Matteo Franceschi, 502231 *mfranceschi94@gmail.com*
Federica Guiducci, 600310 *guifede3@gmail.com*
Lorenzo Mannocci, 518263 *mannocci.lore@gmail.com*
Riccardo Paoletti, 532143 *paoletti.riccardo0@gmail.com*

Analisi del Dataset di Carvana

Progetto Data Mining

Università di Pisa

Anno Accademico 2019/2020

1 Data Understanding

1.1 Data Semantics

Il Dataset preso in analisi è composto da 58.386 record descritti da 34 attributi di diverso tipo. Ogni record rappresenta un acquisto di un veicolo da Carvana, un'azienda che si occupa della rivendita di veicoli usati negli Stati Uniti. Obiettivo delle analisi svolte nel Report è quello di illustrare un modello che sia in grado di prevenire il rischio che l'azienda possa acquistare veicoli non convenienti utilizzando come variabile dipendente l'attributo **IsBadBuy**. È opportuno specificare che nel dizionario fornito sono inseriti due attributi che nel Dataset risultano assenti: **AcquisitionType** (identifica come è stato acquistato il veicolo) e **KickDate** (data in cui il veicolo è stato respinto all'asta).
Nel seguente elenco si fornisce la lista degli attributi e il loro dominio, mentre la tabella in Figura 1 fa riferimento alla classificazione degli attributi per tipo.

- **Auction** - Asta presso la quale è stato effettuato l'acquisto (OTHER, MANHEIM, ADESA);
- **Make** - Casa di produzione del veicolo (ACURA, BUICK, CADILLAC, ..., ecc);
- **Model** - Modello del veicolo (1500 RAM PICKUP 2WD, SPECTRA, ... ecc);
- **Trim** - Livello di Trim (equipaggiamento) del veicolo (EX, STX, SE, ..., ecc);
- **SubModel** - Sottomodello del veicolo(2D CONVERTIBLE, 4D SEDAN EX,...,ecc);
- **Color** - Colore del veicolo (Red, black, ...,ecc);
- **Transmission** - Tipo di cambio(Automatico, Manuale);
- **WheelType** - Il tipo di ruote del veicolo (Covers,Alloy, Special)
- **Nationality** - Nazionalità del veicolo (AMERICAN,..., TOP LINE ASIAN);
- **Size** - La taglia del veicolo (MEDIUM, VAN,...,ecc);
- **TopThreeAmericanName** - Identifica se il veicolo è stato prodotto da uno dei tre maggiori produttori americani (Chrysler, Ford, GM, Other);
- **AUCGUART** - Il livello di garanzia fornito dall'asta per l'acquisto del veicolo (GREEN, YELLOW, RED);
- **VNST** - Stato americano dove il veicolo è stato acquistato (AL, AR,..., ecc) ;
- **PRIMEUNIT** - Identifica se il veicolo avrebbe una domanda maggiore di un acquisto standard (yes,no);
- **VehBCost** - Prezzo del veicolo al tempo dell'acquisto ($N > 0$);
- **WarrantyCost** - Costo della garanzia (durata = 36 mesi, Tassa di Millage = 36K) ($N > 0$);
- **RefId** - Numero univoco (sequenziale) assegnato ai veicoli (1,..,73014);
- **MMRAcquisitionAuctionAveragePrice** - Prezzo di acquisto del veicolo in condizioni medie al tempo dell'acquisto ($N > 0$);
- **MMRAcquisitionAuctionCleanPrice** - Prezzo di acquisto del veicolo in condizioni buone al tempo dell'acquisto ($N > 0$);
- **MMRAcquisitionRetailAveragePrice** - Prezzo di acquisto del veicolo in condizioni medie al dettaglio al tempo dell'acquisto ($N > 0$);
- **MMRAcquisitonRetailCleanPrice** - Prezzo di acquisto del veicolo in condizioni buone al dettaglio al tempo dell'acquisto ($N > 0$);
- **MMRCurrentAuctionAveragePrice** - Prezzo corrente sul mercato del veicolo in condizioni medie ($N > 0$);
- **MMRCurrentAuctionCleanPrice** - Prezzo corrente sul mercato del veicolo in condizioni buone ($N > 0$);
- **MMRCurrentRetailAveragePrice** - Prezzo corrente al dettaglio del veicolo in condizioni medie ($N > 0$);
- **MMRCurrentRetailCleanPrice** - Prezzo corrente al dettaglio del veicolo in condizioni buone ($N > 0$);
- **VehOdo** - Lettura del conta-chilometri del veicolo ($N > 0$);
- **IsBadBuy** - identifica se l'acquisto non è stato conveniente (1, 0);
- **PurchDate** - Data dell'acquisto all'asta (dal 1/10/2009 al 12/30/2010);
- **VehYear** - Anno di produzione del veicolo (2001,...,2010);
- **VehicleAge** - Anni di vita del veicolo (0,...,9);
- **WheelTypeID** - ID del tipo di ruota del veicolo (0,10,20,30);
- **BYRNO** - Numero unico che identifica il compratore (835,...,99761);
- **VNZIP1** - Zipcode che identifica dove è stato effettuato l'acquisto (2764,...,99244);
- **IsOnlineSale** - Specifica se il veicolo è stato originariamente acquistato online (0,1).

CLASSIFICAZIONE		ATTRIBUTO
CATEGORICI	NOMINALI	*Auction, Make, Model, Trim, SubModel, Color, Nationality, TopThreeAmericanName, AcquisitionType, AUCGUART, VNST*
	BINARI	*Transmission, PRIMEUNIT*
	ORDINALI	*Size, WheelType*
NUMERICI	BINARI	*IsBadBuy, IsOnlineSale*
	DISCRETI	*VehYear, VehicleAge, WheelTypeID,BYRNO, VNZIP*
	CONTINUI	*VehBCost, VehOdo, MMRAcquisitionAuctionAveragePrice, MMRAcquisitionAuctionCleanPrice, MMRAcquisitionRetailAveragePrice, MMRAcquisitonRetailCleanPrice, MMRCurrentAuctionAveragePrice, MMRCurrentAuctionCleanPrice, MMRCurrentRetailAveragePrice, MMRCurrentRetailCleanPrice, WarrantyCost*
	DATA	*PurchDate*

Tabella 1: Classificazione degli attributi del Dataset

1.2 Distribuzione delle variabili e statistiche

Di seguito viene riportata la descrizione statistica degli attributi presenti nel Dataset raggruppati secondo le diverse metodologie usate per la loro descrizione e alle caratteristiche in comune.

IsBadBuy, Transmission, IsOnlineSale

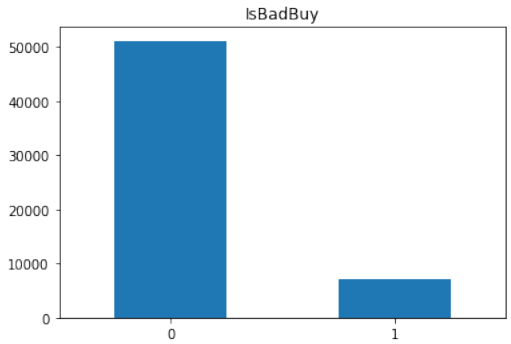

Figura 1: Distruzione IsBadBuy

Queste variabili sono caratterizzate da una distribuzione fortemente sbilanciata:

- **IsBadBuy** - YES(1): 12,4%, NO(0): 87,6%;
- **Trasmission** - AUTO: 96,49%, MANUAL: 3,51%;
- **IsOnlineSale** - YES: 97,4%, NO: 2,6%;

PurchDate, VehYear, VehicleAge

Queste tre variabili esprimono tutte l'età del veicolo, infatti sono accumunate dalla relazione:

$$VehicleAge = Year(PurchDate) - VehYear$$

Nella Figura 2 sono rappresentate le frequenze di acquisto dei veicoli divise per mese. Questo tipo di analisi è stato utile al fine di verificare se fossero o meno presenti significative differenze tra i mesi all'interno dello stesso anno o di anni differenti.

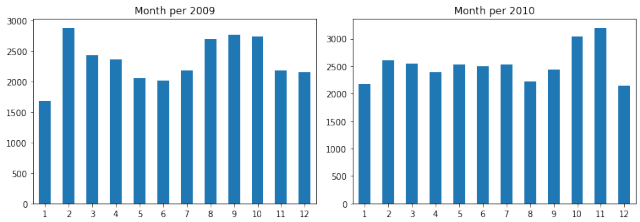

Figura 2: Frequenza di acquisto mensile per gli anni 2009 e 2010

WheelTypeID, WheelType

Ad ogni **WheelType** è stato assegnato un **WheelTypeID**. **WheelTypeID**=0 corrisponde a missing values presenti nella colonna **WheelType**. Come si nota nella Figura 2 non essendo ancora stati gestiti missing values e outliers, compaia nella figura 3 anche il valore 0, tuttavia in **WheelType** il valore non è rappresentato in quanto tali valori (che corrispondono a *nan*) vengono automaticamente esclusi dal grafico. Più nel dettaglio, si osservano le seguenti percentuali: *Alloy* - 49%, *Covers* - 45%, *Special* - 1%.

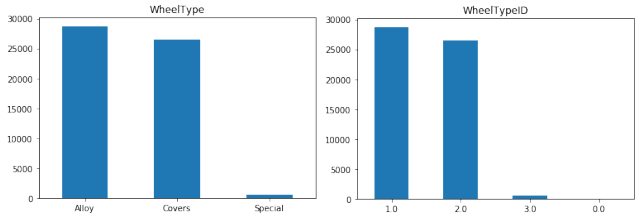

Figura 3: Distruzione degli attributi WheelType e WheelTypeID

AUCTION, Make, Model, Trim, SubModel, Color, Nationality, Size, TopThreeAmericanName

Per gli attributi **Make**, **Model**, **Trim**, **SubModel**, **Color** e **Nationality** al fine di una migliore visualizzazione, si sono distinti i valori maggiormente presenti mentre gli altri sono stati raggruppati (solo a fini illustrativi) in una più ampia categoria denominata *OTHERS* come mostrato nella figura 4, esempio della distribuzione dell'attributo **Make**.

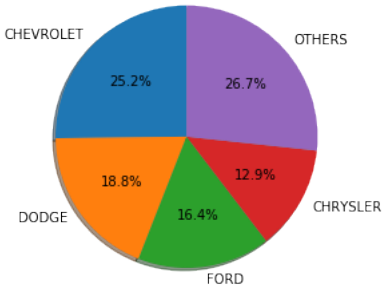

Figura 4: Distruzione dell'attributo Make

Attributi relativi ai prezzi

Gli otto attributi denominati *Price* identificano i prezzi di acquisizione del veicolo in base a determinate condizioni:

- ***Luogo:*** all'asta (Auction) o al mercato al dettaglio (Retail);
- ***Tempo di rilevamento:*** il giorno in cui Carvana ha acquisito la macchina (Acquisition) e il giorno *"corrente"* identificato come il giorno in cui sono stati inseriti nel dataset (Current);
- ***Condizioni del veicolo:*** si dividono tra medie (Average) o ottime (Clean);

La distribuzione di questi attributi risulta di tipo *Normale*, e in egual modo anche quella di: **VehOdo**, **VehBCost** e **WarrantyCost** come illustrato nella Figura 5.

