PREDIKCIA ISCHEMICKEJ CHOROBY SRDCA

Úvod

Podľa údajov Svetovej zdravotníckej organizácie na celom svete každoročne zomrie 12 miliónov ľudí na choroby srdca. Polovica úmrtí v USA a iných rozvinutých krajinách je spôsobená srdcovo-cievnymi ochoreniami. Včasná prognóza kardiovaskulárnych ochorení môže pomôcť pri rozhodovaní o zmene životného štýlu u vysoko rizikových pacientov a následne znížiť komplikácie.

Cieľ

Zámerom tohto výskumu bolo určiť najvýznamnejšie rizikové faktory srdcových ochorení, ako aj predpovedať celkové riziko pomocou logistickej regresie.

Dáta

Dáta pochádzajú z kardiovaskulárnej štúdie obyvateľov mesta Framingham v štáte Massachusetts. Dataset obsahuje informácie o viac ako 4000 pacientoch, zaznamenaných je 15 atribútov, pričom každý z nich predstavuje potenciálny rizikový faktor pre ochorenie ISCH. Atribúty sú rozdelené na demografické, behaviorálne a medicínske. Prehľad uvádzame v tabuľke 1.

Tabuľka 1. Zoznam sledovaných premenných

	PremennáEN	**PremennáSK**	**Datový typ**	**Hodnoty**
Demografické	Male	Pohlavie	Nominálna / binárna	1 (muž) / 0 (žena)
	Age	Vek	Spojitá	Celé kladné čísla
	Education	Vzdelenie	Nominálna	1-4 (1-najnižšia úroveň vzdelania)
Behaviorálne	CurrentSmoker	V súčasnosti fajčí	Nominálna / binárna	1 (fajčí) / 0 (nefajčí)
	CigsPerDay	Počet cigariet za deň	Spojitá	Celé kladné čísla
Lekárske (hist)	BP Meds	Lieky na tlak	Nominálna / binárna	1 (užíval) / 0 (neužíval)
	PrevalentStroke	Cievna mozg.príhoda v minulosti	Nominálna / binárna	1 (prekonal) / 0 (neprekonal)
	PrevalentHyp	Hypertenzia v minulosti	Nominálna / binárna	1 (áno) / 0 (nie)
	Diabetes	Diabetes v minulosti	Nominálna / binárna	1 (áno) / 0 (nie)
Lekárske (súčasné)	TotChol	Hladina cholesterolu	Spojitá	Desatinné kladné čísla
	SysBP	Systolický krvný tlak	Spojitá	Desatinné kladné čísla
	DiaBP	Diastolický krvný tlak	Spojitá	Desatinné kladné čísla
	BMI	Index telesnej hmotnosti	Spojitá	Desatinné kladné čísla
	HeartRate	Srdcová frekvencia	Spojitá	Celé kladné čísla
	Glucose	Hladina glukózy	Spojitá	Celé kladné čísla
Závislá premenná	TenYearCHD	Riziko rozvoja ICHS do 10 rokov	Nominálna / binárna	1 (áno) / 0 (nie)

Datový súbor je dostupný na stránkach kaggle.com: https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/data

Spracovanie dát

V nasledujúcom texte popíšeme jednotlivé analýzy a postupy spracovania dát. Všetky analýzy a modely sme robili v programe Jamovi, pričom sme využili moduly „T-tests“, „Frequencies“ a „Regression“. Vzhľadom k obmedzenosti R v Jamovi sme pri modelovaní logistickej regresie využili aj vlastný kód.

Štatistická významnosť premenných

Aby sme si overili relevantnosť, resp. štatistickú závislosť sledovaných atribútov na závislej premennej, spravili sme t-testy pre metrické premenné a chi kvadrát testy pre nominálne premenné. Výsledky uvádzame v tabuľkách 2 a 3.

