Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra statistiky

Diplomová práce

Datová analýza Covid-19

Matěj Bílek

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Bc. Matěj Bílek

Informatika

Název práce

Datová analýza Covid-19

Název anglicky

Covid-19 data analysis

Cíle práce

Cílem diplomové práce je vytvořit prediktivní model vývoje nemoci Covid-19 pomocí metod strojového učení v jazyce Python.
Diplomová práce má mimo jiné odpovědět na několik otázek: Jak se nemoc šířila ve světě? Lišilo se šíření nemoci podle vyspělosti zemí, přijatých opatření a očkování? Jak Covid ovlivnil ekonomiku? Jaké výsledky ukazují jednotlivé statistické metody a jaké závěry z nich můžeme usuzovat? Jaký se očekává vývoj do budoucna?

Metodika

Teoretická část:

Definice Covid-19, význam sledování statistik vývoje nemoci.

Analýza empirických studií, které se zaměřily na datovou analýzu vývoje nemoci.

Charakteristika programovacího jazyka Python.

Charakteristika pojmu strojové učení (machine learning).

Praktická část:

Výběr reprezentativního datasetu pro šíření Covid-19 ve světě.

Předběžné zpracování vybraných dat pro další použití.

Deskriptivní analýza: "kde a jak se nemoc šířila v čase", trendy v šíření Covid-19 ve světe.

Použití metod strojového učení k tvorbě predikčního modelu v jazyku Python.

Součástí tvorby prediktivního modelu bude popis fungování použitých metod a rozbory kódu v jazyku Python, který vede k porovnání modelů.

Doporučený rozsah práce

60-80 stran bez příloh

Klíčová slova

strojové učení, predikce, Python, modelování

Doporučené zdroje informací

DASGUPTA, Nataraj. *Practical big data analytics : hands-on techniques to implement enterprise analytics and machine learning using hadoop, spark, NoSQL and R.* Birmingham: Packt, 2018. ISBN 978-1783554393.

HAYKIN, Simon S. *Neural networks and learning machines.* Upper Saddle River: Prentice Hall, 2009. ISBN 978-0-13-129376-2.

K, Hemachandran.; KHANRA, Sayantan.; RODRIGUEZ, Raul V.; JARAMILLO, Juan. *Machine Learning for Business Analytics : Real-Time Data Analysis for Decision-Making. [elektronický zdroj] /.* Milton: Productivity Press, 2022. ISBN 9781000615425.

LIU, Yuxi (Hayden). *Python Machine Learning by Example : Build Intelligent Systems Using Python, TensorFlow 2, Pytorch, and Scikit-Learn. [elektronický zdroj] /.* Birmingham: Packt Publishing, Limited, 2020. ISBN 9781800203860.

QUDDUS, Jillur. *Machine learning with apache spark quick start guide : uncover patterns, derive actionable insights, and learn from big data using MLlib.* Birmingham: Packt, 2018. ISBN 978-1789346565.

UNPINGCO, José. *Python for probability, statistics, and machine learning.* New York, NY: Springer Berlin Heidelberg, 2016. ISBN 9783319307152.

WITTEN, I. H.; FRANK, Eibe. *Data mining : practical machine learning tools and techniques.* San Francisco, Calif.: Elsevier Science [distributor], 2005. ISBN 978-0-12-088407-0.

Předběžný termín obhajoby

2024/25 LS – PEF

Vedoucí práce

prof. Ing. Jindřich Špička, Ph.D.

Garantující pracoviště

Katedra statistiky

Elektronicky schváleno dne 16. 06. 2024

doc. Ing. Tomáš Hlavsa, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 04. 11. 2024

doc. Ing. Tomáš Šubrt, Ph.D.

Děkan

V Praze dne 10. 01. 2025

Čestné prohlášení

Prohlašuji, že svou diplomovou práci "Datová analýza Covid-19" jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autor uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 31.3.2025 _____

Poděkování

Rád bych touto cestou poděkoval prof. Ing. Jindřichu Špičkovi, Ph.D. za konzultaci a odborné vedení této diplomové práce, cenné rady a trpělivost.

Datová analýza Covid-19

Abstrakt

Diplomová práce se zabývá tvorbou prediktivních modelů časových řad šíření onemocnění Covid-19. Hlavním cílem práce je vytvořit prediktivní modely zachycující vývoj šíření onemocnění v roce 2025 pomocí metod strojového učení v programovacím jazyce Python. V teoretické části jsou popsána klíčová témata související s datovou analýzou a modelováním, zahrnující charakteristiku onemocnění Covid-19, význam sledování statistik vývoje této nemoci, programovací jazyk Python a metody strojového učení.

Na základě teoretických poznatků je v praktické části vybrán reprezentativní datový soubor, na který jsou aplikovány tři odlišné prediktivní modely v jazyce Python: statistický model Prophet určený pro práci s časovými řadami, model XGBoost využívající rozhodovací stromy v kombinaci s algoritmem gradientního boostingu a model hlubokého učení LSTM založený na rekurentních neuronových sítích. Součástí tvorby modelu je popis důležitých matematických principů, rozbor implementace a vizualizace výsledků predikcí počtu nových případů Covid-19 pro rok 2025. V závěru práce jsou výsledky jednotlivých modelů porovnány a vyhodnoceny, včetně diskuse o jejich predikčních schopnostech, silných a slabých stránkách, limitech a případných překážkách v procesu modelování.

Klíčová slova: strojové učení, prediktivní modely, Python, vizualizace dat, Covid-19, datová analýza, neuronové sítě, časové řady, Python knihovny, hluboké učení

Covid-19 Data Analysis

Abstract

Diploma thesis is focused on developing predictive models for time series forecasting of the spread of Covid-19. The main objective of the thesis is to create predictive models capturing the progression of disease spread in 2025 using machine learning methods in the Python programming language. The theoretical section describes essential topics related to data analysis and modeling, including the characteristics of Covid-19, the significance of monitoring disease progression statistics, the Python programming language, and machine learning methods.

Based on the theoretical insights, a representative dataset is selected in the practical section, and three distinct predictive models are applied in Python: the statistical model Prophet for time series analysis, the XGBoost model using decision trees combined with gradient boosting algorithms, and the LSTM deep learning model based on recurrent neural networks. The creation of each model includes a description of essential mathematical principles, analysis of implementation, and visualization of predictions for the number of new Covid-19 cases for year 2025. Finally, the thesis concludes by comparing and evaluating the results of individual models, including a discussion of their predictive capabilities, strengths and weaknesses, limitations, and potential obstacles encountered during the modeling process.

Keywords: Machine Learning, prediction models, Python, data visualization, Covid-19, data analysis, neural networks, time series, Python libraries, Deep Learning

Obsah

1 Úvod

Onemocnění Covid-19 představuje celosvětový fenomén, který od svého propuknutí na přelomu let 2019 a 2020 zásadním způsobem ovlivnil a dodnes ovlivňuje životy lidí po celém světě. Pandemie zasáhla společnost na všech úrovních – ochromila sociální interakce, narušila pracovní procesy a na řadu měsíců výrazně změnila fungování školství.

Šíření onemocnění a snahy o jeho potlačení se staly zatěžkávací zkouškou, která odhalila míru připravenosti lidstva na globální krize tohoto typu a zároveň vyzdvihla důležitost některých profesí. Nejviditelnějšími představiteli byli lékaři a zdravotnický personál, kteří se v první linii boje s pandemií denně vystavovali riziku nákazy. Neméně důležitým oborem byla epidemiologie, jež při analýze šíření viru využívala podpůrné statistické modely, datovou analýzu a různé predikční nástroje umožňující odhadovat budoucí vývoj pandemie.

Tato diplomová práce se zabývá tvorbou predikčních modelů vývoje šíření nemoci Covid-19 pomocí metod strojového učení realizovaných v programovacím jazyce Python. V teoretické části je popsána charakteristika onemocnění Covid-19, jeho dosavadní vývoj, význam sledování statistik s ním spojených, a jsou analyzovány empirické studie zaměřené na tuto oblast. Dále je představen programovací jazyk Python, který byl zvolen k tvorbě modelů, a také principy strojového učení umožňující uskutečnění predikcí vývoje pandemie.

Praktická část práce je věnována celému procesu modelování. Nejprve je vybrán reprezentativní datový soubor zachycující průběh šíření viru, který je následně statisticky zpracován a připraven k dalším analýzám. Je také provedena deskriptivní analýza popisující trendy ve vývoji nemoci. Poté následuje samotná tvorba prediktivních modelů pomocí metod strojového učení, doplněná o detailní popis použitých technik, implementaci v Pythonu a vyhodnocení výsledků.

Volba tématu této diplomové práce vychází zejména z aktuálnosti problematiky zkoumání onemocnění Covid-19 a jeho vlivů na společnost, dále také ze stále rostoucího významu strojového učení, které dnes tvoří nedílný základ všech současných AI systémů. Důležitou roli při výběru sehrála také popularita a rozšířenost programovacího jazyka Python. Vzdělávání v těchto oblastech je přínosné pro každého, kdo chce držet krok s moderními metodami datové vědy.

2 Cíl práce a metodika

2.1 Cíl práce

Diplomová práce se podrobně věnuje problematice strojového učení a metod využívaných pro tvorbu prediktivních modelů. Hlavním cílem práce je vytvořit prediktivní modely zachycující vývoj šíření nemoci Covid-19 v roce 2025 pomocí metod strojového učení v programovacím jazyce Python. Součástí cíle je i vzájemné porovnání a interpretace výsledků modelů pro ilustraci odlišných přístupů k predikci časových řad a odhadu budoucího šíření viru tohoto onemocnění.

Dílčím cílem je provedení deskriptivní analýzy na vybraném setu dat a zmapování dosavadního šíření nemoci v celosvětovém měřítku. V analýze je posouzen vliv faktorů, jako jsou ekonomická vyspělost jednotlivých zemí, přijatá epidemiologická opatření a míra proočkovanosti populace, na dynamiku šíření této nemoci. Významnou součástí je také zhodnocení ekonomických dopadů pandemie Covid-19.

2.2 Metodika

Metodika diplomové práce je založena na shromáždění, studiu a analýze odborných informačních zdrojů. V teoretické části jsou podrobně popsány klíčové oblasti související s tvorbou prediktivních modelů. Nejprve je definována nemoc Covid-19 a zdůvodněn význam sledování statistik jejího vývoje, které jsou předmětem následného modelování. Dále následuje analýza empirických studií, jež se věnují zkoumání a vyhodnocování průběhu pandemie. Pro lepší názornost a pochopení pozdější praktické implementace je charakterizován programovací jazyk Python společně s jeho knihovnami a moduly nezbytnými pro tvorbu prediktivních modelů. V neposlední řadě je definován a vysvětlen pojem strojového učení (Machine Learning), které hraje klíčovou roli v procesu datového modelování.

Na základě syntézy získaných teoretických poznatků je v praktické části práce vybrán reprezentativní datový soubor zachycující všechny aspekty šíření Covid-19 ve světě. Tento set dat je dále zpracován do podoby vhodné pro tvorbu prediktivních modelů. Nad daty je provedena deskriptivní analýza, která popisuje vývoj šíření nemoci v čase a identifikuje pozorovatelné trendy v datech pomocí grafových a tabulkových výstupů zdrojového kódu.

V návaznosti na analýzu následuje samotná tvorba prediktivních modelů, jejichž součástí je i detailní popis principů použitých metod a podrobný rozbor zdrojového kódu v jazyce Python. V rámci práce jsou aplikovány tři odlišné modely: Prophet – statistický model určený pro práci s časovými řadami, vyvinutý společností Meta; XGBoost – model strojového učení využívající jednoduché rozhodovací stromy s algoritmem gradientního boostingu; a LSTM – model hlubokého učení (Deep Learning), který používá architekturu rekurentních neuronových sítí schopnou zachytit složité vztahy v časových řadách. Závěrečnou fází praktické části je vyhodnocení a porovnání výsledků jednotlivých modelů, doplněné o shrnutí poznatků, limitací a překážek, které se objevily v procesu tvorby prediktivních modelů.

3 Teoretická východiska

Teoretická část diplomové práce popisuje pojmy a témata důležitá pro proces přípravy a tvorby prediktivní modelů pomocí metod Machine Learning. Charakterizuje předmět zkoumání a modelování: virus Covid-19 a všechny informaci s ním spojené, optikou datové vědy a datového modelování. Dále navazuje s analýzou empirických studií, které se na datovou analýzu vývoje nemoci zaměřily nebo zaměřují. V neposlední řadě charakterizuje programovací jazyk Python, pomocí něhož je prediktivní model šíření nemoci v praktické části implementován. Mimo jazyka samotného přináší popis jednotlivých modulů a knihoven nutných pro správnou konstrukci výsledného modelu. Nakonec se věnuje pojmu strojového učení (Machine Learning), jelikož pochopení a aplikace obsažených metod je nezbytné pro tvorbu prediktivního modelu.

3.1 Covid-19 v datové perspektivě

Pandemie COVID-19 představovala bezprecedentní globální výzvu, která si vyžádala rychlou reakci společnosti podloženou efektivním rozhodováním na základě dat. Propuknutí pandemie přímo či nepřímo vedlo k masivnímu shromažďování dat z různých zdrojů, včetně zdravotnických systémů, mobilních aplikací, sociálních sítí a ekonomických ukazatelů. Analýza získaných dat umožnila lépe porozumět šíření viru, identifikovat rizikové faktory a predikovat budoucí vývoj pandemie. Vědecká komunita i technologické společnosti spojily své síly, aby vytvořily složité datové modely schopné přesněji předpovídat epidemické vlny, efektivitu zavedených opatření a další důležité faktory.

3.1.1 Pandemie Covid-19

Pandemie Covid-19, způsobená virem SARS-CoV-2, se poprvé objevila koncem roku 2019 ve městě Wu-han v Číně. Počáteční případy byly spojovány s místním trhem s mořskými plody, což naznačovalo zoonotický[1] původ viru. Virus však rychle prokázal vysokou schopnost mezilidského přenosu vedoucí k jeho globálnímu rozšíření. Světová zdravotnická organizace (WHO) proto 30. ledna 2020 vyhlásila stav globální zdravotní nouze a 11. března 2020 oficiálně označila situaci za pandemii. Do května 2023 bylo celosvětově zaznamenáno přes 750 milionů potvrzených případů a více než 6,9 milionu úmrtí, přičemž skutečný počet nakažených a obětí je pravděpodobně ještě vyšší. V květnu

[1] Virus přenositelný ze zvířete na člověka [2].

2023 WHO prohlásila, že Covid-19 již nepředstavuje globální zdravotní nouzi, čímž fakticky ukončila pandemickou fázi onemocnění [1].

V České republice byla situace obdobně závažná. Ministerstvo zdravotnictví ČR pravidelně zveřejňovalo data o počtu nakažených, hospitalizovaných a zemřelých. K roku 2025 bylo hlášeno více než 4,8 milionů potvrzených případů a 43 tisíc úmrtí spojených s Covid-19 [3].

Pandemie odhalila zranitelnost globálních zdravotnických systémů, zejména v kontextu nedostatečné kapacity nemocnic, omezené dostupnosti osobních ochranných pomůcek a nepřipravenosti na tak rozsáhlou krizovou situaci. Evropská unie v reakci na tuto situaci přijala opatření zaměřená na posílení zdravotní bezpečnosti, koordinaci členských států a zlepšení krizového řízení v oblasti veřejného zdraví [4].

Na úrovni jednotlivých států byly zavedeny různé druhy strategie zvládání pandemie, včetně plošného testování, trasování kontaktů a očkovacích programů. Tyto kroky měly zásadní význam při snižování rychlosti šíření viru a prevenci přehlcení zdravotnických systémů. Podobné přístupy byly implementovány i v České republice, kde v ranějších fázích pandemie probíhala transformace zdravotnických kapacit na covidové lůžkové jednotky, později také intenzivní informační kampaň na podporu očkování [3].

Jedním z nejvýznamnějších opatření byla globální spolupráce vědců při vývoji vakcín a léků. Do konce roku 2021 byly schváleny vakcíny několika různých typů, hlavně však do té doby pouze experimentálních mRNA vakcín, které prokázaly vysokou účinnost v prevenci těžkého průběhu nemoci. Tento nucený zrychlený vývoj vakcín představoval průlom v oblasti biomedicíny [1].

Pandemie rovněž zdůraznila význam veřejného zdravotnictví a potřebu revidovat a zlepšit dlouhodobou připravenost na podobné krize. WHO v této souvislosti vydala doporučení pro budoucí pandemické plány, která zahrnují zlepšení mezinárodní koordinace, posílení zdravotnických kapacit a investice do výzkumu [1].

3.1.2 Význam datové analýzy

Datová analýza hrála během pandemie Covid-19 klíčovou roli při sledování a zvládání jejího šíření. Dostupná data z testování, nemocničních záznamů nebo později také z analýz odpadních vod umožnila odborníkům porozumět dynamice šíření viru a přijímat efektivní protiepidemická opatření. V České republice byla hojně využívána otevřená data

poskytovaná Ministerstvem zdravotnictví, obsahující informace o počtech nakažených, hospitalizovaných a zemřelých. Tato data tvořila základ nejen pro rozhodování o protiepidemických opatřeních, ale také pro tvorbu prediktivních modelů [3].

Významným příkladem využití datové analýzy v Česku byl nástroj MAMES vyvinutý vědci z Masarykovy univerzity, který umožňoval modelovat šíření nemoci pomocí časových řad, demografických dat a informací o mobilitě populace. Tento nástroj napomáhal při identifikaci epidemiologických trendů a hodnocení efektivity přijatých opatření, jako bylo nošení roušek nebo omezení pohybu osob. Podobně jako v České republice, i v dalších zemích umožnila analýza dat porovnat efektivitu různých strategií ke zvládání pandemie. Například Jižní Korea úspěšně využila pokročilé technologie trasování kontaktů, zatímco země s pomalejší reakcí zaznamenaly výrazně vyšší počty úmrtí spojených s virem [5][6].

Schopnost sbírat a analyzovat data v reálném čase umožnila rovněž sledovat epidemiologické ukazatele, jako například reprodukční číslo, a hodnotit dopad zavedených opatření na šíření viru. Data se stala základem globálně dostupných informačních systémů, mezi nimiž vynikl interaktivní panel Johns Hopkins University, který poskytoval detailní mapy a grafy průběhu pandemie a stal se každodenním zdrojem informací pro veřejnost, média a vlády [5][26].

Pandemie rovněž vyzdvihla význam sdílení dat v rámci mezinárodní vědecké komunity. Globální platformy, například systém GISAID, umožnily sledovat genetické mutace viru SARS-CoV-2 a rychle identifikovat nové varianty, čímž napomohly flexibilnějšímu vývoji vakcín a léčiv [7].

Vedle epidemiologických ukazatelů umožnila datová analýza také identifikovat socioekonomické faktory ovlivňující průběh pandemie. Studie ukázaly, že pandemie nejvíce dopadla na zranitelné skupiny obyvatel, jako jsou senioři, nízkopříjmové rodiny a etnické menšiny. Pandemie tak jasně prokázala, že kvalitní data a jejich důsledná analýza jsou nezbytné pro účinné zvládání budoucích globálních zdravotních krizí [6].

3.1.3 Význam prediktivního modelování

Prediktivní modelování se během pandemie Covid-19 ukázalo jako klíčový nástroj zdravotnictví a epidemiologie. Umožňovalo odhadovat šíření infekce, hodnotit účinnost intervenčních opatření a optimálně rozdělovat zdravotnické zdroje. V průběhu pandemie vzniklo velké množství analytických modelů založených na statistických metodách

a strojovém učení, které predikovaly počty případů, hospitalizací či potřebu intenzivní péče. Tyto predikce umožnily zdravotnickým systémům lépe se připravit například navýšením kapacit nemocnic nebo zásob ochranných prostředků. Modely rovněž sloužily k simulaci různých scénářů dopadů opatření, jako jsou karantény nebo očkovací kampaně, a tím pomáhaly efektivně reagovat na průběh pandemie [8].

Díky prediktivnímu modelování mohli odborníci odhadovat epidemiologické ukazatele, jako je reprodukční číslo, sledovat nárůst infekcí nebo odhadovat vrcholy pandemických vln. Tyto znalosti byly důležité například při určování okamžiku zpřísňování či uvolňování opatření, plánování nemocničních kapacit nebo načasování očkovacích kampaní. Renomované vědecké instituce, jako je for Health Metrics and Evaluation (IHME) ve Spojených státech, poskytovaly klíčové predikce, které využívala Světová zdravotnická organizace (WHO) i jednotlivé vlády států. Přestože míra nejistoty těchto modelů byla vzhledem k rychle se měnící situaci značná, predikce průběžně aktualizované na základě nejnovějších dat poskytovaly politikům a zdravotnickým autoritám důležité informace pro přijímání rozhodnutí [8][9][15].

Význam prediktivního modelování přesáhl hranice samotné pandemie Covid-19. Kombinace pokročilé datové analýzy, strojového učení a globální spolupráce se stala základem pro řešení budoucích zdravotnických krizí. Evropská unie zdůraznila důležitost posílení schopnosti predikce a reakce na krizové situace pomocí pokročilých analytických nástrojů a doporučila integraci těchto metod do budoucích pandemických plánů. Schopnost prediktivních modelů identifikovat komplexní vzorce a souvislosti v datech umožnila například efektivní hodnocení účinnosti očkovacích strategií a pomohla lépe porozumět faktorům ovlivňujícím šíření infekcí [9][10].

Celkově pandemie Covid-19 ukázala, že prediktivní modelování je nezbytnou součástí moderního přístupu ke zvládání zdravotnických krizí. Přestože prediktivní metody nejsou dokonalé a vyžadují neustálou aktualizaci a přizpůsobení měnícím se okolnostem, umožňují lepší přípravu na krizové scénáře, efektivnější alokaci zdrojů a informovanější politická rozhodnutí.

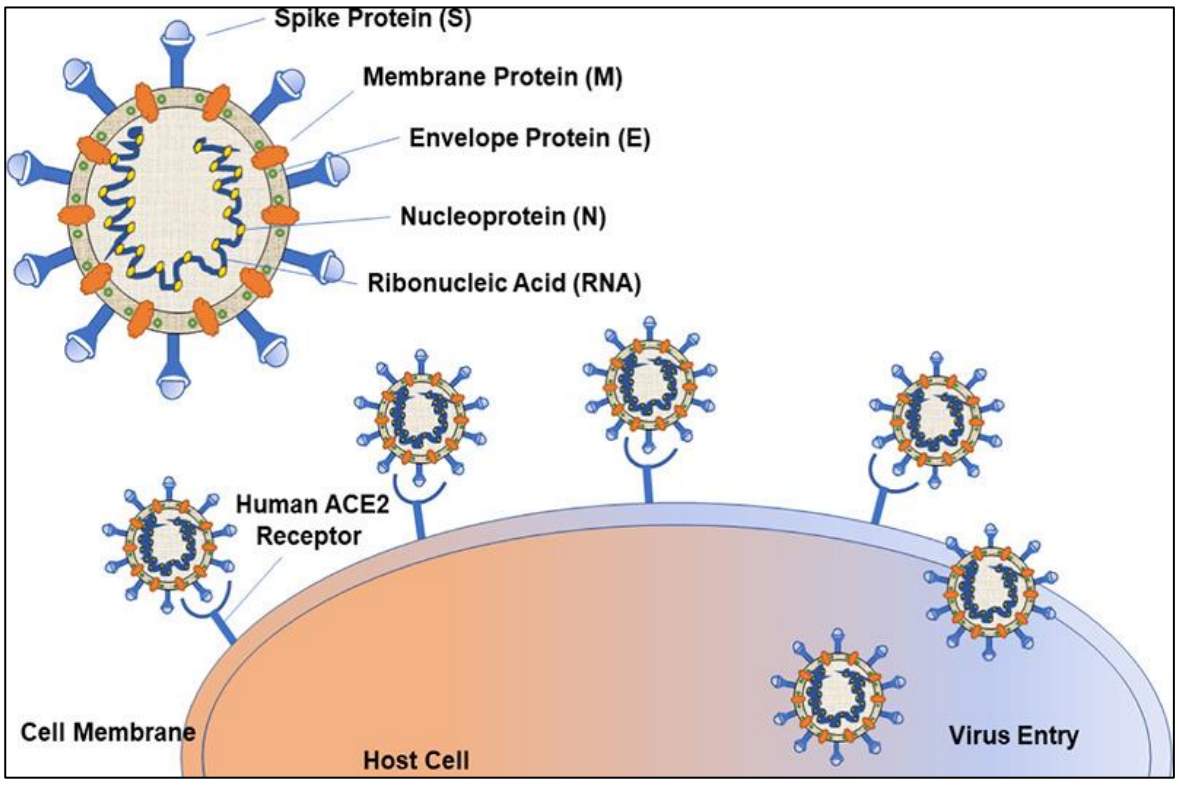
3.2 Charakteristika Covid-19

Onemocnění COVID-19 a jeho původce SARS-CoV-2 způsobily největší zdravotní krizi moderní doby s rozsáhlými dopady na veřejné zdraví, ekonomiku a společnost jako celek. Tato kapitola se zaměřuje na detailní charakteristiku nemoci, včetně jejího původu, biologických vlastností, mechanismu šíření, léčby, prevence a očkování.

3.2.1 Biologické charakteristiky

Virus SARS-CoV-2 (původce nemoci Covid-19) patří do rodiny koronavirů (*Coronaviridae*), které jsou známé svou schopností infikovat různé druhy savců i ptáků. Koronaviry jsou obalené RNA viry charakterizované výraznými výběžky na povrchu obalu, které vytvářejí typický korunovitý vzhled pod elektronovým mikroskopem. Název „coronavirus“ pochází z latinského slova *corona* (koruna) a byl použit pro podobu se vzhledem sluneční koróny (sluneční „atmosféry“) při pozorování dalekohledem. Genom SARS-CoV tvoří jedno-řetězcová RNA, která kóduje čtyři základní strukturální proteiny: spike (S), envelope (E), membrane (M) a nucleocapsid (N), jak je vidět na Obrázku 1 [11].

Obrázek 1 Struktura viru SARS-CoV-2



Zdroj: BBA [11].

Spike protein (S protein) hraje klíčovou roli v mechanismu infekce. Prostřednictvím receptorové vazebné domény (RBD) interaguje s receptorem ACE2[2], který se nachází na povrchu lidských buněk, zejména v plicích, ale i v dalších orgánech jako jsou cévy, ledviny nebo srdce. Vazba spike proteinu na ACE2 receptor vede k fúzi virové a buněčné membrány, čímž dochází k uvolnění virového genomu do buňky a následnému spuštění replikace [12].

Vedle strukturálních proteinů obsahuje genom SARS-CoV-2 také geny pro nestrukturální a pomocné proteiny, které se podílejí na virové replikaci, modulaci buněčné imunitní odpovědi a interakci viru s hostitelskými buňkami. Tento komplexní mechanismus infekce umožňuje viru úspěšně napadat a poškozovat různé tkáně hostitele. Infekce virem SARS-CoV-2 má obvykle inkubační dobu v rozmezí 2 až 14 dní, během které mohou nakažené osoby šířit virus, aniž by vykazovaly jakékoliv příznaky onemocnění. Asymptomatický přenos výrazně urychlil celosvětové rozšíření infekce [11][15].

SARS-CoV-2 se vyznačuje relativně vysokou mutační schopností, typickou pro RNA viry. Na rozdíl od jiných RNA virů však disponuje mechanismem korekce chyb, což do určité míry snižuje jeho mutační rychlost. Přesto v průběhu pandemie vzniklo několik významných subvariant viru. Tyto varianty se vzájemně liší přenosností, klinickou závažností onemocnění a odolností vůči imunitní odpovědi. Mutace ve spike proteinu ovlivňují schopnost viru infikovat lidské buňky a uniknout imunitní ochraně získané očkováním nebo předchozí infekcí. [14].

3.2.2 Varianty viru

Původní varianta viru SARS-CoV-2, identifikovaná ve Wu-chanu na přelomu let 2019 a 2020, měla na koronavirus vysokou přenosnost. To značně napomohlo vzniku globální pandemie. Klinické projevy nemoci sahaly od mírných symptomů (horečka, kašel) po závažné formy, jako je syndrom akutní dechové tísně. Vzhledem k absenci specifické léčby se klíčovými nástroji v boji proti viru stala preventivní opatření a nutnost urychleného vývoje vakcín [11][12].

První významně odlišná varianta se objevila koncem roku 2020. Varianta Alfa, poprvé identifikovaná ve Velké Británii, měla z důvodu mutace na spike proteinu větší přenosnost přibližně o 40-70 % oproti původnímu kmenu. Tato skutečnost vedla k rychlému nárůstu

[2] Angiotenzin-konvertující enzym 2 neboli ACE2 je enzym snižující krevní tlak, který se nachází na povrchu buněk v plicích, tepnách, srdci, ledvinách a střevech. Zároveň slouží jako receptor pro některé lidské koronaviry, např. Sars-CoV-2 [13].

nových případů, dominanci této varianty, a k snížení účinnosti očkování. Dostupné vakcíny si nicméně stále udržely poměrně vysokou ochranu vůči těžkým formám onemocnění [16].

Další variantou byla Beta, původně z Jihoafrické republiky, která obsahovala výrazné mutace, jež částečně umožnily viru uniknout neutralizaci protilátkami získanými prodělánímm infekce či vakcinací. Navzdory těmto mutacím však mRNA vakcíny stále chránily očkované před vážným průběhem onemocnění. Varianta Gamma, identifikovaná v Brazílii, vykazovala podobné mutace jako Beta a dále zvýšila schopnost přenosu a riziko reinfekce [14][11].

V průběhu roku 2021 se objevila varianta Delta, poprvé zaznamenaná v Indii. Tato varianta měla asi dvojnásobnou přenosnost oproti původnímu viru a způsobila výrazný nárůst hospitalizací a úmrtí, zejména u neočkované populace. Data z Anglie ukázala, že varianta Delta mohla zvýšit smrtnost až o 61 % ve srovnání s původním kmenem. Delta se rychle stala dominantní variantou celosvětově [12].

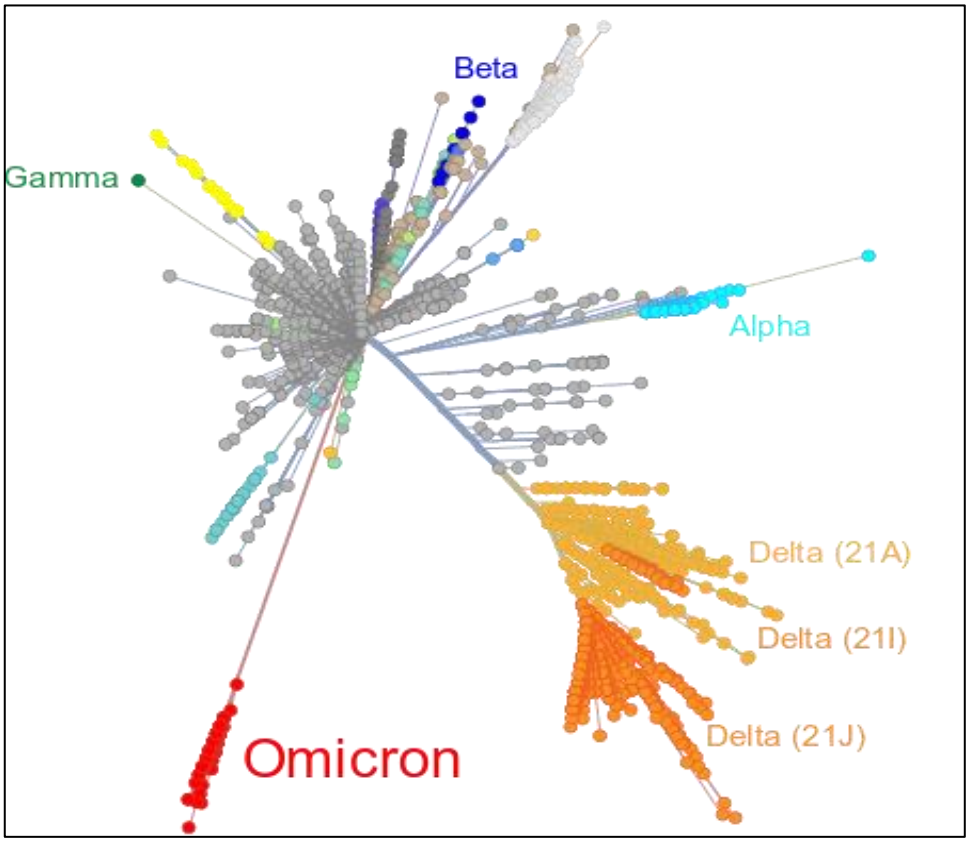
Koncem roku 2021 se objevila varianta Omicron, zjištěná v Jihoafrické republice. Omicron vzbudil globální pozornost kvůli neobvykle vysokému počtu mutací ve spike proteinu, které opět vedly k obcházení tělesné imunity. Přestože Omicron způsobil prudký globální nárůst infekcí, průběh onemocnění byl ve srovnání s Deltou obecně mírnější, zvláště v proočkovaných populacích. Aktualizované vakcíny poskytovaly omezenější ochranu před symptomatickou infekcí, avšak nadále účinně chránily před těžkými průběhy, zejména po podání posilovacích dávek [17].

Omicron se následně štěpil na řadu subvariant, které dominovaly v různých obdobích let 2022–2023, například Kraken, Eris či Pirola. K lednu 2025 mezi cirkulujícími variantami SARS-CoV-2 dominují subvarianty omikronu JN.1 a XEC. Subvarianta JN.1, šířící se rychle v zemích jako Indie, Čína, Spojené království a Spojené státy, nese mutace zvýšující její přenosnost oproti minulým variantám. Podle WHO však současné vakcíny poskytují dostatečnou ochranu a epidemiologické riziko je považováno za nízké. Další subvarianta, XEC vykazuje podobné charakteristiky. Dostupná data nenaznačují závažnější klinické projevy než u předchozích subvariant. Pro sezónu 2024–2025 byly proto vyvinuty aktualizované vakcíny zaměřené na ochranu před novými variantami [17].

Virus SARS-CoV-2 se tedy etabloval jako patogen s řadou cirkulujících variant a subvariant, přičemž odborníci předpokládají, že jeho cirkulace bude nadále pokračovat podobně jako u sezónní chřipky a dalších respiračních virů. Pandemie Covid-19 tak ukázala

význam neustálého sledování genetické diverzity viru a přizpůsobování preventivních i terapeutických strategií nově vznikajícím variantám. Schéma jednotlivých sekvencí a příbuzných větví viru zachycuje Obrázek 2 [17].

