

Témata projektů z M8DM1

Datový soubor Adult

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

Datový soubor Airline

2. Které letecké společnosti jsou nejbezpečnější? A které naopak nejvíce nebezpečné?

Datový soubor Auto

3. Predikujte spotřebu auta v závislosti na zadaných veličinách. Data nejprve rozdělte na tréninková a testovací a poté Váš model vyzkoušejte na testovacích datech.
4. Která auta jsou si nejvíce podobná? Která se naopak nejvíce liší? Dají se mezi auty nalézt nějaké charakteristické skupiny aut?

Datový soubor Bakalari

5. Rozhodněte, jestli je ve všech regionech Francie stejná skladba bakalářských studentů. Kde se tato skladba nejvíce odlišuje od průměru? Jak?

Datový soubor Banknotes

6. Podle výsledků měření rozhodněte, která bankovka je pravá a která falešná.

Datový soubor Boston

7. Odhadněte cenu domu (medián) v Bostonu na základě socio-ekonomických ukazatelů. Data nejprve rozdělte na tréninková a testovací a poté Váš model vyzkoušejte na testovacích datech.
8. Odhadněte koncentraci oxidů dusíku v různých částech Bostonu na základě socio-ekonomických ukazatelů. Data nejprve rozdělte na tréninková a testovací a poté Váš model vyzkoušejte na testovacích datech.

Datový soubor Cars93

9. Dá se pro každého výrobce najít typický vůz, který vyrábí? Jak se různí výrobci liší?
10. Která auta jsou si nejvíce podobná? Která se naopak nejvíce liší? Dají se mezi auty nalézt nějaké charakteristické skupiny aut?

Datový soubor CMC

11. Na základě dat odhadněte druh antikoncepce, který daná žena užívá. Data nejprve rozdělte na tréninková a testovací a poté Váš model vyzkoušejte na testovacích datech.
12. Najděte v datech skupiny žen, které jsou si podobné. Používají stejnou antikoncepci?

Datový soubor Companies

13. Dá se na základě ekonomických ukazatelů firmy rozhodnout, jakým odvětvím se firma zabývá?

Datový soubor Dermatology

14. Na základě symptomů a rodinné historie najděte pacienty s podobným onemocněním.

Datový soubor Fertility

15. Identifikujte rizikové faktory neplodnosti u mužů.

Datový soubor Flags

16. Na základě údajů o zemi a její vlajce predikujte nejrozšířenější náboženství v zemi. Data nejprve rozdělte na tréninková a testovací a poté Váš model vyzkoušejte na testovacích datech.
17. Vlajky kterých zemí jsou si nejvíce podobné? Dají se nalézt skupiny zemí, které mají podobné vlajky?

Datový soubor Geopol

18. Které země jsou si podobné z hlediska ekonomicko-politického? Které se nejvíce liší?
19. Najděte v datech skupiny zemí, které jsou si podobné z hlediska ekonomicko-politického.

Datový soubor Health

20. Porovnejte státy USA z hlediska zdraví. Které státy jsou na tom podobně?
21. Souvisí geografická poloha státu se zdravím?

Datový soubor Heart

22. Na základě údajů o pacientech rozhodněte, jestli daná osoba trpí onemocněním srdce. Data nejprve rozdělte na tréninková a testovací a poté Váš model vyzkoušejte na testovacích datech.

Datový soubor Internet

23. Jaká je situace v ČR z hlediska využívání internetu? Jaké služby nejčastěji lidé využívají?
24. Popište skladbu a charakteristické skupiny, které využívají internet a jednotlivé jeho služby.
25. Přicházíte do ČR jako nová internetová firma. Jakou marketingovou strategii pro svoji revoluční službu „Come'n'Play“ pro poslouchání hudby přes internet byste zvolili?

Datový soubor Jobs

26. Najděte v datech skupiny zemí podobných ve smyslu zaměstnanosti.

Datový soubor Kabelovka

27. Jaké je využívání TV a různých technologií příjmu signálu v ČR?
28. Liší se nějak častý TV divák od běžného? Jak?
29. Nová televizní stanice chce vstoupit na český trh. Jak by se měla profilovat, aby odpovídala častým a typickým potřebám české populace?

Datový soubor Mazlicci

30. Jaká je situace v ČR z pohledu domácích mazlíčků? Kde pro ně lidé nejčastěji nakupují krmení?
31. Kdo je typický chovatel? Liší se chovatelé koček od chovatelů psů?
32. Chcete uvést do prodeje nový produkt pro kočky a psy. Jaka je cílová skupina spotřebitelů? V jaké distribuční síti byste výrobek prodávali?

