13. ROC křivka

Motivace: ROC (Receiver Operating Characteristic Curve) křivky slouží k hodnocení a grafickému znázornění chování klasifikačních pravidel při klasifikaci do dvou tříd. ROC křivka byla poprvé zavedena kolem roku 1950 jako nástroj k vyhodnocení kvality radarových dat – odlišení signálu od šumu. V medicíně se používá nejčastěji pro zhodnocení přesnosti diagnostického testu na základě měření, jehož výsledkem je nějaká spojitá veličina. Stanovíme určitou hodnotu této veličiny (tzv. dělicí bod) a podle něj rozhodneme, zda nastal pozitivní či negativní výsledek testu. ROC křivka graficky znázorňuje chování tohoto klasifikačního pravidla v závislosti na měnícím se dělicím bodu.

Senzitivita a specificita diagnostického testu

Předpokládáme, že máme dvě skupiny objektů – jedna skupina objektů splňuje nějakou podmínku (pozitivní případy), druhá skupina nikoliv (negativní případy). Provedeme diagnostický test, který objekt označí buď jako pozitivní nebo jako negativní.

Zavedeme následující označení:

jev H … objekt je pozitivní
jev $\overline{H}$ … objekt je negativní
jev A … test označí objekt za pozitivní
jev $\overline{A}$ … test označí objekt za negativní

Apriorní pravděpodobnost P(H) se nazývá prevalence a vyjadřuje pravděpodobnost výskytu pozitivních objektů v souboru všech objektů.

Podmíněná pravděpodobnost P(A/H)se nazývá senzitivita a vyjadřuje pravděpodobnost, že test dá kladný výsledek u pozitivního objektu.

Podmíněná pravděpodobnost $P(\overline{A}/\overline{H})$se nazývá specificita a vyjadřuje pravděpodobnost, že test dá záporný výsledek u negativního objektu.

Podmíněná pravděpodobnost $P(\overline{A}/H)$se nazývá falešná negativita a vyjadřuje pravděpodobnost, že test dá záporný výsledek u pozitivního objektu.

Podmíněná pravděpodobnost $P(A/\overline{H})$se nazývá falešná pozitivita a vyjadřuje pravděpodobnost, že test dá kladný výsledek u negativního objektu.

Aposteriorní pravděpodobnost P(H/A) se nazývá prediktivní hodnota pozitivního testu a vyjadřuje pravděpodobnost, že objekt je skutečně pozitivní, když test dopadl pozitivně.

Aposteriorní pravděpodobnost $P(\overline{H}/\overline{A})$ se nazývá prediktivní hodnota negativního testu a vyjadřuje pravděpodobnost, že objekt je skutečně negativní, když test dopadl negativně.

Uvedené podmíněné pravděpodobnosti neznáme, můžeme je pouze odhadnout pomocí výsledků testu, které zapíšeme do tzv. matice záměn:

výsledek testu	podmínka		celkem
	H (pozitivní)	$\overline{H}$ (negativní)	
A (pozitivní)	a	b	a+b
$\overline{A}$ (negativní)	c	d	c+d
celkem	a+c	b+d	n

Odhad senzitivity: $\hat{P}(A/H) = TPF = \frac{a}{a+c}$ (true positive fraction – relativní četnost správně klasifikovaných pozitivních případů).

Odhad specificity: $\hat{P}(\overline{A}/\overline{H}) = TNF = \frac{d}{b+d}$ (true negative fraction – relativní četnost správně klasifikovaných negativních případů).

Odhad falešné negativity: $\hat{P}(\overline{A}/H) = FNF = \frac{c}{a+c}$ (false negative fraction – relativní četnost nesprávně klasifikovaných pozitivních případů).

Odhad falešné pozitivity: $\hat{P}(A/\overline{H}) = FPF = \frac{b}{b+d}$ (false positive fraction – relativní četnost nesprávně klasifikovaných negativních případů).

Je okamžitě zřejmé, že TPF + FNF = 1, TNF + FPF =1.

Odhady prediktivních hodnot pozitivního a negativního testu počítáme podle Bayesova vzorce.

$$\hat{P}(H/A) = PVV = \frac{P(H)\hat{P}(A/H)}{P(H)\hat{P}(A/H) + P(\overline{H})\hat{P}(A/\overline{H})} = \frac{P(H)\cdot TPF}{P(H)\cdot TPF + (1-P(H))(1-TNF)}$$

$$\hat{P}(\overline{H}/\overline{A}) = PVN = \frac{P(\overline{H})\hat{P}(\overline{A}/\overline{H})}{P(\overline{H})\hat{P}(\overline{A}/\overline{H}) + P(H)\hat{P}(\overline{A}/H)} = \frac{(1-P(H))\cdot TNF}{(1-P(H))\cdot TNF + P(H)(1-TPF)}$$

