

**UNIVERZITA MATEJA BELA V BANSKEJ BYSTRICI
FAKULTA PRÍRODNÝCH VIED**

**ANALÝZA ROZSIAHLÝCH ÚDAJOV V DOPRAVE
A MODELOVANIE RIZÍK V INTELIGENTNÝCH
MESTÁCH POMOCOU METÓD UMELEJ
INTELIGENCIE
DIPLOMOVÁ PRÁCA**

5071ACCB-B908-4EEF-8DBC-E5AC933854F5

**UNIVERZITA MATEJA BELA V BANSKEJ BYSTRICI
FAKULTA PRÍRODNÝCH VIED**

**ANALÝZA ROZSIAHLÝCH ÚDAJOV V DOPRAVE
A MODELOVANIE RIZÍK V INTELIGENTNÝCH MESTÁCH
POMOCOU METÓD UMELEJ INTELIGENCIE**

Diplomová práca

5071accb-b908-4eef-8dbc-e5ac933854f5

Študijný program: aplikovaná informatika
Študijný odbor: informatika
Školiace pracovisko: Katedra informatiky
Vedúci práce: doc. Ing. Jarmila Škrinárová, PhD.

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Adam Valašťan
Študijný program: aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: Magisterská záverečná práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Analýza rozsiahlych údajov v doprave a modelovanie rizík v inteligentných mestách pomocou metód umelej inteligencie

Anotácia: Analyzujte dostupné datasety pre účely modelovania dopravy v Smart Cities v súvislosti s počasím a meteorologickými údajmi. Špecifikujte kľúčové úlohy Smart Cities, ktoré súvisia s premávkou, znečistením ovzdušia a meteorologickými údajmi a ich vzájomným prepojením. Navrhnite kritériá kvality pre analýzu riešeného problému. Navrhnite metodiku a urobte analýzy, na báze umelej inteligencie, s cieľom dosiahnuť špecifikované kritériá. Zosumarizujte zistenia a vzájomné korelácie. Navrhnite modely s využitím umelej inteligencie, na základe ktorých bude možné predikovať rizikové situácie z pohľadu znečistenia ovzdušia alebo obmedzenia premávky.

Vedúci: doc. Ing. Jarmila Škrinárová, PhD.
Katedra: KIN FPV - Katedra informatiky
Vedúci katedry: doc. Ing. Jarmila Škrinárová, PhD.
doc. PaedDr. Patrik Voštinár, PhD.

Dátum zadania: 03.04.2025

Dátum schválenia smerovej katedry: 04.04.2025 prof. Dr. Ing. Miroslav Svítek, Dr.
osoba zodpovedná za realizáciu študijného programu

Čestné vyhlásenie

Čestne vyhlasujem, že som túto diplomovú prácu s názvom "Analýza rozsiahlych údajov v doprave a modelovanie rizík v inteligentných mestách pomocou metód umelej inteligencie" spracoval samostatne na základe získaných teoretických vedomostí a že všetku použitú literatúru a ďalšie pramene som v diplomovej práci vyznačil.

Banská Bystrica, 29. apríla 2026

.....
vlastnoručný podpis

PodĎakovanie

Prvotné podĎakovanie patrí mojej školiteľke, doc. Ing. Jarmile Škrinárovej, PhD., za odborné vedenie záverečnej práce, cenné pripomienky, usmernenia a čas, ktorý mi počas jej vypracovania venovala.

Ďalej by som sa chcel podĎakovať Mgr. Adamovi Dudášovi, PhD. za inšpiratívne nápady, odborné konzultácie a nápomoc pri návrhu metodiky spracovania dát. Rovnako ďakujem Bc. Kristíne Mozolíkovej za ochotu a poskytnutie fotografického materiálu zameraného na integráciu senzorov do dopravy v meste Krakov, ktorý prácu vizuálne a fakticky obohatil.

Nakoniec patrí obrovská vďaka mojim rodičom za vytvorenie vhodného prostredia nielen počas písania tejto práce, ale aj počas celého môjho štúdia.

Abstrakt

VALAŠŤAN, Adam: Analýza rozsiahlych údajov v doprave a modelovanie rizík v inteligentných mestách pomocou metód umelej inteligencie [Diplomová práca]. Univerzita Mateja Bela v Banskej Bystrici. Fakulta prírodných vied, Katedra informatiky. Vedúci práce: doc. Ing. Jarmila Škrinárová, PhD. — Stupeň odbornej kvalifikácie: magister. Banská Bystrica: FPV UMB, 2026, 75 strán.

Záverečná práca skúma koncept inteligentných miest (Smart Cities) so zameraním na kritický vzťah medzi mestskou dopravou, kvalitou ovzdušia a meteorologickými faktormi. Hlavným cieľom je návrh a implementácia modelu umelej inteligencie schopného predikovať dopravnú situáciu s hodinovým predstihom na základe historických dát, čo slúži ako nástroj na podporu proaktívneho rozhodovania v mestskom prostredí.

Výskum využíva ucelený dataset z vyťaženej Legerovej ulice v Prahe, ktorý združuje údaje o intenzite dopravy, meteorologických podmienkach a koncentrácii emisií z druhej polovice roka 2020. Exploratívna analýza dát exaktne potvrdila priamu závislosť medzi intenzitou dopravy a koncentráciami oxidov dusíka, ako aj dôležitý ošističujúci efekt rýchlosti vetra, ktorý zabezpečuje rozptyl polutantov. Predspracovanie údajov zahŕňalo ošetrovanie anomálií senzorov a transformáciu časových radov do 24-hodinových sekvencií pomocou metódy posuvného okna.

Samotná predikcia je realizovaná pomocou rekurentnej neurónovej siete s architektúrou Gated Recurrent Unit (GRU), optimalizovanej regularizačnou vrstvou Dropout a mechanizmom EarlyStopping na ochranu pred preučeníím. Model úspešne zachytáva dynamiku časového radu a dokáže spoľahlivo predikovať ranné a poobedné dopravné špičky. Kvantitatívne vyhodnotenie na testovacích dátach preukázalo vysokú presnosť; model dosiahol priemernú absolútnu chybu (MAE) na úrovni 302 vozidiel za hodinu pre intenzitu dopravy a odchýlku 3 km/h pri predikcii priemernej rýchlosti vozidiel. Tieto výsledky dokazujú, že implementácia GRU modelu predstavuje spoľahlivý a v praxi využiteľný nástroj na predikciu dopravy v kontexte Smart Cities.

Kľúčové slová: inteligentné mestá, neurónové siete, dátová analýza, gru, predikcia dopravy

Abstract

VALAŠŤAN, Adam: Analysis of large-scale data in transport and risk modeling in smart cities using artificial intelligence methods [Master's thesis]. Matej Bel University in Banská Bystrica. Faculty of Natural Sciences, Department of Computer Science. Thesis supervisor: doc. Ing. Jarmila Škrinárová, PhD. — Qualification level: Master. Banská Bystrica: FPV UMB, 2026, 75 pages.

This thesis explores the concept of Smart Cities, focusing on the critical relationship between urban traffic, air quality, and meteorological factors. The primary objective is the design and implementation of an artificial intelligence model capable of predicting the traffic situation one hour in advance based on historical data, serving as a tool to support proactive decision-making in urban environments.

The research utilizes a comprehensive dataset from the busy Legerova Street in Prague, which aggregates data on traffic intensity, meteorological conditions, and emission concentrations from the second half of 2020. Exploratory data analysis explicitly confirmed the direct correlation between traffic intensity and nitrogen oxide concentrations, as well as the important cleansing effect of wind speed, which ensures the dispersion of pollutants. Data preprocessing involved handling sensor anomalies and transforming the time series into 24-hour sequences using the sliding window method.

The prediction itself is realized using a recurrent neural network with a Gated Recurrent Unit (GRU) architecture, optimized by a Dropout regularization layer and an EarlyStopping mechanism to protect against overfitting. The model successfully captures the dynamics of the time series and can reliably predict both morning and afternoon traffic peaks. Quantitative evaluation on the test data demonstrated high accuracy; the model achieved a Mean Absolute Error (MAE) of 302 vehicles per hour for traffic intensity and a deviation of 3 km/h when predicting the average vehicle speed. These results prove that the implementation of the GRU model represents a reliable and practically viable tool for traffic prediction within the context of Smart Cities.

Keywords: smart cities, neural networks, data analysis, gru, traffic prediction

Obsah

Zoznam zdrojových kódov	10
Zoznam ilustrácií a tabuliek	10
Úvod	12
1 Smart Cities	13
1.1 Čo sú Smart Cities	13
1.2 Kľúčové úlohy Smart Cities	15
1.2.1 Mestská doprava	15
1.2.2 Kvalita ovzdušia	16
1.2.3 Vplyv počasia	17
1.3 Príklady úspešnej integrácie	18
1.3.1 Mesto Krakov	18
2 Kritériá kvality	21
2.1 Doménové kritériá	21
2.1.1 Časový predstih predikcie	21
2.1.2 Výpočtová efektivita a časová odozva	21
2.1.3 Nasaditeľnosť a reprodukovateľnosť	22
2.2 Kritériá kvality života	22
2.2.1 Plynulosť a intenzita dopravy	22
2.2.2 Limity znečistenia ovzdušia	22
2.2.3 Meteorologické podmienky	23
2.3 Vyhodnocovacie metriky modelu	23
2.3.1 Metrika Root Mean Square Error (RMSE)	24
2.3.2 Metrika Mean Absolute Error (MAE)	24
3 Neurónové siete	25
3.1 Učenie sa neurónových sietí	25
3.1.1 Gradientný zostup a spätné šírenie chyby	26
3.2 Aktivačné funkcie	27
3.3 Rekurentné neurónové siete	30
3.4 Architektúra Gated Recurrent Unit	31
4 Spracovanie a analýza dát	34
4.1 Zber a integrácia dát	34
4.1.1 Dataset z Legerovej ulice	35

4.2	Agregácia a transformácia zozbieraných dát	38
4.2.1	Zjednotenie dát podľa času	38
4.2.2	Predbežná filtrácia a ošetrovanie anomálií	39
4.2.3	Zjednotenie záznamov pomocou odkazov	39
4.2.4	Premenovanie atribútov	40
4.2.5	Finálny export	40
4.3	Predspracovanie dát	41
4.3.1	Analytické zhodnotenie a čistenie anomálií	41
4.3.2	Riešenie chýbajúcich hodnôt	42
4.4	Exploratívna analýza dát	42
4.4.1	Analýza distribúcie dát a skúmanie extrémov	43
4.4.2	Analýza časových závislostí a sezónnosti	47
4.4.3	Korelačná analýza	49
4.4.4	Sumarizácia analýz a kľúčové zistenia	54
5	Modelovanie na báze umelej inteligencie	55
5.1	Príprava dát pre modelovanie	55
5.2	Návrh a architektúra modelu	58
5.3	Tréning modelu	60
5.4	Vyhodnotenie natrénovaného modelu	62
5.4.1	Kvantitatívne vyhodnotenie	62
5.4.2	Vizuálna analýza predikcie	62
	Záver	65
	Zoznam bibliografických odkazov	66
	Prílohy	70

Zoznam zdrojových kódov

1	Importovanie knižníc	56
2	Načítanie dát	56
3	Určenie vstupov a výstupov modelu	56
4	Škálovanie dát	57
5	Tvorba časových okien	57
6	Rozdelenie dát	58
7	Architektúra modelu	59
8	Trénovanie modelu	60
9	Uloženie modelu, škálovačov a histórie	60
10	Spätné škálovanie hodnôt	62