Obrázek 2 Schéma variant viru SARS-CoV-2



Zdroj: Wikimedia Commons [18].

3.2.3 Příznaky a průběh onemocnění

Covid-19 se klinicky projevuje velmi širokým spektrem příznaků, od asymptomatických forem či mírných projevů podobných běžnému nachlazení až po závažné stavy vyžadující hospitalizaci, případně vedoucí k úmrtí. Inkubační doba onemocnění je v průměru 5 až 6 dní, může se však pohybovat v rozmezí 2–14 dní. Mezi nejčastější příznaky patří horečka, suchý kašel, únava, bolesti svalů, bolest hlavy, bolest v krku a dušnost. Významná část pacientů, zejména v počátečním období pandemie, zaznamenala charakteristickou ztrátu čichu a chuti (anosmie a ageuzie). Covid-19 je onemocněním systémovým, a proto se mohou objevit také gastrointestinální příznaky (průjem, nevolnost, zvracení), kožní projevy, zimnice nebo rýma. U většiny nakažených osob (cca 80 %) má nemoc mírný průběh bez nutnosti hospitalizace, zatímco přibližně u 15 % pacientů se

rozvine těžší průběh s pneumonií vyžadující lékařskou péči. Kritické stavy (asi 5 % případů) zahrnují respirační selhání, syndrom akutní dechové tísně (ARDS), multiorgánové selhání, trombózy či sekundární bakteriální infekce [19].

Průběh nemoci zásadně ovlivňují faktory jako věk, přítomnost chronických onemocnění a stav imunity. Riziko vážného průběhu prudce roste u starších pacientů, zejména nad 70 let věku, a u jedinců s komorbiditami, jako jsou cukrovka, hypertenze, obezita, kardiovaskulární choroby či imunodeficitní stavy (HIV infekce, stav po transplantaci orgánů). Ačkoli úmrtí u mladých a zdravých lidí je vzácné, v důsledku enormního počtu nakažených vedla pandemie celosvětově k milionům úmrtí. V některých zemích, jako jsou USA nebo Rusko, dokonce výrazně poklesla průměrná délka života obyvatelstva. Kromě respiračního postižení může Covid-19 vyvolávat i řadu systémových komplikací. Byly zaznamenány případy akutního poškození srdce (myokarditida), ledvinového selhání či neurologických komplikací (encefalopatie, cévní mozková příhoda) [19].

Významným problémem souvisejícím s proděláním infekce Covid-19 je postcovidový syndrom, označovaný také jako „long Covid“. Tento syndrom zahrnuje řadu dlouhodobých zdravotních problémů přetrvávajících týdny až měsíce po akutní fázi onemocnění. Long Covid může postihnout pacienty bez ohledu na závažnost původní infekce – týká se jak pacientů s těžkým průběhem, tak i osob, které prodělaly mírné nebo dokonce asymptomatické formy nemoci. Mezi nejčastější příznaky postcovidového syndromu patří chronická únava, dušnost při námaze, bolesti svalů a kloubů, bolesti na hrudi, přetrvávající ztráta čichu a chuti, problémy se spánkem, kognitivní obtíže (poruchy paměti a koncentrace, označované jako „mozková mlha“) a psychické problémy, zejména deprese a úzkost [20].

Long Covid významně snižuje kvalitu života postižených pacientů a má přímé socioekonomické dopady, neboť postižení lidé často nejsou schopni plně se vrátit do práce a běžných aktivit. Je odhadováno, že syndromem long Covid trpí až 10–20 % pacientů po infekci, celosvětově může jít až o desítky milionů osob. Pacienti často potřebují dlouhodobou léčbu, rehabilitaci a podpůrnou péči, což značně zvyšuje zátěž zdravotních systémů. Mechanismy vzniku postcovidového syndromu dosud nejsou plně objasněny, ale studie ukazují na několik možných příčin, včetně přetrvávajícího chronického zánětu, autoimunitních reakcí, dlouhodobého poškození orgánů a přetrvávání viru či jeho fragmentů v organismu [20].

3.2.4 Epidemiologické charakteristiky

Covid-19 se vyznačuje vysokou infekčností, což ilustruje i jeho základní reprodukční číslo (R_0), které bylo v počátečních fázích pandemie odhadováno mezi 2,2 až 3,5. To znamená, že jeden nakažený člověk v populaci bez imunity a bez opatření průměrně infikoval dva až tři další jedince, což vedlo k exponenciálnímu nárůstu počtu nakažených. Ve srovnání se sezónní chřipkou (R_0 okolo 1,3) byla nakažlivost SARS-CoV-2 výrazně vyšší, přibližovala se spíše původnímu SARS z roku 2003. Během pandemie se infekčnost viru dále zvýšila vznikem nových variant [24].

Přenos SARS-CoV-2 probíhá především prostřednictvím respiračních kapének vznikajících při kašli, kýchání či mluvení. Významnou roli hraje rovněž aerosolový přenos, zejména v uzavřených a špatně větraných prostorách. Kontaminace povrchů hrála ve srovnání s kapénkovým a aerosolovým přenosem méně významnou roli, ačkoliv virus mohl na některých materiálech přežívat několik hodin až dnů. K šíření infekce výrazně přispělo i to, že vysoké procento přenosů pocházelo od osob bez viditelných příznaků (asymptomatičtí jedinci). Obecně lze říci, že nakažlivost byla nejvyšší krátce před nástupem příznaků a v prvních dnech onemocnění, přičemž infekčnost trvala obvykle 7–10 dní, v těžkých případech i déle [25].

Průběh pandemie Covid-19 měl významné regionální rozdíly, jež odrážely demografické, sociální a environmentální faktory jednotlivých oblastí. K důležitým faktorům, které ovlivňovaly dopad pandemie, patřily hustota obyvatelstva, úroveň zdravotnických systémů a včasnost a důslednost přijatých protiepidemických opatření. Pandemie také zdůraznila fenomén „superspreadingu“, kdy menšina nakažených způsobila neúměrně vysoký počet dalších infekcí, například při akcích v uzavřených prostorech [26].

Důležitým ukazatelem epidemiologického dopadu viru byla úmrtnost. Letalita infekce (IFR – infection-fatality ratio) se pohybovala přibližně mezi 0,5 až 1 % Epidemiologická data ukazují, že určité demografické skupiny, například starší či nemocní lidé, čelily vyšší úmrtnosti (až 20 %). U mladších skupin populace byla úmrtnost nižší, avšak dlouhodobé následky, jako je postcovidový syndrom, ovlivnily všechny věkové kategorie. Z epidemiologického hlediska se tak Covid-19 zařadil mezi nejvýznamnější infekční hrozby moderní historie. Relativně vysoká infekčnost, kombinovaná se střední úrovní smrtnosti a schopností rychle mutovat, vytvořila podmínky pro dlouhodobé šíření viru v populaci [24][26].

3.2.5 Diagnostika

Diagnostika infekce SARS-CoV-2, je založena na kombinaci klinických příznaků, laboratorních testů a zobrazovacích metod. Nejspolehlivějším nástrojem pro detekci viru je polymerázová řetězová reakce (PCR). Tento test detekuje přítomnost virové RNA ve vzorcích nejčastěji odebíraných výtěrem z nosohltanu, případně ze slin. Princip PCR spočívá v amplifikaci cílové virové RNA, čímž lze identifikovat i minimální množství viru ve vzorku. PCR testy jsou vysoce citlivé a specifické, dosahují téměř 100% analytické přesnosti, ovšem jejich nevýhodou je potřeba specializovaného laboratorního vybavení, vyškoleného personálu a delší doba zpracování (typicky 12–24 hodin, v krizových situacích však i několik dní) [21].

Alternativní metodou testování jsou antigenní testy, které detekují virové proteiny (antigeny) v odebraných vzorcích. Princip těchto testů spočívá ve vazbě antigenu na specifické protilátky na testovací sadě, které při aktivaci tvoří viditelný proužek, podobně jako u těhotenských testů. Výhodou antigenních testů je rychlost (výsledek do 15–30 minut), jednoduchost použití a možnost aplikace přímo v terénu [21].

Hlavní nevýhodou je však nižší citlivost, která dosahuje přibližně 50–80 % oproti PCR. Z tohoto důvodu může být test falešně negativní zejména u osob s nízkou virovou náloží v počátku či závěru infekce nebo u asymptomatických pacientů. Specificita antigenních testů je naopak vysoká (obvykle 97–99 %), proto je pozitivní antigenní výsledek považován za spolehlivou známku infekčnosti. Z těchto důvodů hrály antigenní testy významnou roli při masovém testování ve školách či na pracovištích a umožňovaly rychlou izolaci infekčních jedinců [21].

Při těžších formách onemocnění s podezřením na zápal plic či jiné komplikace spojené s Covid-19 se využívají také zobrazovací metody, zejména rentgenové vyšetření hrudníku a výpočetní tomografie (CT). CT vyšetření umožňuje detailní posouzení stavu plicní tkáně a detekci typických změn způsobených virovou pneumonií, Diagnostika a výsledky těchto vyšetření mohou sloužit jako podklady pro další léčbu u závažnějších případů [21].

3.2.6 Léčba a prevence COVID-19

Léčba onemocnění Covid-19 se v průběhu pandemie postupně vyvíjela podle narůstajících klinických zkušeností a výsledků studií. U mírných případů, které tvořily většinu infekcí, se léčba soustředila na symptomatické řešení obtíží – doporučoval se klidový

režim, dostatečná hydratace a podávání léků snižujících horečku a bolesti, jako jsou analgetika a antipyretika. Tito pacienti obvykle mohli zvládnout léčbu v domácí izolaci bez další pomoci[22].

Pacienti se středně těžkým a těžkým průběhem vyžadovali specifickou farmakologickou léčbu. Již od jara 2020 probíhalo intenzivní testování antivirotik, zejména remdesiviru, který blokuje virovou replikaci a prokázal určitou schopnost zkrátit dobu hospitalizace, přestože jeho efekt na mortalitu byl omezený. Významnější pokrok představovala antivirotika vyvinutá v roce 2021, především Paxlovid nebo Molnupiravir, která při podání v časné fázi infekce (do pěti dnů od nástupu příznaků) významně snížila riziko hospitalizace či úmrtí až o 89 % [22].

Podpůrná léčba zahrnuje podávání kyslíku pacientům s hypoxií, mechanickou ventilaci u pacientů s respiračním selháním a použití monoklonálních protilátek ke snížení virové nálože u rizikových pacientů. V závažných případech se mohou používat také antikoagulační („protisrážlivé“) terapie k prevenci tromboembolických komplikací, které jsou časté u hospitalizovaných pacientů s COVID-19 [22].

V oblasti prevence onemocnění sehrálo nejdůležitější roli očkování, které od konce roku 2020 dramaticky změnilo dynamiku pandemie. Schválené vakcíny, zejména mRNA vakcíny firem Pfizer-BioNTech a Moderna či vektorové vakcíny od AstraZeneca a Johnson&Johnson, vykázaly velmi vysokou účinnost v prevenci těžkého průběhu nemoci, hospitalizací a úmrtí. Vakcinace stimulovala tvorbu specifických protilátek i buněčnou imunitu, což vedlo k omezení cirkulace viru v populaci a ke snížení dopadu jednotlivých pandemických vln. Přestože ochrana před samotnou infekcí mohla časem klesat (zejména v důsledku nových variant viru jako Omicron), podávání posilovacích (booster) dávek znovu posilovalo imunitní odpověď, zejména u rizikových skupin populace [23].

Podpůrná preventivní opatření byla rovněž zásadní, zvláště v počáteční fázi pandemie před dostupností vakcín. Nošení respirátorů a roušek významně omezilo šíření viru kapénkami a aerosolem, sociální distancování a omezení velkých akcí snížily pravděpodobnost hromadného šíření a pravidelné testování s trasováním kontaktů umožnilo rychleji identifikovat a izolovat nakažené osoby. Důsledné dodržování hygienických pravidel (mytí a dezinfekce rukou) pomáhalo minimalizovat riziko kontaktního přenosu [23].

3.3 Analýza empirických studií

Empirické studie hrají klíčovou roli při získávání důkazů o různých aspektech pandemie COVID-19 a jejich dopadů na společnost. Analýza těchto studií poskytuje cenné informace o šíření viru, účinnosti zavedených opatření, dopadech na zdravotnické systémy a ekonomiku, stejně jako o psychologických a sociálních důsledcích pandemie. Tato kapitola se zaměřuje na přehled klíčových empirických studií, které přispěly k pochopení dynamiky pandemie a pomohly při formulování strategií pro její zvládání.

3.3.1 Přehled a typologie studií

Pandemie Covid-19 vyvolala bezprecedentní nárůst empirických studií, které se zaměřily na široké spektrum aspektů této globální krize. Tyto studie lze dle zaměření rozdělit do několika hlavních kategorií: epidemiologické, klinické, ekonomické a sociálně-behaviorální. Přestože jednotlivé kategorie mají vlastní specifické zaměření, mnohé studie vykazovaly průniky napříč těmito oblastmi [27].

Epidemiologické studie se věnovaly analýze šíření viru, jeho infekčnosti a efektivitě zavedených opatření. Velký význam měl zejména výzkum založený na matematickém modelování, který umožnil předpovědět vývoj pandemie a testovat scénáře různých intervencí. Například studie Ferguson et al. (2020) ukázala, že kombinace opatření, jako je sociální distancování a uzavírání škol, může dramaticky snížit počet infekcí a zátěž nemocnic. Epidemiologický výzkum rovněž zkoumal účinnost konkrétních opatření, jako je nošení roušek či plošné testování populace [28].

Klinické a biomedicínské studie se zaměřovaly na klinické charakteristiky pacientů, rizikové faktory těžkého průběhu či podíl asymptomatických případů. Například studie publikovaná v časopise The Lancet doložila, že antivirotikum remdesivir dokáže zkrátit dobu hospitalizace pacientů s těžkým průběhem Covid-19. Neméně důležitý byl rychlý vývoj a testování vakcín, například vakcína Pfizer-BioNTech vykázala v klinické fázi účinnost přesahující 95 % [29].

Ekonomické studie dokumentovaly dopady pandemie na globální i lokální ekonomiku. Výzkumy organizací jako Mezinárodní měnový fond (IMF) či Světová banka ukázaly, že pandemie způsobila největší ekonomickou krizi od Velké hospodářské krize 30. let, kdy globální HDP poklesl v roce 2020 o přibližně 3,5 %. Tyto studie rovněž hodnotily účinnost fiskálních stimulů a záchranných balíčků vlád, analyzovaly náklady a přínosy různých

strategií zvládání pandemie (např. lockdowny vs. otevřená ekonomika) a mapovaly dopady na specifická odvětví, jako jsou cestovní ruch, gastronomie nebo doprava [30].

Sociální a behaviorální studie se zaměřily na dopady pandemie na každodenní život, psychické zdraví populace a společenské změny. Zkoumaly například vlivy izolace během lockdownů, změny pracovních podmínek a dopady distanční výuky na vzdělávání. Významným tématem bylo také duševní zdraví, například studie Světové zdravotnické organizace (WHO) upozornily na výrazný nárůst výskytu úzkosti a deprese, zejména mezi mladými lidmi a seniory, kteří byli dlouhodobě izolováni od svých komunit [31].

Celkově pandemie Covid-19 vedla k výraznému nárůstu množství i kvality empirických studií, což představuje významný posun v přístupu k řešení globálních krizí. Výsledky tohoto komplexního výzkumu jsou klíčové nejen pro zvládání současné pandemie, ale především pro efektivnější řešení podobných krizových situací v budoucnu [27].

3.3.2 Metody datové analýzy v epidemiologii

Datová analýza sehrála během pandemie Covid-19 zásadní roli při zpracování velkého množství epidemiologických dat. Nejpoužívanějšími metodami byly deskriptivní analýza, analýza časových řad a modely založené na strojovém učení. Deskriptivní analýza umožnila rychlé pochopení základních vzorců a trendů v datech, jako byly geografické rozdíly v šíření infekce nebo dynamika hospitalizací. Identifikace těchto vzorců představovala klíčový krok k efektivnímu plánování protiepidemických opatření [32].

Velký důraz byl kladen na analýzu časových řad používanou při predikci vývoje epidemiologických vln. Pro tyto účely se využívaly jak klasické metody, například ARIMA modely, tak specializované epidemiologické modely typu SEIR (Susceptible-Exposed-Infected-Recovered). Například studie Johns Hopkins University prokázaly, že ARIMA modely dokážou poměrně přesně předpovídat denní počty případů a vrcholy vln, což umožnilo zdravotnickým systémům připravit se na zvýšený nápor pacientů [33].

Strojové učení (ML) přineslo do epidemiologické analýzy další rozměr. Díky schopnosti algoritmů zpracovat komplexní datové sady (včetně údajů o mobilitě populace, meteorologických dat či demografických parametrů) bylo možné identifikovat skryté vzory a vztahy, které tradiční modely často nezachytily. Algoritmy byly úspěšně aplikovány na predikci rizikových faktorů či identifikaci ohnisek nákazy. Studie publikovaná v časopise

Nature doložila, že kombinace strojového učení s klasickou epidemiologií vede k přesnějším predikcím a hlubšímu pochopení dynamiky pandemie [34].

Nedílnou součástí epidemiologické práce během pandemie byla i vizualizace dat. Srozumitelné grafy, mapy a interaktivní dashboardy umožnily rychlou komunikaci důležitých informací jak odborníkům, tak široké veřejnosti. Nejznámějším příkladem byl „Covid-19 Dashboard“ vytvořený Johns Hopkins University, který se stal celosvětově využívaným nástrojem pro sledování pandemie [26].

Důležitou roli během pandemie hrála také mezinárodní spolupráce a sdílení dat mezi institucemi a státy. Platformy jako GISAID umožnily globální sdílení genomických dat viru SARS-CoV-2, což přispělo k rychlé identifikaci a monitorování nových mutací a jejich dopadu na přenosnost a závažnost onemocnění. Tento sdílený přístup k datům významně urychlil vývoj vakcín a antivirotik a zlepšil schopnost globálního systému reagovat na pandemické hrozby [7].

3.3.3 Limity a výzvy při práci s daty

Práce s epidemiologickými daty během pandemie Covid-19 se potýkala s řadou výzev a omezení, které ovlivňovaly efektivitu analýzy a přesnost predikcí šíření nemoci. Jedním z nejvýznamnějších problémů byla neúplnost a nekonzistence datových zdrojů. Rozdílné testovací strategie a kritéria napříč státy vedly k nesrovnatelným údajům o počtu nakažených, hospitalizací či úmrtí. Omezená testovací kapacita zejména v počátečních fázích pandemie způsobila výrazné podhodnocení skutečného počtu případů. Navíc data často obsahovala chyby, duplicity, nekompletní informace či časové nesrovnalosti (například opožděné dohlášení případů), což komplikovalo jejich využití v prediktivním modelování [3][35][36].

Rychle se měnící povaha pandemie znamenala, že i dobře navržené modely se rychle stávaly zastaralými. Například modely kalibrované na variantu Delta měly omezenou predikční schopnost po nástupu varianty Omicron s odlišnou dynamikou přenosu. Proto bylo nutné modely neustále aktualizovat a zahrnovat nové proměnné, jako jsou vakcinační kampaně, sezónní faktory nebo změny chování populace. Studie Masarykovy univerzity ukázala, že implementace specializovaných epidemiologických nástrojů a průběžná aktualizace modelů výrazně zlepšily schopnost predikce a reakce na vývoj pandemie [5].

Etické otázky a ochrana soukromí představovaly další významnou překážku při práci s daty. Například využití digitálních trasovacích aplikací (v ČR aplikace eRouška) narazilo na omezenou ochotu veřejnosti sdílet data kvůli obavám z narušení soukromí. Kromě trasování byly také řešeny problémy s anonymizací zdravotnických dat při výzkumu – agregovaná data chránila soukromí pacientů, avšak snižovala možnosti detailní analýzy [37].

Zpracování a analýza dat také čelily logistickým výzvám. Pandemie vyžadovala rychlou aktualizaci a agregaci dat v reálném čase z velkého množství různorodých zdrojů. Například dashboard Univerzity Johnse Hopkinse agregoval data z více než 400 globálních zdrojů, což vyžadovalo robustní IT infrastrukturu a schopnost průběžného škálování datových platforem. Změny ve vykazování a metodice (například přechody mezi různými definicemi hospitalizací) dále komplikovaly dlouhodobou analýzu dat [26].

Komplexita dat vyžadovala multidisciplinární přístup, který propojil epidemiologické poznatky s ekonomickými a sociologickými daty. Například modelování vlivu lidského chování (mobilita, ochota dodržovat opatření) na dynamiku šíření viru představovalo metodologickou výzvu kvůli obtížné kvantifikaci behaviorálních změn. Multidisciplinární spolupráce mezi epidemiology, datovými vědci, zdravotníky a sociálními vědci proto byla zásadní pro správnou interpretaci dat a rozhodování během pandemie [5].

Navzdory všem uvedeným výzvám pandemie přinesla i významný pokrok ve sběru, analýze a sdílení dat. Mnohé státy zřídily otevřené datové portály, které umožnily transparentní zveřejňování denních statistik a jejich nezávislé ověřování ze strany analytiků a vědců. Pandemie také zdůraznila potřebu robustních datových infrastruktur, které by byly schopny v budoucnu rychle poskytovat kvalitní informace pro zvládání krizí [35].

3.4 Python

Python je jedním z nejpopulárnějších programovacích jazyků současnosti, pro který existuje široké uplatnění v mnoha oblastech, včetně datové analýzy, strojového učení a modelování komplexních systémů, jakým je i šíření onemocnění COVID-19. Díky své jednoduché a čitelné syntaxi, rozsáhlé standardní knihovně a široké komunitě vývojářů se Python stal preferovaným nástrojem nejen pro práci s velkými datovými sadami a pro vývoj pokročilých algoritmů. Kapitola se věnuje představení základních vlastností Pythonu, jeho historie a důvodů, proč je považován za klíčový nástroj v oblasti datové vědy.

3.4.1 Historie a vývoj jazyka Python

Programovací jazyk Python byl vytvořen Guidem van Rossumem na přelomu 80. a 90. let 20. století, přičemž první veřejná verze byla zveřejněna 20. února 1991. Jazyk navazoval na jazyk ABC a od začátku byl navržen jako snadno čitelný, univerzální a uživatelsky přístupný nástroj. Název Python byl inspirován britskou komediální skupinou Monty Python [38].

V průběhu let Python prošel řadou významných změn. Klíčovým milníkem bylo vydání verze Python 2.0 v roce 2000, která zavedla například automatickou správu paměti pomocí funkce „garbage collector“. Dalším zásadním krokem byl příchod verze Python 3.0 v roce 2008, která přinesla významná vylepšení, jako například nativní podporu Unicode či efektivnější práci s řetězci, ale zároveň znamenala nutnost přechodu od předchozích verzí. Přestože podpora Pythonu 2 skončila v roce 2020, díky aktivní podpoře komunity je přechod na Python 3 dnes téměř dokončený [39].

Python je v současnosti otevřeným projektem spravovaným neziskovou organizací Python Software Foundation. Díky aktivní komunitě rychle roste dostupnost knihoven a nástrojů, tedy i pronikání jazyka do mnoha nových odvětví, zejména do oblasti datové vědy, vědeckých výpočtů a umělé inteligence. K rozšíření Pythonu přispěly také webové frameworky jako Django či Flask, které výrazně usnadnily tvorbu webových aplikací. Každoročně vydávané nové verze jazyka (např. 3.11 v roce 2022 a 3.12 v roce 2023) přinášejí průběžná zlepšení výkonu a nové funkcionality, díky čemuž jazyk neustále udržuje krok s technologickým vývojem [40].

Dnes patří Python k nejpoužívanějším a nejpopulárnějším programovacím jazykům. Podle průzkumu Stack Overflow z roku 2021 je oblíbený především díky své flexibilitě, jednoduchosti použití a schopnosti rychle se adaptovat v různých oblastech – od webového vývoje přes systémovou administraci až po pokročilé aplikace ve vědeckých výpočtech, strojovém učení a analýze velkých dat [41].

3.4.2 Charakteristika Pythonu

Python patří mezi nejoblíbenější programovací jazyky díky vlastnostem, které jej činí univerzálním, efektivním a snadno použitelným nástrojem. Jednou z jeho nejvýznamnějších předností je jednoduchá a čitelná syntaxe založená na odsazování bloků místo složených závorek, což usnadňuje psaní a porozumění kódu. Tato vlastnost je ceněna nejen začátečníky, kterým umožňuje rychlejší pochopení a snadný vstup do programování, ale i zkušenými vývojáři, pro které znamená efektivnější práci a vyšší produktivitu. Díky tomu lze v Pythonu řešit problémy s výrazně menším množstvím kódu než například v jazycích Java či C++ [38].

Univerzálnost Pythonu spočívá v jeho podpoře různých programovacích paradigmat – procedurálního, objektově orientovaného i funkcionálního programování. Jazyk je tak schopen přizpůsobit se širokému spektru projektů od webového vývoje přes automatizaci až po komplexní datové analýzy. Python disponuje vestavěnou standardní knihovnou, která pokrývá běžné funkce, ale také obrovským množstvím externích knihoven specializovaných na konkrétní oblasti, jako jsou webové frameworky (např. Django, Flask), datová věda (Scikit-learn, TensorFlow) či zpracování přirozeného jazyka [39].

Díky open-source charakteru Pythonu a aktivní komunitě vývojářů jsou knihovny i samotný jazyk neustále aktualizovány a vylepšovány. Vývojáři z celého světa mohou přispívat k rozvoji jazyka, čímž je zajištěno jeho průběžná inovaci a flexibilita vůči novým technologickým výzvám. Python běží na různých platformách včetně Windows, Linux, MacOS a mikrokontrolerů (MicroPython) [38].

Důležitou výhodou Pythonu je také jeho snadná integrace a rozšiřitelnost. Je-li potřeba vyšší výkon, lze kritické části aplikací napsat v „rychlejších“ kompilovaných jazycích, jako jsou C, C++ nebo Java, a jednoduše je propojit s Pythonem. K dispozici jsou i implementace Pythonu pro jiné platformy, například Jython běžící na JVM nebo IronPython pro .NET [39].

Popularita Pythonu se odráží také v aktivní komunitě uživatelů, kteří sdílejí zdroje, jako jsou tutoriály, příspěvky diskusních fór a každoročně pořádané odborné konference (např. PyCon). Tato komunita poskytuje začínajícím programátorům i pokročilým vývojářům potřebnou podporu a motivuje je k neustálému rozvoji dovedností. Python je také hojně využíván v akademické sféře pro výuku programování [40].

3.4.3 Frameworky a knihovny pro práci s daty

Následující podkapitola obsahuje seznam důležitých frameworků a knihoven, které jsou velmi často využívány pro práci s daty, datovou analýzu, grafické znázornění, či více pokročilé funkce, jako je tvorba modelů strojového učení a konstrukce složitých vícevrstvých neuronových sítí.

3.4.3.1 Tensor Flow

TensorFlow patří mezi nejpopulárnější frameworky pro strojové učení a zejména hluboké neuronové sítě. Byl vyvinut společností Google a umožňuje vytvářet, trénovat a nasazovat pokročilé AI modely díky své flexibilitě a modularitě. TensorFlow využívá optimalizace GPU i specializovaných TPU (Tensor Processing Unit), což výrazně urychluje výpočetní čas nutný pro trénování robustních modelů. Díky aktivní podpoře firmy Google a široké komunitě vývojářů je hojně využíván v průmyslu i akademické sféře. K snadnému použití slouží rozhraní Keras, které je součástí TensorFlow a nabízí vysoce abstraktní API pro jednoduché sestavování neuronových sítí. Během pandemie COVID-19 byl TensorFlow využíván například při analýze lékařských snímků (detekce infekce plic z CT) či v predikčních modelech šíření nákazy pomocí neuronových sítí typu LSTM [38].

- **Použití**: trénování neuronových sítí (CNN, RNN), predikce časových řad, zpracování obrazu a přirozeného jazyka;
- **Hlavní funkce**: tf.keras, TensorFlow Datasets, TensorFlow Serving, TensorFlow Lite;
- **Výhody**: vysoký výkon (GPU/TPU), flexibilní API, silná komunitní podpora [38].

3.4.3.2 Pandas

Pandas je jednou z nejdůležitějších knihoven pro analýzu a manipulaci s tabulkovými daty v Pythonu. Nabízí datové struktury jako DataFrame a Series, které umožňují snadné načítání, zpracování a agregaci dat z různých zdrojů (CSV, Excel, SQL, JSON). Pandas je velmi oblíbený díky své jednoduché, SQL podobné syntaxi a vysokému výkonu. Ve spojení s pandemií Covid-19 se Pandas používal pro agregaci a vizualizaci epidemiologických dat, čištění datových souborů a analýzu časových trendů [42].

- **Použití**: čištění dat, manipulace s časovými řadami, agregace, spojování datasetů;
- **Hlavní funkce**: read_csv(), groupby(), pivot_table(), merge();
- **Výhody**: jednoduchost, výkon, široká komunita [42].

3.4.3.3 NumPy

NumPy (Numerical Python) tvoří základ pro vědecké výpočty v Pythonu. Poskytuje podporu více rozměrových polí (array) a rozsáhlé matematické operace s efektivní implementací v C. NumPy umožňuje rychlé numerické výpočty a je klíčovým stavebním kamenem pro mnoho dalších knihoven (např. Pandas, Scikit-learn). Během pandemie NumPy sloužil například pro statistické analýzy epidemiologických dat a výpočty predikčních metrik [42].

- **Použití**: numerické výpočty, vektorové a maticové operace;
- **Hlavní funkce**: array(), reshape(), mean(), dot();
- **Výhody**: vysoká efektivita, dobrá integrace s jinými knihovnami [42].

3.4.3.4 Matplotlib

Matplotlib je základní knihovnou pro tvorbu grafů v Pythonu. Umožňuje detailní přizpůsobení vizualizací (čárové grafy, sloupcové grafy, histogramy či bodové grafy). V epidemiologii a analýze dat z COVID-19 se hojně používal pro vizualizaci trendů, distribucí a porovnávání regionů v čase. Je flexibilní a snadno použitelný a je možné skrz něj exportovat grafy v mnoha podporovaných formátech [42].

- **Použití**: vizualizace dat a výsledků modelů;
- **Hlavní funkce**: plot(), bar(), hist(), scatter();
- **Výhody**: flexibilita, snadné ovládání, export grafů [42].

3.4.3.5 Scikit-learn

Scikit-learn je jedna z nejpoužívanějších knihoven strojového učení v Pythonu. Nabízí široké spektrum algoritmů pro regresi, klasifikaci, shlukování či redukci dimenzionality. Má přehledné a intuitivní API (*fit(), predict())* a dobře spolupracuje s Pandas a NumPy. V pandemii COVID-19 se využíval například k predikci hospitalizací či analýze rizikových faktorů [42].

- **Použití**: predikční modely, klasifikace dat;
- **Hl. funkce**: train_test_split(), LinearRegression(), RandomForestRegressor();
- **Výhody**: široký výběr algoritmů, jednoduchost použití a nastavení, kvalitní dokumentace [42].

3.4.4 Budoucnost v datových vědách

Python si díky své jednoduchosti, univerzalitě a silnému ekosystému knihoven a frameworků udržuje pozici hlavního programovacího jazyka v datové vědě a umělé inteligenci. Mezi klíčové trendy, které upevňují postavení Pythonu, patří rozšiřování jeho využití v oblasti GPU akcelerace (např. knihovny RAPIDS), zpracování přirozeného jazyka (např. spaCy, NLTK), interoperabilita datových formátů (Apache Arrow) a optimalizace výkonu (např. odstranění Global Interpreter Lock – GIL, PyPy, Cython) [40].

V datové vědě a analýze se Python podle průzkumů etabloval jako primární nástroj pro více než 80 % analytiků a výzkumníků. Zároveň se očekává další posilování kombinace Pythonu a SQL jako hlavních jazyků datových profesionálů. Python navíc funguje jako univerzální skriptovací jazyk pro většinu AI/ML platforem [39].

Díky svým výhodám popsaným výše a robustnímu ekosystému knihoven má Python jasnou budoucnost v oblasti datové vědy a umělé inteligence. Lze očekávat, že bude i nadále preferovanou volbou nejen pro začátečníky, ale i pro odborníky a zachová si své postavení klíčového nástroje pro řešení budoucích technologických a globálních programovacích problémů [38].

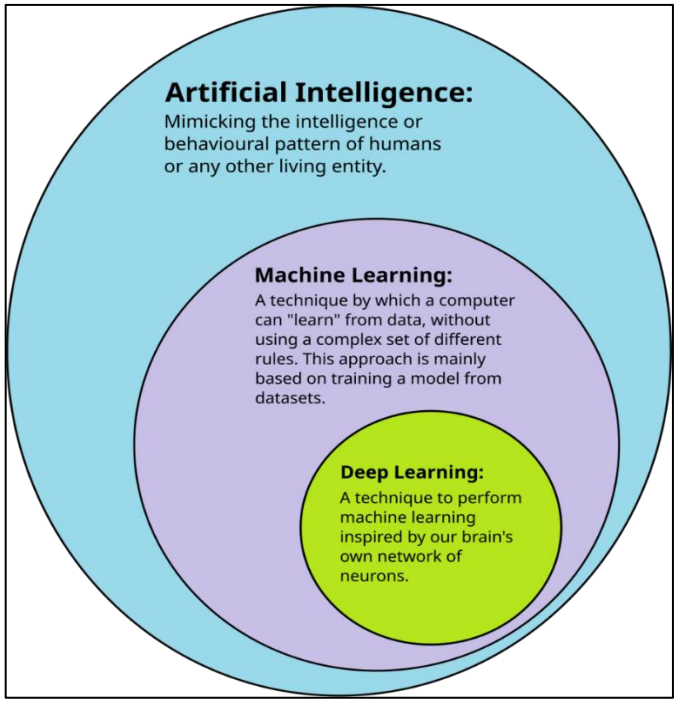
3.5 Machine Learning

Strojové učení (Machine Learning) je dynamicky se rozvíjející oblast umělé inteligence, která umožňuje počítačovým systémům učit se ze zkušeností a dat bez potřeby explicitního programování. Tato technologie se stále častěji uplatňuje v široké škále odvětví, včetně zdravotnictví, finančnictví, marketingu nebo průmyslu. V kontextu pandemie COVID-19 bylo jeho zapojení zásadní při predikci šíření nemoci, identifikaci rizikových faktorů a optimalizaci zdravotnických kapacit.