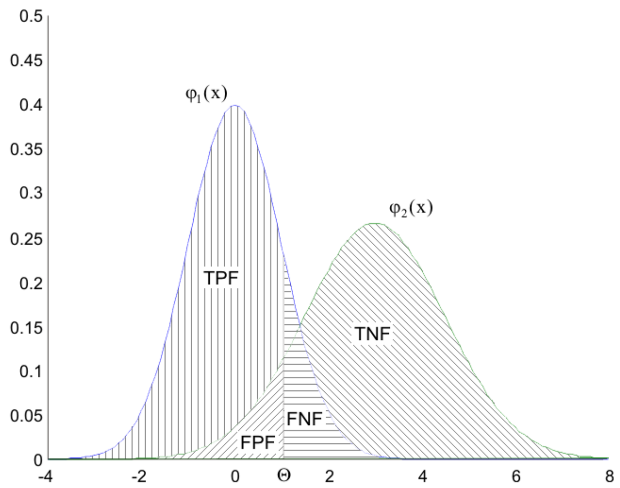
Vidíme, že praktický význam diagnostického testu záleží na prevalenci P(H), odhadu senzitivity TPF a odhadu specificity TNF. Tyto charakteristiky tedy plně určují prediktivní hodnoty PVV a PVN.

Konstrukce ROC křivky

Nechť je diagnostický test založen na nějaké spojité náhodné veličině X (nazývá se prediktor), podle níž chceme objekty klasifikovat buď do 1. skupiny (pozitivní případy) nebo do 2. skupiny (negativní případy). Předpokládáme, že tato veličina se v 1. skupině řídí spojitým rozložením s hustotou pravděpodobnosti $\varphi_1(x)$ a ve 2. skupině spojitým rozložením s hustotou pravděpodobnosti

$\varphi_2(x)$. Pokud veličina X nabude hodnoty nejvýše θ, objekt zařadíme do 1. skupiny, v opačném případě do 2. skupiny.

Ilustrace:



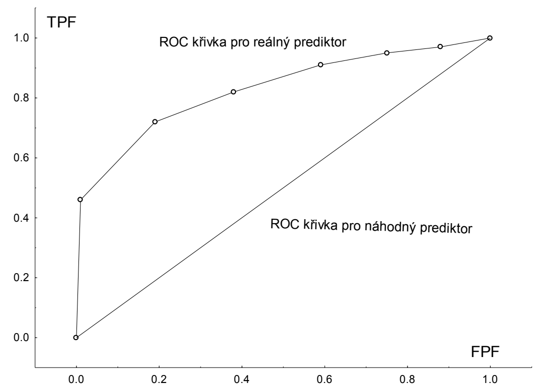
Plocha pod první hustotou ležící nalevo od dělicího bodu θ vyjadřuje podíl správně klasifikovaných pozitivních případů, tedy TPF – odhad senzitivity. Plocha pod první hustotou ležící napravo od dělicího bodu θ vyjadřuje podíl nesprávně klasifikovaných pozitivních případů, tedy FNF – odhad falešné negativity.

Plocha pod druhou hustotou ležící nalevo od dělicího bodu θ vyjadřuje podíl nesprávně klasifikovaných negativních případů, tedy FPF – odhad falešné pozitivity. Plocha pod druhou hustotou ležící napravo od dělicího bodu θ vyjadřuje podíl správně klasifikovaných negativních případů, tedy TNF – odhad specificity.

Posuneme-li dělicí bod θ napravo, zvětšíme odhad senzitivity TPF, ale zmenšíme odhad specificity TNF. Naopak, posuneme-li dělicí bod θ nalevo, zmenšíme TPF a zvětšíme TNF.

Abychom docílili společného nárůstu či poklesu dvojice charakteristik, vezmeme místo odhadu specificity odhad falešné pozitivity FPF = 1 – TNF. Máme tedy dvojici (FPF, TPF). Postupně měníme dělicí bod θ a pro každou hodnotu θ zaznamenáme dvojici (FPF, TPF). Grafickým znázorněním těchto dvojic získáme ROC křivku.

Ilustrace:



Objekty klasifikované náhodným prediktorem jako pozitivní patří ve skutečnosti mezi pozitivní s pravděpodobností P(H) a mezi negativní s pravděpodobností 1 - P(H).

Objekty klasifikované reálným prediktorem jako pozitivní patří ve skutečnosti mezi pozitivní s pravděpodobností vyšší než P(H) a mezi negativní s pravděpodobností nižší než 1 – P(H).

Poznámka: ROC křivku lze použít i v případě, kdy máme k dispozici odhady pravděpodobností příslušnosti objektů k 1. a 2. skupině. Je-li tato pravděpodobnost nanejvýš θ (0<θ<1), zařadíme objekt do 1. skupiny, jinak do 2. skupiny.

Příklad ROC křivky

Máme dvě skupiny osob, přičemž v jedné skupině jsou jedinci trpící určitou chorobou a ve druhé skupině jsou lidé zdraví. V obou skupinách sledujeme hodnoty nějaké veličiny X, podle níž chceme osoby klasifikovat na nemocné a zdravé.

