Teoretické otázky ke zkoušce `M8DM1 Data mining I`:

1. Vysvětlete pojmy data mining a KDD proces.
2. Popište historii a současnost data miningu. Uveďte jeho typické aplikace.
3. Uveďte data miningové metodologie. Popište metodologii SEMMA.
4. Uveďte data miningové metodologie. Popište metodologii CRISP-DM.
5. V čem se liší data mining a statistika. Jaké další disciplíny data mining zahrnuje? Uveďte úlohy, které DM řeší.
6. Popište rozdíl mezi prediktivním a deskriptivním modelováním. Uveďte příklady.
7. Popište rozdíl mezi supervised a unsupervised learning. Uveďte příklady.
8. Co to je relační databáze? Popište některé datové modely.
9. Jaké vztahy mezi tabulkami definujeme v relačních databázích?
10. Vysvětlete pojem OLTP. Jak se liší od OLAP?
11. Definujte datový sklad.
12. Co to je datové tržiště a datová pumpa?
13. Popište datovou kostku a typické operace nad ní.
14. Vysvětlete pojem OLAP. Jak se liší od OLTP?
15. Co to je SQL? K čemu se používá?
16. Co obnáší integrace dat? Jaké problémy při ní mohou nastat?
17. Popište úkoly při čištění dat (ověření a průzkum proměnných).
18. Proč a jak vznikají chybějící pozorování? Jak se tento problém řeší?
19. Co to je imputace dat? Jaké její metody znáte?
20. K čemu slouží transformace proměnných? Uveďte příklady.
21. K čemu se používá kategorizace proměnných? Proč?
22. Proč provádíme redukci datového souboru? Jak to můžeme udělat?
23. Proč se provádí redukce dimenze? Uveďte příklady používaných metod.
24. K čemu slouží exploratorní analýza dat? Jak se provádí?
25. Popište metody jednorozměrné exploratorní analýzy.
26. Popište grafické metody jednorozměrné exploratorní analýzy.
27. Popište metody mnohorozměrné exploratorní analýzy.
28. Popište metody exploratorní analýzy pro kategoriální data.

Matematické (metodologické) otázky ke zkoušce `M8DM1 Data mining I`:

1. **Analýza hlavních komponent**. Popište cíle analýzy hlavních komponent. Jak jsou komponenty konstruovány? V čem spočívá redukce dimenze? Jak se v praxi aplikuje?

2. **Faktorová analýza**. Popište cíle a model faktorové analýzy. Jak se faktory hledají? Jak se v praxi aplikuje? K čemu slouží rotace?

3. **Mnohorozměrné škálování**. Popište úlohu mnohorozměrného škálování. Jaký je rozdíl mezi metrickým a nemetrickým? Popište základní myšlenky metrického škálování a zobrazení v prostoru nízké dimenze. Uveďte hlavní kroky Shepardova - Kruskalova algoritmu.

4. **Kontingenční tabulky**. Popište testy nezávislosti v kontingenčních tabulkách. Jak se v nich měří závislosti? Popište znaménkové schéma. K čemu se používá? Co to je a k čemu se používá korespondenční analýza? Jak se interpretují její výsledky?

5. **Analýza nákupního košíku**. Popište analýzu nákupního košíku. Co to je podpora a spolehlivost? Jak se hledají pravidla pro dvou i víceprvkové množiny? Popište její zobecnění pro negované položky a hierarchické struktury dat.

6. **Hierarchická shluková analýza**. Popište úlohu shlukové analýzy. Uveďte metody hierarchického shlukování. Jak se počítají vzdálenosti mezi pozorováními a jak mezi shluky? Jak se určí výsledný počet shluků?

7. **Nehierarchická shluková analýza**. Popište úlohu shlukové analýzy. V čem se nehierarchické shlukování liší od hierarchického. Popište metodu k-means a k-medoids. Jak se určí výsledný počet shluků?

8. **Lineární regrese**. Popište model lineární regrese. Co to je multikolinearita? Jak se identifikuje a jaké může mít následky? Popište hřebenovou regresi a LASSO. K čemu se tyto metody používají?

9. **Logistická regrese**. Popište model logistické regrese. Co znamenají jednotlivé parametry tohoto modelu? Popište multinomickou regresi.

10. **Logistická regrese**. Co to je logistické skóre? Jak se v logistické regresi odhadují hodnoty závisle proměnné? Co to je ROC a Lorenzova křivka? Uveďte číselné charakteristiky odvozené od těchto křivek.

11. **Rozhodovací stromy**. Jakou úlohu řešíme pomocí rozhodovacích stromů? Popište algoritmy CART a CHAID. K čemu slouží a jak funguje prořezávání?