EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103005/I/2024/36124048422812164

PREDIKCIA ÚMRTIA PACIENTOV COVID-19 NA ZÁKLADE ICH MEDICÍNSKEJ HISTÓRIE VYUŽITÍM ALGORITMOV STROJOVÉHO UČENIA

Diplomová práca

2024 **Bc. Anna Hronská**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

PREDIKCIA ÚMRTIA PACIENTOV COVID-19 NA ZÁKLADE ICH MEDICÍNSKEJ HISTÓRIE VYUŽITÍM ALGORITMOV STROJOVÉHO UČENIA

Diplomová práca

Študijný program: Informačný manažment
Študijný odbor: Ekonómia a manažment
Školiace pracovisko: KŠ FHI - Katedra štatistiky
Vedúci záverečnej práce: Ing. Silvia Komara, PhD.

Bratislava 2024 **Bc. Anna Hronská**

ČESTNÉ VYHLÁSENIE

Čestne vyhlasujem, že túto diplomovú prácu som vypracovala samostatne a že všetka použitá literatúra bola uvedená v tejto práci.

Dátum: .....

POĎAKOVANIE

V prvom rade patrí nesmierna vďaka mojej rodine a blízkym, ktorí pri mne stáli celých 5 rokov štúdia, podporovali ma vo všetkých mojich rozhodnutiach a neľahkých chvíľach. Touto cestou by som sa tiež chcela poďakovať vedúcej diplomovej práce Ing. Silvii Komara, PhD., ktorá svojím prístupom bola značným uľahčením písania tejto práce. Vďaka patrí aj pedagógom z mojich zahraničných štúdií, ktorí mi darovali pevný základ praktických znalostí, z ktorých som mnohé čerpala aj v tejto práci. V neposlednom rade patrí poďakovanie spolužiakom, ktorí mi boli nápomocní počas tejto študijnej jazdy.

ABSTRAKT

HRONSKÁ, Anna: *Predikcia úmrtia pacientov COVID-19 na základe ich medicínskej histórie využitím algoritmov strojového učenia.* – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra štatistiky. – Ing. Silvia Komara, PhD.– Bratislava: FHI EU, 2024, 86 s.

Za hlavný cieľ tejto práce môžeme považovať predikciu úmrtnosti pacientov na ochorenie COVID-19, vzhľadom na ich zdravotný stav a dostupné charakteristiky použitím algoritmov strojového učenia. Diplomová práca je rozdelená do piatich kapitol. V jej obsahu je zahrnutých dvadsať grafov, štrnásť obrázkov a pätnásť tabuliek. V prvej kapitole je preskúmaný súčasný stav problematiky vo forme vymedzenia teoretických pojmov súvisiacich so strojovým učením a nachádza sa tu taktiež priblíženie témy COVID-19 v kontexte s predikciou. Druhá kapitola stanovuje hlavné a čiastkové ciele tejto práce. V tretej kapitole je predstavený proces strojového učenia z metodologického hľadiska. Štvrtá kapitola sa zaoberá výsledkami skúmania cieľov, kde sú konkrétne metódy aplikované vytvorením predikčných modelov strojového učenia, ktoré klasifikujú pacienta či zomrel na ochorenie COVID-19 alebo nie. Záverečná kapitola sumarizuje výsledky riešenia danej problematiky, ktoré boli získané z výskumu a taktiež objasňuje prínosy a nedostatky tejto práce, ktorá by mohla byť v budúcnosti rozšírená.

Kľúčové slová: pandémia, COVID-19, strojové učenie, Python, umelá inteligencia

ABSTRACT

HRONSKÁ, Anna: *Predicting the COVID-19 related death of patients based on the medical history using Machine Learning.* – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Statistics. – Ing. Silvia Komara, PhD. – Bratislava: FHI EU, 2024, 86 p.

As the main objective of this thesis can be considered the prediction of mortality of COVID-19 patients using machine learning algorithms, based on their health status and available characteristics. The thesis is divided into five chapters. In its content is included twenty charts, fourteen figures and fifteen tables. The first chapter reviews the current state of the subject in the form of defining the theoretical concepts related to machine learning and introduces the topic of COVID-19 in the context of prediction. The second chapter sets out the main and sub-objectives of this thesis. The third chapter introduces the machine learning process from a methodological perspective. The fourth chapter discusses the results of the research, where specific methods are applied by creating machine learning prediction models, which classify the patient whether he died of COVID-19 disease or not. The concluding chapter summarizes the results of addressed problem, that were obtained from the research, and clarifies the benefits besides the shortcomings of this work, which could be extended in the future.

Keywords: pandemic, COVID-19, machine learning, Python, artificial intelligence

Obsah

Zoznam ilustrácií a zoznam tabuliek

Zoznam skratiek a značiek

AI	Artificial Intelligence [Umelá inteligencia]
ARIMA	AutoRegressive Integrated Moving Average [Autoregresný integrovaný kĺzavý priemer]
DL	Deep Learning [HĺBkové učenie]
DM	Data Mining [HĺBková analýza údajov]
EDA	Exploratívna dátová analýza
JIS	Jednotka intenzívnej starostlivosti
KDD	Knowledge Discovery in Databases [Objavovanie znalostí v databázach]
KNN	K-Nearest Neighbour [K-najbližší sused]
ML	Machine Learning [Strojové učenie]
SML	Supervised Machine Learning [Strojové učenie s učiteľom]
SVM	Support Vector Machine [Algoritmus podporných vektorov]
UML	Unsupervised Machine Learning [Strojové učenie bez učiteľa]

Úvod

V decembri 2019 došlo k udalosti, ktorá pohla svetom a zapísala sa do svetových dejín. V čínskom meste Wuhan bolo v niekoľkých prípadoch zaznamenané, predtým neznáme ochorenie s príznakmi zápalu pľúc. Neskôr Svetová zdravotnícka organizácia (WHO) označila toto ochorenie ako COVID-19 a vyhlásila globálnu pandémiu. Napriek tomu, že krajiny zaviedli stratégie na spomalenie šírenia „korona" vírusu, svet stále bojuje s následkami. Hospodársky vplyv pandémie bol čoraz zjavnejší, keďže dochádzalo k výraznému spomaleniu globálneho hospodárskeho rastu. Zníženie cestovného ruchu, nedostatok zdravotníckeho personálu, obmedzenia na hraniciach, obmedzenia v službách, prepúšťanie zamestnancov, zníženie produktivity, ako aj bankroty, to všetko prispievalo k zníženiu HDP krajín. Vzhľadom na tieto skutočnosti, sme sa rozhodli túto prácu zamerať na predpovedanie následkov pandémie, aby sme poskytli cenné informácie pre viaceré sektory, s cieľom zlepšiť prípravné kroky v prípade možných pandémií v budúcnosti.

Predmetom tejto práce je predikcia úmrtnosti na ochorenie COVID-19, vzhľadom na lekársku anamnézu pacienta a iných dostupných informácií, ktoré by mohli ovplyvniť následky tohto ochorenia. Predikcia je uskutočňovaná využitím modelov strojového učenia. Okrem iného sa práca zameriava na detekciu vzorov medzi jednotlivými faktormi a úmrtím na toto ochorenie, s cieľom definovať fakty, ktoré by napomohli vo výskume tejto medicínskej problematiky. Rozloženie tejto práce je formované do piatich kapitol, ktoré približujú teoretickú základňu tejto témy, rovnako ako praktickú.

Prvá kapitola práce rozoberá tému strojového učenia v podobe vytýčenia teoretických pojmov. Taktiež je tu zahrnutá aj predikcia ochorenia COVID-19, na základe existujúcej literatúry v tomto odvetví. Druhá kapitola vymedzuje hlavné ciele, spolu s vedľajšími zámermi realizácie tohto výskumu. V tretej kapitole je priblížená metodika výskumu, ktorá opisuje celý proces a preferované metódy strojového učenia, ktoré slúžia na analýzu tejto práce. V štvrtej kapitole sú predstavené výsledky analýzy praktickej časti tejto práce. Môžeme tu nájsť priblíženie celého procesu výskumu, od získania a úpravy dát, až po aplikáciu zvolených metód na vytvorenie predikčných modelov strojového učenia, s cieľom analyzovať danú oblasť. V rámci tejto analýzy, tu vieme nájsť rôzne druhy zistení, explicitne aj vo vizuálnej forme. Záverečná kapitola sumarizuje výskum práce, v podobe zhrnutia interpretácie výsledkov. Vieme tu nájsť aj prínosy práce, rovnako aj možné nedostatky analýzy, kde sú predložené návrhy na zlepšenie v ďalšom dodatočnom výskume v danej oblasti.

1. Súčasný stav riešenej problematiky

V tejto kapitole sú vymedzené teoretické pojmy súvisiace s predikciou a strojovým učením ako takým. Okrem iného sme sa zamerali aj na priblíženie problematiky COVID-19 v súvislosti s predikciou.

1.1. Pojmy spojené s predikciou, vytvorenou využitím strojového učenia

Hĺbkovú analýzu údajov (Data mining = DM) vieme definovať ako zbierku metód aplikovaných na extrakciu použiteľných informácií z rozsiahleho súboru surových dát, v ktorom sa už dané informácie nachádzajú. Komplexnosť dát a ich viacrozmerný charakter určuje to, že bez matematických nástrojov nie je možné nájsť užitočné informácie. Cieľom DM je nachádzať vzory, ktoré sú existujúce, ale dôkladne skryté. Ich skrytosť zabezpečuje veľké množstvo premenných, šum dát alebo komplikácie pri prepojení premenných súčasne jednorozmerným spôsobom. Dôležitým faktom DM je, že neobsahuje žiadny proces učenia. (Amigo, 2021)

Pod pojmom **strojové učenie** (Machine learning = ML), si vieme predstaviť spôsob analýzy údajov, ktorý učí stroje vykonávať učenie zo skúseností, ktoré pre ľudí pôsobí prirodzene. Machine learning algoritmy využívajú výpočtové metódy, ktoré slúžia na nadobudnutie informácií priamo z dát, bez spoľahnutia sa na vopred definovanú rovnicu. S narastajúcim počtom vzoriek, ktoré sú využiteľné na učenie, je výkon algoritmov adaptívne zlepšovaný. Dané algoritmy majú schopnosť nachádzať v údajoch prirodzené vzory, ktoré majú tendenciu zlepšovať predikcie a rozhodnutia. (Štalmachová & Strenitzerová, 2020)

V kontexte s data miningom, machine learningom, štatistikou, dátovou analýzou či dátovou vedou, môže nastať istý chaos v prelínaní definícií. DM čerpá z rôznych odborov, avšak najviac zo štatistiky a teórie machine learningu. **Štatistika** vyzdvihuje matematickú presnosť a vznik teoretického základu skôr, ako je nastolená aplikácia overená praxou. Na druhej strane, machine learning vychádza z počítačovej praxe, ktorá sa zameriava na zisťovanie efektívnych procesov bez formálnych dôkazov. Preto sa ML stretáva skôr s termínom algoritmus, kde naproti tomu štatistika využíva pojem model. (Terek et. al., 2010) **Dátová analýza** je procesom kontroly, transformácie a modelovania údajov so zámerom navrhnutia záverov. Je široko využívaná v mnohých odvetviach s cieľom ponúknuť efektívnejšie riešenia pre obchodné rozhodnutia. Pri pojme **dátová veda** je možné si predstaviť vednú oblasť o systémoch na získavanie poznatkov z dát, ktoré sú

pokračovaním oblastí dátovej analýzy, ako je ML, DM či štatistika, rovnako ako **objavovanie znalostí v databázach** (Knowledge Discovery in Databases = KDD). Z tohto je zrejmé, že ML nemôžeme považovať za jedinú oblasť zaoberajúcu sa dátami na učenie, ktoré sú ďalej využívané. V každom odvetví sa spomínajú takmer identické procesy a techniky, ktoré sa však zaoberajú rôznymi témami, ale majú množstvo presahov. Všetky sú prakticky rovnaké, avšak v každej existuje istý rozdiel alebo významový odtieň. Faktom je, že každý okruh sa historicky odvíja od aplikácie, na ktorú bol vyvinutý. Ich pôvod preto určuje vymedzenie daných pojmov. (Swamynathan, 2017)

ML je druh disciplíny, ktorá sa zameriava na vývoj algoritmov, ktoré sú navrhnuté na aplikáciu dátových súborov, kde medzi hlavné oblasti zámeru patrí klasifikácia, predikcia (regresia), zhlukovanie a zoskupovanie úloh. Dané úlohy sa delia na dve odvetvia: strojové učenie bez a s učiteľom. (Athey, 2019)

Strojové učenie **bez učiteľa** (Unsupervised machine learning = UML) obsahuje identifikáciu zhlukov pozorovaní, ktoré majú spoločné črty, z pohľadu ich vzájomnej závislosti (kovariancie). Tento typ machine learningu môže byť interpretovaný ako „redukcia dimenzionality“, najčastejšie využívaný pre video, text či obrázky. (Athey, 2019) UML má schopnosť nájsť v údajoch vnútorné štruktúry či skryté vzory. Je použiteľný vtedy, keď dáta nie je možné jednoducho klasifikovať. Model analyzuje dáta, ktoré pozostávajú zo vstupných dát neobsahujúcich označené odpovede, a na základe toho dedukuje závery, opisujúce skryté štruktúry v údajoch. Za najpoužívanejšiu techniku je považované zhlukovanie (clustering), ktoré na základe nezaradených dát vytvorí zoskupenia dát využitím skrytých vzorov. (Štalmachová & Strenitzerová, 2020)

Strojové učenie **s učiteľom** (Supervised machine learning = SML), zvyčajne zahŕňa súbor premenných *(X)*, ktoré slúžia na predikciu výstupu *(Y)*. (Athey, 2019) SML je druh strojového učenia, ktorý mapuje vstupné dáta na výstupné, na základe určených vzorových dvojíc „vstup-výstup“. Zámerom je určiť výstupnú premennú, vzhľadom na trénovací model, ktorý zaraďuje trénovacie údaje. Tento typ strojového učenia sa opiera o fakty, kde je známa reálna trieda údajov. Pre priblíženie tohto pojmu, si ho môžeme predstaviť na konkrétnom príklade klasifikácie obrázkov. Ak by sme chceli systém naučiť rozlišovať obrázky vtákov a mačiek, je potrebné naučiť algoritmus ako kategorizovať dané obrázky, a až potom je možné použiť tento algoritmus na nové údaje. Následne model určí zaradenie (vták/mačka) doposiaľ nevidených obrázkov. (Dike et. al., 2018)

Takýto typ SML úlohy je možné vyriešiť pomocou nasledujúcich krokov:

1. Selekcia trénovacieho súboru dát.
2. Zoskupenie trénovacieho súboru dát.
3. Definovanie reprezentatívnej vzorky vstupných dát trénovacej funkcie.
4. Kontrola konfigurácie trénovacej funkcie, spolu s vhodným trénovacím algoritmom.
5. Dokončenie dizajnu a spustenie trénovacieho algoritmu, na zber natrénovaných údajov.
6. Odhadnutie správnosti naučenej úlohy. (Dike et. al., 2018)

Vo všeobecnosti **predikcia** tkvie vo využívaní atribútov na predvídanie alebo odhad neznámych hodnôt iných atribútov. (Terek et. al., 2010) Pri použití tohto termínu je potrebné poznamenať, že podstatou je nezameriavať sa na predpoveď, ale skôr na určenie zaradenia pozorovaní. Zmyslom je, že obe X aj Y sú pozorované v trénovacej množine, kde zámerom je predikovať výsledky Y v nezávislom testovacom súbore, založenom na dosiahnutých hodnotách X pre každú jednotku tohto súboru. Predpokladá sa, že pozorovania sú nezávislé a spoločné rozdelenia pravdepodobností X aj Y sú rovnaké v trénovacom aj testovacom súbore. Tieto predpoklady sú podstatné pre fungovanie väčšiny ML metód. (Athey, 2019)

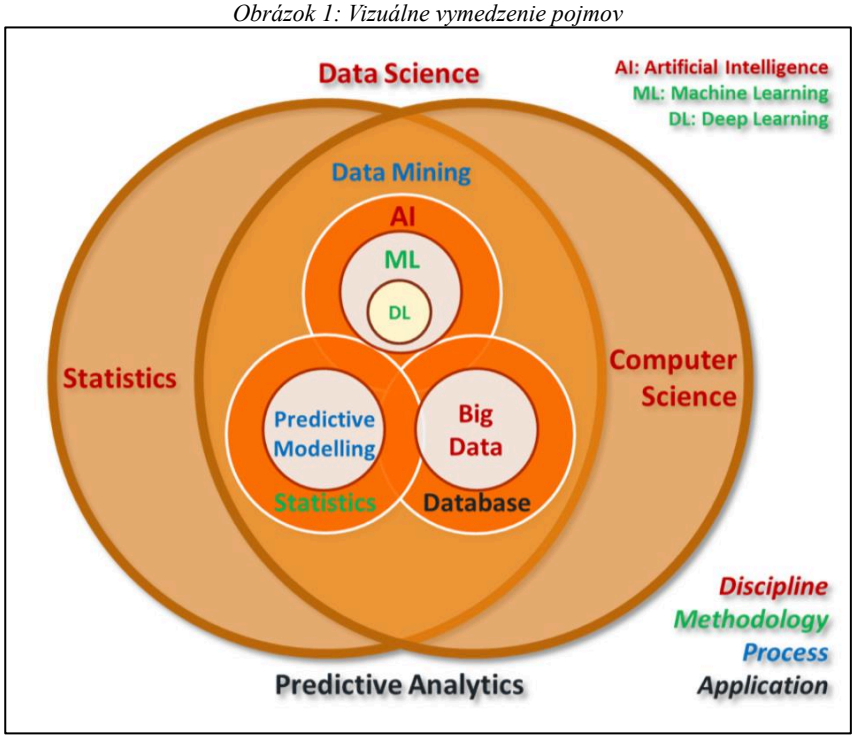
Pod pojmom **vzor** sa poníma lokálny model, ktorý určuje súvislosti v časti pozorovaní. (Terek et. al., 2010) Ide o spoľahlivú vzorku znakov alebo iných pozorovaných dátových charakteristík. Rozpoznávanie vzorov teda reprezentuje odhaľovanie vzorov v dátach. Aj keď sa vzory často vnímajú automaticky, tento postup je náročné definovať. (Héberger, 2008)

V prípade **klasifikácie** je hlavným účelom presne zaradiť pozorovania. Napríklad, ak je výsledkom zviera vyobrazené na obrázku, premenné alebo kovariancie sú pixely v obrázku, kde cieľom je ich správne klasifikovať, a tým docieliť korektne vyobrazené zviera. (Athey, 2019) Klasifikačné modely presne opisujú vzťahy medzi dátami a predikujú hodnoty pre ďalšie pozorovania. Úlohou týchto modelov je naučiť cieľovú premennú, ktorá mapuje jednotlivé atribúty súboru dát X, aby určila jednu konkrétnu preddefinovanú triedu Y. Existujú viaceré klasifikačné techniky, napríklad: stromové metódy, Bayesove klasifikátory, metóda podporných vektorov, umelé neurónové siete (ANN), metódy založené na pravidlách, či uvažovanie na základe pamäti. Pri klasifikácií sú testovacie dáta využívané na určenie presnosti klasifikačných modelov. Ak je táto presnosť akceptovateľná, potom daný model smie byť použitý aj na nové vstupy dát. (Silva & Fonseca, 2017)

S pojmom ML úzko súvisí termín **hĺbkového učenia** (Deep learning = DL), ktorý reprezentuje podskupinu metód ML. Algoritmy hĺbkového učenia sú zväčša založené na poznaných konvolučných neurónových sieťach (CNN) a hĺbkových neurónových sieťach (DNN). Tie majú základnú štruktúru vstupov a výstupov, kde sa využívajú skryté vrstvy, ktoré sa skladajú z tzv. neurónov. Táto komplexnosť spojení umožňuje výber prvkov zo surových dát nezávisle, bez ich predošlého spracovania alebo prípravy. Je však nutné si uvedomiť, že akýkoľvek poznatok, ktorý dokáže neurónová sieť získať zo surových dát, je už obsiahnutá v údajoch, na ktorých sa učí. DNN preto potrebuje veľmi rozsiahle množstvo údajov, aby bola schopná hľadať korektné informácie a riešiť vysoko nelineárne problémy. Taktiež má tendenciu sa nadštandardne prispôsobovať, nakoľko jej návrh je nastavený tak, aby modelovala minimálny rozptyl v údajoch. (Amigo, 2021)

Pojem **umelá inteligencia** (Artificial intelligence = AI) definuje inteligencia, ktorú demonštrujú stroje. V tomto scenári je to termín zastrešujúci všetky predchádzajúce definície, t. j. aplikácia hĺbkovej analýzy údajov, strojového učenia aj hĺbkového učenia. Bez zachádzania do väčších podrobností, AI pripúšťa možnosť logického uvažovania či interakcie algoritmov, so zámerom progresu výsledkov modelov. Kľúčovým úzkym profilom v tomto odbore je však prístupnosť validných dát. V priamej analógií, sa dá považovať ľudský mozog za najefektívnejší stroj uplatňujúci umelú inteligenciu, ktorý má sklon analyzovať informácie prijímané z analytických nástrojov, ktorými sú ľudské zmysly, a následne vie poskytnúť odpovede. Tiež je neustále trénovaný informáciami a edukačnými postupmi, ktoré sú označované ako vzdelávanie, teda v odbornej problematike nami spomínanej, ako „učenie“. (Amigo, 2021)

Obrázok 1: Vizuálne vymedzenie pojmov



Zdroj: (CopperTree Analytics, 2019)

1.2. Problematika predikčného modelovania COVID-19

Pandémia COVID-19 je považovaná za jednu z najnáročnejších zdravotných kríz modernej doby. Vývoj účinných stratégií, s víziou kontroly šírenia SARS-CoV-2, je jeden z hlavných cieľov tvorcov politík. ML a matematické modelovanie sa preukázali ako značne silné nástroje na optimalizáciu a usmerňovanie opatrení. (Sarmiento Varón et. al., 2023)

Pre lepšie priblíženie, SARS-CoV-2 je patogén podnecujúci koronavírusové ochorenie z roku 2019, označované ako COVID-19. Prejavy tohto kvapôčkového ochorenia sú na škále od miernych príznakov chrípky, až po akútny respiračný syndróm. Pri prenose vírusu medzi ľuďmi sa hromadia mutácie, z čoho následkom sú nové varianty, ktoré sú prevládajúce v rôznych krajinách. Vývoj vírusu určili vnútorné (rýchlosť mutácií), ale aj vonkajšie faktory (vek, imunita, stupeň ohrozenia, sociokultúrne faktory). Integrácia údajov ohľadom tohto vírusu je nevyhnutná pri predikcii správania sa, a aj jeho neprestajného šírenia. (Sarmiento Varón et. al., 2023)

Spomínaná téma vytvorila počas pandémie veľký rozruch v oblasti modelovania. Experti sa snažia o pochopenie tejto domény, aby bolo možné predpovedať budúci priebeh pandémie. Avšak, použitie modelov v tejto oblasti zaznamenalo značnú kritiku, založenú na viacerých chybných predpovediach, uskutočnených počas skorých mesiacov pandémie. (Nixon et. al., 2022) Pre ďalší výskum ohľadom tejto problematiky, odborníci neustále

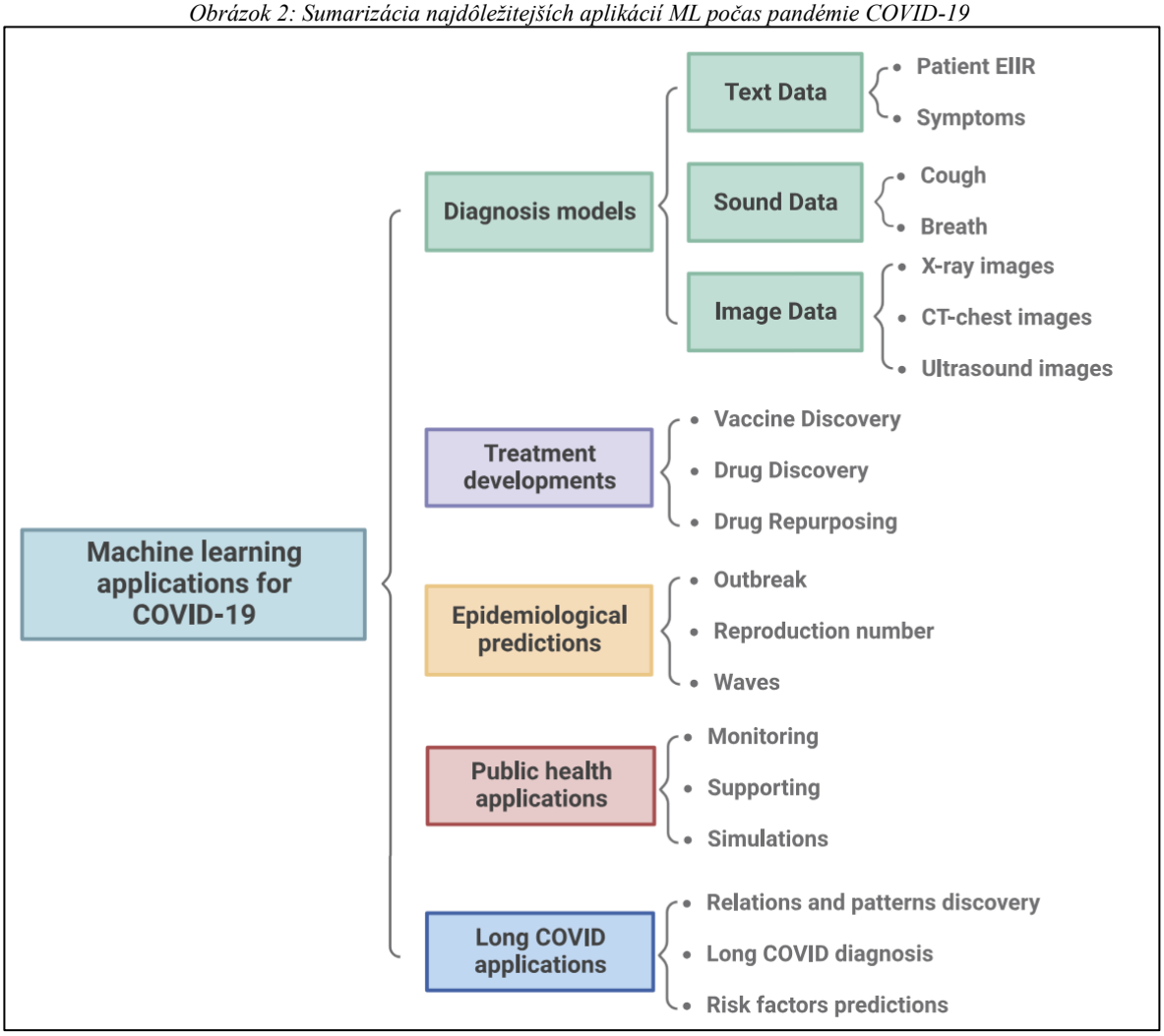
pracujú na vytvorení tzv. „pandemických modelov“, vždy keď sú dostupné nové dáta. Avšak, použitie epidemiologických modelov nie je novou myšlienkou pri predikciách. Takýto druh modelov bol vyvinutý Danielom Bernoullim v roku 1760. V súčasnosti, v odbore výskumu vírusových ochorení je vyvinutých viacero metód, ktoré zahŕňajú štatistické reprezentácie či počítačové simulácie. (Kim et. al., 2021)

Pokrokové ML algoritmy dokážu integrovať a vyhodnotiť obrovské množstvo údajov, ohľadom infikovaných pacientov, za účelom zlepšenia diagnostickej presnosti, vývinu novej a účinnej liečby či identifikácie ľudí s rizikom ochorenia. AI sa v poslednom období prejavila ako sľubný nástroj v oblasti medicíny, nakoľko preukazuje vysokú presnosť spracovania údajov, ktorá zabezpečuje presné rozhodovanie. (Tiwari et. al., 2022) V súlade s profesionálmi, je široko známe, že epidemiologické modely majú primárne ciele v podobe:

- progresu v pochopení rozšírenia vírusu,
- špecifikácií aspektov, ktoré ovplyvňujú rozšírenie vírusu,
- predpovedí v súvislosti s vybavením liečenia infikovaných pacientov s týmto vírusom [napr. ochranné prostriedky, počet lôžok, atď.].

