VYSOKÁ ŠKOLA POLYTECHNICKÁ JIHLAVA

Aplikovaná informatika

DATOVÁ VĚDA: PREDIKCE VÝSKYTU SRDEČNÍHO SELHÁNÍ

Bakalářská práce

Autor práce: Jakub Lukáš Skalický

Vedoucí práce: doc. Dr. Ing. Jan Voráček, CSc.

Jihlava 2026

Vysoká škola polytechnická Jihlava

Tolstého 16, 586 01 Jihlava

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Autor práce:	**Jakub Lukáš Skalický**
Studijní program:	Aplikovaná informatika
Garant studijního programu:	Ing. Lenka Kuklišová Pavelková, Ph.D.
Název práce:	**Datová věda: predikce výskytu srdečního selhání**
Vedoucí práce:	doc. Dr. Ing. Jan Voráček, CSc.
Cíl práce:	Cílem práce je aplikace úplného životního cyklu datové vědy na problematiku predikce výskytu srdečního onemocnění. Půjde při ní zejména o nalezení nejvýznamnějších fyziologických a diagnostických atributů, věrohodně predikujících možné zdravotní komplikace. Naplnění cíle zahrnuje identifikaci problému, rešerši souvisejících zdrojů, návrh a optimalizaci vhodných experimentů, čištění a konsolidaci relevantních dat, jejich explorativní analýzy a induktivní modelování, srozumitelnou vizualizaci výsledků a diskusi možných řešení definovaného problému na základě shromážděných, vyhodnocených a řádně diskutovaných podkladů. Základní datový korpus bude převzat z veřejných zdrojů a student ho může dle vlastních úvah dále modifikovat. Experimenty budou realizovány a distribuovány formou interaktivního online notebooku, využívajícího problémově orientované knihovny jazyka Python. Při vývoji finálního řešení bude využita agilní metodika Rapid Application Development s podporou umělé inteligence.

Abstrakt

Bakalářská práce se zabývá využitím metod datové vědy a strojového učení při predikci kardiovaskulárního onemocnění z klinických a diagnostických údajů pacientů. Teoretická část vymezuje principy životního cyklu datové vědy, techniky explorativní analýzy, klasifikační modely, hodnocení výkonnosti a interpretaci výsledků. Součástí je i přehled vybraných studií zaměřených na predikci srdečních onemocnění. Praktická část vychází z veřejně dostupného klinického datasetu a zahrnuje jednorozměrnou, dvourozměrnou i vícerozměrnou analýzu dat, UMAP projekci, shlukování metodou K-Prototypes, hodnocení významnosti proměnných pomocí Mutual Information a SHAP a porovnání více klasifikačních modelů. Hodnocení je provedeno pomocí více metrik, ROC a Precision–Recall křivek, learning curve, intervalů spolehlivosti a Forest Plotu. Praktická část je doplněna experimenty se změnou velikosti trénovací množiny, chybějícími daty a pravděpodobnostní interpretací výsledků.

Klíčová slova

Datová věda; strojové učení; kardiovaskulární onemocnění; srdeční selhání; predikce; explorativní analýza dat; klasifikační modely; interpretovatelnost modelů; zdravotnictví.

Abstract

This bachelor's thesis focuses on the use of data science and machine learning methods for the prediction of cardiovascular disease based on patients' clinical and diagnostic data. The theoretical part introduces the principles of the data science lifecycle, exploratory analysis techniques, classification models, model evaluation, and result interpretation. It also includes a review of selected studies dealing with heart disease prediction. The practical part is based on a publicly available clinical dataset and includes univariate, bivariate, and multivariate data analysis, UMAP projection, clustering with the K-Prototypes method, feature importance evaluation using Mutual Information and SHAP, and a comparison of multiple classification models. The models are evaluated using several metrics, ROC and Precision–Recall curves, learning curve analysis, confidence intervals, and Forest Plot visualization. The practical part is further extended by experiments with train-test split size, missing data, and probabilistic interpretation of the results.

Keywords

Data science; machine learning; cardiovascular disease; heart failure; prediction; exploratory data analysis; classification models; model interpretability; healthcare.

Prohlašuji, že předložená bakalářská práce je původní a zpracoval/a jsem ji samostatně. Prohlašuji, že citace použitých pramenů je úplná, že jsem v práci neporušil/a autorská práva (ve smyslu zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, v platném znění, dále též „**AZ**“).

Byl/a jsem seznámen/a s tím, že na mou bakalářskou práci se plně vztahuje **AZ**, zejména § 60 (školní dílo).

Podle § 47b zákona o vysokých školách souhlasím se zveřejněním své práce podle Směrnice pro vedení, vypracování a zveřejňování závěrečných prací na VŠPJ, a to bez ohledu na výsledek obhajoby.

Beru na vědomí, že VŠPJ má právo na uzavření licenční smlouvy o užití mé bakalářské práce a prohlašuji, že **s o u h l a s í m** s případným užitím mé bakalářské práce (prodej, zapůjčení apod.).

Jsem si vědom/a toho, že užít své bakalářské práce či poskytnout licenci k jejímu využití mohu jen se souhlasem VŠPJ, která má právo ode mě požadovat přiměřený příspěvek na úhradu nákladů, vynaložených vysokou školou na vytvoření díla (až do jejich skutečné výše), z výdělku dosaženého v souvislosti s užitím díla či poskytnutím licence.

V Jihlavě dne 13. dubna 2026

…….

Podpis studenta/ky

Poděkování

Rád bych poděkoval vedoucímu bakalářské práce za odborné vedení, cenné rady a připomínky, které mi pomohly při zpracování práce. Dále bych chtěl poděkovat své rodině a blízkým za podporu a trpělivost během studia. Poděkování patří všem, kteří mi poskytli pomoc a konzultace a přispěli k úspěšnému dokončení bakalářské práce.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

AI	Artificial Intelligence (umělá inteligence)
AUC	Area Under the Curve (plocha pod křivkou, souhrnná míra výkonnosti modelu)
BMI	Body Mass Index (index tělesné hmotnosti)
CSV	Comma-Separated Values (textový formát tabulkových dat oddělených čárkami)
CV	Cross-Validation (křížová validace)
EDA	Exploratory Data Analysis (explorativní analýza dat)
EKG	Elektrokardiografie / elektrokardiogram (metoda a záznam elektrické aktivity srdce)
CHD	Coronary Heart Disease (ischemická choroba srdeční)
KNN	k-Nearest Neighbors (klasifikační metoda založená na nejbližších sousedech)
LIME	Local Interpretable Model-agnostic Explanations (metoda lokální interpretace predikcí modelů strojového učení)
MAE	Mean Absolute Error (průměrná absolutní chyba)
MCC	Matthews Correlation Coefficient (koeficient pro hodnocení kvality binární klasifikace)
ML	Machine Learning (strojové učení)
MSE	Mean Squared Error (průměrná kvadratická chyba)
PCA	Principal Component Analysis (analýza hlavních komponent)
PR	Precision–Recall (křivka vyjadřující vztah mezi precision a recall)
QRS	komorový depolarizační komplex na elektrokardiogramu
R^2	Coefficient of Determination (koeficient determinace)
RMSE	Root Mean Squared Error (odmocnina z průměrné kvadratické chyby)
ROC	Receiver Operating Characteristic (křivka vyjadřující vztah mezi senzitivitou a falešně pozitivní mírou při různých rozhodovacích prahových hodnotách)
SHAP	SHapley Additive exPlanations (metoda pro interpretaci predikcí modelů strojového učení)
SVM	Support Vector Machine (metoda strojového učení založená na maximálním oddělení tříd)
UMAP	Uniform Manifold Approximation and Projection (metoda nelineární redukce dimenzionality pro vizualizaci a zachování struktury dat)
WHO	World Health Organization (Světová zdravotnická organizace)

Úvod

Srdeční onemocnění představují dlouhodobě jednu z nejzávažnějších příčin úmrtnosti ve světové populaci. Podle údajů Světové zdravotnické organizace způsobují každoročně miliony úmrtí a tvoří významnou část celkové zátěže zdravotnických systémů. Do široké skupiny srdečních onemocnění patří i srdeční selhání, které představuje závažný klinický stav spojený s vysokou nemocností, mortalitou a výrazným dopadem na kvalitu života pacientů. Včasná identifikace rizika závažných srdečních chorob, včetně srdečního selhání, proto představuje zásadní prvek moderní preventivní medicíny a může přispět ke snížení mortality, zlepšení kvality péče a efektivnějšímu využití zdravotnických zdrojů.

Rozvoj metod datové vědy v posledních letech otevřel nové možnosti pro analýzu a predikci zdravotních stavů na základě rozsáhlých datových souborů. Kombinací statistických metod, strojového učení a pokročilých vizualizačních technik lze z dostupných klinických a fyziologických dat odvozovat modely, které pomáhají lékařům lépe rozpoznat rizikové pacienty ještě před vznikem závažných komplikací. Zmíněné přístupy umožňují lépe pochopit vzájemné vztahy mezi jednotlivými zdravotními ukazateli, jako jsou krevní tlak, hladina cholesterolu, věk, pohlaví či přítomnost chronických onemocnění, a jejich vliv na pravděpodobnost vzniku kardiovaskulárních problémů. Spojení medicíny, statistiky a umělé inteligence vytváří rozsáhlý interdisciplinární rámec, který má potenciál významně přispět k personalizaci zdravotní péče.

Cílem bakalářské práce je aplikovat kompletní životní cyklus datové vědy na problematiku predikce výskytu srdečního onemocnění. Výzkum se zaměřuje na identifikaci fyziologických a diagnostických atributů, které mají největší vliv na pravděpodobnost výskytu srdečních chorob, a na vytvoření prediktivního modelu umožňujícího odhadnout riziko zdravotních komplikací na základě vybraných dat. Důraz je kladen na proces zpracování dat od počáteční akvizice až po konečnou interpretaci výsledků, přičemž důležitou roli hraje i transparentnost a srozumitelnost modelu pro budoucí praktické využití.

V posledních letech se metody strojového učení stále častěji využívají v kardiologii k predikci rizika srdečních onemocnění a k podpoře klinického rozhodování. Dosavadní studie ukazují, že moderní algoritmy včetně neuronových sítí, Random Forest či podpůrných vektorových metod mohou dosahovat vysoké přesnosti při analýze klinických a zobrazovacích dat (Krittanawong a kol., 2019; Krittanawong a kol., 2020; Ambale-Venkatesh a kol., 2017). Současně se objevují přístupy, které zdůrazňují potřebu srozumitelné interpretace modelů, což je klíčové pro jejich praktické uplatnění ve zdravotnictví (Ribeiro a kol., 2016). Ačkoli řada prací se zaměřuje především na porovnání jednotlivých algoritmů a jejich predikčního výkonu, bakalářská práce navazuje na jejich závěry a rozšiřuje je o komplexní pohled na celý životní cyklus datové vědy od získání a přípravy dat až po interpretaci a vizualizaci výsledků.

Hlavním přínosem práce je komplexní aplikace principů datové vědy na reálnou zdravotnickou problematiku a demonstrace možností využití strojového učení při odhalování rizikových faktorů spojených se srdečními onemocněními. Výsledkem bude přehledné a opakovatelně použitelné experimentální prostředí, které umožní nejen srovnání různých modelů, ale i hlubší porozumění souvislostem mezi fyziologickými ukazateli a pravděpodobností výskytu srdečních chorob. Přístup založený na analýze dat a modelování může přispět ke zkvalitnění rozhodovacího procesu v klinické praxi a stát se podkladem pro další výzkum v oblasti prediktivní medicíny.

1 Teoretická část

Teoretická část práce představuje základní koncepty, ze kterých vychází současné metody analýzy dat a prediktivního modelování. Práce s daty, ať už ve vědě, průmyslu nebo zdravotnictví, se opírá o soubor principů, které určují, jak data získáváme, organizujeme, zpracováváme a interpretujeme. Pro správné porozumění celému procesu je proto nezbytné nejprve vymezit základní pojmy, zmapovat historický vývoj datové analytiky a objasnit roli, kterou hraje datová věda jako samostatná disciplína.

V posledních desetiletích dochází k významnému posunu od tradičních statistických metod k pokročilým algoritmům strojového učení, které umožňují automatické rozpoznávání vzorců v rozsáhlých a komplexních datových souborech. Zmíněný vývoj přinesl změnu nejen v možnostech analýzy, ale i v samotném pohledu na data jako na klíčový zdroj poznání. Zatímco dříve byla data vnímána především jako prostředek pro popis reality, současné přístupy z nich činí základní stavební jednotku pro modelování budoucích jevů a vytváření prediktivních systémů. Datová věda jako obor vznikla právě v reakci na nová technologická a metodická paradigmata. Jejím charakteristickým rysem je interdisciplinarita, spojuje statistiku, informatiku, matematické modelování i znalost konkrétní aplikační oblasti. Díky tomu umožňuje komplexně chápat strukturu dat a volit adekvátní postupy jejich zpracování. Součástí teoretického rámce datové vědy jsou i principy správné analytické praxe, zahrnující například systematické předzpracování dat, důsledné vyhodnocování kvality modelů či transparentní dokumentaci celého procesu.

Důležitým aspektem současné analytické praxe je rostoucí důraz na srozumitelnost modelů a pochopitelnost výsledků. Přestože moderní algoritmy dokážou dosahovat velmi vysoké přesnosti, jejich využití ve společensky citlivých oblastech, například v medicíně, vyžaduje, aby bylo možné jasně zdůvodnit, jak model k dané predikci dospěl. Zmíněný požadavek ovlivňuje volbu algoritmů, strukturu dat a způsob prezentace výsledků.

1.1 Datová věda a její charakteristika

Datová věda je klíčový interdisciplinární obor, který propojuje principy statistiky, informatiky, matematiky a odborné znalosti oboru s cílem získat z dat nové poznatky a využít je k rozhodování, predikci či optimalizaci procesů. Představuje systematický přístup k analýze dat založený na aplikaci vědeckých principů, algoritmických postupů a metod datové analytiky. Její rychlý rozvoj je úzce spjat se vzestupem digitálních technologií, narůstajícím množstvím dostupných dat a potřebou efektivního využívání daných informací v praxi (Provost a Fawcett, 2013).

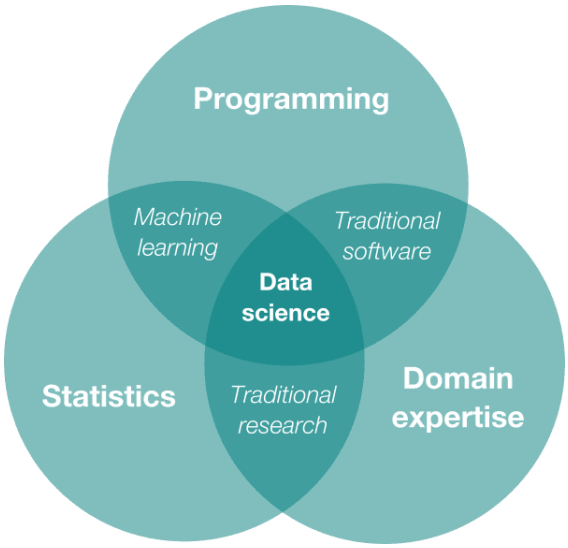
Význam datové vědy narůstá napříč různými oblastmi, včetně zdravotnictví, kde hraje stále větší roli například při analýze zdravotních záznamů, predikci rizikových faktorů, personalizaci medicíny či podpoře klinického rozhodování (Topol, 2019). Vzhledem k tomu, že klinická data bývají složitá, heterogenní a mnohdy neúplná, datová věda poskytuje rámec a metodologii, díky nimž je možné tuto komplexitu překonat a využít data ke zlepšení péče o pacienty. Pro predikci srdečních onemocnění, která patří mezi hlavní příčiny úmrtí ve světě (World Health Organization, 2024), je datová věda obzvláště přínosná, protože umožňuje analyzovat vzorce

chování fyziologických a diagnostických ukazatelů a identifikovat kombinace faktorů, které signalizují zvýšené riziko.

1.1.1 Definice a pojetí datové vědy

Přestože se datová věda dnes řadí mezi standardní vědecké disciplíny, její definice není jednotná. Provost a Fawcett (2013) definují datovou vědu jako „soubor principů a postupů zaměřených na získávání poznatků z dat“, přičemž zdůrazňují, že jejím jádrem je systematická aplikace vědecké metody, která zahrnuje formulaci hypotéz, jejich ověřování na základě empirických dat a vyhodnocování získaných závěrů.

Dhar (2013) doplňuje, že datová věda představuje experimentální vědu založenou na datech, jejímž cílem je získávat zobecnitelné poznatky prostřednictvím matematických a algoritmických postupů. Klade důraz na to, že datová věda spojuje schopnosti kvantitativního analytika a softwarového inženýra a vyžaduje technickou zdatnost, statistickou gramotnost i schopnost vytvářet systematické modely, což je znázorněno na následujícím obrázku.



Obr. 1: Postavení datové vědy mezi statistikou, programováním a doménovou znalostí
Zdroj: Data Tricks (2019)

Současný odborný diskurz vnímá datovou vědu nejen jako technický soubor nástrojů, ale jako strukturovaný metodologický rámec, který se zaměřuje na celý proces práce s daty od jejich získání přes předzpracování až po interpretaci výsledků. Účinná datová věda musí být replikovatelná, transparentní a založená na jasné dokumentaci jednotlivých kroků, díky čemuž je možné zajistit důvěryhodnost a robustnost analytických výstupů.

Jedním ze současných trendů v oblasti datové vědy je zdůrazňování požadavků na etiku, odpovědnost a soulad s regulacemi, zejména při práci s osobními údaji. O’Neil a Schutt (2013) upozorňují, že špatně navržené nebo neadekvátně interpretované modely mohou vést k nespravedlivým rozhodnutím nebo nepřesným závěrům. Datová věda musí být chápána rovněž jako disciplína, která zohledňuje dopady svých výsledků na uživatele i celou společnost.

1.1.2 Interdisciplinární charakter datové vědy

Datová věda se od tradičních analytických disciplín liší svou výraznou interdisciplinaritou. V literatuře se opakovaně uvádí, že její efektivní aplikace vyžaduje propojení několika odborných oblastí, které se vzájemně doplňují (Provost a Fawcett, 2013; Dhar, 2013). Klíčové složky datové vědy jsou rozebrány v následujícím seznamu:

1. Statistika a matematika tvoří základ analytického uvažování. Jsou nezbytné pro porozumění datovým rozdělením, testování hypotéz, tvorbu regresních modelů či odhadování pravděpodobností. Statistické metody určují, jak robustní a spolehlivé závěry lze z dat učinit.
2. Strojové učení umožňuje vytvářet modely, které se adaptují na dostupná data a odhalují v nich skryté vzorce. Krittanawong a kol. (2020) poukazují na to, že algoritmy strojového učení mají díky své flexibilitě zásadní význam pro predikci složitých jevů, jako jsou právě srdeční onemocnění.
3. Efektivní práce s daty není možná bez technických dovedností, jako je manipulace s datovými strukturami, implementace kódu nebo optimalizace pracovních postupů. Python představuje v datové vědě nejčastěji používaný nástroj díky své flexibilitě a rozsáhlému ekosystému knihoven (O'Neil a Schutt, 2013).
4. Odborné znalosti zpracovávaných dat, jelikož bez pochopení oblastí, jako je medicína, biologie či fyziologie, nelze interpretovat výsledky modelů ani navrhnout adekvátní metodologii. Topol (2019) zdůrazňuje, že aplikace datové vědy ve zdravotnictví předpokládá úzkou interdisciplinární spolupráci.
5. Etika, regulace a správa dat, protože práce s osobními zdravotnickými údaji podléhá přísné regulaci, a proto musí analytik dbát na správné nakládání s daty, ochranu soukromí a minimalizaci zkreslení (World Health Organization, 2024).

Interdisciplinární rámec umožňuje komplexní a efektivní zpracování dat, které je nezbytné zejména v oblastech, kde mají výsledky přímý dopad na lidské zdraví a život, pro lepší přehlednost je zobrazen v následující tabulce.

Tab. 1: Hlavní disciplíny tvořící datovou vědu

Disciplína	**Stručný popis**
Statistika a matematika	Základ pro analýzu dat, odhad parametrů, testování hypotéz a kvantifikaci nejistoty.
Strojové učení	Metody umožňující automatické učení modelů z dat a odhalování skrytých vzorů.
Datové inženýrství	Zajištění sběru, ukládání, integrace a efektivního zpracování dat.
Softwarové nástroje	Implementace analytických postupů, automatizace workflow a práce s velkými daty.
Doménová znalost	Porozumění aplikační oblasti umožňující správnou interpretaci výsledků.

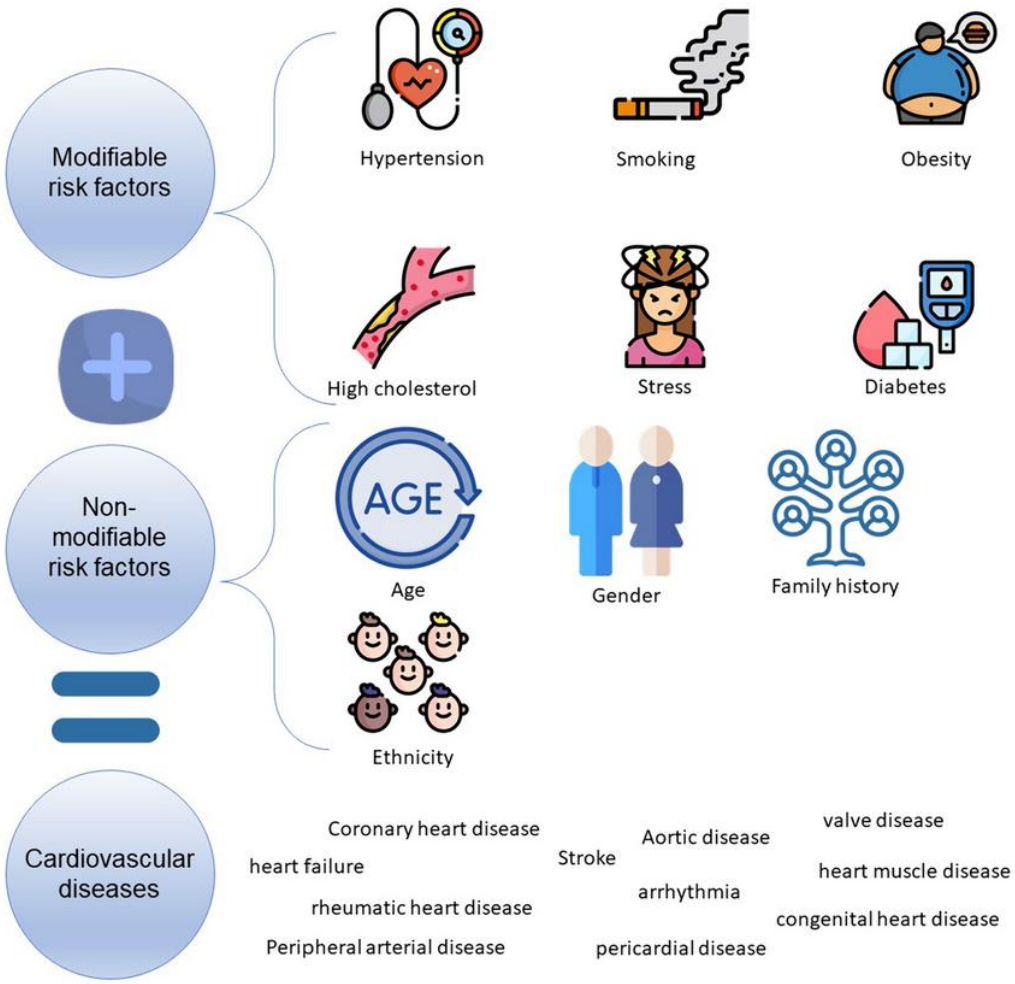
Zdroj: vlastní zpracování dle Dhar, Provost a Fawcett (2013)

1.1.3 Datová věda v medicíně

Zdravotnictví je jednou z oblastí, kde má datová věda potenciál zásadně transformovat praxi. Moderní nemocnice a zdravotnická zařízení generují obrovské množství dat od elektronických zdravotních záznamů přes výsledky laboratorních testů až po signály ze senzorů a lékařských

přístrojů. Datová věda a umělá inteligence představují základní stavební kameny tzv. digitální medicíny, která směřuje k personalizovaným postupům založeným na datech (Topol, 2019).

Jednou z nejvýznamnějších oblastí je predikce srdečních onemocnění. Podle zprávy American Heart Association (Benjamin a kol., 2023) představují kardiovaskulární choroby jednu z nejčastějších příčin úmrtí v celosvětovém měřítku. Díky datové vědě je možné analyzovat vzájemné vztahy mezi rizikovými faktory, jako jsou hypertenze, kouření, diabetes, hladina cholesterolu či BMI, a odhadovat pravděpodobnost budoucích zdravotních komplikací. Následující obrázek znázorňuje dříve zmíněné ovlivnitelné a neovlivnitelné rizikové faktory kardiovaskulárních onemocnění.



Obr. 2: Hlavní rizikové faktory ovlivňující vznik kardiovaskulárních onemocnění

Zdroj: ResearchGate (2024)

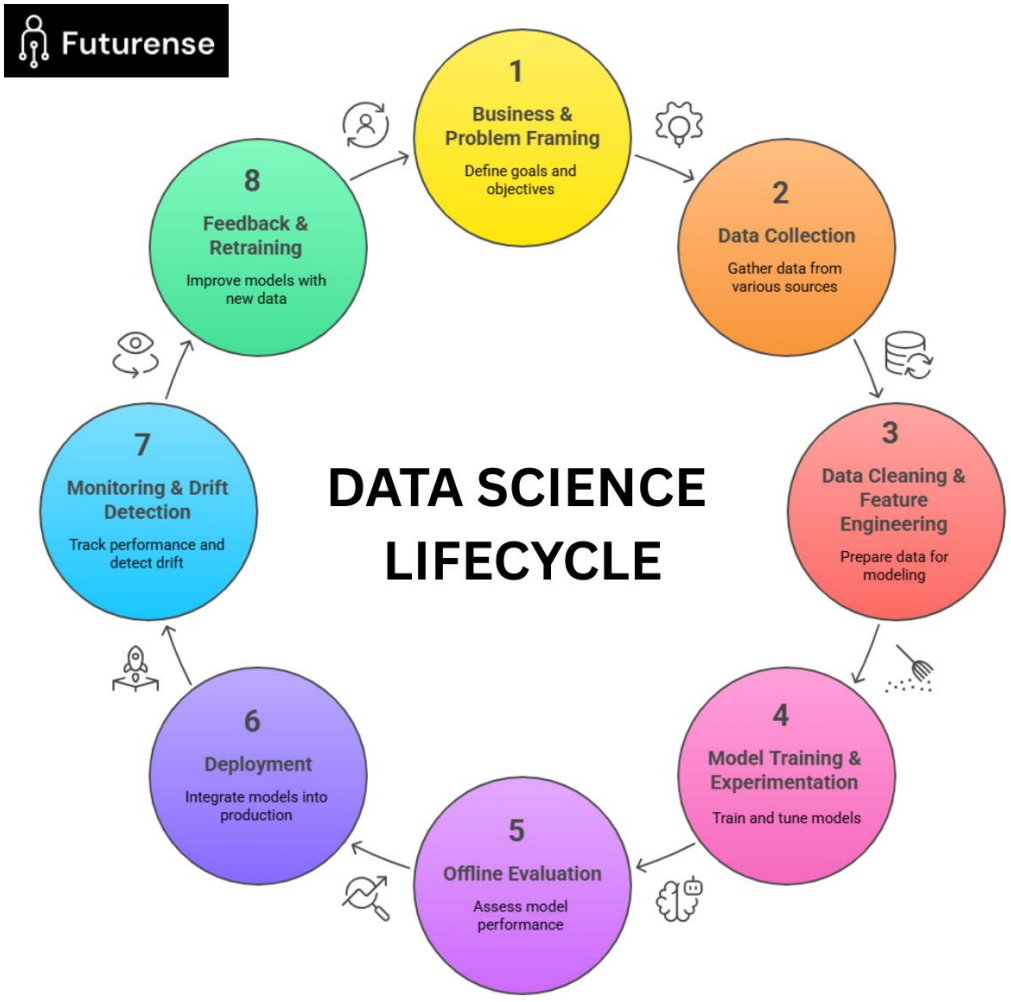
Moderní přístupy využívají metody strojového učení, které umožňují odhalit i nelineární vztahy, jež jsou tradičními statistickými metodami obtížně zachytitelné. Krittanawong a kol. (2019) popisují, že neuronové sítě dokážou dosahovat v diagnostice srdečních onemocnění vysoké přesnosti, zejména díky schopnosti zachytit komplexní vztahy v datech. Naopak interpretovatelné metody, jako je logistická regrese, poskytují lékařům přehlednější informace o tom, jak jednotlivé proměnné ovlivňují výsledek predikce.

Ribeiro a kol. (2016) upozorňují, že praktické nasazení modelů v medicíně vyžaduje vysokou úroveň srozumitelnosti. Metody jako LIME nebo SHAP umožňují vysvětlit, jakými mechanismy strojové učení dochází ke svým závěrům, což je klíčové pro důvěru kliniků i pacientů.

Datová věda v medicíně funguje nejen jako analytický nástroj, ale i jako prostředek k transformaci systémů péče. Využití prediktivních modelů může podpořit včasnou diagnostiku, personalizovanou léčbu a efektivní prevenci srdečních onemocnění.

1.2 Životní cyklus datové vědy

Jedná se o strukturovaný rámec pro systematickou práci s daty, jehož cílem je získání smysluplných a využitelných poznatků. Proces zahrnuje postupné kroky od formulace analytického problému přes sběr, přípravu a analýzu dat až po vytvoření, vyhodnocení a praktické využití modelů. Jednotlivé fáze na sebe úzce navazují a společně vytvářejí iterativní proces, který umožňuje průběžné zpřesňování řešení na základě nových poznatků a dat. Pochopení životního cyklu datové vědy je zásadní nejen pro správnou aplikaci analytických metod, ale i pro zajištění transparentnosti, reprodukovatelnosti a důvěryhodnosti dosažených výsledků. Cyklus je znázorněn na obrázku vizuálně pro lepší pochopení.



Obr. 3: Životní cyklus datové vědy a návaznost jednotlivých fází analytického procesu
Zdroj: Futurense (2026)

1.2.1 Definice problému

Stanovení cílů analýzy a definice problému představují výchozí fázi životního cyklu datové vědy a zásadně ovlivňují celý další průběh analytického procesu. Při definici problému dochází k formulaci otázky, na kterou má datová analýza odpovědět, a k určení očekávaných výstupů, jež mají přinést prakticky využitelné poznatky. Úspěšná datová analýza nezačíná výběrem algoritmů, ale jasným vymezením problému a jeho převodem do analyticky řešitelné podoby (Provost a Fawcett, 2013).

V kontextu datové vědy je problém obvykle formulován jako predikční, deskriptivní nebo explorační úloha. Predikční úlohy se zaměřují na odhad budoucího chování nebo stavu na základě historických dat, zatímco deskriptivní analýzy se soustředí na shrnutí a pochopení existujících vzorců v datech. Správná formulace problému vyžaduje porozumění doméně, dostupným datům i omezením analytických metod, aby bylo zajištěno, že zvolený přístup povede k relevantním a interpretovatelným výsledkům (Dhar, 2013).

Stanovení cílů analýzy úzce souvisí s vymezením problému a slouží k určení, jak bude úspěšnost řešení hodnocena. Cíle mohou zahrnovat například dosažení určité úrovně přesnosti, identifikaci klíčových faktorů ovlivňujících sledovaný jev nebo vytvoření modelu, který je snadno interpretovatelný a prakticky využitelný. V odborné literatuře je kladen důraz na to, aby cíle analýzy nebyly definovány pouze technicky, ale aby reflektovaly širší kontext využití výsledků, včetně etických a společenských dopadů (O'Neil a Schutt, 2013).

Důležitou součástí zmíněné fáze je rovněž volba vhodných hodnoticích kritérií a metrik, které umožní objektivně posoudit kvalitu dosažených výsledků. Nevhodně zvolené metriky mohou vést k mylným závěrům nebo k preferování modelů, které nejsou z hlediska reálného použití optimální. Ribeiro a kol. (2016) upozorňují, že důvěra v analytické modely je podmíněna nejen jejich výkonností, ale i transparentností a srozumitelností jejich výstupů.

Definice problému a cílů analýzy tvoří základní rámec, který určuje směr celého životního cyklu datové vědy. Pečlivé vymezení dané fáze přispívá k efektivní práci s daty, snižuje riziko chybných interpretací a zvyšuje pravděpodobnost, že výsledky analýzy budou přenositelné do praxe a přinesou skutečnou přidanou hodnotu.

1.2.2 Sběr a pochopení dat

Zde jsou identifikovány relevantní datové zdroje a je posuzována jejich vhodnost vzhledem k řešené analytické úloze. Data mohou pocházet z různých zdrojů, například z databází, senzorů, otevřených datových repozitářů nebo historických záznamů, přičemž jejich kvalita a reprezentativnost zásadně ovlivňují výsledky analýzy (Provost a Fawcett, 2013).

Pochopení dat zahrnuje základní průzkumnou analýzu, jejímž cílem je seznámit se se strukturou dat, typy proměnných a jejich rozdělením. Fáze sběru dat umožňuje identifikovat potenciální problémy, jako jsou chybějící hodnoty, nekonzistence nebo nevyváženost tříd, které by mohly negativně ovlivnit následné modelování. Dhar (2013) zdůrazňuje, že důkladné porozumění datům je nezbytným předpokladem pro správnou volbu analytických metod a interpretaci výsledků.

Součástí fáze je rovněž posouzení relevance jednotlivých atributů vzhledem k původně definovanému problému. Ne všechna dostupná data musí být pro analýzu přínosná a jejich

nevhodné zahrnutí může vést ke zvýšení složitosti modelu bez odpovídajícího zlepšení výsledků. Efektivní sběr a pochopení dat vytvářejí pevný základ pro další fáze životního cyklu datové vědy a přispívají k transparentnosti a reprodukovatelnosti celého analytického procesu (O'Neil a Schutt, 2013).

1.2.3 Předzpracování a čištění dat

V praxi bývají data často neúplná, zatížená chybami nebo obsahují hodnoty, které neodpovídají předpokladům použitých analytických metod. Bez adekvátního předzpracování mohou nedostatky vést ke zkresleným výsledkům a snížení výkonnosti modelů (Provost a Fawcett, 2013).

Mezi základní kroky předzpracování patří práce s chybějícími hodnotami, detekce a ošetření odlehlých pozorování a sjednocení datových formátů. Dále se provádějí různé transformace dat, jako je normalizace nebo standardizace numerických proměnných, které jsou důležité zejména pro algoritmy citlivé na měřítko vstupních dat. U kategoriálních proměnných se využívají techniky kódování, jež umožňují jejich zpracování pomocí numerických modelů (Dhar, 2013).

Předzpracování dat není čistě technickou činností, ale vyžaduje porozumění doméně a kontextu dat. Nevhodně zvolený postup může vést ke ztrátě důležitých informací nebo k zavedení systematické chyby do analýzy. Transparentní dokumentace jednotlivých kroků předzpracování je klíčová pro reprodukovatelnost a důvěryhodnost výsledků. Pečlivé čištění a příprava dat zvyšují pravděpodobnost, že následné modely budou stabilní, robustní a schopné generalizace na nová data. (O'Neil a Schutt, 2013).

1.2.4 Analýza a výběr příznaků

Fází životního cyklu datové vědy, která se zaměřuje na identifikaci proměnných nejvíce relevantních pro řešený problém, je práce s příznaky a jejich vhodný výběr. Cílem je zlepšit kvalitu modelu odstraněním redundantních nebo irelevantních atributů a současně snížit složitost datového prostoru. Správně zvolená sada příznaků může významně přispět ke zvýšení přesnosti, stability a srozumitelnosti výsledných modelů (Provost a Fawcett, 2013).

Analýza příznaků často zahrnuje kombinaci statistických metod, vizualizací a doménových znalostí. Pomocí korelační analýzy nebo testování závislostí lze odhalit vztahy mezi jednotlivými proměnnými a cílovou veličinou. Zmíněné postupy umožňují lépe pochopit strukturu dat a podporují informovanou volbu atributů pro modelování (Dhar, 2013).

