

# LUISS



Cattedra

---

RELATORE

---

CORRELATORE

---

CANDIDATO

Anno Accademico



# INDICE

<b>ABSTRACT</b> .....	<b>3</b>
<b>INTRODUZIONE</b> .....	<b>4</b>
<b>CAPITOLO 1. INTRODUZIONE AL MACHINE LEARNING</b> .....	<b>6</b>
<b>1.1 OVERVIEW</b> .....	<b>6</b>
<b>1.2 CARATTERISTICHE</b> .....	<b>9</b>
1.2.1 SUPERVISED LEARNING.....	11
1.2.2 UNSUPERVISED LEARNING .....	13
1.2.3 REINFORCEMENT LEARNING .....	15
<b>1.3 CLUSTERING</b> .....	<b>16</b>
1.3.1 OVERVIEW.....	17
1.3.2 CLUSTERING GERARCHICO .....	19
1.3.3 CLUSTERING NON GERARCHICO (K-MEANS) .....	21
<b>CAPITOLO 2. STATO DELL'ARTE</b> .....	<b>24</b>
<b>2.1. MOTIVI PER L'ACQUISTO E IL CONSUMO DI CAFFÈ</b> .....	<b>24</b>
<b>2.2. APPLICAZIONE DI METODI DI MACHINE LEARNING PER PREDIRE E TESTARE LA QUALITÀ DEI PRODOTTI</b> .....	<b>28</b>
<b>CAPITOLO 3. COFFEE QUALITY DATABASE FROM CQI: ANALISI SPERIMENTALE SULLA QUALITÀ DI ALCUNE TIPOLOGIE DI COFFEE BEANS</b> .....	<b>34</b>
<b>3.1 LINGUAGGIO R: OVERVIEW</b> .....	<b>34</b>
<b>3.2 IL DATASET COFFEE QUALITY DATABASE FROM CQI: CARATTERISTICHE</b> .....	<b>36</b>
<b>3.3 METODOLOGIA</b> .....	<b>38</b>
3.3.1 CLUSTERING GERARCHICO .....	41
3.3.2 CLUSTERING NON GERARCHICO (K-MEANS) .....	46
<b>3.4 DISCUSSIONE DEI RISULTATI</b> .....	<b>57</b>
<b>CONCLUSIONI</b> .....	<b>60</b>
<b>ESTENSIONI FUTURE</b> .....	<b>62</b>
<b>ELENCO DELLE FIGURE</b> .....	<b>63</b>
<b>APPENDICE</b> .....	<b>64</b>
<b>BIBLIOGRAFIA</b> .....	<b>81</b>
<b>SITOGRAFIA</b> .....	<b>84</b>



## ABSTRACT

Ai giorni d'oggi i consumatori cercano sempre di più prodotti che siano in linea con i loro gusti e con le loro aspettative. Per un brand è estremamente importante soddisfare i propri consumatori in modo tale da creare una fedeltà in loro e far sì che li riacquistino. La qualità è per certo un attributo a cui gli acquirenti fanno caso; infatti, secondo uno studio<sup>1</sup> è strettamente legato alla percezione della sicurezza del prodotto in questione. La valutazione della qualità è un processo complesso e spesso il consumatore si ritrova in una situazione di incertezza. La ragione è che molti aspetti del prodotto sono difficili da valutare a causa dell'assenza di informazioni al momento dell'acquisto. Questo accade specialmente nell'ambito dei prodotti alimentari e delle bevande. Per aiutare i consumatori in fase di scelta bisognerebbe che i brand gli dessero maggiori informazioni. Da qui nasce l'obiettivo della presente ricerca, che è quello di valutare, tramite i rating dati da esperti, alcune tipologie di chicchi di caffè. L'analisi sarà effettuata attraverso il clustering di essi sulla base di attributi appartenenti alla macrocategoria della qualità. A questo proposito, la domanda di ricerca è la seguente: quali attributi i brand di caffè possono mettere in luce per ottimizzare la comunicazione dei loro prodotti? Per rispondere alla domanda di ricerca è stato analizzato un dataset che racchiude i punteggi dati da esperti del Coffee Quality Institute ai chicchi di caffè presi in esame. L'analisi è stata effettuata tramite l'algoritmo di clustering, nello specifico sono state applicate entrambe le tipologie: prima il clustering gerarchico e in seguito il k-means nel linguaggio di programmazione R. Le osservazioni sono state suddivise in quattro cluster e in seguito ogni gruppo è stato analizzato per andare a vedere in quali attributi ha ottenuto dei punteggi più alti. Sulla base di quanto ottenuto è stato possibile capire su quali attributi i brand di caffè possono fare leva per dare maggiori informazioni ai consumatori sulla qualità dei loro prodotti.

---

<sup>1</sup> (Sadilek, 2019)

# INTRODUZIONE

Negli ultimi anni, i Big Data hanno creato importanti opportunità nel marketing delle imprese in quanto gli hanno permesso di avere un punto di vista analitico nello sviluppo delle strategie di comunicazione<sup>2</sup>. Ma cosa si intende quando si parla di Big Data? Il termine si riferisce alla raccolta di un'enorme quantità di dati di diverse tipologie e con una struttura complessa basata sul modello di elaborazione e applicazione dei dati stessi. I Big Data sono sempre più presenti nella vita di tutti i giorni e hanno un ruolo importante nello sviluppo futuro delle imprese. È importante per un'azienda stare al passo coi tempi in modo che possa progredire sempre di più e sviluppare strategie sempre più innovative in modo da avere un vantaggio competitivo rispetto alla concorrenza<sup>3</sup>. L'avvento dei Big Data ha portato la necessità di utilizzare il Machine Learning per scoprire modelli e fare previsioni sempre più accurate e tempestive<sup>4</sup>. Il Machine Learning è sempre più utilizzato in moltissimi ambiti, tra questi sta giocando un ruolo importante nella valutazione della qualità dei prodotti. Essa è un attributo estremamente importante per la soddisfazione del consumatore finale e per questo motivo l'ottenimento di risultati più precisi ed accurati, grazie all'utilizzo di questi algoritmi, può dare un grosso aiuto alle aziende. Proprio alla luce di queste informazioni e dell'importanza della qualità come attributo determinante nel settore alimentare e delle bevande, la presente tesi mira ad applicare il Machine Learning alla valutazione della qualità dei chicchi di caffè presi in esame sulla base dei rating dati da esperti del prodotto. L'obiettivo finale di questo studio è quello di comprendere come si potrebbe ottimizzare la comunicazione effettuata dai brand per queste tipologie di caffè mettendo in luce i giusti attributi come punti di forza. Per lo svolgimento di tale studio si utilizza il dataset Coffee Quality database from CQI che si basa sulle valutazioni di chicchi di caffè date da esperti del Coffee Quality Institute. Queste valutazioni sono date a una serie di attributi che sono ritenuti fondamentali dai consumatori e determinanti per la preparazione finale di un buon caffè in tazza. Il lavoro si suddivide in tre capitoli. Nel primo capitolo si introduce il concetto di Machine Learning, un campo che appartiene all'insieme dell'intelligenza artificiale e che dà ai computer l'abilità di apprendere dai dati senza essere esplicitamente programmati. In seguito, si spiegano l'origine e le caratteristiche illustrando le principali tipologie che si

---

<sup>2</sup> (Jabbar, Akhtar, & Dani, 2020)

<sup>3</sup> (Hu, 2018)

<sup>4</sup> (Zhou, Pan, Wang, & Vasilakos, 2017)

suddividono in: apprendimento supervisionato, apprendimento non supervisionato e apprendimento con rinforzo. In seguito, è presente un focus sulla tipologia di apprendimento non supervisionato utilizzata nell'analisi sperimentale dell'elaborato: il clustering. Di questo si presentano le caratteristiche e le due tipologie esistenti: clustering gerarchico e clustering non gerarchico.

Nel secondo capitolo si pone l'attenzione sullo stato dell'arte, quindi una revisione della letteratura sulle applicazioni passate del Machine Learning nel settore delle bevande calde. Si presentano in primo luogo evidenze della letteratura sulle principali ragioni per cui i consumatori bevono il caffè e in seguito evidenze sull'applicazione del Machine Learning per testare la qualità dei prodotti, in particolare delle bevande calde.

Il terzo capitolo si divide in due parti. Nella prima viene presentato il software di programmazione R e il dataset di riferimento per l'analisi. Nella seconda parte viene svolta l'analisi sperimentale e si discutono i risultati che emergono. Infine, si propongono ulteriori estensioni future della presente ricerca.

# CAPITOLO 1. INTRODUZIONE AL MACHINE LEARNING

Il presente capitolo mira a fornire una panoramica del Machine Learning iniziando dalla nascita fino alla sua evoluzione mettendo l'accento anche sull'importanza che ha e sul progresso che ha portato. Prosegue con le caratteristiche e le varie tipologie di apprendimento esistenti. Infine, ci sarà un focus sull'algoritmo di clustering e le due macroaree esistenti.

## 1.1 OVERVIEW

Nel 1956 fu usata per la prima volta l'espressione "Intelligenza Artificiale" (AI). In quel periodo cinque scienziati americani cercarono di fare un esperimento, dato che la computazione stava iniziando ad emergere. La loro intenzione era quella di costruire con successo una macchina che imitasse l'intelligenza umana. Tuttavia, questo sforzo fu un fallimento e da allora l'intelligenza artificiale passò in secondo piano. Solo negli anni '80, con l'avvento dei "sistemi esperti", si ricominciò a parlarne. I sistemi esperti erano utilizzati per trasferire la conoscenza di un esperto in un programma per il computer. Eliza è stato il primo sistema esperto creato. Questo era un sistema con la funzione di psicologo e si poteva parlare con lei. Arthur Samuel è stato uno degli scienziati che hanno condotto l'esperimento. "La scienza che fornisce ai computer la capacità di imparare senza essere istruiti esplicitamente", così Samuel definiva l'intelligenza artificiale<sup>5</sup>. Di conseguenza, lo scienziato definiva l'apprendimento automatico come un insieme di metodi che consentono ai computer di acquisire conoscenze dai dati. Infatti, l'apprendimento automatico impiega algoritmi che assorbono la conoscenza dai dati. Ad esempio, consente di rilevare informazioni anche non identificate senza indicare specificamente dove cercarle<sup>6</sup>. Un nome da tenere a mente quando si parla di intelligenza artificiale è quello di Alan Turing. Egli è infatti considerato il padre dell'informatica e dell'intelligenza artificiale. Negli anni Cinquanta si è chiesto se le macchine potessero davvero pensare. Per rispondere, egli condusse un test noto come Test di Turing per determinare se le macchine potessero esibire un comportamento umano. Le grandi scoperte neurologiche

---

<sup>5</sup> (Machine Learning: Cos'è e perché è importante, 2020)

<sup>6</sup> (Machine learning: Che cos'è e perché è importante)

di quel periodo portarono i ricercatori a chiedersi se fosse possibile creare un cervello elettronico e, in caso affermativo, se una macchina potesse pensare e agire in modo simile all'uomo. Il 1956 è l'anno in cui si pensa sia stata scoperta l'intelligenza artificiale. La ragione è che un gruppo di scienziati si sono riuniti e fecero un'enorme quantità di invenzioni. Da allora, i progressi dell'intelligenza artificiale sono stati notevoli, a partire dal 1996, quando il computer Blue Deep ha sconfitto il campione mondiale di scacchi Garry Kasparov. Negli anni successivi, un'auto a guida autonoma ha vinto la DARPA Grand Challenge nel 2005 e la DARPA Urban Challenge nel 2007<sup>7</sup>. Con il progredire della tecnologia e le innovazioni più sofisticate, questa scienza continua a progredire, portando all'uso di algoritmi sempre più precisi e accurati. In realtà, grazie a questi progressi tecnologici, il Machine Learning ha innescato cambiamenti significativi in numerosi settori. È diventato uno strumento che numerosi settori possono utilizzare per automatizzare le procedure. Le tecnologie che utilizzano l'apprendimento automatico sono fondamentali, soprattutto per l'elaborazione di grandi volumi di dati, perché massimizzano le risorse computazionali e fanno risparmiare molto tempo<sup>8</sup>.

È fondamentale distinguere tra apprendimento automatico e intelligenza artificiale. Mentre l'apprendimento automatico sviluppa sistemi che imparano dai dati, "l'intelligenza artificiale è un termine generico e si riferisce a sistemi o macchine che imitano l'intelligenza umana". Poiché l'apprendimento automatico è un sottoinsieme dell'intelligenza artificiale, tutto ciò che è Machine Learning appartiene all'insieme dell'intelligenza artificiale ma non si può dire il contrario (figura 1)<sup>9</sup>. In altri termini, l'intelligenza artificiale è la branca della ricerca che crea l'architettura necessaria affinché le macchine funzionino in modo simile al cervello umano. Al contrario, l'apprendimento automatico è il metodo che consente alle macchine intelligenti di svilupparsi nel tempo imparando dai dati senza che gli venga detto esplicitamente cosa fare. L'esperienza è ciò che consente alle macchine di migliorare completamente in modo autonomo<sup>10</sup>. Il Deep Learning è un'ulteriore divisione del campo dell'apprendimento automatico. Tutti gli

---

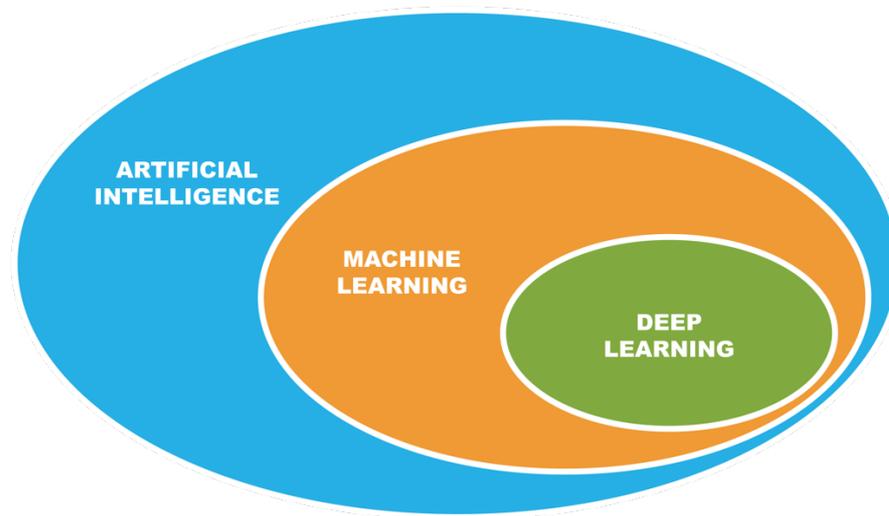
<sup>7</sup> (La storia dell'intelligenza artificiale, da Turing ad oggi, 2018)

<sup>8</sup> (Xu, Zhou, Sekula, & Ding, 2021)

<sup>9</sup> (Cos'è il machine learning?, 2021)

<sup>10</sup> (Artificial intelligence & Machine Learning)

algoritmi che sono modellati sull'organizzazione del cervello umano e comprendono una rete di neuroni artificiali sono definiti Deep Learning <sup>11</sup>.



**Fig.1.** *Relazione tra Intelligenza Artificiale, Machine Learning e Deep Learning*

La ripetizione è un fattore cruciale da tenere in considerazione quando si utilizza l'apprendimento automatico. I modelli devono essere esposti ai dati più volte per potersi adattare da soli. Al giorno d'oggi, l'apprendimento automatico è una realtà comune; infatti, ne incontriamo esempi ogni giorno. Un esempio è la guida autonoma dei veicoli, ma ce ne sono molti altri, tra cui tutti i motori di raccomandazione come quelli utilizzati da Netflix e Amazon<sup>12</sup>. L'informatica, la statistica, la probabilità, l'intelligenza artificiale, la psicologia, le neuroscienze e molti altri campi utilizzano l'apprendimento automatico, che è un campo molto ampio. Creando un modello che rappresenta accuratamente un insieme di dati scelti, l'apprendimento automatico consente di risolvere i problemi. La statistica è riuscita a trasformarsi in una scienza completa che produce teorie statistiche computazionali fondamentali sui processi di apprendimento grazie all'apprendimento

---

<sup>11</sup> (Boldrini, 2021)

<sup>12</sup> (Machine Learning: Che cos'è e perché è importante)

automatico, che nel tempo è diventato un campo sempre più sofisticato che insegna ai computer a imitare il cervello umano<sup>13</sup>. È quindi possibile che l'analisi intelligente dei dati diventi una componente sempre più vitale e necessaria nell'evoluzione della tecnologia, man mano che aumenta la quantità di dati disponibili<sup>14</sup>.

## 1.2 CARATTERISTICHE

Il "mezzo" attraverso cui si manifesta l'intelligenza artificiale può essere descritto come apprendimento automatico. Più nel dettaglio, il computer riceve una raccolta di dati, li analizza e ne ricava conoscenza. Il sistema modifica poi automaticamente gli algoritmi man mano che aumenta la quantità di informazioni che elabora<sup>15</sup>. La ricerca di regolarità statistiche e di altri modelli di dati è un processo che costituisce l'apprendimento. Lo sviluppo di algoritmi di apprendimento automatico mira a simulare il processo di apprendimento umano durante la padronanza di un determinato compito. Questi algoritmi possono anche servire come rappresentazione di quanto sia impegnativo l'apprendimento in ambienti diversi<sup>16</sup>. L'obiettivo dell'apprendimento automatico è scoprire l'accessibilità nascosta nei Big Data. Poiché i dati sono generati e inviati a livello di macchina, il Machine Learning è in grado di ridurre la dipendenza dalla singola traccia estraendo informazioni significative dall'analisi di fonti di dati grandi e distinte. L'apprendimento automatico è in grado di gestire insiemi di dati sempre più ampi perché si adatta alla complessità della gestione dei dati provenienti da molte fonti, nonché alla varietà di variabili e volumi di dati coinvolti. Un sistema di apprendimento automatico è in grado di essere addestrato e di generare risultati di valore tanto più elevato quanto maggiore è il numero di dati che gli vengono presentati. I sistemi di apprendimento automatico sono in grado di trovare e visualizzare modelli nascosti nei dati perché sono liberi dai vincoli del pensiero e della ricerca a livello individuale<sup>17</sup>. Esiste un approccio all'apprendimento automatico che può essere utilizzato, a seconda del tipo di problema da risolvere. Quando si tratta di utilizzare l'apprendimento automatico, la questione cruciale non è se un algoritmo di apprendimento sia migliore di altri, ma piuttosto in quali circostanze un approccio specifico potrebbe essere sensibilmente più efficace di altri su un determinato

---

<sup>13</sup> (Nasteski, An overview of the supervised machine learning methods, 2017)

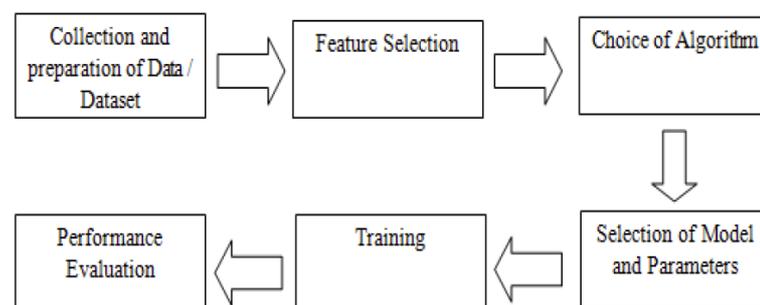
<sup>14</sup> (Osisanwo, et al., Supervised Machine Learning Algorithms: Classification and Comparison, 2017)

<sup>15</sup> (Rezzani A. , Apprendimento automatico e intelligenza artificiale, 2019)

<sup>16</sup> (Nasteski, An overview of the supervised machine learning methods, 2017)

<sup>17</sup> (Osisanwo, et al., Supervised Machine Learning Algorithms: Classification and Comparison, 2017)

problema applicativo. Si dovrebbe pensare di combinare diversi algoritmi per affrontare un problema se si conoscono meglio i vantaggi e gli svantaggi di ciascun approccio. L'obiettivo è quello di sfruttare i vantaggi di un metodo per compensare le carenze di un altro. Se l'obiettivo principale è quello di aumentare l'accuratezza, potrebbe essere difficile trovare un singolo metodo che funzioni bene quanto un buon insieme di metodi<sup>18</sup>. Tra le varie categorie di problemi ci sono le seguenti: problema di classificazione, problema di regressione e problema di clustering. Indipendentemente dall'algoritmo scelto, il modello generico di apprendimento automatico è composto da sei parti principali (Figura 2): 1) raccolta e preparazione dei dati (i dati devono essere in un formato che può essere dato in input all'algoritmo); 2) estrazione delle caratteristiche (selezionare solo i valori che possono essere rilevanti per essere utilizzati come input per l'apprendimento); 3) scelta di un algoritmo; 4) scelta del modello e dei parametri; 5) addestramento del modello; 6) valutazione del modello<sup>19</sup>.



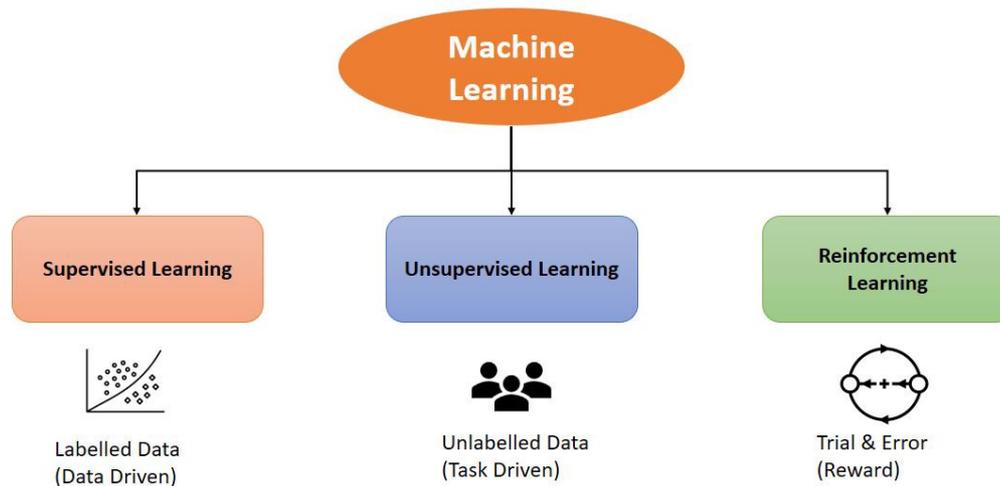
**Fig.2.** Componenti di un generico modello di Machine Learning

L'apprendimento automatico ha il potenziale di risolvere un problema con qualsiasi tipo di dati, e questo è uno dei suoi punti di forza. Questa capacità è resa possibile dal fatto che i dati sono divisi in due sottoinsiemi: l'insieme di addestramento per l'apprendimento

<sup>18</sup> (Osisanwo, et al., Supervised Machine Learning Algorithms: Classification and Comparison, 2017)

<sup>19</sup> (Alzubi, Nayyar, & Kumar, 2018)

del computer e l'insieme di test per valutare il funzionamento dell'algoritmo<sup>20</sup>. A seconda dell'algoritmo utilizzato, si possono distinguere tre tecniche di apprendimento: apprendimento supervisionato, apprendimento non supervisionato e apprendimento per rinforzo.



**Fig.3.** Tipologie di apprendimento del Machine Learning

### 1.2.1 SUPERVISED LEARNING

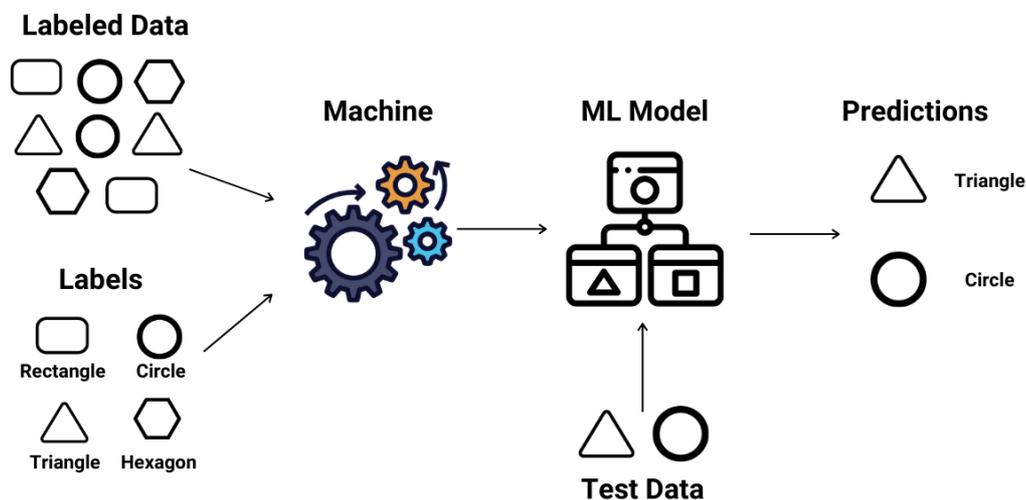
Il metodo più diffuso dei tre, l'apprendimento supervisionato, consiste nel fornire al computer concetti coordinati e precisi, seguiti da modelli e istanze che gli consentono di costruire un vero e proprio database di conoscenze ed esperienze. Ciò significa che tutto ciò che viene richiesto alla macchina è la capacità di fornire risposte basate su esperienze precedentemente codificate<sup>21</sup>. Una funzione dedotta può essere utilizzata per mappare nuovi esempi utilizzando un metodo di apprendimento automatico supervisionato che esamina i dati di formazione<sup>22</sup>. Un data scientist funge da insegnante e insegna

<sup>20</sup> (Velocci, Machine Learning: cos'è e perché è importante, 2020)

<sup>21</sup> (Machine Learning - I diversi apprendimenti di una macchina)

<sup>22</sup> (Sharma & Kumar, A Review on Machine Learning Algorithms, Tasks and Applications, 2017)

all'algoritmo come produrre risultati in questo tipo di apprendimento. L'algoritmo apprende da un set di dati che è stato precedentemente etichettato<sup>23</sup>, dove ogni input ha il suo output corrispondente, che viene utilizzato per insegnare all'algoritmo le regole del modello. La macchina deve aver già incontrato una circostanza simile quando calcola l'output. Il programma fornirà un risultato per analogia con l'esperienza precedente, utilizzando le probabilità corrispondenti alle opzioni potenziali<sup>24</sup>. Affinché il computer trovi una regola in grado di collegare i dati di ingresso con quelli di uscita, vengono forniti in ingresso una serie di dati e di informazioni correlate. In seguito, la macchina applicherà tale regola a ulteriori compiti comparabili<sup>25</sup>. Negli ultimi dieci anni sono state sviluppate numerose tecniche di apprendimento supervisionato nel settore dell'apprendimento automatico. Di fatto, l'apprendimento supervisionato è diventato un argomento cruciale per la ricerca sull'apprendimento automatico. Numerosi approcci di apprendimento supervisionato sono stati utilizzati per l'elaborazione e l'analisi di diversi tipi di dati. La capacità dell'apprendimento supervisionato di annotare i dati di addestramento è una delle sue caratteristiche principali. Nel processo di categorizzazione, le cosiddette etichette sono in realtà etichette di classe. Le metodologie di apprendimento supervisionato utilizzano un'ampia gamma di algoritmi<sup>26</sup>.



**Fig.4.** *Apprendimento supervisionato*

<sup>23</sup> (Clayton, Cos'è il machine learning?, 2021)

<sup>24</sup> (Dotti, Supervised learning, cos'è, esempi di apprendimento supervisionato, 2022)

<sup>25</sup> (Cos'è il Machine learning, come funziona l'apprendimento automatico e quali sono le sue applicazioni, 2022)

<sup>26</sup> (Nasteski, An overview of the supervised machine learning methods, 2017)

L'apprendimento supervisionato prevede una serie di fasi. Per prima cosa viene prodotto il set di dati di addestramento e ogni input viene abbinato a un output sotto la supervisione umana. Il set di dati viene quindi diviso in due parti, una di addestramento e una di test. L'algoritmo viene prima addestrato sul set di dati di addestramento, dove scopre le connessioni tra i dati di input e di output. Grazie a queste relazioni, è possibile fare previsioni utilizzando input diversi da quelli usati per l'addestramento. Il modello viene testato sul set di dati di prova durante la fase di test, che è la fase finale. Confrontando quest'ultimo output con quello precedente e calcolando le metriche, si valuta la qualità dell'output. L'errore deve essere sufficientemente piccolo per essere favorevole. Le classificazioni negli algoritmi supervisionati sono preimpostate. Poiché sono costruite in una modalità limitata, definita dall'uomo, queste classi saranno, di fatto, applicate a un sottoinsieme specifico dei dati. Per creare modelli matematici, l'algoritmo di apprendimento automatico deve identificare i modelli. Poi, il potere predittivo di questi modelli viene misurato rispetto agli indicatori di varianza dei dati<sup>27</sup>.

I metodi di regressione, il cui output è un valore numerico continuo, e gli algoritmi di classificazione, il cui output è un'etichetta, una variabile binaria, sono i due principali sottocampi dell'apprendimento supervisionato. Il tipo di output e i criteri utilizzati per giudicare la qualità del modello sono diversi in queste due macroaree<sup>28</sup>. I modelli di regressione traducono lo spazio di input in un dominio a valori reali. Lo spazio di input viene mappato su classificazioni predeterminate mediante classificatori. Macchine vettoriali di supporto, alberi decisionali, sommatorie probabilistiche, funzioni algebriche, ecc. sono solo alcune opzioni per la modellazione dei classificatori. La classificazione è uno dei modelli più studiati, insieme alla regressione e alla stima delle probabilità, e potrebbe anche avere la maggiore applicazione pratica. Poiché la categorizzazione ha un impatto così importante su altre aree del data mining e sulle sue applicazioni, i potenziali benefici dei progressi in questo campo sono enormi<sup>29</sup>.