Dané predikcie sú potrebné pri podniknutí preventívnych opatrení, za účelom zredukovania dopadu ochorenia. Môžu byť pokladané za prvok asistencie, v súvislosti s verejným zdravotníckym systémom. (Kim et. al., 2021)

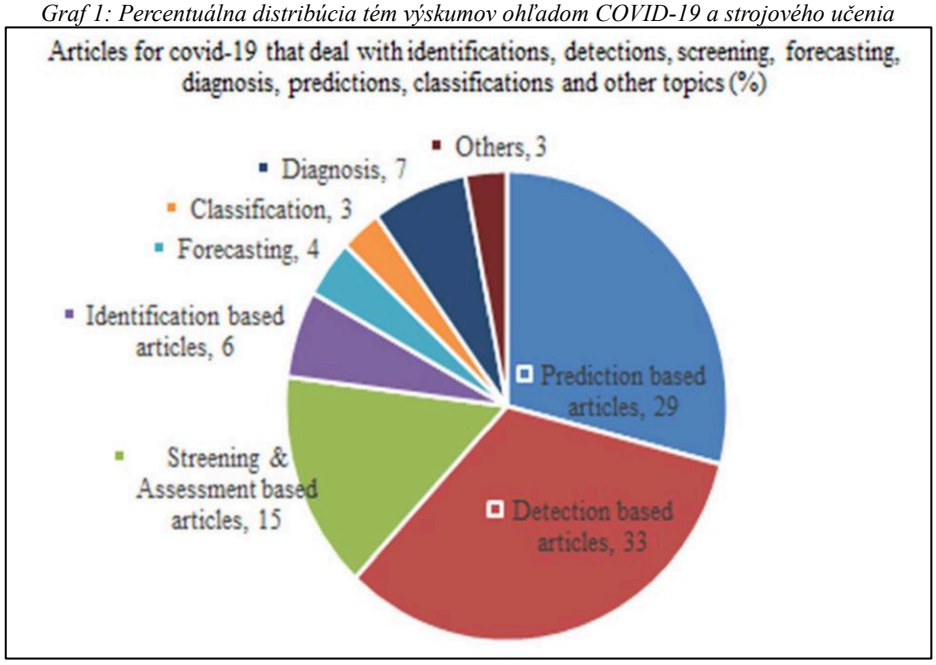
Obrázok 2: Sumarizácia najdôležitejších aplikácií ML počas pandémie COVID-19



Zdroj: (Sarmiento Varón et. al., 2023)

Doposiaľ bolo ohľadom tejto problematiky uskutočnených už viacero výskumov. Podľa Tiwariho rozboru z roku 2022, v ktorej bolo analyzovaných 81 publikácií spomínanej témy, vznikli viaceré zistenia. Bolo zistené, že medzi najčastejšie používané prístupy ML patria algoritmy logistickej regresie, naivnej Bayesovej metódy, algoritmus podporných vektorov, algoritmus najbližšieho suseda a náhodného lesa. Taktiež najväčšie percento publikácií v súvislosti s COVID-19, sa zameriava na predikciu, ako je možné vidieť na nasledujúcom obrázku.

Graf 1: Percentuálna distribúcia tém výskumov ohľadom COVID-19 a strojového učenia



Zdroj: (Tiwari et. al., 2022)

Vo všeobecnosti, COVID-19 v porovnaní s ostatnými infekčnými chorobami predstavuje veľa výziev. Jednou z nich je ľudský faktor, keďže dôsledkom tohto ochorenia je vznik pandémie. Nákaza je rapídna, dovršujúca vysokú mieru mortality. Nakoľko toto ochorenie je pomerne stále neprebádané, skutočnosťou je nedostatok historických dát, ktorý má za následok nárast limitácií modelovania. Najdôležitejším faktorom je nejasnosť správania sa modelu, kvôli kvalite použitého dátového súboru. Nedostatok konzistentnosti a neodrazenie reality sú elementy, ktoré spôsobujú nedôveryhodnosť vyhodnocovaných dát. Mnohé doteraz existujúce výskumy sa spoliehali na malé súbory dát alebo informácie, ktoré nemusia presne odrážať spoločnosť ako celok. Vďaka pandémií vznikol veľký nedostatok spoľahlivých a správne označených dát, čo spôsobilo, že experti využívali na trénovanie modelov menší súbor údajov. Dané údaje boli využité na nárast veľkosti súboru dát, pomocou techník rozšírenia údajov, čo však spôsobuje neadekvátnu generalizáciu modelu. Použitie nevybalansovanej databázy má veľkú tendenciu zvyšovať presnosť. Avšak na to, aby bol model úspešný a efektívny, je potrebné masívne množstvo tréningových údajov. (Kim et. al., 2021; Nixon et. al., 2022; Panjeta et. al., 2024)

Žiaľ, momentálna infraštruktúra neponúka kvalitu dát v tejto oblasti, ktorá by spĺňala požiadavky, keďže ide o stále čerstvú udalosť, na ktorú nikto nebol pripravený. Vznikajúce nedostatky v danom odvetví, sú hlavne kvôli vykazovaciemu systému. Existujú nekonzistentné spôsoby zbierania a zdieľania dát medzi jednotkami zdravotnej starostlivosti, či medzi krajinami. Iné prvky, ktoré môžu byť medzi spomenutými, sú

napríklad skreslený výber vzoriek či nedostatok štandardizácie dát. (Kim et. al., 2021; Nixon et. al., 2022)

V kontexte s DM, ľudské behaviorálne dáta sú taktiež predmetom záujmu pri zisťovaní prenosových vzorov. Táto disciplína je náročná sama o sebe, nakoľko ľudské správanie je nepredvídateľné a ťažko merateľné v reálnom čase. Fundamentálnym pokrokom v tejto oblasti sú sociálne siete, konkrétne zníženie rizika pomocou adresovaných prieskumov. Taktiež je potrebné spomenúť príchod nových mutácií či variantov, a ich významná úloha v počte nových prípadov. Časové a priestorové rozlíšenia sú oblasťami, ktoré nie sú dotiahnuté v dátových systémoch. (Nixon et. al., 2022) Heterogenita údajov nastáva pri použití prístupov identifikácie COVID-19, z dôvodu rozdielov v závažnosti ochorenia a komorbidít, či v rozdieloch demografických údajov pacientov. Napríklad, rozdiely v populácií pacientov, infraštruktúre zdravotnej starostlivosti alebo testovacích postupoch, spôsobujú nevhodné zovšeobecnenie ML modelu z jednej nemocnice v porovnaní s inou nemocnicou alebo krajinou. Dôsledkom týchto skutočností môže pre lekárov vzniknúť problém pri určovaní vhodného postupu liečby pre konkrétneho pacienta. (Panjeta et. al., 2024)

Panjeta (2024) vo svojej štúdii zdôrazňuje hlavne absenciu jednotnosti algoritmov, slúžiacich na detekciu COVID-19, čo predstavuje významnú medzeru vo výskumoch. Môžu byť použité rôzne spôsoby, nielen metódy strojového učenia, ale napríklad aj analýza obrazu či spracovanie prirodzeného jazyka. Toto spôsobuje náročnosť vyhodnocovania výsledkov, nakoľko neexistuje jednotná metodika.

Naopak, Tiwari (2022) zaraďuje interpretáciu ML prístupov medzi hlavné problémy, kvôli zdravotníkom, nakoľko oni sú nútení pochopiť, ktoré charakteristiky sú použiteľné na identifikáciu COVID-19. Aj napriek tomu, že predikcia COVID-19 spolu s diagnostikou a skríningom vo vzťahu s ML preukazuje sľubné výsledky, väčšina týchto modelov nebola doposiaľ otestovaná v reálnom prostredí (v zdravotníckych zariadeniach), aby sa preukázala v boji proti pandémií ako účinná. Práve preto je potrebné v tejto súvislosti prekonať nasledujúce výzvy:

- konzistentnosť bezpečnosti siete so zámerom umožnenia autentickejšej komunikácie a požadovania údajov v sieťach,
- adaptácia na fog computing (využitie okrajových zariadení pri výpočtovom výkone), edge computing (distribuovaný výpočet, ktorý prináša aplikácie

podniku bližšie k dátam), alebo cloud computing (dodanie výpočtových služieb virtuálne vďaka internetu), kvôli veľkému výpočtovému výkonu pri veľkom množstve dát,

- ochrana súkromia a bezpečnosti súvisiaca s údajmi pacienta.

Na základe Kimovej štúdie (2021), existuje niekoľko oblastí výziev, ktoré sa vyskytli s novým typom vírusu. Tie je možné zosumarizovať nasledovne:

1. *Nedostatok štandardizácie*

Dôležité merania sú nekonzistentné v mnohých regiónoch. V problematike zotavovania sa z tejto choroby, sa stretávame s nedostatočnými faktami. Falošná pozitivita, pravdepodobnosť potvrdenia prípadu - nachádza sa tu nedostatok klinickej štandardizácie aplikovanej do praxe.

2. *Absencia unifikovaného dátového vykazovacieho systému*

Súčasné pandemické dáta sa odlišujú v geografickej oblasti, časovom rozlíšení či verejnej dostupnosti. Vyskytujú sa tu signifikantné rozdiely medzi zdrojmi, ktoré poskytujú tieto dáta, čo taktiež ovplyvňuje potreby analýzy na tvorbu modelov.

3. *Problém integrácie*

Dátové súbory z rôznych sfér vedia taktiež slúžiť ako faktor v tejto problematike. Legislatívne nariadenia v oblastiach edukácie, zdravotnej pohotovosti, psychologických faktorov, vládnych ustanoveniach, toto všetko sú typy dát, ktoré sú definované ako aktíva pri identifikovaní rozšírenia choroby. Avšak, metódy získavania informácií sa líšia, preto je integrácia týchto dát signifikantným problémom.

4. *GDPR a transparentnosť*

Primárne údaje o pacientovi obsahujú senzitívne informácie. Preto vzniká ďalší problém v tejto oblasti, keďže autori publikácie nie sú schopní v mnohých prípadoch údaje agregovať a anonymizovať, kvôli čomu mnohé údaje chýbajú, aby mohli slúžiť na výskum.

5. *Neúplná dostupnosť údajov*

Dáta obsahujúce potrebné informácie súvisiace s COVID-19 sú častokrát zverejnené oneskorene, vďaka prebiehajúcim revíziám, či komplikáciám s identifikáciou prípadov.

V dôsledku všetkých vznikajúcich výziev sa vytvára nátlak na tvorcov modelov a dátových analytikov, ktorí sú zodpovední za pravdivú a korektnú interpretáciu výsledkov a zistení. (Nixon et. al., 2022)

2. Ciele práce

Ako hlavný cieľ tejto práce môžeme definovať predikciu úmrtnosti pacientov na ochorenie COVID-19, vzhľadom na ich zdravotný stav a dostupné charakteristiky. Na dosiahnutie tohto cieľu bude využitá metodika strojového učenia. Predikcia v podobe modelov bude uskutočnená pomocou vhodne vybraných algoritmov. Samotná aplikácia predikčného modelovania nastane využitím vybraného softvérového prostredia. Dôvody výberu daných prostriedkov budú na základe zvoleného zdrojového súboru dát. Interpretácia výsledkov bude definovaná metrikami hodnotenia, ktoré vychádzajú zo štatistickej metodológie.

Zámerom tohto výskumu je dosiahnuť také výsledky, ktoré by slúžili ako prvok pri potenciálnom zlepšení prípravy krokov možnej pandémie v budúcnosti. Táto práca by mohla slúžiť ako východisko pre vedcov v danej oblasti, ktorí ju môžu využiť ako podklad pri ďalších štúdiách. Taktiež by sme mohli poskytnúť tieto cenné informácie ako podporu pre rôzne sektory, akými sú medicína, zdravotníctvo, poisťovníctvo, manažment rizík či financie.

Aby sme boli schopní určiť odpoveď na otázku, či pacient zomrie alebo nezomrie na následok ochorenia COVID-19, vzhľadom na asociácie v jeho zdravotnom stave alebo iných vlastnostiach, je nutné si charakterizovať vedľajšie ciele a plán tejto práce štruktúrovať v bodoch:

1. Preštudovať teoretické podloženie témy COVID-19 ohľadom predikcie, možností a výziev, ktoré prináša.
2. Vo všeobecnosti zistiť, aké sú vhodné existujúce metódy strojového učenia, ktorými môžeme dosiahnuť náš hlavný cieľ.
3. Dôkladne zanalyzovať dostupné dátové zdroje, ktoré môžu slúžiť na výskum a zvážiť všetky alternatívy.
4. Získať vhodný dátový súbor a naštudovať si jeho skutočnosti.
5. Upraviť dáta podľa potreby do vhodnej podoby, ktorá bude slúžiť na predikciu.
6. Dáta vizuálne zanalyzovať a určiť fakty, ktoré môžu byť podložené pri samotnom modelovaní a zistení výsledkov.
7. Zvoliť preferované metódy na modelovanie a následne ich nasadiť na vhodne upravený dátový súbor.
8. Interpretovať výsledky modelovania.

9. Zhrnúť všetky dosiahnuté výsledky pri zohľadnení teoretických skutočností, stanovených cieľov a praktických zistení.
10. Identifikovať nedostatky analýzy a predložiť návrhy na zlepšenie pri ďalšom výskume.
11. Definovať prínosy tejto práce.

3. Metodika práce a metódy skúmania

Úloha strojového učenia vychádza zo štandardizácií pre KDD, DM a štatistiku. Proces ML je z praxe rozdelený na 6 častí, avšak každý ML problém je iterovaný podľa potreby. V nasledujúcich podkapitolách si jednotlivé fázy môžeme priblížiť.

3.1. Zber dát

Pri začatí práce je potrebné mať údaje ideálne v dobrej kvalite a presnosti. Dáta sú získavané zväčša z overených zdrojov (voľne dostupné sú napr. data.gov.in, kaggle.com). Táto fáza vyžaduje veľké množstvo času, kapitálu a zdrojov. (GeeksforGeeks, 2023b) Samozrejme dáta môžu byť v rôznych formátoch ako sú databázy, obrázky, textové súbory či zvukové súbory. Populárnou formou získania údajov sú informácie z webu, ktoré sú extrahované formou tzv. „web scrapingu“ (postup analýzy obsahu na internete a proces programového čítania). Pred prípravou dát je potrebné nahrať dáta do prostredia kde sa budú upravovať a to vo vhodnom formáte ako je napr. CSV, čo zabezpečí, aby boli údaje relevantné pre riešenie. (Crabtree, 2023; Lutkevich, 2023)

Najčastejším manuálnym spôsobom aplikácie procesu ML je využitie programovacieho jazyka **Python** a jeho knižníc. Čitateľnosť a jednoduchosť spolu s veľkým množstvom balíčkov ho spravili programovacím jazykom ušitým na mieru pre ML. Python je interpretovaným jazykom nie kompilovateľným, jeho kód môže byť spustený riadok po riadku. Sú viaceré prostredia, kde je možné realizovať písanie a spustenie kódu, kde medzi populárne patrí **Jupyter notebook**. Použitie tohto prostredia umožňuje viaceré funkcionality vrátane písania, spustenia kódu v kombinácii s použitím textových a grafických prvkov. (Zollanvari, 2023)

Ako bolo spomenuté, Python obsahuje zástup balíčkov s open-source knižnicami. Balíky dátovej analýzy poskytujú vedecké a matematické funkcionality slúžiace na transformáciu a čistenie dát. Medzi nimi hrajú hlavnú rolu Pandas, NumPy, Matplotlib a SciPy. Naopak, ML balíky sú slúžiace na extrakciu vzorov použitím ML algoritmov. Scikit-learn je nadstavba SciPy, ktorá je najpoužívanejšou ML knižnicou s prvkami algoritmov pre strojové učenie s učiteľom, aj bez. (Swamynathan, 2017)

3.2. Čistenie údajov

Po časti pochopenia problematiky proces plynule pokračuje prípravou dát. V tejto ide o identifikáciu chýb a ich odstránenie, opravu diskrepancií v dátach, so zámerom

zlepšenia ich kvality a použiteľnosti. Ide o kľúčový krok v celom procese, pretože dáta s veľkým šumom majú často negatívny vplyv na výsledky modelovania a jeho spoľahlivosť. Je známe, že kvalitnejšie údaje prekonávajú prepracované algoritmy. (GeeksforGeeks, 2024a)

Pod túto fázu spadajú viaceré podprocesy, medzi ktoré patrí napr. **klasifikácia údajov**. Vo všeobecnosti sú známe dva typy atribútov, ktoré reprezentujú kvalitatívne a kvantitatívne premenné. Kvalitatívne hodnoty predstavujú kategórie alebo obmeny, ktoré je možné vyjadriť slovne. Tie môžu byť nominálne (nemožno ich usporiadať podľa preferencie) alebo ordinálne (možnosť usporiadania napr. výborný – chválitebný – dobrý – dostatočný – nedostatočný). Kvantitatívne premenné sú číselné premenné, ktoré sú merateľné. Pojem *kategoriálna premenná* môže predstavovať aj kvalitatívne aj kvantitatívne hodnoty, s podmienkou konečného počtu obmien. Táto premenná, obsahujúca len dve konkrétne obmeny (napr. 0/1, teplý/studený) sa nazýva binárna. Je potrebné definovať a modifikovať typy premenných podľa potreby pri postupe ML. (Terek et. al., 2010)

Redukcia údajov je taktiež nevyhnutnou súčasťou prípravy dát. Tento krok zastrešuje:

- *Redukcia počtu atribútov* – ide o selekciu premenných potrebných pre modelovanie, aby bola zvýšená presnosť výsledkov.
- *Redukcia počtu hodnôt* – ako sme už spomínali môže ísť o kategorizáciu hodnôt, teda isté hodnoty budú spojené do nejakej kategórie, ktorá ich bude reprezentovať (napr. intervalové hodnoty).
- *Redukcia počtu prípadov* – pre presnejšie výsledky a prácu pri modelovaní je možné z celkovej množiny údajov vybrať konkrétnu podmnožinu pozorovaní (vytvorenie vzorky). Vzorka môže byť vytvorená rôznymi technikami, ktoré zahŕňajú stratifikáciu, priemerové či prírastkové vyberanie. Okrem iného existujú pozorovania, ktoré majú chybné alebo chýbajúce údaje, a tie sú v rámci procesu odstránené. (Terek et. al., 2010)

Podproces **transformácie dát** figuruje pri premene údajov z danej formy do iného formátu, za účelom vhodného použitia v analýze. Patria sem viaceré techniky:

- *Normovanie* – pri rôznych premenných je škála merania rôzna, čo môže spôsobiť skreslenie zistení a nepresnosť záverov z modelovania. Normovanie hodnôt dosiahne súlad v rôznych typoch jednotiek atribútov v rovnakom

rádovom rozsahu. Metódy, ktoré sa využívajú na túto techniku sú napr. štandardizácia, min-max škálovanie a iné. (Swamynathan, 2017)

- *Kódovanie* – ide o prevod kategoriálnych hodnôt na číselné, za účelom ich využitia v modeloch ML, ktoré pracujú len s číselnými údajmi. Teda napríklad, atribút obsahujúci hodnoty – vysoký/stredný/nízky sa preformátuje na hodnoty 0/1/2, kde je každá hodnota reprezentovaná týmto označením. Tento typ kódovania sa nazýva „label encoding" – kódovanie štítkami a využíva sa zväčša pre ordinálne údaje. Kódovanie pre nominálne údaje sa označuje ako „one-hot encoding". (GeeksforGeeks, 2023a)
- *„Data smoothing" – vyhladenie údajov* – pri kvantitatívnych premenných sa stáva, že majú veľmi veľa rozličných hodnôt, redukcia celkového počtu takýchto hodnôt a považovanie blízkych hodnôt atribútu za rovnaké sa nazýva vyhladenie. (Terek et. al., 2010)
- *Tvorba nových premenných* – v procese je niekedy potrebné na základe jednej premennej odvodiť nový atribút. Experti analyzujú a sami manuálne odvodia novú premennú podľa potreby. Existuje však aj metóda hlavných komponentov, ktorá podľa korelačnej štruktúry premenných vytvára lineárne kombinácie (komponenty). (Terek et. al., 2010)

Čistenie údajov zahŕňa aj dva rozsiahle a významné podprocesy a to: **spracovanie chýbajúcich údajov a odľahlých hodnôt**. Kvôli obšírnosti týchto tém sú ich špecifiká priblížené v nasledujúcich podkapitolách.