X	počet nemocných	počet zdravých
≤ 5	18	1
5,1 - 7	7	17
7,1 - 9	4	36
> 9	3	39
celkem	32	93

Nejprve předpokládáme, že osoby s hodnotou $X \leq 5$ jsou nemocné. Dostaneme matici záměn:

X	počet nemocných	počet zdravých	celkem
≤ 5	18	1	19
> 5	14	92	106
celkem	32	93	125

TPF = 18/32 = 0,56
FPF = 1/93 = 0,01

Nyní posuneme dělicí bod θ tak, že osoby s hodnotou $X \leq 7$ jsou považovány za nemocné. Dostaneme matici záměn:

X	počet nemocných	počet zdravých	celkem
≤ 7	25	18	43
> 7	7	75	82
celkem	32	93	125

TPF = 25/32 = 0,78
FPF = 18/93 = 0,19

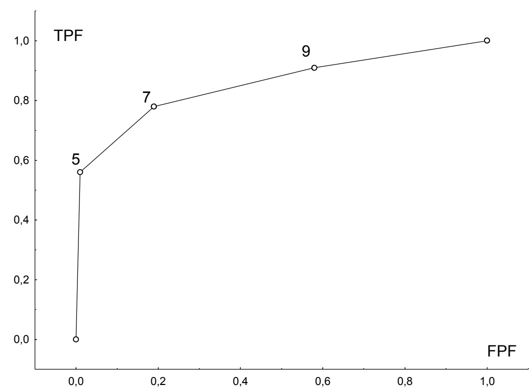
Dělicí bod θ posuneme ještě jednou tak, že osoby s hodnotou $X \leq 9$ jsou považovány za nemocné. Dostaneme matici záměn:

X	počet nemocných	počet zdravých	celkem
≤ 9	29	54	83
> 9	3	39	42
celkem	32	93	125

TPF = 29/32 = 0,91
FPF = 54/93 = 0,58

Hodnoty dělicích bodů a odhady falešné pozitivity a odhady senzitivity napíšeme do tabulky a sestrojíme ROC křivku:

θ	FPF	TPF
5	0,01	0,56
7	0,19	0,78
9	0,58	0,91



Optimální dělicí bod se získá tak, že zjistíme, která hodnota θ odpovídá maximálnímu geometrickému průměru odhadu senzitivity a specificity.

Vlastnosti ROC křivky

a) ROC křivka znázorňuje vztah mezi odhady specificity a senzitivity diagnostického testu. Definiční obor a obor hodnot jsou intervaly $\langle 0,1 \rangle$. Čtverec $\langle 0,1 \rangle \times \langle 0,1 \rangle$ se nazývá ROC prostor.
b) ROC křivka je invariantní k monotónní transformaci výsledku testu (tj. prediktoru X). Znamená to, že závisí pouze na pořadí dat, nikoliv na jejich konkrétním umístění na číselné ose.
c) Teoretická ROC křivka pro náhodný prediktor (tj. pro test s nulovou diskriminační schopností) je diagonála vedoucí z levého dolního rohu do pravého horního rohu ROC prostoru.
d) Teoretická ROC křivka pro test s perfektní diskriminační schopností (tj. test, který každý pozitivní objekt prohlásí za pozitivní a každý negativní objekt za negativní) kopíruje levý horní roh ROC prostoru.
e) Čím blíže je ROC křivka levému hornímu rohu ROC prostoru, tím lepší je diskriminační schopnost testu.
f) Plocha pod ROC křivkou (označovaná jako AUC) může být chápána jako měřítko kvality testu.

Plocha AUC pod ROC křivkou

Velikost A plochy AUC pod ROC křivkou je nejběžnější kvantitativní index popisující ROC křivku. Současně vyjadřuje přesnost testu. Nevyžaduje žádné předpoklady o rozložení prediktoru.

Pro ohodnocení přesnosti testu se používá tabulka:

A	hodnocení
0,9 - 1	výborně
0,8 – 0,9	velmi dobře
0,7 – 0,8	dobře
0,6 - 0,7	dostatečně
0,5 – 0,6	nedostatečně

Velikost A plochy AUC má několik možných interpretací:

a) průměrná hodnota senzitivity pro všechny možné hodnoty specificity
b) průměrná hodnota specificity pro všechny možné hodnoty senzitivity
c) pravděpodobnost, že náhodně vybraný objekt ze skupiny pozitivních objektů bude mít nižší hodnotu prediktoru než náhodně vybraný objekt ze skupiny negativních objektů.

Velikost A plochy AUC se počítá buď neparametricky (např. pomocí lichoběžníkového nebo Simpsonova pravidla) nebo parametricky s použitím metody maximální věrohodnosti.

