**Machine Learning per la Mobilità Elettrica: Stima dell'Autonomia delle Batterie**

Anno Accademico 2024/2025

Margio Antonio - 0512116137

Gragnaniello Francesco - 0512116465

## **IDENTIFICAZIONE DEL PROBLEMA**

La previsione dell'autonomia dei veicoli elettrici è un tema di grande interesse nel settore automobilistico e della mobilità sostenibile. Conoscere in anticipo l'autonomia stimata di un veicolo in base a caratteristiche come l'anno di produzione, il tipo di batteria e il consumo energetico può essere utile sia per i produttori di auto elettriche che per i consumatori.

L'obiettivo di questo progetto è sviluppare un modello di Machine Learning per stimare l'autonomia dei veicoli elettrici utilizzando un dataset opportunamente preprocessato.

## **CONTESTUALIZZAZIONE DELL’IMPORTANZA**

Avere un sistema basato su Machine Learning per la previsione dell'autonomia delle batterie è fondamentale per:

* Ottimizzare la gestione della ricarica e ridurre l'ansia da autonomia.
* Supportare produttori e acquirenti nella scelta del veicolo più adatto alle proprie esigenze.
* Favorire lo sviluppo di infrastrutture di ricarica più efficienti.

## **DESCRIZIONE DEL DATASET**

Abbiamo utilizzato il dataset "Electric Vehicle Dataset" disponibile su Kaggle (https://www.kaggle.com/datasets/yxshee/electric-vehicle-dataset). Questo dataset contiene diverse caratteristiche dei veicoli elettrici, tra cui:

* **Marca e modello**
* **Anno di produzione**
* **Tipo di veicolo elettrico**
* **Autonomia (variabile target)**
* **Prezzo base**
* **Utility elettrica**

Il dataset è stato pulito e preprocessato per eliminare dati mancanti e convertire variabili categoriche in numeriche.

## **PREPROCESSING DEI DATI**

Dopo l'analisi del dataset, è stato necessario eseguire alcune operazioni di prepocessing per preparare i dati per l’addestramento dei modelli di Machine Learning. Sono state eseguite le seguenti operazioni :

* **Gestione dei valori mancanti:** Le righe contenenti valori mancanti nella variabile target sono state rimosse, mentre per le feature restanti i valori mancanti sono stati sostituiti con la media della colonna corrispondente.
* **Encoding delle variabili categoriche:** Le variabili categoriche nel dataset, come Marca, Modello, Tipo di carburante e Trasmissione, devono essere trasformate in formato numerico affinché i modelli possano elaborarle correttamente. Ad esempio, consideriamo la variabile categorica “Marca” nel dataset, che può assumere valori come “Audi”, “BMW”, “Ford”, Toyota. Con il one-hot encoding, questa variabile viene suddivisa in diverse colonne binarie, una per ogni valore unico. Quindi, se una macchina è di marca Audi, la colonna corrispondente alla variabile “Audi” avrà valore 1, mentre le altre colonne avranno valore 0.

| **Marca Originale** | **Audi** | **BMW** | **Ford** | **Toyota** |
| --- | --- | --- | --- | --- |
| **Audi** | 1 | 0 | 0 | 0 |
| **BMW** | 0 | 1 | 0 | 0 |
| **Ford** | 0 | 0 | 1 | 0 |
| **Toyota** | 0 | 0 | 0 | 1 |

In questo modo, ogni valore unico della variabile categoria originale ha una sua colonna binaria corrispondente nel dataset codificato.

* **Suddivisione del dataset:** Il dataset è stato diviso in training set (80%) e test set (20%) per valutare l’accuratezza dei modelli.