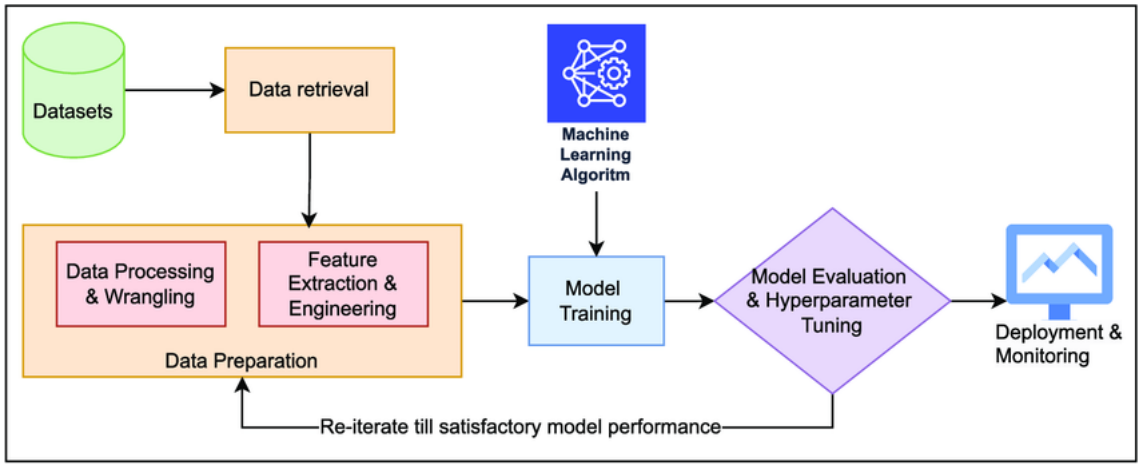
Výběr příznaků může být prováděn manuálně nebo automatizovaně pomocí algoritmických metod. Automatické přístupy zahrnují například filtrační, obalové nebo embedded metody, které hodnotí význam jednotlivých atributů na základě jejich přínosu k predikčnímu výkonu modelu. Volba konkrétní metody závisí na typu dat, použitém algoritmu a požadavcích na srozumitelnost výsledků (Krittanawong a kol., 2020).

Příliš velký počet příznaků může vést k přeučení a ztížit interpretaci výsledků, zatímco příliš agresivní redukce může způsobit ztrátu důležitých informací. Analýza a výběr příznaků tvoří klíčový krok směrem k efektivnímu a prakticky využitelnému řešení.

1.2.5 Modelování a výběr algoritmu

Fáze modelování a výběru algoritmů představuje stěžejní část životního cyklu datové vědy, ve které jsou aplikovány metody strojového učení nebo statistického modelování na připravená data. Cílem fáze je nalézt model, který je schopen co nejlépe zachytit vztahy v datech a splnit předem stanovené cíle analýzy. Volba algoritmu závisí na charakteru problému, typu dat a požadavcích na srozumitelnost výsledků (Provost a Fawcett, 2013).

Modelování obvykle zahrnuje rozdělení dat na trénovací a testovací množinu, případně využití validačních technik, které umožňují objektivně posoudit výkonnost modelu. Během trénování dochází k optimalizaci parametrů modelu na základě trénovacích dat, přičemž je nutné dbát na prevenci přeučení neboli situace, kdy model dosahuje dobrých výsledků na trénovacích datech, ale selhává při generalizaci na nová data (Dhar, 2013). Obrázek představuje obecné schéma strojového učení po nasazení modelu.



Obr. 4: Obecné schéma pipeline strojového učení od přípravy dat po nasazení modelu

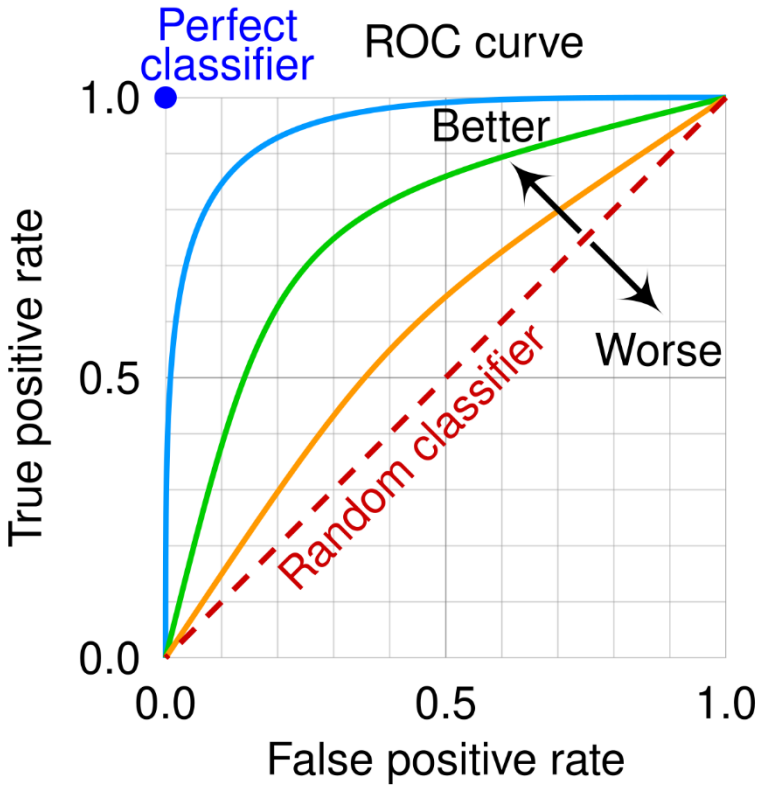
Zdroj: ResearchGate (2022)

V praxi se často testuje více algoritmů, které jsou následně porovnávány na základě zvolených hodnoticích metrik. Popisovaný postup umožňuje identifikovat silné a slabé stránky jednotlivých přístupů a zvolit řešení nejlépe odpovídající dané úloze. V oblasti aplikací strojového učení je rovněž běžné ladění parametrů, které může významně ovlivnit výsledný výkon modelu (Krittanawong a kol., 2019). Modelování není izolovanou činností, ale úzce navazuje na předchozí fáze a často vyžaduje opakované úpravy dat nebo příznaků. Iterativnost je charakteristickým znakem životního cyklu datové vědy a přispívá k postupnému zlepšování kvality výsledného řešení.

1.2.6 Vyhodnocení modelu

Klíčovou fází životního cyklu datové vědy je objektivní posouzení kvality a spolehlivosti vytvořených modelů. V probírané fázi jsou výsledky modelování analyzovány pomocí vhodně zvolených hodnoticích metrik, které odpovídají charakteru řešeného problému. Správná volba metrik je zásadní pro interpretaci výsledků a pro rozhodnutí, zda je model vhodný pro další použití (Provost a Fawcett, 2013). Hodnoticí metriky se liší v závislosti na typu úlohy, například klasifikační nebo regresní. U klasifikačních problémů se často používají metriky jako přesnost, citlivost, specificita nebo plocha pod ROC křivkou, které umožňují komplexnější pohled na

chování modelu než samotná celková přesnost. Dhar (2013) upozorňuje, že spoléhání se na jedinou metriku může vést k zavádějícím závěrům, zejména v případě nevyvážených dat. Na obrázku je znázorněna ROC křivka, která znázorňuje vztah mezi citlivostí a mírou falešně pozitivních výsledků a slouží k hodnocení výkonnosti klasifikačního modelu při různých rozhodovacích prahových hodnotách.



Obr. 5: ROC křivka používaná pro hodnocení výkonnosti klasifikačních modelů
Zdroj: Wikipedie (2025)

Součástí vyhodnocení je rovněž analýza schopnosti modelu generalizovat na dosud neviděná data. K danému účelu se využívají validační techniky, jako je křížová validace, které snižují riziko náhodného zkreslení výsledků. Identifikace přeučení nebo nedostatečného učení modelu umožňuje včasné úpravy modelové architektury nebo návrat k předchozím fázím životního cyklu (O'Neil a Schutt, 2013).

Vyhodnocení modelů poskytuje podklady pro informované rozhodování o dalším směřování analýzy. Výsledky ovlivňují nejen výběr finálního modelu, ale i míru důvěry v jeho použití v praxi a jeho potenciální přínos v dané aplikační oblasti.

1.2.7 Interpretace výsledků

Cílem interpretace výsledků je porozumět výstupům analytických modelů a zasadit je do širšího kontextu řešeného problému. Nestačí pouze dosáhnout vysoké výkonnosti modelu, klíčové je pochopit, jakým způsobem model k jednotlivým predikcím dospívá a jaké faktory je ovlivňují (Provost a Fawcett, 2013).

V uvedené fázi dochází k analýze vztahů mezi vstupními proměnnými a výstupy modelu a k posouzení jejich významu. Proces interpretace může zahrnovat globální pohled na chování

modelu i lokální vysvětlení jednotlivých predikcí. Důraz na transparentnost a srozumitelnost výsledků je zásadní zejména v oblastech, kde rozhodování založené na datech může mít významné dopady, například ve zdravotnictví nebo veřejné správě (Ribeiro a kol., 2016).

O'Neil a Schutt (2013) zdůrazňují, že interpretace výsledků je rovněž nezbytná pro odhalení potenciálních zkreslení nebo nežádoucích důsledků modelu. Bez kritického zhodnocení výstupů může docházet k nesprávným závěrům nebo nekritickému přebírání výsledků, které neodpovídají realitě. Interpretace proto vyžaduje kombinaci technických znalostí, doménového porozumění a kritického myšlení.

Analytická část životního cyklu datové vědy je fází interpretace výsledků uzavřena a vytváří přímý přechod k praktickému využití výsledků. Správně interpretované výstupy zvyšují důvěru v model a umožňují jeho smysluplné zapojení do rozhodovacích procesů.

1.2.8 Nasazení a údržba modelu

Nasazení modelu představuje fázi životního cyklu datové vědy, ve které jsou výsledky analytického procesu převedeny do praktického použití. Model je integrován do konkrétního systému, aplikace nebo rozhodovacího procesu, kde může poskytovat predikce nebo doporučení v reálném čase či v pravidelných intervalech. Úspěšné nasazení vyžaduje nejen technickou připravenost modelu, ale i zohlednění organizačních, etických a uživatelských aspektů jeho využití (Provost a Fawcett, 2013).

Po nasazení modelu je nezbytné průběžně sledovat jeho výkonnost a chování v čase. Vlivem změn v datech nebo v okolním prostředí může docházet k postupnému zhoršování kvality predikcí, což je jev označovaný jako změna datového rozdělení. Pravidelná evaluace modelu na nových datech umožňuje včas identifikovat změny a rozhodnout o nutnosti aktualizace nebo přeučení modelu (Dhar, 2013).

Údržba modelu zahrnuje i dokumentaci použitých postupů, správu verzí a kontrolu reprodukovatelnosti výsledků. Dlouhodobá udržitelnost datových řešení je podmíněna transparentností a odpovědným přístupem k práci s daty. V některých aplikačních oblastech je navíc nutné pravidelně posuzovat etické dopady a rizika spojená s automatizovaným rozhodováním (O'Neil a Schutt, 2013).

1.3 Techniky a nástroje používané v datové vědě

Datová věda využívá soubor metod a nástrojů, které společně pokrývají celý analytický proces od prvotního porozumění datům až po interpretaci výsledků a jejich komunikaci. Volba konkrétních technik závisí na povaze řešeného problému, typu dostupných dat a požadavku na přesnost, robustnost a srozumitelnost výsledků. V prostředí zdravotnických dat hraje zvláštní roli nejen predikční výkon modelů, ale i schopnost vysvětlit jejich rozhodování a převést analytické výstupy do podoby použitelné pro odborníka z aplikační oblasti (Provost a Fawcett, 2013; O'Neil a Schutt, 2013).

1.3.1 Explorativní analýza dat a statistické techniky

Explorativní analýza dat představuje základní součást datové vědy, protože umožňuje porozumět struktuře dat ještě před samotným modelováním. Cílem není pouze popsat datový soubor, ale i odhalit asymetrii rozdělení, odlehlé hodnoty, nerovnoměrné zastoupení kategorií a základní vztahy mezi proměnnými. Dhar (2013) zdůrazňuje, že statistické uvažování je nezbytné pro správnou interpretaci výsledků a pro odlišení skutečných vzorců od náhodných jevů. O'Neil a Schutt (2013) současně upozorňují, že právě důkladné pochopení dat představuje důležitý předpoklad pro správnou volbu analytických metod. Význam explorativní analýzy je zvlášť patrný u predikčních úloh, kde kvalita následného modelování přímo závisí na schopnosti správně rozpoznat vlastnosti vstupních dat a jejich omezení. Steyerberg a kol. (2010) uvádějí, že počáteční porozumění datům a vztahům mezi proměnnými tvoří nezbytný základ pro korektní konstrukci i hodnocení predikčních modelů.

V praxi se explorativní analýza opírá o kombinaci popisných statistik a vizualizačních technik. U numerických proměnných se používají například histogramy, boxploty, testy normality, koeficienty šikmosti nebo korelační koeficienty. U kategoriálních proměnných mají význam tabulky četností, sloupcové grafy a ukazatele síly asociace, například Cramérovo V. Význam uvedených technik spočívá v možnosti získat první systematický vhled do dat, odhalit potenciální problémy a vytvořit základ pro další modelování. V prostředí predikčních úloh je důležitá zároveň schopnost porozumět vztahům mezi proměnnými nejen z hlediska jejich rozdělení, ale i z hlediska jejich budoucího vlivu na kvalitu modelu. Explorativní analýza tak nepředstavuje pouze úvodní popis dat, ale plnohodnotnou analytickou fázi, která významně ovlivňuje všechny navazující kroky datové vědy.

1.3.2 Předzpracování a transformace dat

Nedílnou součástí datové vědy jsou techniky předzpracování a transformace dat. Cílem daných postupů je zajistit, aby vstupní data byla konzistentní, analyticky použitelná a kompatibilní s vybranými algoritmy. O'Neil a Schutt (2013) uvádějí mezi běžné kroky zejména práci s chybějícími hodnotami, standardizaci numerických proměnných, kódování kategoriálních atributů a úpravu struktury datasetu pro modelování.

Význam předzpracování spočívá nejen v technickém zajištění kompatibility s algoritmy, ale i v ovlivnění stability a výkonnosti výsledných modelů. Nevhodně zvolené transformace mohou vést ke ztrátě informace nebo ke zkreslení interpretace, zatímco vhodné předzpracování podporuje robustnost celého analytického postupu. V případě smíšených dat, která obsahují současně numerické i kategoriální proměnné, je navíc důležité zachovat informační hodnotu obou typů atributů.

1.3.3 Redukce dimenzionality a vícerozměrná explorace

S rostoucím počtem proměnných klesá přehlednost datového prostoru a narůstá obtížnost jeho interpretace. Redukce dimenzionality proto představuje důležitou techniku, která umožňuje převést vysokodimenzionální data do jednodušší reprezentace při zachování podstatné části jejich struktury. Vedle lineárních metod se v současné analytické praxi stále častěji používají i nelineární přístupy vhodné pro zachycení složitějších vztahů mezi pozorováními.

Jednou z metod je UMAP neboli Uniform Manifold Approximation and Projection. McInnes a kol. (2018) popisují UMAP jako nelineární metodu redukce dimenzionality založenou na myšlence zachování lokální struktury dat při současné praktické použitelnosti pro reálné datové soubory. Výhodou UMAP je schopnost nabídnout vizualizaci, která umožňuje lépe rozpoznat podobnost mezi pozorováními, částečně oddělené skupiny i přechodové oblasti v datovém prostoru.

1.3.4 Shlukování a segmentace smíšených dat

Metody učení bez učitele umožňují hledat přirozené skupiny objektů bez předem známé cílové proměnné. Shlukování má význam zejména tehdy, když cílem analýzy není pouze predikce, ale i odhalení typických profilů nebo segmentů v datech. Provost a Fawcett (2013) považují segmentaci za důležitý nástroj pro porozumění datové populaci a pro identifikaci vnitřní struktury, která by při čistě dozorovaném modelování nemusela být zřejmá.

U datových souborů obsahujících současně numerické i kategoriální proměnné je vhodné využít algoritmy schopné pracovat se smíšenými daty přímo. Huang (1998) popisuje algoritmus K-Prototypes jako rozšíření K-Means a K-Modes, které kombinuje práci s oběma typy atributů pomocí společné míry nepodobnosti. Význam dané metody spočívá v zachování informační hodnoty kategoriálních proměnných a v možnosti vytvářet interpretovatelné segmenty i v heterogenním datovém prostoru.

1.3.5 Klasifikační modely v datové vědě

Klasifikace představuje základní typ úloh dozorovaného učení, v němž je cílem přiřadit pozorování do předem definovaných tříd. V medicínském kontextu se klasifikace využívá například pro odhad přítomnosti nebo nepřítomnosti onemocnění, identifikaci rizikových pacientů nebo podporu rozhodování při diagnostice. Volba modelu závisí na charakteru dat, požadavku na interpretovatelnost a na schopnosti modelu zachytit složitější vztahy mezi proměnnými. Provost a Fawcett (2013) i O'Neil a Schutt (2013) uvádějí mezi běžné klasifikační přístupy logistickou regresi, rozhodovací stromy, metody nejbližších sousedů a ansámblové modely. V oblasti klinických predikčních úloh má výběr modelu zvláštní význam, protože vedle samotného výkonu je důležitá i transparentnost výsledků a možnost jejich odborné interpretace, což zdůrazňují i Collins a kol. (2015) v souvislosti s korektním vývojem a prezentací predikčních modelů.

Logistická regrese je ceněna zejména pro svou interpretovatelnost a pro schopnost modelovat pravděpodobnost příslušnosti ke třídě. Rozhodovací stromy nabízejí intuitivní pravidlovou strukturu, avšak mohou být citlivé na konkrétní rozdělení dat. Náhodný les představuje robustnější ansámblovou variantu stromových metod, která obvykle lépe zachycuje složitější kombinace proměnných. KNN je založen na podobnosti pozorování v datovém prostoru a Support Vector Machine hledá hranici s maximálním oddělením tříd. V oblasti kardiovaskulární medicíny bývají pružnější modely strojového učení, včetně SVM a ansámblových metod, často spojovány s velmi dobrým predikčním výkonem, přestože mezi studiemi existuje metodická heterogenita. Krittanawong a kol. (2020) uvádějí, že výkonnější modely strojového učení mohou v kardiovaskulární predikci dosahovat velmi dobrých výsledků, avšak skutečná hodnota závisí i na kvalitě validace, interpretaci a klinické použitelnosti dosažených závěrů. Obdobný závěr uvádějí i Sajid a kol. (2021), kteří při vývoji nelaboratorních modelů predikce kardiovaskulárního

onemocnění ukazují, že porovnání konvenčních a strojově učených přístupů má smysl zejména pokud je výkon hodnocen ve vztahu k dostupnosti vstupních proměnných a k praktické použitelnosti výsledného modelu.

1.3.6 Hodnocení výkonnosti modelů

Vyhodnocení klasifikačních modelů nelze opírat o jedinou metriku, protože jednotlivé ukazatele zachycují odlišné aspekty modelového chování. Accuracy, Precision, Recall a F1-score umožňují podrobněji porozumět tomu, jak model pracuje s pozitivními a negativními případy a nakolik je jeho klasifikace vyvážená. Samotná přesnost totiž nemusí poskytovat dostatečně úplný obraz o kvalitě modelu, zejména tehdy, pokud jsou jednotlivé typy chyb z hlediska aplikace různě závažné. V medicínském kontextu může mít například opomenutí skutečně rizikového pacienta výrazně závažnější důsledky než nesprávné zařazení zdravého jedince do rizikové skupiny. Z uvedeného důvodu je vhodné sledovat více ukazatelů současně a chápat hodnocení modelu jako vícerozměrný proces, nikoli jako jednoduché porovnání jediné číselné hodnoty.

Významný je rovněž Matthewsův korelační koeficient, původně zavedený Matthewsem v roce 1975. Zmíněná metrika poskytuje komplexnější pohled na kvalitu binární klasifikace, protože zohledňuje všechny části matice záměn a není omezena pouze na celkový podíl správných odpovědí. Význam MCC je patrný obzvlášť, když je třeba posoudit model vyváženěji a nepřeceňovat výsledek založený pouze na přesnosti nebo F1-score. Na výhody dané metriky upozorňují i Chicco a Jurman (2020), kteří ukazují, že MCC může v řadě binárních úloh poskytovat informativnější a stabilnější hodnocení než běžně používané metriky, zejména pokud je cílem odpovědně porovnávat více modelů mezi sebou.

Velmi důležitou roli hrají také ROC a Precision–Recall křivky. Fawcett (2006) popisuje ROC analýzu jako nástroj umožňující sledovat chování modelu napříč různými rozhodovacími prahy a posuzovat jeho diskriminační schopnost bez fixace na jedinou hranici rozhodnutí. Precision-Recall křivka doplňuje ROC pohled tam, kde je významné především správné zachycení pozitivní třídy. Saito a Rehmsmeier (2015) zdůrazňují, že Precision–Recall křivka může být v některých situacích informativnější než ROC analýza, zejména při hodnocení binárních klasifikátorů v prostředí s menším zastoupením jedné ze tříd. Hodnocení modelů je vhodné doplnit zároveň o learning curve a o intervaly spolehlivosti metrik, protože stabilita výsledků bývá v praktických aplikacích stejně důležitá jako samotná maximální hodnota výkonnostní metriky. Steyerberg a kol. (2010) v dané souvislosti upozorňují, že kvalitní hodnocení predikčního modelu by mělo zahrnovat nejen diskriminační schopnost, ale i širší pohled na robustnost, kalibraci a celkovou interpretaci modelového výkonu.

1.3.7 Významnost proměnných a vysvětlitelnost modelů

S rostoucí komplexitou modelů roste i význam technik zaměřených na interpretaci a vysvětlování výsledků. Srozumitelnost modelů je důležitá nejen pro odborné porozumění, ale i pro budování důvěry v jejich výstupy, zejména v medicínském prostředí. Ribeiro a kol. (2016) upozorňují, že důvěra v model není podmíněna pouze jeho výkonem, ale i schopností objasnit, proč model dospěl k určité predikci. V oblasti zdravotnických aplikací má vysvětlitelnost zvláštní význam, protože modelové rozhodnutí může ovlivňovat odborný úsudek a navazující klinické kroky.

Z uvedeného důvodu nelze výkonnost modelu oddělit od otázky, zda je možné výsledky interpretovat věcně, konzistentně a odborně obhajitelným způsobem.

Vedle obecných principů vysvětlitelné umělé inteligence se v praxi používají i konkrétní metody hodnocení významnosti proměnných. Mutual Information umožňuje posoudit, kolik informace přinášejí jednotlivé atributy o cílové proměnné. Výhodou daného přístupu je schopnost zachytit i nelineární závislosti, které nemusí být dobře popsány běžnou korelační analýzou. Vergara a Estévez (2014) uvádějí, že metody založené na vzájemné informaci patří mezi důležité nástroje výběru příznaků právě díky schopnosti vystihnout informační přínos proměnných bez omezení na čistě lineární vztahy. V prostředí klinických dat má takový přístup zvláštní hodnotu, protože zdravotní stav pacienta bývá výsledkem souběhu více faktorů, jejichž vliv nemusí být jednoduchý ani přímo úměrný.

Významnou roli v interpretaci modelů hraje také metoda SHAP, tedy SHapley Additive exPlanations. Lundberg a Lee (2017) ji představují jako sjednocený rámec pro interpretaci modelových predikcí založený na Shapleyho hodnotách, který umožňuje kvantifikovat příspěvek jednotlivých proměnných k výslednému rozhodnutí modelu. Přínos SHAP spočívá v možnosti nabídnout globální i lokální interpretaci. Globální pohled umožňuje určit, které proměnné mají největší celkový vliv na modelové rozhodování, zatímco lokální interpretace objasňuje, proč model dospěl ke konkrétní predikci u jednotlivého pozorování. Právě spojení obou úrovní vysvětlení činí z SHAP vhodný nástroj pro analytické úlohy, v nichž nestačí model pouze natrénovat, ale je nutné zároveň věrohodně vysvětlit jeho chování.

1.3.8 Pravděpodobnostní interpretace, kalibrace a Brier score

V medicínských úlohách často nestačí pouze přiřadit pacienta do jedné ze dvou tříd. Významnou roli hraje i možnost vyjádřit míru rizika pomocí pravděpodobnostního odhadu. Pravděpodobnostní interpretace modelu umožňuje jemnější pohled na pacientský profil a může sloužit jako doplněk binární klasifikace. V prostředí klinických predikčních modelů má podobný přístup zvláštní význam, protože podporuje rozhodování založené nejen na kategorickém zařazení, ale i na odhadu intenzity rizika (Steyerberg a kol., 2010). Podobný význam zdůrazňují také Sajid a kol. (2021), kteří při vývoji nelaboratorních modelů kardiovaskulární predikce ukazují, že praktická použitelnost modelu nespočívá pouze ve správné klasifikaci, ale i ve schopnosti poskytovat klinicky využitelný odhad rizika.

K hodnocení kvality pravděpodobnostního odhadu se používají například metriky MAE, RMSE a Brier score. Uvedené ukazatele umožňují posoudit, jak přesně model odhaduje pravděpodobnost příslušnosti k pozitivní třídě a jak velké odchylky se mezi predikovanými a skutečnými hodnotami objevují. Brier score představuje důležitou metriku právě pro hodnocení kvality pravděpodobnostních predikcí, protože zohledňuje velikost chyby v celém intervalu od 0 do 1 (Steyerberg a kol., 2010). V prostředí klinických predikčních modelů má podobné hodnocení význam hlavně proto, že umožňuje rozlišit mezi modelem s dobrou diskriminační schopností a modelem, který současně poskytuje i přesnější pravděpodobnostní odhady.

Důležitou součástí pravděpodobnostní interpretace je také kalibrace modelu. Kalibrační křivka porovnává predikované pravděpodobnosti se skutečnou četností pozitivních případů a ukazuje, nakolik jsou modelové odhady realistické. Van Calster a kol. (2016) upozorňují, že dobře

kalibrovaný model nepředstavuje pouze technicky přesný nástroj, ale i předpoklad pro důvěryhodné využití rizikových modelů v praxi. Podobný důraz na klinickou využitelnost pravděpodobnostních predikcí je patrný i ve studii Qian a kol. (2022), která ukazuje význam kombinace diskriminační schopnosti a prakticky využitelného odhadu kardiovaskulárního rizika. Zařazení kalibrace a Brier score je proto zvlášť důležité tehdy, když model neslouží pouze k rozhodnutí o třídě, ale i k odhadu intenzity rizika. Podobný přístup má ve zdravotnictví výrazný význam, protože umožňuje přejít od čistě kategorického rozhodnutí k jemnějšímu vyjádření pravděpodobnosti onemocnění. onemocnění.

1.3.9 Programové nástroje a notebooková prostředí

Praktická realizace uvedených technik je umožněna prostřednictvím specializovaných programovacích jazyků a knihoven. Mezi nejčastěji používané patří Python a R, které podporují manipulaci s daty, statistickou analýzu, strojové učení i vizualizaci. Python je v současné datové vědě zvlášť rozšířený díky své flexibilitě, širokému ekosystému knihoven a podpoře reprodukovatelných analytických workflow (O'Neil a Schutt, 2013).

Důležitou roli hrají i notebooková prostředí, která umožňují spojit kód, textový výklad, tabulky a grafy do jednoho uceleného dokumentu. Význam notebooků spočívá nejen v technickém experimentování, ale i v transparentní prezentaci výsledků a jejich srozumitelné komunikaci širšímu publiku. V oblasti datové vědy notebook nepředstavuje pouze pracovní nástroj analytika, ale i prostředek dokumentace a reprodukce celého postupu.

1.4 Přehled dosavadních studií a přístupů k mému problému

Dosavadní práce zaměřené na predikci srdečních onemocnění se liší především použitým typem dat, volbou algoritmů strojového učení a způsobem hodnocení dosažených výsledků. Část studií se soustředí na klasickou binární klasifikaci přítomnosti nebo nepřítomnosti onemocnění, zatímco jiné přístupy řeší dlouhodobější predikci rizika, například odhad pravděpodobnosti rozvoje koronárního srdečního onemocnění v určitém časovém horizontu. Významnou roli hraje i míra srozumitelnosti modelů, která bývá ve zdravotnictví považována za důležitou podmínku jejich praktického využití.

Cílem kapitoly je podat přehled vybraných existujících přístupů k predikci kardiovaskulárních onemocnění pomocí metod strojového učení. Pozornost je věnována konkrétním implementacím využívajícím reálná klinická data a veřejně dostupné analytické nástroje. Sledovány jsou zejména použité datové sady, zvolené modely, dosažené výsledky, způsob evaluace a úroveň interpretovatelnosti jednotlivých řešení.

Výběr studií vychází z jejich relevance k řešenému problému, dostupnosti implementace a možnosti srovnání s vlastní praktickou částí práce. Získané poznatky umožňují identifikovat opakující se metodické postupy, silné stránky jednotlivých přístupů i omezení, která se v oblasti predikce srdečních onemocnění objevují nejčastěji. Přehled současně vytváří základ pro pozdější porovnání vlastních výsledků s již existujícími řešeními.

1.4.1 Predikce srdečních onemocnění pomocí strojového učení

Jedním z praktických příkladů aplikace metod datové vědy ve zdravotnictví je projekt Heart-Disease-Detection-Using-ML, zaměřený na predikci srdečních onemocnění na základě klinických dat pacientů. Cílem projektu je vytvořit klasifikační modely schopné odhadnout, zda u konkrétního pacienta existuje zvýšené riziko srdeční choroby, a podpořit včasnou diagnostiku a rozhodování ve zdravotnické praxi (Doughty, 2025).

Použitý dataset vychází z veřejně dostupných klinických dat obsahujících základní fyziologické a diagnostické ukazatele, například věk, pohlaví, typ bolesti na hrudi, krevní tlak v klidu, hladinu cholesterolu, maximální dosaženou srdeční frekvenci nebo výsledky elektrokardiografického vyšetření. Cílová proměnná je binární a indikuje přítomnost nebo nepřítomnost srdečního onemocnění. Zmíněná struktura dat odpovídá běžné úloze dozorovaného učení, v níž je k dispozici množina vstupních atributů i známý výstup (Doughty, 2025).

Součástí projektu je předzpracování dat zahrnující odstranění chybějících hodnot, normalizaci numerických atributů a rozdělení dat na trénovací a testovací množinu. Normalizace je důležitá zejména u algoritmů citlivých na měřítko vstupních proměnných, například u metody k-nejbližších sousedů. Projekt dále využívá k-fold cross-validation, která umožňuje robustnější odhad výkonnosti modelů a snižuje riziko přeučení (Doughty, 2025).

Pro samotnou predikci jsou použity algoritmy Support Vector Machine, K-Nearest Neighbors a Naive Bayes. Jednotlivé modely jsou porovnávány pomocí standardních metrik, mezi které patří přesnost, citlivost a F1-skóre. Nejlepších výsledků podle autora dosahuje model založený na algoritmu K-Nearest Neighbors, který vykazuje vysokou schopnost správně identifikovat pacienty se srdečním onemocněním, což je v medicínské aplikaci zvlášť důležité (Doughty, 2025).

Významným prvkem projektu je důraz na srozumitelnost modelových výstupů. Autor využívá metodu SHAP, která umožňuje kvantifikovat vliv jednotlivých vstupních proměnných na výslednou predikci. Použití uvedené metody umožňuje identifikovat atributy s největším dopadem na rozhodování modelu, například typ bolesti na hrudi nebo maximální srdeční frekvenci. Vyšší transparentnost modelu následně podporuje jeho potenciální využitelnost v klinickém prostředí (Doughty, 2025).

Mezi hlavní výhody projektu patří relativní jednoduchost použitých modelů, dobrý výkon na dostupných datech a možnost interpretace výsledků pomocí vysvětlovacích metod. Současně je však nutné zmínit i omezení, především relativně malý rozsah datasetu a absenci validace na nezávislé klinické populaci. Právě takové okolnosti mohou snižovat generalizační schopnost dosažených výsledků a naznačují potřebu dalšího ověřování na rozsáhlejších a diverzifikovanějších datech (Doughty, 2025)

1.4.2 Predikce koronárního srdečního onemocnění na základě dlouhodobých klinických dat

Dalším příkladem využití metod datové vědy v oblasti kardiovaskulárních onemocnění je projekt Coronary-Heart-Disease-Prediction, zaměřený na predikci desetiletého rizika vzniku koronárního srdečního onemocnění. Projekt pracuje s dlouhodobými epidemiologickými daty a porovnává více algoritmů strojového učení s cílem nalézt model s co nejvyšší predikční schopností (Paul, 2020).

Použitá data pocházejí ze studie Framingham Heart Study, která patří mezi nejznámější dlouhodobé epidemiologické studie zaměřené na kardiovaskulární onemocnění. Dataset obsahuje více než čtyři tisíce záznamů a zahrnuje klinické i behaviorální atributy, například věk, pohlaví, systolický a diastolický krevní tlak, hladinu cholesterolu, kouření, diabetes nebo užívání antihypertenziv. Cílovou proměnnou je binární ukazatel, který vyjadřuje, zda se u pacienta v průběhu následujících deseti let rozvinulo koronární srdeční onemocnění (Paul, 2020).

Před modelováním jsou data podrobena předzpracování zahrnujícímu práci s chybějícími hodnotami, standardizaci numerických proměnných a rozdělení datasetu na trénovací a testovací množinu. Autor klade důraz na vyvážené hodnocení modelů a používá více metrik, aby se vyhnul zkresleným závěrům založeným pouze na jediné hodnotě přesnosti (Paul, 2020).

Pro predikci rizika je implementováno několik algoritmů strojového učení, konkrétně logistická regrese, k-nejbližších sousedů, rozhodovací stromy, Random Forest a Gradient Boosting. Výsledky ukazují, že ansámblové metody, zejména Random Forest, dosahují lepšího výkonu než jednodušší lineární modely. Zmíněný závěr naznačuje přítomnost nelineárních vztahů mezi vstupními proměnnými a cílovou proměnnou, které jednodušší modely nedokážou zachytit v plném rozsahu (Paul, 2020).

Významnou součástí projektu je i optimalizace parametrů a následné sestavení ansámblového řešení pomocí metody stacking. Kombinací více klasifikátorů bylo podle autora dosaženo velmi vysoké predikční přesnosti, která se v některých experimentech blíží hranici 95 %. Výsledek podporuje obecný předpoklad, že vhodná kombinace více modelů může vést ke stabilnějším a přesnějším predikcím než použití jednoho samostatného algoritmu (Paul, 2020).

Silnou stránkou projektu je práce s rozsáhlejším a dlouhodobě sledovaným datasetem, který lépe odpovídá reálným klinickým podmínkám. Projekt současně ukazuje přínos ansámblových metod pro řešení komplexních predikčních úloh ve zdravotnictví. Na druhé straně je nutné upozornit na nižší srozumitelnost výsledného modelu, zejména v případě stacking přístupů, a na skutečnost, že vysoká přesnost sama o sobě neznamená automatickou připravenost modelu pro přímé klinické použití bez další externí validace (Paul, 2020).

1.4.3 Predikce srdečního selhání pomocí metod strojového učení

Třetím analyzovaným příkladem je projekt Heart-Failure-Prediction-with-Machine-Learning, který se zaměřuje na predikci srdečního selhání na základě klinických dat pacientů. Srdeční selhání představuje závažný zdravotní stav s vysokou mortalitou, proto je včasná identifikace rizikových pacientů důležitým cílem moderní medicíny. Projekt ukazuje možnosti využití různých algoritmů strojového učení při podpoře klinického rozhodování (Duru, 2022).

Použitý dataset pochází z veřejně dostupného zdroje a obsahuje přibližně devět set záznamů pacientů s dvanácti klinickými atributy. Mezi sledované proměnné patří například věk, pohlaví, krevní tlak, hladina cholesterolu, přítomnost diabetu, kouření nebo výsledky laboratorních vyšetření. Cílovou proměnnou je binární ukazatel indikující, zda u pacienta došlo ke srdečnímu selhání. Datová struktura odpovídá běžné klasifikační úloze, v níž je cílem rozlišit pacienty s vyšším a nižším rizikem závažné kardiovaskulární události (Duru, 2022).