3.5.1 Definice a základní koncepty

Strojové učení (Machine Learning, ML) představuje významnou podmnožinu umělé inteligence (AI), jejíž hlavní vlastností je schopnost počítačových systémů automaticky se učit a zlepšovat na základě zkušeností (dat), aniž by musely být všechny rozhodovací postupy explicitně naprogramovány. Zatímco v tradičním programování programátor ručně definuje přesný algoritmus, jak ze vstupu získat požadovaný výstup, ve strojovém učení se vytváří flexibilní model, který vztah mezi vstupem a výstupem objevuje sám na základě analýzy a interpretace dostupných dat. Strojové učení je nadřazeným pojmem Hlubokému učení (Deep Learning) – jak je vidět na Obrázku 3 [44].

Obrázek 3 Hierarchie AI a její součásti



Zdroj: Wikimedia Commons [45].

Dle klasické definice Toma Mitchella se systém učí z příkladů (zkušenosti) vzhledem k dané úloze a metrice výkonnosti tehdy, když se jeho výkonnost v dané úloze s rostoucím množstvím zkušeností zlepšuje. Jádrem strojového učení jsou algoritmy, které zpracovávají vstupní data a iterativně upravují parametry svých modelů (např. váhy neuronů v neuronových sítích, nebo rozdělovací pravidla rozhodovacích stromů), aby minimalizovaly chybu na trénovacích datech. Pro hledání optimálních parametrů se používají například algoritmy založené na gradientním sestupu [46].

Důležitým principem ML je generalizace – model musí být schopen naučit se obecné vzory, nikoliv pouze memorovat trénovací data. S tím úzce souvisí riziko přeučení (overfitting), kdy model příliš přizpůsobí parametry trénovacím datům a ztrácí schopnost dobře predikovat na nových, dosud neviděných datech. Proto se používají metody validace, například rozdělení dat na trénovací, validační a testovací sady, které umožňují správně posoudit schopnost modelu zobecňovat realitu [46].

Strojové učení má široké spektrum aplikací, včetně zdravotnictví, financí, logistiky a marketingu. Například v medicíně se používá k analýze snímků, diagnostice nemocí a personalizované léčbě. Aspektem nutným pro správnou predikci je kvalita dat, protože špatně připravená nebo zaujatá data mohou negativně ovlivnit výkon a vypovídající sílu modelu. Proto je důležitá pečlivá příprava a předzpracování dat zahrnující odstraňování chyb, normalizaci a transformaci atributů [47].

3.5.2 Deep Learning

Hluboké učení (Deep Learning) je specifickou oblastí strojového učení, která využívá **vícevrstvé** neuronové sítě ke zpracování složitých datových struktur a automatické extrakci hierarchických reprezentací z dat. Na rozdíl od tradičních metod strojového učení, které vyžadují manuální návrh příznaků **(**feature engineering**),** umožňuje hluboké učení samostatně odhalit významné vzory a reprezentace přímo ze surových dat.

Vícevrstvé neuronové sítě jsou volně inspirovány strukturou lidského mozku a skládají se z mnoha vzájemně propojených umělých neuronů. Každý neuron provádí lineární kombinaci vstupů a aplikuje na výsledek nelineární aktivační funkci. Termín „hluboké“ označuje neuronové sítě obsahující zpravidla několik až stovky vrstev, toto zapojení umožňuje zachytit složité nelineární vztahy [48].

Mezi nejvýznamnější architektury hlubokého učení patří:

- **Konvoluční neuronové sítě (CNN):** Specializované na zpracování obrazových dat, CNN používají konvoluční vrstvy k detekci lokálních rysů (hrany, textury), které postupně spojují do komplexnějších tvarů a objektů. Dosáhly revolučních úspěchů v oblastech počítačového vidění, zejména klasifikaci obrazů a detekci objektů. Během pandemie Covid-19 byly CNN efektivně nasazeny například pro detekci příznaků covidové pneumonie na RTG a CT snímcích s vysokou přesností [46].
- **Rekurentní neuronové sítě (RNN)**: RNN jsou vhodné pro analýzu sekvenčních dat (časové řady, texty), neboť udržují vnitřní stav aktualizovaný během zpracování vstupních sekvencí. Klasické RNN však mají problémy s učením dlouhodobých závislostí (mizející gradient), proto byly vyvinuty jejich efektivnější varianty jako LSTM (Long Short-Term Memory) a GRU (Gated Recurrent Units). Tyto sítě jsou efektivní pro strojový překlad, rozpoznávání řeči či predikci časových řad, například epidemického šíření [46].
- **Transformery**: Relativně nová architektura představená roku 2017, využívající mechanismus „self-attention“ místo explicitní rekurence. Díky své schopnosti efektivně zachycovat dlouhé závislosti se transformery staly základem moderních jazykových modelů jako je například asi nejznámější velký jazykový model GPT (Generative Pre-trained Transformers). V kontextu pandemie pomáhaly transformery analyzovat obrovská množství biomedicínských publikací týkajících se Covid-19 [38].

Hluboké učení umožňuje použití stejné architektury napříč různými oblastmi, pokud je k dispozici dostatečné množství dat. Díky škálovatelnosti lze efektivně zpracovávat rozsáhlé datové sady pomocí výkonného hardwaru, jako jsou GPU a TPU. Hluboké učení také vykazuje vysokou přesnost na komplexních rozpoznávacích úlohách, kde značně překonává lidské možnosti v rozpoznávání obrazu, řeči a dalších oblastech [44].

Přes své četné přínosy čelí použití hlubokého učení celé řadě výzev. Tou je například vysoká závislost na kvalitních a anotovaných datech – bez nich modely selhávají. Kromě toho je trénování hlubokých neuronových sítí výpočetně náročné, představuje tedy mnohem vyšší infrastrukturní a peněžní zátěž. Problematická je rovněž interpretovatelnost těchto modelů – jejich rozhodnutí často působí jako „černá skříňka“, kdy není jasné, na základě čeho model predikuje své výsledky. Dalším rizikem je, stejně jako u ML modelů, přeučení (overfitting), kdy model ztrácí schopnost generalizovat na nová data [47].

3.5.3 Druhy strojového učení

Následující kapitola popisuje tři nejčastější druhy strojového učení, které jsou využívány podle povahy řešeného problému a charakteru dat, se kterými je pracováno. Jde o učení s učitelem, učení bez učitele a také mírně odlišné rozšířené učení.

3.5.3.1 Učení s učitelem

Učení s učitelem je nejrozšířenější a nejvíce využívanou formou strojového učení. Využívá označená trénovací data, která obsahují vstupy a k nim přiřazené správné výstupy (tzv. labely). Cílem je naučit model mapovat nové vstupy na správné výstupy tím, že se učí z dostupných příkladů. Model se učí nalézat vzorce ve značených datech a používá je k predikci nových neznámých vstupů [49].

Příklady algoritmů pro učení s učitelem zahrnují regresní a klasifikační úlohy. Klasifikace spočívá v přiřazení vstupních dat do předem definovaných kategorií, což se využívá například při rozpoznávání spamových e-mailů nebo diagnostice nemocí z lékařských snímků. Regrese se naopak zaměřuje na predikci spojitých hodnot, jako je odhad cen nemovitostí nebo předpověď počtu budoucích případů onemocnění na základě historických dat. V rámci učení s učitelem jsou využívány různé algoritmy, například lineární a logistická regrese, rozhodovací stromy a náhodné lesy, metoda podpůrných vektorů (SVM) nebo neuronové sítě [49].

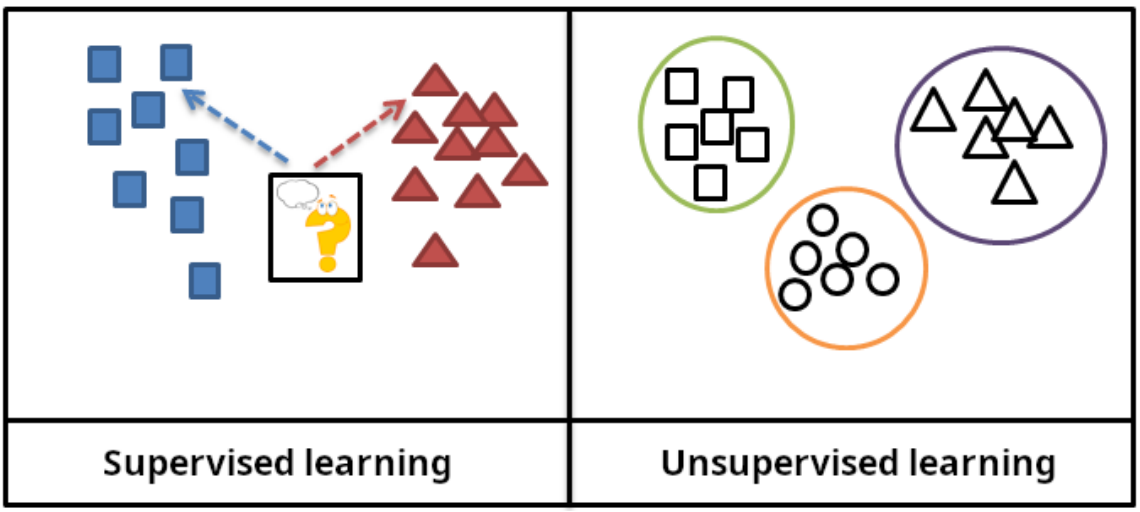
Jednou z hlavních výzev učení s učitelem je potřeba velkého množství kvalitně označených dat. Pokud data nejsou reprezentativní nebo obsahují chyby, může dojít ke zkreslení modelu (tzv. bias), čímž je narušena jeho generalizační schopnost [48].

3.5.3.2 Učení bez učitele

Učení bez učitele (unsupervised learning) se uplatňuje v situacích, kdy nejsou k dispozici označená data nebo je jejich anotace příliš nákladná. V těchto případech modely samostatně objevují skryté struktury a vzory v datech bez předem definovaných labelů. Mezi typické úlohy patří shlukování („clustering“), kde algoritmy jako k-means nebo hierarchické shlukování seskupují data podle vnitřní podobnosti – například segmentace zákazníků podle nákupního chování či klasifikace virových variant během pandemie. Další důležitou oblastí je redukce dimenzionality, například pomocí metody hlavních komponent (PCA), která umožňuje zjednodušit komplexní data a odstranit nadbytečné informace. Učení bez učitele se rovněž využívá při detekci anomálií, kdy jsou identifikovány datové body výrazně se lišící od ostatních, například při hledání neobvyklých finančních transakcí nebo nečekaných kombinací symptomů [44][50].

Jednou z hlavních výzev učení bez učitele je interpretace výsledků, protože model nemusí generovat jednoznačné výstupy a určit, zda jsou nalezené vzory skutečně užitečné nebo jen náhodné. Rozdíl mezi oběma přístupy k učení zachycuje Obrázek 4 [50].

Obrázek 4 Učení s učitelem vs. Učení bez učitele

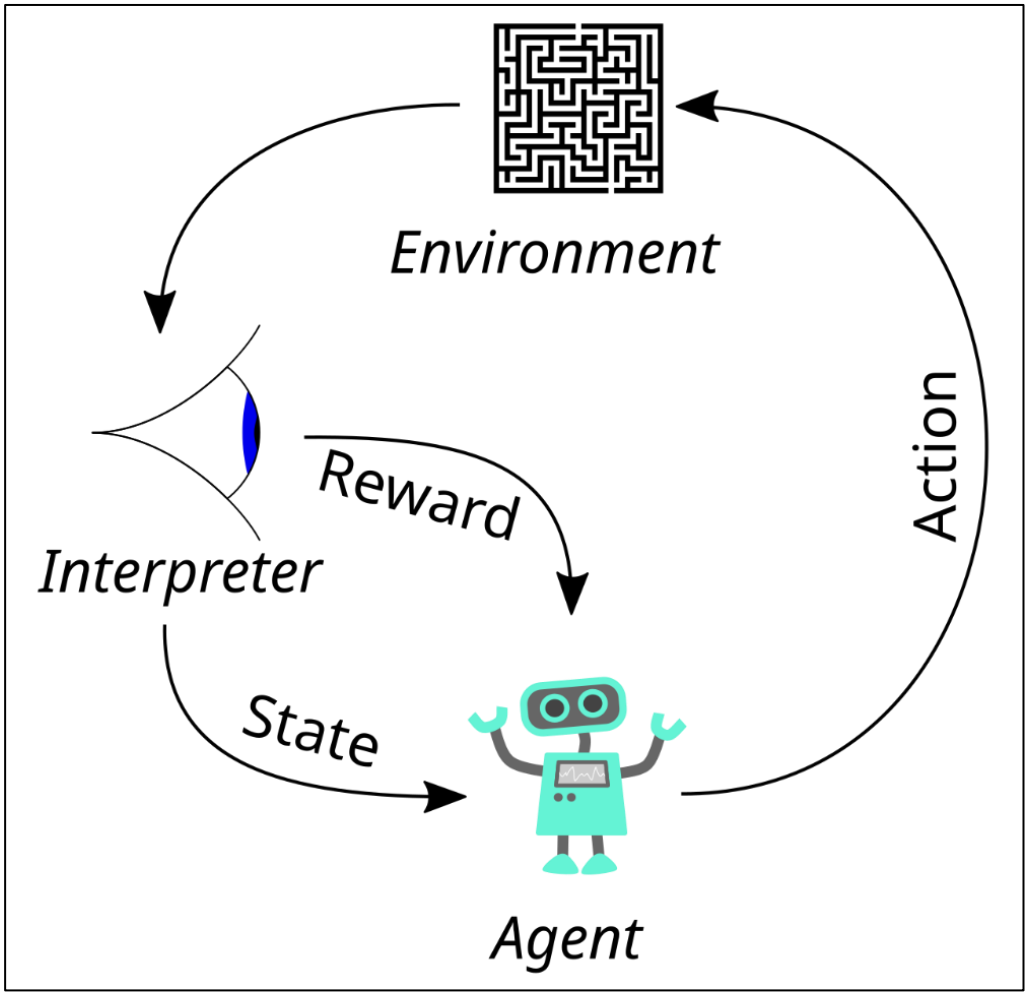


Zdroj: Wikimedia Commons [51].

3.5.3.3 Zesílené učení

Zesílené učení (Reinforcement Learning) nabízí odlišný přístup k učení modelů, který je založen na interakci tzv. agenta s prostředím. Agent provádí akce, které vedou k určitým následkům, a na základě obdržených odměn nebo trestů se učí optimalizovat své chování (Obrázek 5). Cílem agenta je optimalizovat své rozhodování tak, aby maximalizoval dlouhodobý souhrn odměn [52].

Obrázek 5 Schéma průběhu zesíleného učení



Zdroj: Wikimedia Commons [53].

Tento typ učení se využívá v různých oblastech, například při trénování algoritmů pro hraní her a simulací (např. AlphaGo), v robotice a autonomním řízení vozidel, při optimalizaci rozhodovacích procesů v reálném čase (např. v logistice nebo řízení výroby), či v adaptivních systémech. Mezi běžně používané algoritmy zesíleného učení patří Q-learning, Deep Q Networks (DQN) a Proximal Policy Optimization (PPO) [52].

Přestože tento přístup nabízí značný potenciál, čelí, stejně jako v předchozích druzích učení, několika výzvám. Jednou z hlavních je nalezení rovnováhy mezi objevováním nových strategií (exploration) a využíváním již osvědčených postupů (exploitation), tedy technik, které má agent v prostředí k dispozici. Dále je třeba počítat s vysokou výpočetní náročností a nutností spolehlivého běhu, neboť učení často probíhá v komplexních simulovaných prostředích, které jsou závislé na neustálém přístupu k nejnovějším datům z prostředí [47].

3.5.4 Klíčové modely

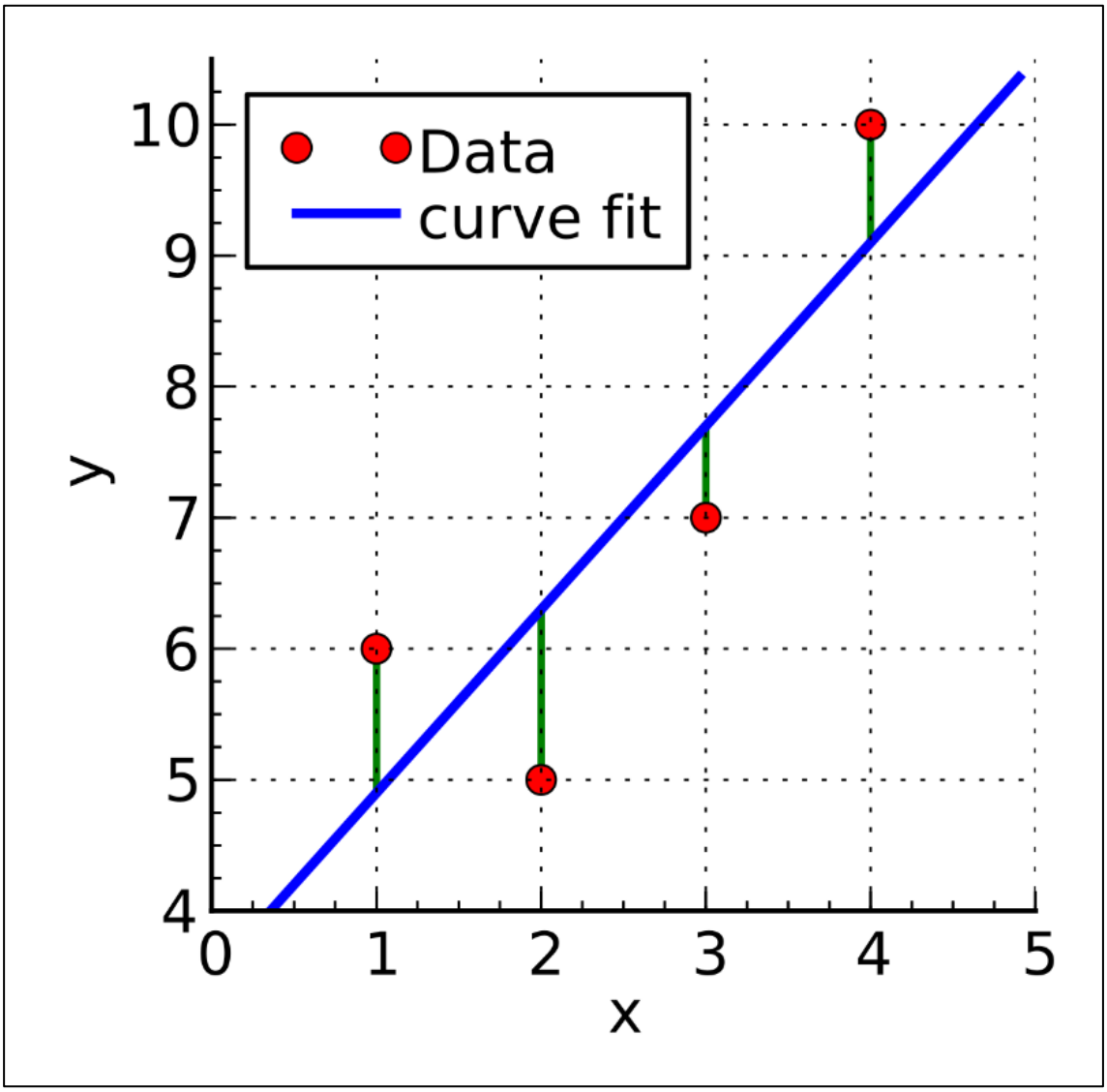
Algoritmy strojového učení představují základní nástroje, které umožňují modelům identifikovat vzorce a provádět predikce. Nejdůležitějším algoritmům se věnuje následující kapitola.

3.5.4.1 Lineární regrese

Lineární regrese je jedním ze základních algoritmů strojového učení a statistiky, který slouží k modelování vztahů mezi závislou proměnnou a jednou či více nezávislými proměnnými. Model předpokládá, že změna vstupních proměnných vede k proporcionální změně výstupní proměnné. Z matematického hlediska se vztah vyjadřuje lineární rovnicí, kde koeficienty jsou optimalizovány metodou nejmenších čtverců. Hlavní předností tohoto přístupu je snadná interpretace, protože získané koeficienty jasně ukazují, jak jednotlivé faktory ovlivňují výslednou hodnotu. Přestože jde o snadno použitelný nástroj, je nutné věnovat pozornost omezením metody – především předpokladu lineární závislosti mezi proměnnými, normálnímu rozdělení dat a absenci multikolinearity (vzájemné závislosti) mezi vstupními proměnnými. Lineární regrese je také poměrně citlivá na odlehlé hodnoty, které zkreslují výsledky modelu [46].

Na obrázku 6 je znázorněna metoda nejmenších čtverců, která se používá k nalezení regresní přímky minimalizací součtu čtverců reziduí. Přímka je proložena mezi naměřenými hodnotami (červené tečky) tak, aby rozdíl mezi predikovanými a skutečnými hodnotami byl co nejmenší. Odchylky (rezidua – zeleně) jsou umocněny na druhou, čímž se penalizují extrémní hodnoty [54].

Obrázek 6 Lineární regresní přímka metody nejmenších čtverců



Zdroj: Wikimedia Commons [54].

Implementace lineární regrese v Pythonu je relativně jednoduchá díky knihovně Scikit-learn. Používají se metody jako *LinearRegression()* pro trénink modelu a *predict()* pro predikci nových hodnot [42].

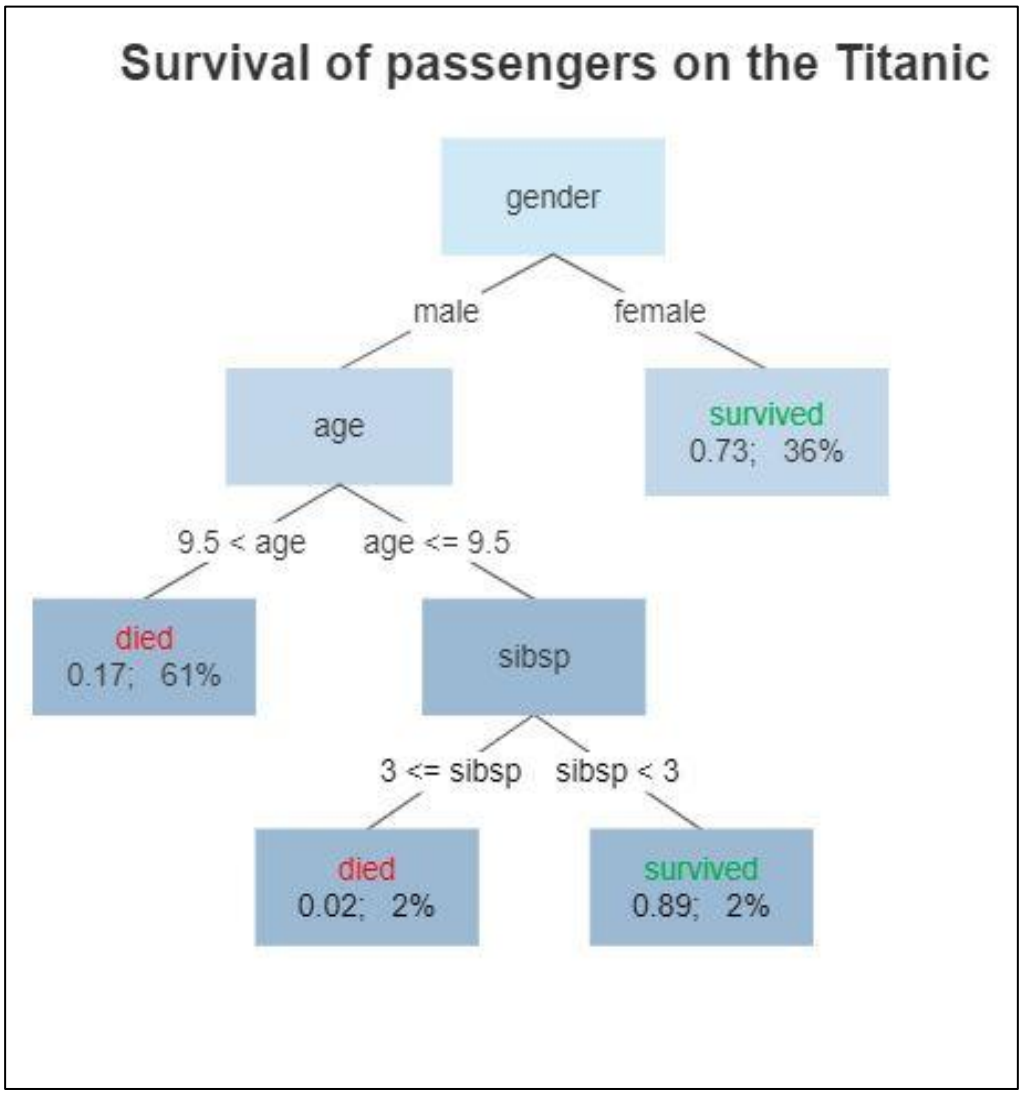
3.5.4.2 Rozhodovací stromy

Rozhodovací stromy jsou algoritmem strojového učení, který k analýze dat používá hierarchickou strukturu rozhodnutí. Každý uzel stromu představuje rozhodnutí založené na hodnotě konkrétního atributu, zatímco větve vedou k možným výsledkům rozhodnutí. Výsledkem je model, který lze snadno interpretovat jako sadu pravidel typu „if-then“ (vnořené cykly *if*). Díky své přehlednosti a srozumitelnosti jsou rozhodovací stromy vhodné pro aplikace, kde je důležitá transparentnost rozhodovacích procesů, tedy na základě jaké podmínky došlo k určitému rozdělení [32][46].

Navzdory své přehlednosti mají rozhodovací stromy některá omezení, z nichž nejvýznamnějším je tendence k přeučení (overfitting), kdy model ztrácí schopnost vyhodnocovat nová vstupní data, na kterých nebyl proveden trénink. Tento problém lze řešit omezením hloubky stromu nebo využitím metod zahrnujících predikci s využitím více stromů (např. Náhodné lesy), které vrací střední hodnotu z výsledků jednotlivých stromu, čímž zlepšují vypovídající hodnotu a přesnost modelu [46].

Obrázek 7 ukazuje rozhodovací strom přeživších na Titanicu. Hodnoty na listech ukazují pravděpodobnost přežití a procentuální zastoupení z celku. Z průchodu stromem je patrné, že větve, které mají největší šanci na přežití, jsou ženy (36 %) a muži pod 9,5 roku s méně než třemi sourozenci (*sibsp*) [55].

Obrázek 7 Ukázka rozhodovacího stromu



Zdroj: Wikimedia Commons [55].

Implementace v Pythonu se realizuje knihovnou Scikit-learn, pomocí tříd *DecisionTreeClassifier()* nebo *DecisionTreeRegressor()* pro klasifikační a regresní úlohy [42].

3.5.4.3 Prophet

Prophet je open-source model vyvinutý společností Meta (dříve Facebook) zaměřený na predikci časových řad. Je navržený tak, aby automaticky modeloval trendy a sezónnost v datech. Je implementován v jazycích R a Python [43].

Matematicky je Prophet založen na aditivním modelu, který rozkládá časovou řadu na tři hlavní složky: trend $g(t)$, sezónnost $s(t)$ a vliv svátků $h(t)$. Trendová složka g(t) modeluje dlouhodobé neperiodické změny a může být reprezentována jako lineární funkce nebo logistická křivka. Sezónní složka $s(t)$ zachycuje periodické vzory, jako jsou roční, týdenní nebo denní sezónnosti, a je modelována pomocí Fourierových řad. Vliv svátků $h(t)$ zohledňuje specifické události, které mohou ovlivnit hodnoty časové řady, například státní svátky nebo speciální akce [56].

Odhad parametrů v modelu Prophet je prováděn pomocí Bayesovského přístupu, konkrétně maximalizací aposteriorní (zjištěné) pravděpodobnosti (MAP). Tento přístup umožňuje začlenit apriorní (předchozí) informace o parametrech a poskytuje robustnost vůči chybějícím datům, odlehlým hodnotám a náhlým změnám trendu. Díky této flexibilitě je Prophet schopen efektivně modelovat složité časové řady s výraznými sezónními efekty a několika sezónami historických dat [56].

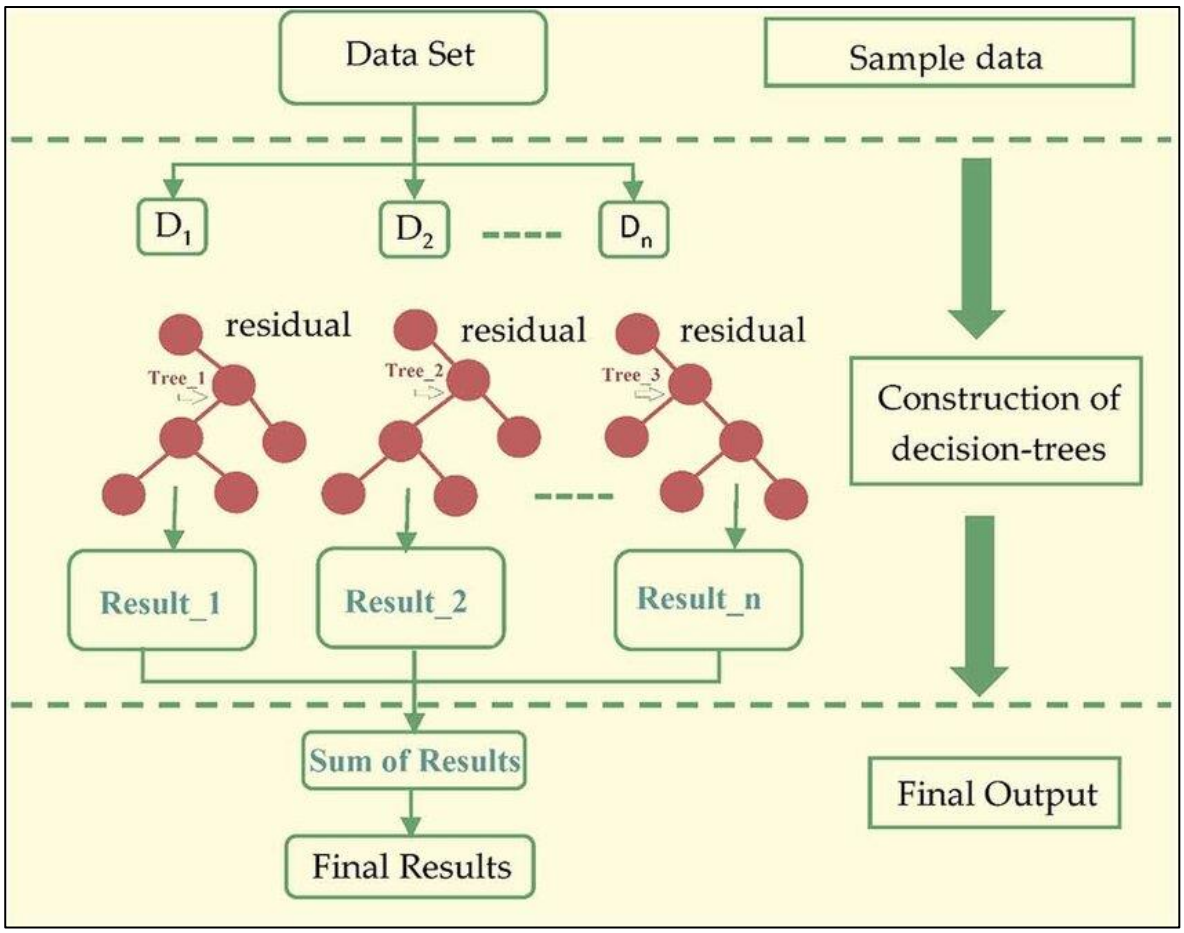
3.5.4.4 XGBoost

XGBoost, zkratka pro eXtreme Gradient Boosting, je model strojového učení navržený pro efektivitu, rychlost a vysoký výkon. Jedná se o open-source implementaci algoritmu gradientního boostingu, která využívá rozhodovací stromy jako základní modely. XGBoost je známý svou schopností zpracovávat velké objemy dat a poskytovat přesné predikce, Z tohoto důvodu byl tento algoritmus oblíbený nástroj v mnoha soutěžích (např. Kaggle) před nástupem složitějších modelů hlubokého učení [57].

Matematicky je XGBoost založen na technice gradientního boostingu, která kombinuje výstupy několika slabých modelů (např. jednoduchých rozhodovacích stromů) za účelem vytvoření silného prediktivního modelu (Obrázek 8). Každý nový strom je trénován tak, aby korigoval chyby předchozích stromů minimalizací specifikované ztrátové funkce pomocí gradientního sestupu. XGBoost navíc implementuje regularizační techniky ke snížení přefitování a zlepšení generalizace modelu [57].

Při použití XGBoost pro predikci časových řad je běžnou praxí převést problém na úlohu dohledu, kde se budoucí hodnoty predikují na základě minulých pozorování. To zahrnuje vytváření zpožděných proměnných a dalších relevantních rysů, které zachycují časové závislosti v datech. XGBoost byl úspěšně použit pro predikci nejen finančních časových řad, často jako hybridní model v kombinaci s tradičními přístupy, jako je ARIMAX, nebo s moderními metodami, jako jsou rekurentní neuronové sítě [57].

Obrázek 8 Princip fungování XGBoost modelu



Zdroj: Applied Artificial Intelligence [58].

