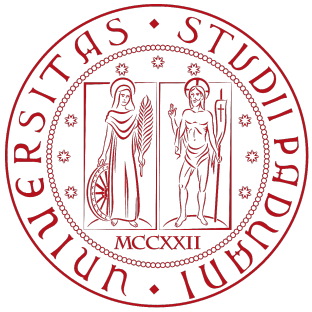
Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in

STATISTICA PER LE TECNOLOGIE E LE SCIENZE



RELAZIONE FINALE

TECNICHE DI WEB SCRAPING PER L'ANALISI DEL MERCATO ONLINE DEI COMPONENTI RICAMBIO DI SISTEMI DI RISCALDAMENTO E RAFFRESCAMENTO ATTRAVERSO METODI DI CLASSIFICAZIONE E TEXT MINING, UN PROGETTO STAGE CON BAXI S.p.A

Relatore: dott. Erlis Ruli

Dipartimento di Scienze Statistiche

Laureando: Mattia Trevisol

Matricola n. 1235642

Anno Accademico 2022/2023

Alla mia famiglia

Ai miei compagni di corso

Ai limiti della legalità

Indice

Sommario

Sommario

Il periodo di stage svolto presso BAXI S.p.A è stata un'occasione per comprendere il funzionamento e l'organizzazione di una grande azienda attiva nel settore della produzione di caldaie e pompe di calore.
Lo studio del mercato online dei componenti di ricambio per sistemi di riscaldamento e raffrescamento rappresenta un'area di grande interesse per le aziende attive in questo settore. In particolare, l'analisi dei prezzi da siti e-commerce di rivendita ricambi può sia fornire informazioni preziose sulla competitività dell'azienda, sia essere di supporto a politiche commerciali mirate ad ottimizzare le vendite.
Il progetto in questione ha previsto l'utilizzo di tecniche di web scraping e text mining per effettuare una comparazione dei prezzi dei componenti di ricambio presenti sui principali siti di e-commerce. Inoltre, è stata effettuata una cluster analysis dei componenti in modo da classificarli in categorie e sottocategorie. Il resto dell'elaborato è strutturato come segue. Il primo capitolo introduce l'azienda Baxi S.p.A. e le principali tecniche utilizzate nel progetto. Nel secondo si affronta il tema dell'estrazione e processazione dei dati. Il terzo capitolo presenta la fase di analisi. Infine nell'ultimo capitolo si riportano alcuni commenti finali.

1 Introduzione

1.1 L'Azienda

In questo primo capitolo si riassume brevemente l'azienda Baxi S.p.A. presso la quale ho conseguito il periodo di stage curriculare dal 11/09/2022 fino al 03/03/2023. Andrò a spiegare la storia dell'azienda, dalla nascita come Smalterie Metallurgiche Venete fino ad oggi, dove la produzione di caldaie ha raggiunto di recente le 600.000 unità annue.

1.1.1 BAXI S.p.A.

Baxi S.p.A. è un'azienda italiana che offre una gamma completa di prodotti e sistemi per la climatizzazione sia invernale che estiva. Lo stabilimento ha sede a Bassano del Grappa, Vicenza, è il più grande del settore a livello europeo, occupa una superficie di 100.000 mq.

(a) Logo all'entrata dell'azienda

(b) Visione dall'alto

Figure 1: Stabilimento di Bassano del Grappa

Baxi S.p.A. ha origini lontane quando nel 1925 la famiglia tedesca Westen fondò lo stabilimento delle Smalterie Metallurgiche Venete, uno dei maggiori stabilimenti di prodotti smaltati come scaldacqua elettrici e vasche da bagno oltre a prodotti per il riscaldamento quali corpi scaldanti in acciaio.

Figure 2: Locandine pubblicitarie delle Smalterie Metallurgiche Venete

Alla fine degli anni 70 l'azienda focalizza la propria produzione nel settore del riscaldamento divenendo uno dei primi stabilimenti ad introdurre gli apparecchi domestici a gas con la produzione di caldaie murali contestualmente all'espansione della rete del gas che avveniva proprio in quel periodo. Nella metà degli anni Ottanta l'azienda prosegue con ottimi risultati consolidando la sua presenza in territorio nazionale; comincia allora anche l'espansione nei mercati esteri. Nel 1999 entra a far parte del gruppo inglese BAXI GROUP, leader in Europa nel settore riscaldamento. Nel 2009 De Dietrich Remeha Group e Baxi Group annunciano la creazione di BDR Thermea.

Figure 3: Logo BDR Thermea Group

Figure 4: Caldaia residenziale duotech marchiata Baxi S.p.A.

Oggi Baxi S.p.A. progetta e produce nello stabilimento più grande del settore a livello europeo per la produzione di caldaie murali. Recenti stime attestano un ammontare di oltre 11 milioni di caldaie prodotte. Baxi offre inoltre una gamma completa di prodotti energeticamente efficienti per garantire massimo comfort e risparmio che si è evoluta negli anni passando dall'offerta di singolo prodotto a sistema integrato con fonti rinnovabili.

Il portfolio prodotti comprende sistemi ibridi, caldaie domestiche, caldaie a condensazione di alta potenza, pompe di calore e climatizzatori sia per contesti residenziali che commerciali, fan coil, bollitori, scaldacqua, sistemi solari, moduli d'utenza. L'ultima novità di prodotto a zero emissioni a firma Baxi è la prima caldaia domestica premiscelata a idrogeno puro prodotto tramite energia rinnovabile senza emissioni di CO e CO2.

Figure 5: Gamma caldaie ad idrogeno di Baxi S.p.A.

1.1.2 BDR THERMEA GROUP

BDR Thermea è un produttore e distributore, leader a livello mondiale, di sistemi e servizi innovativi per il riscaldamento e la produzione di acqua calda, attivo su un mercato il cui volume di vendite supera i 16 miliardi di euro all'anno.

In tutta Europa lavorano per BDR 6.100 addetti, con un fatturato nel 2021 di oltre 2 miliardi di euro. Il Gruppo ha una posizione di punta nel mercato di sei Paesi: UK, Francia, Germania, Spagna, Olanda ed Italia, ed ha una forte presenza nei mercati in rapida crescita dell'Europa dell'Est, di Turchia, Russia, Nord America e Cina. Complessivamente BDR Thermea è attivo in più di 100 Paesi in tutti i continenti. La strategia del Gruppo si basa sull'articolazione delle attività con diversi marchi commerciali e sulla presenza di importanti filiali nazionali nelle economie europee più importanti, tali da consentire una pronta reazione ai cambiamenti nella domanda locale.

BDR Thermea è titolare di alcuni tra i marchi più importanti nel mercato europeo dei prodotti per il riscaldamento. Questi comprendono De Dietrich, Baxi, Remeha, Heatrae Sadia, Brötje, Potterton, Chappée, BaxiRoca and Baymak.

Figure 6: Composizione del gruppo BDR Thermea

1.1.3 Collocazione e descrizione dello stage

Durante il mio periodo di stage curriculare, ho avuto l'opportunità di lavorare presso l'ufficio ricambi di Baxi S.p.A., il quale costituisce una delle branche della parte *service* dell'azienda. Durante il mio periodo di tirocinio, mi sono occupato principalmente delle revisioni dei cataloghi di ricambi di caldaie e pompe di calore. In particolare, ho fornito supporto nell'aggiornamento e approvazione degli Spare Part BOM, ovvero gli esplosi di caldaie, utilizzando il programma aziendale.

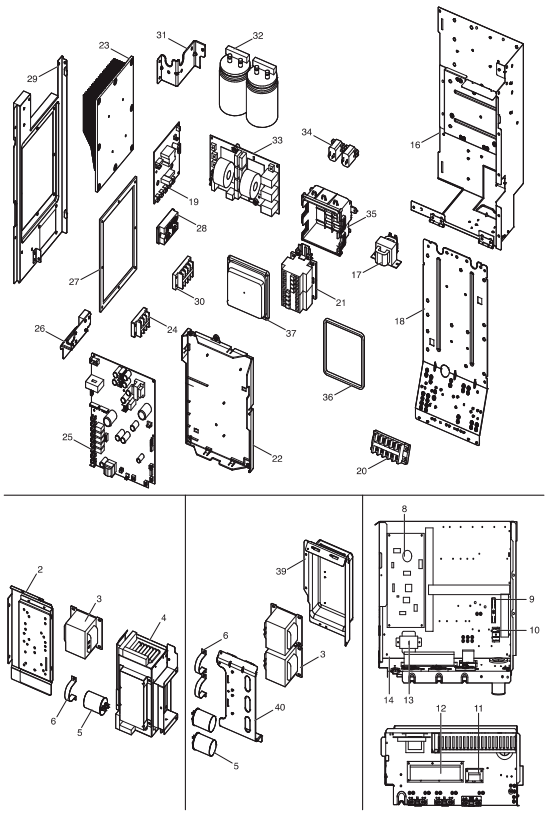


Figure 7: Esploso di una caldaia

1.2 Tecniche e Tecnologie

Questo progetto si concentra sull'analisi di dati estratti tramite tecniche di web scraping, in combinazione con tecniche di text mining e cluster analysis, con l'obiettivo di estrarre informazioni utili a supporto delle decisioni aziendali. Nel seguente capitolo, verranno descritte in dettaglio le tre tecniche utilizzate per l'analisi dei dati, le loro finalità e le modalità di utilizzo nel contesto del progetto.

1.2.1 Web scraping

Il web scraping, o estrazione dati web, è una tecnica di estrazione automatica di informazioni da pagine web. Consiste nell'utilizzo di software per estrarre i dati e le informazioni dalle pagine web in modo automatico, al fine di costruire dataset strutturati a partire da pagine non strutturate o non disponibili in formato digitale. In genere, il web scraping viene utilizzato per raccogliere informazioni di diverso tipo, come prezzi di prodotti, recensioni, notizie, indirizzi email, dati di contatto e così via. Il web scraping è considerato legale quando viene utilizzato per scopi leciti, come la raccolta di dati pubblici o di informazioni disponibili liberamente online per analisi, ricerca, elaborazione di dati, monitoraggio dei prezzi, ecc. In generale, il web scraping è legale se:

- Si accede solo a pagine web pubbliche o a quelle a cui si ha il permesso di accedere;
- Si rispettano le limitazioni imposte dal sito web. Queste informazioni sono verificabili nei termini e condizioni del sito da cui si sta estraendo;
- Non si infrangono le leggi sulla privacy o sui diritti di proprietà intellettuale.

A livello europeo si fa affidamento alle seguenti normative: *Regolamento generale sulla protezione dei dati (GDPR)* che regola la protezione dei dati personali, *Codice della proprietà industriale (CPI)* che garantisce i diritti di proprietà intellettuale in Italia e la *legge sul diritto d'autore* che regola la protezione dei diritti di proprietà intellettuale sui contenuti creativi, come i testi, le immagini e i video. Nel mio progetto di tesi la tecnica del *web scraping* è fondamentale per l'acquisizione del dataset iniziale. Verranno estratti da 11 siti un elenco di componenti ricambio di caldaie. Per ogni articolo verrà estratta la descrizione, il prezzo, a cui poi verrà dedotto dal link anche la marca e il nome del sito.

1.2.2 Text mining

Il *Text Mining* è una tecnica di elaborazione automatica del linguaggio naturale che si occupa di analizzare e comprendere il significato dei testi scritti in forma digitale, al fine di estrarne informazioni utili. Il *Text Mining* si basa sulla combinazione di tecniche di elaborazione del linguaggio naturale, informatica, matematica e statistica, e richiede competenze interdisciplinari per essere applicato con successo. Grazie alla crescente quantità di dati disponibili in forma digitale, il *Text Mining* è diventato un'importante area di ricerca e sviluppo in molte industrie, tra cui il marketing, la finanza, la ricerca scientifica e molte altre.
La tecnica del *Text Mining* verrà applicata nella fase di post-estrazione (essendo fondamentale nell'estrapolazione e pulizia della descrizione e del prezzo dal codice HTML) e sia in fase di classificazione (mi ha permesso di creare e utilizzare le Document Term Matrix).

1.2.3 Cluster Analysis

La Cluster Analysis è una tecnica statistica di analisi multivariata che si occupa di raggruppare oggetti in base alla somiglianza tra di essi. Il suo obiettivo principale è quello di identificare dei cluster o gruppi di oggetti omogenei tra di loro, ma eterogenei rispetto agli oggetti che appartengono ad altri cluster.
La Cluster Analysis nel mio progetto consisterà nella collocazione di componenti estratti tramite web scraping in categorie e sottocategorie. Questa tecnica, una volta valutata la bontà di classificazione, ci permetterà di utilizzare le categorie e le sottocategorie nella fase di analisi.

2 Estrazione e processazione dei dati

Il seguente capitolo illustra i processi decisionali adoperati al fine di ottenere il dataset completo. Nella fattispecie verranno trattate la modalità e gli strumenti che mi hanno guidato nell'estrazione e nella gestione del dato grezzo. Spiegherò nel dettaglio il prelievo dei dati fatto da 11 siti online di vendita di ricambi originali di caldaie e sistemi ibridi e il modo in cui ho rielaborato e curato il dato evitando l'eccessiva perdita di informazione.

2.1 Processo di estrazione del dato

Partiamo dal presupposto che l'estrazione massiva di informazioni da un sito web non è banale, vanno considerati molteplici aspetti e non vi è sempre la certezza di poterlo fare. Vedremo in seguito che sarà necessario un minuzioso studio della conformazione del sito online da cui si vuole trarre informazioni. L'aspetto preliminare da valutare è la scelta del programma da utilizzare per poter effettuare le estrazioni. Molti software (come Python e R per citarne alcuni) mettono a disposizione librerie e specifiche funzioni che permettono all'utente di effettuare rilevazioni di dati da siti web. Per il progetto in questione la mia scelta è ricaduta sul software statistico R sia per un fattore di dimestichezza con il programma e sia per la sua predisposizione alla creazione di grafici necessari per le mie analisi. Per la teoria dell'applicazione del web scraping su R è stato dedicato un paragrafo apposito.

2.1.1 Studio del sito web

Tra gli aspetti cardine da considerare quando si vuole effettuare il prelievo di informazioni online è necessario citare lo studio delle dinamiche del sito desiderato. E' fondamentale infatti capire la conformazione del codice sorgente e, in particolare, il modo in cui lo script si altera in base alle azioni eseguite dall'utente all'interno del sito. Quando una persona si interfaccia ad una qualsiasi pagina online ed effettua una selezione o ricerca, in background la sorgente della pagina corrente si modifica in automatico (per verificare basta tenere aperta la finestra Ispeziona su Chrome). Ad ogni mutazione del codice sorgente corrisponde una conseguente alterazione parziale del link del sito. Come approfondirò in seguito, il link subisce un'estensione di una stringa alfanumerica sulla coda.

Questa permette l'indirizzamento dell'utente ad un altra pagina dove è presente la scelta che ha effettuato. Nei paragrafi successivi verrà dunque approfondita l'applicazione di queste nozioni per il raggiungimento dello scopo di questo processo, cioè l'estrazione delle informazioni grezze dai siti.

2.1.2 Struttura e dinamiche di mutamento dell'URL

Se ipotizzassimo di paragonare l'infinita rete internet con il nostro mondo, non sarebbe un pensiero così utopico. Andando alla radice delle due realtà è possibile rendersi conto che si comportano esattamente nello stesso modo. Un insieme di elementi connessi tra loro, la configurazione classica di un grafo. Nessun'altra rappresentazione di questi due emisferi sarebbe più verosimile di questa figura geometrica. La comunicazione tra esseri umani esiste fin dall'alba dell'umanità. Evoluzione dopo evoluzione si è passati dal semplice gesticolare alla parola. Questa infatti permette la connessione e soprattutto lo scambio di informazioni tra persone. Nel mondo di internet non ci si discosta molto da questa realtà. Le persone, che nel grafo idealizzato rappresenterebbero i nodi, non sono altro che pagine web. La comunicazione umana in tutte le sue sfaccettature e forme, che individua gli archi del grafo, è rappresentata sotto il termine informatico URL (Uniform Resource Locator). Chiamato colloquialmente indirizzo web, è un identificatore univoco utilizzato per individuare una risorsa su Internet. Gli URL sono costituiti da più parti, tra cui un protocollo e un nome di dominio, che indicano a un browser Web come e dove recuperare una risorsa. Gli utenti finali utilizzano gli URL digitando direttamente nella barra degli indirizzi di un browser o facendo clic su un collegamento ipertestuale trovato in una pagina Web, in un'e-mail o da un'altra applicazione. La struttura di un URL è composta generalmente in tre parti.

