# Domaći 2 – klasifikacija

Dat je skup podataka koji sadrži podatke o prisustvu/odsustvu 150 određenih sastojaka za pojedine recepte koji potiču iz različitih zemalja.

**Zadatak**: Primeniti 2 različita tipa klasifikatora koji će na osnovu informacije o prisustvu/odsustvu određenih sastojaka odrediti zemlju porekla za dati recept. Ukupno ima 10566 prikupljenih recepta iz 9 zemalja i za svaki je naznačeno prisustvo (1) ili odustvo (0) svakog od 150 sastojaka.

A. Za svaku klasu analizirati u nekoj formi (po izboru) pojavljivanja određenih sastojaka. Iskomentarisati i uporediti ih za različite klase. Podatke podeliti na trening i test set.
B. Broj svog indeksa podeliti po modulu 5 - dobijeni broj označava klasifikator koji treba da se koristi:
    0. Logistička regresija
    1. kNN
    2. Neuralna mreža
    3. SVM
    4. Stablo odluke (RF)

    Uz označeni klasifikator, odabrati jedan drugi tip klasifikatora po izboru (Naivni Bayes, Logističku regresiju, kNN, Neuralnu mrežu, SVM, RF).

C. Za svaki od klasifikatora uraditi analizu u pogledu evaluacije/procene parametara:
    0. Koristeći metodu unakrsne validacije nad trening setom odrediti optimalne parametre oslanjajući se na željenu meru uspešnosti. Obratiti pažnju da u svakom od podskupova za unakrsnu validaciju bude dovoljan broj uzoraka svake klase.
    1. Za konačno odabrane parametre prikazati i analizirati matricu konfuzije dobijenu akumulacijom matrica iz svake od iteracija unakrsne validacije. Odabrati adekvatnu meru uspešnosti i potom na osnovu matrice konfuzije izračunati prosečnu vrednost za klasifikator, kao i vrednost te mere za svaku klasu.
    2. Klasifikator sa konačno odabranim parametrima obučiti i testirati. Analizirati rezultate u poređenju sa rezultatima unakrsne validacije.
D. Uporediti konačna dva klasifikatora po osetljivosti, specifičnosti i preciznosti, za svaku klasu posebno, kao i prosečnu tačnost klasifikatora, mikro i makro preciznost, osetljivost i F-meru. Iskomentarisati prednosti i mane svakog od klasifikatora.
E. Rezultate prikazati i diskutovati u izveštaju (2-4 strane).
F. **OPCIONO**: Formirati finalne klasifikatore koristeći definisane parametre i **celokupni skup.** Klasifikatore snimiti po uzoru na:
    https://scikit-learn.org/stable/modules/model_persistence.html
    Potom napisati zasebnu skriptu koja učitava klasifikatore i vrši predikciju nad csv ulazom (sa uključenom predobradom ako je korišćena, csv je u istoj formi kao i priloženi podaci). Ova skripta će biti pokrenuta nad našim **skrivenim** test skupom (za pisanje skripte može se iskoristiti segment polaznih podataka). Pojedinačni modeli na osnovu svakog od tipa klasifikatora koji budu imali najbolju F-meru na test skupu biće nagrađeni sa 2 bonus boda.

Za sva eventualna pitanja, nejasnoće ili ako smatrate da je traženo nešto što se ne može uraditi ili deluje preterano zahtevno, obratiti se mailom na ivan.lazic@uns.ac.rs ili putem MSTeams-a. Pri pisanju izveštaja pratiti uputstva koja su data. U izveštajima ne treba objašnjavati kod niti ga prepisivati, akcenat je na analizi baze, objašnjenju metoda, izboru parametara i interpretaciji rezultata. Ako je ostalo bilo šta nejasno povodom pisanja izveštaja, stojim na raspolaganju. Izdvojite dovoljno

vremena za pisanje izveštaja kako biste ga uradili kvalitetno. Putem moodla **najkasnije do 18:00 18.01.2021**. treba predati 4 fajla: skripte koja sadrže kod (.py ili .ipynb, jednu kreiranu za potrebe izveštaja i drugu sa skriveno testiranje), izveštaj (u .pdf formatu) i zipovane snimljene modele. Domaći se radi samostalno – dva ista koda ili dva ista izveštaja dobijaju 0 bodova bez daljeg istraživanja kako su nastali.