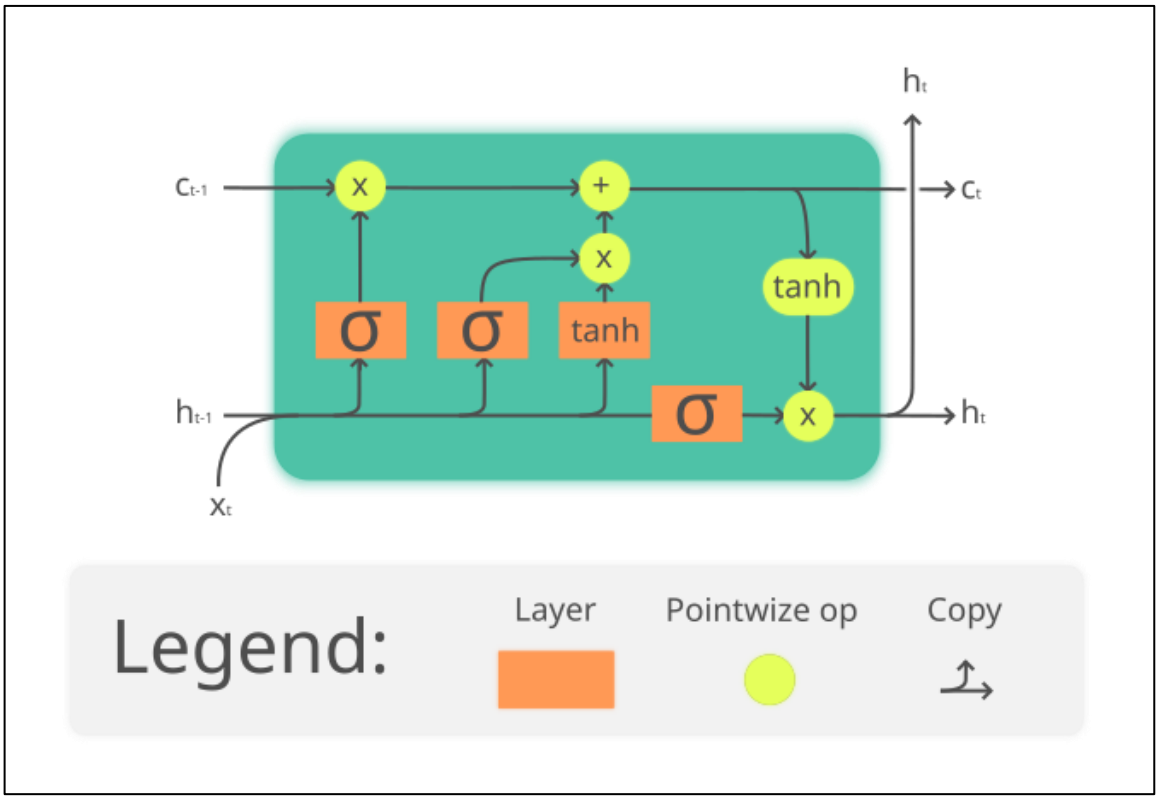
3.5.4.5 LSTM neuronová síť

Long Short-Term Memory (LSTM) je typ rekurentní (zpětné) neuronové sítě (RNN), navržený k překonání omezení tradičních sítí, zejména problému mizejícího gradientu („zapomínání“ informací, které se síť naučila před mnoha iteracemi), který brání efektivnímu učení dlouhodobých závislostí. LSTM sítě jsou schopny uchovávat informace

po dlouhé časové úseky, jsou tedy vhodné pro zpracování sekvenčních dat, jako je zpracování přirozeného jazyka, rozpoznávání řeči nebo predikce časových řad [59].

Architektura LSTM se skládá z buněk (cells) a tří typů bran: zapomínací brány (forget gate), která rozhoduje, které informace z předchozího stavu *c(t-1)* budou vymazány, vstupní brány (input gate), která určuje, jaké nové informace budou uloženy do paměti, a výstupní brány (output gate), která ovlivňuje výstupní skrytý stav *h(t)*. Aktivace těchto bran probíhá pomocí funcí sigmoid (σ) a hyperbolické tangens *(tanh)*. Paměťová buňka *c(t)* uchovává dlouhodobé informace a propojení mezi jednotlivými prvky zajišťují operace násobení, sčítání a kopírování, relevantní informace jsou tedy udrženy v paměti přes více časových úseků [59].

Obrázek 9 Schéma fungování LSTM neuronové sítě



Zdroj: Wikimedia Commons [60].

Díky své schopnosti zachytit dlouhodobé závislosti jsou LSTM sítě široce používány v úlohách, jako je strojový překlad, rozpoznávání řeči a predikce časových řad. Například při predikci časových řad mohou LSTM sítě modelovat komplexní vzory a trendy v datech, což umožňuje přesnější předpovědi [59].

4 Vlastní práce

Praktická část diplomové práce se věnuje procesu tvorby predikčních modelů aplikovaných na vybraný dataset. Modely jsou zde sestaveny a vyhodnoceny s cílem zodpovědět výzkumné otázky stanovené v cíli práce. Praktická část zahrnuje přehled použitého softwaru, zdůvodnění výběru datasetu, přípravu dat pro následnou analýzu, provedení deskriptivní analýzy a interpretaci získaných poznatků. Dále následuje tvorba samotných predikčních modelů, včetně podrobného rozboru implementovaného kódu v programovacím jazyce Python. Závěrem praktické části je provedeno porovnání dosažených výsledků, jejich vyhodnocení a formulace závěrů odpovídajících stanoveným cílům práce.

4.1 Použitý software

Následující podkapitola obsahuje výčet software použitého pro programování, deskriptivní analýzu, vizualizace, trénování predikčních modelů a dosažení interpretovatelných výsledků.

4.1.1 Programovací jazyk

Programovací jazyk použitý pro kompletní praktickou implementaci je Python, konkrétně jeho verze 3.11.11. Tato verze je využívána z důvodu výběru vývojového prostředí, které momentálně zmíněnou verzi používá, ačkoliv nejde o nejaktuálnější dostupnou verzi jazyka. V práci je využíván Python framework TensorFlow a vícero podpůrných knihoven, které byly popsány v kapitole 3.4.3.

4.1.2 Vývojové prostředí

Pro implementaci predikčních modelů v programovacím jazyce Python bylo zvoleno internetové vývojové prostředí Google Colab. Toto prostředí poskytuje uživatelům přístup k virtuálním strojům včetně možnosti využití GPU a TPU jednotek, což výrazně urychluje trénování modelů strojového učení. Google Colab je založen na populární webové aplikaci Jupyter Notebook, která umožňuje efektivní integraci a kombinaci textových a kódových bloků, jež lze spouštět nezávisle na sobě. Tento přístup zajišťuje přehlednost a flexibilitu, obzvláště při práci s více grafy nebo modely, bez nutnosti dělení kódu do samostatných skriptů. Výsledné soubory vytvořené v prostředí Jupyter Notebook používají příponu *.ipynb*.

4.2 Výběr datového setu

Tato kapitola popisuje vybraný set dat, který je dále použit pro zpracování dat a tvorbu prediktivních modelů. Na dataset byly kladeny následující nároky:

- schopnost uspokojivě odpovědět na otázky vytyčené v cíli práce – dataset by měl obsahovat takový objem dat, jejichž úpravou a predikčním modelováním lze dosáhnout interpretovatelných výsledků
- aktualita dat – data by měla být aktuální (alespoň v klíčových faktorech), jelikož v opačném případě nelze účinně predikovat budoucí vývoj, jelikož je nutné chybějící data „dopočítat“, tedy predikce je výrazně méně přesná
- důvěryhodný zdroj dat – zdroji dat by měly být pouze důvěryhodné instituce, které se zabývají problematikou Covid-19 a jejich datasety jsou dostatečně robustní (např. WHO, ECDC, CDC, ÚZIS, John Hopkins University atd.)
- úplnost dat – data by měla být úplná alespoň v takovém rozsahu, aby bylo možné realizovat predikci bez nutnosti velkých zásahů do datasetu, které by snižovaly vypovídající hodnotu výsledků modelů
- strukturovanost dat – dataset by měl být dobře strukturovaný, tedy rozdělený podle zemí a času, se správně pojmenovanými a zdokumentovanými proměnnými
- dostupnost dat – data by měla být snad dostupná, nejlépe jako .csv soubor či API

Na základě kladených nároků byl vybrán datový set zpracovaný neziskovou organizací Global Change Data Lab, v rámci projektu „*Our World in Data*“. Ten obsahuje většinu potřebných informací a metrik pro provedení další analýzy a výsledného modelování [61].

Následující tabulka (Tabulka 1) znázorňuje dostupné metriky v datasetu *Our World in Data*. U každé metriky je uveden zdroj dat a počet zemí, pro které jsou tyto údaje k dispozici. Zahrnuje informace o očkování, testování, hospitalizacích, potvrzených případech a úmrtích, dále také reprodukční číslo, vládní opatření a další proměnné zájmu, jako jsou socioekonomické ukazatele z mezinárodních organizací. Ten kombinuje data z vícero zdrojů v jednom přehledném **datasetu**: obsahuje data z WHO, OECD, World Bank, ECDC doplněné o originální data sesbíraná v rámci OWID a dalších podpůrných zdrojů. Důležité části datasetu budou v další kapitole blíže představeny, včetně jednotlivých proměnných, které budou tvořit deskriptivní analýzu šíření nemoci.

Tabulka 1 Data a jejich zdroje v datasetu OWID

Metrika	Zdroj dat	Počet zemí
Očkování	Oficiální data shromážděná týmem OWID	218
Testování a pozitivita	Oficiální data shromážděná týmem OWID	193
Hospitalizace a JIP	Oficiální data shromážděná týmem OWID	46
Potvrzené případy	WHO COVID-19 data	219
Potvrzená úmrtí	WHO COVID-19 data	219
Reprodukční číslo (R)	Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C	196
Vládní opatření	Oxford COVID-19 Government Response Tracker	185
Další metriky a proměnné	OSN, Světová banka, OECD, IHME atd.	219

Zdroj: OWID [61].

4.3 Deskriptivní analýza

Tato kapitola popisuje fázi práce s daty před zahájením samotného prediktivního modelování. Jde o nezbytnou součást celého procesu datového modelování a datové analýzy, kde budou neošetřená data z vybraného datasetu připravena pro snadnější vizualizaci a modelování a budou analyzovány jejich aspekty, které povedou k pochopení šíření nemoci a jejich důsledků. Kapitola obsahuje pouze výsledky deskriptivní analýzy, celý kód v jazyce Python je vzhledem ke své délce součástí příloh práce (Příloha A). Zobrazená velikost grafů je přizpůsobena velikosti stránky, grafy v plné velikosti jsou součástí zdrojového kódu a jsou také vloženy do příloh práce (Příloha B).

4.3.1 Příprava dat

Ve fázi přípravy jsou data načtena, jsou zjištěny základní informace o datasetu, data jsou prozkoumána a připravena pro další použití. Veškerá data pocházejí z vybraného datasetu Our World in Data (OWID), s výjimkou dat ohledně HDP, které dataset neobsahuje v dostatečné míře pro provedení ekonomické analýzy. Pro doplnění dat byly použity data z datasetu statistického úřadu EU, Eurostat [63]. Příprava dat zahrnuje následující kroky:

1. **Načtení dat**: Data jsou načtena přímo ze stránek OWID a načtena pro potřeby analýzy do dataframe poskytovaného knihovnou Pandas. Datový set obsahuje (převážně) kompletní časové řady údajů týkajících se Covid-19 pro všechny světové regiony.
2. **Struktura dat**: Zde jsou zobrazeny základní informace o povaze datasetu: dataset má tvar (485796, 61), tedy obsahuje 485 796 záznamů (řádků) a 61 proměnných (sloupců). Dále jsou vypsány názvy všech dílčích proměnných.

3. **Klíčové sloupce**: Zde jsou identifikovány důležité proměnné pro další analýzu (Tabulka 2).

Tabulka 2 Klíčové proměnné datasetu OWID

Název proměnné	Datový typ	Popis
country	object	název dané země
date	object	datum záznamu
total_cases	float64	celkový počet případů
new_cases	float64	nový počet případů
new_deaths	float64	nový počet úmrtí
population	float64	populace dané země
gdp_per_capita	float64	HDP na obyvatele
continent	object	kontinent
HDI	float64	Index lidského rozvoje
stringency_index	float64	Index přísnosti opatření
people_vaccinated	float64	počet očkovaných osob

Zdroj: vlastní zpracování, OWID [61].

4. **Chybějící hodnoty**: V tomto kroku jsou identifikovány chybějící hodnoty v klíčových sloupcích. Je zjištěno, že ve sloupci „people_vaccinated“ je téměř 84 % chybějících hodnot, proto nebude v další analýze použit.

5. **Transformace dat**: Zde je datový sloupec převeden z typu „object“ na typ „datetime“ a jsou vytvořeny dodatečné časové proměnné (měsíc, rok a měsíc_rok) pro potřebu podrobnější časové analýzy v pozdějších krocích.

6. **Nová proměnná**: Pro potřebu další analýzy je vytvořena analytická proměnná „míru úmrtnosti“, vypočítaná z celkového počtu případu a celkového počtu úmrtí.

7. **Nahrazení chybějících hodnot**: Zde jsou nahrazeny chybějící hodnoty pomocí metody Pandas knihovny *.ffill()* neboli forward fill (dopředné

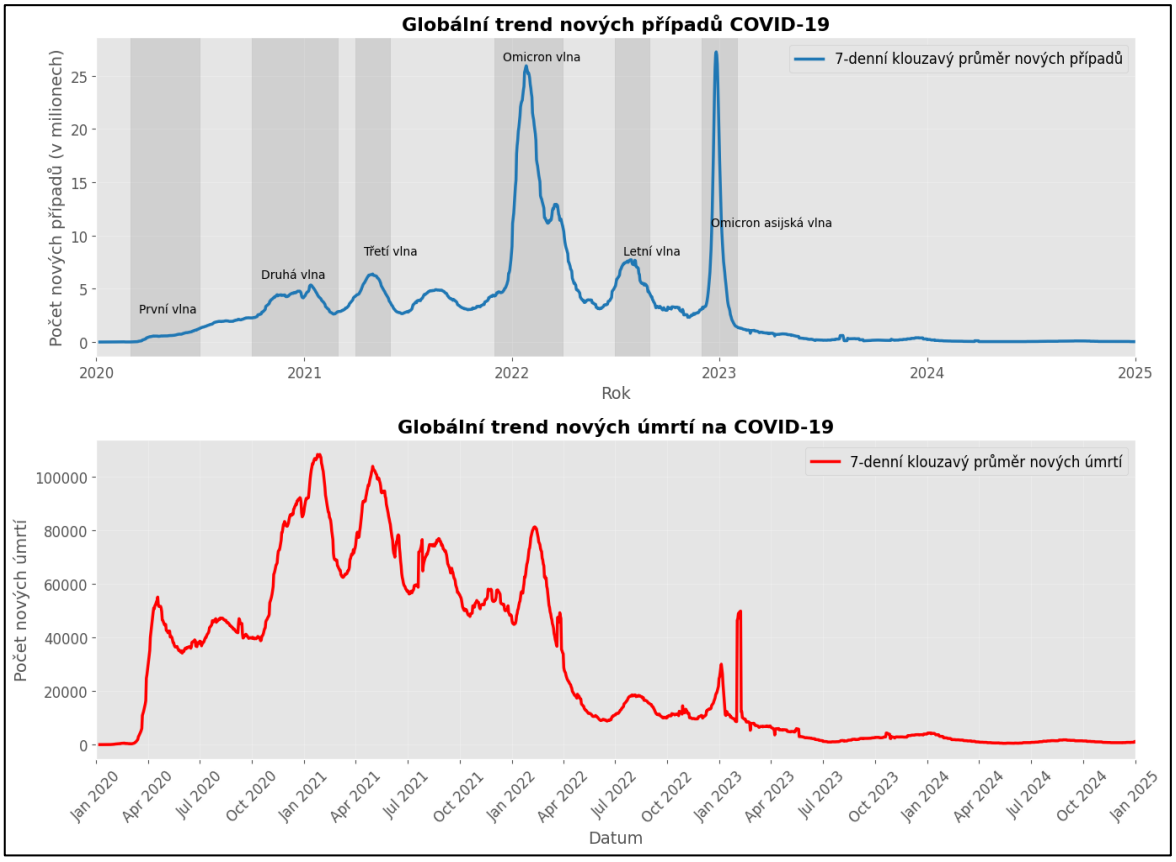
vyplnění) – ta funguje na principu nahrazení chybějících hodnot pomocí poslední známé hodnoty v daném sloupci.

8. **Datová sada nejnovějších údajů**: Na závěr přípravy dat je vytvořena datová sada obsahující nejnovější dostupné informace o daném státu „latest_by_country“.

4.3.2 Analýza šíření Covid-19 ve světě

V této části se deskriptivní analýza věnuje formulací odpovědi na otázku „jak se nemoc šířila ve světě?“. Graf 1 zachycuje globální trendy v počtu nových případů a úmrtí spojených s COVID-19. Ilustruje, jak se nemoc šířila ve větších vlnách, které většinou souvisí s nástupem nové, nakažlivější varianty (např. Delta, Omicron apod.). V horním grafu jsou jednotlivé vlny a délka jejich trvání vyznačeny šedivou barvou. Pro vyhlazení nepravidelností (lokální faktory, nepravidelné reportování atd.) je na graf zanesen 7denní klouzavý průměr nových případů či úmrtí.

Graf 1 Globální trendy případů a úmrtí nemoci COVID-19 (2020–2024)



Zdroj: vlastní zpracování, OWID [61].

Z grafu případů je patrné, že covid se, převážně během let pandemie (2020–2023), šířil ve větších vlnách, které souvisely se sezónními výkyvy (např. podzim v Evropě) nebo s nástupem nových variant. Například jde o nástup varianty Omicron v roce 2022 či rozšíření nové subvarianty v Asii v roce 2023, kdy nové případy dosahovaly v průměru 25 milionů za týden.

Zároveň je možné pozorovat, že počínaje rokem 2023 se globální případy nemoci začaly rapidně snižovat. To může mít několik důvodů: s oficiálním ukončením pandemického stavu v roce 2024 začaly snižovat úsilí věnované na trasování a testování nových případu, tedy i oficiální číslo nových případů začalo rapidně klesat. Dále s nástupem variant s větší nakažlivostí a rychlejším nástupem je procento zachycených případů jenom velmi malé. Skutečné hodnoty budou pravděpodobně o mnoho vyšší, jak ukazují i analýzy výskytu viru Covidu v odpadních vodách, které v ČR provádí např. Výzkumný ústav vodohospodářský TGM Brno, nebo v USA státní úřad CDC [64].

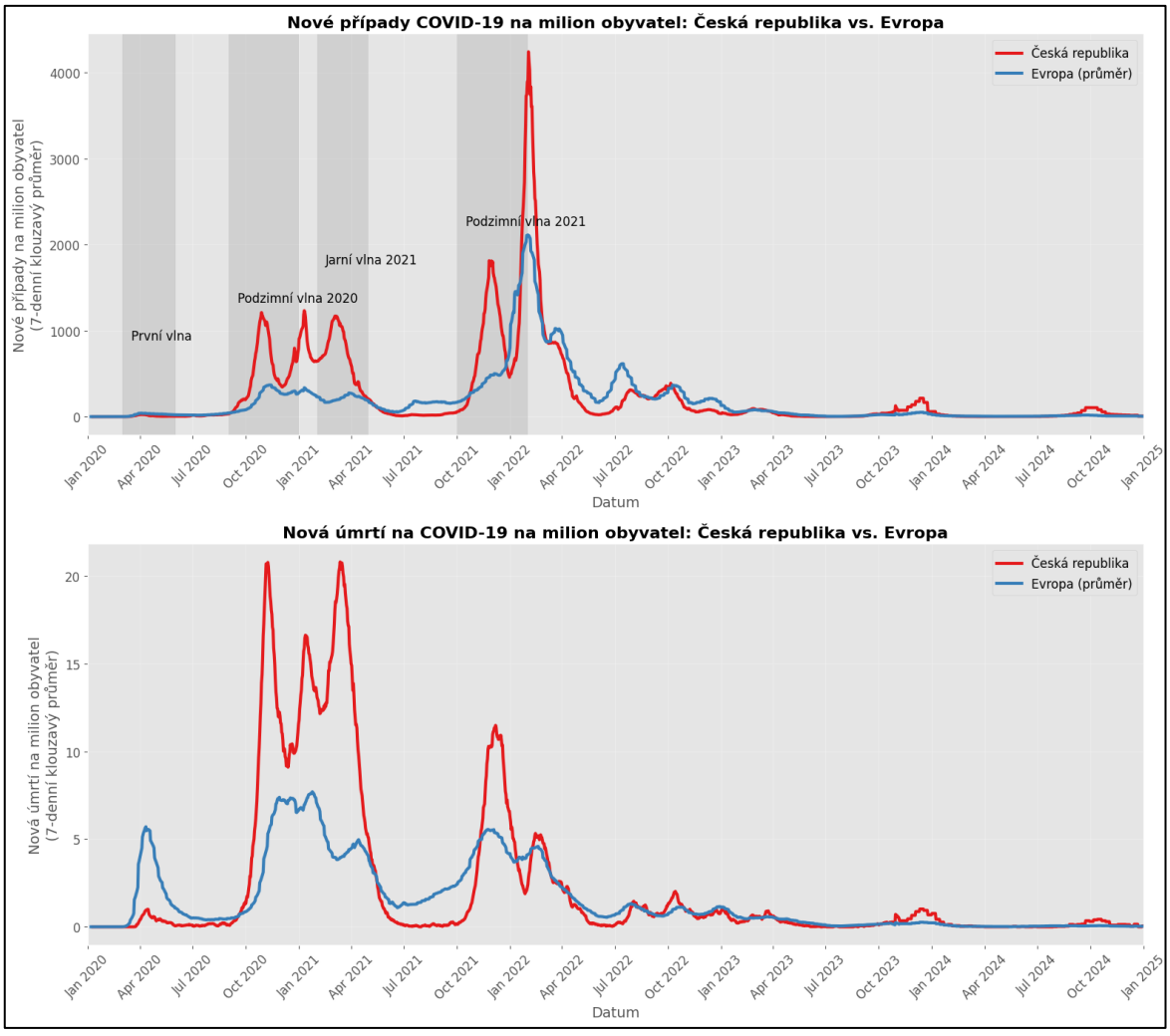
Globální trendy nových úmrtí spojených s virem COVID-19 se pohybovaly v největších číslech spíše v prvních letech pandemie (převážně rok 2021), tedy v době, kdy ještě neproběhla ve většině státech efektivní očkovací strategie. Tendence úmrtí je tedy klesající, s maximem v lednu 2021. Pozdější výrazné vlny nových případů se projevily i na dočasném zvýšení novým úmrtí, nicméně smrtnost se, hlavně v roce 2023 a 2024, snížila na minimum oproti počátkům pandemie.

4.3.3 Šíření nemoci v ČR

Následující kapitola se věnuje šíření nemoci v České republice, novým úmrtím v průběhu pandemie a jejich srovnání se zbytkem Evropy. Přináší nejen srovnání ČR s evropským průměrem, ale také porovnaní s okolními evropskými zeměmi.

Graf 2 popisuje analogicky k prvnímu grafu trendy šíření a úmrtí v Evropě a v České republice. V části zachycující nové případy na milion obyvatel jsou vyznačeny jednotlivé vlny nemoci.

Graf 2 Trendy šíření a úmrtí COVID-19 v Česku a Evropě (2020–2024)



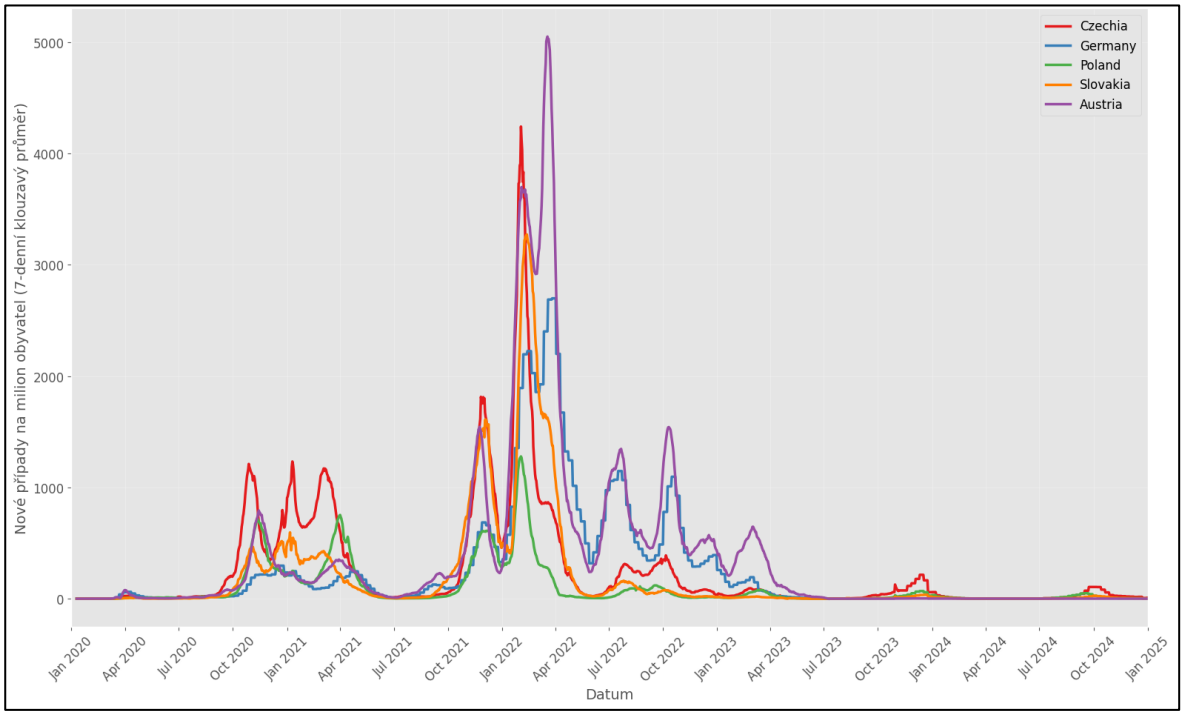
Zdroj: vlastní zpracování, OWID [61].

V grafu nových případů lze pozorovat, že Česká republika měla na poměry Evropy větší výkyvy případů a vyšší maxima, než byl evropský průměr. To mohlo být způsobeno chaotickými vládními opatřeními, menší připraveností státu na pandemii, či zrychlenému

rozvolnění v roce 2021, kde „dvojitá“ vlna v ČR dosáhla svému maxima za celou dobu pandemie. Na části nových úmrtí lze pozorovat obdobný trend, tedy že vyšší počet případů měl za následek vyšší míru úmrtí, převážně v prvních letech pandemie (2020-21), tedy ještě před zahájením očkovací strategie. Dále již ČR příliš nevybočuje z průměru, pouze v rekordní vlně na přelomu rok 2021 úmrtí opět na nějakou dobu stoupla.

Poslední z grafů (Graf 3) nabízí srovnání České republiky se sousedními státy, jmenovitě Německem, Polskem, Slovenskem a Rakouskem. Pro zobrazení je opět použit 7denní klouzavý průměr.

Graf 3 Nové případy COVID-19 na milion obyvatel: ČR a sousední země (2020–24)



Zdroj: vlastní zpracování, OWID [61].

Zde můžeme opět pozorovat výraznější nárůst případů během let 2020 a 2021 oproti ostatním státům. Naopak ve výrazné vlně v zimě 2021 a jaře 2022 dosahovalo nejvyšších průměrných případů Rakousko. Šlo až o 5000 nových případů na milion obyvatel. Také další průběh roku 2022 byl v ČR „klidnější“ než v Německu, či Rakousku. S koncem pandemie je zde, stejně jako ve světě, možné pozorovat značné snížení počtu nahlášených případů, které může být opět způsobeno nižší mírou provedených testů a rychlejším nástupem onemocnění, případně také menší dostupnosti profesionálních (PCR) testů, které by mohly nemoc odhalit.

Statistické shrnutí pro ČR:

- Celkem případů: 4 828 163
- Celkem úmrtí: 43 807
- Počet případů na milion ob.: 452 363 (45 %)
- Počet úmrtí na milion ob.: 4104
- Plně očkovaní (%): 64,61

Tabulka 3 zachycuje porovnání pandemických metrik pro Českou republiku a evropský průměr. Zde je možné znovu pozorovat, že Česká republika patřila v rámci Evropy k jedné z více zasažených zemí, což se projevilo na větším počtu případů, ale hlavně smrtí (na milion obyvatel), které jsou přibližně o 50 % vyšší než evropský průměr. Míra smrtelnosti v ČR se však příliš neliší, pokud je počítáno s celkovým počtem případů pro Evropu a pro ČR.

Tabulka 3 Porovnání statistických údajů ČR s evropským průměrem (2025)

Metrika/Lokalita	Česká republika	Evropský průměr
Počet případů na milion ob.	452 363	402 831
Počet úmrtí na milion ob.	4104	2679
Míra smrtelnosti (%)	0,91	0,83

Zdroj: vlastní zpracování, OWID [61].

Následující kapitola přinesla odpovědi o šíření covidu v Evropě a přímo v České republice, což je důležité pro pochopení nemoci v kontextu místního prostředí a možném porovnání situací v nejbližším okolí.

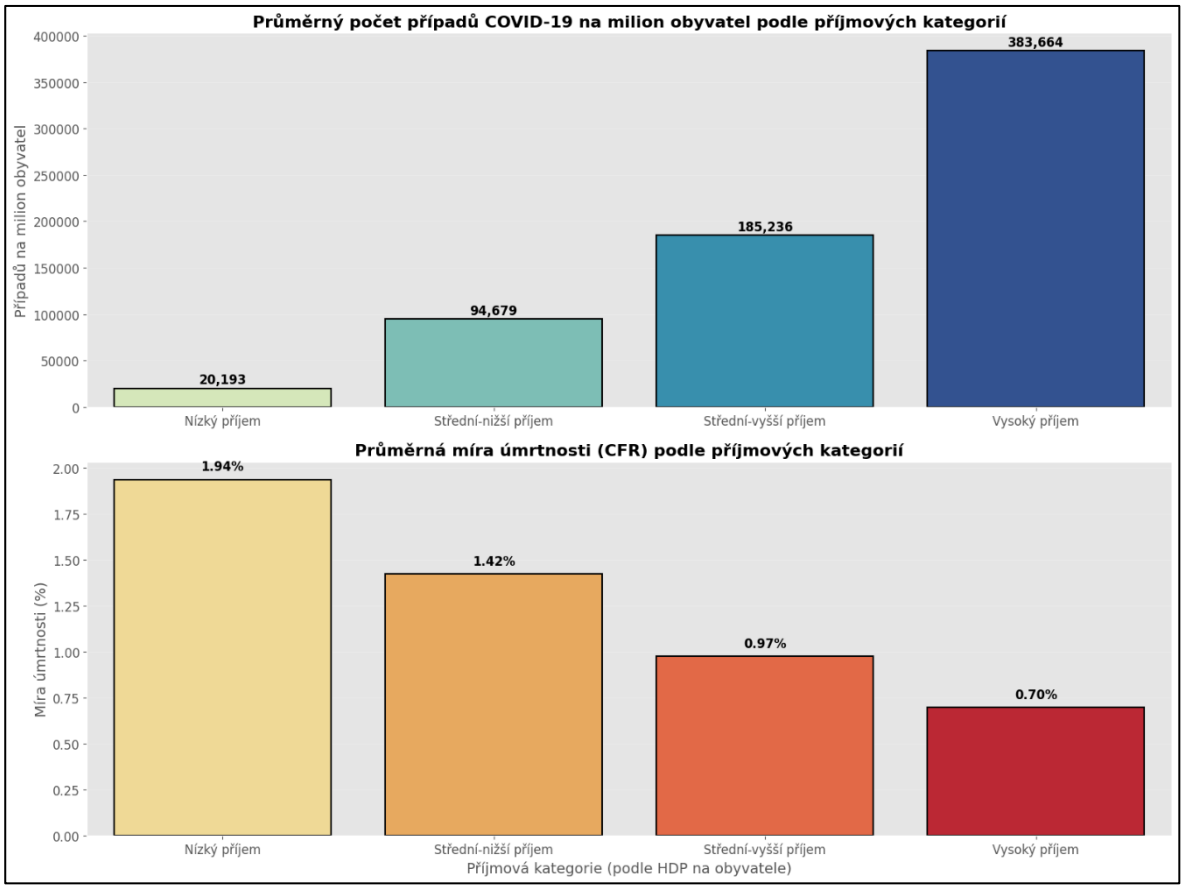
4.3.4 Analýza rozdílů dle vyspělosti, opatření a očkování

Tato kapitola se zabývá analýzou šíření a dopadů viru při rozdělení států do kategorií podle příjmu, přísnosti opatření vyjádřené indexem přísnosti opatření a dle míry proočkovanosti.

Graf 4 se zabývá porovnáním počtu nových případů a míry úmrtnosti napříč státy s různou velikostí příjmů. Příjmové skupiny byly rozděleny následovně podle HDP na obyvatele:

- Nízký příjem: 0–5000$
- Střední nižší příjem: 5000–15 000$
- Střední vyšší příjem: 15 000–30 000$
- Vysoký příjem: 30 000$ a více

Graf 4 Průměrné případy a úmrtí podle příjmových kategorií



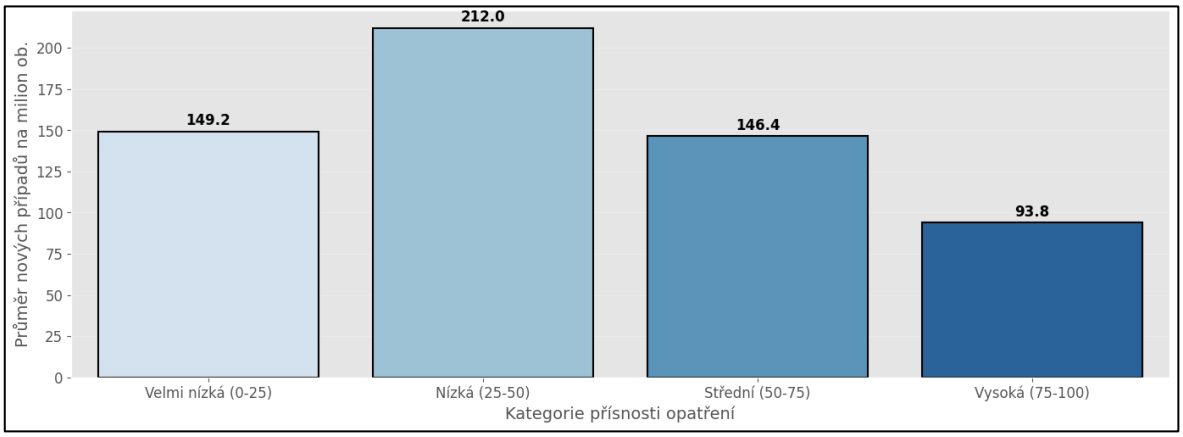
Zdroj: vlastní zpracování, OWID [61].

Z vrchního grafu je patrné, že průměrný počet případů rostl zároveň s příjmem, tedy největší průměrný přírůstek případů vykazuje skupina „Vysoký příjem“. To může být způsobeno na jedné straně zodpovědnějším přístupem k testování a trasování šíření viru, kdy nižší příjmové skupiny států často nepřikládaly testování a hlášení případů příliš velkou váhu, či na straně druhé nižší mírou globalizace a urbanizace v nižších příjmových skupinách, a tedy menší možnosti viru se efektivně rozšířit velkou rychlostí v populaci.

Ve spodním grafu můžeme naopak vidět, že největší míra úmrtnosti na COVID-19 byla ve státech s nejnižšími příjmy. Jde o logický vývoj, kdy státy s nižším příjmem mají méně vyvinutou nemocniční a lékařskou infrastrukturu, nemohou si dovolit investovat do drahých přístrojů a nakupovat přednostně očkovací vakcíny. Také dostupnost lékařské péče bývá většinově nižší než ve vyspělém světě.