Protocollo di rete	**Dominio**	**Percorso**
https://	*my.ecommerce.site.com*	*/ricambi-caldaie-originali*

Table 1: Struttura base di un URL

Il prefisso iniziale della stringa alfanumerica prende il nome di **protocollo di rete**. La sua funzione principale è quella di stabilire una connessione con il server e inviare le pagine HTML al browser dell'utente. In genere i browser web utilizzano per impostazione predefinita il protocollo HTTP. Successivamente troviamo il **dominio**, riferimento univoco che rappresenta una pagina Web (in sto caso *my.ecommerce.site.com*). Infine un **percorso** che fa riferimento a un file o a un percorso sul server Web. Nell'esempio in questione facciamo riferimanto alla stringa "*/ricambi-caldaie-originali*".

A questa composizione si aggiunge un'ulteriore parte riguardante le scelte effettuate dall'utente all'interno della pagina web. Parliamo in particolare di estensioni di URL riguardanti Query di ricerca (es. *Sonda rapida*),

https://www.my.ecommerce.site.com/ricambi-caldaia-originali/?q=Sonda+rapida

filtri e parametri impostati dall'utente (esempio selezione della Marca),

https://www.my.ecommerce.site.com/ricambi-caldaia-originali/Baxi

o scorrimenti di pagina.

https://www.my.ecommerce.site.com/ricambi-caldaia-originali&pg=2

Avendo quindi chiare le mutazioni e le estensioni dell'indirizzo appena descritte, saremo in grado con tecniche computazionali di generare dei link che saranno utili poi all'estrazione delle informazioni.

2.1.3 La sorgente della pagina, il punto di estrazione

Un altro aspetto fondamentale da tener conto per la corretta realizzazione del processo di estrazione è l'analisi della sorgente del sito web desiderato. Riprendendo la metafora usata in precedenza, il sito web è identificabile ed esprimibile sotto il concetto di comunità/famiglia di elementi. Al suo interno infatti le pagine web possiedono una forma simile di URL. A meno dell'estensione del "percorso", la composizione dell'indirizzo web è la medesima.

Esempio:

https://www.my.ecommerce.site.com/ricambi-caldaia-originali/Baxi

https://www.my.ecommerce.site.com/ricambi-caldaia-originali/Riello

Come accennato nell'introduzione al capitolo, ad ogni mutazione dell'URL della pagina c'è una conseguente alterazione del codice sorgente della pagina corrente. Non è mia intenzione scendere in dettaglio nello studio e nell'interpretazione del codice HTML della pagina web, dopotutto sono uno statistico e non un programmatore. Il mio obiettivo è l'ottenimento e l'analisi di dati e non di certo lo studio minuzioso della sintassi e funzionalità della codifica della sorgente della pagina web. Dato che il mio progetto promette di effettuare un'analisi di mercato, le variabili assolutamente necessarie da ottenere sono: la descrizione del componente, il prezzo, il marchio e il sito da dove è stato estratto. Nel codice sorgente delle pagine web mi sono focalizzato unicamente sulla denominazione della classe degli elementi che desideravo estrarre. Unica informazione necessaria poiché, come spiegherò nel commento del codice R, le funzioni messe a disposizione dal software statistico attraverso l'URL passato come parametro in input, ricavano il codice sorgente della pagina e ricercano al suo interno tutte le classi con la denominazione passata in input. La descrizione e il prezzo dei componenti, in base al sito web, verranno estratti con questo metodo, il marchio e il sito saranno invece ricavati dall'URL. In particolare il sito verrà identificato dal dominio, il marchio dal percorso dell'indirizzo web.

- **Estrazione della descrizione dell'articolo:**

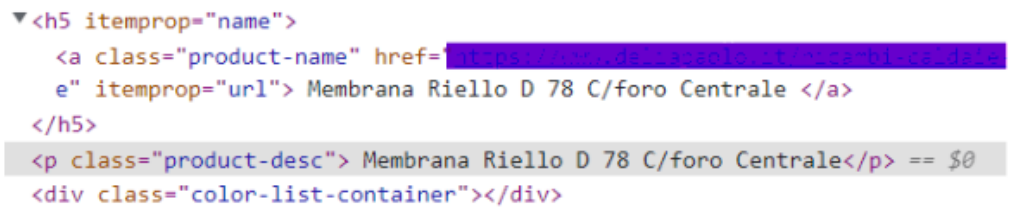


Figure 8: La descrizione del componente dal sito web verrà estratta dalla classe **product-descr**.

- **Estrazione del prezzo dell'articolo:**

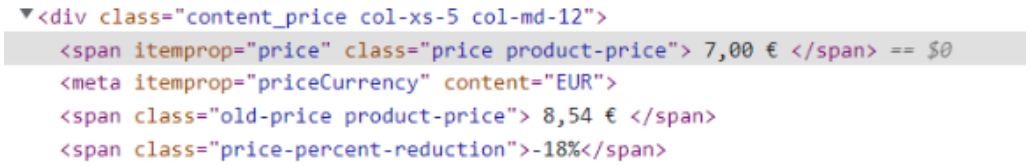


Figure 9: Il prezzo del componente dal sito web verrà estratto dalla classe **price**.

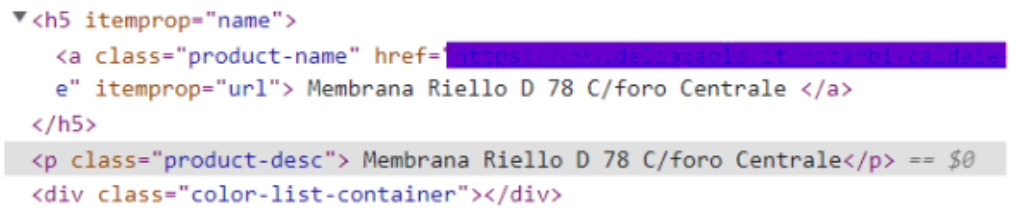
2.1.4 Codice R dell'estrazione

Una volta compresa la teoria della struttura e della meccanica di mutamenti di un sito web, si passa all'effettiva applicazione dei metodi di estrazione nel software statistico R. In questa sezione vedremo infatti come ho gestito la generazione dei link e l'effettiva rilevazione dei dati dalle pagine web. Prima di tutto, per applicare il web scraping in R è necessario installare la libreria *rvest*. Questa infatti contiene una serie di funzioni che permettono di estrarre i dati dal codice HTML della sorgente della pagina. In particolare, tra le tante presenti, utilizzerò *read_html()*. Passando come parametro l'URL della pagina web, la funzione genera come output l'intera l'intera codifica della sorgente della pagina. Per comodità si utilizza anche la funzione *html_text()* per trasformare l'output della funzione di lettura della sorgente da formato HTML in formato text.

```
URL_sito %<% read_html()%<% html_text()
```

Estrarre totalmente il codice sorgente della pagina è il primo passo per arrivare ad ottenere l'informazione che desideriamo estrarre. Necessitiamo ora di una funzione che permetta di rilevare i dati che hanno una determinata tipologia/denominazione comune. Nel sottocapitolo precedente abbiamo visto che nella codifica del sito è presente una struttura in classi degli elementi, dove ognuna è identificata da una determinata denominazione che cambia da sito a sito. La funzione in questione è *html_nodes()*. Una volta passato come parametro *".[NOME DELLA CLASSE]"*, l'utente riceverà come output la codifica HTML del codice sorgente che ha come nome classe quella inserita in input.

Nell'esempio estraiamo la descrizione di un articolo da un sito dalla classe **product-descr**.



In R:

URL_sito %<% read_html() %<% html_nodes(".product-descr") %<% html_text()

E' il momento della generazione dell'URL. Avendo appreso le sue alterazioni in base alle selezioni dell'utente all'interno della pagina, ora siamo in grado di creare un link dinamico idoneo alle nostre esigenze. In particolare, atto a permetterci di effettuare un estrazione massiva.

2.1.5 Macrocodice dell'estrazione

In questo breve capitolo andrò a indicare passo per passo il modo in cui ho codificato l'estrazione usando R studio.

Elenco dei competitors: vettore che contiene la lista dei competitors di Baxi

competitors<-c("riello","beretta","chaffoteaux","ferroli","immergas""lamborghini","junkers","sime","duval","baxi","vaillant")

Creiamo un dataset vuoto che conterrà l'URL, Descrizione e prezzo dei componenti estrazioni

data_completo<-matrix(NA,ncol=3,nrow=50000)

pos=1

Eseguo un ciclo for per ogni competitor

```
for(i in 1:length(competitors)){

  pagina=1
```

URL di partenza ***(sarà pagina 1 del primo competitors dell'elenco)***

```
  link<-paste("https://www.my.ecommerce.site.com/ricambi-ricambi-caldaie/page/",pagina,"/"?filter_marca=",competitors[i],sep="")
```

Eseguo un ciclo for per ogni pagina

```
  while(pagina<=10){          (per ogni pagina del sito relativa al competitors_i prendo le prime 10 pagine di articoli)

      link<-paste("https://www.my.ecommerce.site.com/ricambi-ricambi-caldaie/page/",pagina,"/"?filter_marca=",competitors[i],sep="")
```

Estrazione delle descrizioni **(prendiamo le descrizioni degli articoli della pagina dalla classe product-descr)**

```
      descrizione<-try(link %<% read_html() %<% html_nodes("**.product-descr**") %<% html_text(),silent = TRUE)
```

Vedo se si presentano errori ***(es. se il sito ha meno di 10 pagine di articoli)***

```
      Riga_errore<-0+length(which(testerrore=="404"))+length(which(testerrore=="open.connection"))+length(which(testerrore=="error"))

       te<-(as.tibble(str.lower(descrizione)))%<% unnest_token(output=word,input=value,token ="words")

          %<% group_by(word)%<%summarise(n=n())

       if(Riga_errore==0){          (Se non ci sono errori)
```

Estraggo il prezzo ***(prendiamo i prezzi degli articoli della pagina dalla classe price)***

```
        prezzo<-link %<% read_html() %<% html_nodes("**.price") %<% html_text()**
```

————————-PROCESSO DI PULIZIA DEI DATI ESTRATTI ————————-

Verranno usate funzioni di pulizia di caratteri speciali, eccessivi spazi su descrizione e prezzi estratti. Funzione gsub()

Aggiornamento del dataset

```
        ind<-pos:(pos+(length(descrizione)-1))

        data_completo[ind,c(1,2,3]<-c(link,descrizione,prezzo)

        pos=pos+length(descrizione)

        pagina<-pagina+1

       }else{ pagina<-pagina+1}}}
```

Salvataggio del dataset in formato .csv

```
write.csv(data_completo[non_nulle,],"Estrazione_sito1.csv")
```

2.1.6 Estrazione del dato grezzo, dataset iniziale

Una volta ultimata la processazione del codice otterremo un file in formato csv di migliaia di righe di questa struttura:

Link	Descrizione	Prezzo
https://www.sito1.com/...	Scambiatore Sanitario 16 Piastre...	61.9
https://www.sito1.com/...	Sonda Sensore Temperatura NTC...	10.91
https://www.sito1.com/...	Rubinetto Carico per caldaia...	61.9

Table 2: Estratto del file ottenuto dall'estrazione di dati dal *sito1*

L'estrazione e l'inserimento del link nel nostro dataset serviva solamente per tenere traccia del nome del sito e della marca degli articoli. Una volta ottenute queste informazioni, l'intero URL della pagina non servirà più. Con alcuni semplici passaggi effettuati manualmente su Excel, grazie alla funzione *stringa.estrai*, otteniamo un dataset più chiaro e completo:

Sito	Marca	Descrizione	Prezzo
sito1	Riello	LIMITATORE DI FLUSSO 11LT BLU	3.26
sito1	Riello	TRASFORMATORE	117.92
sito1	Riello	PANNELLO DI COMANDO	56.95

Table 3: Dataset iniziale

L'estrazione da un unico sito non è però sufficiente. Uno dei principali obiettivi di questo progetto, oltre all'analisi di mercato tra competitors, è l'analisi della popolarità dei marchi all'interno dei siti. Per questo motivo ho studiato le dinamiche di 10 siti diversi di rivendita online di ricambi originali di caldaie e, adottando le tecniche illustrate in precedenza, ho ottenuto e creato un dataset unico composto da 4 colonne (sito,marca,descrizione,prezzo) e 20700 righe circa. Il file in questione sarà il nostro punto di partenza.

2.2 Processo di classificazione del dato

Nel seguente capitolo saranno approfondite le modalità e gli strumenti per la collocazione degli articoli estratti tramite web scraping in categorie e sottocategorie. Verrà discussa la regola di classificazione su cui si basa l'algoritmo, i file utilizzati e le tecniche di analisi testuale utilizzate. Una volta ultimato questo processo otterremo il dataset definitivo pronto per effettuare le analisi.

2.2.1 Obiettivi e dati forniti

Riuscire a dare un significato e una collocazione ben precisa a componenti quando non si è esperti in quel determinato ambito non è un lavoro facile. La situazione era difficile da considerare non disponendo né di alcuna nozione riguardante i ricambi per caldaie e né di un' idea delle categorie esistenti in questo mondo. La mia conoscenza del dato da classificare, fino ad ora, si limita unicamente alla sua descrizione. In fase di estrazione, infatti, non vi era la possibilità di reperire informazioni ulteriori sulla sua tipologia. La descrizione dei dati raccolti tramite web scraping, sebbene spesso poco comprensibile e ricca di abbreviazioni, riveste un ruolo fondamentale nella categorizzazione dei dati stessi. Il mio obiettivo era quello di attribuire un'identità precisa agli articoli estratti, utilizzando al massimo l'analisi testuale delle parole presenti nella loro codifica alfanumerica. Tuttavia, questa operazione non sarebbe stata sufficiente per garantire la corretta categorizzazione degli articoli, dal momento che non ero in grado di identificare tutte le categorie pertinenti nel settore dei ricambi Baxi. Per tale motivo, il mio datore di lavoro mi ha fornito un dataset contenente un vasto elenco di articoli di ricambio Baxi già categorizzati, in modo da facilitare l'associazione dei nuovi dati alle categorie corrette.

Cod.Comp.	Descrizione	Cod.Cat.	Categoria
768176100	DOTAZ.TUBI MODULO...	C9999	NESSUN UTILIZZO IN ...
767737900	DOTAZIONE LUNA3...	C0134	RACCORDERIA GENERICO
768458500	DOTAZIONE TUBI ACQUA...	C0134	RACCORDERIA GENERICO
710573200	ASSIEME TUBO ENTRATA...	C0132	TUBO VASO ESPANSIONE

Table 4: Estratto del file Componenti.xlsx

Di questo file utilizzeremo unicamente la colonna **Descrizione**, codifica testuale dei componenti ricambio Baxi e **Categoria** la relativa categoria. L'idea è quella di trovare, per ogni articolo estratto, l'associazione più verosimile con un componente Baxi. Attraverso l'utilizzo di tecniche di Text Mining su *Descrizione* e *Categoria* di questo file, saremo in grado di raggiungere il nostro obiettivo. Per maggior chiarezza la colonna *Descrizione* verrà denominata **Sottocategoria** poiché l'associazione che andremo ad individuare tra componente Baxi e articolo estratto assume il ruolo di sottogruppo di elementi. La descrizione del componente Baxi non andrà affatto a sostituire quella del dato estratto, funge unicamente come informazione ulteriore, esattamente come se fosse una sottocategoria. Il dataset filtrato che andremo ad utilizzare sarà composto da *Categoria* e *Sottocategoria*.

