VALUTAZIONI SUL CONSUMO DI CARBURANTE in Canada

Federico Sgambelluri

01.
SCELTA DEL DATASET

Informazioni sul dataset

Il set di dati selezionato è stato prodotto a partire da dati messi a disposizione dal Governo Canadese per valutare il **consumo di carburante** in base al modello dell'auto e le sue specifiche rispettivamente alle **emissioni di CO_2** stimate per i nuovi veicoli in vendita sul mercato canadese.

Grazie a questo dataset si possono analizzare:

1. **Emissioni di CO_2:** Valutare l'impatto ambientale dei diversi veicoli in termini di emissioni di CO_2.
2. **Confronto dei Veicoli:** Confrontare l'efficienza dei consumi di carburante tra diversi modelli e marche di veicoli.
3. **Ricerca sul Consumo di Carburante:** Fornire dati per la ricerca accademica e per lo sviluppo di modelli predittivi riguardanti il consumo di carburante e le emissioni.
4. **Identificazione dei Veicoli più Efficienti:** Determinare quali marche e modelli di veicoli sono più efficienti dal punto di vista del consumo di carburante.
5. **Trend delle Emissioni di CO_2:** Osservare i trend delle emissioni di CO_2 e come variano tra diversi tipi di veicoli e carburanti.

Colonne dell dataset

01.

Make

Casa automobilistica produttrice

02.

Model

Modello del veicolo

03.

Veichle Class

Segmento del veicolo

04.

Engine Size

La cilindrata in litri del motore

05.

Cylinders

Il numero di cilindri del motore

06.

Fuel Type

Tipo di carburante utilizzato

Colonne dell dataset

07.

Fuel Consuption

Consumi L/100km dichiarati dal Maker in città

08.

Hwy

Consumi L/100km in autostrada

09.

Comb

Consumi L/100km (55% città e 45% hwy)

10.

CO2 Emissions

Emissioni di CO2 g/Km allo scarico per la guida combinata

11.

CO2 Rating

Emission valutate su una scala da 1 a 10

12.

Smog Rating

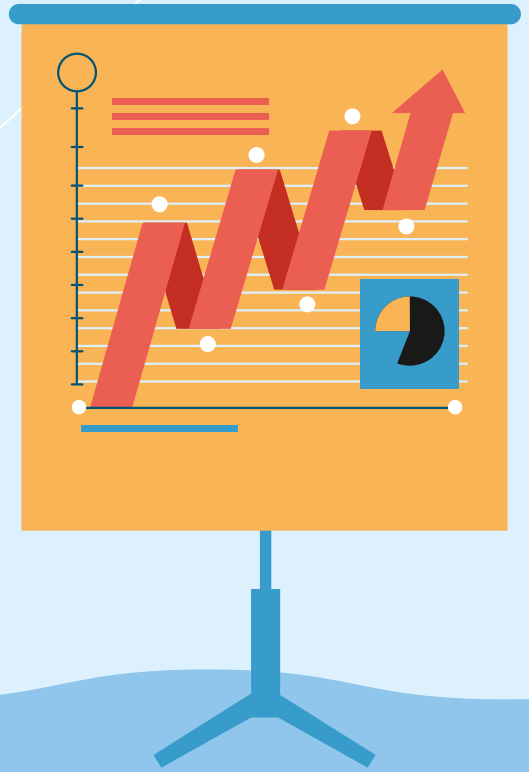
Emissioni di inquinanti che formano lo smog in una scala da 1 a 10

Altre info

Legenda tipo di carburante:

- **X** → Benzina normale
- **Z** → Benzina premium
- **D** → Diesel
- **E** → Etanolo
- **N** → Gas naturale

Link per scaricare il dataset:
https://www.kaggle.com/datasets/imtkaggleteam/fuel-concumption-ratings-2023/data



02.

PRE - PROCESSING

Rimozioni e Pulizia

NaN

```
# Rimuoviamo eventuali NaN
data = data.dropna()
print(data.info())
```

Valori

```
print(data.describe())
```

Il dataset è pulito, non contiene n'è NaN n'è valori fuori posto poichè il numero delle entry non cambia dopo la pulizia, rimanendo a 833

Rimozioni e Pulizia

Colonne

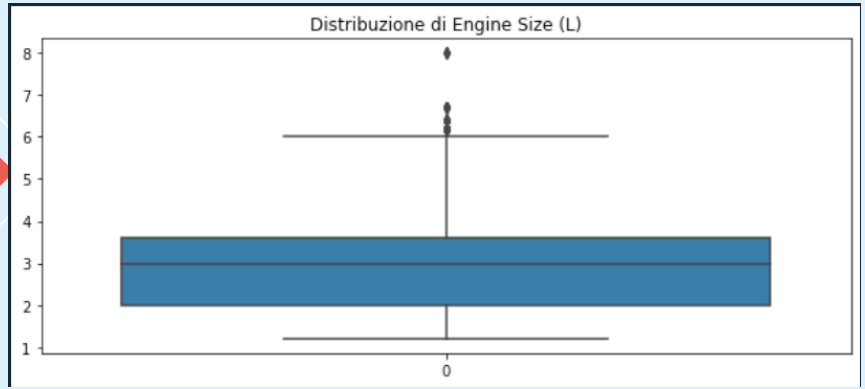
Sono state rimosse le colonne:

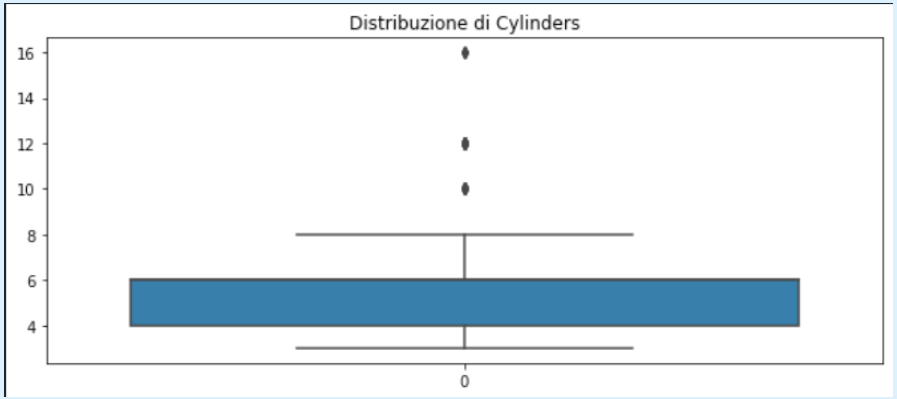
- **Year** → Tutte le auto analizzate sono state prodotte nel 2023
- **Transmission** → Influisce poco sul consumo effettivo del veicolo
- **Comb (mpg)** → Vogliamo effettuare l'analisi con il sistema metrico e non in galloni

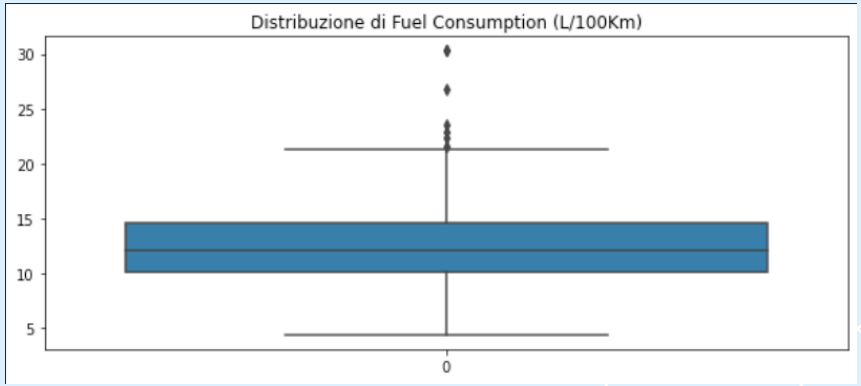
```
#Rimuoviamo le colonne inutili
data=data.drop(columns=["Transmission","Comb(mpg)","Year"])
print(data.info())
```

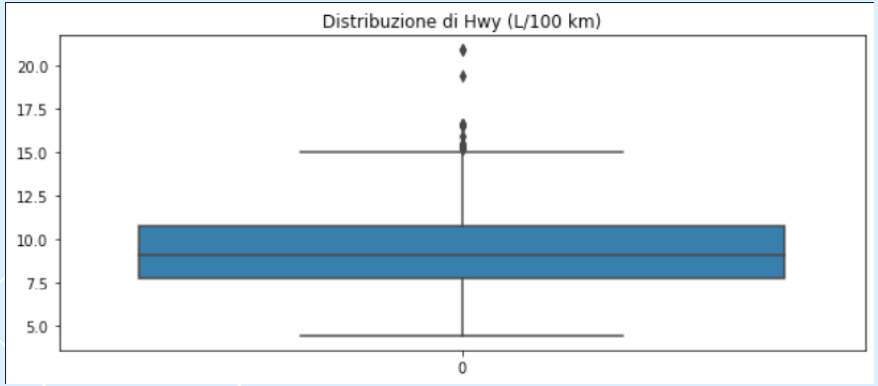
BOXPLOT

Visualizzo meglio la distribuzione delle variaibli **numeriche** per identificare i valori fuori soglia