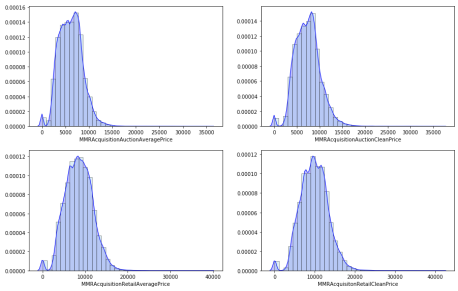

Figura 5: Distribuzione di 4 prezzi dei veicoli

PRIMEUNIT, AUCGUART, VNZIP1, VNST

PRIMEUNIT è presente solo per il 4,6% dei record, così come l'attributo **AUCGUART**, per il quale sono presenti solo luci verdi (la maggior parte) e luci rosse per una piccola percentuale.
L'attributo **VNZIP1** e **VNST** rappresentano entrambi la stessa informazione: il luogo di provenienza del veicolo, rispettivamente alla città e allo Stato. Nella figura 6 viene mostrata la distribuzione dell'attributo **VNST**.

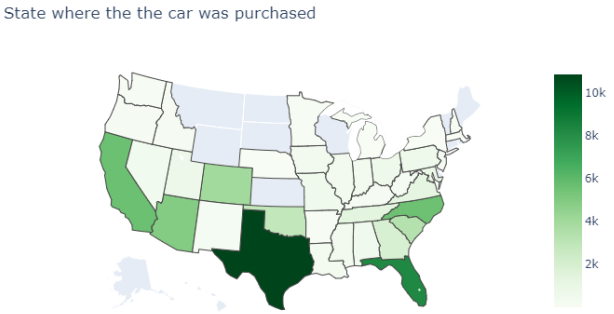

Figura 6: Distribuzione degli Stati in cui vengono acquistati veicoli

BYRNO, RefID

Questi attributi sono stati esclusi dal lavoro svolto in quanto rappresentano numeri univoci identificativi, e non forniscono informazioni utili al tipo di analisi in oggetto.

1.3 Valutazione della Data Quality

1.3.1 Missing Values

La figura 7 mostra la distribuzione di missing values nel dataset. Le zone bianche individuano i valori nulli, e si può notare come molti degli attributi non ne presentino.

La prima cosa che si nota è che per gli attributi **PRIMEUNIT** e **AUCGUART** vi è una presenza di valori nulli pari al 95%. In questa particolare situazione, dove non è possibile determinare una logica per estrapolare dati,

si è preferito eliminare le colonne. Gli altri attributi del Dataset presentano una percentuale di missing values decisamente minore.

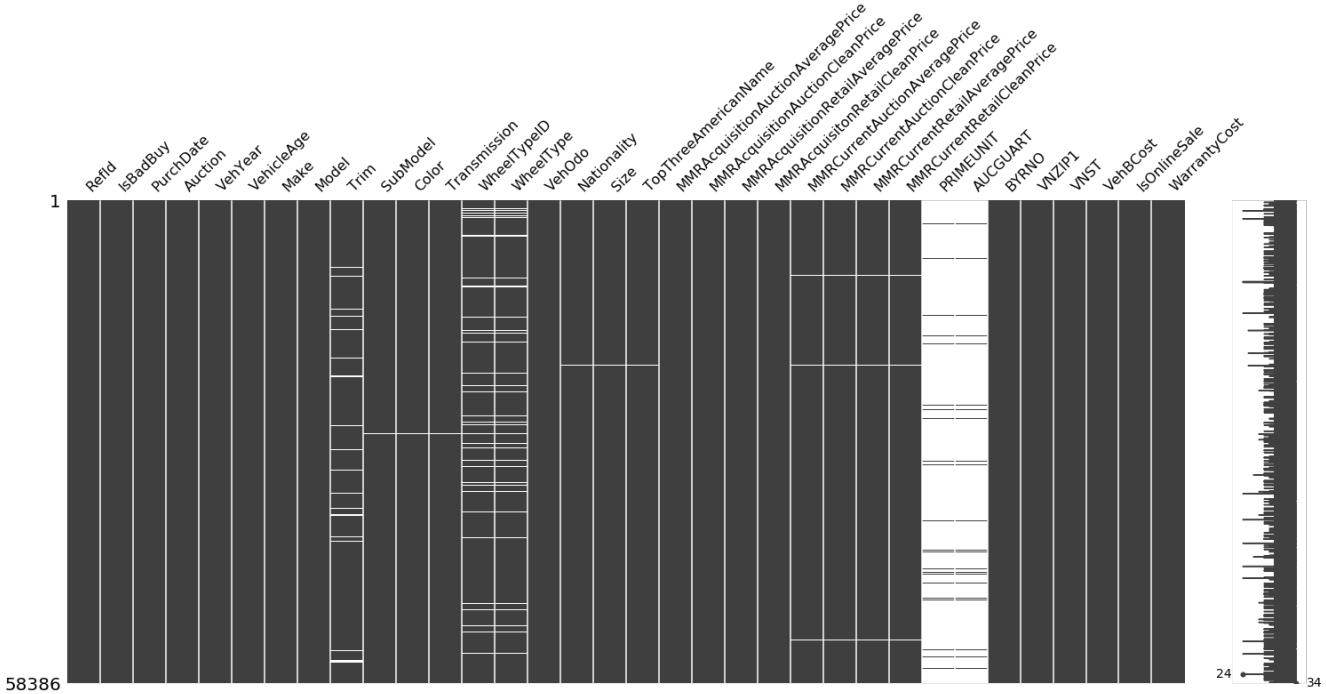

Figura 7: Distribuzione dei missing values

In base a questo rilevamento si è deciso di procedere nel seguente modo: nel caso i valori mancanti possano essere estrapolati da fonti esterne, o in base al valore di altri attributi, si è preferito non eliminare la riga ma sostituire il valore mancante con quello derivato, mentre in caso contrario si è eliminata la riga.

Si descrive ora per ogni attributo come si è proceduto per gestire i valori mancanti:

- **VehicleAge** - non presenta missing values, ma presenta 6 record il cui valore non è uguale alla differenza tra PurchDate e VehYear. Il valore per questi 6 record è stato sostituito con tale differenza;
- **Transmission** - 8 missing values di cui 7 eliminati durante la gestione di altri attributi ed uno con Model = *MONTEGO 3.0L V6 EFI*. Si è osservato che tutti i record con questo modello, presentano Transmission = *AUTO*, si è dunque deciso di adeguare questo record con lo stesso valore;
- **Trim** - contiene 1911 missing values fra i suoi valori. Essendo un livello di accessoria del veicolo che dipende dal produttore, non è possibile ricavarne un valore;
- **Submodel** - presenta 7 missing values. I record corrispondenti sono stati eliminati, in quanto non è possibile derivare il loro valore da altri attributi nel dataset;
- **WheelTypeID, WheelType** - i record in cui entrambi gli attributi erano mancanti sono stati eliminati, in quanto non era possibile ricavare in alcun modo il loro valore;
- **TopThreeAmericanName** - 4 record con valori mancanti. Per correggerli è stato usato il valore dell'attributo *Make*;
- **Nationality** - 4 record con valori mancanti. Si è osservato che tutti i veicoli in questione sono stati prodotti da un'azienda automobilistica posseduta dalla *General Motors*, dunque si è impostato l'attributo *Nationality = AMERICAN*;
- **Size** - 4 record con valori mancanti (in corrispondenza di quelli presenti in TopThreeAmericanName). Per 3 è stato possibile ricavare il valore in base ad altri record aventi SubModel o Model in comune. Per il quarto record non è stato possibile, dunque è stato eliminato;
- **Attributi di prezzo** - presentano dei missing values, che sono stati sostituiti con la media[1], raggruppando per **Modello**, al fine di ridurre al minimo l'errore (modelli di veicoli diversi hanno prezzi completamente diversi).

1.3.2 Outliers

Si descrive ora per ogni attributo come si è proceduto per gestire gli outliers:

- **Transmission** - in un record è presente il valore *Manual*, anziché *MANUAL*. È stato corretto;
- **Attributi di prezzo** - tali prezzi dei veicoli presentano valori pari a 0 o a 1 in molti record, si è quindi ritenuto opportuno sostituire tali valori con la media[1] dei prezzi dei veicoli. La media è stata fatta raggruppando per modello, per ridurre al minimo l'errore. Dalla figura 8 si possono osservare numerosi outliers, che si è deciso di eliminare a causa dei possibili problemi che possono causare nelle analisi successive.

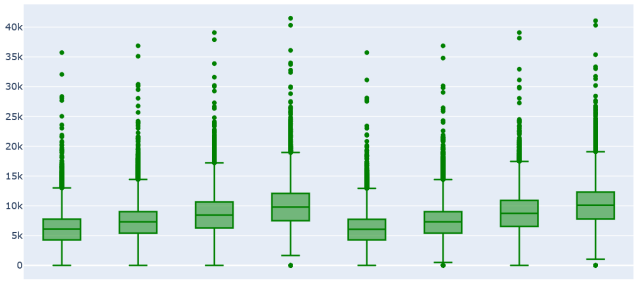

Figura 8: Distribuzione degli outliers nelle 8 variabili di prezzo

1.4 Trasformazione delle variabili

Molti degli attributi del Dataset presentano un elevata granularità e si è ritenuto necessario diminuire ove possibile. In particolare il dominio dell'attributo:

- **Size** è stato ridotto a {0=SMALL($\leq$ 3000lb), 1=MEDIUM(3000 - 3500lb), 2=LARGE($\geq$ 3500lb)}. Le classi sono state raggruppate secondo il peso in libbre, seguendo la categorizzazione effettuata da Wikipedia Vehicle Size Class.
- **Make** è stato raggruppato secondo i principali gruppi automobilistici: *GENERAL MOTORS (GM), CHRYSLER, FORD, HYUNDAI, MITSUBISHI, TOYOTA* e *OTHERS*. I dati sono stati in parte estrapolati dall'attributo **TopThreeAmericanName** e dai siti delle case automobilistiche, non incorporando le fusioni dal 2011 ad oggi.
- **Color** è stato trasformato raggruppando i colori secondo il 2011 DuPont Automotive Color Popularity Report osservando inoltre che le percentuali di distribuzione dei colori sono simili a quelle presenti nel Dataset. Il dominio risulta essere: *SILVER, BLUE, GREY, BLACK, RED, OTHER* e *NOT AVAIL*.
- **VNST**: Sono stati raggruppati secondo macroaree tra: *SOUTH, WEST, MID WEST* e *NORTH EAST*, in base alla divisione fornita da Wikipedia.
- **Nationality**: si è notato che nel dominio *OTHER* sono tutti appartenenti al continente Europeo quindi per rendere più intuitivo l'attributo il dominio è stato cambiato in: *AMERICAN, ASIAN* ed *EUROPEAN*.

I valori dei prezzi dei veicoli sono stati normalizzati utilizzando la normalizzazione **Min-Max** in modo da traslare i valori sulla stessa scala e rendere possibile eseguire algoritmi di clustering in modo efficiente. I seguenti attributi sono stati eliminati:

- **BYRNO, RefId**: perché numeri identificativi e non utili ai fini delle analisi.
- **SubModel, VNZIP1, Trim, Model**: perché aventi una granularità troppo elevata e avendo trovato un riscontro nei dati per poterla diminuire.
- **TopThreeAmericanName,WheelTypeID,VehYear, PurchDate**: perché ridondanti all'interno del Dataset.

1.5 Correlazioni ed eventuale eliminazione di variabili ridondanti

Nella Figura 10 è rappresentata la matrice di correlazione tra gli attributi: **VehicleAge, VehOdo, MMRAcquisitionAuctionAveragePrice, MMRAcquisitionAuctionCleanPrice, MMRAcquisitionRetailAveragePrice, MMRAcquisitonRetailCleanPrice, MMRCurrentAuctionAveragePrice, MMRCurrentAuctionCleanPrice, MMRCurrentRetailAveragePrice, MMRCurrentRetailCleanPrice, VehBCost, WarrantyCost**.

