Predikce diabetu u žen starších 21 let indiánského původu Pima

---

1. Úvod, ukotvení

DM patří mezi nejrozšířenější chronická onemocnění současnosti a představuje významný zdravotní i společenský problém. Včasná identifikace rizikových faktorů může výrazně přispět k prevenci tohoto onemocnění. K jeho vzniku ve většině případů přispívá život ve stresu, nadměrný příjem energie, nevhodné složení potravy a nedostatek pohybu, tedy faktory, které také velmi často vedou k obezitě. Nevhodný životní styl však může uspíšit vznik diabetu a komplikovat průběh onemocnění.

Použitý dataset je dostupný z:
https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?utm_source=chatgpt.com
“Tato datová sada pochází původně z Národního institutu pro diabetes a onemocnění trávicího traktu a ledvin. Cílem této datové sady je diagnosticky předpovědět, zda pacient trpí diabetem, na základě určitých diagnostických měření zahrnutých v datové sadě. Na výběr těchto případů z větší databáze bylo kladeno několik omezení. Zejména všechny pacientky zde jsou ženy starší 21 let indiánského původu Pima.”
Dataset obsahuje i proměnné jako je počet těhotenství, velikost kožní řasy a podobně, já budu pracovat pouze s několika vybranými z nich.

Cílem této práce je zanalyzovat vztah mezi vybranými faktory (např. BMI, věk, hladina glukózy…) a pravděpodobností výskytu diabetu pomocí logistické regrese. Ve své práci použiji jak logistickou regresi, tak deskriptivní statistiku a korelaci mezi proměnnými.

K analyzování bylo využito tabulek Excel a SPSS program.

2. Průběh

Analýza dat byla provedena na datasetu obsahujícím 200 respondentů. Cílem bylo zjistit vztah mezi vybranými proměnnými (BMI, hladina glukózy a věk) a výskytem diabetu.

Nejprve byla provedena deskriptivní statistika jednotlivých proměnných. U proměnné BMI byl zjištěn průměr přibližně 31, přičemž minimální hodnota byla okolo 18 a maximální dosahovala hodnoty přibližně 50. To naznačuje, že se většina respondentů nachází v pásmu nadváhy až obezity.

2.1 Deskriptivní statistika

Proměnná	N	Průměr	sm. odchylka	min.	medián	max.
Glukoza	200	120,38	30,62	70	122,16	200
BMI	200	31,05	7,34	18	30,84	50
Věk	200	46,46	15,45	21	45,50	74
Inzulín	200	113,12	77,78	15	100,12	300

Outcome: 0=NEMÁ, 1=MÁ, V souboru je 61,5 % respondentů s diabetem a 38 % bez diabetu.

U proměnné hladiny glukózy byl zjištěn průměr přibližně 120, minimum kolem 70 a maximum přibližně 200. Rozptyl hodnot ukazuje na značnou variabilitu mezi respondenty.

Průměrný věk respondentů byl přibližně 46 let, přičemž věkové rozmezí sahalo zhruba od 21 do 74 let.

Dále byla analyzována závislá proměnná Outcome, která udává výskyt diabetu (0 =NEMÁ, 1 = MÁ). Pomocí funkce COUNTIF v Excel bylo zjištěno, že byl průměr: 0,615
→ cca 61,5 % má diabetes, 38,5 % nemá

2.2 Korelace s proměnnou Outcome

Pro vztah mezi binární proměnnou Outcome a jednotlivými metrickými proměnnými byl použit bodově-biseriální korelační koeficient.

Proměnná	Korelace r	Sig.	Interpretace
Glukóza	0,398	<0,001	středně pozitivní
BMI	0,024	0,737	slabý vztah
Věk	0,078	0,274	slabý vztah
Inzulín	-0,048	0,498	slabý vztah

Nejsilnější vztah s výskytem diabetu má hladina glukózy ($r = 0,398$; $p < 0,001$). BMI, věk ani inzulín nevykázaly v tomto datasetu statisticky významnou korelaci.

2.3. Logistická regrese

Závislá proměnná: Outcome (0 = bez diabetu, 1 = s diabetem).
Nezávislé proměnné: Glukóza, BMI, Věk.

Test celkového modelu

-2 Log Likelihood	Chi-square	df	Sig.
231,703	34,880	3	<0,001

Model je jako celek statisticky významný ($\chi^2 = 34.880$, $df = 3$, $p < 0,001$), což znamená, že alespoň jedna z prediktorových proměnných významně přispívá k predikci diabetu.

Souhrn modelu

-2 Log Likelihood	Cox & Snell R^2	Nagelkerke R^2	AUC
231,703	0,160	0,217	0,742

Nagelkerke $R^2 = 0,217$ ukazuje, že model vysvětluje přibližně 21,7 % variability výskytu diabetu. AUC = 0,742 značí přijatelnou diskriminační schopnost modelu.

