Task: Predikcija kategorije proizvoda na osnovu naslova

Kurs: Introduction to Machine Learning using Python
Modul: Kompletan ML projekat: Od početka do kraja

Kontekst zadatka

Zamisli da si deo razvojnog tima firme koja se bavi online trgovinom i koja svakodnevno u sistem unosi hiljade novih proizvoda. Pravi izazov: svaki proizvod mora biti tačno i brzo kategorizovan, ali ručna klasifikacija troši dragoceno vreme i povećava mogućnost greške.

Kako bi proces bio efikasniji, tim je odlučio da razvije inteligentan sistem koji će automatski predlagati odgovarajuću kategoriju na osnovu naziva proizvoda koji korisnik unosi.

Tvoje zaduženje je upravo razvoj tog modela – tvoje rešenje biće ključni deo sistema koji svakog dana olakšava posao stotinama kolega i unapređuje korisničko iskustvo na platformi.

Cilj zadatka

Cilj ovog zadatka je da razviješ model mašinskog učenja koji će automatski predlagati odgovarajuću kategoriju za svaki novi proizvod na osnovu njegovog naziva.

Na ovaj način doprinosiš tome da proces unosa proizvoda na online platformi postane brži, jednostavniji i precizniji – bez potrebe za ručnom klasifikacijom i sa manjom mogućnošću greške.

Tvoj model omogućava da svaki novi artikal odmah dobije pravu kategoriju, što ubrzava rad čitavog tima, olakšava pretragu i unapređuje iskustvo korisnika na sajtu.

U isto vreme, stičeš iskustvo kako se kreira i primenjuje ML rešenje za konkretan poslovni izazov, kao deo šireg digitalnog ekosistema.

Zašto je ovaj zadatak važan?

Automatska klasifikacija proizvoda nije samo tehnička inovacija – ona je ključ za efikasno poslovanje svake moderne online trgovine. Ručno razvrstavanje hiljada novih proizvoda oduzima vreme, povećava rizik od greške i usporava ceo proces.

Razvijanjem ovog modela, pokazuješ kako veštine mašinskog učenja mogu rešiti konkretan, realan problem – ubrzati rad tima, smanjiti troškove i obezbediti bolje korisničko iskustvo.

Istovremeno, učiš kako da vodiš kompletan ML projekat: od razumevanja biznis zahteva, preko pripreme podataka, do implementacije rešenja koje tim može odmah da koristi.

Ovaj zadatak ti daje priliku da pokažeš kako tvoje znanje može imati vidljiv, pozitivan uticaj na svakodnevni rad i zadovoljstvo korisnika – baš kao što se očekuje od data stručnjaka u savremenom poslovnom okruženju.

Šta imaš na raspolaganju? – Tvoj toolbox za kompletan ML projekat

Skup podataka

Na raspolaganju ti je realan, bogat skup podataka (`products.csv`) sa više od 30.000 proizvoda iz različitih kategorija.

Za svaki proizvod dobijaš:

- Product ID – jedinstveni identifikator;
- Product Title – naziv proizvoda (npr. Samsung Galaxy A52 128GB);
- Merchant ID – prodavac;
- Category Label – ciljna kategorija (npr. Mobile Phones, Laptops);
- Product Code – interni kod;

- Number of Views – broj pregleda;
- Merchant Rating – ocena prodavca;
- Listing Date – datum postavljanja.

Test proizvodi

Kada razviješ i istestiraš model, možeš ga proveriti i ručno – ubaci jedan od ovih naziva i vidi kako tvoj model reaguje:

naziv proizvoda	kategorija koju očekujemo od modela
iphone 7 32gb gold,4,3,Apple iPhone 7 32GB	Mobile Phones
olympus e m10 mark iii geh use silber	Digital Cameras
kenwood k20mss15 solo	Microwaves
bosch wap28390gb 8kg 1400 spin	Washing Machines
bosch serie 4 kgv39vl31g	Fridge Freezers
smeg sbs8004po	Fridge Freezers

Tvoj zadatak

Tvoj zadatak je da razviješ i podeliš kompletan projekat za automatsku klasifikaciju proizvoda po kategorijama, koristeći realan skup podataka.

