Izveštaj
Klasifikacija ručno napisanih cifara

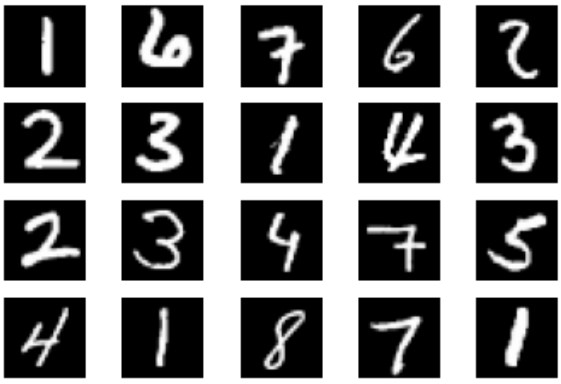
Nikola Šljivkov, IN56-2018, sljivkov@gmail.com
Velibor Vasiljević, IN23-2018, veliborvasiljevic7@gmail.com

1. Uvod

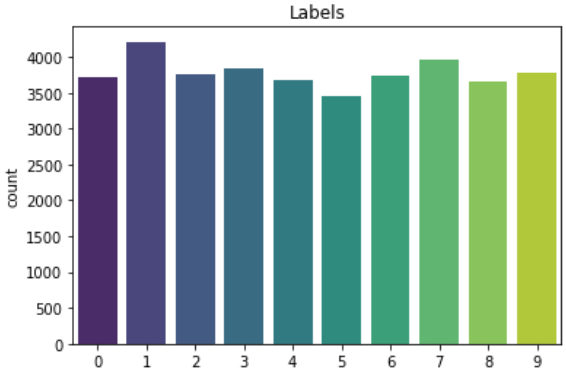
Cilj ovog projekta je da utvrdimo efikasnost klasifikatora prilikom klasifikacije ručno napisanih cifara od 0 do 9. Za ovaj projekat ćemo koristiti bazu podataka koja sadrži 42000 (četrdeset i dve hiljade) trening uzoraka odnosno cifara i 28000 (dvadeset i osam hiljada) test uzoraka. Svaki uzorak je slika formata 28x28 i zbog toga u trening skupu svaki uzorak poseduje 784 atributa plus jedan dodatni atribut koji služi kao labela. Test skup ne sadrži labele.

2. Analiza podataka i pca

Trening skup ne sadrži null vrednosti tako da nismo morali da odradimo čišćenje podataka. Pošto podaci sadrže 784 dimenzije odnosno atributa (Slika 1) koristili smo pca (eng. Principal Component Analysis) kako bi dimenzionalnost uzoraka sveli na 36 radi lakšeg klasifikovanja.



Slika 1. Izgled cifara



Slika 2. Broj cifara u trening skupu

3. Odabir modela

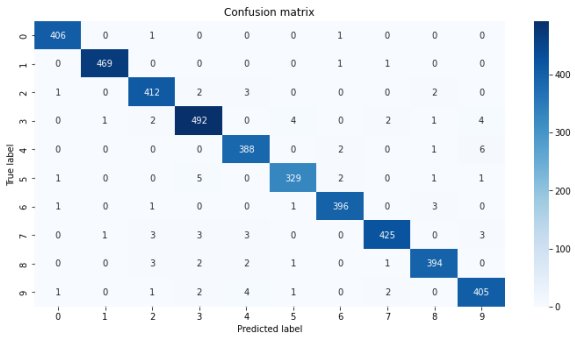
Klasifikatore kojem smo koristili u zadatku su kNN (k-Nearest Neighbors), SVM (Support Vector Machine) i Stablo odluke. Pomoću unakrsne validacije smo trenirali i testirali modele sa različitim parametrima nad trening skupom. Za unakrsnu validaciju skup je podeljen na 5 delova i svaki od tih 5 delova je u jednoj iteraciji bio test skup. Nakon što smo pronašli najbolji model pomoću unakrsne validacije testirali smo ga nad validacionim skupom a potom ga i istrenirali nad tim skupom kako bi povećali preciznost.