### 1.2.2 UNSUPERVISED LEARNING

L'apprendimento non supervisionato, a differenza del precedente, non prevede che le informazioni inserite abbiano delle etichette. È quindi la macchina che va a codificare le

---

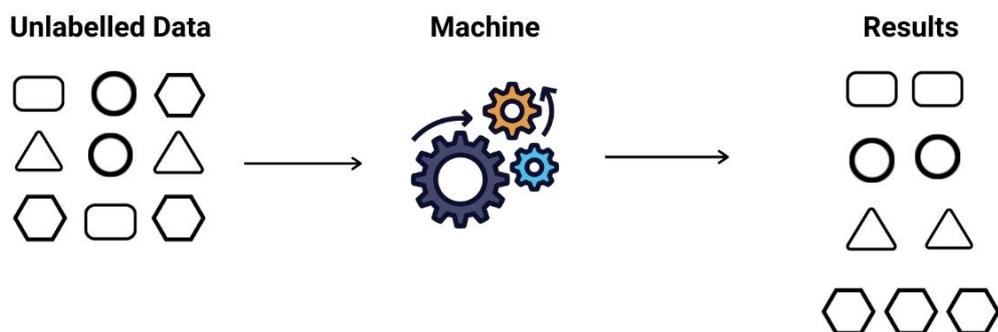
<sup>27</sup> (Nasteski, An overview of the supervised learning methods, 2017)

<sup>28</sup> (Dotti, Supervised learning, cos'è, esempi di apprendimento supervisionato, 2022)

<sup>29</sup> (Nasteski, An overview of the supervised machine learning methods, 2017)

informazioni per fornire l'output<sup>30</sup>. È compito dell'apprendimento automatico quello di dedurre una funzione per rappresentare una struttura nascosta dei dati non etichettati. Poiché gli esempi non sono etichettati, non vi è alcuna valutazione dell'accuratezza della struttura prodotta dell'algoritmo pertinente, il che è un modo per distinguere l'apprendimento non supervisionato dall'apprendimento supervisionato e dall'apprendimento con rinforzo<sup>31</sup>.

In questo tipo di apprendimento si utilizzano solo algoritmi per identificare le relazioni tra i dati. Poiché la parte di addestramento non viene fatta, questo metodo ha il vantaggio di richiedere meno tempo e impegno. Quando si lavora con grandi quantità di dati, per separarli in sottoinsiemi con proprietà simili o per identificare pezzi non necessari, questa tecnica di apprendimento viene spesso utilizzata. Può essere adoperata per cercare modelli nascosti nei dati che sfuggono all'osservazione. Esistono diversi tipi di apprendimento non supervisionato. Il primo tipo sono gli algoritmi di clustering, in cui si raggruppano i dati in cluster internamente omogenei per scoprire la struttura dei dati non categorizzati. Un'altra categoria è rappresentata dalla riduzione della dimensionalità. Queste tecniche consentono di estrarre le caratteristiche più importanti dai dati di partenza o di combinarle con essi per creare componenti indipendenti in presenza di una grande quantità di dati. Questa strategia ha lo svantaggio di produrre risultati difficili da interpretare<sup>32</sup>.



**Fig.5.** *Apprendimento non supervisionato*

<sup>30</sup> (Machine Learning - I diversi apprendimenti di una macchina)

<sup>31</sup> (Sharma & Kumar, A Review on Machine Learning Algorithms, Tasks and Applications, 2017)

<sup>32</sup> (Dotti, UnSupervised learning, cos'è ed esempi di apprendimento non supervisionato, 2022)

I sistemi di raccomandazione sono un esempio di apprendimento non supervisionato. Quando ai consumatori vengono mostrati prodotti simili a quelli acquistati o visti da altri utenti con gusti simili viene rilevata la similarità. Per esempio, si consideri Amazon, che suggerisce ai consumatori i prodotti che altri utenti con gusti simili hanno già acquistato. Anche Netflix si avvale di questa tecnologia per consigliare film basati su ciò che hanno guardato altri utenti che condividono gli stessi gusti. Le aziende utilizzano spesso l'algoritmo di clustering per suddividere la propria clientela in gruppi omogenei, in modo da poterli indirizzare con offerte e promozioni a cui potrebbero essere interessati<sup>33</sup>.

### 1.2.3 REINFORCEMENT LEARNING

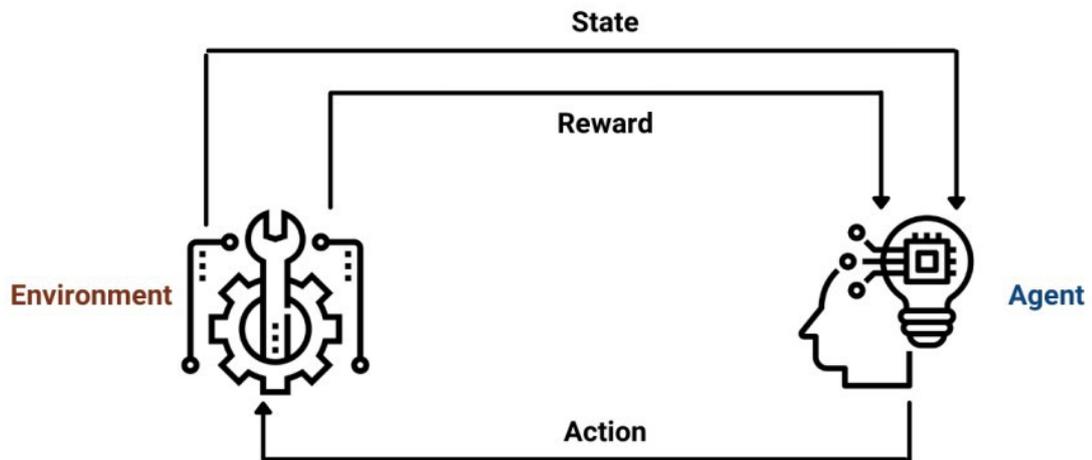
La tecnica di apprendimento automatico più complessa è l'apprendimento per rinforzo. In questo caso, la macchina dispone di strumenti che le consentono di comprendere le proprietà dell'ambiente circostante. La macchina si muove nello spazio prendendo decisioni che le consentono di integrarsi al meglio con l'ambiente circostante. Le auto senza pilota, che percorrono la strada riconoscendo e cercando di evitare gli ostacoli, sono un esempio di questa forma di apprendimento<sup>34</sup>. In questo tipo di apprendimento, la macchina acquisisce nuove informazioni interagendo con l'ambiente circostante e commettendo errori. Per ottenere le ricompense, la macchina deve capire quali comportamenti adottare nell'ambiente circostante. Nel caso in cui la macchina commetta degli errori, viene penalizzata e il suo obiettivo è quello di correggersi. Per rendere possibile questo processo è necessario un rinforzo, ovvero un segnale di ricompensa monetaria che viene attivato quando l'obiettivo viene raggiunto. Senza dare all'agente alcuna istruzione specifica, il rinforzo viene inserito nel processo di addestramento. Attraverso innumerevoli interazioni con l'ambiente circostante, l'agente dovrà dedurlo. L'apprendimento per rinforzo comporta una serie di fasi: in primo luogo, la definizione e la simulazione dell'ambiente operativo dell'agente; in secondo luogo, la definizione del segnale di ricompensa e il suo calcolo; in terzo luogo, la definizione della policy, quindi dell'algoritmo di addestramento; in quarto luogo, l'addestramento; infine,

---

<sup>33</sup> (Dotti, UnSupervised learning, cosè ed esempi di apprendimento non supervisionato, 2022)

<sup>34</sup> (Machine Learning - I diversi apprendimenti di una macchina)

l'implementazione della policy determinata. Nel lungo periodo, l'agente ha l'obiettivo di massimizzare la ricompensa<sup>35</sup>.



**Fig.6.** *Apprendimento con rinforzo*

AlphaGo, un programma per il gioco del Go sviluppato da DeepMind, è una delle applicazioni più note dell'apprendimento per rinforzo. Nel 2016 il coreano Lee Sedol, uno dei migliori giocatori di Go al mondo, è stato sconfitto da questo programma. AWS DeepRacer, un'auto da corsa automatizzata creata da Amazon, è uno dei numerosi sistemi di guida autonoma che troviamo<sup>36</sup>.

### 1.3 CLUSTERING

Nel precedente paragrafo sono state illustrate le varie tipologie di algoritmi di Machine Learning. In questo si farà un focus sul clustering, in quanto è l'algoritmo di apprendimento non supervisionato che verrà utilizzato per l'analisi sperimentale.

---

<sup>35</sup> (Condemi, Reinforcement learning: cos'è, come funziona ed esempi dell'apprendimento per rinforzo, 2021)

<sup>36</sup> (Condemi, Reinforcement learning: cos'è, come funziona ed esempi dell'apprendimento per rinforzo, 2021)

Il clustering è una tecnica di apprendimento non supervisionato che ha lo scopo di raggruppare i dati in cluster che siano omogenei internamente ed eterogenei esternamente. Sebbene non tutti i dati siano identici, essi condividono una serie di caratteristiche. Le tecniche di clustering sono utilizzate nell'apprendimento automatico per identificare modelli nei dati<sup>37</sup>. Un algoritmo di clustering riceve elementi in ingresso e produce in uscita descrizioni dei cluster in cui sono raggruppati gli elementi e i cluster stessi<sup>38</sup>.

### 1.3.1 OVERVIEW

Il clustering è una tecnica di *descriptive analytics* che aiuta ad avere una maggiore comprensione dei propri dati per compiere *data driven decisions*<sup>39</sup>. Nel marketing, la cluster analysis può essere utilizzata per diversi scopi, come le ricerche di mercato, l'individuazione di modelli, la segmentazione dei consumatori in base ai modelli di acquisto, il posizionamento dei prodotti e l'individuazione degli outlier. Conoscere il proprio pubblico è fondamentale per qualsiasi imprenditore che intenda avviare la propria azienda. Ne consegue che capire cosa rende unici i vari gruppi di buyer personas diventa uno strumento cruciale in una strategia di marketing digitale, necessaria nelle aziende di oggi<sup>40</sup>. Le tecniche di clustering possono essere suddivise in due gruppi: gli algoritmi di clustering non gerarchici (noti anche come partizionali) e gli algoritmi di clustering gerarchici. Esistono diversi modi per calcolare la somiglianza tra gli elementi per identificare l'appartenenza ai diversi cluster. Il metodo della distanza euclidea, la cui formula è la seguente, è uno dei più semplici<sup>41</sup>:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^n (x_{i k} - x_{j k})^2}$$

Inoltre, non è obbligatorio che un'osservazione appartenga a un solo cluster. Infatti, si possono distinguere tre tipi di clustering: il clustering esclusivo, che ammette un solo elemento per cluster, il clustering non esclusivo, che permette a un elemento di

---

<sup>37</sup> (La tecnica del clustering)

<sup>38</sup> (Tecniche di clustering, 2015)

<sup>39</sup> (Provino, Clustering: tutto quello che devi sapere | Edizione 2021, 2020)

<sup>40</sup> (Provino, Clustering: tutto quello che devi sapere | Edizione 2021, 2020)

<sup>41</sup> (Tecniche di clustering, 2015)

appartenere a più di un cluster, e il fuzzy clustering, in cui ogni elemento è presente in ogni cluster<sup>42</sup>. Un algoritmo di clustering si svolge in cinque fasi. La prima fase è la selezione delle caratteristiche; le caratteristiche devono essere scelte con cura in modo da fornire il maggior numero possibile di informazioni sull'assegnazione. L'obiettivo fondamentale è diminuire le informazioni e ridurre la ridondanza. In secondo luogo, è necessario quantificare quanto le caratteristiche di due vettori siano simili o meno. Dopo aver misurato la somiglianza, si sceglie un determinato algoritmo di clustering per mostrare la suddivisione dei dati. Per giungere a conclusioni appropriate, è necessario convalidare i risultati acquisiti con l'algoritmo di clustering e giudicarli<sup>43</sup>.

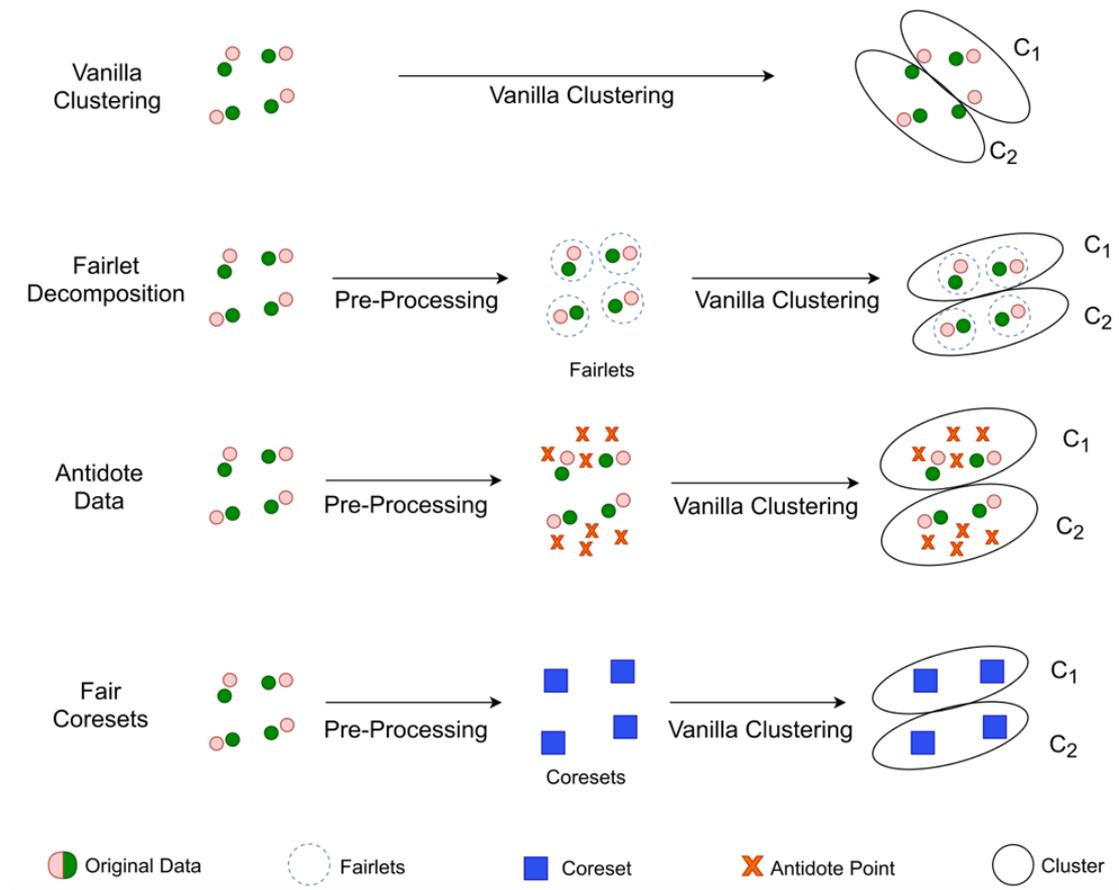
Da una ricerca<sup>44</sup> emerge che un tema molto attuale riguardante il clustering e il Machine Learning in generale è la ricerca e la proposta di approcci che garantiscano il rispetto dei vincoli di equità. Questo tema è particolarmente complicato nel caso specifico del clustering, poiché non ci sono etichette nei dati e gli insiemi di addestramento e di test sono identici. Per questo motivo, esistono numerose teorie distinte sull'equità del clustering, con vari studi che scelgono diverse metriche o ne propongono di nuove. Il metodo della decomposizione *fairlet* dei cluster viene proposto nel clustering non gerarchico per bilanciare i cluster in uscita. La decomposizione in *fairlet* è un approccio di preelaborazione che calcola micro-cluster in cui l'equità è garantita. La ricerca utilizza quindi il cluster *fairlet* come un nuovo set di dati modificato rispetto all'originale. Dato che i *fairlet* stessi sono equi, si ottengono risultati di clustering approssimativamente equi. In seguito, sono stati creati i *coreset*, un'evoluzione dei *fairlet* che utilizza una metodologia più pratica. Un *coreset* è una sintesi di un dato insieme di punti che approssima efficacemente la funzione di costo per ogni potenziale soluzione candidata.

---

<sup>42</sup> (Provino, Clustering: tutto quello che devi sapere | Edizione 2021, 2020)

<sup>43</sup> (Zou, 2020)

<sup>44</sup> (Chhabra, Masalkovaite, & Mohapatra, 2021)



**Fig.7.** *Approcci di preelaborazione per garantire l'equità dei cluster*

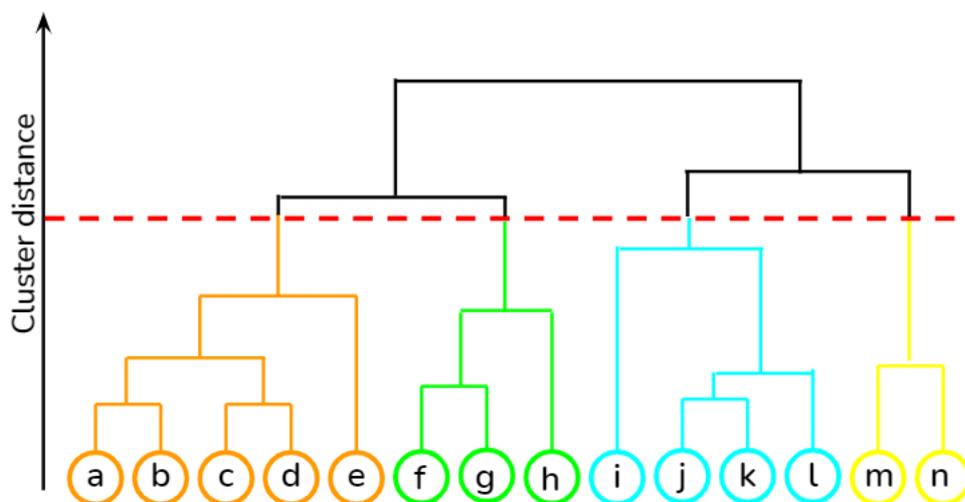
### 1.3.2 CLUSTERING GERARCHICO

Un processo chiamato clustering gerarchico divide gli oggetti in cluster in base alla loro somiglianza. Il prodotto finale è una collezione di cluster, dove ogni cluster è identificato dagli altri in base agli oggetti che sono approssimativamente associati ad esso<sup>45</sup>. In altre parole, un insieme di dati viene ripetutamente suddiviso in cluster sempre più piccoli. Come input viene utilizzato un grafico ponderato con i pesi dei bordi che rappresentano le somiglianze o le differenze a coppie tra i punti di dati. Un clustering gerarchico è rappresentato da un albero radicato in cui ogni foglia rappresenta un punto dati e ogni nodo interno rappresenta un cluster che contiene le sue foglie discendenti. Un tema fondamentale nell'analisi dei dati è che il clustering gerarchico è spesso utilizzato per

<sup>45</sup> (What is Hierarchical Clustering?)

classificare, preelaborare e analizzare enormi insiemi di dati. Grazie al clustering gerarchico, è possibile individuare gruppi nelle reti sociali a diversi livelli socioeconomici, determinare l'albero evolutivo della vita o dividere un'immagine digitale in regioni distinte di diversa granularità. In diverse aree di ricerca, tra cui l'apprendimento automatico, l'analisi dei Big Data e la bioinformatica, la creazione di algoritmi affidabili ed efficaci per il calcolo dei gruppi gerarchici è fondamentale<sup>46</sup>.

Per costruire schemi di classificazione, il clustering gerarchico agglomerativo è stato il metodo preferito<sup>47</sup>. Esso impiega la seguente procedura per individuare i cluster: inizialmente ogni elemento del set di dati forma un cluster separato. Poi, a ogni iterazione, vengono accoppiati i cluster più simili. L'iterazione termina quando tutti gli oggetti sono raggruppati in un unico cluster in cui tutti i componenti sono simili. Un grafico chiamato dendrogramma serve a rappresentare le numerose iterazioni.



**Fig.8.** *Rappresentazione del dendrogramma*

<sup>46</sup> (Cohen-Addad, Kanade, Mallmann-Trenn, & Mathieu, Hierarchical Clustering: Objective Functions and Algorithms, 2019)

<sup>47</sup> (Murtagh & Contreras, Algorithms for hierarchical clustering: an overview, II, 2017)

L'utente definisce quindi un punto di taglio sull'asse y del dendrogramma, che stabilisce il numero di cluster<sup>48</sup>. Da un punto di vista teorico, il clustering gerarchico ha ricevuto molta meno attenzione rispetto al clustering basato su partizioni. Un obiettivo ben definito viene spesso minimizzato nel contesto del clustering basato su partizioni e le strategie per farlo sono state al centro della ricerca per almeno tre decenni. Tuttavia, in termini di algoritmi impiegati nella pratica, il clustering gerarchico è stato esaminato a un livello più procedurale. Tali algoritmi possono essere ampiamente classificati in due categorie: euristiche agglomerative che costruiscono l'albero dei cluster dal basso verso l'alto, ad esempio *average-linkage*, *single-linkage* e *complete-linkage*, ed euristiche divisive che costruiscono l'albero dall'alto verso il basso, ad esempio *bisection k-means*, *recursive sparsest cut* e così via. Uno dei motivi per cui lo studio teorico del clustering gerarchico è rimasto indietro rispetto a quello del clustering basato su partizioni può essere l'assenza di una funzione obiettivo chiaramente definita<sup>49</sup>.

L'analisi dei dati, interfacce utente interattive, archiviazione e recupero e riconoscimento di modelli sono alcuni dei fattori che hanno spinto l'adozione del clustering gerarchico. Con radici negli anni '60 e '70, le tecniche di clustering gerarchico vengono costantemente ampliate da nuove sfide. Si tratta di una classe di algoritmi che sono essenziali per risolvere numerosi problemi significativi. Gli algoritmi di clustering gerarchico continueranno a essere importanti per molto tempo, data la loro ampia diffusione in diversi settori applicativi<sup>50</sup>.

### 1.3.3 CLUSTERING NON GERARCHICO (K-MEANS)

Il clustering non gerarchico, o partizionale, è l'altro tipo di clustering. Si decide innanzitutto il numero di cluster che si desidera avere in questa forma di raggruppamento (k). Poi, assegnando elementi a caso a ciascun cluster, si dividono gli elementi in k cluster. Successivamente, si determinano i centroidi di ciascun cluster e si utilizza l'errore quadratico per determinare la distanza degli elementi dal centroide. L'errore totale, che è il valore da ridurre, viene calcolato sommando gli errori di ciascun cluster. Una volta minimizzato l'errore o stabilizzati gli elementi, questi vengono ridistribuiti in base al

---

<sup>48</sup> (Rezzani A. , Tecniche di clustering, 2015)

<sup>49</sup> (Cohen-Addad, Kanade, Mallmann-Trenn, & Mathieu, Hierarchical Clustering: Objective Functions and Algorithms, 2019)

<sup>50</sup> (Murtagh & Contreras, Algorithms for hierarchical clustering: an overview, II, 2017)

centroide più vicino. K-means è un esempio di clustering non gerarchico<sup>51</sup>. Come si stabilisce il numero di cluster a priori? È fondamentale comprendere il dominio e il mercato di riferimento prima di determinare il numero ideale di cluster utilizzando la conoscenza dei dati già in possesso. È importante che il k-means separi il set di dati in k cluster discreti e non sovrapposti. Ciò indica che i componenti di un cluster sono distinti da quelli presenti in altri cluster e comparabili tra loro<sup>52</sup>. Esistono due metodi per scegliere i centroidi iniziali: il primo in cui si sceglie un campione casuale di osservazioni come punto di partenza; il secondo segue i passi del primo metodo, ma sostituisce elementi casuali da un set di dati diverso. Tuttavia, la convergenza non è certa al termine dei cicli multipli; è probabile che, pur riducendo l'errore, si rimanga su un minimo locale anziché arrivare al minimo globale. I centroidi selezionati hanno un impatto sulle prestazioni dell'algoritmo. Pertanto, può essere utile testare l'algoritmo con diverse condizioni di partenza per vedere quale produce l'errore più basso<sup>53</sup>.

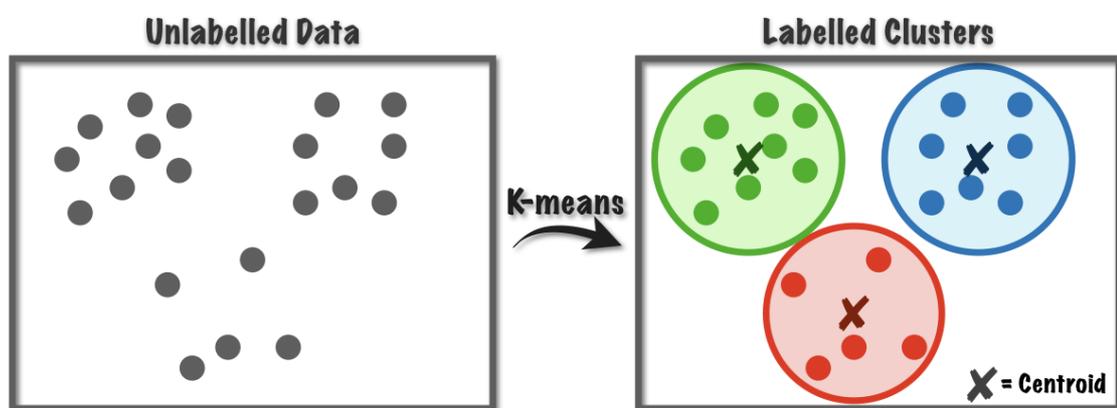


Fig.9. Clustering k-means

Non esiste una tecnica di clustering superiore alle altre; per selezionare quella più adatta alle vostre esigenze, è fondamentale comprendere il vostro mercato di riferimento, la vostra organizzazione, il vostro obiettivo e il set di dati che avete a disposizione. Queste

<sup>51</sup> (Rezzani A. , Tecniche di clustering, 2015)

<sup>52</sup> (Provino, K-means Clustering, 2019)

<sup>53</sup> (Provino, K-means Clustering, 2019)

informazioni consentono di selezionare l'algoritmo migliore per il lavoro<sup>54</sup>. Per valutare i risultati del clustering si possono utilizzare gli indici di validità interna e di validità esterna SSE (Sum of Squared Error).

---

<sup>54</sup> (Provino, Clustering: tutto quello che devi sapere | Edizione 2021, 2020)

## CAPITOLO 2. STATO DELL'ARTE

Il presente capitolo fornisce una revisione della letteratura relativa agli attributi del caffè che i consumatori ritengono più importanti e alle ragioni principali per cui lo acquistano e lo consumano. Successivamente si riporta un'analisi delle applicazioni di metodi di Machine Learning nel mondo del caffè e delle bevande calde in generale.

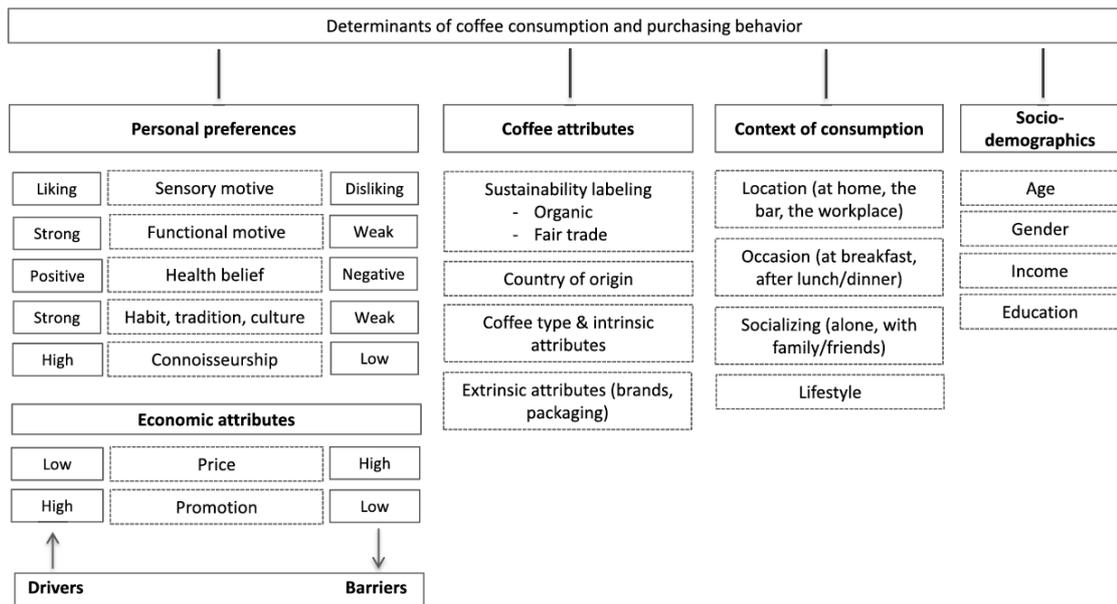
### 2.1. MOTIVI PER L'ACQUISTO E IL CONSUMO DI CAFFÈ

Il caffè è una delle bevande più consumate al mondo e assume un ruolo rilevante nella cultura del consumatore da moltissimi anni. Bere un caffè non significa solo consumare una bevanda. Si tratta di piacere, esperienza, stile di vita e stato sociale. Come vediamo nella vita di tutti i giorni, il caffè ha assunto un ruolo estremamente significativo. Inoltre, ai giorni d'oggi il mercato del caffè è sempre più competitivo. Infatti, cresce continuamente il numero di brand che cercano di conquistare, e in seguito fidelizzare, il maggior numero di consumatori possibile. È infatti da qui che nasce l'esigenza delle aziende di produrre un caffè sempre più vicino ai bisogni dei consumatori.

Ma quali sono i fattori che determinano il gradimento di un cliente per la sua tazza di caffè? Secondo le ricerche,<sup>55</sup> uno dei fattori più significativi che influenzano il consumo di caffè e l'acquisto del prodotto sono le sue caratteristiche. Le proprietà sensoriali del prodotto, come il gusto, l'aroma e il profumo, sono alcune delle considerazioni più importanti per i clienti.

---

<sup>55</sup> (Samoggia & Riedel, Coffee consumption and purchasing behavior review: Insights for further research, 2018)



**Fig.10.** Fattori che influenzano il comportamento del consumatore verso il caffè

Queste caratteristiche sono essenziali per bere la bevanda. Come tutte le bevande contenenti caffeina, il sapore del caffè è molto importante per il cliente, che lo associa spesso a emozioni piacevoli. In particolare, le persone bevono spesso caffè quando socializzano con gli amici o la famiglia o si rilassano. Il gusto e l'odore hanno un'influenza significativa nel rendere l'esperienza del consumatore ancora più piacevole anche in questa situazione. Ci sono numerosi motivi per gustare una tazza di caffè. In realtà, oltre a essere una fonte di intrattenimento, viene spesso consumato durante le ore di lavoro o quando si va incontro alle difficoltà che ognuno di noi affronta quotidianamente.

