# Fashion MNIST klasifikacija

Isidora Tulumbić IN52/2019, itulumbic@gmail.com

Sara Marković IN46/2019, smarkovic61@yahoo.com

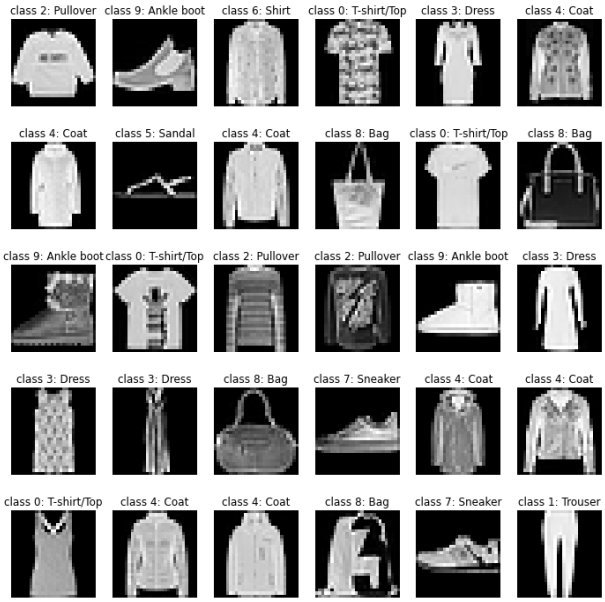
## I. UVODA

Fashion MNIST predstavlja skup podataka kreiran 2017. godine od strane istraživača i od tada je popularan resurs koji se koristi u mašinskom učenju. Ovaj skup podataka se koristi kao zamena za originalni MNIST skup podataka, koji se sastoji od ručno pisanih brojeva, a smatra se nedovoljno kompleksnim za savremene algoritme mašinskog učenja.

## II. BAZA PODATAKA

Ova baza podataka se sastoji od 70 000 *grayscale* slika odeće, veličine 28*28 piksela, koje su razvrstane u 10 kategorija (Slika 1): ***Majica, Džemper, Pantalone, Haljina, Kaput, Sandale, Košulja, Patike, Torba i Čizme***. Svaka kategorija poseduje po 7 000 slika. Svakom pikselu je dodeljenja vrednost, koja ukazuje na to koliko je piksel svetao, odnosno taman, gde veće vrednosti ukazuju na tamnije nijanse sive boje. Ove vrednosti se kreću od 0 do 255.

Podaci su podeljeni na trening i test skup, gde trening skup sadrži 60 000 slika, a test skup 10 000 slika. Trening i test skupovi imaju po 785 obeležja i sva obeležja su numerička.




**Slika 1**: Prikaz odevnih predmeta klasa i kategorija kojim pripadaju

## III. ANALIZA PODATAKA

### A. *Predobrada nedostajućih podataka*

Analizom je utvrđeno da u bazi nema nedostajućih podataka, kao ni podataka koji imaju netipične vrednosti za ovakav skup podataka.

### B. *Razdvajanje podataka*

Na početku, podaci su razdvojeni na dva niza: jedan koji sadrži podatke o slikama i drugi koji sadrži korespodentne oznake. Razdvajanje omogućava korišćenje slika kao ulaznih parametara za klasifikacioni algoritam i pruža adekvatne oznake za obuku i evaluaciju.

### C. *Podela na trening i validacione podatke*

Kod mašinskog učenja važno je imati odvojen skup podataka za obuku modela i skup podataka za evaluaciju performansi modela, kako bi se procenila sposobnost modela da generalizuje naučeno na novim, neviđenim podacima.

**Trening skup** podataka se koristi za obuku klasifikacionog modela i sadrži 75% podataka od ukupnog skupa podataka. Algoritam koristi ovaj skup da bi naučio veze između slika i njihovih oznaka, kako bi bio u stanju da klasifikuje nove, nepoznate slike.