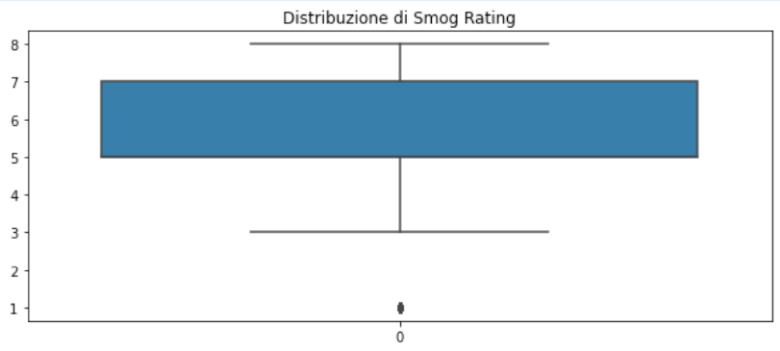


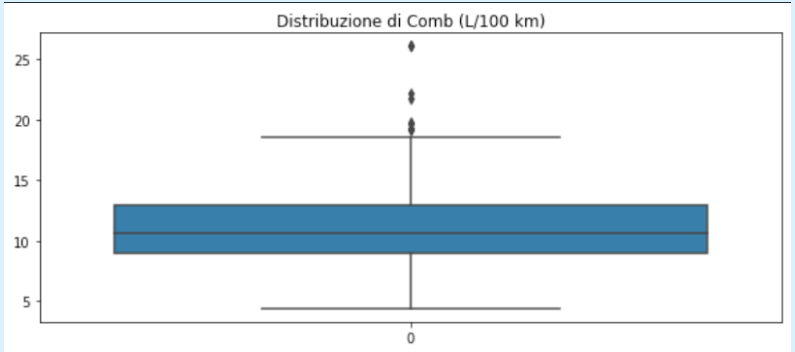


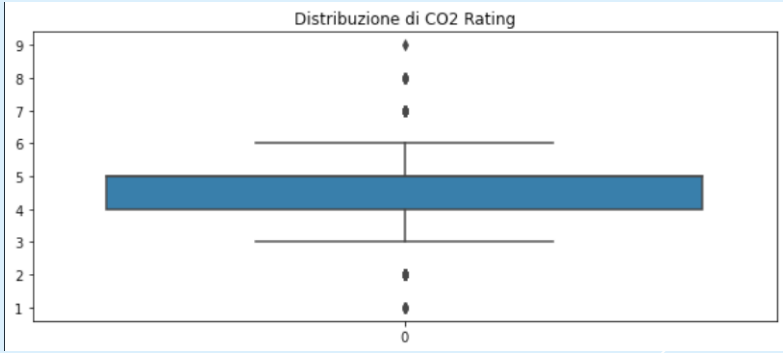


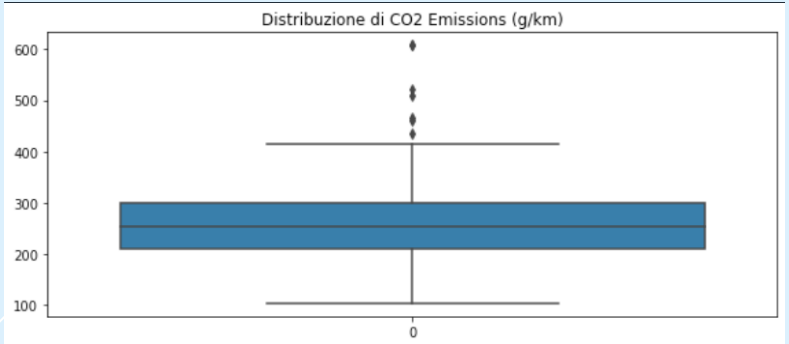


BOXPLOT


Distribuzione di Smog Rating
8
7
6
5
4
3
2
1
0


Distribuzione di Comb (L/100 km)
25
20
15
10
5
0


Distribuzione di CO2 Rating
9
8
7
6
5
4
3
2
1
0


Distribuzione di CO2 Emissions (g/km)
600
500
400
300
200
100
0

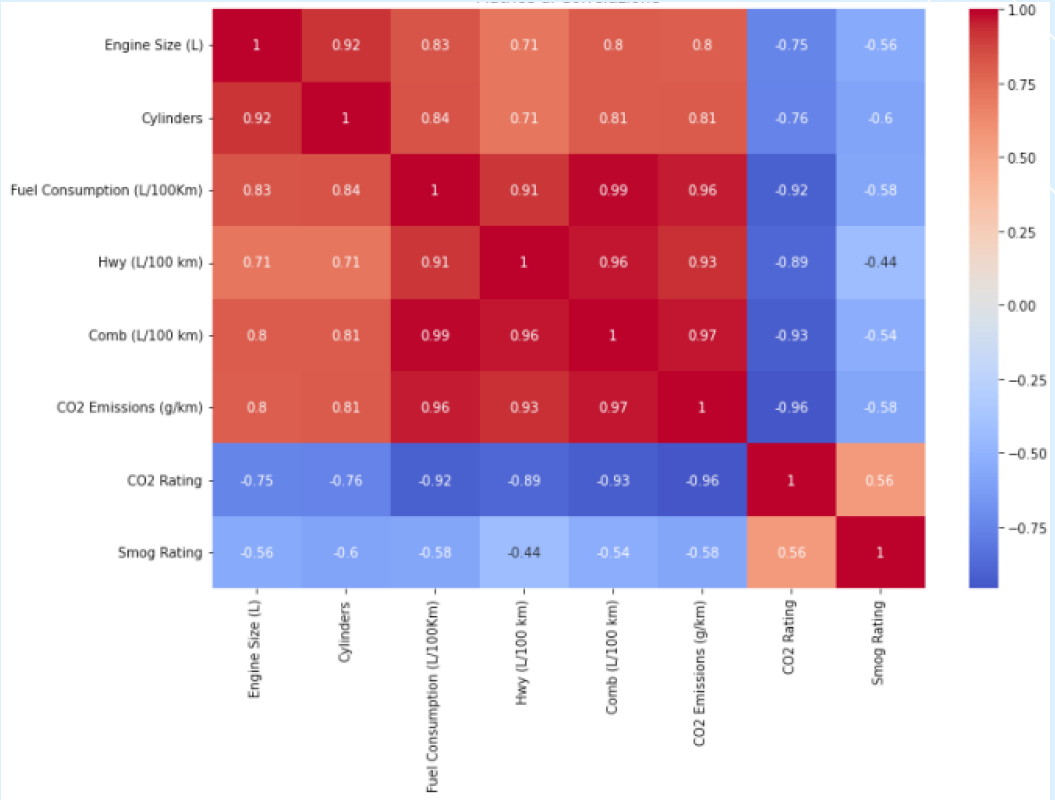
03. EXPLORATORY DATA ANALYSIS (EDA)

Matrice di correlazione

Ci permette di misurare la relazione lineare tra due o più variabili numeriche, tramite il coefficiente di correlazione di Pearson.

Valori del coefficiente di correlazione:
+1: Correlazione positiva perfetta. Quando una variabile aumenta, anche l'altra aumenta in modo lineare.
0: Nessuna correlazione lineare. Le variabili non mostrano alcuna relazione lineare.
-1: Correlazione negativa perfetta. Quando una variabile aumenta, l'altra diminuisce in modo lineare.

0.7 - 1.0 (o -0.7 a -1.0): Correlazione forte
.**0.3 - 0.7** (o -0.3 a -0.7): Correlazione moderata.
0.0 - 0.3 (o 0.0 a -0.3): Correlazione debole.



Matrice di correlazione

Correlazioni Positive Forti

- **Fuel Consumption (L/100km) e Comb (L/100 km)**: Queste due variabili hanno una correlazione molto alta (0.99), indicando che quando il consumo di carburante aumenta in un contesto, aumenta anche nell'altro. Questo è atteso, dato che entrambe misurano il consumo di carburante in diverse condizioni di guida.
- **Fuel Consumption (L/100km) e CO2 Emissions (g/km)**: Anche qui, c'è una correlazione molto alta (0.96), suggerendo che un maggior consumo di carburante è strettamente legato a maggiori emissioni di CO2.
- **Engine Size (L) e Cylinders**: La correlazione di 0.92 indica che i veicoli con motori più grandi tendono ad avere più cilindri.

Attenzione:
Analizzerò la seconda correlazione più alta, perché la prima non è molto significativa in quanto il consumo misto è frutto della media tra consumo in città e consumo in autostrada, quindi prenderò per le analisi futura la seconda correlazione forte, ovvero **Fuel Consumption (L/100km) e CO2 Emissions (g/km)**:

Matrice di correlazione

Correlazioni Negative Forti

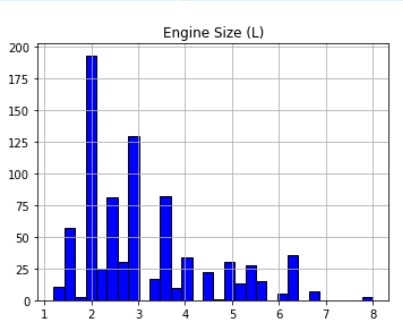
- **CO2 Rating e CO2 Emissions (g/km)**: Anche qui, la correlazione è molto negativa (-0.96), il che significa che veicoli con emissioni di CO2 più alte hanno un CO2 Rating peggiore
- **CO2 Rating e Fuel Consumption (L/100km)**: La correlazione negativa (-0.92) mostra che veicoli con un rating migliore (valore più basso per CO2 Rating) tendono a consumare meno carburante.

Correlazioni Moderate

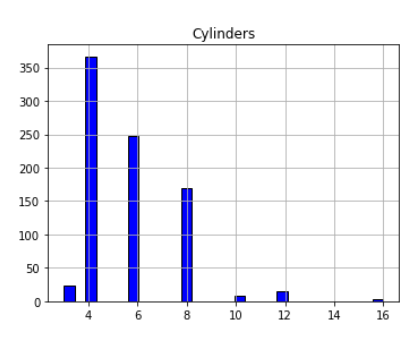
- **Engine Size (L) e Fuel Consumption (L/100km)**: La correlazione di 0.83 indica una relazione moderatamente forte, suggerendo che motori più grandi tendono a consumare più carburante.
- **Engine Size (L) e CO2 Emissions (g/km)**: La correlazione di 0.80 indica che i motori più grandi tendono a emettere più CO2.
- **Smog Rating e CO2 Rating**: La correlazione di 0.56, seppur non fortissima, indica che un miglior rating delle emissioni di CO2 è moderatamente associato a un miglior rating delle emissioni di smog.

