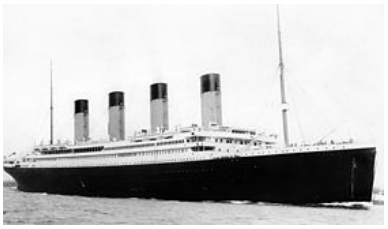
1) Titanic

Popis datového souboru:



Datový soubor *titanic.csv* obsahuje 1 045 případů a 8 proměnných. V datovém souboru jsou postupně popsány: jméno, pohlaví (0=žena, 1=muž), věk, počet sourozenců/partnerů na palubě, počet rodičů/dětí na palubě, cena lístku, třída (1. až 3.) a zda jedinec přežil (0=nepřežil, 1=přežil) havárii Titaniku v roce 1912.

Cílem analýzy je určit, které charakteristiky nejlépe predikují, zda jedinec přežil. Posledních pět řádků datového souboru využijte pouze pro predikci přežití u posledního úkolu.

1) Interpretujte podle výsledků logistické regrese:
   - Které parametry na hladině významnosti 0,05 nejsou významně asociovány s přežitím?
   - Kdo měl vyšší pravděpodobnost přežití:
     Muži nebo ženy?
     Mladší nebo starší?
     Cestující s vysokou nebo nízkou cenou lístku?
     Cestující v 1. nebo 3. třídě?

2) Uveďte proměnné, které Vám doporučí dopředná selekce. Které tři parametry budou mít v analýze největší diskriminační schopnost? Uveďte Wilksovo lambda modelu s těmito třemi parametry.

3) Apriorní pravděpodobnosti byly nastaveny proporcionálně nebo rovnoměrně?

4) Proveďte diskriminační analýzu na proměnných, které doporučila dopředná selekce. Doplňte tabulku níže a interpretujte správnost klasifikace. Uveďte celkový počet špatně klasifikovaných osob, a která skupina je klasifikována s větší a která s menší přesností. Jaká je celková přesnost klasifikace?

Skutečnost	Klasifikace dle diskriminační analýzy: 0	Klasifikace dle diskriminační analýzy: 1
0		
1		

5) U 932. a 1009. jedince interpretujte, jak diskriminační analýza na základě posteriorní pravděpodobnosti klasifikovala přežití a zda tato klasifikace byla správná.

6) Klasifikujte na základě sestaveného modelu zbývajících 5 osob (posledních 5 řádků v souboru). Uveďte výsledek klasifikace.

2) Znalosti

Popis datového souboru:

Datový soubor *znalosti.txt* obsahuje 239 případů a 10 proměnných. Proměnné biologie až dějepis popisují výsledky testů (počet bodů) středoškolských studentů. Na základě subjektivního pohledu učitele (na vzdělanost studentů) byli studenti zařazeni do jedné z pěti kategorií v proměnné znalosti (1 největší vzdělanost, 5 nejmenší vzdělanost), čtyři studenti (v posledních řádcích) nejsou do těchto kategorií zařazeni.

Cílem analýzy je určit, které předměty nejlépe predikují znalost studenta hodnocenou vyučujícím. Poslední čtyři řádky datového souboru využijte pouze pro predikci zařazení studenta do jedné z kategorií znalostí.

1) Vykreslete a vložte boxplot pro výsledky testů předmětů v jednotlivých skupinách studentů (využijte nastavení par(mfrow=c(3,3)), aby se grafy vykreslily v jenom okně). Uveďte předmět, který bude mít v analýze největší diskriminační schopnost (jakou kvalifikaci bude mít učitel, který studenty hodnotil).

2) Které proměnné vyřadí dopředná selekce z modelu? Na základě výsledků dopředné selekce uveďte, které tři parametry budou mít v analýze největší diskriminační schopnost. Uveďte Wilksovo lambda modelu se všemi vybranými proměnnými.

3) S jakou pravděpodobností student dostane hodnocení 1 a 3?

4) Doplňte tabulku níže a interpretujte správnost klasifikace. Uveďte celkový počet špatně klasifikovaných studentů, a která skupina je klasifikována s největší a která s nejmenší přesností. Jaká je celková přesnost klasifikace?

Skutečnost	Klasifikace dle diskriminační analýzy 1	2	3	4	5
1					
2					
3					
4					
5					

5) Kolik kanonických os můžete vykreslit a jak lze tento počet odvodit? Kolik musíte vykreslit os, abyste popsali alespoň 90% variability mezi skupinami.

6) U 1. a 2. studenta interpretujte, jak diskriminační analýza na základě posteriorní pravděpodobnosti klasifikovala studenta a zda tato klasifikace byla správná.

7) Klasifikujte na základě sestaveného modelu zbývající studenty (poslední 4 studenti v souboru) do kategorií znalosti. Uveďte, do jakých skupin diskriminační analýza studenty klasifikovala.