Před samotným modelováním jsou data podrobena standardním krokům předzpracování, které zahrnují kontrolu chybějících hodnot, kódování kategoriálních proměnných a normalizaci

numerických atributů. Projekt klade důraz na systematické porovnání jednotlivých algoritmů, přičemž dataset je rozdělen na trénovací a testovací část a výkonnost modelů je hodnocena pomocí více metrik, zejména přesnosti, citlivosti a F1-skóre (Duru, 2022).

Pro predikci srdečního selhání je implementováno několik algoritmů strojového učení, mezi které patří logistická regrese, Support Vector Machine, K-Nearest Neighbors, Random Forest, Gaussian Naive Bayes a XGBoost. Výsledky experimentů ukazují, že stromové a ansámblové metody, zejména Random Forest a XGBoost, dosahují nejlepších hodnot sledovaných metrik. Uvedené modely jsou schopny zachytit složitější nelineární vztahy mezi klinickými proměnnými, které mohou být u jednodušších lineárních přístupů přehlédnuty (Duru, 2022).

Součástí projektu je i snaha zvýšit srozumitelnost modelů, například prostřednictvím analýzy důležitosti jednotlivých atributů. Daný postup umožňuje identifikovat faktory s největším vlivem na predikci srdečního selhání a poskytuje cenné informace nejen datovým analytikům, ale i zdravotnickým pracovníkům. Srozumitelnost má v uvedeném kontextu zásadní význam, protože podporuje důvěru v model a usnadňuje jeho případné začlenění do klinického rozhodování (Duru, 2022).

Přínosem projektu je komplexní porovnání širšího spektra algoritmů strojového učení a praktická demonstrace jejich využití při predikci závažného kardiovaskulárního onemocnění. Omezení spočívají především v relativně malém rozsahu datasetu a v absenci externí validace na nezávislé populaci pacientů. Získané výsledky je proto nutné interpretovat opatrně a dále ověřovat v rozsáhlejších klinických studiích (Duru, 2022).

1.4.4 Diagnostika koronárního onemocnění tepen pomocí porovnání více algoritmů

Další relevantní příspěvek k problematice kardiovaskulární predikce představuje studie Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms. Garavand a kol. (2022) porovnávají více klasifikačních algoritmů při diagnostice koronárního onemocnění tepen na základě klinických vyšetřovacích znaků. Studie pracuje s 303 záznamy a s klinicky vybranými atributy, přičemž autoři zdůrazňují snahu o vytvoření diagnostického modelu založeného na dostupných a relativně levných vyšetřeních.

Do porovnání byly zařazeny mimo jiné SVM, Random Forest, logistická regrese, KNN, Naive Bayes a další klasifikátory. Nejlepších výsledků dosáhly podle autorů především modely SVM a Random Forest, zatímco KNN vykazoval nižší efektivitu. Autoři zároveň používají více hodnoticích ukazatelů, zejména AUC, F-measure, MCC a ROC, což zvyšuje vypovídací hodnotu porovnání a umožňuje posoudit modely z více hledisek než pouze podle jediné hodnoty přesnosti.

Přínos studie spočívá hlavně v potvrzení, že i při využití tabulkových klinických dat lze dosáhnout dobrých diagnostických výsledků bez nutnosti použití obrazových nebo laboratorně velmi náročných vstupů. Současně se ukazuje, že rozdíly mezi algoritmy nejsou zanedbatelné a že výběr modelu výrazně závisí na charakteru dat i na použitém způsobu evaluace. Z metodického hlediska je cenný i důraz na porovnání více algoritmů nad jedním problémem, protože podobný přístup umožňuje lépe rozlišit, zda dobrý výsledek vychází z vlastností datasetu, nebo z konkrétní modelové architektury.

Určité omezení studie souvisí s velikostí datového souboru a s absencí širší externí validace. Autoři sice dosahují velmi dobrých výsledků, avšak přenositelnost modelu do odlišného klinického prostředí zůstává otevřenou otázkou. Další omezení vyplývá z toho, že analýza je soustředěna především na diagnostický výkon, zatímco menší pozornost je věnována hlubší interpretaci rozdílů mezi jednotlivými modely nebo širšímu zhodnocení kalibrace pravděpodobnostních odhadů.

1.4.5 Predikce srdečního selhání po akutním infarktu myokardu pomocí více modelů strojového učení

Z hlediska oficiálního zadání práce je zvlášť významná studie Development and comparison of machine learning-based models for predicting heart failure after acute myocardial infarction. Li a kol. (2023) se zaměřují na predikci vzniku srdečního selhání po akutním infarktu myokardu a porovnávají sedm algoritmů strojového učení s cílem nalézt nejvhodnější model pro klinické použití. Studie je cenná hlavně proto, že propojuje medicínsky závažný problém s přímým srovnáním více modelových přístupů a současně pracuje s externí validací.

Autoři porovnávají sedm algoritmů a uvádějí, že cílem bylo vytvořit optimální model pro predikci srdečního selhání u pacientů po akutním infarktu myokardu. Podstatnou předností studie je právě srovnávací rámec, protože umožňuje posoudit nejen absolutní výkon jednotlivých modelů, ale i jejich stabilitu a potenciální využitelnost v klinické praxi. Externí validační přístup zároveň posiluje důvěryhodnost výsledků více než běžné jednorázové rozdělení dat na trénovací a testovací část.

Význam studie spočívá i v tematické blízkosti vůči problému řešenému v bakalářské práci. Na rozdíl od obecnějších projektů zaměřených na přítomnost srdečního onemocnění se zde analyzuje užší a klinicky velmi závažný fenotyp, tedy srdeční selhání po akutní ischemické události. Podobný směr výzkumu ukazuje, že metody strojového učení nacházejí uplatnění nejen při obecné kardiovaskulární klasifikaci, ale i při přesněji vymezených klinických scénářích. Z rešeršního hlediska jde o důležitý most mezi obecnou predikcí srdečních onemocnění a přesnějším zaměřením na srdeční selhání.

Určité omezení vyplývá z faktu, že i při použití silnější validace zůstává model svázán s konkrétním klinickým kontextem a s vlastnostmi zvolených kohort. Přenos výsledků do jiných zdravotnických prostředí proto vyžaduje další ověření. Přesto jde o metodicky hodnotnou studii, protože ukazuje, že srovnání více algoritmů v kombinaci s validací a klinickou interpretací poskytuje podstatně pevnější základ pro závěry než izolované hodnocení jednoho modelu.

1.4.6 Predikce mortality u pacientů se srdečním selháním pomocí strojového učení

Další odbornou studií relevantní k řešené problematice je práce Machine learning-based model for predicting 1 year mortality of hospitalized patients with heart failure od Tohyamy a kol. (2021). Studie se zaměřuje na predikci jednoroční mortality u hospitalizovaných pacientů se srdečním selháním a představuje příklad využití metod strojového učení v úzce vymezeném, klinicky velmi závažném kontextu. Tematické zaměření studie je důležité zejména z hlediska návaznosti na oficiální zadání práce, protože přenáší pozornost od obecné predikce srdečního onemocnění k přesněji definovanému problému srdečního selhání.

Autoři pracují s klinickými daty hospitalizovaných pacientů a porovnávají model založený na strojovém učení s konvenčnějšími predikčními přístupy. Cílem není pouze dosáhnout vysoké predikční úspěšnosti, ale také ověřit, zda lze pomocí širší kombinace klinických znaků vytvořit model schopný přesnější stratifikace rizika. Význam studie spočívá ve snaze zachytit komplexnější vztahy mezi proměnnými, které mohou být u tradičních jednodušších přístupů částečně potlačeny (Tohyama a kol., 2021).

Metodický přínos práce je patrný hlavně v klinickém zaměření predikční úlohy. Na rozdíl od obecnějších heart disease datasetů je zde cílem odhad závažného klinického výsledku u pacientů již zatížených konkrétní diagnózou. Predikční model nepracuje pouze s obecným rozlišením přítomnosti nebo nepřítomnosti srdeční patologie, ale s detailněji vymezenou prognostickou úlohou. Podobné pojetí je důležité i z metodického hlediska, protože ukazuje, že strojové učení nachází uplatnění nejen při diagnostické klasifikaci, ale i při odhadu budoucího klinického průběhu.

Omezení studie souvisejí především s vazbou modelu na konkrétní klinickou populaci a na strukturu dostupných vstupních proměnných. Přesto jde o významný příspěvek do rešeršní části práce, protože rozšiřuje spektrum analyzovaných přístupů o seriózní odbornou studii přímo navázanou na problematiku srdečního selhání a současně ukazuje, že klinicky orientované predikční modely mohou být stavěny i na relativně standardních zdravotnických datech (Tohyama a kol., 2021).

1.4.7 Závěrečné srovnání studií

Zařazené práce se liší použitým typem dat, vymezením cílové úlohy i zvolenými algoritmy, avšak spojuje je snaha zvýšit kvalitu včasné identifikace pacientů s vyšším kardiovaskulárním rizikem. GitHub projekty analyzované v podkapitolách 1.4.1 až 1.4.3 představují prakticky orientované ukázky využití běžně dostupných klinických dat a standardních metod strojového učení. Odborné publikace Garavanda a kol. (2022), Li a kol. (2023) a Tohyamy a kol. (2021) naproti tomu přinášejí metodicky pevnější rámec, založený na systematičtějším porovnání modelů, klinicky přesnějším vymezení problému a v některých případech i na důslednějším validačním postupu.

První skupina analyzovaných prací zdůrazňuje především praktickou implementaci modelů, základní předzpracování dat a porovnání výkonu několika klasifikátorů. Přínos podobných projektů spočívá hlavně v názorné demonstraci toho, že i nad běžně dostupnými klinickými daty lze vytvářet funkční klasifikační řešení. Odborné publikace však posouvají pozornost více k metodické robustnosti, širšímu souboru hodnoticích ukazatelů a silnější klinické obhajitelnosti výsledků. Garavand a kol. (2022) ukazují, že i nad relativně omezeným klinickým datasetem mohou SVM a Random Forest dosahovat velmi dobrých výsledků, Li a kol. (2023) potvrzují význam srovnání více algoritmů při predikci srdečního selhání po akutním infarktu myokardu a Tohyama a kol. (2021) rozšiřují pohled směrem k prognostickému modelování mortality u pacientů se srdečním selháním.

Společným rysem všech analyzovaných studií je práce se standardními klinickými proměnnými a využití klasických i pokročilejších metod strojového učení. Opakovaně se objevují algoritmy jako logistická regrese, KNN, SVM, Random Forest nebo další ansámblové přístupy. Výsledky napříč literaturou potvrzují, že tabulková klinická data mohou poskytovat dostatečný predikční signál pro úspěšné modelování kardiovaskulárních stavů. Literatura však současně upozorňuje

na opakující se omezení, mezi něž patří menší rozsah datasetů, absence externí validace, metodická nejednotnost evaluace a nižší srozumitelnost komplexnějších modelů. Zvláštní pozornost si zaslouží rozdíl mezi diagnostickou a prognostickou úlohou. Zatímco část studií se soustředí na rozlišení přítomnosti onemocnění, jiné práce, například Tohyama a kol. (2021), řeší odhad budoucího klinického vývoje, což klade vyšší nároky na interpretaci i validaci.

Rešeršní přehled vytváří důležitý rámec pro vlastní praktickou část práce. Na jedné straně potvrzuje, že použití více klasifikačních algoritmů, důraz na interpretovatelnost a práce s klinickými atributy odpovídají běžné výzkumné praxi v kardiovaskulární predikci. Na straně druhé ukazuje prostor pro rozšíření analytického rámce o vícerozměrnou exploraci bez učitele, profilování shluků, statistické hodnocení stability modelů a pravděpodobnostní interpretaci výsledků. Právě v uvedených bodech bude praktická část práce na dosavadní literaturu navazovat a současně ji částečně rozšiřovat.

2 Praktická část

Praktická část práce byla navržena jako aplikace kompletního životního cyklu datové vědy na úlohu predikce srdečního onemocnění. Zvolený přístup kombinuje explorativní analýzu dat, učení bez učitele, učení s učitelem, interpretaci modelů a návrh interaktivního notebooku určeného pro koncového uživatele. Cílem nebylo pouze dosažení co nejvyšší predikční úspěšnosti, ale i vytvoření srozumitelného a opakovatelně použitelného analytického řešení, které propojí medicínskou interpretaci s metodami datové vědy. Kompletní notebook je k dispozici z následujícího odkazu: https://www.kaggle.com/code/jakubskalick/datov-v-da-predikce-v-skytu-srde-n-ho-selh-n.

2.1 Metodika

Kapitola metodika vymezuje postup, kterým byla praktická část práce realizována, a současně vytváří vazbu mezi teoretickými východisky a konkrétní analytickou implementací. Pozornost je věnována charakteristice použitého datového souboru, způsobu jeho zpracování i výběru metod použitých v jednotlivých fázích životního cyklu datové vědy. Popsaný postup zahrnuje přípravu a předzpracování dat, explorativní analýzu, metody učení bez učitele, řízenou analýzu významnosti příznaků, klasifikační modelování, doplňkový regresní experiment i návrh interaktivních prvků notebooku. Smyslem kapitoly není prezentace samotných výsledků, ale systematické vysvětlení, jakým způsobem bylo k výsledkům dosaženo a proč byly zvoleny právě dané analytické postupy.

2.1.1 Výzkumný přístup a metodologický rámec

Metodologický rámec vychází z principů datově orientovaného experimentování a z požadavku na iterativní vývoj řešení. Jednotlivé kroky na sebe navazovaly v logice formulace problému, porozumění datům, čištění a transformace dat, explorace vztahů, segmentace pacientů, návrhu predikčních modelů, jejich vyhodnocení a interpretace. Zmíněné pojetí odpovídá obecné logice životního cyklu datové vědy i doporučené struktuře bakalářské práce, v níž má metodika zajišťovat opakovatelnost a jednoznačný přehled použitých postupů.

Výběr metod vycházel ze dvou kritérií. První kritérium představovala věcná vhodnost pro řešený problém. Druhé kritérium představovala interpretovatelnost výsledků pro čtenáře orientovaného na zdravotnickou problematiku. Z uvedeného důvodu byla do analýzy zařazena vedle výkonnějších modelů i metoda logistické regrese a vedle samotné evaluace výkonu i postupy, které umožňují vysvětlit význam jednotlivých proměnných. Význam srozumitelnosti modelů v medicíně zdůrazňují i práce zahrnuté do rešerše.

2.1.2 Charakteristika a zdroj použitých dat

Analýza byla provedena nad veřejně dostupným datovým souborem zaměřeným na predikci srdečního onemocnění. Datová sada obsahuje klinické a diagnostické charakteristiky pacientů, mezi něž patří věk, pohlaví, typ bolesti na hrudi, klidový krevní tlak, hladina cholesterolu, informace o glykemii nalačno, výsledek klidového EKG, maximální dosažená srdeční frekvence, přítomnost zátěžové anginy, hodnota Oldpeak a sklon ST segmentu (úsek elektrokardiografické

křivky mezi komplexem QRS a vlnou T, která souvisí s repolarizací komor). Cílová proměnná HeartDisease je binární a vyjadřuje přítomnost nebo nepřítomnost srdečního onemocnění. Zmíněná struktura odpovídá standardní úloze dozorovaného učení. Podobný typ dat i úloh zmiňuje i dosavadní text práce v části věnované přehledu existujících řešení.

Použitý soubor byl zvolen ze tří důvodů. Zaprvé obsahuje proměnné s přímou klinickou interpretací. Zadruhé je dostatečně kompaktní pro přehlednou demonstraci celého analytického cyklu v rámci bakalářské práce. Zatřetí umožňuje současně provést základní exploraci, shlukování i klasifikační modelování bez nutnosti řešit vysoce specializované formy dat, například obrazové nebo časové signály. Ve vztahu k formulovanému cíli práce proto představuje vhodný kompromis mezi srozumitelností a analytickou hodnotou.

2.1.3 Analytické prostředí a použitý software

Praktická realizace byla provedena v prostředí Python notebooku publikovaného na platformě Kaggle. Volba uvedeného prostředí souvisela s několika požadavky současně: s podporou práce s datovými soubory, s možností průběžné vizualizace výsledků, s veřejným sdílením notebooku a s možností opětovného spuštění na cizím účtu. Součástí metodiky proto není pouze výběr analytických metod, ale i návrh formy prezentace výsledků, aby byl notebook čitelný i pro uživatele, který není specialistou na datovou analytiku.

Základ technologického řešení tvořily knihovny pandas a numpy pro práci s daty, matplotlib a seaborn pro vizualizace, scikit-learn pro předzpracování dat, klasifikační modelování, křížovou validaci a hodnocení modelů, dále UMAP-learn pro redukci dimenze, kmodes pro algoritmus K-Prototypes a SHAP pro interpretaci významnosti příznaků. Volba Pythonu a specializovaných knihoven odpovídá charakteristice současné datové vědy, v níž hrají hlavní roli otevřené a reprodukovatelné nástroje podporující experimentování a interpretaci výsledků.

2.1.4 Příprava a předzpracování dat

První krok praktické části spočíval v načtení dat, kontrole jejich struktury a rozdělení proměnných na numerické, kategoriální a binární. Uvedený krok byl důležitý nejen pro základní orientaci v datech, ale i pro správné nastavení následného předzpracování, protože různé typy proměnných vyžadují odlišný způsob zacházení.

Následovala kontrola chybějících hodnot, datových typů a základních deskriptivních charakteristik. Soubor nevyžadoval rozsáhlou imputaci, hlavní pozornost se proto soustředila na konzistenci dat, kontrolu potenciálně problematických hodnot a přípravu proměnných pro další analytické kroky. U numerických proměnných byla sledována přítomnost odlehlých hodnot a tvar rozdělení, u kategoriálních proměnných četnost jednotlivých kategorií a případná nerovnoměrnost jejich zastoupení.

Před modelováním byly numerické proměnné v příslušných pipeline standardizovány a kategoriální proměnné kódovány pomocí One-Hot Encoding. Uvedený postup byl zvolen proto, že část použitých modelů je citlivá na měřítko vstupních proměnných a současně vyžaduje numerickou reprezentaci kategoriálních atributů. V rámci metodiky šlo o řízené předzpracování navázané na konkrétní rodiny modelů, nikoli o mechanickou transformaci všech dat jednotným způsobem.

2.1.5 Explorativní analýza dat

Explorativní analýza byla členěna do několika navazujících úrovní. První úroveň zahrnovala jednorozměrnou analýzu numerických, kategoriálních a binárních proměnných. U numerických atributů byly využity deskriptivní statistiky, histogramy, boxploty, testy normality a koeficienty šikmosti. U kategoriálních a binárních proměnných byly zpracovány relativní četnosti, sloupcové grafy a testy dobré shody. Smyslem uvedeného kroku nebylo pouze popsat datový soubor, ale ověřit jeho klinickou smysluplnost, rozpoznat výrazně asymetrické rozdělení, odhalit nápadně nerovnoměrné kategorie a připravit půdu pro interpretaci následných modelů.

Druhá úroveň explorace spočívala ve dvourozměrné analýze vztahů mezi proměnnými. U dvojic numerických proměnných byly použity Pearsonovy a Spearmanovy korelace, u kombinací kategoriálních a numerických proměnných korelační poměr η a boxploty, u kategoriálních proměnných vůči cílové proměnné Cramérovo V a relativní četnosti. Výběr prezentovaných dvojic se soustředil na nejvýraznější a klinicky interpretovatelné vazby, zejména na proměnné spojené se zátěžovým vyšetřením, bolestí na hrudi a změnami ST segmentu.

Třetí úroveň explorace měla vícerozměrný charakter. Byly využity korelační heatmapy, pairplot numerických proměnných a následně i rozšířená explorativní analýza pomocí redukce dimenze. Uvedený postup umožnil ověřit, že datový soubor obsahuje smysluplné vztahy, ale současně nevykazuje jednoduchou lineární separaci tříd. Instrukce k BP výslovně doporučují zařadit do explorativní části nejen popisné statistiky a jednorozměrné grafy, ale i dvourozměrné a vícerozměrné vizualizace.

2.1.6 Rozšířená explorativní analýza a učení bez učitele

Rozšířená EDA byla zaměřena na vícerozměrnou strukturu pacientských profilů. Pro redukci dimenze byla použita metoda UMAP, jejímž cílem nebyla konstrukce prediktoru, ale zobrazení podobnosti pacientů v latentním prostoru. UMAP byl zvolen proto, že dokáže zachytit i nelineární strukturu dat a lépe než čistě lineární projekce odhalit případné přechodové oblasti a částečně oddělené skupiny pacientů.

Na projekci následně navázalo shlukování smíšených dat pomocí algoritmu K-Prototypes, který je vhodný pro soubory obsahující současně numerické i kategoriální proměnné. Volba uvedeného algoritmu odpovídala struktuře analyzovaného datasetu, v němž by použití čistě numerického shlukovacího postupu vedlo ke ztrátě informace obsažené v kategoriálních a binárních znacích. Smyslem shlukování nebyla tvorba diagnostické klasifikace, ale odhalení opakujících se pacientských profilů.

Interpretace shluků byla doplněna profilováním. U numerických proměnných byly porovnány průměry, u kategoriálních proměnných modální hodnoty a zastoupení cílové proměnné. Dále byly použity boxploty pro vybrané klinicky významné atributy a jednoduchý rozhodovací strom nad číslem shluku jako cílovou proměnnou. Uvedený krok umožnil převést anonymní čísla shluků do klinicky čitelnějšího popisu skupin pacientů. Zařazení učení bez učitele do metodiky odpovídá doporučené struktuře práce, v níž má shluková analýza tvořit samostatnou fázi po explorativní části a před finálním modelováním.

2.1.7 Supervised EDA a výběr významných proměnných

Před samotným klasifikačním modelováním byla provedena řízená analýza významnosti příznaků se znalostí cílové proměnné. Použity byly dvě vzájemně se doplňující metody: Mutual Information a SHAP. Mutual Information sloužila k odhadu informačního přínosu jednotlivých proměnných vůči binárnímu cíli. SHAP pak umožnil interpretovat globální význam proměnných a směr jejich vlivu na predikovanou pravděpodobnost onemocnění.

Zařazení uvedené fáze mělo dvojí účel. První účel představovalo ověření, zda se jako významné znovu objeví proměnné, které již byly důležité v explorativní a vícerozměrné analýze. Druhý účel spočíval v přípravě interpretace finálních modelů, aby výkon klasifikátorů nebyl prezentován izolovaně od obsahu použitých atributů.

2.1.8 Klasifikační modelování

Hlavní predikční úloha měla charakter binární klasifikace. Do srovnání bylo zařazeno pět modelů reprezentujících odlišné principy učení: logistická regrese, rozhodovací strom, náhodný les, KNN a SVM. Výběr pokrývá lineární přístup, stromovou metodu, ansámblový model, instanční klasifikátor a metodu založenou na maximálním oddělení tříd. Zjištěné spektrum odpovídá doporučení vyzkoušet několik principiálně odlišných tříd modelů.

Modelování bylo realizováno prostřednictvím pipeline, v nichž bylo předzpracování dat svázáno s konkrétním modelem. Rozdělení dat na trénovací a testovací část bylo provedeno se zachováním podílu tříd. Pro ladění hyperparametrů byla využita křížová validace a Grid Search. Nastavená metodika umožnila porovnávat modely za srovnatelných podmínek a současně minimalizovat riziko, že výsledek bude dán nevhodným jednorázovým nastavením parametrů.

Vyhodnocení výkonu se neopíralo o jedinou metriku. Sledovány byly zejména Accuracy, Precision, Recall, F1-score, MCC, ROC AUC a PR AUC, dále matice záměn, ROC a Precision–Recall křivky a learning curve pro nejúspěšnější model. Zařazeno bylo rovněž statistické hodnocení metrik pomocí opakované křížové validace, 95% intervalů spolehlivosti a Forest Plotu.

2.1.9 Regresní experiment

Vedle základní klasifikační úlohy byl realizován i experiment založený na převodu binární klasifikace na regresní interpretaci. Uvedený postup odpovídá doporučeným experimentům, které připouštějí převedení binárního výstupu na spojitou hodnotu vyjadřující pravděpodobnost příslušnosti k pozitivní třídě. V praxi byla využita logistická regrese a predikce pravděpodobnosti predict_proba, kdy cílem nebylo přiřadit pouze třídu 0 nebo 1, ale odhadnout míru rizika srdečního onemocnění v intervalu od 0 do 1.

Hodnocení regresního experimentu vycházelo z metrik MAE, RMSE, R^2, Brier score a ROC AUC. Zařazeny byly i kalibrační grafy, histogram predikovaných pravděpodobností, ukázky individuálních predikcí a statistické hodnocení pomocí intervalů spolehlivosti a Forest Plotu. Smyslem experimentu bylo ověřit, zda lze problém interpretovat nejen jako klasifikační rozhodnutí, ale i jako odhad intenzity rizika, což je pro zdravotnický kontext často užitečnější než čistě binární odpověď.

2.1.10 Interaktivní prvky notebooku

Součástí povinné základní úlohy byl i návrh interaktivity. Instrukce požadují možnost práce s vlastním datovým souborem, možnost zadání jednoho vlastního vzorku a parametrickou interaktivitu alespoň u jedné metody. Praktická realizace se opírala o prostředí Kaggle a byla navržena s ohledem na omezení podpory widgetů v běžném zobrazení notebooku. Proto byla finální interaktivita řešena konzervativněji, ale reprodukovatelně.

Datová interaktivita zahrnovala možnost načíst vlastní CSV soubor se stejnou strukturou proměnných a získat nad ním predikce naučeného modelu. Druhá část umožňovala ruční zadání hodnot jednoho pacienta a výpočet pravděpodobnosti srdečního onemocnění. Parametrická interaktivita byla řešena změnou vybraných hyperparametrů logistické regrese a opětovným přepočtem výstupních metrik. Uvedené řešení sice není postaveno primárně na grafických widgetech, ale splňuje hlavní funkční požadavek: uživatel může měnit vstupní data nebo parametry a sledovat dopad na výsledky.

2.1.11 Strukturovaný přehled použitých metod napříč životním cyklem datové vědy

Závěr metodické kapitoly tvoří souhrnný přehled fází a metod aplikovaných v praktické části práce a v doprovodném notebooku:

1. Formulace problému: predikce přítomnosti srdečního onemocnění; doplňkově odhad pravděpodobnosti rizika.
2. Zdroj a porozumění datům: veřejný dataset; identifikace proměnných, jejich typů a cílové proměnné.
3. Předzpracování dat: kontrola chybějících hodnot, kontrola konzistence, standardizace numerických atributů, One-Hot Encoding kategoriálních proměnných, rozdělení na trénovací a testovací část.
4. Explorativní analýza dat: deskriptivní statistiky, histogramy, boxploty, četnosti kategorií, testy normality, šikmost, korelace, Cramérovo V, korelační poměr η, heatmapy, pairplot.
5. Učení bez učitele: UMAP pro redukci dimenze, K-Prototypes pro shlukování smíšených dat, profilování shluků, doplňkový rozhodovací strom nad číslem shluku.
6. Supervised EDA: Mutual Information, SHAP.
7. Učení s učitelem – klasifikace: logistická regrese, rozhodovací strom, náhodný les, KNN, SVM; ladění hyperparametrů; křížová validace; matice záměn; ROC, PR, learning curve; intervaly spolehlivosti a Forest Plot.
8. Experiment nad rámec základní úlohy: převod klasifikace na regresní interpretaci prostřednictvím pravděpodobnosti pozitivní třídy; metriky MAE, RMSE, R^2, Brier score, ROC AUC; kalibrační vyhodnocení; intervaly spolehlivosti a Forest Plot.
9. Interaktivita notebooku: nahrání vlastního CSV souboru, zadání jednoho pacienta, ruční parametrická změna vybraného modelu.

Zařazení souhrnného přehledu použitých fází a metod přispívá k větší přehlednosti metodické kapitoly a poskytuje ucelený obraz o struktuře a logice praktické části práce.

2.1.12 Použití AI při tvorbě řešení

Při návrhu notebooku i při průběžném ladění textových interpretací byly využity nástroje generativní umělé inteligence. Role AI spočívala především v návrhu alternativních implementačních variant, ve formulaci textových komentářů k jednotlivým částem notebooku, v konzultaci interpretace výstupů a v úpravách struktury notebooku, aby byl srozumitelnější pro doménového experta. Konečný výběr metod, jejich kombinace, způsob evaluace a výsledná interpretace však byly prováděny autorsky a průběžně ověřovány v kontextu řešeného problému. Instrukce výslovně požadují, aby použití AI při generování notebooku i kódu bylo v práci přiznáno a řádně popsáno.

2.2 Výsledky analýz

Kapitola předkládá hlavní zjištění získaná při zpracování datového souboru zaměřeného na predikci srdečního onemocnění. Pozornost je věnována výsledkům explorativní analýzy, vícerozměrnému pohledu na strukturu pacientských profilů, významnosti vstupních proměnných, porovnání klasifikačních modelů i doplňkovému regresnímu experimentu. Výklad sleduje stejnou logiku jako praktická realizace v notebooku, avšak důraz je kladen především na interpretaci analytických výstupů a na jejich význam pro řešený medicínský problém.

2.2.1 Výsledky explorativní analýzy dat

Explorativní analýza byla zaměřena na základní popis datového souboru, identifikaci hlavních charakteristik jednotlivých proměnných a ověření, zda data vykazují klinicky smysluplnou strukturu. Cílem nebylo pouze shrnout rozdělení hodnot, ale i posoudit, zda vstupní proměnné nesou potenciálně využitelnou informaci pro další analytické kroky.

2.2.2 Jednorozměrná analýza numerických proměnných

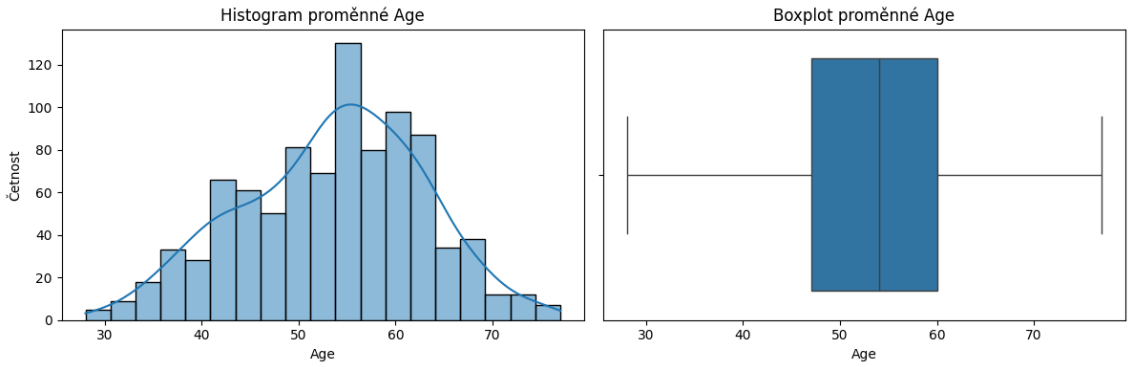
Jednorozměrná analýza numerických proměnných nepředstavuje pouze úvodní statistický popis, ale i první ověření odborné smysluplnosti dat. Jednorozměrná analýza numerických proměnných poskytla základní přehled o rozložení fyziologických a diagnostických ukazatelů v analyzovaném souboru. Posuzován byl zejména rozsah hodnot, přítomnost asymetrie, odlehlých pozorování a případná odchylka od normálního rozdělení. Smyslem nebylo pouze statisticky popsat jednotlivé atributy, ale i ověřit, zda jejich rozdělení odpovídá očekávatelnému klinickému obrazu pacientské populace a zda nevykazují nápadné nebo obtížně interpretovatelné chování.

V následujících obrázcích jsou zobrazeny výsledky jednorozměrné analýzy vybraných numerických proměnných, vždy ve formě histogramu a boxplotu. Součástí interpretace je i zohlednění asymetrie rozdělení, variability hodnot a přítomnosti extrémnějších pozorování.

Tab. 2: Četnost numerických proměnných

	count	mean	std	min	25 %	50 %	75 %	max
Age	918.0	53.510893	9.432617	28.0	47.0	54.0	60.0	77.0
RestingBP	918.0	132.538126	17.990127	80.0	120.0	130.0	140.0	200.0
Cholesterol	918.0	243.204793	53.401297	85.0	214.0	237.0	267.0	603.0
MaxHR	918.0	136.809368	25.460334	60.0	120.0	138.0	156.0	202.0
Oldpeak	918.0	0.887364	1.066570	-2.6	0.0	0.6	1.5	6.2

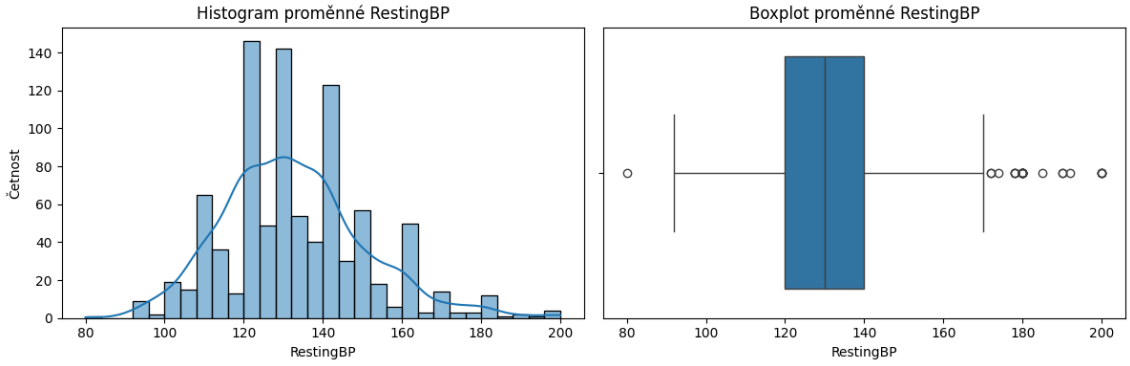
Zdroj: Vlastní zpracování (2026)



Obr. 6: Rozdělení hodnot proměnné Age (histogram a boxplot)

Zdroj: Vlastní zpracování (2026)

Proměnná Age vykazuje rozdělení blízké přibližně symetrickému tvaru, bez výrazné deformace nebo extrémních odlehlých hodnot. Histogram naznačuje koncentraci pacientů zejména ve středních a vyšších věkových kategoriích, což odpovídá očekávání, že analyzovaný soubor bude obsahovat převážně osoby ve věku, v němž se kardiovaskulární obtíže objevují častěji. Boxplot současně ukazuje relativně kompaktní rozložení bez nápadně izolovaných pozorování. Z interpretačního hlediska jde o proměnnou, která působí stabilně a dobře použitelně pro další modelování.