Další hledisko, ze kterého je možné posuzovat přírůstky nových případů, je přísnost opatření v dané zemi. To je vyjádřeno Indexem přísnosti vytvořeným týmem vědců při Oxfordské univerzitě v rámci programu „Oxford COVID-19 Government Response Tracker“. Index rozděluje státy podle přísnosti zavedených opatření na potlačení šíření viru na stupnici 1-100 bodů. V následujícím grafu (Graf 5) byly státy rozděleny do skupiny podle hodnoty indexu přísnosti opatření na čtyři skupiny.

Graf 5 Průměrný počet nových případů COVID-19 podle přísnosti opatření

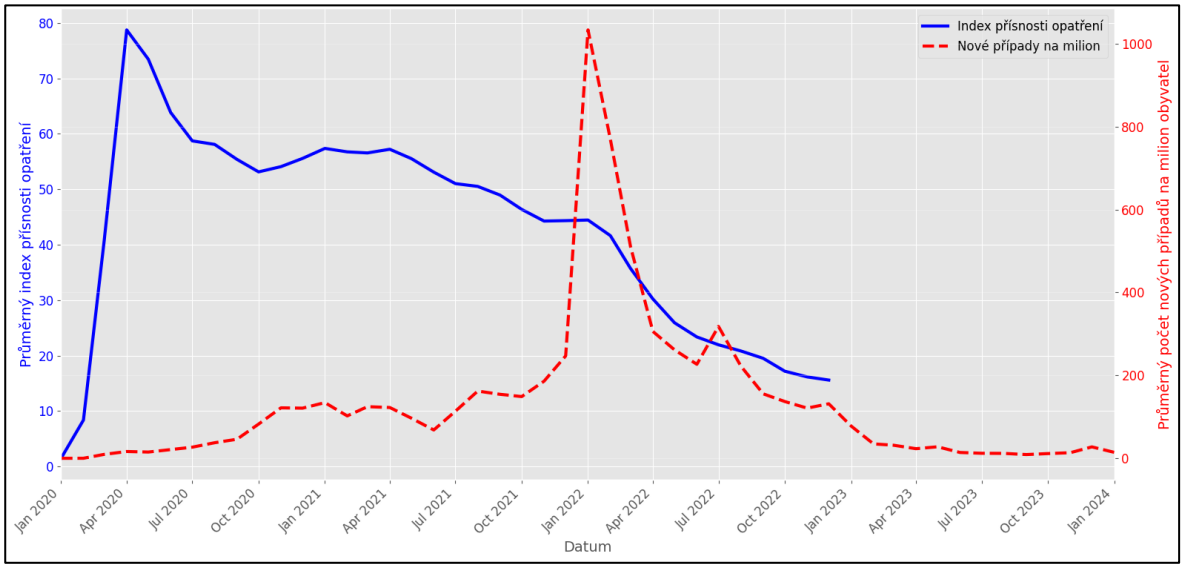


Zdroj: vlastní zpracování, OWID [61].

V grafu je možné pozorovat, že s přibývající přísností opatření klesal průměr nových případů, zavedení účinných a smysluplných opatření tedy přispělo k snížení počtu nových případů nemoci ve skupině států. Pouze státy s velmi nízkou hodnotou Indexu přísnosti tento trend nepotvrzují, zde ale může jít opět o zkreslení vlivem nedostatečného trasování nemoci.

Následující graf (Graf 6) navazuje na předchozí téma Indexu přísnosti opatření. Zachycuje vztah mezi zprůměrovanou hodnotu indexu od ledna 2020 do ledna 2023 a globálním průměrným počtem nových případů na milion obyvatel (2020–2024).

Graf 6 Vztah mezi Indexem přísnosti opatření a novými případy COVID-19 ve světě



Zdroj: vlastní zpracování, OWID [61].

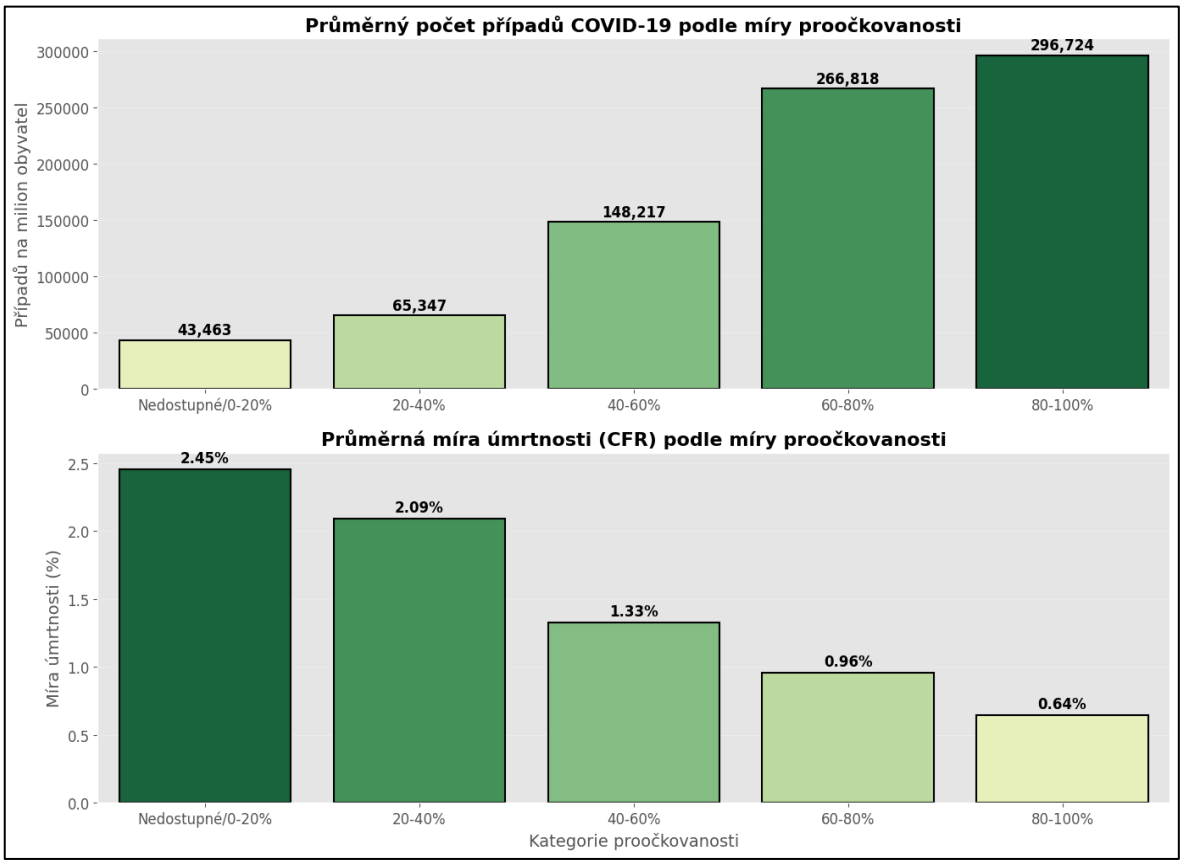
V grafu je možné pozorovat, že v počátcích pandemie, kdy státy nasadily velmi přísné opatření, se počet nových případů držel relativně nízko, s postupem času, kdy státy začaly postupně uvolňovat opatření (převážně v roce 2022), se počet případů strmě zvýšil, aby dosáhl svého maxima v lednu 2022.

Nárůst případů je možné vysvětlit tím, že virus měl při snížených opatřeních mnohem větší šanci se šířit mezi širší populací, jelikož byly znovuotevřeny školy, zrušeny omezení setkávání, sníženy různé preventivní mechanismy proti šíření nemoci, a další. Toto širší rozvolnění pravděpodobně přišlo příliš brzy a umožnilo viru se šířit s nevídanou rychlostí. S tím souvisí i vznik nových, nakažlivějších mutací, které znovu uspíšily šíření viru.

I přes další rozvolňovaní ale trend nových případů s postupem roku 2022 a 2023 dále klesal. Index přísnosti opatření přestal být aktualizován v lednu roku 2023, jelikož už nezbývalo mnoho státu, které by měly stále zavedená rozsáhlá opatření (s výjimkou např. Číny). Tento optický pokles nových případů je možné opět přisoudit snížení testování a trasování s vyhlášením konce pandemie v květnu roku 2023.

Poslední sada grafů (Graf 7) rozděluje státy do skupin dle míry proočkovanosti (%) a zkoumá průměr případů na milion obyvatel a míru úmrtnosti v jednotlivých skupinách. Míra proočkovanosti většinou úzce koresponduje s vyspělostí daného státu, je tedy možné očekávat podobné výsledky, jako v případě analýzy na základě příjmových skupin.

Graf 7 Globální průměr případů a úmrtnosti podle míry proočkovanosti



Zdroj: vlastní zpracování, OWID [61].

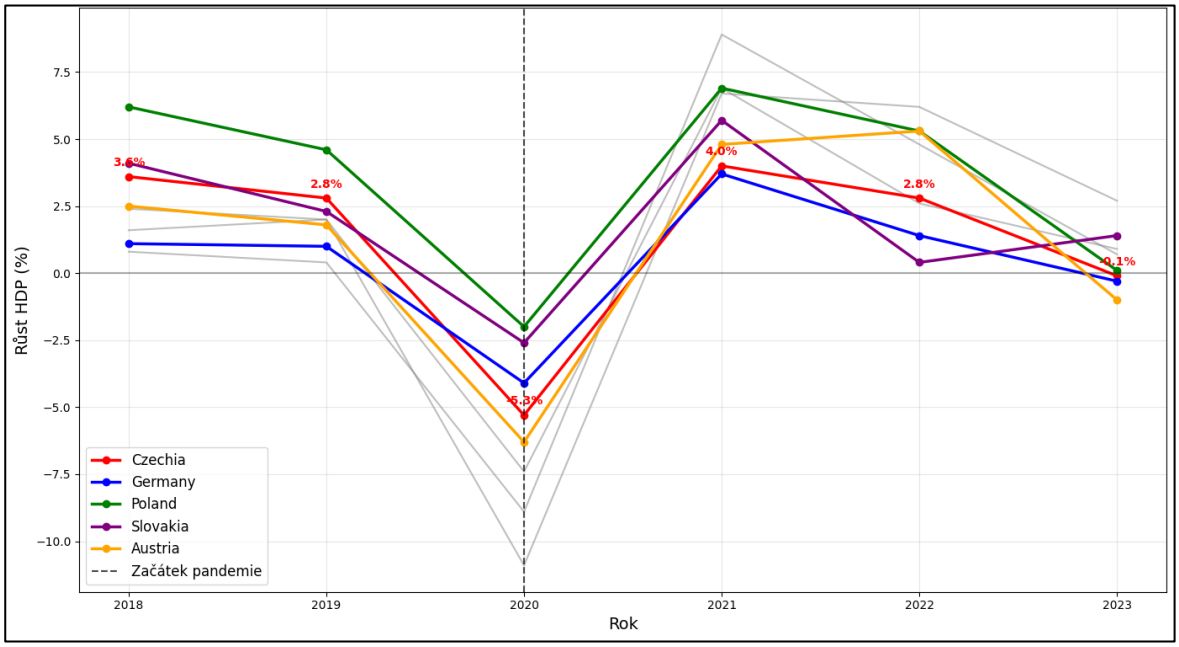
Jak bylo řečeno výše, průměrný počet případů na milion obyvatel i míra úmrtnosti kopírují trend pozorovaný v rámci analýzy příjmových skupin. Nejvíce případů paradoxně vykazují státy, kde je míra proočkovanosti nejvyšší. To může být způsobeno nejpřesnějším hlášením nových případů, či největší mírou globalizace a urbanizace v této skupině, a tedy i možností rychlého šíření viru. Naopak míra úmrtnosti je v této skupině nejnižší, z grafu tedy vyplývá, že očkovací strategie a plná proočkovanost obyvatel vede ke značnému snížení úmrtnosti, a to až o dva procentní body oproti státům, které očkovací strategii nemají. Dále k menší úmrtnosti pravděpodobně přispěl rozvinutější systém zdravotnictví a dostupnější zdravotní péče, kterou státy s efektivní očkovací strategií vykazují.

4.3.5 Analýza ekonomických dopadů

V poslední části deskriptivní analýzy je zkoumáno, jaký měla pandemie COVID-19 dopad na světovou ekonomiku skrze sledování vývoje růstu HDP. Pro názornost je vybrána pouze Česká republika a další evropské státy, na kterých jsou změny ilustrovány. Pro doplnění dat o hodnoty HDP v jednotlivých letech byl použit dataset Eurostatu, který obsahuje přesné roční údaje všech zkoumaných států [63].

Graf 8 zachycuje změny v meziročním růstu HDP před pandemií (2018–2019), v pandemických letech (2020–2022) a v postpandemickém roce 2023. Česká republika a její sousedící státy jsou zobrazeny barevně, ostatní státy z evropského výběru jsou sdruženy do šedivých čar.

Graf 8 Vývoj růstu HDP v ČR a okolních zemích (2018–2023)

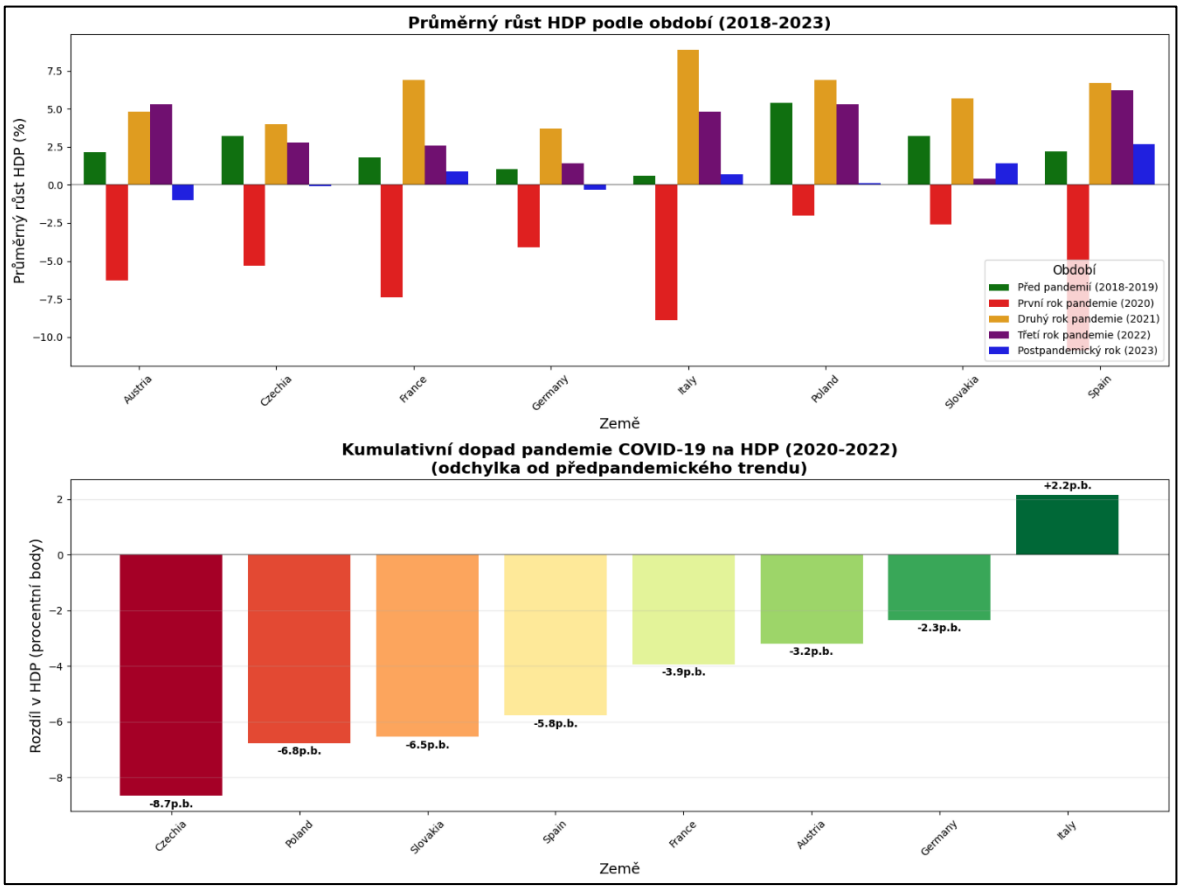


Zdroj: vlastní zpracování, Eurostat [63].

V grafu je jasně patrné, že první rok pandemie se negativně podepsal na všech vybraných ekonomikách, které bez výjimky vykazují negativní růst HDP oproti předchozím rokům. Například Česká republika vykázala v roce 2020 růst HPD -6,3 %. V následujícím roce se všem sledovaným státům podařilo strmý pokles zastavit a vykazovaly opět růst, nicméně dlouhodobý ekonomický trend byl spíše klesající a státy jako Česká republika nebo Německo měly problém se vrátit na předpandemickou úroveň. Touto problematikou se dále zabývá následující graf.

Ekonomickou analýzu uzavírá poslední dvojice grafů (Graf 9) týkající se průměrného růstu HDP rozděleného podle období před pandemií, letech pandemie a postpandemického roku (vrchní část). Spodní část se zabývá znázorněním kumulativního dopadu pandemie na jednotlivé vybrané státy pomocí odchylky HDP od předpandemického trendu (p.b.).

Graf 9 Změny HDP pro vybrané země během pandemie COVID-19



Zdroj: vlastní zpracování, Eurostat [63].

V grafu průměrného růstu HDP je možné pozorovat stejný propad ekonomik evropských států v prvním roce pandemie jako v Grafu 8. Největší propad zaznamenalo Španělsko a Itálie. Naproti tomu v grafu odchylek je možné pozorovat, že největší odchylku od předpandemického trendu měla Česká republika, pro kterou pandemie přinesla nečekaný konec ekonomického růstu, což mělo silný dopad na ekonomickou situaci následujících let. Podobná situace nastala i v sousedním Polsku. Naopak u států, jejichž ekonomika vykazovala v letech před začátkem pandemie pouze velmi malý růst (např. Německo nebo Itálie), není kumulativní dopad pandemie tak znatelný a Itálie vykazuje v tomto datasetu dokonce kladný rozdíl 2,2 p.b.

4.3.6 Shrnutí deskriptivní analýzy

Deskriptivní analýza měla za úkol popsat vhodným způsobem dataset OWID a na základě informací v něm obsažených poskytnout odpovědi na otázky spojené s pandemií a jejím vývojem.

První část analýzy se zabývala načtením a transformací dat do požadovaného tvary využitého v popisných kapitolách. První popisná kapitola sledovala globální trendy nových případů a úmrtí a přinesla odpověď na otázku ohledně šíření a dopadů nemoci v minulých letech až do současnosti (k 1.1. 2025). Byly zde také popsaný jednotlivé vlny nemoci. Druhá kapitola se věnovala stejným trendům pouze omezeným na Českou republiku a nabídla trendy a rozdíl v šíření a dopadech nemoci v ČR a Evropě.

Další z kapitol přiblížila šíření a dopady nemoci podle příjmových kategorií („vyspělosti), přijatých opatření (reprezentovaných Indexem přísnosti) a míry proočkovanosti. Bylo zjištěno, že ačkoliv vyspělé země vykazovaly více případů, vyšší míra přijatých opatření a proočkovanosti dokázala razantně snížit úmrtnost. Poslední blok se zabýval dopady Covidu a jeho šíření na ekonomiku vybraných států. Pomocí měřítka HDP na obyvatele bylo ilustrováno, jak Covid-19 negativně ovlivnil vývoj HDP a poslal ekonomiky většiny států do recese.

4.4 Tvorba predikčních modelů

Fáze tvorby predikčních modelů následuje po fázi deskriptivní analýzy, kde byla dostupná data prozkoumána, popsána a transformována, na základě čehož mohly být formulovány odpovědi na některé otázky týkající se šíření viru a jeho působení na různé společenské aspekty v minulosti. Pro pochopení možného chování viru do budoucna jsou v následující kapitole sestrojeny tři prediktivní modely časových řad, z nichž každý používá odlišného přístupu k predikování budoucího vývoje. Model Prophet zastupuje spíše tradiční statistické modely pro práci s časovými řadami, model XGBoost zastupuje modely strojového učení a využívá v jádru jednoduchý princip rozhodovacích stromů s korekčními prvky, a model LSTM neuronových sítí přináší pohled na věc z hlediska hlubokého učení (Deep Learning) simulujícího činnost lidského mozku při vyhodnocování dat.

4.4.1 Model Prophet

Prvním modelem použitým pro predikci je statistický model a knihovna v jazyce Python s názvem Prophet. Jde o nástroj vyvinutý společností Meta s důrazem na jednoduchost použití pro převážně krátkodobé a střednědobé predikce časových řad. Více se modelu Prophet věnuje kapitola 3.5.5.3.

4.4.1.1 Princip fungování modelu

Prophet je založen na aditivním modelu, který rozkládá časovou řadu na čtyři složky podle vzorce:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$$

kde:

- g(t) představuje trendovou složku
- s(t) reprezentuje sezónní komponentu
- h(t) zohledňuje vliv svátků či jiným událostí
- ε(t) je chybový člen zachycující náhodné a nevysvětlené odchylky

Trendová složka g(t) zachycuje dlouhodobý vývoj časové řady. Prophet nabízí dva druhy trendu: lineární a logistický. Zde je využit logistický trend z důvodu nestability lineárního trendu, který na základě minulých výkyvů nemoci dává velmi nepřesné výsledky.

Sezónní složka s(t) je modelována pomocí Fourierovy řady, tedy využívá k aproximaci periodických jevů harmonické funkce sinus a cosinus. V práci sezónní složka mapuje trendy covidového šíření v rámci roku, tedy by měla například zachytit opakující se zimní vlny. Sezónní složka je vypočítána podle následujícího vzorce:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos \frac{2n\pi t}{P} + b_n \sin \frac{2n\pi t}{P} \right)$$

kde:

- P je perioda (např. roční, měsíční apod.)
- N je počet harmonických složek

Složka události h(t) umožňuje do modelu zahrnout vliv různých specifických událostí, například vliv státních svátků apod. Pro potřeby práce není s touto složkou dále pracováno. Poslední složka ε(t) pouze reprezentuje náhodné výkyvy a nevysvětlené chování modelu, tedy není v rámci přípravy modelování specificky nastavována. U náhodné složky se předpokládá normální rozdělení reziduí se středem v nule.

4.4.1.2 Průběh modelování

Následující kapitola se věnuje rozboru funkcionalit zdrojového kódu modelu Prophet (Zdrojový kód 1) napsaného v programovacím jazyce Python ve webovém vývojovém prostředí Google Colab.

Před zahájením predikce jsou importovaný využívané knihovny:

- knihovna Prophet obsahující potřebnou logiku modelu nutnou k predikci;
- knihovna Matplotlib použitá pro vykreslení výsledků modelování;
- knihovna Pandas, jejíž DataFrame je použit pro uchovávání dat – tato knihovna zde není volána, jelikož byla použita již pro přípravu dat v předchozí kapitole.

Nejdříve jsou data připravena před zahájením modelování. Je načten očištěný dataset z části prediktivní analýzy, který má nahrazeny chybějící a nulové hodnoty. Data jsou agregována na denní úroveň podle sloupců „date“ (datum) a „new_cases“ (nové případy). Sloupce jsou dále přejmenovány podle požadavků modelu na „ds“ (datová složka) a „y“ (cílová proměnná). Na sloupec s novými případy je aplikován 7denní klouzavý průměr s centrováním, což má za úkol vyhladit denní fluktuace a poskytnout lepší predikci a přehled o budoucím trendu.

V druhé části jsou přidána omezení pro trénovací data, od března 2020 do května 2023. Březen 2020 byl vybrán jako krajní bod z důvodu zachycení prvního znatelného stoupání v datech a květen 2023 byl vybrán z důvodu oficiálního ukončení globálního pandemického stavu WHO. Kvůli tomu jsou data po tomto datumu často podhodnocená, a ne příliš věrná realitě, což zkresluje predikci. V tomto bodě je také nastavena hodnota „cap“ určující maximum případů, které může model předpovědět. Tato hodnota zde slouží spíše jako doporučení a příliš neovlivňuje výslednou podobu modelu. Zde je nastaveno omezení na 20 milionů případů.

V části třetí je přikročeno k nastavení hyperparametrů modelu. Jde o tyto parametry:

- growth = 'logistic' – model používá logistický růst
- changepoint_prior_scale = 0.05 – určuje flexibilitu modelu při detekci změn trendu, zde je použita výchozí
- n_changepoints = 30 – definuje počet potenciálních změn trendu uvažovaných modelem
- seasonality_mode = 'additive' – aditivní režim sezónosti určuje, že sezónní efekty jsou přičteny k trendu
- yearly_seasonality = True – zapíná roční sezónnost, což umožňuje modelu zachycovat opakující se vzory v datech
- weekly_seasonality = False – týdenní sezónnost se pro potřeby tohoto modelu vypnuta, jelikož není třeba sledovat výkyvy případů v rámci dnů v týdnu

V posledním kroku je přistoupeno k samotnému trénování modelu pomocí metody *fit()*. Je provedena predikce budoucího šíření na 1000 dnů (od května 2023), tedy do února 2026. Model je omezen, aby nezobrazoval záporné predikce, které vzhledem k povaze zkoumaných dat nedávají smysl.

Zdrojový kód 1 Predikce modelem Prophet

```
# Predikce modelem Prophet
from prophet import Prophet
import matplotlib.pyplot as plt
from matplotlib.ticker import FuncFormatter
import matplotlib.dates as mdates

print("\n## PREDIKCE šíření COVID-19 Prophet modelem")

# 1. Příprava dat před modelováním
df_daily = (
    df_clean.groupby('date')['new_cases'].sum().reset_index()
    .rename(columns={'date': 'ds', 'new_cases': 'y'})
)
df_daily['y'] = df_daily['y'].rolling(window=7, center=True,
min_periods=1).mean()

# 2. Trénovací data a omezení
df_train = df_daily[(df_daily['ds'] >= '2020-03-01') & (df_daily['ds'] <=
'2023-05-01')].copy()
df_train['cap'] = 20_000_000

# 3. Nastavení hyperparametrů modelu
model = Prophet(
    growth='logistic',
    changepoint_prior_scale=0.05,
    n_changepoints=30,
    seasonality_mode='additive',
    yearly_seasonality=True,
    weekly_seasonality=False,
)
model.fit(df_train)

# 4. Predikce budoucího šíření
future = model.make_future_dataframe(periods=1000)
future['cap'] = 20_000_000
forecast = model.predict(future)
forecast[['yhat', 'yhat_lower', 'yhat_upper']] = forecast[['yhat',
'yhat_lower', 'yhat_upper']].clip(lower=0)
```

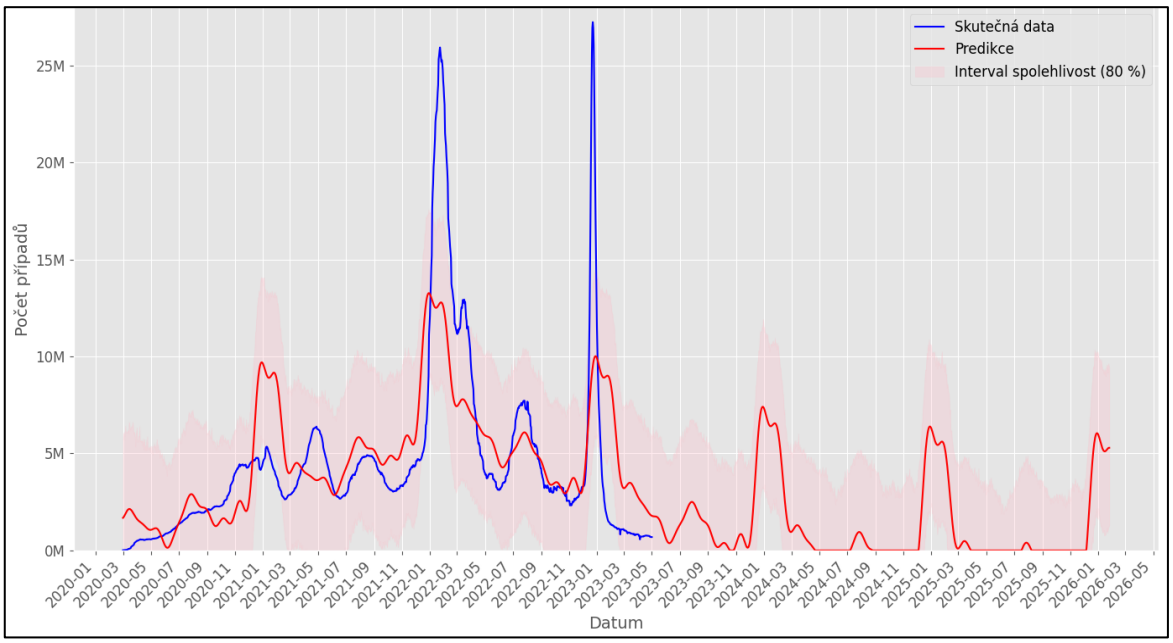
Zdroj: vlastní zpracování

Po provedení predikční části jsou pomocí knihovny Matplotlib vykresleny výsledky modelu, kterým se věnuje následující kapitola.

4.4.1.3 Výsledky

V grafu 10 je zobrazena dlouhodobá predikce případů COVID-19 pomocí modelu Prophet. Skutečná data jsou zobrazena modře, predikce červeně a růžová plocha představuje 80% interval spolehlivosti. Model se učil na pandemických datech od března 2020 do května 2023. Použití tohoto rozsahu je z důvodu, že v případě využití dat až do současnosti se model nedokázal „vyrovnat“ s velmi malými výkyvy případů a predikoval neinterpretovatelné výsledky. Graf v plné velikosti je dostupný v Příloze C.1.

Graf 10 Predikce vývoje globálních nových případů COVID-19 modelem Prophet



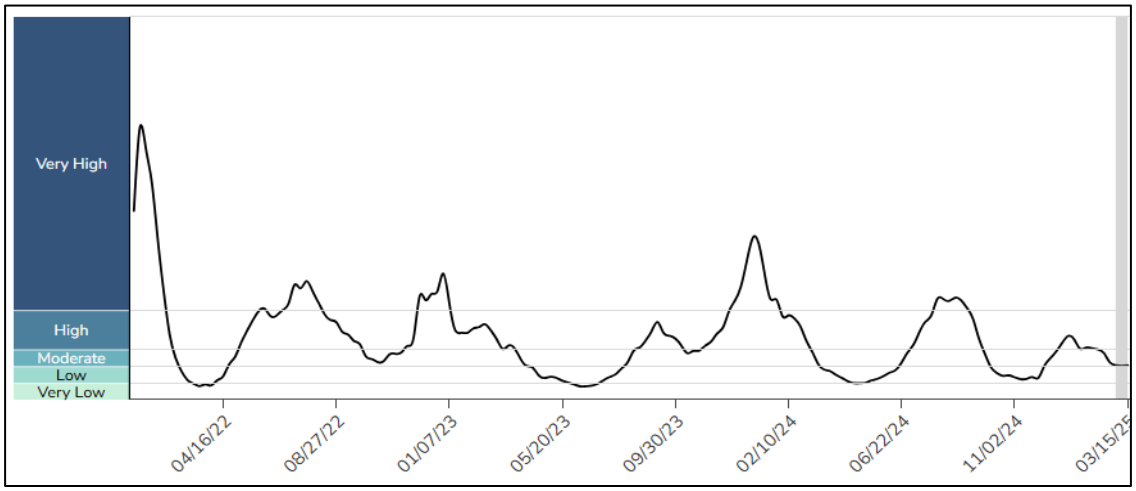
Zdroj: vlastní zpracování, OWID [61].

Je patrné, že model dokáže s relativní přesností replikovat skutečný vývoj, s výjimkou období charakterizovaných prudkým nárůstem případů. Tyto náhlé změny, například v důsledku šíření nové varianty viru, model nezachycuje, jelikož při predikci nepočítá s reálnými externími faktory. V těchto obdobích tedy predikuje výrazně nižší nárůst případů, než jaký je pozorován ve skutečných datech. Na základě trénovacích dat (značených modře) model identifikuje určité sezónní vzorce, které následně využívá k predikci budoucího vývoje. Výsledkem je předpoklad klesající tendence výskytu případů mimo zimní sezónu, přičemž v zimních měsících (na severní polokouli, která významně ovlivňuje celkový trend) očekává model výraznější nárůst. Letní vlny jsou naopak v modelu znatelně slabší a postupně se blíží k nulovým hodnotám.

Ačkoli model predikuje vyšší počet případů v zimním období, než jaký je ve skutečnosti zaznamenán, je třeba vzít v úvahu, že reálná data po ukončení pandemie mohou být významně podhodnocená. Tato data tedy nejsou plně vypovídající pro detailní mapování vývoje vln. Při srovnání s koncentrací viru SARS-CoV-2 v odpadních vodách, která může lépe odrážet skutečný výskyt onemocnění v populaci, je patrné, že model správně predikoval zimní vlnu na přelomu let 2023 a 2024. Rovněž předpokládal podobný vývoj i na přelomu let 2024 a 2025, avšak skutečný počet případů byl v tomto období nižší než predikce.

V případě letních vln si model vede méně přesně – přestože interval spolehlivosti predikce připouští mírné nárůsty případů v letních měsících, samotná střední predikce je oproti reálným hodnotám výrazně podhodnocena. Například nástup varianty JN.1 v červenci 2024 (viz Graf 11) model zachytil pouze částečně, což je důsledkem skutečnosti, že nepracuje s informacemi o výskytu nových variant. Výsledný graf tak ilustruje limity modelu při predikci dlouhodobého vývoje, zejména v kontextu náhlých epidemiologických změn.

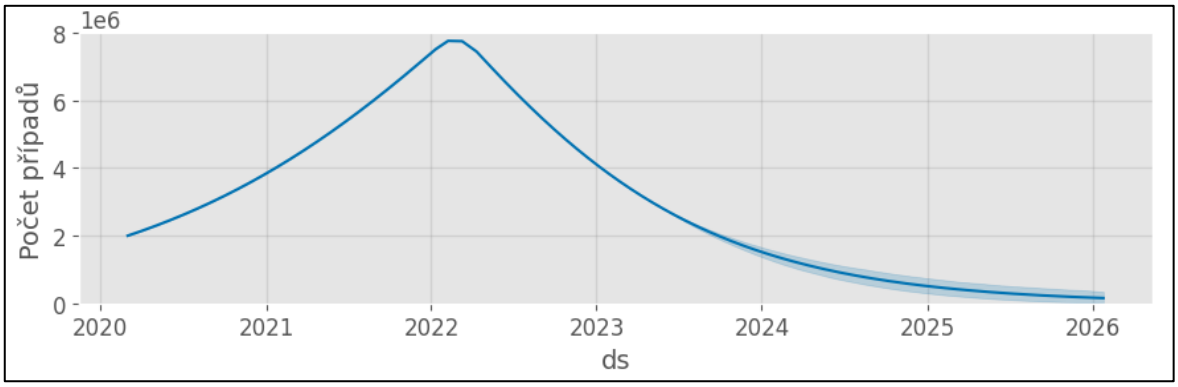
Graf 11 Koncentrace viru COVID-19 v odpadních vodách (USA, 2020–24)



Zdroj: CDC [64].