## **MODELLI UTILIZZATI**

Sono stati addestrati tre modelli di Machine Learning per la regressione dell'autonomia dei veicoli elettrici:

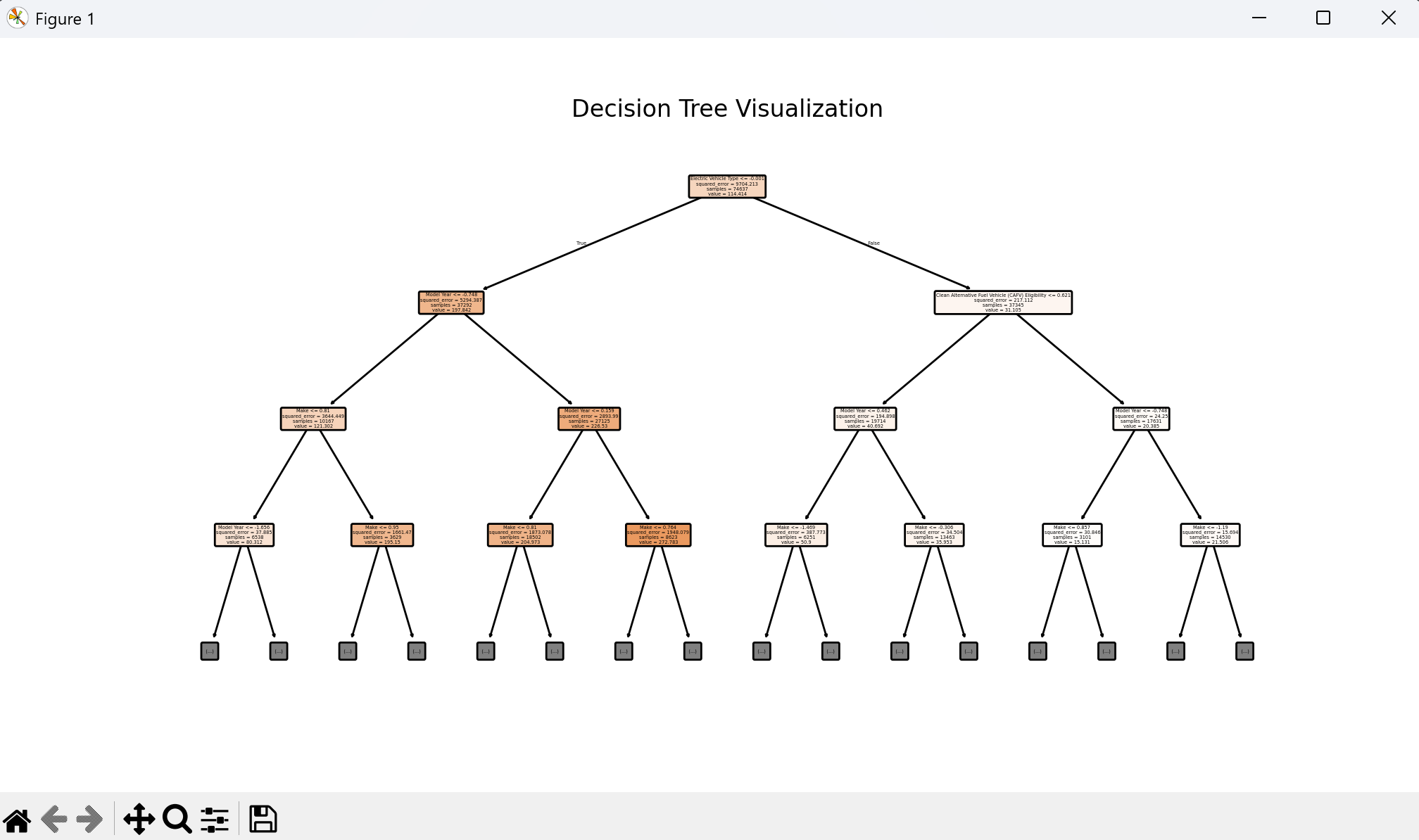
1. **Decision Tree**
2. **Random Forest**

Per ciascun modello sono stati calcolati i tempi di addestramento e le seguenti metriche di valutazione:

* **MAE (Mean Absolute Error):** Errore medio assoluto.
* **MSE (Mean Squared Error):** Errore quadratico medio.
* **R² Score:** Indica quanto bene il modello spiega la variabilità dell'autonomia.