Projekat treba da sadrži:

- trenirani model sačuvan u .pkl formatu;
- Python skript za treniranje i čuvanje modela;
- Python skript za interaktivno testiranje: korisnik unosi naziv proizvoda, model predviđa kategoriju;
- Jupyter radne sveske sa kompletnom analizom, inženjeringom karakteristika, treniranjem i evaluacijom modela;
- javno dostupan GitHub repozitorijum sa celim projektom, uključujući:
    - sve relevantne skriptove i sveske,
    - jasno napisan README sa uputstvom za korišćenje i

testiranje modela,
- dokumentaciju koda i logičnu organizaciju projekta.

Napomena: Očekuje se da svaka komponenta bude jasno dokumentovana i spremna za dalje korišćenje ili razvoj u timu.

Roadmap za rešavanje zadatka

Pristupi ovom zadatku kao pravi član data tima – svaki korak vodi te korak bliže rešenju koje ceo tim može odmah da koristi!

1. Kreiraj GitHub repozitorijum.
   - Postavi projekat javno, kao što bi to uradio u pravom timu (naziv, opis, inicijalni README).

2. Kloniraj repozitorijum lokalno.
   - Ako želiš i lokalni rad, obezbedi sinhronizaciju između svog računara i remote repozitorijuma.

3. Kreiraj jednu ili više radnih svezaka (Colab/Jupyter).
   - Ovde eksperimentišeš, analiziraš i obrađuješ podatke.

4. Učitaj i istraži podatke.
   - Razumi strukturu skupa i detektuj potencijalne probleme ili nedoslednosti.

5. Pripremi i očisti podatke.
   - Reši prazne vrednosti, standardizuj podatke i pripremi ih za modeliranje.

6. Istraži mogućnosti inženjeringa karakteristika.
    - Probaj različite feature – svaki dodatak treba da ima svrhu i potencijalnu vrednost za model.

7. Uporedi performanse više ML algoritama.
    - Ispitaj različite modele i nađi najbolji balans između preciznosti, robustnosti i interpretabilnosti.

8. Treniraj i sačuvaj finalni model.
    - Izaberi najbolje rešenje, treniraj finalni model i sačuvaj ga za kasniju upotrebu.

9. Pravovremeno pravi checkpointe / commituj promene.
    - Nakon svakog većeg koraka, napravi commit – važno je da tvoj rad bude transparentan i pregledan kroz istoriju projekta.

10. Kreiraj završne skriptove:
    - train_model.py – logika za treniranje i čuvanje modela (bazirano na tvojim analizama i testiranju);
    - predict_category.py – logika za učitavanje modela i interaktivno testiranje (unos naslova proizvoda, dobijanje kategorije).

11. Organizuj projekat i dodaj README.md.
    - Neka struktura repozitorijuma bude logična, a README jasan i upotrebljiv za svakog člana tima.

Na kraju, proveri da li bi tvoj projekat mogao da koristi bilo koji član tima – bez dodatnih objašnjenja. To je moćan signal profesionalnosti!

Resursi i pomoć

Ako zapneš ili želiš više da istražiš, koristi ove smernice kao polaznu tačku – a najbolje uvide i mini-failove obavezno zapiši za tim!

- Inženjering karakteristika
  Ne zadržavaj se samo na Product Title – eksperimentiši sa dodatnim karakteristikama (feature engineering) koje mogu poboljšati model:
    - broj reči ili karaktera u naslovu;
    - prisustvo brojeva ili specijalnih znakova;
    - da li naslov sadrži naziv brenda ili pojmove napisane velikim slovima (USB, LED…);
    - dužina najduže reči…
      Ove male karakteristike mogu biti ključne za precizniju kategorizaciju!
      Preporuka: dokumentuj zapažanja – šta ima smisla, šta nije dalo rezultat – kako bi ceo tim mogao da razume tvoj proces.

- Evaluacija modela
  Model evaluiraj pomoću sledećih metrika:
    - tačnost (accuracy);
    - klasifikacioni izveštaj (precision, recall, F1 score po kategorijama);
    - matrica zabune (idealno i vizualizovana – to olakšava timu da vidi gde model greši).