La prima cosa che si nota è che la correlazione dei prezzi è molto elevata: raggiunge livelli anche dello 0.99. Da questo è ragionevole pensare di mantenere un solo prezzo (**MMRAcquisitionAuctionAveragePrice**) e creare quattro nuovi attributi ottenuti calcolando la differenza fra, rispettivamente, quelli denominati *Current*, e i loro corrispettivi denominati *Acquisition*. Un esempio è l'attributo denominato **MMRDiffAuctionAveragePrice** ottenuto dalla differenza tra **MMRCurrentAuctionAveragePrice** e **MMRAcquisitionAuctionAveragePrice**, in modo da mettere in relazione due prezzi che possano rappresentare il dataset in maniera più sintetica. In particolare, una differenza negativa sta a significare che il prezzo corrente del veicolo è minore del prezzo a cui Carvana lo ha acquistato: il veicolo si è quindi svalutato.

[1] Media Aritmetica svolta su tutti i valori che non siano 1 o 0.

L'attributo **VehicleAge** è invece correlato negativamente con tutti i prezzi: l'aumento dell'età del veicolo corrisponde infatti ad una diminuzione dei prezzi. Tale correlazione è mostrata anche nella figura 9 dalla quale si può evincere come l'età sia un attributo che condiziona molto il valore di un veicolo, e che sia quindi molto importante per valutarne la convenienza di acquisto.

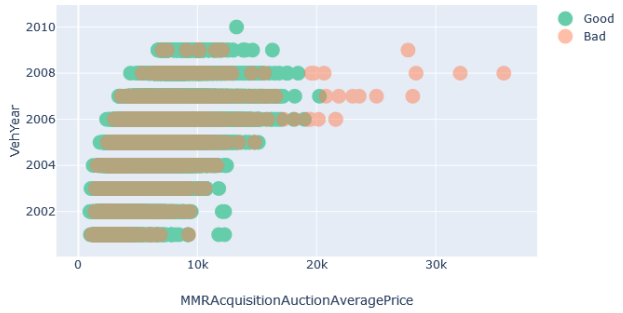

Figura 9: Correlazione fra VehYear e MMRAcquisitionAuctionAveragePrice

Infine è importante sottolineare come gli attributi **VehOdo** e **WarrantyCost** abbiano una correlazione positiva: un veicolo che abbia percorso più chilometri è presumibilmente più usurato e quindi sarà più costoso assicurarlo. Il fenomeno è confermato anche dalla correlazione positiva che si riscontra tra **WarrantyCost** e **VehicleAge**.

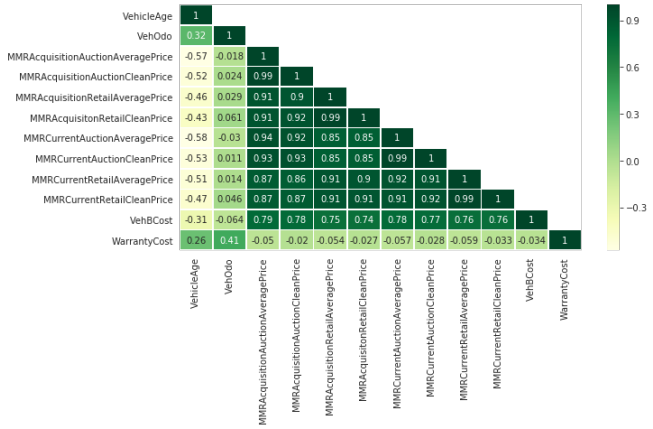

Figura 10: Matrice di correlazione

2 Clustering

In questa sezione si è cercato di raggruppare insiemi di acquisti che condividono alcune caratteristiche in modo da individuare correlazioni inaspettate e scoprire caratteristiche interessanti, che non si sono riscontrate nel Data Understanding. Gli algoritmi di clustering utilizzati sono il **K-Means**, il **DBSCAN** e l'**Hierarchical Clustering**. Si procede ora descrivendo le analisi svolte e i risultati ottenuti.

2.1 K-Means

Data l'alta correlazione dei prezzi fra di loro, si è deciso di svolgere diverse prove, sia mantenendo un prezzo solo, ossia **MMRAcquisitionAuctionAveragePrice**, sia utilizzando tutte e quattro le differenze definite nel precedente paragrafo, sia solamente con **MMRDiffAuctionAveragePrice**. I risultati sono stati riportati all'interno della Tabella 2

Min - Max		**SSE**	**Silhouette**
K = 8	All Differences	1188.27	0.21
K = 6	One Difference	1089.91	0.26
K = 7	One Price	1089.88	0.23
Z - Scaler			
K = 8	All Differences	211529.89	0.17
K = 6	One Difference	143291.89	0.19
K = 7	One Price	1036677.84	0.23

Tabella 2: Valori di Support e Silhouette in base agli attributi e alla normalizzazione

Il risultato migliore si ottiene utilizzando la normalizzazione *MIN-MAX* e solo una differenza, per questo motivo le successive analisi faranno riferimento a questo caso.

2.1.1 Scelta di K

K è stato selezionato attraverso i metodi *Elbow* e *Silhouette*, eseguendo l'algoritmo per un range di K fra 2 e 50. Se ne mostra il risultato nelle figure 11 e 12.

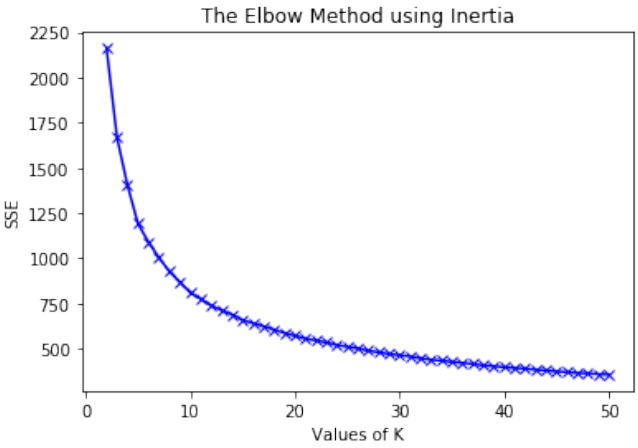

Figura 11: Elbow method application

Dalla figura 12 si può vedere come i valori migliori di K si attestino fra 4 e 6, e grazie alla figura 11 è possibile vedere come per K=6 il valore del SSE sia minore. Si è dunque optato per **K=6**.

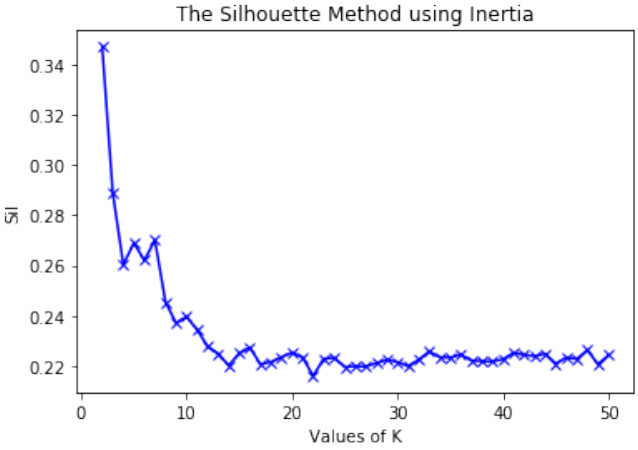

Figura 12: Silhouette method application

2.1.2 Contenuto dei cluster

Si elenca ora la dimensione di ogni cluster:

- $C0 \rightarrow 8745$;
- $C1 \rightarrow 13964$;
- $C2 \rightarrow 6424$;
- $C3 \rightarrow 5527$;
- $C4 \rightarrow 10805$;
- $C5 \rightarrow 10325$.

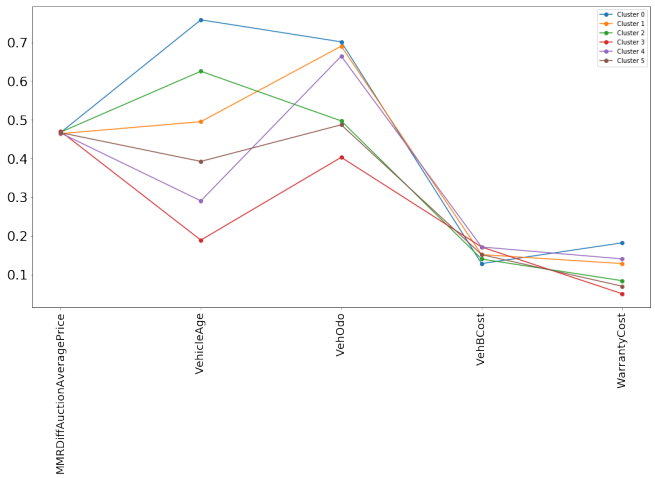

Figura 13: Coordinates Plot

Nella figura 13 si possono osservare alcuni trend dei cluster:

- i cluster 0 e 2 contengono in media i veicoli più anziani e che hanno percorso molti chilometri. Per questi veicoli il costo di acquisizione **VehBCost** è tra i più bassi, mentre il costo della garanzia **WarrantyCost** è tra i più alti per il cluster 0 mentre resta basso per il cluster 2, cosa che non ci si aspetterebbe;
- i cluster 3, 4 e 5 contengono in media veicoli giovani e il **WarrantyCost** invece più basso della media per il 3 e il 5, invece più alto per il 4.
- si nota come per tutti i cluster il valore di **MMRDiffAuctionAveragePrice** sia pressocché lo stesso, così come quello di **VehBCost**.

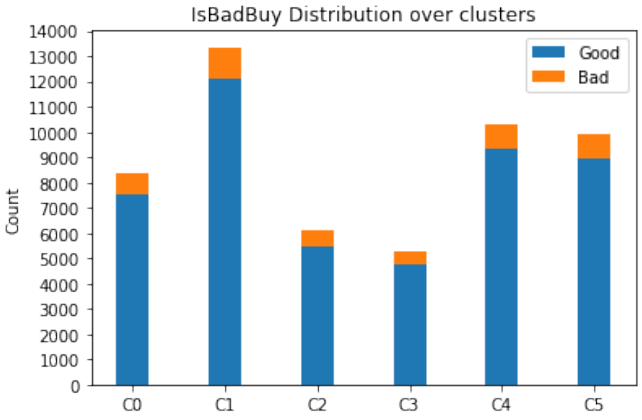

Figura 14: Distribuzione attributo Bad Buys sui Cluster

Come si vede dalla figura 14, nessun cluster spicca sugli altri per la quantità di *Good* o *Bad Buys*, infatti l'andamento è lo stesso per tutti, con un'alta percentuale di *Good Buys* e una minima di *Bad Buys*.

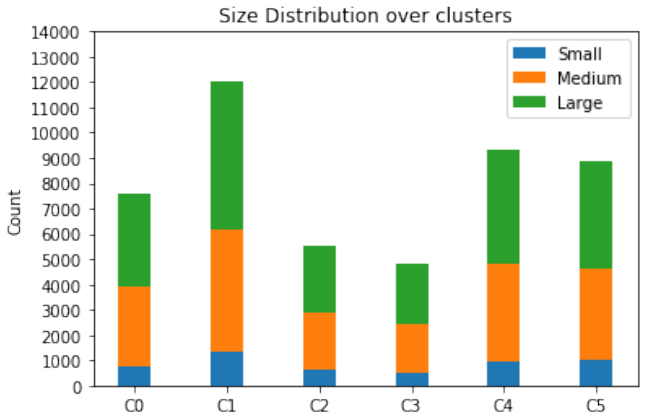

Figura 15: Distribuzione di Size sui Clusters

Dalla figura 15, si vede come anche la distribuzione dell'attributo **Size** sia la stessa per tutti i cluster, con una bassa percentuale di veicoli di piccola taglia, una percentuale alta di veicoli grandi e una percentuale media di veicoli di taglia media.