Analisi Univariate

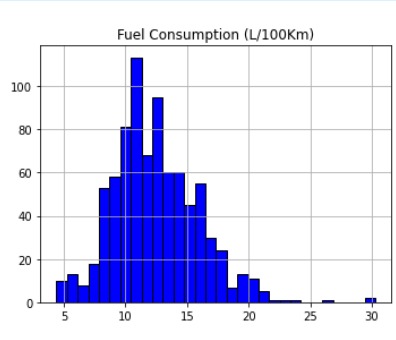
Analisi statistica di una **singola variabile** alla volta con l'obiettivo di descrivere e comprendere: la distribuzione, la centralità e la dispersione di ciascuna variabile



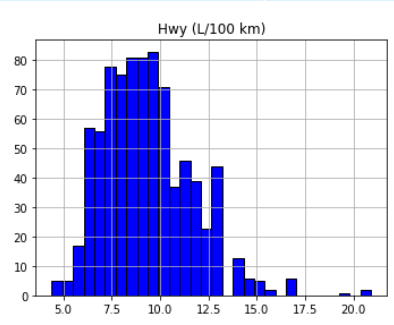
La maggior parte dei veicoli ha motori di cilindrata compresa tra 2 e 3 litri



La maggior parte dei veicoli ha motori con 4, 6 o 8 cilindri



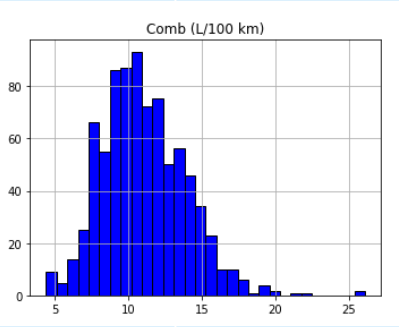
La maggior parte dei veicoli consuma in città tra i 5 e i 15 litri per 100km con un picco di 10-12 litri



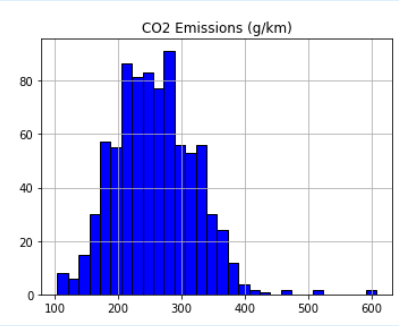
La maggior parte dei veicoli consuma in autostrada tra i 5 e i 12 litri per 100km

Analisi Univariate

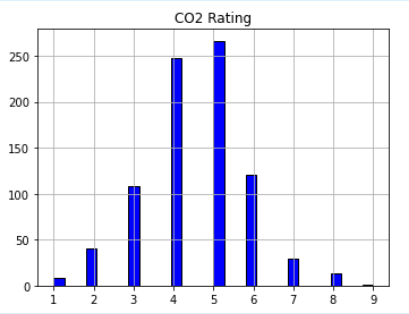
Analisi statistica di una **singola variabile** alla volta con l'obiettivo di descrivere e comprendere: la distribuzione, la centralità e la dispersione di ciascuna variabile



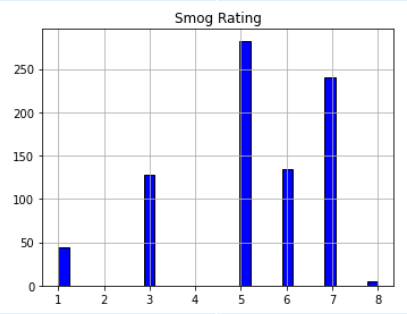
La distribuzione è simile al consumo di carburante in città, con un picco tra i 10 e 15 litri



La maggior parte dei veicoli emette tra 100 e 300 g/km di CO_2



La maggior parte dei veicoli ha un rating tra 4 e 6 con un picco a 5

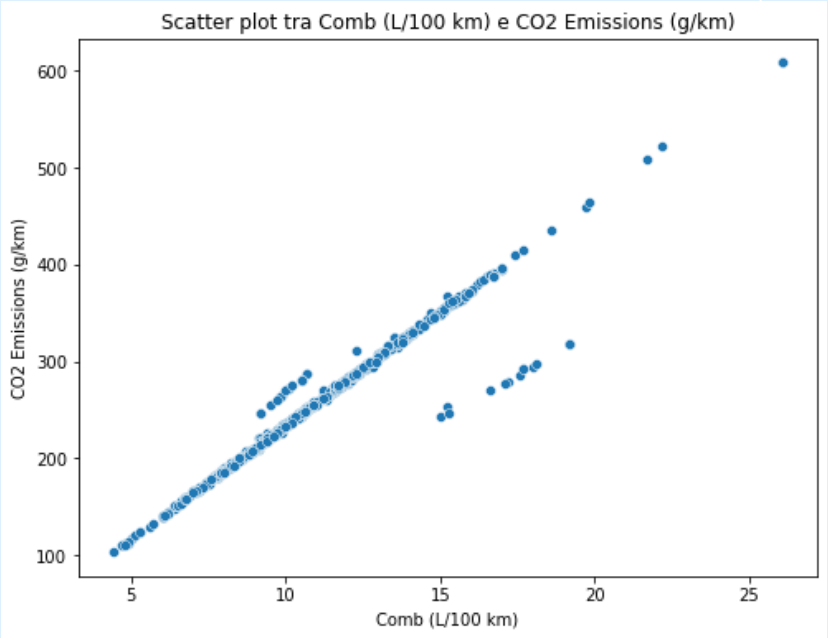


La maggior parte dei veicoli ha un rating compresto tra 3 e 7 con picco a 5

Analisi Bivariate

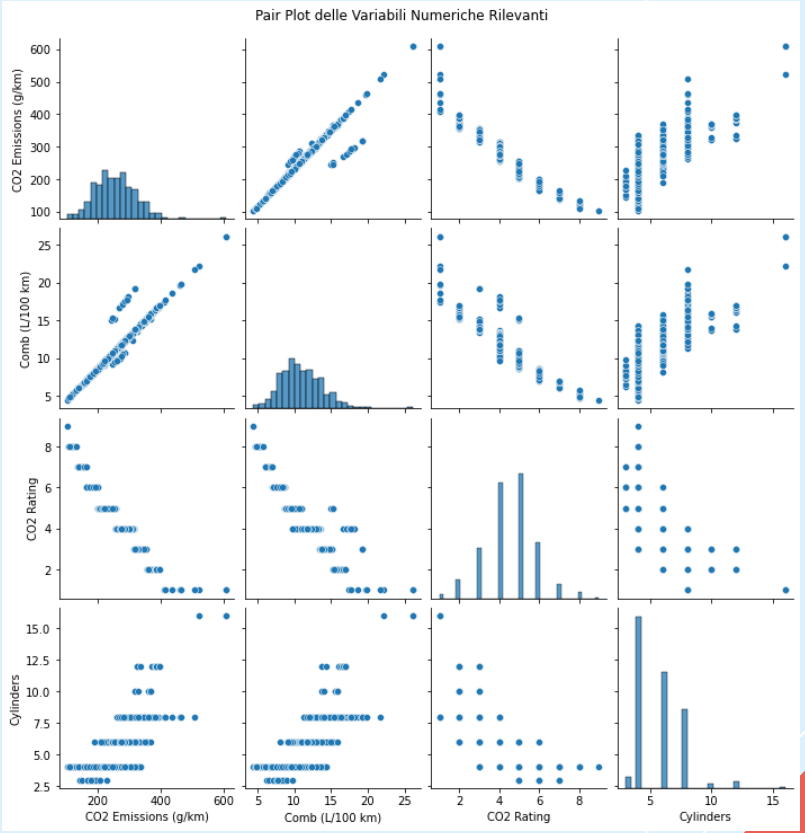
Analisi statistica che si concentra sull'**analisi di due variabili** contemporaneamente per esplorare la relazione tra di esse con l'obiettivo di determinare una relazione tra le due variabili

Il grafico mostra una chiara **relazione lineare positiva** tra **consumo di carburante ed emissioni**, quando aumenta il cosnumo aumentano le emissioni e viceversa. Gli **outliers** ovvero i punti fuori dal comune mostrano come consumo di carburante superiore a 25L/100km superi 500g/km



Analisi Multivariata

L'analisi **multivariata** è una tecnica statistica che esamina **più variabili contemporaneamente** per esplorare le relazioni tra di esse. In questo contesto, l'obiettivo principale è determinare come diverse variabili influenzano le emissioni di CO_2

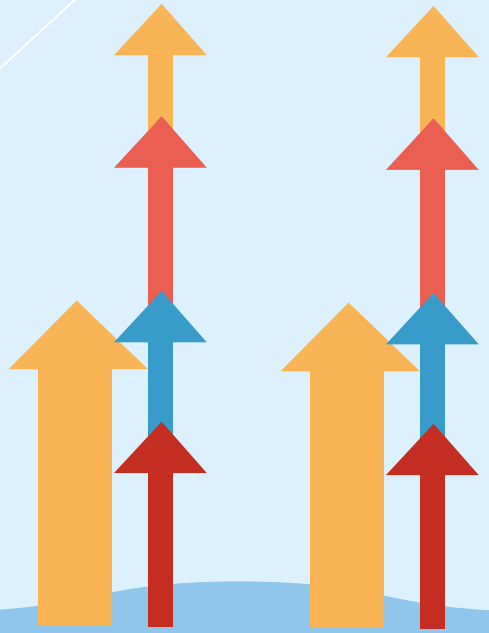


Relazione tra CO_2 Emissions (g/km) e Comb (L/100 km):
All'aumentare del consumo di carburante, aumentano anche le emissioni di CO_2..

Relazione tra CO_2 Emissions (g/km) e CO_2 Rating:
Esiste una relazione lineare negativa: un rating di CO_2 migliore (più basso) è associato a emissioni di CO_2 più basse, confermando l'efficacia del rating come indicatore delle performance ambientali dei veicoli.

Relazione tra CO_2 Emissions (g/km) e Cylinders:
Veicoli con un numero maggiore di cilindri tendono ad avere emissioni di CO_2 più alte.

Outliers:
Alcuni veicoli con un consumo di carburante combinato superiore a 25L/100km possono superare emissioni di 500g/km

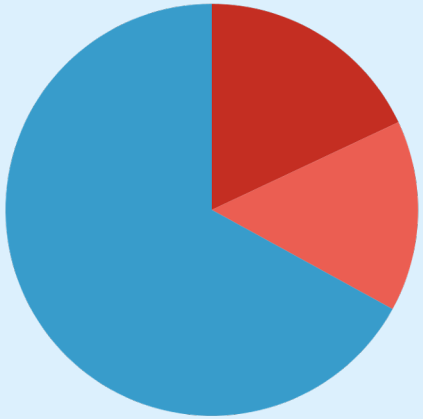


04.