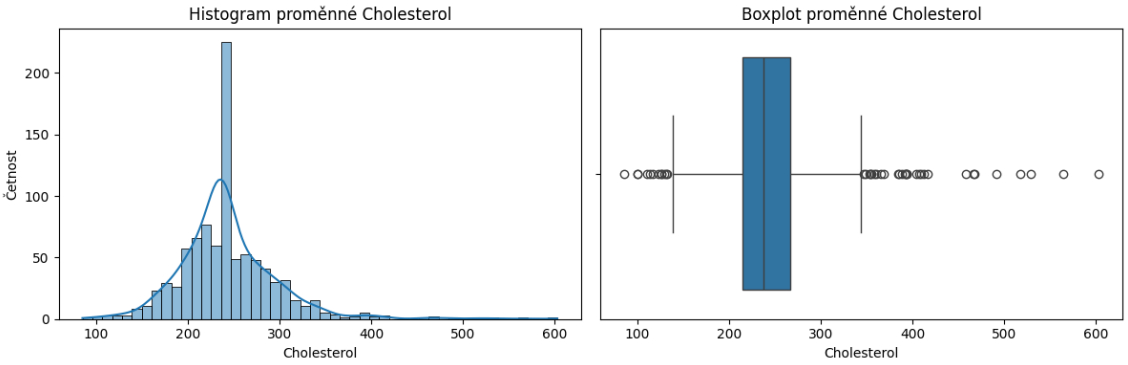


Obr. 7: Rozdělení hodnot proměnné RestingBP (histogram a boxplot)

Zdroj: Vlastní zpracování (2026)

Proměnná RestingBP vykazuje širší rozptyl hodnot a mírnou pravostrannou asymetrii. Histogram ukazuje, že většina pacientů se soustřeďuje v běžném rozmezí klidového krevního tlaku, avšak směrem k vyšším hodnotám je patrný delší ocas rozdělení. Boxplot potvrzuje přítomnost

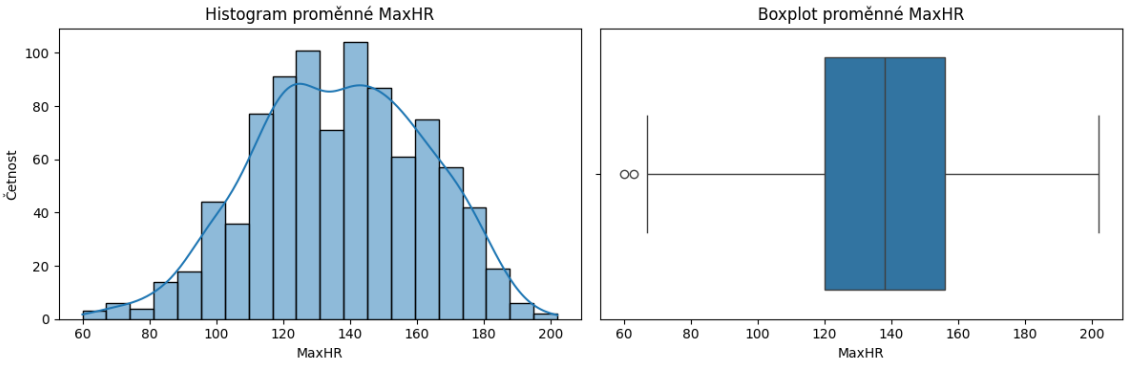
několika vyšších odlehlých hodnot. Zmíněný výsledek je z klinického hlediska dobře interpretovatelný, protože zvýšený krevní tlak patří mezi významné rizikové faktory kardiovaskulárních onemocnění a v populaci pacientů lze očekávat i menší skupinu osob s výrazně vyššími hodnotami.



Obr. 8: Rozdělení hodnot proměnné Cholesterol (histogram a boxplot)

Zdroj: Vlastní zpracování (2026)

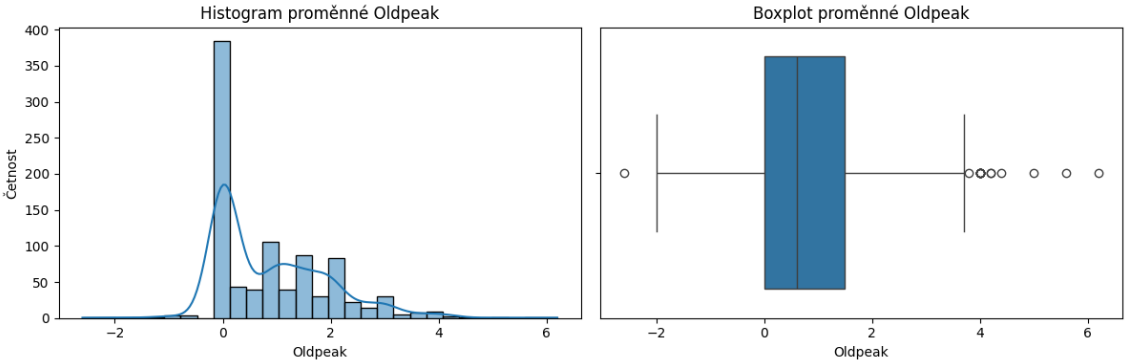
Proměnná Cholesterol vykazuje výraznější pravostrannou asymetrii a současně i větší množství odlehlých hodnot. Histogram ukazuje koncentraci hodnot v nižší a střední části rozmezí, zatímco menší část pacientů dosahuje výrazně vyšších koncentrací cholesterolu. Boxplot potvrzuje přítomnost četnějších extrémnějších pozorování na horním konci rozdělení. Při interpretaci je však třeba zohlednit, že část atypického chování dané proměnné souvisí i s předchozím ošetřením nulových hodnot, které byly v datovém souboru nahrazeny průměrnou hodnotou. Daný zásah ovlivnil výsledný tvar rozdělení a přispěl k, že se v okolí průměrné hodnoty vytváří výraznější koncentrace pozorování. Z analytického hlediska proto proměnná odráží nejen přirozenou biologickou variabilitu, ale i dopad zvoleného způsobu předzpracování dat, který je při další interpretaci nutné brát v úvahu.



Obr. 9: Rozdělení hodnot proměnné MaxHR (histogram a boxplot)

Zdroj: Vlastní zpracování (2026)

Proměnná MaxHR působí z hlediska rozdělení relativně vyváženě a blíží se přibližně symetrickému tvaru. Histogram ukazuje poměrně plynulé rozložení hodnot kolem střední části rozsahu a boxplot neodhaluje výrazný problém s odlehlými pozorováními. Ve srovnání s jinými numerickými proměnnými se jedná o ukazatel s méně deformovaným rozdělením, což může být výhodné z hlediska jeho další interpretace i modelového využití. Z klinického pohledu je důležité, že maximální dosažená srdeční frekvence vykazuje dostatečnou variabilitu napříč pacienty.



Obr. 10: Rozdělení hodnot proměnné OldPeak (histogram a boxplot)

Zdroj: Vlastní zpracování (2026)

Proměnná Oldpeak vykazuje nejvýraznější pravostrannou asymetrii ze sledovaných numerických proměnných. Histogram ukazuje vysokou koncentraci nízkých hodnot a postupné řídnutí směrem k vyšším hodnotám, zatímco boxplot potvrzuje přítomnost několika vyšších extrémnějších pozorování. Zmíněný výsledek je metodicky významný, protože ukazuje, že proměnná nenese informaci rovnoměrně po celém rozsahu, ale že vyšší hodnoty se vyskytují pouze u části pacientů. Z medicínského hlediska je daná vlastnost důležitá, protože může odrážet rozdílnou intenzitu zátěžových změn mezi méně rizikovými a více rizikovými profily.

Tab. 3: Testy normality a šikmosti numerických proměnných

Proměnná	Šikmost	Test normality – statistika	Test normality – p-hodnota
Age	-0.1959	14.3873	0.0008
RestingBP	0.6080	63.0376	0.0000
Cholesterol	1.4470	315.1115	0.0000
MaxHR	-0.1444	15.8325	0.0004
Oldpeak	1.0229	139.0609	0.0000

Zdroj: Vlastní zpracování (2026)

Tabulka testů normality a koeficientů šikmosti uvedené grafické závěry dále potvrzuje. U části proměnných se rozdělení přibližuje symetrickému tvaru, avšak zejména u proměnných Cholesterol a Oldpeak je patrná výraznější pravostranná asymetrie. Odchylka od normálního rozdělení nepředstavuje sama o sobě nedostatek dat, ale spíše informaci o jejich skutečné struktuře. V klinickém kontextu je přirozené, že některé fyziologické a diagnostické ukazatele nevykazují ideálně normální průběh, protože odrážejí různorodé zdravotní stavy pacientů.

Celkově jednorozměrná analýza numerických proměnných ukázala, že analyzovaný soubor nevytváří jednotný a homogenní obraz pacientské populace. Naopak potvrzuje přítomnost přirozené variability mezi pacienty, která se projevuje rozdílným rozsahem hodnot, různou mírou asymetrie i odlišnou četností extrémnějších pozorování. Z metodického hlediska jde o důležité zjištění, protože podporuje využití analytických přístupů, které dokážou pracovat s heterogenní strukturou dat a nejsou omezeny pouze na jednoduché lineární předpoklady. Z věcného hlediska pak výsledky potvrzují, že datový soubor obsahuje reálně využitelný signál pro další analýzu a následné modelování rizika srdečního onemocnění.

2.2.3 Jednorozměrná analýza kategoriálních a binárních proměnných

Jednorozměrná analýza kategoriálních a binárních proměnných byla zaměřena na rozložení jednotlivých klinických kategorií v analyzovaném souboru. Pozornost byla věnována zejména četnosti výskytu jednotlivých hodnot, jejich relativnímu zastoupení a případné nerovnoměrnosti distribuce. Smyslem nadcházející části nebylo pouze formálně popsat zastoupené kategorie, ale především získat základní přehled o struktuře pacientské populace z hlediska klinicky významných znaků a posoudit, které charakteristiky jsou v souboru dominantní a které se vyskytují spíše okrajově.

V následujících obrázcích a tabulce jsou zobrazeny výsledky jednorozměrné analýzy kategoriálních a binárních proměnných. Grafické výstupy zachycují četnost jednotlivých kategorií, zatímco souhrnná tabulka poskytuje přehled o jejich absolutním a relativním zastoupení v celém souboru.

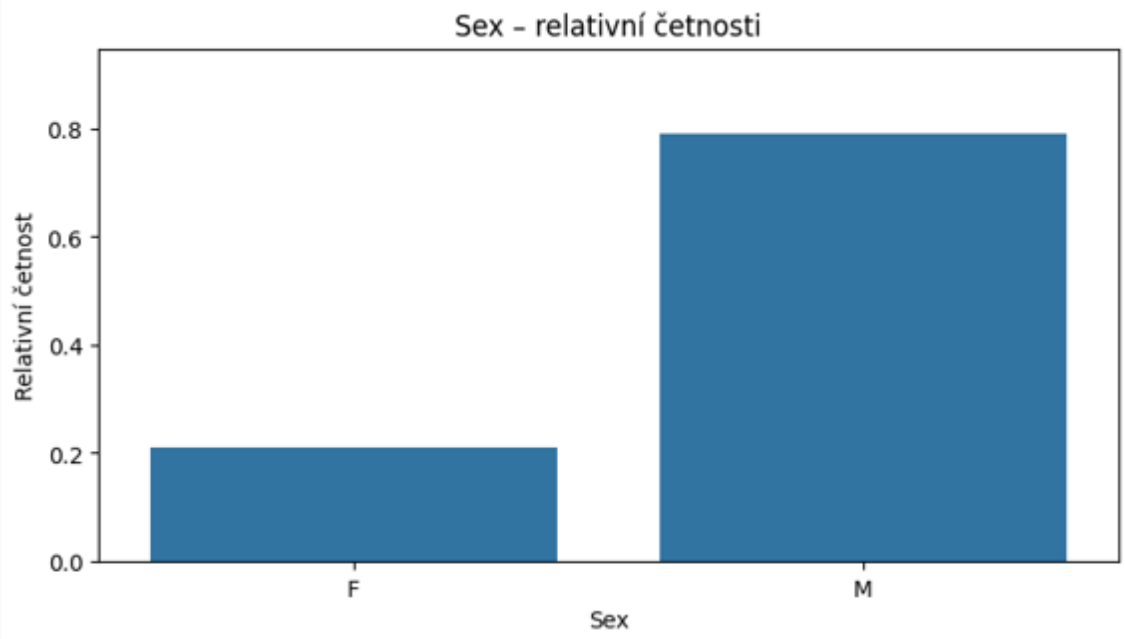
Tab. 4: Četnost kategoriálních a binárních proměnných

Proměnná	Počet kategorií	Nejčetnější kategorie	Relativní četnost nejčetnější kategorie	Nejméně četná kategorie	Relativní četnost nejméně četné kategorie	Chi2 statistika	p-hodnota
Sex	2	M	0.7898	F	0.2102	308.305	0.0000
ChestPainType	4	ASY	0.5403	TA	0.0501	473.154	0.0000
RestingECG	3	Normal	0.6013	ST	0.1939	296.810	0.0000
ExerciseAngina	2	N	0.5959	Y	0.4041	33.742	0.0000
ST_Slope	3	Flat	0.5011	Down	0.0686	296.359	0.0000
FastingBS	2	0	0.7669	1	0.2331	261.546	0.0000
HeartDisease	2	1	0.5534	0	0.4466	10.461	0.0012

Zdroj: Vlastní zpracování (2026)

Rozložení kategoriálních proměnných ukázalo, že pacientská populace není z hlediska klinických znaků zastoupena rovnoměrně. U části atributů převažuje jedna nebo několik dominantních kategorií, zatímco jiné hodnoty se objevují méně často. Získaný výsledek je z analytického hlediska důležitý, protože naznačuje, že některé proměnné budou v dalších krocích pravděpodobně nést vyšší informační hodnotu než jiné. Nerovnoměrné zastoupení kategorií přitom nemusí být problémem, pokud odpovídá charakteru sledované populace a klinickému kontextu.

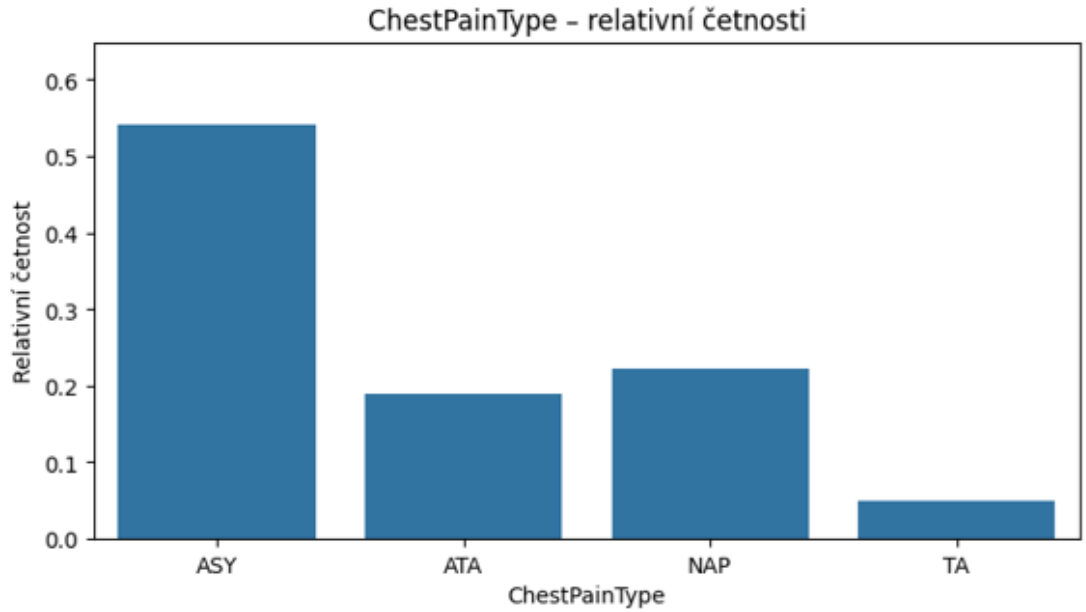
Z pohledu interpretace je důležité, že kategoriální a binární proměnné nepůsobí jako nahodile rozdělené atributy, ale vytvářejí smysluplný obraz analyzované skupiny pacientů. Četnosti jednotlivých kategorií poskytují první orientaci v četněji zastoupených typech klinických profilů a naznačují, které charakteristiky mohou být z hlediska dalšího modelování obzvlášť významné. Již nyní je patrné, že datový soubor obsahuje užitečnou strukturální informaci, kterou bude možné dále využít při analýze vztahů k cílové proměnné i při následném klasifikačním modelování.



Obr. 11: Relativní četnost proměnné Pohlaví

Zdroj: Vlastní zpracování (2026)

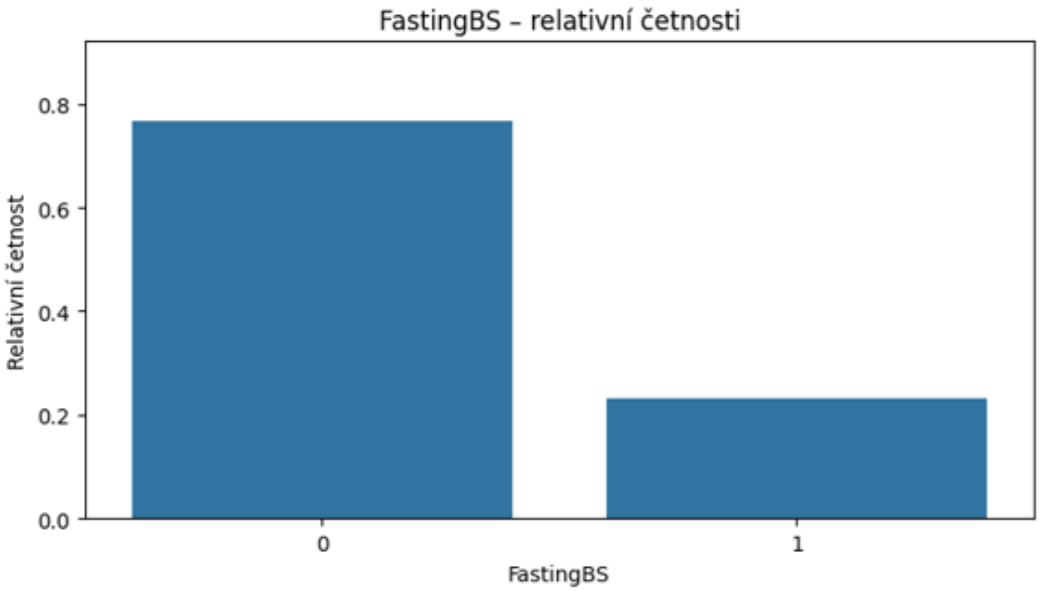
Proměnná Pohlaví ukazuje zastoupení pohlaví v analyzovaném souboru. Nerovnoměrné rozdělení proměnné je důležité nejen z hlediska popisu dat, ale i z hlediska další interpretace výsledků, protože pohlaví bývá v kardiologii spojováno s rozdílným rizikovým profilem i odlišným klinickým průběhem onemocnění. Výsledek proto naznačuje, že proměnná může v dalších částech analýzy nést významnou doplňující informaci.



Obr. 12: Relativní četnost proměnné ChestPainType

Zdroj: Vlastní zpracování (2026)

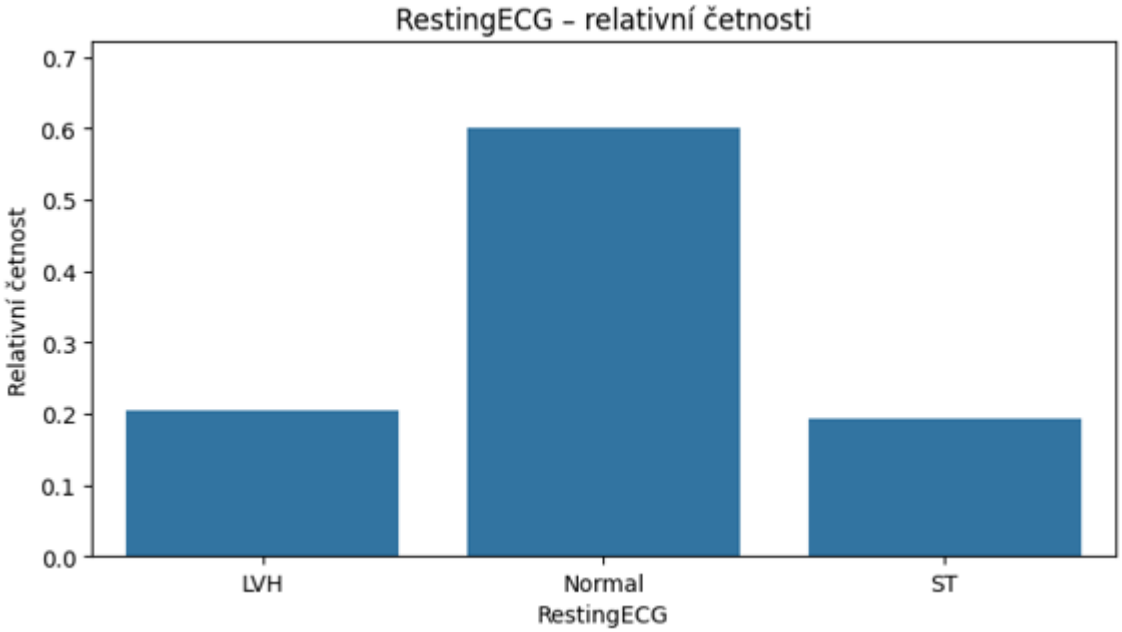
Proměnná ChestPainType zachycuje různé typy bolesti na hrudi, které představují klinicky velmi významný diagnostický ukazatel. Rozložení kategorií ukazuje, které typy obtíží se v souboru vyskytují častěji a které jsou naopak méně zastoupené. Zjištěné rozdělení je důležité, protože různé formy bolesti na hrudi mohou souviset s odlišnou pravděpodobností srdečního onemocnění a později se mohou výrazně projevit i v modelování.



Obr. 13: Relativní četnost proměnné FastingBS

Zdroj: Vlastní zpracování (2026)

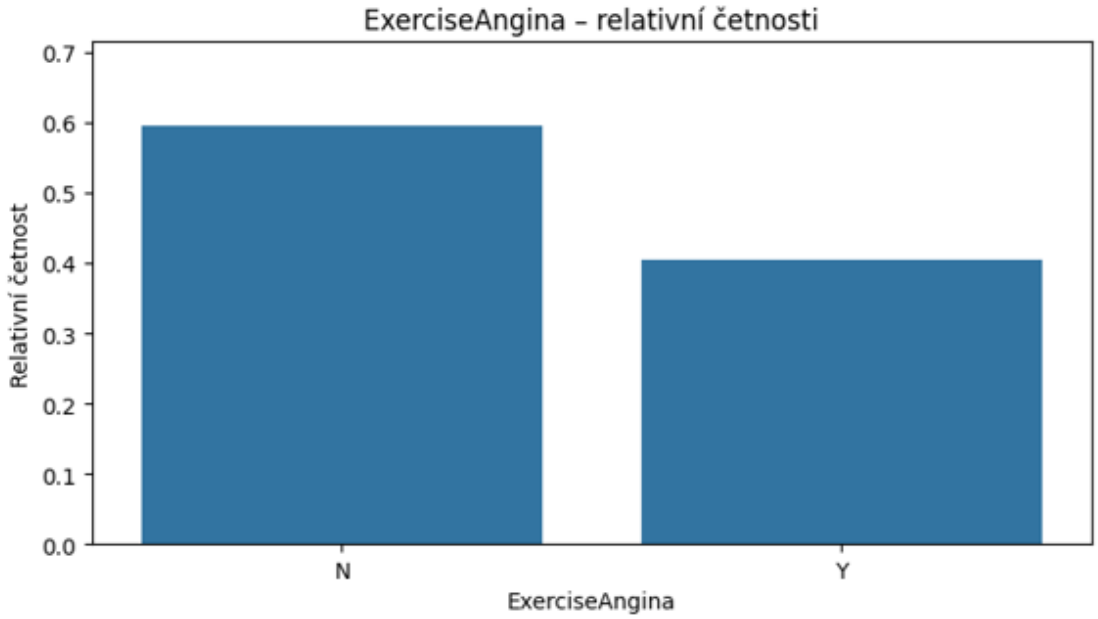
Proměnná FastingBS představuje binární informaci o zvýšené hladině glukózy nalačno. Přestože jde o jednoduchý atribut s pouze dvěma kategoriemi, jeho rozložení je klinicky relevantní, protože zvýšená glykemie patří mezi důležité metabolické faktory spojené s kardiovaskulárním rizikem. Četnost jednotlivých hodnot proto napovídá, jak výrazně je v analyzované populaci zastoupen metabolický rizikový profil.



Obr. 14: Relativní četnost proměnné RestingECG

Zdroj: Vlastní zpracování (2026)

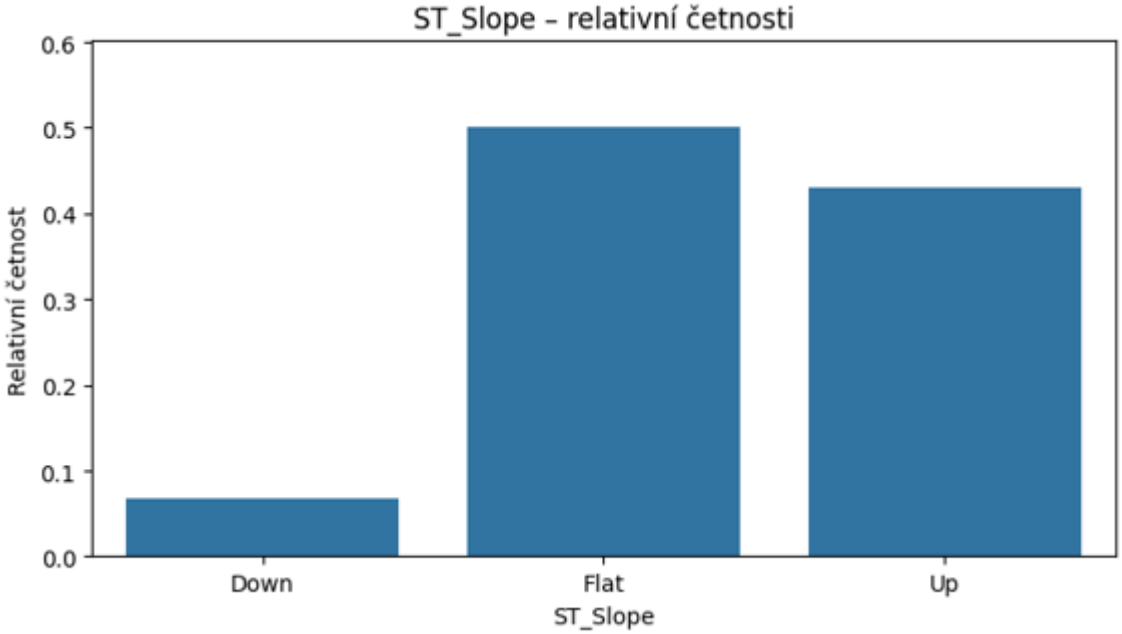
Proměnná RestingECG popisuje výsledek klidového EKG a přináší informaci o základním elektrickém projevu srdeční činnosti. Rozdělení kategorií poskytuje přehled o četnosti normálních a odchylných nálezů v analyzovaném souboru. Získané informace jsou důležité nejen pro popis souboru, ale i pro další interpretaci, protože odchylky na EKG mohou souviset s přítomností srdeční patologie.



Obr. 15: Relativní četnost proměnné RestingECG

Zdroj: Vlastní zpracování (2026)

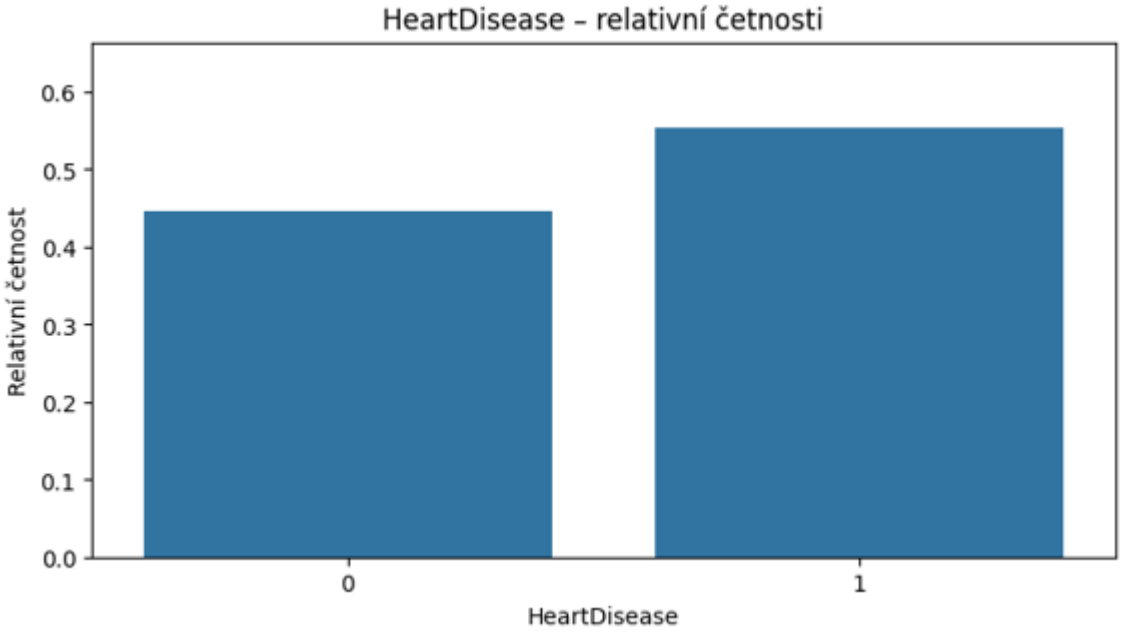
Proměnná ExerciseAngina zachycuje přítomnost nebo nepřítomnost anginy při zátěži. Z klinického hlediska jde o velmi významný symptom, protože potíže vyvolané fyzickou zátěží často souvisejí s omezeným prokrvením srdeční svaloviny. Rozložení kategorií proto představuje důležitý orientační údaj o zastoupení symptomu v analyzované populaci.



Obr. 16: Relativní četnost proměnné ST_Slope

Zdroj: Vlastní zpracování (2026)

Proměnná ST_Slope popisuje sklon ST segmentu a patří mezi klíčové diagnostické ukazatele v kardiologickém kontextu. Četnost jednotlivých kategorií ukazuje, které typy nálezu se v souboru vyskytují častěji. Vzhledem k významu dané proměnné lze očekávat, že její rozložení bude důležité i pro následnou analýzu vztahů s cílovou proměnnou a pro klasifikační modelování.



Obr. 17: Relativní četnost proměnné HeartDisease

Zdroj: Vlastní zpracování (2026)

Proměnná HeartDisease představuje cílový znak a její rozdělení je důležité pro celkové pochopení charakteru úlohy. Poměr mezi pozitivními a negativními případy ukazuje, zda je klasifikační problém vyvážený, nebo zda bude při modelování nutné věnovat zvýšenou pozornost případné nerovnováze tříd. V analyzovaném souboru informace současně poskytuje základní přehled o tom, jak výrazně je srdeční onemocnění ve sledované populaci zastoupeno.

Celkově jednorozměrná analýza kategoriálních a binárních proměnných ukázala, že datový soubor zachycuje klinicky smysluplně členěnou populaci pacientů. Nerovnoměrné zastoupení kategorií přitom nelze chápat pouze jako statistickou vlastnost dat, ale i jako odraz reálné struktury sledovaného souboru. Probíraná část poskytuje důležitý základ pro další kroky, zejména pro posouzení vztahů mezi kategoriálními proměnnými a cílovým znakem a pro interpretaci významnosti jednotlivých atributů v následném modelování.

2.2.4 Dvourozměrná analýza numerických proměnných

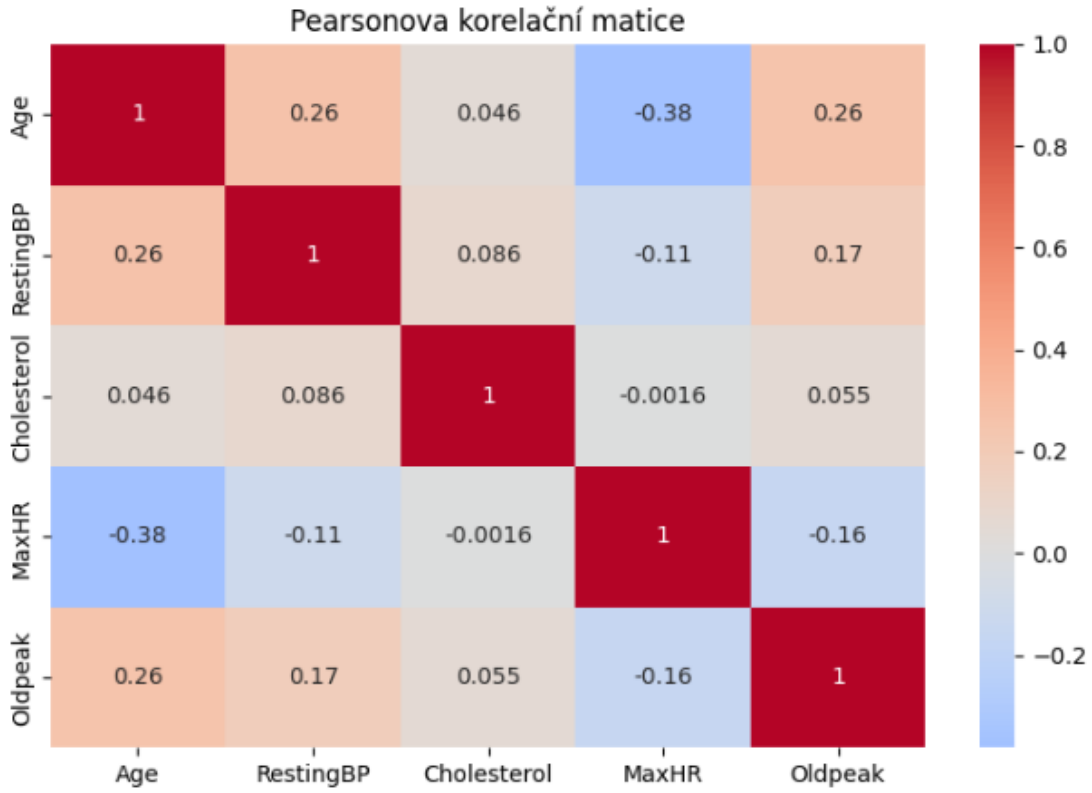
Dvourozměrná analýza numerických proměnných byla zaměřena na posouzení vzájemných vztahů mezi číselnými ukazateli zahrnutými v datovém souboru. Pozornost směřovala zejména na sílu a směr lineárních i monotónních vazeb, na možnost existence redundantních informací a na celkovou strukturu numerické části dat. Cílem nebylo pouze zjistit, zda mezi proměnnými existuje statistická souvislost, ale posoudit, zda zjištěné vazby dávají smysl z hlediska medicínské interpretace a zda mohou ovlivnit pozdější modelování.

V následujících tabulkách a grafech jsou zobrazeny výsledky dvourozměrné analýzy numerických proměnných, zejména korelační koeficienty, heatmapa korelačních vztahů a vícerozměrný pairplot.