Il paese di origine del caffè è un altro elemento che la maggior parte dei clienti considera, secondo un altro studio<sup>56</sup>. Questa caratteristica può essere utilizzata per distinguere le torrefazioni rispetto alle preferenze dei clienti. Diverse aziende dell'industria del caffè sfruttano questa qualità. Starbucks è un esempio<sup>57</sup> in cui il fattore origine assume una rilevanza significativa, perché l'azienda pone una forte enfasi sulle origini del prodotto

<sup>56</sup> (Samoggia & Riedel, Consumers' Perceptions of Coffee Health Benefits and Motives for Coffee Consumption and Purchasing, 2019)

<sup>57</sup> (Bookman, 2013)

come differenziazione del caffè. Infatti, l'idea è quella di creare il profilo perfetto del consumatore in base ai suoi gusti, trovare il caffè perfetto per lui confrontando le sue preferenze con il caffè della provenienza che gli si avvicina di più per poi fornirgli l'esperienza migliore durante il consumo della bevanda.

La salubrità del prodotto da acquistare e consumare è un altro elemento cruciale per i consumatori. Da una ricerca<sup>58</sup> emerge che i consumatori stanno diventando più esigenti e consapevoli. Grazie alla digitalizzazione, le persone hanno accesso a un maggior numero di informazioni sui prodotti, compresa la possibilità di confrontarli tra loro e di leggere le recensioni di altri clienti. Grazie all'informazione, le persone diventano più consapevoli di ciò che desiderano e si preoccupano maggiormente delle questioni relative ai pericoli per la salute. Cominciano quindi ad aspettarsi prodotti di qualità sempre più elevata e più sani. Ci sono ancora molte caratteristiche del caffè che le persone non conoscono. Ad esempio, non si conoscono i vantaggi e i benefici che il caffè può apportare alla salute delle persone. Negli ultimi anni, le aziende hanno cercato anche alternative direttamente legate alla salute, nel tentativo di distinguere i propri prodotti. Da una ricerca<sup>59</sup> emerge che uno dei motivi principali per cui le persone non bevono caffè è proprio la sua salubrità. Uno studio<sup>60</sup> dimostra che solo il 25% dei 250 partecipanti all'analisi riteneva che bere caffè avesse effetti benefici sulla salute.

	Negative Perception %	Positive Perception %	Total	ANOVA	p-Value
Total <sup>a</sup>	75.2	24.8	100		
Perception of health effect of coffee (average) <sup>a</sup>	2.29	3.70	2.91	0.000	***
Standard deviations	0.500	0.484	0.762		

**Fig. 11.** *Percezione dei consumatori sugli effetti benefici del caffè*

La maggior parte delle persone non sa che il caffè è un prodotto molto benefico, come provato in una ricerca<sup>61</sup>. Infatti, contiene minerali e antiossidanti salutari che possono

<sup>58</sup> (Petrescu, Vermeir, & Petrescu-Mag, 2019)

<sup>59</sup> (Samoggia, Del Prete, & Argenti, Functional Needs, Emotions, and Perceptions of coffee Consumers and non-Consumers, 2020)

<sup>60</sup> (Samoggia & Riedel, Consumers' Perceptions of Coffee Health Benefits and Motives for Coffee Consumption and Purchasing, 2019)

<sup>61</sup> (Wachamo, 2017)

aiutare le nostre condizioni fisiche. Nello studio sono citati numerosi studi che dimostrano che i consumatori di caffè hanno una probabilità significativamente inferiore di sviluppare alcune malattie importanti. Per esempio, può ridurre il tasso di mortalità per alcuni tipi di cancro.

La capacità del caffè di influenzare le emozioni è un altro aspetto che ne influenza l'assunzione e lo distingue da altre bevande. La decisione di un consumatore di acquistare un prodotto è spesso influenzata dalle sensazioni che il prodotto suscita in lui. Una ricerca<sup>62</sup> dimostra l'efficacia del caffè in questo senso. Infatti, ha il potere di influenzare il benessere delle persone e la loro sicurezza nelle relazioni sociali. Le emozioni e gli stati d'animo delle persone possono variare grazie al caffè. Il caffè è in grado di modificare le emozioni e l'umore delle persone. Oltre al semplice atto di bere la bevanda, che per molte persone è una routine quotidiana, il caffè è anche un'occasione di interazione sociale, può essere consumato in occasioni speciali ed è un alimento considerato nutriente grazie ai benefici per la salute indicati nel paragrafo precedente.

Secondo una ricerca<sup>63</sup>, anche il grado di *familiarity* verso il prodotto ha un ruolo significativo nell'acquisto e nel consumo di caffè. In questo senso, la comunicazione assume una funzione importante e si trasforma in uno strumento cruciale per le aziende che cercano di aumentare la consapevolezza e la preferenza dei consumatori per i loro prodotti. I clienti consumano il caffè per abitudine e per piacere. Questi elementi hanno un impatto sull'aroma, sul sapore e su altre qualità sensoriali della bevanda. Di conseguenza, è fondamentale per i marchi far sapere ai consumatori come i loro prodotti eccellono in relazione a questi attributi. Un'altra ricerca<sup>64</sup> ha esaminato come, dal punto di vista dell'acquirente, conoscere la qualità di un prodotto possa influenzare il desiderio di acquistarlo. Nello studio sono stati utilizzati due metodi per valutare il grado di accettazione del caffè da parte dei consumatori: un blind test e attraverso la fornitura di informazioni sulla qualità del prodotto. La conoscenza delle proprietà sensoriali del prodotto ha sicuramente un impatto sul comportamento del consumatore. Infatti, un consumatore valuta il prodotto con maggiore attenzione quando gli vengono fornite

---

<sup>62</sup> (Samoggia, Del Prete, & Argenti, Functional Needs, Emotions, and Perceptions of Coffee Consumers and Non-Consumers, 2020)

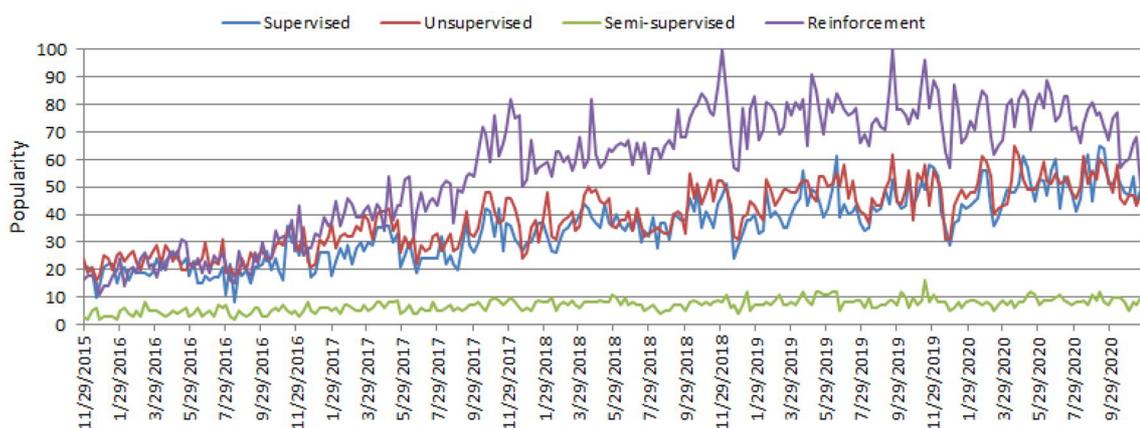
<sup>63</sup> (Samoggia & Riedel, Consumers' Perceptions of Coffee Health Benefits and Motives for Coffee Consumption and Purchasing, 2019)

<sup>64</sup> (Bemfeito, et al., 2021)

informazioni sulla qualità del caffè. I consumatori, invece, non valutano la distinzione tra un caffè di qualità eccellente e uno di qualità inferiore quando non ricevono informazioni sulla qualità del prodotto. Questi risultati sottolineano ancora una volta l'importanza del fatto che le aziende forniscano ai consumatori questo tipo di informazioni nell'ambito dei loro piani di marketing, al fine di incrementare le vendite dei prodotti e di fornire loro le migliori esperienze di consumo possibili.

## 2.2. APPLICAZIONE DI METODI DI MACHINE LEARNING PER PREDIRE E TESTARE LA QUALITÀ DEI PRODOTTI

Data l'importanza degli attributi menzionati sopra per il consumatore nell'ambito delle bevande calde e la continua competitività delle aziende nel voler andare incontro ai bisogni del consumatore, nasce la necessità di valutare la qualità del caffè, e delle bevande calde in generale, in un modo sempre più preciso ed accurato. Sebbene i metodi utilizzati in precedenza fossero abbastanza efficaci, presentavano spesso degli svantaggi, tra cui la soggettività, i prezzi più elevati e l'imprecisione. A causa di questi inconvenienti, numerose aziende si sono impegnate nella ricerca di approcci di apprendimento automatico nel tentativo di creare procedure più obiettive, accurate ed efficienti in termini di tempo. L'analisi dei dati ha visto un rapidissimo aumento dell'uso dell'intelligenza artificiale e dell'apprendimento automatico in molti contesti diversi.<sup>65</sup>



**Fig. 12.** Crescita nel mondo del tasso di popolarità dei vari algoritmi di Machine Learning

<sup>65</sup> (Sarker, Machine learning: Algorithms, Real-World Applications and Research Directions, 2021)

Poiché l'apprendimento automatico ha dato un contributo così significativo alla produzione e alla valutazione della qualità dei prodotti nel settore delle bevande, negli ultimi anni ha registrato una crescita e un'innovazione significative<sup>66</sup>. In particolare, la tecnologia ha permesso di ottenere risultati più rapidi, corretti e oggettivi. Grazie alla capacità di prevedere i desideri dei clienti e gli attributi di qualità dei prodotti, utili per la creazione di articoli di qualità superiore e di conseguenza scelti dai consumatori, il machine learning ha svolto un ruolo importante in questo senso. I metodi precedenti avevano più limiti, perché richiedevano molto più tempo e non potevano produrre risultati immediati. Le tecniche più arcaiche impiegavano macchinari molto costosi o valutavano le qualità delle bevande utilizzando i sensi umani che fornivano conclusioni soggettive. Questo ha portato alla necessità di automatizzare alcune operazioni. Uno studio<sup>67</sup> ha dimostrato che caratteristiche come il sapore e il profumo sono spesso utili indicatori di qualità. L'intelligenza artificiale è stata utilizzata per costruire alcune nuove tecniche di valutazione delle caratteristiche di qualità delle bevande. Queste tecniche consentono ai ricercatori di ottenere dati in modo molto più rapido, migliorando al contempo l'obiettività, l'accuratezza e l'affidabilità. Si tratta di metodi che utilizzano varie fonti di dati, algoritmi informatici e modelli di apprendimento automatico per prevedere la qualità dei prodotti finiti in base a vari fattori misurabili. Un esempio di automatizzazione di questi processi è dato da uno studio<sup>68</sup> in cui i ricercatori hanno cercato di valutare con precisione l'intensità e gli aromi del caffè utilizzando un naso elettronico portatile e a basso costo insieme a modelli di apprendimento automatico (Machine Learning). Nel campo del caffè, questi modelli si sono rivelati molto utili sia per i ricercatori che per le aziende e i clienti. Possono essere incorporati in modo specifico nelle macchine da caffè per valutare eventuali alterazioni della fragranza causate da influenze esterne che alterano il processo di infusione. Incorporando questa tecnica nelle macchine commerciali, è anche possibile garantire la qualità e la freschezza dei chicchi di caffè e regolare la qualità nel punto vendita, in modo che i clienti possano individuare gli odori sgradevoli.

---

<sup>66</sup> (Gonzalez Viejo, Torrico, Dunshea, & Fuentes, Emerging Technologies Based on Artificial Intelligence to Assess the Quality and Consumer Preference of Beverages, 2019)

<sup>67</sup> (Gonzalez Viejo, Torrico, Dunshea, & Fuentes, Emerging Technologies Based on Artificial Intelligence to Assess the Quality and Consumer Preference of Beverages, 2019)

<sup>68</sup> (Gonzalez Viejo, Tongson, & Fuentes, Integrating a Low-Cost Electronic Nose and Machine Learning Modelling to Assess Coffee Aroma Profile and Intensity, 2021)

Anche la previsione della qualità di vari tipi di bevande calde si è rivelata un uso molto efficace dell'apprendimento automatico. Infatti, uno studio lo ha utilizzato per creare un metodo semplice ed efficace per valutare il livello di fermentazione del tè nero utilizzando le caratteristiche elettriche delle foglie di tè.<sup>69</sup> In realtà, la fermentazione viene utilizzata per valutare la qualità del prodotto. Ancora una volta, i risultati imparziali erano impossibili con approcci rudimentali. In realtà, il colore e il profumo del tè venivano modificati per indicare il livello di fermentazione. Essendo largamente influenzati dall'esperienza e dall'ambiente, ciò aumentava la possibilità di ottenere conclusioni soggettive. Lo studio suggerisce di utilizzare il clustering gerarchico per dividere i campioni in base al grado di fermentazione e Random Forest e Support Vector Machine per creare modelli discriminanti per il grado di fermentazione, al fine di automatizzare questo processo e renderlo più obiettivo e accurato. Si è scoperto che queste tecniche hanno un tasso di accuratezza incredibilmente elevato. Un'altra ricerca<sup>70</sup> ha dimostrato l'uso di immagini computerizzate per valutare i cambiamenti degli indicatori di qualità e dei pigmenti durante la fermentazione del tè. Studiando la correlazione tra le caratteristiche cromatiche delle foto e gli indici di qualità, questo studio ha estratto nove fattori di colore e ha dimostrato un rapido approccio quantitativo per valutare la qualità della fermentazione. Di conseguenza, queste due metriche divergono notevolmente in varie fasi della fermentazione, pur essendo significativamente collegate. Essendo più adatti a rappresentare relazioni analitiche, i modelli non lineari Random Forest e Support Vector Machine sono quelli che meglio descrivono il legame tra gli attributi del colore dell'immagine e gli indici di qualità. Con questi metodi è possibile monitorare anche altri aspetti della qualità del tè nero.

L'apprendimento automatico è stato utilizzato anche nel settore delle bevande calde per individuare e quantificare in modo più rapido e preciso l'adulterazione dei prodotti. Una ricerca<sup>71</sup> ha applicato tecniche di categorizzazione di dati complessi, come Random Forest e Reti neurali artificiali. In realtà, queste tecniche sono efficaci perché riescono a cogliere in modo più preciso e accurato le relazioni non lineari tra le osservazioni. Queste tecniche sono state impiegate nello studio per identificare il tipo di adulterazione del latte.

---

<sup>69</sup> (Zhu, et al., Application of machine learning algorithms in quality assurance of fermentation process of black tea-based on electrical properties, 2019)

<sup>70</sup> (Dong, et al., 2018)

<sup>71</sup> (Farah, et al., 2021)

I consumatori attribuiscono grande importanza all'autenticità di alimenti e bevande, che ora può essere studiata in modo più approfondito con l'uso di tecniche di apprendimento automatico. Questi metodi hanno il vantaggio di essere adattabili a diversi formati di dati e di essere resistenti alla presenza di outlier e dati mancanti.

Machine learning	Control	Starch	Formaldehyde	Whey	Urea	Total
Recognition capacity (%) for training set						
RF	100.0	100.0	100.0	100.0	100.0	100.0
GBM	100.0	100.0	100.0	100.0	100.0	100.0
MLP	100.0	100.0	100.0	100.0	100.0	100.0
Prediction capacity (%) for test set						
RF	100.0	100.0	100.0	100.0	50.0	88.5
GBM	100.0	100.0	100.0	100.0	100.0	100.0
MLP	100.0	100.0	100.0	100.0	100.0	100.0

**Fig. 13.** Capacità di predizione dei metodi di Machine Learning di rilevare l'adulterazione nel latte

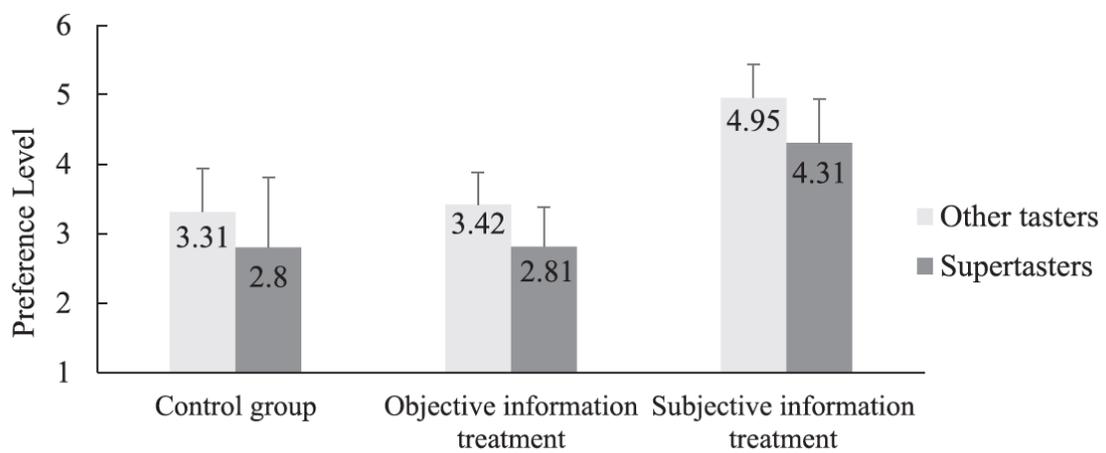
Secondo un altro studio<sup>72</sup>, le tecniche di apprendimento automatico sono necessarie per prevedere la qualità del latte vaccino, poiché riducono gli errori quadratici medi rispetto al più convenzionale approccio dei minimi quadrati parziali. Gli autori suggeriscono in particolare di utilizzare la spettroscopia a medio infrarosso, che si è rivelata un metodo economico, rapido e non distruttivo per determinare la quantità di grasso, proteine e lattosio presenti nei campioni di latte. Questo studio ha apportato un contributo significativo in quanto una migliore previsione della qualità del latte può essere applicata alla produzione di tutti i suoi derivati e del latte stesso, con il risultato di ottenere prodotti finali migliori. Inoltre, è possibile implementare ulteriori miglioramenti dell'accuratezza senza alcun costo aggiuntivo.

Come è evidente, l'apprendimento automatico trova numerosi impieghi nel settore delle bevande calde. Questi studi dimostrano anche che il machine learning è un metodo valido e affidabile per valutare la qualità dei prodotti. Secondo una ricerca<sup>73</sup> per "informazione

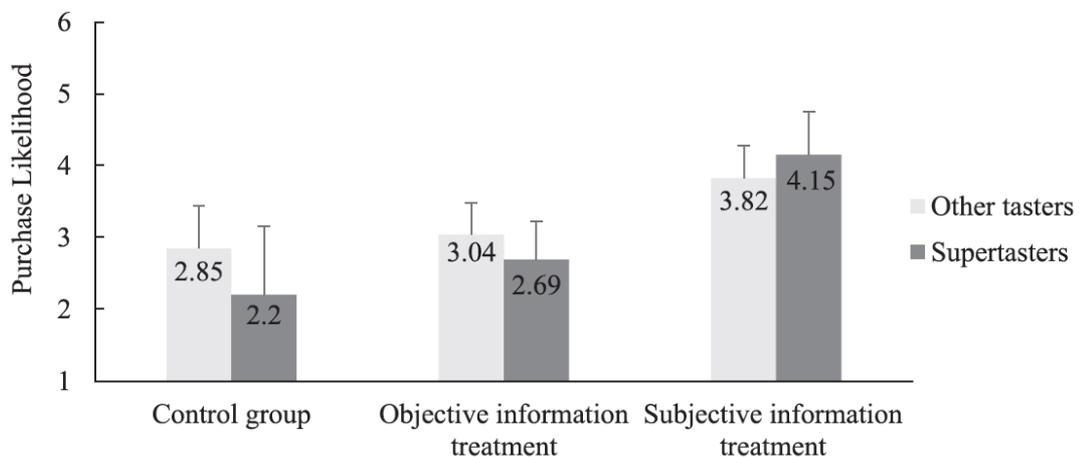
<sup>72</sup> (Frizzarin, et al., 2021)

<sup>73</sup> (Li, Streletskaia, & Gomez, 2019)

soggettiva" si intende il modo in cui i consumatori che non hanno familiarità con termini scientifici o tecnici apprezzano le rappresentazioni metaforiche delle qualità estetiche di un prodotto. Questo concetto include i consigli degli esperti forniti ai clienti con minore esperienza di prodotti. Le "informazioni oggettive", invece, si riferiscono alle caratteristiche di base del prodotto, come il luogo di origine, la varietà, il modo in cui è stato realizzato, ecc. Questo studio dimostra come la percezione soggettiva dei consumatori del sapore e del profumo del caffè influenzi le loro opinioni sul gusto e sulla preferenza del prodotto.



**Fig. 14.** *L'impatto delle informazioni legate agli attributi sensoriali sul livello di preferenza dei consumatori*



**Fig. 15.** *L'impatto delle informazioni legate agli attributi sensoriali sull'intenzione d'acquisto dei consumatori*

Da qui nasce lo scopo di questa ricerca, che ha infatti l'intento di concentrarsi sul caffè e nello specifico andare a clusterizzare varie tipologie di chicchi di caffè sulla base dei rating dati da esperti del Coffee Quality Institute (CQI) riguardo agli attributi ritenuti più importanti dai consumatori. Lo scopo è quello di andare a vedere su quali attributi i chicchi di caffè analizzati sono più forti e quindi su cui i brand possono fare leva nella comunicazione dei loro prodotti.

## CAPITOLO 3. COFFEE QUALITY DATABASE FROM CQI: ANALISI SPERIMENTALE SULLA QUALITÀ DI ALCUNE TIPOLOGIE DI COFFEE BEANS

Nei precedenti capitoli sono state illustrate le caratteristiche del Machine Learning mettendo in luce la rilevanza nell'ambito del marketing. Successivamente sono stati riportati studi riguardo agli attributi che i consumatori ritengono più rilevanti e alcune applicazioni di Machine Learning nell'ambito del caffè e delle bevande calde.

Il presente capitolo fornisce in primo luogo un'overview del linguaggio di programmazione R, successivamente viene descritto il dataset utilizzato per l'analisi sperimentale "*Coffee Quality Database from CQI*". Su questo verrà effettuata un'analisi sperimentale di clustering sulla qualità delle tipologie di chicchi di caffè osservate sulla base degli attributi valutati dagli esperti. Infine, si commenteranno i risultati andando a vedere le caratteristiche di ogni cluster e come i brand potrebbero utilizzare questi risultati a loro vantaggio.

### 3.1 LINGUAGGIO R: OVERVIEW

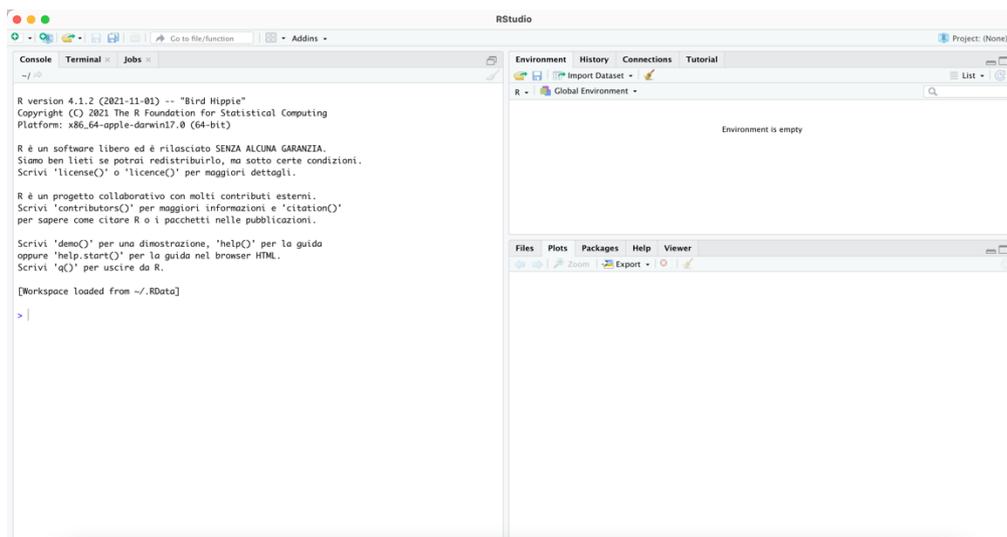
R è un linguaggio di programmazione e un ambiente software open source utilizzato per l'analisi dei dati e la soluzione di problemi matematici e statistici. Ross Ihaka e Robert Gentleman hanno creato R all'inizio degli anni '90 come ispirandosi al già esistente linguaggio di programmazione S. Dieci anni dopo è stata resa disponibile la versione iniziale autorizzata. Il suo status di open source consente a chiunque di scaricarlo, utilizzarlo e modificarlo, rendendo accessibili a tutti le modifiche apportate.<sup>74</sup>

Viene utilizzato soprattutto per la formazione di grafi, per i calcoli statistici, per la gestione e la visualizzazione dei dati. Il software funziona tramite un'interfaccia a riga di comando e consente il collegamento di codice C, C++ e Fortran per attività computazionalmente impegnative. Il codice C può anche modificare direttamente gli oggetti R.<sup>75</sup>

---

<sup>74</sup> (Minini, Cos'è R)

<sup>75</sup> (Cos'è R)



**Fig. 16.** *Interfaccia linguaggio di programmazione R Studio*

Il linguaggio di programmazione R è stato sviluppato per l'utilizzo di calcoli statistici, matematici e vettoriali. Viene utilizzato principalmente nell'apprendimento automatico e nell'analisi dei dati. Per programmare in R è necessario uno script di testo con funzioni; il programma esegue le funzioni e produce i risultati<sup>76</sup>. Le variabili vettori, matrici, liste e dataframe, che riflettono i componenti fondamentali dei dati, costituiscono gli oggetti dello studio. R è incredibilmente flessibile, il che lo rende la scelta migliore quando si devono estrapolare, analizzare e interagire con dati di varie forme e creare vari tipi di grafici<sup>77</sup>.

Altri sistemi, tra cui Microsoft Excel, MySQL, Microsoft Access, Oracle, SAS, SPSS e altri, sono integrati e interoperabili con R. Molte aziende ne hanno riconosciuto il potenziale e lo stanno utilizzando. Ne sono un esempio Bank of America, Ford, Facebook e il New York Times. L'inferenza è la motivazione principale per l'utilizzo di R nel mondo degli affari. Il linguaggio di programmazione R consente in particolare di accertare l'importanza o l'influenza che determinate variabili hanno su un determinato fenomeno. R può determinare quali variabili e quanto contribuiscono alla spiegazione di un fenomeno. La previsione è un altro modo in cui il linguaggio di programmazione può

---

<sup>76</sup> (Minini, Cos'è R)

<sup>77</sup> (La programmazione R: che cos'è, 2019)

essere utile alle organizzazioni. Infatti, consente di stimare un valore in base a una serie di fattori. Ad esempio, si può stimare un prezzo in base alle qualità di un consumatore. È anche possibile combinare questi due metodi, ad esempio stimando un'inferenza prima di fare una previsione<sup>78</sup>.

R è in continua evoluzione e crescita e il gruppo di sviluppatori che lavora al progetto CRAN (Comprehensive R Archive Network) è particolarmente attivo nella produzione di software e pacchetti liberi.<sup>79</sup>.

### 3.2 IL DATASET COFFEE QUALITY DATABASE FROM CQI: CARATTERISTICHE

Il dataset Coffee Quality database from CQI contiene le recensioni di 1312 chicchi di caffè arabica e 28 chicchi di caffè robusta da parte di recensori qualificati del Coffee Quality Institute. L'arabica e la robusta sono le tipologie di caffè più utilizzate nelle miscele in commercio. Cosa differenzia queste due tipologie di chicchi di caffè? In primo luogo, la forma del chicco, in quanto quello dell'arabica ha una forma allungata, mentre la robusta ha una forma più arrotondata. Un altro elemento di differenziazione è il luogo d'origine, infatti l'arabica cresce meglio a quote elevate e per questo ha anche un costo più alto, mentre la robusta cresce a quote più basse. Infine, la robusta contiene una quantità maggiore di caffeina rispetto all'arabica. Questi elementi insieme a tanti altri hanno un'influenza importante sul sapore in tazza. Il chicco di caffè può essere valutato tramite diversi sensi. Il primo è la vista, infatti la tonalità di colore che va dal giallo ocre, al nocciola, al marrone sono definite dall'intensità. Un altro senso importante è il gusto, dove si riconosce la dolcezza del caffè sulla punta della lingua, mentre sulla parte inferiore di essa si sente l'amarrezza. L'amarrezza del caffè è data dalla quantità di caffeina, da qui si deduce che la robusta, avendo un quantitativo maggiore di caffeina, è più amara, mentre l'arabica è più acida. Tramite il gusto, un altro attributo che si può valutare è il corpo, definito come la densità, la pesantezza in bocca. Tramite l'olfatto, si può capire il livello di tostatura e il livello di conservazione del chicco. Infine, tra i sensi importanti si trova anche il retrofatto, quindi tutto il bouquet di profumi che possono sentirsi dopo.

---

<sup>78</sup> (R Software, cosa è e a cosa serve questo software statistico)

<sup>79</sup> (La programmazione R: che cos'è, 2019)



**Fig. 17.** *Rappresentazione dell'arabica e della robusta*

Gli attributi considerati nel dataset di riferimento sono i seguenti: l'aroma, il sapore, il retrogusto, l'acidità, la corposità, l'equilibrio, l'uniformità, la pulizia della tazza, la dolcezza, l'umidità, i difetti, il metodo di lavorazione, il colore, la specie (arabica/robusta), il proprietario, il paese d'origine, il nome dell'azienda agricola, il numero di lotto, il mulino, l'azienda, l'altitudine e la regione<sup>80</sup>. Il Coffee Quality Institute (CQI) è un'organizzazione no profit che lavora internazionalmente per migliorare la qualità del caffè e la vita delle persone che lo producono. Il Coffee Quality Institute ha un ruolo importante nella comunità del caffè e per avere il massimo impatto si avvale di un piano strategico per poter servire al meglio la comunità, rimanere concentrati sul futuro e lavorare nel modo più efficiente possibile<sup>81</sup>. Gli attributi che verranno presi in considerazione nell'analisi sono i seguenti:

- Aroma: l'odore del caffè quando viene infuso con acqua calda
- Aftertaste: la lunghezza e la piacevolezza della tazza
- Balance: l'entità della sinergia tra gli aromi presenti

---

<sup>80</sup> (Coffee Quality Database, 2018)

<sup>81</sup> (Coffee Quality Institute)

- Uniformity: indica se una tazza è diversa dalle altre, è importante per garantire la qualità costante del campione
- Cup cleanliness: è penalizzante se si percepisce una nota che ha un effetto negativo tra il primo assaggio e il retrogusto dei successivi assaggi
- Sweetness: un caffè non difettoso, amaro, astringente e sgradevole
- Flavor: il gusto percepito
- Body: indica la corposità, intesa come la percezione tattile e la persistenza sul palato
- Acidity: indica l'acidità del caffè, ma anche la piacevolezza e la delicatezza dell'acidità dopo il primo sorso di caffè in bocca.