Zoznam ilustrácií a tabuliek

1	Infraštruktúra Smart City (Zdroj: [1])	14
2	Diagram fáz konceptu inteligentného monitorovania mesta Krakov	18
3	Zobrazenie kvality ovzdušia pomocou koncentrácie prachových častíc PM10 v dopravnom prostriedku na zobrazovacej jednotke. Osobné údaje boli cenzúrované. Zdroj: Mozolíková, K. (2026)	19
4	Ukážka zjednodušenej neurónovej siete	25
5	Gradientný zostup	26
6	Lineárna aktivačná funkcia	27
7	Sigmoidálna aktivačná funkcia	28
8	Funkcia hyperbolického tangensu	28
9	Aktivačná funkcia ReLU	29
10	Rekurentná neurónová sieť	30
11	Model bunky GRU (Zdroj: [2])	31
12	Geografická lokalizácia Legerovej ulice (zvýraznené červenou farbou) ako súčasť hlavného dopravného ťahu (Zdroj: [3])	35
13	Použité hárky datasetu z Legerovej ulice	35
14	Ukážka dát hárku "Doprava (R203005)"	36
15	Ukážka dát hárku "HODIN Teplota a srážky (Karlovy)"	36
16	Ukážka dát hárku "PM10"	37
17	Ukážka dát hárku "PM2.5"	37
18	Ukážka dát hárku "CO"	37
19	Ukážka dát hárku "NO2"	37
20	Ukážka dát hárku "NOx"	37
21	Ukážka dát hárku "ID_Value_type"	37

22	Ukážka abnormálnych dát hárku "CO"	38
23	Ukážka hodnôt atribútu Datetime hárku "Dataset1"	39
24	Ukážka filtrovaných hodnôt hárku "Doprava (R203005)"	39
25	Ukážka prepojených hodnôt v hárku "Dataset1"	40
26	Ukážka hodnôt vo finálnom hárku "Dataset2"	40
27	Ukážka vyexportovaného datasetu vo formáte .csv	41
28	Histogram distribúcie teploty vzduchu	43
29	Histogram distribúcie kvality ovzdušia	44
30	Histogram distribúcie dopravy	44
31	Boxplot bez odľahlých hodnôt	45
32	Boxplot s odľahlými hodnotami	46
33	Priemerný denný cyklus intenzity dopravy a emisie NO2	47
34	Priemerná intenzita dopravy počas týždňa a priemerná koncentrácia PM10 počas týždňa	48
35	Mesačný vývoj koncentrácie PM10	49
36	Spearmanova korelačná matica	50
37	Vplyv intenzity dopravy na koncentráciu oxidov dusíka (NOx)	51
38	Vplyv rýchlosti vetra na koncentráciu prachových častíc PM10	51
39	Vplyv teploty vzduchu na koncentráciu PM2.5	52
40	Vplyv zrážok na koncentráciu PM10	52
41	Vplyv PM2.5 na koncentráciu PM10	53
42	Vplyv NO2 na koncentráciu NOx	53
43	Architektúra modelu	58
44	Vývoj chybovej krivky počas tréningu	61
45	Porovnanie predikcie a skutočnosti intenzity dopravy a priemernej rýchlosti vozidiel	63
46	Histogram distribúcie chýb modelu	64
47	Výstup skriptu v konzole	72
1	Ukážka pomocného atribútu REAL_VALUE pre hárky "CO"	39
2	Ukážka doplnenia chýbajúcich hodnôt pre atribút oxid uhoľnatý (CO)	42

Úvod

Rastúca urbanizácia a život v moderných mestách so sebou prinášajú kritické výzvy, predovšetkým v podobe preťaženej dopravy a zhoršenej kvality ovzdušia. Odpoveďou na tieto problémy je koncept inteligentných miest (Smart Cities), ktorý využíva dáta zo senzorických sietí a umelú inteligenciu na moderné riadenie verejnej infraštruktúry. Keďže emisie zo spaľovacích motorov v kombinácii s nepriaznivými meteorologickými javmi (ako sú teplotné inverzie) predstavujú mimoriadne vážne riziko pre verejné zdravie, je pre samosprávy kľúčové tieto hrozby nielen monitorovať, ale s dostatočným časovým predstihom aj predikovať.

Z tohto dôvodu je hlavným cieľom predkladanej diplomovej práce návrh a implementácia predikčného modelu na báze umelej inteligencie, ktorý dokáže predpovedať vývoj dopravnej situácie. Pre úspešné modelovanie využívame komplexný dátový súbor z pražskej Legerovej ulice, ktorý priamo prepája údaje o plynulosti dopravy, meteorologických podmienkach a koncentrácii znečisťujúcich látok v ovzduší.

Štruktúra práce je navrhnutá tak, aby čitateľa plynule previedla od teórie k praktickej časti. Po definovaní technologického rámca Smart Cities nasleduje hĺbková exploratívna analýza dát, ktorá skúma vzájomné závislosti medzi dopravou, počasím a smogovými situáciami. Záverečná časť práce detailne opisuje postup návrhu, tréningu a vyhodnotenia architektúry rekurentnej neurónovej siete typu Gated Recurrent Unit (GRU). Zámerom tejto aplikácie je preukázať, že implementovaný model umelej inteligencie predstavuje spoľahlivý nástroj pre proaktívne riadenie dopravy a zvyšovanie kvality života v meste.

V texte sú použité anglické slová, ktoré sú v oblasti informatiky zaužívané alebo nie sú plne preložené do slovenčiny. Všetky obrázky a tabuľky v tejto práci sú vytvorené autorom, okrem výnimiek, ktoré sú riadne vyznačené.

1 Smart Cities

Život v moderných mestách prináša okrem ekonomických a sociálnych príležitostí aj obrovské výzvy. Ako mestské oblasti rastú, čelia čoraz väčšiemu tlaku na zastaranú infraštruktúru, hustotu dopravy a predovšetkým na životné prostredie. Odpoveďou na tieto naliehavé globálne problémy je prechod k modelu takzvaných inteligentných miest - Smart Cities. Táto úvodná kapitola slúži ako vstupná brána do problematiky, v ktorej nielen zadefinujeme teoretický a technologický rámec inteligentných miest, ale predovšetkým ukážeme, prečo je integrácia dát a technológií nevyhnutná pre udržateľnú budúcnosť.

V nasledujúcich podkapitolách postupne prevedieme čitateľa od základných definícií až po reálne aplikácie v praxi. V podkapitole Čo sú Smart Cities predstavíme samotný koncept Smart Cities a jeho technologické piliere. Vysvetlíme, ako prepojenie fyzickej infraštruktúry, senzorických sietí (IoT) a umelej inteligencie umožňuje mestám prejsť na moderné rozhodovanie založené na dátach.

Následne sa v podkapitole Kľúčové úlohy Smart Cities detailne zameriame na tie najkritickejšie úlohy, ktorým dnešné samosprávy čelia. Rozoberieme úzky a často veľmi nebezpečný vzťah medzi preťaženou mestskou dopravou a zhoršenou kvalitou ovzdušia. Do tohto dynamického systému navyše zasadíme vplyv meteorologických javov, akými sú napríklad teplotné inverzie, ktoré dokážu lokálnu situáciu rapídne zhoršiť.

Aby sme nezostali len pri teoretických konceptoch, v podkapitole Príklady úspešnej integrácie presunieme pozornosť z teórie do reality. Na úspešnom príklade európskej metropoly, ako je poľský Krakov, ukážeme, že využívanie dát o doprave a ovzduší nie je len hudbou budúcnosti. Je to už dnes funkčný nástroj, ktorý vďaka včasnému varovaniu a odvážnym urbanistickým zmenám preukázateľne chráni verejné zdravie a radikálne zvyšuje kvalitu života obyvateľov.

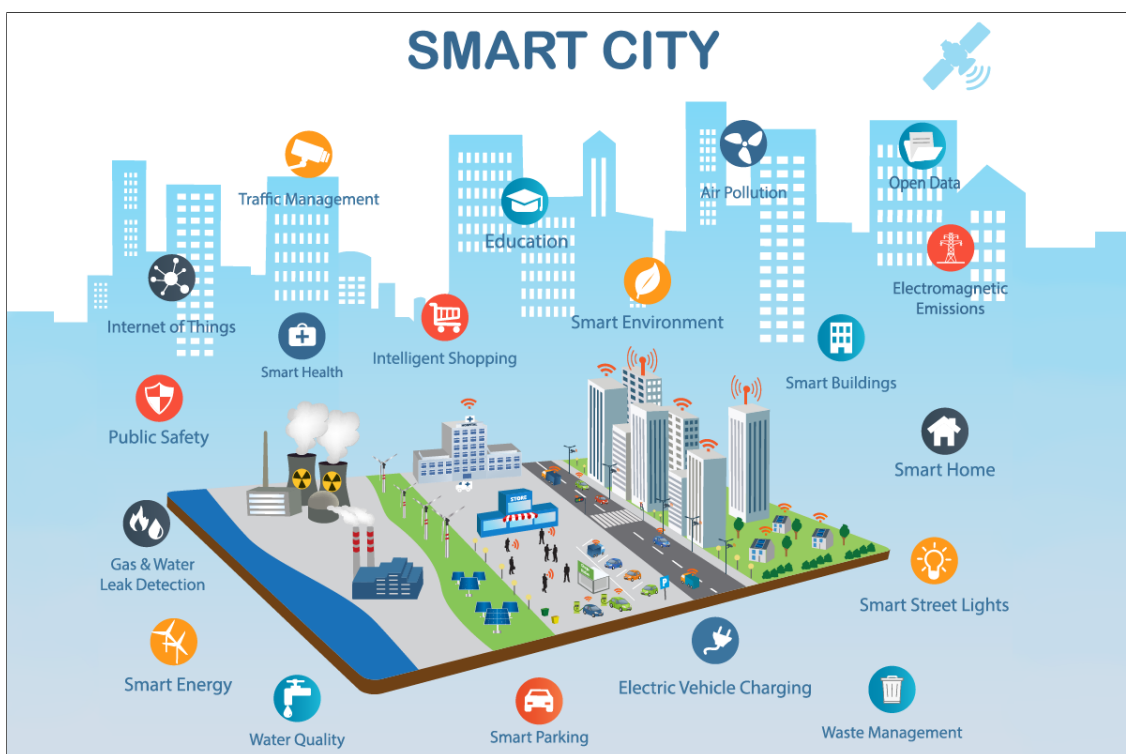
1.1 Čo sú Smart Cities

Koncept inteligentných miest (Smart Cities) predstavuje komplexnú reakciu na rastúcu urbanizáciu a s ňou spojené výzvy v oblasti riadenia zdrojov a infraštruktúry, napríklad riadenie elektrickej energie a dopravnej infraštruktúry.

V súčasnosti neexistuje celosvetovo jednotná definícia tohto pojmu, no medzinárodné inštitúcie aj odborná literatúra sa zhodujú na jeho základných princípoch. Európska komisia definuje inteligentné mesto ako "inteligentné mesto alebo komunita (ktorá) sa zameriava na zlepšenie životných podmienok svojich obyvateľov, podnikov, návštevníkov, organizácií a správcov poskytovaním digitálnych služieb, ktoré prispievajú k lepšej kvalite života" [4]. Na národnej úrovni tento prístup zdôrazňuje aj Ministerstvo investícií, regionálneho rozvoja a informatizácie SR, ktoré koncept vníma ako prepojenie

viacerých spoločenských oblastí - od dopravy a energetiky až po životné prostredie [5].

Z akademického hľadiska je kľúčovou a najčastejšie citovanou prácou model, ktorý predstavil Giffinger a kol. [6]. Ten inteligentné mesto kategorizuje do šiestich hlavných kategórií: inteligentná ekonomika (Smart Economy), inteligentní ľudia (Smart People), inteligentná správa vecí verejných (Smart Governance), inteligentná mobilita (Smart Mobility), inteligentné životné prostredie (Smart Environment) a inteligentný spôsob života (Smart Living). V kontexte tejto práce sú pre nás zaujímavé predovšetkým oblasti Smart Mobility a Smart Environment, v rámci ktorých sa analyzujú dopady dopravy a meteorologických javov na znečistenie ovzdušia.