Poslední grafem modelu je predikce dlouhodobého trendu. Zde model předpovídá stabilní pokles případů od roku 2022 (který je vrcholem pandemie) dále do budoucnosti. Tato predikce se shoduje s trendem pozorovaným na reálných datech, nicméně nelze usuzovat, že se případy budou v realitě doopravdy neustále snižovat, vzhledem k velmi proměnlivé povaze viru. Graf trendu modelu je uveden v milionech případů (1e6).

Graf 12 Dlouhodobý trend šíření COVID-19 ve světě podle Prophet



Zdroj: vlastní zpracování, OWID [61].

4.4.2 Model XGBoost

Druhým modelem v pořadí je model strojového učení XGBoost. Ten má stejně jako Prophet svou vlastní knihovnu pro jazyk Python, pomocí které je modelování provedeno. Model využívá ve svém jádru konstrukci mnoha jednoduchých rozhodovacích stromů, které jsou obohaceny o metodu gradientního boostingu, což vede k lepší paměti modelu.

4.4.2.1 Princip fungování modelu

XGBoost model je sestaven jako součet mnoha „slabých“ rozhodovacích stromů. Cílem je minimalizovat celkovou chybu modelu pomocí postupného přidávání stromů, které „opravují“ chyby předchozích stromů. Matematicky je model pro i-tý vzorek vyjádřen následovně:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

kde $\hat{y}_i^{(t)}$ je predikce po t-té iteraci a $f_t(x_i)$ je strom přidaný v iteraci t.

Celkovým cílem je minimalizovat následující objektivní funkci:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)$$

kde:

- $(y_i, \hat{y}_i)$ je ztrátová funkce – měří, jak dobře model predikuje cílové hodnoty;
- $\Omega(f_t)$ je regularizační člen – penalizuje složitost stromu, předchází přeučení stromu.

Postup učení je založen na gradientním sestupu:

1. Ve fázi inicializace je zahájena predikce s počáteční hodnotou, např. průměrem cíle.
2. V každé další iteraci je spočtena první (gradient) a druhá (hessian) derivace, ty ukazují směr a velikost aktuální chyby.
3. Každý další strom je trénován tak, aby aproximoval záporný gradient, tedy aby se po jeho přidání vylepšila predikce modelu.
4. Model je aktualizován jako součet předchozích stromů a nově natrénovaného stromu.

4.4.2.2 Průběh modelování

Následující kapitola se věnuje rozboru zdrojového kódu (Zdrojový kód 2–4) v jazyce Python, který vede ke konstrukci a predikci modelu XGBoost. Stejně jako v předchozí ukázce modelu Prophet byl tento model vyvíjen v prostředí Google Colab.

Kód začíná načtením knihoven potřebných pro predikci:

- knihovna Matplotlib použitá pro vykreslení výsledků modelování;
- knihovna Pandas, jejíž DataFrame je použit pro uchovávání dat;
- knihovna XGBoost, která obsahuje samotnou implementaci a logiku prediktivního modelu
- knihovna NumPy, pro implementaci Fourierových členů

V prvním kroku jsou seskupena globální data dle dní a je provedena korekce přírůstku nových případů po oficiálním konci pandemie (květen 2023), jelikož, jak již bylo zmíněno při interpretaci předchozího modelu Prophet, oficiální sesbíraná data se neshodují s nepřímými vlivy viru, např. s mírou jeho zastoupení v odpadních vodách velkých měst.

V další fázi je definována funkce, která vytváří příznaky na základě původních dat a aplikuje je na připravenou agregaci. Funkce obsahuje:

- kalendářové proměnné (den, měsíc, rok)
- Zpožděné hodnoty (lag features)
- klouzavé průměry (za posledních 7 dnů)
- hodnota řady před rokem (lag_365)
- Fourierovy členy (simulace roční sezónnost pomocí sin a cos funkcí)

V následujícím roce je přikročeno k trénování samotného modelu. Trénovací data jsou omezena do 1. ledna 2025. Vstupní proměnné modelu jsou datum (date) a počet případů (daily_cases) a cílová proměnná je nazvána „y“ (denní nové případy). Model je natrénován pomocí knihovny XGBoost metodou *XGBRegressor()*.

V poslední části je provedena rekurzivní predikce na 365 dnů. Jsou připravena historická data, ze kterých je vycházeno při predikci. Pomocí cyklu *for* je prováděna predikce postupně den po dni. V každé iteraci cyklu je provedeno:

- vygenerování příznaků pro další den v pořadí
- vypočtení zpožděných hodnot (lagů)
- výpočet sezónních složek podle dnu v roce (Fourierovy členy)
- predikce nových případů pro další den pomocí natrénovaného modelu
- přidání vytvořené predikce do historie a její použití pro další predikci
- uložení predikce do výstupu

Zdrojový kód 2 1. fáze predikce modelem XGBoost

```
import pandas as pd
import numpy as np
import xgboost as xgb
import matplotlib.pyplot as plt
from sklearn.metrics import mean_absolute_error, mean_squared_error

# 1. Příprava denních dat
df_global = df.groupby('date')['new_cases'].sum().reset_index()

# Zesílení podhodnocených hodnot po konci pandemie (Květen 2023)
cutoff_date = pd.to_datetime("2023-05-01")
df_global.loc[df_global['date'] >= cutoff_date, 'new_cases'] *= 4
df_daily = df_global.rename(columns={'new_cases': 'daily_cases'})

# 2. Inženýrství příznaků
def create_daily_features(df, lags=7, roll_windows=[3, 7], fourier_order=3):
    df = df.copy()
    df = df.sort_values('date').reset_index(drop=True)

    # Kalendářové příznaky
    df['dayofweek'] = df['date'].dt.dayofweek
    df['month'] = df['date'].dt.month
    df['dayofyear'] = df['date'].dt.dayofyear

    # Zpožděné hodnoty (lag features)
    for lag in range(1, lags + 1):
        df[f'lag_{lag}'] = df['daily_cases'].shift(lag)

    # Klouzavé průměry
    for window in roll_windows:
        df[f'roll_mean_{window}'] =
df['daily_cases'].shift(1).rolling(window=window).mean()

    # Hodnota před rokem (roční sezónnost)
    df['lag_365'] = df['daily_cases'].shift(365)

    # Fourierovy členy
    df = df.set_index('date')
    for k in range(1, fourier_order + 1):
        df[f'sin_{k}'] = np.sin(2 * np.pi * k * np.arange(len(df)) / 365)
        df[f'cos_{k}'] = np.cos(2 * np.pi * k * np.arange(len(df)) / 365)
    df = df.reset_index()

    return df.dropna().reset_index(drop=True)
df_features = create_daily_features(df_daily)
```

Zdroj: vlastní zpracování

Zdrojový kód 3 2. fáze predikce modelem XGBoost

```
# 3. Trénování modelu
forecast_start_date = pd.to_datetime("2025-01-02")
train = df_features[df_features['date'] < forecast_start_date]

features = [col for col in train.columns if col not in ['date',
'daily_cases']]
X_train, y_train = train[features], train['daily_cases']

model = xgb.XGBRegressor(n_estimators=300, learning_rate=0.05, max_depth=5,
random_state=10)
model.fit(X_train, y_train)

# 4. Rekurzivní predikce
forecast_days = 365
history = df_features[df_features['date'] < forecast_start_date].copy()
future_dates = []
future_preds = []

for i in range(forecast_days):
    last_row = history.iloc[-1:].copy()
    next_date = pd.Timestamp(last_row['date'].values[0]) +
pd.Timedelta(days=1)
    future_dates.append(next_date)

    new_row = {
        'date': next_date,
        'dayofweek': next_date.dayofweek,
        'month': next_date.month,
        'dayofyear': next_date.dayofyear,
    }

    # Zpožděné hodnoty
    for lag in range(1, 8):
        new_row[f'lag_{lag}'] = history['daily_cases'].iloc[-lag]

    # Klouzavé průměry
    for window in [3, 7]:
        new_row[f'roll_mean_{window}'] = history['daily_cases'].iloc[-
window:].mean()

    # Hodnota před rokem
    if len(history) >= 365:
        new_row['lag_365'] = history['daily_cases'].iloc[-365]
    else:
        new_row['lag_365'] = history['daily_cases'].mean()
```

Zdroj: vlastní zpracování

Zdrojový kód 4 3. fáze predikce modelem XGBoost

```
    # Fourierovy členy
    t = len(history)
    for k in range(1, 4):
        new_row[f'sin_{k}'] = np.sin(2 * np.pi * k * t / 365)
        new_row[f'cos_{k}'] = np.cos(2 * np.pi * k * t / 365)

    input_df = pd.DataFrame([new_row])
    pred = model.predict(input_df[features])[0]
    new_row['daily_cases'] = pred
    future_preds.append(pred)

    history = pd.concat([history, pd.DataFrame([new_row])],
ignore_index=True)
```

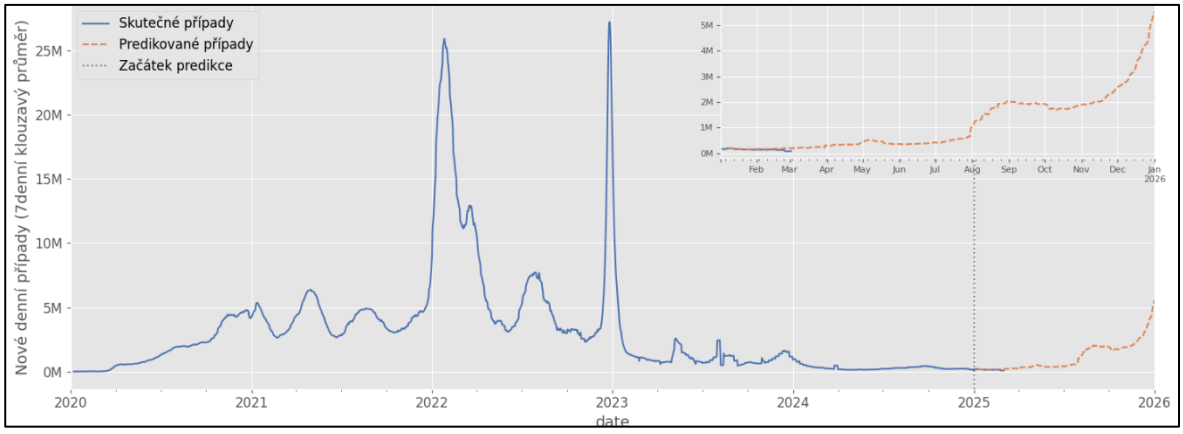
Zdroj: vlastní zpracování

Tímto způsobem je realizována dlouhodobá predikce nových případů COVID-19. Výsledky predikce jsou následně vykresleny pomocí knihovny Matplotlib a rozebrány v následující kapitole výsledky.

4.4.2.3 Výsledky

Graf 13 zobrazuje predikci modelem XGBoost od ledna 2025 do ledna 2026. Graf je z důvodu větší názornosti zobrazen v sedmidenním klouzavém průměru nových případů COVID-19 ve světě. Skutečné naměřené počty případů znázorňuje modrá křivka (datující od ledna 2020 do ledna 2025) a predikované počty případů znázorňuje přerušovaná oranžová křivka. Graf je dostupný v plné velikosti v Příloze C.2.

Graf 13 Predikce vývoje globálních nových případů modelem XGBoost



Zdroj: vlastní zpracování, OWID [61].

Z grafu je patrné, že model se pokouší predikovat určité sezónní trendy, nicméně nepředpokládá vlnu na přelomu roku 2024 a 2025, což je dáno nízkou mírou potvrzených případů v těsné minulosti. Očekává nicméně výraznější nárůst případů v září 2025, což podle příkladů z minulých let (např. roku 2024) může být způsobeno začátkem školního roku a šíření viru mezi dětmi a mladistvými, kteří jej roznáší mezi své rodinné příslušníky. Model také predikuje výraznější prosincovou vlnu, což pravděpodobně způsobila pandemická maxima předchozích let (2022 a 2023) v okolí prosince, na které se model „naučil“. Stejně jako u předchozího modelu je nutné zdůraznit, že ačkoliv reálně reportovaná data ukazuji klesající trend a žádné výraznější vlny v posledním roce, faktory jako prevalence viru Covidu v odpadních vodách mluví o značně vyšším počtu reálných případů – čísla predikovaná modelem se tedy pohybují v realistických rámcích (ačkoliv „zimní“ vlna ke konci roku 2025 je pravděpodobně nadhodnocena vlivem předchozích výkyvů).

4.4.3 Model LSTM neuronové sítě

Třetí konstruovaný model je vytvořen pomocí architektury LSTM (Long Short-Term Memory). Jde o pokročilý AI model založený na odvětví Deep Learning (Hluboké učení), které se specializuje na tvorbu, trénink a validaci umělých neuronových sítí. Architektura modelu využívá znalostí z oblastí nervové soustavy a činnosti mozku pro efektivní učení na vstupních datech. Model LSTM je obzvlášť vhodný pro práci s časovými řadami, jelikož „nezapomíná“ trendy, které se udály mnoho iterací zpět. Obecnému principu fungování LSTM sítí se věnuje kapitola 3.5.5.5.

4.4.3.1 Princip fungování modelu

LSTM model neuronové sítě je založen na sekvenčním zpracování dat neuronovou sítí, tedy každé časové okno je přímo závislé na předchozím stavu sítě. Síť dokáže uchovávat a udržovat informace pomocí vnitřního stavu jednotlivých neuronů a regulovat tok informací v síti pomocí řídících bran.

Matematické vyjádření modelu LSTM buňky v čase t, pro vstupní vektor x_t, předchozí skrytý stav h_{t-1} a předchozí stav buňky c_{t-1}, váhovou matici pro vstupní vektor W a váhovou matici U pro předchozí skrytý stav, je následující:

- Zapomínací brána f_t – rozhoduje o udržení či smazání informace:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

kde σ je sigmoidní (logistická) funkce (výstup v intervalu 0–1);

- Vstupní brána i_t – určuje, které informace vložit do buněčného stavu:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c)$$

kde $\tilde{c}_t$ je kandidát na vstup do paměti a tanh je hyperbolický tangens (-1 až 1);

- Aktualizace buněčného stavu – kombinace starého a nového obsahu:

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t}$$

kde c_t je aktuální vnitřní paměť buňky a $\odot$ je prvkový součin matic;

- Výstupní brána o_t – řídí, co bude výstupem buňky:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

kde h_t je skrytý stav, který je posílán do následující buňky.

4.4.3.2 Průběh modelování

Následující kapitola je věnována rozboru zdrojového kódu (Zdrojový kód 5–7) konstrukce, trénování a evaulace modelu rekurentních neuronových sítí LSTM pomocí Python frameworku TensorFlow. Opět bylo použito webové vývojové prostředí Google Colab.

V prvním kroku jsou opět načteny potřebné knihovny pro práci s daty (Pandas), maticemi (NumPy), grafy (Matplotlib) a škálování (ScikitLearn). Tyto knihovny jsou doplněny o robustní framework TensorFlow, který obsahuje moduly a jednotlivé součástí modelu (vrstvy) potřebné k sestavení LSTM neuronové sítě.

Před zahájením predikce jsou data o počtu nových případů agregována podle dne. Data použitá pro učení jsou vybrána z výrazných pandemických let 2021 a 2022. Data po konci pandemie (květen 2023) jsou mírně nadhodnocena, jako v případě předchozího modelu. Nové případy jsou převedeny na 7denní klouzavý průměr z důvodu vyhlazení denního šumu

v datech. Také jsou přidány sezónní složky (den, měsíc) a Fourierovy složky, reprezentující sezónní výkyvy.

V dalším kroku jsou přidány trendové složky zachycující změnu a tempo změny nových případů. Tyto složky (příznaky) jsou poté použity při škálování vstupní a výstupních hodnot pomocí metod *StandardScaler()* a *MinMaxScaler()*. Hodnoty vstupu jsou škálovány z důvodu odlišných měřítek jednotlivých příznaků a výstupní hodnoty jsou škálovány z důvodů zpřesnění predikce v daném rozsahu.

V dalším kroku je definována délka predikce a okno, které model využívá jako základ učení. Vzhledem k velikosti datového setu je vybráno okno jednoho roku. Poté je okno sekvenčně posouváno přes trénovací data pomocí cyklu *for*.

V části číslo 3 je vytyčena architektura sítě (jednotlivé vrstvy a jejich parametry):

- vstupní vrstva přijímající škálovaná data
- 1D konvoluční vrstva zachycující lokální trendy
- dvě LSTM vrstvy pro trénování časových závislostí
- dvě Dropout (zahazovací) vrstvy typu Monte Carlo (proti přeučení sítě)
- výstupní vrstva (vektor s 365 hodnotami, pro každý den v roce 2025)

Model využívá optimalizační funkci ADAM (vhodnou pro velké datasety a komplexní modely), která optimalizuje styl učení a pomáhá efektivně korigovat bias sítě. Jako ztrátová funkce je zvolena MSE (střední kvadratická chyba). Po definování architektury je možné přistoupit k trénování modelu pomocí *model.fit()*. Pro zefektivnění učícího procesu jsou do kódu přidány automatické stop funkce, které zastaví trénování, pokud se ztrátová funkce nesnižuje s další iterací.

V posledním kroku je přikročeno k samotné predikci s využitím dříve definované Monte Carlo Dropout vrstvy. Je prováděno více predikcí najednou a generován interval spolehlivost, ve kterém se nachází 80 % všech „odchylek“ od predikce. Na závěr je vygenerován predikční Dataframe, tedy tabulka s predikovanými denními hodnotami za rok 2025 společně s intervaly spolehlivosti.

Zdrojový kód 5 1. fáze predikce modelem LSTM neuronové sítě

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler, MinMaxScaler
import tensorflow as tf
from tensorflow.keras.models import Model
from tensorflow.keras.layers import Input, LSTM, Dense, Dropout, Conv1D
from tensorflow.keras.callbacks import EarlyStopping, ReduceLROnPlateau

# 1. Příprava dat
df = df[(df['date'] >= '2021-01-01') & (df['date'] <= '2022-12-31')].copy()
df = df_clean.groupby('date')['new_cases'].sum().reset_index()
df.loc[df['date'] >= '2023-05-01', 'new_cases'] *= 4
df['cases_7d'] = df['new_cases'].rolling(7, center=True,
min_periods=1).mean()

# Sezónnost
df['dayofyear'] = df['date'].dt.dayofyear
df['month'] = df['date'].dt.month

# Fourierovy složky
df['sin1'] = np.sin(2 * np.pi * df['dayofyear'] / 365)
df['cos1'] = np.cos(2 * np.pi * df['dayofyear'] / 365)
df['sin2'] = np.sin(4 * np.pi * df['dayofyear'] / 365)
df['cos2'] = np.cos(4 * np.pi * df['dayofyear'] / 365)
df['sin3'] = np.sin(6 * np.pi * df['dayofyear'] / 365)
df['cos3'] = np.cos(6 * np.pi * df['dayofyear'] / 365)

# Trendové složky
df['delta'] = df['cases_7d'].diff().fillna(0)
df['pct_change'] = df['cases_7d'].pct_change().fillna(0)
df['roll7'] = df['cases_7d'].rolling(7).mean().fillna(0)
df['roll30'] = df['cases_7d'].rolling(30).mean().fillna(0)

features = ['cases_7d', 'delta', 'pct_change', 'roll7', 'roll30',
            'sin1', 'cos1', 'sin2', 'cos2', 'sin3', 'cos3', 'month']
df = df.dropna().reset_index(drop=True)

# Škálování
scaler_X = StandardScaler()
scaler_y = MinMaxScaler()
df_scaled = df.copy()
df_scaled[features] = scaler_X.fit_transform(df[features])
df_scaled['cases_7d_scaled'] = scaler_y.fit_transform(df[['cases_7d']])
```

Zdroj: vlastní zpracování

Zdrojový kód 6 2. fáze predikce modelem LSTM neuronové sítě

```
# 2. Sekvence
window_size = 365
horizon = 365
X, y = [], []

for i in range(window_size, len(df_scaled) - horizon):
    seq_x = df_scaled.loc[i - window_size:i - 1, features].values
    seq_y = df_scaled.loc[i:i + horizon - 1, 'cases_7d_scaled'].values
    if len(seq_y) == horizon:
        X.append(seq_x)
        y.append(seq_y)

X = np.array(X)
y = np.array(y)
print(f"X: {X.shape}, y: {y.shape}")

# 3. Architektura a trénování modelu
class MCDropout(Dropout):
    def call(self, inputs, training=None):
        return super().call(inputs, training=True)

input_layer = Input(shape=(window_size, len(features)))
x = Conv1D(filters=64, kernel_size=3, activation='relu',
padding='same')(input_layer)
x = LSTM(64, return_sequences=True)(x)
x = MCDropout(0.3)(x)
x = LSTM(32)(x)
x = MCDropout(0.3)(x)
x = Dense(128, activation='relu')(x)
output = Dense(horizon, activation='relu')(x)

model = Model(inputs=input_layer, outputs=output)
model.compile(optimizer='adam', loss='mse')

callbacks = [
    EarlyStopping(patience=20, restore_best_weights=True),
    ReduceLROnPlateau(patience=10, factor=0.5, verbose=1)
]

model.fit(X, y, epochs=100, batch_size=32, validation_split=0.2,
callbacks=callbacks, verbose=1)
```

Zdroj: vlastní zpracování

Zdrojový kód 7 3. fáze predikce modelem LSTM neuronové sítě

```
# 4. Predikce
forecast_start = pd.to_datetime("2025-01-01")
idx = df_scaled[df_scaled['date'] == forecast_start].index[0]
X_input = df_scaled.loc[idx - window_size:idx - 1,
features].values.reshape(1, window_size, len(features))

mc_runs = 100
mc_predictions = []

for _ in range(mc_runs):
    pred = model(X_input, training=True).numpy()[0]
    pred = scaler_y.inverse_transform(pred.reshape(-1, 1)).flatten()
    pred = np.maximum(pred, 0)
    mc_predictions.append(pred)

mc_predictions = np.array(mc_predictions)
mean_pred = mc_predictions.mean(axis=0)
lower_bound = np.percentile(mc_predictions, 10, axis=0)
upper_bound = np.percentile(mc_predictions, 90, axis=0)