Sottocategoria	Categoria
DOTAZ.TUBI MODULO SOLARE	NESSUN UTILIZZO IN GARANZIA
DOTAZIONE LUNA3 SOLAR PLUS	RACCORDERIA GENERICO
DOTAZIONE TUBI ACQUA	RACCORDERIA GENERICO
ASSIEME TUBO ENTRATA VASO ESPANS.	TUBO VASO ESPANSIONE

Table 5: Dataset filtrato

Per raggiungere il mio obiettivo e riuscire a ottenere il dataset pronto per le analisi userò una particolare tecnica di Text Mining: la Document Term Matrix (abr. "DTM").

2.2.2 Document Term Matrix

Si definisce *Document Term Matrix* (trad. Matrice documento-termine) matrice che descrive la frequenza dei termini che ricorrono in una raccolta di documenti. In genere è comune incontrare la trasposizione dove i documenti sono le colonne e le parole sono le righe. E' un particolare strumento di Text Mining disponibile in R Studio installando la libreria "tm" e eseguendo la funzione DocumentTermMatrix.
Nel nostro caso, questa tecnica ci permetterà di creare tre matrici fondamentali, la loro intersezione renderà possibile l'ottenimento di una sottocategoria e di una categoria per il componente estratto dai siti web.

La prima Document Term Matrix riguarda le sottocategorie, avremo come "documenti" (righe della DTM) le descrizioni dei componenti Baxi e come colonne l'insieme di tutte le parole degli articoli della colonna Sottocategoria della tabella *Table 2*. Discorso analogo alle altre due DTM che ho creato di cui una è inerente alla Categoria (seconda colonna della tabella *Table 2* e l'altra riguarda le descrizioni degli articoli estratti.

anello	**posteriore**	**sx**	**accessori**	**idraulici**
1	1	1	0	0
0	0	0	1	1
0	0	0	1	0
0	1	0	0	0

Table 6: Estratto della DTM delle sottocategorie

Una volta ultimata la creazione delle DTM disponiamo di tutti gli strumenti necessari per la messa in pratica delle Regola di classificazione su cui si baserà il nostro algoritmo.

2.2.3 Regola di classificazione

Per l'implementazione di un qualsiasi algoritmo, anche il più semplice, necessitiamo di uno schema logico, atto a gestire e prevedere tutte le eccezioni. Nel nostro caso, dato che dovremo processare e classificare oltre 20000 descrizioni di componenti, dovremo strutturare un piano d'azione studiato nei minimi dettagli onde evitare interruzioni dell'elaborazione ed eventuali errori. Lo schema logico di questo progetto ho deciso di denominarlo *Regola di classificazione*. Questa fonderà come principio cardine l'intersezione tra le Document Term Matrix precedentemente create. Per una più semplice comprensione, la descrizione degli step di questa regola sarà affiancata da una o più rappresentazioni grafiche. Per la spiegazione del piano di classificazione, dato che lo stesso procedimento verrà eseguito per ognuna delle righe della DTM delle estrazioni, mi limiterò nella descrizione della prima.

Il primo step della *Regola di classificazione* consiste nell'intersezione tra la DTM delle estrazioni e l'intera DTM delle sottocategorie. In particolare rileviamo dalla prima riga della matrice delle estrazioni la denominazione di colonna dei valori di celle diversi da zero. In questo modo otterremo l'elenco delle parole della nostra estrazione che saranno ricercate nella DTM delle sottocategorie.

Nell'esempio otteniamo le parole "anello", "posteriore" e "sinistro", nonché le componenti della descrizione del primo dato estratto.

anello	**posteriore**	**sx**	**accessori**	**idraulici**
1	1	1	0	0

Table 7: Estratto della prima riga della DTM delle estrazioni

Una volta aver filtrato dalla Document Term Matrix delle sottocategorie unicamente le colonne di uguale denominazione alle parole che compongono la codifica del componente estratto. Successivamente per una migliore interpretazione, integriamo alla matrice delle sottocategorie appena filtrata con l'inserimento di una prima colonna, posizionata all'inizio della tabella, con la denominazione completa della sottocategoria e di una seconda colonna alla fine con il valore della somma delle frequenze per riga. Essendo una matrice binaria la somma equivarrebbe a sapere il numero di parole che la sottocategoria ha in comune con la descrizione del primo dato estratto.

Categoria	**Sottocategoria**	**anello**	**posteriore**	**sx**	**somma**
Categoria X1	anello posteriore sx	1	1	1	3
Categoria X2	accessori idraulici 300 lt	0	0	0	0
Categoria X3	accessorio posteriore	0	1	0	1

Table 8: Estratto della matrice risultato

In prossimità del massimo della colonna troviamo la categoria e la sottocategoria più verosimile all'articolo estratto. Nell'esempio la sua descrizione è riconducibile alla sottocategoria *anello posteriore sx* e categoria *Categoria X1*. Non sempre però si ottiene il massimo in corrispondenza di un'unica sottocategoria, potrebbero esserci numerose sottocategorie che contengono la stessa frequenza di parole in comune con il dato estratto. E' una situazione abbastanza frequente, necessitiamo quindi di fare ulteriori considerazioni.

Ipotizziamo di avere come valore massimo della colonna **somma** pari a 2, in questo caso filtriamo la matrice precedente tenendo unicamente le righe con valore pari a 2 nell'ultima colonna.

Categoria	**Sottocategoria**	**anello**	**posteriore**	**sx**	**somma**
Categoria X1	*Sottocategoria Y1*	1	1	0	**2**
Categoria X2	*Sottocategoria Y2*	1	0	1	**2**
Categoria X3	*Sottocategoria Y3*	0	1	1	**2**

Table 9: Caso di molteplicità di soluzione

La scelta di quale sottocategoria delle rimanenti assegnare al dato non può avvenire in maniera del tutto casuale. In questa eccezione svolgerà infatti un ruolo fondamentale la terza matrice creata, la DTM delle categorie. Per trovare o quantomeno restringere la mole di papabili sottocategorie, analizziamo le frequenze di parole della descrizione del dato estratto ricorrenti nelle categorie. Attraverso un processo analogo alle sottocategorie, troveremo la categoria più verosimile tra quelle rimaste. Filtriamo per riga in base alle categorie rimaste nella matrice *Table 6* e per colonna in base alle parole della descrizione della prima estrazione. Successivamente aggiungiamo a questa matrice filtrata due ulteriori colonne. La prima che conterrà la somma di tutte le frequenze per riga e la seconda che, anch'essa per riga, conta il numero di frequenze diverse da zero così da capire quante parole della descrizione sono presenti in quella categoria. Otterremo una matrice del tipo:

Categoria	**anello**	**posteriore**	**sx**	**Somma**	**Conta**
Categoria X1	10	1	17	28	3
Categoria X2	17	18	0	35	2
Categoria X3	1	1	3	5	3

Table 10: Matrice delle categorie

Analogamente a prima scegliamo come categoria definitiva quella che ha valore massimo nella colonna **Conta**. Se il massimo non si trovasse in corrispondenza di un'unica riga, faremo riferimento alla colonna **Somma**. Sceglieremo infatti come categoria definitiva quella che ha, come somma delle frequenze di parole in comune con il dato estratto, valore massimo.

Se si dovesse verificare un ulteriore ambiguità cioè non avendo ancora la certezza di avere una e unica categoria definitiva, la scelta di quest'ultima ricadrà in maniera casuale. Analogamente a prima scegliamo come categoria definitiva quella che ha valore massimo nella colonna **Conta**. Se il massimo non si trovasse in corrispondenza di un'unica riga, faremo riferimento alla colonna **Somma**. Sceglieremo infatti come categoria definitiva quella che ha, come somma delle frequenze di parole in comune con il dato estratto, valore massimo. Se si dovesse verificare un ulteriore ambiguità cioè non avendo ancora la certezza di avere una e unica categoria definitiva, la scelta di quest'ultima ricadrà in maniera casuale. Analogamente a prima scegliamo come categoria definitiva quella che ha valore massimo nella colonna **Conta**. Se il massimo non si trovasse in corrispondenza di un'unica riga, faremo riferimento alla colonna **Somma**. Sceglieremo infatti come categoria definitiva quella che ha, come somma delle frequenze di parole in comune con il dato estratto, valore massimo. Se si dovesse verificare un ulteriore ambiguità cioè non avendo ancora la certezza di avere una e unica categoria definitiva, la scelta di quest'ultima ricadrà in maniera casuale. Analogamente a prima scegliamo come categoria definitiva quella che ha valore massimo nella colonna **Conta**. Se il massimo non si trovasse in corrispondenza di un'unica riga, faremo riferimento alla colonna **Somma**. Sceglieremo infatti come categoria definitiva quella che ha, come somma delle frequenze di parole in comune con il dato estratto, valore massimo. Se si dovesse verificare un ulteriore ambiguità cioè non avendo ancora la certezza di avere una e unica categoria definitiva, la scelta di quest'ultima ricadrà in maniera casuale. Analogamente a prima scegliamo come categoria definitiva quella che ha valore massimo nella colonna **Conta**. Se il massimo non si trovasse in corrispondenza di un'unica riga, faremo riferimento alla colonna **Somma**. Sceglieremo infatti come categoria definitiva quella che ha, come somma delle frequenze di parole in comune con il dato estratto, valore massimo. Se si dovesse verificare un ulteriore ambiguità cioè non avendo ancora la certezza di avere una e unica categoria definitiva, la scelta di quest'ultima ricadrà in maniera casuale. Analogamente a prima scegliamo come categoria definitiva quella che ha valore massimo nella colonna **Conta**. Se il massimo non si trovasse in corrispondenza di un'unica riga, faremo riferimento alla colonna **Somma**. Sceglieremo infatti come categoria definitiva quella che ha, come somma delle frequenze di parole in comune con il dato estratto, valore massimo.

Se si dovesse verificare un ulteriore ambiguità cioè non avendo ancora la certezza di avere una e unica categoria definitiva, la scelta di quest'ultima ricadrà in maniera casuale.

Una volta aver ultimato la scelta della categoria definitiva, filtriamo la matrice delle sottocategorie *Table 6* tenendo unicamente le sottocategorie che rientrano in quella categoria. Ipotizziamo di aver ottenuto come categoria definitiva per il dato finale la categoria *Categoria X1*.

Categoria	**Sottocategoria**	**anello**	**posteriore**	**sx**	**Somma**
Categoria X1	*Sottocategoria Y1*	1	1	0	2

Table 11: Caso di unicità di soluzione

Nella maggior parte dei casi, come nell'esempio in figura, il risultato che otteniamo è un'unica riga con rispettivamente categoria e sottocategoria definitiva per il dato estratto. In rare situazioni può capitare di ottenere nuovamente più di un risultato. Per eliminare ogni ambiguità la scelta della sottocategoria verrà fatta casualmente.

Applicando questa regola per ogni riga della Document Term Matrix delle estrazioni, otterremo il dataset iniziale su cui basare le nostre analisi.

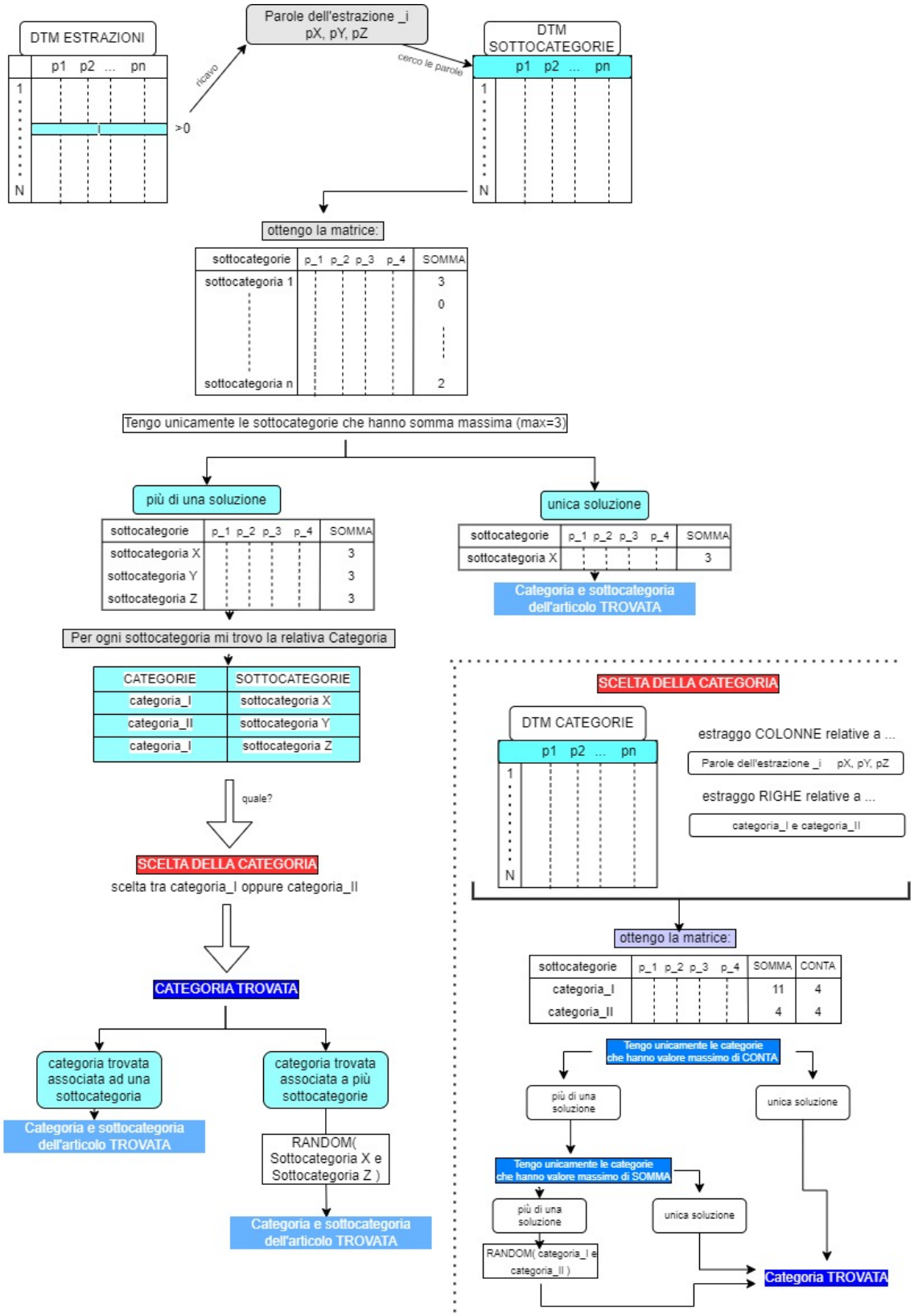


Figure 10: Schema logico della regola di classificazione

2.3 Valutazione della bontà di classificazione

In questo paragrafo andremo ad analizzare e a dare un giudizio alla classificazione dei ricambi di caldaie estratti dai siti web. L'algoritmo di classificazione ci ha permesso di selezionare per ciascuno dei componenti estratti una categoria e una sottocategoria, sarà nostro dovere ora valutare quanto bene va questo algoritmo e quanto sia affidabile. Ovviamente per tempi e costi è impossibile trovare la vera categoria e sottocategoria di ogni componente, dovremo prendere un campione. Sarà nostro compito quindi prelevare un campione e valutare su questo la bontà di classificazione unicamente delle categorie.

2.3.1 Piano di campionamento

Prima di illustrare il piano di campionamento per l'ottenimento di un campione di estrazioni è necessario conoscere in linea teorica i piani di campionamento utilizzati nel progetto.

