Predikce diabetu u kmene Pima – Role věku a genetické zátěže: ROC analýza a cut-off skóre

Úvod

Diabetes mellitus 2. typu patří mezi nejčastější metabolická onemocnění. Vede k akutním i chronickým komplikacím, které významně zhoršují kvalitu života pacientů a zvyšují jejich úmrtnost. V počátečních fázích je charakterizován relativním nedostatkem inzulinu, což vede ke zhoršenému využití glukózy a zvýšení její hladiny v krvi, tzv. **hyperglykémii**. Onemocnění má **genetický základ**, přičemž jeho rozvoj je ovlivňován také vnějšími faktory (Škrha, 2009).

Bennett et al. v roce 1971 píše, že výsledky glukózových tolerančních testů ukazují, že původní Američané z kmene Pima jsou **populací s dosud nejvyšší zaznamenanou prevalencí diabetu**. Při použití konzervativních kritérií dosahovala u osob **starších 35 let** přibližně **50 %**.

Mojí hlavní otázkou, na kterou jsem si chtěla odpovědět bylo: **Jak dobře dokážeme predikovat diabetes melitus 2. typu, když známe pouze rodinnou anamnézu (tedy genetické riziko)?** Pro lepší představu jsem použila další proměnou **Věk** a srovnala obě **ROC křivky**. Z literatury víme, že věk je jedním z **podstatných prediktorů cukrovky** (Jain et al., 2024). Proto mě zajímalo, jestli Genetická zátěž jeho **hodnotu AUC** (Area Under the Curve) překoná, a bude tak dosahovat vyšší diskriminační schopnosti. Cílem této práce je také pro obě proměnné stanovit **cut-off skóre**, které by kvantifikovalo hranici, nad kterou už bychom mohli předpokládat **rizikovost pro vznik diabetu**. Pro úplnost jsem také otestovala nulovou hypotézu o tom, že hodnota genetické zátěže nedokáže rozlišovat mezi skupinami (rizikoví/nerizikoví) pomocí **Mann-Whitneyho U-testu**.

Metoda

Pro účely **ROC analýzy** (Receiver Operating Characteristic) jsem využila veřejně dostupná data ze stránky **Kaggle**. Původním zdrojem je **National Institute of Diabetes and Digestive and Kidney Diseases**. Jako hlavní zdroj je uveden článek Smitha et al. (1988), pracující s těmito daty.

Všichni zařazení respondenti byly **ženy mezi 21 a 81 lety (N=768)**. Všechny pocházejí z **kmene Pima** (původní obyvatelé Ameriky). **Prevalence** (podíl osob, u nichž se vyskytlo onemocnění) byla **35 %**.

Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=788.

Dataset obsahuje několik medicínských **prediktorů** a jednu **kriteriální (závislou) proměnnou**, Outcome. Prediktory zahrnují počet těhotenství respondentky, její BMI, hladinu inzulinu, věk a další charakteristiky.

Jak z názvu vyplývá, **pro účely ROC analýzy byly zvoleny následující proměnné:**

- **Genetická zátěž (Diabetes pedigree function):**

  Jedná se o funkci vypočítanou podle **speciálně vyvinutého vzorce**. Poskytuje souhrnné informace o **výskytu diabetes mellitus u příbuzných** a o **genetickém vztahu těchto příbuzných k dané osobě**. Hodnota ukazatele DPF roste s počtem příbuzných, u nichž se diabetes mellitus objevil, s nižším věkem, ve kterém u nich onemocnění vzniklo, a s vyšším podílem sdílené genetické informace. Naopak klesá s rostoucím počtem příbuzných, kteří diabetem netrpí, s vyšším věkem, kterého dosáhli při posledním vyšetření, a rovněž s vyšším podílem sdílených genů.

- **Věk (Age)**
- **Outcome** (Závislá proměnná)

Pro úplnost dodávám, že **cut-off skóre představuje hranici**, při jejímž překročení je výsledek považován za pozitivní, zatímco při jejím nedosažení je výsledek hodnocen jako negativní.