## **DECISION TREE**

Un albero di decisione (Decision Tree) è un algoritmo di apprendimento supervisionato, utilizzato sia per la classificazione che per i compiti di regressione. Ha una struttura ad albero, che consiste in un nodo radice, rami , nodi interni e nodi foglia. Ogni nodo interno rappresenta l’esito del test, e ogni nodo foglia rappresenta un’etichetta di classe(decisione presa dopo il calcolo di tutti gli attributi). Gli alberi di decisione cercano di trovare la migliore divisione per suddividere i dati. Tuttavia, possono essere soggetti a problemi come il sovradattamento. Il modello ottiene queste prestazioni:



## 

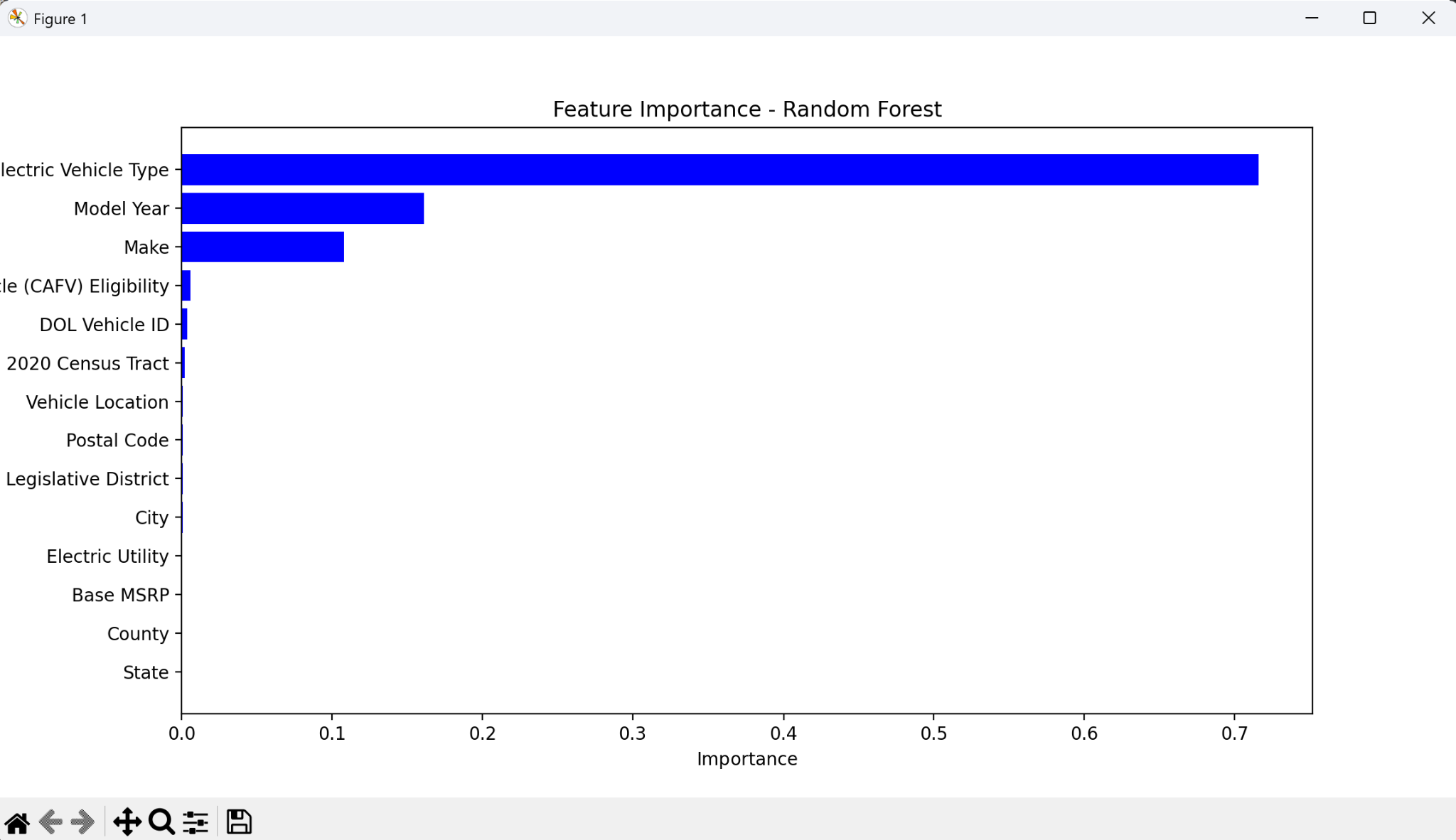
## 

## 

## 

## **RANDOM FOREST**

Random Forest è un algoritmo di apprendimento automatico comunemente utilizzato, che combina l’output di più alberi di decisione per raggiungere un singolo risultato. E’ un’estensione del metodo di bagging in quanto utilizza sia il bagging che la casualità delle caratteristiche per creare una foresta di alberi di decisione non correlati tra loro. La Random Forest può essere utilizzata sia per problemi di classificazione che di regressione.Questo algoritmo mantiene la sua accuratezza formando un insieme di alberi di decisione. Il modello ottiene queste prestazioni:



## **RISULTATI E ANALISI**

Dopo l’addestramento, i modelli sono stati testati su dati nuovi. Di seguito le prestazioni ottenute:

| **Modello** | **MAE** | **MSE** | **R² Score** | **Tempo di addestramento** |
| --- | --- | --- | --- | --- |
| Decision Tree | 3.75 | 155.5 | 0.98 | 0.42s |
| Random Forest | 3.59 | 92.0 | 0.99 | 30.45s |

Il modello **Random Forest** risulta il migliore rispetto al **Decision Tree**, per i seguenti motivi:

✅ **Errore più basso**

* **MAE** (Mean Absolute Error) più basso: **3.60 vs 3.76**, indica che le predizioni sono mediamente più vicine ai valori reali.
* **MSE** (Mean Squared Error) più basso: **92.09 vs 155.5**, quindi il modello fa meno errori grandi.

✅ **R² Score più alto**

* **0.99 vs 0.98**, significa che **Random Forest spiega meglio la variabilità dell'autonomia dei veicoli** rispetto al Decision Tree.

✅ **Maggiore robustezza e generalizzazione**

* **Random Forest** utilizza più alberi per ridurre il rischio di overfitting rispetto a un singolo Decision Tree.
* Questo si riflette in predizioni più stabili e accurate su nuovi dati.

❗ **Contro di Random Forest**

* Tempo di addestramento molto più lungo (**30.45s vs 0.42s**), ma in questo caso il miglioramento della precisione compensa il costo computazionale.

**Conclusione  
Random Forest è il modello migliore** perché offre maggiore accuratezza e capacità di generalizzazione, anche se richiede più tempo per l'addestramento.

## **CONCLUSIONI**

I risultati di questo progetto dimostrano come il Machine Learning possa essere utilizzato efficacemente per la previsione dell'autonomia dei veicoli elettrici.

Il modello **Random Forest** ha mostrato le migliori prestazioni, confermando la sua robustezza nella gestione di dati complessi. L'analisi della feature importance ha evidenziato che **l'anno di produzione e il tipo di veicolo** sono variabili chiave nella determinazione dell'autonomia.

Ulteriori miglioramenti potrebbero includere l’utilizzo di modelli più avanzati come le reti neurali e l’ottimizzazione degli iperparametri tramite tecniche di tuning avanzate.