Splitting

Divisione dataset

Dividiamo il dataset in **3 parti distinte** per creare, addestrare e valutare i modelli di machine learning



70 %

TRAINING SET

583 campioni

15 %

VALIDATION SET

125 campioni

15 %

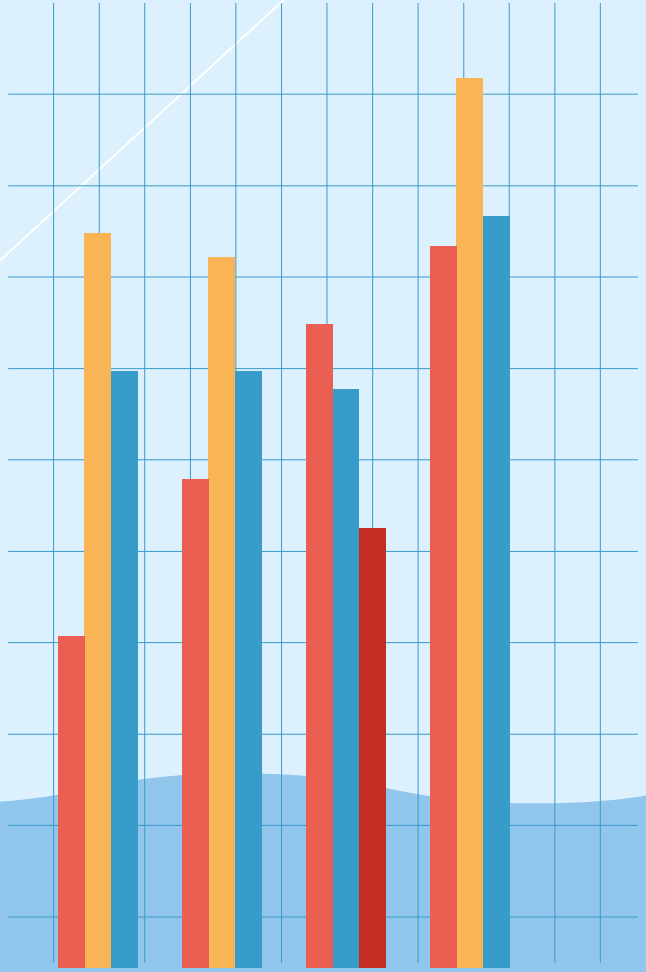
TEST SET

125 campioni

Training Set: Utilizzato per addestrare il modello. Il mdoello apprende le caratteristiche dei dati e costruisce una relazione tra le variabili di input (features) e quella di output (target)

Validation set: Utilizzato per valutare le performance del modello durante la fase di addestramento

Test Set: Utilizzato per la valutazione finale del modello, fornisce una stima imparziale delle performance del modello su dati completamente novi



05.

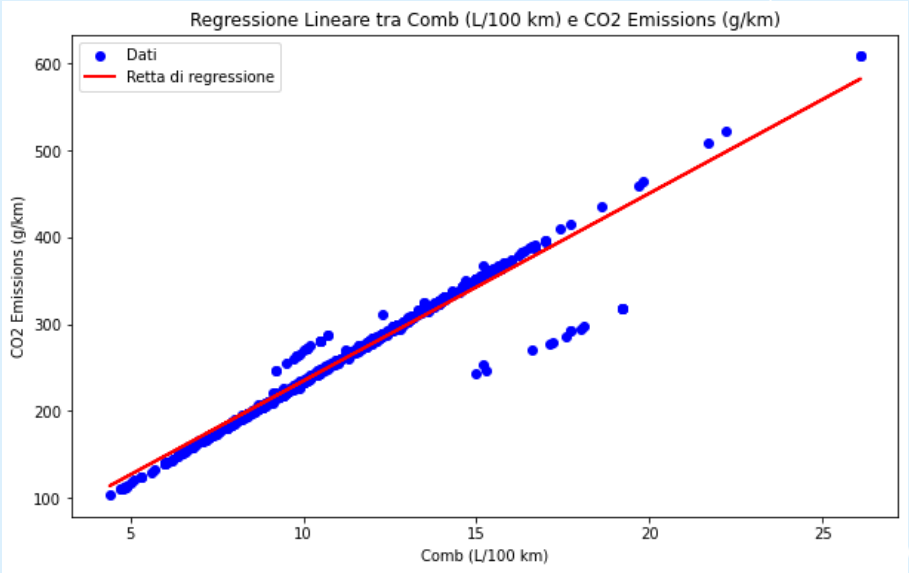
Regressione

Lineare

Regressione Lineare - positiva

Utilizziamo la **regressione lineare semplice** per modellare la relazione tra una variabile dipendente e una o più variabili indipendenti.
Siccome la consegna richiede **due coppie** di variabili, ne prendo una con altra correlazione **positiva** e una con alta correlazione **negativa**, per costruire, addestrare e valutare il modello di regressione lineare.

Coppia di variabili con alta correlazione **positiva**: **Comb (L/100km) e CO2 Emission (g/km)**



Regressione Lineare - positiva

Relazione Lineare:

Il grafico mostra una relazione lineare tra il consumo di carburante combinato e le emissioni di CO_2. Quando il consumo di carburante aumenta, aumentano anche le emissioni di CO_2. Questo è rappresentato dalla retta di regressione in rosso.

Retta di Regressione:

La pendenza della retta indica che per ogni incremento di 1 L/100 km nel consumo di carburante, c'è un aumento corrispondente nelle emissioni di CO_2.
La retta di regressione si adatta bene ai dati, il che suggerisce che il modello lineare è appropriato per descrivere la relazione tra queste variabili.

Distribuzione dei Punti:

I punti dati (in blu) sono distribuiti abbastanza vicini alla retta di regressione, indicando che la variabilità delle emissioni di CO_2 è ben spiegata dal consumo di carburante combinato.
I punti che si discostano dalla retta di regressione, indicano la presenza di outliers o variabilità non spiegata dal modello lineare semplice.

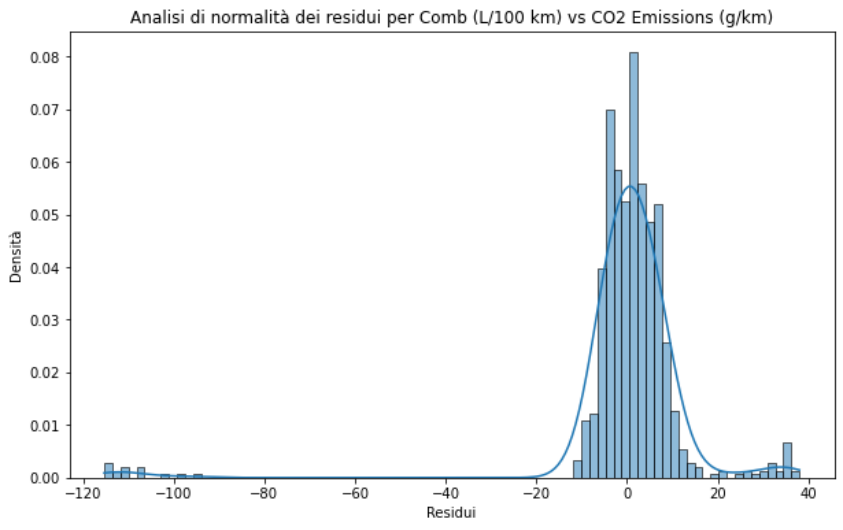
Outliers:

Possono rappresentare veicoli con caratteristiche uniche o potrebbero essere errori nei dati.
Gli outliers possono influenzare la stima della retta di regressione e dovrebbero essere analizzati separatamente per comprendere le loro cause.

Analisi di normalità

Utilizziamo la **regressione lineare** per modellare la relazione tra una variabile dipendente e una o più variabili indipendenti.
Nel nostro caso prendiamo una coppia di **variabili con altra correlazione positiva** (Comb – CO2), per costruire, addestrare e valutare il modello di regressione lineare.

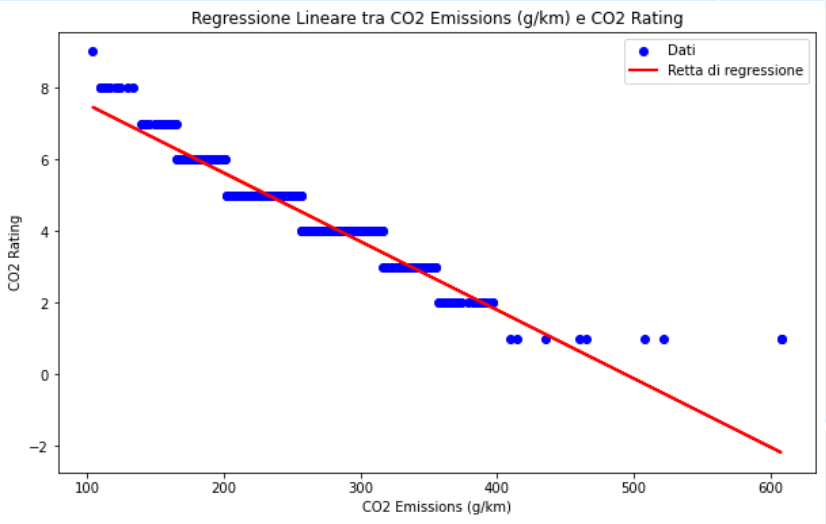
L'istogramma rappresenta **l'analisi di normalità dei residui** del modello di regressione lineare sopradescritto.
Notiamo che ci sono degli **outliers significativi**, che possono indicare che in alcuni casi il modello non reagisce bene, la maggior parte dei residui stanno nella norma, quindi **il modello è in grado di spiegare buona parte della variabilità dei dati**



Regressione Lineare - negativa

Utilizziamo la **regressione lineare semplice** per modellare la relazione tra una variabile dipendente e una o più variabili indipendenti.
Siccome la consegna richiede **due coppie** di variabili, ne prendo una con altra correlazione **positiva** e una con alta correlazione **negativa**, per costruire, addestrare e valutare il modello di regressione lineare.

Coppia di variabili con alta correlazione **negativa**: **CO2 Emission (g/km) e CO2 Rating**



Regressione Lineare - negativa

Relazione Lineare: Il grafico mostra una relazione lineare negativa tra le emissioni di CO_2 (g/km) e il rating di CO_2. Quando le emissioni di CO_2 aumentano, il rating di CO_2 diminuisce. Questo è rappresentato dalla retta di regressione in rosso.