È possibile scaricare il dataset al seguente link: <https://www.kaggle.com/datasets/volpato/coffee-quality-database-from-cqi>.

### 3.3 METODOLOGIA

In questo paragrafo viene illustrata la metodologia eseguita su R per l'analisi sperimentale sulla qualità dei chicchi di caffè presenti nel dataset. Verranno applicate entrambe le tipologie di clustering: clustering gerarchico e clustering non gerarchico (k-means).

In primo luogo, si scaricano i seguenti pacchetti:

```
library(dplyr)
library(ggplot2)
library(caret)
library(lattice)
library(factoextra)
library(NbClust)
```

In seguito, si permette a R di leggere il dataset da analizzare e sui cui si andrà a svolgere l'algoritmo di clustering. Si va a vedere, inoltre, la struttura del dataset in questione:

```
coffee <- read.csv("merged_data_cleaned.csv")
str(coffee)
coffee
```

Si assegnano a piacimento i nomi delle variabili che corrispondono alle colonne del dataset in modo che siano possibilmente rappresentativi.

```
colnames(coffee) = c("ID", "Species", "Owner", "CountryOfOrigin",  
"FarmName", "LotNumber", "Mill", "ICONumber", "Company", "Altitude",  
"Region", "Producer", "NumberOfBags", "BagWeight", "InCountryPartner",  
"HarvestYear", "GradingDate", "Owner1", "Variety", "ProcessingMethod",  
"Aroma", "Flavor", "Aftertaste", "Acidity", "Body", "Balance",  
"Uniformity", "CleanCup", "Sweetness", "CupperPoints",  
"TotalCupPoints", "Moisture", "CategoryOneDefects", "Quakers",  
"Color", "CategoryTwoDefects", "Expiration", "CertificationBody",  
"CertificationAddress", "CertificationContact", "UnitOfMeasurement",  
"AltitudeLowMeters", "AltitudeHighMeters", "AltitudeMeanMeters")  
str(coffee)
```

Si rimuovono in seguito le colonne che non interessano per l'analisi. In questo caso, verranno tenute solo le variabili relative agli attributi su cui verrà sviluppato l'algoritmo di clustering:

```
coffee$ID = NULL  
coffee$Owner = NULL  
coffee$FarmName = NULL  
coffee$LotNumber = NULL  
coffee$Mill = NULL  
coffee$ICONumber = NULL  
coffee$Company = NULL  
coffee$Altitude = NULL  
coffee$Region = NULL  
coffee$Producer = NULL  
coffee$NumberOfBags = NULL  
coffee$BagWeight = NULL  
coffee$InCountryPartner = NULL  
coffee$HarvestYear = NULL  
coffee$GradingDate = NULL  
coffee$Owner1 = NULL  
coffee$Variety = NULL  
coffee$ProcessingMethod = NULL  
coffee$Moisture = NULL  
coffee$CategoryOneDefects = NULL  
coffee$CategoryTwoDefects = NULL
```

```

coffee$Quakers = NULL
coffee$Expiration = NULL
coffee$CertificationBody = NULL
coffee$CertificationAddress = NULL
coffee$CertificationContact = NULL
coffee$UnitOfMeasurement = NULL
coffee$AltitudeLowMeters = NULL
coffee$AltitudeHighMeters = NULL
coffee$AltitudeMeanMeters = NULL
coffee$Species = NULL
coffee$CountryOfOrigin = NULL
coffee$Color = NULL
coffee$CupperPoints = NULL
coffee$TotalCupPoints = NULL
str(coffee)

```

```

> str(coffee)
'data.frame': 1339 obs. of 9 variables:
 $ Aroma      : num  8.67 8.75 8.42 8.17 8.25 8.58 8.42 8.25 8.67 8.08 ...
 $ Flavor     : num  8.83 8.67 8.5 8.58 8.5 8.42 8.5 8.33 8.67 8.58 ...
 $ Aftertaste: num  8.67 8.5 8.42 8.42 8.25 8.42 8.33 8.5 8.58 8.5 ...
 $ Acidity    : num  8.75 8.58 8.42 8.42 8.5 8.5 8.5 8.42 8.42 8.5 ...
 $ Body       : num  8.5 8.42 8.33 8.5 8.42 8.25 8.25 8.33 8.33 7.67 ...
 $ Balance    : num  8.42 8.42 8.42 8.25 8.33 8.33 8.25 8.5 8.42 8.42 ...
 $ Uniformity: num  10 10 10 10 10 10 10 10 9.33 10 ...
 $ CleanCup   : num  10 10 10 10 10 10 10 10 10 10 ...
 $ Sweetness  : num  10 10 10 10 10 10 10 9.33 9.33 10 ...

```

**Fig. 18.** *Struttura del dataset in seguito all'eliminazione delle variabili non utili ai fini dell'analisi*

Tra le osservazioni è presente un outlier, rappresentato dall'osservazione 1311, di conseguenza si provvede ad eliminarlo dal dataset e si crea un nuovo dataset di riferimento denominato `coffee_new`. Si richiede in seguito al software la struttura del nuovo dataset senza l'osservazione eliminata:

```

coffee_new <- coffee[-1311,]
str(coffee_new)

```

```

> str(coffee_new)
'data.frame':  1338 obs. of  9 variables:
 $ Aroma      : num  8.67 8.75 8.42 8.17 8.25 8.58 8.42 8.25 8.67 8.08 ...
 $ Flavor     : num  8.83 8.67 8.5 8.58 8.5 8.42 8.5 8.33 8.67 8.58 ...
 $ Aftertaste: num  8.67 8.5 8.42 8.42 8.25 8.42 8.33 8.5 8.58 8.5 ...
 $ Acidity    : num  8.75 8.58 8.42 8.42 8.5 8.5 8.5 8.42 8.42 8.5 ...
 $ Body       : num  8.5 8.42 8.33 8.5 8.42 8.25 8.25 8.33 8.33 7.67 ...
 $ Balance    : num  8.42 8.42 8.42 8.25 8.33 8.33 8.25 8.5 8.42 8.42 ...
 $ Uniformity: num  10 10 10 10 10 10 10 10 9.33 10 ...
 $ CleanCup   : num  10 10 10 10 10 10 10 10 10 10 ...
 $ Sweetness  : num  10 10 10 10 10 10 10 9.33 9.33 10 ...

```

**Fig. 19.** *Struttura del dataset in seguito all'eliminazione dell'outlier*

### 3.3.1 CLUSTERING GERARCHICO

Viene applicato in primo luogo il clustering gerarchico. Come primo step, si calcolano le distanze con il metodo delle distanze euclidee e si applica l'algoritmo di clustering gerarchico con il metodo di Ward.

```

distances = dist(coffee_new[1:9], method = "euclidean")
clusterCoffee = hclust(distances, method = "ward.D")

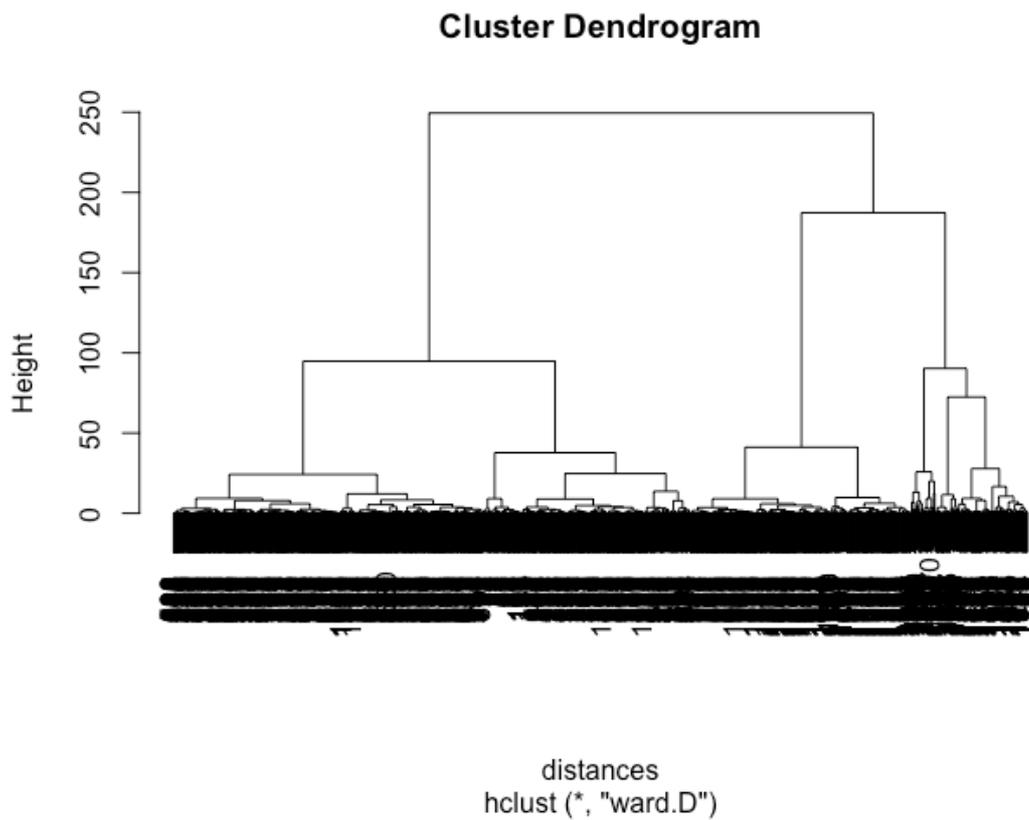
```

In seguito, si richiede al software di eseguire la funzione per la creazione del dendrogramma:

```

plot(clusterCoffee)

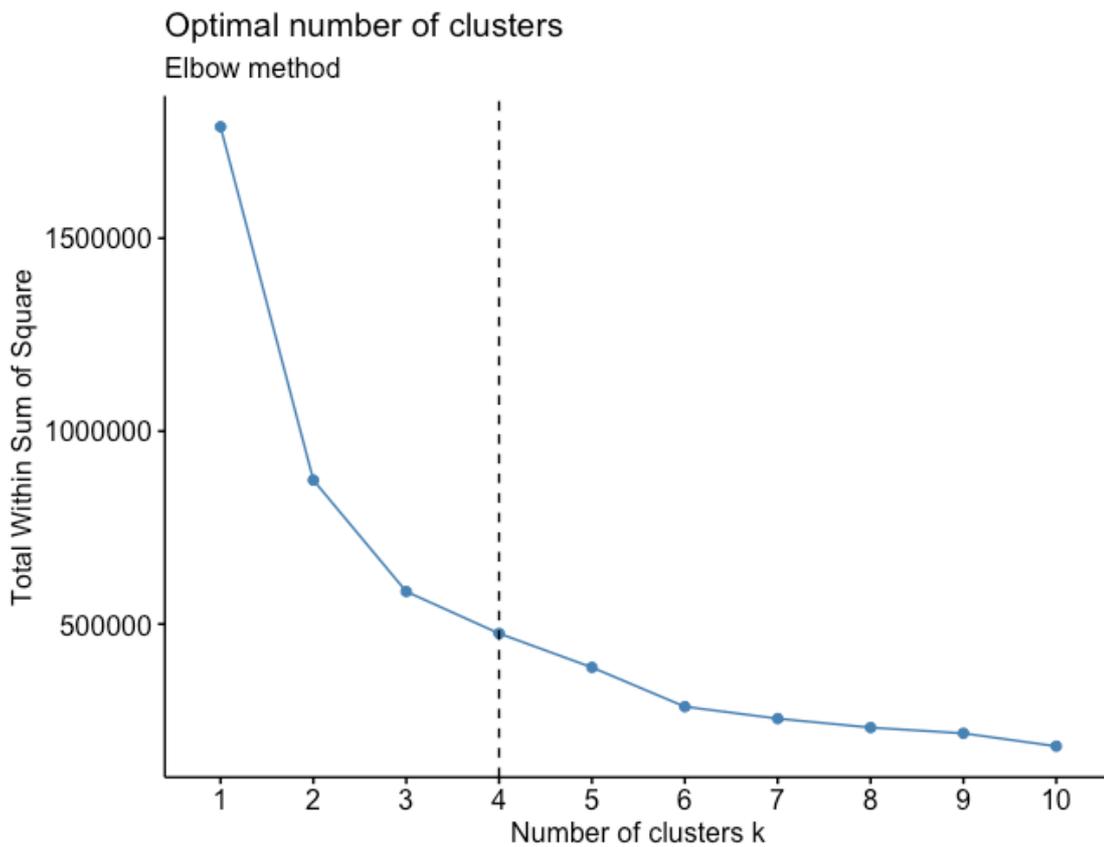
```



**Fig. 20.** *Rappresentazione del dendrogramma*

Si applica l'elbow method per andare a calcolare il numero ottimo di cluster da formare per l'analisi. Come si evince dal grafico, il numero ottimo di cluster è pari a quattro:

```
fviz_nbclust(scale(dist(coffee_new)), kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2) +
  labs(subtitle = "Elbow method")
```

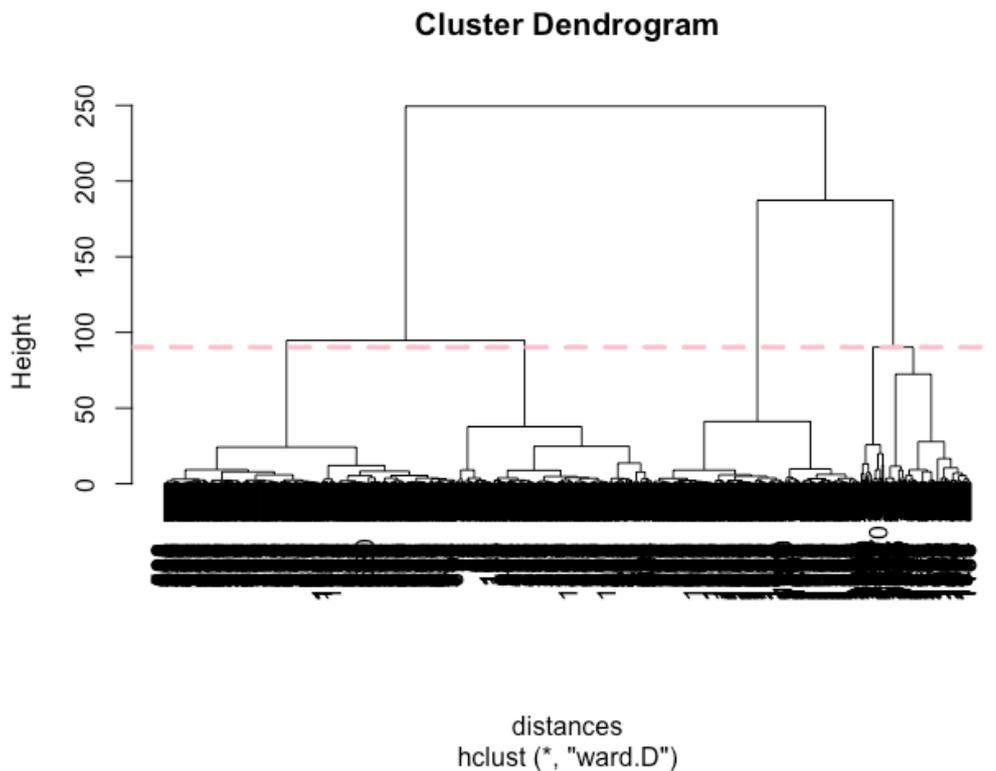


**Fig. 21.** Applicazione dell'Elbow Method

In seguito, si assegnano le osservazioni ai vari cluster con altezza del taglio pari a 90.2 e si ottiene il seguente output:

```
clusterGroups = cutree(clusterCoffee, k = 4)
clusterGroups[1:1000]
subset(clusterCoffee$height, clusterCoffee$height>90.2)
View(clusterGroups)

> subset(clusterCoffee$height, clusterCoffee$height>90.2)
[1] 90.24577 94.72453 187.34375 249.46417
```



**Fig. 22.** Dendrogramma con il taglio ad  $h=90.2$  per l'ottenimento dei quattro cluster

Si va a calcolare la media dei punteggi assegnati a ciascun attributo all'interno di ognuno dei quattro cluster:

```
tapply(coffee_new$Aroma,clusterGroups,mean)
tapply(coffee_new$Flavor,clusterGroups,mean)
tapply(coffee_new$Aftertaste,clusterGroups,mean)
tapply(coffee_new$Acidity,clusterGroups,mean)
tapply(coffee_new$Body,clusterGroups,mean)
tapply(coffee_new$Balance,clusterGroups,mean)
tapply(coffee_new$Uniformity,clusterGroups,mean)
tapply(coffee_new$CleanCup,clusterGroups,mean)
tapply(coffee_new$Sweetness,clusterGroups,mean)
```

```
> tapply(coffee_new$Aroma,clusterGroups,mean)
```

```
1      2      3      4
```

```
7.842523 7.445217 7.609467 7.336319
```

```
> tapply(coffee_new$Flavor,clusterGroups,mean)
```

```

      1      2      3      4
7.857570 7.321848 7.573996 7.258667
> tapply(coffee_new$Aftertaste,clusterGroups,mean)
      1      2      3      4
7.752150 7.204348 7.452172 7.128551
> tapply(coffee_new$Acidity,clusterGroups,mean)
      1      2      3      4
7.840498 7.429402 7.564344 7.290145
> tapply(coffee_new$Body,clusterGroups,mean)
      1      2      3      4
7.791651 7.435761 7.550881 7.280580
> tapply(coffee_new$Balance,clusterGroups,mean)
      1      2      3      4
7.859751 7.341848 7.584283 7.222058
> tapply(coffee_new$Uniformity,clusterGroups,mean)
      1      2      3      4
9.907664 9.024728 10.000000 9.994174
> tapply(coffee_new$CleanCup,clusterGroups,mean)
      1      2      3      4
9.997913 8.872609 10.000000 9.992232
> tapply(coffee_new$Sweetness,clusterGroups,mean)
      1      2      3      4
9.937445 9.131522 10.000000 9.994174

```

Si procede con la creazione di quattro subset con all'interno solo le osservazioni assegnate a ciascuno di essi e si richiede al software di mostrarne le rispettive strutture:

```

cluster1 = subset(coffee_new, clusterGroups==1)
cluster2 = subset(coffee_new, clusterGroups==2)
cluster3 = subset(coffee_new, clusterGroups==3)
cluster4 = subset(coffee_new, clusterGroups==4)

```

```

str(cluster1)
str(cluster2)
str(cluster3)
str(cluster4)

```

```
> str(cluster1)
```

```

'data.frame':   321 obs. of  9 variables:
 $ Aroma   : num  8.67 8.75 8.42 8.17 8.25 8.58 8.42 8.25 8.67 8.08 ...
 $ Flavor   : num  8.83 8.67 8.5 8.58 8.5 8.42 8.5 8.33 8.67 8.58 ...

```

```

$ Aftertaste: num  8.67 8.5 8.42 8.42 8.25 8.42 8.33 8.5 8.58 8.5 ...
$ Acidity      : num  8.75 8.58 8.42 8.42 8.5 8.5 8.5 8.42 8.42 8.5 ...
$ Body        : num  8.5 8.42 8.33 8.5 8.42 8.25 8.25 8.33 8.33 7.67 ...
$ Balance     : num  8.42 8.42 8.42 8.25 8.33 8.33 8.25 8.5 8.42 8.42 ...
$ Uniformity: num  10 10 10 10 10 10 10 10 9.33 10 ...
$ CleanCup   : num  10 10 10 10 10 10 10 10 10 10 ...
$ Sweetness  : num  10 10 10 10 10 10 10 9.33 9.33 10 ...

```

```
> str(cluster2)
```

```

'data.frame':    184 obs. of  9 variables:
 $ Aroma      : num  8.33 8.08 7.92 7.83 8.17 7.75 7.83 7.75 7.67 7.83 ...
 $ Flavor     : num  8.42 8.17 8.17 7.83 8.17 8 7.75 8 7.58 7.83 ...
 $ Aftertaste: num  7.83 7.92 8.17 7.83 7.83 7.92 7.67 7.67 7.5 7.67 ...
 $ Acidity    : num  8.33 8 8.42 7.67 7.67 8.08 7.75 8.08 7.67 7.83 ...
 $ Body       : num  8 8.08 8.08 8 8.08 7.75 7.83 7.83 7.92 7.92 ...
 $ Balance    : num  8.25 8 8.17 8.17 7.75 8 8.5 7.92 8 7.83 ...
 $ Uniformity: num  9.33 9.33 8.67 9.33 9.33 9.33 9.33 9.33 9.33 9.33 ...
 $ CleanCup  : num  9.33 9.33 8.67 9.33 9.33 9.33 9.33 9.33 9.33 9.33 ...
 $ Sweetness : num  10 10 9.33 10 9.33 9.33 9.33 9.33 10 9.33 ...

```

```
> str(cluster3)
```

```

'data.frame':    488 obs. of  9 variables:
 $ Aroma      : num  7.83 7.75 7.83 8.17 7.75 7.83 7.92 8 7.92 7.83 ...
 $ Flavor     : num  7.58 7.58 7.83 7.75 7.83 7.75 7.75 7.5 7.75 7.75 ...
 $ Aftertaste: num  7.58 7.33 7.58 7.58 7.58 7.58 7.67 7.5 7.67 7.5 ...
 $ Acidity    : num  7.83 7.83 7.75 7.67 7.83 7.67 7.67 7.67 7.67 7.83 ...
 $ Body       : num  7.75 7.58 7.75 7.83 7.67 7.75 7.5 8 7.83 7.83 ...
 $ Balance    : num  7.67 8.08 7.75 7.58 7.75 7.83 7.67 7.83 7.92 7.75 ...
 $ Uniformity: num  10 10 10 10 10 10 10 10 10 10 ...
 $ CleanCup  : num  10 10 10 10 10 10 10 10 10 10 ...
 $ Sweetness : num  10 10 10 10 10 10 10 10 10 10 ...

```

```
> str(cluster4)
```

```

'data.frame':    345 obs. of  9 variables:
 $ Aroma      : num  7.58 7.67 7.25 7.42 7.25 7.58 7 7.42 7.58 7.5 ...
 $ Flavor     : num  7.58 7.67 7.25 7.5 7.33 7.67 7.08 7.67 7.58 7.5 ...
 $ Aftertaste: num  7.17 7 7 7.33 7.25 7.58 7 7.5 7.5 7.17 ...
 $ Acidity    : num  7.17 7.08 7.08 7.5 7.5 7.33 7.17 7.08 7.17 7.42 ...
 $ Body       : num  7 7 7 7.25 7.42 7 7 7.5 7.42 7.5 ...
 $ Balance    : num  7 7 7 7.42 7.5 7.58 7 7.42 7.42 7.5 ...
 $ Uniformity: num  10 10 10 10 10 10 10 10 10 10 ...
 $ CleanCup  : num  10 10 10 10 10 10 10 10 10 10 ...
 $ Sweetness : num  10 10 10 10 10 10 10 10 10 10 ...

```

### 3.3.2 CLUSTERING NON GERARCHICO (K-MEANS)

Per eseguire l'algoritmo di clustering k-means, si comincia con una preelaborazione dei dati attraverso due processi: la normalizzazione e la standardizzazione. Come è visibile dall'output, a seguito di questo processo, le variabili avranno media uguale a zero e standard deviation uguale a 1:

```
preProcessParams = preProcess(coffee_new, method =c("scale","center"))
preProcessParams
preProcesscoffee = predict(preProcessParams, coffee_new)
```

```
> sd(preProcesscoffee$Aroma)
[1] 1
> sd(preProcesscoffee$Flavor)
[1] 1
> sd(preProcesscoffee$Aftertaste)
[1] 1
> sd(preProcesscoffee$Acidity)
[1] 1
> sd(preProcesscoffee$Body)
[1] 1
> sd(preProcesscoffee$Balance)
[1] 1
> sd(preProcesscoffee$Uniformity)
[1] 1
> sd(preProcesscoffee$CleanCup)
[1] 1
> sd(preProcesscoffee$Sweetness)
[1] 1
> sd(preProcesscoffee$CupperPoints)
[1] 1
> sd(preProcesscoffee$TotalCupPoints)
[1] 1
```

```
> summary(preProcesscoffee)
```

Aroma		Flavor		Aftertaste		Acidity		Body		Balance		Uniformity	
Min.	:-7.88933	Min.	:-4.2359	Min.	:-3.53011	Min.	:-7.1790	Min.	:-7.9370	Min.	:-6.42942	Min.	:-7.9191
1st Qu.	:-0.48229	1st Qu.	:-0.5743	1st Qu.	:-0.44708	1st Qu.	:-0.6621	1st Qu.	:-0.6274	1st Qu.	:-0.54756	1st Qu.	0.3252
Median	: 0.02418	Median	: 0.1580	Median	: 0.03821	Median	: 0.1211	Median	:-0.0751	Median	:-0.06683	Median	: 0.3252
Mean	: 0.00000	Mean	: 0.00000	Mean	: 0.00000	Mean	: 0.00000	Mean	: 0.00000	Mean	: 0.00000	Mean	: 0.00000
3rd Qu.	: 0.56230	3rd Qu.	: 0.6560	3rd Qu.	: 0.49496	3rd Qu.	: 0.6538	3rd Qu.	: 0.4772	3rd Qu.	: 0.64013	3rd Qu.	: 0.3252
Max.	: 3.72770	Max.	: 3.8196	Max.	: 3.60654	Max.	: 3.7869	Max.	: 3.4335	Max.	: 3.46794	Max.	: 0.3252
CleanCup		Sweetness											
Min.	:-13.7600	Min.	:-15.3987										
1st Qu.	: 0.2202	1st Qu.	: 0.2453										
Median	: 0.2202	Median	: 0.2453										
Mean	: 0.00000	Mean	: 0.00000										
3rd Qu.	: 0.2202	3rd Qu.	: 0.2453										
Max.	: 0.2202	Max.	: 0.2453										