**Validacioni skup** podataka se koristi za evaluaciju performansi modela tokom treninga i sadrži 25% podataka od ukupnog skupa. Nakon svake epohe obučavanja model se testira na validacionom setu kako bi se utvrdila preciznost i drugi oblici performansi. Validacioni skup pomaže i praćenju generalizacije modela i otkrivanju prenaučenosti modela.

### D. *Standardizacija*

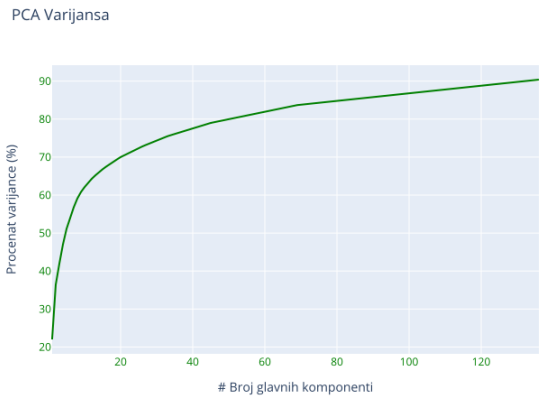
Nakon podele podataka, izvršena je standardizacija podataka. Neke od prednosti standardizacije su: **Stabilnosti algoritma** gde se smanjuje efekat velikih vrednosti piksela koji dovode do nestabilnosti algoritma, zatim **uklanjanje razlike u skalama**, gde standardizacija postavlja sve vrednosti piksela na isti opseg, što pomaže modelu da nauči bitne karakteristike bez obzira na veličinu piksela i **na ubrzanje konvergencije** ka optimalnoj vrednosti parametara. Standardizacijom je postignuta

normalizacija vrednosti piksela u slikama. Srednja vrednost slika se postavlja na nulu (*mean* = 0), a standardna devijacija na jedan (*standard deviation* = 1). Normalizacija je primenjena na trening, test i validacioni skup.

### *E. Redukcija dimenzija*

Poslednji korak koji je izvršen pre same obuke modela je redukcija dimenzija. Redukcija dimenzija je postupak smanjivanja dimenzionalnosti skupa podataka. Kada je reč o Fashion MNIST skupu podataka, slike su dimenzija 28*28 piksela, što znači da svaka slika ima 784 piksela. To čini ovu bazu podataka relativno malom, ali se redukcija ipak može primeniti kako bi se smanjila složenost modela i poboljšala efikasnost procesa obuke modela.

U ovom radu korišćena je PCA (*Principal Component Analysis*) metoda (Slika 2). Ona detektuje kombinacije atributa koje nose najveći deo varijabilnosti u podacima. PCA transformiše originalne atribute u manji broj novih atributa.




**Slika 2**: PCA varijansa – grafički prikaz

Kada se pogleda ovaj grafikon, vidi se koliko ukupne varijanse mogu da se objasne korišćenjem različitog broja glavnih komponenti.

## IV. OBUKA MODELA

U ovom radu korišćene su sledeće tri metode za klasifikaciju: *Gaussian Naive Bayes*, *Random Forest Classifier* i *XGBoost algoritam*. U nastavku rada biće predstavljene sve tri metode i međusobno upoređivanje istih.

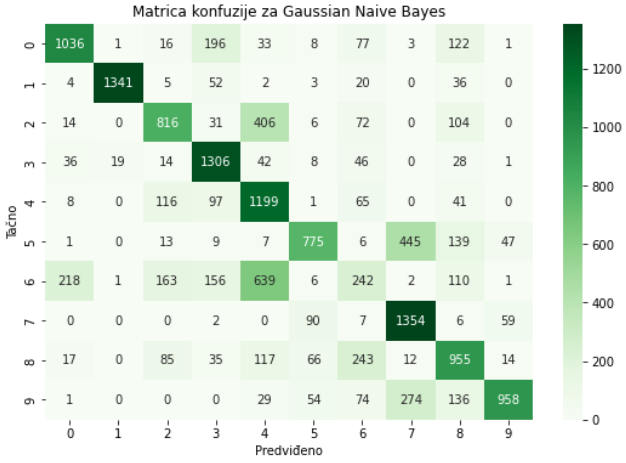
### *A. Gaussian Naive Bayes*

*Gaussian Naive Bayes* se zasniva na Bayesovoj teoremi koja se koristi za izračunavanje verovatnoće događaja, imajući unapred informacije o tom događaju.