Datový soubor Mushroom

33. Na základě údajů o houbách rozhodněte, jestli je daná houba jedovatá nebo jedlá. Data nejprve rozdělte na tréninková a testovací a poté Váš model vyzkoušejte na testovacích datech.

Datový soubor Postoje

34. Analyzujte výroky = charakteristiky dosavadního života. Jaké výroky jsou si „blízko“ a jaké jsou naopak opačné? Lze výroky zredukovat do menšího počtu tvrzení?
35. Vytvořte segmentaci populace podle analyzovaných výroků.
36. Lze vysledovat nějaký vztah mezi výroky a demografickými charakteristikami?

Datový soubor Protein

37. Najděte v datech skupiny zemí podobných ve smyslu spotřeby bílkovin.

Datový soubor Records

38. Najděte v datech skupiny zemí s podobnými atletickými rekordy.

Datový soubor Saheart

39. Odhalte rizikové faktory pro onemocnění srdce.
40. Lze v datech odhalit podobné skupiny lidí? Které z nich mají problémy s onemocněním srdce?

Datový soubor Salaries

41. Najděte faktory, na kterých závisí výše průměrného akademického platu.

Datový soubor Sleep

42. Na základě údajů o spánku zvířat najděte nejvíce podobná/nepodobná zvířata. Dají se mezi zvířaty nalézt nějaké skupiny podobných zvířat?
43. Predikujte, jestli je dané zvíře nebezpečné (index 3–5) či nikoliv (index 1–2) na základě údajů o spánku zvířat.

Datový soubor Sport

44. Popište využívání různých druhů sportu v ČR. Které sporty jsou provozovány klienty současně?
45. Existuje rozdíl mezi sportováním bohatších a chudších klientů? Jak se bohatší klienti odlišují od běžné populace?
46. Existují v populaci nějaké typické skupiny věnující se určitým sportům?

Datový soubor States

47. Najděte americké státy, ve kterých je podobná situace z hlediska vzdělání. Existuje nějaký stát, ve kterém je situace diametrálně odlišná od zbytku USA?

Datový soubor StudentPerformance

48. Jak se liší studenti školy Gabriel Pereira (GP) od studentů školy Mousinho da Silveira (MS)?
49. Můžeme v datech najít nějaké typické skupiny studentů? Jsou tyto skupiny stejné na obou školách?
50. Na základě dat predikujte známku studenta na konci 3. období (G3). Uvažujte modely, které uvažují, resp. neuvažují proměnné popisující známky na konci 1. a 2. období (G1 a G2).
51. Existuje souvislost mezi konzumací alkoholu a sociálními a školními atributy?

Datový soubor Swiss

52. Najděte faktory, na kterých závisí plodnost v Švýcarsku.
53. Najděte švýcarské provincie, které jsou si podobné z hlediska socio-ekonomických ukazatelů. Je u nich podobnost i z hlediska plodnosti?

Datový soubor TimeBudget

54. Najděte v datech skupiny lidí, které tráví podobně svůj volný čas.
55. Které skupiny lidí tráví svůj volný čas nejvíce rozdílným způsobem?

Datový soubor Tips

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

Datový soubor TVSales

57. Porovnejte americké televizní stanice z finančního hlediska. Které stanice jsou na tom podobně?
58. Modelujte prodejní cenu pomocí ostatních finančních faktorů.

Datový soubor USCrime

59. Porovnejte státy USA z hlediska kriminality? Které státy jsou na tom podobně?
60. Souvisí geografická poloha státu s jeho kriminalitou?

Datový soubor VolnyCas

61. Jaký podíl volného času věnuje průměrný člověk různým činnostem? Liší se v trávení volného času muži od žen? Jakou roli hraje věk?
62. Liší se v nějakých aspektech lidé trávící různě svůj volný čas? Lze nějak popsat typického jedince věnujícího se různým volno-časovým aktivitám?

Datový soubor Wine

63. Na základě výsledků chemické analýzy odhalte různé typy vína. Která vína jsou si nejvíce podobná? Čím se naopak nejvíce liší?
64. Na základě výsledků chemické analýzy odhadněte obsah alkoholu v daném víně.

Datový soubor WineQualityRed

65. Na základě výsledků chemické analýzy predikujte kvalitu červeného vína. Data nejprve rozdělte na tréninková a testovací a poté Váš model vyzkoušejte na testovacích datech.

Datový soubor WineQualityWhite

66. Na základě výsledků chemické analýzy predikujte kvalitu bílého vína. Data nejprve rozdělte na tréninková a testovací a poté Váš model vyzkoušejte na testovacích datech.

Datový soubor Zoo

67. Na základě vlastností zvířat najděte nejvíce podobná/nepodobná. Dají se mezi zvířaty nalézt nějaké skupiny podobných zvířat?