forecast_dates = pd.date_range(start=forecast_start, periods=horizon)
forecast_df = pd.DataFrame({
    'date': forecast_dates,
    'mean': mean_pred,
    'lower': lower_bound,
    'upper': upper_bound
}).set_index('date')
```

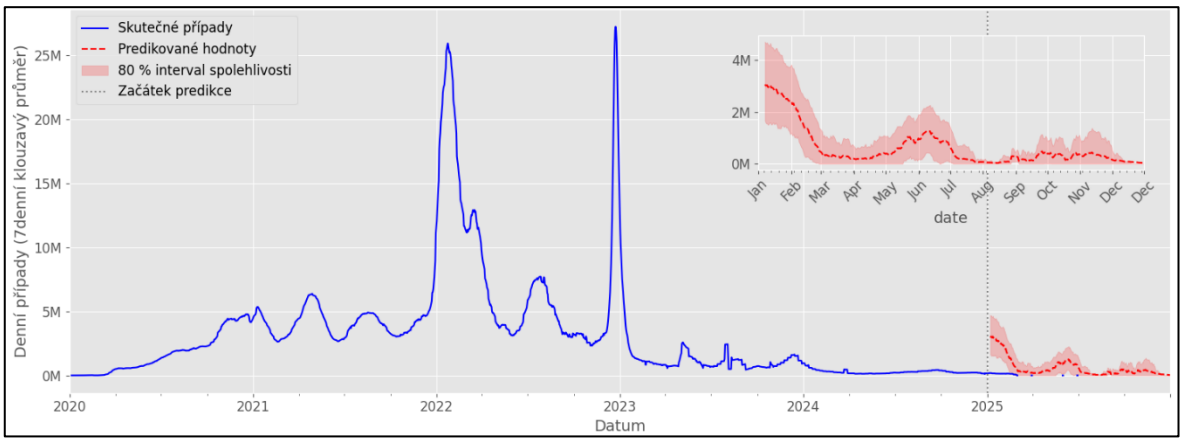
Zdroj: vlastní zpracování

Tímto způsobem je realizována dlouhodobá predikce nových případů COVID-19 pomocí neuronové Long Short-Term Memory. Výsledky predikce jsou následně vykresleny pomocí knihovny Matplotlib a rozebrány v podkapitole výsledky.

4.4.3.3 Výsledky

Graf 14 zachycuje predikci modelem LSTM neuronové sítě pro rok 2025. Graf je z důvodu větší názornosti zobrazen v sedmidenním klouzavém průměru nových případů COVID-19 ve světě. Skutečné naměřené počty případů znázorňuje modrá křivka (datující od ledna 2020 do ledna 2025) a predikované počty případů znázorňuje přerušovaná červená křivka. Růžové plochy tvoří 80 % interval spolehlivosti modelu. Graf je dostupný v plné velikosti v Příloze C.3.

Graf 14 Predikce nových globálních případů COVID-19 pomocí LSTM sítě



Zdroj: vlastní zpracování, OWID [61].

Na tvaru výsledné predikované křivky lze pozorovat, že předpokládá silnější zimní vlnu na začátku roku 2025, ačkoliv skutečná data z konce roku 2024 tomu nenasvědčují (nicméně je nutné opět zmínit jejich možné podhodnocení v důsledku minimálního testování). Tento vývoj predikce je pravděpodobně založen na analýze výrazných lednových vln minulých let modelem – jejich význam byl uložen do paměti modelu a byla na něm založena tato predikce.

Model dále ve své predikci předpovídá méně významnou vlnu okolo měsíce června – i tento trend lze pozorovat v předcházejících letech, například roky 2021, 2022 i 2023 mají menší vlnu v průběhu letních měsíců. Tato vlna chybí pouze v roce 2024, což ale může být způsobeno nedostatečnou dostupností dat. Dle grafu prevalence viru Covid-19 v odpadních vodách (Graf 11) lze i v tomto roce pozorovat nárůst výskytu viru v letních měsících.

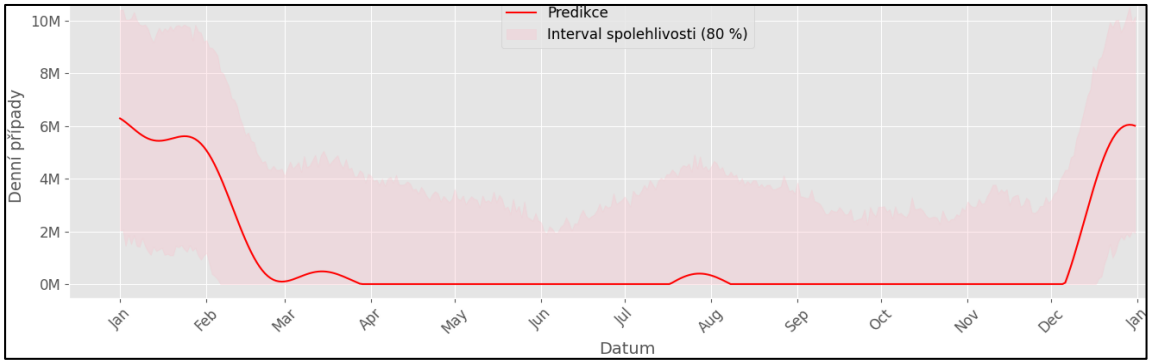
Na predikovaných datech je možné pozorovat vytrvale klesající trend případů – model bere tento trend v potaz, a proto predikuje ke konci roku 2025 už pouze nižší přírůstky případů.

5 Zhodnocení výsledků

Na základě principů a technologií popsaných v teoretické části práce byl ve vlastní části práce uskutečněn proces tvorby prediktivních modelů. Nejdříve byl vybrán reprezentativní dataset „Our World in Data“, který v sobě sdružoval potřebné datové zdroje renomovaných organizací (např. WHO, CDC, ECDC apod.) zabývajících se sběrem pandemických dat. V rámci něj byla provedena deskriptivní analýza popisující a mapující průběh, vývoj a dopady pandemie. V rámci této analýzy byl vyšetřován průběh šíření ve světě, šíření omezené na Českou republiku, dopady viru na státy podle vyspělosti, přijatých opatření či očkovanosti. Na závěr analýzy proběhla analýza ekonomických dopadů na základě dat z iniciativy Eurostat.

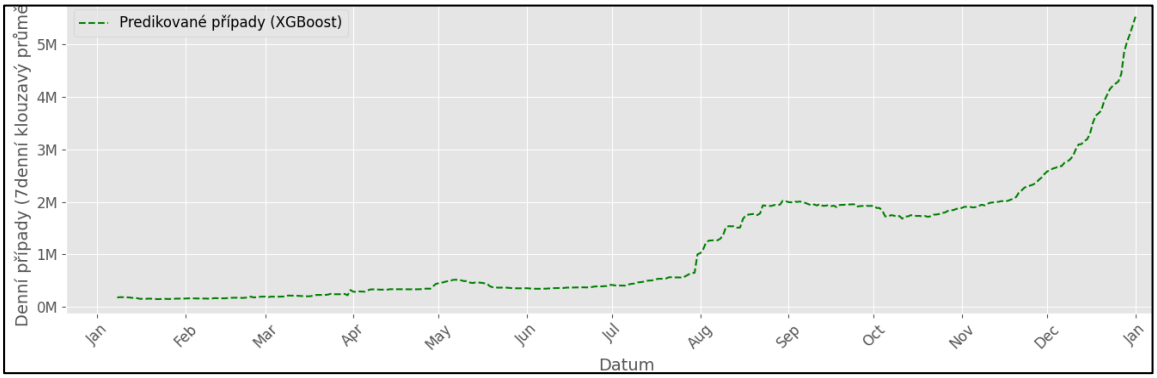
Zjištěné informace a podklady v rámci prediktivní analýzy byly využity pro tvorbu tří Machine Learning modelů dlouhodobé predikce počtu nových případů Covid-19, jejichž výsledky jsou zachyceny na grafech níže (Graf 15–17).

Graf 15 Prophet – Predikce globálních případů COVID-19 pro rok 2025



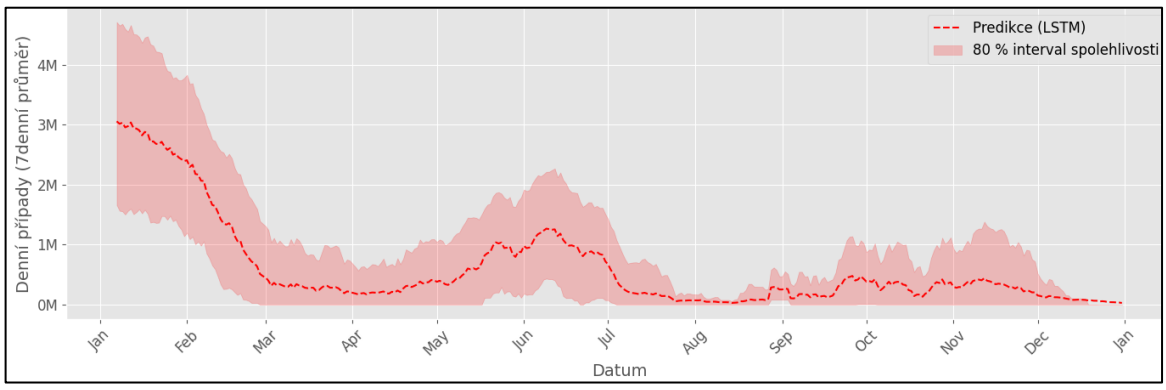
Zdroj: vlastní zpracování

Graf 16 XGBoost – Predikce globálních případů COVID-19 pro rok 2025



Zdroj: vlastní zpracování

Graf 17 LSTM – Predikce globálních případů COVID-19 pro rok 2025



Zdroj: vlastní zpracování

Na výsledné podobě grafů predikcí se projevují principy výpočtu jednotlivých modelů. Prophet, jakožto poměrně jednoduchý model zachycující sezónnost, byl naučen vnímat opakující se vlivy větších vln na přelomu roků a několika menších v průběhu, které opakuje pro všechny budoucí predikované roky. Tento fenomén ukazuje určité limitace a zvážení použití tohoto modelu pro dlouhodobou predikci, jelikož výsledky mohou být značně zavádějící a model nemusí „pochopit“ datovou strukturu. V tomto případě je nicméně finální podoba predikce relativně uspokojující, jelikož trendová i sezónní složka modelu poskytuje interpretovatelné a reálné výsledky.

Predikce modelem XGBoost nabízí mírně odlišné výsledky, jelikož předpokládá téměř nulový počet případů na začátku roku. Pozorované chování modelu je dáno podhodnocením dat těsně předcházejícím predikci, které model bere v potaz s největší silou. Vlivem sezónních proměnných je možné pozorovat vzestupnou tendenci v rámci roku, kde model predikuje zářijový nárůst případů a jejich vytrvalý růst až do konce roku. U XGBoost bylo možné pozorovat jeho „zmatení“ nad předchozími daty v největším měřítku. Bez vícestupňové úpravy hyperparametrů (vnitřního nastavení modelu) a přidání sezónních vlivů konvergovala většina modelu k nule a neposkytovala žádné interpretovatelné výsledky. Stojí opět za zvážení k vhodnosti použití tohoto modelu pro dlouhodobou predikci. Nicméně i přes nedostatky je prezentované chování i počty predikovaných případů realistické a interpretovatelné na základě chování viru v minulosti.

Poslední z modelů poskytuje asi nejvíce uvěřitelné výsledky, jelikož zvládá zachytit sezónní vlny i klesající trend bez obtíží popisovaných u předchozích modelů. Předpokládaný vývoj je uvěřitelný a shoduje se s předchozím vývojem časové řady. Toto zjištění koresponduje s faktem, že modely LSTM neuronových sítí jsou nejvíce vhodné

ke střednědobým a dlouhodobým predikcím časových řad a při správném nastavení hyperparametrů dokáží velmi spolehlivě zachytit jemné nuance a změny v datech.

Vyšší spolehlivost modelu pramení z jeho velmi vysoké výpočetní složitosti a nutnosti použití grafické karty pro provedení procesu trénování a predikce. Oproti ostatním modelům trval výpočet řádově déle. Model také obsahuje výrazně větší množství upravovatelných parametrů (počet vrstev, počet neuronů, optimalizační funkci, ztrátovou funkci, typ vrstev apod.) než konkurenční modely. Možným budoucím projektem může být další vylepšování modelu a ladění všech dostupných parametrů na optimální meze.

Je důležité zmínit také překážky procesu tvorby výše zmíněných prediktivních modelů. Vzhledem k povaze dat z posledního roku, kdy valná většina států hlásí pouze velmi nízké či nulové hodnoty podpůrných statistik o viru a jeho vlivu (např. počet testů, počet očkování apod.), bylo velmi složité zařadit do modelu odpovídající regresory tak, aby ovlivnily výsledky či poskytovaly uspokojivé závěry.

Tento nedostatek a možný problém v modelování byl zjištěn až důkladnou deskriptivní analýzou použitého datového setu OWID. Absence významných regresorů byla u modelů suplována přidáním většího počtu různých sezónních proměnných, které měly za úkol tyto vlivy alespoň částečně do modelu přinést. Nadstavbovým projektem nad touto diplomovou prací by mohla aplikace jednotlivých metod strojového učení na data jednotlivých možných prediktoru a jejich „dokreslení“ pomocí vícestupňové predikce. Výsledný komplexní model by značně zvýšil vypovídající hodnotu stávajícího modelu.

Další překážkou ve tvorbě spolehlivého modelu jsou skutečnosti a události, které model vzhledem ke vstupním datům nemá šanci predikovat. V tématu Covid-19 jde většinou o výskyt nové, velmi agresivní varianty viru, které způsobí velmi znatelný výkyv v datech. Model se sice učí na předchozích výkyvech a dokáže s mírnou přesností predikovat, kdy by mohlo k další vlně dojít, nicméně tento způsob predikce budoucích vln je velmi omezený a nespolehlivý. Model sice dokáže spolehlivě zachytit sezónnost a celkový trend dat, nicméně lokální maxima a „chaotické“ chování jsou pro něj neuchopitelné.

Principy a implementace výsledných modelů a nabytých zkušeností a znalostí problematiky strojového a hlubokého učení lze prakticky využít pro téměř libovolnou predikci časových řad a údajů. Moderní IT přístupy jsou neoddělitelně spjaty s využitím AI téměř ve všech odvětvích a profesích, jde tedy o velmi aktuální téma a cennou zkušenost pro budování kariéry v informačních technologiích.

6 Závěr

Předmětem této diplomové práce byla problematika tvorby prediktivních modelů šíření viru SARS-CoV-2 v programovacím jazyce Python. Pro uvedení do tématu prediktivních modelů byly představeny základní pojmy důležité pro pochopení dalších procesů. Byly popsány charakteristiky onemocnění Covid-19 v kontextu datové vědy a byl nastíněn vliv pandemie na přístup k datovým analýzám. V rámci charakteristik viru byly zmíněny základní biologické vlastnosti, varianty viru, příznaky nemoci, průběh onemocnění, epidemiologické charakteristiky i možné dlouhodobé následky nemoci. Dále se práce zabývala metodami diagnostiky, léčby a prevence nemoci a byla provedena analýza empirických studií zkoumajících různé aspekty pandemie.

Po definici předmětu zkoumání následoval popis technologií a metod aplikovaných v praktické části práce na data spojená se šířením nemoci, konkrétně programovací jazyk Python a strojové učení (Machine Learning). Charakteristika jazyka Python zahrnovala stručný popis, historii a vývoj, významné frameworky a knihovny pro práci s daty, a závěrem byly zmíněny jeho budoucí trendy a možnosti využití. V oblasti strojového učení byly popsány základní koncepty tohoto oboru, jeho specializace – hluboké učení (Deep Learning), různé druhy strojového učení a jejich aplikace na reálné problémy. Rovněž byly popsány klíčové modely, které byly využity pro predikce v praktické části.

Na základě popsaných technologií a metod byl vybrán reprezentativní datový set z projektu „Our World in Data“, který sdružuje informace z několika renomovaných zdrojů jako WHO, CDC nebo ECDC. Tento dataset byl následně využit v deskriptivní analýze, která odpověděla na otázky ohledně šíření nemoci a vztahů mezi jednotlivými faktory pandemie.

Po důkladné přípravě a analýze dat následovala fáze predikce pomocí modelů Prophet, XGBoost a LSTM neuronové sítě. Každá část zahrnovala popis fungování modelu podložený matematickým vyjádřením a detailní postup konstrukce modelu v Pythonu s využitím frameworku TensorFlow a dalších knihoven. Výsledné predikce nových případů Covid-19 pro rok 2025 byly analyzovány prostřednictvím grafů a dalších výstupů. Finální porovnání výsledků, kvalitu jednotlivých modelů a jimi generované predikce shrnula kapitola Zhodnocení výsledků, která také popsala přednosti, nedostatky a možná praktická využití modelů.

Výsledný projekt může sloužit jako úvod do řešení náročných úloh pomocí metod strojového učení, hlubokého učení a obecně umělé inteligence. Nabyté znalosti lze využít v širokém spektru profesí souvisejících s datovou vědou, od datového analytika až po vývojáře komplexních AI systémů. Práce nabízí odborný vhled a praktickou ukázku, která může být oporou pro budoucí projekty a experimenty.

Zdrojový kód deskriptivní analýzy a predikčních modelů je dostupný jako příloha práce, a také na platformě Github. K jeho zobrazení a úpravám se doporučuje webové vývojové prostředí Google Colab, které bylo při tvorbě zdrojového kódu použito. Kód obsahuje implementaci modelů a rovněž zdrojové vykreslení všech tabulek a grafů použitých v této diplomové práci.

7 Seznam použitých zdrojů

[1] WORLD HEALTH ORGANIZATION. *Weekly epidemiological update on COVID-19–31 December 2021*. WHO [online]. 2021 [cit. 2025-01-12]. Dostupné z: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---21-december-2021.

[2] LEKARSKE.SLOVNIKY.CZ. *Zoonotický* [online]. [cit. 2025-01-12]. Dostupné z: https://lekarske.slovniky.cz/pojem/zoonoticky.

[3] KOMENDA M., PANOŠKA P., BULHART V., ŽOFKA J., BRAUNER T. et al. *COVID-19: Přehled aktuální situace v ČR*. Onemocnění aktuálně [online]. Praha: Ministerstvo zdravotnictví ČR, 2020 [cit. 2025-01-12]. Dostupné z: https://onemocneni-aktualne.mzcr.cz/covid-19.

[4] RADA EVROPSKÉ UNIE. *Lessons learned in health: The Council approves conclusions*. Consilium [online]. 2020 [cit. 2025-01-12]. Dostupné z: https://www.consilium.europa.eu/en/press/press-releases/2020/12/18/covid-19-lessons-learned-in-health-the-council-approves-conclusions/.

[5] MASARYKOVA UNIVERZITA**.** *Matematici a biostatistici z MU vytvořili unikátní nástroj pro modelování epidemií* [online]. Brno: Masarykova univerzita, 2020 [cit. 2025-01-13]. Dostupné z: https://www.em.muni.cz/veda-a-vyzkum/14118-matematici-a-biostatistici-z-mu-vytvorili-unikatni-nastroj-pro-modelovani-epidemii.

[6] SOCIOLOGICKÝ ÚSTAV AV ČR**.** *Výzkumná zpráva zaměřená na analýzy fyziologických a sociálních rizik nákazy COVID-19* [online]. Praha: Sociologický ústav AV ČR, 2021 [cit. 2025-01-13]. Dostupné z: https://www.soc.cas.cz/cz/publikace/vyzkumna-zprava-zamerena-na-analyzy-fyziologickych-socialnich-rizik-nakazy-covid-19.

[7] GISAID. *Global Initiative on Sharing Avian Influenza Data*. GISAID [online]. 2025 [cit. 2025-01-14]. Dostupné z: https://www.gisaid.org/.

[8] ILANI, M. A. et al. *COVID-19 Probability Prediction Using Machine Learning: An Infectious Approach*. Arxiv [online]. 2024 [cit. 2025-01-15]. Dostupné z: https://arxiv.org/abs/2408.12841.

[9] ČSKB. *Big data, strojové učení a umělá inteligence v klinické laboratoři*. Klinická biochemie [online]. 2022 [cit. 2025-01-15]. Dostupné z: https://www.cskb.cz/wp-content/uploads/2022/12/KBM_3_2022_Friedecky-bigdata-92.pdf.

[10] EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL. *Pandemic preparedness plan*. ECDC [online]. 2024 [cit. 2025-01-15]. Dostupné z: https://www.ecdc.europa.eu/en/news-events/ecdc-issues-recommendations-strengthening-emergency-and-pandemic-preparedness-planning.

