità e competenza.

Un ringraziamento particolare va al mio correlatore Ing. Alberto Danese e la collega Ing. Marta Toschi per avermi supportata e seguita nella realizzazione del progetto

Il ringraziamento più grande va a tutta la mia famiglia, mio marito e i miei figli, che mi hanno sopportata e sostenuta in questi anni di studio.