Proiect

Pachete Software

Autori: Codorean Andrei si Cazaceanu Octavian
Seria: C
Grupa: 1089

1. Scopul Proiectului:

1.1 Prezentare pe scurt:

Proiectul urmărește două scopuri principale:

- **Explorarea și Înțelegerea Datelor:** Analiza unui set de date despre smartphone-uri si laptop-uri, preluate prin metode de web scraping de pe platforma evomag.ro. Aceasta include curățarea datelor, extragerea specificațiilor relevante și pregătirea lor pentru analize ulterioare.
- **Vizualizarea Interactivă și Modelare Predictivă:** Crearea unei aplicații web interactive folosind Streamlit pentru a vizualiza diverse aspecte ale datelor (distribuția prețurilor, corelații între caracteristici, analiza brandurilor), cat si simplificarea procesului de testare a modelelor(xgboost) si algoritmilor(clusterizare) din punct de vedere al explorarii hyperparametrilor.

1.2 Sample set de date:

{
"timestamp": "2024_11_13_22_34",
"name": "Nou! Laptop Gaming ASUS TUF A15 FA507NUR (Procesor AMD Ryzen\u2122 7 7435HS (16M Cache, up to 4.50 GHz), 15.6\" Full HD 144Hz, 16GB, 512GB SSD, NVIDIA GeForce RTX 4050 @6GB, No OS, Negru/Gri)",
"price": 5599.99,
"rating": 0,
"number_of_reviews": 0,
"is_in_stoc": 1,
"url": "https://www.evomag.ro/portabile-laptopuri-notebook/asus-laptop-gaming-asus-tuf-a15-fa507nur-procesor-amd-ryzen-7-7435hs-16m-cache-up-to-4.50-ghz-15.6-full-hd-144hz-16gb-512gb-ssd-nvidia-geforce-rtx-4050-6gb-no-os-negru-gri-4186074.html",
"product_code": "ASFA507NUR-LP104",
"online_mag": "evomag",
"specifications": {
"Model": "TUF A15 FA507",
"Tip Laptop": "Gaming",
"Familie procesor": "AMD Ryzen\u2122 7",
"Numar nuclee": "8",
"Model procesor": "7435HS",
"Clock Speed (MHz)": "3100",

"Max Turbo Frequency (MHz)": "4500",
"Smart Cache (Kb)": "16384",
"Tehnologie fabricatie": "6 nm",
"Diagonala": "15.6\"",
"Rezolutie maxima": "1920x1080",
"Tip display": "LED backlight",
"TouchScreen": "Nu",
"Rata Refresh": "144 Hz",
"Placa video dedicata": "NVIDIA GeForce RTX 4050 ",
"Tip memorie": "DDR5",
"Camera Web": "Da, HD",
"Memorie video(MB)": "6144 MB dedicata",
"Capacitate memorie": "16384 MB",
"Format memorie instalata": "2x8192",
"Frecventa memorie": "4800 MHz",
"Capacitate maxima memorie": "16384 MB",
"Sloturi memorie": "2",
"Capacitate SSD": "512GB",
"Interfata HDD / SSD": "M.2 PCIe",
"Unitate optica": "Nu are",
"Audio": "Dolby Audio",
"Difuzoare": "Difuzoare stereo Microfoane duale",
"Retea cu Fir": "Gigabit Ethernet(10/100/1000Mbs)",
"Retea Wireless": "802.11 ax 2x2",
"Bluetooth": "Da, 5.3",
"Total porturi USB": "2 x USB 3.2 Type C Gen 2 2 x USB 3.2 Type A Gen 2",
"Retea (RJ-45)": "1 x Mufa RJ-45 (LAN Ethernet)",
"Iesire Audio": "1 x Mufa Casti/Boxe",
"HDMI": "1 x High-Definition Multimedia Interface",
"Tehnologie": "Lithium-Ion",
"Numar celule": "Baterie 4 cell",
"Sistem operare": "Fara Sistem de Operare",
"Lungime (mm)": "354",
"Latime (mm)": "251",
"Inaltime (mm)": "24.9",
"Dimensiuni(mm)": "354 x 251 x 22.4 - 24.9 mm",
"Greutate(Kg)": "2.2",
"Culoare": "Negru-Gri",
"Tastatura numerica": "Da",
"Tastatura iluminata": "Da"
},
"manufacturer": "ASUS",
"category": "Laptopuri / Notebook"

},

2. Componenta Streamlit:

Proiectul nostru de streamlit s-a concretizat intr-un dashboard centralizat care permite studiul datelor colectate intr-o maniera comfortabila. Fiecare sectiune este reprezentata de o functie, care contine subfunctii.

2.1 Prezentare Generală:

- Afișează un eșantion din date și statistici descriptive pentru coloanele numerice.
- Prezintă informații detaliate despre fiecare coloană, inclusiv tipul de date și procentajul valorilor lipsă.