Obrázok 1: Infraštruktúra Smart City (Zdroj: [1])

Na obrázku 1 je zobrazený príklad infraštruktúry Smart City. Pre naplnenie vízie inteligentného mesta a prechod na proaktívne riadenie sú nevyhnutné technologické inovácie. Caragliu a kol. vo svojej práci zdôrazňujú, že investície do tradičnej aj modernej informačno-komunikačno technologickej (IKT) infraštruktúry sú predpokladom pre udržateľný rast a vysokú kvalitu života [7]. Samotný technologický rámec mesta stojí predovšetkým na týchto pilieroch [8]:

- Internet vecí: Rozsiahle senzorické siete nasadené v mestskom prostredí, ktoré v reálnom čase agregujú údaje o priepustnosti dopravy či parametroch kvality ovzdušia.
- Spracovanie veľkých dát: Efektívna integrácia kvantitatívnych dát do databázových

štruktúr s cieľom odhaľovať skryté korelácie (napríklad vplyv zrážok a teploty na plynulosť premávky).

- Strojové učenie a umelá inteligencia: Analytické metódy schopné predikovať vývoj budúcich krízových stavov na základe historických dát.

Prepojením fyzickej infraštruktúry s digitálnym svetom vzniká ekosystém, ktorý umožňuje rozhodovanie založené na dátach. V reálnom svete to znamená, že pomocou nazbieraných dát zo senzorov (fyzická infraštruktúra) s použitím umelej inteligencie (digitálny svet) sa budeme môcť lepšie rozhodovať pri riešení problémov. Vďaka tomuto prístupu dokážu samosprávy dostatočne presne modelovať riziká a pružne reagovať na zmeny, čím naplňajú samotnú podstatu konceptu Smart City.

1.2 Kľúčové úlohy Smart Cities

Hoci koncept inteligentného mesta zasahuje do množstva sfér spoločenského života, jeho primárnym cieľom je zabezpečiť udržateľný rozvoj a vysokú kvalitu života v husto obývaných mestských oblastiach. Na dosiahnutie tohto cieľa sa moderné samosprávy zameriavajú na riešenie kritických výziev, medzi ktoré patrí predovšetkým optimalizácia mestskej mobility a ochrana životného prostredia. Tieto dve oblasti sú vo výraznej miere vzájomne prepojené a spoločne s meteorologickými faktormi vytvárajú komplexný dynamický systém, ktorého riadenie patrí k najdôležitejším úlohám Smart Cities.

1.2.1 Mestská doprava

Pre správne pochopenie a následné modelovanie mestskej mobility je nevyhnutné vychádzať z definícií dopravného inžinierstva. Základnou skúmanou jednotkou je intenzita dopravy, ktorá je definovaná ako počet cestných vozidiel alebo chodcov, ktorý prejde určitým priečnym rezom pozemnej komunikácie alebo jej časti za zvolené časové obdobie. V rámci tejto práce analyzujeme hodinovú intenzitu vozidiel [9].

Ďalším parametrom je priepustnosť, alebo kapacita komunikácie. Ide o maximálny počet vozidiel, ktoré môžu daným úsekom prejsť za ideálnych podmienok. Maximálny počet vozidiel, ktoré môžu prejsť za jednotku času priečnym rezom jazdného pruhu alebo jazdného pásu cesty jedným smerom pri vopred definovaných podmienkach [9].

Nasledujúci výraz, ktorý budeme používať je dopravný prúd, ktorý je zadefinovaný ako sled všetkých vozidiel, cyklistov alebo chodcov pohybujúcich sa v pruhu za sebou alebo v pruhoch vedľa seba tým istým dopravným smerom [9]. V tejto práci sa nebudeme zaoberať cyklistami alebo chodcami.

S narastajúcim stupňom urbanizácie a zvyšujúcim sa počtom individuálnej motorovej dopravy však dochádza k preťaženiu komunikačnej siete. Neefektívna doprava a pravidelné dopravné zápchy majú podľa Európskej environmentálnej agentúry

(EEA) priamy negatívny vplyv na kvalitu života obyvateľov, pričom z ekonomického a sociálneho hľadiska spôsobujú stratu času, zvyšujú mieru stresu a generujú nadmerný hluk [10].

Najzávažnejšou externalitou mestskej dopravy je však jej priamy dopad na kvalitu ovzdušia. Spaľovacie motory sú primárnym zdrojom emisií priamo v dýchacej zóne obyvateľov. V režime plynulej premávky sú emisie vozidiel optimalizované, avšak v situáciách s častým zastavovaním a rozbiehaním (tzv. stop-and-go efekt v kolónach) dochádza k ich exponenciálnemu nárastu [10]. Úlohou inteligentného mesta je preto nasadzovať systémy inteligentného riadenia dopravy (Intelligent Transport Systems - ITS), ktoré dokážu pomocou dátových analýz plynulosť premávky optimalizovať.

1.2.2 Kvalita ovzdušia

Podľa odbornej literatúry vieme definovať znečistenie ovzdušia ako stav, ktorý nastal v určitom čase a na určitom mieste ako dôsledok súhrnu celého komplexu procesov, na základe prítomnosti zdroja znečisťujúcich látok. V meste to môžu byť napríklad doprava alebo spaľovanie fosílnych palív [11].

Čisté ovzdušie je základným predpokladom pre zdravý život v meste. Znečistenie ovzdušia predstavuje v súčasnosti jedno z najväčších environmentálnych rizík pre verejné zdravie na globálnej úrovni. Dlhodobá expozícia znečistenému mestskému ovzdušiu má preukázateľne negatívny vplyv na ľudský organizmus. Odborné štúdie spájajú zhoršenú kvalitu ovzdušia so zvýšeným výskytom kardiovaskulárnych ochorení a chronických respiračných problémov. Medzinárodná agentúra pre výskum rakoviny (International Agency for Research on Cancer - IARC), ktorá je súčasťou World Health Organization (WHO), navyše oficiálne klasifikovala znečistenie vonkajšieho ovzdušia ako preukázateľne karcinogénne pre ľudí (Skupina 1), pričom najužšie prepojenie bolo dokázané s rakovinou pľúc [12].

Globálnou autoritou, ktorá na základe vedeckých a medicínskych dôkazov určuje bezpečné limity pre kvalitu ovzdušia, je WHO. Jej aktualizované usmernenia o kvalite ovzdušia z roku 2021 slúžia ako referenčný rámec pre národné legislatívy s cieľom minimalizovať zdravotné riziká [13]. Pre účely monitorovania a modelovania kvality ovzdušia sa v praxi sledujú viaceré kľúčové metriky a polutanty [13]:

- **Tuhé prachové častice (Particle matters - PM10 a PM2.5):** Drobné častice (s priemerom menším ako 10, resp. 2,5 mikrometra), ktoré vznikajú zo spaľovacích procesov, ale aj z oteru pneumatík o vozovku a trenie brzdového obloženia. Sú obzvlášť nebezpečné, pretože dokážu prenikať hlboko do dýchacích ciest a až do krvného obehu.
- **Oxidy dusíka (NOx):** Patria sem najmä oxid dusnatý (NO) a oxid dusičitý (NO₂). Ich dominantným zdrojom v mestách je doprava, konkrétne vozidlá s dieselovými

motormi.

- **Prchavé organické zlúčeniny (Volatile Organic Compounds - VOC):** Široká skupina chemických látok, ktoré sa vyparujú do ovzdušia pri izbovej teplote. Patria sem napríklad zložky benzínu, oxidované zlúčeniny z výfukových plynov, ale aj organické chemikálie, ktoré sa dávajú ako prísady do výrobkov pre domácnosť - nátery, čistiace prostriedky, prostriedky na riad, ... [11].
- **Oxid uhoľnatý (CO) a oxid uhličitý (CO₂):** Oxid uhoľnatý je jedovatý plyn vznikajúci pri nedokonalom spaľovaní a takisto sa podieľa na tvorbe fotochemického smogu. Oxid uhličitý v mestskom prostredí slúži ako dôležitý indikátor intenzity spaľovacích procesov a hustoty dopravy.

Termín fotochemický smog vieme pomenovať dve základné kategórie mestského smogu [11]:

- **Zimný (redukčný, resp. Londýnsky smog) smog:** Vzniká prevažne v hmlistých dňoch, často pri teplotných inverziách. Je tvorený zmesou oxidu siričitého a tuhých prachových častíc (PM₁₀, PM_{2.5}), ktorých zdrojom je primárne lokálne vykurovanie a výfukové plyny.
- **Letný (Losangeleský, oxidačný, fotochemický) smog:** Vzniká za dňoch intenzívneho slnečného svitu, kedy pôsobí UV žiarenie na exhaláty zo spaľovacích motorov, najmä NO_x a CO .

1.2.3 Vplyv počasia

Pri komplexnej analýze mestského prostredia nie je možné izolovať dopravu a znečistenie od meteorologických podmienok. Počasie funguje v tomto systéme ako mimoriadne silný modifikátor, ktorý priamo ovplyvňuje plynulosť dopravy aj koncentráciu škodlivín [11]:

- **Rýchlosť a smer vetra:** Rýchlosť (meraná v metroch za sekundu, m/s) a vektor smeru vetra priamo determinujú horizontálny rozptyl polutantov. Kým silný vietor atmosféru čistí, bezvetrie je primárnym katalyzátorom pre vznik smogových epizód.
- **Teplota vzduchu:** Teplota (°C) a jej vertikálny profil v atmosfére definujú termickú konvekciu - schopnosť ohriateho vzduchu stúpať a prirodzene odvetrávať znečistenie.
- **Úhrn zrážok:** Úhrn zrážok (meraný v milimetroch, mm) zabezpečuje proces takzvanej mokrej depozície (vymývania atmosféry), kedy sa jemné prachové častice nabaľujú na kvapky vody a padajú na zemský povrch.

Meteorologické javy ovplyvňujú správanie dopravného prúdu. Zrážky alebo námraza významne znižujú priepustnosť ciest, predlžujú brzdné dráhy a znižujú priemernú rýchlosť vozidiel. Tým priamo prispievajú k vzniku zápch a následnému nárastu emisií [11].

A takisto meteorologické javy rozhodujú o tom, či sa vyprodukované emisie rozptýlia, alebo zostanú koncentrované v uliciach. Kľúčovú rolu tu zohráva rýchlosť vetra a teplotné zvrstvenie atmosféry. Najnebezpečnejším javom pre mestské ovzdušie je teplotná inverzia - meteorologická situácia, kedy je stabilná vrstva teplého vzduchu usadená nad chladnejším vzduchom pri zemi. Týmto sa znemožňuje vertikálny rozptyl škodlivín, čo vedie k ich zrýchlenému hromadeniu a vzniku nebezpečných smogových situácií [11]. Z pohľadu konceptu Smart City je preto kľúčové integrovať dopravné dáta s meteorologickými údajmi pre presnú predikciu týchto hrozieb.

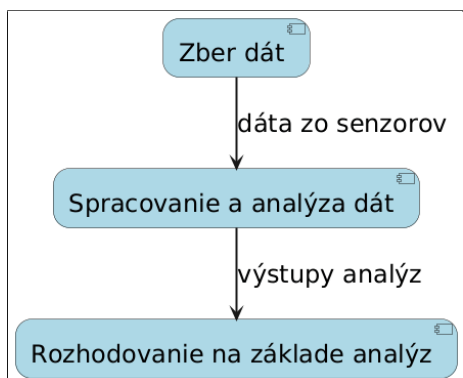
1.3 Príklady úspešnej integrácie

Implementácia konceptov Smart City v praxi prináša viditeľné a hmatateľné výsledky v podobe lepšej informovanosti obyvateľov a efektívnejšieho riadenia krízových situácií. Mnohé európske metropoly už dnes využívajú prepojenie senzorických sietí s verejnou infraštruktúrou, čím demonštrujú praktický význam prediktívneho modelovania dopravy a kvality ovzdušia.

My sme si vybrali jednu z mnohých úspešných implementácií, konkrétne mesto Krakov v Poľsku.

1.3.1 Mesto Krakov

Mesto Krakov predstavuje jeden z najprogresívnejších príkladov integrácie Smart City technológií v regióne strednej Európy. Strategický prístup mesta sa nesústreďuje len na izolované technické riešenia, ale je postavený na komplexnom programe ochrany ovzdušia, ktorý využíva dáta ako primárny nástroj na podporu rozhodovania samosprávy aj obyvateľov.

