VÝBĚR PREDIKTIVNÍHO MODELU PRO MALIGNÍ A BENIGNÍ NÁDORY PRSU

Teoretické ukotvení

Zhoubný (maligní) nádor prsu je u žen nejrozšířenější formou rakoviny. V poslední době nádorů přibývá, ale stejně tak se postupně vyvíjí léčba, tudíž je úmrtnost na tento typ nádorového onemocnění menší něž v dřívějších letech (Chodounská & Lojková, 2026). Přesto je ale včasná a správná diagnostika zásadní. K tomu se využívají různé charakteristiky nádorů či jejich buněk jako je velikost samotného nádoru, jestli prostupuje do jiné tkáně a podobně (Rakha et al., 2008). Novější výzkumy se také zaměřují na různé metody zabývající se vlastnostmi nádorových kmenových buněk, které mají vliv na růst nádoru i na jeho odolnost vůči terapii (Khan et al., 2025).

Zájem zkoumání charakteristik nádorových buněk tedy není nijak nový. V tomto projektu vycházíme z datasetu publikovaného v článku z roku 1993, kde ve svém výzkumu pracovali s 10 charakteristikami jader nádorových buněk (William Wolberg, 1993). Cílem tohoto projektu bylo vytvořit prediktivní model s určitým počtem proměnných, charakteristik jader nádorových buněk, jejichž pomocí bychom mohli odhadnout, zda bude daný nádor maligní či benigní.

Data

Původní data byla získána na základě analýzy snímků buněk z nádorů prsů získaných pomocí metody „fine aspiration needle“, která je méně intenzivní než celková biopsie, která se využívá k diagnostice malignity nádorů. Hodnoty vybraných charakteristik nebyly posuzovány subjektivně lékaři, ale matematickým algoritmem (William Wolberg, 1993).

Původní dataset obsahuje 10 proměnných, jejichž původní název z datasetu i český popis jsme vložili do přehledné tabulky č.1. Při analýze dat i prezentaci výsledků v tabulkách byly ponechány názvy proměnných v angličtině. Proměnné, se kterými jsme později pracovali v rámci menšího modelu jsou pro lepší přehlednost v tabulce vyznačeny tučně. (Popisy proměnných byly zkráceny pro účely této zprávy a tabulky.)

Tabulka 1: původní proměnné a jejich český popis

Proměnná (charakteristiska nádoru)	Český popis
radius	poloměr jádra
texture	**textura jádra**
perimeter	obvod jádra
area	plocha jádra
smoothness	**hladkost obrysu jádra, rozdíl mezi poloměrem a průměrnou délkou přímek, které ho obklopují**
compactness	kompaktnost buněčných jader, poměr obvodu k ploše
concavity	počet a výraznost prohlubní v jádru
concave points	počet bodů, kde se se hranice nádoru prohýbá dovnitř
symmetry	**symetrie**
fractal dimension	„zubatost" obrysu jádra (Mandelbrotova aproximace)

Analýza dat

Původním záměrem práce bylo vytvořit prediktivní model za pomoci binomiální logistické regrese, v níž bychom využili všech 10 proměnných, které se nachází v původním datasetu. Po vytvoření prvního modelu a ověření potenciální multikolinearity pomocí variance inflation factor (VIF) jsme ale museli 7 z 10 původních proměnných vyřadit, jelikož se jejich hodnoty VIF pohybovaly nad hodnotou 5, kterou jsme si stanovili jako hraniční. Všechny hodnoty jsou znázorněny v tabulce č.2. Proměnné, se kterými jsme pracovali v rámci dalšího modelu jsou zvýrazněny tučně.

Tabulka 2: ověření multikolinearity v původním modelu s 10 proměnnými

Proměnná (charakteristiska nádoru)	VIF
radius	899.522
texture	**1.806**
perimeter	698.983
area	129.559
smoothness	**4.373**
compactness	15.281
concavity	5.259
concave points	5.856
symmetry	**1.839**
fractal dimension	9.788

Po vyřazení velké části proměnných jsme vytvořili prediktivní model, u kterého jsme pro větší přesnost vypočítali hodnoty VIF znovu. Mezi zvolenými proměnnými se už nenacházela žádná multikolinearita. Výsledky ověření multikolinearity jsou prezentovány v tabulce č. 3.

Tabulka 3: ověření multikolinearity v modelu se 3 proměnnými

Proměnná (charakteristiska nádoru)	VIF
texture	1.155
smoothness	1.368
symmetry	1.248

Výsledky logistické regrese tohoto modelu nám ukázaly, že všechny 3 proměnné jsou pro nás signifikantní, což ukazují hodnoty v tabulce č. 4.

Tabulka 4: výsledky analýzy modelu se 3 proměnnými

	Odhad parametru	St. chyba	Waldova statistika	p
(Intercept)	-15.203	1.355	-11.217	<0.001
texture	0.286	0.030	9.509	<0.001
smoothness	63.057	9.951	6.337	<0,001
symmetry	15.713	4.732	3.321	<0.001

Problém ale nastává v momentě, kdy chceme spočítat i tzv. odds ratios (OR; pravděpodobnosti šancí) pro náš model. Některé výsledné hodnoty jsou příliš vysoké (OR(smoothness) =2.429068e+27; OR(symmetry) = 6666203), což nám ukazuje numerickou nestabilitu, se kterou nemůžeme v rámci této zprávy pracovat. Z tohoto důvodu jsme se rozhodli vytvořit výsledný prediktivní model s jediným prediktorem, který se ukázal jako dost stabilní a statisticky významný. Výsledky analýzy jsou uvedeny v následující tabulce.