Ovaj algoritam se zasniva na pretpostavci da su podaci iz svake klase normalno raspoređeni, tj. da distribuiraju prema Gausovoj (normalnoj) raspodeli. Nakon pripreme podatka i izdvajanja intenziteta piksela koji se koriste kao ulazni podaci za model, za svaku klasu računaju se parametri Gausove raspodele, tj. srednja vrednost i standardna devijacija. Zatim se za svaku klasu računa verovatnoća da slika pripada toj klasi korišćenjem Bayesove teoreme. Na kraju se slika klasifikuje u klasu sa najvećom verovatnoćom.

Kada je reč o Fashion MNIST klasifikaciji, *Gaussian Naive Bayes* postiže tačnost od 67% i na trening i na validacionom skupu podataka, dok je vreme predviđanja 0.8 sekunde.

Na slici 3 (Slika 3) nalazi se matrica konfuzije koja prikazuje koliko je puta svaka klasa tačno klasifikovana i koliko puta je pogrešno klasifikovana. Broj redova i kolona u matrici odgovara broju klasa u skupu podatka.




**Slika 3**: Matrica konfuzije kod *Gaussian Naive Bayes* klasifikacije

Iz matrice konfuzije se može zaključiti da najčešće dolazi do pogrešne predikcije kod kategorije 6 tj. kategorije Košulja, gde od 1538 uzoraka u 1296 slučajeva, algoritam pravi grešku. U najvećem broju slučajeva algoritam umesto klase Košulja predviđa klasu 4 tj. klasu Kaput (639 slučajeva). Najbolje predviđanje algoritam ima za kategoriju 7 tj. kategoriju Patike, gde od 1518 uzoraka u 1354 slučajeva algoritam daje tačnu pretpostavku.

### *B. Random Forest Classifier*

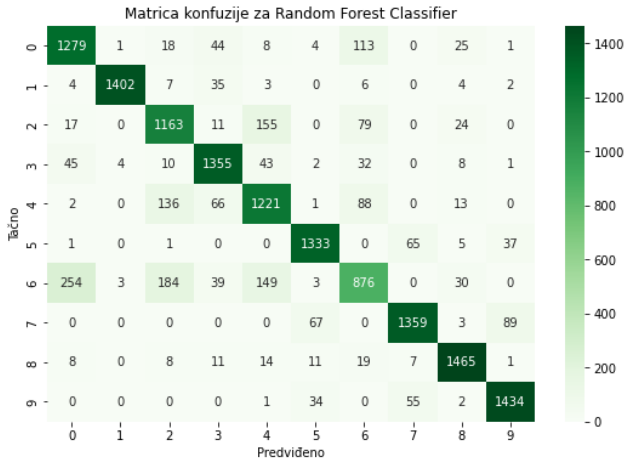
*Random Forest Classifier*, kao što mu i sam naziv kaže, sadrži veliki broj individualnih stabala odluke koji funkcionišu kao ansambl (skup većeg broja stabala koji zajednički donose odluke).

Nakon prikupljanja i pretprocesiranja podataka izvršava se obuka modela na osnovu trening skupa. Svako stablo se formira na osnovu podskupa trening podataka koji se slučajno odabiru iz slučajno izabranog podskupa atributa. Na kraju se vrši klasifikacija koju svako stablo vrši nezavisno i za svaki ulazni uzorak odlučuje o njegovoj klasi. Konačna klasifikacija se vrši na osnovu glasanja svih stabala, gde klasa sa najvećim brojem glasova postaje konačna klasifikacija.