Binormální model

Empirická ROC křivka se často vyhlazuje teoretickou ROC křivkou procházející body (FPF, TPF). Tato křivka je založena na předpokladu, že výsledek testu (tj. náhodná veličina X – prediktor) se v 1. skupině řídí rozložením $N(\mu_1, \sigma_1^2)$ a ve 2. skupině rozložením $N(\mu_2, \sigma_2^2)$. Hovoříme o tzv. binormálním modelu. Binormální model je plně určen dvěma parametry a, b, kde $a = \frac{\mu_2 - \mu_1}{\sigma_2}, b = \frac{\sigma_1}{\sigma_2}$. Protože parametry $\mu_1, \mu_2, \sigma_1, \sigma_2$ neznáme, nahradíme je odhady M_1, M_2, S_1, S_2, kde M_1, M_2 jsou výběrové průměry v 1. a 2. skupině a S_1, S_2 jsou výběrové směrodatné odchylky v 1. a 2. skupině. Po dosazení výběrových charakteristik do vzorců pro parametry a, b dostaneme jejich odhady $\hat{a}, \hat{b}$. Teoretická ROC křivka je pak určena dvojicemi bodů $(1 - \Phi(\theta), 1 - \Phi(b\theta - a))$, kde Φ je distribuční funkce standardizovaného normálního rozložení N(0,1) a $\theta \in (-\infty, \infty)$ představuje všechny možné hodnoty dělicího bodu.

V binormálním modelu lze snadno testovat hypotézu o diskriminační schopnosti testu. Testujeme H_0: A = 0,5 proti H_1: A > 0,5. Označme $\hat{A}$ odhad velikosti plochy A a $\hat{D}(\hat{A})$ rozptyl tohoto odhadu. Testová statistika $Z = \frac{\hat{A} - 0{,}5}{\sqrt{\hat{D}(\hat{A})}}$ se v případě platnosti nulové hypotézy asymptoticky řídí rozložením N(0,1). Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti α, když $|Z| \geq u_{1-\alpha}$, kde u_α je α-kvantil rozložení N(0,1). Přitom odhady $\hat{A}$ a $\hat{D}(\hat{A})$ se počítají podle vzorců:

$$\hat{A} = \Phi\left(\frac{\hat{a}}{\sqrt{1+\hat{b}^2}}\right),\quad \hat{D}(\hat{A}) = f^2\hat{D}(\hat{a}) + g^2\hat{D}(\hat{b}) + 2fg\hat{C}(\hat{a},\hat{b}),\ \text{kde}$$

$$f = \frac{\exp\left(-\frac{\hat{a}^2}{2(1+\hat{b}^2)}\right)}{\sqrt{2\pi(1+\hat{b}^2)}},\quad g = -\frac{\hat{a}\hat{b}\exp\left(-\frac{\hat{a}^2}{2(1+\hat{b}^2)}\right)}{\sqrt{2\pi(1+\hat{b}^2)^3}},$$

$\hat{D}(\hat{a}) = \frac{n_2(\hat{a}^2+2) + 2n_1\hat{b}^2}{2n_1n_2}, \hat{D}(\hat{b}) = \frac{(n_1+n_2)\hat{b}^2}{2n_1n_2}, \hat{C}(\hat{a},\hat{b}) = \frac{\hat{a}\hat{b}}{2n_1}$, přičemž n_1 je počet objektů v 1. skupině a n_2 je počet objektů ve 2. skupině.

Příklad:

V r. 1997 nastoupilo do magisterského studia na PřF MU v Brně 390 studentů.
Důležité sledované proměnné:
pohlaví,
počet bodů u přijímací zkoušky,

typ absolvované střední školy (gymnázium, střední průmyslová škola, ostatní),
zaměření studia (odborné, učitelské),
studijní program (fyzika, chemie, geologie, matematika, biologie, aplikovaná matematika, geografie),
studijní průměr,
úspěch u prvního termínu státní závěrečné zkoušky.
Hodnoty všech proměnných se podařilo získat u 236 studentů.

Popis datového souboru pomocí tabulek četností:

	Tabulka četností:pohlaví (magistri1997.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
muž	109	109	46,2	46,2
žena	127	236	53,8	100,0