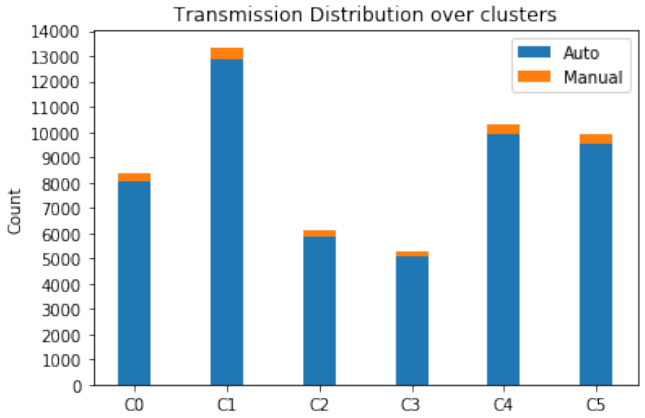

Figura 16: Distribuzione di Transmission sui Clusters

Similmente a quanto visto per gli altri attributi, come si vede dalla figura 16, anche per **Transmission** abbiamo che i cluster rispettano tutti la stessa distribuzione senza particolari differenze.

2.1.3 Sintesi dei risultati

Riassumendo i risultati ottenuti, si può affermare che attributi come **MMRDiffAcquisitionAveragePrice** e **VehBCost** contribuiscono meno alla classificazione delle tuple, mentre altri come **VehicleAge**, **VehOdo** e **WarrantyCost** siano molto importanti in questo processo, tuttavia si è visto come questi soli non siano sufficienti, infatti la distribuzione di **IsBadBuy** nei cluster è uniforme, così come per **Transmission** e **Size**.

2.2 DBSCAN

Per quanto riguarda il DBSCAN si è proceduto con le seguenti modalità:

1. Secondo quanto riportato da diversi paper, vi è la regola empirica per cui si sceglie come punto di riferimento per il parametro $min\_samples = 2 * DIM^2$, dove DIM è il numero di attributi utilizzati per il calcolo della distanza;
2. Si è poi visualizzato le distanze restituite da *KNN algorithm*, così da poter trovare graficamente l'*epsilon* corretto;
3. Quest'ultimo punto è stato iterato diverse volte, con valori di k diversi: si è potuto notare che al variare di k, *epsilon* varia minimamente, da considerarsi pressoché costante;
4. Visto il risultato osservato al punto 3, individuato graficamente un range dei possibili valori di epsilon, si è proceduto con degli esperimenti con *min_samples* fissato secondo la regola empirica del punto 1, e *epsilon* che varia all'interno del range sopra trovato;
5. Si è potuto trovare in questo modo il valore ottimo di epsilon, utilizzando come funzione obiettivo la seguente:
   - massimizzare il **Silhouette Score**
   - ma allo stesso tempo scartare tutti quei valori di *epsilon* che producono:
     * un cluster gigante con tutti i punti al suo interno
     * numerosi cluster ma con pochi punti al loro interno.
6. Trovato l'*epsilon* ottimo, si è proceduto a delle iterazioni con *epsilon* fissato, al variare di *min_samples*, così da poter trovare il valore ottimo di quest'ultimo (in quanto è vero che al variare di *min_samples*, ovvero k nel *KNN*, *epsilon* varia di pochissimo, ma non è vero che i risultanti cluster siano quelli ottimi). Per raggiungere tale obiettivo, si è utilizzato la stessa funzione obiettivo presentata al punto 5.

Tale procedimento si è ripetuto in tre casi, con attributi in input diversi:

1. Mantenendo le 4 differenze dei prezzi;
2. La sola differenza **MMRDiffAuctionAveragePrice**;
3. Un solo prezzo: **MMRAcquisitionAuctionAverage**.

[2]Sander, Jörg; Ester, Martin; Kriegel, Hans-Peter; Xu, Xiaowei (1998). *Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery.*

Infine il procedimento dei 6 punti sopra elencati e le 3 prove con gli attributi sopra specificati sono stati ripetuti sia con una normalizzazione *MinMax*, che con una *Standard*. Facendo notare che è stato possibile utilizzare il primo tipo di normalizzazione in quanto gli outliers erano stati eliminati in precedenza.

Nella Tabella 3 sono riportati alcuni risultati degli esperimenti fatti (non tutti, per motivi di spazio): in particolar modo questi sono i risultati finali, dopo aver ottimizzato *epsilon* e *min_samples*, nei vari casi sopra esposti. "Nessun risultato" significa che non è stato possibile trovare un combinazione di *epsilon* e *min_samples* che desse dei risultati validi (sempre un un unico grande cluster).

Normalizzazione	All differeces		One Difference		One Price	
MIN - MAX	eps	0.12	**eps**	**0.12**	eps	0.12
	min_samples	55	**min_samples**	**15**	min_samples	55
	Silhoutte	- 0.005636	**Silhoutte**	**0.0273224**	Silhoutte	- 0.0288784
Z- Scaler	eps	1.2	eps	1.1	eps	0.5
	Nessun Risultato		Nessun Risultato		min_samples	15
	Nessun Risultato		Nessun Risultato		Silhoutte	- 0.1173888

Tabella 3: Risultati finali, dopo aver ottimizzato epsilon e min_samples, con normalizzazione MinMax e Standard, con diversi tipi di attributi

Visti i risultati ottenuti si è deciso di procedere per una più approfondita analisi con una normalizzazione *MinMax* e mantenendo gli attributi del punto 2 sopra citato (casistica evidenziata nella tabella).

I cluster così trovati hanno le seguenti caratteristiche:

- Labels dei cluster: [-1, 0, 1, 2, 3, 4, 5, 6, 7, 8]
- Numero dei punti nei cluster: [443, 6328, 11886, 12863, 9595, 2332, 3270, 5820, 1552, 356]

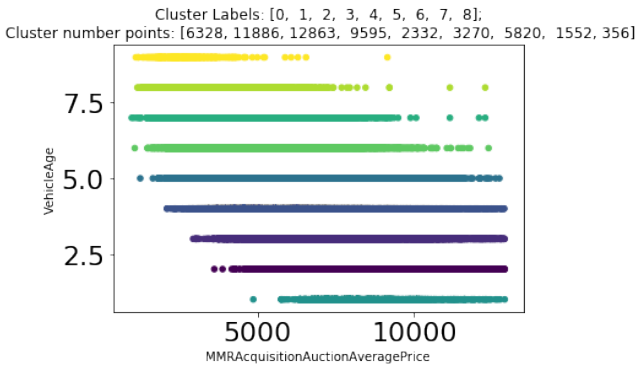

Figura 17: Cluster distribution su VehicleAge e MMRAcquisitionAuctionAveragePrice

Si può subito notare dal grafico in Figura 17, che **VehicleAge** è l'attributo determinante per la formazione dei cluster in *DBSCAN*. Questo ci permette comunque di osservare come i veicoli più vecchi abbiano un **MMRAcquisitionAuctionAveragePrice** più basso e quelle più nuove un prezzo più alto. Anche il numero di macchine con una **VehicleAge** media è maggiore rispetto al numero di auto più nuove o più vecchie.

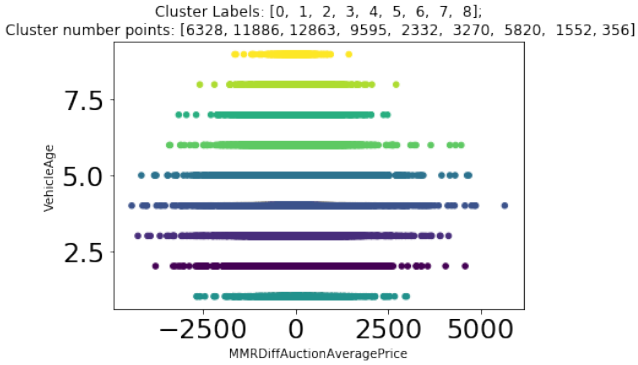

Figura 18: Cluster distribution su VehicleAge e MMRDiffAuctionAveragePrice

Del grafico in Figura 18 vale la pena osservare come sui veicoli con una **VehicleAge** media, la **MMRDiffAuctionAveragePrice** sia maggiore, così come la volatilità del prezzo d'asta.

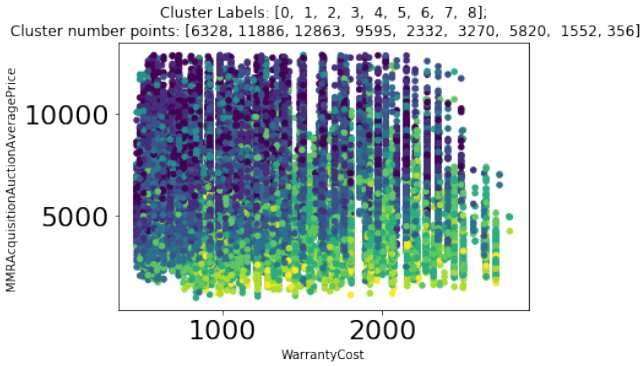

Figura 19: Cluster distribution su WarrantyCost e MMRAcquisitionAveragePrice

Dal grafico in Figura 19 è possibile notare invece come i veicoli più vecchi (colore giallo e verde chiaro), abbiano un maggiore **WarrantyCost**, che a sua volta è inversamente proporzionale al **MMRAcquisitionAveragePrice**.

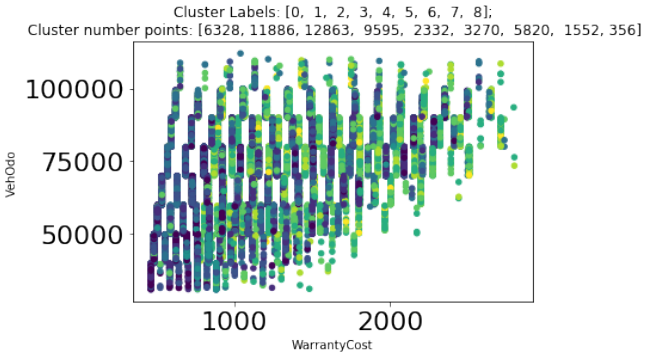

Figura 20: Cluster distribution su WarrantyCost e VehOdo

Infine dal grafico in Figura 20 si osserva che i veicoli più vecchi hanno ovviamente un maggior chilometraggio, aspetto che influisce sul **WarrantyCost** in maniera proporzionale.

2.2.1 Sintesi dei risultati

La forte dipendenza del *DBSCAN* clustering dall'attributo **VehicleAge** non permette di ottenere ulteriori informazioni; dunque nonostante un Silhouette Score positivo, il *DBSCAN* non è da ritenere un buon algoritmo di clustering per il dataset in questione: infatti rispetto agli altri attributi, non è possibile trovare una separazione netta dei cluster.

Metodo	Disanza	N. Cluster	Dimensione dei Cluster	Silhouette
Single	*Euclidea*	10	54494, 2, 2, 1, 1, 1, 1, 1, 1, 1	0.01378
Average	*Euclidea*	5	54467, 24, 8, 5, 1	0.12806
Complete	*Euclidea*	2	28825, 25680	0.20489
Ward	*Euclidea*	2	31647, 22858	0.22671
Single	*Manhattan*	10	54494, 2, 2, 1, 1, 1, 1, 1, 1, 1	0.01378
Average	*Manhattan*	5	46201, 8296, 4, 2, 1	0.12806
Complete	*Manhattan*	2	47960, 6545	0.20489
Ward	*Manhattan*	2	43402, 11103	0.22671

Tabella 4: Risultati delle analisi per diversi livelli di taglio e di distanza

Come si può notare dalla tabella è evidente che i risultati ottenuti sono simili sia utilizzando la ***Manhattan*** che la ***Euclidean distance*** (la Silhouette con la distanza euclidea è minore rispetto alla *Manhattan* per le caratteristiche proprie delle diverse definizioni). Le differenze maggiori che si riscontrano sono però nella dimensione dei cluster ottenuti: utilizzando la *Manhattan Distance* si rileva uno sbilanciamento nella distribuzione dei cluster in particolare riguardo ai metodi *Complete* e *Ward*. Con la distanza Euclidea risultano invece più bilanciati. Nella figura 21 e nella figura 22 sono rappresentati i dendrogrammi in riferimento rispettivamente alla distanza *Manhattan* e a quella *Euclidea* con il metodo del *linkage complete*, dove si può notare con maggiore accuratezza il fenomeno.

