

Mašīnmācīšanās: Sirds imputācijas

Problēmas apraksts

Lemon AI noslēdza līgumu ar prestižo Rīgas Kardioloģijas centru. Ivana uzdevums bija sakārtot slimnīcas vēsturiskos pacientu datus, lai apmācītu modeli, kas var prognozēt sirds slimības klātbūtni.

Tomēr slimnīcas datubāzes serveris bija ārkārtīgi vecs un kļūdainš. Datu eksportēšanas laikā aptuveni 10% vērtību tika bojātas un zaudētas (aizstātas ar `NaN`).

Turklāt galvenā kardioloģe bija ļoti stingra: viņa uzskatīja, ka nespēja konstatēt pacientu ar sirds slimību ir **3 reizes sliktāk** nekā kļūdains diagnosticēt veselam cilvēkam.

"Ja palaidīsim garām vienu slimu pacientu, tās ir dzīvības risks," viņa izskaidroja Ivanam. "Viltus pozitīvais rezultāts vienkārši nozīmē vairāk pārbažu. Uzskatu, ka viltus negatīvais rezultāts (false negative) ir 3 reizes sliktāks nekā viltus pozitīvais (false positive)."

Ivanam bija jāsakārto bojātie dati un jāizveido modelis, kas atbilst galvenās kardioloģes stingrajām drošības prasībām. Šis bija viņa grūtākais izaicinājums līdz šim — problēma ar reālām sekām.

Uzdevums

Jums ir dota pacientu veselības rādītāju datu kopa:

- `X_train.csv`: Pazīmes (features) apmācības kopai. Satur trūkstošas vērtības.
- `y_train.csv`: Mērķa apzīmējumi (target labels) apmācības kopai (`1` = sirds slimība, `0` = nav sirds slimības).
- `X_test.csv`: Pazīmes testa kopai. Satur trūkstošas vērtības.

Jūsu uzdevums ir prognozēt mērķi testa kopai.

Norādes

- Priekšapstrāde: Apskatiet `sklearn.impute`, lai apstrādātu `NaN` vērtības CSV failos.
- Optimizācija: Pievērsiet īpašu uzmanību vērtēšanas formulai.
- Var noderēt šis ceļvedis no scikit-learn: <https://scikit-learn.org/stable/modules/impute.html>

Vērtēšana

Iesniegums tiek vērtēts, pamatojoties uz svērto precizitātes metriku. Formula:

$$Score = \frac{3 \cdot TP + TN}{3 \cdot (TP + FN) + (TN + FP)}$$

Kur:

- **TP (True Positive, patiesi pozitīvs)**: Prognozējāt slimību, un viņiem patiesām tā ir. (Svērts x3)
- **TN (True Negative, patiesi negatīvs)**: Prognozējāt veselu, un viņi patiesām ir veseli.
- **FP (False Positive, viltus pozitīvs)**: Prognozējāt slimību, bet viņi ir veseli.
- **FN (False Negative, viltus negatīvs)**: Prognozējāt veselu, bet viņiem patiesām ir slimība. (Tas vissvairāk kaitē rezultātam, jo samazina skaitītāja augstas vērtības locekli).

Būtībā rezultāts reprezentē precizitāti uz datu kopu, kurā katrs pozitīvs pacients ir dublēts 3 reizes.

Gala punktu skaits par uzdevumu ir $(\text{score} - \text{baseline}) / (\text{max_score} - \text{baseline})$.