Retta di Regressione: La pendenza della retta indica che per ogni aumento nelle emissioni di CO_2, c'è una diminuzione corrispondente nel rating di CO_2. La retta di regressione si adatta bene ai dati, il che suggerisce che il modello lineare è appropriato per descrivere la relazione tra queste variabili.

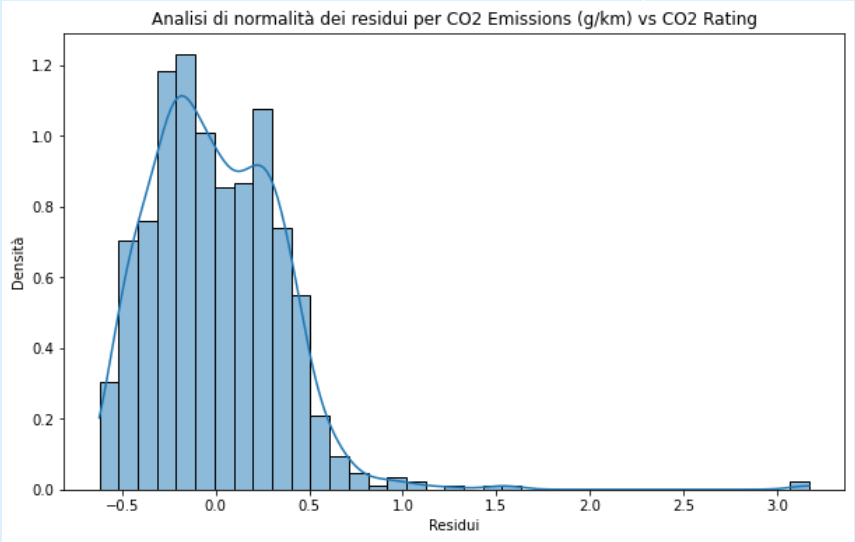
Distribuzione dei Punti: I punti dati (in blu) sono distribuiti abbastanza vicini alla retta di regressione, indicando che la variabilità del rating di CO_2 è ben spiegata dalle emissioni di CO_2. I punti che si discostano dalla retta di regressione indicano la presenza di outliers o variabilità non spiegata dal modello lineare semplice.

Outliers: Gli outliers possono rappresentare veicoli con caratteristiche uniche o potrebbero essere errori nei dati. Gli outliers possono influenzare la stima della retta di regressione e dovrebbero essere analizzati separatamente per comprendere le loro cause..

Analisi di normalità

Utilizziamo la **regressione lineare** per modellare la relazione tra una variabile dipendente e una o più variabili indipendenti.
Nel nostro caso prendiamo una coppia di variabili con **altra correlazione negativa** (Comb – CO2), per costruire, addestrare e valutare il modello di regressione lineare.

L'istogramma mostra che il modello **predice accuratamente i dati** poiché la maggior parte dei residui è situata sullo zero. Ci sono anche qui **outliers**, ma non sono devianti, il **modello è accettabile**



Regressione Lineare

Coefficiente → Pendenza della retta di regressione

Intercetta → Intersezione con asse Y

R^2 → Compreso tra 0 e 1, più è vicino ad 1 più suggerisce un buon modello

MSE → Mean Squared Error, misura la media dei quadrati degli errori, più è basso, più il modello è accurato

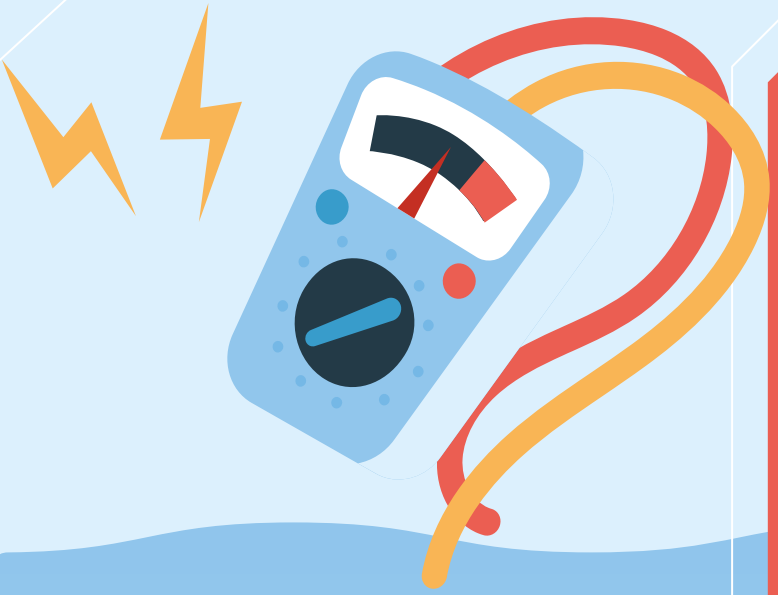
Regressione Lineare

Correlazione Positiva

Coefficiente	21.57
Intercetta	19.14
R^2	0.93
MSE	273.3

Correlazione Negativa

Coefficiente	-0.02
Intercetta	9.44
R^2	0.92
MSE	0.131



06.

Addestramento del Modello

Modelli di classificazione

Regressione Logistica

E' uno dei più semplici classificatori lineari, che **permette** tramite l'addestramento di dati, **di predire** la **variabile targhet** tramite la funzione:

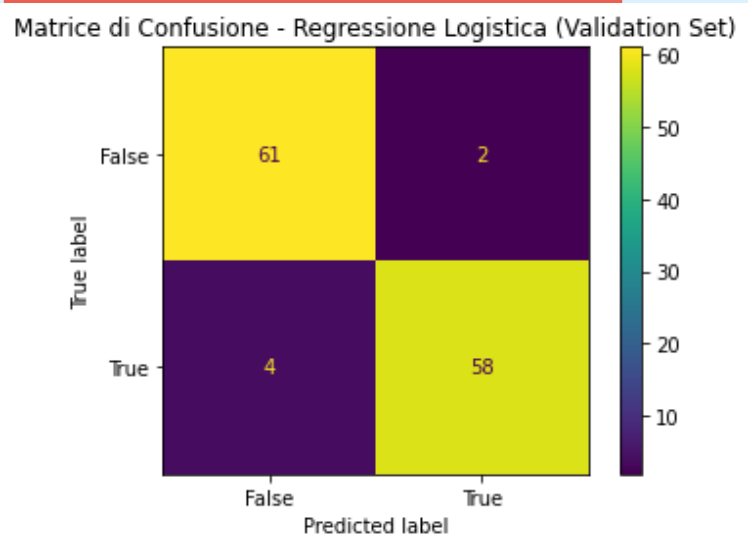
$$h_\theta(x_1, x_2) = \frac{1}{1 + e^{-(ax_1 + bx_2 + c)}}.$$

SVM

Support Vectrom Machines, è un altro classificatore che utilizza la **curva separatrice** per migliorare l'algoritmo di separazione. La forma di tale funzione detta **kernel function** rende linearmente separabili le caratteristiche del dataset

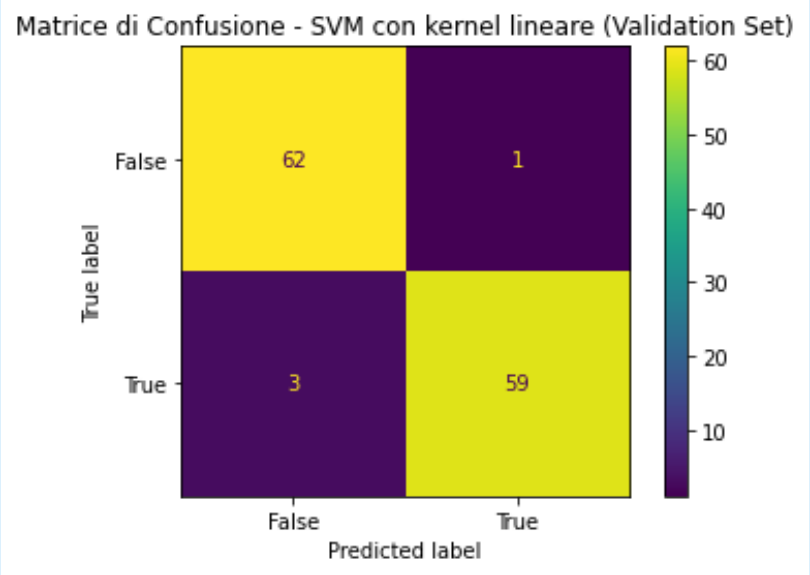
Matrici di confusione

Regressione Logistica



Accuracy: 95%

SVM



Accuracy: 97%

Matrici di confusione

Analisi

- **True Negatives (TN)**: 63
Campioni che erano effettivamente negativi (classe False) e che il modello ha correttamente predetto come negativi.
- **False Positives (FP)**: 0
Campioni che erano effettivamente negativi (classe False) ma che il modello ha predetto erroneamente come positivi (classe True).
- **False Negatives (FN)**: 0
Campioni che erano effettivamente positivi (classe True) ma che il modello ha predetto erroneamente come negativi (classe False).
- **True Positives (TP)**: 62
Campioni che erano effettivamente positivi (classe True) e che il modello ha correttamente predetto come positivi.

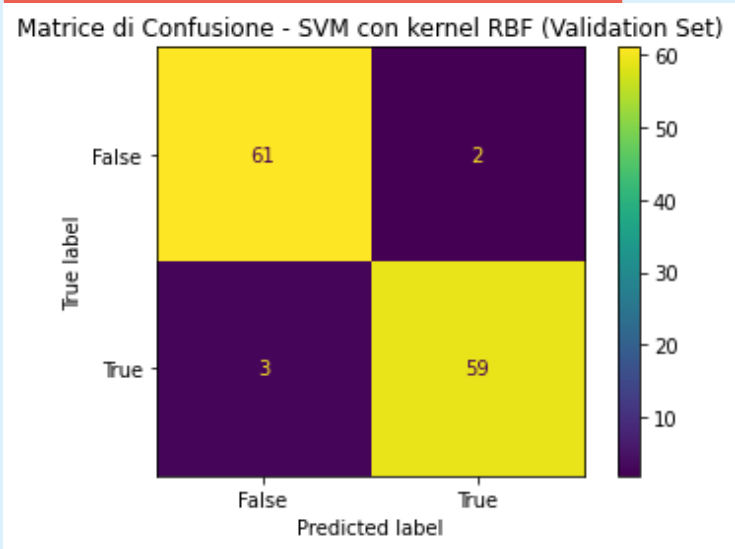
Considerazioni

Ottima Accuratezza:
Le matrici di confusione mostrano che il modello ha una quasi perfetta accuratezza sul validation set. Tutte le predizioni sono corrette, non ci sono né FP né FN.