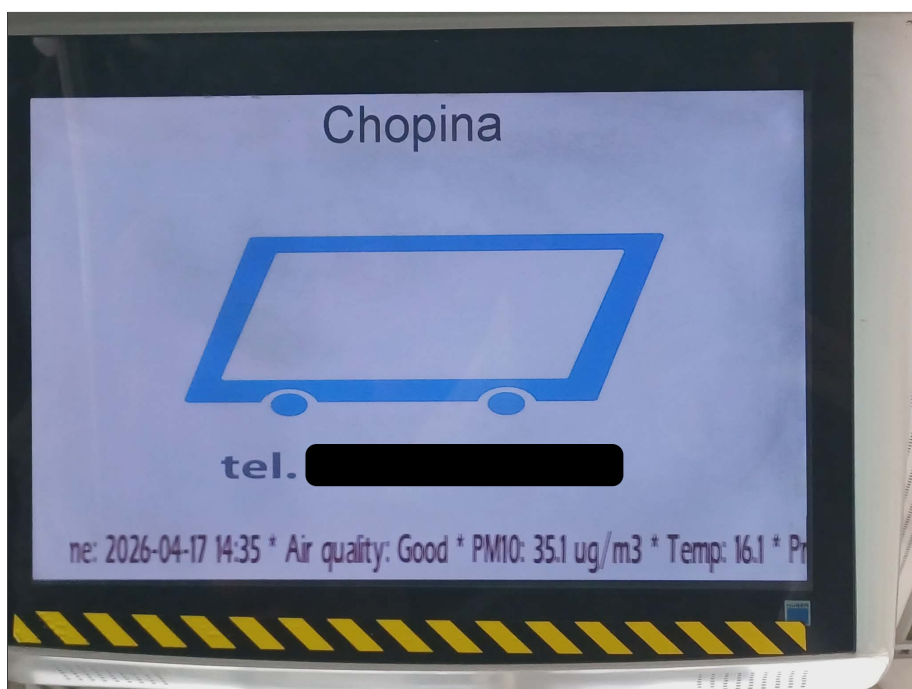


Obrázok 2: Diagram fáz konceptu inteligentného monitorovania mesta Krakov

Celkový koncept inteligentného monitorovania v Krakove funguje v uzatvorenom cykle, ktorý možno rozdeliť do troch fáz, ako to vidíme na obrázku 2:

- **Zber dát:** Mesto disponuje jednou z najhustejších sietí monitorovacích staníc, ktoré v reálnom čase merajú koncentrácie prachových častíc PM10 a PM2.5 [14].
- **Analýza a spracovanie dát:** Zozbierané dáta sú integrované do centrálného systému, kde sa korelujú s dopravnými prúdmi a meteorologickými údajmi. Tieto analýzy poskytujú mestskému dispečerom znalosti o aktuálnom vývoji smogových situácií.
- **Zber údajov:** Výsledné znalosti sa transformujú do konkrétnych opatrení, ako je bezplatná mestská hromadná doprava počas kritických dní s cieľom viacej využívať hromadnú dopravu namiesto osobného auta alebo sprísnenie kontroly v nízkoemisných zónach.

Práve na základe dlhodobého zberu a analýzy dát o zložení dopravného prúdu a jeho vplyve na emisie bolo v Krakove rozhodnuté o zavedení nízkoemisných zón (po poľsky: Strefy Czystego Transportu - SCT) [14]. Tieto nízkoemisné zóny zakazujú vstup dieselovým vozidlám, ktoré boli vyrobené pred rokom 2010. Toto strategické rozhodnutie je jasným príkladom toho, ako spracovanie rozsiahlych údajov vedie k exaktným urbanistickým zásahom chrániacim verejné zdravie obyvateľov.



Obrázok 3: Zobrazenie kvality ovzdušia pomocou koncentrácie prachových častíc PM10 v dopravnom prostriedku na zobrazovacej jednotke. Osobné údaje boli cenzúrované. Zdroj: Mozolíková, K. (2026)

Dôležitou súčasťou stratégie je vizualizácia dát priamo vo verejnom priestore. Informovanosť občanov je zabezpečená prostredníctvom vizuálnych informačných tabúl umiestnených na hlavných dopravných uzloch a priamo v prostriedkoch verejnej dopravy. Na obrázku 3 je zobrazené, že kvalita ovzdušia je dobrá (anglicky Good), hodnota PM10 je $35,1 \mu g/m^3$ a teplota 16,1 (pravdepodobne °C).

Tento detailný pohľad na informačnú jednotku v dopravnom prostriedku potvrdzuje, že krakovský model mení dáta na okamžité využiteľnú informáciu. Obyvatel' získava znalosť o kvalite ovzdušia v reálnom čase, čo mu umožňuje urobiť vedomé rozhodnutie o obmedzení pohybu vonku alebo o voľbe alternatívneho spôsobu dopravy. Týmto sa naplňa hlavná úloha Smart City - využitie technológií na zvýšenie kvality života prostredníctvom dátami podložených rozhodnutí.

2 Kritériá kvality

V tejto kapitole sa venujeme definovaniu kritérií kvality pre hodnotenie predikčných modelov a ich praktické nasadenie v meste. Najskôr predstavíme doménové požiadavky, ktoré musia modely spĺňať pre reálne využitie v prostredí Smart City. Následne popíšeme kritériá kvality života, ktoré hodnotia, ako nasadené systémy prispievajú k udržiavaniu bezpečného a zdravého prostredia pre obyvateľov. Na záver zadefinujeme vyhodnocovacie metriky modelov.

2.1 Doménové kritériá

Špičkové výsledky technických metrík samy osebe nezaručujú, že bude model v reálnom prostredí Smart City použiteľný. Pre praktické nasadenie v meste musia vytrénované modely spĺňať aj špecifické doménové požiadavky. Zvolili sme si časový predstih predikcie, výpočtová efektivita a časová odozva a nasaditeľnosť a reprodukovateľnosť.

2.1.1 Časový predstih predikcie

Základným zmyslom integrácie umelej inteligencie do riadenia mesta je proaktivita. Predikovať vznikajúcu dopravnú zápchu alebo nebezpečnú smogovú inverziu minútu pred jej reálnym vznikom je pre potreby mesta bezpredmetné, keďže samospráva nestihne vykonať žiadne preventívne opatrenia.

Kľúčovým kritériom je preto časový predstih predikcie. Pre potreby tejto práce sme stanovili cieľ vytrénovať model tak, aby na základe historických dát z predchádzajúcich 24 hodín dokázal s vysokou presnosťou predikovať rizikovú situáciu s časovým predstihom minimálne 1 hodinu vopred. Takýto časový rámec poskytuje mestským systémom dostatok času na reakciu, napríklad na úpravu intervalov na semaforochoch, zníženie povolenej rýchlosti na obchvatoch alebo včasné varovanie občanov.

2.1.2 Výpočtová efektivita a časová odozva

V kontexte ITS a riadenia dopravy v reálnom čase je kritickým doménovým parametrom nielen presnosť modelu, ale aj rýchlosť, akou dokáže systém vygenerovať novú predikciu po prijatí aktuálnych senzorických dát. Skrátka chceme mať výpočet čo najrýchlejší. Kritérium výpočtovej efektivity hodnotí náročnosť modelu pri takzvanej inferencii (produkčnej prevádzke).

2.1.3 Nasaditeľnosť a reprodukovateľnosť

Pre praktické využitie predikčných modelov v doméne dopravného inžinierstva je nevyhnutné, aby bol proces predikcie izolovaný od procesu tréningu. Kritérium nasaditeľnosti hodnotí, či je možné navrhnutú architektúru po jej optimalizácii trvalo uložiť (vrátane všetkých normalizačných parametrov a váh neurónovej siete) a následne ju nasadiť do produkcie.

Splnenie tohto kritéria garantuje reprodukovateľnosť výsledkov na nových, nezávislých dátových sadách bez nutnosti opätovného tréningu siete, čo je kľúčový princíp pre moderné inžinierstvo strojového učenia [15].

2.2 Kritériá kvality života

V kontexte inteligentných miest a podpory rozhodovania sa úspešnosť riadenia nehodnotí výlučne technickými parametrami použitých informačných systémov, ale predovšetkým tým, ako tieto systémy prispievajú k udržaniu bezpečného a zdravého prostredia pre obyvateľov. Aby sme vedeli analyzovať vznikajúce riziká a vyhodnotiť efektivitu nášho predikčného modelu, je nevyhnutné exaktne zadať hranice, kedy je kvalita služby a života v meste na dobrej úrovni a kedy sa, naopak, stáva kritickou.

Tieto doménové kritériá kvality života sme rozdelili do troch vzájomne prepojených oblastí: plynulosť dopravy, kvalita ovzdušia a vplyv meteorologických modifikátorov.

2.2.1 Plynulosť a intenzita dopravy

Kvalita života v oblasti mestskej mobility sa neodvíja len od samotného počtu vozidiel (intenzity), ale od plynulosti dopravného toku. Podľa štandardov dopravného inžinierstva rozlíšujeme rôzne stupne kvality dopravy [9].

- **Plynulá premávka:** Nastáva vtedy, keď sa intenzita dopravy pohybuje hlboko pod maximálnou kapacitou komunikácie. Vozidlá si udržiavajú priemernú rýchlosť blízku maximálnej povolenej rýchlosti (napríklad nad 40 km/h v obci). Z hľadiska emisií ide o optimálny stav, keďže spaľovacie motory pracujú plynulo.
- **Dopravná zápcha:** Nastáva v momente, keď hodinová intenzita prekročí priepustnosť úseku. Rýchlosť vozidiel rapídne klesá (často pod 15 km/h) a vzniká nežiaduci stop-and-go efekt. Tento stav nielenže znižuje komfort cestovania, ale vedie k exponenciálnemu nárastu výfukových emisií [10].

2.2.2 Limity znečistenia ovzdušia

Pre definovanie kritérií kvality života v tejto oblasti vychádzame z prísnych limitov stanovených WHO, ktoré určujú maximálne bezpečné denné koncentrácie jednotlivých

polutantov [13].

- **Prachové častice PM10 a PM2.5:** Tieto častice predstavujú najväčšie riziko, pretože prenikajú hlboko do dýchacej sústavy. Kvalita života sa považuje za vysokú (kvalitné ovzdušie), ak 24-hodinový priemer neprekročí $45 \mu\text{g}/\text{m}^3$ pre PM10 a $15 \mu\text{g}/\text{m}^3$ pre PM2.5 [13]. Akékoľvek hodnoty nad touto hranicou indikujú zhoršenú kvalitu ovzdušia, pričom extrémne hodnoty (napríklad presahujúce $100 \mu\text{g}/\text{m}^3$) už predstavujú závažnú smogovú epizódu [11].
- **Oxid uhoľnatý (CO):** Tento limit je stanovený v odlišných jednotkách - miligram na meter kubický (mg/m^3). Bezpečný 24-hodinový limit je stanovený na $4 \text{ mg}/\text{m}^3$ [13], alebo $4000 \mu\text{g}/\text{m}^3$.
- **Oxid dusičitý (NO2):** Keďže jeho hlavným zdrojom v mestách je doprava, jeho hodnoty sú silným indikátorom dopravného zaťaženia. Bezpečný 24-hodinový limit je stanovený na $25 \mu\text{g}/\text{m}^3$ [13].

2.2.3 Meteorologické podmienky

Okrem extrémneho počasia, meteorologické faktory samy o sebe nezhoršujú kvalitu života, ale fungujú ako silné modifikátory, ktoré rozhodujú o tom, aký bude mať vplyv dopravy dopad na kvalitu ovzdušia.

- **Rýchlosť vetra:** Pre zachovanie dobrej kvality ovzdušia je nutná dostatočná rýchlosť vetra (zvyčajne nad 3 m/s), ktorá zabezpečuje horizontálny rozptyl emisií [16]. Naopak, rýchlosť vetra blízka nule (bezvetrie) predstavuje z hľadiska kvality života kritický stav. Pri absencii vetra sa všetky emisie z dopravy kumulujú priamo na úrovni ulice.
- **Teplotné inverzie:** Hoci teplotu samotnú nevieme ovplyvniť, jej vývoj (najmä prudké ochladenia kombinované s bezvetrím v zimných mesiacoch) je pre naše modely hlavným indikátorom pre vznik zimného smogu, kedy sa znečistenie uzatvára pri zemskom povrchu [11].

