KNN – K Nearest Neighbors

* Esența algoritmului KNN
* Soluții pentru evitarea egalității (tie) in selectarea clasei
* Metoda cotului în selectarea valorii lui K
* Selectarea valorii lui K prin cross-validation
* Necesitatea scalarii caracteristicilor
* Algoritmul de elaborare a modelului KNN în Scikit-Learn:
* importul datelor
* vizualizarea informatie despre date (info())
* vizualizarea descrierii statistice a datelor (describe())
* vizualizarea numărului de clase din coloana label (value\_counts())
* vizualizarea grafica a numărul de valori in fiecare clasa a coloanei eticheta (countplot)
* vizualizarea grupării claselor in funcție de 2 caracteristici (scatterplot)
* vizualizarea poziționării și distribuției punctelor in funcție de 2 caracteristici (pairplot())
* crearea setului X si y
* crearea seturilor de train si de test
* scalarea datelor
* importul si crearea modelului KNN de bază
* ajustarea modelului pe datele de training
* realizarea predicției
* determinarea acurateti modelului (accuracy\_score())
* determinarea confusion matrix a modelului (confusion\_matrix())
* vizualizarea grafica a confusion matrix a modelului (plot\_confusion\_matrix())
* elaborarea classificasion report a modelului (classification\_report())
* Selectarea a valorii K a modelului KNN în Scikit-Learn prin metoda cotului:
* crearea unei liste goale ce va păstra valorile erorilor
* crearea unei bucle for ce permite crearea, ajustarea, predicția si determinarea erori pentru diferiți K
* afișarea grafica a dependentei erorii de valoarea k
* selectarea valoare minima a lui K ce asigura o valoarea minima a erorii
* Necesitatea pipeline
* Selectarea a valorii K a modelului KNN în Scikit-Learn prin cross validation:
* crearea unui model KNN implicit
* crearea unui obiect de scalare implicit
* crearea unei liste cu tupluri cu doua elemente - denumirea (string) si obiectul procesului ce se vor include in pipeline
* importul si crearea unui obiect pipeline
* crearea unei liste cu valorile lui K
* crearea unui dicționar cu hiperparametrii ce vor fi variați in obiectul pipeline (cheile fac referința si la procesele din pipeline separate prin \_\_)
* crearea modelului gridsearch
* ajustarea modelului pe datele de training fără scalare
* verificarea celor mai buni hiperparametri ai pipeline
* realizarea predicției pe datele de test fără scalare
* vizualizarea grafica a confusion matrix a modelului
* elaborarea classification report a modelului