- **Campionamento multistadio**

  Il campionamento multistadio (o a più stadi) è un metodo di campionamento che viene spesso utilizzato in situazioni in cui è necessario campionare una popolazione che è suddivisa in unità di campionamento più piccole, ad esempio in un'indagine demografica in cui i dati vengono raccolti su base regionale. In un campionamento multistadio, la popolazione viene suddivisa in gruppi più piccoli, ad esempio in regioni geografiche. Viene quindi selezionato un campione di regioni, ad esempio attraverso un campionamento casuale semplice o stratificato, e successivamente in ognuna di queste regioni viene selezionato un campione di unità di campionamento più piccole, ad esempio delle famiglie o degli individui. Il vantaggio del campionamento multistadio è che consente di ridurre i costi e il tempo necessari per condurre il sondaggio, in quanto il campionamento viene effettuato su unità di campionamento più piccole rispetto all'intera popolazione. Inoltre, il campionamento multistadio può anche consentire di ridurre la varianza del campione, se il campionamento viene effettuato in modo adeguato.

  Nel nostro caso applicheremo un campionamento multistadio a tre stadi: categorie, sottocategorie e componenti estratti.

- **Campionamento a probabilità variabili**

  Un piano di campionamento a probabilità variabili è una tecnica di campionamento che assegna una probabilità di selezione diversa a ciascun elemento dell'unità di campionamento. In altre parole, gli elementi dell'unità di campionamento non hanno tutti la stessa probabilità di essere selezionati nel campione. Il piano di campionamento a probabilità variabili è spesso utilizzato quando l'unità di campionamento è suddivisa in categorie e si desidera selezionare un numero diverso di elementi da ciascuna categoria.

  Questo campionamento verrà utilizzato per il primo e il secondo stadio, rispettivamente per Categoria a Sottocategoria.

- **Campionamento casuale semplice (CCS)**

  Il campionamento casuale semplice è una tecnica di campionamento in cui ogni elemento dell'unità di campionamento ha la stessa probabilità di essere selezionato nel campione. In altre parole, ogni elemento è selezionato in modo casuale e indipendente dagli altri elementi.l campionamento casuale semplice è spesso considerato uno dei metodi più rigorosi di campionamento, in quanto garantisce che ogni elemento dell'unità di campionamento abbia la stessa probabilità di essere selezionato. Questo campionamento verrà utilizzato per l'ultimo stadio: i componenti del dataset.

- **Applicazione del piano di campionamento**

  Come accennato in precedenza utilizzeremo un campionamento a tre stadi. Il primo stadio è la categoria a cui verrà applicato un piano di campionamento a probabilità variabili. Ho selezionato una probabilità di estrazione per ciascuna delle 108 categorie basata sul numero di componenti del dataset in ognuna delle categorie.

CATEGORIE	mi	Pi = Mi / M
Categoria 1	10	10 / M = P1
Categoria 2	20	20 / M = P2
Categoria 3	12	12 / M = P3
Categoria ...	...	...
Somma	**1000 (M)**	**1**

Table 12: Probabilità di estrazione di ogni categoria

Il mio obiettivo è quello di estrarre un campione di 1000 componenti. Per la generazione di quanti componenti estrarre da ogni categoria mi sono affidato ad una tecnica computazionale che si basa sulla generazione di valori con supporto discreto e finito.

```
prob <- c(P1, P2, P3, ...)

categorie<-c(1:108)

xx<- sample(categorie, size = 1000, replace = TRUE, prob = prob)

table(xx)
```

CATEGORIA 1	CATEGORIA 2	CATEGORIA 3	CATEGORIA ...
n1	n2	n3	...

Table 13: Numero di componenti da estrarre da ogni categoria

Il secondo stadio è la sottocategoria. Anche in questo caso utilizzeremo un piano di campionamento a probabilità variabili. Eseguo un ciclo per ogni categoria. Dentro ognuna dovranno essere scelte le sottocategorie,in maniera proporzionale al numero di componenti presenti in quella sottocategoria, da cui poi estrarre le osservazioni e creare il campione.

Ipotizziamo di essere nella seconda categoria e che essa è composta da 3 sottocategorie:

CATEGORIA x	**mi**	**pi = mi / nx**
Sottocategoria 1	16	16 / nx = p1
Sottocategoria 2	20	20 / nx = p2
Sottocategoria 3	10	10 / nx = p3
Sottocategoria ...	...	...
Somma	**nx**	**1**

Table 14: Probabilità di estrazione di ogni sottocategoria all'interno della categoria x

Analogamente a prima affidandoci alla tecnica computazionale che si basa sulla generazione di valori con supporto discreto e finito otteniamo il numero di osservazioni da prelevare da ciascuna sottocategoria.

SOTTOCATEGORIA 1	**SOTTOCATEGORIA 2**	**SOTTOCATEGORIA ...**
s1	s2	...

Table 15: Numero di componenti da estrarre da ogni sottocategoria

Infine abbiamo l'ultimo stadio, i componenti. I componenti verranno estratti usando il campionamento casuale semplice senza reinserimento dove ogni osservazione ha la stessa probabilità di essere estratta. Utilizzando la funzione sample() con la particolare opzione di replace=F.

2.3.2 Ottenimento del campione e Matrice di confusione

Una volta applicato il piano di campionamento all'intero elenco dei componenti, otteniamo un campione abbastanza rappresentativo della realtà. La mia idea è quella di ottenere una lista di 1000 unità dove, con l'aiuto dei colleghi di Baxi (persone da anni nel settore di ricambi caldaie), andrò a trovare le vere categorie dei componenti in questione. Per ogni elemento del campione avrò dunque una *Categoria generata dall'algoritmo* e una *Categoria vera*. Otteniamo dunque come dataset composto da queste colonne per la valutazione della bontà di classificazione.

CATEGORIE VERE	CATEGORIE ALGORITMO
CategoriaV X	*CategoriaA Z*
CategoriaV Y	*CategoriaA Z*
CategoriaV Z	*CategoriaA Z*

Table 16: Estratto del campione

La classificazione che è stata fatta in questo progetto, dato che non vengono valutate unicamente due classi, è una **Multiclass classification**. La classificazione multiclasse è una tecnica di apprendimento automatico che viene utilizzata per classificare gli oggetti in più di due classi. Nel nostro caso questo tipo di classificazione ci consente di assegnare ad un componente una delle molte possibili classi cioè le categorie.
Per la valutazione di quanto bene abbiamo classificato i vari articoli necessitiamo di uno strumento molto importante: la *Matrice di misclassificazione*.

La ***matrice di misclassificazione*** , o *matrice di errore* o *matrice di confusione*, è uno strumento utilizzato per valutare le prestazioni di un modello di classificazione, e permette di stimare quanto bene il modello classifica correttamente le diverse categorie di oggetti. In particolare ci da l'idea di quante volte il modello ha predetto correttamente o erroneamente l'appartenenza di un componente a una specifica classe.

CATEGORIE	*Categoria 1*	*Categoria 2*	*Categoria 3*
Categoria 1	0.02	0	0
Categoria 2	0	0.015	0
Categoria 3	0.001	0	0.012

Table 17: Estratto della matrice di confusione

Nel nostro caso per ottenere questa matrice dovremo restringere il nostro dataset poichè necessitiamo di avere le stesse categorie in entrambe le colonne. Le categorie vere saranno in numero uguale a quelle stabilite dall'algoritmo. Questo perchè la matrice di misclassificazione dev'essere quadrata. Il dataset ottenuto da questa cernita di osservazioni si ristringe di circa 50 componenti.

2.3.3 Valutazione della classificazione

La matrice di confusione permette di ottenere facilmente diverse metriche su cui basarsi per valutare la bontà di classificazione. In particolare analizzeremo: il Tasso di errore, l'accuratezza, la precisione, la recall e il punteggio F. Sarà inoltre accennata anche l'AUC, indice derivante dalla creazione della curva ROC.

- **Error rate**

  Il primo indice che andremo ad analizzare è il *tasso di errore*. L'error rate è una metrica che misura la percentuale di errore delle previsione sul totale delle istanze. Sommando tutti i valori della matrice di misclassificazione non presenti nella sua diagonale otteniamo il valore di questo indice.
  Per il campione considerato:

$$ERR = 0.1629863 \approx 16\%$$

  Questo valore ci permette di dire che l'algoritmo di classificazione classifica in maniera errata il 16% circa dei componenti. Ciò significa che circa 2 componenti su 10 l'algoritmo li classifica in maniera errata.

- **Accuracy**

  L'accuratezza è una metrica che misura la percentuale di elementi classificati correttamente dal modello di classificazione. Questo indice si calcola o sommando la diagonale della matrice di misclassificazione oppure:

$$ACC = 1 - ERR = 1 - 0.1629863 = 0.8370137 \approx 84\%$$

  Il nostro algoritmo di classificazione classifica correttamente circa l'84% dei componenti del campione in esame. Questo indice, insieme a l'Error rate, permette di avere una visione generale di quanto è buona la classificazione. Con le prossime metriche si andrà più nel particolare, otterremo infatti un singolo valore per ogni categoria.

- **Precision**

  La precisione (precision) è una metrica di valutazione delle prestazioni di un modello di classificazione multiclasse, che misura la frazione di istanze classificate come positive che sono effettivamente positive rispetto a tutte le istanze classificate come positive, indipendentemente dalla loro vera classe. Nel nostro caso abbiamo 61 classi (categorie). Per ogniuna di esse il valore di precision è pari a:

  $$precisione_i = \frac{TP_i}{TP_i + FP_i}$$

  dove TP_i rappresenta il numero di casi correttamente classificati come appartenenti alla classe i (veri positivi) e FP_i rappresenta il numero di casi erroneamente classificati come appartenenti alla classe i (falsi positivi). In una matrice di confusione multiclasse, l'indice di precisione può essere calcolato come la media pesata delle precisioni delle singole classi:

  $$precisione = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i} = 0.7912403 \approx 79\%$$

  dove N rappresenta il numero totale di classi. Il valore di precision pari a 79% significa che in media il 8 istanze su 10 vengono classificate correttamente dall'algoritmo di classificazione. Per citare un esempio di interpretazione consideriamo la categoria *guarnizione*. Su tutti i componenti del campione che il modello ha classificato come appartenenti alla classe *guarnizione*, l'86% di essi sono effettivamente corretti. L'algoritmo distingue abbastanza correttamente la categoria *guarnizione* dalle altre classi.

- **Recall**

  La recall, anche chiamata sensibilità o true positive rate, è una metrica di valutazione di un modello di classificazione che indica la capacità del modello di identificare correttamente le istanze positive (es. campioni che appartengono a una specifica classe) rispetto al totale di istanze positive presenti nel dataset.

  $$Recall(sensibilita') = \frac{TP}{TP + FN}$$

  Nel nostro caso la Recall sarà un vettore di 61 valori, uno per ogni categoria/classe. Per comodità ho rappresentato in un grafico i valori di questo indice per ogni categoria.

Come vediamo dal grafico notiamo che la maggior parte dei componenti all'interno della categorie viene categorizzato correttamente. Riprendendo l'esempio di prima, possiamo constatare che l'algoritmo ha classificato correttamente il 77% dei componenti appartenenti alla classe *guarnizione*.

- **F1 score**

  Nella classificazione multiclasse, l'F1 score è una metrica che combina precisione e richiamo per ogni classe. L'F1 score per una classe specifica viene calcolato come la media armonica di precisione e recall. Questo indice infatti bilancia queste due metriche, dando maggior peso alle classi che hanno sia precisione che richiamo elevato.

$$F1\ score = \frac{1}{C}\sum_{i=1}^{C}\frac{2 \times precision_i \times recall_i}{precision_i + recall_i}$$

  Dove C è il numero di classi, $precision\_i$ è la precisione per la classe i e $recall\_i$ è il richiamo per la classe i. Il valore di F1 score varia tra 0 (peggior risultato possibile) e 1 (risultato perfetto). Un punteggio F1 elevato indica che il modello ha una buona capacità di classificare correttamente gli elementi in ogni classe.

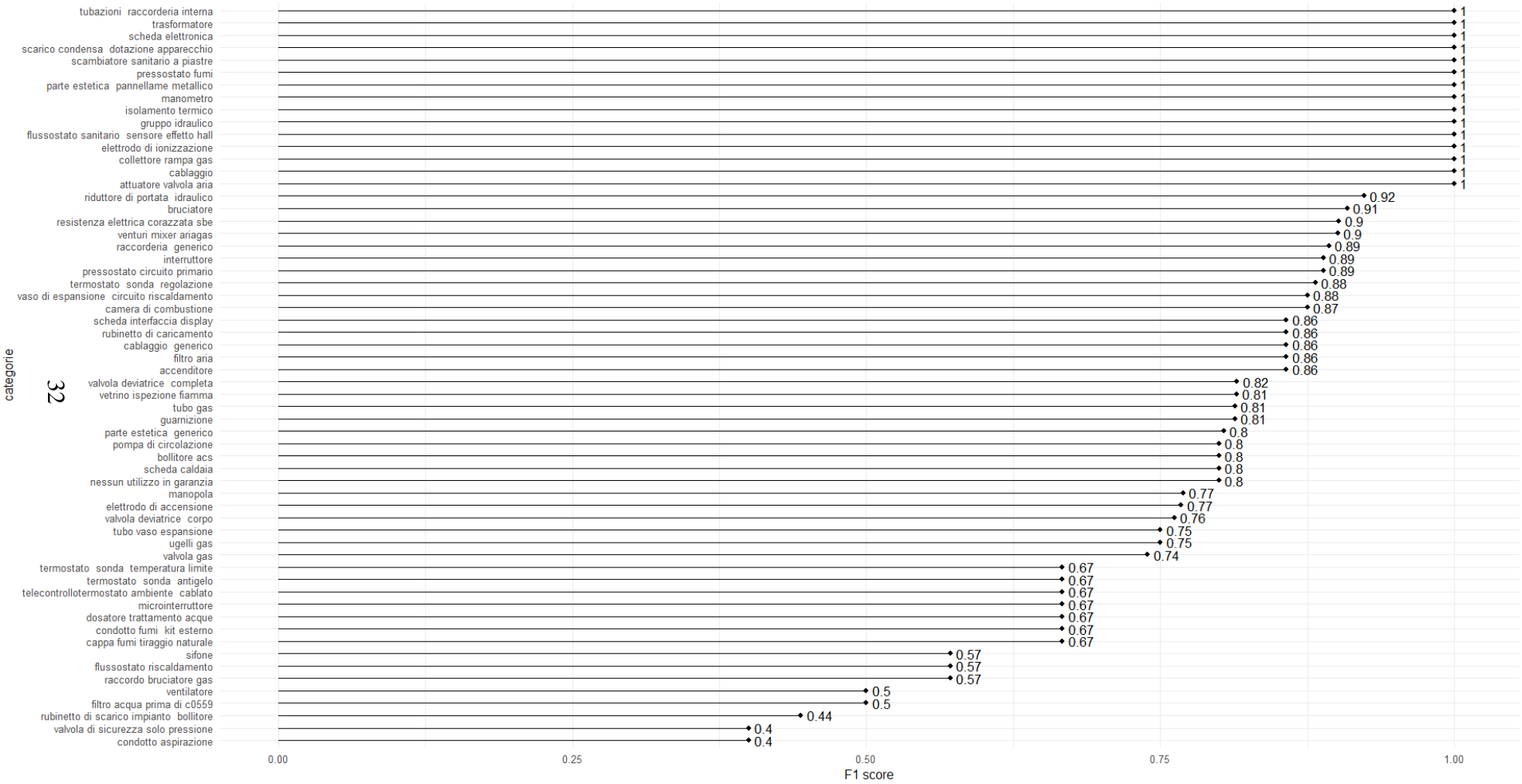
  Per la categoria *guarnizione* abbiamo un valore di F1 score di 0.81. Questo valore ci permette di dire che, per il campione considerato, l'algoritmo di classificazione ha ottenuto un buon equilibrio tra precisione e recall per questa categoria. L'algoritmo di clustarizzazione è stato in grado di identificare correttamente la maggior parte delle istanze appartenenti alla categoria *guarnizione* (valore alto di recall) e di evitare di classificare erroneamente altre istanze come appartenenti alla categoria *guarnizione* (valore alto di precisione), il che si traduce in un punteggio alto valore di F1.