2.3 Vyhodnocovacie metriky modelu

Keďže predikcia znečistenia ovzdušia alebo dopravného zdržania predstavuje regresnú úlohu (model predikuje spojitú číselnú hodnotu, nie kategóriu), na vyhodnotenie presnosti modelov strojového učenia využijeme štandardné regresné metriky.

Zvolili sme si dve metriky: Root Mean Square Error (RMSE) a Mean Absolute Error (MAE).

2.3.1 Metrika Root Mean Square Error (RMSE)

Odmocnina zo strednej kvadratickej chyby (Root Mean Square Error - RMSE) je jedna z najpoužívanejších metrík na hodnotenie kvality regresných modelov. Metrika kvantifikuje štandardnú odchýlku chýb predikcie. Matematicky je definovaná vzorcom (1):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (1)$$

kde n je celkový počet pozorovaní, y_i je skutočná nameraná hodnota i -teho pozorovania a $\hat{y}_i$ je hodnota i -teho pozorovania predikovaná modelom.

V kontexte predikcie extrémnych javov (smogové situácie, rozsiahle dopravné kolapsy) je táto metrika kľúčová. Tým, že sa rozdiel medzi skutočnosťou a predikciou pred samotným spriemerovaním umocňuje na druhú, metrika RMSE exponenciálne penalizuje väčšie odchýlky [17].

Ak model pri bežnej premávke urobí chybu 2 minúty, metrika to zohľadní mierne. Ak by však model pri predikcii uštelil a predikoval by zdržanie o 40 minút kratšie, než je realita, RMSE túto hrubú chybu okamžite a výrazne prejaví na celkovom skóre modelu.

2.3.2 Metrika Mean Absolute Error (MAE)

Stredná absolútna chyba (Mean Absolute Error - MAE) slúži ako dôležitý doplnok k RMSE. Na rozdiel od RMSE nerieši kvadratickú penalizáciu extrémov, ale poskytuje priemer absolútnych hodnôt chýb. MAE sa vypočíta podľa vzorca (2):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

kde n je celkový počet pozorovaní, y_i je skutočná nameraná hodnota i -teho pozorovania a $\hat{y}_i$ je hodnota i -teho pozorovania predikovaná modelom.

Hlavnou výhodou tejto metriky je jej vysoká miera interpretovateľnosti pre koncového používateľa (napríklad zamestnanca krízového riadenia mesta) [17]. MAE totiž vyjadruje priemernú chybu predikcie priamo v pôvodných reálnych jednotkách.

Ak model predikuje úroveň znečistenia PM2.5, pričom bol natrénovaný v jednotkách mikrogram na meter kubický ($\mu\text{g}/\text{m}^3$) a dosiahne skóre MAE hodnoty 5, znamená to, že sa model v priemere mýli o $5 \mu\text{g}/\text{m}^3$.

3 Neurónové siete

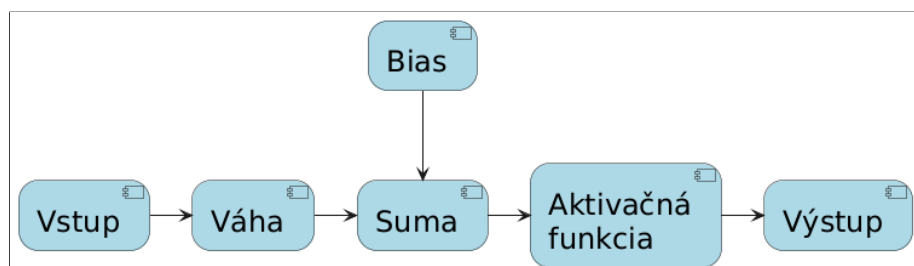
Schopnosť umelých neurónových sietí učiť sa a zovšeobecňovať vedomosti na nové dáta je v jadre založená na matematickej optimalizácii. Celý tento proces spočíva v hľadaní takých parametrov, ktoré čo najviac minimalizujú chybu modelu pri predikcii. Táto kapitola postupne predstaví kľúčové stavebné prvky a architektúry, na ktorých je toto učenie postavené.

V úvode sa zameriame na samotný mechanizmus trénovania. Oboznámime sa s algoritmami gradientného zostupu a spätného šírenia chyby, prostredníctvom ktorých si model iteratívne upravuje svoje nastavenia. Následne si priblížime rolu aktivačných funkcií, ktoré do výpočtov zavádzajú nevyhnutnú nelinearitu, aby bola sieť schopná riešiť aj zložité úlohy.

V druhej časti kapitoly prejdeme od základov k spracovaniu sekvenčných dát prichádzajúcich v chronologickom slede. Popíšeme si rekurentné neurónové siete (RNN), ktoré na rozdiel od štandardných architektúr disponujú vnútornou pamäťou pre udržiavanie kontextu. Záver kapitoly bude venovaný pokročilej hradlovej architektúre Gated Recurrent Unit (GRU), pri ktorej si rozoberieme, ako pomocou vnútorných brán efektívne rieši limitácie tradičných rekurentných sietí.

3.1 Učenie sa neurónových sietí

Schopnosť neurónových sietí učiť sa, extrahovať komplexné nelineárne príznaky a zovšeobecňovať z trénovacích dát na dosiaľ nevidené vzorky je založená na matematickej optimalizácii [18]. Učenie sa umelých neurónových sietí nie je ničím iným ako hľadaním takých hodnôt voľných parametrov (synaptických váh a prahov), ktoré minimalizujú rozdiel medzi skutočným a predikovaným výstupom systému pre danú množinu vstupných dát [18].



Obrázok 4: Ukážka zjednodušenej neurónovej siete

Proces učenia sa s učiteľom možno rozdeliť na dve fázy: dopredný prechod (forward pass) a spätný prechod (backward pass) spojený s aktualizáciou parametrov [19]. Na obrázku 4 v doprednom prechode vstup prechádza cez vrstvy siete, pričom v každom neuróne dochádza k výpočtu lineárnej kombinácie vstupov a váhových koeficientov,

ku ktorej sa pripočíta prahová hodnota (bias). Táto suma je následne transformovaná nelineárnou aktivačnou funkciou a stáva sa vstupom pre nasledujúcu vrstvu [18]. Vzťah je vyjadrený rovnicou (3):

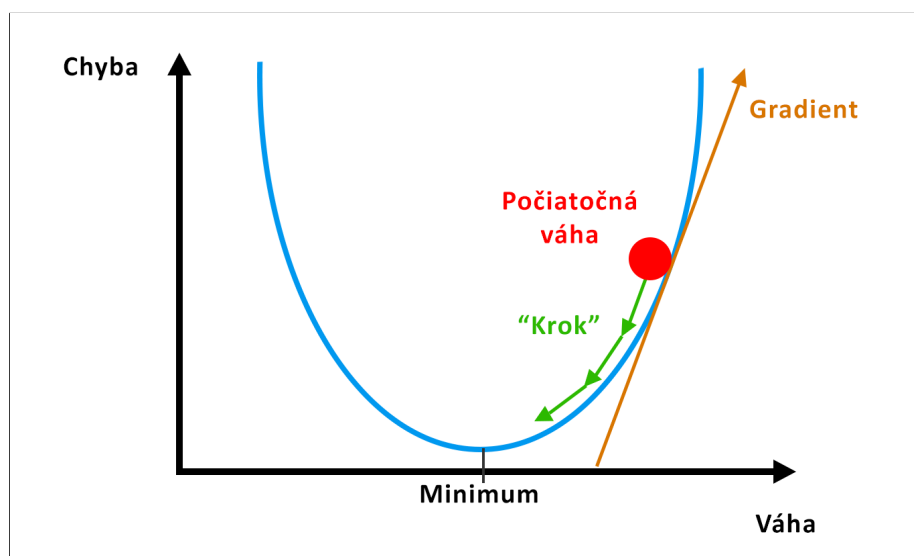
$$y = f \left(\sum_{i=1}^n w_i x_i + b \right) \quad (3)$$

kde y je výstup, f je aktivačná funkcia, n je počet vstupov do neurónu, w_i je váha pre i -te spojenie, x_i je vstup pre i -te spojenie a b je bias.

Na konci tohto reťazca sieť vygeneruje výstup, ktorý sa porovnáva s požadovaným výstupom pomocou stratovej funkcie, napríklad strednej kvadratickej chyby (Mean Squared Error - MSE) pre regresné úlohy [18].

3.1.1 Gradientný zostup a spätné šírenie chyby

Keď neurónová sieť vypočíta výstup a následne svoju chybu pomocou stratovej funkcie, začína sa proces učenia. Jej cieľom je nastaviť svoje parametre (váhy a biasy) tak, aby bola táto chyba čo najmenšia [18]. Na to slúži algoritmus gradientného zostupu. Gradient matematicky ukazuje, ktorým smerom chyba rastie najrýchlejšie. Keďže my chceme chybu naopak minimalizovať, upravujeme parametre presne v opačnom smere, než ukazuje gradient [20].



Obrázok 5: Gradientný zostup

Pri týchto úpravách hrá kľúčovú úlohu rýchlosť učenia sa (learning rate). Je to nastavenie, ktoré určuje veľkosť "kroku" pri hľadaní minima chyby [20]. Ak je tento krok príliš malý, učenie je neefektívne a trvá extrémne dlho. Ak je naopak príliš veľký, sieť môže minimum preskočiť a nikdy sa nenaučí to, čo má [21].

Na to, aby sa tieto zmeny dali vôbec spočítať vo veľkých sieťach s viacerými vrstvami, sa používa algoritmus spätného šírenia chyby. Tento algoritmus aplikuje

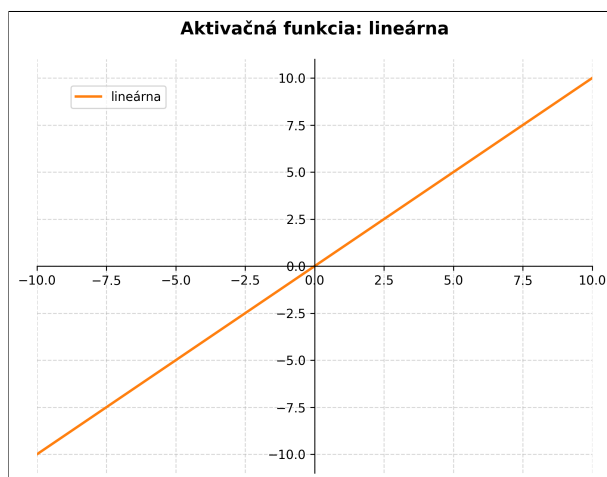
matematické reťazové pravidlo priamo na štruktúru neurónovej siete [18].

V praxi to funguje tak, že vypočítaná chyba sa prenáša od konca siete (výstupu) krok za krokom späť na začiatok siete (vstup). Každá vrstva prevezme informáciu o chybe z nasledujúcej vrstvy, upraví ju podľa svojich vlastností a pošle ju o vrstvu nižšie [19]. Tento prístup je veľmi užitočný, pretože rozdelí jeden obrovský a zložitý problém na sériu menších a matematicky jednoduchých krokov [18].

3.2 Aktivačné funkcie

Každý neurón v modeli transformuje lineárnu kombináciu svojich vstupov pomocou takzvanej aktivačnej funkcie. V kontexte hlbokého učenia a obzvlášť modelovania sekvencií zohráva výber správnej aktivačnej funkcie ôložitú rolu. Ak by neurónová sieť obsahovala len identické, lineárne transformácie, bez ohľadu na počet skrytých vrstiev, tak by bola v konečnom dôsledku celá jej architektúra matematicky ekvivalentná jedinej lineárnej vrstvy [18]. Takýto systém by nebol schopný aproximovať komplexné nelineárne ohraničenia rozhodovacích tried, ani extrahovať abstraktné javy z dát [22].