3.2.1. Chýbajúce údaje

Pri príprave dát môže nastať problém chýbajúcich údajov v súbore dát. Niektoré techniky ML sú odolné voči chýbajúcim údajom, iné však potrebujú všetky dáta. (Terek et. al., 2010) Sú viaceré prístupy ako riešiť takúto situáciu:

- *Odstránenie pozorovaní s chýbajúcimi dátami* - je možné takéto riadky jednoducho vymazať. Tento spôsob je prijateľný a účinný, ak majú pozorovania s chýbajúcimi hodnotami zanedbateľný počet s porovnaním celkového počtu údajov. (Swamynathan, 2017)
- *Manuálne doplnenie hodnôt* – v tomto prípade sa takéto pozorovania fyzicky skontrolujú a podľa úvahy analytika sa doplnia. V praxi je takýto spôsob

využiteľný len pri malom počte takýchto prípadov s chýbajúcimi hodnotami. (Terek et. al., 2010)

- *Automatická náhrada chýbajúcich hodnôt* – táto technika je pre rozsiahly súbor údajov najbežnejšou. Prázdna množina sa môže nahradiť konkrétne zadanou konštantou, hodnotou aritmetického priemeru, modusom či mediánom. (Swamynathan, 2017; Terek et. al., 2010)
- *Použitie predikčného modelu* – táto technika je pokročilejšia, keďže ide o natrénovanie modelu na existujúcich hodnotách a použitie modelu na predikciu chýbajúcich hodnôt. (Swamynathan, 2017)

3.2.2. Odľahlé hodnoty

Odchýlky sú mimoriadne malé alebo veľké hodnoty a ich detekcii je pripisovaná istá dôležitosť. (Terek et. al., 2010) Podľa kontextu je na rozhodnutí experta, či chce odstrániť takéto odľahlé hodnoty alebo zmeniť ich podobu, ktorá by minimalizovala vplyv na analýzu. (GeeksforGeeks, 2024a) Kvôli spoľahlivosti modelu je potrebné ich zisťovanie, pretože často spôsobujú poškodenie modelovacieho systému. Na ich odhalenie sa zväčša využívajú rôzne štatistické metódy. (Zhao et. al., 2008) Terek (2010) opisuje vo svojej literatúre 2 metódy a to:

- *Métóda založená na mediánovej absolútnej odchýlke (MADN),*
- *Metóda založená na kvartilovom rozpätí (IQR).*

V praxi je dlhodobo veľkou vizuálnou pomôckou na zobrazenie medzikvartilového rozpätia – **box plot**. Jedná sa o grafickú reprezentáciu distribúcie dát. Zobrazuje kvartily, potenciálne odľahlé pozorovania aj medián danej premennej. Stredová čiara v „boxe“ reprezentuje medián, kde samotný „box“ predstavuje **IQR**. Čiary vytŕčajúce z daného boxu („fúzy“) dosahujú najextrémnejšie „neodľahlé“ hodnoty, ktoré tvoria 1,5 násobok *IQR*. Hodnoty nachádzajúce sa za hranicou „fúzov“ sú označované ako odľahlé pozorovania. (GeeksforGeeks, 2024a)

Z-skóre v súvislosti s odľahlými dátami predstavuje štatistický pojem, ktorý pomáha porozumieť, či je údaj menší alebo väčší ako priemer a taktiež jeho vzdialenosť od priemeru, teda špecifikuje počet štandardných odchýlok bodu vzdialeného od priemeru.

$$\text{Z} - \text{skóre} = \frac{\text{bod} - \text{priemer}}{\text{štandardná odchýlka}}$$

Prakticky zaužívané, ak daná hodnota *Z-skóre* je vyššia ako 3, tak je zrejmé, že daný bod je odľahlým údajom. (GeeksforGeeks, 2024b)

3.3. Exploratívna dátová analýza (EDA)

Exploratívna dátová analýza (EDA) predstavuje skúmanie pozorovaní s cieľom porozumenia vzťahov a vzorov. Bežne sa vykonáva pred štatistickými analýzami alebo samotným modelovaním. (GeeksforGeeks, 2023c) EDA je analýza, ktorá slúži na porozumenie dát pomocou rôznych sumarizačných a vizualizačných techník. (Swamynathan, 2017) Tento spôsob je považovaný za veľmi efektívny pri manipulácií s dátami, za účelom získania nevyhnutných odpovedí. Tak pomáha dátovým expertom objavovať vzory, anomálie alebo overovať hypotézy či predpoklady. V prvom rade sa používa na lepšie pochopenie atribútov súboru dát a vzťahov medzi nimi ale taktiež dokáže odhaliť zistenia nad rámec zadania. Schopnosť definovania vhodných štatistických techník potrebných pre prácu s dátami, je taktiež predmetom záujmu. (IBM, 2024)

Je tiež využiteľná pri čistení údajov, pretože kontroluje chyby a nezrovnalosti. Môže zahŕňať metódy imputácie pozorovaní, chýbajúcich či odľahlých hodnôt. Okrem iného, hodnotí kvalitu údajov. Vie posúdiť spoľahlivosť kontrolou integrity a presnosti záznamov. (GeeksforGeeks, 2023c)

Podľa Swamynathana (2017) existujú dve roviny EDA: jednorozmerná a viacrozmerná. IBM (2024) považuje za najjednoduchšiu formu jednorozmernú analýzu, ktorá nevyužíva grafické prvky, pretože sa aplikuje výhradne len pri dátových súboroch obsahujúcich jednu premennú. Nezaoberá sa vzťahmi ani príčinami, pretože jej zámerom je opísať dáta a nachádzať zákonitosti. Na druhej strane, jednorozmerná analýza s grafickými prvkami poskytuje celkový obraz o dátach. Medzi bežné typy vizualizačných metód patria histogramy a box ploty pre priblíženie distribúcie údajov. Multivariačné (viacrozmerné) techniky vznikajú pri viac ako jednej premennej. Negrafická analýza sa zaoberá zobrazením vzťahov medzi atribútmi využitím rôznych štatistík, naopak vizualizačná analýza využíva pre zobrazenie vzťahov dát najmä stĺpcové diagramy.

GeeksforGeeks (2023c) predstavuje ešte ďalšie typy EDA. Analýza časových radov sa uplatňuje na dáta obsahujúce časovú zložku. Modeluje prvky usporiadané v časovom horizonte. Využíva sa napríklad aj autokorelačná analýza či ARIMA. Do EDA by bolo možné taktiež zaradiť vyššie spomínané analýzy odchýlok či chýbajúcich údajov.

3.4. Rozdelenie dát

Keď analytici usúdia, že údaje sú pripravené na modelovanie, nastáva fáza rozdelenia dát. Induktívne učenie je založené na odhadoch neznámych asociácií v systéme, ktoré sa odvádzajú z množiny prípadov. (Terek et. al., 2010) Celková množina údajov je rozdelená na trénovaciu a testovaciu časť, ktorá slúži na vyhodnotenie. Pri meraní efektivity modelu nastáva potreba overenia natrénovaného modelu na trénovacích dátach. Na to slúži hodnotiaci postup naučeného modelu na testovacích údajov. (Zollanvari, 2023)

Zväčša sa v tejto časti procesu okrem rozdelenia pozorovaní udeje aj rozdelenie premenných. Cieľová premenná je atribút, ktorý sa model snaží predikovať. Je označovaná aj ako závislá premenná, premenná „*y*“, „**target**“ alebo „label“. Ide o najdôležitejší atribút, ktorý určuje na akom type modelovania sa bude pracovať, no nie vždy je jednoduché definovať, ktorá premenná bude slúžiť ako cieľová. (Bruehl, 2023)

3.5. Modelovanie

Táto fáza obsahuje výber modelovacích techník a zostavovanie modelu. (Hotz, 2023) V procese ML je táto fáza vykonávaná adekvátnymi algoritmami. Nakoľko cieľ tejto práce napovedá o klasifikácií pacienta, predstavíme si metodológiu klasifikačných ML algoritmov.

SML algoritmy pri klasifikácií sa využívajú na riešenie problémov vzhľadom na efektívnejšiu prácu s údajmi. Používajú umelé vedomie, ktoré bežne prijíma a zlepšuje štruktúry. Podnecujú schopnosti pomenovať dáta zahŕňajúce modely získavania informácií. (Butt et. al., 2020) Rôzne typy klasifikátorov sú v procese ML využívané, kde následne najefektívnejší algoritmus je určený s využitím dátového súboru, jeho atribútov a počtu pozorovaní. (Osisanwo et. al., 2017) V praktickej časti práce sa využíva 6 druhov algoritmov, ktoré sú opísané v nasledujúcich podkapitolách.

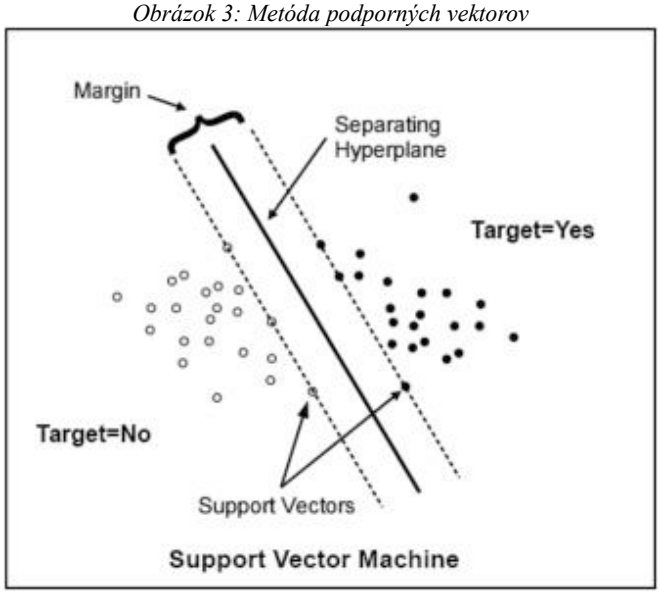
3.5.1. Algoritmus podporných vektorov (SVM)

Klasifikátor podporných vektorov [Support Vector Machine = SVM] je algoritmus predstavený Vapnikom v roku 1995. Podstatou je využívať presnosť na generalizáciu chýb. Je vytvorená „hyperplocha“, ktorá rozdeľuje údaje na 2 kategórie. (Abdar et. al., 2015) Hyperplochu definuje hranica a podporné vektory. (Rady & Anwar, 2019) Táto hranica je určená ako veľkosť priestoru medzi týmito dvoma kategóriami, na ktoré je rozdelená. Geometricky reprezentuje najkratšiu vzdialenosť medzi dvoma dátovými bodmi, ktoré sú

najbližšie k hyperrovine. Najlepšie riešenie SVM je to, ktoré maximalizuje hranicu aj keď je takýchto hyperrovín veľa. (Wu et. al., 2008)

Ak je riešenie nelineárne, teda ak je nemožné manuálne určiť danú hyperplochu, algoritmus využíva kernel. Kernel [jadro] je funkcia, ktorá transformuje nízko-dimenzionálny vstupný priestor na vysoko-dimenzionálny priestor. Inými slovami, vytvorí postup zložitej úlohy na oddelenie údajov podľa definovaných označení. (Sen et. al., 2020)

Obrázok 3: Metóda podporných vektorov



Zdroj: (Wu et. al., 2008)

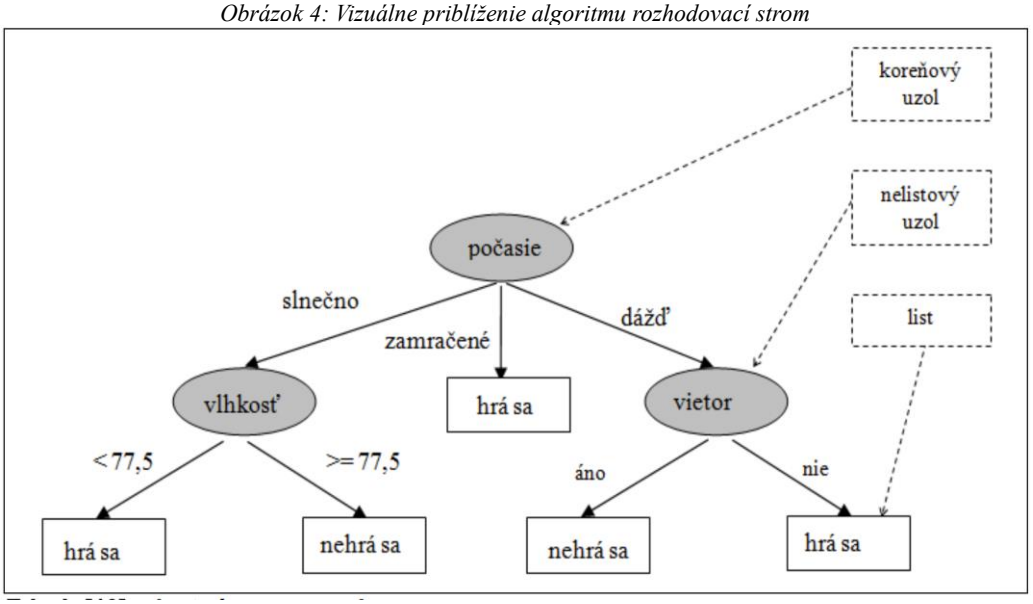
Trénovanie modelu môže byť časovo náročné, presnosť tohto algoritmu je príkladná v oblasti modelovania náročných nelineárnych rozhodovacích hraníc. Je oveľa menej náchylný na pretrénovanie. (Rady & Anwar, 2019) *Pretrénovanie* je pojem definovaný ako nežiadúce správanie ML, ktoré nastáva, keď model určuje presné predikcie pre trénovaciu množinu, ale nie pre novú množinu. Tým pádom poskytuje nepresné predpovede a je nespoľahlivý. (Amazon Web Services, 2024) SVM algoritmus je možné využiť na klasifikáciu v rôznych oblastiach, napríklad rozpoznávanie číslic, objektov či identifikácie hovoriacich. (Rady & Anwar, 2019)

3.5.2. Rozhodovací strom

Algoritmus rozhodovacieho stromu je uvedený ako typický príklad algoritmu založeného na logike. Využíva sa v regresii aj klasifikácií. Ide o generovanie sekvencií na základe pravidiel, ktoré keď sú dodržiavané, tak dosahujú zaradenie neoznačených údajov. (Sen et. al., 2020) Tieto sekvencie sa nazývajú uzly rozhodovacieho stromu, kde jeden uzol reprezentuje vlastnosť cieľovej premennej, ktorá má byť klasifikovaná. Vetvy uzlov

prestavujú hodnoty, ktoré môže uzol nadobudnúť. Na základe týchto hodnôt premenných sa tieto uzly trieda, čo sa označuje ako rozhodovacie pravidlo. V ML sa využíva tento algoritmus ako prediktívny model mapujúci pozorovania so zámerom klasifikácie cieľovej premennej. (Osisanwo et. al., 2017)

Obrázok 4: Vizuálne priblíženie algoritmu rozhodovací strom



Zdroj: (Labudová, 2017)

Ako pri predchádzajúcom algoritme, funkcionalita rozhodovacieho stromu sa zameriava na spojité aj kategorické premenné. (Sen et. al., 2020) Ak je „target" kategoriálny, vtedy hovoríme o klasifikačnom strome. Ak je naopak spojitou premennou, rozhodovací strom je označovaný ako regresný. Strom začína generovať vetvenie v tzv. koreňovom uzle. (Terek et. al., 2010) Rozhodovacie pravidlá vytvárajú súbor, ktorý popisuje konkrétny profil pre každý uzol. Typickým formátom pravidla je „if-then" [ak-potom], kde podmienku predstavuje vlastnosť kategórie a následok reprezentuje názov kategórie, alebo iné pravidlo, ktoré sa ďalej bude vetviť. (Khan et. al., 2010) Uzol, ktorý nie je ďalej vetvený, je označovaný ako list. (Terek et. al., 2010) Premenná, ktorá sa vetví na konkrétnej úrovni môže byť vyberaná rôznymi postupmi. Pri klasifikačnom strome sa využíva entropia, informačný zisk, Giniho index alebo chi-kvadrát test. (Labudová, 2017)

Entropia je metrikou miery rozptylu dát (náhodnosti údajov) alebo miery „nečistoty". (Jairi, 2021) Vzťah tejto metriky je vyjadrený ako:

$$Entropia = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Nečistota je v tomto vzťahu označená ako log_2 pravdepodobnosti kategórie (p_i). Index i vyjadruje počet prípustných kategórií. (Seth, 2023)

Alternatívou výpočtu môže byť Giniho index. Tento index zovšeobecňuje nečistotu rozptylu, no môže byť považovaný aj za očakávanú chybovosť. Má o niečo silnejší vrchol pri totožných pravdepodobnostiach dvoch tried. Výhodou je jeho spojitosť s varianciami. (Brown & Myles, 2009)

Vzťah Giniho indexu je reprezentovaný:

$$\text{Giniho index}_{\,\mathrm{D}} \;=\; 1 \;-\sum_{\mathrm{i\ do\ m}} \mathrm{p_i^2}$$

V tomto vzťahu $\mathrm{p_i}$ hovorí o pravdepodobnosti, že súbor údajov v D patrí do triedy $\mathrm{C_i}$, ktorá sa odhaduje pomocou $|\mathrm{C_i}, D| \,/\, |D|$. Daný súčet je vyjadrený pre m tried. (Gangadhar & Rangaswamy, 2018)

Veľkou výhodou klasifikačného stromu je zrozumiteľnosť pravidiel a preto je uprednostňovaný v mnohých aplikáciách. Medzi jeho silné stránky sa radí práca s chýbajúcimi hodnotami, či rýchlosť natrénovania modelu. (Abdar et. al., 2015) Najväčším rizikom implementovania klasifikačného stromu je prispôsobenie sa trénovacím dátam, kde vzniká veľký počet alternatív stromu. Tento algoritmus je vytvorený tak, aby kategorizácia trénovacích údajov bola efektívna, čo však spôsobuje pokles výkonnosti klasifikácie pri testovacích dátach. Taktiež pri veľkom súbore dát môže vzniknúť zložitá štruktúra stromu. (Khan et. al., 2010)

3.5.3. Náhodný les

Na báze rozhodovacieho stromu bol vytvorený nový algoritmus, náhodný les alebo „random forest“ (RF). Funkcionalita klasifikátora je založená na väčšom počte rozhodovacích stromov, ktorých frekvencia výstupov určí klasifikáciu cieľovej premennej. (Lahiri et. al., 2020) Do úvahy sú brané všetky pravdepodobnosti klasifikačných stromov, z ktorých je odvodená celková pravdepodobnosť. (Shaw et. al., 2020) RF je vhodným prostriedkom pri nadmernom prispôsobovaní sa svojej trénovacej množine. (Lahiri et. al., 2020) V zásade sa zväčša používa pre súbory dát s vysokou dimenziou, kde sú jednotlivé atribúty veľmi zašumené a nie stacionárne náhodné. (Shaw et. al., 2020)

Pri tomto algoritme je náhodne vyberaná podmnožina pozorovaní a atribútov, z ktorých je vytvorených viacero nezávislých stromových modelov. Pri vetvení stromu je používaná len podmnožina atribútov a práve preto sú stromy viac nekorelované. (Swamynathan, 2017) Existuje viacero techník, ktoré slúžia na rast stromov. Náhodný výber

selekcie, ktorý uzly rozdeľuje náhodne spomedzi najlepších rozdelení je jedným z prístupov. (Breiman, 2001) Ďalšou metódou je tzv. „bagging“, ktorý je realizovaný iteračne. Ide totiž o náhodný výber s opakovaním, kde je obsahom trénovacia množina. Rozsah výberu je zväčša totožný s rozsahom trénovacej množiny. V každom kroku je využitá jedna ML metóda (v tomto prípade klasifikačný strom). (Terek et. al., 2010)

Podľa Breimana (2001) je RF definovaný ako klasifikačný algoritmus zložený zo setu klasifikačných stromov, kde sú náhodné vektory nezávisle a identicky rozdelené, kde je každý strom jednotkovým hlasom pre najpopulárnejšiu triedu vstupu.

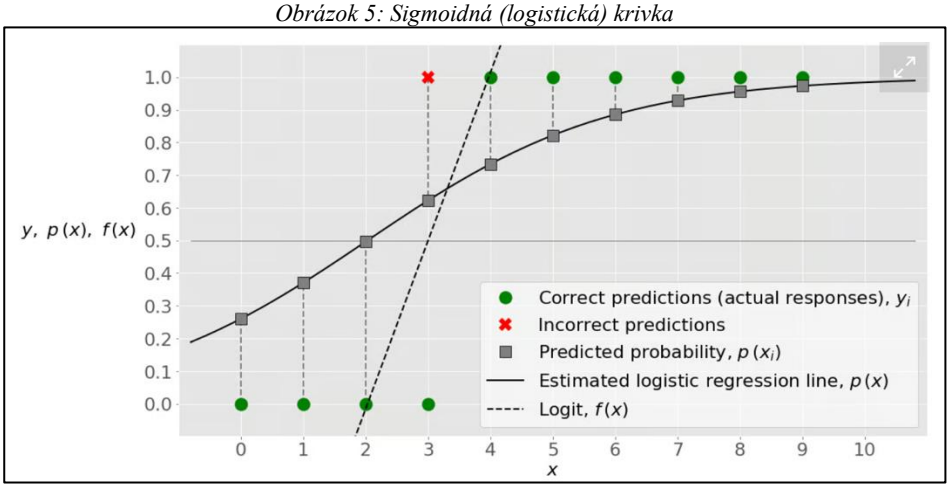
3.5.4. Logistická regresia

Klasifikátor logistickej regresie vytvára model, ktorý využíva binárnu závislú premennú. Výstupná premenná je reprezentovaná kódovanými hodnotami, povedzme, že 1 pre úspech a 0 pre neúspech. (Chaturvedi et. al., 2020) Pravdepodobnosť výstupu je vypočítaná na základe súboru atribútov. Je predpokladané, že *x* reprezentuje danú vlastnosť a *y* je výstupnou premennou. (Lahiri et. al., 2020)

$$\log\left(\frac{\mathrm{p(X)}}{1-\mathrm{p(X)}}\right) = \beta_0 + \beta_1 \mathrm{X} + \beta_\mathrm{i} \mathrm{X_n}$$

Ľavú stranu rovnice definuje logistická funkcia (nazývaná aj ako sigmoidná, pretože vizuálne tvorí krivku, ktorá pripomína písmeno S), kde výraz $(p(X)/(1-p(X)))$ je označovaný ako šanca. Tá hovorí o pomere pravdepodobnosti úspechu, vzhľadom k pravdepodobnosti neúspechu. (Lahiri et. al., 2020; Jurafsky & Martin, 2023) Na pravej strane rovnice sa nachádzajú β_i regresné koeficienty, ktoré sú spojené s referenčnou skupinou a vysvetľujúcimi atribútmi X. β_0 predstavuje referenčnú skupinu, ktorá tvorí referenčnú úroveň každej premennej. (Sperandei, 2014)

Obrázok 5: Sigmoidná (logistická) krivka



Zdroj: (Stojiljković, 2020)

Vo všeobecnosti je funkcionalita logistickej regresie veľmi podobná lineárnej, kde je však závislá premenná binomická. Pri tomto druhu algoritmu je jednoduchšie pracovať s viacerými vysvetľujúcimi premennými súčasne, a to je jeho veľkou výhodou. Totiž v procese je zohľadňovaná kovariancia medzi premennými a nie sú pozorované nezávisle. Pri použití logistickej regresie môžu vzniknúť aj problémy, a to hlavne s výberom premenných. Niekedy sa pre výskumníkov môže zdať viacero premenných ako štatisticky významných, avšak môže sa stať, že model s veľkým počtom premenných klesá na štatistickej sile. (Sperandei, 2014)

3.5.5. K-najbližší sused (KNN)

KNN je klasifikačný algoritmus, ktorého podstata je založená na analógii jedného prípadu s druhým. Blízke pozorovanie s druhým sa nazýva „sused“. (Abdar et. al., 2015) Tento prístup je okamžite učenlivý, pretože určuje kategóriu na základe najbližšieho priestoru pozorovaní v trénovacej sade. Bod, ktorý je potrebné zaradiť, je v priestore priradený určitej kategórii, ktorá sa vyskytuje najčastejšie v k-najbližších trénovacích dátach. Bežne sa využíva na výpočet vzdialenosti tzv. „euklidovská vzdialenosť“. (Khan et. al., 2010)

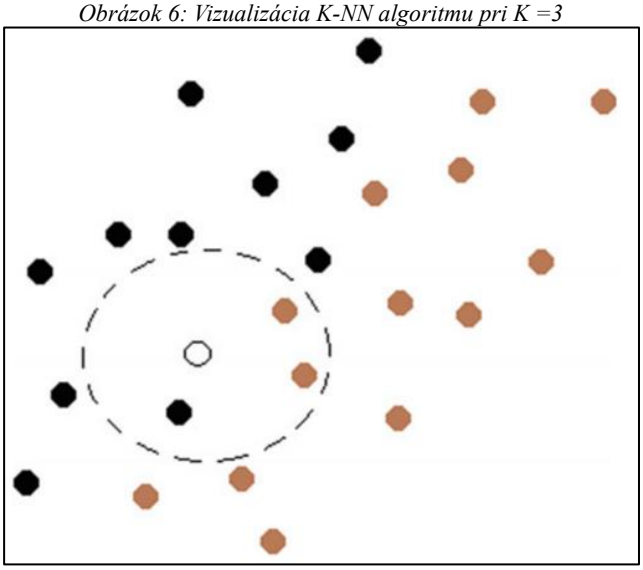
Euklidovská vzdialenosť je považovaná za najpoužívanejšiu mieru vzdialenosti, ktorá je daná vzťahom:

$$Euklidovská\ vzdialenosť\ = \sqrt{i\ -\ k^2\ +\ j\ -\ l^2}$$

V danom vzťahu je matematicky vyjadrená vzdialenosť dvoch bodov, ktoré majú priestorové indexy $[i, j]$ a $[k, l]$. (Merchant et. al., 2023)

K v KNN predstavuje počet najbližších susedov, ktorí sa budú brať do úvahy pre zaradenie pozorovania. Ak by napríklad bolo k = 3, tak bod, ktorý treba zaradiť bude vizualizovaný tak, že tri susedné dátové body v najbližšom kruhu rozhodnú o označení tohto bodu. Platí, že ak K má vyššiu hodnotu, hranica sa stáva hladšia a môže sa stať, že všetko sa stane jednou triedou. (Sen et. al., 2020)

Obrázok 6: Vizualizácia K-NN algoritmu pri K =3



Zdroj: (Sen et. al., 2020)

Tento algoritmus má veľmi ľahkú implementáciu a je účinný pri veľkom počte trénovacích dát, avšak má veľmi dlhý čas validácie. Používa totiž všetky pozorovania pri výpočte vzdialenosti. Presnosť zaradenia sa znižuje, ak sú dáta nachádza šum. Je nutné správne určiť hodnotu K, aby chybovosť algoritmu bola znížená. (Khan et. al., 2010; Sen et. al., 2020) Okrem iného tento klasifikátor vie priebežne určiť hodnoty cieľovej premennej, podľa priemerných alebo mediánových hodnôt najbližšieho suseda. (Abdar et. al., 2015)

3.5.6. Naive Bayes

Ďalším veľmi dôležitým klasifikačným prístupom je naivná Bayesova metóda. Je to jednoduchý algoritmus, ktorý nevyžaduje žiadne schémy na odhad parametrov a je ľahko použiteľný na obrovský súbor dát. (Wu et. al., 2008) Medzi výhody tohto algoritmu sa radí dostatočnosť malého množstva trénovacích dát na odhadnutie potrebných klasifikačných parametrov. (Khan et. al., 2010) Jeho výnimočnosťou je tiež expresná rýchlosť a efektivita. (Sen et. al., 2020) Na druhej strane sa nezaraďuje medzi najpresnejšie vyhodnocovacie algoritmy. (Wu et. al., 2008)

Jeho schopnosť spočíva vo vytváraní pravdepodobnosti pre každé pozorovanie, kde však vychádza z faktu, že atribúty sú navzájom nezávislé. (Sen et. al., 2020) Je založený na odhade, zložený z acyklických grafov jedného neodpozorovaného uzla, na základe niekoľkých pozorovaných uzlov. (Osisanwo et. al., 2017)

3.6. Hodnotenie modelov

V ML je pripisovaný znamenitý význam odhadovaniu výkonnosti modelu, kde táto časť je označovaná ako hodnotenie modelu. Miera jeho užitočnosti závisí od predikčnej výkonnosti. Od rôznych techník ohodnotenia sa následne odvíja **proces výberu modelu**. (Zollanvari, 2023)

Hodnotiace metriky sú významnou rolou v dosiahnutí optimálnych výsledkov klasifikačných algoritmov. Preto je dôležité zvoliť vhodné techniky hodnotenia, ktoré slúžia pri danom druhu modelovania. (Hossin & Sulaiman, 2015)

Pri binárnych klasifikačných úlohách je optimálnym riešením **konfúzna matica** (alebo matica zámen). (Hossin & Sulaiman, 2015) Ide o tabuľku, ktorá opisuje výkon modelu. (Swamynathan, 2017) Na hlavnej diagonále matice sa nachádzajú prípady, ktoré model správne určil ako target. Ostatné údaje predstavujú počet nesprávne klasifikovaných prípadov. (Terek et. al., 2010) Riadky tejto tabuľky reprezentujú predikované hodnoty, stĺpce zas opisujú počty skutočnej triedy. Označenia TP (True Positive) a TN (True Negative) hovoria o počte pozitívnych a negatívnych prípadov, ktoré sú správne určené. Naopak, FP (False Positive) a FN (False Negative) sú nesprávne určenými pozitívnymi a negatívnymi pozorovaniami. (Hossin & Sulaiman, 2015)

Tabuľka 1: Konfúzna matica pre binárny target

Určenie modelom	**Reálna hodnota**		**Spolu**
	+	-	
+	TP	FP	TP + FP
-	FN	TN	FN + TN
Spolu	TP + FN	FP + TN	TP + FP + FN + TN

Zdroj: Vytvorené autorkou podľa (Terek et. al., 2010, 66 s.)

