Tubes I Bagian A: Eksplorasi library Decision Tree Learning pada Jupyter Notebook

Tugas besar 1 ini dikerjakan berkelompok, terdiri atas 4 mahasiswa (boleh gabungan mahasiswa K1, K2, dan K3). Terdapat 4 bagian yang akan dikerjakan, dengan deadline yang berbeda.

Lakukan eksplorasi scikit-learn pada Jupiter Netbook dan bacalah dokumentasinya:https://jupyter-notebook.readthedocs.io/en/stable/notebook.html dan http://scikit-learn.org/stable/documentation.html. Jupyter Notebook (http://jupyter.org/) memudahkan kita untuk membuat dan men-share dokumen yang merupakan gabungan dari live code, equation, visualisasi dan catatan. Jupyter dapat digunakan untuk visualisasi, pembersihan dan data transformasi, statistical model dan machine learning. Scikit-learn merupakan library machine learning pada bahasa python.

1. Tulislah script dalam bahasa python pada satu notebook untuk melakukan task berikut ini:
    1. Membaca dataset standar iris dan dataset play-tennis (dataset eksternal dalam format csv). Gunakanlah sklearn.datasets untuk membaca dataset iris. Untuk membaca dataset csv, gunakanlah Python Data Analysis Library http://pandas.pydata.org/.
    2. Melakukan pembelajaran DTL dengan DecisionTreeClassifier (http://scikit-learn.org/stable/modules/tree.html), dan Id3Estimator (https://github.com/svaante/decision-tree-id3) untuk dataset iris dan play-tennis dengan memanggil method fit(data,target) untuk semua data (full training), dan menampilkan modelnya.
        1. Jika diperlukan encoding data kategorikal, gunakanlah library LabelEncoder.
        2. Untuk menampilkan model pohonnya, gunakan method export_text.
2. Pelajari algoritma pada hal 56 buku Machine Learning Tom Mitchell, lalu carilah persamaan dan perbedaan algoritma tersebut dengan kedua library DecisionTreeClassifier dan Id3Estimator dalam hal berikut ini. *Jawaban tetap dituliskan dalam notebook yang akan dikumpulkan*.
    1. Penentuan atribut terbaik
    2. Penanganan label dari cabang setiap nilai atribut
    3. Penentuan label jika examples kosong di cabang tersebut

    4. Penanganan atribut kontinu
    5. Penanganan atribut dengan missing values
    6. Pruning dan parameter confidence
3. Tugas dikumpulkan berupa hasil download notebook dalam dua format yaitu file .ipynb dan pdf. Hanya salah satu anggota kelompok saja yang *uplad* file tugas pada website kuliah ini. Penamaan file yang dikumpulkan: Tubes1A_[NIM salah satu anggota].zip.
4. Pengumpulan yang terlambat tidak diperbolehkan, batas akhir adalah hari Senin, 3 Februari 2020 jam 06.00 pagi (waktu situs kuliah ini).

Tubes I Bagian B: Implementasi Decision Tree Learning

Tugas besar 1 ini dikerjakan berkelompok, terdiri atas 4 mahasiswa (boleh gabungan mahasiswa K1, K2, dan K3). Terdapat 4 bagian yang akan dikerjakan, dengan deadline yang berbeda.

1. Implementasi modul myID3 (sesuai hal 56), dan myC45 (isu2 DTL) sesuai buku Machine Learning Tom Mitchell.
    1. Overfitting training data dengan post pruning. Gunakanlah 20% training data untuk data validasi.
    2. Continuous-valued attribute: information gain dari kandidate.
    3. Alternative measures for selecting attributes: gain ratio.
    4. Handling missing attribute value: most common target value.
2. Lakukan pembelajaran DTL dengan myID3 dan myC45 untuk dataset iris dan play-tennis untuk semua data (full training), dan menampilkan modelnya.
3. Deliverables: a) source code, b) laporan berisi penjelasan implementasi, hasil eksekusi (langkah 2), perbandingan dengan hasil DTL sklearn dan Id3Estimator (bagian A), dan pembagian tugas setiap anggota kelompok.
4. Hanya salah satu anggota kelompok saja yang *uplad* file tugas pada website kuliah ini. Penamaan file yang dikumpulkan: Tubes1B_[NIM salah satu anggota].zip.
5. Pengumpulan yang terlambat tidak diperbolehkan, batas akhir adalah hari Jumat, 14 Februari 2020 jam 06.00 pagi (waktu situs kuliah ini).