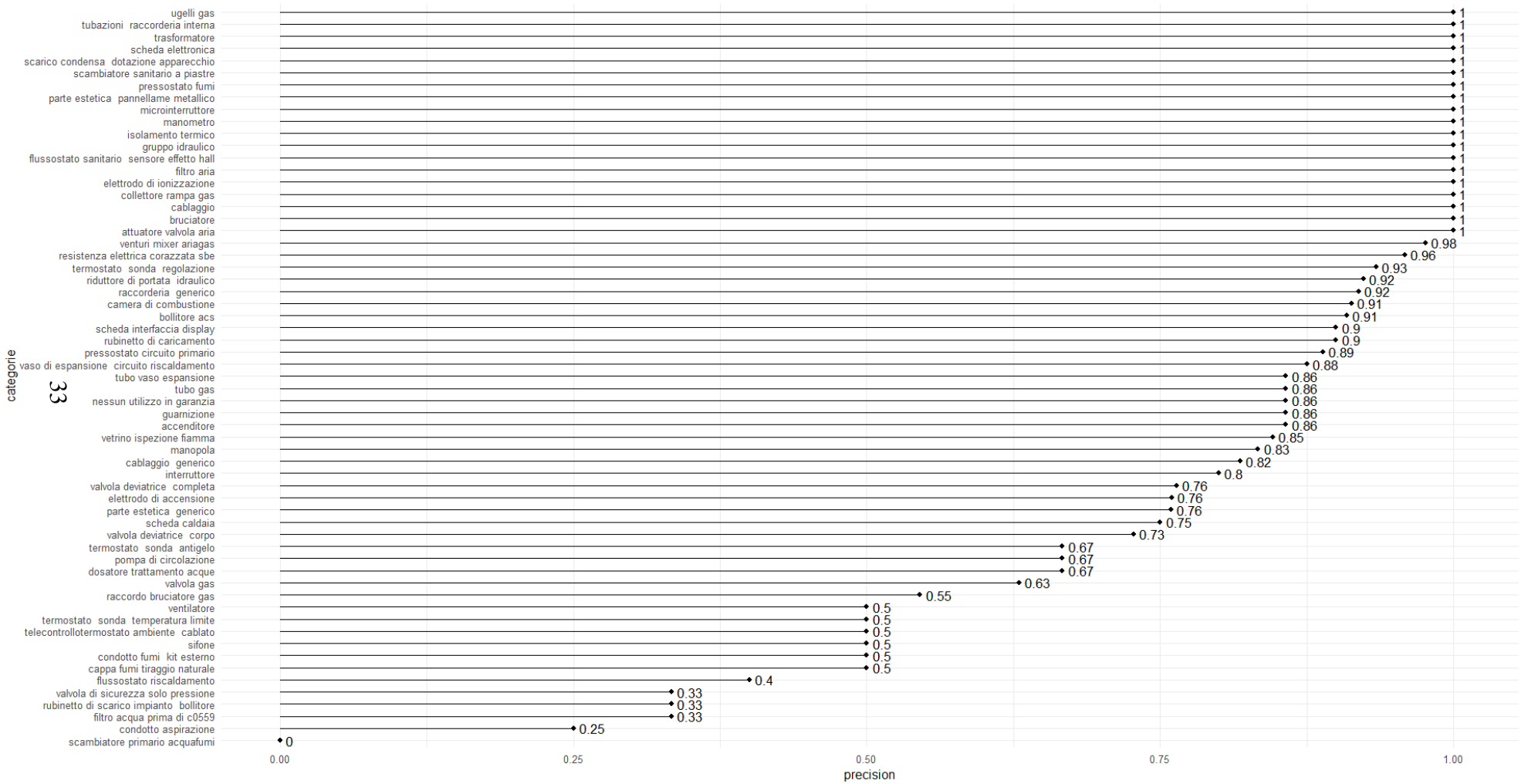
- **AUC**

  Un altro valore da tenere in considerazione per quanto riguarda la classificazione multiclasse, seppur non ottenuto dalla matrice di misclassificazione, è il valore dell'**AUC**. L'AUC (Area Under the Curve) è una misura comune della qualità del modello di classificazione e rappresenta l'area sottesa dalla curva ROC (Receiver Operating Characteristic). L'intervallo di valori di AUC varia da 0 a 1, dove un valore di 1 indica un modello perfetto e un valore di 0.5 indica un modello completamente casuale.

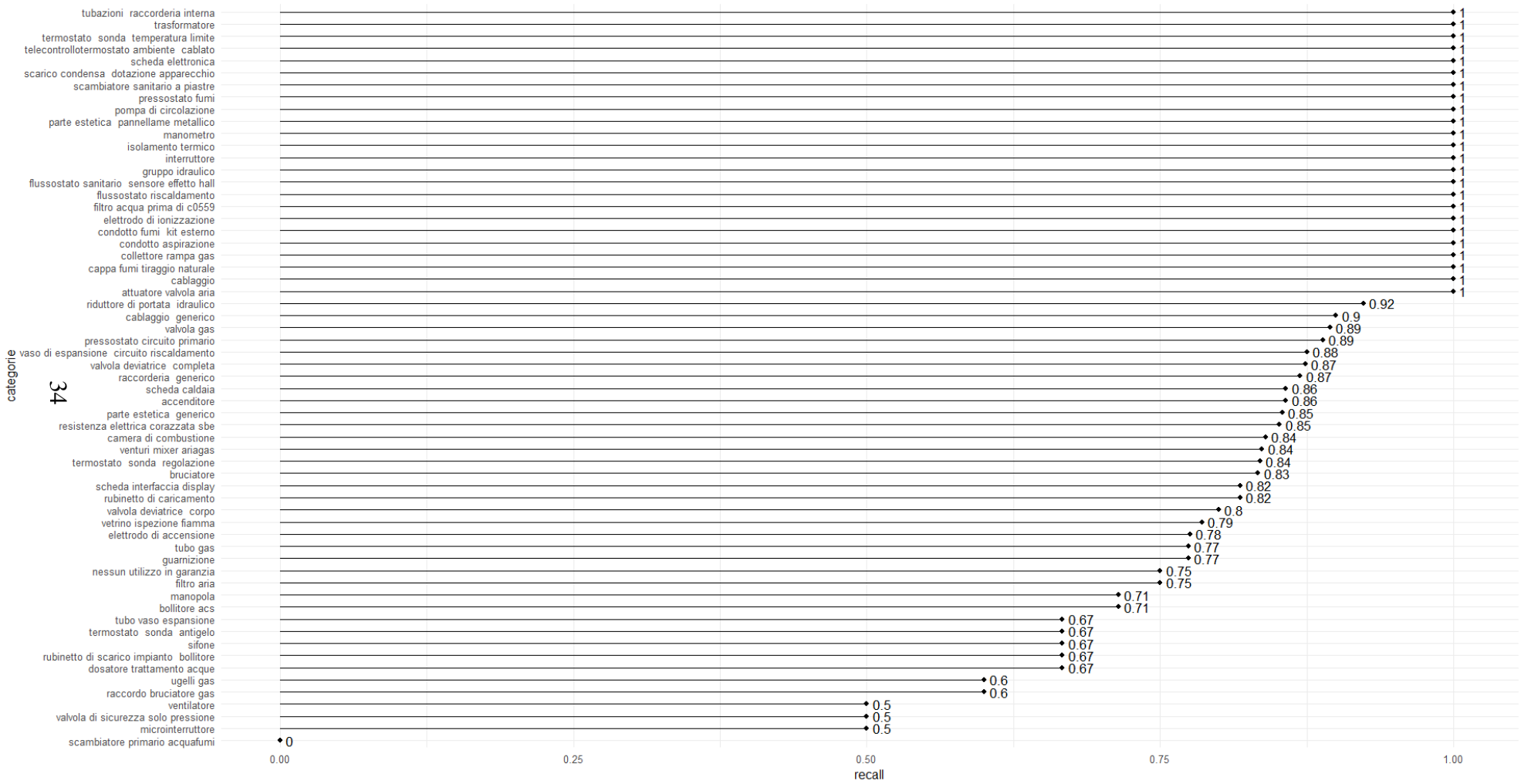
  Nel nostro caso il valore di AUC è pari a 0.9138. Questo suggerisce che l'algoritmo di classificazione, per il campione considerato, ha una buona capacità di distinguere le categorie.

In conclusione, alla luce dei risultati degli indici possiamo constatare che l'algoritmo di classificazione ha classificato in maniera molto buona. Ovviamente non possiamo averne la più completa certezza poichè l'analisi della bontà di classificazione è stata eseguita su un campione che, sebbene abbastanza rappresentativo, non rappresenta l'intera popolazione.


tubazioni raccorderia interna 1
trasformatore 1
scheda elettronica 1
scarico condensa dotazione apparecchio 1
scambiatore sanitario a piastre 1
pressostato fumi 1
parte estetica pannellame metallico 1
manometro 1
isolamento termico 1
gruppo idraulico 1
flussostato sanitario sensore effetto hall 1
elettrodo di ionizzazione 1
collettore rampa gas 1
cablaggio 1
attuatore valvola aria 1
riduttore di portata idraulico 0.92
bruciatore 0.91
resistenza elettrica corazzata sbe 0.9
venturi mixer ariagas 0.9
raccorderia generico 0.89
interruttore 0.89
pressostato circuito primario 0.89
termostato sonda regolazione 0.88
vaso di espansione circuito riscaldamento 0.88
camera di combustione 0.87
scheda interfaccia display 0.86
rubinetto di caricamento 0.86
cablaggio generico 0.86
filtro aria 0.86
accenditore 0.86
valvola deviatrice completa 0.82
vetrino ispezione fiamma 0.81
tubo gas 0.81
guarnizione 0.81
parte estetica generico 0.8
pompa di circolazione 0.8
bollitore acs 0.8
scheda caldaia 0.8
nessun utilizzo in garanzia 0.8
manopola 0.77
elettrodo di accensione 0.77
valvola deviatrice corpo 0.76
tubo vaso espansione 0.75
ugelli gas 0.75
valvola gas 0.74
termostato sonda temperatura limite 0.67
termostato sonda antigelo 0.67
telecontrollotermostato ambiente cablato 0.67
microinterruttore 0.67
dosatore trattamento acque 0.67
condotto fumi kit esterno 0.67
cappa fumi tiraggio naturale 0.67
sifone 0.57
flussostato riscaldamento 0.57
raccordo bruciatore gas 0.57
ventilatore 0.5
filtro acqua prima di c0559 0.5
rubinetto di scarico impianto bollitore 0.44
valvola di sicurezza solo pressione 0.4
condotto aspirazione 0.4
categorie
32
0.00
0.25
0.50
0.75
1.00
F1 score


categorie
precision
0.00
0.25
0.50
0.75
1.00
ugelli gas 1
tubazioni raccorderia interna 1
trasformatore 1
scheda elettronica 1
scarico condensa dotazione apparecchio 1
scambiatore sanitario a piastre 1
pressostato fumi 1
parte estetica pannellame metallico 1
microinterruttore 1
manometro 1
isolamento termico 1
gruppo idraulico 1
flussostato sanitario sensore effetto hall 1
filtro aria 1
elettrodo di ionizzazione 1
collettore rampa gas 1
cablaggio 1
bruciatore 1
attuatore valvola aria 1
venturi mixer ariagas 0.98
resistenza elettrica corazzata sbe 0.96
termostato sonda regolazione 0.93
riduttore di portata idraulico 0.92
raccorderia generico 0.92
camera di combustione 0.91
bollitore acs 0.91
scheda interfaccia display 0.9
rubinetto di caricamento 0.9
pressostato circuito primario 0.89
vaso di espansione circuito riscaldamento 0.88
tubo vaso espansione 0.86
tubo gas 0.86
nessun utilizzo in garanzia 0.86
guarnizione 0.86
accenditore 0.86
vetrino ispezione fiamma 0.85
manopola 0.83
cablaggio generico 0.82
interruttore 0.8
valvola deviatrice completa 0.76
elettrodo di accensione 0.76
parte estetica generico 0.76
scheda caldaia 0.75
valvola deviatrice corpo 0.73
termostato sonda antigelo 0.67
pompa di circolazione 0.67
dosatore trattamento acque 0.67
valvola gas 0.63
raccordo bruciatore gas 0.55
ventilatore 0.5
termostato sonda temperatura limite 0.5
telecontrollotermostato ambiente cablato 0.5
sifone 0.5
condotto fumi kit esterno 0.5
cappa fumi tiraggio naturale 0.5
flussostato riscaldamento 0.4
valvola di sicurezza solo pressione 0.33
rubinetto di scarico impianto bollitore 0.33
filtro acqua prima di c0559 0.33
condotto aspirazione 0.25
scambiatore primario acquafumi 0

33


tubazioni raccorderia interna
trasformatore
termostato sonda temperatura limite
telecontrollotermostato ambiente cablato
scheda elettronica
scarico condensa dotazione apparecchio
scambiatore sanitario a piastre
pressostato fumi
pompa di circolazione
parte estetica pannellame metallico
manometro
isolamento termico
interruttore
gruppo idraulico
flussostato sanitario sensore effetto hall
flussostato riscaldamento
filtro acqua prima di c0559
elettrodo di ionizzazione
condotto fumi kit esterno
condotto aspirazione
collettore rampa gas
cappa fumi tiraggio naturale
cablaggio
attuatore valvola aria
riduttore di portata idraulico
cablaggio generico
valvola gas
pressostato circuito primario
vaso di espansione circuito riscaldamento
valvola deviatrice completa
raccorderia generico
scheda caldaia
accenditore
parte estetica generico
resistenza elettrica corazzata sbe
camera di combustione
venturi mixer ariagas
termostato sonda regolazione
bruciatore
scheda interfaccia display
rubinetto di caricamento
valvola deviatrice corpo
vetrino ispezione fiamma
elettrodo di accensione
tubo gas
guarnizione
nessun utilizzo in garanzia
filtro aria
manopola
bollitore acs
tubo vaso espansione
termostato sonda antigelo
sifone
rubinetto di scarico impianto bollitore
dosatore trattamento acque
ugelli gas
raccordo bruciatore gas
ventilatore
valvola di sicurezza solo pressione
microinterruttore
scambiatore primario acquafumi
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
0.92
0.9
0.89
0.89
0.88
0.87
0.87
0.86
0.86
0.85
0.85
0.84
0.84
0.84
0.83
0.82
0.82
0.8
0.79
0.78
0.77
0.77
0.75
0.75
0.71
0.71
0.67
0.67
0.67
0.67
0.67
0.6
0.6
0.5
0.5
0.5
0
0.00
0.25
0.50
0.75
1.00
recall
categorie

3 Fase di analisi

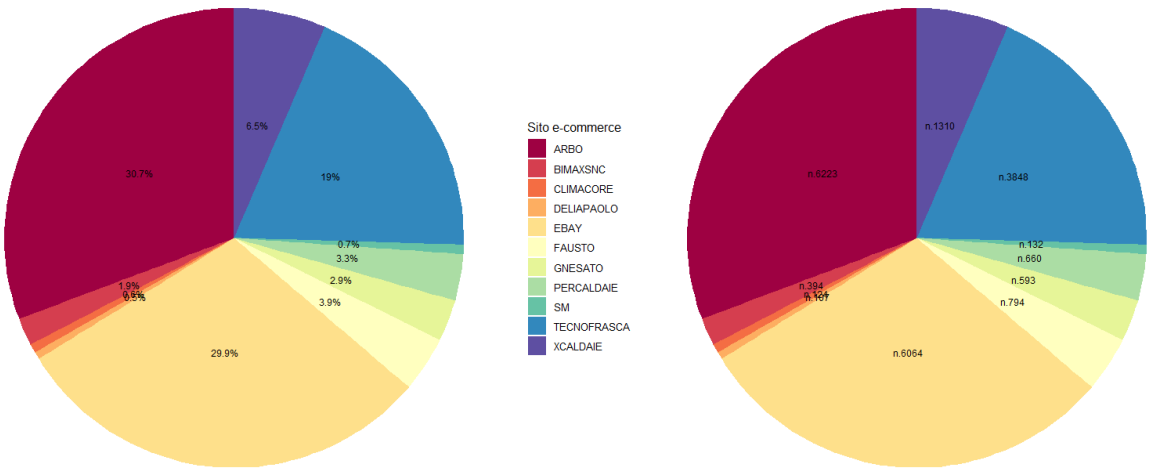
Per analizzare la mole di dati disponibili, è importante considerare diversi fattori, tra cui la qualità dei dati e la loro dimensione. Nel caso specifico, il dataset originale era composto da 20700 componenti, ma alcuni di essi non avendo una categoria e sottocategoria associata, sono stati esclusi. Questo ha portato all'ottenimento di un file di 20249 articoli. In generale, una volta definiti i dati da analizzare, il lavoro del Data Analyst consiste nel sfruttare al massimo il dataset a propria disposizione e ricavare più informazioni interessanti possibili. Sarà quindi necessaria, per comprendere la natura delle variabili considerate, un'accurata analisi esplorativa. Solamente dopo uno studio preliminare dei dati a disposizione, si potrà procedere ad analisi più complesse. In particolare sarà effettuata un'analisi di identificazione del marchio nel mercato e un analisi di prezzi di un determinato componente: la scheda elettronica.

