Natural Language Processing Homework nr. 5

Even during the last workshop we showed you how to create a Machine Learning Model and solve it using a problem from out day-to-day life, now you are going to solve a pure Machine Learning problem, not one from NLP. This will help you to better understand the Machine Learning field and it's utility in the real world tasks.

Now you are going to build a model that will try to diagnose if it has or not a heart disease. Data is just a bunch of medical records from patients with and without heart disease, like: blood pressure or age and the target column which the model should predict.

link to the data set - UCI Heart Disease UCI.

Task 1. Create a model that will diagnose a heart disease.

1. Import the data:
   Import the data using read_csv() function implemented in pandas library.
2. Analyse the data:
   Analyse data using such methods as info() and describe() applied on your data set (df) - df.info(), df.describe().
3. Eliminate the outliers:
   IsolationForest is a model for detecting outliers. Import it and create an object of this algorithm, fit it on your data (don't forget to convert it to a numpy array) then use it to predict whether it's an outlier or not. After that you will get an array of 1 (ones) and -1 (minus ones), 1 meaning a normal sample, while -1 means an outlier, which you should eliminate.
4. Choose some classification algorithms:
   We strongly recommend you to try such algorithms as LogisticRegression, RandomForest and SVM, but feel free to try others too.
5. Train the model:
   Create an object of the algorithm class, fit the data, and predict whatever the patients are ill or not. Don't forget to split the data into train and test subset, and shuffle the rows before the split.
6. Evaluate the model:
   Look at the accuracy and confusion matrix, and write down the conclusions.

Chiar și în timpul ultimului workshop v-am arătat cum să vă creați propriul model și să încercați să rezolvați o problemă de bază care apare în viața noastră de zi cu zi. Deci, acum, va trebui să vă confruntați cu o problemă pură de Machine Learning, nu este legată de NLP, dar vă va ajuta să înțelegeți mai bine ideea de bază a ML și utilitatea sa în sarcinile lumii reale.
Pentru astăzi veți încerca să aflați dacă un pacient are sau nu boli de inimă. Datele constau dintr-o grămadă de caracteristici care descriu fiecare pacient, de exemplu: tensiunea arterială sau vârsta și coloana target pe care trebuie să o preziceți.

link pentru dataset - UCI Heart Disease UCI

Task 1. Creați un model care să diagnostice bolile cu inima.

1. **Importează datasetul**
   Importează datasetul prin read_csv functia implementată în librăria pandas.
2. **Analizează datele**
   Puteți utiliza metode precum info() și describe() aplicate pe datasetul vostru (df) - df.info(), df.describe().
3. **Scoate valorile aberante**
   IsolationForest este un model de detectare anormal, care ajută la găsirea valorilor aberante în date. Importați și apoi creați o instanță a clasei de algoritm, faceți fit() pe datele voastre în mod specific pe versiunea matricei numpy a datasetului și apoi preziceți valorile aberante print predict(), după ce faceți prezicerea, veți primi doua tipuri de valori, 1 și -1, 1 reprezintă exemplul normal și -1 reprezintă valoarea aberantă, după ce cunoașteți care sunt valorile aberante, trebuie să le scoateți.
4. **Alege câteva algoritmi de clasificare**
   Algoritmii de bază pentru sarcinile de clasificare ar fi LogisticRegression, RandomForest, SVM, dar poți liber să încerci și altele.
5. **Crează modelul**
   Creați instanțe de clasă de algoritmi, faceți fit și preziceți target-ul pentru datele de testare. Desigur nu uita să amesteci rândurile datasetului și să-l împarți în setul de antrenare și testare.
6. **Evaluează modelul**
   Vedeți ce precizie și matrice de confuzie vă oferă fiecare model și trageți concluzii.