Tabuľka 2. Výsledky Welch testov pre dva nezávislé výbery

	Statistic	df	p
sysBP	-12.02	784	<.001
glucose	-4.84	640	<.001
totChol	-5.00	832	<.001
diaBP	-8.25	800	<.001
BMI	-4.48	816	<.001
heartRate	-1.47	878	0.141
cigsPerDay	-3.49	838	<.001
age	-15.58	916	<.001

Tabuľka 3. Výsledky Chi kvadrát testov pre dva nezávislé výbery

	Statistic	df	p	Odds ratio
male	33.1	1	<.001	1.64
education	31.9	3	<.001	-
currentSmoker	1.6	1	0.205	1.11
BPMeds	32.0	1	<.001	2.89
prevalentStroke	16.2	1	<.001	4.44
prevalentHyp	134	1	<.001	2.68
diabetes	40.1	1	<.001	3.38

Zo štatistík, p-hodnôt a u nominálnych premenných aj z odd ratio môžeme odčítať, že na hladine významnosti 5 % nie sú štatisticky významné atribúty heartRate a currentSmoker. Z uvedených

výsledkov môžeme súdiť, že tieto dva atribúty nebudú spoľahlivo predpovedať budúci výskyt srdečného ochorenia. S ostatnými atribútmi budeme pracovať v logistickej regresii.

Logistická regresia

Model bol vytvorený pomocou logistickej regresie; predikovali sme pravdepodobnosť vzniku ochorenia ISCH (závislá premenná: TenYearCHD) v závislosti na 13 vybraných premenných (viz výsledok welchovho a chi kvadrát testu). Po backward eliminácii zostalo v modeli osem významných prediktorov: pohlavie, vek, počet vyfajčených cigariet denne, výskyt mŕtvice, hypertenzia, celkový cholesterol, systolický krvný tlak a hladina glukózy.

Pri „backward" metóde sú na začiatku (v prvej iterácii) do modelu zaradené všetky nezávislé premenné, v každej ďalšej iterácii sú postupne odstraňované premenné s vysokými p-hodnotami (nad 0,05), čím sa docieli vyššia predikčná schopnosť modelu. V našom prípade sme k 8 prediktorom došli po 5 iteráciách, viz tabuľku č. 4.

Tabuľka 4. Logistická regresia (finálny model po 5.iterácii)

Predictor	Estimate	SE	Z	p	Odds ratio
male1	0.555132	0.107021	5.187	<0.001	1.742
age	0.065512	0.006441	10.171	<0.001	1.068
cigsPerDay	0.019641	0.004180	4.698	<0.001	1.020
prevalentStroke1	0.749423	0.483675	1.549	<0.001	2.116
prevalentHyp1	0.226882	0.135075	1.680	<0.001	1.255
totChol	0.002229	0.001122	1.986	<0.001	1.002
sysBP	0.014268	0.002857	4.995	<0.001	1.014
glucose	0.007319	0.001673	4.374	<0.001	1.007

Hodnota Odds ratio (OR) pre pohlavie (male1) má hodnotu 1,742, čo znamená, že muži majú o 74,2 % vyššiu pravdepodobnosť vzniku ochorenia v porovnaní so ženami. Pravdepodobnosť ochorenia sa s každým rokom veku zvyšuje o 6,8 %. Každá ďalšia cigareta denne zvyšuje riziko ochorenia o 2 %. Výskyt mŕtvice má OR = 2,116, čo naznačuje, že ľudia s anamnézou mozgovej príhody majú 2,1-násobne vyššie riziko vzniku ISCH. Ľudia s hypertenziou majú o 25,5 % vyššiu pravdepodobnosť ochorenia. Celkový cholesterol má OR = 1,002, čo naznačuje len mierny, ale štatisticky významný vplyv na riziko. Systolický krvný tlak (OR = 1,014); každé zvýšenie o 1 mmHg zvyšuje riziko ochorenia o 1,4 %. A nakoniec, hladina glukózy s OR = 1,007 znamená, že každé zvýšenie hladiny glukózy o 1 jednotku zvyšuje pravdepodobnosť ochorenia o 0,7 %.