3.1 Analisi esplorative

L'analisi esplorativa delle variabili considerate è molto importante. Nel dataset considerato sono presenti 7 variabili: *Sito, Compagnia, Marca, Categoria, Sottocategoria, Descrizione e Prezzo*. Per questa analisi preliminare saranno valutate le prime 5.

- **Variabile *SITO***

  Il ***Sito*** è una variabile qualitativa nominale con supporto finito composto da 11 elementi. Rappresenta il sito e-commerce da cui è stata prelevata l'unità statistica in fase di estrazione.



Nella rappresentazione grafica ho riportato la distribuzione delle frequenze (e relative frequenze percentuali) del supporto della variabile *Sito* all'interno del dataset. I siti ,che popolano il grafico a torta, sono i principali e-commerce di componenti e ricambi originali di caldaie. Facendo una rapida ricerca su Google (es. *"ricambi caldaie originali"*), tra le prime dieci soluzioni proposte, compariranno almeno cinque di questi siti. Analizzando gli spicchi del grafico, notiamo che la maggior parte dei componenti estratti (circa 16100 che corrispondono al 79.6% dell'intero dataset) sono stati rilevati dai siti *Ebay, Arbo e Tecnofrasca*. Come vedremo nelle analisi successive, questi tre siti si dimostreranno i principali su cui affidarsi in caso di ricerca di un determinato ricambio. Successivamente troviamo con una percentuale tra il 2 e il 4% l'uno, i siti *Climacore, Faustoricambi, Gnesato e Ricambipercaldaie*. I siti *smricambi, Climacore e Deliapaolo* che, con meno di 200 componenti l'uno, compongono il restante 1.8% del dataset.

- **Variabili *MARCA* e *COMPAGNIA***

  La variabile ***MARCA*** è una variabile qualitativa nominale con supporto composto da 12 elementi. Come da denominazione, corrisponde alla marca del componente estratto. Nel nostro contesto questa variabile assumerà successivamente il nome di *COMPETITOR*, poichè comprende effettivi rivali e concorrenti di Baxi nel mercato ricambi originali. I competitors su cui mi sono incentrato sono *Ariston, Chaffoteaux, Junkers, Ferroli, Immergas, Lamborghini, Beretta, Riello, Sime, Vaillant, Hermann Saunier Duval*e infine *Baxi*. In aggiunta ho inserito nel supporto un ulteriore elemento "Altro". Questo comprende marchi o poco conosciuti oppure di cui ho ottenuto troppe poche informazioni (es. marchio Caleffi per citarne uno).

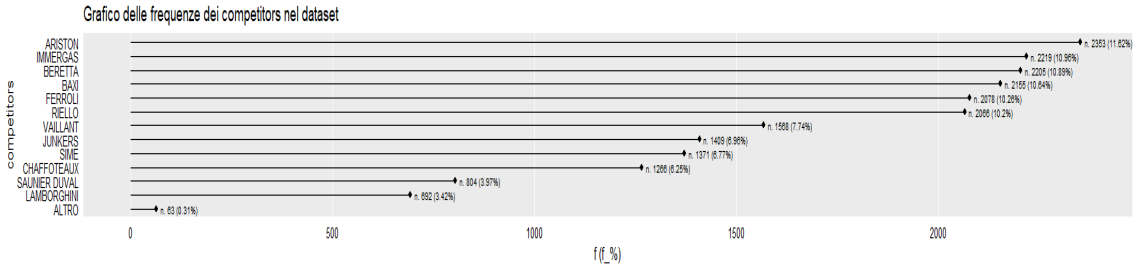
  La variabile ***COMPAGNIA*** è anch'essa una variabile qualitativa nominale e corrisponde al gruppo aziendale di cui fa parte il competitor.

  Prima di introdurre il suo supporto è fondamentale conoscere alcuni ricorsi storici relativi ai marchi:

  - Negli anni 90' il gruppo Riello ufficializza l'acquisizione di Beretta;
  - Nel 2001 la società Ariston conclude l'acquisizione di alcune società e marchi nel settore del riscaldamento e dei bruciatori, in particolare il gruppo Chaffoteaux;

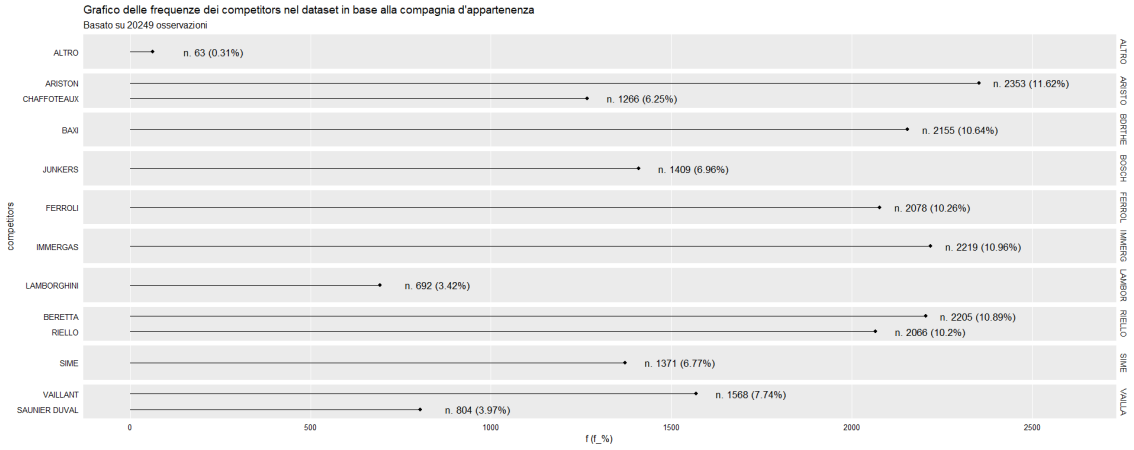
- Il gruppo Vaillant commercializza da oltre 70 anni caldaie e ricambi sia sotto propria denominazione e sia con quella di Hermann Saunier Duval;
- Nel 2009 De Dietrich Remeha Group e Baxi Group annunciano la creazione di BDR Thermea;
- Nel 2018 il colosso Bosch acquisisce definitivamente lo storico marchio Junkers.

Il supporto della variabile *COMPAGNIA* sarà dunque composto da 9 elementi: il gruppo *BDR Thermea, Bosch, Ariston, Immergas, Lamborghini, Riello, Sime, Ferroli e Vaillant*. Per avere un idea generale della distribuzione dei competitors, all'interno del nostro dataset, ho creato questo grafico di frequenza.



Come possiamo vedere, i competitors più presenti nel nostro dataset sono *Ariston, Immergas,* i marchi del gruppo *Riello e Ferroli* che insieme compongono circa il 65% del dataset. Successivamente abbiamo una buona presenza dei marchi *Vaillant, Junkers, Sime e Chaffoteaux* con una percentuale attorno al 6-8% l'uno. A seguire i restanti.

Andando più nel particolare, ho creato una seconda rappresentazione grafica. Nel grafico in questione, ho valutato la distribuzione delle frequenze dei competitors nel dataset raggruppati per Compagnia.

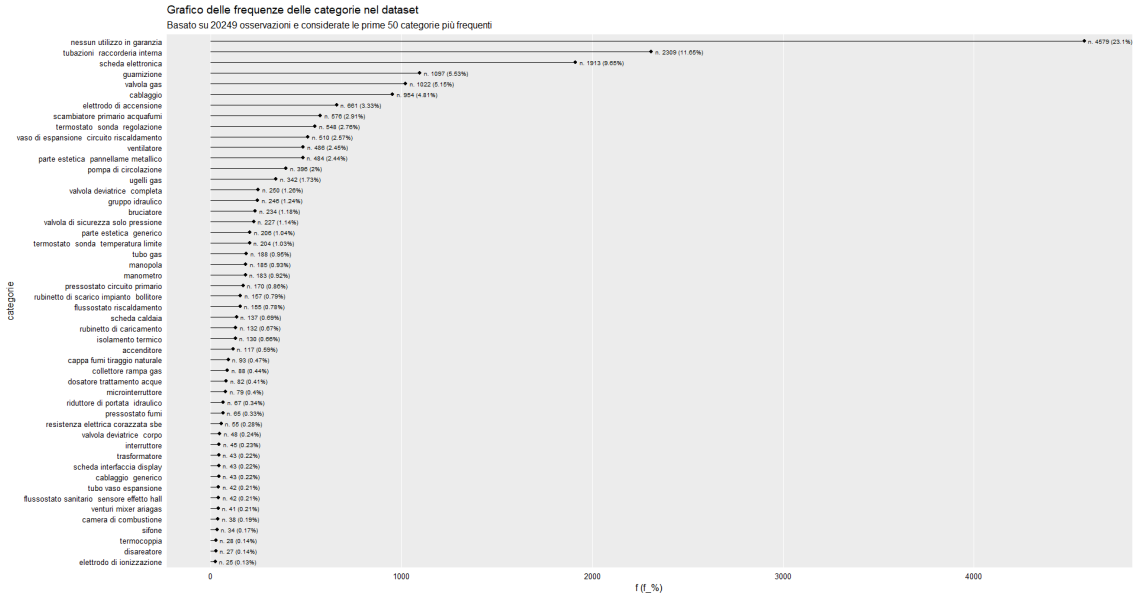


Dal grafico è possibile ricavare la classifica delle compagnie, in base alle frequenze delle marche che lo compongono, nel dataset: Gruppo Riello (21.09%), Ariston (17.87%), Vaillant (11.71%), Immergas (10.96%),BDR Thermea (10.64%), Ferroli (10.26%), Bosch (6.96%), Sime (6.77%), Lamborghini (3.42%) e Altro (0.31%). Guardando all'interno delle compagnie possiamo vedere che la mole di dati estratti per le marche che le compongono, non sono tutte uguali. Notiamo in particolare nei gruppi Ariston e Vaillant è presente un netto dislivello tra i marchi che le compongono. Da ciò possiamo dedurre che all'interno di queste compagnie c'è la presenza di una marca principale (che compone il 65% circa del gruppo) e di una secondaria. In generale vediamo che la principale è quella con la stessa denominazione del gruppo. Discorso diverso per la compagnia Riello dove i due marchi (Beretta e Riello) occupano entrambi il 50% del gruppo.

- **Variabili *CATEGORIA* e *SOTTOCATEGORIA***

  Le variabili categoria e sottocategoria le conosciamo molto bene. Il processo di classificazione ci ha permesso di classificare il dato estratto dando vita a queste due colonne del dataset.

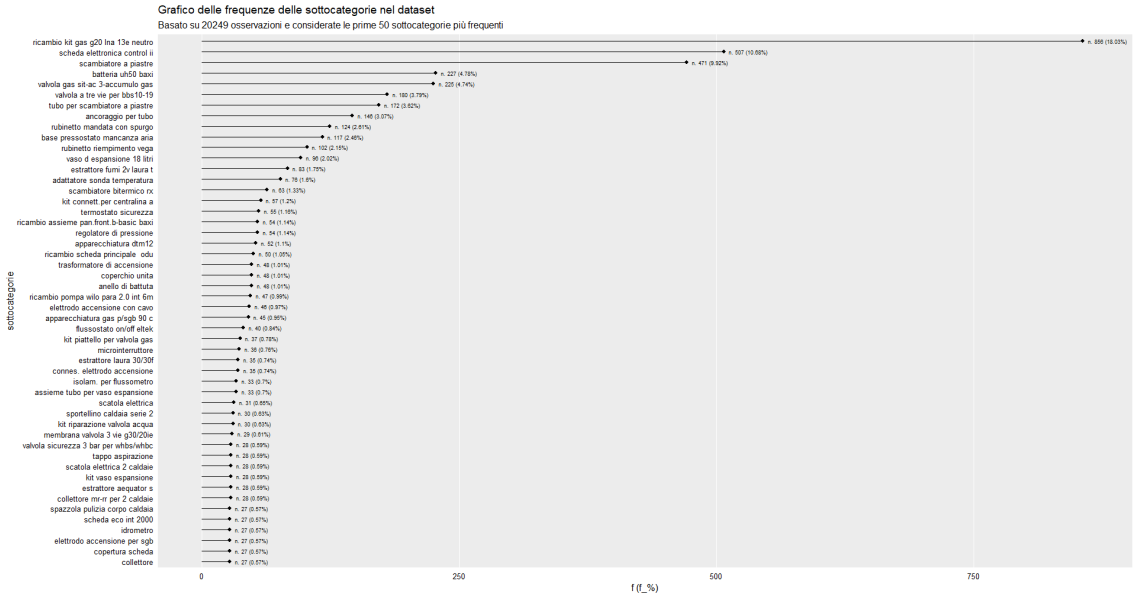
  La variabile ***Categoria*** è una variabile qualitativa nominale con supporto composto da 119 elementi. Rappresenta il macrogruppo di componenti ricambio Baxi assegnato all'estrazione. Nella seguente rappresentazione grafica rappresentiamo la distribuzione delle frequenze delle categorie nel dataset.



Come possiamo notare la categoria più frequente è ”nessun utilizzo in garanzia” che compone circa il 23% dell’intero dataset. Questa è una categoria molto generica, al suo interno è possibile trovare qualsiasi tipo di componente. In fase di analisi successive non andremo ad analizzarla proprio per questo motivo, la possiamo considerare come una ”nessuna categoria”. Subito dopo troviamo ”tubazioni raccorderia interna” (11.65%) e ”scheda elettronica” (9.65%). A mio parere risultato abbastanza prevedibile poichè, in genere, nei mercati e-commerce queste due categorie sono molto vendute. Essendo componenti soggetti a veloce usura, la vendita e la domanda di questi tipi di componenti è molto frequente ed elevata. Da segnalare inoltre che anche categorie come valvole a gas e cablaggi compongono anche loro buona parte del dataset.

Passiamo ora alla valutazione delle sottocategorie.

