

Predikce rizika onemocnění srdce

Dataset získán z: <https://www.kaggle.com/datasets/noeysislearning/framingham-heart-study>

Cílem této práce je identifikace klíčových rizikových faktorů, které přispívají k rozvoji ischemické choroby srdeční v desetiletém horizontu. Analýza vychází z dat prestižní **Framingham heart study**, což je jedna z nejvýznamnějších longitudinálních studií v historii medicíny. Logistická regrese byla zvolena jako optimální statistický nástroj pro modelování binární závislé proměnné (přítomnost či absence onemocnění) na základě demografických ukazatelů a faktorů životního stylu.

Datový soubor

Analýzovaný dataset pochází z longitudinální kohortové studie **Framingham heart study**, zaměřené na identifikaci rizikových faktorů kardiovaskulárních onemocnění. Studie sleduje zdravotní stav účastníků v čase a umožňuje analyzovat vliv demografických, klinických a behaviorálních faktorů na rozvoj srdečních onemocnění.

Datový soubor byl získán z webové stránky Kaggle a obsahuje výpovědi 4 240 respondentů, z nichž 644 (15,2 %) prodělalo koronární srdeční příhodu během 10letého sledování. Protože je závislá proměnná binární (výskyt onemocnění ano/ne), byl k predikci vytvořen model založený na **logistické regresi**.

Popis proměnných

V modelu kvůli přehlednosti pracujeme se čtyřmi proměnnými, ačkoliv v původním datasetu pracovali s více faktory. Pro analýzu byly vybrány proměnné:

- **Ten Year CHD (závislá proměnná):** Binární proměnná udávající, zda se u pacienta v následujících deseti letech rozvinula ischemická choroba srdeční.
 - 1 = onemocnění se vyskytlo,
 - 0 = onemocnění se nevyskytlo.
- **Věk (age):** Spojitá proměnná vyjádřená v letech. Věk je v medicíně považován za jeden z nejsilnějších nekontrolovatelných rizikových faktorů, neboť s rostoucím věkem dochází k přirozené degradaci kardiovaskulárního systému.
- **Pohlaví (male):** Kategorická proměnná, která v modelu vystupuje skrze referenční kategorii 0 (ženy). Obecně je známo, že muži čelí vyššímu riziku v ranějším věku, zatímco u žen riziko prudce stoupá po menopauze.
- **Kouření (current smoker):** Kategorická proměnná indikující životní styl pacienta v době sběru dat.
 - 1 = aktivní kuřák,
 - 0 = nekuřák. Kouření je zásadním ovlivnitelným faktorem, který prokazatelně poškozuje cévní stěny a urychluje aterosklerotické procesy

Data a další informace o této zprávě jsou dostupné na adrese:

<https://dostal.vyzkum-psychologie.cz/stat4?i=639>

Výsledky logistické regrese

Tabulka 1 shrnuje statistické parametry modelu. Všechny sledované prediktory jsou statisticky vysoce významné ($p < 0,001$).

Tabulka 1: Statistické parametry logistického modelu pro predikci „Ten Year CHD“

Prediktor	Koeficient (B)	Standardní chyba	Waldova stat.	p-hodnota	Poměr šancí (OR)	95% Konfidenční interval pro OR
Intercept		0,296	392,997			
Věk (age)	0,081	0,005	216,265	< 0,001	1,084	[1,073; 1,096]
Nekuřák (0)	-0,326	0,047	12,093	< 0,001	0,722	[0,601; 0,867]
Žena (0)	-0,493	0,045	29,396	< 0,001	0,611	[0,511; 0,730]

Pro správné pochopení modelu definujeme význam jednotlivých hodnot v tabulce výsledků:

- **Koeficient B (estimate):** Udává směr vztahu. Pokud je hodnota kladná, faktor riziko zvyšuje; pokud je záporná, faktor riziko snižuje. Samotné číslo je však v logaritmické stupnici, což je pro běžnou praxi hůře srozumitelné.
- **Standardní chyba (S.E.):** Určuje stabilitu přesnosti našeho odhadu. Čím nižší je tato hodnota, tím přesněji model daný vliv vypočítal.
- **Waldova statistika a p-hodnota:** Určují statistickou významnost. Waldova statistika ukazuje sílu vlivu dané proměnné na model – čím vyšší číslo, tím důležitější je daný faktor pro výslednou předpověď.
- **Poměr šancí (Odds Ratio – OR):** Tento ukazatel má nejlepší interpretační potenciál. Převádí logaritmický koeficient B na srozumitelné číslo. Hodnota poměru šancí se interpretuje jako násobek šance:
 - **OR > 1:** Šance na onemocnění roste (např. 1,08 znamená nárůst o 8%).
 - **OR < 1:** Šance na onemocnění klesá (např. 0,72 znamená pokles o 28%).
 - **OR = 1:** Faktor nemá na nemoc žádný vliv.

Interpretace tabulky 1

Nejsilnějším rizikovým faktorem v modelu je věk, což potvrzuje i vysoká hodnota Waldovy statistiky (216,26). Model predikuje, že s každým dalším uplynulým rokem života se šance na rozvoj onemocnění zvyšuje o **8,4 %** (OR = 1,084). Úzký interval spolehlivosti (1,073–1,096) naznačuje vysokou přesnost tohoto odhadu.

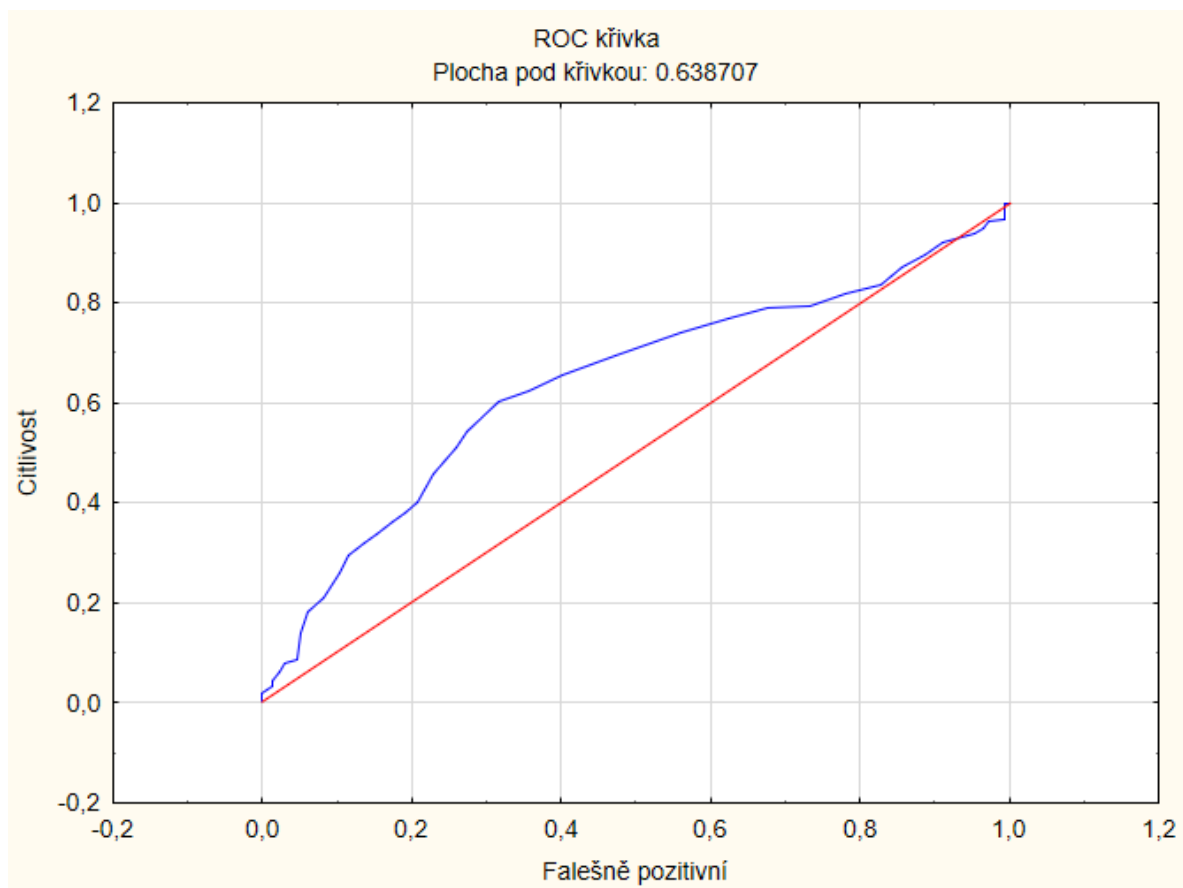
Významnými prediktory se ukázaly být i faktory životního stylu a demografické ukazatele.