**Fig. 23.** Riassunto delle variabili in seguito alla normalizzazione e standardizzazione

In seguito, si applica il clustering k-means e si va a vedere quali elementi appartengono a quali cluster:

```
kr = kmeans (x = preProcesscoffee, centers = 4)
subset(preProcesscoffee, kr$cluster == 1)
subset(preProcesscoffee, kr$cluster == 2)
subset(preProcesscoffee, kr$cluster == 3)
subset(preProcesscoffee, kr$cluster == 4)
```

```
> subset(preProcesscoffee, kr$cluster == 1)
      Aroma      Flavor Aftertaste      Acidity      Body      Balance Uniformity      CleanCup      Sweetness
993  0.30906443  0.65601993  0.7518769  0.1211325  0.18479764  0.41390108 -6.5382143  0.2202462  0.2452905
1053 0.81552887 -0.07629666  0.7518769  0.4031116  0.99697489 -0.06682778 -6.5382143  0.2202462  0.2452905
1131 0.81552887  0.42167862  0.4949573  0.6537597  0.99697489 -1.00000733 -6.5382143  0.2202462  0.2452905
1150 0.02417819 -0.07629666  0.7518769  0.6537597  0.47718145  0.15939757 -3.7969778 -2.5758009  0.2452905
1193 0.02417819  0.42167862  0.0382115 -0.3801637  0.47718145  0.41390108 -2.4160541 -1.6391251 -2.1545352
1220 1.10041512  0.15804465  0.4949573  1.1863870  1.28935870 -0.06682778 -3.7969778 -4.4351721  0.2452905
1227 -0.22905403  0.65601993  0.4949573  0.9044078  1.28935870  1.12085528 -5.1779014 -2.5758009 -2.1545352
1230 -1.02040471 -0.80861325 -1.1607463 -0.1295156 -0.62737961 -0.29305312 -3.7969778 -0.7164296  0.2452905
1232 -0.22905403 -0.07629666 -0.2187080  0.1211325 -0.07509908 -0.29305312 -2.4160541 -1.6391251 -2.1545352
1233 -0.48228625 -0.57427194 -0.2187080 -0.1295156  0.47718145  0.64012642 -3.7969778 -1.6391251 -0.9636442
1236 -0.22905403 -1.30658853 -0.2187080 -1.1634390 -0.07509908 -0.06682778 -6.5382143  0.2202462  0.2452905
1238 0.02417819 -0.57427194  0.2665844 -0.3801637 -0.07509908 -0.54755664 -3.7969778 -1.6391251 -0.9636442
1243 -0.48228625 -1.04295456 -1.1607463 -1.1634390  0.47718145 -1.00000733 -2.4160541 -0.7164296 -0.9636442
1251 0.02417819  0.42167862  0.4949573  0.4031116 -0.33499580 -0.29305312 -1.0557413 -0.7164296 -6.9722305
1254 -0.22905403 -0.57427194  0.0382115  0.1211325  0.18479764 -1.70696153 -3.7969778 -2.5758009  0.2452905
1259 -1.55852318 -1.54092984 -1.1607463 -0.9127909 -0.62737961 -0.77378198 -2.4160541 -1.6391251  0.2452905
1260 -0.48228625  0.65601993  0.0382115  0.4031116  1.28935870  0.86635177 -2.4160541 -6.3085236 -0.9636442
1261 0.30906443 -0.57427194 -0.6754539 -2.7299896 -1.14717305 -0.29305312 -1.0557413  0.2202462 -5.7632957
1265 -0.22905403 -1.54092984 -1.6460387 -0.1295156 -1.14717305 -0.54755664 -2.4160541 -1.6391251 -2.1545352
1267 -1.02040471 -1.30658853 -1.6460387 -1.1634390 -0.62737961 -1.96146504 -2.4160541 -1.6391251 -0.9636442
1268 -0.22905403 -0.57427194 -0.2187080  0.1211325  0.47718145 -0.77378198 -3.7969778 -2.5758009 -3.3634700
1269 -0.48228625 -1.54092984 -1.6460387 -1.4454181 -1.95935031 -1.96146504 -5.1779014  0.2202462  0.2452905
1274 -1.27363693 -1.54092984 -1.6460387 -0.1295156 -1.14717305 -1.48073618 -3.7969778 -2.5758009  0.2452905
1275 1.10041512  0.15804465 -0.6754539  0.1211325 -0.62737961 -1.00000733 -3.7969778 -2.5758009 -3.3634700
1277 -0.22905403 -0.80861325 -1.3891192 -1.6960662 -1.14717305 -1.96146504 -3.7969778 -2.5758009  0.2452905
1278 -1.27363693 -1.77527114 -2.1027846 -2.2286934 -0.88727633  0.64012642  0.3251824 -5.3718479  0.2452905
1279 -1.02040471 -2.03890511 -2.1027846 -1.4454181 -1.14717305 -1.96146504  0.3251824 -4.4351721  0.2452905
1280 -1.27363693 -1.04295456 -1.6460387 -1.1634390 -0.33499580 -2.41391573 -2.4160541 -3.5124766  0.2452905
1281 -1.81175540 -1.77527114 -1.8744116 -1.9467143 -1.14717305 -1.96146504  0.3251824 -4.4351721  0.2452905
1282 -1.55852318 -2.27324642 -2.3597041 -1.6960662 -1.69945358 -2.18769039 -2.4160541 -1.6391251  0.2452905
1283 -0.76717250 -2.27324642 -2.1027846 -0.9127909 -0.07509908 -1.48073618  0.3251824 -6.3085236  0.2452905
1284 -2.60310608 -3.00556301 -3.5301153 -1.6960662 -0.88727633 -2.18769039 -3.7969778  0.2202462  0.2452905
1285 -2.60310608 -2.50758773 -3.5301153 -1.1634390 -1.69945358 -2.66841925 -3.7969778  0.2202462  0.2452905
1286 -1.55852318 -1.54092984 -1.6460387 -0.1295156 -0.62737961 -1.96146504 -3.7969778 -3.5124766  0.2452905
1287 -0.22905403 -0.80861325 -1.1607463 -0.9127909 -0.33499580 -0.54755664 -3.7969778 -1.6391251 -5.7632957
1288 -3.39445677 -3.50353829 -2.5880770 -0.1295156 -0.62737961 -1.96146504 -2.4160541  0.2202462 -2.1545352
1289 -1.81175540 -2.77122170 -2.1027846 -2.2286934 -1.43955686 -1.96146504  0.3251824 -4.4351721  0.2452905
1290 -2.06498762 -3.23990432 -3.5301153 -0.6621428 -0.88727633 -2.18769039  0.3251824 -4.4351721  0.2452905
1291 0.30906443 -1.54092984 -0.6754539  0.1211325 -0.62737961 -0.54755664  0.3251824 -10.0272662  0.2452905
1292 -1.27363693 -1.04295456 -0.6754539  0.1211325  0.47718145 -1.70696153 -1.0557413 -4.4351721 -5.7632957
1293 -1.02040471 -2.77122170 -2.3597041 -1.6960662 -1.69945358 -3.12086994 -3.7969778 -2.5758009  0.2452905
1294 -1.81175540 -1.04295456 -1.6460387 -1.1634390 -1.69945358 -1.00000733 -3.7969778 -2.5758009 -3.3634700
1295 -1.55852318 -1.30658853 -1.3891192 -1.4454181 -1.95935031 -1.70696153 -3.7969778 -2.5758009 -3.3634700
```

```

1296 1.35364734 -0.57427194 0.9802498 -0.6621428 -0.33499580 0.41390108 -6.5382143 -4.4351721 -5.7632957
1297 -2.06498762 -1.77527114 -1.3891192 -0.1295156 0.18479764 -1.00000733 -6.5382143 -4.4351721 -0.9636442
1298 -2.06498762 -2.27324642 -0.9323734 -1.1634390 -0.62737961 -2.41391573 0.3251824 -6.3085236 -2.1545352
1299 -3.93257523 -3.00556301 -3.0733694 -2.2286934 -3.87608862 -2.89464459 0.3251824 -4.4351721 0.2452905
1300 -2.06498762 -1.77527114 -1.3891192 -1.9467143 -1.14717305 -1.48073618 0.3251824 -10.0272662 0.2452905
1301 -3.39445677 -2.50758773 -2.8164499 -1.1634390 -0.62737961 -2.89464459 -3.7969778 -5.3718479 0.2452905
1302 -2.06498762 -1.54092984 -1.6460387 -1.9467143 -0.33499580 -1.70696153 -7.9191379 -5.3718479 0.2452905
1303 -1.81175540 -1.54092984 -1.6460387 -1.6960662 -0.62737961 -1.96146504 -7.9191379 -5.3718479 0.2452905
1304 -2.85633830 -3.00556301 -3.5301153 -2.7299896 -2.25173412 -3.82782414 -3.7969778 -2.5758009 -3.3634700
1305 -1.81175540 -3.97222091 -3.5301153 -2.7299896 -3.32380809 -3.82782414 -3.7969778 -2.5758009 -3.3634700
1306 -1.81175540 -3.50353829 -3.5301153 -3.2626168 -2.77152756 -3.82782414 -3.7969778 -2.5758009 -3.3634700
1307 -1.55852318 -2.03890511 -3.3017424 -0.3801637 -0.88727633 -2.18769039 0.3251824 -13.7599890 0.2452905
1308 -2.60310608 -2.77122170 -2.8164499 -2.7299896 -1.43955686 -2.41391573 -1.0557413 -5.3718479 -6.9722305
1309 -1.02040471 -2.77122170 -3.0733694 -4.0458921 -3.58370481 -4.08232765 -7.9191379 -5.3718479 -6.9722305
1310 -0.22905403 -2.50758773 -2.1027846 0.4031116 -0.62737961 -2.41391573 -3.7969778 -11.9006177 -15.3986862
1337 -0.76717250 -0.57427194 -0.6754539 -0.3801637 -0.07509908 -1.00000733 -1.0557413 -0.7164296 -4.4100105
1338 -0.48228625 -2.03890511 -1.8744116 -1.1634390 -0.88727633 -1.48073618 -1.0557413 -0.7164296 -5.0234998
1339 -2.60310608 -2.50758773 -2.5880770 -2.2286934 -1.95935031 -1.96146504 -1.0557413 -0.7164296 -5.7632957

```

```
> subset(preProcesscoffee, kr$cluster == 2)
```

```

      Aroma      Flavor Aftertaste  Acidity      Body      Balance Uniformity CleanCup Sweetness
180 -0.76717250 0.15804465 1.2086227 0.6537597 0.73707817 1.12085528 0.3251824 0.2202462 0.2452905
186 0.56229665 0.15804465 -0.2187080 0.9044078 0.18479764 1.57330597 0.3251824 0.2202462 0.2452905
193 0.81552887 0.65601993 0.9802498 -0.3801637 0.99697489 0.41390108 0.3251824 0.2202462 0.2452905
213 1.10041512 0.65601993 0.7518769 0.4031116 -0.07509908 0.41390108 0.3251824 0.2202462 0.2452905
234 0.02417819 0.42167862 0.7518769 0.4031116 0.73707817 0.86635177 0.3251824 0.2202462 0.2452905
242 0.81552887 0.65601993 0.7518769 0.6537597 -0.62737961 0.64012642 0.3251824 0.2202462 0.2452905
245 0.30906443 0.42167862 1.2086227 0.1211325 0.18479764 0.86635177 0.3251824 0.2202462 0.2452905
247 0.56229665 0.89036123 0.2665844 0.9044078 0.18479764 0.41390108 0.3251824 0.2202462 0.2452905
248 0.30906443 0.65601993 0.7518769 0.4031116 0.47718145 0.64012642 0.3251824 0.2202462 0.2452905
250 0.30906443 0.42167862 0.2665844 -0.1295156 0.47718145 0.41390108 0.3251824 0.2202462 0.2452905
251 0.56229665 0.65601993 0.2665844 0.1211325 0.47718145 1.12085528 0.3251824 0.2202462 0.2452905
252 0.02417819 0.89036123 0.7518769 -0.9127909 1.54925542 0.64012642 0.3251824 0.2202462 0.2452905
258 0.02417819 0.65601993 0.4949573 0.6537597 0.18479764 0.64012642 0.3251824 0.2202462 0.2452905
259 -0.22905403 0.65601993 0.4949573 -0.1295156 0.18479764 -0.06682778 0.3251824 0.2202462 0.2452905
260 0.81552887 0.89036123 0.4949573 1.1863870 -0.33499580 -0.06682778 0.3251824 0.2202462 0.2452905
261 0.30906443 0.65601993 0.2665844 0.4031116 0.47718145 0.86635177 0.3251824 0.2202462 0.2452905
262 0.02417819 0.42167862 0.7518769 0.6537597 0.18479764 0.86635177 0.3251824 0.2202462 0.2452905
267 1.10041512 0.42167862 0.2665844 0.9044078 0.18479764 0.41390108 0.3251824 0.2202462 0.2452905
268 0.81552887 0.65601993 0.4949573 0.1211325 0.73707817 0.64012642 0.3251824 0.2202462 0.2452905
272 0.30906443 0.89036123 0.4949573 0.4031116 0.73707817 0.86635177 0.3251824 0.2202462 0.2452905
273 -0.22905403 0.42167862 0.4949573 0.9044078 0.73707817 0.64012642 0.3251824 0.2202462 0.2452905
275 1.10041512 0.15804465 0.2665844 0.9044078 0.18479764 0.41390108 0.3251824 0.2202462 0.2452905
277 0.30906443 0.42167862 0.9802498 0.4031116 -0.07509908 0.64012642 0.3251824 0.2202462 0.2452905
278 0.56229665 0.65601993 0.2665844 0.6537597 0.99697489 0.15939757 0.3251824 0.2202462 0.2452905
279 0.56229665 0.42167862 0.2665844 0.6537597 0.73707817 0.41390108 0.3251824 0.2202462 0.2452905
280 -1.02040471 0.15804465 0.0382115 1.4370351 0.73707817 0.86635177 0.3251824 0.2202462 0.2452905
281 0.30906443 0.65601993 0.7518769 0.6537597 -0.33499580 0.64012642 0.3251824 0.2202462 0.2452905
282 0.56229665 0.65601993 0.7518769 0.4031116 0.18479764 0.41390108 0.3251824 0.2202462 0.2452905
284 0.81552887 -0.31063797 0.2665844 0.9044078 0.73707817 0.86635177 0.3251824 0.2202462 0.2452905
285 0.02417819 0.89036123 -0.2187080 1.4370351 0.18479764 0.64012642 0.3251824 0.2202462 0.2452905
286 1.10041512 -0.07629666 -0.2187080 0.4031116 2.10153596 -0.06682778 0.3251824 0.2202462 0.2452905
287 -0.22905403 0.42167862 0.4949573 0.6537597 0.99697489 0.86635177 0.3251824 0.2202462 0.2452905
288 0.30906443 0.15804465 0.4949573 0.9044078 0.18479764 0.41390108 0.3251824 0.2202462 0.2452905
289 -0.22905403 0.89036123 0.2665844 0.9044078 0.99697489 0.41390108 0.3251824 0.2202462 0.2452905
290 0.30906443 1.38833651 0.7518769 -0.3801637 0.47718145 -0.06682778 0.3251824 0.2202462 0.2452905
291 0.30906443 0.15804465 0.2665844 0.6537597 0.73707817 0.64012642 0.3251824 0.2202462 0.2452905
292 0.81552887 0.89036123 0.4949573 0.4031116 -0.07509908 0.15939757 0.3251824 0.2202462 0.2452905
294 -0.22905403 0.89036123 0.4949573 0.1211325 0.73707817 0.86635177 0.3251824 0.2202462 0.2452905
296 0.56229665 0.15804465 0.2665844 0.1211325 0.73707817 0.64012642 0.3251824 0.2202462 0.2452905
297 0.30906443 0.42167862 0.4949573 0.6537597 0.47718145 0.41390108 0.3251824 0.2202462 0.2452905
298 0.81552887 0.89036123 0.4949573 0.1211325 0.18479764 0.15939757 0.3251824 0.2202462 0.2452905
299 0.30906443 0.42167862 0.4949573 0.1211325 0.47718145 0.86635177 0.3251824 0.2202462 0.2452905
300 1.10041512 0.65601993 0.2665844 0.4031116 0.73707817 -0.06682778 0.3251824 0.2202462 0.2452905

```

301	0.81552887	0.65601993	0.7518769	-0.1295156	0.18479764	0.41390108	0.3251824	0.2202462	0.2452905
302	0.56229665	0.42167862	0.2665844	0.4031116	0.18479764	0.86635177	0.3251824	0.2202462	0.2452905
303	1.35364734	1.15399521	0.4949573	0.6537597	-0.62737961	-0.06682778	0.3251824	0.2202462	0.2452905
305	0.30906443	0.42167862	0.7518769	1.1863870	-0.07509908	0.41390108	0.3251824	0.2202462	0.2452905
306	0.30906443	-0.31063797	0.0382115	-0.3801637	1.54925542	2.28026018	0.3251824	0.2202462	0.2452905
307	0.02417819	0.42167862	0.9802498	0.4031116	0.47718145	0.41390108	0.3251824	0.2202462	0.2452905
308	0.56229665	0.42167862	0.9802498	0.1211325	-0.33499580	1.34708063	0.3251824	0.2202462	0.2452905
309	-1.02040471	-0.31063797	0.7518769	0.1211325	0.99697489	2.05403483	0.3251824	0.2202462	0.2452905
310	0.81552887	0.15804465	0.4949573	-0.1295156	0.73707817	0.41390108	0.3251824	0.2202462	0.2452905
311	0.30906443	0.15804465	0.4949573	0.4031116	0.73707817	0.64012642	0.3251824	0.2202462	0.2452905
312	0.81552887	0.89036123	0.4949573	-0.6621428	-0.33499580	0.41390108	0.3251824	0.2202462	0.2452905
313	0.56229665	0.42167862	-0.4470809	-0.9127909	0.18479764	2.76098903	0.3251824	0.2202462	0.2452905
314	0.81552887	0.42167862	0.9802498	0.4031116	0.18479764	-0.06682778	0.3251824	0.2202462	0.2452905
315	0.30906443	0.89036123	0.2665844	0.4031116	0.47718145	-0.06682778	0.3251824	0.2202462	0.2452905
316	1.60687956	0.42167862	0.7518769	0.1211325	-0.88727633	0.41390108	0.3251824	0.2202462	0.2452905
318	0.30906443	0.65601993	0.7518769	0.4031116	0.99697489	0.15939757	0.3251824	0.2202462	0.2452905
319	0.02417819	0.65601993	0.4949573	0.6537597	0.18479764	0.15939757	0.3251824	0.2202462	0.2452905
320	-0.22905403	0.42167862	-0.2187080	1.1863870	0.99697489	0.41390108	0.3251824	0.2202462	0.2452905
321	-0.76717250	0.42167862	0.7518769	-0.1295156	0.47718145	-0.06682778	0.3251824	0.2202462	0.2452905
322	-0.22905403	0.65601993	0.4949573	0.9044078	0.47718145	0.15939757	0.3251824	0.2202462	0.2452905
323	-0.22905403	0.65601993	0.2665844	0.6537597	0.47718145	0.41390108	0.3251824	0.2202462	0.2452905
325	0.81552887	0.65601993	0.7518769	0.1211325	0.18479764	0.41390108	0.3251824	0.2202462	0.2452905
326	0.30906443	0.42167862	0.7518769	0.4031116	-0.62737961	0.64012642	0.3251824	0.2202462	0.2452905
327	0.30906443	0.42167862	0.0382115	0.4031116	1.54925542	0.64012642	0.3251824	0.2202462	0.2452905
328	0.81552887	0.42167862	-0.2187080	0.4031116	0.47718145	0.86635177	0.3251824	0.2202462	0.2452905
329	0.56229665	0.65601993	0.4949573	0.9044078	-1.14717305	0.64012642	0.3251824	0.2202462	0.2452905
331	0.30906443	1.88631179	0.2665844	-0.3801637	0.18479764	0.86635177	0.3251824	0.2202462	0.2452905
332	0.30906443	-0.07629666	0.2665844	-0.6621428	0.99697489	2.05403483	0.3251824	0.2202462	0.2452905
333	1.89176580	0.42167862	0.4949573	0.1211325	-0.33499580	0.41390108	0.3251824	0.2202462	0.2452905
334	0.30906443	-0.07629666	-0.2187080	1.4370351	0.47718145	0.41390108	0.3251824	0.2202462	0.2452905
336	0.02417819	0.42167862	0.4949573	0.1211325	0.47718145	0.15939757	0.3251824	0.2202462	0.2452905
337	0.56229665	0.65601993	0.2665844	0.9044078	0.18479764	0.15939757	0.3251824	0.2202462	0.2452905
338	0.56229665	0.65601993	0.4949573	0.1211325	0.47718145	0.15939757	0.3251824	0.2202462	0.2452905
339	0.30906443	0.15804465	0.4949573	0.1211325	0.47718145	0.64012642	0.3251824	0.2202462	0.2452905
340	0.30906443	0.65601993	0.4949573	0.4031116	0.73707817	0.15939757	0.3251824	0.2202462	0.2452905
341	0.56229665	0.65601993	0.4949573	0.1211325	-0.07509908	0.41390108	0.3251824	0.2202462	0.2452905
342	0.56229665	0.42167862	0.4949573	0.4031116	-0.07509908	0.41390108	0.3251824	0.2202462	0.2452905
343	0.30906443	0.42167862	0.0382115	0.6537597	0.73707817	0.41390108	0.3251824	0.2202462	0.2452905
344	0.30906443	0.15804465	0.7518769	0.4031116	0.47718145	0.41390108	0.3251824	0.2202462	0.2452905
345	0.30906443	0.15804465	0.7518769	0.1211325	0.47718145	0.64012642	0.3251824	0.2202462	0.2452905
346	0.56229665	0.15804465	0.2665844	0.1211325	0.73707817	0.86635177	0.3251824	0.2202462	0.2452905
347	0.56229665	0.65601993	0.9802498	0.6537597	-0.07509908	0.15939757	0.3251824	0.2202462	0.2452905
348	0.30906443	0.42167862	0.4949573	0.4031116	0.73707817	0.15939757	0.3251824	0.2202462	0.2452905
349	0.81552887	0.42167862	0.4949573	0.1211325	0.47718145	0.15939757	0.3251824	0.2202462	0.2452905
350	0.30906443	0.15804465	0.4949573	0.1211325	0.47718145	0.64012642	0.3251824	0.2202462	0.2452905
351	0.81552887	0.65601993	0.2665844	0.6537597	-0.07509908	0.15939757	0.3251824	0.2202462	0.2452905
352	0.56229665	0.15804465	0.4949573	0.6537597	0.73707817	-0.06682778	0.3251824	0.2202462	0.2452905
354	1.10041512	-0.31063797	0.2665844	0.4031116	-0.07509908	0.41390108	0.3251824	0.2202462	0.2452905
355	0.56229665	0.15804465	0.4949573	-0.1295156	0.73707817	0.64012642	0.3251824	0.2202462	0.2452905
356	0.30906443	0.65601993	0.2665844	0.9044078	0.18479764	0.15939757	0.3251824	0.2202462	0.2452905
357	0.02417819	0.15804465	0.4949573	0.1211325	0.73707817	0.41390108	0.3251824	0.2202462	0.2452905
358	0.56229665	0.42167862	0.4949573	0.4031116	0.47718145	0.15939757	0.3251824	0.2202462	0.2452905
359	-1.02040471	0.42167862	0.4949573	1.1863870	0.99697489	0.15939757	0.3251824	0.2202462	0.2452905
360	0.30906443	0.65601993	0.2665844	0.1211325	0.73707817	0.41390108	0.3251824	0.2202462	0.2452905
361	0.30906443	-0.31063797	0.7518769	0.9044078	0.18479764	0.41390108	0.3251824	0.2202462	0.2452905
362	1.35364734	0.15804465	-0.4470809	0.1211325	1.28935870	0.15939757	0.3251824	0.2202462	0.2452905
363	0.02417819	0.65601993	0.2665844	0.9044078	-0.33499580	0.41390108	0.3251824	0.2202462	0.2452905
364	0.56229665	0.15804465	-0.2187080	-0.1295156	0.99697489	0.86635177	0.3251824	0.2202462	0.2452905
365	0.30906443	-0.07629666	0.7518769	-0.1295156	0.47718145	0.41390108	0.3251824	0.2202462	0.2452905
366	0.02417819	0.65601993	0.0382115	0.4031116	0.73707817	0.41390108	0.3251824	0.2202462	0.2452905
367	0.02417819	0.42167862	0.4949573	0.6537597	0.18479764	0.64012642	0.3251824	0.2202462	0.2452905
368	0.81552887	0.15804465	-0.4470809	-0.1295156	-0.07509908	-0.29305312	0.3251824	0.2202462	0.2452905
369	0.30906443	0.65601993	0.7518769	0.6537597	0.18479764	-0.06682778	0.3251824	0.2202462	0.2452905
370	0.56229665	0.42167862	0.7518769	0.6537597	-0.62737961	-0.06682778	0.3251824	0.2202462	0.2452905
371	0.81552887	-0.07629666	0.2665844	1.4370351	0.47718145	-0.06682778	0.3251824	0.2202462	0.2452905
372	0.30906443	0.15804465	0.4949573	0.4031116	0.47718145	0.41390108	0.3251824	0.2202462	0.2452905
373	-0.22905403	-0.07629666	0.9802498	-0.1295156	1.54925542	0.41390108	0.3251824	0.2202462	0.2452905
374	-0.22905403	0.42167862	0.4949573	1.1863870	0.18479764	0.41390108	0.3251824	0.2202462	0.2452905

```

> subset(preProcesscoffee, kr$cluster == 3)

```

	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	CleanCup	Sweetness
1	3.47446717	3.8196276	3.6065383	3.7868610	3.17360993	2.5347637	0.3251824	0.2202462	0.2452905
2	3.72769939	3.3509450	3.1212458	3.2542338	2.91371321	2.5347637	0.3251824	0.2202462	0.2452905
3	2.68311649	2.8529697	2.8928729	2.7529376	2.62132940	2.5347637	0.3251824	0.2202462	0.2452905
4	1.89176580	3.0873110	2.8928729	2.7529376	3.17360993	2.0540348	0.3251824	0.2202462	0.2452905
5	2.14499802	2.8529697	2.4075805	3.0035857	2.91371321	2.2802602	0.3251824	0.2202462	0.2452905
6	3.18958092	2.6186284	2.8928729	3.0035857	2.36143268	2.2802602	0.3251824	0.2202462	0.2452905
7	2.68311649	2.8529697	2.6359534	3.0035857	2.36143268	2.0540348	0.3251824	0.2202462	0.2452905
8	2.14499802	2.3549944	3.1212458	2.7529376	2.62132940	2.7609890	0.3251824	0.2202462	-0.9636442
9	3.47446717	3.3509450	3.3496187	2.7529376	2.62132940	2.5347637	-1.0557413	0.2202462	-0.9636442
10	1.60687956	3.0873110	3.1212458	3.0035857	0.47718145	2.5347637	0.3251824	0.2202462	0.2452905
11	1.89176580	3.3509450	2.4075805	3.0035857	0.73707817	1.8278095	0.3251824	0.2202462	0.2452905
12	2.14499802	2.6186284	2.1792076	2.4709585	1.80915214	1.8278095	0.3251824	0.2202462	0.2452905
13	1.60687956	3.3509450	2.6359534	2.7529376	1.54925542	1.5733060	0.3251824	0.2202462	0.2452905
14	2.39823024	2.6186284	1.9222880	2.2203104	2.36143268	1.3470806	0.3251824	0.2202462	0.2452905
15	2.14499802	2.3549944	3.1212458	2.2203104	3.43350665	3.4679432	-1.0557413	0.2202462	-0.9636442
16	1.35364734	2.8529697	3.3496187	1.9696623	2.10153596	1.3470806	0.3251824	0.2202462	0.2452905
17	2.39823024	2.1206531	1.2086227	0.6537597	3.17360993	2.5347637	0.3251824	0.2202462	0.2452905
18	1.89176580	2.3549944	2.4075805	2.4709585	2.91371321	2.2802602	-1.0557413	0.2202462	-0.9636442
19	2.68311649	2.1206531	1.9222880	1.9696623	1.28935870	1.3470806	0.3251824	0.2202462	0.2452905
20	1.89176580	1.8863118	1.6939151	1.9696623	1.80915214	2.2802602	0.3251824	0.2202462	0.2452905
21	1.35364734	2.1206531	1.9222880	3.0035857	2.36143268	1.3470806	0.3251824	0.2202462	0.2452905
22	1.60687956	2.1206531	1.6939151	1.9696623	1.54925542	2.2802602	0.3251824	0.2202462	0.2452905
23	1.89176580	2.1206531	2.1792076	1.4370351	0.99697489	1.8278095	0.3251824	0.2202462	0.2452905
24	2.14499802	2.3549944	2.1792076	1.9696623	0.99697489	1.8278095	0.3251824	0.2202462	0.2452905
25	2.68311649	1.8863118	1.4655422	1.9696623	2.62132940	1.3470806	0.3251824	0.2202462	0.2452905
26	2.93634871	2.8529697	1.6939151	1.4370351	1.54925542	1.3470806	0.3251824	0.2202462	0.2452905
27	0.81552887	2.1206531	1.9222880	1.9696623	2.10153596	1.8278095	0.3251824	0.2202462	0.2452905
28	2.68311649	1.8863118	2.1792076	1.9696623	0.99697489	1.1208553	0.3251824	0.2202462	0.2452905
29	1.89176580	1.6226778	1.9222880	1.4370351	1.80915214	1.3470806	0.3251824	0.2202462	0.2452905
30	1.35364734	1.3883365	1.6939151	2.2203104	1.54925542	1.8278095	0.3251824	0.2202462	0.2452905
31	1.10041512	2.1206531	1.6939151	2.4709585	1.54925542	1.5733060	0.3251824	0.2202462	0.2452905
32	2.68311649	1.8863118	2.1792076	1.4370351	0.18479764	1.3470806	0.3251824	0.2202462	0.2452905
33	2.93634871	1.8863118	1.6939151	0.6537597	1.54925542	1.3470806	0.3251824	0.2202462	0.2452905
34	1.89176580	0.8903612	1.6939151	1.6876832	0.99697489	1.3470806	0.3251824	0.2202462	0.2452905
35	1.35364734	1.6226778	1.4655422	1.4370351	1.80915214	1.5733060	0.3251824	0.2202462	0.2452905
36	1.60687956	1.3883365	1.6939151	2.2203104	1.28935870	1.1208553	0.3251824	0.2202462	0.2452905
37	1.60687956	1.8863118	1.6939151	1.1863870	1.28935870	0.8663518	0.3251824	0.2202462	0.2452905
38	2.14499802	2.6186284	1.9222880	0.6537597	0.47718145	0.8663518	0.3251824	0.2202462	0.2452905
39	1.10041512	0.6560199	1.9222880	1.6876832	1.80915214	1.8278095	0.3251824	0.2202462	0.2452905
40	0.30906443	1.8863118	1.6939151	1.4370351	2.10153596	1.5733060	0.3251824	0.2202462	0.2452905
41	0.81552887	2.3549944	1.2086227	2.2203104	0.18479764	1.1208553	0.3251824	0.2202462	0.2452905
42	1.89176580	1.1539952	1.4655422	0.6537597	2.62132940	1.3470806	0.3251824	0.2202462	0.2452905