3.1. SVM

Parametre koje smo podešavali su bili "širina" margine i funkcija margine. Širine koje smo varirali su 1, 5 i 10, a funkcije su radijalna i polinomijalna. Nakon upoređivanja modela sa različitim parametrima utvrdili smo da je najbolji model sa 'širinom' margine 5 i radijalnom funkcijom, koji je imao preciznost 0.98 nad validacionim skupom. U tabeli 1 možemo videti kako je 'širina' margine utica na preciznost. Svi modeli u tabeli su koristili radijalnu funkciju.

SVM modeli	Procenat uspesnosti modela
C = 1	0,97
C = 5	0,98
C = 10	0.98

Tabela 1. Rezultati SVM modela



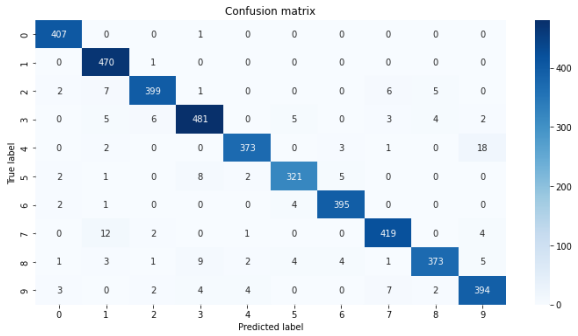
Slika 3. Matrica konfuzije za SVM

3.2. kNN

Parametre koje smo podešavali su metrika i broj komšija prilikom klasifikacije. Za metriku smo varirali euklidsku, chebyshevljevu i canberrovu, a za broj komšija 1, 5 i 10. Najveću preciznost smo dobili za euklidsku metriku kada je broj komšija 5. Preciznost tog modela nad validacionim skupom je 0.96. U tabeli 2 možemo videti kako je promena metrike uticala na preciznost. Svi modeli u tabeli su koristili broj komšija 5.

Metrika kNN modela	Procenat uspesnosti modela
Euklid	0,97
Chebyshev	0,96
Canberra	0.95

Tabela 2. Rezultati kNN modela



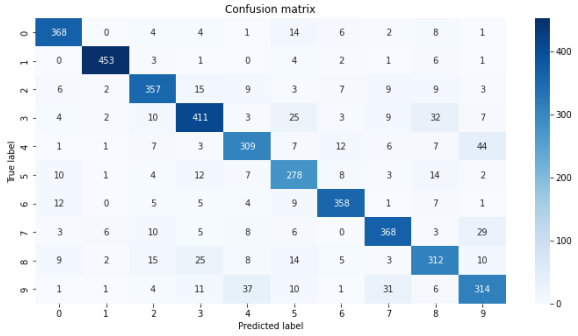
Slika 4. Matrica konfuzije za kNN

3.3. Decision Tree

Za treći tip modela smo izabrali stablo odluke. Varirali smo dubinu podele stabla i meru nečistoće podele. Za dubinu podele stabla smo varirali 25, 50 i adaptivno (model sam odlucuje dubinu). Za meru nečistoće podele koje smo koristili entropiju i Ginijev indeks diverziteta. U tabeli 3 možemo videti da promena dubine podele nije uticala u nekoj većoj meri na preciznost našeg modela. Svi modeli u tabeli su koristili entorpiju kao meru nečistoće. Preciznost nad validacionim skupom je 0.83.

Modeli stabla odluke	Procenat uspesnosti modela
Dubina = 25	0,83
Dubina = 50	0,83
Dubina = None (adaptivno)	0.83

Tabela 3. Rezultati modela stabla odluke



Slika 5. Matrica konfuzije za Decision Tree

4. Zaključak

Kada smo dobili matrice konfuzije i perciznosti naših modela možemo videti gde su se dešavale najčešće greške. Matrice konfuzije pokazuju da su modeli najčešće grešili prilikom klasifikacije broja 3, kojeg su pomešali sa brojevima 5 i 8, kao i kod broja 4, kojeg su pomešali sa brojem 9. Na slici 6 možemo da vidimo neke od brojeva koje su loše klasifikovali.

Od sva 3 modela koje smo testirali najbolje se pokazao SVM model, neznatno lošije se pokazao kNN model dok kod stabla odluke možemo videti dosta lošije rezultate naspram prethodna dva.

Slika 6. Loše klasifikovani brojevi