NEPSI: Is the price range of a mobile phone predictable ?

Lo scopo di questo progetto è quello di analizzare e studiare le caratteristiche comuni dei prodotti di telefonia mobile presenti sul mercato al fine di sviluppare un modello di previsione dei prezzi. Il modello NEPSI potrà aiutare il produttore a ottimizzare la pianificazione delle vendite e, al contempo, guidare i consumatori nell'acquisto di un dispositivo possa soddisfare le proprie esigenze.

Per raggiungere questo obiettivo, abbiamo analizzato un dataset contenente 2000 prodotti di telefonia mobile. Ogni prodotto è descritto attraverso 21 diversi attributi, classificando ogni prodotto nella corrispondente fascia di prezzo. Questi dati ci hanno permesso di applicare algoritmi di supervised machine learning per poter redigere il nostro prodotto.

Esplorando il dataset (sezione: EDA) abbiamo potuto verificarne la sua integrità e, al contempo osservare come le diverse fasce di prezzo sono distribuite su tutto il dataset.

Infine una matrice di correlazione ci ha permesso di analizzare le mutue relazioni.

Da tale ispezione si è potuto ricavare che le caratteristiche maggiormente correlate sono: Ram, potenza della batteria, larghezza e larghezza dei pixel.

Data l’analisi di cui sopra e al fine di agevolare produttori e consumatori nella valutazione dei parametri più rilevanti per l'investimento in un determinato prodotto, abbiamo preprocessato (sezione: Preprocessing) il dataset mediante l'analisi delle componenti principali (Principal Component Analysis, PCA). In questo modo, siamo stati in grado di concentrare il dataset su solo le caratteristiche che descrivono il dataset al 95% di varianza, riducendo il numero di attributi necessari per la predizione del prezzo. L'analisi ha confermato che la RAM, la risoluzione dello schermo (PIXEL) e la capacità della batteria sono attributi rilevanti per collocare un prodotto in una specifica fascia di prezzo.

Al fine di convalidare ulteriormente la scelta dei predittori, dal dataset completo sopra menzionato, si e’ deciso di utilizzare un algoritmo (sezione: Decision tree) di Machine Learning supervisionato, il modello ad albero decisionale (Decision Tree) sul dataset completo (sub sezione: Over all data). Successivamente ad un'ottimizzazione (sub sezione: Over all data-Parameter tuning) dei relativi parametri si e’ applicato il modello il quale ci ha permesso di ottenere un' accuratezza della previsione del prezzo, compresa tra l'80% e l'85%, con una percentuale di True Positive per le quattro classi compresa tra l'80% e il 95% (sub sezione: Over all data-train-analysis) . È importante sottolineare che, tra gli attributi più rilevanti per il modello, per tale previsione, sono risultati gli stessi attributi precedentemente ottenuti tramite PCA. Pertanto si è potuto ingegnerizzare e processarlo come da protocollo. Questa ottimizzazione ha permesso di ottenere un incremento dell'accuratezza vicino al 90%, con una percentuale di True Positive compresa tra il 90% e il 96%.

Successivamente, al fine di perfezionare ulteriormente la classificazione, abbiamo applicato delle tecniche ensemble (sezione: Ensemble), sempre basate su modelli di albero decisionale. Si e’ replicata la sequenza di esperimenti di cui sopra, ottenendo ulteriore conferma sugli attributi più rilevanti e sulla maggior efficienza dei modelli allenati con il dataset ingegnerizzato piuttosto che con il dataset integrale. Mediante un diagramma a barre, sono stati valutati contemporaneamente i tre modelli rilevandone una simile performance (RandomForest (RF), Bagging (BG) e AdaBoost (ADA)). I modelli hanno evidenziato un incremento dell’efficienza delle previsioni,rispetto al singolo albero decisionale, permettendo di superare la soglia dell'accuratezza del 90%.

In aggiunta, abbiamo affiancato un modello di regressione, il Support Vector Machine (SVM) (sezione SVM). Questo modello ha permesso l'applicazione di una tecnica di regressione su un dataset non lineare, non concessa invece dai modelli di regressione lineare classici. Dall’analisi è emerso che, il modello SVM ha mostrato le migliori performance sull'intero dataset (SVM\_T), superando un accuratezza della previsione del 97% rispetto ad una performance leggermente inferiore, con un decremento dei percentuali, quando applicato al dataset ingegnerizzato (SVM\_P).

Tuttavia la percentuale dei True Positive e’ stata valutata, per tutte le classi, superiore al 90%.

Nonostante la possibilità di poter utilizzare modelli che possano permettere la conoscenza di unicamente tre caratteristiche per la predizione del prezzo, si è deciso di utilizzare il modello più performante (SVM\_T).

In ogni caso vi è libertà di scelta, in base a conoscenza e disponibilità, dell’inserimento di qualsivoglia indicatore.

La dove non vi ci sarà indicazione, verrà automaticamente scelto un valore medio per tale caratteristica.

Nella sezione NEPSI model-Use it, può essere utilizzato il modello Nepsi.

Verrà inizialmente messa a supporto una lista delle caratteristiche da dover considerare con i relativi valori medi.

Guida all’uso:

Nepsi(＜caratteristica＞=＜valore＞,＜caratteristica＞=＜valore＞)

Inserire in elenco, separati da una virgola (come da esempio):

＜caratteristica＞, l’indicatore della caratteristica per la quale si vuole indicare un valore

＜valore＞, il valore per la caratteristica scelta

LIBRERIE USATE:   
Numpy, Pandas, Matplotlib, Seaborn, Sklearn

ALGORITMI USATI:

PCA, DECISION TREE, ENSAMBLE (Decision tree based algorithm), SVM

Autori: Guglielmo Tedeschi, Manuel Naviglio, Alessio Franchi