Modello Ben Addestrato:
Un'accuratezza quasi perfetta sul validation set indica che il modello è ben addestrato e generalizza bene ai dati non visti, almeno per questo set di validazione.

Matrici di confusione

SVM – Kernel RBF



Accuracy: 97%

Alta Accuratezza:
Il modello SVM con Kernel RBF ha una accuratezza identica al modello con kernel lineare

Modello Ben Addestrato:
Quindi il modello SVM con Kernel RBF mostra comunque una buona accuratezza, indicando che è ben addestrato e generalizza bene sui dati non visti, almeno per il validation set.

Errore Limitato:
I pochi errori presenti potrebbero indicare alcuni casi limite che il modello non è riuscito a classificare correttamente. Questo potrebbe essere dovuto alla complessità del modello o alla natura dei dati.

Grafico di separazione (Decision Boundary)

Il grafico di separazione (o **decision boundary**) visualizza come un modello di classificazione divide lo spazio delle caratteristiche per distinguere le diverse classi.
Mostra la **linea** che il modello ha appreso per separare le diverse classi nel set di dati, **colorando** le **aree** in base alla classe (**reale o predetta**) nel modello per vedere chiaramente quali aree appartengono a ciascuno spazio sovrapponendo i punti dei dati.
Lo scopo è visualizzare come il modello sta classificando i dati ed è utile per **diagnosticare** se il modello sta facendo **errori**.
L'ho utilizzato come alternativa alla matrice di confusione.
Avendola applicata nell'SVM, vediamo le differenze tra i kernel:

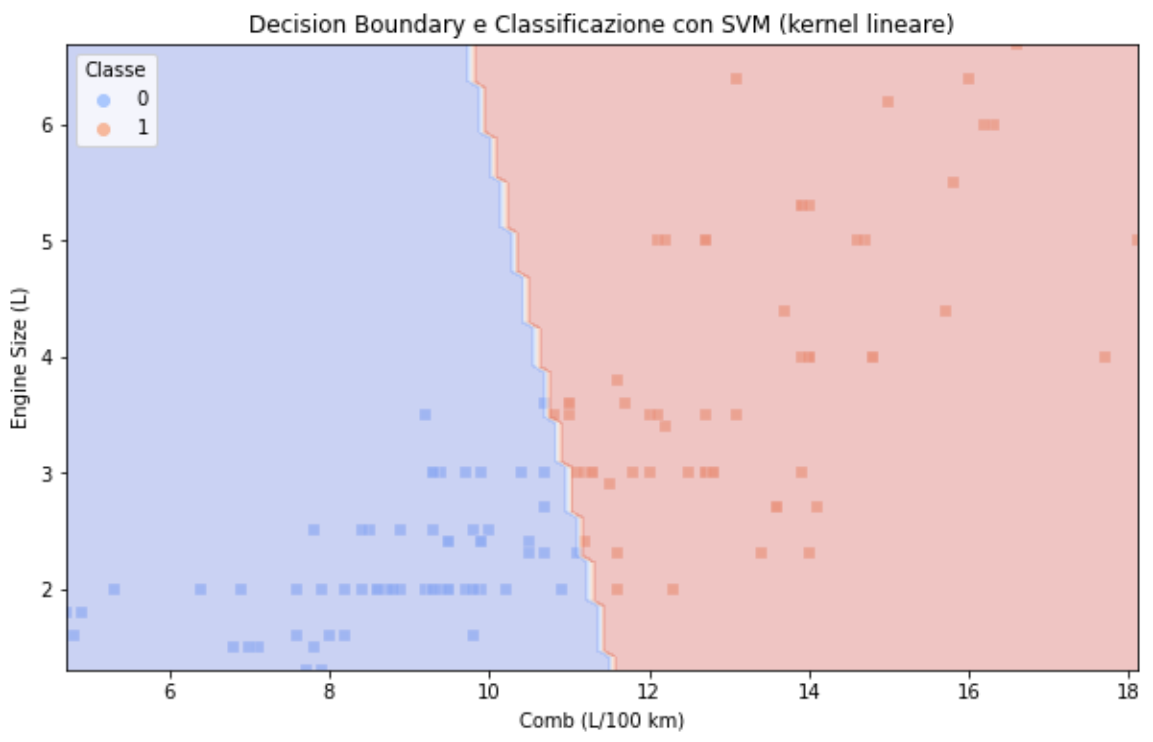
Kernel Lineare:

- Decision boundary lineare.
- Adatto per problemi in cui le classi sono chiaramente separabili da una linea retta.
- Meno flessibile per catturare relazioni complesse nei dati.

Kernel RBF:

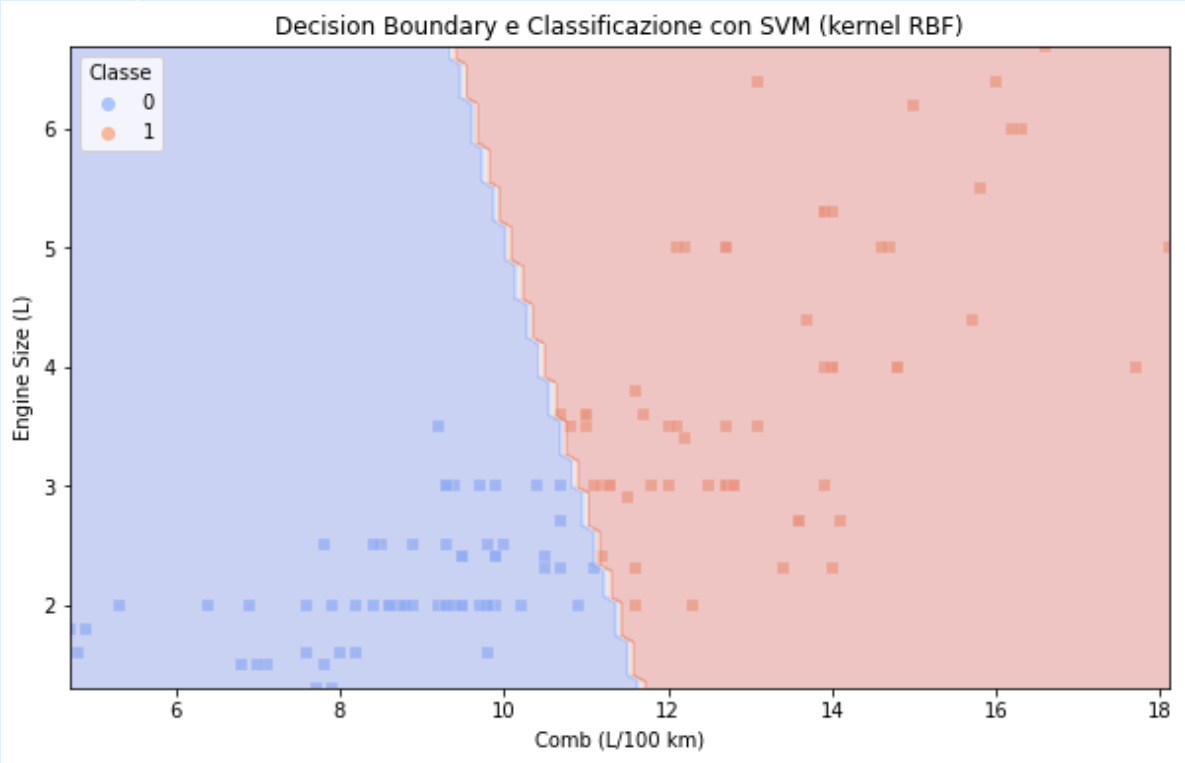
- Decision boundary non lineare.
- Adatto per problemi in cui le classi hanno confini complessi e non lineari.
- Più flessibile e in grado di adattarsi meglio a dati complessi.

Grafico di separazione – kernel lineare



Questo **grafico di separazione** fornisce una visione chiara di come il modello **SVM con kernel lineare** classifica i dati in base alle caratteristiche Comb (L/100 km) e Engine Size (L).
Notiamo un' **ottima predizione** poiché tutti i punti blu sono nell'area blu e tutti i punti rossi nell'area rossa, anche se essendo lineare non è preciso come uno non lineare

Grafico di separazione – kernel RBF



Questo **grafico di separazione** fornisce una visione chiara di come il modello **SVM con kernel RBF** classifica i dati in base alle caratteristiche Comb (L/100 km) e Engine Size (L). La decision boundary non lineare riflette la capacità del modello di gestire complessità nei dati, **migliorando la separazione** delle classi rispetto a un modello lineare. Le predizioni sono Corrette.

07.