2.3 Hierarchical

Il terzo metodo utilizzato al fine della Cluster Analysis è il gerarchico. Le analisi sono state svolte applicando la tipologia *Agglomerative* sulla base di diverse definizioni di distanza (**Euclidean** e **CityBlock**) e di *cluster proximity (Single, Average, Complete e Ward)*. In questo caso sono stati utilizzati i seguenti attributi: **VehicleAge, VehOdo, MMRDiffAuctionAveragePrice, VehBCost** e **WarrantyCost**. Infine si vuole specificare che tutte le prove fatte sono state basate sul dataset completo, senza l'utilizzo di un sampling, grazie all'utilizzo di un Cloud esterno (Colab di Google).

Nella tabella 4 sono rappresentati i principali risultati delle analisi svolte in particolare in riguardo alla Silhouette corrispondente al taglio effettuato e alle dimensioni dei cluster ottenuti.

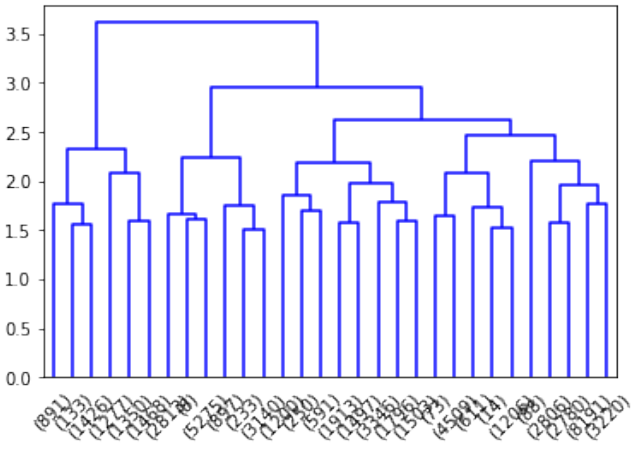

Figura 21: Complete Link attraverso la Manhattan Distance

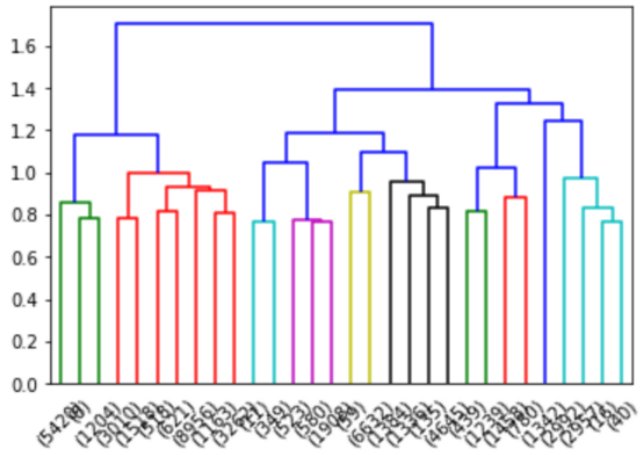

Figura 22: Complete Link attraverso la Euclidean Distance

Per i motivi sopracitati le successive osservazioni saranno riportate solo in riferimento alla distanza Euclidea.

2.3.1 Single Link

Il metodo del **Single link** produce un dendrogramma che non evidenzia nessun tipo di raggruppamento significativo: anche tagliando l'albero all'altezza di un numero di cluster pari a 100 si ottengono comunque 99 cluster contenenti meno di 5 record e uno molto grande contenente i restanti. Nonostante quindi una Silhouette, con taglio all'altezza di 0.3 e numero di cluster pari a due, piuttosto alta (pari a ben 0.32925) non si è ritenuto opportuno svolgere ulteriori analisi.

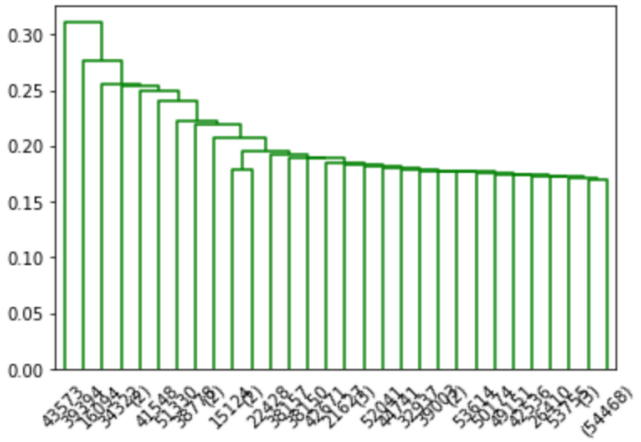

Figura 23: Dendrogramma Single Link

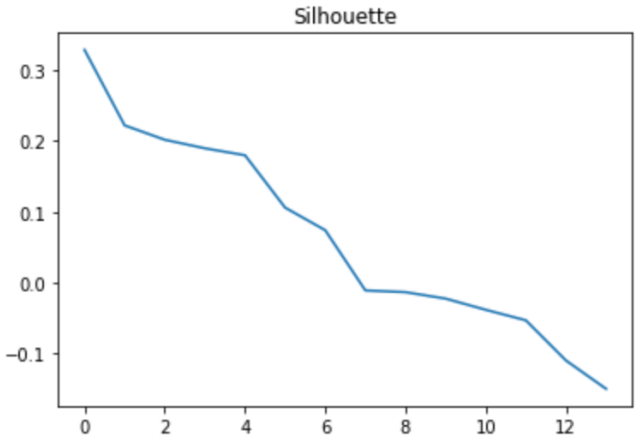

Figura 24: Andamento della Silhouette per il Single Link

2.3.2 Group Average

Con questo metodo si ottiene il livello di Silhouette maggiore pari a 0.47 con un taglio all'altezza di 0.8 ottenendo due cluster composti da 8 e da 54497 record. Data la natura di questi cluster si è provato a tagliare il dendrogamma in un punto diverso, in particolare all'altezza di 0.55, punto in cui la silhouette ha un picco positivo che corrisponde a una divisione del dataset in 7 classi da: 43743, 10696, 28, 24, 8, 5, 1 record. Ponendo quindi maggiore attenzione alle prime due classi con un maggior numero di record si è notato che:

- Il cluster da 43743 risulta composto da veicoli medio-giovani, con un chilometraggio medio basso e con una maggiore presenza di differenze di prezzo intorno allo zero oltre che un ***WarrantyCost*** medio basso. In questo cluster risulta una maggiore presenza di veicoli "*buoni*";
- Il cluster da 10696 risulta composto da veicoli molto vecchi con un alto livello di garanzie, maggiore chilometraggio e differenze di prezzo spostate maggiormente verso livelli sotto lo zero. In questo caso sono principalmente presenti veicoli con **IsBadBuy** pari a 1.

2.3.3 Complete e Ward

Le analisi svolte con metodi **Complete e Ward** sono state quelle che hanno prodotto i risultati con cluster maggiormente equilibrati nonostante la Silhouette non risultasse la migliore. Ponendo l'attenzione sul metodo del Ward, che ha ottenuto un livello di Silhouette maggiore, si possono osservare dei risultati interessanti tagliando il dendrogramma a sei cluster:

- Classe 1: veicoli in medie condizioni
- Classe 2: veicoli in medie-cattive condizioni
- Classe 3: veicoli in cattive condizioni

- Classe 4: veicoli in medie-buone condizioni
- Classe 5: veicoli in buone condizioni
- Classe 6: veicoli in ottime condizioni

La figura 25 e la figura 26 rappresentano la distribuzione delle classi sull'attributo **VehicleAge**, rispettivamente con 6 e 4 cluster. Si può notare, in riferimento al raggruppamento in 6 cluster, che:

- I cluster 1 e 2 possono essere racchiusi in uno solo composto da veicoli in medie-cattive condizioni
- I cluster 5 e 6 possono essere racchiusi tra i veicoli in medie-buone condizioni
- I cluster 4 e 3 rimangono invariati e corrispondono rispettivamente a veicoli in medie-ottime e in ottime condizioni.

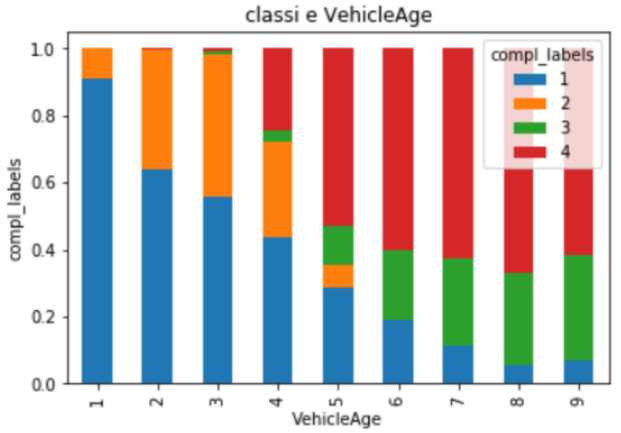

Figura 25: VehicleAge su 4 Cluster

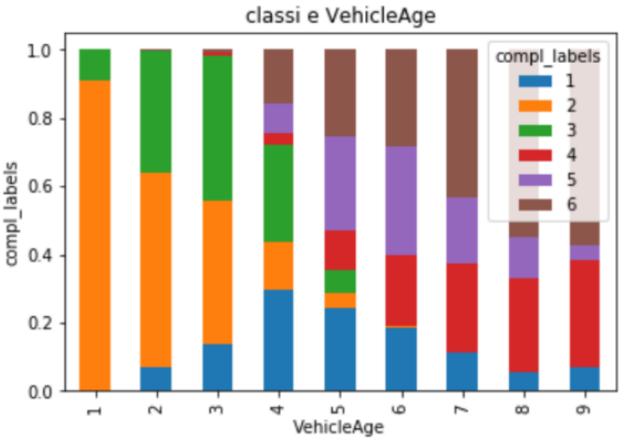

Figura 26: VehicleAge su 6 Cluster

2.3.4 Sintesi dei risultati

In conclusione, si può affermare che nonostante un dataset molto ampio, attraverso il metodo gerarchico utilizzando in particolare il **Ward e Complete linkage**, si è riusciti ad arrivare a una buona clusterizzazione che permette di distinguere in maniera abbastanza approfondita e a seconda del taglio diverse classi che descrivono in che condizioni versano i diversi veicoli acquistati da Carvana.

2.4 Valutazione finale del Clustering

Dall'analisi compiuta, risulta evidente come, secondo il valore della Silhouette, il *K-Means* e lo *Hierarchical* producano i risultati migliori, mentre quella ottenuta con il *DBSCAN* sia notevolmente più bassa, poco maggiore di zero.

Andando a considerare invece, la qualità dei cluster ottenuti, i risultati migliori sono riscossi attraverso lo *Hierachical Clustering* che riesce a individuare cluster eterogenei di veicoli sulla base della **VehicleAge**, dei chilometri percorsi (**VehOdo**), delle differenze di prezzo e in parte anche sulla base dell'attributo target **IsBadBuy**.

3 Association Rules

In questa sezione verranno analizzate le principali regole di associazioni che possono essere individuate all'interno del dataset e le applicazioni di esse per la determinazione e la sostituzione dei Missing Values e nella classificazione dell'attributo target dell'analisi al fine di poter differenziare i record *IsBadBuy_0* da quelli *IsBadBuy_1*.