Pro každou cut-off hodnotu byly stanoveny následující ukazatele:

- **TP (true positives):** počet jedinců **správně identifikovaných** jako osoby se **zvýšeným rizikem diabetu**
- **FP (false positives):** počet jedinců **chybně označených** jako osoby se **zvýšeným rizikem diabetu**
- **TN (true negatives):** počet jedinců **správně identifikovaných** jako osoby **bez zvýšeného rizika diabetu**
- **FN (false negatives):** počet jedinců **chybně označených** jako osoby **bez zvýšeného rizika diabetu**

Další užité **pomocné ukazatele** byly:

- **Senzitivita** (citlivost) vyjadřuje **schopnost testu správně identifikovat jedince se zvýšeným rizikem diabetu**, tedy pravděpodobnost, že cut-off hodnota správně označí osoby, u nichž se diabetes skutečně vyskytuje.

- **Specificita** naopak udává **schopnost testu správně rozpoznat jedince bez zvýšeného rizika diabetu**, tedy pravděpodobnost, že cut-off hodnota správně identifikuje osoby bez diabetu.
- Na základě senzitivity a specificity byla pro jednotlivé cut-off hodnoty vypočtena **Youdenova statistika (J)**, která slouží k určení optimální hranice. **Vyšší hodnota této statistiky značí lepší diskriminační schopnost daného cut-off skóre.** Tento ukazatel však **předpokládá vyvážené zastoupení obou skupin** (tj. prevalenci kolem 50 %), což nemusí odpovídat reálným datům.

- Proto byla doplněna také **statistika I**, která **zohledňuje skutečné zastoupení obou skupin** v souboru. **Vyšší hodnota této statistiky rovněž ukazuje na vhodnější cut-off skóre**, avšak s ohledem na reálnou prevalenci sledovaného jevu.

Výsledky

Následující tabulka shrnuje popsané ukazatele. Jelikož však může genetická zátěž nabývat velkého počtu hodnot (tabulka má 528 řádků), uvádím pro srovnání pouze nejvyšší hodnoty J a I s korespondujícími cut-off skóry. Celou podobu dat naleznete v Excel souboru. Proměnná Věk je v Tabulce 2 prezentována už standartním způsobem.

	Cut-off	TP	FP	TN	FN	Senzitivita	Specificita	1-Senz	1-Spec	J	I
Youdenovo J	0,502	126	150	350	142	47 %	70 %	53 %	30 %	0,170	61,9792 %
Statistika I	1,114	25	16	484	243	9 %	97 %	91 %	3 %	0,061	66,2760 %

Tab. 1: Cut-off skóre: Genetická zátěž

Z Tabulky 1 je patrné, že na základě J statistiky bychom mohli za **nejvhodnější cut-off skóre** považovat hodnotu **0,502**. Statistika I ale říká, že vhodnější by mohla být hodnota genetické zátěže **1,114**. To je poměrně velký rozdíl mezi oběma ukazateli, daný právě předpokládanou prevalencí u Youdenovy statistiky.

Cut-off	TP	FP	TN	FN	Senzitivita	Specificita	1-Senz	1-Spec	J	I
21	268	500	0	0	100%	0%	0%	100%	0,000	34,896%
22	263	442	58	5	98%	12%	2%	88%	0,097	41,797%
23	252	381	119	16	94%	24%	6%	76%	0,178	48,307%
24	245	350	150	23	91%	30%	9%	70%	0,214	51,432%
25	237	312	188	31	88%	38%	12%	62%	0,260	55,339%
26	223	278	222	45	83%	44%	17%	56%	0,276	57,943%
27	215	253	247	53	80%	49%	20%	51%	0,296	60,156%
28	207	229	271	61	77%	54%	23%	46%	0,314	62,240%
29	197	204	296	71	74%	59%	26%	41%	0,327	64,193%
30	184	188	312	84	69%	62%	31%	38%	0,311	64,583%
31	178	173	327	90	66%	65%	34%	35%	0,318	65,755%
32	165	162	338	103	62%	68%	38%	32%	0,292	65,495%

Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=788.