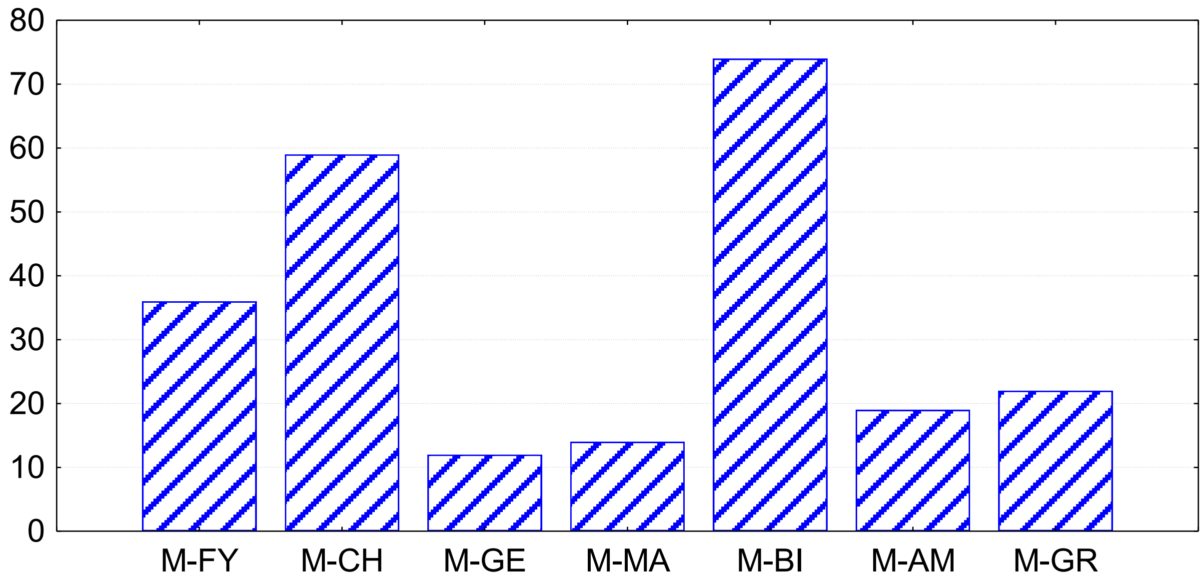
	Tabulka četností:program (magistri1997.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
M-AM	19	19	8,1	8,1
M-BI	74	93	31,4	39,4
M-FY	36	129	15,3	54,7
M-CH	59	188	25,0	79,7
M-GR	22	210	9,3	89,0
M-MA	14	224	5,9	94,9
M-GE	12	236	5,1	100,0

	Tabulka četností:priznak (magistri1997.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
odbornik	166	166	70,3	70,3
učitel	70	236	29,7	100,0

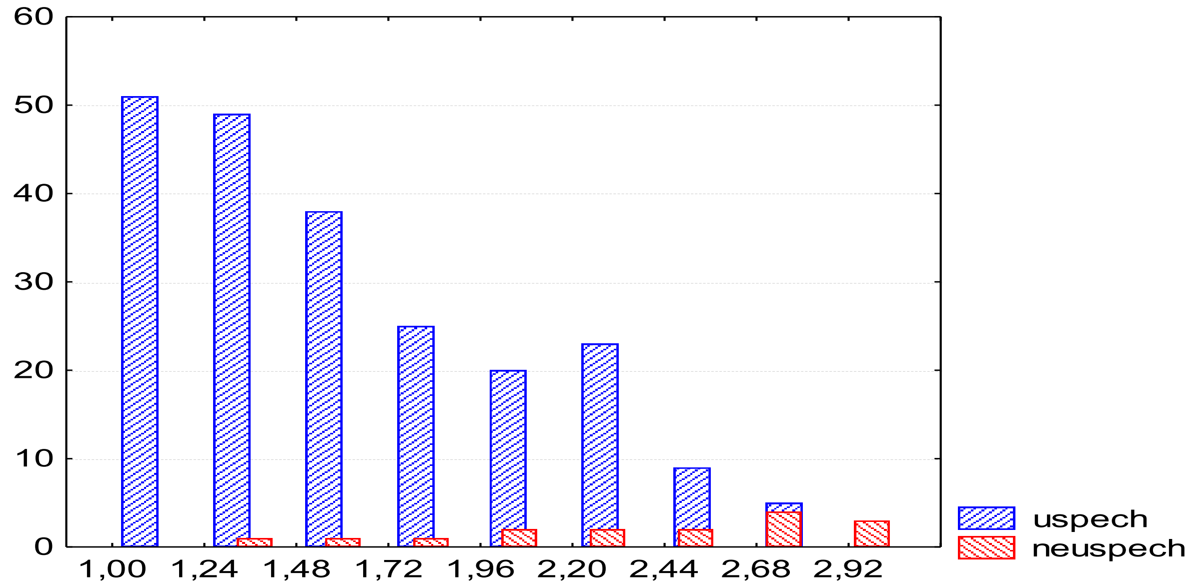
	Tabulka četností:SZZ_1 (magistri1997.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
uspech u 1. terminu SZZ	220	220	93,2	93,2
neuspech u 1. terminu SZZ	16	236	6,8	100,0

	Tabulka četností:SS_typ (magistri1997.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
Gymnázium	211	211	89,40678	89,4068
SPŠ	13	224	5,50847	94,9153
ostatní	12	236	5,08475	100,0000