3.1 Preparazione del Dataset

Gli attributi presi in considerazione nell'analisi sono: **IsBadBuy, Auction, Make, Color, Transmission, WheelType, Nationality, Size, IsOnlineSale, VNST, VehicleAge, VehOdo, VehBCost, WarrantyCost, MMRDiffAuctionAveragePrice**. In particolare, per gli attributi numerici quali **VehicleAge, VehOdo, VehBCost, WarrantyCost e MMRDiffAuctionAveragePrice** è stato applicato il metodo di discretizzazione per individuare 3 intervalli con la stessa frequenza in modo da permettere all'algoritmo *Apriori* di funzionare nel modo corretto. Sempre per lo stesso motivo, il dataset è stato trasformato in uno di tipo *transactional*.

3.2 Estrazione dei frequent itemset e osservazioni sugli itemset al variare del supporto minimo

Il problema principale in questa prima fase è stata l'individuazione del parametro di supporto minimo più idoneo a rappresentare il dataset.

Il metodo utilizzato è stato quello di mettere in relazione il numero di itemset prodotti, con la soglia di supporto minimo, facendola variare da un minimo di 0.05 a un massimo 0.95. Si ottiene una curva che presenta un *"gomito"* fra i valori 0.10 e 0.20, come si può osservare dalla figura 27.

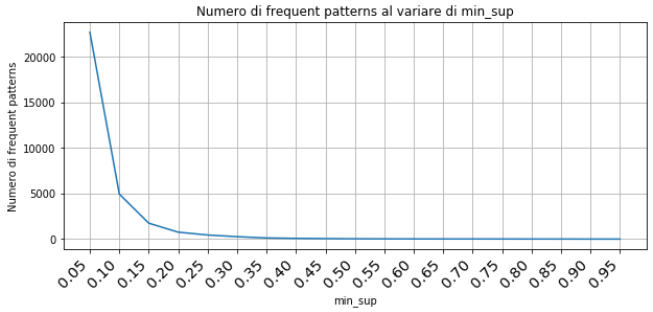

Figura 27: Numero dei Frequent Itemset al variare del min supp

Nella tabella 5 vengono mostrati il numero di itemset in relazione alla soglia di supporto minimo impostata, in particolare per i valori nel gomito della curva.

Supp_min	Numero Itemset
0.10	4946
0.15	1747
0.20	761

Tabella 5: Numero dei frequent itemset individuati per il min_supp compreso tra 0.1 e 0.20

Si è deciso di adottare un supporto minimo pari a 0.10 in modo da poter comprendere un maggior numero di itemset frequenti. I frequent itemset sono riportati nelle tabelle 6 e 7.

Come si può notare nelle tabelle 6 e 7 l'itemset di lunghezza pari a 4 condivide tutti gli elementi con gli itemsets di lunghezza 2 e 3, aggiungendo *Nationality_American* e *IsBadBuy_0*. Sono quindi molto frequenti i veicoli che hanno un cambio automatico, comprati originariamente online, Americani e che alla fine sono risultati buoni acquisti.

Lunghezza	Support	Itemset
2	0.939427	(Transmission_AUTO, IsOnlineSale_0)
3	0.850093	(IsBadBuy_0, Transmission_AUTO, IsOnlineSale_0)
4	0.719213	(IsBadBuy_0, Transmission_AUTO, IsOnlineSale_0, Nationality_AMERICAN)
5	0.477565	(IsBadBuy_0, VNST_SOUTH, Transmission_AUTO, IsOnlineSale_0, Nationality_AMERICAN)
6	0.273881	(IsBadBuy_0, VNST_GEO_SOUTH, Transmission_AUTO, IsOnlineSale_0, Auction_MANHEIM, Nationality_AMERICAN)
7	0.150109	(IsBadBuy_0, VNST_GEO_SOUTH, Transmission_AUTO, IsOnlineSale_0, WheelType_Alloy, Auction_MANHEIM, Nationality_AMERICAN

Tabella 6: lista degli itemsets più frequenti con lunghezze da 2 a 7

Lunghezza	Support	Itemset
2	0.939427	(Transmission_AUTO, IsOnlineSale_0)
2	0.881224	(IsBadBuy_0, IsOnlineSale_0)
2	0.872224	(IsBadBuy_0, Transmission_AUTO)
3	0.850093	(IsBadBuy_0, Transmission_AUTO, IsOnlineSale_0)
3	0.794075	(Transmission_AUTO, IsOnlineSale_0, Nationality_AMERICAN)
3	0.738810	(IsBadBuy_0, IsOnlineSale_0, Nationality_AMERICAN)

Tabella 7: tre itemset più frequenti con lunghezza compresa tra 2 e 3

Si può notare che gli itemset più frequenti ottenuti sono correlati al forte sbilanciamento dei diversi attributi nel dataset tra cui *IsOnlineSale_0* (presente per il 97%) e *Transmission_AUTO* (presente per il 96%).

3.3 Estrazione delle regole di associazione con diversi valori di confidence e discussione sulle regole più interessanti

In questa sezione la *min_conf* è stata fatta variare tra un minimo di 0.05 a un massimo di 0.95: nella figura 28 è riportata la relazione tra il numero delle *Assosiciation Rule* e il differente valore di *min_conf* che si è applicato per ottenerle.

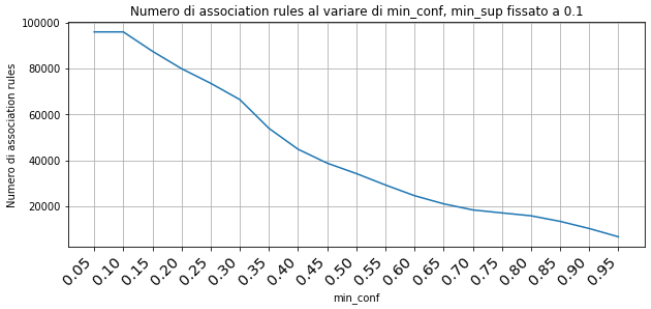

Figura 28: Numero AR per diversi livelli di confidence minima

Nella figura 29 viene riportato il totale di regole per i diversi valori di *min_conf* da 0.8 a 1 in modo da individuare con maggior dettaglio quale sia l'andamento delle Association Rule.

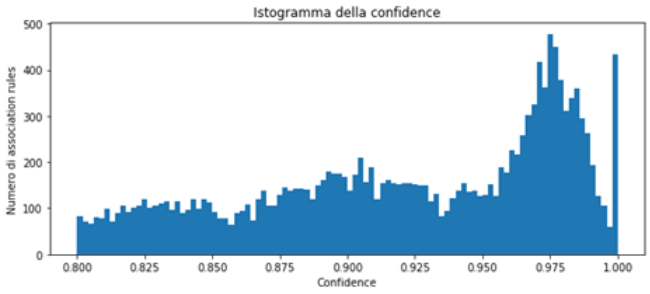

Figura 29: Confidence maggiore dello 0.8

L'alto numero di regole individuato con una confidence pari a 1 sono quelle che hanno come conseguente *Nationality_AMERICAN* e come antecedente una combinazione di marche americane (attributo **Make**) con gli altri attributi: sono regole tautologiche, in quanto sempre vere (ad esempio $(Make\_GM) \rightarrow (Nationality\_AMERICAN)$ o $(Make\_CHRYSLER, Color\_SILVER) \rightarrow (Nationality\_AMERICAN)$).

Come si nota vi è un picco del numero di *AR* a partire da 0.95: si è dunque deciso di analizzare il valore di *Lift* delle regole ottenute.

Nella figura 30 è riportato l'istogramma contenente i diversi livelli di *Lift* che si ottengono dalle regole ottenute ponendo i parametri del supporto e della *confidence* minimi pari rispettivamente allo 0.1 e allo 0.8.

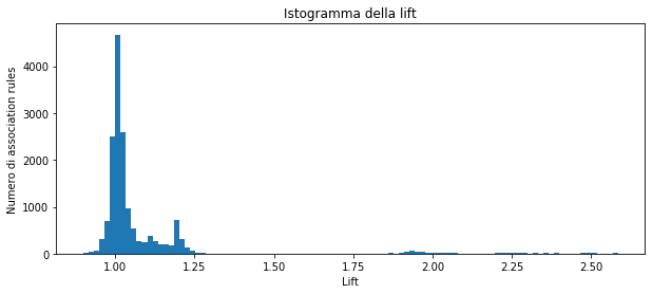

Figura 30: Distribuzione del valore di Lift in base alle Association Rules

Si presenta una forte concentrazione di regole intorno a un *Lift* pari a 1, segno evidente della presenza di regole con antecedenti e conseguenti indipendenti. Molto più interessanti sono invece le regole con una *Lift* maggiore di 1.5, riportate nella tabella 31.

	antecedents	consequents	lift
13420	(DiscretizationVehBCost_2, Nationality_AMERICA...	(Size_2, Transmission_AUTO)	2.304786
15494	(DiscretizationVehBCost_2, IsOnlineSale_0, Nat...	(Size_2, Transmission_AUTO)	2.303702
8838	(DiscretizationVehBCost_2, Nationality_AMERICA...	(Size_2, Transmission_AUTO)	2.301514
13405	(DiscretizationVehBCost_2, IsBadBuy_0, Nationa...	(Size_2, Transmission_AUTO)	2.300119
13390	(DiscretizationVehBCost_2, IsBadBuy_0, IsOnlin...	(Size_2, Transmission_AUTO)	2.281756
8831	(DiscretizationVehBCost_2, IsOnlineSale_0, Dis...	(Size_2, Transmission_AUTO)	2.279977
8810	(DiscretizationVehBCost_2, IsBadBuy_0, Discret...	(Size_2, Transmission_AUTO)	2.278459
3768	(DiscretizationVehBCost_2, DiscretizationWarra...	(Size_2, Transmission_AUTO)	2.277123
13416	(DiscretizationVehBCost_2, IsOnlineSale_0, Nat...	(Size_2)	2.270340
15487	(DiscretizationVehBCost_2, IsOnlineSale_0, Nat...	(Size_2)	2.269486
8815	(DiscretizationVehBCost_2, Nationality_AMERICA...	(Size_2)	2.269250
13356	(DiscretizationVehBCost_2, IsOnlineSale_0, Nat...	(Size_2)	2.268308
8836	(DiscretizationVehBCost_2, Nationality_AMERICA...	(Size_2)	2.267013
3761	(DiscretizationVehBCost_2, Nationality_AMERICA...	(Size_2)	2.265970
13401	(DiscretizationVehBCost_2, Nationality_AMERICA...	(Size_2)	2.265841
8794	(DiscretizationVehBCost_2, IsBadBuy_0, Nationa...	(Size_2)	2.264715
13386	(DiscretizationVehBCost_2, IsOnlineSale_0, IsB...	(Size_2)	2.250465
9826	(DiscretizationVehicleAge_0, Nationality_AMERI...	(Size_2, Transmission_AUTO)	2.249857
8780	(DiscretizationVehBCost_2, IsBadBuy_0, IsOnlin...	(Size_2)	2.249035
8829	(DiscretizationVehBCost_2, Transmission_AUTO, ...	(Size_2)	2.248642
4488	(DiscretizationVehicleAge_0, DiscretizationWar...	(Size_2, Transmission_AUTO)	2.247383
3755	(DiscretizationVehBCost_2, IsOnlineSale_0, Dis...	(Size_2)	2.247339
8808	(DiscretizationVehBCost_2, IsBadBuy_0, Transmi...	(Size_2)	2.246996

Figura 31: Regole più frequenti con un Lift maggiore di 1.5

3.4 Sostituzione dei missing values tramite AR e valutazione della relativa accuratezza

Gli attributi che, dopo aver applicato tutte le trasformazioni, presentano valori mancanti sono mostrati nella tabella 8.