Tab. 5: Četnost kategoriálních a binárních proměnných

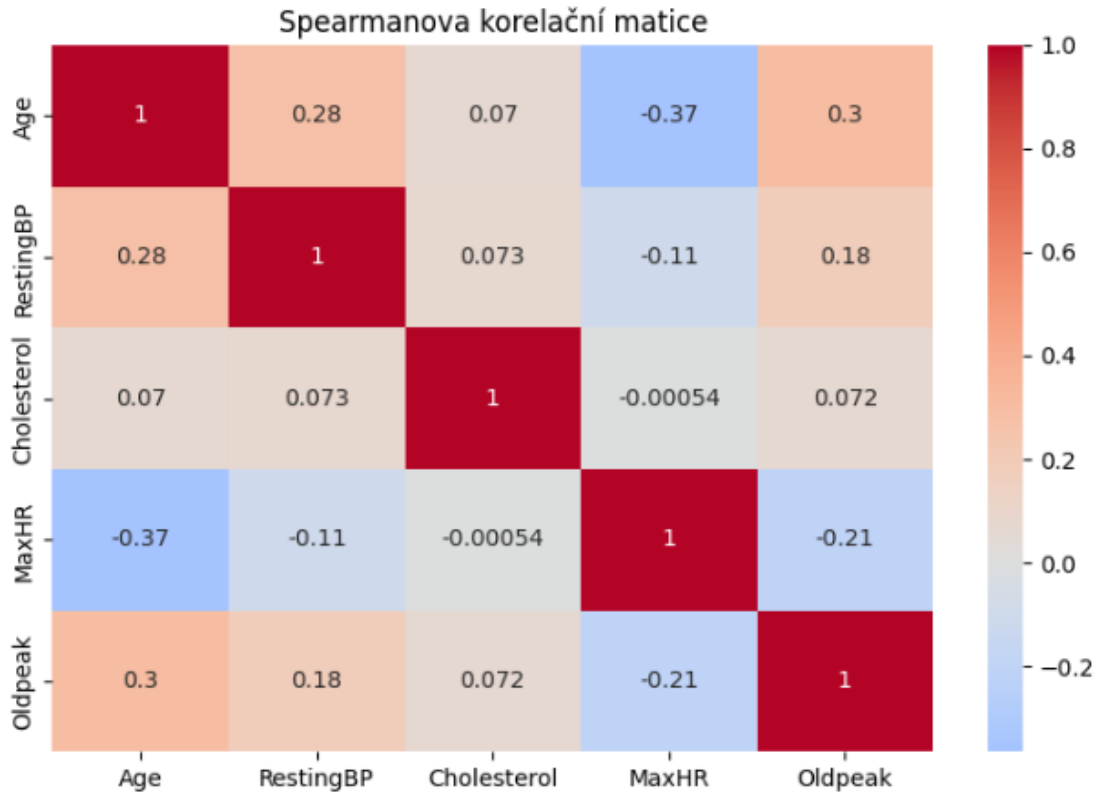
Proměnná 1	Proměnná 2	Pearson r	Pearson p-hodnota	Spearman ρ	Spearman p-hodnota
Age	MaxHR	-0.3820	0.0000	-0.3650	0.0000
Age	Oldpeak	0.2586	0.0000	0.2983	0.0000
Age	RestingBP	0.2631	0.0000	0.2801	0.0000
MaxHR	Oldpeak	-0.1607	0.0000	-0.2051	0.0000
RestingBP	Oldpeak	0.1741	0.0000	0.1771	0.0000
RestingBP	MaxHR	-0.1098	0.0009	-0.1062	0.0013
RestingBP	Cholesterol	0.0861	0.0091	0.0734	0.0262
Cholesterol	Oldpeak	0.0545	0.0989	0.0717	0.0298
Age	Cholesterol	0.0455	0.1681	0.0703	0.0331
Cholesterol	MaxHR	-0.0016	0.9614	-0.0005	0.9869

Zdroj: Vlastní zpracování (2026)



Obr. 18: Heatmapa Pearsonovy korelační matice

Zdroj: Vlastní zpracování (2026)

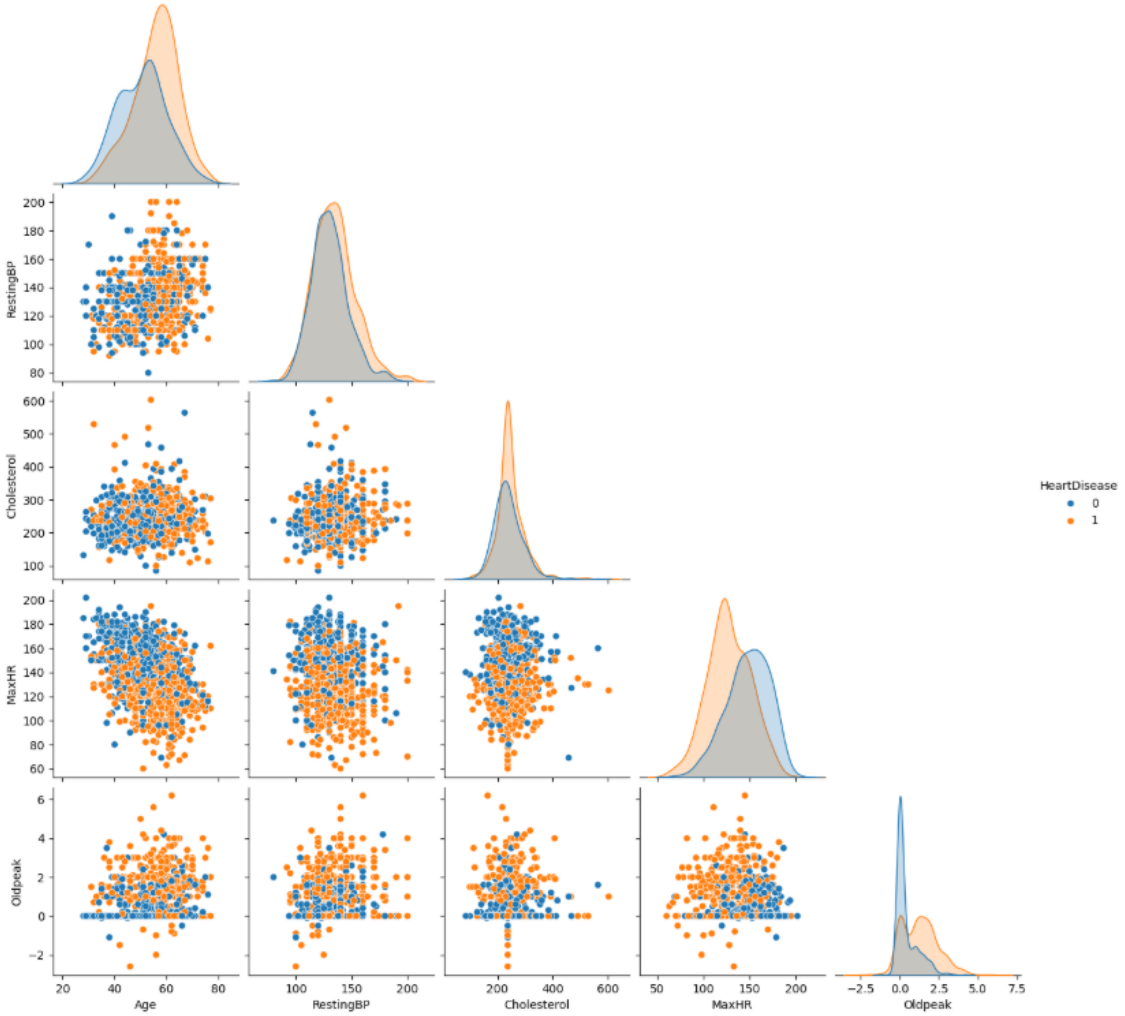


Obr. 19: Heatmapa Spearmanovy korelační matice

Zdroj: Vlastní zpracování (2026)

Korelační analýza ukázala převážně slabší až středně silné vztahy mezi sledovanými numerickými proměnnými. Výsledek naznačuje, že numerická část dat neobsahuje vysokou míru redundance, což znamená, že více proměnných pravděpodobně nezachycuje stejný jev pouze v odlišné podobě. Z metodického hlediska jde o příznivý závěr, protože jednotlivé ukazatele pravděpodobně přinášejí doplňující se informace o zdravotním stavu pacientů a není nutné je chápat jako pouhé varianty téhož měření.

Heatmapa korelačních vztahů poskytuje souhrnný pohled na strukturu numerické části dat a umožňuje rychle rozpoznat dvojice proměnných s výraznější vazbou. Význam zjištěných vztahů spočívá i ve skutečnosti, že většina vazeb nepůsobí izolovaně ani natolik silně, aby bylo možné zdravotní profil pacienta uspokojivě vysvětlit prostřednictvím jediné číselné proměnné. Přítomnost spíše slabších a středně silných souvislostí odpovídá charakteru klinických dat, v nichž zdravotní stav zpravidla vzniká jako výsledek souběhu více faktorů. Informační hodnota datasetu proto pravděpodobně nevychází z jednotlivých proměnných posuzovaných odděleně, ale z jejich kombinace a vzájemného působení.



Obr. 20: Pairplot numerických proměnných

Zdroj: Vlastní zpracování (2026)

Pairplot současně doplňuje korelační koeficienty o vizuální pohled na tvar vztahů, rozptyl hodnot a případné nelinearity. Spojení numerické tabulky s grafickou reprezentací je důležité, protože samotná hodnota korelace nemusí zachytit všechny aspekty vztahu mezi proměnnými. Vizuální kontrola proto představuje vhodné doplnění čistě statistického vyhodnocení.

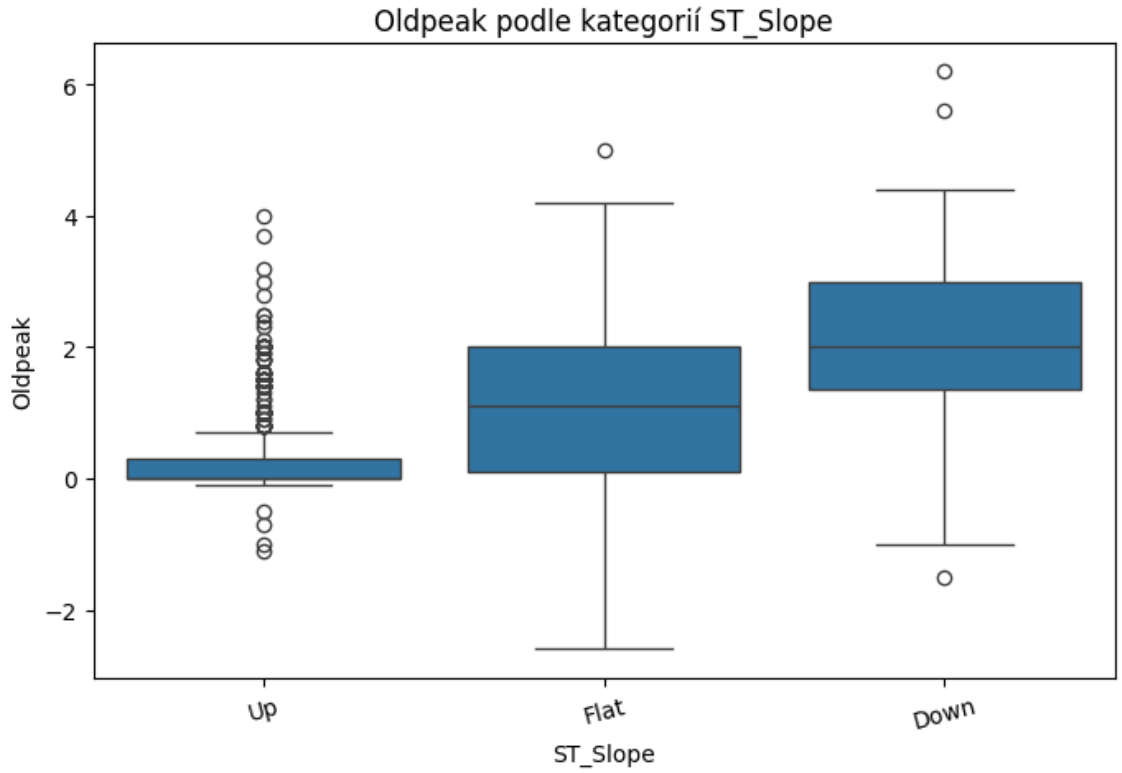
Celkové výsledky dvourozměrné analýzy numerických proměnných podporují předpoklad, že následné modelování bude úspěšnější při současném využití více atributů než při spoléhání na jednotlivé izolované ukazatele. Zároveň potvrzují, že numerické proměnné zachycují různé dimenze zdravotního stavu pacientů a že jejich společná interpretace má větší hodnotu než prosté posuzování každé proměnné samostatně.

2.2.5 Vztahy mezi kategoriálními a numerickými proměnnými

Další krok explorativní analýzy byl zaměřen na propojení kategoriálních klinických znaků s číselně vyjádřenými fyziologickými a diagnostickými ukazateli. Cílem bylo posoudit, zda vybrané kategorie souvisejí se systematicky odlišnými hodnotami numerických proměnných a zda zjištěné rozdíly nesou klinicky interpretovatelný význam. Pozornost byla soustředěna na trojici

vztahů s nejvyššími hodnotami korelačního poměru, které představují nejvýraznější vazby mezi kategoriálními a numerickými proměnnými v analyzovaném souboru.

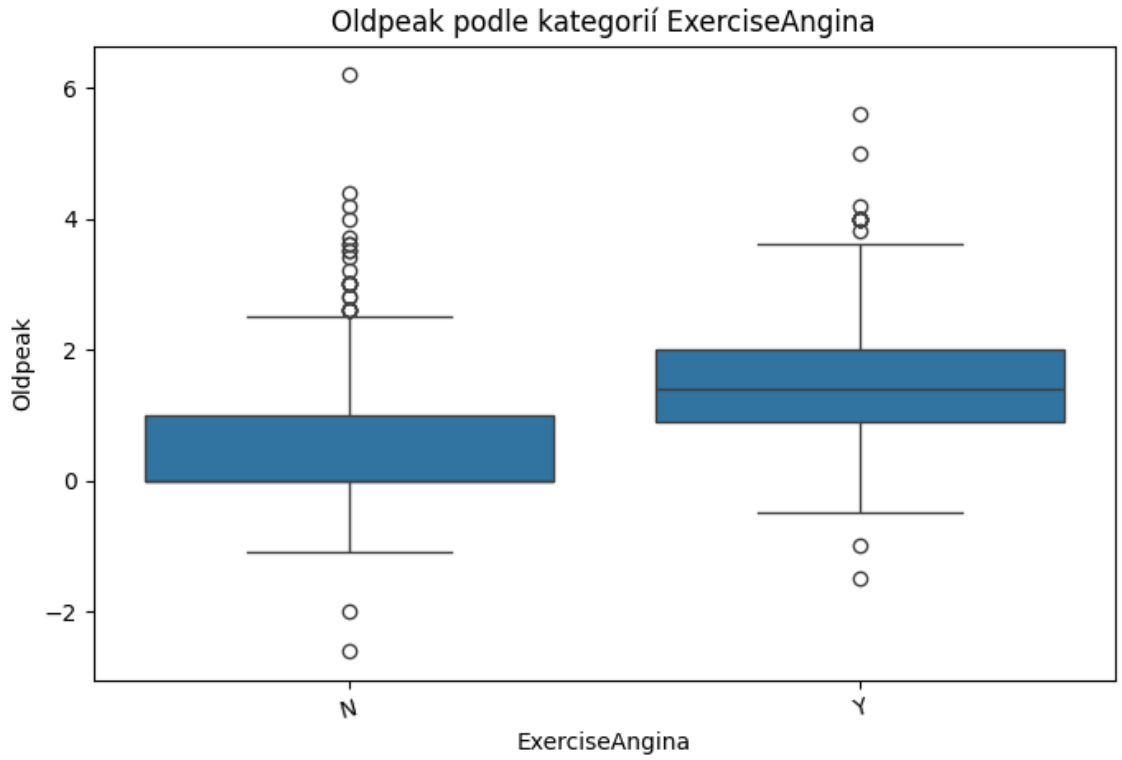
V následujících grafech jsou zobrazeny tři vybrané vztahy mezi kategoriálními a numerickými proměnnými, které vykázaly nejvyšší korelační hodnoty.



Obr. 21: Boxplot proměnné Oldpeak dle ST_Slope

Zdroj: Vlastní zpracování (2026)

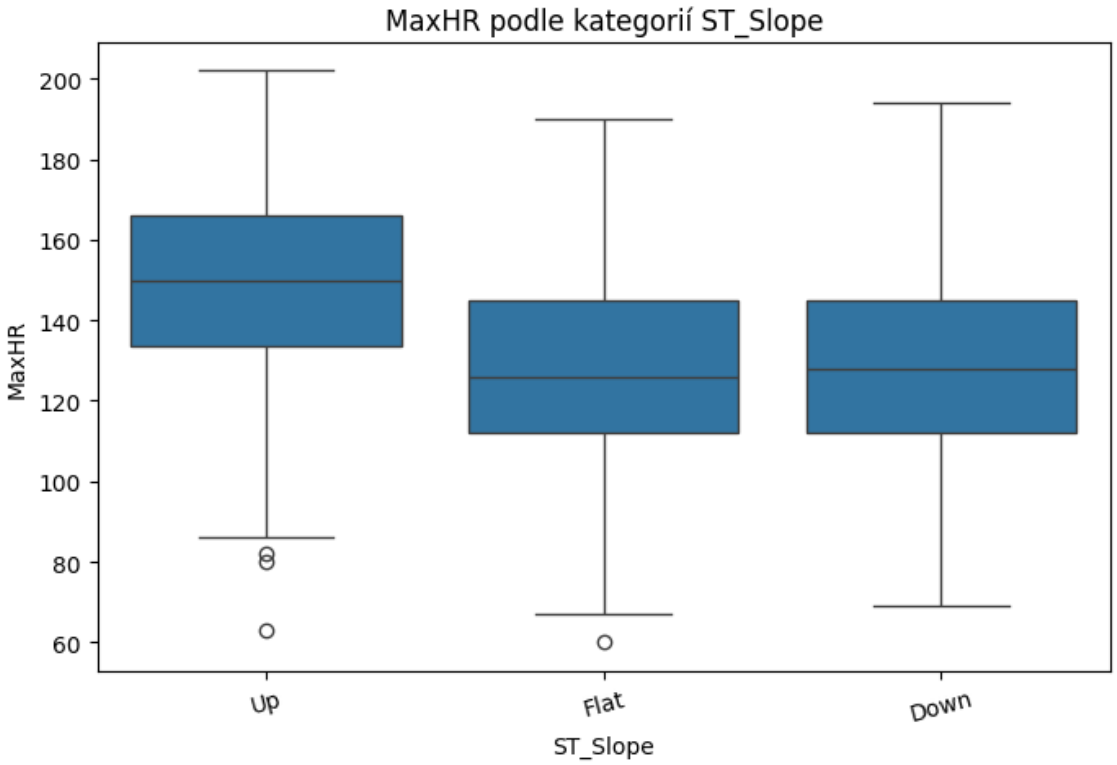
Spojení proměnných Oldpeak a ST_Slope představuje nejvýraznější vazbu v rámci celé části analýzy. Výsledek ukazuje, že jednotlivé kategorie sklonu ST segmentu souvisejí s odlišným rozložením hodnot zátěžového ukazatele Oldpeak. Z klinického hlediska jde o dobře interpretovatelný vztah, protože oba znaky souvisejí s reakcí srdce na zátěž a s možnými ischemickými změnami. Výraznější rozdíly mezi kategoriemi podporují předpoklad, že proměnná ST_Slope nese významnou diagnostickou informaci a že její role nebude omezena pouze na deskriptivní úroveň, ale projeví se i v navazujícím modelování.



Obr. 22: Boxplot proměnné Oldpeak dle ExerciseAngina

Zdroj: Vlastní zpracování (2026)

Výrazná vazba byla zjištěna i mezi proměnnými Oldpeak a ExerciseAngina. Rozdělení hodnot ukazuje, že přítomnost anginy při zátěži souvisí s odlišnou úrovní zátěžových změn zachycených proměnnou Oldpeak. Výsledek je odborně logický, protože propojuje subjektivně vnímaný symptom s číselně vyjádřeným fyziologickým projevem srdeční zátěže. Analytický význam spočívá především ve skutečnosti, že proměnná ExerciseAngina nepůsobí jako izolovaný binární údaj, ale jako znak, který ve spojení s Oldpeak pomáhá lépe rozlišit odlišné klinické profily pacientů.



Obr. 23: Boxplot proměnné MaxHR dle ST_Slope

Zdroj: Vlastní zpracování (2026)

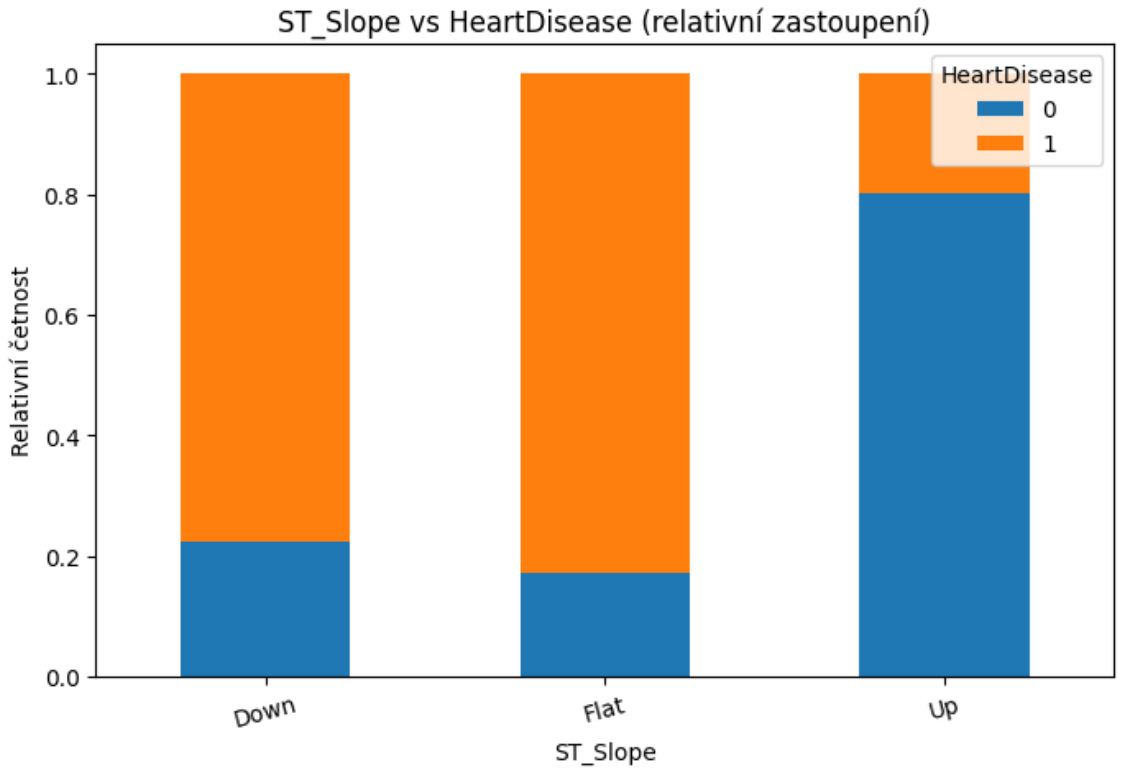
Silnější souvislost byla pozorována rovněž mezi proměnnými MaxHR a ST_Slope. Výsledek naznačuje, že jednotlivé kategorie sklonu ST segmentu souvisejí s rozdílnou maximální dosaženou srdeční frekvencí. Z interpretačního hlediska jde o významné propojení elektrokardiografického nálezu se zátěžovou výkonností organismu. Nejde pouze o statistickou souvislost, ale o kombinaci znaků, které společně popisují funkční stav kardiovaskulárního systému. Právě propojení kvalitativního a kvantitativního ukazatele zvyšuje hodnotu datového souboru pro další analytické kroky.

Vybrané vztahy ukazují, že část kategoriálních proměnných získává plný význam až ve spojení s numerickými ukazateli. Samostatné posouzení jednotlivých atributů by nedokázalo zachytit všechny souvislosti, které se objevují při společné interpretaci. Výsledky proto potvrzují, že smíšená povaha dat představuje v analyzovaném souboru důležitou výhodu, protože umožňuje zachytit jemnější rozdíly mezi pacientskými profily a vytváří vhodný základ pro vícerozměrné zpracování i následné modelování.

2.2.6 Vztahy mezi kategoriálními proměnnými a cílovou proměnnou

Další část explorativní analýzy byla zaměřena na vztahy mezi kategoriálními proměnnými a cílovou proměnnou HeartDisease. Cílem bylo posoudit, které klinické kategorie vykazují nejsilnější asociaci s přítomností srdečního onemocnění a které znaky proto mohou mít zvláštní význam pro následnou interpretaci i modelování. Pro vyhodnocení síly vztahu byl použit koeficient Cramérovo V, vhodný pro posouzení asociace mezi kategoriálními proměnnými.

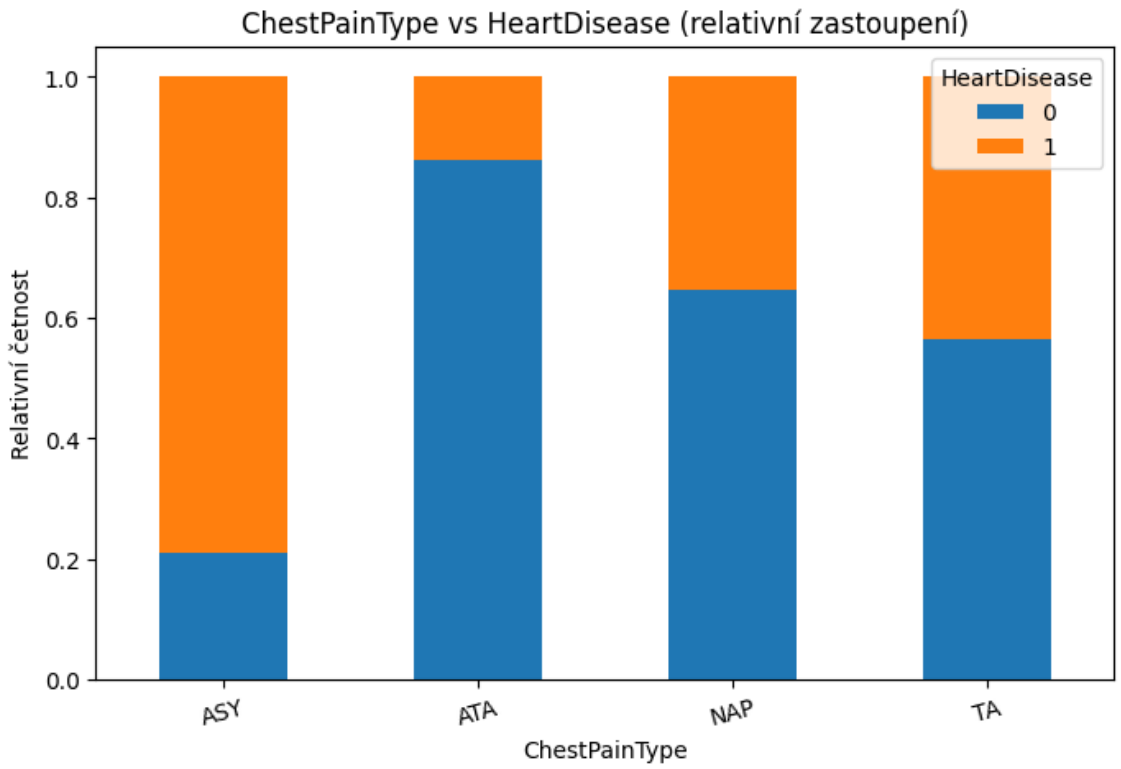
V následujících grafech jsou zobrazeny tři vztahy mezi kategoriálními proměnnými a cílovou proměnnou, které vykázaly nejvyšší hodnoty Cramérova V.



Obr. 24: Relativní zastoupení proměnné ST_Slope vůči HeartDisease

Zdroj: Vlastní zpracování (2026)

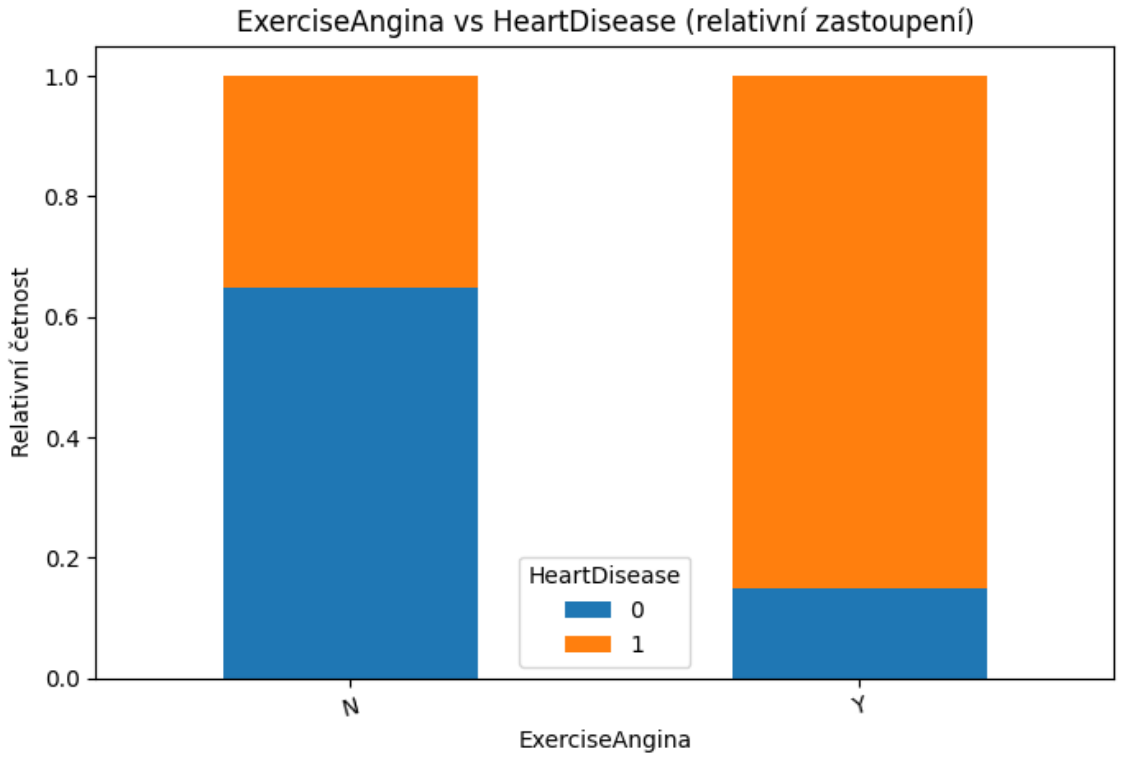
Nejsilnější kategoriální asociace byla zjištěna mezi proměnnými ST_Slope a HeartDisease. Rozložení tříd ukazuje, že jednotlivé kategorie sklonu ST segmentu nejsou ve vztahu k cílové proměnné zastoupeny rovnoměrně a že některé z nich se pojí s výrazně vyšším podílem pozitivních případů. Z klinického hlediska jde o vysoce významný výsledek, protože změny ST segmentu patří mezi důležité ukazatele srdeční ischemie a dalších poruch srdeční činnosti. Silná asociace proto podporuje předpoklad, že proměnná ST_Slope bude hrát zásadní roli i v navazujícím klasifikačním modelování.



Obr. 25: Relativní zastoupení proměnné ChestPainType vůči HeartDisease

Zdroj: Vlastní zpracování (2026)

Výrazná vazba byla zjištěna rovněž mezi proměnnými ChestPainType a HeartDisease. Zastoupení pozitivních a negativních případů se mezi jednotlivými typy bolesti na hrudi zřetelně liší, což naznačuje, že různé formy bolesti nemají stejnou diagnostickou závažnost. Výsledek odpovídá i medicínské logice, protože charakter bolesti na hrudi představuje jeden z klíčových symptomů při hodnocení kardiovaskulárního rizika. Analytická hodnota proměnné spočívá ve schopnosti odlišovat pacienty s rozdílnou pravděpodobností onemocnění již na úrovni základní klinické charakteristiky.



Obr. 26: Relativní zastoupení proměnné ExerciseAngina vůči HeartDisease

Zdroj: Vlastní zpracování (2026)

Silná asociace byla pozorována i mezi proměnnými ExerciseAngina a HeartDisease. Přítomnost anginy při zátěži je spojena s vyšším zastoupením pacientů se srdečním onemocněním, zatímco absence tohoto symptomu se častěji pojí s negativní třídou. Výsledek je odborně dobře obhajitelný, protože zátěžová angina představuje klinicky významný projev omezeného prokrvení srdeční svaloviny. V kontextu datové analýzy jde o důležitý znak, který propojuje symptomatický popis pacienta s cílovou proměnnou a zvyšuje interpretační sílu celého modelového rámce.

Výsledky kategoriální analýzy vůči cílové proměnné potvrzují, že nejvýznamnější asociace se soustřeďují do proměnných úzce spojených se zátěžovým vyšetřením, změnami ST segmentu a charakterem bolesti na hrudi. Dané zjištění má zásadní význam pro další analytické kroky, protože ukazuje, že následné modelování nebude stavět na nahodilých vztazích, ale na atributech s jasným klinickým obsahem. Silnější asociace zároveň podporují předpoklad, že právě uvedené proměnné budou patřit mezi nejdůležitější prediktory i v rámci supervised EDA a finálních klasifikačních modelů.

2.2.7 Vícerozměrná explorativní analýza bez učitele

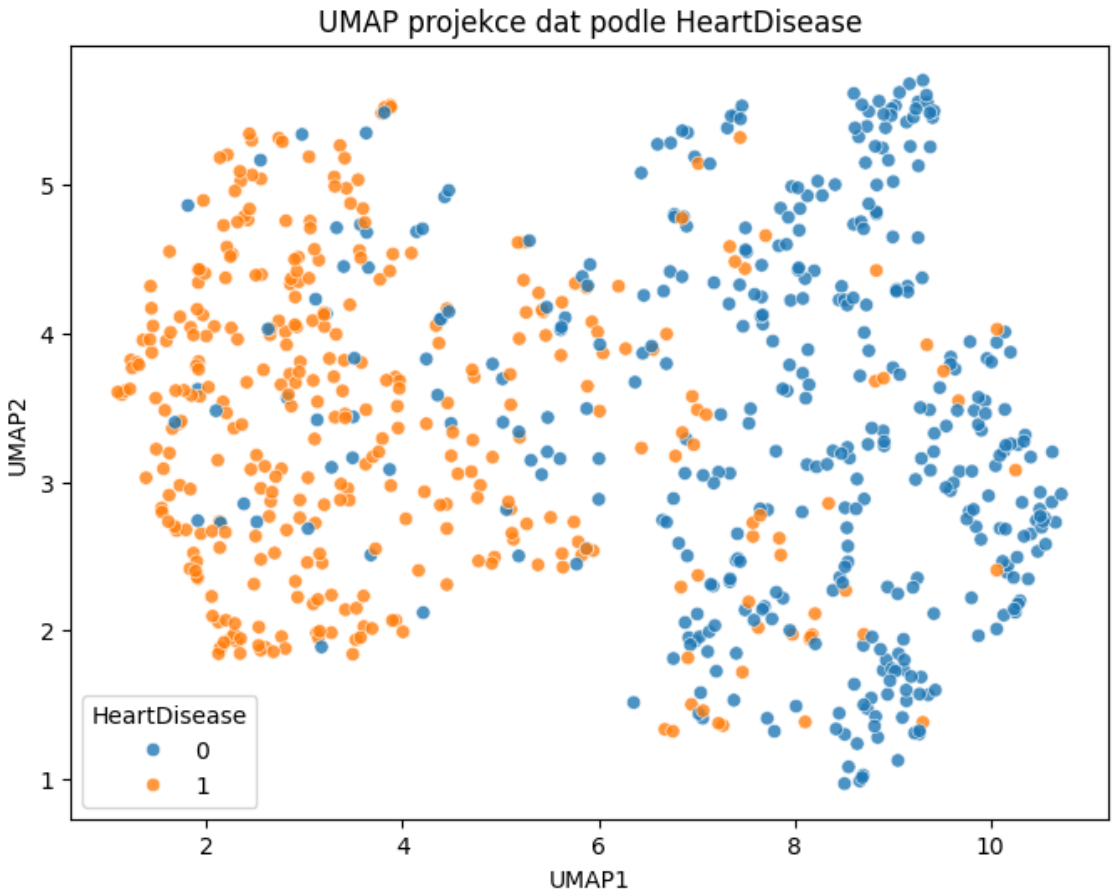
Vícerozměrná explorativní analýza bez učitele byla zařazena s cílem ověřit, zda se v datech objevují přirozené skupiny pacientů s podobnými charakteristikami a zda lze v prostoru více proměnných rozpoznat vnitřní strukturu, která nemusí být zřejmá při jednorozměrné nebo dvourozměrné analýze. Na rozdíl od předchozích kroků zde nešlo o přímé posouzení vztahu k cílové proměnné, ale o zkoumání organizace datového prostoru bez využití předem známého označení tříd.

Pozornost byla soustředěna na dvě hlavní části. Nejprve byla použita metoda UMAP pro projekci vícerozměrných dat do dvourozměrného prostoru. Následně bylo provedeno shlukování pomocí algoritmu K-Prototypes, který umožňuje pracovat se smíšeným typem dat obsahujícím numerické i kategoriální proměnné. Výsledky mají význam především pro porozumění vnitřní heterogenitě pacientské populace a pro identifikaci profilů, které mohou být důležité i z hlediska následného modelování.

2.2.8 UMAP projekce dat podle cílové proměnné

Prvním krokem vícerozměrné explorativní analýzy bez učitele byla projekce dat pomocí metody UMAP. Cílem bylo převést vícerozměrný prostor vstupních proměnných do zjednodušené dvourozměrné podoby a vizuálně posoudit míru podobnosti mezi pacienty. Přestože UMAP nevyužívá cílovou proměnnou při samotné konstrukci projekce, barevné odlišení tříd umožňuje následně sledovat, zda se pacienti s různým výskytem srdečního onemocnění v projekci alespoň částečně seskupují.

V následujícím obrázku je zobrazena UMAP projekce dat rozlišená podle cílové proměnné HeartDisease.