Aktivačné funkcie zavádzajú do výpočtového grafu nevyhnutnú nelinearitu. Z matematického hľadiska musia spĺňať podmienku diferencovateľnosti (alebo aspoň spojitú diferencovateľnosť po častiach), nakoľko sa ich prvá derivácia priamo zúčastňuje vo výpočtoch reťazového pravidla počas algoritmu spätného šírenia chyby [18].



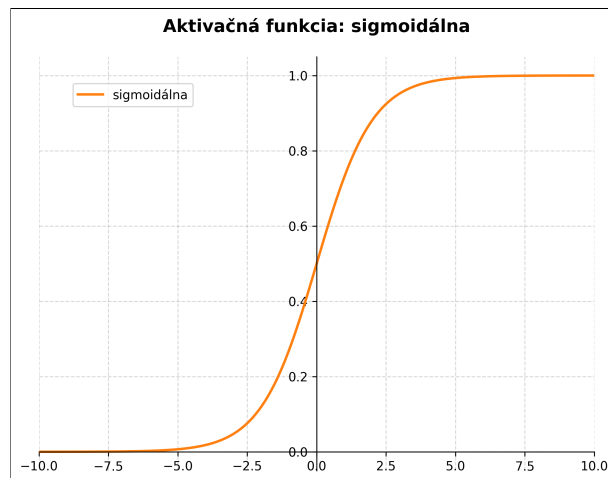
Obrázok 6: Lineárna aktivačná funkcia

Transformácia lineárnej aktivačnej funkcie je priamo úmerná vstupu, ako to vidíme na obrázku 6 vďaka čomu je derivácia konštantná. Lineárna funkcia je vyjadrená rovnicou (4):

$$f(x) = x \quad (4)$$

Ak sa však lineárna funkcia použije v skrytých vrstvách, neurónová sieť prichádza o schopnosť abstrakcie a jej výpočtová sila sa obmedzuje len na lineárnu regresiu [23].

Z tohto dôvodu sa využíva primárne len vo výstupnej vrstve pre predikciu spojitých hodnôt.

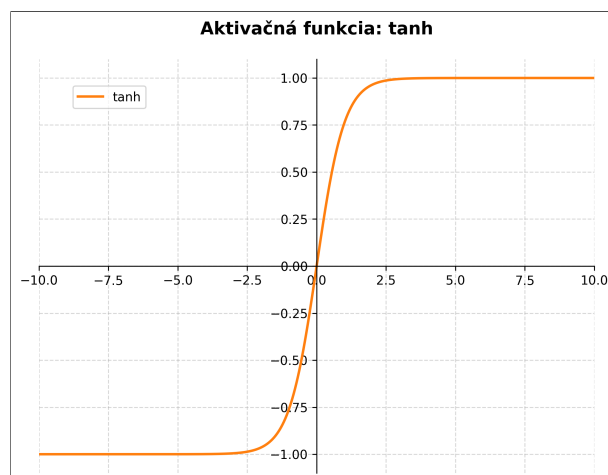


Obrázok 7: Sigmoidálna aktivačná funkcia

Sigmoidálna, alebo logistická funkcia, označovaná σ predstavuje hladkú a zhora aj zdola obmedzenú nelineárnu transformáciu. Sigmoidálna funkcia je zobrazená na obrázku 7 a je vyjadrená rovnicou (5):

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

Jej hlavným nedostatkom je, že výstup nie je centrovaný okolo nuly, čo spomaľuje konvergenciu počas tréningu. Okrem toho je výrazne náchylná na problém miznúceho gradientu (vanishing gradient). Pre okrajové hodnoty vstupu derivácia prudko klesá k nule, čím dochádza k saturácii a následnému zastaveniu učenia [22].



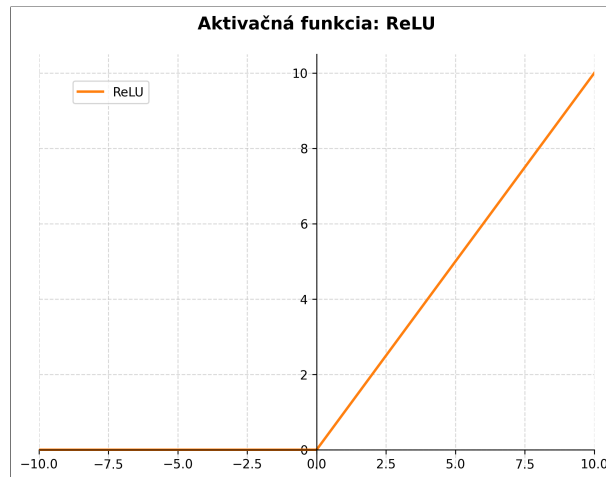
Obrázok 8: Funkcia hyperbolického tangensu

Funkcia hyperbolického tangensu (tanh) čiastočne rieši nevýhody logistickej

funkcie tým, že jej výstup je centrováný do nuly. Je vyjadrená nasledovnou rovnicou (6):

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (6)$$

Vďaka tomu sa váhy pri učení aktualizujú omnoho plynulejšie v porovnaní so sigmoidou [24]. Napriek lepšej konvergencii však stále podlieha extrémnej saturácii pri príliš vysokých alebo nízkych hodnotách vstupu. To opäť vedie k problému miznúcich gradientov, čo spôsobuje ťažkosti predovšetkým v hlbokých sieťach [25].



Obrázok 9: Aktivačná funkcia ReLU

Aktivačná funkcia ReLU (Rectified Linear Unit) je definovaná predpisom (7):

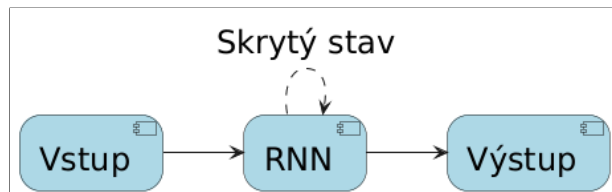
$$\text{ReLU}(x) = \max(0, x) \quad (7)$$

ReLU rieši problém miznúceho gradientu pre kladné hodnoty, čo z nej robí štandard pre moderné hlboké siete [24]. Pri zápornom vstupe je jej derivácia nulová. To spôsobuje problém takzvaných "mŕtvych neurónov" (dying ReLU), ktoré natrvalo prestanú prepúšťať chybový signál, čím sa úplne zastaví ich vlastná aktualizácia a z procesu učenia sa vyradia [24].

V tradičných architektúrach spracovávajúcich časové rady a sekvencie (akými sú RNN) zohrával hyperbolický tangens ($\tanh$) dlhodobu dominantnú rolu. Pri moderných hradlových architektúrach, akými sú GRU a LSTM, sa z dôvodu presnej logiky riadenia priepustnosti informácie vo vnútorných bránach (gates) využíva logistická funkcia. Výstupný interval poskytuje ideálnu aproximáciu pravdepodobnostného otvárania (hodnoty blízko 1) a zatvárania (hodnoty blízko 0) toku dát pamäťovým aparátom [18].

3.3 Rekurentné neurónové siete

Limitácie dopredných neurónových sietí sa prejavujú v situáciách, kde dáta prichádzajú v chronologickom a logickom slede. Kým štandardný perceptrón alebo viacvrstvová sieť spracováva každý predložený vstup len na základe jeho okamžitej charakteristiky, sekvencie (vety, hudba, časové rady trhov) vyžadujú kontext [18]. Ak sa napríklad snažíme predpovedať nasledujúce slovo vo vete, model musí nutne evidovať prítomnosť všetkých predošlých slov [26].



Obrázok 10: Rekurentná neurónová sieť

Rekurentné neurónové siete (Recurrent Neural Network - RNN) elegantne riešia tento problém zavedením vnútornej pamäte, takzvaného skrytého stavu, ktorý nesie neustále aktualizovaný kontext o celej histórii spracovaných časových krokov [18]. Ukážka je na obrázku 10. RNN je teda systém, ktorý reaguje na aktuálny vstup a históriu toho, čo "videl" doteraz [26].

Formálne je štandardná RNN definovaná tak, že spracováva vstupnú sekvenciu premenlivej dĺžky, podľa rovnice (8):

$$h_t = f(h_{t-1}, x_t) \quad (8)$$

kde t je časový krok, h_t je skrytý (hidden) stav v časovom kroku t , f je aktualizácia funkcia, h_{t-1} je skrytý stav predošlého kroku t a x_t je aktuálny vstup kroku t .

V tradičnej architektúre je aktualizácia funkcia realizovaná ako hladká nelineárna transformácia, najčastejšie hyperbolický tangens alebo logistická funkcia, čo je ukázané v rovnici (9):

$$h_t = \tanh(Wx_t + Uh_{t-1} + b) \quad (9)$$

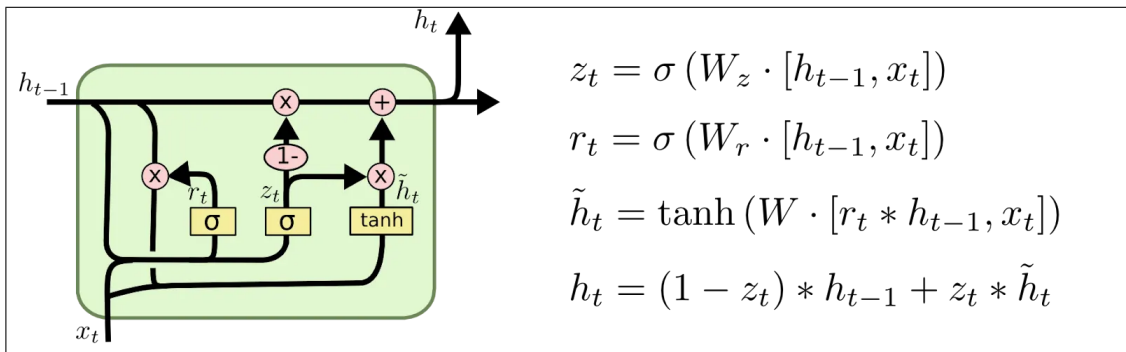
kde t je časový krok, h_t je skrytý stav v časovom kroku t , $\tanh$ je aktualizácia funkcia, W je matica váh pre aktuálny vstup x_t , U je matica váh pre predošlý skrytý stav h_{t-1} a b je bias [18].

Vďaka tomuto odovzdávaniu skrytého stavu naprieč celou sekvenciou dochádza k distribuovanému kódovaniu času. Rekurentná sieť vďaka tomu slúži ako univerzálny nástroj na mapovanie sekvencií a dokáže fungovať ako generatívny model, ktorý predikuje celkovú pravdepodobnostnú distribúciu spracovávanej postupnosti [18].

3.4 Architektúra Gated Recurrent Unit

Gated Recurrent Unit (GRU) bola predstavená pôvodne v rámci modelovania štatistického strojového prekladu v systéme RNN Encoder-Decoder [18]. Táto architektúra poskytuje zjednodušenie zložitosti LSTM buniek s priamym zameraním na elimináciu problému miznúceho gradientu pomocou princípu hradlových jednotiek (gating mechanisms) [2].

Hlavnou zmenou oproti LSTM je odstránenie dedikovaného bunkového stavu (cell state - c_t). Model GRU všetky pamäťové toky - históriu aj aktuálne kontexty - zlúčil a spravuje prostredníctvom jediného toku - skrytého stavu h_t [18]. Prúd informácií z minulosti chráni pred exponenciálnym potlačením dvoch zladených nelineárnych prepínačov: brána resetovania (Reset Gate) a brána aktualizácie (Update Gate) [18]. Tieto brány sa v každom kroku samostatne rozhodujú o tom, aký objem informácií je zbytočný a bude vyradený a aký kontext z minulosti sa odovzdá do výstupu pre budúcnosť [18].



Obrázok 11: Model bunky GRU (Zdroj: [2])

Na obrázku 11 je zobrazený všeobecný model bunky GRU. Pre plné pochopenie schopnosti modelu učiť sa z dlhých signálov je nutné dekonštruovať matematickú štruktúru výpočtového grafu vo vnútri GRU bunky. Prechod z času do času podmienený príchodom nového vstupu prebieha v štyroch naväzujúcich fázach:

Fáza 1: Brána resetovania, vo vzorci (10) označovaná ako r_t , rozhoduje o tom, akú časť minulej pamäte (teda predchádzajúceho skrytého stavu h_{t-1}) by mala sieť pri výpočte novej informácie ignorovať [2]. V praxi to napríklad znamená, že pri spracovaní jazyka alebo zvuku dokáže model "zabudnúť" na kontext predchádzajúcej vety a plne sa sústrediť na začiatok novej myšlienky.

