

Témata projektů z M8DM1

Datový soubor Adult

1. Predikujte, zda příjem dané osoby překročí 50 000 USD. Data nejprve rozdělte na trénovací a testovací a poté váš model vyzkoušejte na testovacích datech.

Datový soubor Adult2

2. Predikujte, zda příjem dané osoby překročí 50 000 USD. Data nejprve rozdělte na trénovací a testovací a poté váš model vyzkoušejte na testovacích datech.

Datový soubor Airline

3. Které faktory ovlivňují celkovou spokojenost zákazníků leteckých společností?
4. Pro daného zákazníka predikujte, zda bude spokojen s celkovými službami letecké společnosti. Data nejprve rozdělte na trénovací a testovací a poté váš model vyzkoušejte na testovacích datech.

Datový soubor Auto

5. Predikujte spotřebu auta v závislosti na zadaných veličinách. Na kterých z nich nejvíce závisí? Data nejprve rozdělte na trénovací a testovací a poté váš model vyzkoušejte na testovacích datech.
6. Která auta jsou si nejvíce podobná? Která se naopak nejvíce liší? Dají se mezi auty nalézt nějaké charakteristické skupiny aut?

Datový soubor Boston

7. Predikujte cenu domu (medián) v Bostonu na základě socio-ekonomických a dalších ukazatelů. Na jakých atributech závisí? Data nejprve rozdělte na trénovací a testovací a poté váš model vyzkoušejte na testovacích datech.
8. Predikujte koncentraci oxidů dusíku v různých částech Bostonu na základě socio-ekonomických a dalších ukazatelů. Na jakých atributech závisí? Data nejprve rozdělte na trénovací a testovací a poté váš model vyzkoušejte na testovacích datech.

Datový soubor Cars93

9. Která auta jsou si nejvíce podobná? Která se naopak nejvíce liší? Dají se mezi auty nalézt nějaké charakteristické skupiny aut?

Datový soubor Estonia

10. Které faktory ovlivnily to, zda člověk přežil potopení lodi Estonia?
11. Pro daného pasažéra predikujte, zda přežil nebo ne.

Datový soubor Flags

12. Na základě údajů o vlajce dané země predikujte nejrozšířenější náboženství v zemi.
13. Vlajky kterých zemí jsou si nejvíce podobné? Dají se nalézt skupiny zemí, které mají podobné vlajky?

Datový soubor German

14. Najděte faktory ovlivňující to, zda bude klient splácet půjčku.

Datový soubor Internet

15. Jaká je situace v ČR z hlediska využívání internetu? Jaké služby nejčastěji lidé využívají? Které služby jsou používány současně?
16. Popište skladbu lidí a charakteristické skupiny, které využívají internet a jednotlivé jeho služby. Můžeme mezi nimi pozorovat nějaké rozdíly?
17. Liší se využívání internetových služeb mezi muži a ženami? Mezi bohatšími a chudšími? Napříč věkovými kategoriemi? Podle jiných aspektů?

Datový soubor Kabelovka

18. Jaké je využívání TV a různých technologií příjmu signálu v ČR? Liší se nějakým způsobem podle socio-demografických charakteristik?
19. Jak vypadá častý (pravidelný) divák TV? Jak se liší od běžného diváka?
20. Nová televizní stanice chce vstoupit na český trh. Jak by se měla profilovat, aby odpovídala častým a typickým potřebám české populace?
21. Analyzujte uvedené výroky. Dají se vypořádat nějaké vztahy mezi jednotlivými výroky? Jak souvisí se sledováním TV?
22. Analyzujte uvedené výroky v závislosti na socio-demografických charakteristikách.

Datový soubor Mazlicci

23. Jaká je situace v ČR z pohledu domácích mazlíčků? Kde pro ně lidé nejčastěji nakupují krmivo? Dají se nalézt nějaké zákonitosti, ve kterých obchodech lidé nakupují?
24. Kdo je typický chovatel? Liší se nějak chovatelé koček od chovatelů psů?
25. Můžeme v datech identifikovat skupiny osob, které mají podobné nakupovací návyky? Čím se mezi sebou odlišují?

Datový soubor Postoje

26. Analyzujte výroky = charakteristiky dosavadního života. Jaké výroky jsou si podobné a jaké jsou naopak opačné? Lze výroky zredukovat do menšího počtu tvrzení?
27. Analyzujte výroky = charakteristiky dosavadního života. Dají se vypořádat nějaké vztahy mezi jednotlivými výroky?
28. Najděte v populaci skupiny lidí s podobnými postoji podle analyzovaných výroků.
29. Lze vysledovat nějaký vztah mezi výroky a demografickými charakteristikami?

Datový soubor Saheart

30. Odhalte rizikové faktory pro onemocnění srdce.

Datový soubor Sport

31. Popište využívání různých druhů sportu v ČR. Dají se v populaci vypořádat nějaké vztahy mezi provozovanými sporty?
32. Existuje rozdíl mezi sportováním bohatších a chudších občanů? Jak se vůbec bohatší lidé odlišují od běžné populace?
33. Existují v populaci nějaké typické skupiny věnující se určitým sportům?

Datový soubor StudentPerformance

34. Jak se liší studenti školy Gabriel Pereira (GP) od studentů školy Mousinho da Silveira (MS)?
35. Můžeme v datech najít nějaké typické skupiny studentů? Jsou tyto skupiny stejné na obou školách?
36. Existuje souvislost mezi konzumací alkoholu a sociálními a školními atributy?

Datový soubor Tips

37. Na čem záleží velikost dýška? Jaký zákazník dá největší a jaký nejmenší dýško?

Datový soubor VolnyCas

38. Jaký podíl volného času věnuje průměrný člověk různým činnostem? Liší se v trávení volného času muži od žen? Jakou roli hrají další ukazatele?
39. Liší se v nějakých aspektech lidé trávící různě svůj volný čas? Lze nějak popsat typického jedince věnujícího se různým volnočasovým aktivitám?

40. Existuje nějaký rozdíl, jak lidé tráví svůj volný čas ve všední a volný den?
U kterých jedinců jsou rozdíly nejmenší (největší)?

Datový soubor Wholesale

41. Na základě útrat za jednotlivé produkty najděte skupiny podobných klientů jistého velkoobchodu.
42. Na základě dostupných informací rozhodněte, zda se klient daného velkoobchodu věnuje maloobchodu nebo je subdodavatelem hotelu/kavárny/restaurace.

Datový soubor Wine

43. Na základě výsledků chemické analýzy odhalte různé typy vína. Která vína jsou si nejvíce podobná? Která se naopak nejvíce liší? Čím?
44. Na základě výsledků chemické analýzy predikujte obsah alkoholu v daném víně. Na čem závisí?