La variabile ***Sottocategoria*** è una variabile qualitativa nominale con supporto composto da 5843 elementi. Questa variabile come visto in precedenza, questa corrisponde all’associazione di un articolo del listino dei ricambi Baxi all’estrazione. Per comodità è stata chiamata ”sottocategoria”. Nell’analisi grafica in sottoimpressione, analogamente alle categorie, ho rappresentato le sottocategorie più frequenti all’interno del dataset.



Dal grafico possiamo notare che la sottocategoria che prevale all'interno del dataset è il "ricambio kit gas g20 lna 13e neutro" con una frequenza percentuale di circa 18%. E' un particolare kit di trasformazione metano presente all'interno delle caldaie Baxi. Nel nostro dataset questa sottocategoria corrisponde alla categoria "ugelli gas". Successivamente riscontriamo un alta frequenza di una particolare scheda elettronica "scheda elettronica control ii" con 10.68% e dello "scambiatore a piastre" con quasi 10%.

Figure 11: kit gas g20 lna 13e neutro

3.2 Analisi per l'identificazione del marchio nel mercato

- **Identificazione del marchio attraverso le parole chiave delle descrizioni**

  In questa sezione andremo a vedere le parole chiave più associate ai vari marchi e proveremo ad interpretare il significato. Prendendo le descrizioni relative ad ogni marchio utilizzerò una particolare tecnica di Text mining semplice e immediata per l'analisi delle parole chiave: il Wordcloud. Le parole più usate saranno quelle più grandi e, viceversa, quelle meno usate saranno più piccole.

Figure 12: Sar.Duval

Figure 13: Riello

Figure 14: Ferroli

Figure 15: Sime

Figure 16: Lamborghini

Figure 17: Junkers

Figure 18: Immergas

Figure 19: Vaillant

Figure 20: Chaffoteaux

Figure 21: Beretta

Figure 22: Baxi

Figure 23: Ariston

Figure 24: Altro

Commento:

Dando uno sguardo generale all'intera rappresentazione grafica, notiamo che il termine più ricorrente nei wordcloud dei competitors è "caldaia/e". Nella maggior parte delle descrizioni degli annunci alla fine della denominazione del componente viene aggiunta la parola caldaia. Un altro aspetto interessante da notare sono i termini "codice" e "ex" che compaiono in diverse nuvole. Questi due termini fanno riferimento al codice prodotto che nella maggior parte delle volta affianca la denominazione dell'articolo. E' possibile notare inoltre, che in alcune nuvole dei competitors, come Saunier Duval e Chaffoteaux, sono presenti parti della denominazione del marchio. La denominazione completa dei due marchi è rispettivamente "Hermann - Sarnier Duval" e "Chaffoteaux - &maury". In altre situazioni, come nel caso di Junkers, Ariston e molti altri, è possibile intravedere tra i termini il nome della compagnia di riferimento (es. Bosch per Junkers) o della marca sorella (es. Chaffoteaux per Ariston). Guardando con attenzione i wordcloud dei marchi Lamborghini e Ferroli noto una certa somiglianza. Oltre al fatto che nelle nuvole di entrambi i competitor prevale il termine "kit", ho individuato tra i termini di Lamborghini il nome del competitor Ferroli. Incuriosito dal fatto e tramite varie ricerche ho scoperto che parte della produzione e del commercio dei ricambi (nello specifico la ditta Calor) è stata comprata dal gruppo veronese Ferroli.

Un altro aspetto da notare nell'interpretazione dei wordcloud, è la presenza di linee di caldaie dei vari competitors. In particolare vediamo che in Baxi sono presenti le linee "ocean" e "luna eco 3", in Immergas "maior eolo", "vitrix" e "nike eco 4", in Vaillant "mag" e "vmw", in Lamborghini "TAURA 24 MCS" e infine "Alixia", "Alexia" e "Pigma" per Chaffoteaux.

Dando uno sguardo agli effettivi componenti di una caldaia, riscontro l'elevata frequenza di termini come "valvola", "scheda" (tendenzialmente affiancate da aggettivi come "elettronica" o di "accensione") e "scambiatore". In generale, tutti i competitors posseggono tra le parole più frequenti questi tre termini. In maniera minore ma comunque da citare sono i componenti ricambio "elettrodo" e "vaso di espansione". In conclusione, a parte qualche piccola eccezione, i marchi all'interno dei siti ecommerce di rivendita ricambi offrono bene o male gli stessi tipi di prodotti.

- **Analisi della popolarità e dimensione del sito e-commerce in base agli annunci di vendita**

  In questa analisi di mercato, ho valutato come i vari competitors si distribuiscono all'interno dei vari siti e-commerce.

Figure 25: Sar.Duval

Figure 26: Riello

Figure 27: Ferroli

Figure 28: Sime

Figure 29: Lamborghini

Figure 30: Junkers

Figure 31: Immergas

Figure 32: Vaillant

Figure 33: Chaffoteaux

Figure 34: Beretta

Figure 35: Baxi

Figure 36: Ariston

Figure 37: Altro

Tramite questa analisi posso ottenere due importanti informazioni. La prima riguarda la popolarità e la dimensione del sito. Attraverso lo studio delle frequenze posso capire quale sito è più conosciuto e utilizzato dai compratori rispetto ad altri. In genere un sito e-commerce misura la propria popolarità e grandezza in base alla mole di annunci di vendita presenti nel sito.

La seconda informazione invece è legata al lato cliente. In questo modo sto valutando la probabilità di trovare un ricambio di un determinato marchio in base ai siti che ho preso in esame. Immedesimandomi in un acquirente medio di ricambi, ed essendo interessato ad un determinato componente di una particolare marca, andrò a valutare in quali siti navigare per avere più possibilità di ottenere il ricambio desiderato.

Commento:

Dando uno sguardo generale ai worcloud otteniamo subito la prima informazione da ricercare. I siti più popolari e più ricorrenti nei wordcloud dei vari competitors sono Ebay, Arbo e Tecnofrasca. La loro quantità di ricambi offerti e commercializzati è nettamente superiore rispetto ad altri siti e-commerce. Alla luce dei risultati della rappresentazione grafica, se fossi interessato ad un determinato competitor, i siti su cui essere sicuro di trovare un vasto assortimento di ricambi sono quelli sopra citati. Per alcuni marchi però la distribuzione e la mole di ricambi dei tre siti sopra citati non è la stessa. In particolare se vogliamo un ricambio di marca Junkers o Vaillant otterremo un vasto assortimento unicamente in Arbo e Ebay, due dei tre siti più popolari. Se fossi interessato al marchio Chaffoteaux avrei più probabilità di ottenere il ricambio che desidero navigando nel sito Teecnofrasca, se non dovessi trovarlo ripiegherei in Arbo. Per altri tipi di marchi come Sarnier Duval, Lamborghini, il cliente farà riferimento unicamente ad un unico sito, rispettivamente a Ebay per il primo e Arbo per il secondo. In conclusione, se è nostro interesse un ricambio di marchi non presenti tra quelli principali, andremo sul sito di Deliapaolo.

- **Analisi della popolarità dei competitors nel mercato online**

  In questa terza analisi andremo a valutare numerosi aspetti relativi al rapporto tra competitors e siti e-commerce di ricambi caldaie. Gli obiettivi di questa parte di analisi sono 2. Il primo è di capire e misurare quanto un determinato competitor è popolare e influente all'interno dei vari siti e-commerce. Riuscendo a trarre alcune conclusioni su quanto un marchio sia sponsorizzato da un particolare sito.

Il secondo obiettivo è quello di capire quanto un determinato sito è assortito. In base alla valutazione delle frequenze dei competitors , all'interno dei vari siti, riusciamo a capire se un determinato sito e-commerce permette l'acquisto di svariate marche oppure no. In un sito ecommerce è molto importante la presenza di vari tipi di marchi. Potrà succedere che alcuni siti commercializzino solamente un particolare gruppo ristretto di marche. Utilizzando sempre la tecnica wordcloud, andremo ad analizzare e a rispondere agli obbiettivi che ci siamo posti.

Figure 38: Fausto

Figure 39: Gnesato

Figure 40: Ebay

Figure 41: Ricambipercaldaie

Figure 42: Ricambixcaldaie

Figure 43: sm

Figure 44: Tecnofrasca

Figure 45: Bimaxsnc

Figure 46: Climacore

Figure 47: Deliapaolo

Figure 48: Arbo

Commento:

La rappresentazione grafica ci mette davanti 11 worcloud relativi ai siti ecommerce di vendita di ricambi di caldaie originali. A livello di vastità di repertorio in ambito marchi, i siti Arbo, sm, Ricambixcaldaie, Ebay, Ricambipercaldaie e Faustoricambi offrono un buon assortimento. In particolare i componenti ricambio che vengono commercializzati in questi siti, sono distribuiti per marca in maniera equilibrata. Altri siti come Gnesato e Deliapaolo prediligono la vendita di unicamente una cerchia molto ristretta di marche. Partendo dai siti con un buon assortimento e varietà di marchi, andremo a valutare la presenza e l'assenza dei vari marchi.

- **Arbo**: commercializza ricambi di un pò tutti i marchi. Sono poco forniti di ricambi a marchio Chaffoteaux e noto dell'assenza del marchio Saunier Duval.
- **sm**: ben assortito, non predilige alcun competitor in particolare. Non vengono forniti ricambi di marchio Saunier Duval e Lamborghini.
- **Ricambixcaldaie**: ben assortito, commercializzano maggiormente i marchi Baxi, Immergas, Ferroli Ariston e Beretta. Non presenti nel sito i marchi Vaillant e Lamborghini.
- **Ebay**: fornisce una vasta gamma di ricambi dei marchi più importanti, non presenti però Chaffoteaux, Sime e Lamborghini.
- **Ricambipercaldaie**: sito ben assortito di marche, non predilige di alcuna marca in particolare. Non è presente il marchio Lamborghini e Saunier Duval.
- **Faustoricambi**: ben fornito di ricambi di svariate marche. Non commercializzati articoli a marchio Lamborghini, Junkers e Saunier Duval.
- **Climacore**: sito ben assortito, presenti quasi tutte le marche. Scarseggi la commercializzazione del marchio sime ed è assente il marchio Lamborghini.
- **Bimaxsnc**: Notiamo che in questo sito il marchio che viene commercializzato di più è Immergas. E' comunque presente uno svariato assortimento di altri marchi anche se è da sottolineare una poca presenza di prodotti Chaffoteaux, Sime e Ariston.

- **Gnesato**: vengono commercializzati unicamente alcuni marchi come Sarnier Duval e Vaillant, ma in particolare il marchio Chaffoteaux.
- **Tecnofrasca**: predilige la vendita di ricambi forniti dalle compagnie Ariston e Riello.E' abbastanza equilibrato nella commercializzazione di ricambi di altre marche tranne Lamborghini, Saunier Duval, Junkers e Vaillant.
- **Deliapaolo**: predilige la vendita di marchi minori e del marchio Beretta. Non fornisce ricambi di altri marchi.

3.3 Analisi di mercato della scheda elettronica

In questo capitolo, effettueremo un'analisi dei prezzi per confrontare i vari competitor presenti nel dataset. Dato l'ingente quantità di informazioni disponibili, sceglieremo accuratamente una sottocategoria di particolare interesse per condurre un'analisi dettagliata. Da questa analisi, trarrò alcune conclusioni importanti. Tuttavia, prima di iniziare, ci tengo a specificare alcune considerazioni preliminari.

Premessa importante

Il mio obiettivo attraverso questa tesi non è quello di fornire sentenze definitive, poiché ci sono numerose variabili che potrebbero compromettere le analisi:

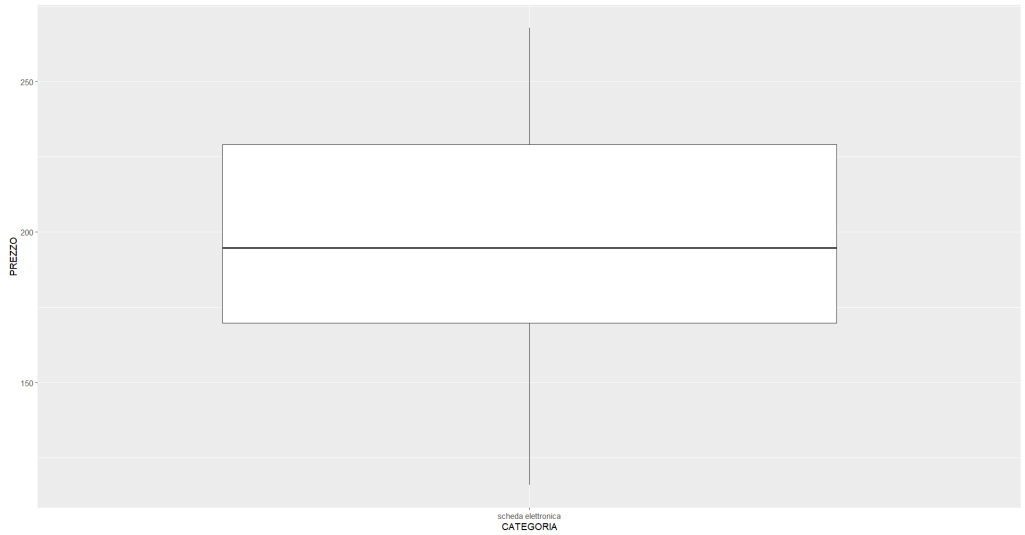
- *I dati all'interno dei siti e-commerce non sono sempre gli stessi e vengono continuamente aggiornati con annunci contenenti descrizioni e prezzi differenti;*
- *Il mercato online non è rappresentativo della realtà, quindi una sentenza nel mercato online potrebbe essere completamente smentita nel vero mercato dei ricambi originali;*
- *La presenza di annunci creati da venditori non ufficiali che possono alterare i prezzi e renderli quindi diversi da quelli di listino in base alle loro necessità;*
- *La mancanza di uno strumento o tecnica in grado di garantire al 100% che un determinato componente di una marca sia esattamente lo stesso di un altro marchio. Non esiste modo di vedere tramite codice ricambio se un componente di una marca è compatibile con uno di un altra;*

- *La classificazione sebbene molto efficiente, non è perfetta. Inoltre viene particolarmente compromessa dall'incomprensibilità delle descrizioni. Per via del limite di 30 caratteri dato dal gestionale, viene infatti abbreviata, talvolta anche con errori.*

Infine, il mondo della termoidraulica non è la mia area di competenza, quindi potrebbero esserci delle limitazioni nella mia capacità di comprendere completamente i dati e le informazioni presenti.

Boxplot e analisi preliminari

Prima di scegliere la sottocategoria da analizzare è di fondamentale importanza l' identificazione della categoria. Dalla premessa si evince il fatto che la classificazione dei componenti, seppur molto efficiente, non è perfetta. Per la scelta della categoria, mi sono affidato alle conclusioni riguardanti la bontà della cluster analysis. In particolare, ho scelto una categoria che, durante la valutazione della classificazione, ha ottenuto l'indice di F1 score più alto. Tra le papabili, ho deciso di analizzare la categoria "scheda elettronica". Questa è composta da 1913 componenti e 350 sottocategorie. La scheda elettronica della caldaia, detta anche scheda madre, rappresenta il cervello elettronico dell'apparecchio e gestisce tutte le funzioni della macchina, regolandone anche il funzionamento. Esistono svariati tipi di schede a seconda della tecnologia del macchinario. Una volta individuata la categoria possiamo focalizzarci sulla scelta della sottocategoria. Il "grafico delle frequenze delle sottocategorie" creato e interpretato per l'analisi esplorativa della variabile "sottocategoria" mi ha permesso di avere una visione chiara delle papabili scelte. Alla fine la categoria che ho deciso di analizzare è "scheda elettronica control ii", seconda sottocategoria più frequente del dataset con 507 componenti. Innanzitutto, per disporre di un riepilogo visivo della variabilità dei prezzi di questa scheda nel dataset, ho creato un Boxplot e una tabella con i valori che lo compongono.