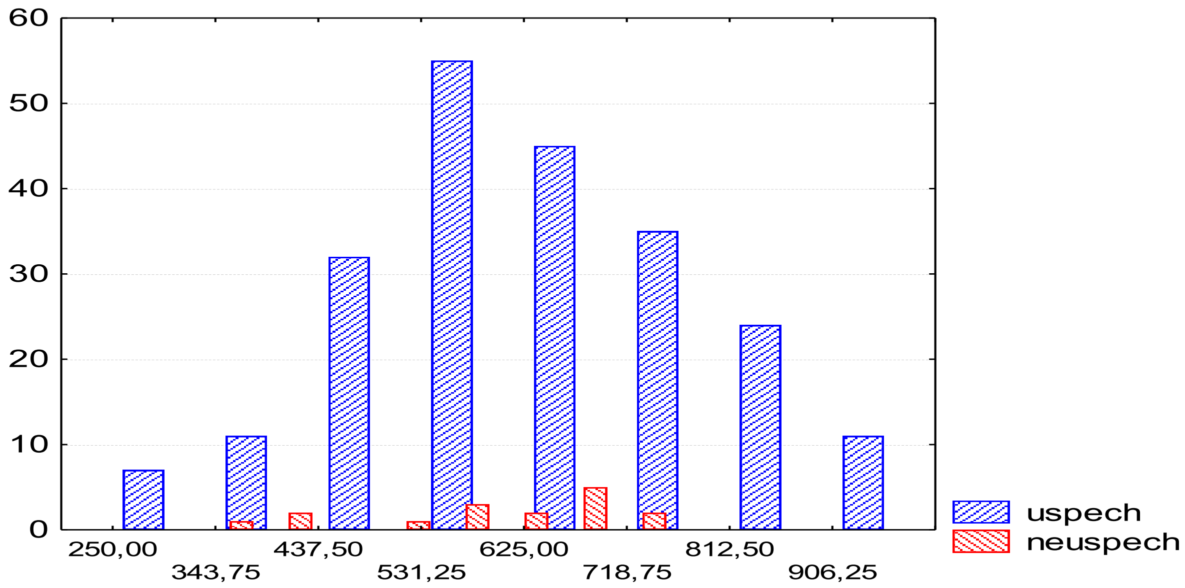
Sloupkový diagram studijních programů



Histogram průměrného prospěchu ve skupině úspěšných a neúspěšných studentů



Histogram počtu bodů u přijímací zkoušky ve skupině úspěšných a neúspěšných studentů

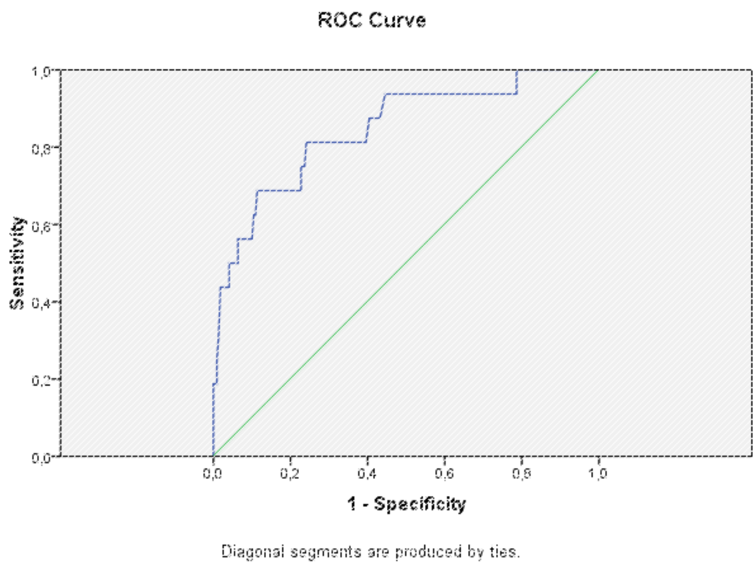


Konstrukce ROC křivky je implementována v systému SPSS, v systému STATISTICA se nachází pouze v modulu Neuronové sítě.

Nyní budeme za prediktor úspěchu u 1. termínu SZZ považovat průměrný prospěch dosažený za celou dobu studia.

Analyze – ROC Curve – Test Variable prumer, State Variable SZZ_1, Value of State Variable 1, v Display zaškrtneme ROC Curve, With Diagonal reference line, Standard error and confidence interval, Coordinate points of the ROC Curve – OK.

Průběh ROC křivky



Pomocí souřadnic ROC křivky vypočteme geometrický průměr specificity a senzitivity. Maximální hodnota tohoto průměru je dosažena pro prospěch = 1,99.

Velikost A plochy AUC pod ROC křivkou s 95% asymptotickým intervalem spolehlivosti