Vektor tejto brány sa počíta pomocou logistickej funkcie, označovanej σ podľa vzorca (10):

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (10)$$

kde predstavujú W_r a U_r matice váh, ktoré sa model učí počas tréningového algoritmu a b_r označuje bias [18]. Keďže sa na výpočet používa logistická funkcia, výstupné hodnoty r_t ležia vždy v intervale od 0 do 1. Tento výstup funguje ako filter: ak brána vygeneruje hodnoty blízke nule, prísun informácií z minulosti sa prakticky preruší

(zresetuje). Naopak, ak sa hodnoty blížia k jednotke, tak minulý kontext prechádza do ďalšej fázy výpočtu v podstate bez zmien [18].

Fáza 2: Výpočet kandidátskeho skrytého stavu, rovnica (11). V tomto kroku sieť generuje takzvanú kandidátsku pamäťovú zložku ($\tilde{h}_t$). Táto zložka predstavuje potenciálny nový obsah, ktorý sa môže zapísať do skrytého stavu. Práve tu sa naplno prejavuje vplyv brány resetovania r_t , ktorá filtruje historický skrytý stav (h_{t-1}) pomocou elementárneho (Hadamardovho) súčinu, označovaného symbolom $\odot$:

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \quad (11)$$

Tento mechanizmus zásadne ovplyvňuje pamäť siete. Ak je hodnota z brány resetovania blízka nule ($r_t \approx 0$), výraz reprezentujúci minulosť kolabuje. Nová kandidátska zložka pamäte $\tilde{h}_t$ sa tak v tomto prípade vygeneruje výhradne na základe aktuálneho bezprostredného vstupu x_t [18].

Na výpočet sa aplikuje nelineárna funkcia $\tanh$, ktorá slúži na normalizáciu hodnôt. Kandidátsky stav sa ňou stlačí do ohraničeného intervalu $(-1, 1)$. Je to kľúčové ochranné opatrenie - predchádza sa tým kumulatívne rastu hodnôt do extrémnych čísel, čím sa zachováva celková numerická stabilita a integrita siete počas dlhších sekvencií [18].

Fáza 3: Brána aktualizácie (označovaná ako z_t) plní kľúčovú úlohu pri udržiavaní dlhodobej pamäte, kým brána resetovania riadi formovanie nového kontextu. Rozhoduje o tom, v akom pomere si sieť bez zmeny ponechá staré informácie z minulosti a do akej miery ich prepíše novým kandidátskym stavom [18].

Výpočet tejto brány prebieha paralelne s predchádzajúcimi krokmi podľa vzorca (12):

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (12)$$

Podobne ako pri resetovacej bráne, aj na tento výpočet sa využíva logistická funkcia. Výsledné hodnoty tohto vektora sa tak opäť nachádzajú v intervale $(0, 1)$.

Fáza 4: Na záver aktuálneho časového kroku sieť realizuje prechod od historického skrytého stavu h_{t-1} k novému finálnemu stavu h_t . Tento proces prebieha prostredníctvom plynulej konvexnej lineárnej interpolácie podľa vzorca (13):

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (13)$$

Ak je signál dlhodobého kontextu pre model mimoriadne dôležitý a potrebuje si ho uchovať, brána aktualizácie vygeneruje hodnotu z_t blížiacu sa k nule [18]. V takom prípade interpolácia prideli koeficient blízky jednotke historickému stavu (výraz $1 - z_t$). Výsledkom je faktické prekopírovanie pôvodnej pamäte h_{t-1} priamo do cieľového vektora h_t , ku ktorému sa pridá len takmer nulová zložka nového kandidátskeho stavu $\tilde{h}_t$.

Tento princíp aditívneho (lineárneho) dopĺňania zmien sa zásadne líši od tradičných

RNN sietí, ktoré svoj skrytý stav neustále kompletne a nelineárne prepisujú. Vďaka sčítavaniu sa v architektúre vytvára takzvaná skratka pre tok učiaceho signálu. Keď sa následne pri spätnej propagácii šíri chybový signál z budúcnosti hlboko do minulosti, gradienty dokážu plynule pretekať práve cez tieto aditívne väzby. Nedochádza tak k ich opakovanému násobeniu, ktoré by ich postupne degradovalo, čím sa model efektívne chráni pred problémom miznúcich gradientov [18].

4 Spracovanie a analýza dát

Úspešné modelovanie pomocou neurónových sietí si vyžaduje kvalitné a správne pripravené vstupné dáta. Táto kapitola popisuje metodiku práce s dátami, od ich prvotného získania až po hĺbkovú analýzu, ktorá predchádza samotnému trénovaniu umelej inteligencie. Namiesto zložitého spájania viacerých datasetov sme pre našu prácu využili ucelený dataset z pražskej Legerovej ulice, ktorý obsahuje všetky tri skúmané domény - údaje o doprave, meteorologických podmienkach a kvalite ovzdušia.

V úvodných častiach kapitoly sa zameriame na technickú integráciu a transformáciu tohto datasetu. Keďže surové dáta z meracích staníc obsahovali výpadky a nereálne chybové hodnoty, ukážeme si proces ich filtrácie, agregácie do jednotnej časovej osi a štandardizácie pre programátorské prostredie. V časti predspracovania dát si zdôvodníme kľúčové rozhodnutie skrátiť dataset len na druhú polovicu roka, čo bolo nevyhnutné na odstránenie dlhých, chýbajúcich blokov meraní a zachovanie plynulej časovej osi, ktorú vyžadujú sekvenčné neurónové siete.

Záverečná časť kapitoly je venovaná exploratívnej analýze, ktorá slúži na pochopenie vnútorných vzťahov v spracovaných dátach. Prostredníctvom vizualizácií a korelačnej analýzy preskúmame distribúciu premenných, identifikujeme extrémny a zhodnotíme denné a sezónne cykly. Potvrdíme si dôležité súvislosti fungovania mestského prostredia, napríklad priamy vplyv intenzity dopravy na emisie oxidov dusíka, či výrazný čistiaci efekt vetra a zrážok na kvalitu ovzdušia. Získané poznatky nám tak poskytnú základ pre lepšie rozhodovanie sa.

4.1 Zber a integrácia dát

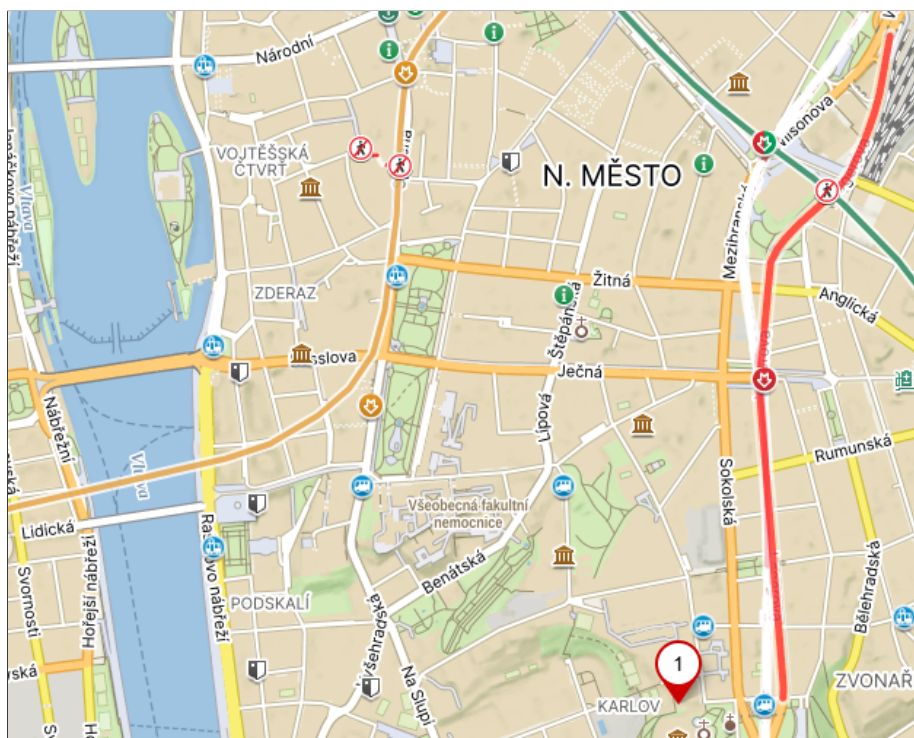
V prostredí Smart Cities je jednou z najväčších analytických výziev samotné získanie ucelených dát. Informácie o priepustnosti dopravy, meteorologických podmienkach a kvalite ovzdušia sú štandardne evidované v oddelených, často navzájom nekompatibilných informačných systémoch rôznych štátnych inštitúcií (napríklad hydrometeorologické ústavy vs. správy ciest).

Pre účely tejto záverečnej práce sme však využili inžiniersky efektívnejší prístup. Namiesto zložitého získavania a párovania dát z rôznorodých externých zdrojov sme zvolili prácu s jedným, vysoko komplexným datasetom. Tento dataset v sebe už natívne združoval všetky tri kľúčové domény nevyhnutné pre našu analýzu - údaje o doprave, meteorologické údaje aj údaje o kvalite ovzdušia.

Využitie takto predpripraveného, uceleného dátového súboru eliminovalo riziko vzniku chýb pri spájaní rôznych geografických lokalít a umožnilo nám sústrediť analytickú kapacitu priamo na fázu vnútorného predspracovania, agregácie údajov a následného modelovania pomocou umelej inteligencie.

4.1.1 Dataset z Legerovej ulice

Dataset "Legerova ulica" predstavuje vysoko špecifický dátový súbor zachytávajúci dopravnú a emisnú situáciu v Prahe počas roka 2020. Bol vytvorený pre pedagogické a výskumné účely a je dostupný v rámci univerzitného kurzu Soft Computing [27]. Dataset kombinuje dáta z priameho monitorovania ulice s meteorologickými údajmi z príľahlej a relevantnej meracej stanice (Praha-Karlov).



Obrázok 12: Geografická lokalizácia Legerovej ulice (zvýraznené červenou farbou) ako súčasť hlavného dopravného ťahu (Zdroj: [3])

Legerova ulica, ako dôležitá súčasť Severojužnej magistrály v Prahe, patrí medzi najvyťaženejšie dopravné ťahy v meste a historicky vykazuje jedny z najvyšších hodnôt znečistenia ovzdušia v strednej Európe [16]. Na obrázku 12 je zobrazená geografická lokalizácia Legerovej ulice zvýraznená červenou farbou. Takisto je na mape vyznačená meracia stanica (Praha-Karlov) označená bodom č. 1.

Doprava (R203005)	HODIN Teplota a srážky (Karlov)	PM10	PM2.5	CO	NO2	NOx	ID_Value_Type
-------------------	---------------------------------	------	-------	----	-----	-----	---------------

Obrázok 13: Použité hárky datasetu z Legerovej ulice

Zdrojové dáta sú poskytované vo formáte tabuľkového zošita (typ súboru .xlsx). Dataset sám o sebe obsahuje viacero hárkov z ktorých si vyberieme len niektoré. Na obrázku 13 sú zobrazené hárky, ktoré sme v tejto práci použili. Každý z týchto použitých hárkov obsahuje údaje o dátume a čase. Hárky o počasí a doprave majú niektoré atribúty v českom jazyku. Použité hárky môžeme rozdeliť do viacerých kategórií:

- **Dopravné údaje:** Je to hárok "Doprava (R203005)". Hárok obsahuje atribúty pre intenzitu vozidiel v jednotkách počet vozidiel za hodinu a priemernú rýchlosť vozidiel meranú v kilometroch za hodinu (km/h). Tieto údaje umožňujú sledovať hustotu a plynulosť cestnej premávky v hodinových intervaloch. Ukážka dát je na obrázku 14.

