Minta ZH 2 – Big Data Architektúrák és elemző módszerek

A feladat osztályozási modellt készíteni az alábbi paraméterek alapján:

- Készítsünk osztályozást arra, hogy az adott sorban a dolgozó túlórázott-e?
- Használd a 'hr-employee.csv' filet inputnak!
    - (elérés: http://ggombos.web.elte.hu/oktatas/BigDataArchitekturaEsElemzo/GY/zh/)

1. Az adathalmazból a következő oszlopokat használd: (1 pont)
    a. Age, Attrition, DistanceFromHome, OverTime, TotalWorkingYears, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole
2. Készíts egy új oszlopot, aminek a neve 'VeryOvertime' legyen és a következő képpen számolódjon ki: (1 pont)
    a. OverTime alapján, ha yes akkor 1, különben 0.
3. A model elkészítéséhez az adathalmaz train és test halmazra a következő képpen válaszd szét:
    a. 80/20 vagy Cross-validációt alkalmazva
4. A karakteres oszlopokat alakítsd át megfelelő formátumba! (1 pont)
5. A hiányzó adatokat helyére rakd az oszlop átlagát vagy mediánját! (1 pont)
6. A következő feladatok közül választhatsz, hogy melyiket valósítod meg: (3 pont)
    a. Használod a 'DecisionTree' és 'Naive-Bayes' algoritmusokat, amelyeket betanítva megmutatod melyik ad jobb eredményt.
    b. Használod a 'SVM' és 'RandomForest' algoritmusokat, amelyeket betanítva megmutatod melyik ad jobb eredményt.
    c. Használod a Bagging módszert és legalább 2 paraméter változtatásával megmutatod, hogy jobb eredményt lehet elérni.
7. Plotolási feladatok: (4 pont)
    a. Hisztogram segítségével rakd ki a következő oszlopot: 'Age'!
    b. Hisztogram segítségével rakd ki a következő oszlopot: 'DistanceFromHome'!
    c. Pivot table segítségével rakd ki a következő oszlopok közötti összefüggést: 'YearsAtCompany, OverTime'!
    d. Scatter plot segítségével rakd ki a következő oszlopok közötti összefüggést: 'YearsAtCompany, Age'!