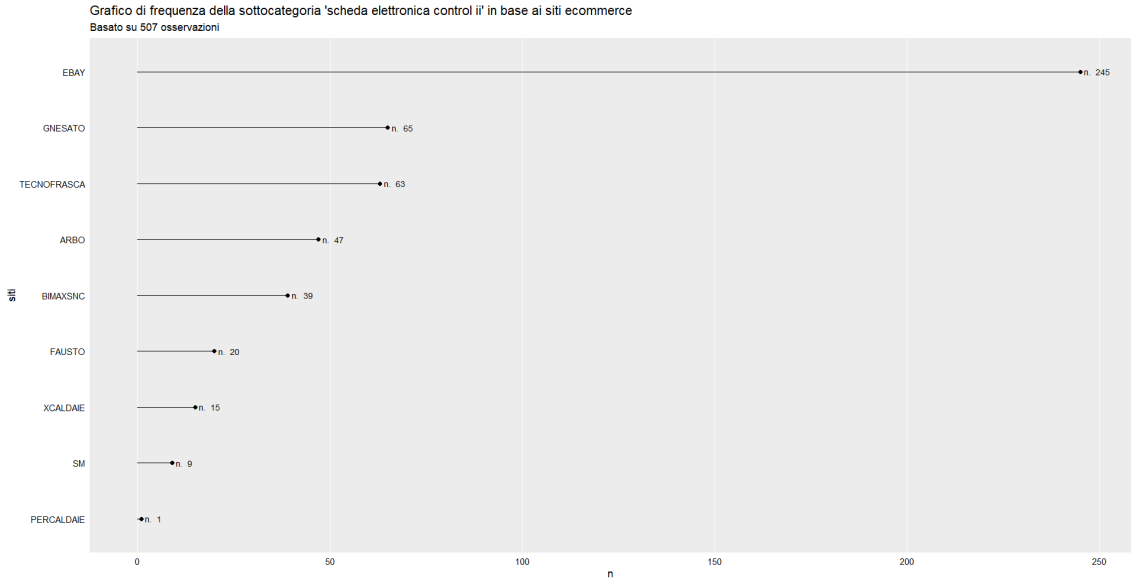


Min.	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
116.0	169.8	194.8	196.8	229.1	268.0

Table 18: Valori del Boxplot

Come possiamo notare, la maggior parte dei dati (il 50% centrale) si concentra tra 169.8 e 229.1, con una mediana di 194.8. La media, tuttavia, è leggermente più alta rispetto alla mediana, il che indica una distribuzione dei dati leggermente asimmetrica verso destra. Inoltre, ci sono alcuni valori che si discostano molto dalla maggioranza dei dati, vengono evidenziati dalla lunghezza delle ”baffi” del boxplot. Questo è segno che potrebbero esserci degli outliers ma, nel caso in questione, non sono presenti. Per un accurata analisi preliminare è opportuno verificare le frequenze dei componenti in base al marchio e in base al sito.

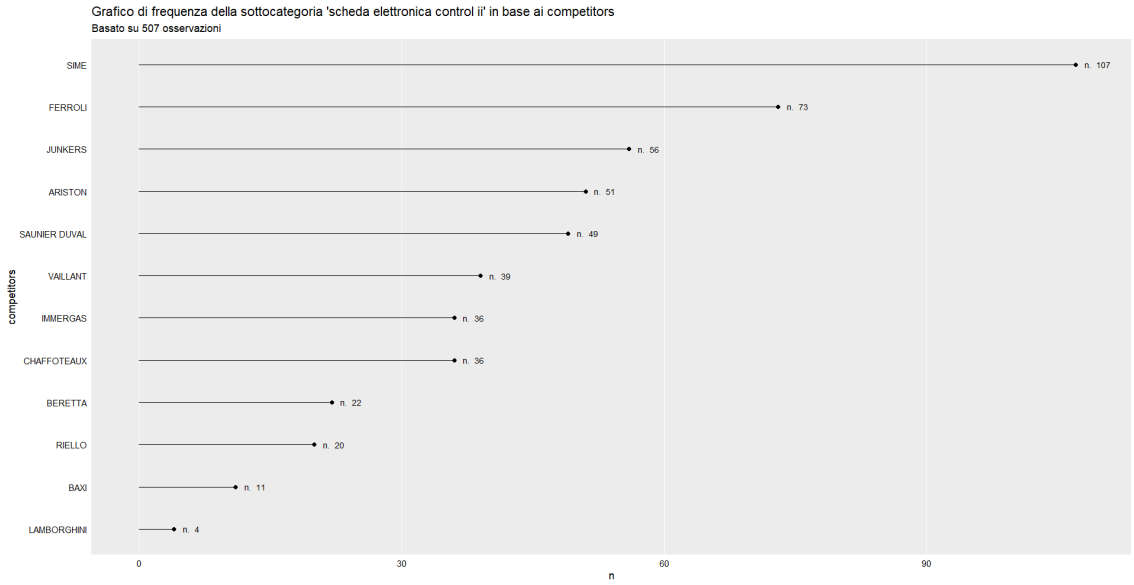
- **Rappresentazione delle frequenze dei componenti in base al sito e-commerce**



Con l'analisi in questione vediamo dove i componenti sono stati estratti con maggior frequenza. In modo tale da capire in che sito vengono commercializzate di più le schede elettroniche, in particolare la scheda "scheda elettronica control ii".
Come possiamo vedere, la maggior parte dei schede (circa il 48 %) sono state estratte dal famoso sito Ebay. Successivamente segue Gnesato con 13% e Tecnofrasca con 12%. Il sito che fornisce meno schede di questa sottocategoria è il sito "ricambipercaldaie" dove è stato estratto solo un componente.

- **Rappresentazione delle frequenze dei componenti in base ai competitors**
  Una volta analizzata la provenienza dei dati, valutiamo sotto quali marchi queste schede vengono commercializzate. Nel grafico il questione vediamo la frequenza di schede di tipo "scheda elettronica control ii" in base alla marca.

Dal grafico notiamo che, la main competitor di Baxi in ambito di schede elettroniche di tipo "scheda elettronica control ii" è la marca veronese Sime. Con una percentuale di circa 21%. A seguire troviamo Ferroli con 14% e Junkers, Ariston e Saunier Duval con una frequenza percentuale tra il 10 e 12% l'uno. Il sito che commercializza meno schede in questione è Lamborghini con unicamente 4 componenti.



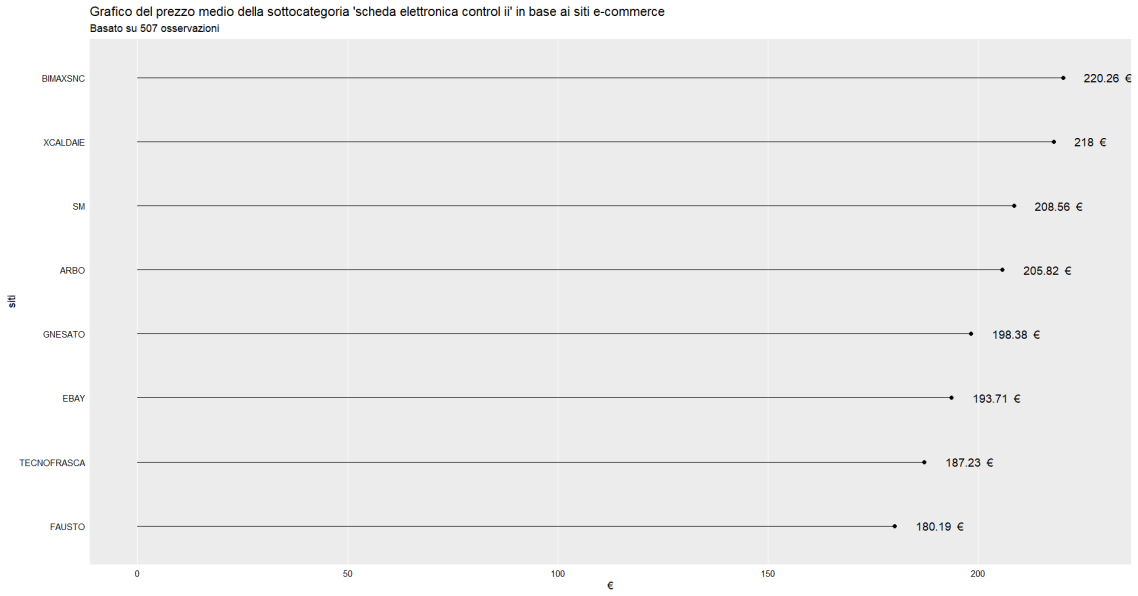
Analisi dei prezzi

Una volta analizzata la provenienza e la distribuzione di questo tipo di schede in base al competitor è il momento di affettuare una vera e propria analisi dei prezzi. In particolare valuteremo il prezzo medio di questi componenti.

- **Analisi dei prezzi in base al sito**

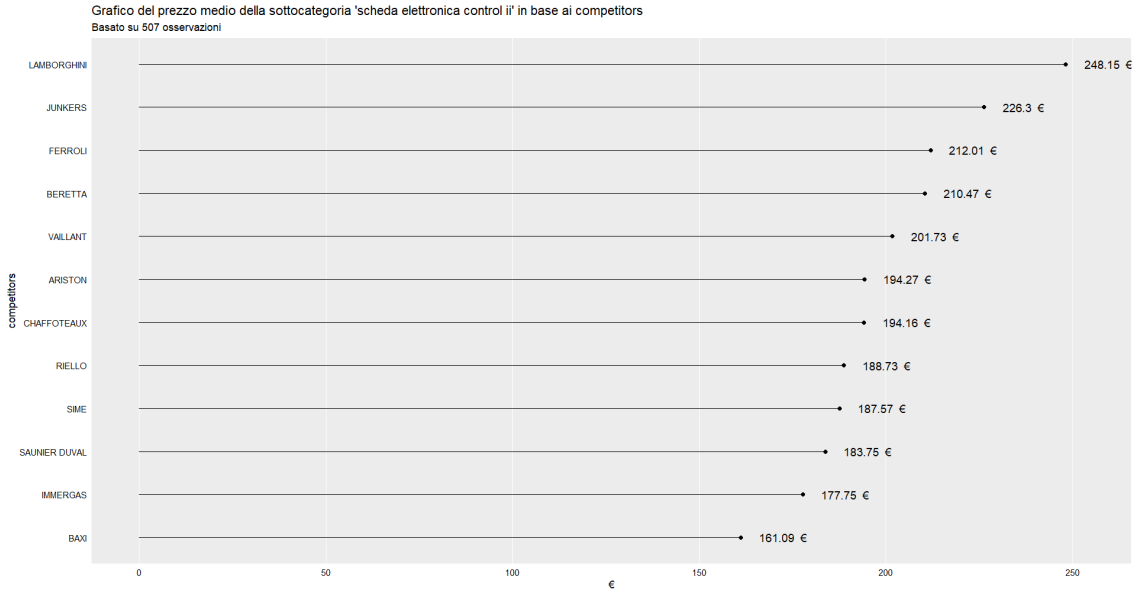
  La rappresentazione grafica in sottoimpressione rappresenta il prezzo medio della sottocategoria ”scheda elettronica control ii” in base ai siti ecommerce. Il prezzo di una scheda elettronica di una caldaia è variabile e dipendente esclusivamente dalla marca e modello caldaia specifica. In genere il prezzo di una scheda elettronica si aggira tra i 150 e i 200 euro.

  All’interno dei vari siti vediamo che questa particolare scheda elettronica ha costo che si aggira tra i 180 e 220 €. Il siti che commercializzano mediamente questo articolo con il prezzo più alto sono il sito BIMAXSNC (220 € circa) e RICAMBIXCALDAIE (218 €). Tra i siti dove mediamente viene offerta una soluzione di prezzo più vantagiosa abbiamo FaustoRicambi (180 €) e Tecnofrasca (188 € circa). Rispetto al range di prezzo stabilito all’inizio riscontriamo che all’interno dei vari siti ecommerce il prezzo di questo componente aumenta. Questo può essere dato dal fatto che i prezzi vengono gonfiati per ottenere un discreto margine di guadagno da parte dei creatori di annunci.



- **Analisi dei prezzi in base ai competitors**

  All'interno del listino Baxi è la sottocategoria che stiamo analizzando corrisponde ad un articolo con prezzo pari a 160,21 €. Avendo questa informazione abbiamo la possibilità di effettuare confronti con i prezzi dei vari competitors.



Dal grafico notiamo che nessun altro competitor offre un prezzo vantagioso come quello di Baxi. Notiamo in particolare che Lamborghini e Junkers sono i competitors che commercializzano questa particolare scheda ad un prezzo decisamente alto. Rispettivamente 248.15 € e 226,30€.

Vediamo che in generale il prezzo di questo componente viene venduto dai competitors mediamente a 200€ circa, mentre il marchio Baxi a 160€ circa. Possiamo concludere che per quanto riguarda questo particolare ricambio, Baxi offre il prezzo più conveniente di tutti.

4 Conclusioni

In conclusione, il presente elaborato ha permesso l'ideazione e la creazione di un'analisi di mercato completa e approfondita dei ricambi di caldaie. La realizzazione di un progetto di questa portata richiede una pianificazione attenta e una gestione efficiente delle risorse, nonché un'approfondita conoscenza dei concetti e degli strumenti statistici necessari per l'analisi dei dati. Grazie alla raccolta di dati, composta da oltre 20500 componenti, è stato possibile ottenere una dinamica generale dei prezzi nel settore termoidraulico, con importanti informazioni utili per il miglioramento del posizionamento di Baxi S.p.A. nel mercato online. L'esperienza lavorativa presso Baxi S.p.A. è stata fondamentale per la realizzazione di questo progetto e ha permesso di sviluppare una conoscenza abbastanza approfondita del settore. Sono grato per l'opportunità che mi è stata concessa e spero di poter applicare le competenze acquisite in futuro, contribuendo alla crescita e allo sviluppo di altre aziende e progetti.

Sitografia:

- https://www.baxi.it/
- https://www.ricambipercaldaie.it/
- https://www.smricambi.com/it
- https://www.ebay.it/
- https://www.gnesato.com/ricambi-caldaie/
- https://www.ricambixcaldaie.com/
- https://www.tecnofrasca.it/
- https://www.deliapaolo.it/
- https://www.tecnofrasca.it/
- https://www.faustoricambi.it/
- https://www.arbo.it/
- https://www.bimaxsnc.com/
- https://climacore.it/
- https://www.bosch-thermotechnology.com/it/it/residenziale/junkers/
- https://www.lifegate.it/vaillant-una-storia-allinsegna-della-sostenibilita/
- https://techtarget.com/
- https://ggplot2.tidyverse.org/
- https://cran.r-project.org/package=ggplot2/
- https://r-graph-gallery.com/ggplot2-package.html/

Bibliografia:

- Favero G. LE SMALTERIE DOPO LE SMALTERIE. Castelfranco Veneto: Linea Grafica; 2003.
- Cortese G. BAXI 1925-2015. Bassano del Grappa: Editrice Artistica Bassano; 2015.

Ringraziamenti

Questi favolosi tre anni a Santa Caterina si sono conclusi. La scelta di studiare statistica devo dire si è rivelata quella giusta. Ho appreso nozioni importanti che spero di sfruttare nel mondo lavorativo che mi attende. Un percorso stupendo che mi ha permesso di conoscere un sacco di bellissime persone. In primis ringrazio i miei genitori Stefano e Mariacristina e i miei fratelli Tommaso ed Elisa, grazie al loro sostegno mi è stato permesso di frequentare e sostenere questa università. Si un attimo un attimo ora tocca a voi *Stat Bombers* aka *Gamers*. Ringrazio di cuore il dottori Raul Zanatta, Marco Rudelli e Fabio Frassetto. Grazie a voi lo studio della statistica è stato più semplice e divertente come una partita a Stumble Guys. A seguire ringrazio tutti i miei amici di statistica, la mia compagnia e i miei parenti.