***Valutazione del Prezzo delle Auto***

Anno Accademico: 2024/2025

Francesco Casillo-0512118276

***1. Identificazione del Problema***

Il progetto ha l’obiettivo di sviluppare un sistema di Machine Learning per \*\*classificare il prezzo delle auto in due categorie: Economica e Costosa .

Determinare il prezzo di un'auto in base alle sue caratteristiche è un problema fondamentale nel settore automobilistico. Un sistema automatico di previsione del prezzo può essere utile per concessionari, acquirenti e venditori, offrendo una stima rapida e affidabile del valore di un veicolo.

I modelli utilizzati in questo progetto sono:

- Decision Tree

- Random Forest

L’apprendimento sarà supervisionato con Batch Learning.

***2. Importanza del Problema***

Il prezzo di un'auto dipende da diversi fattori, tra cui:

* - Marca e modello
* - Anno di produzione
* - Dimensione del motore
* - Tipo di carburante(Diesel, Benzina, Ibrido, Elettrico)
* - Trasmissione (Automatica, Manuale, Semi-Automatica)
* - Chilometraggio
* - Numero di proprietari precedenti

Automatizzare il processo di valutazione del prezzo può migliorare l'efficienza del mercato automobilistico, riducendo il tempo necessario per stimare il valore di un'auto e minimizzando il rischio di valutazioni soggettive o errate.

***3. Descrizione del Dataset***

Il dataset utilizzato è il Car Price Dataset , che contiene informazioni su diverse caratteristiche delle auto, tra cui:

* - Brand : Marca del veicolo
* - Model : Modello
* - Year: Anno di produzione
* - Engine\_Size: Dimensione del motore
* - Fuel\_Type: Tipo di carburante
* - Transmission: Tipo di trasmissione
* - Mileage: Chilometraggio
* -Doors: Numero di porte
* - Owner\_Count: Numero di proprietari precedenti
* - Price: Prezzo dell'auto

Per facilitare la classificazione, la variabile target Price è stata trasformata in due categorie:

- Economica: Prezzo inferiore alla mediana

- Costosa : Prezzo superiore o uguale alla mediana

***4. Preprocessing dei Dati***

Le seguenti operazioni di preprocessing sono state applicate al dataset:

1. Gestione dei Valori Mancanti : Rimozione o imputazione di dati mancanti.

2. Encoding delle Variabili Categorical: Conversione di variabili testuali in numeri (One-Hot Encoding o Label Encoding).

3. Normalizzazione delle Variabili Numeriche: Scaling delle feature numeriche per migliorare le prestazioni dei modelli.

4. Split del Dataset: Divisione in Training Set (80%) e Test Set (20%).

***5. Modelli Utilizzati***

Dopo il preprocessing, sono stati addestrati due modelli di Machine Learning per la classificazione del prezzo delle auto:

5.1 Decision Tree

Un albero decisionale è un algoritmo che suddivide i dati in sottoinsiemi basati sulle caratteristiche più significative, creando una struttura ad albero per prendere decisioni.

5.2 Random Forest

La Random Forest è un insieme di più alberi decisionali, che migliora la precisione delle previsioni combinando i risultati di più alberi.

***6. Metriche di Valutazione***

Le prestazioni dei modelli sono state valutate utilizzando le seguenti metriche:

- Accuracy: Proporzione di previsioni corrette.

- Precision: Percentuale di auto classificate come "Costose" che lo sono realmente.

- Recall: Percentuale di auto "Costose" correttamente identificate.

- F1-Score: Media armonica tra Precision e Recall.

***7. Risultati e Grafici***

**7.1 Matrice di Confusione**

Una matrice di confusione è stata utilizzata per analizzare le prestazioni dei modelli e individuare eventuali errori di classificazione.

**7.2 Distribuzione del Prezzo delle Auto**

Un grafico a barre mostra la distribuzione delle auto classificate come "Economiche" e "Costose".

**7.3 Impatto delle Caratteristiche sul Prezzo**

- Boxplot del chilometraggio rispetto al prezzo.

- Grafico a barre della trasmissione rispetto al prezzo.

-Istogramma della distribuzione delle categorie di prezzo

- Grafico a barre del carburante rispetto al prezzo.

***8. Confronto tra i Modelli***

| Modello | Accuracy | Precision | Recall | F1-Score |
| --- | --- | --- | --- | --- |
| Decision Tree | 0,9145 | 0,917 | 0,9166 | 0,9168 |
| Random Forest | 0,95 | 0,96 | 0,95 | 0,95 |

Dal confronto emerge che il modello Random Forest ha ottenuto risultati migliori rispetto al Decision Tree, grazie alla sua capacità di ridurre l'overfitting e di gestire meglio la variabilità del dataset.

**9. Conclusioni e Sviluppi Futuri**

Il progetto ha dimostrato che è possibile classificare il prezzo delle auto con un buon livello di accuratezza utilizzando tecniche di Machine Learning.

**Possibili sviluppi futuri:**

- Aggiungere più feature (es. stato dell’auto, optional, posizione geografica).

- Utilizzare modelli più avanzati, come XGBoost o Deep Learning.

- Migliorare il preprocessing con tecniche di feature engineering avanzate.

***10. Implementazione Python***

L'implementazione completa del progetto, inclusi il preprocessing dei dati, l'addestramento dei modelli e la valutazione delle prestazioni, è disponibile nel codice Python allegato.