33	156	155	345	112	58%	69%	42%	31%	0,272	65,234%
34	146	148	352	122	54%	70%	46%	30%	0,249	64,844%
35	142	138	362	126	53%	72%	47%	28%	0,254	65,625%
36	137	133	367	131	51%	73%	49%	27%	0,245	65,625%
37	127	127	373	141	47%	75%	53%	25%	0,220	65,104%
38	121	114	386	147	45%	77%	55%	23%	0,223	66,016%
39	111	108	392	157	41%	78%	59%	22%	0,198	65,495%
40	108	99	401	160	40%	80%	60%	20%	0,205	66,276%
41	102	92	408	166	38%	82%	62%	18%	0,197	66,406%
42	89	83	417	179	33%	83%	67%	17%	0,166	65,885%
43	82	72	428	186	31%	86%	69%	14%	0,162	66,406%
44	71	70	430	197	26%	86%	74%	14%	0,125	65,234%
45	66	67	433	202	25%	87%	75%	13%	0,112	64,974%
46	58	60	440	210	22%	88%	78%	12%	0,096	64,844%
47	51	54	446	217	19%	89%	81%	11%	0,082	64,714%
48	47	52	448	221	18%	90%	82%	10%	0,071	64,453%
49	46	48	452	222	17%	90%	83%	10%	0,076	64,844%
50	43	46	454	225	16%	91%	84%	9%	0,068	64,714%
51	38	43	457	230	14%	91%	86%	9%	0,056	64,453%
52	33	40	460	235	12%	92%	88%	8%	0,043	64,193%
53	26	39	461	242	10%	92%	90%	8%	0,019	63,411%
54	22	38	462	246	8%	92%	92%	8%	0,006	63,021%
55	18	36	464	250	7%	93%	93%	7%	-0,005	62,760%
56	17	33	467	251	6%	93%	94%	7%	-0,003	63,021%
57	15	32	468	253	6%	94%	94%	6%	-0,008	62,891%
58	14	28	472	254	5%	94%	95%	6%	-0,004	63,281%
59	11	24	476	257	4%	95%	96%	5%	-0,007	63,411%
60	9	23	477	259	3%	95%	97%	5%	-0,012	63,281%
61	7	20	480	261	3%	96%	97%	4%	-0,014	63,411%
62	6	19	481	262	2%	96%	98%	4%	-0,016	63,411%
63	4	17	483	264	1%	97%	99%	3%	-0,019	63,411%
64	4	13	487	264	1%	97%	99%	3%	-0,011	63,932%
65	4	12	488	264	1%	98%	99%	2%	-0,009	64,063%
66	4	9	491	264	1%	98%	99%	2%	-0,003	64,453%
67	2	7	493	266	1%	99%	99%	1%	-0,007	64,453%
68	1	5	495	267	0%	99%	100%	1%	-0,006	64,583%
69	1	4	496	267	0%	99%	100%	1%	-0,004	64,714%
69	1	4	496	267	0%	99%	100%	1%	-0,004	64,714%
70	1	2	498	267	0%	100%	100%	0%	0,000	64,974%
72	0	2	498	268	0%	100%	100%	0%	-0,004	64,844%
81	0	1	499	268	0%	100%	100%	0%	-0,002	64,974%