43	2.68311649	1.1539952	1.6939151	0.6537597	1.28935870	1.3470806	0.3251824	0.2202462	0.2452905
44	1.60687956	1.8863118	0.9802498	1.6876832	0.73707817	0.8663518	0.3251824	0.2202462	0.2452905
45	1.60687956	1.6226778	1.6939151	1.4370351	0.99697489	1.5733060	0.3251824	0.2202462	0.2452905
46	2.39823024	1.3883365	1.6939151	1.4370351	0.73707817	1.3470806	0.3251824	0.2202462	0.2452905
47	1.60687956	1.6226778	1.2086227	1.1863870	1.80915214	1.1208553	0.3251824	0.2202462	0.2452905
48	0.56229665	1.1539952	1.2086227	1.6876832	1.80915214	1.8278095	0.3251824	0.2202462	0.2452905
49	1.10041512	1.8863118	1.2086227	3.0035857	-0.07509908	1.1208553	0.3251824	0.2202462	0.2452905
50	2.14499802	1.3883365	1.2086227	1.6876832	1.54925542	1.1208553	0.3251824	0.2202462	0.2452905
51	0.56229665	1.6226778	1.6939151	2.2203104	1.54925542	1.3470806	0.3251824	0.2202462	0.2452905
52	2.39823024	1.8863118	0.9802498	1.4370351	0.73707817	0.8663518	0.3251824	0.2202462	0.2452905
53	1.10041512	1.6226778	1.4655422	1.6876832	1.80915214	0.8663518	0.3251824	0.2202462	0.2452905
54	1.89176580	1.8863118	1.4655422	1.6876832	0.99697489	0.6401264	0.3251824	0.2202462	0.2452905
55	1.35364734	1.3883365	1.6939151	1.9696623	0.99697489	1.3470806	0.3251824	0.2202462	0.2452905
56	1.35364734	1.3883365	1.4655422	1.6876832	0.99697489	1.3470806	0.3251824	0.2202462	0.2452905
57	0.81552887	1.6226778	1.4655422	1.6876832	1.54925542	1.1208553	0.3251824	0.2202462	0.2452905
58	0.81552887	1.3883365	1.6939151	1.9696623	0.99697489	1.3470806	0.3251824	0.2202462	0.2452905
59	1.35364734	1.3883365	1.2086227	2.4709585	0.99697489	0.8663518	0.3251824	0.2202462	0.2452905
60	0.81552887	1.3883365	1.2086227	1.9696623	1.28935870	1.3470806	0.3251824	0.2202462	0.2452905
61	1.35364734	1.3883365	1.6939151	1.6876832	1.28935870	0.6401264	0.3251824	0.2202462	0.2452905
62	2.39823024	2.6186284	1.2086227	2.4709585	1.54925542	2.0540348	-1.0557413	-0.7164296	0.2452905
63	1.10041512	1.8863118	1.6939151	1.1863870	0.73707817	0.8663518	0.3251824	0.2202462	0.2452905
64	0.81552887	1.3883365	1.6939151	0.6537597	1.28935870	1.8278095	0.3251824	0.2202462	0.2452905
65	0.81552887	1.3883365	1.4655422	0.6537597	1.80915214	1.3470806	0.3251824	0.2202462	0.2452905
66	1.35364734	0.6560199	1.2086227	0.9044078	0.47718145	2.9872144	0.3251824	0.2202462	0.2452905
67	1.10041512	1.1539952	1.2086227	1.6876832	0.99697489	1.1208553	0.3251824	0.2202462	0.2452905
68	0.81552887	1.3883365	1.2086227	1.1863870	0.99697489	1.5733060	0.3251824	0.2202462	0.2452905
69	2.39823024	1.1539952	0.7518769	1.1863870	0.47718145	1.1208553	0.3251824	0.2202462	0.2452905
70	1.35364734	0.8903612	0.7518769	2.4709585	0.99697489	1.3470806	0.3251824	0.2202462	0.2452905
71	0.56229665	0.8903612	1.2086227	1.9696623	1.28935870	0.8663518	0.3251824	0.2202462	0.2452905
72	1.60687956	1.3883365	1.2086227	0.4031116	1.54925542	0.8663518	0.3251824	0.2202462	0.2452905
73	1.10041512	1.1539952	0.9802498	1.1863870	0.73707817	1.1208553	0.3251824	0.2202462	0.2452905
74	1.35364734	1.3883365	1.6939151	0.4031116	0.73707817	1.3470806	0.3251824	0.2202462	0.2452905
75	1.60687956	0.8903612	1.2086227	1.6876832	0.99697489	0.8663518	0.3251824	0.2202462	0.2452905
76	1.35364734	1.3883365	1.6939151	0.9044078	0.99697489	0.6401264	0.3251824	0.2202462	0.2452905
77	0.81552887	0.6560199	1.2086227	0.4031116	0.99697489	2.9872144	0.3251824	0.2202462	0.2452905
78	1.10041512	1.6226778	1.6939151	1.1863870	1.54925542	1.5733060	0.3251824	0.2202462	-0.9636442
79	1.35364734	0.1580446	0.4949573	0.4031116	1.80915214	2.7609890	0.3251824	0.2202462	0.2452905
80	0.81552887	1.6226778	0.9802498	1.6876832	1.28935870	0.6401264	0.3251824	0.2202462	0.2452905
81	1.35364734	1.1539952	1.4655422	1.6876832	0.99697489	0.6401264	0.3251824	0.2202462	0.2452905
82	0.81552887	0.8903612	1.4655422	1.6876832	0.73707817	0.8663518	0.3251824	0.2202462	0.2452905
83	1.10041512	0.8903612	0.9802498	2.4709585	0.73707817	0.6401264	0.3251824	0.2202462	0.2452905
84	1.35364734	1.1539952	0.9802498	1.4370351	1.28935870	0.8663518	0.3251824	0.2202462	0.2452905
85	1.89176580	1.1539952	1.2086227	1.4370351	0.18479764	0.8663518	0.3251824	0.2202462	0.2452905
86	0.02417819	1.6226778	1.4655422	0.9044078	1.54925542	1.1208553	0.3251824	0.2202462	0.2452905
87	0.56229665	0.8903612	0.7518769	1.4370351	1.80915214	0.8663518	0.3251824	0.2202462	0.2452905
88	2.14499802	1.3883365	0.7518769	1.1863870	0.73707817	0.4139011	0.3251824	0.2202462	0.2452905
89	0.56229665	1.3883365	0.4949573	1.4370351	1.54925542	1.1208553	0.3251824	0.2202462	0.2452905
90	1.89176580	0.8903612	0.4949573	1.6876832	1.54925542	0.6401264	0.3251824	0.2202462	0.2452905
91	2.68311649	1.3883365	0.0382115	1.4370351	1.28935870	1.1208553	-1.0557413	0.2202462	0.2452905
92	0.02417819	0.8903612	1.2086227	1.1863870	0.99697489	1.8278095	0.3251824	0.2202462	0.2452905
93	0.81552887	1.3883365	0.9802498	0.6537597	0.99697489	0.8663518	0.3251824	0.2202462	0.2452905
94	1.60687956	1.8863118	1.4655422	1.4370351	1.80915214	1.3470806	-1.0557413	-0.7164296	0.2452905
95	0.30906443	1.3883365	1.2086227	1.4370351	1.28935870	0.8663518	0.3251824	0.2202462	0.2452905
96	-0.22905403	1.1539952	2.4075805	0.9044078	1.28935870	0.6401264	0.3251824	0.2202462	0.2452905
97	1.89176580	1.1539952	0.9802498	0.6537597	0.47718145	0.6401264	0.3251824	0.2202462	0.2452905
98	1.35364734	1.1539952	0.9802498	1.1863870	0.73707817	0.8663518	0.3251824	0.2202462	0.2452905
99	0.81552887	1.3883365	0.4949573	1.1863870	1.28935870	0.8663518	0.3251824	0.2202462	0.2452905
100	0.81552887	1.3883365	1.2086227	0.6537597	0.47718145	0.8663518	0.3251824	0.2202462	0.2452905
101	1.89176580	0.8903612	1.4655422	0.6537597	0.47718145	0.8663518	0.3251824	0.2202462	0.2452905
102	0.56229665	0.8903612	0.9802498	1.6876832	1.54925542	0.6401264	0.3251824	0.2202462	0.2452905
103	1.35364734	1.3883365	1.2086227	0.9044078	0.47718145	0.4139011	0.3251824	0.2202462	0.2452905
104	1.10041512	1.8863118	0.9802498	1.6876832	-0.33499580	0.8663518	0.3251824	0.2202462	0.2452905
105	0.56229665	1.1539952	0.9802498	0.6537597	1.54925542	1.3470806	0.3251824	0.2202462	0.2452905
106	0.81552887	1.6226778	0.7518769	1.4370351	0.73707817	0.4139011	0.3251824	0.2202462	0.2452905
107	1.10041512	1.1539952	0.7518769	0.9044078	1.28935870	0.8663518	0.3251824	0.2202462	0.2452905
108	1.10041512	0.8903612	0.9802498	0.9044078	0.73707817	1.3470806	0.3251824	0.2202462	0.2452905
109	0.30906443	0.8903612	1.6939151	0.1211325	1.54925542	1.1208553	0.3251824	0.2202462	0.2452905
110	1.35364734	0.8903612	1.2086227	1.1863870	0.73707817	0.6401264	0.3251824	0.2202462	0.2452905
111	0.56229665	1.1539952	1.2086227	0.6537597	0.73707817	1.3470806	0.3251824	0.2202462	0.2452905

```

> subset(preProcesscoffee, kr$cluster == 4)

```

	Aroma	Flavor	Aftertaste	Acidity	Body	Balance	Uniformity	CleanCup	Sweetness
353	0.02417819	0.15804465	-0.6754539	-1.1634390	-1.69945358	-1.48073618	0.3251824	0.2202462	0.2452905
380	0.30906443	0.42167862	-1.1607463	-1.4454181	-1.69945358	-1.48073618	0.3251824	0.2202462	0.2452905
637	-1.02040471	-0.80861325	-1.1607463	-1.4454181	-1.69945358	-1.48073618	0.3251824	0.2202462	0.2452905
736	-1.81175540	-1.30658853	-1.1607463	-1.1634390	-1.69945358	-1.48073618	0.3251824	0.2202462	0.2452905
815	-0.76717250	-1.54092984	-0.9323734	-0.3801637	-0.07509908	0.41390108	0.3251824	0.2202462	0.2452905
839	-1.27363693	0.15804465	-1.6460387	-0.1295156	0.73707817	-0.77378198	0.3251824	0.2202462	0.2452905
840	0.02417819	-1.54092984	-1.3891192	-1.9467143	0.47718145	-0.06682778	0.3251824	0.2202462	0.2452905
861	-1.81175540	-1.30658853	-0.2187080	-0.1295156	-0.07509908	-0.54755664	0.3251824	0.2202462	0.2452905
872	-0.48228625	-0.31063797	-0.9323734	-0.3801637	-0.62737961	-0.29305312	0.3251824	0.2202462	0.2452905
875	-1.27363693	-0.57427194	-0.2187080	-0.3801637	-0.07509908	-0.54755664	0.3251824	0.2202462	0.2452905
877	0.56229665	0.65601993	0.9802498	0.4031116	-7.38469435	0.86635177	0.3251824	0.2202462	0.2452905
880	-0.76717250	-0.57427194	0.2665844	-0.9127909	-0.33499580	-0.54755664	0.3251824	0.2202462	0.2452905
882	-1.27363693	-0.31063797	-0.2187080	-0.3801637	-0.62737961	-0.29305312	0.3251824	0.2202462	0.2452905
883	-0.22905403	-0.57427194	0.2665844	-0.1295156	-1.14717305	-1.00000733	0.3251824	0.2202462	0.2452905
884	-0.22905403	-1.04295456	-0.6754539	-1.1634390	-0.07509908	-1.25451084	0.3251824	0.2202462	0.2452905
912	-0.48228625	-0.31063797	-0.2187080	-0.6621428	-0.88727633	-0.54755664	0.3251824	0.2202462	0.2452905
913	-0.22905403	-0.07629666	-0.2187080	-0.3801637	-1.14717305	-0.77378198	0.3251824	0.2202462	0.2452905
916	-1.02040471	-0.31063797	-0.4470809	-0.1295156	-0.62737961	-0.54755664	0.3251824	0.2202462	0.2452905
917	0.81552887	-0.57427194	-0.9323734	-1.4454181	-0.62737961	-0.06682778	0.3251824	0.2202462	0.2452905
922	-0.48228625	-0.57427194	-0.4470809	-0.3801637	-0.33499580	-0.54755664	0.3251824	0.2202462	0.2452905
923	-0.22905403	-0.31063797	-0.4470809	-0.9127909	-0.62737961	-0.54755664	0.3251824	0.2202462	0.2452905
928	-0.48228625	-0.31063797	-0.2187080	-1.1634390	-0.88727633	-0.29305312	0.3251824	0.2202462	0.2452905
932	-0.76717250	-0.07629666	-0.2187080	-0.3801637	-1.14717305	-0.54755664	0.3251824	0.2202462	0.2452905
935	0.02417819	-0.57427194	-0.9323734	-0.6621428	-0.88727633	-0.54755664	0.3251824	0.2202462	0.2452905
936	-0.48228625	-0.57427194	-0.9323734	0.1211325	-0.62737961	-0.77378198	0.3251824	0.2202462	0.2452905
937	-0.76717250	-0.07629666	0.2665844	-0.6621428	-2.25173412	-0.29305312	0.3251824	0.2202462	0.2452905
940	0.02417819	-1.04295456	-0.2187080	-0.3801637	-1.43955686	-0.29305312	0.3251824	0.2202462	0.2452905
941	-1.02040471	-0.07629666	0.0382115	-1.1634390	-0.62737961	-0.54755664	0.3251824	0.2202462	0.2452905
943	0.30906443	-0.57427194	-0.4470809	-1.1634390	-0.88727633	-0.29305312	0.3251824	0.2202462	0.2452905
946	-0.48228625	-0.57427194	-0.4470809	-1.1634390	-0.88727633	-0.29305312	0.3251824	0.2202462	0.2452905
949	0.56229665	-0.57427194	0.0382115	-1.6960662	-0.62737961	-0.77378198	0.3251824	0.2202462	0.2452905
951	-0.22905403	-0.57427194	-0.2187080	-1.1634390	-0.62737961	-0.29305312	0.3251824	0.2202462	0.2452905
952	-0.22905403	-0.07629666	-1.1607463	-1.1634390	-0.07509908	-0.54755664	0.3251824	0.2202462	0.2452905
954	-1.02040471	-0.57427194	-0.2187080	-0.6621428	-0.62737961	-0.29305312	0.3251824	0.2202462	0.2452905
955	0.02417819	-0.31063797	-0.6754539	-0.3801637	-1.14717305	-1.00000733	0.3251824	0.2202462	0.2452905
956	-1.27363693	-0.57427194	0.0382115	-0.9127909	-0.33499580	-0.54755664	0.3251824	0.2202462	0.2452905
957	-1.55852318	-0.57427194	0.0382115	-0.6621428	-0.62737961	-0.06682778	0.3251824	0.2202462	0.2452905
958	0.56229665	-0.57427194	-0.9323734	-0.6621428	-0.62737961	-0.77378198	0.3251824	0.2202462	0.2452905
959	-0.76717250	-0.57427194	-0.6754539	-0.6621428	-0.07509908	-0.54755664	0.3251824	0.2202462	0.2452905
960	0.02417819	-0.31063797	-0.4470809	-0.6621428	-0.88727633	-0.77378198	0.3251824	0.2202462	0.2452905
961	-0.76717250	-0.80861325	-0.6754539	-0.3801637	0.18479764	-0.54755664	0.3251824	0.2202462	0.2452905
964	-1.27363693	-0.31063797	-0.6754539	-1.4454181	-1.95935031	1.82780949	0.3251824	0.2202462	0.2452905
965	-1.55852318	-0.07629666	-0.2187080	-1.1634390	-0.33499580	-0.77378198	0.3251824	0.2202462	0.2452905
966	-0.48228625	-1.04295456	-0.6754539	-0.1295156	-0.62737961	-0.77378198	0.3251824	0.2202462	0.2452905
967	-0.76717250	-0.31063797	-0.6754539	-0.9127909	-0.33499580	-0.54755664	0.3251824	0.2202462	0.2452905

968	-0.76717250	-1.04295456	-0.6754539	-0.1295156	-0.88727633	-0.29305312	0.3251824	0.2202462	0.2452905
969	-0.22905403	-0.07629666	-0.4470809	-0.9127909	-0.88727633	-0.77378198	0.3251824	0.2202462	0.2452905
970	-0.22905403	-0.57427194	0.2665844	0.1211325	-1.69945358	-0.77378198	0.3251824	0.2202462	0.2452905
972	-0.76717250	-0.07629666	-0.6754539	-1.1634390	0.18479764	-1.00000733	0.3251824	0.2202462	0.2452905
973	0.30906443	-1.04295456	-0.6754539	-0.6621428	-0.07509908	-1.00000733	0.3251824	0.2202462	0.2452905
974	0.56229665	0.15804465	-0.9323734	-0.9127909	-1.43955686	-0.77378198	0.3251824	0.2202462	0.2452905
975	-0.22905403	-0.07629666	-0.2187080	-1.1634390	-1.43955686	-1.00000733	0.3251824	0.2202462	0.2452905
976	-0.48228625	-0.57427194	-1.1607463	-0.3801637	-0.62737961	-0.54755664	0.3251824	0.2202462	0.2452905
977	-2.34987386	-0.31063797	0.0382115	-1.1634390	-0.07509908	-0.29305312	0.3251824	0.2202462	0.2452905
978	-1.02040471	-0.80861325	-0.4470809	-1.4454181	0.18479764	-0.06682778	0.3251824	0.2202462	0.2452905
979	0.02417819	-1.04295456	-0.6754539	-0.6621428	-0.62737961	-0.77378198	0.3251824	0.2202462	0.2452905
980	0.02417819	-0.80861325	-0.2187080	-1.4454181	-0.62737961	-0.54755664	0.3251824	0.2202462	0.2452905
981	-0.76717250	-1.30658853	-0.2187080	-0.9127909	-0.62737961	0.15939757	0.3251824	0.2202462	0.2452905
982	-0.48228625	-0.31063797	-0.6754539	-0.3801637	-0.62737961	-0.77378198	0.3251824	0.2202462	0.2452905
983	-0.48228625	-0.80861325	-0.9323734	0.6537597	-0.33499580	-1.00000733	0.3251824	0.2202462	0.2452905
984	-0.48228625	-0.80861325	-0.2187080	-0.9127909	-0.88727633	-0.29305312	0.3251824	0.2202462	0.2452905
985	-0.48228625	-0.57427194	-0.4470809	-0.3801637	-0.07509908	-1.00000733	0.3251824	0.2202462	0.2452905
987	-0.48228625	-0.57427194	-0.6754539	-0.6621428	-0.33499580	-1.00000733	0.3251824	0.2202462	0.2452905
989	-0.76717250	-0.57427194	-0.2187080	-0.3801637	-0.62737961	-0.54755664	0.3251824	0.2202462	0.2452905
990	-0.48228625	-1.04295456	-0.2187080	-0.3801637	-0.88727633	-0.54755664	0.3251824	0.2202462	0.2452905
991	-0.48228625	-0.57427194	-0.2187080	-0.9127909	-0.88727633	-0.54755664	0.3251824	0.2202462	0.2452905
992	-0.48228625	-0.80861325	-0.6754539	-0.9127909	-0.88727633	0.15939757	0.3251824	0.2202462	0.2452905
994	-1.02040471	-0.31063797	-0.2187080	-0.6621428	-1.14717305	-0.54755664	0.3251824	0.2202462	0.2452905
995	-0.76717250	-0.80861325	-0.2187080	-0.3801637	-0.88727633	-0.29305312	0.3251824	0.2202462	0.2452905
996	-0.76717250	-0.57427194	-0.4470809	-0.3801637	-0.88727633	-0.77378198	0.3251824	0.2202462	0.2452905
997	-0.48228625	-0.57427194	-0.4470809	-0.3801637	-1.14717305	-0.77378198	0.3251824	0.2202462	0.2452905
998	-0.48228625	-0.31063797	-0.4470809	0.1211325	-1.95935031	-1.00000733	0.3251824	0.2202462	0.2452905
999	-1.02040471	-0.80861325	-0.6754539	-0.1295156	-0.62737961	-1.00000733	0.3251824	0.2202462	0.2452905
1000	-0.48228625	-1.30658853	-0.9323734	-0.6621428	-0.88727633	0.15939757	0.3251824	0.2202462	0.2452905
1001	-0.76717250	-1.04295456	-0.6754539	0.4031116	-0.62737961	-1.00000733	0.3251824	0.2202462	0.2452905
1002	-0.48228625	-0.80861325	-0.6754539	-0.1295156	-0.88727633	-1.00000733	0.3251824	0.2202462	0.2452905
1003	0.30906443	-0.07629666	0.0382115	-7.1789934	-0.62737961	0.64012642	0.3251824	0.2202462	0.2452905
1004	-0.22905403	-0.80861325	-1.1607463	-0.9127909	-0.07509908	-0.77378198	0.3251824	0.2202462	0.2452905
1005	-1.02040471	-1.04295456	-0.6754539	-0.9127909	-0.07509908	-0.54755664	0.3251824	0.2202462	0.2452905
1006	-0.48228625	-0.80861325	-0.4470809	-0.6621428	-0.88727633	-1.00000733	0.3251824	0.2202462	0.2452905
1008	-0.76717250	-0.57427194	-0.9323734	-0.3801637	-1.14717305	-0.54755664	0.3251824	0.2202462	0.2452905
1009	-0.76717250	-1.77527114	-0.6754539	-0.6621428	-0.33499580	-0.29305312	0.3251824	0.2202462	0.2452905
1010	-0.22905403	-1.77527114	-0.2187080	-0.6621428	-0.62737961	-0.54755664	0.3251824	0.2202462	0.2452905
1011	-0.76717250	-0.31063797	-0.4470809	-0.3801637	-0.88727633	-1.00000733	0.3251824	0.2202462	0.2452905
1012	-1.55852318	-0.57427194	-0.2187080	-0.6621428	-0.33499580	-0.77378198	0.3251824	0.2202462	0.2452905
1014	-0.76717250	-1.04295456	-0.4470809	-0.6621428	-0.62737961	-0.77378198	0.3251824	0.2202462	0.2452905
1018	-1.27363693	-0.31063797	-1.1607463	0.1211325	-0.62737961	-1.48073618	0.3251824	0.2202462	0.2452905
1019	-0.48228625	-1.30658853	-0.6754539	-1.1634390	0.18479764	-0.77378198	0.3251824	0.2202462	0.2452905
1020	-0.76717250	-1.04295456	-0.6754539	-0.9127909	-0.33499580	-0.54755664	0.3251824	0.2202462	0.2452905
1021	-1.02040471	-0.57427194	-0.4470809	-0.9127909	-0.88727633	-0.54755664	0.3251824	0.2202462	0.2452905
1022	1.10041512	0.42167862	-0.2187080	-1.4454181	-1.43955686	-1.25451084	0.3251824	0.2202462	0.2452905
1023	-1.27363693	-0.31063797	-0.6754539	-0.9127909	-0.88727633	-0.54755664	0.3251824	0.2202462	0.2452905
1024	-0.76717250	-1.30658853	-0.4470809	-1.6960662	-0.07509908	-0.29305312	0.3251824	0.2202462	0.2452905
1025	0.02417819	-0.31063797	-1.3891192	-0.1295156	-0.88727633	-1.00000733	0.3251824	0.2202462	0.2452905
1026	-0.48228625	-0.80861325	-0.4470809	-0.9127909	-0.88727633	-0.77378198	0.3251824	0.2202462	0.2452905
1027	-1.02040471	-1.04295456	-0.9323734	-1.4454181	-1.43955686	1.57330597	0.3251824	0.2202462	0.2452905
1028	-0.48228625	-0.80861325	-0.4470809	-0.9127909	-0.88727633	-0.77378198	0.3251824	0.2202462	0.2452905
1029	0.02417819	-0.57427194	-0.9323734	-0.1295156	-1.14717305	-0.77378198	0.3251824	0.2202462	0.2452905
1030	0.30906443	0.15804465	-0.2187080	-0.6621428	-1.14717305	-1.48073618	0.3251824	0.2202462	0.2452905
1031	-0.48228625	-0.57427194	-0.6754539	-1.1634390	-1.14717305	-0.77378198	0.3251824	0.2202462	0.2452905
1032	0.30906443	-0.80861325	-0.9323734	-0.6621428	-1.14717305	-0.77378198	0.3251824	0.2202462	0.2452905
1033	-1.27363693	-1.30658853	-0.4470809	-0.3801637	-0.88727633	-0.54755664	0.3251824	0.2202462	0.2452905
1034	-0.76717250	-0.80861325	-0.2187080	-0.6621428	-0.88727633	-1.00000733	0.3251824	0.2202462	0.2452905
1035	-0.22905403	-1.04295456	-0.2187080	-0.9127909	-1.14717305	-1.00000733	0.3251824	0.2202462	0.2452905
1036	-0.48228625	-0.57427194	-1.1607463	-0.9127909	-0.62737961	-0.06682778	0.3251824	0.2202462	0.2452905
1037	-0.76717250	-0.57427194	-1.1607463	-2.2286934	-0.62737961	0.41390108	0.3251824	0.2202462	0.2452905
1038	-1.27363693	-0.80861325	-0.9323734	-1.1634390	-0.88727633	0.15939757	0.3251824	0.2202462	0.2452905
1039	-2.06498762	-0.07629666	-1.1607463	0.1211325	-0.07509908	-1.25451084	0.3251824	0.2202462	0.2452905
1040	-0.22905403	-0.57427194	-1.1607463	-0.3801637	-1.14717305	-1.00000733	0.3251824	0.2202462	0.2452905
1041	-1.02040471	-0.80861325	-0.6754539	-0.9127909	-0.62737961	-0.77378198	0.3251824	0.2202462	0.2452905
1042	-0.22905403	-0.57427194	-1.1607463	-0.3801637	-0.88727633	-1.00000733	0.3251824	0.2202462	0.2452905

In seguito, si rappresentano graficamente i quattro cluster formati:

```
fviz_cluster(object = kr,
             data = preProcesscoffee)
```

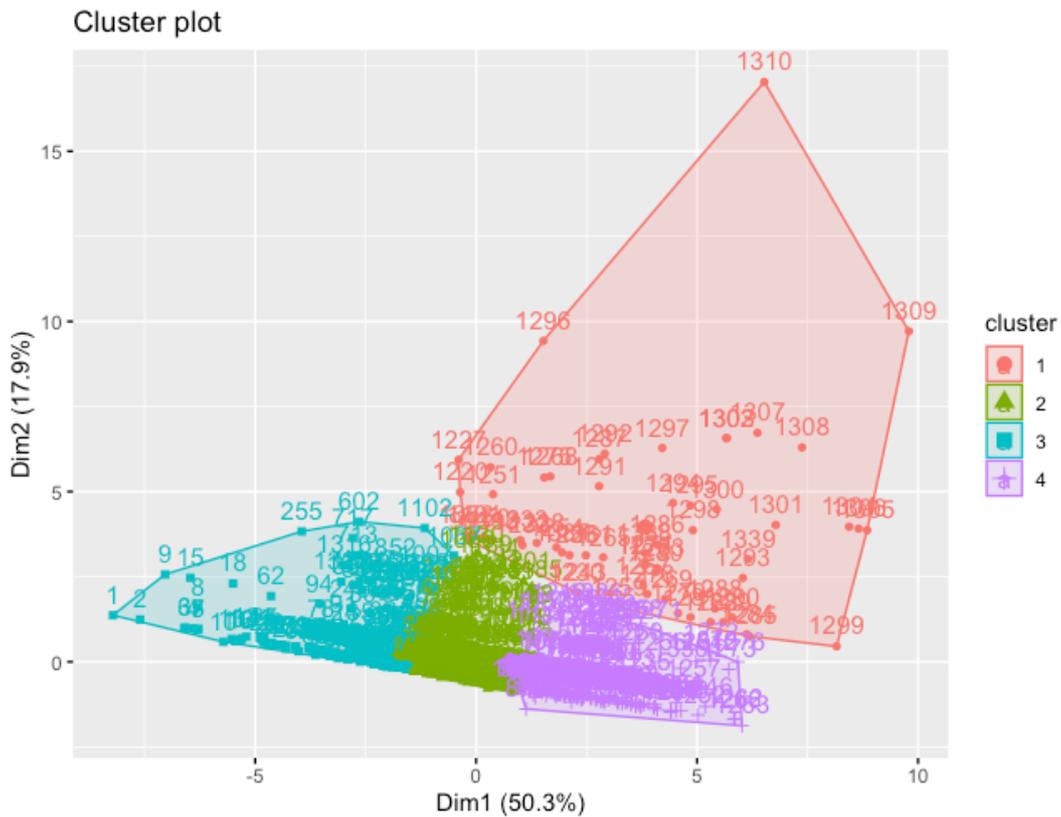


Fig. 24. Suddivisione delle osservazioni in cluster tramite l'algoritmo di clustering non gerarchico k-means

Si assegnano successivamente le osservazioni ai vari cluster:

```
cluster_1 <- subset(preProcesscoffee, kr$cluster == 1)
cluster_2 <- subset(preProcesscoffee, kr$cluster == 2)
cluster_3 <- subset(preProcesscoffee, kr$cluster == 3)
cluster_4 <- subset(preProcesscoffee, kr$cluster == 4)
```

Si richiede poi al software il riassunto delle principali caratteristiche a livello statistico dei quattro cluster formati:

summary(cluster\_1)

```
> summary(cluster_1)
      Aroma      Flavor      Aftertaste      Acidity      Body      Balance      Uniformity
Min.   :-3.9326  Min.   :-3.9722  Min.   :-3.5301  Min.   :-4.0459  Min.   :-3.8761  Min.   :-4.0823  Min.   :-7.9191
1st Qu.:-1.8118  1st Qu.:-2.2732  1st Qu.:-2.1028  1st Qu.:-1.6961  1st Qu.:-1.1472  1st Qu.:-2.1877  1st Qu.:-3.7970
Median :-1.0204  Median :-1.5409  Median :-1.6460  Median :-0.9128  Median :-0.6274  Median :-1.4807  Median :-3.7970
Mean   :-1.0323  Mean   :-1.4415  Mean   :-1.3938  Mean   :-0.9328  Mean   :-0.7131  Mean   :-1.3829  Mean   :-3.0303
3rd Qu.:-0.2291  3rd Qu.:-0.5743  3rd Qu.:-0.2187  3rd Qu.:-0.1295  3rd Qu.:-0.0751  3rd Qu.:-0.5476  3rd Qu.:-1.0557
Max.   : 1.3536  Max.   : 0.6560  Max.   : 0.9802  Max.   : 1.1864  Max.   : 1.2894  Max.   : 1.1209  Max.   : 0.3252

      CleanCup      Sweetness
Min.   :-13.7600  Min.   :-15.3987
1st Qu.: -4.4352  1st Qu.: -3.3635
Median :-2.5758  Median :  0.2453
Mean   :-3.1863  Mean   :-1.7904
3rd Qu.: -1.6391  3rd Qu.:  0.2453
Max.   :  0.2202  Max.   :  0.2453
```

summary(cluster\_2)

```
> summary(cluster_2)
      Aroma      Flavor      Aftertaste      Acidity      Body      Balance      Uniformity
Min.   :-7.88933  Min.   :-1.54093  Min.   :-1.38912  Min.   :-2.47934  Min.   :-2.51163  Min.   :-1.96147  Min.   :-3.7970
1st Qu.:-0.22905  1st Qu.:-0.07630  1st Qu.:-0.21871  1st Qu.:-0.38016  1st Qu.:-0.33500  1st Qu.:-0.29305  1st Qu.:  0.3252
Median : 0.02418  Median : 0.15804  Median : 0.03821  Median : 0.12113  Median :-0.07510  Median :-0.06683  Median : 0.3252
Mean   : 0.04482  Mean   : 0.09789  Mean   : 0.08001  Mean   : 0.01232  Mean   : 0.05601  Mean   : 0.10571  Mean   : 0.1935
3rd Qu.: 0.30906  3rd Qu.: 0.42168  3rd Qu.: 0.49496  3rd Qu.: 0.40311  3rd Qu.: 0.47718  3rd Qu.: 0.41390  3rd Qu.: 0.3252
Max.   : 1.89177  Max.   : 1.88631  Max.   : 1.20862  Max.   : 1.96966  Max.   : 2.62133  Max.   : 2.98721  Max.   : 0.3252

      CleanCup      Sweetness
Min.   :-4.4352  Min.   :-5.7633
1st Qu.: 0.2202  1st Qu.: 0.2453
Median : 0.2202  Median : 0.2453
Mean   : 0.1784  Mean   : 0.1255
3rd Qu.: 0.2202  3rd Qu.: 0.2453
Max.   : 0.2202  Max.   : 0.2453
```

summary(cluster\_3)

```
> summary(cluster_3)
      Aroma      Flavor      Aftertaste      Acidity      Body      Balance      Uniformity
Min.   :-1.0204  Min.   :-0.3106  Min.   :-0.4471  Min.   :-1.1634  Min.   :-1.1472  Min.   :-0.5476  Min.   :-6.5382
1st Qu.: 0.5623  1st Qu.: 0.6560  1st Qu.: 0.7519  1st Qu.: 0.6538  1st Qu.: 0.4772  1st Qu.: 0.6401  1st Qu.: 0.3252
Median : 0.8155  Median : 0.8904  Median : 0.9802  Median : 0.9044  Median : 0.9970  Median : 0.8664  Median : 0.3252
Mean   : 1.0107  Mean   : 1.1091  Mean   : 1.1170  Mean   : 1.1272  Mean   : 1.0034  Mean   : 1.0536  Mean   : 0.1376
3rd Qu.: 1.3536  3rd Qu.: 1.3883  3rd Qu.: 1.4655  3rd Qu.: 1.4370  3rd Qu.: 1.2894  3rd Qu.: 1.3471  3rd Qu.: 0.3252
Max.   : 3.7277  Max.   : 3.8196  Max.   : 3.6065  Max.   : 3.7869  Max.   : 3.4335  Max.   : 3.4679  Max.   : 0.3252

      CleanCup      Sweetness
Min.   :-4.4352  Min.   :-5.76330
1st Qu.: 0.2202  1st Qu.: 0.24529
Median : 0.2202  Median : 0.24529
Mean   : 0.1125  Mean   :-0.08468
3rd Qu.: 0.2202  3rd Qu.: 0.24529
Max.   : 0.2202  Max.   : 0.24529
```

summary(cluster\_4)

```
> summary(cluster_4)
      Aroma      Flavor      Aftertaste      Acidity      Body      Balance      Uniformity
Min.   :-4.4390 Min.   :-4.2359 Min.   :-3.5301 Min.   :-7.1790 Min.   :-7.9370 Min.   :-6.4294 Min.   :-3.79698
1st Qu.:-1.2736 1st Qu.:-1.3066 1st Qu.:-1.3891 1st Qu.:-1.4454 1st Qu.:-1.2203 1st Qu.:-1.4807 1st Qu.: 0.32518
Median :-0.7672 Median :-1.0430 Median :-0.9324 Median :-0.9128 Median :-0.8873 Median :-1.0000 Median : 0.32518
Mean   :-0.8990 Mean   :-1.0320 Mean   :-1.0099 Mean   :-0.9635 Mean   :-0.9805 Mean   :-1.0056 Mean   : 0.04436
3rd Qu.:-0.4823 3rd Qu.:-0.5743 3rd Qu.:-0.4471 3rd Qu.:-0.3802 3rd Qu.:-0.6274 3rd Qu.:-0.5476 3rd Qu.: 0.32518
Max.   : 1.3536 Max.   : 0.6560 Max.   : 0.9802 Max.   : 0.6538 Max.   : 0.7371 Max.   : 1.8278 Max.   : 0.32518
CleanCup      Sweetness
Min.   :-3.5125 Min.   :-3.8146
1st Qu.: 0.2202 1st Qu.: 0.2453
Median : 0.2202 Median : 0.2453
Mean   : 0.1341 Mean   : 0.1670
3rd Qu.: 0.2202 3rd Qu.: 0.2453
Max.   : 0.2202 Max.   : 0.2453
```

Infine, si guardano per ogni cluster i punteggi medi attribuiti a ciascun attributo:

```
preProcesscoffee %>%
  mutate (cluster = kr$cluster) %>%
  group_by (cluster) %>%
  summarise (Aroma = mean (Aroma),
            Aftertaste = mean (Aftertaste),
            Balance = mean (Balance),
            Uniformity = mean (Uniformity),
            CleanCup = mean (CleanCup),
            Sweetness = mean (Sweetness),
            Flavor = mean (Flavor),
            Body = mean (Body),
            Acidity = mean (Acidity))

# A tibble: 4 × 10
  cluster Aroma Aftertaste Balance Uniformity CleanCup Sweetness Flavor Body Acidity
  <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     1 -1.03 -1.39 -1.38 -3.03 -3.19 -1.79 -1.44 -0.713 -0.933
2     2 0.0448 0.0800 0.106 0.194 0.178 0.126 0.0979 0.0560 0.0123
3     3 1.01 1.12 1.05 0.138 0.112 -0.0847 1.11 1.00 1.13
4     4 -0.899 -1.01 -1.01 0.0444 0.134 0.167 -1.03 -0.980 -0.963
```

### 3.4 DISCUSSIONE DEI RISULTATI

Nella sezione precedente si è mostrato come è stata effettuata l'analisi per le due tipologie di algoritmi di clustering sulle valutazioni date dagli esperti del Coffe Quality Institute (CQI) a varie tipologie di chicchi di caffè sulla base dei seguenti attributi: "Aroma", "Aftertaste", "Balance", "Uniformity", "CleanCup", "Sweetness", "Flavor", "Body" e "Acidity".