Na základe konfúznej matice sa dajú určiť viaceré štatistiky. **Presnosť modelu [accuracy]** sa meria pomerom korektných predikcií v súvislosti s celkovým počtom vyhodnotených prípadov. (Hossin & Sulaiman, 2015)

$$\text{Presnosť} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Od presnosti modelu sa odvíja určenie zhodnosti alebo **precíznosti modelu [precision].** Reprezentuje počet percent správne určených pozitívnych predikcií z celkového počtu pozitívnych predikcií. (Swamynathan, 2017)

$$\text{Precíznosť} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Ďalšou metrikou, ktorá sa zaoberá správnymi určeniami je **úplnosť [recall].** Táto metrika vyjadruje pravdepodobnosť toho, že pozorovanie je správne klasifikované v porovnaní s celkovým počtom reálnych hodnôt. V niektorých literatúrach je táto metrika označovaná aj ako senzitívnosť. (Zollanvari, 2023)

$$\text{Úplnosť} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Na základe váženého priemeru precíznosti a úplnosti vie byť určené tzv. **F1 – skóre**. (Swamynathan, 2017)

$$\text{F1} - \text{skóre} = \frac{2 * (\text{precíznosť} * \text{úplnosť})}{(\text{precíznosť} + \text{úplnosť})}$$

4. Výsledky práce

Táto časť práce sa zameriava na praktické priblíženie procesu splnenia cieľov, ktoré sme si stanovili. Proces nadobudnutia stanovených cieľov je objasnený v jednotlivých podkapitolách.

4.1. Získanie dát

Pre cieľ tejto práce je potrebné mať konkrétne údaje o pacientoch infikovaných COVID-19, ktoré obsahujú špecifické informácie o ich zvykoch a anamnéze. Vzhľadom na súčasné bezpečnostné opatrenia a najmä nariadenie GDPR neexistuje v tejto oblasti veľa možností. Preto po dlhodobom hľadaní vhodného súboru údajov, bol v našej analýze použitý súbor údajov, ktorý zverejnilo Ministerstvo zdravotníctva v Mexiku na oficiálnej stránke mexickej vlády.

Tento konkrétny dátový set obsahuje množstvo anonymizovaných informácií týkajúcich sa pacientov. Preto je potrebné spomenúť, že analýza je platná najmä pre Mexiko, prípadne Severnú Ameriku. Dátum zverejnenia databázy je 14. apríla 2020 za účelom dostupnosti informácií pre používateľov, ktorí ich potrebujú. (Secretaría de Salud, 2024)

Obsahuje informácie o každom jednotlivcovi testovanom na tento typ vírusu v Mexiku. Zahŕňa rôzne demografické informácie týkajúce sa národnosti, miesta pobytu, používania jazyka alebo veku. Okrem toho sú tu uvedené faktory, ktoré sú pre analýzu nevyhnutné, predchádzajúce lekárske záznamy o mnohých ochoreniach vrátane astmy, cukrovky, obezity, hypertenzie atď. Okrem iného ponúka presnú časovú os, kedy sa pacient dostal na ošetrovacie oddelenie, dátum úmrtia (ak nastalo), dátum prvých príznakov. Napokon sú tu zahrnuté polia o hospitalizácii, predchádzajúcom zápale pľúc, potrebe intubácie, liečbe na jednotke intenzívnej starostlivosti a výsledkoch testov COVID-19.

Pôvodný dátový súbor obsahuje 3 868 396 pozorovaní a má 40 premenných. Počas prípravných fáz pred modelovaním bol upravovaný do adekvátnych podôb. Detaily ohľadom premenných použitých na modelovanie sú priblížené v ďalších kapitolách.

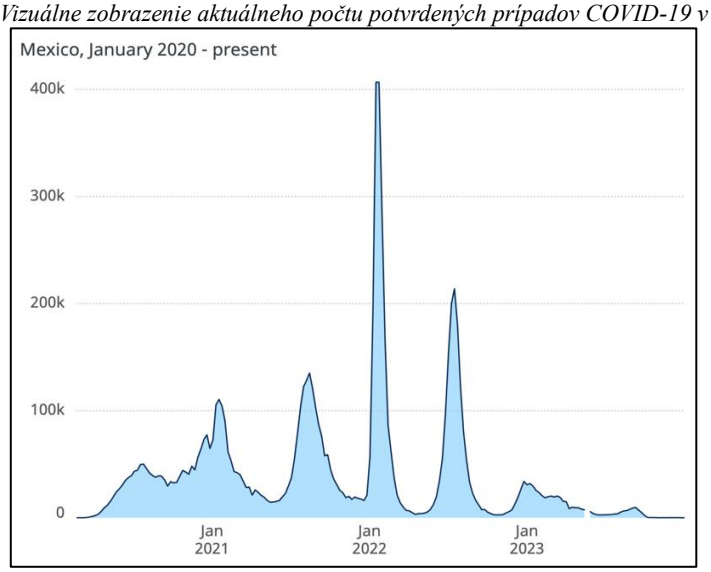
Nakoľko dáta použité v tejto práci sú z lokality, ktorá je pre nás neznáma ohľadom problematiky COVID-19, rozhodli sme sa preskúmať skutočnosti tejto témy pre adekvátny teoretický základ pri určení výsledkov nášho výskumu.

4.1.1. Poznatky o situácií COVID-19 v Mexiku

Kolekcia údajov použitá v tejto práci je z určitého časového obdobia počas celej pandémie. Tento výsek údajov prispieva k celkovému počtu potvrdených prípadov v Mexiku, ktoré sú uvedené nižšie.

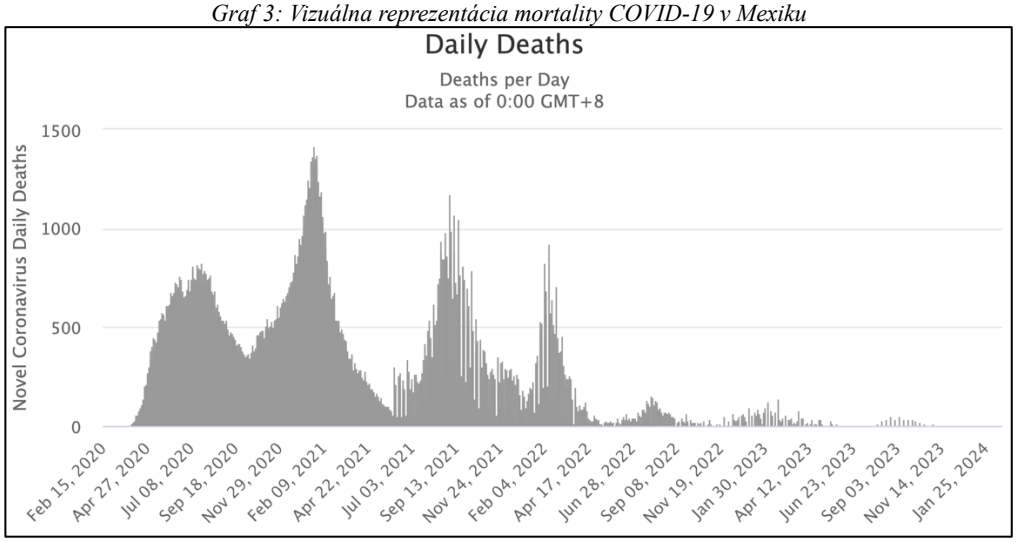
Podľa Svetovej zdravotníckej organizácie - WHO (2024) bolo v Mexiku od začiatku pandémie, teda od januára 2020, podľa poslednej obnovy dát zo 4.2.2024, vyše 7 702 809 potvrdených prípadov COVID-19 spolu s 334 958 úmrtiami. Doposiaľ bolo zaznamenaných 222 miliónov podaných vakcín.

Graf 2: Vizuálne zobrazenie aktuálneho počtu potvrdených prípadov COVID-19 v Mexiku



Zdroj:(WHO, 21.2.2024)

Graf 3: Vizuálna reprezentácia mortality COVID-19 v Mexiku



Zdroj: (Worldometer, 21.2.2024)

Mexiko je federatívnym štátom zloženým z 32 čiastočne samosprávnych štátov. Prijaté opatrenia každej štátnej vlády hrajú signifikantnú rolu v kontrole vypuknutia nákazy

COVID-19 v celej krajine. Avšak, spôsob akým sú nariadenia implementované, tak ako ich vážnosť a rýchlosť, sa medzi štátmi líši. (Knaul et. al., 2021)

Latinská Amerika sa stala globálnym epicentrom SARS-CoV2 vírusovej infekcie a taktiež mortality COVID-19 v lete 2020. Mexiko sa snažilo o nápravu situácie nasadením nových obmedzení keď sa rozmohol variant Omikron. Vláda prijala niekoľko opatrení proti šíreniu tohto ochorenia, ktoré zahŕňajú:

- *Lockdowny:* Mexická vláda zaviedla lockdown v oblastiach krajiny, kde počet prípadov bol alarmujúci. Tento pojem zahŕňa zatvorenie nie až tak fundamentálnych podnikov aj obmedzenie pohybu obyvateľov v zasiahnutých oblastiach.
- *Sociálny dištanc:* Vláda implementovala nariadenia dodržiavania dostatočnej vzdialenosti kvôli zníženiu nákazy pri obsiahlych zhromaždeniach či vo verejných priestoroch.
- *Povinné rúška:* Nariadenie nosenia povinných ochranných masiek vo verejných priestoroch, ktoré zahŕňajú aj obchodné strediská či hromadnú dopravu, taktiež prišlo do platnosti.
- *Dopravné obmedzenia:* Zatvorenie hraníc pre nie podstatné presuny a vyžadovanie negativity na COVID-19 pri vstupe do krajiny sa stalo nežiadúcim.
- *Zákaz vychádzania:* Určité oblasti nastavili zákazy vychádzania kvôli obmedzeniu pohybu v istých časoch počas dňa. (Knaul et. al., 2021)

Mexická vláda predstavila tzv. semaforový systém, kde podľa kategórie farby je všeobecne známe pre daný štát, ktoré opatrenia sú sprísnené. Kategórie boli nastavené podľa 4 farieb (červená, oranžová, žltá, zelená), ktoré boli určené podľa počtu aktuálnych pozitívnych prípadov, obsadených lôžok a iných. Ohľadom reakcie jednotlivých mexických štátov na pandémiu nie je doposiaľ dostatok informácií. Na mieste je poukázať na veľký rozdiel v oblasti stupňov závažnosti. Pokiaľ čo v jednom štáte mohli byť otvorené kluby a bary, naopak v inom boli tieto aktivity zakázané. Niektorí experti kritizovali mexickú vládu v jej reakcii na počiatočne fázy pandémie. Preto mohlo dôjsť k určitým oneskoreniam, čo sa týka podstúpenia komplexných testov, ktoré mohli prispieť k veľkému počtu prípadov. (Knaul et. al., 2021)

Podľa štúdie García-Lópeza z roku 2022 je známe, že do daného roku boli prezentované 4 epidemiologické vlny. Najväčšie pandemické vrcholy infikovaných pacientov boli uvedené:

- 18.6.2020 (57,73 prípadov na milión obyvateľov)
- 10.1.2021 (126,74 prípadov na milión obyvateľov)
- 10.8.2021 (151,66 prípadov na milión obyvateľov)
- 15.1.2022 (462,01 prípadov na milión obyvateľov).

Podľa priemerných hodnôt najvyššieho a najnižšieho výskytu nárazov bola určená prvá vlna v období od 15.5.2020 do 3.10.2020, kde maximálny počet úmrtí bol denne v priemere 0,63 prípadov na milión rezidentov. Táto vlna bola zasiahnutá Alfa variantom a jeho mutáciami. Trend výskytu nových prípadov pokračoval neustále, kde hneď po konci prvej vlny začala druhá vlna, ktorá trvala od 3.10.2020 do 19.4.2021, počas ktorej začala aj vakcinácia COVID-19. V tomto období sa rozšíril variant B.1.1.519 a taktiež najvyšší stupeň mortality bol identifikovaný a to denne 1,13 prípadu na milión. Tretia vlna prišla 13.6.2021 do 29.10.21, kde Beta variant nahradil Delta, ktorý bol najrýchlejší v súvislosti so šírením. V tomto období bol zaznamenaný doposiaľ najväčší rekord v počte prípadov, ale za to úmrtnosť poklesla. Štvrtá vlna bola dominovaná Omikronom, ktorý začal 23.12.2021, kde bol najväčší počet infikovaných za deň ale úmrtnosť klesla ešte viac.

4.2. Príprava dát

Pred vytvorením modelov sme podnikli kroky potrebné pre prípravu dát na postupné spracovanie do modelov. Dané kroky sú priblížené v nasledujúcich bodoch.

1. *Redukcia premenných*

Ako sme už spomenuli v predošlej podkapitole, pôvodný dátový súbor použitý v tejto práci obsahuje 40 rôznych premenných. Význam, ktorý daná premenná reprezentuje je predstavený v zdroji dátového súboru. Využitie všetkých dostupných atribútov v modelovaní môže byť výpočtovo náročné a neefektívne, nakoľko nie všetky premenné sú významovo potrebné pri vytváraní modelov v tejto práci. Existuje viacero dôvodov nepotrebnosti atribútov, ako: nevyužitie atribútov s odľahlými hodnotami (napr. premenná tehotenstvo pre mužské pohlavie), atribúty neovplyvňujúce predikovanú premennú (napr. dátum prvých symptómov, dátum prijatia do nemocnice) alebo iné nekonzistentné a nepresné premenné. Po dôkladnej analýze sme sa rozhodli vymazať nasledujúce údaje:

- Dátum poslednej aktualizácie databázy = irelevantnosť pri modelovaní,
- ID registrácie = unikátna premenná zbytočná pre náš dátový súbor,
- Špecifiká zdravotníckych zariadení v Mexiku = detaily zdravotníckeho systému danej lokácie, ktoré sa nevzťahujú na analýzu, na ktorú je táto práca zameraná (vplyv prostredia podľa kvality a dostupných prostriedkov na predikciu by mohol byť predmetom skúmania v rozšírení tejto práce v budúcnosti, avšak na túto problematiku je potrebný pevný informačný podklad, ktorý nespadá do rozsahu aktuálnej práce),
- Identifikačné údaje pacienta (miesto narodenia, miesto pobytu, pôvod) = zameranie sa na komunity či pôvod nákazy by mohli byť podrobnejšie preštudované avšak taktiež nespadajú pod predmet záujmu práce,
- Dátum prvých symptómov = rozhodli sme sa tejto časovej analýze nevenovať a zamerať sa na iné dátumové premenné, ktoré majú vplyv na predikciu,
- Tehotenstvo,
- Potvrdenie pôvodu odobratej vzorky = ohľadom tejto informácie existujú v dátovom súbore iné atribúty, ktoré sú využívané.

Vďaka výmazu spomínaných premenných sme dátový súbor zredukovali na 22 atribútov, ktoré boli naďalej využívané.

2. *Premenovanie premenných*

Nakoľko zdrojový dátový súbor je mexického pôvodu, názvy atribútov sú v španielskom jazyku. Pre efektívnu manipuláciu s dátami sme sa rozhodli pomocou príkazu „replace“, nahradiť názvy atribútov anglickými označeniami podľa významu.

3. *Vytvorenie nových premenných*

Nové atribúty boli vytvorené kvôli určeniu predikcie, ktoré sú potrebné v ďalšom procese.

- *COVID_POS* – vytvorili sme nový atribút pomocou funkcie, ktorú sme si definovali. Z existujúcej premennej COVID_CASE sme odvodili tento atribút, ktorý bude hovoriť o tom, či je alebo nie je pacient pozitívny podľa podmienky funkcie.

- *DEATH* – vytvorená nová premenná, ktorá reprezentuje úmrtie na COVID-19. Taktiež je odvodená z vytvorenej funkcie, na základe premennej DATE_DEATH, kde dátum vo formáte „9999-99-99“ všeobecne reprezentuje, že dátum smrti nenastal. Táto kategoriálna premenná hovorí o tom, či úmrtie nastalo alebo nenastalo.
- *LAB_POS, AG_POS* – vytvorené na základe premenných RESULT_LAB a RESULT_AG, z ktorých sú nové kategoriálne premenné definované či nastala pozitivita na laboratórnom alebo antigénovom teste[1].
- *HOSP* – vytvorenie novej premennej, ktorá hovorí či pacient bol hospitalizovaný. Na jej vytvorenie je použitá funkcia, kde sa využíva atribút P_TYPE, ktorý reprezentujú kategórie: 1 – ambulantný pacient, 2 – hospitalizovaný pacient alebo 99 – nešpecifikované. Ak sa teda tento atribút rovná 2, vznikne nová binárna kategoriálna premenná HOSP.

Nami vytvorené atribúty nám nahradili potrebu atribútov, z ktorých boli odvodené, preto sme sa rozhodli pôvodné atribúty vymazať. V danom momente bol stav premenných v počte, ktorý bol vhodný na ďalšie použitie. V nasledujúcej tabuľke sú opísané jednotlivé premenné, ktoré boli využívané pri predikcii v tejto práci.

Tabuľka 2: Opis premenných použitých pri modelovaní

No.	Názov premennej	Opis	Formát / Kategórie
0	SEX	Pohlavie pacienta	1 – ženské, 2 – mužské
1	DATE_ENTRY	Dátum prijatia pacienta na ošetrovaciu jednotku	dátum v tvare „RRRR-MM-DD“
2	INTUBED	Identifikácia stavu pacienta či potreboval intubáciu[2]	1 – áno, 2 – nie, 97 – neaplikovateľné
3	PNEUMONIA	Identifikácia či pacientovi bola diagnostikovaná pneumónia[3]	1 – áno, 2 – nie, 99 – nešpecifikované

[1] Test rýchlej diagnostiky detekciou antigénu proteínu vírusu, kde určenie výsledku trvá 15-30 minút na mieste odobratia vzorky. (District of Columbia Government, 2024)
[2] Proces zavádzania rúrky do priedušnice vzhľadom na dýchaciu nedostatočnosť s nutnosťou napojenia na umelú pľúcnu ventiláciu. (Wikiskripta, 2018)
[3] Zápal pľúc, infekcia s možnosťou zasiahnutia oboch pľúc naplnením tekutinou, ktorej príznakovosť spočíva v kašli, horúčke a prípadných problémoch s dýchaním. (National Heart, Lung, and Blood Institute, 2022)

4	AGE	Vek pacienta	numericky v rokoch
5	DIABETES	Identifikácia či pacientovi bol diagnostikovaný diabetes[4]	1 – áno, 2 – nie, 98 – ignorované
6	COPD	Identifikácia či pacientovi bola diagnostikovaná chronická obštrukčná choroba pľúc[5]	1 – áno, 2 – nie, 98 – ignorované
7	ASTHMA	Identifikácia či pacientovi bola diagnostikovaná astma[6]	1 – áno, 2 – nie, 98 – ignorované
8	IM_SUPPR	Identifikácia či pacientovi bola diagnostikovaná imunosupresia[7]	1 – áno, 2 – nie, 98 – ignorované
9	HYPERTENS	Identifikácia či pacientovi bola diagnostikovaná hypertenzia[8]	1 – áno, 2 – nie, 98 – ignorované
10	OTHER_DIS	Identifikácia či pacientovi bolo diagnostikované iné ochorenie	1 – áno, 2 – nie, 98 – ignorované
11	CARDIO	Identifikácia či pacientovi bolo diagnostikované kardiovaskulárne ochorenie[9]	1 – áno, 2 – nie, 98 – ignorované
12	OBESITY	Identifikácia či pacientovi bola diagnostikovaná obezita[10]	1 – áno, 2 – nie, 98 – ignorované
13	RENAL_CHRON	Identifikácia či pacientovi bolo diagnostikované chronické zlyhanie obličiek	1 – áno, 2 – nie, 98 – ignorované
14	SMOKING	Identifikácia či pacient bol fajčiar	1 – áno, 2 – nie, 98 – ignorované
15	OTHER_CON	Identifikácia či pacient bol v kontakte s iným prípadom COVID-19	1 – áno, 2 – nie, 99 – nešpecifikované
16	ICU	Identifikácia či pacient bol prijatý na Jednotku Intenzívnej Starostlivosti (JIS)	1 – áno, 2 – nie, 97 – neaplikovateľné, 99 – nešpecifikované

[4] Cukrovka, chronické ochorenie kvôli nedostatku produkcie inzulínu v pankrease. (WHO, 2023b)
[5] COPD, pojem pre skupinu ochorení pľúc, ktoré spôsobujú ťažkosti s dýchaním. (National Health Service - NHS, 2023)
[6] Chronické ochorenie pľúc sťažujúce dýchanie. (WHO, 2023c)
[7] Stav imunitného systému, kedy nastáva jeho dysfunkcia a vedie telo k zvýšenej náchylnosti ochorenia. (Avian Immunology, 2014)
[8] Vysoký krvný tlak, nastáva keď tlak v cievach je príliš vysoký od štandardu. Ide o bežné ochorenie, ktoré je veľmi závažné pri neliečení. (WHO, 2023a)
[9] Všeobecné pomenovanie ochorenia postihujúce cievy alebo srdce. (NHS, 2022)
[10] Komplexná diagnóza spojená s nadmerným množstvom telesného tuku. (Mayo Clinic Staff, 2023)

17	COVID_POS	Identifikácia pacienta na pozitivitu COVID-19	1 – áno, 2 – nie
18	DEATH	Identifikácia či pacient zomrel	1 – áno, 2 – nie
19	LAB_POS	Identifikácia či pacient bol pozitívny na laboratórnom teste	1 – áno, 2 – nie
20	AG_POS	Identifikácia či pacient bol pozitívny na antigénovom teste	1 – áno, 2 – nie
21	HOSP	Identifikácia či pacient bol hospitalizovaný	1 – áno, 2 – nie

Zdroj: Vytvorené autorkou podľa (Secretaría de Salud, 2024)

4. *Zmena dátových typov*

Atribúty, ktoré sme nami vytvorili sú dátového typu *object*, teda majú hodnoty určené ako textové. Nakoľko obsahujú číselné hodnoty bolo na mieste unifikovať tieto atribúty, spolu s doteraz existujúcimi, na dátový typ *int*. V jazyku Python existuje aj dátový typ *category*, ktorý reprezentuje kategoriálne hodnoty. V tomto momente bol náš dátový súbor ešte používaný na ďalšie analýzy a na manipuláciu s ním je praktickejšie využiť hodnoty *int,* preto dané kategoriálne premenné boli typovo zmenené až pred samotným modelovaním.