Kada je reč o Fashion MNIST klasifikaciji, *Random Forest Classifier* postiže tačnost od 100% na trening skupu i 86% na validacionom skupu podataka, dok je vreme

predviđanja 69 sekundi.

Na slici 4 (Slika 4) nalazi se matrica konfuzije kod ovog algoritma.




**Slika 4**: Matrica konfuzije kod *Random Forest* klasifikacije

Iz matrice konfuzije se može zaključiti da najčešće dolazi do pogrešne predikcije kod kategorije 6 tj. kategorije Košulja, gde od 1538 uzoraka u 662 slučajeva, algoritam pravi grešku. U najvećem broju slučajeva algoritam umesto klase Košulja predviđa klasu 0 tj. klasu Majica (254 slučajeva). Najbolje predviđanje algoritam ima za kategoriju 8 tj. kategoriju Torba, gde od 1544 uzoraka u 1465 slučajeva algoritam daje tačnu pretpostavku.

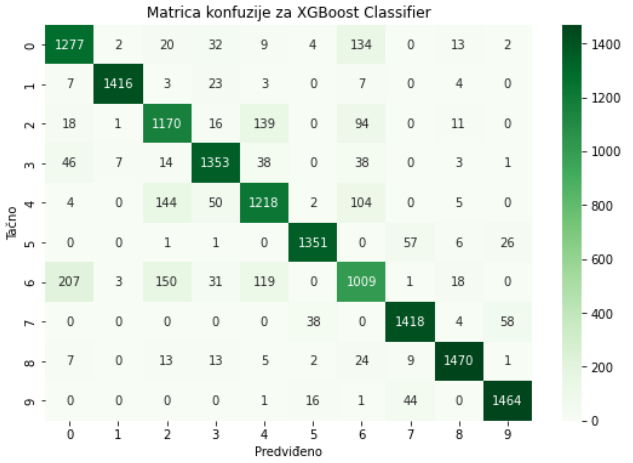
### C. *XGBoost algoritam*

*XGBoost* algoritam takođe spada u kategoriju ansambl metoda i zasniva se na tehnici pojačavanja, koja kombinuje više slabih modela kako bi se postigao snažan model za klasifikaciju. Razlika u odnosu na *Random Forest Classifier* jeste što se kod ovog algoritma stabla grade sekvencijalno što znači da je svako naredno stablo zavisno od ishoda poslednjeg stabla.

Nakon prikupljanja i pretprocesiranja podataka vrši se obuka modela. XGBoost model se obučava iterativno. Svaka iteracija se naziva "runda" ili "*boosting* runda". U svakoj rundi se konstruiše stablo odlučivanja koje se dodaje na prethodno konstruisano stablo radi poboljšanja klasifikacije. Algoritam je fokusiran na instance koje su bile pogrešno klasifikovane u prethodnim rundama. Krajnja klasifikacija se vrši na osnovu kombinacije rezultata svih stabala.

Kada je reč o Fashion MNIST klasifikaciji, *XGBoost algoritam* postiže tačnost od 100% na trening skupu i 88% na validacionom skupu podataka, dok je vreme predviđanja 487 sekunde.

Na slici 5 (Slika 5) nalazi se matrica konfuzije kod ovog algoritma.




**Slika 5**: Matrica konfuzije kod *XGBoost algoritma*

Iz matrice konfuzije se može zaključiti da najčešće dolazi do pogrešne predikcije kod kategorije 6 tj. kategorije Košulja, gde od 1538 uzoraka u 529 slučajeva, algoritam pravi grešku. U najvećem broju slučajeva algoritam umesto klase Košulja predviđa klasu 0 tj. klasu Majica (207 slučajeva). Najbolje predviđanje algoritam ima za kategoriju 8 tj. kategoriju Torba, gde od 1544 uzoraka u 1470 slučajeva algoritam daje tačnu pretpostavku.

