

MA012 – zadání projektu

S pomocí *R* vyřešte níže zadané úkoly. Řešení odevzdejte zpracované ve formátu PDF (např. exportem z jakéhokoliv textového editoru) do Odevzdávárny v ISu a přiložte i kód pro *R*. V PDF dokumentu u jednotlivých úkolů zdůvodněte volbu metody, popište svůj postup, uveďte výsledky a grafy (kde je to možné) a svůj závěr okomentujte. Nezapomeňte na kontrolu podmínek pro dané metody.

Do Odevzdávárny vložte své řešení nejpozději 2 dny před termínem Vaší zkoušky. Po písemné části zkoušky bude následovat krátká ústní část nad Vaším projektem o použitých metodách, způsobu řešení a dosažených výsledcích.

I.

Pracujte se souborem `nadory.csv` ze studijních materiálů v ISu. Soubor obsahuje anonymizované lékařské údaje.

S využitím vhodných statistických metod – Pearsonův korelační koeficient, Spearmanův korelační koeficient, jednoduché třídění, anebo dvojné třídění – vyřešte následující úkoly:

1. Zjistěte, zda povrch těla pacientů (proměnná `Povrch`) závisí na hmotnosti pacienta (proměnná `Hmotnost`).
2. Zjistěte, zda povrch těla pacientů závisí na věku pacienta (proměnná `Vek`).
3. Zjistěte, zda povrch těla pacientů závisí na pohlaví pacienta (proměnná `Pohlavi`).
4. Zjistěte, zda povrch těla pacientů závisí na typu nádoru (proměnná `Nador`) a pohlaví pacienta.
5. Zjistěte, zda povrch těla pacientů závisí na typu nádoru a na tom, zda pacient bydlí ve městě, anebo na vesnici (proměnná `Obec`).
6. Zjistěte, zda povrch těla pacientů závisí na okresu (proměnná `Okres`), ze kterého pochází; berte přitom v úvahu jen ty okresy, ze kterých bylo alespoň 24 pacientů.
7. Zjistěte, zda povrch těla pacientů závisí na okresu, ze kterého pochází, a zda pacient bydlí ve městě, anebo na vesnici; berte přitom v úvahu jen ty okresy, ze kterých bylo alespoň 24 pacientů.

II.

Pracujte se souborem `titanic.csv` ze studijních materiálů v ISu. Soubor obsahuje údaje o počtech přeživších (`zivi`), mrtvých (`mrtvi`) a celkových počtech pasažérů (`celkem`) na lodi Titanic, v závislosti na třídě (`trida`: 1., 2., 3., anebo `posadka`), pohlaví (`pohlavi`) a věku (`vek`: `dite`, anebo `dosp` `dospely`).

Pomocí logistické regrese modelujte pravděpodobnost přežití (= úspěch) pomocí tří regresorů (`trida`, `pohlavi`, `vek`). Vyzkoušejte postupně do modelu zařazovat/nezařazovat tyto regresory. Takto naleznete kombinaci regresorů, které jsou pro modelování pravděpodobnosti přežití významné. Ve které z kategorií regresorů bylo vhodné se nacházet z pohledu vyšší šance na přežití?