Hyperparameter Tuning

Validation VS K-Fold

Eseguiamo il tuning degli iperparametri per il modello SVM utilizzando la tecnica Grid Search che ci permette di testare una combinazione di diversi valori per gli iperparametri e scegliere la combinazione che produce i risultati migliori, utilizziamo due approcci:

Validation Set Test

Come abbiamo visto nel punto 4 dividiamo il dataset in: *Training Set, Validation Set e Test Set*

Vantaggi: semplice e rapido

Svantaggi: instabile se validation set è piccolo, modello finale impreciso se la divisione non è ottimale

K-Fold Cross-Validation

Suddivido il dataset in K parti o folds, addestrando il modello k volte utilizzando k-1 folds per l'addestramento

Vantaggi: utilizzo tutti i campioni e aumento la stabilità, ottimo per dataset di piccole dimensioni

Svantaggi: computazionalmente costoso

(Non lo faremo per questo progetto poiché non lo abbiamo visto a lezione)

Risultati

SVM Ottimizzato (Validation Set)

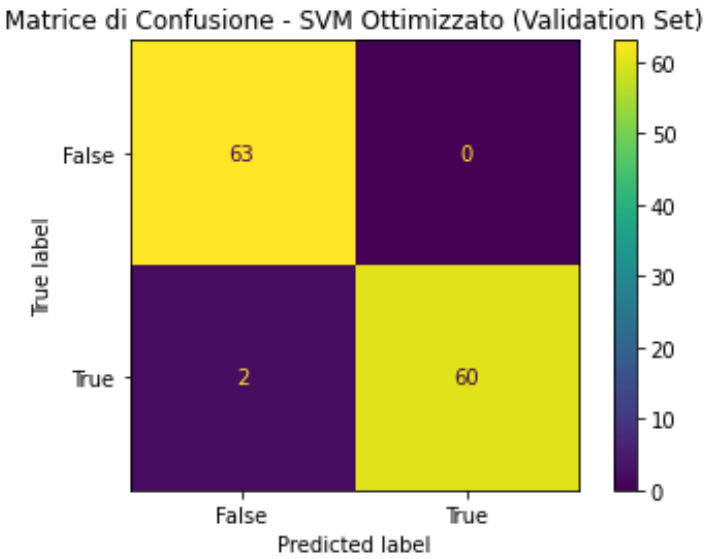
```
Fitting 5 folds for each of 32 candidates, totalling 160 fits
I migliori iperparametri sono: {'C': 10, 'gamma': 0.1, 'kernel': 'rbf'}

Valutazione del modello ottimizzato SVM (Validation Set):
Accuracy: 0.984
              precision    recall  f1-score   support

           0       0.97      1.00      0.98        63
           1       1.00      0.97      0.98        62

    accuracy                           0.98       125
   macro avg       0.98      0.98      0.98       125
weighted avg       0.98      0.98      0.98       125
```

Performance Elevata: Il modello SVM ottimizzato ha mostrato un'ottima performance con alti valori di precision, recall e F1-score per entrambe le classi.

Bilanciamento delle Classi: Il supporto è abbastanza bilanciato tra le due classi, il che aiuta il modello a mantenere buone metriche di performance.

Robustezza del Modello: L'alta accuratezza e i valori uniformemente alti di precision, recall e F1-score suggeriscono che il modello è robusto e generalizza bene sui dati di validazione.

Matrice di confusione



Accuracy: 98%

Risultati

Precision: indica la proporzione di veri positivi tra tutte le predizioni positive. Una precisione alta per entrambe le classi suggerisce che il modello fa pochi errori quando predice una classe positiva.

Recall: misura la proporzione di veri positivi che sono stati correttamente identificati dal modello. Una recall alta per entrambe le classi significa che il modello riesce a catturare la maggior parte delle istanze positive reali.

F1-Score: è la media armonica di precision e recall. Valori alti indicano che il modello bilancia bene precision e recall per entrambe le classi.

Support: indica il numero di occorrenze di ciascuna classe nel dataset di validazione.

Accuracy: rappresenta la proporzione di predizioni corrette sul totale delle predizioni. Un'accuratezza del 98.4% indica che il modello SVM ottimizzato ha performato molto bene sul validation set.

Macro Avg: calcola la media di precision, recall e F1-score trattando tutte le classi in modo uguale, indipendentemente dal numero di campioni per classe.

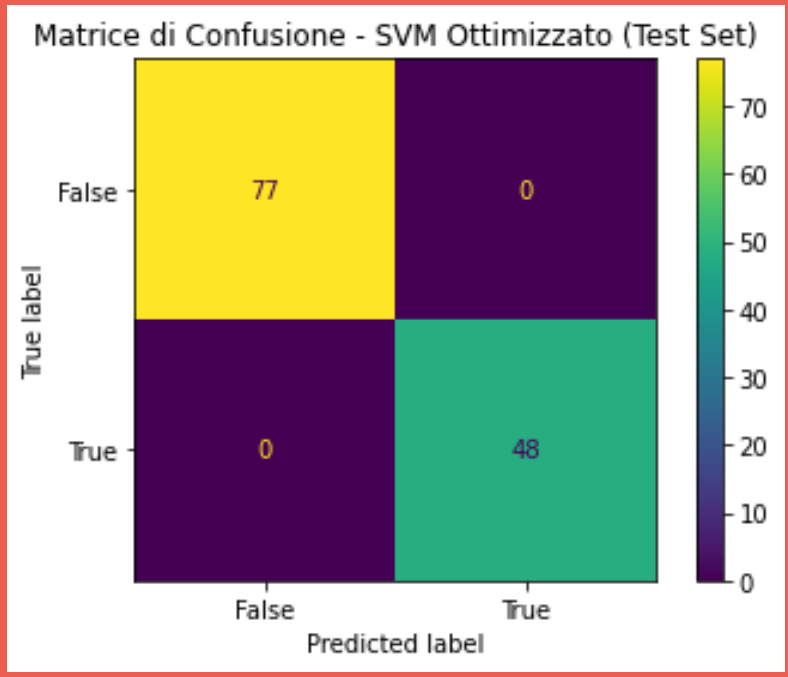
Weighted Avg: calcola la media ponderata di precision, recall e F1-score, tenendo conto del supporto di ciascuna classe.

08.

Valutazione delle Performance

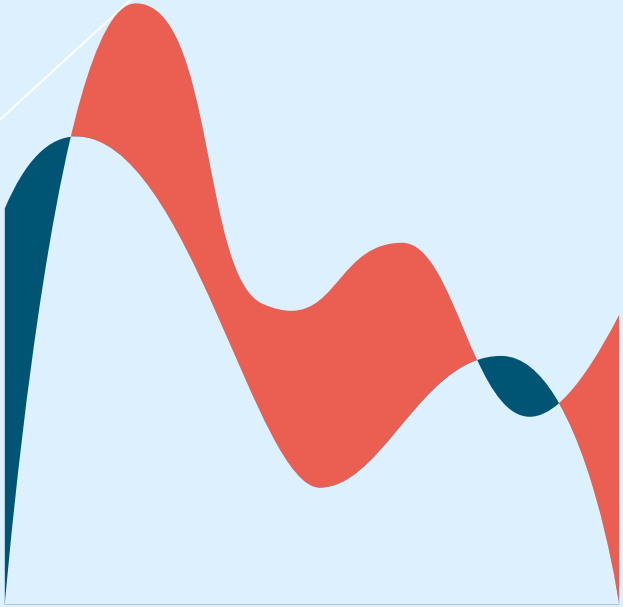
Performance Set Test

La matrice di confusione dell'SVM ottimizzato sul test set mostra che tutte le 77 istanze negative (False) e tutte le 48 istanze positive (True) sono state **classificate correttamente**, indicando un'accuratezza perfetta sul test set.




```
Valutazione finale del modello ottimizzato SVM sul test set:
Accuracy: 1.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        77
           1       1.00      1.00      1.00        48

    accuracy                           1.00       125
   macro avg       1.00      1.00      1.00       125
weighted avg       1.00      1.00      1.00       125
```

Questi risultati indicano che i modelli di regressione logistica e SVM hanno ottenuto una prestazione **eccellente** sia sul validation set che sul test set, con una classificazione perfetta di tutte le istanze.



09.
Studio statisico
dei risultati

Statistica Inferenziale

Media Accuracy **99%**

Media delle accuratezze ottenute dal modello SVM sulle diverse suddivisioni dei dati durante le 10 iterazioni di training e testing.
Un'accuratezza media del 99,64% indica una **performance eccellente**

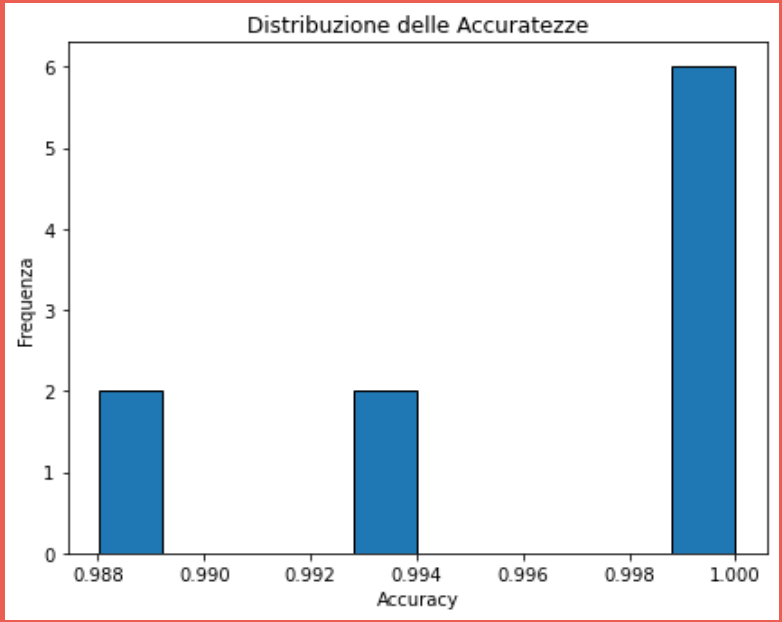
Deviazione Standard **0,4%**

La deviazione standard misura la variabilità delle accuratezze ottenute in ciascuna delle 10 iterazioni.
Una deviazione standard di 0,0048 indica che le **performance** del modello sono molto **consistenti** tra le diverse iterazioni

Intervallo Confidenza **95%**

Questo intervallo indica quanto siamo sicuri che la media dell'accuratezza trovata rappresenti la vera media di accuratezza del modello su nuove suddivisioni dei dati, nel nostro caso corrisponde a **(0.9934-0.9994)**

Statistica Descrittiva - istogramma

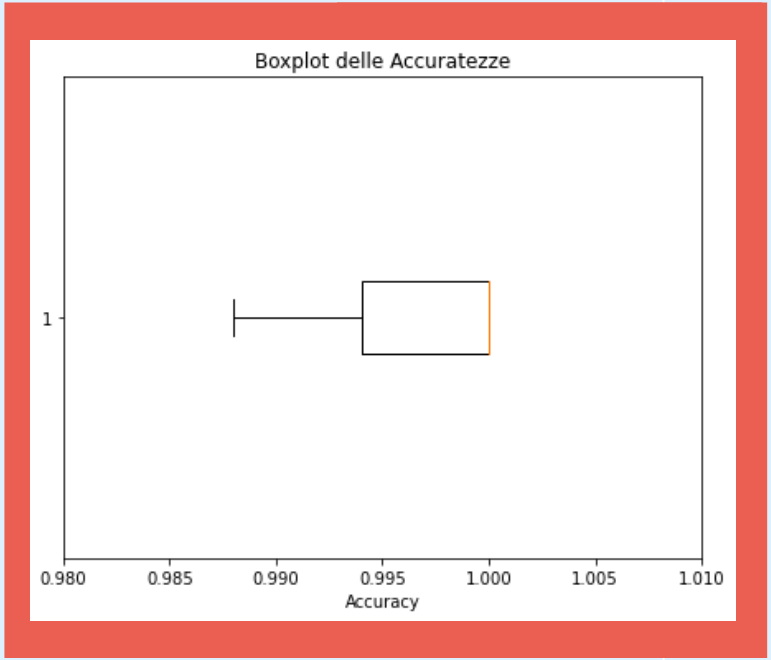


L'istogramma mostra che la maggior parte delle accuratezze si trova molto vicino al valore di 1.0, con una leggera variazione tra circa 0.988 e 1.0.