Area Under the Curve

Test Result Variable(s):prumer

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,846	,055	,000	,739	,953

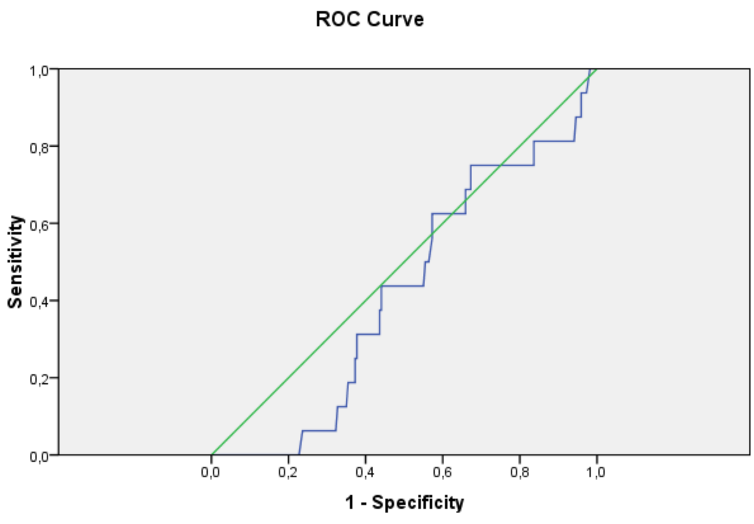
a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Na asymptotické hladině významnosti 0,05 testujeme nulovou hypotézu, že velikost plochy pod ROC křivkou je 0,5 (tj. průměrný prospěch nemá žádný vliv na výsledek studenta u prvního termínu státní závěrečné zkoušky) proti alternativní hypotéze, že velikost plochy pod ROC křivkou je větší než 0,5 (tj. průměrný prospěch má jistou predikční schopnost vzhledem k výsledku studenta u prvního termínu státní závěrečné zkoušky). Protože 95% asymptotický interval spolehlivosti pro A neobsahuje číslo 0,5, nulovou hypotézu zamítáme na asymptotické hladině významnosti 0,05.

Dále budeme za prediktor úspěchu u 1. termínu SZZ považovat počet bodů, které student získal u přijímací zkoušky.

Průběh ROC křivky



Diagonal segments are produced by ties.

Velikost A plochy AUC pod ROC křivkou s 95% asymptotickým intervalem spolehlivosti

Area Under

Test Result Variable(s):body

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,420	,063	,287	,297	,543

The test result variable(s): body has at least one tie between the positive act

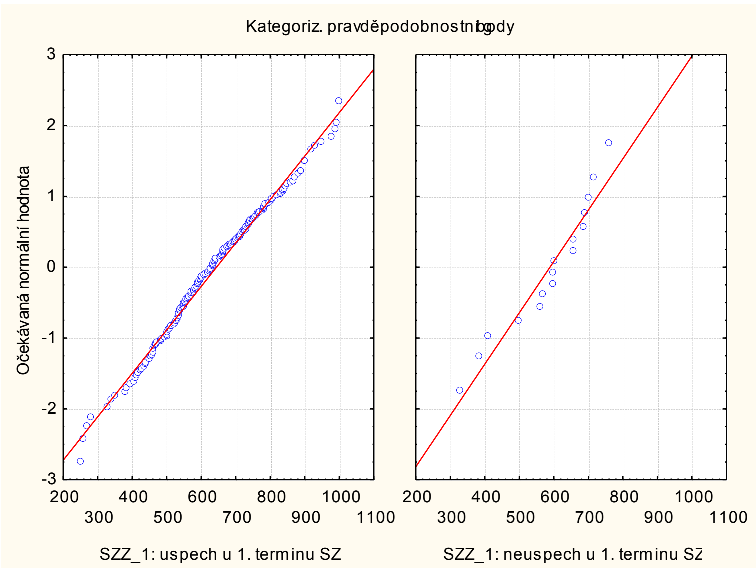
a. Under the nonparametric assumption

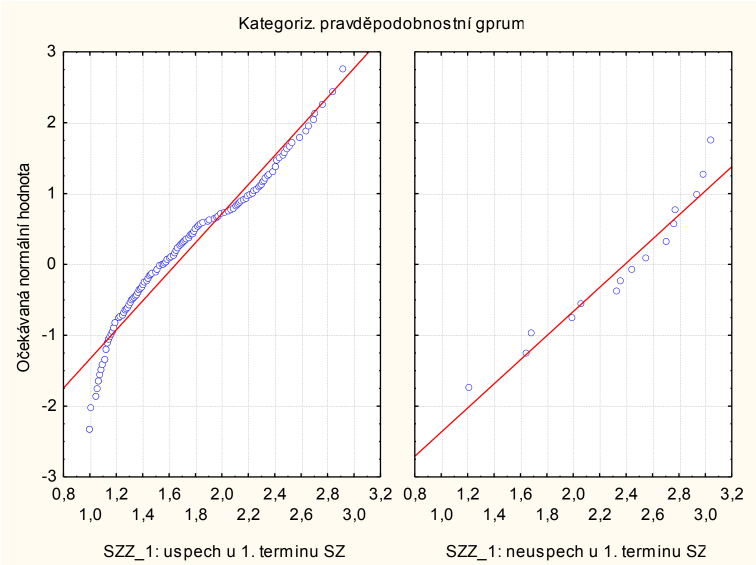
b. Null hypothesis: true area = 0.5

Na asymptotické hladině významnosti 0,05 testujeme nulovou hypotézu, že velikost plochy pod ROC křivkou je 0,5 (tj. počet bodů u přijímací zkoušky nemá žádný vliv na výsledek studenta u prvního termínu státní závěrečné zkoušky) proti alternativní hypotéze, že velikost plochy pod ROC křivkou je větší než 0,5 (tj. počet bodů u přijímací zkoušky má jistou predikční schopnost vzhledem k výsledku studenta u prvního termínu státní závěrečné zkoušky). Protože 95% asymptotický interval spolehlivosti pro A obsahuje číslo 0,5, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Další sledované proměnné by mohly zpřesnit predikci úspěchu studenta u prvního termínu státní závěrečné zkoušky. Provedeme diskriminační analýzu se spojitými proměnnými PRUMER a BODY a kategorizovanými proměnnými POHLAVÍ, TYP STUDIA, TYP SŠ, PROGRAM.