5. *Chýbajúce dáta*

Predmetom záujmu je overenie, či existujú v dátovom súbore chýbajúce údaje. Využitím jazyku Python sme vypísali sumu chýbajúcich hodnôt v rámci každej premennej.

Obrázok 7: Chýbajúce hodnoty premenných

```
SEX             0
DATE_ENTRY      0
INTUBED         0
PNEUMONIA       0
AGE             0
DIABETES        0
COPD            0
ASTHMA          0
IM_SUPPR        0
HYPERTENS       0
OTHER_DIS       0
CARDIO          0
OBESITY         0
RENAL_CHRON     0
SMOKING         0
OTHER_CON       0
ICU             0
COVID_POS       0
DEATH           0
LAB_POS         0
AG_POS          0
HOSP            0
dtype: int64
```

Zdroj: Vytvorené autorkou

Ako je viditeľné z obrázku 7, v dátovom súbore sa fyzicky nenachádzali žiadne prázdne miesta, ktoré by potrebovali riešenie ďalšieho postupu.

Pri overení počtu obmien jednotlivých pozorovaní môžeme konštatovať, že jednotlivé kategoriálne premenné mali v celku 5 typov kategórií (1, 2, 97, 98, 99), ako sme si predstavili už aj v tabuľke 2. Kategóriou 98 a 99 boli reprezentované nešpecifikované alebo ignorované hodnoty – teda tieto hodnoty nemajú žiadnu výpovednú hodnotu, svojim spôsobom sú to kategórie reprezentujúce chýbajúce informácie.

Obrázok 8: Zobrazenie obmien a počtu pozorovaní pri každej obmene

	SEX	INTUBED	PNEUMONIA	DIABETES	COPD	ASTHMA	IM_SUPPR	HYPERTENS	OTHER_DIS	CARDIO
1	2001715.0	66806.0	374992.0	418275.0	40690.0	100040.0	37093.0	563200.0	75288.0	60608.0
2	1866681.0	450122.0	3477717.0	3437751.0	3816193.0	3757014.0	3819530.0	3293613.0	3775501.0	3796317.0
97	NaN	3340218.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
98	NaN	NaN	NaN	12370.0	11513.0	11342.0	11773.0	11583.0	17607.0	11471.0
99	NaN	11250.0	15687.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Zdroj: Vytvorené autorkou

Z obrázku 8 môžeme vidieť, že v celku pri každej premennej, táto kategória reprezentuje približne 0,5 % pozorovaní zo všetkých pozorovaní danej premennej, preto sme sa rozhodli z dátového súboru pozorovania obsahujúce tieto hodnoty odstrániť.

Obrázok 9: Zobrazenie obmien a počtu pozorovaní pri každej obmene po odstránení hodnôt

	SEX	INTUBED	PNEUMONIA	DIABETES	COPD	ASTHMA	IM_SUPPR	HYPERTENS	OTHER_DIS	CARDIO
1	1867439.0	47930	302859.0	367676.0	33482.0	91535.0	31238.0	498696.0	61438.0	52385.0
2	1716685.0	360985	3281265.0	3216448.0	3550642.0	3492589.0	3552886.0	3085428.0	3522686.0	3531739.0
97	NaN	3175209	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Zdroj: Vytvorené autorkou

Po odstránení hodnôt 98 a 99, ako je viditeľné na obrázku 9, bol dátový súbor v poriadku ohľadom chýbajúcich dát, nakoľko kategória 97 (neaplikovateľnosť) je špecifickou pre JIS a intubáciu. Tieto dve premenné sú špecifikované len v prípade, ak pacient bol hospitalizovaný. Za týchto okolností nie je možné aplikovať hodnotu 1 alebo 2.

Obrázok 10: Zobrazenie obmien premenných ICU a INTUBED

	INTUBED	ICU
97	3175209	3175209
2	360985	373723
1	47930	35192

Zdroj: Vytvorené autorkou

Teda kategória 97 reprezentovala všetkých nehospitalizovaných pacientov, a ak by tieto pozorovania boli vymazané, prišli by sme o všetky pozorovania nehospitalizovaných pacientov a mali by sme len vzorku hospitalizovaných pozorovaní. Preto sme sa rozhodli túto kategóriu zanechať.

Po podstúpení všetkých krokov súvisiacich s úpravou dátového súboru bol počet riadkov zmenený na 3 584 124 a počet atribútov zostal na 22. V tomto tvare bol dátový súbor pripravený na ďalšiu analýzu.

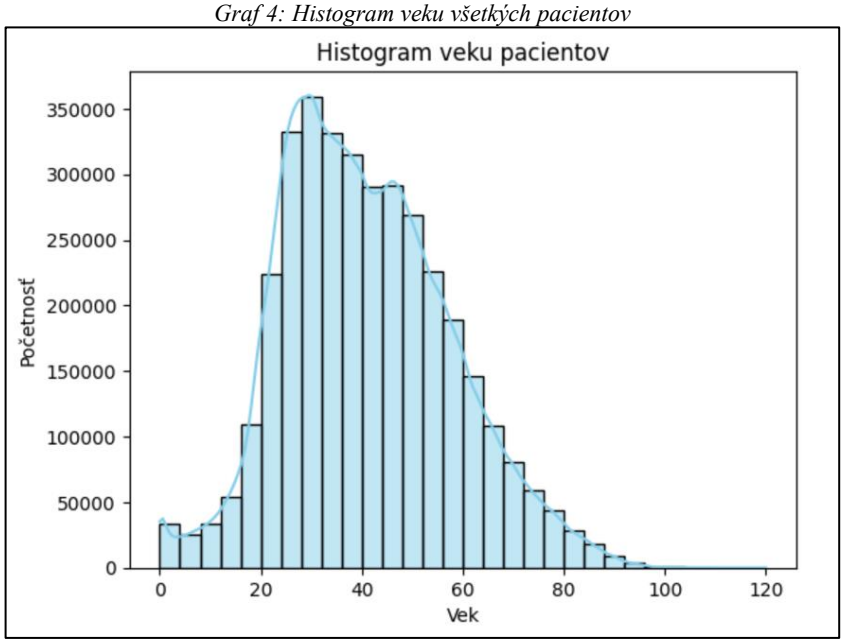
4.3. EDA

Pre identifikáciu vzorov v našom dátovom súbore sme sa zaoberali aj vizualizáciou v časti EDA. Definované skutočnosti sú naďalej prospešné pre konkrétne nasadenie modelu. V nasledujúcich podkapitolách si rozoberieme jednotlivé oblasti záujmu, ktoré boli vizualizované.

4.3.1. Aká je distribúcia veku pacientov?

Predmetom pozornosti pred samotným modelovaním bolo presnejšie porozumenie dát, preto sme si chceli priblížiť vekovú škálu pacientov všetkých pozorovaní, z ktorých bola uskutočňovaná predikcia. Na to bol použitý vizualizačný prostriedok histogramu pre určenie početnosti.

Graf 4: Histogram veku všetkých pacientov



Zdroj: Vytvorené autorkou

Pri distribúcii veku môžeme vidieť, že najväčšia početnosť pozorovaní je v intervale od 20 do 40 rokov, z celkového počtu záznamov, teda pri pacientoch, ktorí podstúpili testovanie. Najmenšia distribúcia je v intervale od 80+. Pri histograme je možné vidieť zošikmenie a špicatosť. Na grafe 4 je očividné, že nastáva pravostranné kladné zošikmenie. Z toho vyplýva, že modus je menší ako medián a priemer, teda môžeme tvrdiť, že prevažujú nižšie hodnoty z celkovej škály. Danú metriku sme si potvrdili aj matematicky pomocou kódu.

Obrázok 11: Miera šikmosti a špicatosti veku

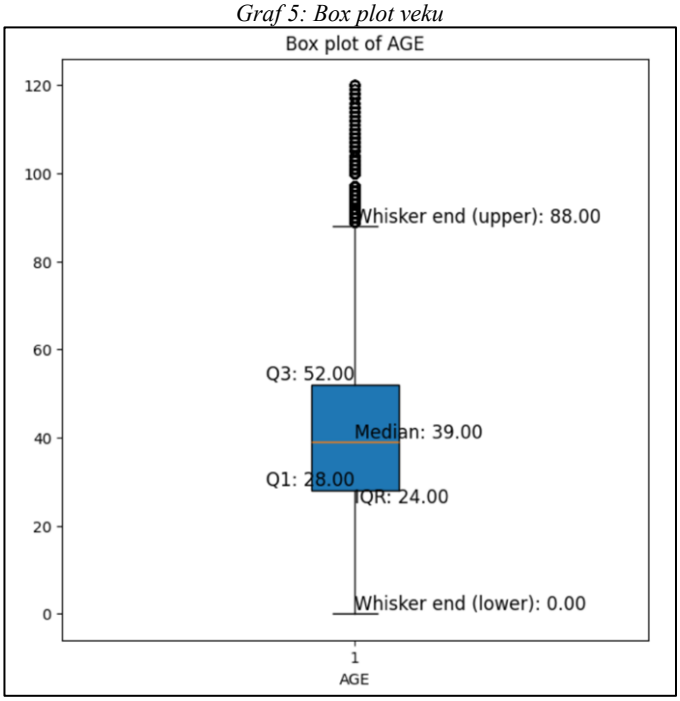
```
Skewness: 0.3545487862039228 ( Positive )
Kurtosis: -0.07187142388474133 ( Platykurtic )
```

Zdroj: Vytvorené autorkou

Pri definovaní špicatosti to na základe vizualizácie nie je až tak jasné, preto sme si pomohli výpočtom, ktorý definuje rozdelenie veku ako negatívnu (platykurtickú) špicatosť, nakoľko jej hodnota je menšia ako 3. Rozdelenie dát je v skutočnosti plochšie ako normálne rozdelenie, teda nemá sklon k extrémnym výkyvom hodnôt.

V rámci skúmania rozdelenia hodnôt premennej veku si môžeme špecifikovať, či existujú odľahlé hodnoty. Na to slúži vizualizácia pomocou box plotu.

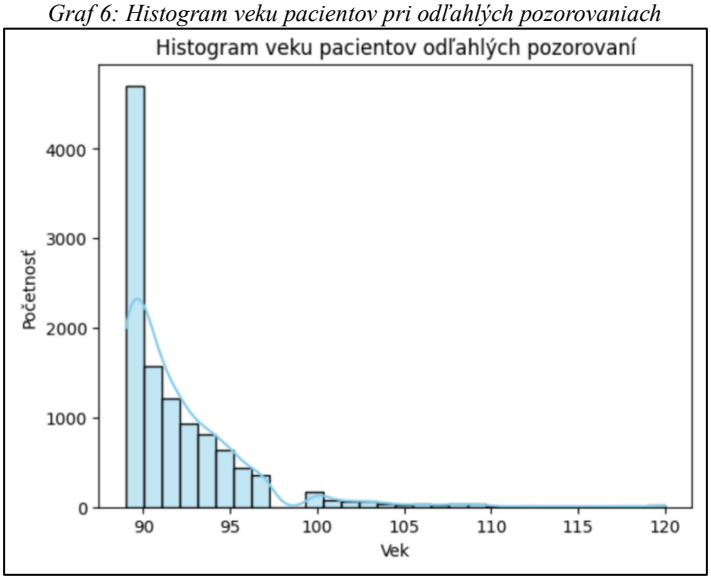
Graf 5: Box plot veku



Zdroj: Vytvorené autorkou

Ako je možné vidieť na grafe 5, tak mediánová hodnota je okolo 39. veku života. Medzikvartilové rozpätie je rovné 24 a nachádza sa v rozmedzí od 28 do 52. *Hodnoty približne od 88. roku sa nachádzajú za hranicou fúzov a sú označené ako odľahlé hodnoty*. Avšak je potrebné poznamenať, že atribút veku je dôležitým aspektom pri našej predikcii. Mohlo by sa stať, že keby odstránime všetky pozorovania, kde je vek vyšší ako 88, tak by sme prišli o vzácne dáta pre predikciu úmrtia (nakoľko predpokladáme, že úmrtie častejšie nastalo pri vyššom veku). Pre lepšie porozumenie, sme si odľahlé hodnoty veku vizuálne zobrazili.

Graf 6: Histogram veku pacientov pri odľahlých pozorovaniach



Zdroj: Vytvorené autorkou

Na základe *IQR* bolo detegovaných 11 322 odľahlých hodnôt. Ako môžeme vidieť, najväčšia početnosť odľahlých hodnôt je vo vekovom intervale do 95. roku. Avšak sú tu isté pozorovania až do veku 120. Pozorovania na vekovom intervale medzi 110. a 120. rokom sa nám zdajú príliš vysoké, preto sme sa rozhodli túto skutočnosť preveriť.

Podľa dostupného zdroja sme zistili, že od začiatku roku 2020 až doposiaľ, zomrelo v Mexiku 34 ľudí, ktorí sú na zozname supercenteriánov (ľudia, ktorí sa dožili viac ako 110 rokov). Zatiaľ najstaršia overená osoba, ktorá kedy žila v Mexiku, zomrela vo veku 113 rokov a 225 dní. (Gerontology Wiki, 2024)

Po vyfiltrovaní odľahlých pozorovaní len na tie, ktoré majú vek vyšší ako 110, sme zistili, že ide o 92 riadkov. Po overení skutočností z dostupnej literatúry, vieme s určitosťou potvrdiť, že na vekovom intervale medzi 110. a 120. rokom, ide o viaceré chybné pozorovania. Nakoľko by všetky pozorovania po dosiahnutom veku 100, mohli byť zavádzajúce, rozhodli sme sa ich vymazať. Dátový súbor, ktorý bude naďalej využívaný bez týchto hodnôt, obsahuje 3 583 633 pozorovaní (491 pozorovaní bolo vymazaných).

Rozhodli sme sa však overiť si odľahlé hodnoty veku aj pomocou štandardnej odchýlky využitím *Z-skóre*.

Obrázok 12: Výpočet Z-skóre pri atribúte vek

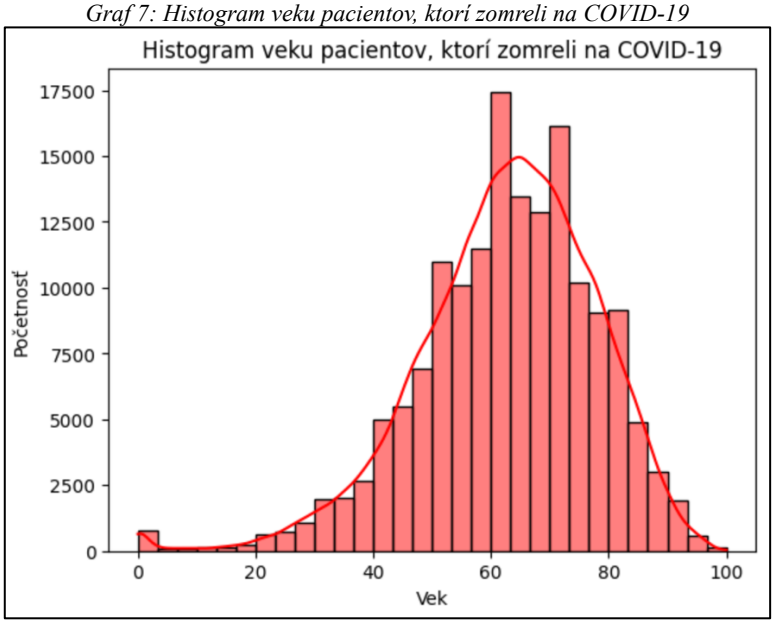
```
Mean age: 40.630236996060795
Standard deviation of age: 16.45144556428273
```

Zdroj: Vytvorené autorkou

Z výpočtov sme určili, že priemerným vekom je pri zaokrúhlení približne 41 rokov a štandardná odchýlka je približne 16. Podľa *Z-skóre* nebolo identifikované žiadne odľahlé pozorovanie, ktoré by potrebovalo výmaz.

Nakoľko predmetom modelovania je detekcia úmrtia pacientov na COVID-19, zaujíma nás aká je distribúcia pacientov, ktorí na toto ochorenie zomreli, preto sme si vytvorili histogram pre takýchto pacientov.

Graf 7: Histogram veku pacientov, ktorí zomreli na COVID-19



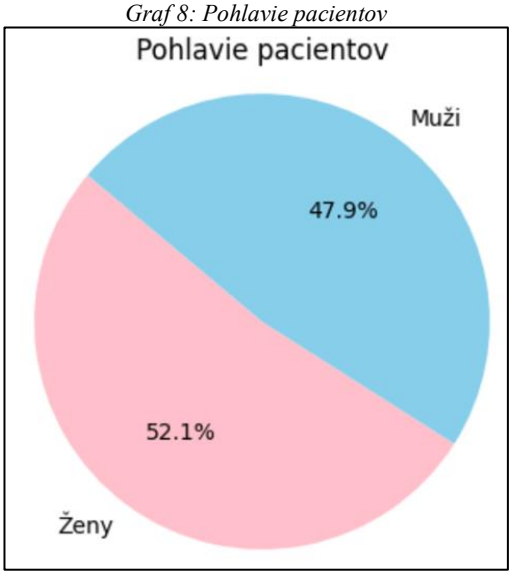
Zdroj: Vytvorené autorkou

Je zrejmé z grafu 7, že distribúcia je ľavostranne negatívne zošikmená, teda je tu istý vzor objavovania sa vyšších hodnôt zo škály. Najviac pozorovaní je v intervale od 60 - 80. Po výpočtoch sme zistili, že modus je v 65. roku života, medián je rovný 64 a *IQR* je v intervale (54;73). Z tohto je zrejmé, že *tendencia úmrtia na toto ochorenie je v dôchodcovskom veku,* čím potvrdzujeme svoj predošlý predpoklad.

4.3.2. Aká je distribúcia pohlavia pacientov?

Pri vizualizácií pohlavia je efektívne použiť koláčový graf s percentuálnym zastúpením.

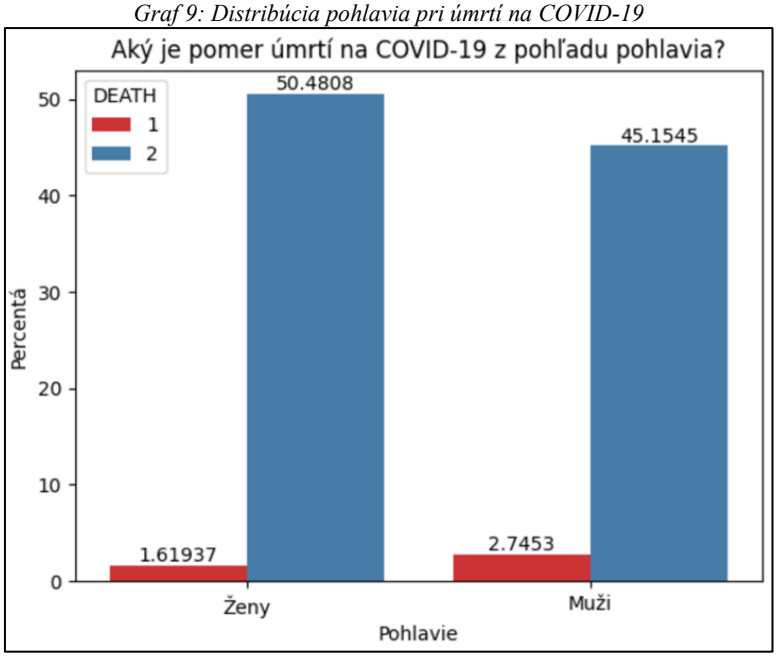
Graf 8: Pohlavie pacientov



Zdroj: Vytvorené autorkou

Z grafu 8 je zrejmé, že v dátovom súbore sa nachádza väčšie zastúpenie žien ako mužov. Aká by bola však distribúcia pohlaví v ohľade na úmrtie COVID-19 sme si priblížili v nasledovnom grafe.

Graf 9: Distribúcia pohlavia pri úmrtí na COVID-19



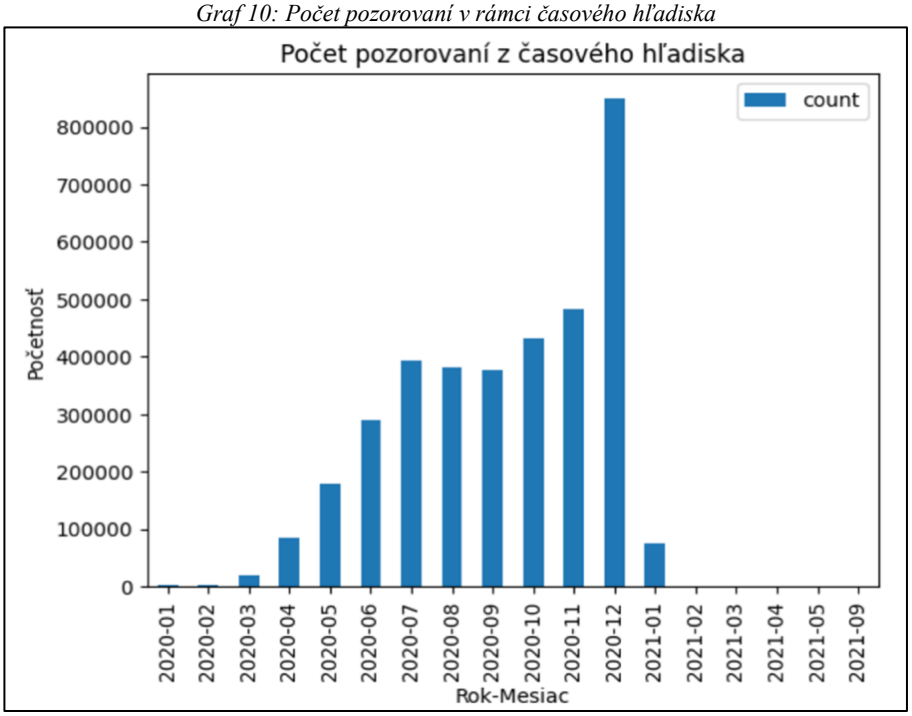
Zdroj: Vytvorené autorkou

V danom stĺpcovom diagrame je viditeľné, že medzi prežitím a úmrtím na COVID-19 existuje v danom dátovom súbore istý nepomer. Z celkového počtu pozorovaní tvorí približne *96 % tých, ktorí prežili na COVID-19 a zvyšné 4 %, ktorí umreli*. Pri distribúcií pohlaví je faktom, že *zomrelo väčšie percento mužov na toto ochorenie ako žien*.

4.3.3. V akom časovom ohraničení sa pohybujú dané pozorovania?

Dôležitým faktorom pri určení výsledkov je časové hľadisko, počas ktorého boli dáta zaznamenávané. Je nutné zohľadniť skutočnosti, ktoré prebiehali počas daného obdobia akými sú napríklad výskyt mutácií, vládne opatrenia, fáza rozšírenia a mnohé iné. Na časové hľadisko bol využitý typ vizualizácie „countplot“.

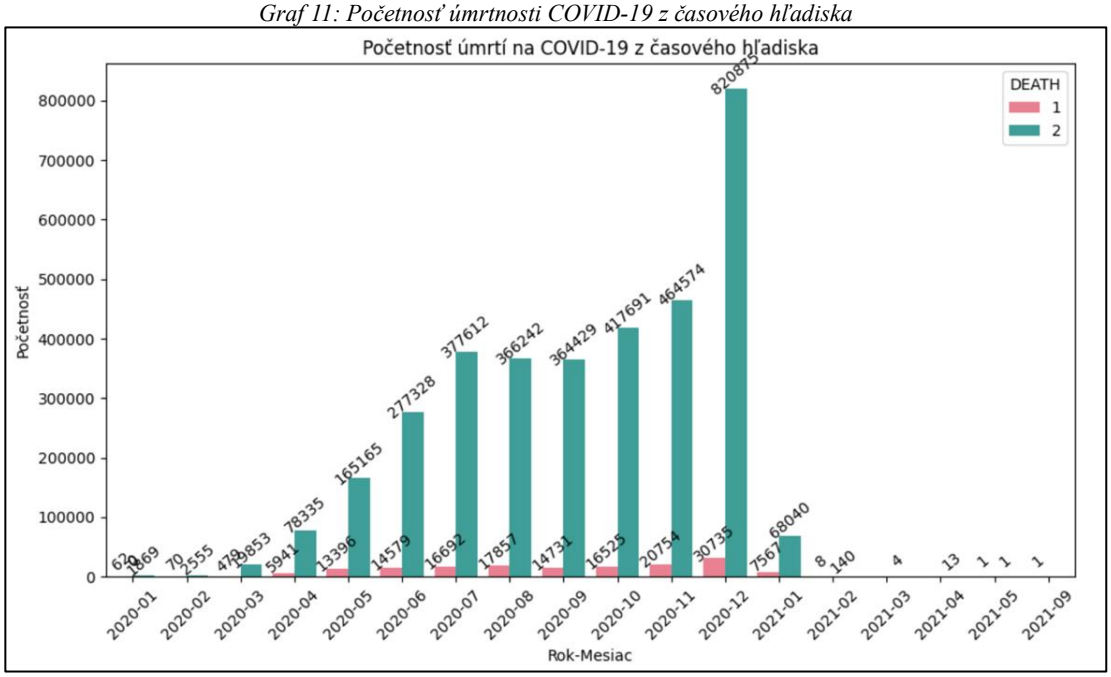
Graf 10: Počet pozorovaní v rámci časového hľadiska



Zdroj: Vytvorené autorkou

Je zrejmé, že pozorovania zaznamenané v tomto dátovom súbore sú v časovom intervale od januára 2020 do septembra 2021. Môžeme konštatovať, že počet pozorovaní má v jednotlivých mesiacoch stúpajúcu tendenciu až do decembra 2020, kde od nového roku 2021 je významný pokles hodnôt. Avšak najväčšia distribúcia pozorovaní je v decembri 2020, kde počet pozorovaní siaha až nad 800-tisíc, čo predstavuje viac ako 20 % z celkového počtu záznamov. Môžeme vidieť, že od júla do novembra 2020 sa hodnoty pohybujú v okolí intervalu od 400 do 500-tisíc záznamov. Najnižší počet záznamov je od februára do septembra 2021, kde je počet pozorovaní v porovnaní so zvyškom v zanedbateľných hodnotách.