Attributi	Valori Mancanti
Color	84
Transmission	8
WheelType	2577
Nationality	4
Size	4
MMRDiffAuctionAveragePrice	245

Tabella 8: Regole più frequenti

Provando a sostituire tali valori mancanti, e confrontandoli con le sostituzioni svolte nella sezione del Data Understanding, si ha:

- **Transmission:** tutte le regole trovate implicano *Transmission_AUTO*, per tutti gli 8 valori mancanti. Cosa molto probabile considerando che quasi il 97% dei dati ha *Transmission_AUTO*. La sostituzione in questo caso è risultata corretta con quella già svolta.

- **Nationality:** i 4 valori mancanti sono stati sostituti con *American*. La correttezza dell'analisi è stata verificata dal fatto che l'attributo **Nationality** è derivabile dall'attributo **Make**.

- **Size:** tutti i record mancanti sono stati sostituiti con 2 (corrispondente a veicoli classificati come *LARGE*). In questo caso la sostituzione non è risultata coerente con quella svolta nella gestione dei missing values svolta nella sezione di Data Understanding. Due record erano stati sostituiti con il valore 1 (*Medium*), mentre altri due non presentano informazioni per verificare la bontà della sostituzione. Si deve comunque tener presente che è possibile che alcune caratteristiche di un veicolo di medie dimensioni siano molto simili a quelle di un veicolo più grande.

- **MMRDiffAuctionPrice, WheelType e Color:** non sono presenti regole che riguardano gli attributi in questione e per questo motivo non si è stati in grado di sostituirli.

3.5 Previsione della variabile target e valutazione dell'accuratezza

Usando le regole prodotte nelle precedenti analisi dovrebbe risultare possibile predire le due classi dell'attributo target: IsBadBuy = 1 (Yes) e IsBadBuy = 0 (No).

Non è risultato possibile individuare regole che contenessero nel conseguente *IsBadBuy_1*. Al contrario, per la classe *No* sono state individuate 5103 association rules in grado di descrivere veicoli che risultino buoni acquisti. Di seguito nella tabella 9 si riportano alcune regole utilizzate per predire *IsBadBuy_0*, con la relativa accuratezza della singola regola. Si può osservare come essa sia alta per tutte le regole, risultato che ci si poteva aspettare vista la distribuzione della classe.

Antecedenti	Conseguente	Accuracy
DiscretizationVehicleAge_0, Auction_OTHER	IsBadBuy_0	0,935201401
DiscretizationVehicleAge_0, DiscretizationMMRDiffAuctionAveragePrice_1	IsBadBuy_0	0,905817175
DiscretizationVehicleAge_0, VNSTGEO_SOUTH	IsBadBuy_0	0,818068923
Transmission_AUTO, IsOnlineSale_0, DiscretizationVehicleAge_0, Size_1, DiscretizationWarrantyCost_1	IsBadBuy_0	0,892086331
Transmission_AUTO, IsOnlineSale_0, DiscretizationVehicleAge_0, WheelType_Covers, Nationality_AMERICAN, VNSTGEO_SOUTH	IsBadBuy_0	0,967340591

Tabella 9: Accuracy di alcune Association Rule

È chiaro infatti come questo genere di classificazione riesca a classificare i veicoli solo come "Good Buy" e non riesca a percepire le sfumature che caratterizzano invece i veicoli "Bad Buy". Questo fenomeno, come discusso anche nella sezione 4, è attribuito a un forte sbilanciamento del dataset a favore della classe *NO*: si è stati quindi costretti a procedere verso un bilanciamento del dataset in modo da consentire all'algoritmo di individuare regole riguardanti il valore YES. Il bilanciamento in questa fase è stato svolto attraverso un *Up-sampling*, ovvero un bilanciamento che prevede una duplicazione casuale delle osservazioni della classe di minoranza (nel caso in questione della classe Yes) al fine di rafforzarne il segnale. Attraverso il bilanciamento e dopo una nuova esecuzione dell'algoritmo *Apriori* e applicando le regole ottenute per la predizione delle classi si ottengono comunque risultati non positivi: anche ponendo un bilanciamento, l'algoritmo non riesce a trovare regole aventi come conseguente *IsBadBuy_1*,riuscendo quindi a classificare i veicoli solo come "Good Buy".

4 Classification

In questa sezione verranno analizzati i diversi modelli di classificazione che permettono di predire le classi 0 (definita come *No*) e 1 (definita come *Yes*) dell'attributo **IsBadBuy**. Nella prima parte ci si focalizzerà sulla scelta dei parametri da utilizzare e in particolar modo sulla distribuzione sbilanciata delle classi e dei metodi per far fronte a tale problema. La seconda e terza parte invece, saranno dedicate all'analisi, all'implementazione e alla validazione dei diversi modelli di classificazione. Nell'ultima parte si prenderanno delle decisioni in merito al modello più idoneo alla predizione.

4.1 Ottimizzazione e scelta dei parametri

Gli attributi individuati come rilevanti al fine la classificazione sono: **Make, Transmission, Auction, Color, Nationality, VNST, WheelType, VehicleAge, VehOdo, VehBCost, WarrantyCost, MMRDiffAuctionAveragePrice** e infine, l'attributo target **IsBadBuy**. Al fine di garantire una più facile costruzione del modello, gli attributi numerici (**VehicleAge, VehOdo, VehBCost, WarrantyCost, MMRDiffAuctionAveragePrice**) sono stati discretizzati in modo da ottenere colonne di uguale frequenza e poi trasformati grazie al ***LabelEncoder*** per garantire l'ordine intrinseco dei numeri, mentre per gli attributi categorici è stato utilizzato il ***OneHotEncoder*** che garantisce che non ci siano classi ordinali. Un punto importante da dover analizzare è infine la distribuzione dell'attributo target **IsBadBuy**: come si è notato dalle precedenti analisi, la colonna in oggetto presenta un forte sbilanciamento dovuto a un'eccessiva presenza di valori della classe *No* rispetto a quella *Yes*. Questo forte sbilanciamento rischia di essere dannoso per l'applicazione degli algoritmi, tanto da portare il modello a riuscire a classificare solo la classe

dei *No* e a tralasciare del tutto l'altra. Per questi motivi è risultato necessario ricorrere a due tipi di bilanciamento:

- ***Up – Sampling:*** nel quale si prevede una duplicazione casuale delle osservazioni della classe di minoranza al fine di rafforzarne il segnale.

- ***Down – Sampling:*** implica la rimozione casuale delle osservazioni dalla classe maggioritaria per impedire al suo segnale di dominare sull'algoritmo di apprendimento.

Tutti gli algoritmi sono stati comunque eseguiti sia con il dataset non bilanciato sia con il dataset bilanciato nei due modi sopraindicati in modo da avere un quadro il più completo possibile.

4.2 Apprendimento dei diversi algoritmi di classificazione e validazione dei modelli

Gli algoritmi di classificazione che sono stati utilizzati in queste analisi sono il *Decision Tree*, il *Random Forest* e il *KNN*. Per ognuno di questi algoritmi sono stati applicati diversi parametri e diversi criteri in modo da evitare l'overfitting; Nella tabella 10 sono stati riportati i parametri e le prove che sono state fatte:

Parametri	Valori
Criterio	*Gini , Entropy*
min_sample_leaf	1 - 10.000
min_sample_split	2 - 10.000
max_depth	*None* , 0 - 40
splitter	*best*

Tabella 10: criteri utilizzati per la classificazione

4.2.1 Decision Tree Algorithm

Per valutare i diversi risultati ottenuti con l'utilizzo del *Decision Tree* come modello di classificazione è stata presa in considerazione la misura del *F1-Score*, che incorpora sia il parametro *Recall* che il parametro *Precision*. Nella tabella 11 sono riportati i principali risultati ottenuti nelle varie analisi svolte.

Criterion	Parametri			Non bilanciato F-1 Score	Up - Sampling F-1 Score	Down - Sampling F-1 Score
Gini	*min_sample_split*	25	No	0.95	0.78	0.73
	min_sample_leaf	50				
	max_depth	None	Yes	0.01	0.21	0.23
	min_sample_split	300	No	0.95	0.75	0.75
	min_sample_leaf	150				
	max_depth	None	Yes	0.00	0.23	0.24
	min_sample_split	**5000**	**No**	**0.95**	**0.79**	**0.56**
	min_sample_leaf	**2000**				
	max_depth	**None**	**Yes**	**0.00**	**0.25**	**0.22**
Entropy	*min_sample_split*	500	No	0.95	0.76	0.75
	min_sample_leaf	250				
	max_depth	None	Yes	0.00	0.25	0.24
	min_sample_split	**5000**	**No**	**0.95**	**0.79**	**0.53**
	min_sample_leaf	**2000**				
	max_depth	**None**	**Yes**	**0.00**	**0.25**	**0.22**
	min_sample_split	700	No	0.95	0.76	0.76
	min_sample_leaf	200				
	max_depth	None	Yes	0.00	0.25	0.24

Tabella 11: Migliori risultati ottenuti sulla base dell'F1-Score per il Decision Tree

Osservando la tabella si può notare come in tutti i casi, nel dataset non bilanciato lo *F1-Score* sia il più alto e pari al 95%. La spiegazione del fenomeno è che sia nel training set che nel test set sono presenti maggiormente valori della classe *No* (circa il 90%), che portano a uno sbilanciamento dei dataset. Andando più nel dettaglio, nella maggior parte dei casi si ottiene un modello di classificazione che riesce a predire solo *IsBadBuy_No* risultando insensibile a *IsBadBuy_Yes*, in contrapposizione con l'obiettivo che ci si è preposto.

Come evidenziato nella tabella, i migliori risultati si ottengono utilizzando il dataset bilanciato attraverso l'*Up-Sampling* sia nel caso dell'indice *Gini* che con l'*Entropia* attuando un *pre-pruring* con i parametri *min sample_split* = 5000 e *min sample leaf* = 2000, nonostante comunque non risultino molto distanti rispetto a quelli ottenuti con le altre prove.

Validazione dei risultati

Di seguito (tabella 12) sono messi a confronto i due risultati migliori con i diversi indici – *Gini* e *Entropy* – confrontando anche le altre misure al fine di determinare il risultato migliore.

	Accuracy Train	Accuracy Test	ROC - Curve	Classe	Precision	Recall
Entropy	0.6242	0.6684	0.63	No	0.94	0.68
				Yes	0.16	0.58
Gini	0.6242	0.6684	0.63	No	0.94	0.68
				Yes	0.16	0.58

Tabella 12: Confronto tra i due risultati migliori ottenuti nel Decision Tree

Le unità di misura risultano quindi identiche in entrambi i casi: entrambe riescono a individuare l'albero di decisione migliore. Di seguito (tabella 32) è riportata

anche la rappresentazione della *ROC Curve* relativa al modello.

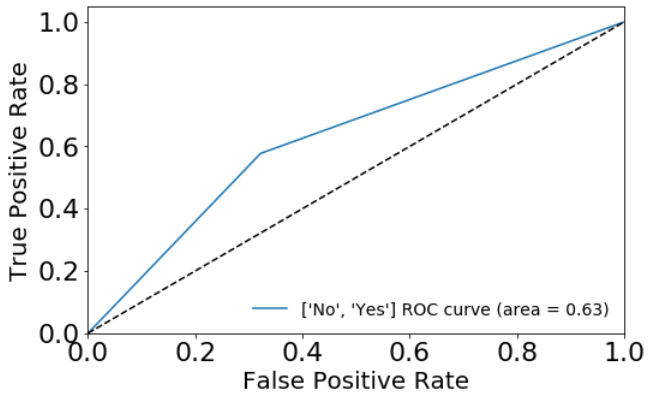

Figura 32: ROC Curve relativa all'albero di decisione

4.2.2 Random Forest

Le stesse modalità di indagine utilizzate per il *Decision Tree* sono state applicate nell'algoritmo del *Random Forest*. Di seguito è riportata una tabella riguardante i tre migliori risultati ottenuti applicando i diversi parametri per ogni indice di *split* (*Entropy* e *Gini*). Nel caso del *Random Forest* è stato tenuto conto anche di un altro parametro per evitare l'*overfitting*: *n_estimator*.