In primo luogo, è stata svolta l'analisi di clustering gerarchico. In seguito all'eliminazione dell'*outlier* individuato, il dataset conteneva un totale di 1.338 osservazioni. Come suggerito dall'Elbow Method queste osservazioni sono state suddivise in 4 cluster, il primo con 321 osservazioni, il secondo con 184 osservazioni, il terzo con 488 osservazioni e il quarto con 345 osservazioni. Considerando il valore medio dei rating degli attributi per ogni cluster, il primo risulta avere un punteggio medio maggiore nei seguenti attributi: aroma, flavor, aftertaste, acidity, body e balance. Il terzo cluster, invece, registra un punteggio medio più alto per gli attributi uniformity, cleancup e sweetness. Questa analisi può risultare utile ai brand di caffè per capire, in base a che caffè utilizzano nelle loro miscele, su quali attributi possono puntare nella loro comunicazione e quindi quali di essi possono utilizzare come elementi di differenziazione. In questo caso, gli esperti valutano i caffè appartenenti al cluster 1 come migliori in termini di aroma, sapore, retrogusto, acidità, corposità ed equilibrio e sinergia tra gli attributi sapore, retrogusto, acidità e corposità. Giudicano invece migliore il cluster 3 per la dolcezza del caffè, l'uniformità intesa come consistenza dell'aroma delle diverse tazze e la *cup cleanliness* intesa come caratteristica tipica di un caffè privo di macchie e difetti, quindi privo di note dure e terrose. Di conseguenza, considerando i punteggi dati dagli esperti agli attributi, l'importanza che questi attributi hanno per i consumatori, e l'effetto che l'informazione soggettiva ha sui consumatori in termini di preferenza per il prodotto e aumento dell'intenzione d'acquisto, un brand che si posiziona nel primo cluster nella comunicazione del proprio caffè potrebbe mettere in luce un buon aroma e odore quando viene infuso con l'acqua calda, un sapore ottimo quando entra in contatto con la lingua, un retrogusto piacevole in modo che rimanga in bocca un buon sapore per un tempo prolungato, un caffè acido al punto giusto che rende l'esperienza di berlo molto piacevole al palato e infine un caffè che può essere considerato corposo. Inoltre, possono far riferimento al fatto che il loro caffè vanta una perfetta sinergia tra gli aromi presenti. Coloro che invece si posizionano nel terzo cluster possono vantare un caffè dolce, con una buona consistenza dell'aroma in diverse tazze così che la qualità del caffè sia garantita anche nella degustazione delle tazze successive e un caffè ben conservato, privo di note dure, terrose e sgradevoli che possano avere un effetto negativo tra il primo sorso e i successivi.

Dall'analisi di clustering non gerarchico con il metodo del k-means, risulta che il cluster 2 ha un punteggio medio maggiore per uniformità intesa come consistenza dell'aroma

delle diverse tazze per garantire una qualità costante e *cup cleanliness* intesa come un caffè privo di macchie e difetti, si riferisce ai sapori, in particolare privi di note dure e terrose. Il cluster 3 si differenzia per il suo aroma garantendo un buonissimo odore nel momento in cui il caffè viene infuso con l'acqua calda. Inoltre, altri elementi di differenziazione per questi caffè sono il retrogusto, l'equilibrio tra sapore, retrogusto, acidità e corposità in modo che formino una combinazione sinergica, sapore, corposità e acidità e il cluster 4 per la dolcezza del caffè. In caso di clustering non gerarchico i brand che si posizionano nel cluster 2 nella loro comunicazione possono vantare come punto di forza del loro caffè una buona uniformità, quindi una buona consistenza dell'aroma in diverse tazze così da garantire un caffè di qualità a ogni assaggio e un caffè pulito, privo di macchie e difetti e privo di note dure e terrose. I brand del cluster 3 possono invece vantare un caffè con un aroma unico nel momento in cui viene infuso con acqua calda, un retrogusto piacevole al palato e durevole così che questa gradevolezza sia prolungata, un gusto percepito unico e piacevole, una corposità gradevole al palato e un'acidità piacevole e delicata. Inoltre, si possono differenziare grazie a un equilibrio perfetto tra sapore, retrogusto, acidità e corposità in modo da formare una sinergia speciale tra questi attributi ritenuti importanti da parte del consumatore. Infine, i brand appartenenti al quarto cluster possono concentrarsi sulla dolcezza del caffè senza la presenza di note troppo amare e sgradevoli.

## CONCLUSIONI

Il presente lavoro di tesi ha contribuito a fornire una spiegazione teorica di cos'è il Machine Learning e l'utilità che esso può avere per le aziende. Vengono successivamente spiegati i motivi del consumo di caffè da parte dei consumatori e l'oggettività che gli algoritmi di Machine Learning possono portare nell'analisi della qualità dei prodotti per giustificare le implicazioni manageriali della ricerca. Il lavoro si concentra prevalentemente sul settore alimentare, vengono infatti successivamente illustrati vari esempi di applicazioni del Machine Learning nell'ambito delle bevande calde per testarne la qualità. L'ultimo capitolo si pone l'obiettivo di indagare sul dataset di riferimento gli attributi più rilevanti per le tipologie di chicchi di caffè prese in analisi tramite l'utilizzo dei due principali algoritmi di clustering: il clustering gerarchico e il k-means. Il dataset di riferimento è Coffee Quality database from CQI, che raccoglie i rating dati da esperti del Coffee Quality Institute su vari attributi legati proprio al caffè. Tramite questa analisi è stato possibile clusterizzare questi attributi sulla base dei rating per andare a capire i punti di forza di ognuno di questi chicchi di caffè. Questa analisi può essere d'aiuto ai brand per ottimizzare la loro comunicazione andando a evidenziare gli attributi dove sono più forti e utilizzarli come vantaggio competitivo sulla concorrenza. Ciò è possibile perché i consumatori si fidano delle opinioni degli esperti e perché questi attributi sono ritenuti estremamente importanti dai consumatori e determinanti per la loro scelta di acquistare e consumare un determinato brand di caffè rispetto ad altri. Le osservazioni sono state suddivise in quattro cluster, ognuno dei quali contenente gli attributi che hanno riportato un punteggio medio maggiore nell'analisi. I consumatori di oggi sono sempre più informati grazie all'evento della tecnologia ed è sempre più facile per loro comparare in modo rapido ed efficace i vari brand di un determinato prodotto. Da qui nasce l'importanza per le aziende di avvicinarsi ai consumatori utilizzando gli attributi che cercano e che ritengono più importanti come punto di forza. Infatti, per valutare i prodotti, la qualità rimane un attributo fondamentale e determinante, ecco perché l'utilizzo di questi attributi può fornire un vantaggio che i brand possono sfruttare per ottenere la fiducia del consumatore. Dall'algoritmo di clustering gerarchico emerge che i cluster da prendere come riferimento sono il primo e il terzo. Nello specifico, quelli appartenenti al primo cluster hanno ottenuto punteggi più alti negli attributi aroma, odore e sapore in seguito al processo di infusione con acqua calda, un ottimo sapore, retrogusto, acidità al punto giusto e corposità. Inoltre, questi caffè sono caratterizzati da un'ottima sinergia tra

gli aromi presenti. I caffè appartenenti al terzo cluster, invece, sono caratterizzati da una dolcezza al punto giusto, con una buona consistenza dell'aroma in diverse tazze in modo che la qualità del caffè sia più duratura e mantenuta anche nelle tazze successive, una buona conservazione del caffè, privo di note sgradevoli al gusto. Dall'analisi di clustering non gerarchico, invece, emerge che i cluster di riferimento sono il secondo, il terzo e il quarto. Nello specifico, i caffè appartenenti al secondo cluster sono ritenuti più forti nell'uniformità e dall'assenza di note terrose e sgradevoli. Quelli del terzo cluster si differenziano per l'aroma, il retrogusto piacevole e durevole, il sapore e un buon grado di corposità e acidità. Inoltre, questi caffè presentano un equilibrio perfetto tra sapore, retrogusto, acidità e corposità. Infine, i caffè appartenenti al quarto cluster presentano il giusto grado di dolcezza.

## ESTENSIONI FUTURE

Il presente elaborato punta a trovare un modo per ottimizzare la comunicazione dei brand di caffè sulla base delle preferenze dei consumatori tramite l'analisi di attributi prettamente correlati alla qualità. Il dataset utilizzato però lascia spazio a più algoritmi che possono essere utilizzati per costruire ulteriori analisi. Si potrebbe infatti applicare un altro algoritmo al dataset di riferimento, come una regressione lineare. Nello specifico, prendendo gli attributi come variabili indipendenti si potrebbe costruire una retta di regressione e stimare l'intenzione di acquisto come variabile dipendente per ogni tipologia di caffè.

La presente ricerca è stata svolta su una specifica categoria di prodotto, il caffè. Ulteriori ricerche potrebbero sviluppare il medesimo studio applicato ad altri dataset contenenti punteggi relativi ad attributi legati ad altre tipologie di prodotti. Un esempio può essere rappresentato dai cibi surgelati. La credenza maggiore su questi prodotti è che siano veloci da preparare ma che non siano per nulla salutari. Bisogna però sapere che non è così per tutti e un'analisi di qualità su alcuni di questi prodotti potrebbe fornire ai brand una maniera per dare ai consumatori maggiore visibilità degli effetti benefici che questi prodotti possono avere.

Inoltre, un'estensione futura per la presente ricerca riguarda la salubrità del caffè. In particolare, questo prodotto può essere consumato spesso anche in momenti di convivialità con amici e parenti. C'è però un gran numero di persone che pensa che il caffè abbia effetti nocivi sulla salute e quindi preferisce evitarne il consumo. Ciò che non si sa è che il caffè può avere delle proprietà benefiche sulla salute delle persone se consumato con moderazione. Potrebbe essere interessante approfondire questo tema e che i brand diano ai consumatori maggiore visibilità anche su questo. Potrebbe essere infatti una maniera per accrescere la loro customer base e fidelizzare maggiormente i clienti esistenti.

Infine, dalla ricerca sono emersi alcuni attributi legati al caffè che possono essere utilizzati dai brand nella loro comunicazione. Questa si potrebbe ampliare andando ad analizzare quali mezzi di comunicazione sono ritenuti più efficaci per questo tipo di prodotto in modo da poter capire dove utilizzare questi attributi e i punti di forza che i brand possono avere su di essi.

## ELENCO DELLE FIGURE

1. Relazione tra Intelligenza Artificiale, Machine Learning e Deep Learning
2. Componenti di un generico modello di Machine Learning
3. Tipologie di apprendimento del Machine Learning
4. Apprendimento supervisionato
5. Apprendimento non supervisionato
6. Apprendimento con rinforzo
7. Approcci di preelaborazione per garantire l'equità dei cluster
8. Rappresentazione del dendrogramma
9. Clustering k-means
10. Fattori che influenzano il comportamento del consumatore verso il caffè
11. Percezione dei consumatori sugli effetti benefici del caffè
12. Crescita nel mondo del tasso di popolarità dei vari algoritmi di Machine Learning
13. Capacità di predizione dei metodi di Machine Learning di rilevare l'adulterazione nel latte
14. L'impatto delle informazioni legate agli attributi sensoriali sul livello di preferenza dei consumatori
15. L'impatto delle informazioni legate agli attributi sensoriali sull'intenzione d'acquisto dei consumatori
16. Interfaccia linguaggio di programmazione R
17. Rappresentazione dell'arabica e della robusta
18. Struttura del dataset in seguito all'eliminazione delle variabili non utili ai fini dell'analisi
19. Struttura del dataset in seguito all'eliminazione dell'outlier
20. Rappresentazione del dendrogramma
21. Applicazione dell'Elbow Method
22. Dendrogramma con il taglio ad  $h=90.2$  per l'ottenimento dei quattro cluster
23. Riassunto delle variabili in seguito alla normalizzazione e standardizzazione
24. Suddivisione delle osservazioni in cluster tramite l'algoritmo di clustering non gerarchico k-means

## APPENDICE

### SCRIPT R:

```
library(dplyr)
library(ggplot2)
library(caret)
library(lattice)
library(factoextra)
library(NbClust)

coffee <- read.csv("merged_data_cleaned.csv")
str(coffee)
coffee

colnames(coffee) = c("ID", "Species", "Owner",
"CountryOfOrigin", "FarmName", "LotNumber", "Mill",
"ICONumber", "Company", "Altitude", "Region", "Producer",
"NumberOfBags", "BagWeight", "InCountryPartner",
"HarvestYear", "GradingDate", "Owner1", "Variety",
"ProcessingMethod", "Aroma", "Flavor", "Aftertaste",
"Acidity", "Body", "Balance", "Uniformity", "CleanCup",
"Sweetness", "CupperPoints", "TotalCupPoints", "Moisture",
"CategoryOneDefects", "Quakers", "Color",
"CategoryTwoDefects", "Expiration", "CertificationBody",
"CertificationAddress", "CertificationContact",
"UnitOfMeasurement", "AltitudeLowMeters",
"AltitudeHighMeters", "AltitudeMeanMeters")
str(coffee)

coffee$ID = NULL
coffee$Owner = NULL
coffee$FarmName = NULL
coffee$LotNumber = NULL
```

```
coffee$Mill = NULL
coffee$ICONumber = NULL
coffee$Company = NULL
coffee$Altitude = NULL
coffee$Region = NULL
coffee$Producer = NULL
coffee$NumberOfBags = NULL
coffee$BagWeight = NULL
coffee$InCountryPartner = NULL
coffee$HarvestYear = NULL
coffee$GradingDate = NULL
coffee$Owner1 = NULL
coffee$Variety = NULL
coffee$ProcessingMethod = NULL
coffee$Moisture = NULL
coffee$CategoryOneDefects = NULL
coffee$CategoryTwoDefects = NULL
coffee$Quakers = NULL
coffee$Expiration = NULL
coffee$CertificationBody = NULL
coffee$CertificationAddress = NULL
coffee$CertificationContact = NULL
coffee$UnitOfMeasurement = NULL
coffee$AltitudeLowMeters = NULL
coffee$AltitudeHighMeters = NULL
coffee$AltitudeMeanMeters = NULL
coffee$Species = NULL
coffee$CountryOfOrigin = NULL
coffee$Color = NULL
coffee$CupperPoints = NULL
coffee$TotalCupPoints = NULL
str(coffee)

coffee_new <- coffee[-1311,]
```

```
str(coffee_new)
```

```
###CLUSTERING GERARCHICO###
```

```
distances = dist(coffee_new[1:9], method = "euclidean")  
clusterCoffee = hclust(distances, method = "ward.D")  
plot(clusterCoffee)
```

```
fviz_nbclust(scale(dist(coffee_new)), kmeans, method =  
"wss") +  
  geom_vline(xintercept = 4, linetype = 2) +  
  labs(subtitle = "Elbow method")
```

```
clusterGroups = cutree(clusterCoffee, k = 4)  
clusterGroups[1:1000]  
subset(clusterCoffee$height, clusterCoffee$height>90.2)  
View(clusterGroups)
```

```
abline(h=90.2,col="pink",lty=2,lwd=3)
```

```
tapply(coffee_new$Aroma,clusterGroups,mean)  
tapply(coffee_new$Flavor,clusterGroups,mean)  
tapply(coffee_new$Aftertaste,clusterGroups,mean)  
tapply(coffee_new$Acidity,clusterGroups,mean)  
tapply(coffee_new$Body,clusterGroups,mean)  
tapply(coffee_new$Balance,clusterGroups,mean)  
tapply(coffee_new$Uniformity,clusterGroups,mean)  
tapply(coffee_new$CleanCup,clusterGroups,mean)  
tapply(coffee_new$Sweetness,clusterGroups,mean)
```

```
cluster1 = subset(coffee_new, clusterGroups==1)  
cluster2 = subset(coffee_new, clusterGroups==2)  
cluster3 = subset(coffee_new, clusterGroups==3)  
cluster4 = subset(coffee_new, clusterGroups==4)
```

```
str(cluster1)
str(cluster2)
str(cluster3)
str(cluster4)
```

```
cluster1 = subset(clusterCoffee, clusterGroups==1)
cluster2 = subset(clusterCoffee, clusterGroups==2)
cluster3 = subset(clusterCoffee, clusterGroups==3)
cluster4 = subset(clusterCoffee, clusterGroups==4)
```

### ###CLUSTERING K-MEANS###

```
preProcessParams = preProcess(coffee_new, method =
c("scale", "center"))
preProcessParams
preProcesscoffee = predict(preProcessParams, coffee_new)
```

```
sd(preProcesscoffee$Aroma)
sd(preProcesscoffee$Flavor)
sd(preProcesscoffee$Aftertaste)
sd(preProcesscoffee$Acidity)
sd(preProcesscoffee$Body)
sd(preProcesscoffee$Balance)
sd(preProcesscoffee$Uniformity)
sd(preProcesscoffee$CleanCup)
sd(preProcesscoffee$Sweetness)
sd(preProcesscoffee$CupperPoints)
sd(preProcesscoffee$TotalCupPoints)
summary(preProcesscoffee)
```

```
kr = kmeans (x = preProcesscoffee, centers = 4)
subset(preProcesscoffee, kr$cluster == 1)
subset(preProcesscoffee, kr$cluster == 2)
```

```
subset(preProcesscoffee, kr$cluster == 3)
subset(preProcesscoffee, kr$cluster == 4)
```

```
fviz_cluster(object = kr,
              data = preProcesscoffee)
```

```
kr$cluster
```

```
cluster_1 <- subset(preProcesscoffee, kr$cluster == 1)
cluster_2 <- subset(preProcesscoffee, kr$cluster == 2)
cluster_3 <- subset(preProcesscoffee, kr$cluster == 3)
cluster_4 <- subset(preProcesscoffee, kr$cluster == 4)
```

```
summary(cluster_1)
summary(cluster_2)
summary(cluster_3)
summary(cluster_4)
```

```
preProcesscoffee %>%
  mutate(cluster = kr$cluster) %>%
  group_by(cluster) %>%
  summarise(Aroma = mean(Aroma),
            Aftertaste = mean(Aftertaste),
            Balance = mean(Balance),
            Uniformity = mean(Uniformity),
            CleanCup = mean(CleanCup),
            Sweetness = mean(Sweetness),
            Flavor = mean(Flavor),
            Body = mean(Body),
            Acidity = mean(Acidity))
```

# RIASSUNTO

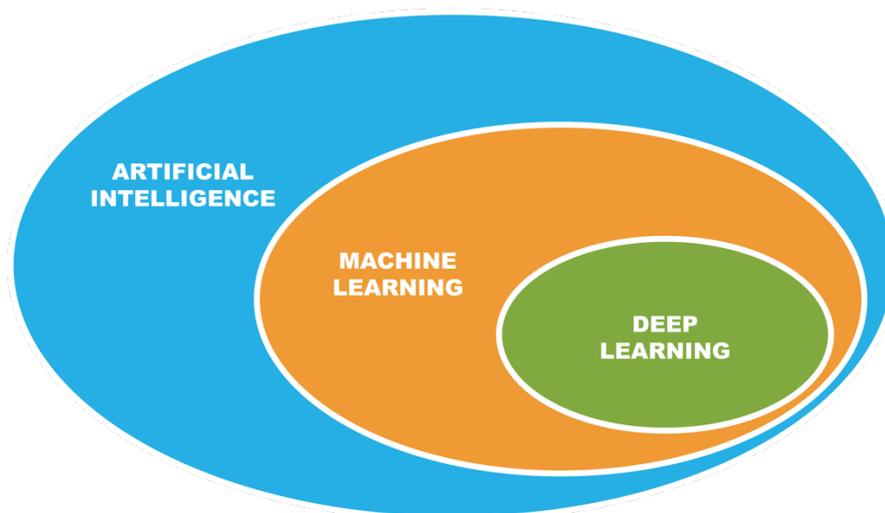
## INDICE

<i>CAPITOLO 1. INTRODUZIONE AL MACHINE LEARNING .....</i>	<i>70</i>
<i>CAPITOLO 2. STATO DELL'ARTE .....</i>	<i>73</i>
<i>CAPITOLO 3. COFFEE QUALITY DATABASE FROM CQI: ANALISI SPERIMENTALE SULLA QUALITÀ DI ALCUNE TIPOLOGIE DI COFFEE BEANS.....</i>	<i>76</i>
<i>BIBLIOGRAFIA .....</i>	<i>81</i>
<i>SITOGRAFIA .....</i>	<i>84</i>

## **CAPITOLO 1. INTRODUZIONE AL MACHINE LEARNING**

Il termine intelligenza artificiale è stato coniato nel 1956 quando alcuni scienziati provarono a svolgere un esperimento che consisteva nel creare una macchina che simulasse l'intelligenza artificiale. Questo esperimento non ebbe successo, infatti si tornò a parlare di intelligenza artificiale solo negli anni '80 con i sistemi esperti, il primo dei quali fu Eliza, uno psicologo con cui si poteva parlare. Questo esperimento fu condotto da alcuni scienziati, tra cui Arthur Samuel. Egli definì il Machine Learning come un insieme di tecniche che permettono ai computer di imparare dai dati. Un nome da tenere a mente quando si parla di intelligenza artificiale è quello di Alan Turing. Negli anni '50 egli si chiese se le macchine potevano pensare e per rispondere a questa domanda effettuò il cosiddetto Test di Turing. Dal 1956 in poi furono numerose le invenzioni e le scoperte. L'intelligenza artificiale infatti subì un grande sviluppo, gli esempi più evidenti sono la macchina Blue Deep che sconfisse Garry Kasparov, campione degli scacchi, i veicoli a guida autonoma, uno dei quali vinse la DARPA Grand Challenge e un altro che vinse la DARPA Urban Challenge. Gli algoritmi dell'intelligenza artificiale diventano negli anni sempre più precisi ed accurati e infatti l'AI è un campo in continua evoluzione.

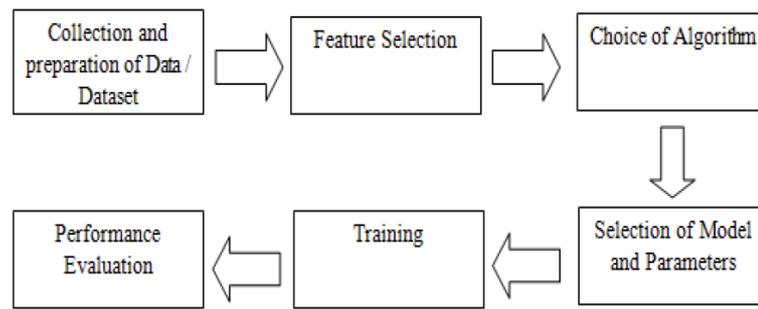
La differenza tra intelligenza artificiale e Machine Learning è che la prima si riferisce a sistemi o macchine che imitano l'intelligenza umana, il secondo invece è un sottoinsieme dell'intelligenza artificiale che crea sistemi che imparano dai dati. All'interno dell'insieme del machine learning c'è un altro sottoinsieme, il Deep Learning. Esso fa riferimento a tutti quegli algoritmi che si ispirano alla struttura del cervello umano ed è formato da una rete di neuroni artificiali (figura 1).



**Fig.1.** *Relazione tra Intelligenza Artificiale, Machine Learning e Deep Learning*

Il Machine Learning è estremamente diffuso ai giorni d'oggi, degli esempi evidenti sono la guida autonoma nelle auto, ma anche tutti i sistemi di raccomandazione come quelli di Amazon e Netflix. Si pensa che l'apprendimento automatico diventerà sempre più importante e necessario per il progresso della tecnologia.

Il Machine Learning funziona nel seguente modo: la macchina riceve un insieme di dati, li analizza e apprende da essi. In seguito, gli algoritmi vengono modificati automaticamente dalla macchina all'aumentare delle informazioni che essa prende in ingresso. Esso è in grado di estrarre informazioni importanti dall'analisi di fonti di dati grandi e distinte. È adatto alla complessità della gestione di dati di origine diversa ed è in grado di gestire insieme di dati sempre più grandi. Infatti, più dati vengono forniti a un algoritmo di Machine Learning, più questa è in grado di essere addestrata e di produrre conseguenze di livello superiore. In base al problema da risolvere c'è uno specifico metodo di Machine Learning che può essere applicato. È importante notare che non ci sono algoritmi superiori ad altri, semplicemente bisogna capire quale di essi si adatta meglio al problema in questione. Una volta compreso questo, si può anche valutare l'idea di integrare più di un algoritmo per risolvere un problema. Esistono due categorie principali di problema: problema di regressione, problema di classificazione e problema di clustering. Il generico modello di Machine Learning è formato da sei componenti principali (figura 2), indipendentemente dall'algoritmo utilizzato: 1) raccolta e preparazione dei dati; 2) estrazione delle features; 3) scelta dell'algoritmo; 4) selezione del modello e dei parametri; 5) addestramento del modello; 6) valutazione del modello.



**Fig.2.** Componenti di un generico modello di Machine Learning

Uno dei punti di forza del Machine Learning è la generalizzazione, quindi la sua capacità di risolvere un problema con qualsiasi tipo di dato. Questa capacità è data dal fatto che i dati vengono suddivisi in due sottoinsiemi: il training set e il test set, il primo per allenare il computer e il secondo per valutare il funzionamento dell'algoritmo. A seconda dell'algoritmo utilizzato si distinguono tre diversi metodi di apprendimento: apprendimento supervisionato, apprendimento non supervisionato e apprendimento per rinforzo.

L'apprendimento supervisionato è il più utilizzato tra i tre e consiste nel fornire alla macchina dei modelli ed esempi che consentano la costruzione di un vero e proprio database di informazioni ed esperienze. Questo significa che la macchina deve semplicemente essere in grado di dare risposte sulla base di esperienze già codificate. L'algoritmo apprende da un set di dati già etichettato, dove per ogni input c'è il rispettivo output che viene usato per insegnare all'algoritmo le regole del modello. L'apprendimento non supervisionato, a differenza del precedente, non prevede che le informazioni abbiano delle etichette. È quindi la macchina che va a codificare le informazioni per fornire l'output. L'apprendimento per rinforzo è il più complesso tra i tre. In questo caso, la macchina viene munita di strumenti in grado di comprendere le caratteristiche dell'ambiente che la circonda. La macchina si muove nello spazio effettuando scelte che le permettano il miglior adattamento nell'ambiente dove si muove. La macchina impara tramite l'interazione con l'ambiente tramite un processo di trial & error. È la macchina che deve scoprire quali azioni intraprendere in modo da ricevere poi delle ricompense. Nel caso in cui la macchina commette degli errori riceve delle punizioni

e il suo obiettivo diventa quello di imparare dall'errore commesso. Il rinforzo è il segnale numerico di ricompensa che si attiva quando l'obiettivo viene raggiunto.