Graf 11: Početnosť úmrtnosti COVID-19 z časového hľadiska



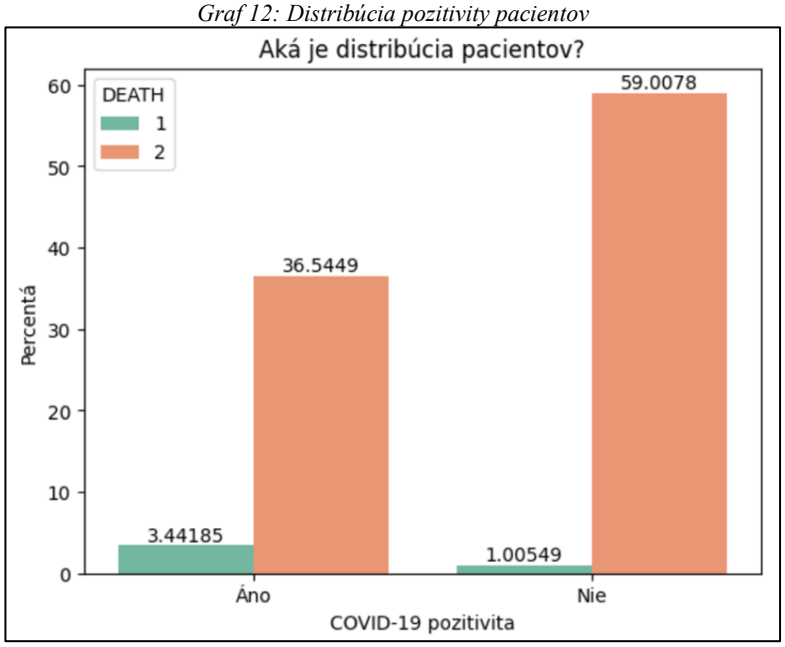
Zdroj: Vytvorené autorkou

Ako sme definovali v predchádzajúcej podkapitole, v danom dátovom súbore je detegovaný značný nepomer medzi záznamami, kde pacient umrel na dané ochorenie a neumrel. Z grafu 11 je značné, že *najväčšia úmrtnosť je zaznamenaná* v mesiaci s najvyšším počtom celkových pozorovaní a to *v decembri 2020 pri viac ako 30-tisíc úmrtiach*. Najmenej úmrtí je zaznamenaných v mesiacoch od februára 2021. Najvyššie percento úmrtí v pomere úmrtie/prežitie v rámci mesiaca je v januári 2021, kde úmrtnosť tvorí až 10 % celkového počtu záznamov v tomto mesiaci.

4.3.4. Aké sú charakteristiky ohľadom pozitivity pacientov a jej dôsledkov?

V predchádzajúcich podkapitolách bolo definované, že existuje významný nepomer medzi distribúciou pacientov, ktorí prežili na toto ochorenie a tými, ktorí neprežili. Domnievame sa, že pri pozitivite bude taktiež istý nepomer, preto je na mieste overenie použitím vizualizácie.

Graf 12: Distribúcia pozitivity pacientov



Zdroj: Vytvorené autorkou

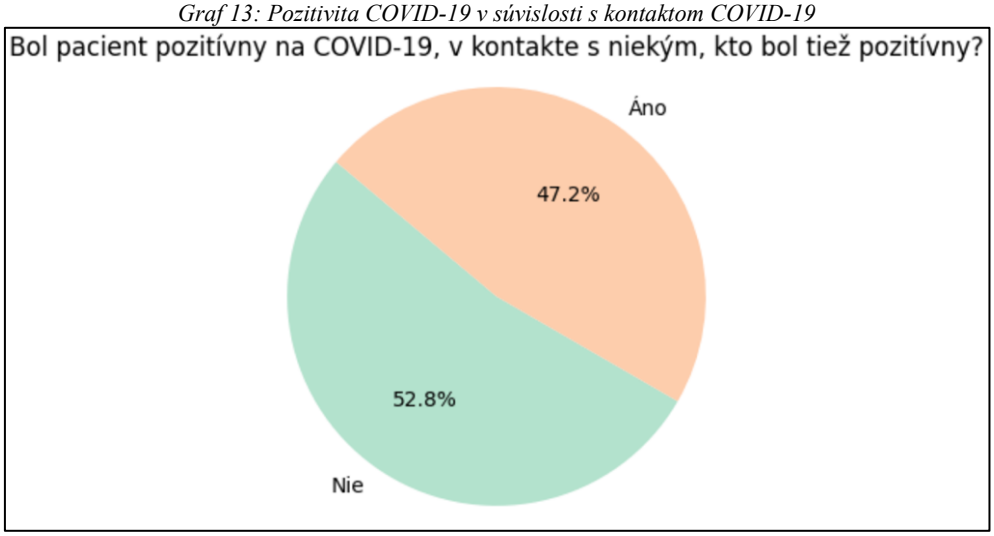
Ako je viditeľné na grafe 12, nepomer pri pozitivite COVID-19 nie je taký znamenitý ako pri úmrtiach. Približne 60 % pacientov z celkového počtu záznamov bolo identifikovaných ako negatívnych na ochorenie COVID-19, pričom *pri zvyšných približne 40 % sa potvrdila pozitivita.* Z celkového počtu záznamov bolo identifikovaných 36,5 % pacientov, ktorí boli pozitívni a prežili, čo tvorí približne 91 % z celkového počtu pozitívnych. Teda inými slovami, *9 % pozitívnych pacientov neprežilo COVID-19* a zvyšok prežil. Zaujímavým zistením je fakt, že úmrtie nastalo aj pri pacientoch, ktorým test nepotvrdil pozitivitu COVID-19 a to približne pri 1 % záznamov z celku. Túto skutočnosť sme sa snažili zanalyzovať z doteraz dostupných zdrojov.

Podľa Sebireho (2020) nie je také ľahké určiť úmrtie, kde dôvodom bol COVID-19. V najjednoduchšom predpoklade by bolo možné identifikovať každé úmrtie kde bol potvrdený pozitívny test vírusu Sars-Cov-2 ako úmrtie na COVID-19. Avšak existujú tu isté problémy ako napríklad náhodnosť infikovania vírusom, ktorý je nesúvisiaci so skutočnou príčinou úmrtia, po ktorej by sa za iných okolností ani nepátralo. Ďalším problémom je, že pacient nemusí byť osobitne testovaný na toto ochorenie, aj keď bolo dôvodom úmrtia. V inom prípade môže nastať falošná negativita v závislosti od odberu vzoriek a ich načasovania či testovania protilátok, ktoré sa rutinne nevykonáva. Klinické príznaky COVID-19 sa môžu v podobnosti prelínať s inými infekciami, ako je napríklad chrípka, sepsa, zlyhanie srdca a iné, čo má taktiež veľký vplyv na určenie dôvodu úmrtia.

Epositova štúdia (2023) sa zaoberá prípadom, kde 83-ročný muž trpel viacerými ochoreniami, akými sú chlopňové ochorenie srdca, srdcové zlyhanie, diabetes, zlyhanie obličiek či chronická obštrukčná choroba pľúc. Jeho prijatie do nemocnice bolo so symptómami dýchavičnosti a negatívnym testom na Sars-Cov-2. Po 11. dňoch hospitalizácie bol pacientovi vykonaný druhý molekulárny test, nakoľko symptómy dýchavičnosti sa zhoršili a tento test potvrdil pozitivitu COVID-19. Pri tomto pacientovi sa potvrdilo nozokomiálne[11] získanie vírusu, kde pacient po 18. dňoch hospitalizácie zomrel. Časový horizont a faktor nozokomiálnej nákazy sú dôležitými aspektami.

Preštudované fakty ohľadom úmrtnosti pri sledovaní COVID-19 aj pre pacientov s negatívnym testom a nepotvrdením tohto ochorenia, budeme zohľadňovať pri vyvodení záverov ohľadom riešenej problematiky.

Graf 13: Pozitivita COVID-19 v súvislosti s kontaktom COVID-19

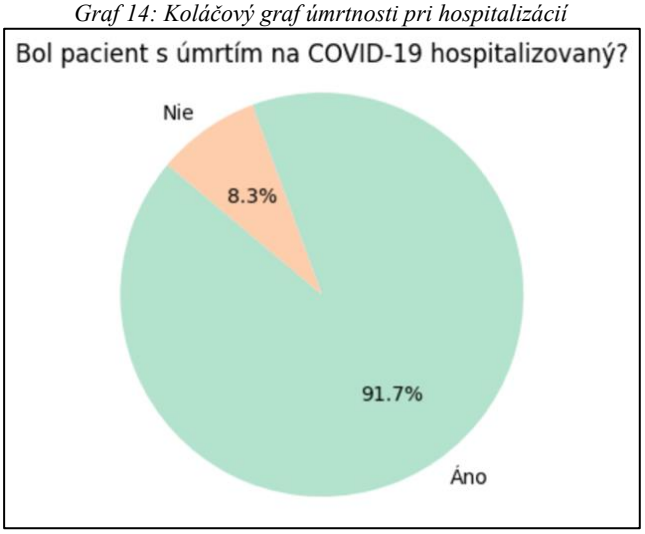


Zdroj: Vytvorené autorkou

Vďaka premenným v dátovom súbore je taktiež možné si overiť, či pôvod pozitivity mohol byť dôsledkom kontaktu s iným pozitívnym pacientom. Potvrdilo sa nám, že pri 47,2 % pozitívnych pacientov nastal kontakt a pri zvyšku nenastal. Môžeme tvrdiť, že *pri približne polovici pacientov mohol nastať prenos vírusu od inej osoby potvrdenej COVID-19, nakoľko kontakt s takouto osobou nastal.*

[11] Nemocničné nakazenie, ktoré vzniká pri pobyte pacienta v zdravotníckom prostredí dôsledkom kontaktu so zdravotníkom alebo pacientom, pri nedostatočne dodržanej hygiene. (Penta Hospitals, 2018)

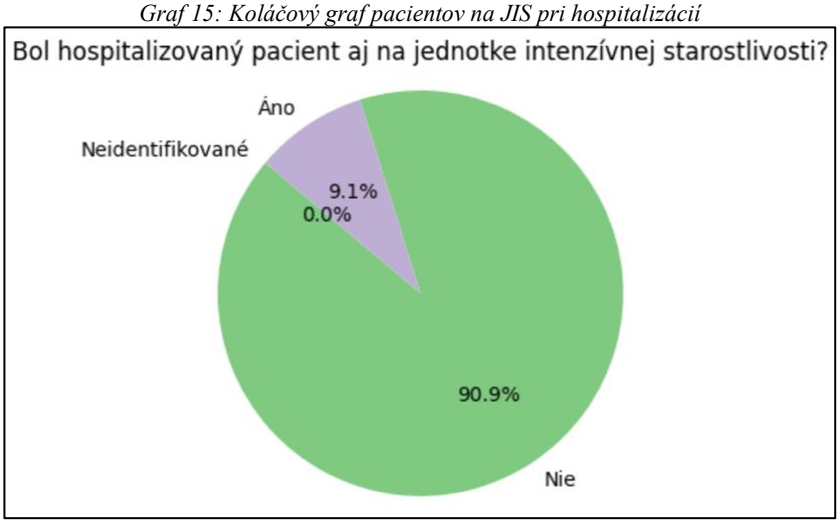
Graf 14: Koláčový graf úmrtnosti pri hospitalizácií



Zdroj: Vytvorené autorkou

Na predstavenom grafe 14 je zrejmé, že úmrtnosť na toto ochorenie úzko súvisí s hospitalizáciou pacientov. *Takmer 92 % všetkých pozorovaných pozitívnych pacientov, ktorí umreli na toto ochorenie bolo hospitalizovaných.* Skutočnosti, ktoré tomuto napovedajú sú jednak, že až pri stave pacienta v zhoršenom štádiu ochorenia, ktoré si vyžadovalo hospitalizáciu nasledovalo úmrtie a taktiež môžeme tvrdiť, že pri väčšine pacientov, ktorých ochorenie končilo úmrtím v danom období, bola poskytnutá zdravotná pomoc vo forme hospitalizácie.

Graf 15: Koláčový graf pacientov na JIS pri hospitalizácií

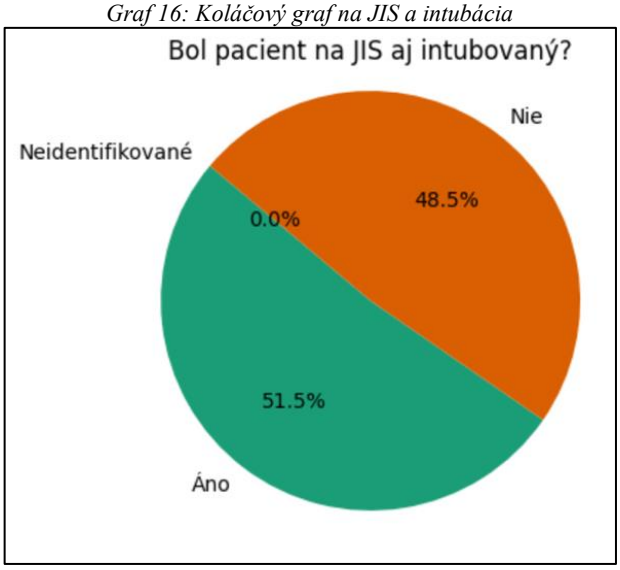


Zdroj: Vytvorené autorkou

Pri objavovaní vzťahu hospitalizácia verzus jednotka intenzívnej starostlivosti (JIS) sme zistili, že približne 9 % z hospitalizovaných pacientov skončilo na JIS. Teda neexistuje tu vzor, kde by hospitalizovaní pacienti vo väčšine skončili na JIS. Avšak môžu tu vstupovať

viaceré faktory, ako napríklad nedostatok kapacity JIS, preto nie je možné jednoznačne určiť záver či nastala alebo nenastala potreba tejto starostlivosti pri hospitalizovaných pacientoch.

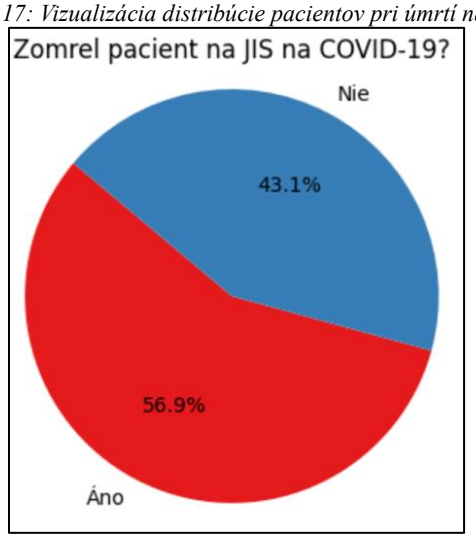
Graf 16: Koláčový graf na JIS a intubácia



Zdroj: Vytvorené autorkou

Z koláčového grafu 16 je zrejmé, že intubácia nastala približne pri 52 % hospitalizovaných na JIS. Vyvodenie záveru na základe tohto zistenia je možné tvrdením, že *približne každý druhý pacient, ktorý bol hospitalizovaný na JIS potreboval napojenie na umelú pľúcnu ventiláciu.*

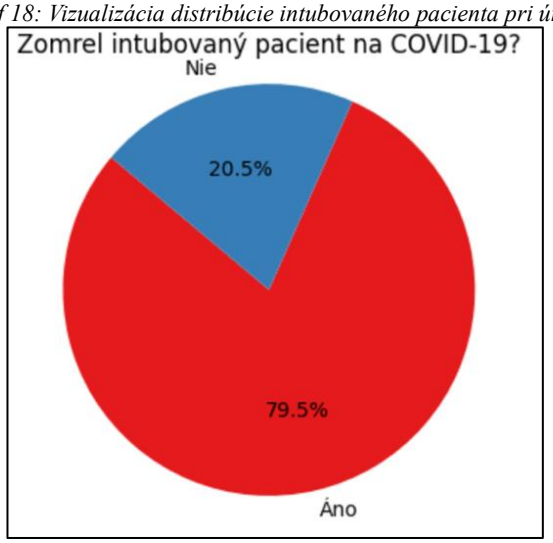
Graf 17: Vizualizácia distribúcie pacientov pri úmrtí na JIS



Zdroj: Vytvorené autorkou

Istým prekvapivým zistením bolo, že približne 40 % zo všetkých pacientov, ktorí boli na JIS sa podarilo dostať z tohto ochorenia. Avšak *viac ako polovica pacientov, ktorá bola na COVID-19 hospitalizovaná na JIS, umrela.*

Graf 18: Vizualizácia distribúcie intubovaného pacienta pri úmrtí



Zdroj: Vytvorené autorkou

Pri intubácii pacienta je jasné, že pri približne 80 % týchto pacientov došlo k úmrtiu. Môžeme tvrdiť, že *drvivá väčšina pacientov napojená na umelú pľúcnu ventiláciu zomrela na COVID-19*.

4.3.5. Aké sú charakteristiky v premenných medicínskej histórie pacientov?

V skúmanom dátovom súbore existujú atribúty ohľadom predošlého zdravotného stavu pacienta, z ktorých je možné určiť isté závery ohľadom vplyvu na pozitivitu či úmrtie pacienta. V nasledujúcich tabuľkách sme zosumarizovali distribúciu špecifikovaných pacientov podľa ochorenia alebo zdravotnej kondície, ktorej bol pacient vystavovaný.

Tabuľka 3: Distribúcia exitovaných pacientov podľa premenných medicínskej histórie

Názov premennej	**Áno**	**Nie**
Pneumónia	71,5 %	28,5 %
Hypertenzia	44,5 %	55,5 %
Diabetes	38,2 %	61,8 %
Obezita	21,4 %	78,6 %
Fajčenie	8,1 %	91,9 %
Chronické zlyhanie obličiek	7,8 %	92,2 %
Iné ochorenie	5,6 %	94,4 %
Kardiovaskulárne ochorenie	5,5 %	94,5 %

Chronická obštrukčná choroba pľúc	4,7 %	95,3 %
Imunosupresia	2,6 %	97,4 %
Astma	1,8 %	98,2 %

Zdroj: Vytvorené autorkou

Pri skúmaní úmrtia pacientov v súvislosti s predošlou medicínskou históriou sme zistili, že najväčšia súvislosť existuje s atribútom ochorenia pneumónie. Pri danom ochorení je zasiahnutá oblasť pľúc, čo dáva istú významnosť pri COVID-19, ktorý je akútnym respiračným ochorením. Teda môžeme tvrdiť, že *väčšina pacientov, ktorá zomrela na COVID-19 mala diagnostikovanú pneumóniu.* Ďalšími signifikantnými ochoreniami sú hypertenzia a diabetes, ktoré sa vyskytli u menej ako polovice exitovaných pacientov, ale stále predstavujú podstatnú časť z ich celkového počtu. Je možné konštatovať, že pacienti diagnostikovaní na tieto ochorenia sú viac ohrození pri pozitivite COVID-19, a to najmä pri pneumatických pacientoch.

V záujme overiť si aké ochorenia majú prevahu pri pozitívnych pacientoch vo všeobecnosti, sme si taktiež vytvorili vizualizácie, ktorých percentuálne distribúcie sú zhrnuté v nasledovnej tabuľke.

Tabuľka 4: Distribúcia pozitívnych pacientov podľa premenných medicínskej histórie

Názov premennej	**Áno**	**Nie**
Hypertenzia	17 %	83 %
Obezita	15,2 %	84,8 %
Pneumónia	13,8 %	86,2 %
Diabetes	13,2 %	86,8 %
Fajčenie	7,5 %	92,5 %
Astma	2,4 %	97,6 %
Iné ochorenie	1,8 %	98,2 %
Kardiovaskulárne ochorenie	1,6 %	98,4 %
Chronické zlyhanie obličiek	1,4 %	98,6 %
Chronická obštrukčná choroba pľúc	1,1 %	98,9 %

Imunosupresia	0,8 %	99,2 %

Zdroj: Vytvorené autorkou

Vzhľadom na väčší počet pozorovaní, ktoré sa zaoberajú pozitivitou COVID-19 v porovnaní s pozorovaniami, kde nastalo úmrtie, môžeme vidieť, že percentá výskytu predošlej medicínskej histórie razantne poklesli. *Nemôžeme tvrdiť, že existujú isté väzby medzi predtým diagnostikovanou kondíciou a pozitívne potvrdeným testom na COVID-19.* Nakoľko ide o respiračné vírusové ochorenie prenášajúce sa kvapôčkami, medicínsky stav pacienta definovaný vyššie spomínanými ochoreniami, nie je dôležitým faktorom pri prenose, čo sa nám potvrdilo aj pri percentuálnej distribúcii. Najväčšie percentuálne zastúpenie pri pozitivite mali atribúty hypertenzie, obezity a pneumónie, všetky však pod 20 %. Nastáva tu pravdepodobnosť, že vo všeobecnosti osoby, ktoré majú totožne zachované podmienky a faktory, tak ako osoby pozorované v dátovom súbore (demografia, prostredie, vek atď.), majú najväčšiu tendenciu byť v danej medicínskej kondícií podľa vyššie priblíženej percentuálnej distribúcie, bez ohľadu na to, či boli pozitívny na COVID-19 alebo nie.

4.4. Modelovanie

Príprava dát a analýza samotného súboru došla do takého stavu, že plynule proces tejto práce prešiel na časť modelovania. Pred samostatným nasadením modelov sme vykonali potrebné náležitosti, ktoré sú priblížené v nasledovných bodoch.

1. *Vytvorenie vzorky*

Nakoľko dátový súbor obsahuje cez 3,5 milióna pozorovaní, výpočtová zložitosť algoritmov by bola veľmi veľká pre bežný procesor počítača. Preto je na mieste vytvoriť reprezentatívnu vzorku dátového súboru, ktorá bude slúžiť na modelovanie. Ako sme mohli vidieť pri vizualizácií distribúcie pacientov, ktorí zomreli na COVID-19 (4,5 %) a ktorí prežili (95,5 %) (Graf 9), existuje tu značný nepomer ohľadom danej charakteristiky pacienta. Ak by bola vytvorená náhodná vzorka 10 000 pozorovaní na základe pôvodného pomeru, tak by približne 9 500 pozorovaní reprezentovalo preživších a 450 pozorovaní reprezentovalo zomrelých. V tomto prípade by sa stalo, že model by určil výsledok s veľmi vysokou úspešnosťou, ale zároveň by tým vznikla veľmi malá presnosť zaradenia a model by boli úplne nepresný, nakoľko z počtu 450 by v testovacom modeli bol veľmi malý počet predikovaných pozorovaní zomrelých pacientov. Preto sme sa rozhodli vytvoriť vzorku

s upraveným pomerom, aby mali modely väčšiu presnosť zaradenia oboch tried a lepšiu schopnosť učiť sa. Vytvorená vzorka obsahuje 50 000 pozorovaní, ktoré sú náhodné premiešané. Podmienkou vytvorenia tejto vzorky je, že obsahuje 20 000 náhodných pozorovaní, kde pacient zomrel na COVID-19 a 30 000 náhodných pozorovaní, kde pacient prežil. Táto vzorka je ďalej použitá v procese.

2. *Nastavenie dátových typov*

Ako sme spomenuli pri príprave dát, tak okrem premenných *AGE* a *DATE_ENTRY* sú všetky premenné kategoriálne, aj keď ich kategórie majú číselnú reprezentáciu. Do tejto chvíle bol dátový typ pre tieto premenné nastavený na *int,* pre lepšiu manipuláciu aj pri vizualizácií. Teraz je dátový typ zmenený na *category,* pre korektné zaobchádzanie s dátami pri modelovaní.

3. *Nastavenie rolí*

Nakoľko cieľom tejto práce je predikcia úmrtia pacientov, tak je potrebné určiť cieľovú premennú. V tomto prípade je premenná *DEATH* označená ako atribút „*y*" – závislá premenná. Premenné okrem *DEATH* a *DATE_ENTRY* sú vyseparované od ostatných a sú označené ako vysvetľujúce premenné „*X*".

4. *Rozdelenie dát*

Keďže pri modelovaní nie je vhodné použiť rovnaké údaje na trénovanie aj na testovanie, údaje sú rozložené do dvoch zložiek. Z tohto dôvodu je stanovené, že 70 % údajov je určených na učenie sa a 30 % na testovanie. Z celkovej modelovacej vzorky 50 000 pozorovaní bude tvoriť testovaciu zložku 15 000 náhodných pozorovaní a 35 000 pozorovaní tvorí trénovaciu zložku.