Obr. 27: UMAP projekce dat podle HeartDisease

Zdroj: Vlastní zpracování (2026)

UMAP projekce ukazuje, že pacienti netvoří zcela náhodně rozptýlený soubor bodů, ale že v datovém prostoru existuje určitá vnitřní struktura. Současně však nelze pozorovat dokonale ostré oddělení všech bodů podle cílové proměnné. Přítomnost částečně oddělených oblastí

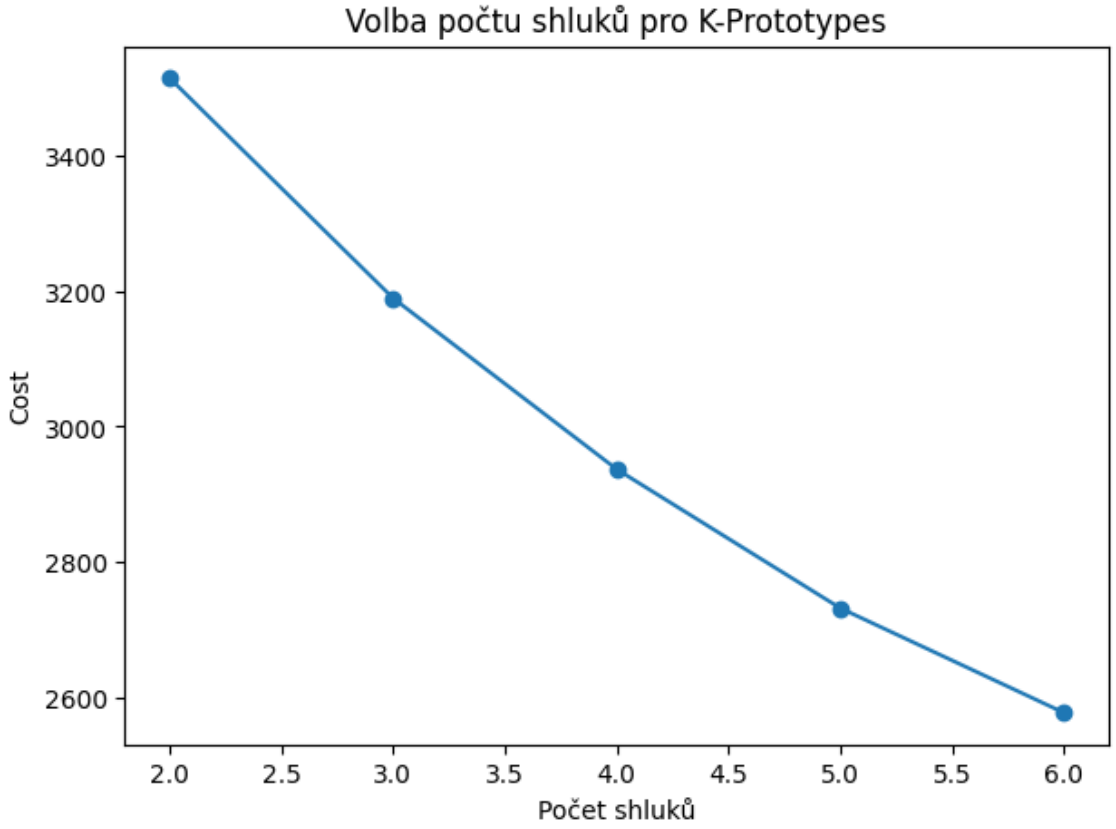
vedle překryvu mezi třídami odpovídá charakteru medicínských dat, v nichž se různé klinické profily často vzájemně prolínají.

Význam projekce spočívá především v potvrzení, že datový soubor obsahuje vícerozměrný signál, který nelze plně zachytit pouze jednoduchými párovými vztahy. Zřetelné oblasti vyšší koncentrace bodů naznačují existenci podobných pacientských profilů, zatímco překryv mezi třídami upozorňuje na složitost problému a na nemožnost jednoduchého oddělení pacientů pomocí několika málo znaků. UMAP projekce proto vytváří vhodný základ pro navazující shlukovou analýzu.

2.2.9 Volba počtu shluků

Před samotným finálním shlukováním bylo nutné určit vhodný počet shluků, který by umožnil zachytit vnitřní strukturu dat bez nadměrného zjednodušení i bez umělého vytváření příliš fragmentovaných skupin. Smyslem dané části bylo nalézt kompromis mezi interpretovatelností výsledků a schopností modelu rozlišit odlišné pacientské profily. Volba počtu shluků byla provedena na základě porovnání hodnot nákladové funkce a s ohledem na následnou klinickou čitelnost výsledných segmentů.

V následujícím grafu je zobrazen průběh hodnot vybraného kritéria pro různé počty shluků.



Obr. 28: Volba počtu shluků pro K-Prototypes

Zdroj: Vlastní zpracování (2026)

Průběh hodnot ukazuje, že s rostoucím počtem shluků dochází ke zlepšování vnitřní homogenity skupin, avšak pouze do určité míry. Po překročení vhodného počtu segmentů se přínos dalšího dělení snižuje a jednotlivé skupiny začínají být méně přehledné z hlediska následné interpretace.

Výběr finálního počtu shluků proto nebyl založen pouze na technickém optimu, ale i na snaze zachovat smysluplný a odborně uchopitelný popis pacientských profilů.

Z metodického hlediska je volba počtu shluků důležitá, protože výrazně ovlivňuje následnou podobu celé segmentace. Příliš nízký počet skupin by vedl ke ztrátě části variability mezi pacienty, zatímco příliš vysoký počet by mohl vytvořit segmenty obtížně interpretovatelné a málo stabilní. Zvolený počet shluků proto představuje analyticky i věcně obhajitelný kompromis mezi přesností a srozumitelností.

2.2.10 Interpretace profilů shluků

Samotné rozdělení dat do shluků neumožňuje bez další analýzy pochopit, jaké typy pacientů jednotlivé segmenty představují. Ze zmiňovaného důvodu bylo finální shlukování doplněno profilováním, které porovnává numerické i kategoriální charakteristiky zvlášť pro každý shluk. Cílem bylo převést výsledek algoritmu do podoby, která je věcně interpretovatelná a použitelná pro popis odlišných pacientských profilů.

V následujících tabulkách jsou uvedeny profily jednotlivých shluků, zejména průměry numerických proměnných, modální hodnoty kategoriálních proměnných a zastoupení cílové proměnné.

Tab. 6: Průměry numerických proměnných v jednotlivých shlucích

Cluster	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
0	58.89	148.80	277.72	129.20	1.90
1	46.79	128.15	241.50	158.56	0.29
2	56.72	125.94	218.86	122.12	0.93

Zdroj: Vlastní zpracování (2026)

Tab. 7: Typické (modální) kategorie v jednotlivých shlucích

Cluster	Sex	ChestPainType	RestingECG	ExerciseAngina	ST_Slope	FastingBS
0	M	ASY	Normal	Y	Flat	0
1	M	ATA	Normal	N	Up	0
2	M	ASY	Normal	Y	Flat	0

Zdroj: Vlastní zpracování (2026)

Tab. 8: Typické (modální) kategorie v jednotlivých shlucích

HeartDisease	0	1
Cluster		
0	0.206	0.794
1	0.813	0.187
2	0.367	0.633

Zdroj: Vlastní zpracování (2026)

Profilování ukazuje, že jednotlivé shluky nepředstavují nahodilé skupiny, ale soubory pacientů s odlišnou kombinací klinických znaků. Rozdíly se projevují u numerických ukazatelů, například věku, zátěžových parametrů nebo srdeční frekvence i u kategoriálních atributů souvisejících

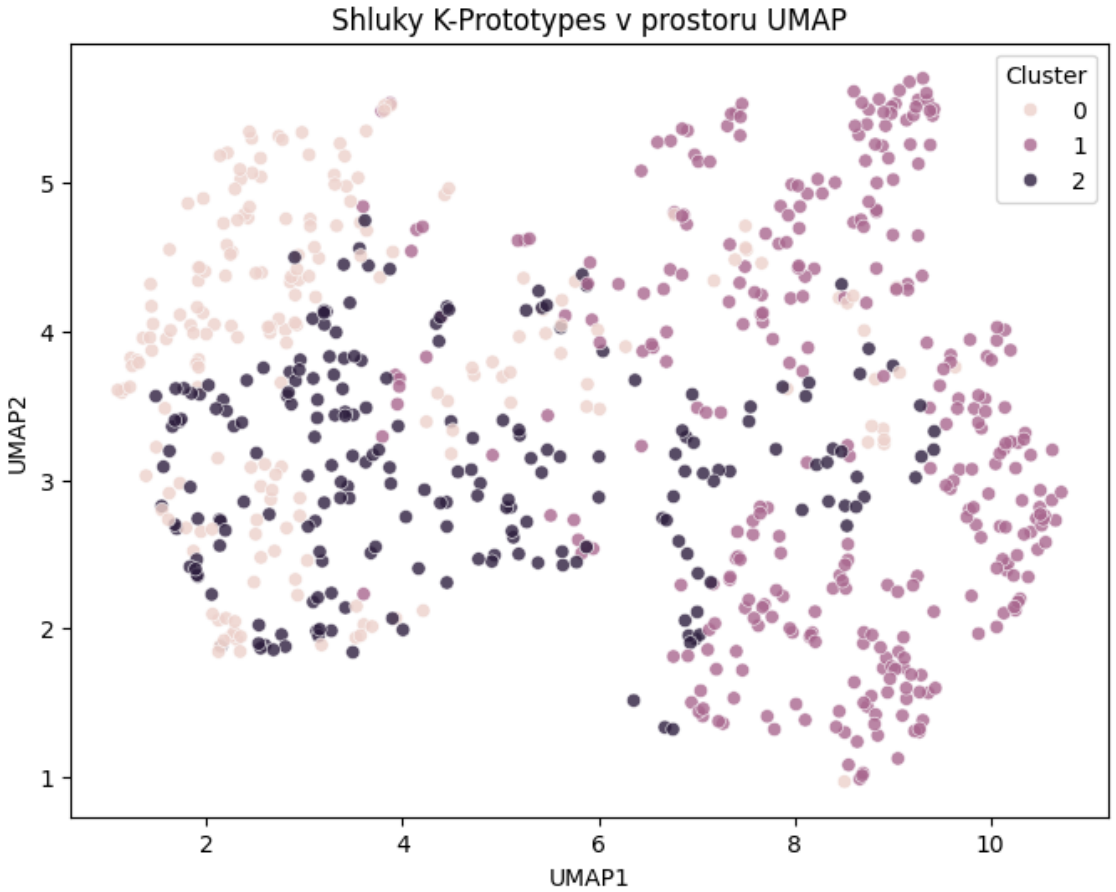
s typem bolesti na hrudi, zátěžovou anginou nebo morfologií ST segmentu. Význam výsledku spočívá v možnosti chápat shluky jako odlišné typy pacientských profilů, nikoli pouze jako technické výstupy algoritmu.

Zvláštní důležitost má porovnání shluků z hlediska zastoupení cílové proměnné. Výrazné rozdíly v podílu pacientů se srdečním onemocněním mezi segmenty potvrzují, že segmentace zachycuje klinicky relevantní heterogenitu pacientské populace. Profilování poskytuje důležitý most mezi explorativní analýzou a následným modelováním, protože ukazuje, jaké kombinace znaků se častěji pojí s vyšším nebo nižším rizikem onemocnění.

2.2.11 Vizualizace shluků v prostoru UMAP

Po vytvoření a interpretaci finálních shluků byla segmentace promítnuta zpět do UMAP prostoru, aby bylo možné vizuálně posoudit, jak se jednotlivé skupiny rozkládají v redukované vícerozměrné projekci. Cílem dané části bylo ověřit, zda profilované shluky odpovídají alespoň částečně zřetelným oblastem v datovém prostoru a zda mezi nimi existují relativně kompaktní jádra nebo naopak plynulé přechody.

V následujícím obrázku je zobrazena UMAP projekce dat rozlišená podle výsledného čísla shluku.



Obr. 29: Shluky K-Prototypes v prostoru UMAP

Zdroj: Vlastní zpracování (2026)

Vizualizace ukazuje, že vytvořené shluky mají v redukovaném prostoru určitou prostorovou soudržnost, avšak mezi skupinami se současně objevují i přechodové oblasti. Získaný výsledek odpovídá charakteru medicínských dat, v nichž bývá hranice mezi různými pacientskými profily

spíše plynulá než ostře oddělená. Přítomnost kompaktnějších částí projekce přesto podporuje závěr, že shlukování zachytilo relevantní vícerozměrnou strukturu dat.

Z interpretačního hlediska má vizualizace význam především díky možnosti propojit technický výstup segmentace s intuitivněji čitelným obrazem datového prostoru. Umožňuje ověřit, že segmenty nevznikly pouze jako abstraktní matematický konstrukt, ale že mají i určitou vizuální oporu v rozmístění pacientů. Výsledek posiluje důvěru v segmentaci jako v analytický krok, který odhaluje reálně přítomnou heterogenitu pacientské populace.

2.3 Supervised EDA a významnost proměnných

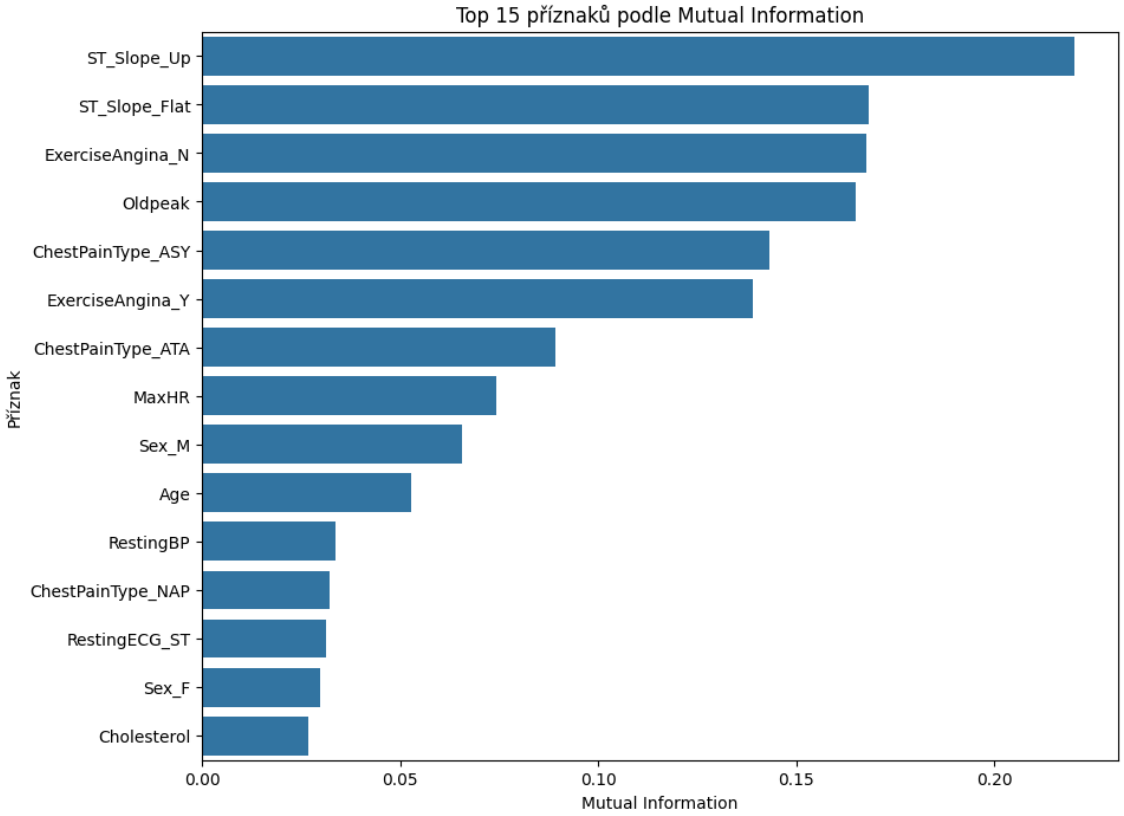
Po dokončení explorativní analýzy a shlukování následovala řízená analýza významnosti proměnných se znalostí cílové proměnné HeartDisease. Smyslem dané fáze bylo určit, které atributy nesou největší podíl predikční informace a které proměnné mají největší význam pro rozlišení pacientů s přítomným a nepřítomným srdečním onemocněním. Na rozdíl od předchozích kroků již nešlo pouze o popis vnitřní struktury dat, ale o posouzení vztahu vstupních proměnných k cílovému znaku v kontextu následného modelování.

Pro řízenou analýzu významnosti byly použity dvě vzájemně se doplňující metody. První metodou byla Mutual Information, která hodnotí, kolik informace jednotlivé proměnné přinášejí o cílové proměnné. Druhou metodou byla SHAP, která umožňuje interpretovat význam proměnných z hlediska jejich vlivu na výsledné predikce modelu. Spojení obou přístupů umožňuje získat přehled o globální důležitosti atributů a zároveň detailnější vhled do směru a síly jejich vlivu.

2.3.1 Významnost proměnných podle Mutual Information

První část supervised EDA byla založena na vyhodnocení významnosti proměnných pomocí metody Mutual Information. Cílem bylo zjistit, které vstupní atributy obsahují největší množství informace relevantní pro predikci cílové proměnné. Výsledek dané analýzy je důležitý zejména pro ověření, zda se mezi nejvýznamnějšími příznaky objeví stejné proměnné, které již v předchozích fázích vykazovaly výraznější vazby k cílovému znaku nebo k důležitým klinickým charakteristikám.

V následujícím grafu jsou zobrazeny hodnoty významnosti proměnných podle Mutual Information.



Obr. 30: Graf top 15 příznaků podle Mutual Information

Zdroj: Vlastní zpracování (2026)

Hodnoty Mutual Information ukazují, že predikční informace není mezi proměnné rozložena rovnoměrně, ale soustřeďuje se do omezeného počtu atributů. Nejvýznamnější roli hrají proměnné spojené se zátěžovým vyšetřením, změnami ST segmentu, přítomností anginy při zátěži a charakterem bolesti na hrudi. Výsledek je metodicky cenný, protože potvrzuje soulad mezi předchozí explorativní analýzou a řízeným hodnocením významnosti příznaků.

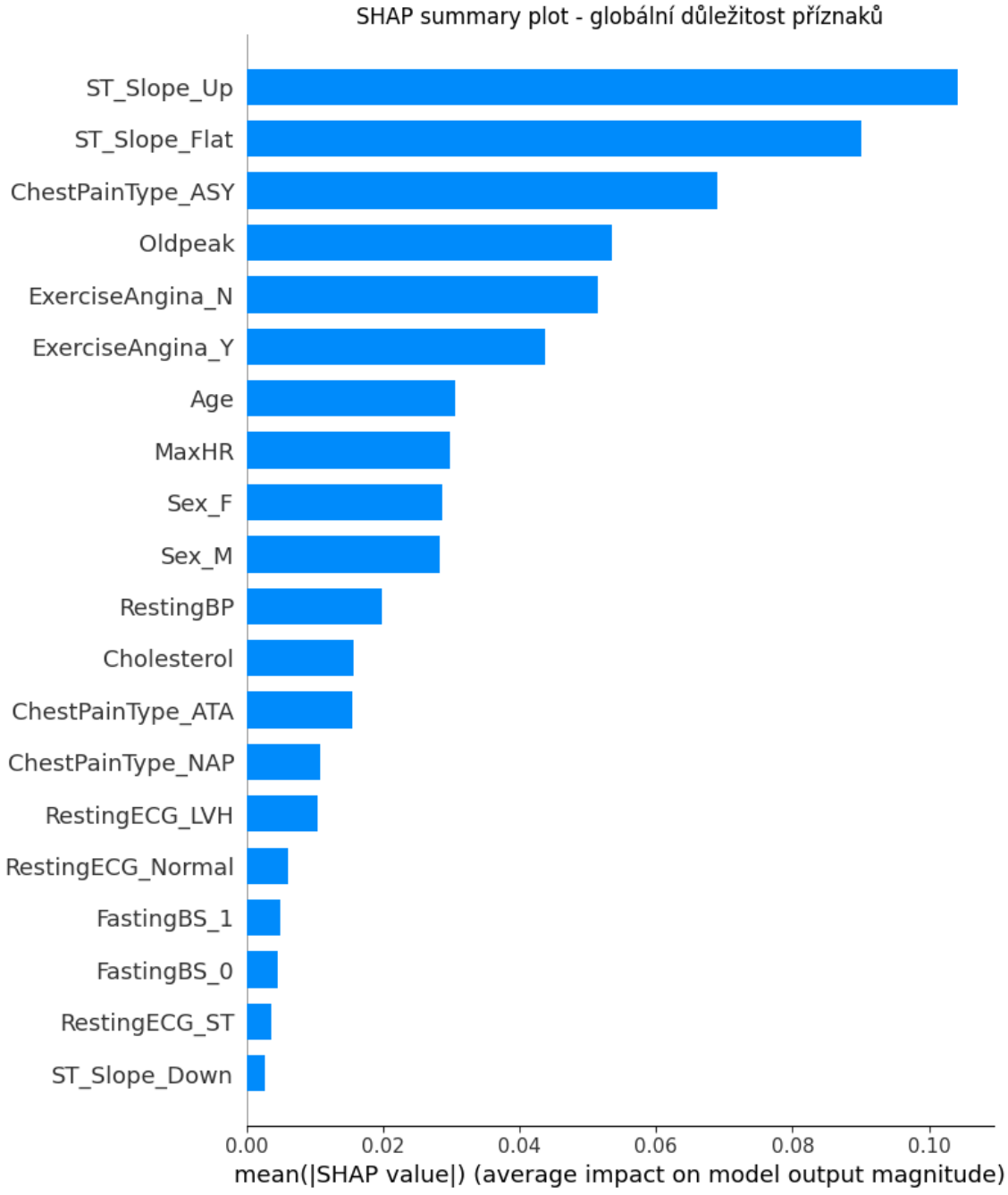
Význam výsledku spočívá i ve skutečnosti, že podstatná část predikční síly datového souboru je soustředěna do klinicky dobře interpretovatelných proměnných. Prediktivní modely proto nestaví na obtížně vysvětlitelných nebo náhodně se jevících atributech, ale na znacích s jasným medicínským obsahem. Právě zmíněná skutečnost zvyšuje důvěryhodnost dalšího modelování a usnadňuje následnou interpretaci výsledků.

Přední pozice proměnné ST_Slope potvrzuje význam změn ST segmentu pro rozlišení pacientů s vyšším a nižším rizikem srdečního onemocnění. Vysoké postavení proměnných ExerciseAngina a ChestPainType podporuje závěr, že klinické projevy související se zátěží a bolestí na hrudi patří mezi nejdůležitější nositele predikční informace. Významnější role proměnných Oldpeak nebo MaxHR současně ukazuje, že číselně měřené zátěžové ukazatele doplňují kategoriální klinické znaky o další důležitou vrstvu informace.

2.3.2 Interpretace významnosti proměnných pomocí SHAP

Druhá část supervised EDA byla zaměřena na interpretaci významnosti proměnných pomocí metody SHAP. Zatímco Mutual Information poskytuje přehled o množství informace, kterou jednotlivé atributy nesou o cílové proměnné, SHAP umožňuje lépe porozumět, jak konkrétní hodnoty proměnných ovlivňují výsledné predikce modelu. Význam dané metody spočívá zejména v možnosti propojit modelový výstup s odborně srozumitelnou interpretací vlivu jednotlivých znaků.

V následujícím obrázku je zobrazen souhrnný SHAP graf významnosti proměnných.



Obr. 31: Souhrnný SHAP graf

Zdroj: Vlastní zpracování (2026)

Souhrnný SHAP graf potvrzuje, že mezi nejdůležitější proměnné patří atributy související se zátěžovým vyšetřením, sklonem ST segmentu, přítomností zátěžové anginy a typem bolesti na hrudi. Shoda s výsledky Mutual Information zvyšuje důvěru v interpretaci významnosti, protože různé analytické přístupy vedou k obdobnému pořadí klíčových prediktorů. Z metodického hlediska jde o silný argument pro stabilitu výsledků a pro věrohodnost následného modelování.

Význam SHAP spočívá i v možnosti zachytit směr působení jednotlivých proměnných. U části atributů lze pozorovat, že vyšší nebo nižší hodnoty zvyšují pravděpodobnost pozitivní třídy, zatímco u kategoriálních proměnných se projevuje rozdílný vliv jednotlivých kategorií. Z hlediska interpretace je důležité, že model nepracuje s proměnnými jako s izolovanými značkami, ale jako s atributy, jejichž konkrétní hodnota mění výsledný odhad rizika.

Proměnná ST_Slope se mezi nejvýznamnějšími příznaky objevuje opakovaně, což potvrzuje její zásadní roli v predikci srdečního onemocnění. Význam proměnné ExerciseAngina ukazuje, že přítomnost zátěžové anginy představuje silný signál směrem k pozitivní třídě. Výsledky pro ChestPainType současně naznačují, že jednotlivé typy bolesti na hrudi nemají stejný vliv na výsledné rozhodování modelu, což odpovídá klinické zkušenosti s rozdílnou diagnostickou závažností různých forem bolesti.

2.3.3 Porovnání výsledků Mutual Information a SHAP

Zařazení dvou metod významnosti proměnných umožňuje posoudit, nakolik jsou výsledky konzistentní napříč odlišnými analytickými přístupy. Mutual Information vyjadřuje obecný informační přínos vstupních atributů vůči cílové proměnné, zatímco SHAP zachycuje význam proměnných již v kontextu konkrétního modelového rozhodování. Společné posouzení obou výstupů proto poskytuje komplexnější pohled na význam jednotlivých znaků.

Srovnání obou metod ukazuje vysokou míru shody v identifikaci klíčových proměnných. Opakovaný výskyt atributů ST_Slope, ExerciseAngina, ChestPainType, případně dalších zátěžových ukazatelů potvrzuje, že rozhodující predikční informace je soustředěna do klinicky významných znaků. Výsledek podporuje závěr, že následné klasifikační modely budou stavět na atributech s jasným odborným obsahem, nikoli na nahodilých nebo obtížně interpretovatelných proměnných.

Konzistence mezi Mutual Information a SHAP má význam i pro další části práce. Vysoká shoda mezi metodami snižuje riziko, že by významnost některého atributu byla pouze artefaktem jedné konkrétní analytické techniky. Analýza významnosti proměnných proto vytváří pevný základ pro interpretaci klasifikačního modelování a zároveň usnadňuje navázání na diskusní část práce, v níž bude možné výsledky porovnat s poznatky uvedenými v rešerši.

2.4 Výsledky klasifikačního modelování

Po dokončení explorativní analýzy, vícerozměrného průzkumu dat a řízeného hodnocení významnosti proměnných následovalo klasifikační modelování. Hlavním cílem bylo porovnat více algoritmů odlišného typu a určit, které přístupy nejlépe rozlišují pacienty s přítomným a nepřítomným srdečním onemocněním. Vedle samotného výkonu byla sledována i stabilita výsledků, aby bylo možné posoudit nejen maximálně dosaženou metriku, ale i robustnost jednotlivých modelů při různých rozděleních dat.

Do srovnání byly zařazeny modely logistické regrese, rozhodovacího stromu, náhodného lesa, KNN a SVM. Výběr pokrývá lineární přístup, stromový model, ansámblovou metodu, instanční klasifikátor a metodu založenou na maximálním oddělení tříd. Hodnocení vycházelo z více metrik současně, protože jediný ukazatel by nepostačoval k zachycení všech důležitých vlastností modelu. Pozornost proto směřovala na přesnost klasifikace, schopnost zachytit pozitivní případy, globální diskriminační výkon i stabilitu napříč opakovanými běhy.

2.4.1 Souhrnné porovnání klasifikačních modelů

První krok klasifikační evaluace spočíval v souhrnném porovnání všech testovaných modelů pomocí základních klasifikačních metrik. Smyslem dané části bylo získat přehled o modelech dosahujících nejlepších výsledků při běžném hodnocení na testovací množině a o míře rozdílů mezi jednotlivými přístupy.

V následující tabulce jsou uvedeny hlavní výsledky klasifikačních modelů podle vybraných metrik.

Tab. 9: Hlavní výsledky klasifikačních modelů

Model	Best CV ROC AUC	Test Accuracy	Test Precision	Test Recall	Test F1-score	Test ROC AUC	Test PR AUC	Test MCC
SVM	0.920	0.8641	0.8812	0.872	0.876	0.938	0.944	0.725
KNN	0.922	0.9130	0.9216	0.921	0.921	0.937	0.920	0.824
Logistická regrese	0.922	0.8913	0.8868	0.921	0.903	0.932	0.942	0.779
Náhodný les	0.932	0.8967	0.8807	0.941	0.910	0.931	0.936	0.791
Rozhodovací strom	0.904	0.8207	0.8710	0.794	0.830	0.889	0.879	0.644

Zdroj: Vlastní zpracování (2026)

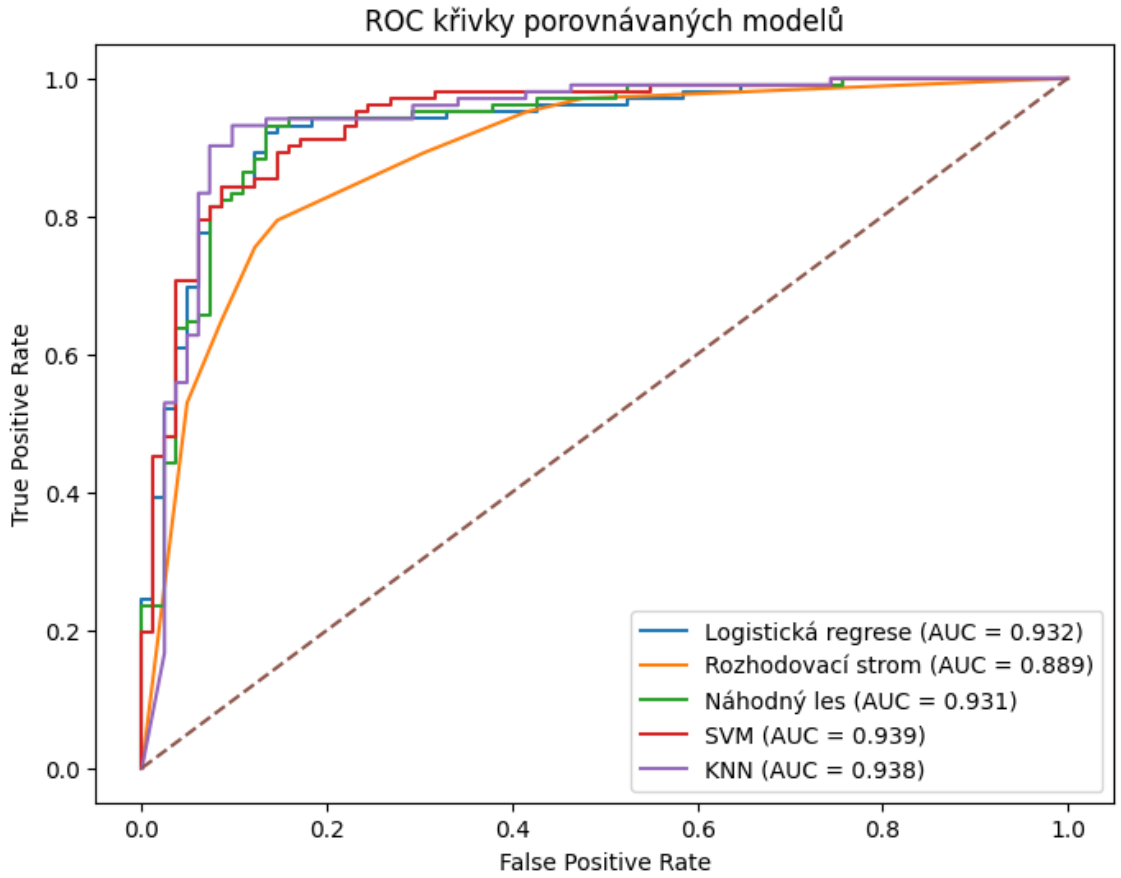
Souhrnné porovnání ukazuje, že mezi nejlepšími modely nejsou zanedbatelné, avšak ani extrémně výrazné rozdíly. Výsledek naznačuje přítomnost relativně silného predikčního signálu v datech, který dokáže zachytit více typů klasifikátorů. Významnější rozdíly se objevují zejména mezi nejúspěšnější skupinou modelů a rozhodovacím stromem, jehož výkon bývá v podobných úlohách často omezen nižší schopností zachytit složitější strukturu dat bez podpory ansámblových postupů.

Z interpretačního hlediska je důležité, že výkon modelů nelze posuzovat pouze podle jedné metriky. Některé algoritmy mohou dosahovat vyšší celkové přesnosti, zatímco jiné lépe rozlišují pozitivní případy nebo vykazují vyšší globální diskriminační schopnost. Právě z tohoto důvodu byla evaluace vedena víceúrovňově a nebyla redukována na jednorázové pořadí modelů podle jediného ukazatele.

2.4.2 ROC křivky a diskriminační schopnost modelů

Významnou část evaluace představuje porovnání ROC křivek, které zachycují diskriminační schopnost modelů napříč různými rozhodovacími prahy. Smyslem dané části bylo posoudit, nakolik si jednotlivé klasifikátory udržují schopnost odlišit pozitivní a negativní případy bez fixace na jediné konkrétní nastavení rozhodovací hranice.

V následujícím grafu jsou zobrazeny ROC křivky klasifikačních modelů.



Obr. 32: Graf porovnání ROC křivek

Zdroj: Vlastní zpracování (2026)

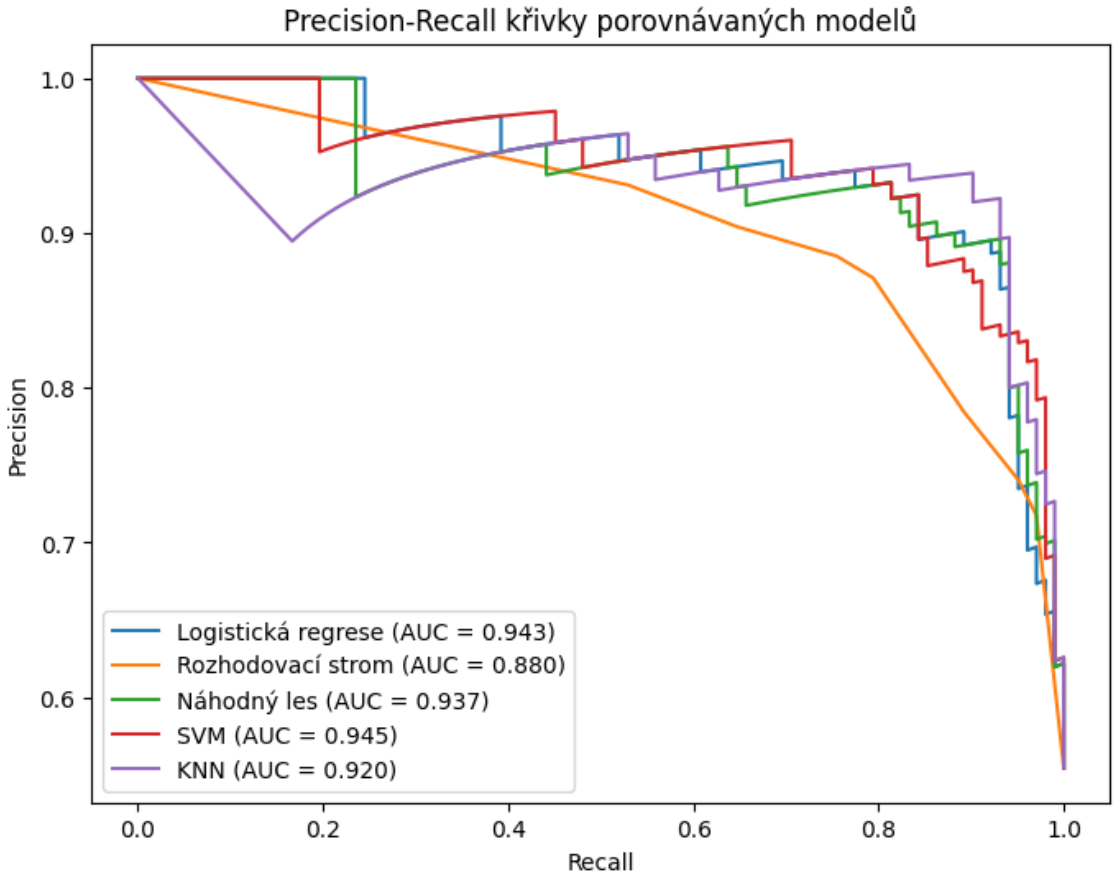
ROC křivky potvrzují, že více modelů dosahuje velmi dobré diskriminační schopnosti. Výraznější oddálení od diagonály náhodného klasifikátoru ukazuje, že modely dokážou rozlišovat mezi pacienty s přítomným a nepřítomným srdečním onemocněním výrazně lépe než náhodné rozhodování. Z hlediska celkového pořadí je důležité posoudit nejen absolutní hodnotu AUC, ale i podobnost křivek mezi nejlepšími algoritmy.