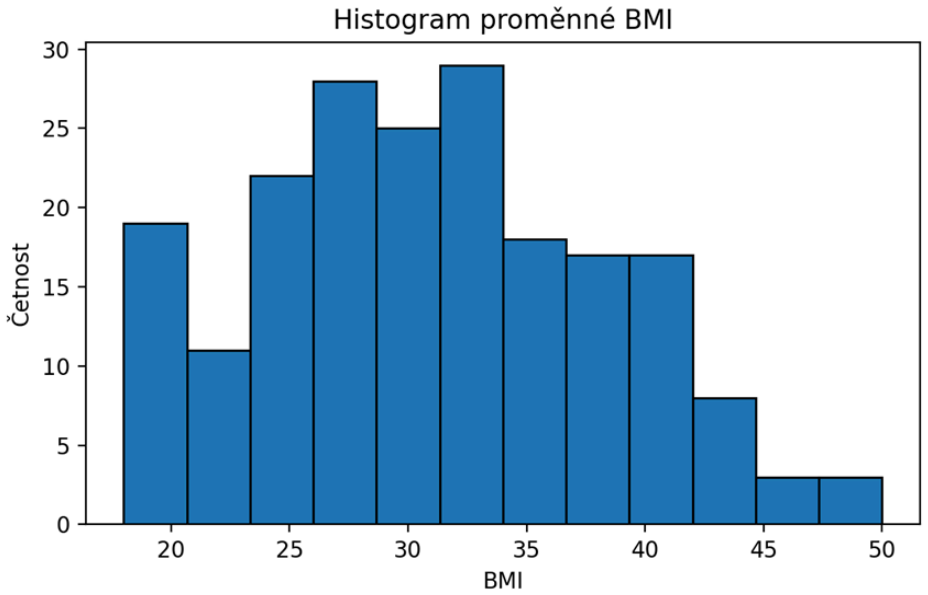
Proměnné v rovnici

Proměnná	B	S.E.	Wald	df	Sig.	Exp(B)	
Konstanta	-3,836	1,086	12,472	1	<0,001	0,022	
Glukóza	0,031	0,006	27,247	1	<0,001	1,032	
BMI	0,011	0,022	0,278	1	0,598	1,011	
Věk	0,007	0,010	0,404	1	0,525	1,007	

Rovnice modelu: logit(p) = -3,836 + 0,031·Glukóza + 0,011·BMI + 0,007·Věk.
Interpretace: při zvýšení glukózy o 1 jednotku se šance na výskyt diabetu zvýší přibližně 1,032× (tj. o 3,2 %), pokud ostatní proměnné zůstávají stejné. BMI ani věk v tomto modelu nebyly statisticky významné.

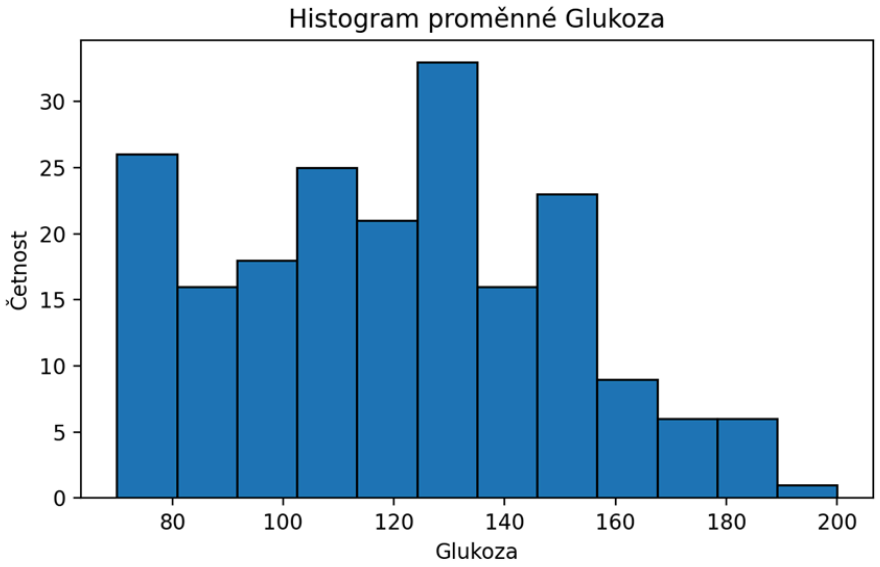
4. Grafy

BMI je koncentrováno hlavně v pásmu nadváhy a obezity, přibližně mezi 25 a 35.