Nekuřáci vykazují signifikantně nižší riziko ve srovnání s aktivními kuřáky (OR = 0,72; $p < 0,001$).

Podobně byl zjištěn nižší výskyt u ženského pohlaví (kategorie 0) oproti mužům (OR = 0,61; $p < 0,001$).

Data a další informace o této zprávě jsou dostupné na adrese:

<https://dostal.vyzkum-psychologie.cz/stat4?i=639>



Pro posouzení celkové predikční schopnosti modelu byla sestrojena ROC křivka (Graf 1). Plocha pod křivkou (**AUC**) dosahuje hodnoty **0,639**. Tato hodnota indikuje střední diskriminační schopnost modelu. Konkrétně to znamená, že existuje 63,9% pravděpodobnost, že model při náhodném výběru jednoho nemocného a jednoho zdravého jedince přiřadí vyšší riziko skutečně nemocnému.

Vzhledem k tomu, že model pracuje s minimálním počtem proměnných (pouze věk, pohlaví a kouření) a nezahrnuje klinické parametry (jako krevní tlak či cholesterol), lze tuto míru přesnosti považovat za uspokojivý základ. „**Citlivost**“ modelu zobrazuje schopnost identifikace skutečně ohrožených jedinců, zatímco osa **falešné positivity** vyjadřuje míru nesprávně přidělených diagnóz (zdravého model označí za nemocného).

Tabulka 2 - klasifikační matice

	predikována nemoc	predikováno zdraví	procento úspěšných predikcí
skutečně nemocní	328	198	62,3574144
skutečně zdraví	178	321	64,3286573

Konečná validace modelu byla provedena pomocí klasifikační matice (viz tabulka 2). Výsledky ukazují, že model dosahuje celkové úspěšnosti klasifikace **63,3 %**, přičemž vykazuje vyváženou schopnost identifikovat jak rizikové pacienty (**62,4 %**), tak jedince bez nálezu (**64,3 %**).

Ve spojení s hodnotou **AUC (0,64)** lze konstatovat, že model má střední predikční sílu. I přes absenci detailních klinických dat (krevní tlak, cholesterol) dokáže tato kombinace věku, pohlaví a kouření zvýšit šanci na správnou identifikaci pacienta v riziku oproti náhodnému odhadu. Tento

Data a další informace o této zprávě jsou dostupné na adrese:

<https://dostal.vyzkum-psychologie.cz/stat4?i=639>

model tak představuje středně efektivní nástroj pro rychlou primární stratifikaci rizik v běžné ambulantní praxi.

Závěr

Pomocí statistického programu STATISTICA byl vytvořen model predikce desetiletého rizika ischemické choroby srdeční na základě dat ze studie Framingham Heart Study (získáno z platformy Kaggle). Predikce byla realizována pomocí logistické regrese, jelikož závislá proměnná měla binární charakter (výskyt onemocnění ano/ne).

Za statisticky významné prediktory lze považovat věk, pohlaví a kuřácký status ($p < 0,001$).

Nejsilnějším rizikovým faktorem se ukázal věk – s každým dalším rokem života se šance na vznik onemocnění zvyšuje o 8,4 %. Mužské pohlaví a aktivní kouření byly rovněž spojeny s vyšším rizikem rozvoje ischemické choroby srdeční.

Klasifikační matice ukázala, že model správně identifikoval přibližně 63 % pacientů, přičemž dosahoval vyvážené úspěšnosti při rozlišování nemocných i zdravých jedinců. Hodnota AUC (0,639) indikuje střední diskriminační schopnost modelu. Přestože model pracuje pouze s omezeným počtem proměnných, představuje funkční základ pro primární stratifikaci kardiovaskulárního rizika.

Data a další informace o této zprávě jsou dostupné na adrese:

<https://dostal.vyzkum-psychologie.cz/stat4?i=639>