### D. *Međusobno upoređivanje*

Kada je u pitanju tačnost na trening skupu, sa slike ispod (Slika 6) može se videti da *Random Forest Classifier* i *XGBoost algoritam* postižu tačnost od 100% u odnosu na *Gaussian Naive Bayes* koji postiže tačnost od 67%.

| Algoritam | Tačnost na trening skupu |
|---|---|
| Gaussian Naive Bayes | 67% |
| Random Forest Classifier | 100% |
| XGBoost | 100% |

**Slika 6**: Tačnost algoritama na trening skupu

Dalje, kada je u pitanju tačnost na validacionom skupu, sa slike ispod (Slika 7) može se videti da *XGBoost algoritam* postiže najbolju tačnost od 88% u odnosu na *Random Forest Classifier* koji postiže nešto manju tačnost od 86%, dok *Gaussian Naive Bayes* postiže najmanju tačnost od 67%.

| Algoritam | Tačnost na validacionom skupu |
|---|---|
| Gaussian Naive Bayes | 67% |
| Random Forest Classifier | 86% |
| XGBoost | 88% |

**Slika 7**: Tačnost algoritama na validacionom skupu

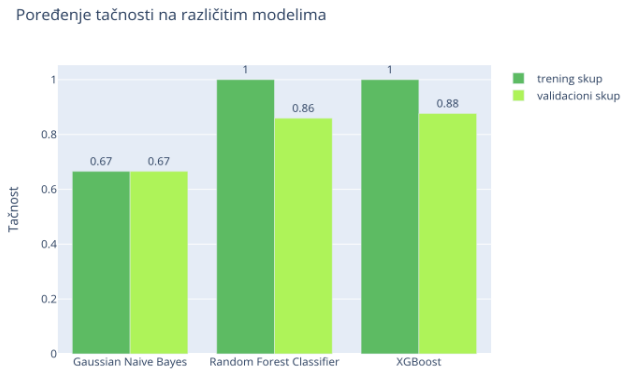
Što se tiče vremena predviđanja na slici ispod jasno je da najmanje vremena zahteva *Gaussian Naive Bayes* – 0.8

sekunde, a najviše *XGBoost* od čak 487 sekunde.

| Algoritam | Vreme predviđanja (sekunde) |
|---|---|
| Gaussian Naive Bayes | 0.8 |
| Random Forest Classifier | 69 |
| XGBoost | 487 |

**Slika 8**: Vreme predviđanja

Na slici 9 (Slika 9) su prikazane uporedne vrednosti za tačnost algoritama na trening i validacionom skupu podataka.




**Slika 9**: Tačnost algoritama na trening i validacionom skupu

Vreme izvršavanja kod *Gaussian Naive Bayes* za trening model je linearno u odnosu na broj instanci u skupu podataka, dok je za klasifikaciju pojedinačnih instanci konstantno. U proseku, složenost je **O(n*d)**, gde je n broj instanci, a d broj obeležja. Kod *Random Forest Classifier* vreme izvršavanja raste sa povećanjem broja stabala i broja obeležja. Složenost je u proseku **O(m*$n$*log(n))**, gde je m broj stabala, a n broj instanci. U opštem slučaju, *XGBoost* je sporiji od *Gaussian Naive Bayes-a* i *Random Forest Classifier-a*, ali može biti mnogo precizniji. Složenost je u proseku **O(n*$d^2$)**, gde je n broj instanci, a d broj obeležja.

## V. Zaključak

Kada se sve sumira, može se zaključiti da sva tri algoritma postižu veću tačnost na trening skupu u odnosu na validacioni skup, što može ukazivati na natprilagođavanje modela. *XGBoost* i *Random Forest Classifier* postižu više tačnosti na validacionom skupu u odnosu na *Gaussian Naive Bayes*, što pokazuje da su ova dva algoritma bolja za klasifikaciju Fashion MNIST skupa podataka.

Takođe, može se zaključiti da najbolje performanse postiže *XGBoost* algoritam kada je u pitanju Fashion MNIST skup podataka.