

## Student Performance Data Set (Tanulói teljesítmény-adatsor)

Ez az adatsor két portugál középiskola tanulóinak teljesítményét mutatja. Az érdemjegy előrejelzése a feladat a megadott attribútumok, prediktorok alapján. Az adatattribútumok közé tartoznak a tanulók demográfiai, szociális és iskolával kapcsolatos jellemzői. Az adatgyűjtés iskolai jelentések és kérdőívek segítségével történt.

### Az eredeti forrás a következő:

<http://archive.ics.uci.edu/ml/datasets/Student+Performance>

### Adatkészlet információ:

# A "Student-Mathematics.mwx (matematika tantárgy) és "Student-Portuguese.mwx" (portugál nyelvi tantárgy) adathalmazok attribútumai:

1 iskola - a tanuló iskolája (bináris: "GP" - Gabriel Pereira vagy "MS" - Mousinho da Silveira).

2 sex - a tanuló neme (bináris: "F" - nő vagy "M" - férfi)

3 életkor - a tanuló életkora (numerikus: 15-től 22-ig)

4 cím - a hallgató lakcímének típusa (bináris: "U" - városi vagy "R" - vidéki)

5 famsize - családméret (bináris: "LE3" - legfeljebb 3 vagy "GT3" - 3-nál nagyobb)

6 Pstatus - a szülők együttélési státusza (bináris: "T" - együtt élő vagy "A" - külön élő).

7 Medu - az anya iskolai végzettsége (numerikus: 0 - nincs, 1 - általános iskolai végzettség (4. osztály), 2 - 5-9. osztály, 3 - középfokú végzettség vagy 4 - felsőfokú végzettség).

8 Fedu - az apa iskolai végzettsége (numerikus: 0 - nincs, 1 - általános iskolai végzettség (4. osztály), 2 - 5-9. osztály, 3 - középiskolai végzettség vagy 4 - felsőfokú végzettség).

9 Mjob - az anya munkája (nominális: "tanár", "egészségügyi" ellátással kapcsolatos, polgári "szolgáltatások" (pl. közigazgatási vagy rendőrségi), "otthon" vagy "egyéb").

10 Fjob - az apa foglalkozása (névleges: "tanár", "egészségügyi" ellátással kapcsolatos, polgári "szolgáltatások" (pl. közigazgatási vagy rendőrségi), "otthon" vagy "egyéb")

11 ok - az iskolaválasztás oka (névleges: közel az "otthonhoz", az iskola "hírneve", a "kurzus" preferenciája vagy "egyéb").

12 gyám - a tanuló gyámja (névleges: "anya", "apa" vagy "egyéb")

13 utazási idő - az otthonról az iskolába történő utazás ideje (numerikus: 1 - <15 perc, 2 - 15-30 perc, 3 - 30 perc és 1 óra között, vagy 4 - >1 óra).

- 14 studytime - heti tanulási idő (numerikus: 1 - <2 óra, 2 - 2-5 óra, 3 - 5-10 óra, vagy 4 - >10 óra)
- 15 failures - a korábbi órák sikertelenségeinek száma (numerikus: n, ha  $1 \leq n < 3$ , egyébként 4)
- 16 schoolsup - extra oktatási támogatás (bináris: igen vagy nem)
- 17 famsup - családi oktatási támogatás (bináris: igen vagy nem)
- 18 paid - extra fizetett órák a tanfolyam tárgyán belül (matematika vagy portugál) (bináris: igen vagy nem)
- 19 activities - tanórán kívüli tevékenységek (bináris: igen vagy nem)
- 20 óvoda - óvodába járt (bináris: igen vagy nem)
- 21 felsőoktatás - felsőoktatásban akar részt venni (bináris: igen vagy nem)
- 22 internet - otthoni internet-hozzáférés (bináris: igen vagy nem)
- 23 romantikus - romantikus kapcsolatban él (bináris: igen vagy nem)
- 24 famrel - családi kapcsolatok minősége (numerikus: 1-től - nagyon rossz 5-ig - kiváló)
- 25 freetime - szabadidő az iskola után (numerikus: 1-től - nagyon alacsonyan - 5-ig - nagyon magasan)
- 26 goout - barátokkal való együttlét (numerikus: 1-től - nagyon alacsony szintig 5-ig - nagyon magas szint)
- 27 Dalc - munkanapi alkoholfogyasztás (numerikus: 1 - nagyon alacsony és 5 - nagyon magas között)
- 28 Walc - hétvégi alkoholfogyasztás (numerikus: 1 - nagyon alacsony és 5 - nagyon magas között)
- 29 egészség - jelenlegi egészségi állapot (numerikus: 1 - nagyon rossz és 5 - nagyon jó között)
- 30 hiányzások - iskolai hiányzások száma (numerikus: 0-tól 93-ig)

# Ezek a jegyek a válaszváltozók, és a matematikához vagy a portugálhoz kapcsolódnak:

- 31 G1 - az első időszak jegye (numerikus: 0-tól 20-ig)
- 31 G2 - a második időszak jegye (numerikus: 0-tól 20-ig)
- 32 G3 - záró jegy (numerikus: 0-tól 20-ig, a válaszváltozó)

### **Elvégzendő feladatok:**

1.

Készítsen korrelogramot a matematika és a portugál tantárgyak numerikus prediktorai között.

Melyik prediktor korrelál a legerősebben a válaszváltozóval?

2.

A Students-Matematics munkalap adatainak felhasználásával készítsen néhány diagramot, amelyekből megtudhatja, hogy mely kategorikus változó (vagy a Minitabban kategorikus változóként is értelmezhető numerikus változó) értékei lehetnek meghatározóak a G3 regressziós fájának megalkotásakor.

3.

A matematika G3-as osztálya esetében konstruáljon regressziós fát az összes numerikus és kategorikus változót felhasználva a fa konstrukciójában. Használja az összes megfigyelést képzési adathalmazként.

Mi lesz az optimális modell ebben az esetben?

4.

Ismételje meg a fa növesztését, de ezúttal távolítsa el a G1 és G2 változókat a prediktorok közül.

Milyen változást tapasztal?

5.

A Students-Portugál adatfájlhoz építsen regressziós fát az összes prediktor (beleértve a G1 és G2 változót is) felhasználásával, a teljes adatfájlt használva képzési adatfájlként.

Mit tapasztal?

6.

Most nézze meg, hogy mennyit változik a modell illeszkedési jósága 10-szeres keresztvalidálás esetén.

7.

A matematika esetében nézze meg a 10-szeres kereszt-validálás eredményeit is.

8.

Távolítsa el a G1 és G2 változókat a prediktorok közül.

Mit tapasztal?

9.

A Students-Portuguese adatfájlhoz hozzon létre egy regressziós fát az összes prediktor (beleértve a G1 és G2 változót is) felhasználásával, 10-szeres kereszt-validálást alkalmazva.

10.

Ugyanerre az adatállományra állítson fel egy regressziós fa modellt úgy, hogy csak a G1 és G2 változókat használja prediktorokként.

Csökkenet vagy nőtt az R-négyzet?

Hogyan lehetséges ez?