Tabulka 5: výsledky analýzy modelu s jediným prediktorem (textura)

	Odhad parametru	St. chyba	Waldova statistika	Odds ratio	p
(Intercept)	-5.126	0.526	-9.738	0.006	<0.001
texture	0.235	0.026	8.975	1.264	<0.001

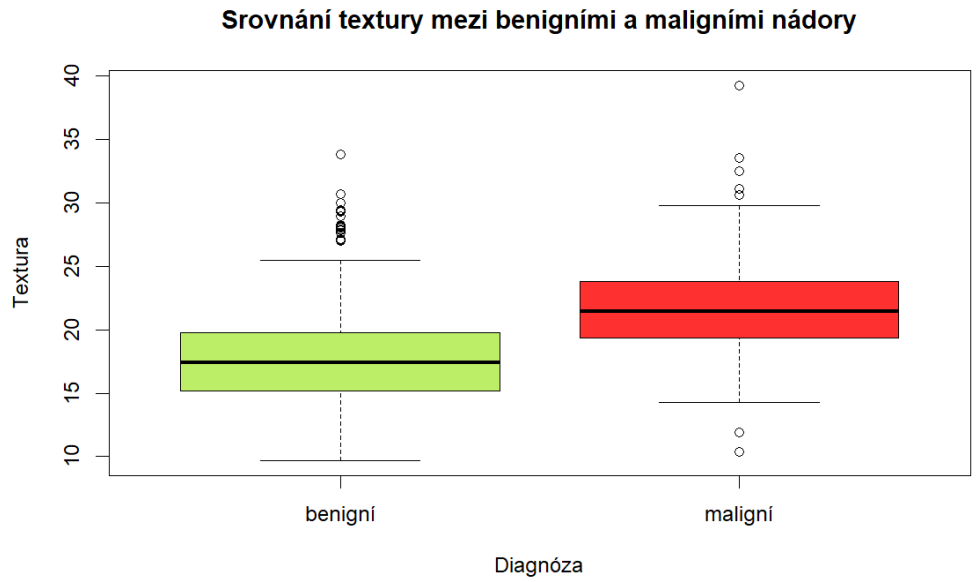
Proměnná textura je v této situaci signifikantní a na 5% hladině významnosti se její odds ratio (pravděpodobnost šance) pohybuje v intervalu [1.203-1.333]. Na základě našich výsledků můžeme říct, že s každou přibývající jednotkou v rámci textury se zvyšuje riziko malignity nádoru prsu 1,264krát (tj. o 26 % více než předtím). Zároveň jsme také spočítali R2 našeho modelu – viz tabulka č. 6. Tyto hodnoty bohužel vychází slabě, což bylo ale očekáváno na základě zmenšení modelu a odebrání téměř všech proměnných.

Tabulka 6: hodnoty R2 u modelu s jediným prediktorem (textura)

	R^2
McFadden	0.288
Cox & Snell	0.316
NagelkerkeR2	0.432

Pro lepší přehlednost uvádíme ještě krabicový graf, ve kterém je znázorněn vliv textury na výslednou diagnózu nádoru prsu jako maligní či benigní. Jsou zde patrné vyšší hodnoty v textuře u maligních nádorů, což koresponduje s našimi výsledky (odds ratio textury je vyšší než 1).

Graf 1: Srovnání proměnné textura mezi benigními a maligními nádory



Diskuze a závěr

Velkým omezením našeho výsledného modelu jsou nízké hodnoty R2 (viz tabulka č. XX), které jsme ale očekávali vzhledem ke zmenšení počtu našich proměnných. Celkově je největší nevýhodou to, že jsme z modelu s 10 proměnnými postupně přešli na model s jediným prediktorem. Naopak předností našeho postupu je, že jsme pracovali poctivě a nepřehlíželi problémy s našimi dřívějšími modely. Díky tomu je náš výsledný prediktor poměrně stabilní.

Naše výsledky se tedy mohou zdát na první pohled neúspěšné, ale opak je pravdou. Cílem této práce nebylo vytvořit co nejpřesnější a nejvhodnější prediktivní model, ale spíše ukázat postup, který můžeme v takovéto situace využít. Pokud bychom se snažili o výběr či tvorbu již zmíněného „nejpřesnějšího a nejvhodnějšího modelu“, přesahovalo by to rámec této zprávy a vyžadovalo by to náročnější metody a postupy. Naše výsledky ukazují nejen proces, ale i zdánlivý neúspěch ve statistickém výzkumu, který je v realitě běžný a musíme s ním počítat. I přesto se nám ale povedlo vytvořit model s jedním prediktorem, který je pro nás statisticky významný.

Seznam literatury

Chodounská, H., & Lojková, R. (2026, 20. ledna). *Nádorů přibylo, ale léčba se zlepšila*. Statistika a My – Český statistický úřad. https://statistikaamy.csu.gov.cz/nadoru-pribylo-ale-lecba-se-zlepsila

Khan, A. Q., Touseeq, M., Rehman, S., Tahir, M., Ashfaq, M., Jaffar, E., & Abbasi, S. F. (2025). Advances in breast cancer diagnosis: A comprehensive review of imaging, biosensors, and emerging wearable technologies. Frontiers in Oncology, 15, 1587517. https://doi.org/10.3389/fonc.2025.1587517

Rakha, E. A., Reis-Filho, J. S., & Ellis, I. O. (2008). Basal-Like Breast Cancer: A Critical Review. *Journal of Clinical Oncology*, *26*(15), 2568–2581. https://doi.org/10.1200/JCO.2007.13.1748

William Wolberg, O. M. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B