2.2 Analiza Prețurilor:

- Oferă vizualizări interactive (histograme, box plot-uri) pentru distribuția generală a prețurilor și identificarea valorilor extreme.
- Permite analiza detaliată a prețurilor în funcție de brand, cu statistici agregate pentru brandurile de top selectate.
- Formule importante:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

2.3 Corelația Caracteristicilor:

- Include o matrice de corelație (heatmap) și grafice scatter plot interactive pentru a explora relațiile dintre diverse specificații numerice.
- Permite colorarea punctelor din scatter plot în funcție de brand.

2.4 Analiza Brandurilor:

- Prezintă distribuția brandurilor de top prin grafice de tip bară și pie (cotă de piață).
- Analizează distribuția unor caracteristici categorice (ex: sistem de operare) pentru brandurile de top, folosind heatmap-uri și tabele.

2.5 Predicția Prețurilor:

- Furnizează o interfață pentru antrenarea unui model de regresie XGBoost, cu selecția caracteristicilor și configurarea hiperparametrilor.
- Afișează performanța modelului (metrici, importanța caracteristicilor, grafic actual vs. prezis).
- Include o secțiune interactivă "Încearcă Modelul" pentru predicții pe baza inputului utilizatorului.

2.6 Clustering:

- Oferă o interfață pentru segmentarea smartphone-urilor folosind K-Means, cu selecția caracteristicilor și configurarea parametrilor.
- Prezintă graficele "Elbow" și Silhouette pentru a ajuta la determinarea numărului optim de clustere (k).
- După aplicarea K-Means, afișează vizualizări ale clusterelor (PCA sau 2D), numărul de elemente per cluster, valorile medii ale caracteristicilor și profiluri radar.
- Opțional, prezintă top producători și distribuția prețurilor per cluster.

2.7 Filtre Interactive:

- Conține un slider pentru selectarea intervalului de preț.
- Include un widget multi-selecție pentru filtrarea după brand.
- Afișează dinamic numărul de produse corespunzătoare filtrelor active.

2.8 Descărcarea Datelor:

- Oferă un buton pentru descărcarea setului de date filtrat în format CSV.

3. Componenta SAS

3.1 Data prep:

Am creat:

- Capacitate_SSD_GB si Capacitate_RAM_MB care reprezinta alternative numerice la alte coloane prezente in setul de date, ceea ce ne permite sa le analizam.
- Price_Category, care reprezinta distribuirea produselor in 3 subcategorii in functie de pret pentru a observa mai usor distribuirea laptopurilor, dar si pentru a remarca relatii cu alte variabile

Setul de date a trebui recombinat, deoarece sas l-a impartit in root(metadatele produselor) si specificatii(datele fiecarui produs)

3.2 Analiza datelor:

S-a efectuat analiza distribuției de frecvență pentru variabile cheie precum:

- **Familie_procesor**: Pentru a vedea popularitatea diferitelor tipuri de procesoare.
- **Sistem_operare**: Pentru a identifica cota de piață a sistemelor de operare.
- **TouchScreen**: Pentru a determina proporția laptopurilor cu ecran tactil.
- **Tastatura_iluminata**: Pentru a vedea răspândirea acestei caracteristici.
- **manufacturer**: Pentru a analiza cotele de piață ale producătorilor.
- **Price_Category**: Pentru a înțelege distribuția laptopurilor pe segmente de preț.
- Combinații (de ex., **manufacturer*Price_Category**): Pentru a vedea cum se poziționează producătorii pe diferitele segmente de preț.
- Această analiză a inclus generarea de tabele de frecvență și grafice de bare (diagrame de frecvență).

Analiza Prețurilor și a Specificațiilor Medii:

- S-a examinat modul în care prețul variază în funcție de diferite caracteristici:
    - Distribuția prețurilor a fost analizată pentru fiecare **Familie_procesor** și **Sistem_operare** (folosind box plots).
    - Prețurile medii și mediane au fost calculate pentru laptopurile cu/fără **TouchScreen** și cu/fără **Tastatura_iluminata**.
- Au fost calculate specificațiile medii (preț, RAM, SSD) pentru fiecare **manufacturer** pentru a compara ofertele acestora.
- Prețul mediu a fost analizat și pentru combinații de **manufacturer** și **Price_Category.**

Vizualizări Specifice:

- **Histograma Prețurilor**: Pentru a vizualiza distribuția generală a prețurilor laptopurilor.
- **Histograma Capacității SSD**: Pentru a vedea distribuția capacităților de stocare SSD.
- **Bubble Chart (Preț vs. RAM)**: Pentru a explora relația dintre preț, capacitatea RAM, capacitatea SSD (ca mărime a bulei) și producător (ca grupare/culoare), oferind o perspectivă multidimensională.