Význam ROC analýzy spočívá především v možnosti sledovat chování modelu v celém rozsahu prahů. V klinickém prostředí nemusí být optimální používat vždy stejnou rozhodovací hranici, protože preference mezi vyšším záchytem nemocných pacientů a omezením falešně pozitivních nálezů se může měnit. Model s vysokou výkonností napříč prahy proto představuje obecně robustnější řešení než model, který dosahuje dobrého výsledku pouze v úzkém bodě nastavení.

2.4.3 Precision–Recall křivky

Vedle ROC analýzy byly vyhodnoceny i Precision–Recall křivky, které poskytují doplňující pohled na kvalitu klasifikace se zaměřením na vztah mezi přesností pozitivních predikcí a záchytem pozitivních případů. Hodnota zmíněné metody spočívá v citlivosti na chování modelu u pozitivní třídy, což může být v medicínských úlohách zvlášť důležité.

V následujícím grafu jsou zobrazeny Precision–Recall křivky klasifikačních modelů.



Obr. 33: Graf porovnání Precision-Recall křivek

Zdroj: Vlastní zpracování (2026)

Precision–Recall křivky ukazují, nakolik si modely dokážou zachovat vysokou přesnost pozitivních predikcí při současném růstu záchytu pozitivních případů. Výsledek doplňuje ROC analýzu o pohled, který je více zaměřen na praktické rozlišení pozitivní třídy. Vyšší plocha pod PR křivkou naznačuje lepší schopnost modelu identifikovat nemocné pacienty bez výrazného nárůstu falešně pozitivních klasifikací.

Z interpretačního hlediska je význam PR křivek spojen se zvýrazněním modelů vhodných pro situace, v nichž je důležité soustředit se především na pozitivní třídu. V prostředí zdravotnictví může mít právě zmíněné vyhodnocení vyšší praktickou hodnotu než metriky orientované pouze na celkovou přesnost klasifikace.

2.4.4 Statistické hodnocení výkonnosti modelů

Bodové hodnoty metrik byly doplněny statistickým hodnocením založeným na opakované křížové validaci a na výpočtu intervalů spolehlivosti. Cílem dané části bylo ověřit, zda výkonnost modelů zůstává stabilní i při různých rozděleních dat a zda rozdíly mezi algoritmy nejsou založeny pouze na jednom konkrétním běhu.

V následující tabulce jsou uvedeny průměrné hodnoty vybraných metrik a jejich intervaly spolehlivosti.

Tab. 10: Průměrné hodnoty vybraných metrik a jejich intervaly spolehlivosti

Model	Accuracy_mean	Accuracy_CI_lower	Accuracy_CI_upper	ROC_AUC_mean	ROC_AUC_CI_lower	ROC_AUC_CI_upper
Náhodný les	0.87006	0.82710	0.91293	0.92701	0.88608	0.96061
SVM	0.86548	0.81622	0.90749	0.92463	0.88976	0.95719
Logistická regrese	0.85902	0.81443	0.90760	0.92258	0.88267	0.95426
KNN	0.86978	0.82166	0.92236	0.92216	0.87856	0.95685
Rozhodovací strom	0.83568	0.78901	0.88572	0.89666	0.85230	0.93762

Zdroj: Vlastní zpracování (2026)

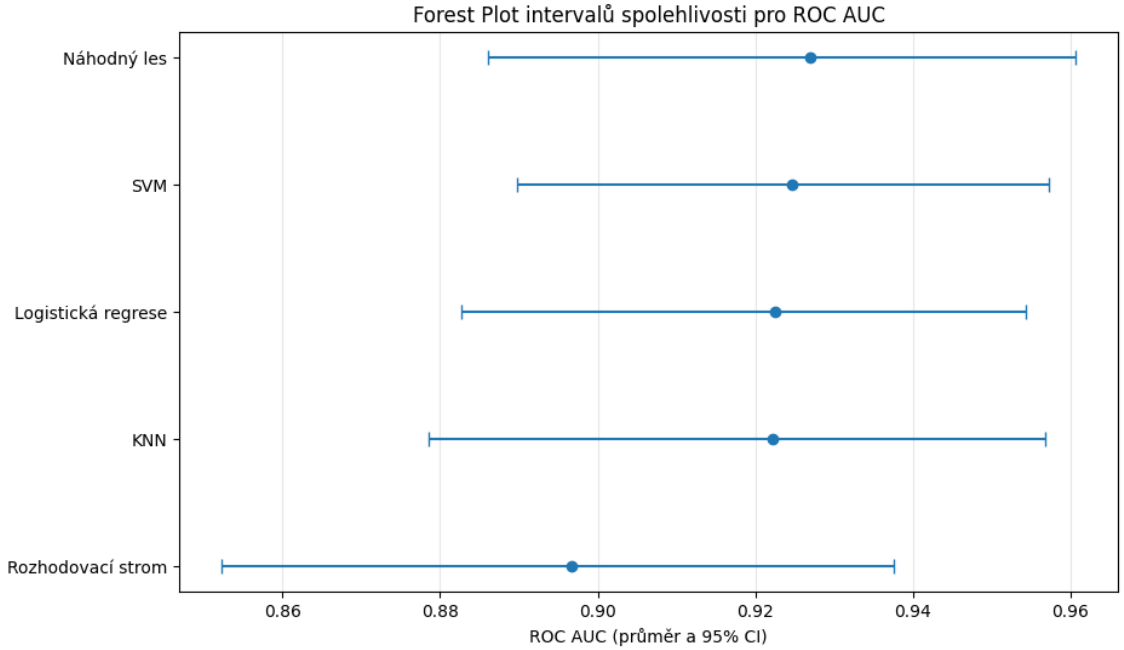
Tabulka ukazuje, že nejúspěšnější modely dosahují nejen vysokých průměrných hodnot metrik, ale i relativně úzkých intervalů spolehlivosti. Zmíněný výsledek podporuje závěr, že dobrý výkon není dán pouze příznivým jednorázovým rozdělením dat, ale že se opakuje i při dalších bězích. Z metodického hlediska jde o důležitý krok, protože interpretace modelů založená výhradně na jednom testovacím výsledku by neposkytovala dostatečný obraz o jejich stabilitě.

Význam intervalů spolehlivosti spočívá i v možnosti odpovědnějšího porovnání modelů. Překrývání intervalů u nejlepších klasifikátorů naznačuje, že rozdíly mezi nimi nemusejí být natolik výrazné, aby bylo možné bez dalšího prohlásit jediný model za jednoznačně nejlepší. Konečné rozhodnutí proto musí zohlednit nejen maximální metriku, ale i robustnost, interpretovatelnost a praktickou použitelnost.

2.4.5 Forest Plot intervalů spolehlivosti

Pro lepší vizuální porovnání stability modelů byl použit Forest Plot, který zobrazuje průměrnou hodnotu vybrané metriky a její interval spolehlivosti. Grafická forma usnadňuje porovnání nejen výkonnosti modelů, ale i míry nejistoty spojené s jejich hodnocením.

V následujícím grafu je zobrazen Forest Plot intervalů spolehlivosti klasifikačních modelů.



Obr. 34: Forest Plot intervalů spolehlivosti pro ROC AUC

Zdroj: Vlastní zpracování (2026)

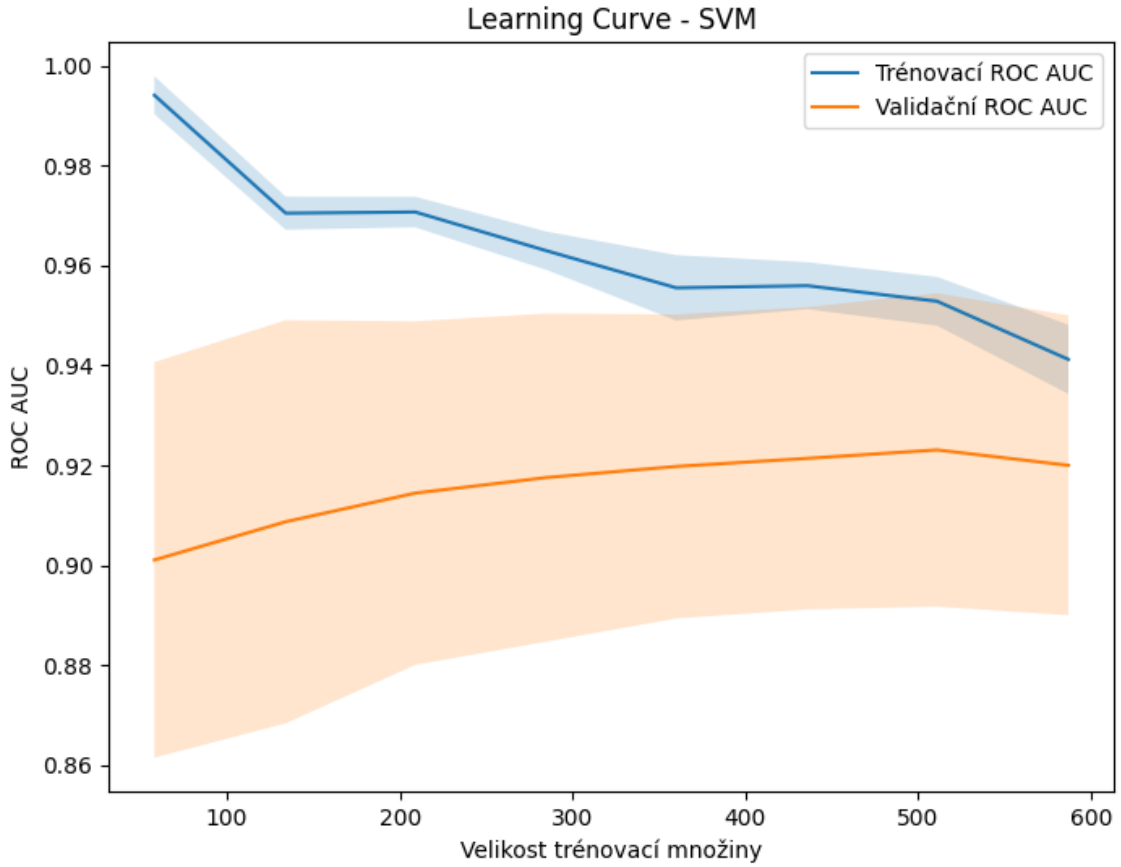
Forest Plot umožňuje rychle rozpoznat, které modely vykazují nejen vysoký výkon, ale i stabilnější chování. Užší intervaly spolehlivosti ukazují menší citlivost na konkrétní rozdělení dat, zatímco širší intervaly naznačují vyšší variabilitu výsledků. Z interpretačního hlediska jde o důležitý doplněk tabulkového hodnocení, protože umožňuje přehledněji posoudit, zda vysoká hodnota metriky není spojena s vyšší nejistotou.

Praktický význam Forest Plotu spočívá ve zvýraznění rozdílu mezi výkonností a stabilitou. Model s nepatrně nižší průměrnou hodnotou může být v některých situacích vhodnější než model s nejvyšším průměrem, pokud současně vykazuje konzistentnější chování napříč opakovanými běhy. Daný pohled je pro odpovědné hodnocení modelů ve zdravotnickém kontextu obzvlášť důležitý.

2.4.6 Learning curve nejúspěšnějšího modelu

Součástí klasifikační evaluace byla i learning curve, vytvořená pro nejúspěšnější model. Cílem bylo posoudit, jak se mění trénovací a validační výkon v závislosti na velikosti trénovací množiny a zda model vykazuje známky přeučení, podučení nebo prostor pro další zlepšení při větším objemu dat.

V následujícím grafu je zobrazena learning curve nejúspěšnějšího klasifikačního modelu.



Obr. 35: Graf Learning Curve modelu SVM

Zdroj: Vlastní zpracování (2026)

Průběh learning curve modelu SVM ukazuje, že s rostoucí velikostí trénovací množiny postupně klesá trénovací hodnota ROC AUC z velmi vysoké úrovně přibližně 0,99 na hodnoty okolo 0,94, zatímco validační ROC AUC naopak mírně roste z přibližně 0,90 k hodnotám kolem 0,92. Vývoj obou křivek naznačuje, že model při menším objemu dat dosahuje na trénovací množině velmi vysokého výkonu, avšak část výkonu není plně přenositelná na validační data. S postupným navyšováním počtu trénovacích vzorků se rozdíl mezi oběma křivkami zmenšuje, což svědčí o lepší generalizační schopnosti modelu.

Graf současně ukazuje, že i při největší velikosti trénovací množiny přetrvává určitá mezera mezi trénovací a validační křivkou. Přítomnost dané mezery naznačuje mírnou tendenci k přeučení, avšak nikoli v extrémní podobě. Validační výkon zůstává stabilně vysoký a nejeví známky výrazného zhoršování, což podporuje závěr, že model SVM je pro daný problém vhodný a dokáže zachytit podstatnou část predikční informace obsažené v datech. Současně lze usuzovat, že další zlepšení by mohlo přinést buď větší množství trénovacích dat, nebo jemnější optimalizace hyperparametrů, protože validační křivka se ve vyšších velikostech trénovací množiny ještě zcela neustálila na zjevné maximální hodnotě.

2.4.7 Přehled klíčových zjištění z klasifikačního modelování

Klasifikační modelování ukázalo, že analyzovaný datový soubor obsahuje dostatečně silný predikční signál pro úspěšné rozlišování pacientů s přítomným a nepřítomným srdečním

onemocněním. Více modelů dosáhlo velmi dobrých výsledků, což podporuje závěr, že rozhodující informace je skutečně obsažena ve vstupních proměnných a není závislá pouze na jediné konkrétní metodě. Výsledky současně ukázaly, že mezi nejlepšími klasifikátory nemusí být rozdíly natolik výrazné, aby bylo možné rozhodnout výhradně podle jedné bodové metriky.

Z metodického hlediska má největší význam kombinace více úrovní evaluace. Souhrnné metriky, matice záměn, ROC a PR křivky, intervaly spolehlivosti, Forest Plot i learning curve dohromady vytvářejí podstatně přesnější obraz o chování modelů než kterýkoli z nástrojů samostatně. Výsledková část proto poskytuje základ nejen pro výběr nejvhodnějšího modelu, ale i pro navazující diskusi nad důvody jeho úspěšnosti a nad vazbou mezi výkonem modelu a významností jednotlivých proměnných.

2.5 Výsledky experimentů

Vedle hlavní klasifikační úlohy byly do praktické části zařazeny doplňkové experimenty, jejichž cílem bylo ověřit stabilitu zvoleného řešení při změně vybraných podmínek a rozšířit pohled na řešený problém. Pozornost byla soustředěna na tři experimentální směry: změnu velikosti trénovací a testovací množiny, práci s chybějícími daty a regresní interpretaci binární cílové proměnné. Zařazení experimentů umožňuje lépe posoudit citlivost analytického postupu na změnu vstupních předpokladů a současně ukazuje, že datový soubor lze interpretovat více než jedním způsobem.

2.5.1 Experiment se změnou velikosti trénovací a testovací množiny

První experiment byl zaměřen na ověření, jak se mění výsledky klasifikačního modelování při odlišném rozdělení dat na trénovací a testovací část. Smyslem nebylo hledat jediné správné rozdělení, ale posoudit, zda modely vykazují při změně velikosti trénovací množiny výrazné kolísání výkonu, nebo zda si udržují relativně stabilní výsledky.

V následující tabulce jsou uvedeny výsledky experimentu se změnou velikosti trénovací a testovací množiny.

Tab. 11: Výsledky experimentu se změnou velikosti trénovací a testovací množiny

Train/Test	Accuracy	F1-score	ROC AUC	MCC
85:15	0.8478	0.8609	0.9391	0.6930
80:20	0.8641	0.8768	0.9388	0.7254
75:25	0.8870	0.9000	0.9453	0.7712
70:30	0.8913	0.9038	0.9457	0.7797

Zdroj: Vlastní zpracování (2026)

Výsledky ukazují, že klasifikační výkon modelu se s rostoucím podílem trénovacích dat postupně zlepšuje. Nejnižších hodnot sledovaných metrik bylo dosaženo při rozdělení 85:15, zatímco nejlepší výsledky vykazuje rozdělení 70:30, kde model dosahuje nejvyšší Accuracy (0,8913), F1-score (0,9038), ROC AUC (0,9457) i MCC (0,7797). Průběh hodnot naznačuje, že model dokáže z větší trénovací množiny efektivněji využít dostupnou predikční informaci a dosahuje stabilnějšího a celkově lepšího výkonu.

Z interpretačního hlediska je důležité, že změna poměru trénovacích a testovacích dat nemění pouze jednu dílčí metriku, ale promítá se konzistentně napříč všemi sledovanými ukazateli. Výsledek podporuje závěr, že výkon klasifikace je citlivý na objem dat dostupných pro učení a že dostatečně rozsáhlá trénovací množina představuje důležitý předpoklad pro kvalitní generalizaci modelu. Z praktického pohledu jde o cenný doplněk hlavní evaluace, protože ukazuje, že kvalita výsledků není oddělena od množství dat, která má model při učení skutečně k dispozici.

2.5.2 Experiment s chybějícími daty

Druhý experiment byl zaměřen na situaci, kdy datový soubor neobsahuje kompletní informace pro všechny pacienty. Cílem bylo ověřit, jak se změna kvality vstupních dat promítne do výsledků modelování a zda zvolený analytický postup dokáže pracovat i s méně úplným datasetem bez zásadního poklesu výkonu.

V následující tabulce jsou uvedeny výsledky experimentu s chybějícími daty.

Tab. 12: Porovnání výsledků hlavní analýzy s výsledky experimentu chybějících hodnot

Scénář	Accuracy	F1-score	ROC AUC	PR AUC	MCC
Hlavní analýza	0.8641	0.8768	0.9388	0.9447	0.7254
Data s chybějícími hodnotami	0.8641	0.8804	0.9250	0.9303	0.7245

Zdroj: Vlastní zpracování (2026)

Výsledky experimentu ukazují, že modelování je vůči neúplnosti vstupních údajů relativně odolné. Ve srovnání s hlavní analýzou zůstala hodnota Accuracy beze změny (0,8641), zatímco F1-score se mírně zvýšil z 0,8768 na 0,8804. Naopak u metrik ROC AUC, PR AUC a MCC došlo k mírnému poklesu, konkrétně na hodnoty 0,9250, 0,9303 a 0,7245. Výsledky proto naznačují, že přítomnost chybějících dat nevede k zásadnímu narušení klasifikační schopnosti modelu, avšak částečně oslabuje jeho globální diskriminační výkon.

Z interpretačního hlediska je důležité, že pokles výkonu se neprojevuje rovnoměrně ve všech metrikách. Zatímco klasifikace při konkrétním rozhodovacím prahu zůstává prakticky stabilní, souhrnné metriky hodnotící schopnost oddělovat třídy napříč prahy vykazují určité zhoršení. Praktický význam spočívá ve skutečnosti, že navržený model si uchovává použitelný výkon i při nižší kvalitě vstupních dat, což je v prostředí klinických dat cenné, protože úplnost záznamů nebývá v reálné praxi samozřejmostí.

2.5.3 Regresní interpretace klasifikační úlohy

Třetí experiment byl založen na převodu binární klasifikace na odhad pravděpodobnosti příslušnosti k pozitivní třídě. Namísto konečného přiřazení hodnoty 0 nebo 1 byl sledován spojitý výstup v intervalu od 0 do 1, vyjadřující odhad míry rizika srdečního onemocnění. Zvolený přístup má praktický význam zejména ve zdravotnickém kontextu, kde může být užitečnější znát nejen kategorické rozhodnutí, ale i intenzitu odhadovaného rizika.

V následující tabulce jsou uvedeny hlavní výsledky regresního experimentu podle vybraných metrik.

Tab. 13: Výsledky regresního experimentu podle vybraných metrik

Metrika	Hodnota
MAE	0.201362
MSE	0.096432
RMSE	0.310536
R^2	0.609658
Brier score	0.096432
ROC AUC	0.929938

Zdroj: Vlastní zpracování (2026)

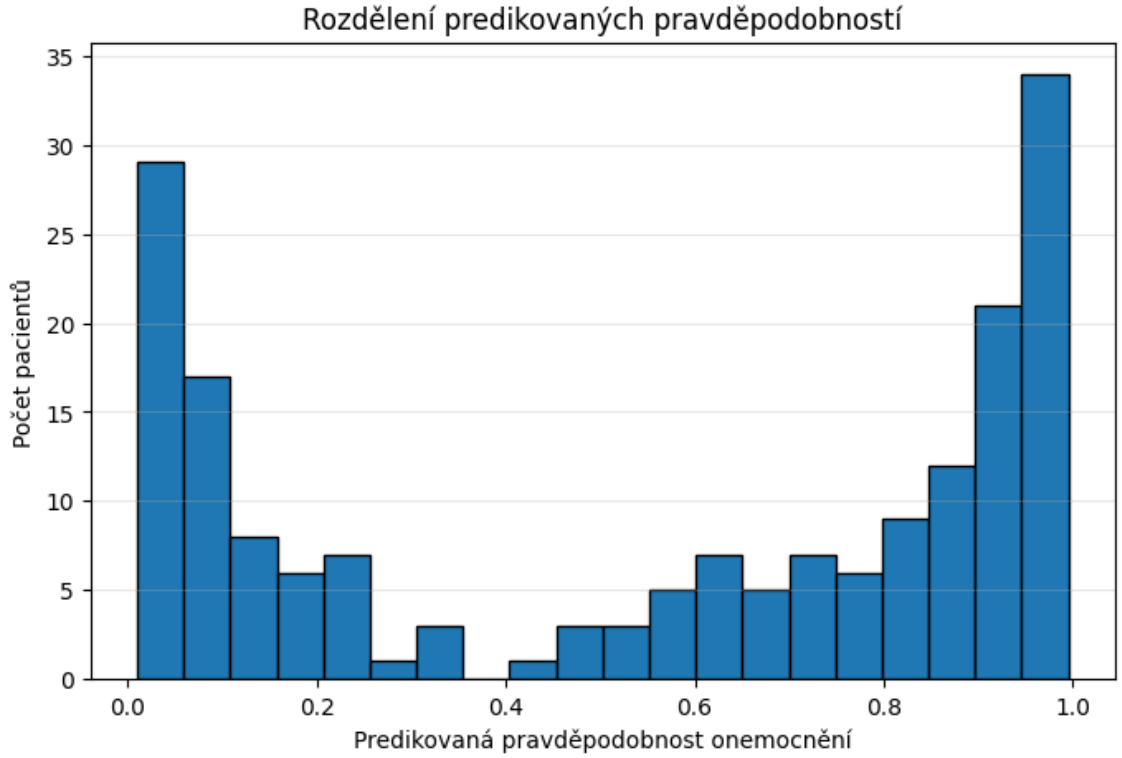
Dosažené hodnoty ukazují, že model poskytuje poměrně přesné pravděpodobnostní odhady a současně si zachovává velmi dobrou schopnost rozlišovat mezi pozitivními a negativními případy. Hodnota MAE 0,201 znamená, že průměrná absolutní odchylka predikované pravděpodobnosti od skutečné hodnoty 0 nebo 1 zůstává relativně nízká, zatímco RMSE 0,311 potvrzuje, že ani po silnějším zohlednění větších chyb nedochází k výraznému zhoršení výsledku. Hodnota R^2 0,610 naznačuje, že model dokáže vysvětlit podstatnou část variability cílové proměnné i při jejím regresním pojetí.

Významný je i Brier score 0,096, který ukazuje dobrou kvalitu pravděpodobnostního odhadu, a zejména ROC AUC 0,930, potvrzující velmi dobrou diskriminační schopnost modelu i v rámci regresní interpretace. Výsledek podporuje závěr, že binární problém lze chápat nejen jako rozhodnutí mezi třídou 0 a 1, ale i jako odhad intenzity rizika. Regresní experiment proto nerozšiřuje klasifikační úlohu pouze formálně, ale přináší prakticky cenný pohled na míru pravděpodobnosti srdečního onemocnění u jednotlivých pacientů.

2.5.4 Rozdělení predikovaných pravděpodobností

Další část regresního experimentu byla zaměřena na rozložení predikovaných pravděpodobností. Cílem bylo posoudit, zda model přiřazuje pacientům hodnoty rozprostřené napříč intervalem rizika, nebo zda se většina výstupů soustřeďuje pouze do úzkého pásma.

V následujícím grafu je zobrazen histogram predikovaných pravděpodobností.



Obr. 36: Graf rozdělení predikovaných pravděpodobností

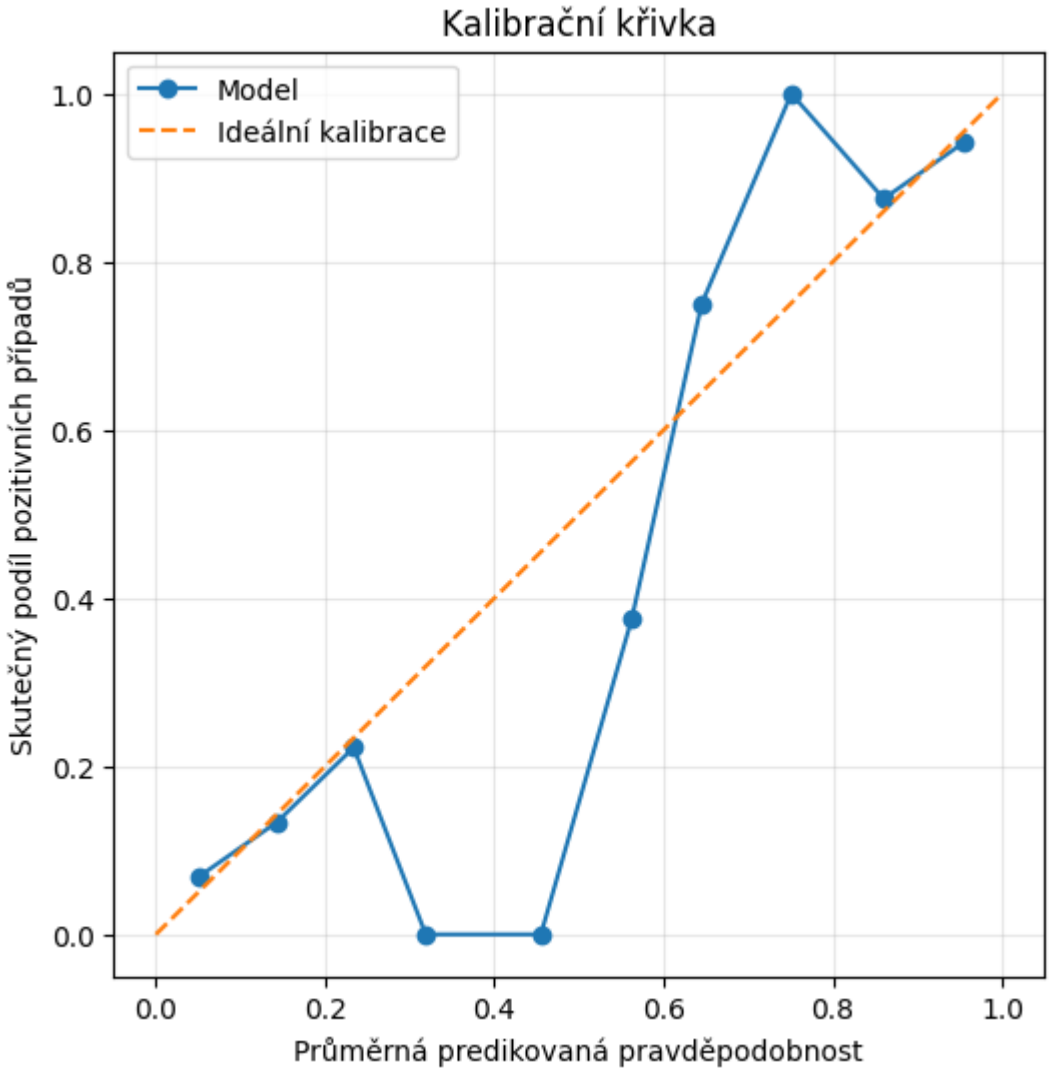
Zdroj: Vlastní zpracování (2026)

Rozdělení pravděpodobností ukazuje, že model nepřiděluje všem pacientům podobné hodnoty, ale vytváří rozumně členěné rozložení od nízkého po vysoké riziko. Praktická hodnota spočívá v možnosti rozlišit nejen dvě krajní skupiny, ale i pacienty s hraničním nebo středním rizikovým profilem. Zjištěná vlastnost dává regresní interpretaci význam z hlediska klinického rozhodování.

2.5.5 Kalibrace pravděpodobnostního odhadu

Vedle samotné velikosti chyby byla posouzena rovněž kalibrace modelu neboli míra souladu mezi predikovanou pravděpodobností a skutečným výskytem pozitivních případů. Kalibrace představuje důležitý aspekt zejména v situaci, kdy je model používán jako nástroj pro odhad rizika.

V následujícím grafu je zobrazena kalibrační křivka regresního experimentu.



Obr. 37: Graf kalibrační křivky regresního experimentu

Zdroj: Vlastní zpracování (2026)

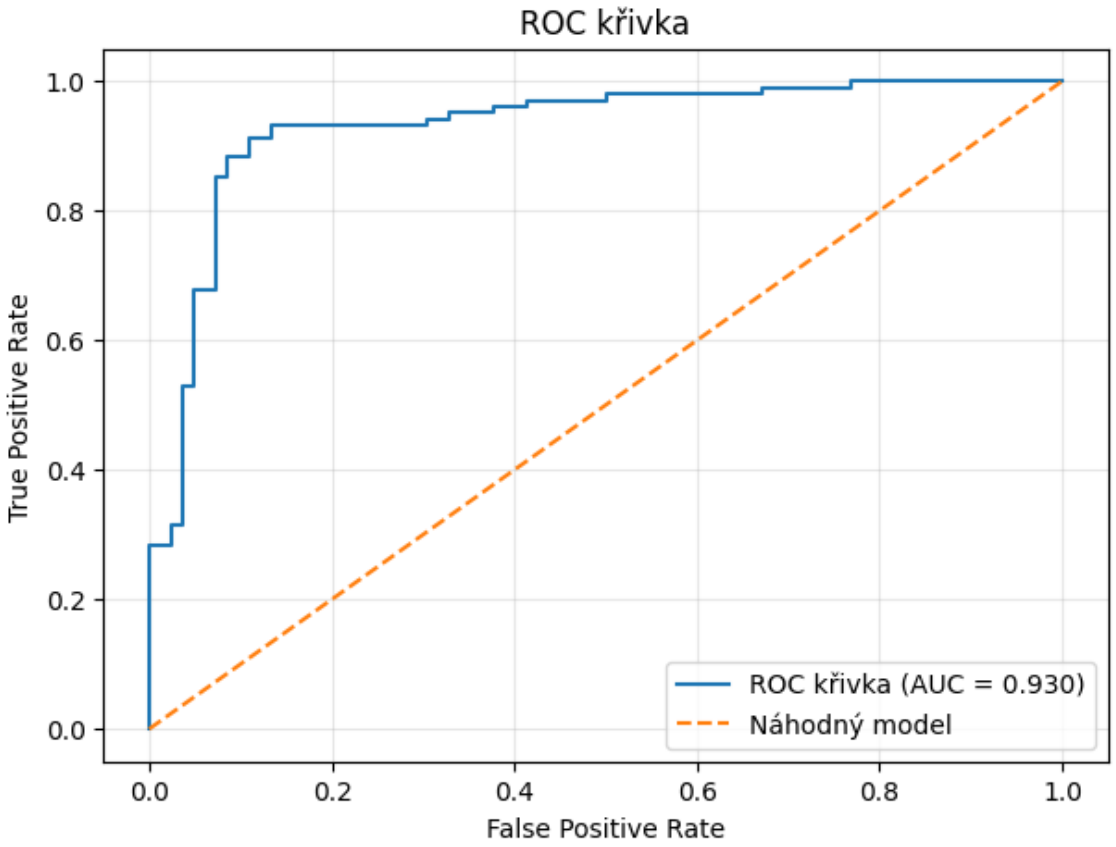
Kalibrační křivka ukazuje, nakolik se odhadnuté pravděpodobnosti přibližují skutečné četnosti onemocnění v jednotlivých intervalech. V nižších a vyšších částech rozsahu se body v několika případech pohybují poměrně blízko ideální diagonály, což podporuje závěr o použitelnosti modelu pro pravděpodobnostní interpretaci. Ve střední části intervalu se však objevují výraznější odchylky, které naznačují, že odhad rizika není ve všech pásmech stejně přesný.

Význam grafu spočívá v rozlišení mezi dobrou diskriminační schopností a přesnou kalibrací pravděpodobností. Model dokáže pacienty velmi dobře rozlišovat z hlediska rizika, avšak některé intervaly ukazují, že číselná hodnota predikované pravděpodobnosti nemusí vždy přesně odpovídat skutečnému podílu pozitivních případů. Výsledná pravděpodobnost proto musí být chápána jako užitečný modelový odhad, nikoli jako absolutně přesná veličina.

2.5.6 ROC analýza regresního experimentu

Přestože experiment pracuje se spojitým výstupem, zůstává důležité posoudit i diskriminační schopnost modelu mezi oběma třídami. Z daného důvodu byla do hodnocení zařazena i ROC křivka a hodnota AUC.

V následujícím grafu je zobrazena ROC křivka regresního experimentu.



Obr. 38: Graf ROC křivky regresního experimentu

Zdroj: Vlastní zpracování (2026)

Výsledek potvrzuje, že i při regresní interpretaci si model zachovává velmi dobrou schopnost rozlišovat mezi pacienty s přítomným a nepřítomným srdečním onemocněním. Hodnota AUC a výrazné oddálení křivky od diagonály náhodného modelu ukazují, že model dokáže pozitivním případům přiřazovat vyšší pravděpodobnosti než případům negativním s vysokou mírou spolehlivosti. Graf podporuje závěr, že regresní pojetí úlohy neoslabuje klasifikační kvalitu modelu, ale naopak ji doplňuje o možnost vyjádřit míru rizika spojitou pravděpodobnostní hodnotou.

2.5.7 Statistické hodnocení regresního experimentu

Podobně jako u klasifikačních modelů bylo i u regresního experimentu zařazeno statistické hodnocení pomocí opakované validace a intervalů spolehlivosti. Cílem bylo ověřit, zda dosažené výsledky zůstávají stabilní i při různých rozděleních dat.

V následující tabulce jsou uvedeny průměrné hodnoty metrik regresního experimentu a jejich intervaly spolehlivosti.