Criterion	Parametri			Non bilanciato F-1 Score	Up - Sampling F-1 Score	Down - Sampling F-1 Score
Gini	*n_estimator*	100				
	min_sample_split	500	No	0.95	0.75	0.73
	min_sample_leaf	250	Yes	0.00	0.25	0.24
	max_depth	None				
	n_estimator	100				
	min_sample_split	300	No	0.95	0.76	0.74
	min_sample_leaf	150	Yes	0.00	0.25	0.25
	max_depth	None				
	n_estimator	**300**				
	min_sample_split	**200**	**No**	**0.95**	**0.78**	**0.76**
	min_sample_leaf	**30**	**Yes**	**0.00**	**0.25**	**0.25**
	max_depth	**None**				
Entropy	*n_estimator*	100				
	min_sample_split	500	No	0.95	0.75	0.73
	min_sample_leaf	250	Yes	0.00	0.25	0.24
	max_depth	None				
	n_estimator	**100**				
	min_sample_split	**300**	**No**	**0.95**	**0.76**	**0.74**
	min_sample_leaf	**150**	**Yes**	**0.00**	**0.25**	**0.25**
	max_depth	**None**				
	n_estimator	300				
	min_sample_split	200	No	0.95	0.75	0.75
	min_sample_leaf	30	Yes	0.00	0.25	0.25
	max_depth	None				

Tabella 13: Migliori risultati ottenuti sulla base dell'F1-Score per il Random Forest

Anche in questo caso i risultati del dataset non bilanciato risultano completamente privi di significato in quanto non sensibili alla classe "Yes" e quindi non saranno in considerazione anche nelle successive analisi. Come nel Decision Tree anche qui i risultati ottenuti sono tutti simili gli uni agli altri.
I migliori risultati ottenuti sono quelli con il Gini index che utilizzano il dataset con Up-Sampling e con l'F1-Score pari a 0.78 e a 0.25 rispettivamente nelle classi "No" e "Yes". Simili risultati si ottengono utilizzando come indice l'Entropia come evidenziato nella tabella 13.

Validazione dei risultati

Analizzando le altre misure del Random Forest che generano il miglior risultato (F1-Score pari a 0.78 e 0.25) si ottengono dei buoni valori soprattutto in relazione alla ROC Curve. I dati riguardanti le misure sono riportati nella tabella 14, mentre la figura 33 rappresenta la ROC Curve generata dal modello.

	Accuracy Train	Accuracy Test	ROC - Curve	Classe	Precision	Recall
Gini	0.682	0.655	0.64	No	0.95	0.66
				Yes	0.13	0.61

Tabella 14: Misure ottenute dal risultato migliore del Random Forest

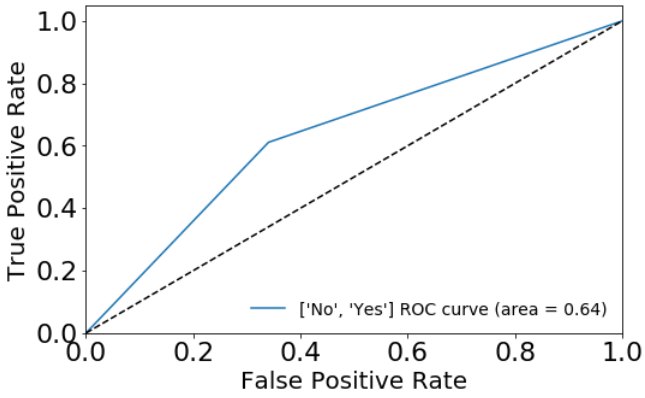

Figura 33: ROC Curve relativa al Random Forest

4.2.3 KNN

Per il KNN è stato utilizzato il parametro *n_neighbors* con diversi livelli di profondità: 5 – 10 – 20 – 30. Per identificare il risultato migliore, si è fatto riferimento alla misura dell'F1-Score come nelle precedenti analisi. I risultati ottenuti sono stati riportati nella tabella 15.

Parametri		Classi	Non bilanciato F-1 Score	Up - Sampling F-1 Score	Down - Sampling F-1 Score
n_neighbors	5	No	0.95	0.82	0.68
		Yes	0.06	0.19	0.21
n_neighbors	10	No	0.95	0.80	0.75
		Yes	0.01	0.20	0.22
n_neighbors	20	No	0.95	0.78	0.74
		Yes	0.00	0.21	0.22
n_neighbors	**30**	**No**	**0.95**	**0.77**	**0.73**
		Yes	**0.00**	**0.22**	**0.23**

Tabella 15: Risultati ottenuti sulla base dell'F1-Score per il KNN

Come si può notare, a differenza dei casi precedenti, non si riesce a individuare un metodo di split migliore degli altri: l'F1-Score che massimizza lo Yes nel KNN risulta essere il peggiore (seppur sempre molto positivo), fra tutti gli algoritmi; nonostante ciò si riscontra una grande somiglianza nei valori.
Al fine di ridurre tale differenza il più possibile, si è scelta la prova, che risulta con un F1-Score pari a 0.73 e 0.23, ottenuta ponendo il parametro n_neighbors pari a 30 e utilizzando il dataset bilanciato attraverso il Down-Sampling.

Validazione dei risultati

I risultati delle diverse misure che si sono ottenute dalla prova migliore sono riportati nella tabella 16 insieme alla figura 34 che rappresenta la relativa ROC – Curve.

Accuracy Train	Accuracy Test	ROC - Curve	Classe	Precision	Recall
0.649	0.601	0.61	No	0.94	0.60
			Yes	0.14	0.61

Tabella 16: Misure ottenute dal risultato migliore ottenuto nel KNN

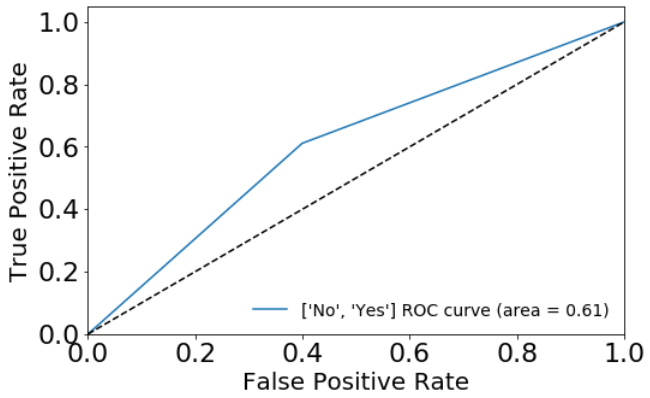

Figura 34: ROC Curve relativa ottenuta dal risultato migliore nel KNN

Dalla tabella si nota che attraverso tale metodo non si riescono a ottenere buoni risultati nonostante siano comunque in linea con quelli precedenti. In ogni caso una maggiore analisi di confronto verrà fatta nella Sezione 4.2.5.

4.2.4 Interpretazione dell'albero di decisione

Come stabilito nella sezione 4.2.1 i due migliori modelli sono quelli che attuano un Pre-Pruning con *min_leaf* e *min_split* pari rispettivamente a 2000 e 5000, applicando sia l'indice di Gini che l'Entropy. Parte dell'albero che si ottiene è mostrato nella figura 35.

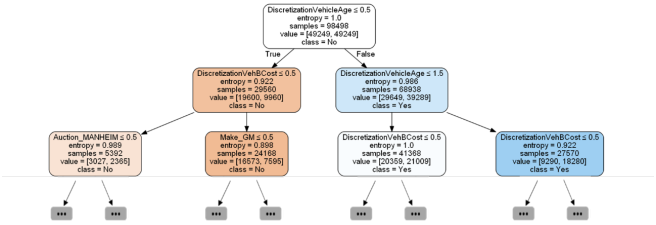

Figura 35: Albero di decisione trovato

Come si nota dalla figura 35 e dalla figura 36 che rappresenta l'importanza degli attributi nella costruzione dell'albero, i principali criteri di split si basano sulle caratteristiche riguardanti l'anno del veicolo, il costo di acquisizione pagato per il veicolo al momento dell'acquisto (VehBCost) e il modello della marca del veicolo (in particolare GM).

Inoltre, si può osservare che il primo split riguarda la suddivisione fra veicoli nuovi e medi-vecchi, che poi sono a loro volta divisi al secondo livello nel ramo destro del nodo radice. A questo punto in tutte e tre le ramificazioni si splitta sull'attributo VehBCost, suddividendo i veicoli con un basso valore dagli altri. L'albero continua per molti altri livelli, che per motivi di spazio non sono mostrati in figura.

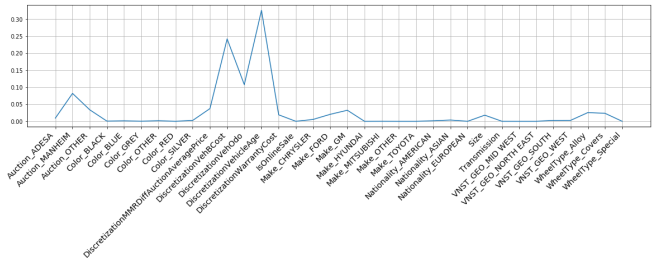

Figura 36: Distribuzione delle classi sulla base dell'importanza

4.2.5 Miglior modello di previsione

Algoritmo	Accuracy Train	Accuracy Test	ROC - Curve	Classe	Precision	Recall	F1 - Score	Dataset
Decision Tree	0.6242	0.6684	0.63	No	0.94	0.68	0.79	Up - Sampling
				Yes	0.16	0.58	0.25	
Random Forest	0.682	0.655	0.64	No	0.95	0.66	0.78	Up - Sampling
				Yes	0.13	0.61	0.25	
KNN	0.649	0.601	0.61	No	0.94	0.60	0.73	Down - Sampling
				Yes	0.14	0.61	0.23	

Tabella 17: Confronto tra i tre diversi modelli

La tabella 17 riporta tutte le misure principali dei tre migliori modelli dei diversi algoritmi di decisione: Decision Tree, Random Forest e KNN. Dalla tabella si vede chiaramente come il KNN risulti il modello relativamente peggiore rispetto agli altri. Il Decision Tree e il Random Forest risultano essere molto simili in termini di risultati ottenuti, ma nel caso in cui si scegliesse solo in relazione

all'F1-Score si otterrebbe come migliore classificazione quella relativa al Decision Tree. Tuttavia è opportuno confrontare anche le altre misure:

- ***ROC – Curve:*** risulta di poco migliore nel Random Forest;
- ***Accuracy Train/Test:*** nel Decision Tree risulta accadere un fenomeno particolare, infatti si riscontra una maggiore accuracy nel test anziché nel Training cosa che in genere non accade;
- ***Recall:*** risulta migliore nel Random Forest;
- ***Precision:*** risulta migliore rispetto ai No nel Random Forest, mentre per gli Yes nel Decision Tree.

Sulla base delle considerazioni precedenti è stato scelto come miglior modello di classificazione il Random Forest che riesce a predire meglio la classe Yes, valore critico in base all'obiettivo che ci si era preposti. Come si vede anche nelle Confusion Matrix sotto riportate si vede come effettivamente i TruePositive siano maggiori nel caso della Random Forest rispetto al Decision Tree.

		Valori Predetti	
		No	Yes
Valori Reali	No	8136	4210
	Yes	515	792

Figura 37: Confusion Matrix Random Forest

		Valori Predetti	
		No	Yes
Valori Reali	No	8370	3976
	Yes	552	755

Figura 38: Confusion Matrix Decision Tree