Nejprve orientačně ověříme předpoklad normality spojitých proměnných BODY a PRUMER.





Vidíme, že normalita proměnné PRUMER je v obou skupinách poněkud porušena.

Zvolíme krokovou dopřednou metodu.

Tabulka výsledků pro proměnné zařazené do modelu:

	Výsledky diskriminační funkční analýzy (magistri1997.sta) krok 3, poč. prom. v modelu: 3; grupovací: SZZ_1 (2 skup) Wilks. lambda: ,85566 přibliž F (3,232)=13,045 p< ,0000					
N=236	Wilks. Lambda	Parc. Lambda	F na vyj (1,232)	Úroveň p	Toler.	1-toler. R^2
prumer	0,992297	0,862301	37,04753	0,000000	0,796787	0,203213
body	0,860719	0,994121	1,37207	0,242658	0,824355	0,175645
priznak	0,859759	0,995231	1,11182	0,292783	0,953561	0,046439

Tabulka výsledků pro proměnné nezařazené do modelu:

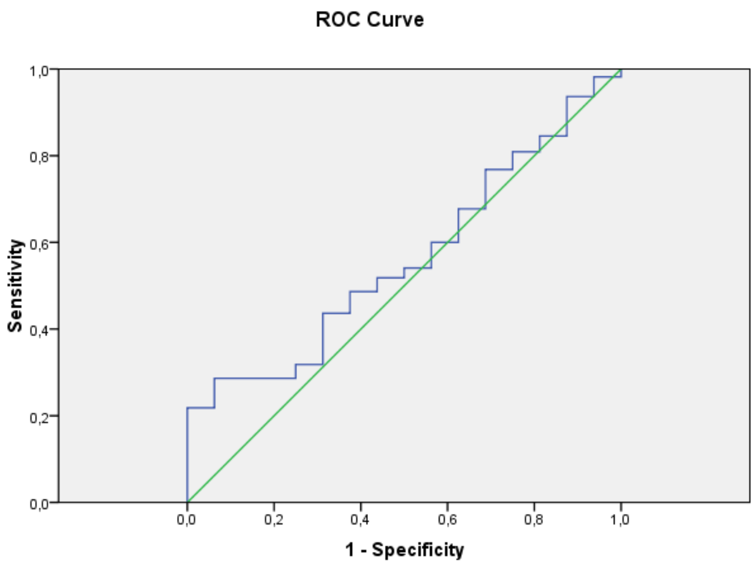
	Proměnné aktuálně neobsažené v modelu (magistri1997.sta) SV pro vš. F-testy: 1,231					
N=236	Wilks. Lambda	Parc. Lambda	F na zahr.	Úroveň p	Toler.	1-toler. R^2
pohlaví	0,855441	0,999746	0,058697	0,808782	0,963711	0,036289
program	0,855164	0,999422	0,133523	0,715141	0,975874	0,024126
SS_typ	0,855641	0,999979	0,004805	0,944794	0,946056	0,053944

Klasikační matice získaná resubstituční metodou:

	Klasifikační matice (magistri1997.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace		
Skup.	% správnýc	uspech u 1. terminu SZZ p=,93220	neuspech u 1. terminu SZZ p=,06780
uspech u 1. terminu SZZ	99,09091	218	2
neuspech u 1. terminu SZZ	18,75000	13	3
Celkem	93,64407	231	5

Klasifikace je daleko úspěšnější u studentů, kteří uspěli u 1. termínu SZZ než u těch, kteří neuspěli.

Pro každého studenta zjistíme aposteriorní pravděpodobnost, s jakou je zařazen do skupiny úspěšných studentů a sestrojíme ROC křivku.



Plocha AUC pod ROC křivkou s 95% intervalem spolehlivosti:

Area Under the Curve

Test Result Variable(s):ap

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,562	,065	,406	,435	,689

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Na asymptotické hladině významnosti 0,05 testujeme nulovou hypotézu, že velikost plochy pod ROC křivkou je 0,5 (tj. zvolené proměnné charakterizující studenta nemají žádný vliv na jeho výsledek u prvního termínu státní závěrečné zkoušky) proti alternativní hypotéze, že velikost plochy pod ROC křivkou je větší než 0,5 (tj. dané proměnné mají jistou predikční schopnost vzhledem k výsledku studenta u prvního termínu státní závěrečné zkoušky). Protože 95% asymptotický interval spolehlivosti pro A obsahuje číslo 0,5, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.