Tra gli algoritmi dell'apprendimento non supervisionato si trova il clustering. Il clustering è una tecnica che ha lo scopo di raggruppare i dati in cluster che siano omogenei internamente ed eterogenei esternamente. La similarità tra gli elementi per determinarne l'appartenenza ai vari cluster può essere calcolata con diversi metodi, uno dei più semplici è il metodo della distanza euclidea. Gli algoritmi di clustering si dividono in due categorie: algoritmi di clustering gerarchico e algoritmi di clustering non gerarchico. Il clustering gerarchico è una suddivisione ricorsiva di un insieme di dati in cluster sempre più piccoli. Esso è rappresentato da un albero radicato in cui ogni foglia rappresenta un punto dati e ogni nodo interno rappresenta un cluster contenente le sue foglie discendenti. L'altra tipologia è il clustering non gerarchico. In questa tipologia si definisce prima il numero di cluster che si vuole avere ( $k$ ) e poi si dividono gli elementi in  $k$  cluster assegnando a ciascun gruppo gli elementi casualmente. Si calcolano in seguito i centroidi di ciascun cluster e si calcola la distanza tra gli elementi e il centroide. Successivamente si riassegnano gli elementi al centroide più vicino e si ripetono questi passaggi finché l'errore non viene minimizzato oppure finché gli elementi non si stabilizzano. Un esempio di clustering non gerarchico è il k-means.

## **CAPITOLO 2. STATO DELL'ARTE**

Il caffè è una delle bevande più consumate al mondo e assume un ruolo rilevante nella cultura del consumatore da anni. Bere un caffè non significa solo consumare una bevanda, ma si tratta di piacere, esperienza, stile di vita e stato sociale. La soddisfazione dei consumatori per questa bevanda è principalmente determinata dagli attributi sensoriali, come il gusto, l'aroma e l'odore. Essendo una bevanda consumata anche in momenti di convivialità, la soddisfazione legata a questi attributi rende l'esperienza ancora più piacevole. Altri fattori che influenzano l'acquisto e il conseguente consumo di caffè sono il paese d'origine, la salubrità del prodotto, che in molti casi è contestata e il fatto che il caffè influenza le emozioni. Quest'ultimo fattore è estremamente importante dato che molte volte la scelta di un prodotto da parte del consumatore è proprio guidata dalle emozioni che questo prodotto gli porta. Il caffè è molto forte in questo, infatti ha la capacità di influenzare le emozioni di chi lo consuma in termini di benessere, ma anche

di sicurezza nelle relazioni sociali. Un ultimo fattore che influenza il consumo del caffè è il grado di *familiarity* verso il prodotto. In questo senso la comunicazione assume un ruolo rilevante e diventa uno strumento necessario per i brand che vogliono far conoscere i loro prodotti e farsi scegliere dal consumatore. I consumatori cercano l'abitudine e il piacere nel bere il caffè. Questi fattori sono legati alle caratteristiche sensoriali della bevanda, come odore, gusto e aroma. Di conseguenza, diventa importante per i brand comunicare al consumatore i punti di forza dei loro prodotti collegati a queste proprietà. Inoltre, è stato provato che ricevere informazioni sulla qualità di un prodotto può influenzare la propensione del consumatore a consumarlo.

Data l'importanza degli attributi menzionati sopra nell'ambito delle bevande calde e la continua competitività delle aziende nel voler andare incontro ai bisogni del consumatore, nasce la necessità di valutarne la qualità in un modo sempre più preciso ed accurato. Le tecniche utilizzate in precedenza si sono rese estremamente utili a questo scopo ma spesso potevano avere delle limitazioni come la soggettività, costi elevati e l'imprecisione. Considerando queste limitazioni, molti studi si sono dedicati a sperimentare tecniche di Machine Learning per sviluppare metodi sempre più veloci in termini di tempo, oltre che più oggettivi e più accurati. In questo senso, il Machine Learning ha avuto un ruolo esponenziale grazie alla sua capacità di prevedere i bisogni dei consumatori e i tratti qualitativi dei prodotti, utili allo sviluppo di prodotti di maggiore qualità e quindi preferiti dai consumatori. Ci sono numerosi esempi di studi che hanno provato con successo a prevedere la qualità dei prodotti finali tramite l'utilizzo di modelli di apprendimento automatico. In uno di essi, i ricercatori hanno stimato l'intensità e gli aromi del caffè tramite l'utilizzo di un naso elettronico economico e portatile abbinato a una modellazione di apprendimento automatico. Questo modello è risultato molto utile nel mondo del caffè non solo per i ricercatori ma anche per le aziende e i consumatori. Questo, infatti, può essere integrato nelle macchine da caffè per valutare possibili cambiamenti dell'aroma causati da fattori esterni che influenzano il processo di infusione. Questo sistema può essere utilizzato per garantire la qualità e la freschezza dei chicchi di caffè e per controllare la qualità all'interno del punto vendita in modo che i consumatori possano trovare gli aromi indesiderati.

Il Machine Learning è risultato molto utile anche per prevedere la qualità di altre tipologie di bevande calde. Uno studio, infatti, l'ha utilizzato per sviluppare un metodo rapido ed

efficace per rilevare il grado di fermentazione del tè nero sulla base delle proprietà elettriche delle foglie del tè. La fermentazione è infatti utilizzata per determinare la qualità del prodotto. Lo studio propone di utilizzare il clustering gerarchico per suddividere i campioni basandosi sul grado di fermentazione e Random Forest e Support Vector Machine per costruire modelli di discriminazione del grado di fermentazione. Il tasso di accuratezza di questi metodi è risultato essere estremamente alto. Un altro studio ha mostrato come delle immagini computerizzate possono essere utilizzate per la valutazione dei cambiamenti degli indici di qualità e del pigmento durante la fermentazione del tè. Questo studio ha rivelato un metodo quantitativo rapido per valutare la qualità della fermentazione attraverso l'estrazione di nove variabili cromatiche analizzando la relazione tra le caratteristiche cromatiche delle immagini e gli indici di qualità. Il risultato è che questi due parametri hanno differenze significative in periodi diversi della fermentazione ma sono anche significativamente correlati.

Un altro scopo per cui il Machine Learning è stato utilizzato nell'ambito delle bevande calde è per rilevare e quantificare l'adulterazione del prodotto in un modo più rapido ed efficiente. Una ricerca ha utilizzato metodi adatti per la classificazione di dati complessi come Random Forest e Reti Neurali Artificiali. Questi metodi possono avere buone prestazioni grazie alla loro capacità di catturare le relazioni non lineari tra le osservazioni con una maggiore accuratezza e robustezza. Lo studio ha utilizzato questi metodi per rilevare la tipologia di adulterazione nel latte. Grazie al Machine Learning è quindi possibile indagare l'autenticità degli alimenti e delle bevande in modo più oggettivo.

Un ultimo studio ha rivelato come per predire la qualità del latte di vacca serva un metodo di apprendimento automatico per diminuire l'errore quadratico medio. Nello specifico, gli autori propongono l'utilizzo della spettroscopia del medio infrarosso, la quale è risultata una tecnica a basso costo, rapida e non distruttiva per la determinazione dei quantitativi di grassi, proteine e lattosio nei campioni di latte.

Uno studio definisce "informazione soggettiva" la fruizione di descrizioni delle proprietà di un prodotto a consumatori che non ne hanno familiarità. Rientra in questa definizione l'opinione di esperti che viene data ai consumatori con meno esperienza sul prodotto. Lo studio prova che l'informazione soggettiva sugli attributi sensoriali del caffè porta i

consumatori a una maggiore preferenza verso il prodotto e ne influenza positivamente l'intenzione di acquisto.

### **CAPITOLO 3. COFFEE QUALITY DATABASE FROM CQI: ANALISI SPERIMENTALE SULLA QUALITA' DI ALCUNE TIPOLOGIE DI COFFEE BEANS**

R è un linguaggio di programmazione e un ambiente software open source che serve per risolvere problemi matematici e statistici e per l'analisi dei dati. Viene principalmente utilizzato per la manipolazione e visualizzazione dei dati, calcoli statistici e generazione di grafici. R viene utilizzato nell'analisi dei dati e nel Machine Learning. Per programmare in R occorre avere uno script di testo con le funzioni e il programma le esegue fornendo il risultato. Viene utilizzato principalmente quando la quantità di dati da analizzare è estremamente grande e sono tante le aziende che ne hanno intuito le potenzialità e le sfruttano. Uno dei punti di forza di R è l'inferenza. Nello specifico, permette di stabilire il peso o contributo che alcune variabili hanno su un determinato fenomeno. È infatti in grado di dire quali variabili spiegano un fenomeno e in che misura. Un altro punto di forza è la previsione. È infatti in grado di stimare un valore sulla base di alcune variabili, per esempio si potrebbe stimare un prezzo sulla base delle caratteristiche di un consumatore.

Per l'analisi sperimentale della presente ricerca viene utilizzato il dataset Coffee Quality database from CQI. Questo contiene le recensioni di 1340 chicchi di caffè da parte di recensori qualificati del Coffee Quality Institute. Gli attributi considerati nel dataset di riferimento sono i seguenti: l'aroma, il sapore, il retrogusto, l'acidità, la corposità, l'equilibrio, l'uniformità, la pulizia della tazza, la dolcezza, l'umidità, i difetti, il metodo di lavorazione, il colore, la specie (arabica/robusta), il proprietario, il paese d'origine, il nome dell'azienda agricola, il numero di lotto, il mulino, l'azienda, l'altitudine e la regione. Per l'analisi verranno presi in considerazione i seguenti: aroma, aftertaste, balance, uniformity, cup cleanliness, sweetness, flavor, body e acidity.

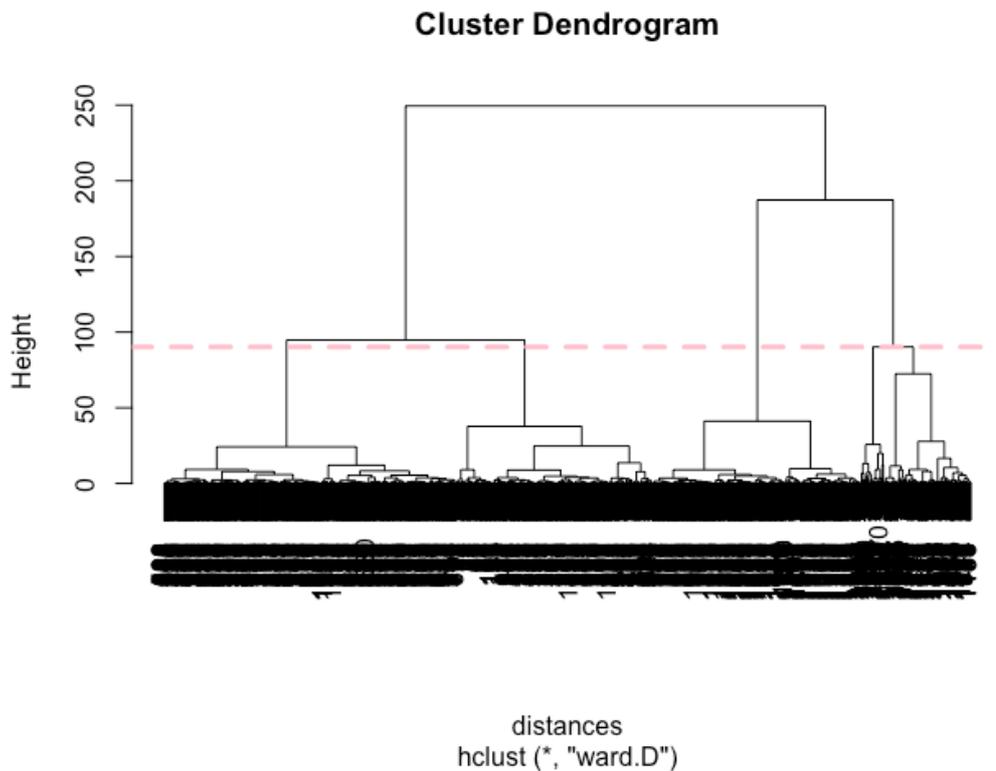
L'analisi consisterà nell'applicare le due tipologie di clustering (gerarchico e non gerarchico) per suddividere i chicchi di caffè in gruppi sulla base dei punteggi ottenuti nei vari attributi presi in considerazione. In primo luogo, si scaricano i pacchetti che servono, si permette ad R Studio di leggere il dataset da analizzare e si assegnano a

piacimento i nomi alle variabili che corrispondono alle colonne del dataset in modo che siano rappresentativi. In seguito, si rimuovono le colonne che non interessano per l'analisi e si eliminano eventuali outlier. Dopo di che si calcolano le distanze con il metodo delle distanze euclidee e si applica l'algoritmo di clustering gerarchico con il metodo di Ward. Si va poi a creare il dendrogramma per poter visualizzare i cluster tramite un grafico e ad applicare l'elbow method per calcolare il numero ottimo di cluster da formare per l'analisi. Da qui si evince che il numero ottimale di cluster è pari a quattro. Si assegnano in seguito le osservazioni ai vari cluster tagliando il dendrogramma. Infine, si calcola la media dei punteggi assegnati a ciascun attributo all'interno di ognuno dei quattro cluster e si procede con la creazione di sottoinsiemi con all'interno solo le osservazioni assegnate a ciascuno di essi.

Si applica in seguito il k-means iniziando da un processo di preelaborazione dei dati attraverso la normalizzazione e la standardizzazione. In seguito, si applica il k-means e si va a vedere quali elementi appartengono a quali cluster. Si rappresentano poi graficamente i cluster formati e si assegnano le osservazioni a questi ultimi. Infine, si guardano per ogni cluster i punteggi medi attribuiti a ciascun attributo.

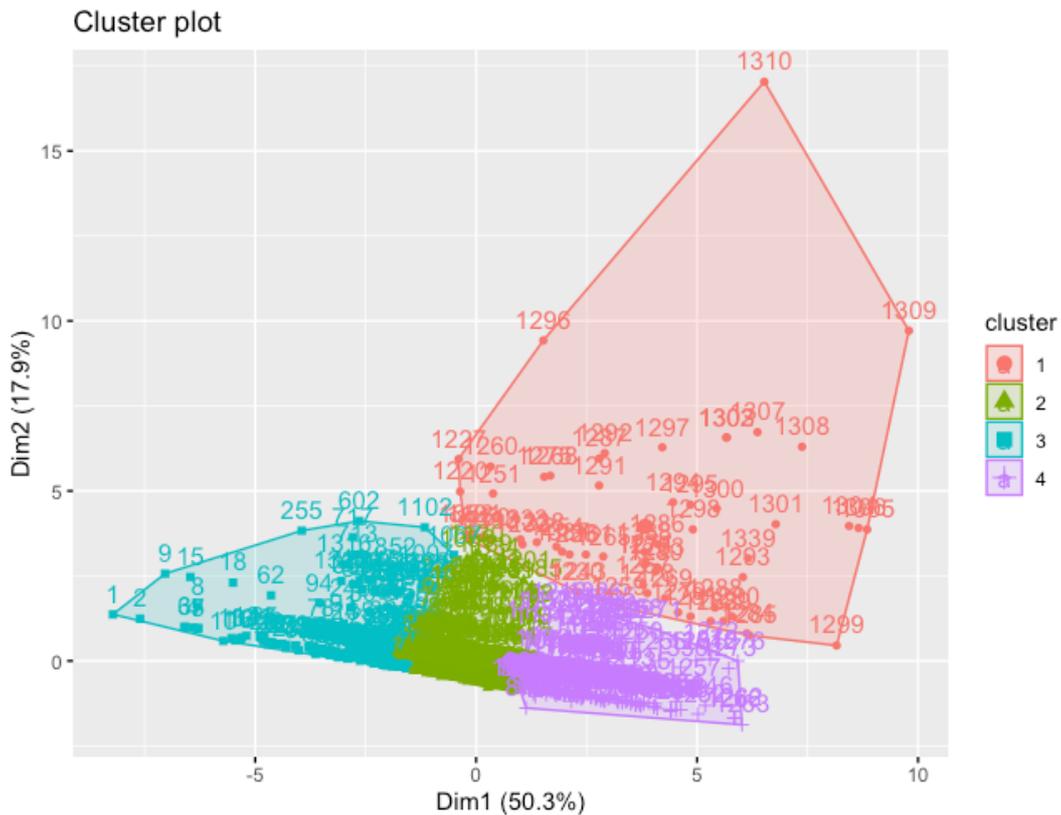
Questa analisi può risultare utile ai brand di caffè per capire, in base a che caffè utilizzano nelle loro miscele, su quali attributi possono puntare nella loro comunicazione e quindi quali di essi possono utilizzare come elementi di differenziazione. Dall'analisi di clustering gerarchico, considerando i punteggi dati dagli esperti agli attributi, l'importanza che questi attributi hanno per i consumatori, e l'effetto che l'informazione soggettiva ha sui consumatori in termini di preferenza per il prodotto e aumento dell'intenzione d'acquisto, un brand che si posiziona nel primo cluster nella comunicazione del proprio caffè potrebbe mettere in luce un buon aroma e odore quando viene infuso con l'acqua calda, un sapore ottimo quando entra in contatto con la lingua, un retrogusto piacevole in modo che rimanga in bocca un buon sapore per un tempo prolungato, un caffè acido al punto giusto che rende l'esperienza di berlo molto piacevole al palato e infine un caffè che può essere considerato corposo. Inoltre, possono far riferimento al fatto che il loro caffè vanta una perfetta sinergia tra gli aromi presenti. Coloro che invece si posizionano nel terzo cluster possono vantare un caffè dolce, con una buona consistenza dell'aroma in diverse tazze così che la qualità del caffè sia garantita anche nella degustazione delle tazze successive e un caffè ben conservato, privo

di note dure, terrose e sgradevoli che possano avere un effetto negativo tra il primo sorso e i successivi (figura 3).



**Fig. 3.** Dendrogramma con il taglio ad  $h=90.2$  per l'ottenimento dei quattro cluster

Dall'analisi di clustering non gerarchico con il metodo del k-means i brand che si posizionano nel cluster 2 nella loro comunicazione possono vantare come punto di forza del loro caffè una buona uniformità, quindi una buona consistenza dell'aroma in diverse tazze così da garantire un caffè di qualità a ogni assaggio e un caffè pulito, privo di macchie e difetti e privo di note dure e terrose. I brand del cluster 3 possono invece vantare un caffè con un aroma unico nel momento in cui viene infuso con acqua calda, un retrogusto piacevole al palato e durevole così che questa gradevolezza sia prolungata, un gusto percepito unico e piacevole, una corposità gradevole al palato e un'acidità piacevole e delicata. Inoltre, si possono differenziare grazie a un equilibrio perfetto tra sapore, retrogusto, acidità e corposità in modo da formare una sinergia speciale tra questi attributi ritenuti importanti da parte del consumatore. Infine, i brand appartenenti al quarto cluster possono concentrarsi sulla dolcezza del caffè senza la presenza di note troppo amare e sgradevoli.



**Fig. 24.** *Suddivisione delle osservazioni in cluster tramite l'algoritmo di clustering non gerarchico k-means*

L'analisi della presente ricerca prevede l'utilizzo dell'algoritmo di clustering ma il dataset di riferimento lascia spazio a più algoritmi che possono essere utilizzati per ulteriori analisi. Si potrebbe infatti applicare un altro algoritmo, come una regressione lineare andando a calcolare l'intenzione di acquisto dei chicchi di caffè utilizzando gli attributi come variabili indipendenti. Inoltre, la medesima analisi potrebbe essere effettuata su ulteriori prodotti, come i cibi surgelati. La credenza maggiore su questi prodotti è che siano veloci da preparare ma che non siano per nulla salutari. Bisogna però sapere che non è così per tutti e un'analisi di qualità su alcuni di questi prodotti potrebbe fornire ai brand una maniera per dare ai consumatori maggiore visibilità degli effetti benefici che questi prodotti possono avere. Inoltre, un'estensione futura per la presente ricerca riguarda la salubrità del caffè. In particolare, questo prodotto può essere consumato spesso anche in momenti di convivialità con amici e parenti. C'è però un gran numero di persone che pensa che il caffè abbia effetti nocivi sulla salute e quindi preferisce evitarne il consumo. Ciò che non si sa è che il caffè può avere delle proprietà benefiche sulla salute delle persone se consumato con moderazione. Infine, dalla ricerca sono emersi alcuni

attributi legati al caffè che possono essere utilizzati dai brand nella loro comunicazione. Questa si potrebbe ampliare andando ad analizzare quali mezzi di comunicazione sono ritenuti più efficaci per questo tipo di prodotto in modo da poter capire dove utilizzare questi attributi e i punti di forza che i brand possono avere su di essi.

## BIBLIOGRAFIA

- Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, 1-16.
- Bemfeito, C., Guimaraes, A., Lima de Oliveira, A., Andrade, B., Frossard de Paula, L., & Pimenta, C. (2021). Do consumers perceive sensory differences by knowing information about coffee quality? *LWT - Food Science and Technology*, 1-7.
- Bookman, S. (2013). Branded Cosmopolitanisms: "Global" Coffee Brands and the Co-creation of "Cosmopolitan Cool". *Cultural Sociology*, 56-72.
- Chhabra, A., Masalkovaite, K., & Mohapatra, P. (2021). An Overview of Fairness in Clustering. *IEEE Access*, 130698-130720.
- Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., & Mathieu, C. (2019). Hierarchical Clustering: Objective Functions and Algorithms. *Journal of the ACM*, 1-42.
- Dong, C., Liang, G., Hu, B., Yuan, H., Jiang, Y., Zhu, H., & Qi, J. (2018). Prediction of Congou Black Tea Fermentation Quality Indices from Color Features Using Non-Linear Regression Methods. *Scientific Reports*, 1-11.
- Farah, J., Cavalcanti, R., Guimaraes, J., Balthazar, C., Coimbra, P., Pimentel, T., . . . Cruz, A. (2021). Differential scanning calorimetry coupled with machine learning technique: An effective approach to determine the milk authenticity. *Food Control*, 1-9.
- Frizzarin, M., Gormley, I., Berry, D., Murphy, T., Casa, A., Lynch, A., & S., M. (2021). Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. *Journal of Dairy Science*, 7438-7447.
- Gonzalez Viejo, C., Tongson, E., & Fuentes, S. (2021). Integrating a Low-Cost Electronic Nose and Machine Learning Modelling to Assess Coffee Aroma Profile and Intensity. *Sensors*, 1-15.
- Gonzalez Viejo, C., Torrico, D. D., Dunshea, F. R., & Fuentes, S. (2019). Emerging Technologies Based on Artificial Intelligence to Assess the Quality and Consumer Preference of Beverages. *Beverages*, 1-25.
- Hu, Y. (2018). Marketing and Business Analysis in the Era of Big Data. *American Journal of Industrial and Business Management*, 1747-1756.

- Jabbar, A., Akhtar, P., & Dani, S. (2020). Real-time big data processing for instantaneous marketing decisions: A problematization approach. *Industrial Marketing Management*, 558-569.
- Li, J., Streletskaia, N., & Gomez, M. (2019). Does taste sensitivity matter? The effect of coffee sensory tasting information and taste sensitivity on consumer preferences. *Food Quality and Preference*, 447-451.
- Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *WIREs Data Mining and Knowledge Discovery*, 1-16.
- Osisanwo, F., J.E.T., A., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, 128-138.
- Petrescu, D., Vermeir, I., & Petrescu-Mag, R. (2019). Consumer Understanding of Food Quality, Healthiness, and Environmental Impact: A Cross-National Perspective. *International Journal of Environmental Research and Public Health*, 1-20.
- Sadilek, T. (2019). Perception of Food Quality by Consumers: Literature Review. *European Research Studies Journal*, 57-67.
- Samoggia, A., & Riedel, B. (2018). Coffee consumption and purchasing behavior review: Insights for further research. *Appetite*, 70-81.
- Samoggia, A., & Riedel, B. (2019). Consumers' Perceptions of Coffee Health Benefits and Motives for Coffee Consumption and Purchasing. *Nutrients*, 1-21.
- Samoggia, A., Del Prete, M., & Argenti, C. (2020). Functional Needs, Emotions, and Perceptions of coffee Consumers and non-Consumers. *Sustainability*, 1-23.
- Sarker, I. H. (2021). Machine learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 1-21.
- Sharma, D., & Kumar, N. (2017). A Review on Machine Learning Algorithms, Tasks and Applications. *International Journal of Advanced Research in Computer Engineering & Technology*, 1548-1552.
- Wachamo, H. (2017). Review on Health Benefit and Risk of Coffee Consumption. *Medicinal & Aromatic Plants*, 1-12.
- Xu, Y., Zhou, Y., Sekula, P., & Ding, L. (2021). Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment*, 1-13.
- Zhou, L., Pan, S., Wang, J., & Vasilakos, A. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, 350-361.

- Zhu, H., Liu, F., Ye, Y., Chen, L., Liu, J., Gui, A., . . . Dong, C. (2019). Application of machine learning algorithms in quality assurance of fermentation process of black tea-based on electrical properties. *Journal of Food Engineering*, 165-172.
- Zou, H. (2020). Clustering Algorithm and Its Application in Data Mining. *Wireless Personal Communications*, 21-30.

## SITOGRAFIA

- Artificial intelligence & Machine Learning*. Tratto da Data Lab: <https://www.bigdata-lab.it/2edizione/corsi/artificial-intelligence-machine-learning/>
- Bock, T. *What is Hierarchical Clustering?* Tratto da Displayr: <https://www.displayr.com/what-is-hierarchical-clustering/>
- Boldrini, N. (2021, Novembre 24). *Deep Learning, cos'è l'apprendimento profondo, come funziona e quali sono i casi di applicazione*. Tratto da ai4business: <https://www.ai4business.it/intelligenza-artificiale/deep-learning/deep-learning-cose/>
- Clayton, R. (2021). *Cos'è il machine learning?* Tratto da Oracle: <https://www.oracle.com/it/data-science/machine-learning/what-is-machine-learning/>
- Coffee Quality Database*. (2018). Tratto da GitHub: <https://github.com/jldbco/coffee-quality-database/blob/master/README.md>
- Coffee Quality Institute*. Tratto da Coffee Quality Institute: <https://www.coffeeinstitute.org/about/strategy>
- Condemi, J. (2021, Novembre 22). *Reinforcement learning: cos'è, come funziona ed esempi dell'apprendimento per rinforzo*. Tratto da bigdata4innovation: <https://www.bigdata4innovation.it/intelligenza-artificiale/reinforcement-learning-cose-come-funziona-ed-esempi-dellapprendimento-per-rinforzo/>
- Cos'è il Machine learning, come funziona l'apprendimento automatico e quali sono le sue applicazioni*. (2022, Giugno 12). Tratto da ai4business: <https://www.ai4business.it/intelligenza-artificiale/machine-learning/machine-learning-cosa-e-applicazioni/>
- Cos'è R*. Tratto da GeekandJob: <https://www.geekandjob.com/wiki/r>
- Dotti, P. (2022, Gennaio 13). *Supervised learning, cos'è, esempi di apprendimento supervisionato*. Tratto da ai4business: <https://www.ai4business.it/intelligenza-artificiale/supervised-learning-cose-esempi-di-apprendimento-supervisionato/>
- Dotti, P. (2022, Gennaio 27). *UnSupervised learning, cos'è ed esempi di apprendimento non supervisionato*. Tratto da ai4business:

- <https://www.ai4business.it/intelligenza-artificiale/unsupervised-learning-cose-ed-esempi-di-apprendimento-non-supervisionato/>
- La programmazione R: che cos'è.* (2019, Settembre 19). Tratto da Digital New Academy: <https://www.digitalnewacademy.com/cosa-e-la-programmazione-r-e-a-cosa-serve/>
- Machine Learning - I diversi apprendimenti di una macchina.* Tratto da [https://www.intelligenzaartificiale.it/machine-learning/#I\\_diversi\\_apprendimenti\\_di\\_una\\_macchina](https://www.intelligenzaartificiale.it/machine-learning/#I_diversi_apprendimenti_di_una_macchina)
- Machine learning: Che cos'è e perchè è importante.* Tratto da sas: [https://www.sas.com/it\\_it/insights/analytics/machine-learning.html](https://www.sas.com/it_it/insights/analytics/machine-learning.html)
- Minini, A. *Cos'è R.* Tratto da andreaminini: <https://www.andreaminini.com/programmazione/r/>
- Minini, A. *La tecnica del clustering.* Tratto da andreaminini: <https://www.andreaminini.com/ai/machine-learning/clustering>
- Nasteski, V. (2017). *An overview of the supervised learning methods.* Tratto da ResearchGate: [https://www.researchgate.net/profile/Vladimir-Nasteski/publication/328146111\\_An\\_overview\\_of\\_the\\_supervised\\_machine\\_learning\\_methods/links/5c1025194585157ac1bba147/An-overview-of-the-supervised-machine-learning-methods.pdf](https://www.researchgate.net/profile/Vladimir-Nasteski/publication/328146111_An_overview_of_the_supervised_machine_learning_methods/links/5c1025194585157ac1bba147/An-overview-of-the-supervised-machine-learning-methods.pdf)
- Provino, A. (2019, Novembre 15). *K-means Clustering.* Tratto da andreaprovino: <https://andreaprovino.it/k-means-clustering/>
- Provino, A. (2020, Novembre 15). *Clustering: tutto quello che devi sapere | Edizione 2021.* Tratto da andreaprovino: <https://andreaprovino.it/clustering/>
- R Software, cosa è e a cosa serve questo software statistico.* Tratto da Strateg.ee: <https://www.strateg.ee/r-software-cosa-e-a-cosa-serve-questo-software-statistico-r/>
- Rezzani, A. (2015, Gennaio 5). *Tecniche di clustering.* Tratto da Data Skills: <https://www.dataskills.it/tecniche-di-clustering/#gref>
- Rezzani, A. (2019, Novembre 28). *Apprendimento automatico e intelligenza artificiale.* Tratto da Data Skills: <https://www.dataskills.it/apprendimento-automatico-e-intelligenza-artificiale/#gref>
- Stecher, M. T. (2018, Novembre 21). *La storia dell'intelligenza artificiale, da Turing ad oggi.* Tratto da cyberlaws: <https://www.cyberlaws.it/2018/la-storia-dellintelligenza-artificiale-da-turing-ad-oggi/>

Velocci, S. (2020, Dicembre 14). *Machine Learning: Cos'è e perchè è importante*.

Tratto da start2impact: <https://www.start2impact.it/blog/programmazione/cose-machine-learning/>