Po podstúpení určitých pred-modelovacích náležitostí, sme implementovali modely strojového učenia. V tejto časti práce si priblížime vyhodnotenie výsledkov jednotlivých modelov. Pri porovnaní modelov sme využívali kľúčové metriky konfúznej matice, presnosti modelu, precíznosti modelu, úplnosti modelu a F1-skóre. Týmito metrikami sme porovnávali 6 spomínaných algoritmov, ktorých vyššie spomínané metriky sme zhrnuli do nasledovnej tabuľky.

Tabuľka 5: Vyhodnotenie metrík algoritmov

Algoritmus	Presnosť	Precíznosť	Úplnosť	F1-skóre
Logistická regresia	93,28 %	0,91	0,92	0,92
KNN	92,74 %	0,90	0,92	0,91
Náhodný les	92,33 %	0,89	0,92	0,90
SVM	92,09 %	0,89	0,92	0,90
Naive Bayes	91,05 %	0,86	0,93	0,89
Rozhodovací strom	90,93 %	0,88	0,90	0,89

Zdroj: Vytvorené autorkou

Detaily ohľadom výkonnosti modelov sme priblížili v nasledujúcich podkapitolách, osobitne pre jednotlivé algoritmy.

4.4.1. Logistická regresia

Najlepšie vyhodnoteným algoritmom bola logistická regresia. Pre tento algoritmus bol použitý príkaz *sklearn.linear_model.LogisticRegression()* z knižnice Sckit-learn, z ktorej boli použité príkazy modelovania všetkých algoritmov. Parameter klasifikátora „*max_iter*“ sme nastavili na hodnotu 10 000. Tento parameter hovorí o tom, že optimalizačný proces algoritmu má maximálny počet iterácií nastavených na 10 000. Ostatné parametre sú nastavené predvolene.

Konfúzna matica, ktorú vytvoril tento algoritmus vyzerala nasledovne:

Tabuľka 6: Konfúzna matica pre logistickú regresiu

Určenie modelom	Reálna hodnota		Spolu
	+	-	
+	5 450	531	5 981
-	477	8 542	9 019
Spolu	5 927	9 073	15 000

Zdroj: Vytvorené autorkou

Hodnota 93,28 % označuje presnosť modelu, teda percento všetkých korektne klasifikovaných prípadov. Z celkového počtu 15 000 bolo predikovaných správne 13 992 pozorovaní.

Precíznosť modelu hovorí o tom, ako často je model presný pri predikcii úmrtia, ktoré nastalo zo všetkých predikovaných hodnôt úmrtia. Pri modeli logistickej regresie môžeme

tvrdiť, že pre triedu 1 (potvrdené úmrtie) je táto metrika na úrovni 0,91. Z počtu 5 981 bolo korektne predikovaných 5 450 hodnôt.

Úplnosť modelu je pre triedu 1 rovné 0,92. Inými slovami, 92 % prípadov, kde nastalo úmrtie, bolo predikovaných korektne (5 450 z 5 927).

F1-skóre je pre triedu 1 predikovaného atribútu na úrovni 0,92, teda vyvážená miera medzi presnosťou a úplnosťou je 92 %.

4.4.2. KNN

Druhým najefektívnejším algoritmom je KNN. Pri použití *sklearn.neighbors.KNeighborsClassifier()* boli ponechané pôvodné nastavenia. V predvolených parametroch bol počet najbližších susedov rovný 5, kde sa pre výpočet vzdialenosti využila euklidovská vzdialenosť.

V nasledujúcej tabuľke je priblížená konfúzna matica výsledkov pre KNN algoritmus.

Tabuľka 7: Konfúzna matica pre KNN, k = 5

Určenie modelom	**Reálna hodnota**		**Spolu**
	+	-	
+	5 444	606	6 050
-	483	8 467	8 950
Spolu	5 927	9 073	15 000

Zdroj: Vytvorené autorkou

Pri týchto parametroch bola presnosť modelu identifikovaná na 92,74 %, kde z 15 000 pozorovaní bolo správne určených 13 911.

Precíznosť modelu bola na úrovni 0,9 pri predikcii potvrdeného úmrtia. Ak by sme to interpretovali v porovnaní s logistickou regresiou, tak tu nastal pokles o jedno percento.

KNN algoritmus určil úplnosť na 92 %, kde sa tomuto percentu zomrelých pacientov, potvrdilo úmrtie predikciou, čo je v porovnaní s predchádzajúcim algoritmom na rovnakej úrovni.

Harmonický priemer reprezentovaný F1-skóre je 91 %.

Nakoľko je daný algoritmus veľmi špecifický ohľadom nastavení, zaujímalo nás, ako by sa model správal pri odlišnom nastavení počtu najbližších susedov. Keď sme *k* nastavili na hodnotu menšiu ako bolo predvolené, teda *k = 3,* model sa správal nasledovne:

Tabuľka 8: Vyhodnotenie metrík modelu KNN, k = 3

Algoritmus	Presnosť	Precíznosť	Úplnosť	F1-skóre
KNN, k = 3	92,3 %	0,90	0,91	0,90

Zdroj: Vytvorené autorkou

Ako je viditeľné, tak pri danom počte najbližších susedov celková výkonnosť modelu klesá. Presnosť poklesla o približne pol percenta, precíznosť ostáva na tej istej úrovni. Úplnosť aj F1-skóre poklesli o 1 %.

Taktiež sme skúsili nastaviť model pre *k = 8*, kvôli overeniu našich tvrdení. Výsledky sú priblížené v nasledovnej tabuľke.

Tabuľka 9: Vyhodnotenie metrík modelu KNN, k = 8

Algoritmus	Presnosť	Precíznosť	Úplnosť	F1-skóre
KNN, k = 8	93,05 %	0,89	0,94	0,91

Zdroj: Vytvorené autorkou

Tvrdenie sa nám potvrdilo pri tomto nastavení, môžeme konštatovať, že s rastúcim počtom zvolenia najbližších susedov celková presnosť modelu stúpa. Avšak precíznosť modelu poklesla, teda zo všetkých predikovaných hodnôt, že úmrtie nastalo, bolo korektne určených 89 % týchto hodnôt. Na druhej strane, pri úplnosti vidíme značný nárast v porovnaní s oboma predchádzajúcimi nastaveniami algoritmu, ale aj v porovnaní s algoritmom logistickej regresie. Až 94 % skutočných hodnôt úmrtia patrilo predikovaným pacientom, ktorí dané ochorenie neprežili. F1-skóre ostáva na rovnakej úrovni s pôvodným nastavením.

4.4.3. Náhodný les

Ďalším implementovaným algoritmom bol *RandomForestClassifier()*. Základné hyperparametre tohto algoritmu sú: počet stromov = 100, kritérium kvality delenia = Giniho index, maximálna hĺbka delenia = pokým všetky listy nie sú čisté, minimálny počet vzoriek na delenie = 2, minimálny počet vzoriek na list = 1. Všetky zvyšné parametre majú pôvodné nastavenie.

Na základe týchto parametrov nám model vyprodukoval nasledovnú konfúznu maticu:

Tabuľka 10: Konfúzna matica pre náhodný les

Určenie modelom	**Reálna hodnota**		**Spolu**
	+	-	
+	5 432	655	6 087
-	495	8 418	8 913
Spolu	5 927	9 073	15 000

Zdroj: Vytvorené autorkou

Po spracovaní konfúznej matice sme dospeli k nasledovným tvrdeniam:

Presnosť sa pohybuje okolo 92,33 %, model má slušný potenciál.

Precíznosť klesla na úroveň 0,89, no naopak úplnosť modelu je na úrovni 0,92, teda 5 432 predikovaných hodnôt z 5 927 bolo korektne určených v súvislosti s reálnymi úmrtiami pacientov.

F1-skóre je pre tento model na úrovni 0,9.

Aj pri tomto algoritme sme skúsili rôzne nastavenia a dospeli sme k nasledovným výsledkom:

Tabuľka 11: Vyhodnotenie metrík rôznych nastavení pre náhodný les

Počet stromov	**Kritérium delenia**	**Maximálna hĺbka**	**Presnosť**	**Precíznosť**	**Úplnosť**	**F1-skóre**
50	gini	10	93,71 %	0,90	0,95	0,92
500	gini	10	93,73 %	0,89	0,95	0,92
50	entropy	10	93,67 %	0,89	0,95	0,92
100	entropy	auto	92,30 %	0,89	0,92	0,90
100	entropy	3	92,48 %	0,89	0,93	0,91
1000	entropy	3	92,38 %	0,89	0,93	0,91

Zdroj: Vytvorené autorkou

Ako je viditeľné v tomto prípade, metriky modelu sa razantne nemenia pri zmene nastavení. Môžeme však konštatovať, že zmena kritéria delenia nie je zásadnou z pohľadu výsledku metriky. Na druhej strane, čím je maximálna hĺbka nižšia, tým sú metriky nepresnejšie.

4.4.4. SVM

Využitím *svm.SVC()* sme v našej praktickej práci aplikovali algoritmus podporných vektorov (SVM). Parameter *kernel* sme nastavili na "*linear*", teda lineárne jadro. Rozhodovacia hranica je lineárnou funkciu, čiže v prípade dvojrozmerných dát ide o priamku a v inom prípade ide o hyperrovinu. Ostatné parametre boli ponechané pôvodne.

Konfúzna matica modelu SVM bola nasledovná:

Tabuľka 12: Konfúzna matica pre SVM algoritmus

Určenie modelom	**Reálna hodnota**		**Spolu**
	+	-	
+	5 426	685	6 111
-	501	8 388	8 889
Spolu	5 927	9 073	15 000

Zdroj: Vytvorené autorkou

Na základe tejto matice boli vyhotovené nasledujúce zistenia:

Presnosť modelu je 92,09 %, v tomto prípade 13 814 z celkového počtu 15 000.

Precíznosť modelu je na úrovni 0,89 a úplnosť na 0,92.

Harmonický priemer precíznosti a úplnosti – F1-skóre je 90 %.

Ako môžeme vidieť, výsledky tohto algoritmu sa pohybujú zhruba na rovnakom stupni ako pri algoritme Náhodného lesa.

4.4.5. Naive Bayes

Predposledným využitím klasifikátorom bol Naive Bayes, príkazom *GaussianNB()*. Pri pôvodných parametroch nám vyšla nasledovná konfúzna matica s výsledkami modelu:

Tabuľka 13: Konfúzna matica pre Naive Bayes

Určenie modelom	**Reálna hodnota**		**Spolu**
	+	-	
+	5 515	931	6 446
-	412	8 142	8 554
Spolu	5 927	9 073	15 000

Zdroj: Vytvorené autorkou

Daný model nám potvrdil úspešnosť v miere 91,05 %, číselne 13 657 prípadov z celkového počtu 15 000 bolo korektne predikovaných v porovnaní s reálnymi prípadmi.

Precíznosť modelu bola na úrovni 0,86, čo je zo všetkých vykonaných modelov najnižšia úspešnosť. Až 931 prípadov model označil ako úmrtie, aj keď v skutočnosti úmrtie nenastalo.

Úplnosť modelu hovorí, že 93 % všetkých reálnych pozorovaní úmrtia predikoval model správne (5 515 z 5 927). Miera úplnosti je pri tomto modeli naopak najvyššou zo všetkých zrealizovaných modelov.

F1-skóre je na štandardnej úrovni 0,89.

4.4.6. Rozhodovací strom

Posledným využitím klasifikačným algoritmom bol rozhodovací strom, kde bol použitý príkaz *DecisionTreeClassifier()*. Daný model sa pri vyhodnotení javil s najslabším potenciálom. Spomínaný algoritmus má značnú mieru možností nastavenia parametrov. V našom prípade sme aplikovali pôvodné nastavenia kde: kritérium delenia = Giniho index, stratégia rozdelenia v uzle = vyberie najlepšie rozdelenie, maximálna hĺbka = rozšírenie uzlov pokiaľ všetky listy nie sú čisté, maximálny počet listov = ľubovoľný.

V tomto ponímaní bola konfúzna matica nasledovne zaznamenaná:

Tabuľka 14: Konfúzna matica algoritmu rozhodovací strom

Určenie modelom	**Reálna hodnota**		**Spolu**
	+	-	
+	5 308	742	6 050
-	619	8 331	8 950
Spolu	5 927	9 073	15 000

Zdroj: Vytvorené autorkou

Presnosť je reprezentovaná 90,93 %, oproti najlepšie vyhodnotenému modelu logistickej regresie je to o 2,35% menej.

Metrika precíznosti napovedá, že 88 % predikovaných úmrtí bolo zaradených korektne, čo je lepšia výkonnosť ako pri predchádzajúcom algoritme.

Úplnosť je reprezentovaná úrovňou 0,9, teda zo všetkých úmrtí, ktoré nastali sa správne predikovalo 90 %.

Miera medzi precíznosťou a úplnosťou reprezentovaná F1-skóre je totožná ako pri predchádzajúcom modeli, čo je 0,89.

Nakoľko klasifikátor náhodný lesa sa odvíja od problematiky tohto algoritmu, naším predmetom záujmu je taktiež otestovať jeho výkonnosť pri inom nastavení parametrov. Výsledné metriky sme zosumarizovali do nasledujúcej tabuľky.

Tabuľka 15: Výsledné metriky rôznych parametrov modelu použitím rozhodovacieho stromu

Kritérium delenia	Rozdelenie	Maximálna hĺbka	Presnosť	Precíznosť	Úplnosť	F1-skóre
entropy	best	auto	91,12 %	0,88	0,90	0,89
entropy	random	10	93,36 %	0,89	0,95	0,92
gini	random	10	93,25 %	0,89	0,94	0,92
gini	best	3	93,30 %	0,89	0,95	0,92
entropy	best	3	93,30 %	0,89	0,95	0,92

Zdroj: Vytvorené autorkou

Nedá sa poprieť, že pri inom nastavení je výkonnosť modelu na úrovni najefektívnejších modelov, ktoré boli použité. Ako je zrejmé, obmena kritéria delenia nezohráva významnú úlohu v zmene úspešnosti. Javí sa, že nastavenie presnej hĺbky, do ktorej sa má strom vetviť, je prospešnejšie vo výkone modelu.

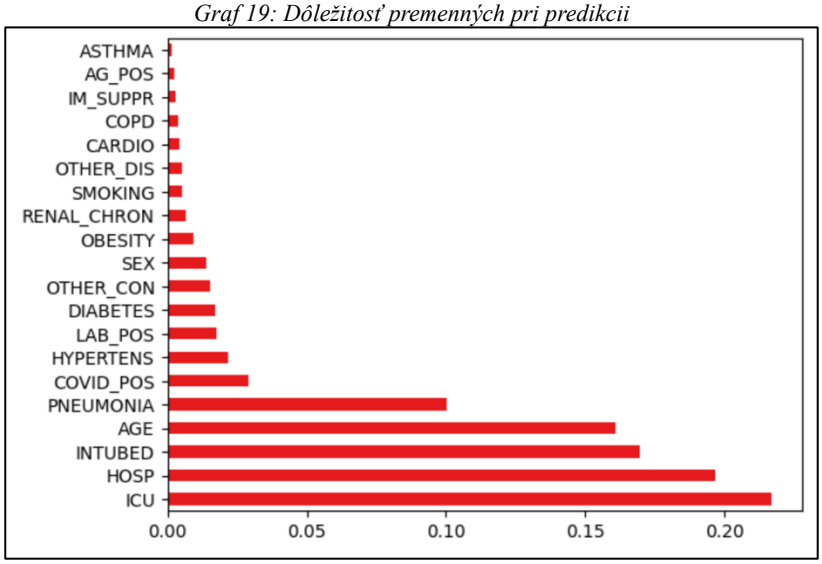
4.5. Dodatočné zistenia z predikcie

Nasadením modelov v tejto práci, sme získali nové hodnoty premenných – predikované hodnoty. Z týchto novo zistených dát je možné sa dozvedieť viaceré dôležité informácie.

Prvou oblasťou, ktorou sme sa zaoberali je dôležitosť jednotlivých premenných a ich vplyv na modelovanie. Využitím atribútu *feature_importances_* v jazyku Python, je pri klasifikačných stromoch a lesoch možné zistiť práve spomínané skutočnosti. Príkaz vypočítava skóre jednotlivých premenných, ktoré s rastúcim trendom predstavuje vyšší vplyv na predikčný model.

Použitím modelu Náhodného lesa a neskôr nadobudnutých predikcií, sme získali nasledovný výstup:

Graf 19: Dôležitosť premenných pri predikcii

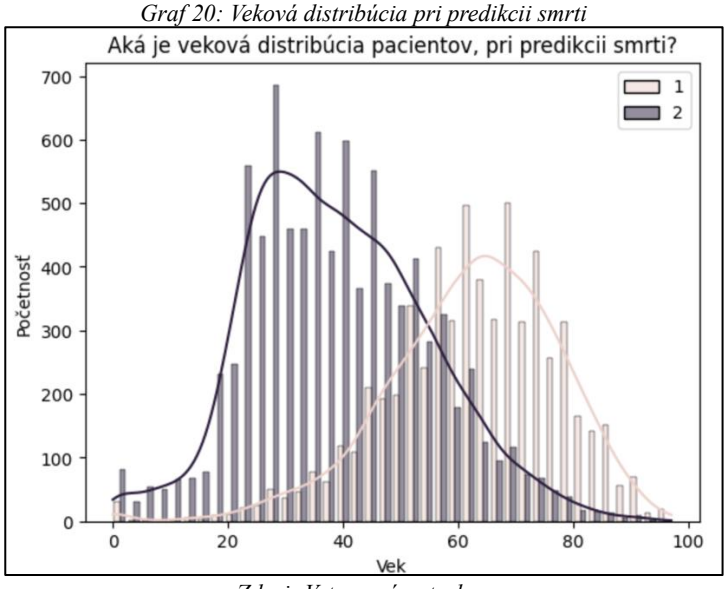


Zdroj: Vytvorené autorkou

Z grafu je zrejmé, že najväčšia dôležitosť je pripisovaná premennej *ICU*. Významnú dôležitosť predstavujú premenné reprezentujúce hospitalizáciu, intubáciu, vek a pneumóniu. Zvyšné premenné nepredstavujú až takú znamenitú významnosť. Táto skutočnosť by mohla byť využitá pri prípadnom ďalšom využití tejto práce a bolo by na zvážení používateľov či by chceli využiť premenné s nízkym skóre, ktoré môžu negatívne ovplyvňovať výkonnosť modelu.

Témou, ktorú sme chceli preanalyzovať je distribúcia predikovaných hodnôt vzhľadom na premenné. Samostatnou premennou, na ktorú sme sa zamerali je vek pri predikcii smrti. Na vizualizáciu sme použili histogram, kde na osi *x* sú použité hodnoty veku z testovacej sady a na osi *y* sú predikované hodnoty smrti. Výsledný histogram je viditeľný na nasledujúcom grafe.

Graf 20: Veková distribúcia pri predikcii smrti



Zdroj: Vytvorené autorkou

Z vytvoreného histogramu je možné konštatovať, že pri predikcii pacientov, ktorí zomreli, sú hodnoty veku vyššie ako pri predikcii pacientov, ktorí prežili. V porovnaní s reálnymi dátami, ktoré sme analyzovali v časti EDA (podkapitola 4.3.1), môžeme spozorovať, že najviac hodnôt predikcii veku pacientov, ktorí prežili sa kumuluje v totožnom intervale pôvodných dát (Graf 4). Pri hodnotách veku pacientov, ktorí neprežili, môžeme vyvodiť záver, že dáta sa vyskytujú v podobných intervaloch, avšak je zrejmé, že špicatosť rozdelenia dát je negatívnejšia (plochšia) ako pri pôvodných dátach (Graf 7). To samozrejme môže byť spôsobené aj nižšou celkovou početnosťou predikovaných dát. Obecne môžeme tvrdiť, že distribúcia veku sa pri predikcii pohybuje obdobne v porovnaní s odrazením reality.

Po dôkladnom preštudovaní výsledkov nás zaujalo, či by bolo možné po daných zisteniach, aplikovať nasadenie nového vstupu na modelovanie, kde by bolo možné vyhodnotiť zaradenie do tried atribútu smrti. V tomto prípade, sme využili model logistickej regresie a pôvodné trénovacie dáta. Parametre modelu logistickej regresie boli nastavené totožne, ako pri modelovaní v podkapitole 4.4.1. Na určenie predikovanej triedy a pravdepodobnosti, sme použili knižničné funkcie pre logistickú regresiu *predict()* a *predict_proba()*.

Najprv sme určili nový vstup – hodnoty premenných pre pacienta, kde následne model na základe zistených asociácií identifikoval, aká ja pravdepodobnosť či pacient prežije alebo nie. Výsledky funkcie sú priblížené na nasledovných dvoch príkladoch:

1. *Aká je pravdepodobnosť, že 33 ročný muž, ktorý je astmatikom a je pozitívny na COVID-19 podľa antigénového testovania, ktorý bol v kontakte s inou pozitívnou osobou a je taktiež fajčiarom, na toto ochorenie nezomrie?*

Obrázok 13: Predikovaná hodnota prvého vstupu

```
Predicted class: [2]
Predicted probabilities: [[0.36659699 0.63340301]]
```

Zdroj: Vytvorené autorkou

Funkcia nám dala výstup, ktorý hovorí o tom, že takýto pacient by dané ochorenie prežil. Táto predikcia je podložená 63,3 % pravdepodobnosťou. Pravdepodobnosť, že by ochorenie neprežil je 36,7 %. Výsledky výstupu potvrdzujú zistenia z predikcií, kde hodnoty, ktoré sme zadali do vstupu nie sú významné v celkovej predikcii a ani nemali vysoké percento výskytu pri predikcii smrti.

2. *Aká je pravdepodobnosť, že 80 ročná žena s COVID-19, ktorá je na toto ochorenie hospitalizovaná a má vysoký krvný tlak, prežije?*

Obrázok 14: Predikovaná hodnota druhého vstupu

```
Predicted class: [1]
Predicted probabilities: [[0.94225147 0.05774853]]
```

Zdroj: Vytvorené autorkou

V tomto prípade môžeme odpozorovať, že takýto pacient by pravdepodobne neprežil s 94,2 % pravdepodobnosťou. Šanca, že by pacient prežil je len 5,8 %. Veková charakteristika a prezencia hospitalizácie napovedajú, že táto predikcia je korektná.

Myslíme si, že daný formát, ktorí sme vytvorili má vysoký potenciál pre ďalšie využitie tejto práce. Vytvorená funkcia založená na modeloch predikcie môže byť použitá v rámci aplikácie pre potreby získania znalostí v oblasti medicíny.

5. Diskusia

Nakoľko sme sa v analytickej časti tejto diplomovej práce venovali preskúmaniu rôznych spôsobov predikcie na základe vstupných údajov a jej následnému porovnaniu výsledkov z mnohých modelov, získali sme dôležité poznatky. Dané výsledky nám napomáhajú zistiť, ktoré metódy sú najúčinnejšie pre nami priblíženú situáciu. Modely strojového učenia sme vytvorili použitím algoritmov a prístupov, ktoré sme preskúmali v metodickej časti tejto práce.

Máme nové poznatky o fungovaní rôznych algoritmov strojového učenia na našich dátach a zistili sme, ktoré z týchto algoritmov sú najúčinnejšie. To nám poskytuje základnú predstavu o tom, aké algoritmy by sme mali využiť v nasledujúcich projektoch a otvára nám dvere pre ďalší výskum a vylepšenie týchto algoritmov.

5.1. Sumarizácia práce

V tejto diplomovej práci sme mali za úlohu vytvoriť model, ktorý sa naučí rozpoznávať prípady úmrtí spojené s COVID-19 a predikovať ich na základe údajov, ktoré sme mali k dispozícii. Dáta, ktoré sa nám podarilo získať obsahovali anonymizované údaje o pacientoch vrátane záznamov o faktoroch, ktoré môžu ovplyvniť priebeh ochorenia COVID-19. Tieto údaje sú z oficiálnej stránky ministerstva zdravotníctva v Mexiku, teda analýza je platná najmä pre túto lokalitu. Skutočnosti ohľadom situácie COVID-19 v Mexiku, ktoré slúžia ako teoretický podklad pri poskytnutí záverov, boli preskúmané. Údaje sme vhodne pripravili na presné modelovanie. Okrem iného sme ich pomocou vizualizácií podrobne zanalyzovali a vyvodili závery. V súvislosti s problémom, na ktorom sme pracovali, sme vytvorili a porovnali 6 rôznych ML modelov: Logistická regresia, K-najbližší sused, Náhodný les, Metóda podporných vektorov, Naive Bayes a Rozhodovací strom. Pre ich posúdenie sme využili bežné postupy hodnotenia výkonnosti modelov strojového učenia, ako sú presnosť (accuracy), precíznosť (precision), úplnosť (recall) a F1-skóre. Výsledky modelov sme taktiež priblížili pomocou konfúznej matice. Spomínané hodnoty sme vypočítali pomocou špeciálnych funkcií, ktoré poskytuje knižnica Scikit-Learn v jazyku Python.