Predaja zadatka

Na kraju zadatka, tvoj javni GitHub repozitorijum treba da sadrži:

- skup podataka (products.csv) korišćen za treniranje modela;
- bar jednu .ipynb radnu svesku sa kompletnom analizom i razvojem rešenja (ili više svezaka sa jasno podeljenim fazama);
- Python skriptove: train_model.py (za treniranje) i predict_category.py (za interaktivno testiranje modela);
- jasan i pregledan README.md fajl sa uputstvima za pokretanje i testiranje projekta.

Kako predaješ zadatak?

- Proveri da li je repozitorijum javan i kompletan.
- Pregledaj README – da li svaki korisnik može da pokrene tvoj projekat prateći uputstva?
- Pošalji link do svog GitHub repozitorijuma instruktoru kursa putem predviđene forme na platformi za učenje.

Brza provera pre predaje:

- Da li je svaki korak analize jasno prikazan i komentarisan tvojom logikom?
- Može li neko iz tima nastaviti rad ili odmah koristiti tvoj model bez dodatnog objašnjenja?
- Radi li kod (sveske i skriptovi) bez greške kada se pokrene od nule?
- Da li je struktura projekta pregledna, a dokumentacija korisna i jasna?

Pre predaje, pitaj sebe: Mogu li ja (ili bilo ko iz tima) za mesec dana, bez gledanja dodatnih materijala, lako nastaviti razvoj na osnovu ovog projekta?

Kriterijumi za ocenjivanje

Tvoj projekat će biti ocenjen na osnovu sledećih kriterijuma:

Kriterijum	Šta to znači u praksi?	Udeo u oceni
Organizacija i kompletan GitHub repozitorijum	Projekat je jasno strukturisan, sadrži sve tražene fajlove i može se lako koristiti ili dodatno razvijati.	20%

Analiza i čišćenje podataka	Jasno prikazana analiza podataka i temeljno očišćen skup, sa objašnjenjem svakog koraka.	20%
Inženjering karakteristika	Primećene su korisne dodatne karakteristike koje poboljšavaju model i odluke su dokumentovane.	20%
Treniranje i uporedna evaluacija više modela	Trenirana i unapređena su najmanje dva modela. Jasno su prikazani njihovi rezultati i obrazložen izbor finalnog rešenja.	15%
Implementacija i funkcionalnost predict_category.py	Skript je funkcionalan, precizno predviđa kategorije i može se lako koristiti za testiranje novih naslova.	15%
Dokumentacija i objašnjenje odluka	README i komentari su jasni i korisni i omogućavaju timu da brzo razume i koristi projekat.	10%

Način ocenjivanja

Tvoj rad će biti ocenjen u rasponu od 0 do 5 zvezdica, u zavisnosti od procentualnog ispunjenja kriterijuma iz prethodne sekcije.

Ova skala ti daje jasnu povratnu informaciju gde se nalaziš na putu od samostalne tehničke vežbe do projekta koji tim može odmah koristiti i nadograđivati.

Rasponi i značenja ocena

(0–59%): Nedovoljno.

Osnovni zahtevi nisu ispunjeni: kod ne funkcioniše, zadatak je nepotpun ili nedovoljno dokumentovan. Ovo je prilika da probaš ponovo, koriguješ pristup i obratiš više pažnje na čitljivost i upotrebljivost.

(60–69%): Delimično.

Većina koraka je započeta, ali nedostaje jasna veza između faza, komentari su minimalni, a analiza nije u celini razumljiva ili upotrebljiva drugima.

(70–79%): Zadovoljava.

Većina kriterijuma je ispunjena, kod radi i analiza je ispravna, ali rad još nije spreman za timsku upotrebu bez dorada ili dodatnih objašnjenja.

(80–89%): Dobro.

Kod je jasan i dobro komentarisan, analiza ima smisla, a interaktivno testiranje radi bez greške. Projekat može koristiti i drugi član tima bez većih problema.

(90–95%): Odlično.

Sve funkcioniše bez greške, komentari i dokumentacija su na visokom nivou, a tvoje objašnjenje pokazuje razumevanje šire slike i spremnost za rad u timu.

(96–100%): Izuzetno!

Tvoj projekat je referentno rešenje: može odmah biti primenjen u timu, jasan je, edukativan, lako se koristi i nadograđuje. Komentari i objašnjenja su izuzetno korisni i mogu inspirisati druge.

Cilj ove skale nije samo proći, već razviti profesionalan pristup i isporučiti rešenje koje može doprineti celom timu.