[11] NAQVI, A. A. T., FATIMA, K., MOHAMMAD, T., FATIMA, U., SINGH, I. K et al. *Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach* [obrázek]. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 2020, roč. 1866, č. 10, s. 165878. [cit. 2025-01-15]. Dostupné z DOI: https://www.doi.org/10.1016/j.bbadis.2020.165878.

[12] HARRISON, A. G., LIN, T., WANG, P. *Mechanisms of SARS-CoV-2 Transmission and Pathogenesis*. Trends in Immunology [online]. prosinec 2020, roč. 41, č. 12, s. 1100–1115. [cit. 2025-01-15]. Dostupné z DOI: https://www.doi.org/10.1016/j.it.2020.10.004.

[13] NZIP. *angiotenzin-konvertující enzym 2* [online]. [cit. 2025-01-16]. Dostupné z: https://www.nzip.cz/rejstrikovy-pojem/1580.

[14] ECDC**.** *SARS-CoV-2 variants of concern as of 2021* [online]. Stockholm: ECDC, 2021 [cit. 2025-01-17]. Dostupné z: https://www.ecdc.europa.eu/en/covid-19/variants-concern.

[15] WORLD HEALTH ORGANIZATION. *Transmission of SARS-CoV-2* [online]. WHO, 2021 [cit. 2025-01-17]. Dostupné z: https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions.

[16] CDC COVID-19 RESPONSE TEAM. *Emergence of SARS-CoV-2 B.1.1.7* Lineage — United States, December 29, 2020–January 12, 2021. *Morbidity and Mortality Weekly Report (MMWR)* [online]. 2021, roč. 70, č. 3, s. 95–99 [cit. 2025-01-17]. Dostupné z DOI: https://www.doi.org/10.15585/mmwr.mm7003e2.

[17] VITIELLO, A., et al. *Advances in the Omicron variant development*. Journal of Internal *Medicine* [online]. 2022, roč. 292, č. 1, s. 81–90 [cit. 2025-01-17]. Dostupné z DOI: https://www.doi.org/10.1111/joim.13453.

[18] SOUPVECTOR**.** *SARS-CoV-2 radial distance tree*. *Wikimedia Commons* [Obrázek]. 2021 [cit. 2025-01-17]. Dostupné z: https://commons.wikimedia.org/w/index.php?curid=112983798.

[19] CENTERS FOR DISEASE CONTROL AND PREVENTION**.** *Risk Factors for Severe COVID-19 Illness* [online]. Atlanta: CDC, 2021 [cit. 2025-01-17]. Dostupné z: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html.

[20] NZIP. *Postcovidový syndrom* [online]. [cit. 2025-01-19]. Dostupné z: https://www.nzip.cz/clanek/1400-postcovidovy-syndrom.

[21] CDC**.** *Diagnostic Testing for COVID-19* [online]. Atlanta: CDC, 2021 [cit. 2025-01-19]. Dostupné z: https://www.cdc.gov/covid/testing/index.html.

[22] WHO**.** *COVID-19 Clinical Management: Living Guidance* [online]. WHO, 2021 [cit. 2025-01-19]. Dostupné z: https://www.who.int/publications/i/item/WHO-2019-nCoV-clinical-2021-2.

[23] EUROPEAN MEDICINES AGENCY**.** *COVID-19 Vaccines: Authorised Medicines* [online]. EMA, 2022 [cit. 2025-01-19]. Dostupné z: https://www.ema.europa.eu/en/human-regulatory-overview/public-health-threats/coronavirus-disease-covid-19/covid-19-medicines.

[24] ECDC**.** *COVID-19 Pandemic Overview* [online]. Stockholm: ECDC, 2021 [cit. 2025-01-19]. Dostupné z: https://www.ecdc.europa.eu/en/covid-19.

[25] WHO. *Modes of Transmission of SARS-CoV-2* [online]. Geneva: WHO, 2020 [cit. 2025-01-19]. Dostupné z: https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions.

[26] JOHN HOPKINS UNIVERSITY**.** *COVID-19 Global Dashboard* [online]. Baltimore: JHU, 2021 [cit. 2025-01-19]. Dostupné z: https://coronavirus.jhu.edu/map.html.

[27] MARTIN, K., PULLIN, M., CLARKE, R., WILSON, R., KLEIJNEN, J. *The quality of COVID-19 systematic reviews during the coronavirus pandemic*. Systematic Reviews [online]. 2024, roč. 13, č. 1, s. 52 [cit. 2025-01-19]. Dostupné z DOI: https://doi.org/10.1186/s13643-024-02552-x.

[28] FERGUSON, N. M., et al. *Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand*. Imperial College London [online dokument]. 2020 [cit. 2025-01-19]. Dostupné z: https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf.

[29] BEIGEL, J. H., et al. *Remdesivir for the Treatment of COVID-19 — Final Report*. The Lancet [online]. 2020, roč. 395, č. 10238, s. 1569–1578 [cit. 2025-01-20]. Dostupné z DOI: https://www.doi.org/10.1056/NEJMoa2007764.

[30] INTERNATIONAL MONETARY FUND. *World Economic Outlook: The Great Lockdown*. Washington D.C. [online]. 2020 [cit. 2025-01-20]. Dostupné z: https://www.imf.org/en/Publications/WEO/Issues/2020/04/14/weo-april-2020.

[31] WHO. *Mental Health and COVID-19*. WHO [online]. 2021 [cit. 2025-01-20]. Dostupné z: https://www.who.int/teams/mental-health-and-substance-use/mental-health-and-covid-19.

[32] KANNAN, H., KHANRA, S., RODRIGUEZ, R., JARAMILLO, J. *Machine Learning for Business Analytics: Real-Time Data Analysis for Decision-Making*. [e-kniha]. Milton: Productivity Press, 2022. ISBN 9781000615425.

[33] JAVID, A., LIANG, X., VENKITARAMAN, A., CHATTERJEE, S. *Predictive Analysis of COVID-19 Time-series Data from Johns Hopkins University*. arXiv preprint [online]. 2020 [cit. 2025-01-20]. Dostupné z: https://arxiv.org/abs/2005.05060.

[34] RASTOGI, V., JAIN, A. *COVID-19: Nature, Outbreak and role of Machine Learning*. International Journal of Advanced Research in Engineering and Technology (IJARET) [online]. Prosinec 2020, roč. 11, č. 12, s. 2387–2395 [cit. 2025-01-20]. DOI: 10.34218/IJARET.11.12.2020.225. Dostupné z: http://iaeme.com/Home/issue/IJARET?Volume=11&Issue=12.

[35] EVROPSKÝ PARLAMENT**.** *Zpráva o pandemii onemocnění COVID-19: získané zkušenosti a doporučení pro budoucnost* [online]. 2023 [cit. 2025-01-21]. Dostupné z: https://www.europarl.europa.eu/doceo/document/A-9-2023-0217_CS.html.

[36] JOHN HOPKINS UNIVERSITY**.** *Predicting COVID-19 Waves Using Time Series Analysis* [online]. Baltimore, 2020 [cit. 2025-01-21]. Dostupné z: https://arxiv.org/abs/2005.05060.

[37] EUROPEAN DATA PROTECTION BOARD**.** *Guidelines on Data Privacy in COVID-19 Tracking* [online]. Brussels, 2020 [cit. 2025-01-21]. Dostupné z: https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_20200420_contact_tracing_covid_with_annex_en.pdf.

[38] LIU, Y. *Python Machine Learning by Example: Build Intelligent Systems Using Python, TensorFlow 2, Pytorch, and Scikit-Learn*. [e-kniha]. Birmingham: Packt Publishing, Limited, 2020. ISBN 9781800203860.

[39] UNPINGCO, J. *Python for Probability, Statistics, and Machine Learning*. New York, NY: Springer Berlin Heidelberg, 2016. ISBN 978-3319307152.

[40] WITTEN, I. H., FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Elsevier Science, 2005. ISBN 978-0-12-088407-0.

[41] STACK OVERFLOW. *Developer Survey Results 2021*. *Stack Overflow* [online]. 2021 [cit. 2025-01-21]. Dostupné z: https://insights.stackoverflow.com/survey/2021.

[42] ISLAM SUJAN, N. *Top 5 Machine Learning Libraries in Python*. Towards Data Science [online]. 22. listopadu 2018 [cit. 2025-01-22]. Dostupné z: https://medium.com/data-science/top-5-machine-learning-libraries-in-python-e36e3e0e02af.

[43] BROWNLEE, J. *Time Series Forecasting With Prophet in Python*. Machine Learning Mastery [online]. 8. května 2020 [cit. 2025-01-22]. Dostupné z: https://www.machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/.

[44] DASGUPTA, N. *Practical big data analytics: hands-on techniques to implement enterprise analytics and machine learning using hadoop, spark, NoSQL and R*. Birmingham: Packt, 2018. ISBN 978-1783554393.

[45] AVIMANYU786**.** *How deep learning is a subset of machine learning and how machine learning is a subset of artificial intelligence* (AI). Wikimedia Commons [Obrázek]. 2020 [cit. 2025-01-23]. Dostupné z: https://commons.wikimedia.org/wiki/File:AI-ML-DL.png.

[46] HAYKIN, S. *Neural networks and learning machines*. Upper Saddle River: Prentice Hall, 2009. ISBN 978-0-13-129376-2.

[47] QUDDUS, J. *Machine learning with apache spark quick start guide: uncover patterns, derive actionable insights, and learn from big data using MLlib*. Birmingham: Packt, 2018. ISBN 978-1789346565.

[48] GOODFELLOW, I., BENGIO, Y., COURVILLE, A. *Deep Learning*. Cambridge: MIT Press, 2016. ISBN 9780262035613.

[49] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009. ISBN 9780387848570.

[50] MURPHY, K.P. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012. ISBN 9780262018029.

[51] BALKISS.HAMAD**.** *Difference between supervised and unsupervised learning*. Wikimedia Commons [Obrázek]. 2018 [cit. 2025-01-24]. Dostupné z: https://commons.wikimedia.org/wiki/File:Supervised_and_unsupervised_learning.png.

[52] SUTTON, R.S.; BARTO, A.G. *Reinforcement Learning: An Introduction*. Cambridge: MIT Press, 2018. ISBN 9780262039246.

[53] MEGAJUICE. *Components in a typical Reinforcement Learning (RL) system.* Wikimedia Commons [Obrázek]. 2017 [cit. 2025-01-24]. Dostupné z: https://commons.wikimedia.org/wiki/File:Reinforcement_learning_diagram.svg.

[54] KRISHNAVEDALA**.** Linear least square example. *Wikimedia Commons* [Obrázek]. 2011 [cit. 2025-01-24]. Dostupné z: https://commons.wikimedia.org/wiki/File:Linear_least_squares_example2.svg.

[55] GILGOLDM. *Decision Tree*. Wikimedia Commons [Obrázek]. 2020 [cit. 2025-01-24]. Dostupné z: https://commons.wikimedia.org/wiki/File:Decision_Tree.jpg.

[56] CHOUDHARY, A. *Generate Quick and Accurate Time Series Forecasts using Facebook's Prophet*. [online]. AnalyticsVidhya. 2025. [cit. 2025-02-15]. Dostupné z: https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-prophet-python-r/.

[57] CHOUHBI, K. *XGBoost: Introduction.* [online]. Medium. 2019 [cit. 2025-02-15]. Dostupné z: https://medium.com/xgboost-all-you-need-to-know/introduction-7c4d8f94bbe1.

[58] AHMED, S. et. al. *The Deep Learning ResNet101 and Ensemble XGBoost Algorithm with Hyperparameters Optimization Accurately Predict the Lung Cancer*. Applied Artificial Intelligence [obrázek]. 2023, 37(1). [cit. 2025-02-15]. Dostupné z DOI: https://doi.org/10.1080/08839514.2023.2166222.

[59] AI PLUS INFO. *Introduction to Long Short-Term Memory (LSTM).* [online]. 2023. [cit. 2025-02-16]. Dostupné z: https://www.aiplusinfo.com/blog/introduction-to-long-short-term-memory-ltsm/.

[60] CHEVALIER, G. *The LSTM Cell in Artificial Neural Networks (ANNs).* Wikimedia Commons [obrázek]. 2018 [cit. 2025-02-16]. Dostupné z: https://commons.wikimedia.org/wiki/File:The_LSTM_Cell.svg.

[61] OUR WORLD IN DATA. *COVID-19 Dataset by Our World in Data.* [online]. 2024. [cit. 2025-02-17]. Dostupné z: https://github.com/owid/covid-19-data/.

[62] JUPYTER. *Project Jupyter*. [online]. 2025. [cit. 2025-02-18]. Dostupné z: https://jupyter.org/.

[63] EUROSTAT. *Real GDP growth rate – volume.* [online]. 2025. [cit. 2025-03-05]. Dostupné z: https://ec.europa.eu/eurostat/databrowser/view/tec00115/default/table?lang=en&category=t_na10.t_nama10.t_nama_10_ma.

[64] CDC. *Wastewater COVID-19 National and Regional Trends*. [online]. 2025. [cit. 2025-03-10]. Dostupné z: https://www.cdc.gov/nwss/rv/COVID19-nationaltrend.html#about-wastewater-viral-activity-level.

8 Seznam obrázků, tabulek, grafů a zkratek

8.1 Seznam obrázků

8.2 Seznam tabulek

8.3 Seznam grafů

8.4 Seznam zdrojových kódů

8.5 Seznam použitých zkratek

Zkratka	Význam	Překlad
WHO	World Health Organization	Světová zdravotnická organizace
ECDC	European Centre for Disease Prevention and Control	Evropské středisko pro prevenci a kontrolu nemocí
CDC	Centers for Disease Control and Prevention	Střediska pro kontrolu a prevenci nemocí
ML	Machine Learning	Strojové učení
AI	Artificial Intelligence	Umělá inteligence
API	Application Programming Interface	Rozhraní pro programování aplikací
DL	Deep Learning	Hluboké učení
LSTM	Long Short-Term Memory	Dlouhá krátkodobá paměť
OWID	Our World in Data	Náš svět v datech
GPU	Graphics Processing Unit	Grafický procesor (Grafická karta)
TPU	Tensor Processing Unit	Tensorový procesor

Přílohy

Příloha A – Zdrojový kód deskriptivní analýzy a prediktivních modelů v jazyce Python (vlastní tvorba)

Zdrojový kód je přiložen do příloh práce v systému IS. Obsahuje deskriptivní analýzu a jednotlivé prediktivní modely, včetně vykreslení veškerých tabulek a grafů obsažených v praktické části práce, pokud není uvedeno jinak. Github repozitář (viz níže) slouží jako alternativní úložiště výsledného kódu. Pro spuštění souboru *.ipynb* použijte vývojové prostředí Google Colab.

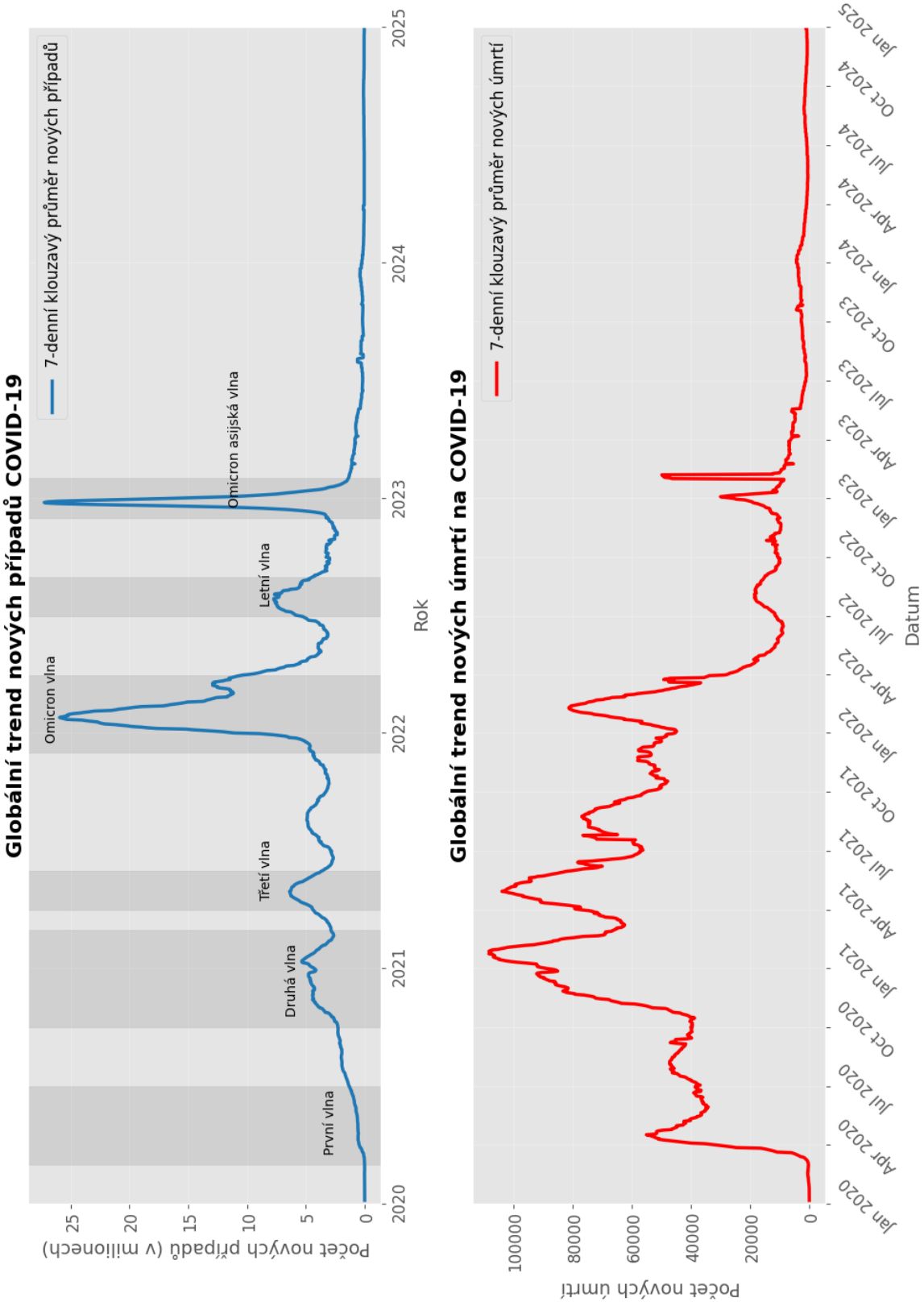
- Název souboru: Bilek_zdrojovy_kod.ipynb

Odkaz na Github repozitář: https://github.com/MatejBilek99/DP_Predikce_2025/

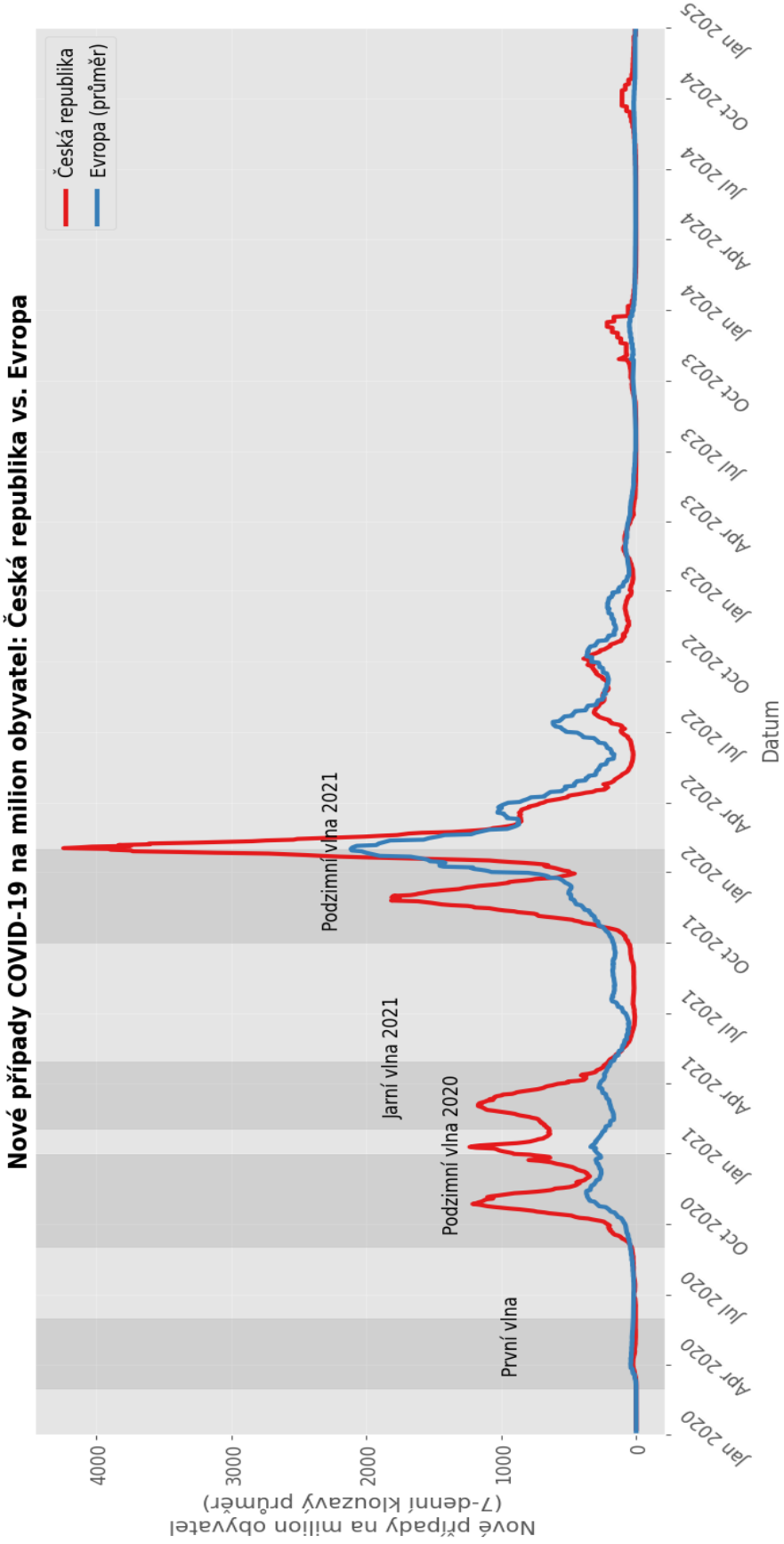
Odkaz na vývojové prostředí Google Colab: https://colab.research.google.com/

Příloha B – Grafy deskriptivní analýzy (vlastní tvorba)

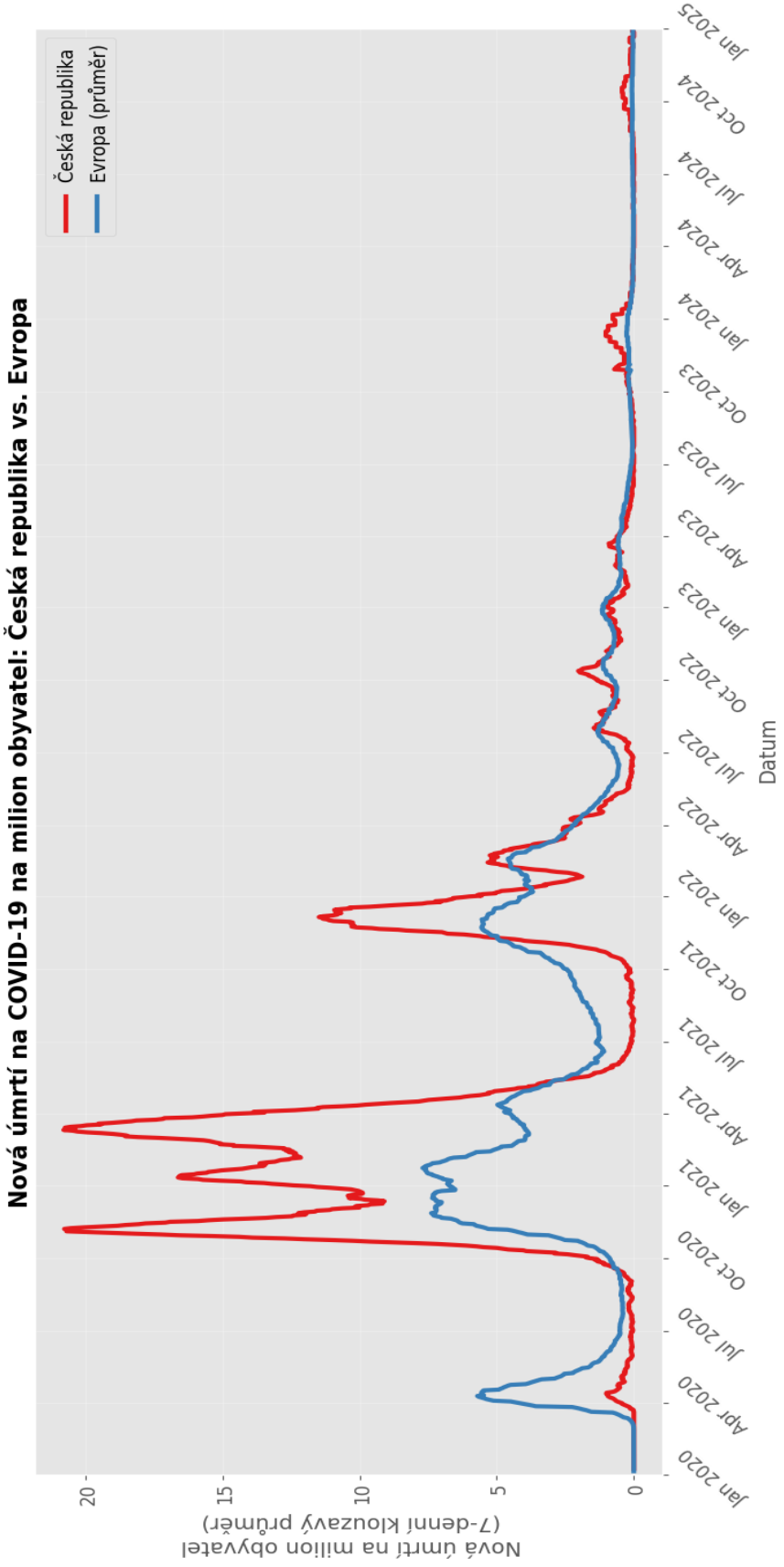
B.1 Globální trendy případů a úmrtí nemoci COVID-19 (2020–2024) [s. 53]



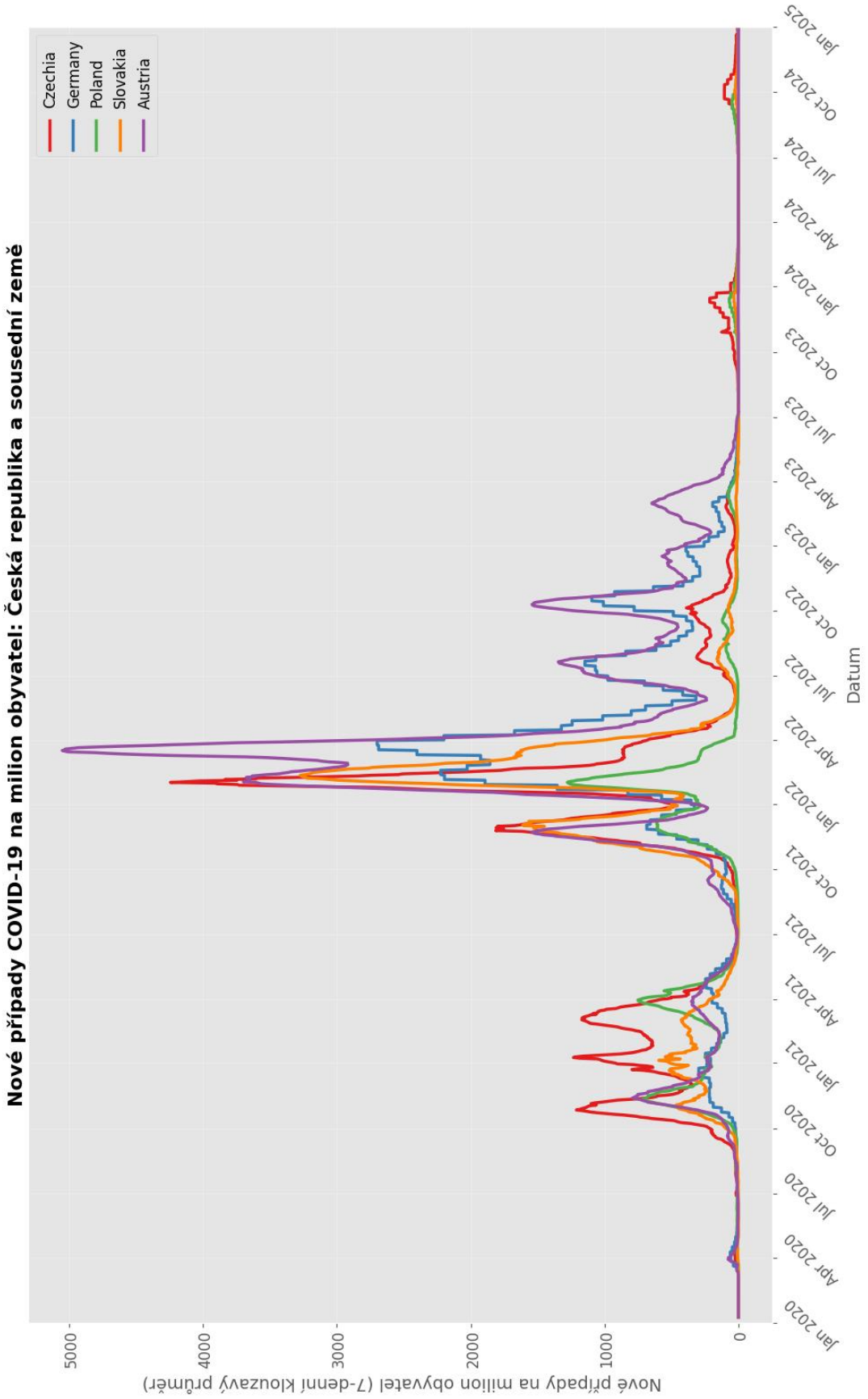
B.2 Trendy šíření COVID-19 v Česku a Evropě (2020–2024) [s. 55]



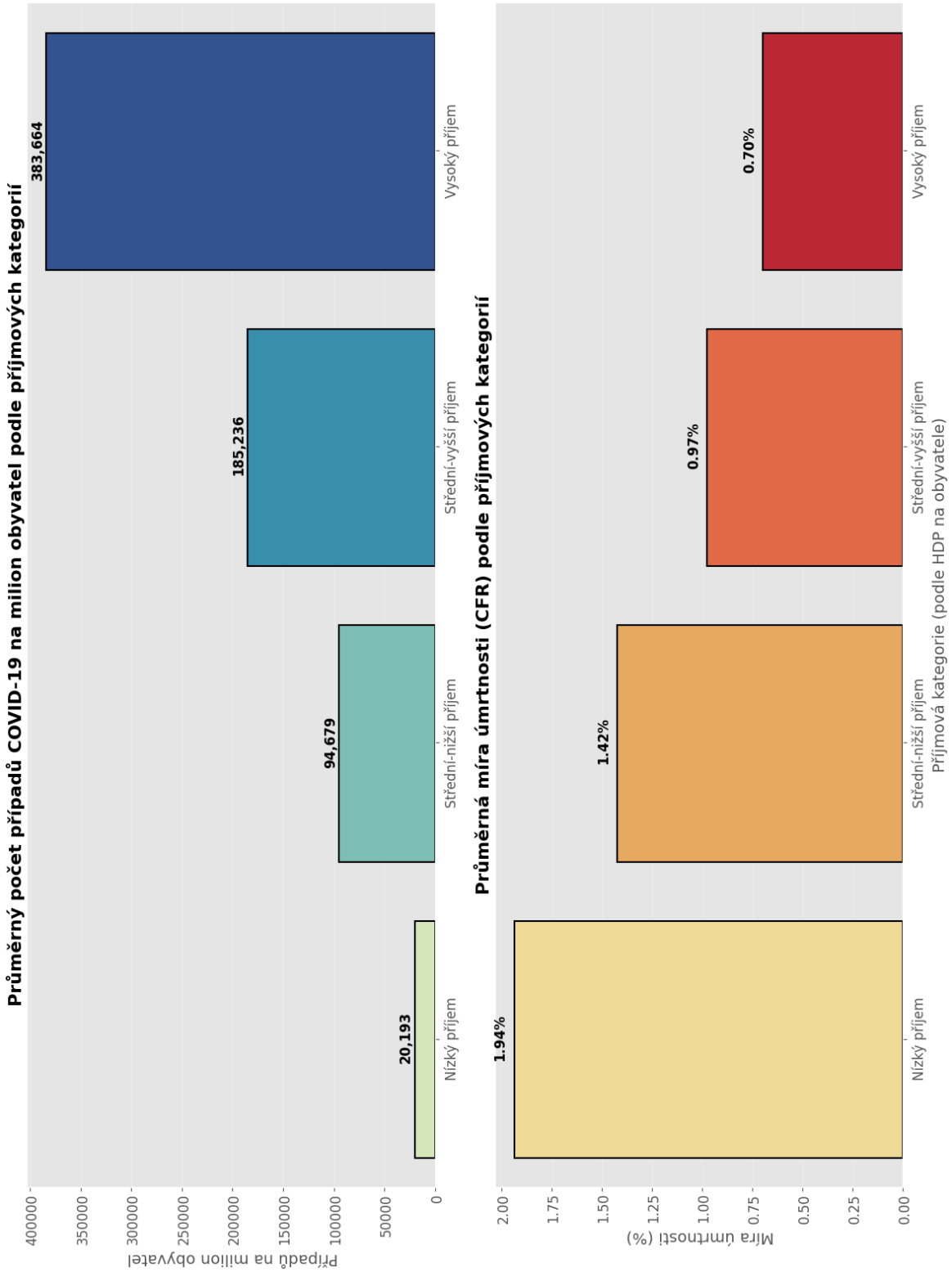
B.3 Trendy úmrtí COVID-19 v Česku a Evropě (2020–2024) [s. 55]



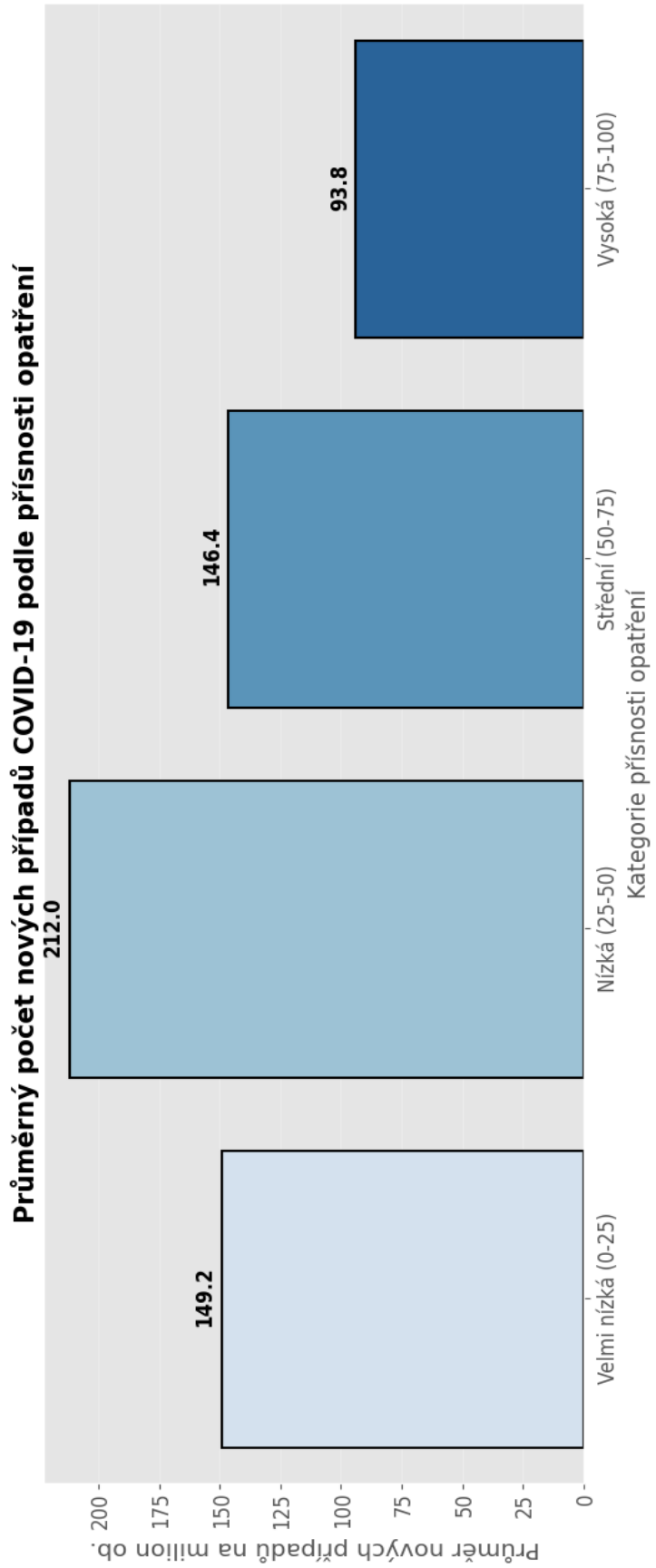
B.4 Nové případy COVID-19: ČR a sousední země (2020–24) [s. 56]



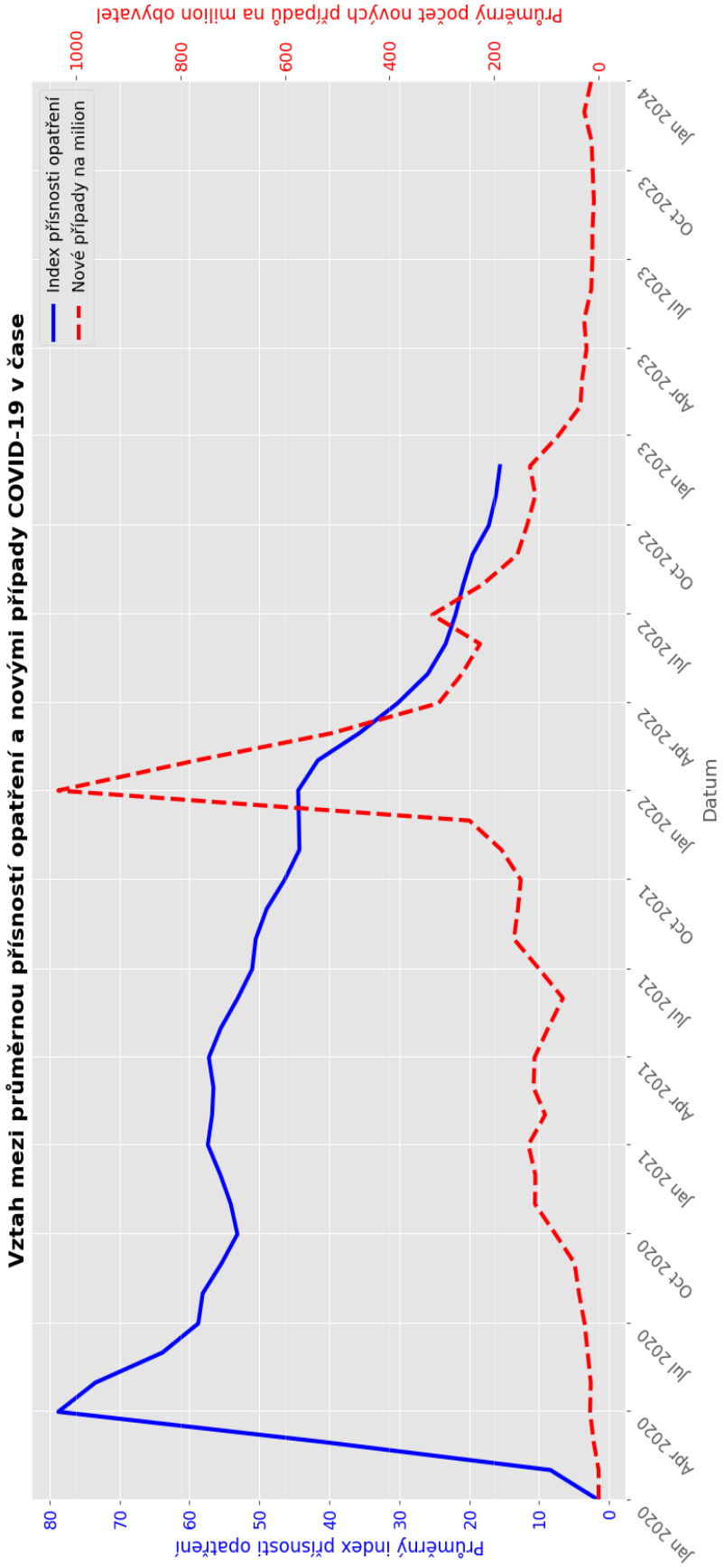
B.5 Průměrné případy a úmrtí podle příjmových kategorií [s. 58]



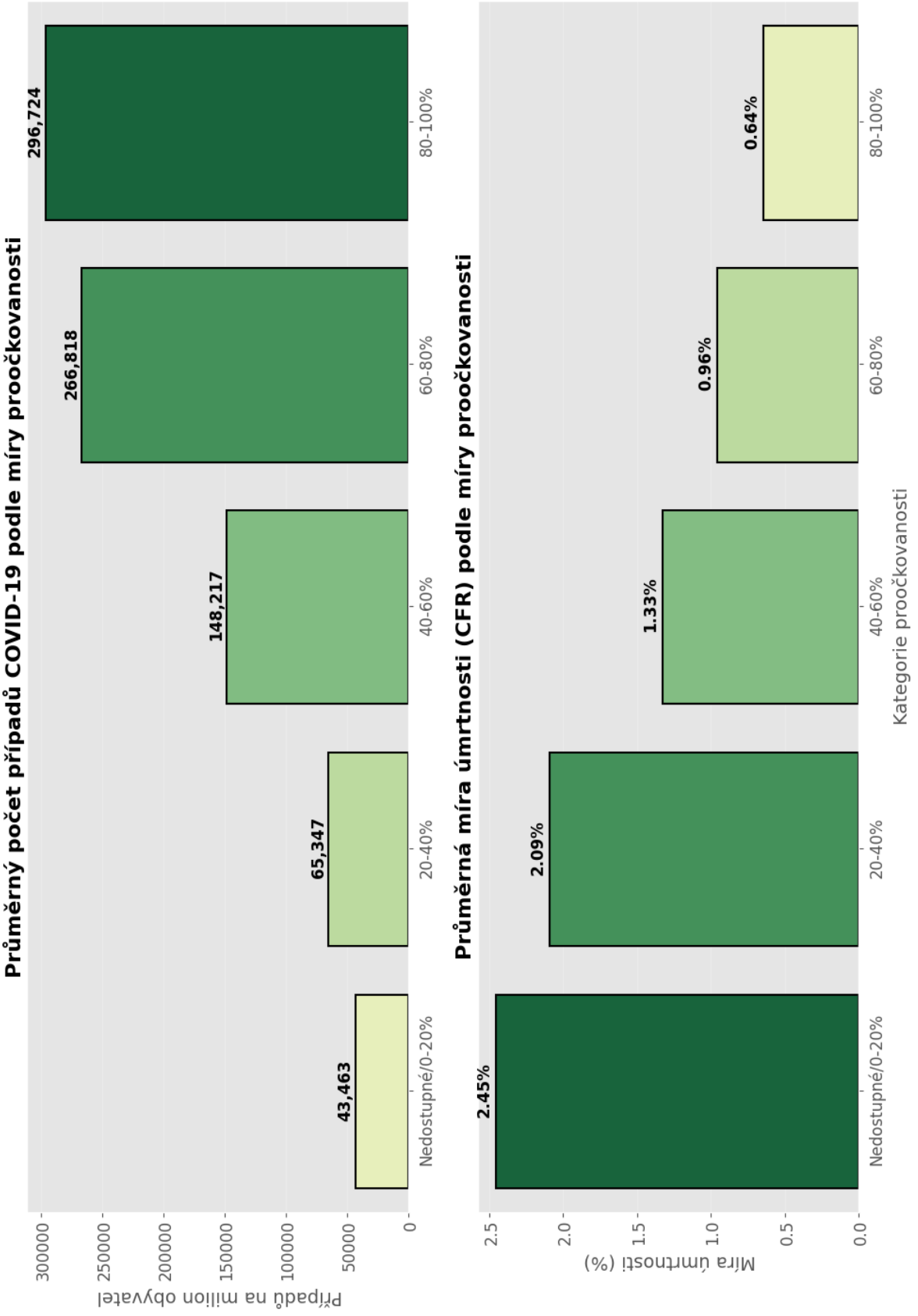
B.6 Průměrný počet nových případů COVID-19 podle přísnosti opatření [s. 59]



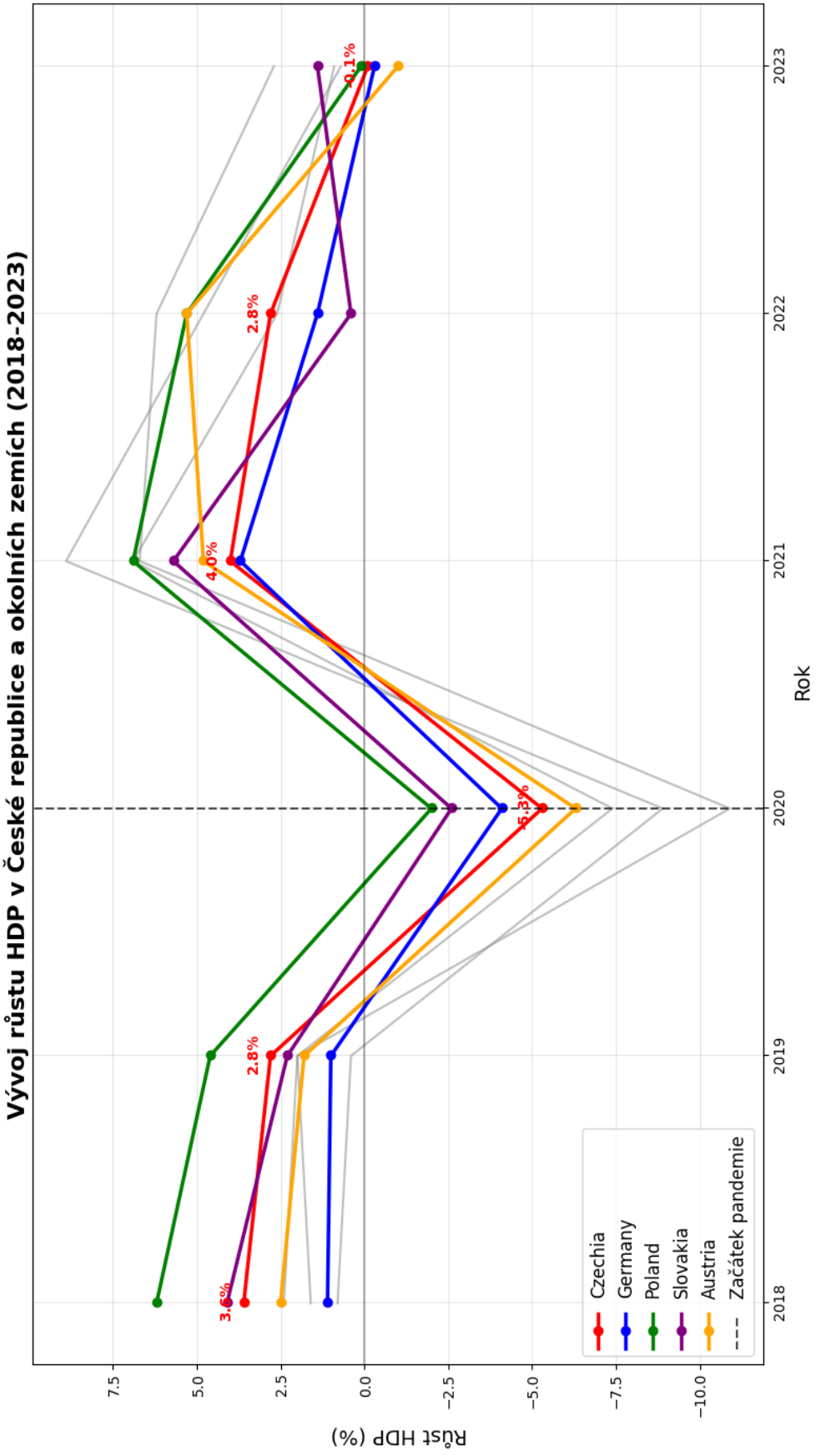
B.7 Vztah mezi Indexem přísnosti opatření a novými případy COVID-19 [s. 60]



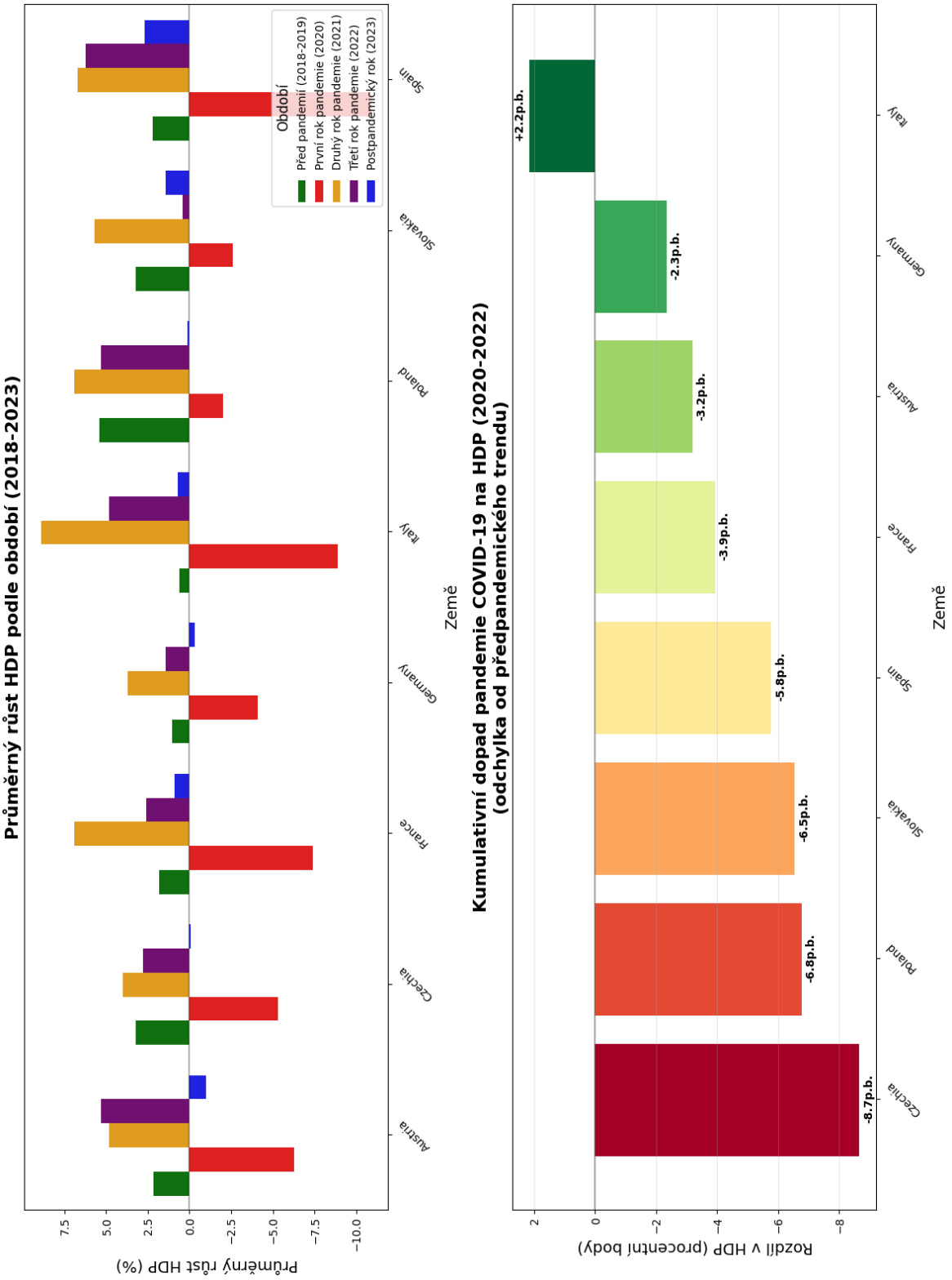
B.8 Globální průměr případů a úmrtnosti podle míry proočkovanosti [s. 61]



B.9 Vývoj růstu HDP v ČR a okolních zemích (2018-2023) [s. 62]

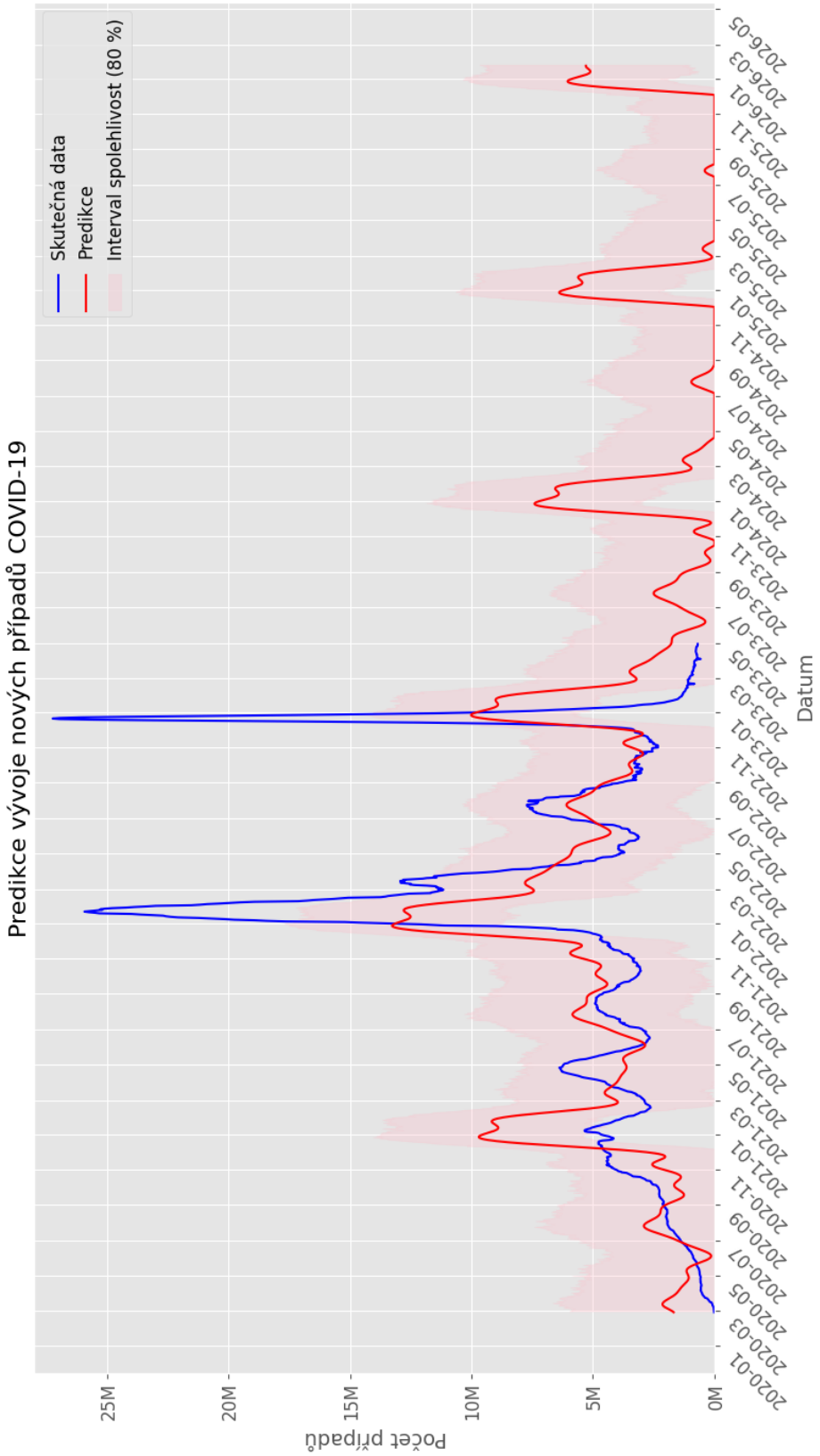


B.10 Změny HDP pro vybrané země během pandemie COVID-19 [s. 63]

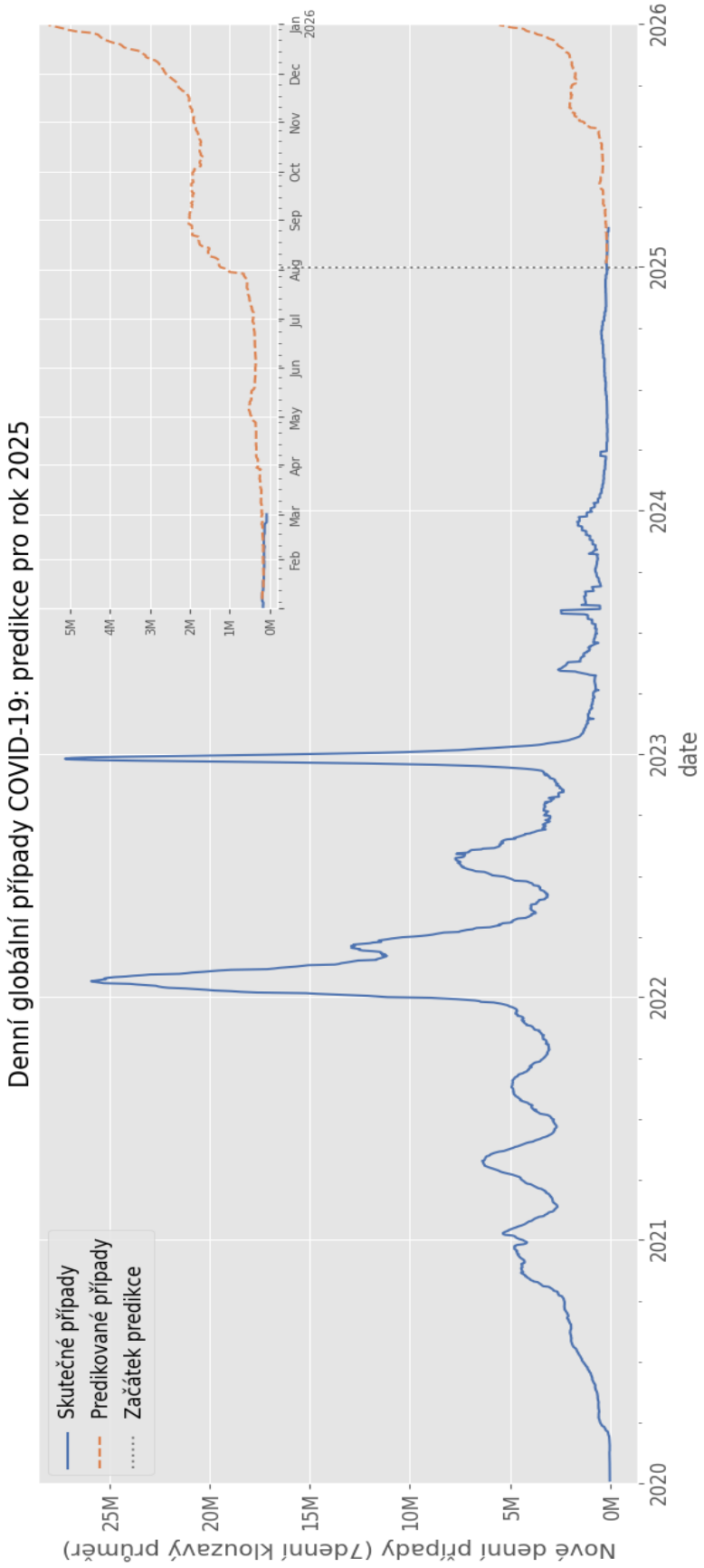


Příloha C: Grafy výsledků prediktivních modelů (vlastní tv.)

C.1 Predikce vývoje globálních nových případů COVID-19 modelem Prophet [s. 69]



C.2 Predikce vývoje globálních nových případů modelem XGBoost [s. 76]



C.3 Predikce nových globálních případů COVID-19 pomocí LSTM sítě [s. 83]