Kvalita modelu

Pseudo R^2 = 0,1171, čo naznačuje, že model vysvetľuje približne 11,7 % variability v dátach. AIC modelu je 2776, pričom nižšie hodnoty AIC naznačujú lepšie prispôsobenie modelu pri zohľadnení

zložitosti. Deviance modelu je 2758, čo znamená, že model sa lepšie prispôsobil dátam v porovnaní s nulovým modelom.

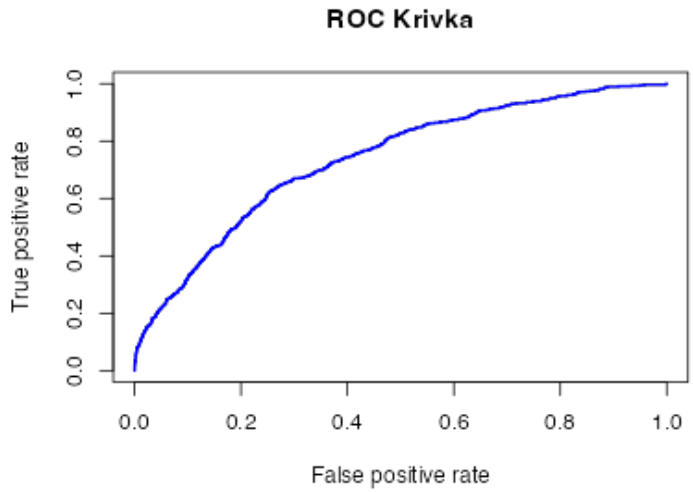
Tabuľka 5. Presnosť modelu (senzitivita a specificita)

	Predicted		
Observed	**0**	**1**	**% Correct**
0	3081	18	99.4
1	512	46	8.24

cut-off = 0.5

Senzitivita (Recall) je veľmi nízka, čo znamená, že model zachytil len 8,24 % skutočných pozitívnych prípadov (osôb s ochorením). Inými slovami, väčšinu ľudí s ochorením model nesprávne klasifikoval ako zdravých. Špecificita je extrémne vysoká, t.j. model správne identifikoval 99,42 % zdravých jedincov. Model teda dáva prednosť správnej klasifikácii zdravých osôb, ale zanedbáva pozitívne prípady, čo môže byť problematické pri diagnostike rizika ochorenia. Je zrejmé, že náš model je veľmi konzervatívny a takmer vždy predpokladá, že osoba nemá ochorenie, čím minimalizuje falošné pozitíva, ale zvyšuje falošné negatíva. V praxi by takýto model nebol ideálny na skríningové účely, pretože veľa ľudí s reálnym rizikom ochorenia by nebolo identifikovaných. Riešením môže byť zníženie rozhodovacieho prahu (cut-off value) z 0,5 na nižšiu hodnotu (napr. 0,3), čím by sa zvýšila senzitivita za cenu mierneho zníženia špecificity.

Graf 1. ROC krivka



AUC (Area Under Curve) je 0,7379, t.j., že model má strednú až dobrú prediktívnu schopnosť (hodnoty nad 0,7 sa považujú za prijateľné).

Zoznam použitých zdrojov a literatúry

Dileep070. (n.d.). *Heart disease prediction using logistic regression* [Dataset]. Kaggle. Retrieved February 28, 2025, from https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/data

Fox, J., & Weisberg, S. (2023). *car: Companion to Applied Regression*. [R package]. Retrieved from https://cran.r-project.org/package=car.

R Core Team (2024). *R: A Language and environment for statistical computing*. (Version 4.4) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from CRAN snapshot 2024-08-07).

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T., Unterthiner, T., & Ernst, F. G. M. (2020). *ROCR: Visualizing the Performance of Scoring Classifiers*. [R package]. Retrieved from https://cran.r-project.org/package=ROCR.

The jamovi project (2024). *jamovi*. (Version 2.6) [Computer Software]. Retrieved from https://www.jamovi.org.