Questo indica che il modello è altamente preciso nella maggior parte delle iterazioni.

Statistica Descrittiva - boxplot

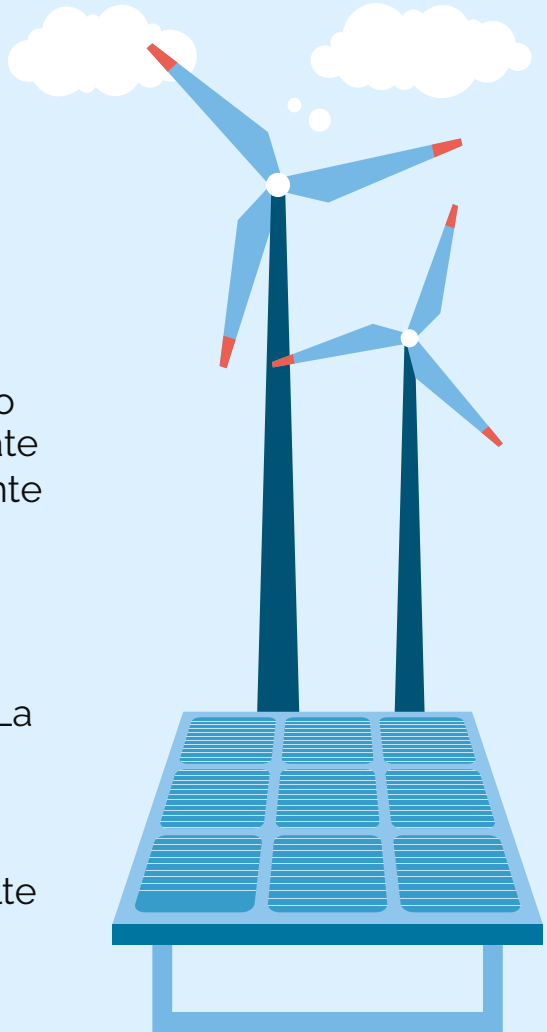
Il boxplot conferma che la mediana dell'accuratezza è vicina a 1.0.
L'intervallo interquartile (IQR) è molto stretto, indicando che le performance del modello sono consistenti.
Non ci sono outliers visibili, il che suggerisce che non ci sono iterazioni con accuratezze significativamente inferiori rispetto alle altre.



CONCLUSIONI - Analisi del Dataset

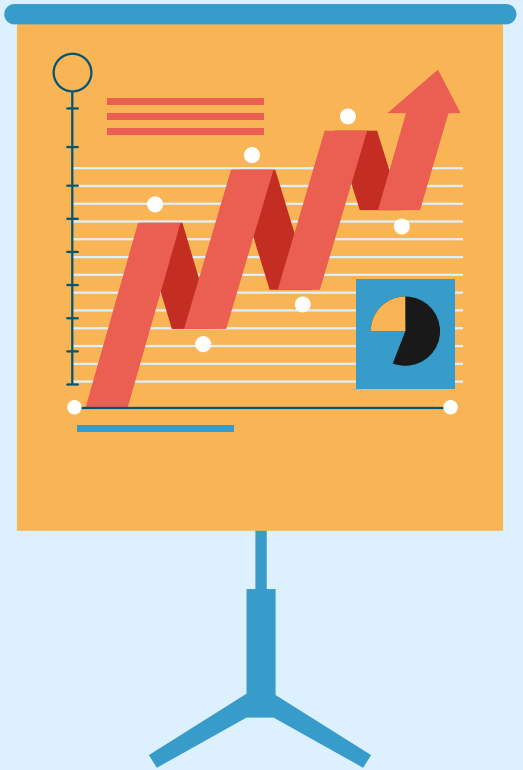
Nel progetto abbiamo osservato diverse relazioni e punti chiave riguardanti il consumo di carburante e le emissioni di CO_2 dei veicoli.

1. **Correlazione tra Consumo di Carburante ed Emissioni di CO_2**: Abbiamo identificato una forte correlazione positiva tra il consumo di carburante (misurato in L/100 km) e le emissioni di CO_2 (misurate in g/km). Questo indica che un aumento del consumo di carburante porta a un aumento delle emissioni di CO_2.
2. **Correlazione tra Cilindrata del Motore e Numero di Cilindri**: La correlazione di 0.92 indica che i veicoli con motori più grandi tendono ad avere più cilindri.
3. **Correlazione Negativa tra CO_2 Rating e Consumo di Carburante**: La correlazione negativa (-0.92) mostra che veicoli con un rating migliore tendono a consumare meno carburante.
4. **Correlazione Negativa tra CO_2 Rating ed Emissioni di CO_2**: Una correlazione di -0.96 indica che veicoli con emissioni di CO_2 più alte hanno un CO_2 Rating peggiore.



CONCLUSIONI – Regressione Lineare

1. **Relazione Lineare Positiva tra Consumo di Carburante e Emissioni di CO_2**: Il modello di regressione lineare tra il consumo di carburante combinato e le emissioni di CO_2 ha mostrato che per ogni incremento nel consumo di carburante, c'è un aumento corrispondente nelle emissioni di CO_2. Questo è rappresentato dalla retta di regressione con un R^2 di 0.93, indicando che il modello spiega bene la variabilità dei dati.
2. **Relazione Lineare Negativa tra Emissioni di CO_2 e CO_2 Rating**: Il modello di regressione lineare ha mostrato una relazione lineare negativa tra le emissioni di CO_2 e il rating di CO_2. La pendenza della retta indica che per ogni aumento nelle emissioni di CO_2, c'è una diminuzione corrispondente nel rating di CO_2. L'R^2 di 0.92 suggerisce che il modello lineare è appropriato per descrivere questa relazione.



CONCLUSIONI - Modelli di Classificazione

I modelli di classificazione, tra cui la Regressione Logistica e il Support Vector Machine (SVM), hanno mostrato prestazioni eccellenti nel predire le emissioni di CO_2 superiori alla media. Entrambi i modelli hanno ottenuto una precisione molto elevata sia sul validation set che sul test set.

1. **Regressione Logistica**: Questo modello ha ottenuto un'accuratezza del 100% sul validation set, dimostrando di essere altamente efficace nella classificazione dei dati. La matrice di confusione mostra che non ci sono falsi positivi né falsi negativi, indicando una classificazione perfetta delle istanze.
2. **SVM con Kernel Lineare**: Anche questo modello ha ottenuto un'accuratezza del 100% sul validation set, confermando l'ottima capacità del modello nel separare correttamente le classi. La matrice di confusione riflette una perfetta separazione delle istanze, senza errori di classificazione.
3. **SVM con Kernel RBF**: Questo modello ha mostrato una leggera diminuzione nell'accuratezza (99%) sul validation set, con un solo errore di classificazione. Tuttavia, l'accuratezza rimane estremamente alta, suggerendo che il modello è robusto e generalizza bene sui dati non visti.
4. **SVM Ottimizzato**: Dopo il tuning degli iperparametri, il modello SVM ottimizzato ha raggiunto un'accuratezza del 98.4% sul validation set e del 100% sul test set. La matrice di confusione per il test set mostra una classificazione perfetta, indicando che il modello è ben addestrato e altamente efficiente.

CONCLUSIONI – Valutazione della Performance

La valutazione delle performance dei modelli addestrati ha dimostrato risultati eccellenti:

1. **Validation Set**: I modelli di Regressione Logistica e SVM hanno ottenuto accuratezze estremamente elevate, dimostrando una grande capacità di generalizzare sui dati non visti durante l'addestramento. L'SVM con kernel lineare e la regressione logistica hanno raggiunto il 100% di accuratezza, mentre l'SVM con kernel RBF ha ottenuto il 99%, con un singolo errore di classificazione.
2. **Test Set**: Il modello SVM ottimizzato ha ottenuto un'accuratezza del 100% sul test set, dimostrando che la combinazione di iperparametri scelta è altamente efficace. La matrice di confusione conferma che tutte le istanze sono state classificate correttamente, senza errori.
3. **Robustezza del Modello**: L'analisi statistica delle accuratezze ha mostrato una media dell'accuratezza del 99.64% con una deviazione standard molto bassa, indicando che le performance del modello sono consistenti e robuste.

CONCLUSIONI

In conclusione, l'analisi dei dati sul consumo di carburante e sulle emissioni di CO_2 dei veicoli ha evidenziato:

1. **Relazioni Chiare tra Variabili**: Le analisi di correlazione e regressione lineare hanno evidenziato relazioni significative tra il consumo di carburante e le emissioni di CO_2, nonché tra le emissioni di CO_2 e il rating di CO_2.
2. **Efficacia dei Modelli di Classificazione**: I modelli di classificazione sviluppati, tra cui la Regressione Logistica e gli SVM, hanno dimostrato elevate capacità di predizione e generalizzazione, con accuratezze prossime al 100%.
3. **Validità delle Metodologie**: Le metodologie utilizzate, inclusi il pre-processing dei dati, l'EDA, la regressione lineare e la classificazione, si sono rivelate appropriate ed efficaci per raggiungere gli obiettivi del progetto.

Federico Sgambelluri

Matricola: 0001068826
E-mail:
federico.sgambelluri@studio.unibo.it