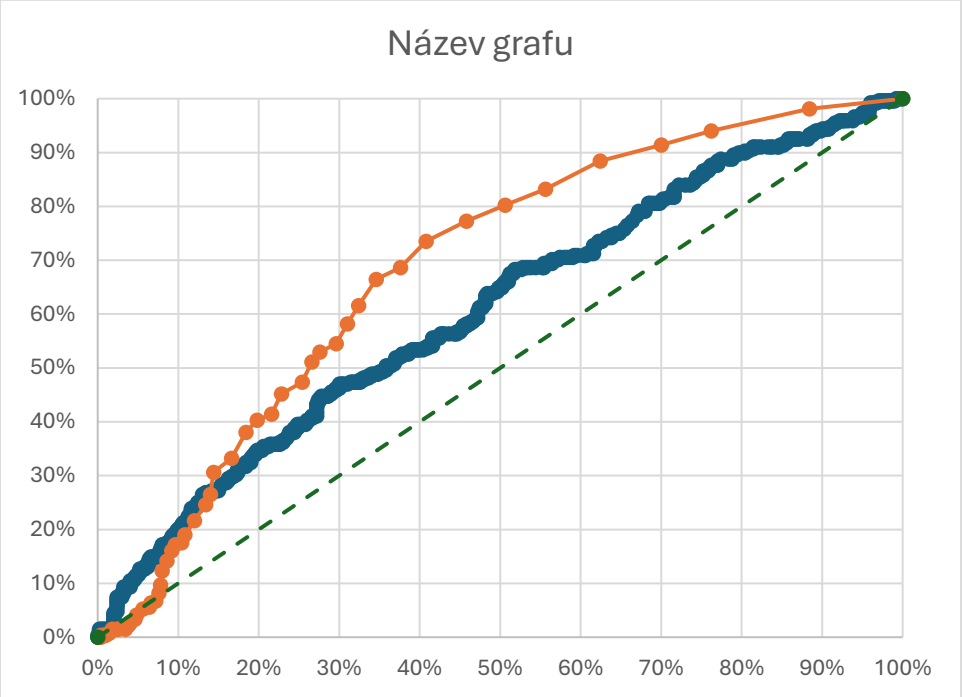
Tab. 2: Cut-off skóre: Věk

Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=788.

Z Tabulky 2 vyplývá, že statistika J stanovuje **nejlepší cut-off** na **29 let** a statistika I na **41** nebo **43 let**.

Níže vkládám také graf **ROC křivek**. Oranžovou barvou je vyznačena proměnná Věk a modrou Genetická zátěž. Čím více se křivky odchlipují směrem k levému hornímu rohu, tím více je test (proměnná) efektivní při predikci diabetu. Zelená přerušovaná čára znázorňuje efektivitu náhodného rozřazování (např. hod mincí).

*Graf **1:** ROC křivky*



Plochu pod křivkou označujeme jako **AUC**. Tento ukazatel nám říká, jak dobře daná proměnná dokáže rozlišit mezi skupinami (nemocní vs. zdraví). Hodnota AUC pro **Věk** je **0,687**, což značí **poměrně dobrou schopnost predikce**. Pro **Genetickou zátěž** je to **0.606**, což vypovídá spíše o **střední diskriminační schopnosti**. **Nulovou hypotézu o tom, že hodnota genetické zátěže nedokáže rozlišovat mezi skupinami**, tedy, že nedokáže predikovat rozvoj diabetu, můžeme ale na základě **Mann-Whitneyho U-testu** s vysokou statistickou jistotou **zamítnout** ($U = 52769$ při $p < 0{,}001$).

Závěr

Výsledky ukázaly, že **hodnota AUC Genetické zátěže hodnotu AUC pro Věk nepřekonala**. Analýza podle očekávání **potvrdila dobrou diskriminační schopnost u proměnné Věk**, ale u **Genetické zátěže ukázala na predikční potenciál pouze střední**. Také v rámci Grafu 1 můžeme vidět, že se oranžová křivka pro Věk odchlipuje od úhlopříčky více, než je tomu u modré křivky Genetické zátěže. Pomocí Mann-Whitneyho U-testu však můžeme **zamítnout**

nulovou hypotézu, která říká, že hodnota genetické zátěže nedokáže rozlišovat mezi skupinami rizikových a nerizikových osob.

Pro obě proměnné se také povedlo stanovit **cut-off skóry**, které kvantifikují hranici, nad kterou už bychom mohli předpokládat **rizikovost pro vznik diabetu.** Pro **Věk** by touto hranicí mohlo být **29 let (J)** nebo **41** či **43 let (I)**. Pro **Genetickou zátěž** je to buď **0,502 (J) nebo 1,114 (I).** Které z těchto hranic by byly zvoleny, závisí na typu a účelu případné intervence. Preventivní programy mohou upřednostňovat nižší cut-ff hodnotu, zatímco pro epidemiologické studie může být vhodnější zvolit hranici vyšší.

Seznam literatury:

Bennett, P. H., Burch, T. A., & Miller, M. (1971). Diabetes mellitus in american (pima) indians. *The Lancet*, 298(7716), 125-128. https://doi.org/https://doi.org/10.1016/S0140-6736(71)92303-8

Jain, R., Tripathi, N. K., Pant, M., Anutariya, C., & Silpasuwanchai, C. (2024). Investigating Gender and Age Variability in Diabetes Prediction: A Multi-Model Ensemble Learning Approach. *IEEE Access*, *12*(1), 71535-71554. https://doi.org/10.1109/ACCESS.2024.3402350

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Symposium on Computer Applications in Medical Care*, 261–265. https://doi.org/0195-4210/88/0000/0261$01.00

Škrha, J. (2009). *Diabetologie*. Galén.