5.2. Interpretácia výsledkov a zistení

Výsledky práce, ktoré boli priblížené v predchádzajúcej kapitole, môžeme interpretovať v súvislosti s faktami ohľadom COVID-19. Na základe precíznej exploratívnej analýzy, vieme zrekapitulovať nasledovné zistenia:

- Sklon k úmrtiu majú pozitívni pacienti na COVID-19 vo veku nad 60 rokov.
- Na toto ochorenie zomrelo viac ľudí mužského pohlavia ako ženského.
- Viac ako 90 % pozitívnych pacientov, ktorí umreli, boli hospitalizovaní.
- Približne každý druhý pacient, ktorý bol hospitalizovaný na JIS, vyžadoval napojenie na umelú pľúcnu ventiláciu.
- Viac ako polovica pozitívnych pacientov na JIS zomrela na toto ochorenie.
- Približne 80 % pacientov napojených na umelú pľúcnu ventiláciu zomrelo na COVID-19.
- Viac ako 70 % pacientov, ktorí zomreli na COVID-19, malo diagnostikovanú pneumóniu.
- Môžeme tvrdiť, že neexistujú väzby medzi pozitivitou COVID-19 a inou diagnostickou kondíciou pacienta.

Okrem iného, sme zistili, že najväčšia úmrtnosť bola zaznamenaná v mesiaci december v roku 2020 pri viac ako 30-tisíc úmrtiach. V kontexte s lokalitou Mexiko, kde boli dáta zaznamenávané, vieme po naštudovaní faktov ohľadom tejto problematiky potvrdiť tento fakt. V tomto období bola prezentovaná druhá epidemiologická vlna, kde bol počas celej pandémie zaznamenaný najvyšší stupeň mortality, čo spôsobil veľmi rozšírený Beta variant tohto ochorenia. Vakcinácia na COVID-19 začala až v nasledujúcom mesiaci. Tieto skutočnosti je potrebné zohľadňovať v rámci generalizácie výsledkov, ktoré sme nadobudli.

Výsledky modelovania nám ukázali, že modely logistickej regresie a KNN sa javia ako najpresnejšie pri 93,3 % a 92,7 %. Avšak metrika úplnosti bola najvyššia pri algoritme Naive Bayes v hodnote 0,93. Vo všeobecnosti môžeme tvrdiť, že hodnoty metrík všetkých modelov sa pohybovali okolo 90 %, čo je vysoká úspešnosť. Vidíme, že neexistuje jednoznačný víťaz, ktorý by súčasne vo všetkých metrikách výrazne exceloval. Toto jasne ilustruje, aké komplexné a náročné je strojové učenie.

Je treba vnímať, že naša vzorka dát na modelovanie bola vybalansovaná v pomere: 60 % pozorovaní, kde pacient prežil a 40 % pozorovaní, kde pacient zomrel. Z tejto vzorky

boli náhodne použité dáta na trénovanie aj testovanie. Ak by bol použitý pomer distribúcie pacientov ohľadom úmrtia na COVID-19 z pôvodného dátového súboru, úspešnosť výsledkov modelovania by zrejme nebola taká efektívna.

Z dodatočných zistení ohľadom predikcie, sme sa dozvedeli, že najväčšia dôležitosť je pripisovaná premennej opisujúcej, či pacient bol hospitalizovaný na JIS. Okrem iného, sú veľmi významné premenné reprezentujúce hospitalizáciu, intubáciu, vek a pneumóniu. Tento fakt je dôkazom, že dané premenné predstavujú väzbu s predikovanou premennou a pozitívne ovplyvňujú výkonnosť modelu.

Taktiež obecne vieme tvrdiť, že distribúcia veku sa pri predikcii pohybuje obdobne v porovnaní so skutočnými dátami. Zistili sme, že pacient, ktorému je predikovaná smrť na COVID-19 má taktiež predikovanú aj prezenciu intubácie a hospitalizácie na JIS. Vidíme tu súvislosť medzi významnosťou premenných a prezenciou pozitívnych hodnôt týchto premenných pri predikcii.

5.3. Prínosy a budúcnosť práce

Podľa vyššie uvedeného zhodnotenia, môžeme konštatovať, že našej práci sa dobre darilo pri používaní a porovnávaní rôznych metód strojového učenia na dátach, ktoré sme si vybrali.

Naše úsilie prináša cenné skúsenosti získané používaním a porovnávaním rôznych algoritmov strojového učenia. Taktiež prispievame dôležitými skutočnosťami formou detailnej analýzy a vysvetľovania výsledkov, čo pomáha objasniť silné a slabé stránky klasifikačného modelovania a poskytuje hodnotný základ pre budúce výskumné práce. Získali sme zásadné poznatky o fungovaní strojového učenia na našich údajoch a získali sme aj dôležité informácie týkajúce sa špecifických aspektov súvisiacich s COVID-19.

Najdôležitejším prínosom tejto práce je jej potenciálne využitie v medicíne. Aj keď je náš model momentálne len prototypom a nie hotovým produktom, má veľký potenciál stať sa dôležitou súčasťou pokročilých inteligentných systémov v medicínskom prostredí. Dokázali sme na základe nového vstupu určiť pravdepodobnosť, či by pacient s danými faktormi prežil alebo neprežil toto ochorenie. Ak by sme pokračovali vo vývoji tejto aplikácie (napr. vo forme softvéru alebo webu), zvyšovali by sme šance, že by sa takýto model mohol stať nevyhnutným nástrojom pre zdravotníkov pri záchrane ľudských životov. To by vďaka jeho schopnosti predikovať možné komplikácie včas, pomáhalo pri prevencii nežiadúcich dôsledkov.

Táto práca má tiež potenciál byť využitá aj v iných odvetviach, ktoré sme si definovali v cieľoch tejto práce. To by mohlo pomôcť aj pri zlepšení príprav pred možnou budúcou pandémiou. Tieto odvetvia zahŕňajú napríklad finančný sektor, poisťovníctvo či manažment rizík. Na to aby sa vedeli tieto sektory lepšie pripraviť, navrhujeme napríklad detegovať finančnú spôsobilosť klienta, ktorý má špecifikovaný profil pri vývoji ochorenia. Taktiež by bolo vhodné pre poisťovateľov investovať do dát a modelovania, vďaka ktorému by vedeli poistencom, ktorí sú prispievateľmi k mortalite COVID-19 svojím chorobným spektrom alebo nezdravým životným štýlom, odôvodniť zvyšovanie nákladov. V týchto prípadoch, by možno museli byť modifikované zdrojové dáta pre využitie nášho modelovania. Model by bol tak natrénovaný na nových dátach, ktoré sú pre danú úlohu relevantné.

Na druhej strane naša práca má isté medzery, ktoré by mohli byť ďalej preskúmané. Model vytvorený v tejto práci, nezahŕňa fázy pandémie, jednotlivé opatrenia, ani mutácie a ich dôsledky pre závažnosť priebehu choroby. Pokročilejšie modely dokážu zohľadniť takéto faktory, čo umožňuje presnejšie predikcie. Teoretické znalosti o pandémií pochádzajú z renomovaných oficiálnych zdrojov, avšak žiadny nezávislý odborník v lekárskej oblasti nebol účasťou na tvorbe a hodnotení modelu. Budúca práca by mohla zahŕňať zapojenie takýchto expertov na vytvorenie nového modelu alebo na detailnejšie preskúmanie existujúcich modelov.

Okrem iného, lokalita z ktorej pochádzajú dáta je taktiež faktorom nedostatkov. Skúmaný zdravotnícky systém v Mexiku je nám cudzí, čo môže viesť k potenciálne prehliadnutým záverom. Zapojenie odborníkov v tejto oblasti by dalo budúcej práci väčšiu aplikovateľnosť, kde by taktiež mohli byť využité demografické dáta tejto problematiky, ktoré neboli pre nás relevantné. Tu vidíme istý priestor pre oficiálne zdroje zdravotníckych informácií na Slovensku, ktoré by sa mohli podieľať na zbere a dostupnosti obdobných dát, ktoré by mohli slúžiť na takýto výskum. Vďaka nim by sme dokázali získať znalosti pre našu krajinu a tým pomôcť lokálnym vedcom, medicínskym odborníkom, zdravotníckemu systému, vládnym nariadeniam či iným podnikateľským sektorom.

Záver

Táto práca venovala pozornosť tvorbe predikčných modelov strojového učenia na klasifikáciu úmrtia pacientov pozitívnych na COVID-19 vzhľadom na ich vek, pohlavie, zdravotný stav a históriu anamnéz. Takisto bolo predmetom objaviť vzory medzi atribútmi, ktoré by mohli negatívne ovplyvňovať dôsledky tejto nákazy.

Keďže ide o špecifickú medicínsku tému, zistili sme, že dáta slúžiace na takýto druh výskumu sú najväčšou výzvou. Dostupnosť dát zo zdravotníckych zariadení ohľadom citlivých informácií pacienta nie je samozrejmosťou. Po dôkladnom rešerši sa nám podarilo získať dáta z oficiálnych vládnych zdrojov Mexika, ktoré sú zdieľané pre verejnosť. Vďaka týmto dátam, sme si naštudovali skutočnosti ohľadom COVID-19 v Mexiku. Zobrali sme na vedomie, že výsledky našej analýzy sú generalizované prioritne pre túto lokalitu.

Na to, aby sme predikčné modely mohli vytvoriť, sme získané dáta upravili do vhodnej podoby využitím zaužívaných metodologických postupov strojového učenia, prameniacich zo štandardizácií dátovej vedy, data miningu a štatistiky. V rámci tohto kroku sme dáta dôkladne zanalyzovali, na čo sme využili rôzne typy vizualizácií. Potvrdilo sa nám, že pacient, ktorý trpí nákazou na COVID-19 má tendenciu umrieť na toto ochorenie, ak je v dôchodcovskom veku. Taktiež k tomu prispieva aj štádium ochorenia, v ktorom je pacient hospitalizovaný alebo napojený na umelú pľúcnu ventiláciu. Dokázali sme identifikovať, že pacient diagnostikovaný na pneumóniu má vysoké riziko úmrtia pri pozitivite COVID-19.

V kontexte modelovania sme porovnali šesť algoritmov strojového učenia: logistickú regresiu, KNN, náhodný les, SVM, naive bayes a rozhodovací strom. Výsledky hodnotenia boli najpozitívnejšie pri logistickej regresii a KNN s presnosťou predikcií 93,3 % a 92,7 %, v tomto poradí. Avšak všetky algoritmy sa nám javili ako veľmi efektívne pre nami nasadenú vzorku dát. Vytvorili sme prototyp modelu, ktorý by dokázal určiť pravdepodobnosť prežitia pacienta, na základe nového vstupu dát. Myslíme si, že tento prototyp je vhodný na budúci vývoj pre vytvorenie hotového produktu vhodného pre odborníkov v tejto oblasti.

Odvážime sa tvrdiť, že hlavné aj čiastkové ciele sme dokázali splniť na úspešnej úrovni. Myslíme si, že táto práca objasnila využitie problému tohto druhu na metódach strojového učenia a prispela svojou hodnotou v rôznych sektoroch. Vidíme tu však aj značný priestor na zlepšenie, kde by mohla byť práca v budúcnosti prebádaná a to hlavne v pokročilejších modeloch, ktoré by brali do úvahy aj iné dôležité aspekty pandémie. Taktiež by model mohol byť použitý na nových dátach, ktoré by bolo možné získať z overených zdrojov zdravotníckych informácií v našej lokalite.

Zoznam použitej literatúry

Abdar, M., Niakan Kalhori, S.R., Sutikno, T., Ibnu Subroto, I.M., & Arji, G. (2015). Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. *International Journal of Electrical and Computer Engineering (IJECE),* 5(6), 1569-1576. Retrieved from http://iaesjournal.com/online/index.php/IJECE

Amazon Web Services. (2024). What is overfitting? Retrieved March 7, 2024 from https://aws.amazon.com/what-is/overfitting/

Amigo, J. M. (2021). Data Mining, Machine Learning, Deep Learning, Chemometrics: Definitions, Common Points, and Trends (Spoiler Alert: VALIDATE your models!). *Brazilian Journal of Analytical Chemistry*, 8(32), 22–38. https://doi.org/10.30744/brjac.2179-3425.AR-38-2021

Athey, S. (2019). The Impact of Machine Learning on Economics. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda* (pp. 507–547). University of Chicago Press. Retrieved from http://www.nber.org/chapters/c14009

Avian Immunology. (2014). Immunosuppression. In *Avian Immunology (Second Edition). ScienceDirect*. https://www.sciencedirect.com/topics/neuroscience/immunosuppression

Breiman, L. (2001). Random Forests. *Machine Learning, 45,* 5–32. https://doi.org/10.1023/A:1010933404324

Brown, S. D., & Myles, A. J. (2009). Decision Tree Modeling in Classification. In S. D. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive Chemometrics* (Vol. 3, pp. 541-569). Elsevier. https://doi.org/10.1016/B978-044452701-1.00025-9

Bruehl, C. (2023, November 15). The Hardest Part: Defining A Target For Classification. *Towards Data Science.* https://towardsdatascience.com/the-hardest-part-defining-a-target-for-classification-872bdeac131b

Butt, U.A., Mehmood, M., Shah, S.B.H., Amin, R., Shaukat, M.W., Raza, S.M., Suh, D.Y., & Piran, M.J. (2020). A Review of Machine Learning Algorithms for Cloud Computing Security. *Electronics*, 9(9), 1379. https://doi.org/10.3390/electronics9091379

CopperTree Analytics. (2019, August 27). Fundamental Series on Building Analytics: Artificial Intelligence, Machine Learning, Predictive Analytics, Deep Learning… What's the Difference? Retrieved from https://www.coppertreeanalytics.com/fundamental-series-on-building-analytics-artificial-intelligence-machine-learning-predictive-analytics-deep-learning-whats-the-difference/

Crabtree, M. (2023, July). What is Machine Learning? Definition, Types, Tools & More. *DataCamp.* Retrieved from https://www.datacamp.com/blog/what-is-machine-learning

Dike, H., Zhou, Y., & Deveerasetty, K. K. (2018). Unsupervised Learning Based On Artificial Neural Network: A Review. In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)* (pp. 1–6). IEEE. https://doi.org/10.1109/CBS.2018.8612259

District of Columbia Government. (2024). What Is an Antigen Test? [Webpage]. Retrieved April 1, 2024 on https://coronavirus.dc.gov/page/what-is-an-antigen-test

Esposito, M., Cocimano, G., Vanaria, F., Sessa, F., & Salerno, M. (2023). Death from COVID-19 in a Fully Vaccinated Subject: A Complete Autopsy Report. *Vaccines,* 11(1), 142. https://doi.org/10.3390/vaccines11010142

Gangadhar, S., & Rangaswamy, S. (2018). Machine Learning. In V. N. Gudivada & C. R. Rao (Eds.), *Handbook of Statistics* (Vol. 38, pp. 197-228). Elsevier. https://doi.org/10.1016/bs.host.2018.07.004

García-López, R., Laresgoiti-Servitje, E., Lemus-Martin, R., Sanchez-Flores, A., & Sanders-Velez, C. (2022). The New SARS-CoV-2 Variants and Their Epidemiological Impact in Mexico. *mBio*, 13(5), e01060-21. https://doi.org/10.1128/mbio.01060-21

GeeksforGeeks. (2023a). *Label Encoding in Python.* Retrieved March 2, 2024 from https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/?ref=lbp

GeeksforGeeks. (2023b). *ML | Understanding Data Processing*. Retrieved March 1, 2024 from https://www.geeksforgeeks.org/ml-understanding-data-processing/?ref=lbp

GeeksforGeeks. (2023c). *What is Exploratory Data Analysis?* Retrieved March 2, 2024 from https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/?ref=lbp

GeeksforGeeks. (2024a). *ML | Overview of Data Cleaning. Data Cleansing Introduction.* Retrieved March 1, 2024 from https://www.geeksforgeeks.org/data-cleansing-introduction/?ref=lbp

GeeksforGeeks. (2024b). *Z score for Outlier Detection – Python*. Retrieved March 1, 2024 from https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/

Gerontology Wiki. (2024, April 11*). List of Mexican supercentenarians*. Retrieved April 28, 2024 from https://gerontology.fandom.com/wiki/List_of_Mexican_supercentenarians

Héberger, K. (2008). Chemoinformatics—multivariate mathematical–statistical methods for data evaluation. *Medical Applications of Mass Spectrometry*, 141-169. https://doi.org/10.1016/B978-044451980-1.50009-4.

Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP),* 5(2). https://doi.org/10.5121/ijdkp.2015.5201

Hotz, N. (2023, January 19). What is CRISP-DM? *Data Science Project Management*. Retrieved February 27, 2024 from https://www.datascience-pm.com/crisp-dm-2/

Chaturvedi, V., Pramanik, A., Ghosh, S., Bhadury, P., & Mondal, A. (2020). A Supervised Approach to Analyse and Simplify Micro-texts. In J. K. Mandal & D. Bhattacharya (Eds.), *Emerging Technology in Modelling and Graphics* (Advances in Intelligent Systems and Computing 937). Springer. https://doi.org/10.1007/978-981-13-7403-6_23

IBM. (2024, February 26). *What is exploratory data analysis (EDA)?* Retrieved February 28, 2024 from https://www.ibm.com/topics/exploratory-data-analysis

Jairi, I. (2021, December 19). A Simple Explanation of Entropy and Information Gain | Decision Tree Classification | Machine Learning. Medium. Retrieved April 9, 2024 from https://medium.com/@jairiidriss/a-simple-explanation-of-entropy-and-information-gain-decision-tree-classification-machine-f7273c3a6f19

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.).* Stanford University. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology, 1*(1). https://doi.org/10.4304/jait.1.1.12-20

Kim, M., Gu, Z., & Yu, S. (2021). Methods, Challenges, and Practical Issues of COVID-19 Projection: A Data Science Perspective. *Philosophies of Data Science,* 19(2), 219–242. https://doi.org/10.6339/21-JDS1013

Knaul, F., Arreola-Ornelas, H., Porteny, T., Touchton, M., Sánchez-Talanquer, M., Méndez, Ó., Chertorivski, S., Ortega, S., Chudnovsky, M., Kuri, P., & the group from the Observatory for the Containment of COVID-19 in the Americas. (2021). Public health policies to combat COVID-19 in Mexico's states. *PLOS ONE*, 16(6), e0251722. https://doi.org/10.1371/journal.pone.0251722

Labudová, V. (2017). Rozhodovacie stromy ako prediktívna modelovacia technika. *Slovenská štatistika a demografia, 27(*3), 60-76. Retrieved from https://ssad.statistics.sk/SSaD/wp-content/files/3_2017/3_2017_clanok_6_Labudova.pdf

Lahiri, R., Dey, S., Roy, S., & Nag, S. (2020). Detection of Pulsars Using an Artificial Neural Network. In J. K. Mandal & D. Bhattacharya (Eds.), *Emerging Technology in Modelling and Graphics* (Advances in Intelligent Systems and Computing 937). Springer. https://doi.org/10.1007/978-981-13-7403-6_23

Lutkevich, B. (2023, September 11). How to scrape data from a website. *TechTarget*. Retrieved March 1, 2024 from https://www.techtarget.com/whatis/feature/How-to-scrape-data-from-a-website

Mayo Clinic Staff. (2023, July 22). Obesity. *Mayo Clinic*. Retrieved March 30, 2024 on https://www.mayoclinic.org/diseases-conditions/obesity/symptoms-causes/syc-20375742

Merchant, F. A., Shah, S. K., & Castleman, K. R. (2023). Object Measurement. In F. A. Merchant & K. R. Castleman (Eds.), *Microscope Image Processing (Second Edition)* (pp. 153-175). Academic Press. DOI: 10.1016/B978-0-12-821049-9.00017-4

National Heart, Lung, and Blood Institute. (2022, March 24). Pneumonia: What Is Pneumonia? Retrieved March 30, 2024 from https://www.nhlbi.nih.gov/health/pneumonia

National Health Service (NHS). (2022, April 22). Cardiovascular disease. Retrieved March 30, 2024 on https://www.nhs.uk/conditions/cardiovascular-disease/

National Health Service (NHS). (2023, April 11). Chronic obstructive pulmonary disease (COPD). Retrieved March 30, 2024 on https://www.nhs.uk/conditions/chronic-obstructive-pulmonary-disease-copd/

Nixon, K., Jindal, S., Parker, F., Marshall, M., Reich, N. G., Ghobadi, K., Lee, E. C., Truelove, S., & Gardner, L. (2022). Real-time COVID-19 forecasting: challenges and opportunities of model performance and translation. *The Lancet. Digital health*, 4(10), e699–e701. https://doi.org/10.1016/S2589-7500(22)00167-4

Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology (IJCTT),* 48(3), 126. DOI: 10.14445/22312803/IJCTT-V48P126

Panjeta, M., Reddy, A., Shah, R., & Shah, J. (2024). Artificial intelligence enabled COVID-19 detection: techniques, challenges and use cases. *Multimed Tools Appl,* 83, 4639–4666. https://doi.org/10.1007/s11042-023-15247-7

Penta Hospitals. (2018, April 25). Viete, čo sú nozokomiálne nákazy? Retrieved April 1, 2024 on https://pentahospitals.sk/nozokomialne-nakazy/

Rady, E.-H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked,* 15, 100178. https://doi.org/10.1016/j.imu.2019.100178

Sarmiento Varón, L., González-Puelma, J., Medina-Ortiz, D., Aldridge, J., Alvarez-Saravia, D., Uribe-Paredes, R., & Navarrete, M. A. (2023). The role of machine learning in health policies during the COVID-19 pandemic and in long COVID management. *Frontiers in Public Health*, 11, Article 1140353. https://doi.org/10.3389/fpubh.2023.1140353

Sebire, N. (2020, May 29). Why it is hard to answer, 'So how many people have died of COVID-19?' [LinkedIn article]. LinkedIn. Retrieved March 31, 2024 on

https://www.linkedin.com/pulse/why-hard-answer-so-how-many-people-have-died-covid-19-neil-sebire/

Secretaría de Salud. (2024, March 5). Datos Abiertos Bases Históricas - Dirección General de Epidemiología. *Gobierno de México*. Retrieved March 12, 2024 from https://www.gob.mx/salud/documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In J. K. Mandal & D. Bhattacharya (Eds.), *Emerging Technology in Modelling and Graphics* (Advances in Intelligent Systems and Computing 937). Springer. https://doi.org/10.1007/978-981-13-7403-6_23

Seth, N. (2023, November 03). Entropy in Machine Learning: Definition, Examples and Uses. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/11/entropy-a-key-concept-for-all-data-science-beginners/

Shaw, B., Suman, A. K., & Chakraborty, B. (2020). Wine Quality Analysis Using Machine Learning. In J. K. Mandal & D. Bhattacharya (Eds.), *Emerging Technology in Modelling and Graphics* (Advances in Intelligent Systems and Computing 937). Springer. https://doi.org/10.1007/978-981-13-7403-6_23

Silva, C. S., & Fonseca, J. M. (2017). Educational Data Mining: A Literature Review. In *Europe and MENA Cooperation Advances in Information and Communication Technologies* (pp. 87–94). Advances in Intelligent Systems and Computing. https://doi.org/10.1007/978-3-319-46568-5_9

Sperandei S. (2014). Understanding logistic regression analysis. *Biochemia medica, 24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Stojiljković, M. (2020, January 13). Logistic Regression in Python. *Real Python*. Retrieved March 10, 2024 from https://realpython.com/logistic-regression-python/

Swamynathan, M. (2017). Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python (1st ed.). *Apress Berkeley*. https://doi.org/10.1007/978-1-4842-2866-1

Štalmachová, K., & Strenitzerová, M. (2020). Umelá inteligencia, strojové učenie a trh práce. *Pošta, Telekomunikácie a Elektronický Obchod,* 15(2), 52–58. https://doi.org/10.26552/pte.C.2020.2.7

Terek, M., Horníková, A., & Labudová, V. (2010). *Hĺbková analýza údajov.* Iura Edition.

Tiwari, S., Chanak, P., & Singh, S. K. (2022). A Review of the Machine Learning Algorithms for Covid-19 Case Analysis. *IEEE transactions on artificial intelligence,* 4(1), 44–59. https://doi.org/10.1109/TAI.2022.3142241

Wikiskripta. (2018, June 28). *Endotracheální intubace.* Retrieved March 29, 2024 on https://www.wikiskripta.eu/w/Endotrache%C3%A1ln%C3%AD_intubace

World Health Organization. (2023a). *Hypertension*. Retrieved March 30, 2024 on https://www.who.int/news-room/fact-sheets/detail/hypertension

World Health Organization. (2023b). *Diabetes.* Retrieved March 30, 2024 on https://www.who.int/news-room/fact-sheets/detail/diabetes

World Health Organization. (2023c). *Asthma.* Retrieved March 30, 2024 on https://www.who.int/news-room/fact-sheets/detail/asthma

World Health Organization. (2024, February 21). *COVID-19 cases | WHO COVID-19 dashboard.* Retrieved on https://data.who.int/dashboards/covid19/cases?m49=484&n=c

Worldometer. (2024, February 21). *Mexico COVID – Coronavirus Statistics – Worldometer.* Retrieved on https://www.worldometers.info/coronavirus/country/mexico/

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems,* 14(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

Zhao, W., Bhushan, A., Santamaria, A. D., Simon, M. G., & Davis, C. E. (2008). Machine Learning: A Crucial Tool for Sensor Design. *Algorithms, 1*(2), 130-152. https://doi.org/10.3390/a1020130

Zollanvari, A. (2023). *Machine Learning with Python: Theory and Implementation (1st ed.)*. Springer Cham. https://doi.org/10.1007/978-3-031-33342-2