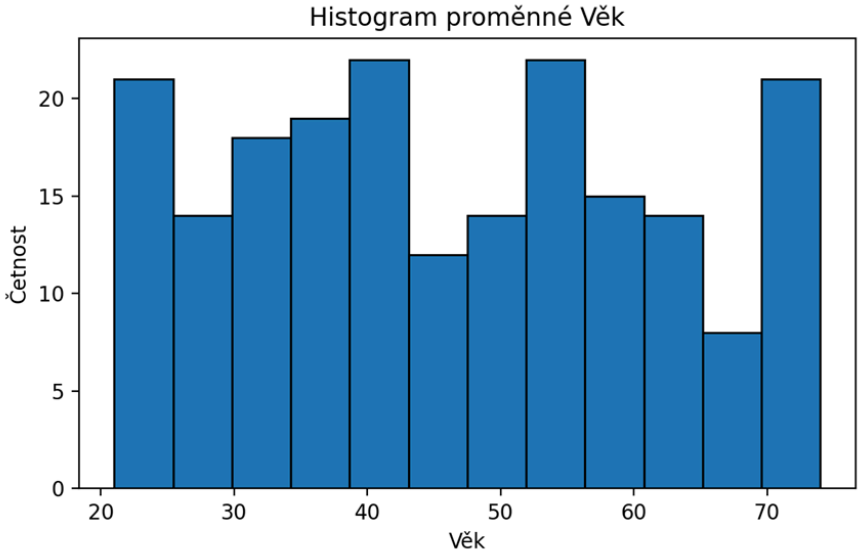
Obrázek č.1

Rozdělení hodnot glukózy je poměrně široké, nejvyšší četnost se soustřeďuje zhruba mezi 100 a 140.



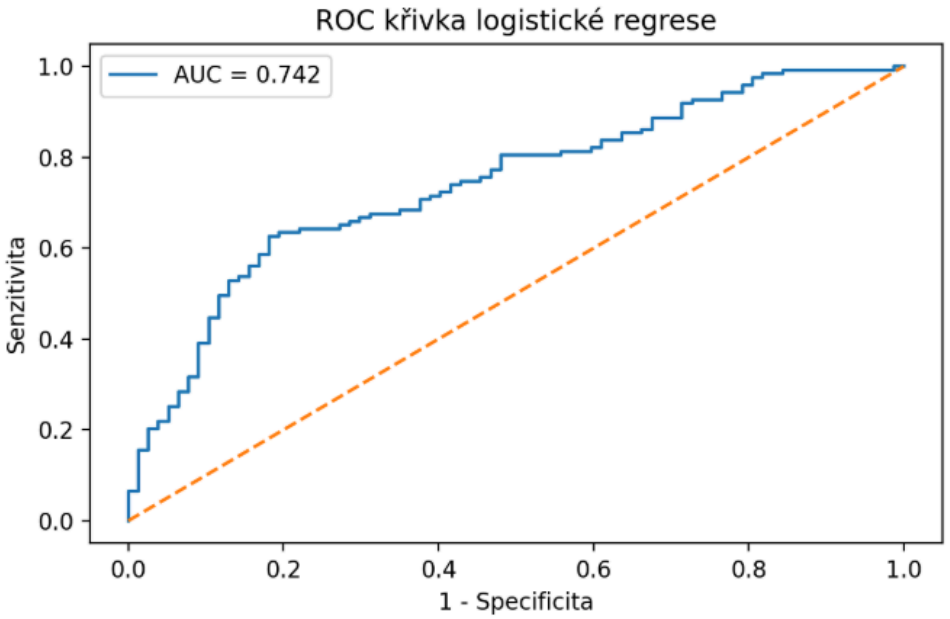
Obrázek č.2

Věk respondentů pokrývá široké rozmezí



Obrázek č.3

ROC křivka s AUC 0,742 potvrzuje přijatelnou schopnost modelu odlišovat osoby s diabetem a bez diabetu.



Obrázek č.4

5. Závěr a diskuze

Korelační analýza i logistická regrese ukázaly, že nejsilnějším prediktorem výskytu diabetu je v tomto datasetu hladina glukózy. Bodově-biseriální korelace mezi glukózou a Outcome dosáhla hodnoty r = 0,398 (p < 0,001). V logistickém modelu byla glukóza jedinou statisticky významnou proměnnou (B = 0,031; p < 0,001; OR = 1,032). BMI a věk sice byly do modelu zahrnuty, ale jejich vliv se jako statisticky významný nepotvrdil. Celková úspěšnost klasifikace činila 66,5 % a AUC 0,742, což znamená, že model má přijatelnou, predikční schopnost.

Omezením je relativně malý rozsah použitého datasetu a absence dalších potenciálně významných proměnných, jako je fyzická aktivita nebo genetické predispozice.

Doporučuji dále literaturu, která se dané problematice věnuje, využít můžete nějakou z mých zdrojů níže.

6. Zdroje

Diabetes mellitus. Online. 2026. Dostupné z: https://www.ikem.cz/cs/diabetes-mellitus/a-4443/ . [cit. 2026-04-09].

UCI Machine Learning Repository. (2019). *Pima Indians Diabetes Database*. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Zimmet, P., Alberti, K. G. M. M., & Shaw, J. (2001). Global and societal implications of the diabetes epidemic. *Nature, 414*(6865), 782–787.

American Diabetes Association. (2023). *Classification and diagnosis of diabetes: Standards of medical care in diabetes—2023*. Diabetes Care, 46(Supplement_1), S19–S40.

Hu, F. B., Manson, J. E., Stampfer, M. J., Colditz, G., Liu, S., Solomon, C. G., & Willett, W. C. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England Journal of Medicine, 345*(11), 790–797.