Tab. 14: Statistické hodnocení regresního experimentu

Metrika	Průměr	CI_2.5 %	CI_97.5 %
MAE	0.205424	0.176816	0.236932
RMSE	0.322602	0.280075	0.366990
R2	0.576716	0.454833	0.682813
Brier	0.104614	0.078442	0.134682
ROC_AUC	0.924830	0.887554	0.957283

Zdroj: Vlastní zpracování (2026)

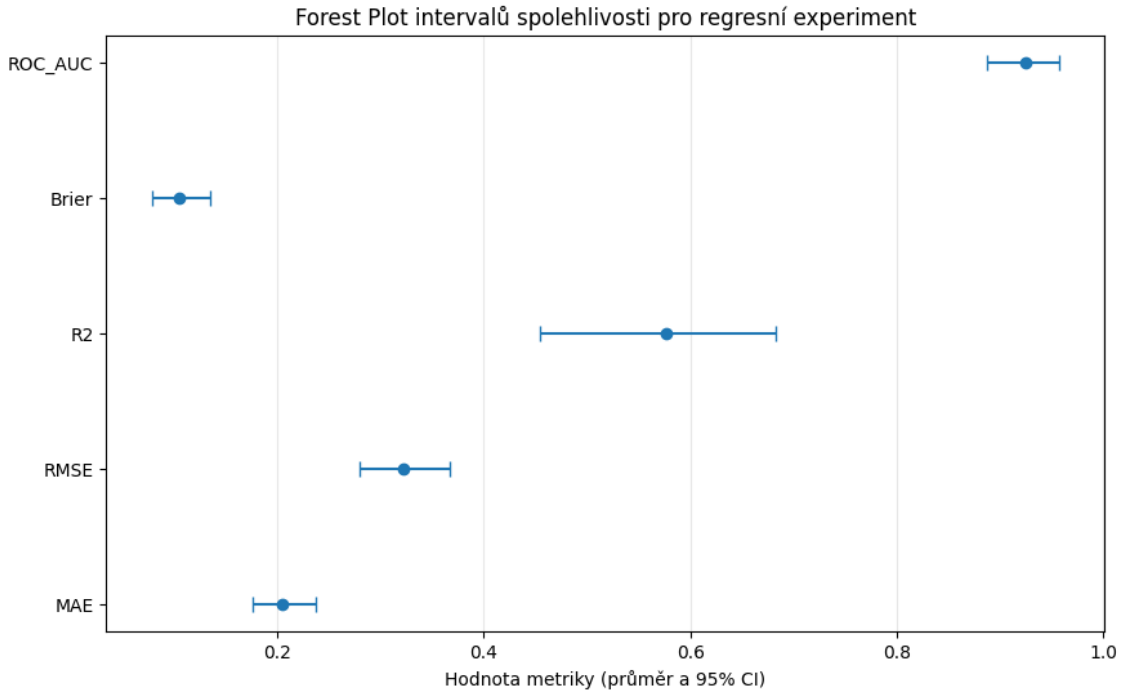
Tabulka ukazuje, že výkon regresního experimentu není založen pouze na jednom příznivém běhu, ale zůstává relativně stabilní i při opakovaném hodnocení. Průměrná hodnota MAE a RMSE potvrzují, že chyba pravděpodobnostního odhadu zůstává na rozumné úrovni i při různých rozděleních dat. Hodnota R^2 současně naznačuje, že model dokáže vysvětlit podstatnou část variability cílové proměnné i v regresním pojetí. Podobně rovněž Brier score ukazuje, že pravděpodobnostní odhad si udržuje dobrou kvalitu napříč opakovanými běhy.

Významnou roli hrají i samotné intervaly spolehlivosti. Poměrně úzké rozpětí u většiny metrik naznačuje, že výsledky nejsou výrazně citlivé na konkrétní rozdělení dat a že model si uchovává stabilní diskriminační schopnost i kvalitu pravděpodobnostního odhadu. Zvlášť důležitá je metrika ROC AUC, jejíž průměrná hodnota 0,925 spolu s intervalem přibližně od 0,888 do 0,957 potvrzuje, že velmi dobrá rozlišovací schopnost modelu není ojedinělým výsledkem, ale opakovaně se potvrzuje i v dalších běžích analýzy.

2.5.8 Forest Plot regresního experimentu

Pro přehlednější vizualizaci stability regresních metrik byl použit Forest Plot. Graf umožňuje zobrazit průměrné hodnoty vybraných metrik spolu s jejich intervaly spolehlivosti a usnadňuje rychlé posouzení variability výsledků.

V následujícím grafu je zobrazen Forest Plot regresního experimentu.



Obr. 39: Forest Plot intervalů spolehlivostí pro regresní experiment

Zdroj: Vlastní zpracování (2026)

Forest Plot přehledně ukazuje průměrné hodnoty metrik regresního experimentu spolu s jejich 95% intervaly spolehlivosti. Graf potvrzuje, že všechny sledované ukazatele zůstávají v rozumném rozmezí i při opakovaném hodnocení. Nízké hodnoty MAE, RMSE a Brier score podporují závěr o poměrně přesném pravděpodobnostním odhadu, zatímco vyšší hodnota ROC AUC znamená, že model si současně uchovává velmi dobrou schopnost odlišovat pozitivní a negativní případy. Hodnota R^2 navíc naznačuje, že model vysvětluje významnou část variability cílové proměnné i při regresním pojetí úlohy.

Význam grafu spočívá především ve vizuálním doplnění tabulkového hodnocení a v možnosti rychle posoudit stabilitu jednotlivých metrik. Poměrně úzké intervaly spolehlivosti potvrzují, že výsledky nejsou výrazně závislé na jednom konkrétním rozdělení dat a že regresní interpretace vykazuje dostatečně konzistentní výkon. Forest Plot podporuje závěr, že převod klasifikační úlohy na pravděpodobnostní odhad rizika nepředstavuje pouze formální experiment, ale metodicky smysluplné rozšíření hlavní analytické linie.

2.6 Diskuse výsledků

Kapitola diskuse výsledků je zaměřena na interpretaci hlavních analytických poznatků v širších souvislostech řešeného problému. Smyslem není pouze zopakovat dříve uvedené výsledky, ale objasnit jejich význam, zhodnotit vhodnost použitých metod a vyvodit závěry důležité pro predikci srdečních onemocnění i pro samotnou metodologii datové vědy. Pozornost je věnována i limitům zvoleného řešení, vztahu mezi notebookem a textem práce a přínosu provedených experimentů.

2.6.1 Význam hlavních zjištění z explorativní analýzy

Explorativní analýza ukázala, že datový soubor obsahuje klinicky smysluplnou a dostatečně bohatou strukturu pro následné modelování. Už jednorozměrná analýza numerických proměnných odhalila, že rozdělení fyziologických a diagnostických ukazatelů není jednotné ani ideálně normální, což odpovídá charakteru zdravotnických dat. Asymetrie, rozdílný rozsah hodnot i výskyt extrémnějších pozorování nepředstavují v daném kontextu nedostatek, ale spíše informaci o heterogenitě pacientské populace. Právě přítomnost různorodých profilů vytváří prostor pro použití metod, které umějí pracovat s více zdroji variability současně.

Zvláštní význam měla analýza kategoriálních a binárních proměnných. Nerovnoměrné zastoupení některých kategorií ukázalo, že klinické znaky nejsou v datasetu rozloženy náhodně, ale že část atributů vystupuje jako výraznější součást typického pacientského profilu. Zmíněný výsledek je důležitý nejen pro popis dat, ale i pro interpretaci následného modelování. Kategorie zastoupené častěji a současně odborně významné mohou nést rozhodující část predikční informace, zatímco málo zastoupené hodnoty vyžadují při interpretaci větší opatrnost.

Dvourozměrná analýza dále ukázala, že vztahy mezi proměnnými existují, avšak většinou nepůsobí izolovaně ani natolik silně, aby bylo možné problém predikce vysvětlit několika jednoduchými pravidly. Právě daný poznatek je metodicky důležitý. Predikce srdečního onemocnění se zde neukazuje jako úloha založená na jednom dominantním ukazateli, ale jako problém vznikající ze souběhu více klinických faktorů. Daný obraz dobře odpovídá medicínské realitě, v níž bývá rizikový stav podmíněn kombinací symptomů, zátěžových ukazatelů a základních fyziologických charakteristik.

2.6.2 Vícerozměrná struktura dat a význam shlukování

Rozšířená explorativní analýza bez učitele potvrdila, že datový soubor nelze chápat pouze jako soubor izolovaných proměnných, ale jako prostor, v němž se jednotlivé klinické znaky kombinují do opakujících se pacientských profilů. Projekce pomocí UMAP ukázala, že pozorování nejsou rozmístěna náhodně a že v datech existuje vnitřní vícerozměrná struktura. Současně nebylo patrné dokonale ostré oddělení všech skupin, což je v případě zdravotnických dat logické. Riziko srdečního onemocnění obvykle netvoří několik absolutně separovaných kategorií, ale spíše kontinuum profilů s částečnými překryvy.

Význam shlukování pomocí K-Prototypes nespočíval pouze v technickém rozdělení dat do několika segmentů. Podstatný přínos vznikl až ve chvíli, kdy byly shluky profilovány prostřednictvím průměrů numerických proměnných, dominantních kategorií a zastoupení cílové proměnné. Zmiňovaný postup umožnil převést anonymní čísla shluků do podoby interpretovatelných pacientských skupin. V metodickém smyslu jde o důležitý krok, protože bez profilování by segmentace zůstala pouze matematickým výsledkem bez odborného obsahu.

Segmentace ukázala, že pacientskou populaci lze rozdělit do několika smysluplných profilů odlišných věkem, zátěžovými charakteristikami, typem bolesti na hrudi nebo zastoupením srdečního onemocnění. Analýza poskytuje doplňkový pohled, který klasifikační modelování samo o sobě nenabízí. Klasifikátor odpovídá na otázku, zda pacient spadá do pozitivní nebo negativní třídy, zatímco shlukování ukazuje, jaké typy pacientů se v datech vůbec vyskytují.

2.6.3 Význam supervised EDA pro interpretaci modelů

Řízená analýza významnosti proměnných představuje jednu z klíčových částí celé práce. Výsledky Mutual Information i SHAP ukázaly vysokou míru shody v identifikaci nejdůležitějších atributů. Mezi klíčové proměnné opakovaně patřily znaky spojené se zátěžovým vyšetřením, změnami ST segmentu, přítomností anginy při zátěži, typem bolesti na hrudi a vybranými fyziologickými ukazateli. Zjištěná konzistence mezi několika analytickými přístupy výrazně zvyšuje důvěryhodnost výsledků.

Z věcného hlediska má uvedené zjištění zásadní význam. Predikční informace není rozptýlena mezi všechny proměnné přibližně rovnoměrně, ale koncentruje se do užší skupiny klinicky velmi dobře interpretovatelných znaků. Modely proto nestaví na obtížně vysvětlitelných kombinacích náhodných atributů, ale na znacích, které odpovídají i klinické intuici. Zjištěný soulad mezi statistickým výsledkem a odborným významem proměnných je v oblasti zdravotnické analýzy velmi důležitý, protože podporuje důvěru v budoucí použitelnost modelu.

Za vlastní přínos práce lze považovat i propojení supervised EDA s předchozí explorativní analýzou. Významné proměnné se neobjevily až v samotném modelování, ale postupně vystupovaly již v jednorozměrné, dvourozměrné i vícerozměrné části. Návaznost mezi jednotlivými analytickými kroky podporuje závěr, že význam atributů nevyplynul až z obtížně interpretovatelného modelového výstupu, ale byl patrný již v předchozích fázích analýzy.

2.6.4 Diskuse klasifikačního modelování

Klasifikační modelování potvrdilo, že analyzovaný datový soubor obsahuje dostatečně silný predikční signál pro úspěšné rozlišování pacientů s přítomným a nepřítomným srdečním onemocněním. Souhrnná tabulka metrik, ROC křivky, Precision–Recall křivky i statistické hodnocení pomocí intervalů spolehlivosti ukázaly, že více modelů dosahuje velmi dobrých výsledků. Významný poznatek spočívá ve skutečnosti, že nejlepší výkonnost nebyla soustředěna pouze do jednoho izolovaného algoritmu. Silný výkon vykázaly SVM, KNN, logistická regrese i náhodný les, což naznačuje, že rozhodující informace o riziku onemocnění je v datech skutečně přítomna a není vázána výhradně na jeden konkrétní modelový princip.

Přesto lze mezi modely rozlišit několik metodicky důležitých rozdílů. SVM se ukázalo jako velmi silný kandidát z hlediska globální diskriminační schopnosti, což odpovídá charakteru dat, v nichž hranice mezi třídami nejsou čistě lineární, ale zároveň nejsou natolik nepravidelné, aby vyžadovaly extrémně komplexní model. KNN dosáhlo velmi dobrých výsledků při konkrétním rozhodovacím prahu, což ukazuje, že lokální podobnost pacientských profilů nese významnou rozlišovací informaci. Náhodný les potvrdil schopnost zachytit kombinace více proměnných bez nutnosti silných předpokladů o tvaru vztahů. Logistická regrese pak prokázala, že i relativně jednoduchý a dobře interpretovatelný model dokáže na daných datech dosáhnout velmi konkurenceschopného výkonu.

Právě výsledek logistické regrese je z metodického hlediska mimořádně zajímavý. Přítomnost nelinearit a složitějších vazeb v datech sice naznačuje výhodu flexibilnějších metod, avšak výkon logistické regrese ukazuje, že podstatná část predikční informace je zachytitelná i lineárněji pojatým přístupem. V prostředí bakalářské práce zaměřené nejen na výkon, ale i na

srozumitelnost a interpretovatelnost, jde o velmi cenný závěr. Logistická regrese představuje rozumný kompromis mezi vysvětlitelností a predikční silou.

Rozhodovací strom naopak vykázal slabší výkon než ostatní modely. Výsledek lze interpretovat jako důsledek omezené schopnosti jednoduché stromové struktury zachytit jemnější kombinace znaků v prostředí s částečně se překrývajícími pacientskými profily. Slabší výsledek stromu současně potvrzuje, že problém predikce srdečního onemocnění nelze v analyzovaném souboru redukovat na několik málo jednoduchých pravidel. Odpověď na klinickou otázku zjevně vyžaduje složitější kombinaci atributů.

2.6.5 Význam statistického hodnocení modelů

Za důležitý přínos práce lze považovat i doplnění bodových metrik o intervaly spolehlivosti a Forest Plot. Zařazení opakované validace a intervalů spolehlivosti umožnilo posoudit, zda se vysoká výkonnost modelů potvrzuje opakovaně, nebo zda jde pouze o důsledek příznivého jednorázového rozdělení dat.

Výsledky ukázaly, že nejlepší modely nevykazují pouze vysoké průměrné hodnoty metrik, ale i relativně úzké intervaly spolehlivosti. Zásadní zjištění spočívá v překryvu intervalů mezi nejlepšími modely. Zmíněný výsledek brání příliš zjednodušenému závěru, že jediný model má jednoznačně dominantní postavení. Rozhodování o vhodnosti modelu proto nemá stát pouze na nejvyšší hodnotě jedné metriky, ale i na stabilitě, interpretovatelnosti a praktické použitelnosti. Z akademického hlediska jde o podstatný metodický moment, protože posouvá práci od prostého „žebříčku algoritmů“ k odpovědnějšímu hodnocení modelové kvality.

2.6.6 Diskuse experimentů

Experiment se změnou velikosti trénovací a testovací množiny ukázal, že výkon modelu se s rostoucím objemem trénovacích dat zlepšuje. Daný výsledek je očekávatelný, avšak analyticky cenný, protože potvrzuje, že model skutečně umí využít dodatečnou informaci přítomnou ve větší trénovací množině. Výsledek současně naznačuje, že další rozšíření datasetu by mohlo přinést ještě mírné zlepšení generalizační schopnosti. V širším kontextu jde o důležitý argument proti interpretaci dosažených výsledků jako definitivního maxima výkonu.

Experiment s chybějícími daty ukázal relativní odolnost modelování vůči neúplnosti vstupních údajů. Část metrik zůstala téměř nezměněna, zatímco jiné vykázaly mírné zhoršení. Dedukce z výsledku je metodicky důležitá: neúplnost dat nemusí okamžitě vést k selhání modelu, avšak může oslabit jemnější diskriminační výkon. V praktickém prostředí zdravotnictví, kde kompletnost záznamů nebývá samozřejmostí, jde o významné zjištění. Analýza získává vyšší realističnost a není omezena pouze na ideálně vyčištěný laboratorní scénář.

Regresní experiment rozšířil hlavní klasifikační úlohu o pravděpodobnostní interpretaci rizika. Výsledek ukázal, že binární problém lze chápat nejen jako rozhodnutí mezi třídami, ale i jako odhad míry rizika srdečního onemocnění. Hodnoty MAE, RMSE, Brier score a ROC AUC potvrdily, že model poskytuje použitelný spojitý výstup bez ztráty diskriminační schopnosti. Kalibrační křivka současně ukázala, že kvalita pravděpodobnostního odhadu není ve všech pásmech intervalu stejná, což je důležité pro odpovědnou interpretaci výsledků. Praktický přínos

regresního experimentu spočívá v tom, že umožňuje přejít od čistě kategorického rozhodnutí k jemnějšímu vyjádření rizika, které může být v medicínské praxi přínosnější.

2.6.7 Přínosy práce ve vztahu k tématu

Komplexní propojení více fází datové vědy nad jedním medicínsky interpretovatelným problémem představuje jednu z nejvýraznějších hodnot celé práce. Praktická část nezůstala u porovnání několika klasifikátorů, ale spojila explorativní analýzu, shlukování, řízené hodnocení významnosti proměnných, klasifikaci, pravděpodobnostní experimenty a interaktivní prezentaci výstupů. Vystavěný analytický rámec poskytuje hlubší pohled na problematiku predikce srdečních onemocnění než samotné porovnání výkonu algoritmů.

Významnou roli hraje rovněž důsledný důraz na interpretaci. Význam proměnných nebyl odvozován pouze z výkonu finálního modelu, ale z návaznosti mezi EDA, supervised EDA a klasifikačním modelováním. Práce tím ukazuje, že nejdůležitější znaky lze identifikovat konzistentně napříč více metodami. Zjištěný výsledek podporuje nejen důvěryhodnost analytického postupu, ale i jeho potenciální použitelnost v prostředí, kde je vysvětlitelnost modelu zásadní podmínkou přijetí modelového řešení.

Pozornost byla věnována i podobě samotného notebooku a způsobu prezentace analytických výstupů. Vedle kódu byla systematicky budována interpretace výsledků, komentáře ke grafům a přehledná struktura jednotlivých kroků. Praktická část proto neslouží pouze jako interní experimentální prostředí, ale i jako komunikační nástroj převádějící datově analytické výstupy do podoby srozumitelné uživateli bez hlubší specializace v oblasti strojového učení.

2.6.8 Omezení a metodické poznámky

Navzdory dosaženým výsledkům je nutné zohlednit několik omezení. Datový soubor obsahuje omezený počet proměnných a nepostihuje všechny klinické, laboratorní ani zobrazovací charakteristiky, které mohou v reálném prostředí ovlivňovat riziko srdečního onemocnění. Výsledky proto nelze chápat jako univerzálně přenosné na všechny klinické populace bez další validace.

Další omezení souvisí s charakterem použitých dat. Analýza vychází z tabulkového datasetu, který umožňuje přehlednou demonstraci metod datové vědy, avšak neobsahuje časový rozměr, detailní průběh klinických měření ani komplexnější diagnostické informace. Predikce proto pracuje s redukovaným obrazem pacienta. V odborné praxi by bylo vhodné propojit podobný přístup s rozsáhlejšími zdroji dat, případně s multimodálními informacemi.

Metodické omezení se týká i interpretace regresního experimentu. Převod binární klasifikace na pravděpodobnostní odhad rizika je analyticky užitečný, avšak neznamená přímý klinický odhad budoucí události v epidemiologickém smyslu. Pravděpodobnost vyjadřuje modelový odhad příslušnosti k pozitivní třídě v rámci použitého datasetu. Při praktickém využití je proto nutné odlišovat mezi statistickým skóre a skutečným klinickým rozhodnutím.

2.6.9 Vztah výsledků k literatuře

Dosažené Dosažené výsledky jsou v obecné rovině v souladu s literaturou zaměřenou na využití strojového učení v kardiologii. Krittanawong a kol. (2020) uvádějí, že metody strojového učení mohou v oblasti kardiovaskulární predikce dosahovat velmi dobrého výkonu a že mezi silné kandidáty často patří SVM, ansámblové metody a další flexibilní modely. Podobný obraz přinesla i realizovaná analýza, v níž velmi dobrých výsledků dosáhly SVM, náhodný les, KNN i logistická regrese. Shoda s literaturou je patrná také při porovnání s prakticky orientovanými projekty popsanými v rešeršní části, v nichž se mezi úspěšnými modely opakovaně objevují K-Nearest Neighbors, Random Forest, Support Vector Machine nebo logistická regrese (Doughty, 2025; Paul, 2020; Duru, 2022). Obdobný závěr přináší také studie Garavanda a kol. (2022), která při diagnostice koronárního onemocnění tepen potvrzuje velmi dobrý výkon metod SVM a Random Forest.

Shoda s literaturou se neprojevuje pouze na úrovni výkonu modelů, ale i na úrovni významných atributů. V předložené práci hrály klíčovou roli proměnné související se zátěžovým vyšetřením, změnami ST segmentu, anginou při zátěži a charakterem bolesti na hrudi. Podobné okruhy proměnných se objevují i v analyzovaných projektech a odborných studiích, v nichž mezi důležité prediktory patří symptomy, EKG ukazatele, srdeční frekvence a další základní klinické charakteristiky pacientů (Doughty, 2025; Duru, 2022). Výsledky podporují závěr, že největší predikční význam nemají náhodné nebo obtížně interpretovatelné atributy, ale klinicky dobře obhajitelné znaky spojené s ischemickou zátěží a funkční odpovědí srdce. Současně se potvrzuje význam interpretovatelnosti modelů, která bývá v medicínském prostředí považována za důležitou podmínku důvěry ve výsledné rozhodování modelu (Ribeiro a kol., 2016; Krittanawong a kol., 2019).

Vedle shody s literaturou je však nutné objasnit i rozdíly. Některé publikace nebo projekty uvádějí vyšší výsledné metriky, než jakých bylo dosaženo v předložené práci, což však nemusí automaticky znamenat slabší kvalitu navrženého řešení. První vysvětlení souvisí s odlišností datových souborů. Rozsah datasetu, složení vstupních proměnných, zastoupení tříd i kvalita předzpracování mohou výsledné metriky výrazně ovlivnit. Projekt Paula (2020) například pracuje s rozsáhlejšími dlouhodobými epidemiologickými daty a využívá také ansámblové techniky včetně stackingu, což může vést k vyšší přesnosti, avšak za cenu nižší srozumitelnosti modelu. Naproti tomu předložená práce staví na vyvážení výkonu, interpretovatelnosti a metodické přehlednosti.

Druhý zdroj rozdílů souvisí se způsobem evaluace. Část běžně dostupných projektů staví závěry především na jediné testovací množině a na omezeném souboru metrik. V předložené práci bylo hodnocení rozšířeno o ROC AUC, PR AUC, MCC, learning curve, intervaly spolehlivosti a Forest Plot. Přísnější a víceúrovňová evaluace může vést k uměřenějšímu, ale metodicky spolehlivějšímu obrazu výkonu modelů. Vyšší číslo uvedené v jiné práci proto nemusí samo o sobě znamenat lepší nebo robustnější model, pokud není doplněno o stabilitu výsledků a širší interpretaci.

Třetí vysvětlení vyplývá ze samotného cíle práce. Řada praktických projektů usiluje především o maximalizaci jedné nebo několika výkonových metrik. V předložené práci byl kladen důraz rovněž na porozumění datům, profilování pacientských skupin, vysvětlení významnosti proměnných a pravděpodobnostní interpretaci výsledků. Zařazení UMAP, K-Prototypes, Mutual

Information, SHAP, kalibrační křivky nebo regresního experimentu sice nerozšiřuje výkon klasifikátorů přímo, ale významně rozšiřuje interpretační hodnotu celé analýzy. Rozdíl oproti části publikovaných nebo veřejně sdílených řešení proto nespočívá pouze v dosažených metrikách, ale i v odlišném analytickém záměru.

Významný rozdíl oproti části běžně dostupných projektů spočívá rovněž v metodickém rozsahu celé práce. Řada praktických řešení zůstává převážně u předzpracování dat, natrénování několika klasifikátorů a porovnání základních metrik výkonu. Předložená práce vedle klasifikační úlohy zahrnuje rovněž vícerozměrnou explorativní analýzu bez učitele, profilování shluků, supervised EDA, statistické hodnocení pomocí intervalů spolehlivosti a Forest Plotu i regresní interpretaci binární cílové proměnné. Větší důraz je kladen také na propojení výkonu, stability a interpretovatelnosti výsledků. V uvedeném rozšíření analytického rámce lze spatřovat jeden z hlavních přínosů práce vzhledem k tématu i vzhledem k porovnávaným projektům.

Závěr

Bakalářská práce byla zaměřena na využití metod datové vědy při predikci srdečních onemocnění na základě klinických a diagnostických údajů pacientů. Hlavní přínos spočívá v propojení více fází datové vědy v rámci jednoho uceleného analytického řešení, které zahrnuje přípravu a předzpracování dat, explorativní analýzu, vícerozměrnou analýzu bez učitele, řízené hodnocení významnosti proměnných, klasifikační modelování, doplňkové experimenty i interaktivní prvky notebooku.

Explorativní analýza ukázala, že datový soubor obsahuje klinicky smysluplné vztahy a přirozenou heterogenitu pacientské populace. Významnou roli hrály především proměnné související se zátěžovým vyšetřením, změnami ST segmentu, přítomností anginy při zátěži a charakterem bolesti na hrudi. Vícerozměrná explorativní analýza bez učitele dále potvrdila existenci rozpoznatelných pacientských profilů, které bylo možné segmentovat a interpretovat pomocí shlukování. Právě propojení deskriptivního, segmentačního a predikčního pohledu představuje důležitou hodnotu práce, protože umožňuje nahlížet na problém nejen jako na čistě klasifikační úlohu, ale i jako na analýzu vnitřní struktury pacientských dat.

Za významnou součást práce lze považovat i návrh notebooku určeného pro čtenáře orientovaného na řešený problém, nikoli pouze na technickou stránku datové analytiky. Vedle samotného výpočtu byla systematicky budována textová interpretace výstupů, přehledná struktura jednotlivých částí a možnost interakce s daty a modelem. Praktická část plní nejen roli experimentálního prostředí, ale i roli komunikačního nástroje, který převádí analytické výstupy do podoby srozumitelné uživateli bez hlubší specializace v oblasti strojového učení.

Stanovený cíl práce byl naplněn. Byly aplikovány hlavní fáze životního cyklu datové vědy na konkrétní zdravotnický problém a vzniklo ucelené řešení umožňující analyzovat datový soubor, interpretovat jeho strukturu a vytvořit modely pro predikci srdečního onemocnění. Praktická část prokázala, že použitý dataset obsahuje dostatečně silný predikční signál, který lze využít několika modelovými přístupy. Současně bylo ověřeno, že klasifikační úlohu lze doplnit o pravděpodobnostní pohled prostřednictvím regresního experimentu a že navržený postup je dostatečně stabilní i při změně některých podmínek. Naplnění cílů je podpořeno i šíří realizovaných analytických kroků. Práce nezůstala pouze u porovnání několika algoritmů, ale zahrnula rovněž shlukování pacientských profilů, řízené hodnocení významnosti příznaků, statistické hodnocení výkonnosti modelů pomocí intervalů spolehlivosti a experimenty zaměřené na změnu velikosti trénovací množiny, práci s chybějícími daty a regresní interpretaci.

Vedle dosažených výsledků je však nutné zohlednit i omezení práce. Použitý datový soubor obsahuje omezený počet atributů a nepokrývá všechny možné faktory ovlivňující riziko srdečního onemocnění. Analýza pracuje s tabulkovými klinickými daty a nezahrnuje například časový průběh stavu pacienta, detailnější laboratorní markery, zobrazovací nálezy ani širší kontext zdravotnické dokumentace. Výsledky proto nelze bez další validace přímo zobecnit na jiné populace nebo klinická prostředí.

Omezení se týká i samotných modelových výstupů. Predikce pravděpodobnosti v regresním experimentu představuje modelový odhad v rámci dostupného datasetu, nikoli přímou prognózu budoucí klinické události. Přestože výkonnost modelů dosáhla velmi dobrých hodnot, interpretace musí zůstat opatrná a respektovat skutečnost, že práce byla provedena nad jedním

konkrétním veřejným souborem dat. Omezením je rovněž prostředí notebooku, v němž se interaktivita částečně liší podle způsobu spuštění a platformy.

Další výzkum by mohl směřovat k ověření navrženého postupu na rozsáhlejších a různorodějších datech. Přínosné by bylo zejména použití datasetů obsahujících širší spektrum klinických proměnných, laboratorních ukazatelů, případně časových nebo obrazových dat. Zmíněné rozšíření by umožnilo přesněji posoudit, zda se zjištěné vztahy a významnost jednotlivých atributů potvrzují i v komplexnějším datovém prostředí.

Jiný směr rozvoje představuje rozšíření modelové části o další algoritmy, například gradient boosting metody nebo neuronové sítě, a jejich porovnání s použitými modely. Přínosné by bylo i hlubší rozpracování kalibrace pravděpodobnostních odhadů, externí validace na nezávislém souboru a detailnější analýza robustnosti vůči neúplnosti nebo změně kvality vstupních dat. Z metodického hlediska by bylo zajímavé ověřit, do jaké míry lze propojit klasifikační a pravděpodobnostní pohled na problém s dalšími formami interpretace modelového rozhodování.

Smysluplný rozvoj nabízí i samotná podoba notebooku, jehož další úpravy by mohly směřovat k stabilnějšímu uživatelskému rozhraní, širší podpoře vlastních vstupních dat a přímějšímu využití v prostředí určeném pro doménového experta. Dlouhodobější perspektiva práce proto nespočívá pouze v dalším rozšiřování použitých metod, ale i v postupném přibližování analytického řešení praktickému využití. Právě propojení predikční výkonnosti, interpretovatelnosti a srozumitelné prezentace výsledků představuje hlavní předpoklad, aby podobně koncipované datově analytické nástroje mohly v budoucnu sloužit jako užitečná podpora při hodnocení kardiovaskulárního rizika.

Seznam použité literatury

AMBALE-VENKATESH, B., Xiaoming YANG, Chao WU, Kiang LIU a kol. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circulation Research. 2017, 121(9), s. 1092–1101. [online]. [cit. 2025-11-19]. Dostupné z: https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.117.311312

COLLINS, Gary S., Johannes B. REITSMA, Douglas G. ALTMAN a Karel G. M. MOONS. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Annals of Internal Medicine. 2015, 162(10), s. 735–736. [online]. [cit. 2026-04-09]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/25984857/

DHAR, Vasant. Data Science and Prediction. Communications of the ACM. 2013, 56(12), s. 64-73. [online]. [cit. 2025-11-19]. Dostupné z: https://dl.acm.org/doi/10.1145/2500499

DOUGHTY, Jack. Heart Disease Detection Using Machine Learning. GitHub [online]. [cit. 2026-04-09]. Dostupné z: https://github.com/jvckdough/Heart-Disease-Detection-Using-ML

DURU, Ilhan Cem. Heart Failure Prediction with Machine Learning. GitHub [online]. [cit. 2026-04-09]. Dostupné z: https://github.com/icemduru/Heart-Failure-Prediction-with-Machine-Learning

FAWCETT, Tom. An introduction to ROC analysis. Pattern Recognition Letters. 2006, 27(8), s. 861–874. [online]. [cit. 2026-04-09]. Dostupné z: https://doi.org/10.1016/j.patrec.2005.10.010

GARAVAND, Ali, Cirruse SALEHNASAB, Somayeh BARKHORDARI, Mahboobeh NIKPAY a Nazanin KHODAEI. Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms. Journal of Healthcare Engineering. 2022, 2022, 5359540. [online]. [cit. 2026-04-09]. Dostupné z: https://pmc.ncbi.nlm.nih.gov/articles/PMC9596250/

HUANG, Zhexue. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery. 1998, 2, s. 283–304. [online]. [cit. 2026-04-09]. Dostupné z: https://link.springer.com/article/10.1023/A:1009769707641

CHICCO, Davide a Giuseppe JURMAN. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020, 21, 6. [online]. [cit. 2026-04-09]. Dostupné z: https://pmc.ncbi.nlm.nih.gov/articles/PMC6941312/

KRITTANAWONG, Charat, Haipeng ZHANG, Zhen WANG a kol. Artificial intelligence in cardiovascular medicine. Scientific Reports. 2020, 10, 16057. [online]. [cit. 2025-11-19]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/32994452/

KRITTANAWONG, Charat, Kaitlyn W. JOHNSON, Robert S. ROSENSON a kol. Deep learning for cardiovascular medicine: a practical primer. European Heart Journal. 2019, 40(25), s. 2058–2073. [online]. [cit. 2025-11-19]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/30815669/

LI, Xuewen, Chengming SHANG, Changyan XU, Yiting WANG, Jiancheng XU, Yaxin LI a Qi ZHOU. Development and comparison of machine learning-based models for predicting heart

failure after acute myocardial infarction. BMC Medical Informatics and Decision Making. 2023, 23, 223. [online]. [cit. 2026-04-09]. Dostupné z: https://link.springer.com/article/10.1186/s12911-023-02240-1

LUNDBERG, Scott M. a Su-In LEE. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017. [online]. [cit. 2026-04-09]. Dostupné z: https://proceedings.neurips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

MATTHEWS, Brian W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta. 1975, 405(2), s. 442–451. [online]. [cit. 2026-04-09]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/1180967/

McINNES, Leland, John HEALY a James MELVILLE. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. 2018. [online]. [cit. 2026-04-09]. Dostupné z: https://arxiv.org/abs/1802.03426

O'NEIL, Cathy a Rachel SCHUTT. Doing Data Science. Sebastopol: O'Reilly Media, 2013. ISBN 978-1-4493-5865-9.

PAUL, Indrashis. Coronary Heart Disease Prediction. GitHub [online]. [cit. 2026-04-09]. Dostupné z: https://github.com/indrapaul824/Coronary-Heart-Disease-Prediction

PROVOST, Foster a Tom FAWCETT. Data Science for Business. Sebastopol: O'Reilly Media, 2013. ISBN 978-1-4493-6136-9.

QIAN, Xiang, Shuxia LUO, Yajie WANG, Li YANG, Ying WANG, Xinyi TANG, Lili LI, Wenqiang LIU, Gulibahaer YISIHAKE, Ainiwaer AILI, Xinquan ZHAO a Yining YANG. A Cardiovascular Disease Prediction Model Based on Routine Physical Examination Indicators Using Machine Learning Methods: A Cohort Study From Southern Xinjiang. Frontiers in Cardiovascular Medicine. 2022, 9, 898359. [online]. [cit. 2025-12-17]. Dostupné z: https://pmc.ncbi.nlm.nih.gov/articles/PMC9247206/

RIBEIRO, Marco Tulio, Sameer SINGH a Carlos GUESTRIN. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD '16, San Francisco. New York: ACM, 2016, s. 1135–1144. [online]. [cit. 2025-11-19]. Dostupné z: https://dl.acm.org/doi/10.1145/2939672.2939778

SAITO, Takaya a Marc REHMSMEIER. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 2015, 10(3), e0118432. [online]. [cit. 2026-04-09]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/25738806/

SAJID, Mirza Rizwan, Bader A. ALMEHMADI, Waqas SAMI, Mansour K. ALZAHRANI, Noryanti MUHAMMAD, Christophe CHESNEAU, Asif HANIF, Arshad Ali KHAN a Ahmad SHAHBAZ. Development of Nonlaboratory-Based Risk Prediction Models for Cardiovascular Diseases Using Conventional and Machine Learning Approaches. International Journal of Environmental Research and Public Health. 2021, 18(23), 12586. [online]. [cit. 2025-12-17]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/34886312/

STEYERBERG, Ewout W., Andrew J. VICKERS, Nancy R. COOK, Thomas GERDS, Mithat GONEN, Nancy OBuchowski, Michael J. PENCINA a Michael W. KATTAN. Assessing the performance

of prediction models: a framework for traditional and novel measures. Epidemiology. 2010, 21(1), s. 128–138. [online]. [cit. 2026-04-09]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/20010215/

TOHYAMA, Takeshi, Tomomi IDE, Masataka IKEDA, Hidetaka KAKU, Nobuyuki ENZAN, Shouji MATSUSHIMA, Kouta FUNAKOSHI, Junji KISHIMOTO a Koji TODAKA. Machine learning-based model for predicting 1 year mortality of hospitalized patients with heart failure. ESC Heart Failure. 2021, 8(5), s. 4077–4085. [online]. [cit. 2025-12-17]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/34390311/

TOPOL, Eric. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York: Basic Books, 2019. ISBN 978-1-5416-4470-6.

VAN CALSTER, Ben, Dirk NIEBOER, Yvonne VERGOUWE, Bavo DE COCK, Michael J. PENCINA a Ewout W. STEYERBERG. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology. 2016, 74, s. 167–176. [online]. [cit. 2026-04-09]. Dostupné z: https://pubmed.ncbi.nlm.nih.gov/26772608/

VERGARA, Jorge R. a Pablo A. ESTÉVEZ. A review of feature selection methods based on mutual information. Neural Computing and Applications. 2014, 24(1), s. 175–186. [online]. [cit. 2026-04-09]. Dostupné z: https://link.springer.com/article/10.1007/s00521-013-1368-0

WORLD HEALTH ORGANIZATION. Cardiovascular diseases (CVDs). WHO Fact Sheet, 2024. [online]. [cit. 2025-11-19]. Dostupné z: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)