Klassifitseerimisülesanne: tüüpide tuvastamine

Andmete ettevalmistamine
Koodi aluseks võtsin faili Klassifitseerimine_naide_b.py. Esmalt eemaldasin mitmeid veerge, sealhulgas 'ID', 'TäisAadress', 'Maakond', 'KOV_voi_LinnaOsa', 'Tänav_Hoone_Nr', 'Korgus_Umardatud', 'MaaAluste_Korruste_Arv', kuna need andmed olid liialt spetsiifilised või olid juba mujal esindatud. Näiteks 'ID' on ainult unikaalne identifikaator, mis ei paku klassifitseerimiseks olulist teavet. "TäisAadress", "Maakond", "Tänav_Hoone_Nr" või muud sarnased veerud ka võivad olla vähem olulised, seega nad võivad tekitada ainult müra. "Korgus" ja "Korgus_Umardatud", võib olla mõistlik eelistada ühte ja teine eemaldada.

Valisin ennustavaks muutujaks veeru "Energy_Certificate_Class". See klassifikatsioon annab ülevaate hoone energiatõhususe tasemest, mis on oluline keskkonna- ja ressursisäästlikkuse seisukohast ning oluline ehitustööstuses ja kinnisvarasektoris. Samuti, hoone energiatõhususe klass mõjutab otseselt hoone väärtust.

Eksperimenteerimine klassifitseerimismeetoditega
Kasutasin mitut klassifitseerimismeetodit, sealhulgas RandomForestClassifier, DecisionTreeClassifier ja KNeighborsClassifier.

RandomForestClassifier ja KNeighborsClassifier saavutasid täpsuse 0.3917525773195876, mis näitab, et need mudelid olid piisavalt head, kuid mitte väga täpsed.
DecisionTreeClassifier oli vähem täpne, umbes 0.26.
Proovisin ka samu klassifikatsioonimeetodeid teiste veergudega nagu "Maksimaalne_Korruste_Arv" (täpsus: 0.86) ja "Ruumide_Arv" (täpsus: 0.41). Tulemused näitasid, et lihtsamad tunnused nagu "Maksimaalne_Korruste_Arv" ja "Ruumide_Arv" andsid suurema täpsuse kui keerulisemad tunnused nagu "Energy_Certificate_Class".

Proovisin kasutada Weka tarkvara, et kontrollida tulemused, kuid ei saanud hästi aru kuidas seda kasutada. Weka tundub potentsiaalselt sobivat antud ülesannete jaoks, kuid vajab põhjalikumat õppimist ja seadistuste muutmist, et optimeerida väljundit. Kuid aga klassifitseerimismudeli ise Pythoni abil loomine, minu meelest, võimaldab rohkem kontrolli mudeli parameetrite üle.

Järeldus
Järeldusena võib öelda, et lihtsamad ja otstarbekamad veerud nagu "Maksimaalne_Korruste_Arv" pakuvad paremat klassifitseerimise täpsust võrreldes keerulisemate tunnustega nagu "Energy_Certificate_Class". Oluline on valida ennustamiseks sobivaimad tunnused, mis pakuvad parimat tasakaalu täpsuse ja mudeli keerukuse vahel. Samuti võib erinevate klassifitseerimismeetodite võrdlemine anda parema ülevaate andmete struktuurist

ja sobivusest. Samuti klassifitseerimise jaoks võib olla vajalik täpsem lähenemine või täiustatud masinõppe mudel, et eristada erinevaid materjale ja nende omadusi palju paremini.

Õppisin selle harjutuse käigus mitmeid olulisi aspekte:

- Andmete ettevalmistamine: Oluline on teha eeltööd, et valida sobivad veerud ja eemaldada mittevajalikud, mis võivad mudeli täpsust mõjutada.
- Tunnuste valik: Ennustava muutuja valikul tuleb arvestada selle olulisust ja mõju analüüsitavale probleemile. Näiteks veerg "Energy_Certificate_Class" osutus keerulisemaks ennustada võrreldes lihtsamate tunnustega nagu "Maksimaalne_Korruste_Arv".
- Klassifitseerimismeetodite kasutamine: Eksperimenteerides erinevate klassifitseerimismeetoditega, sain aimu nende tugevustest ja nõrkustest erinevates olukordades.
- Täpsuse hindamine: Mudeli täpsuse hindamine on oluline, et valida parim mudel konkreetse andmekogumi jaoks.