START_TIME [date]	START_TIME [month]	START_TIME [day]	START_TIME [hour]	START_TIME [hour]	Intenzita vozidel	průměrná rychlost
24.7.2020	7	24	4:00:00	4	0	-
24.7.2020	7	24	5:00:00	5	1300	10,32
24.7.2020	7	24	6:00:00	6	1164	10,87

Obrázok 14: Ukážka dát hárku "Doprava (R203005)"

- **Meteorologické údaje:** Je to hárok "HODIN Teplota a srážky (Karlovy)", ktorý sleduje hodinové záznamy obsahujúce: smer a rýchlosť vetra, smer a rýchlosť nárazového vetra, úhrn srážok a teplotu vzduchu. Smer vetra je meraný v stupňoch ($^{\circ}$), smer maximálneho nárazu vetra je tiež meraný v stupňoch ($^{\circ}$), priemerná rýchlosť vetra je meraná v metroch za sekundu (m/s), rýchlosť maximálneho nárazu vetra je taktiež meraná v metroch za sekundu (m/s), úhrn zrážok je meraný v milimetroch (mm) a teplota vzduchu je meraná v stupňoch celzia ($^{\circ}\text{C}$). Ukážka dát je na obrázku 15.

START_TIME [month]	START_TIME [day]	START_TIME [date]	START_TIME [hour]	směr větru (stupně)	směr maximálního nárazu větru (stupně)	průměrná rychlost větru (m/s)	rychlost maximálního nárazu větru (m/s)	úhrn srážek (mm)	teplota vzduchu ($^{\circ}\text{C}$)
01	1	1.1.2020	00:00	283	285	2,9	4,6	0	4,7
01	1	1.1.2020	01:00	268	259	2,7	4,2	0	4,1
01	1	1.1.2020	02:00	260	250	1,5	2,9	0	3,4

Obrázok 15: Ukážka dát hárku "HODIN Teplota a srážky (Karlovy)"

- **Kvalita ovzdušia:** Dataset obsahuje samostatné hárky pre kľúčové znečisťujúce látky: prachové častice (PM10 a PM2.5), oxid uhoľnatý (CO), oxid dusičitý (NO₂) a oxidy dusíka (NO_x). Pre každý z týchto hárkov sme pridali ukážku - obrázok 16, obrázok 17, ..., obrázok 20. Každý z týchto hárkov takisto obsahuje atribút ID_VALUE_TYPE. Keďže v tomto datasete (Legerova ulica) neboli bližšie popísané jednotky v ktorých boli veličiny pre kvalitu ovzdušia merané, tak sme ich prevzali z datasetu OpenISKO [28], z ktorých boli tieto veličiny prevzaté. Každá znečisťujúca látka bola meraná v jednotke mikrogram na meter kubický ($\mu\text{g}/\text{m}^3$). Takže pre hárok "PM10" a atribút VALUE s hodnotou 65 to znamená $65 \mu\text{g}/\text{m}^3$.
- **Dodatočné hárky:** Hárok "ID_Value_type" vysvetľuje číselné kódy atribútu ID_VALUE_TYPE v hárkoch o kvalite ovzdušia. Napríklad hodnota 9 znamená Verifikované dáta a 7 zase Chybový kód verifikovaných dát. Hárok je zobrazený na obrázku 21.

START_TIME [day]	START_TIME [hour]	ID_VALUE_TYPE	VALUE
1.1.2020	0:00:00	9	65
1.1.2020	1:00:00	9	35
1.1.2020	2:00:00	9	26

Obrázok 16: Ukážka dát hárku "PM10"

START_TIME [day]	START_TIME [hour]	ID_VALUE_TYPE	VALUE
1.1.2020	0:00:00	9	56
1.1.2020	1:00:00	9	29
1.1.2020	2:00:00	9	21

Obrázok 17: Ukážka dát hárku "PM2.5"

START_TIME [day]	START_TIME [hour]	ID_VALUE_TYPE	VALUE
1.1.2020	0:00:00	9	407
1.1.2020	1:00:00	9	379
1.1.2020	2:00:00	9	381

Obrázok 18: Ukážka dát hárku "CO"

START_TIME [day]	START_TIME [hour]	ID_VALUE_TYPE	VALUE
1.1.2020	0:00:00	9	35,6
1.1.2020	1:00:00	9	33,1
1.1.2020	2:00:00	9	32,5

Obrázok 19: Ukážka dát hárku "NO2"

START_TIME [day]	START_TIME [hour]	ID_VALUE_TYPE	VALUE
1.1.2020	0:00:00	9	54,90
1.1.2020	1:00:00	9	49,50
1.1.2020	2:00:00	9	53,80

Obrázok 20: Ukážka dát hárku "NOx"

ID_VALUE_
5 Vyřazeno pro nevěrohodnost
6 Chybový kód operativních dat
7 Chybový kód verifikovaných dat
8 Operativní data
9 Verifikovaná data
10 Substituce hodnoty < det. limit - operativní data
11 Substituce hodnoty < det. limit - verifikovaná data

Obrázok 21: Ukážka dát hárku "ID_Value_type"

Pri predbežnej exploratívnej analýze tohto datasetu sme si všimli počiatočné problémy, ktoré ovplyvňujú proces čistenia dát:

- **Chýbajúce a abnormálne hodnoty:** V dôsledku výpadkov senzorov alebo chýb pri prenose sa v dátach nachádzajú chýbajúce pozorovania. Výpadky však nie sú označené štandardnými nulovými formátmi (ako napríklad NULL alebo NaN), ale sú do nich pevne dosadené nereálne extrémne hodnoty (napríklad -5000, -5003), ktoré sú zobrazené na obrázku 22, alebo neštandardné textové znaky (mínus), ktoré sú zobrazené na predošlom obrázku 14. Takéto nereálne hodnoty sa nachádzali

v každom hárku, okrem hárku s meteorologickými údajmi. Pred samotným modelovaním je nutné tieto anomálie identifikovať a ošetriť.

START_TIME [day]	START_TIME [hour]	ID_VALUE_TYPE	VALUE
1.1.2020	16:00:00	9	36
1.1.2020	17:00:00	7	-5000
1.1.2020	18:00:00	7	-5000

Obrázok 22: Ukážka abnormálnych dát hárku "CO"

- **Prechodný rok:** Dataset pokrýva časové obdobie roka 2020. Rok 2020 je priestupný rok. Takže časové rady obsahujú celkovo 366 dní (vrátane 29. februára).

Tento dataset je ideálny na analýzu vplyvu intenzity dopravy na lokálne znečistenie ovzdušia, pričom zohľadňuje dôležité meteorologické faktory ako smer vetra, rýchlosť vetra a zrážky. Dataset obsahuje surové dáta, verifikované hodnoty ale aj anomálie, ktoré sťažujú analýzu. Každý hárok obsahuje 8748 riadkov dát.

4.2 Agregácia a transformácia zozbieraných dát

Keďže dáta z Legerovej ulice boli distribuované naprieč viacerými hárkami, proces agregácie prebiehal v prostredí tabuľkového procesora. Cieľom bolo vytvoriť zjednotený súbor s jednotnou časovou osou a prečistenými hodnotami. Tento proces bol realizovaný v nasledujúcich nadväzujúcich krokoch: časové zjednotenie, predbežná filtrácia a ošetrovanie anomálií, zjednotenie záznamov pomocou odkazov, premenovanie atribútov a následne finálny export.

4.2.1 Zjednotenie dát podľa času

Základom pre zlúčenie dát bolo vytvorenie jednotnej časovej osi. Pre tento účel bol vytvorený nový pracovný hárok "Dataset1". Časový vektor sa preberá z jednej z hlavných zdrojových tabuliek, ktorá slúžila ako referenčná os na zabezpečenie synchronizácie záznamov. Pôvodné časové formáty (napríklad 00:00:00 1.1.2020, alebo rozdelené formáty do dátumu 1.1.2020 a času 00:00:00) neboli vhodné pre algoritmické spracovanie, preto bol vytvorený nový atribút s názvom "Datetime". Pomocou textových funkcií tabuľkového procesora boli dátum a čas prekonvertované do medzinárodného štandardu ISO 8601 podľa nasledujúceho vzorca (vzorec je zalomený, kvôli pretečeniu na nový riadok):

```
=TEXT(Tabuľka8[Date]; "yyyy-mm-dd") & "T"
& TEXT(Tabuľka8[Time]; "hh:mm:ss")
```

Výsledkom bola štandardizovaná textová reprezentácia času, ktorá slúžila ako index pre ďalšie prepájanie. Ukážka atribútu Datetime je na obrázku 23.

Datetime
2020-01-01T00:00:00
2020-01-01T01:00:00
2020-01-01T02:00:00

Obrázok 23: Ukážka hodnôt atribútu Datetime hárku "Dataset1"

4.2.2 Predbežná filtrácia a ošetrovanie anomálií

Pre každý meraný atribút obsahujúci anomálie sme v zdrojových hárkoch vytvorili pomocný atribút. Úlohou tohto atribútu bolo vyfiltrovať chybné stavy senzorov (ako napríklad spomínané hodnoty -5000, -5003, znak mínus "-"). V prípade, že bunka obsahovala platnú nameranú hodnotu, atribút ju prevzal; v prípade detekcie anomálie bola bunka nahradená prázdny reťazcom (""). Tento krok výrazne uľahčil neskoršie algoritmické doplnenie chýbajúcich dát. Ukážka pomocného atribútu REAL_VALUE je zobrazená v tabuľke 1.

Tabuľka 1: Ukážka pomocného atribútu REAL_VALUE pre hárk "CO"

Date	Time	VALUE	REAL_VALUE
1.1.2020	16:00:00	608	608
1.1.2020	17:00:00	787	787
1.1.2020	18:00:00	-5003	
1.1.2020	19:00:00	649	649

START_TIME [date]	START_TIME [month]	START_TIME [day]	START_TIME [hour]	START_TIME [hour]2	Intenzita vozidel	průměrná rychlost	real_vehicle_intensity	real_vehicle_average_speed
24.7.2020	7	24	4:00:00	4	0	-		
24.7.2020	7	24	5:00:00	5	1300	10,32	1300	10,3222222
24.7.2020	7	24	6:00:00	6	1164	10,87	1164	10,8666667

Obrázok 24: Ukážka filtrovaných hodnôt hárku "Doprava (R203005)"

Pre ukážku o doprave uvádzame obrázok 24. V tomto prípade sme museli uvažovať nad oboma atribútmi, keďže hodnota 0 pre intenzitu vozidiel môže naozaj nastať.

4.2.3 Zjednotenie záznamov pomocou odkazov

Po príprave časovej osi a filtrácii chýb nasledoval proces samotného zlučovania. Do hárku "Dataset1" sme postupne importovali hodnoty z vybraných hárkov pomocou odkazov. Napríklad prenos prefiltrovaných hodnôt oxidu uhoľnatého z hárku "CO" bol realizovaný odkazom:

=Tabuľka4[@[REAL_VALUE]]

Datetime	směr větru (stupně)	průměrná rychlost větru (m/s)	směr maximálního nárazu větru (stupně)	rychlost maximálního nárazu větru (m/s)	úhrn srážek (mm)	teplota vzduchu (°C)	PM10	PM2.5	CO	NO2	NOx	Intenzita vozidel	průměrná rychlost
2020-01-01T00:00:00	283	2,9	285	4,6	0	4,7	65	56	407	35,6	54,90		
2020-01-01T01:00:00	268	2,7	259	4,2	0	4,1	35	29	379	33,1	49,50		
2020-01-01T02:00:00	260	1,5	250	2,9	0	3,4	26	21	381	32,5	53,80		

Obrázok 25: Ukážka prepojených hodnôt v hárku "Dataset1"

Tento prístup zabezpečil, že v jednom riadku priradenom ku konkrétnej hodine sa ocitli všetky relevantné nezávislé premenné. Na obrázku 25 si všimneme zjednotené hodnoty.

4.2.4 Premenovanie atribútov

Pre zabezpečenie kompatibility datasetu s programovacími jazykmi (v ktorých prebiehalo tréovanie modelov) sme pripravili dáta do finálneho pracovného hárku "Dataset2". V tejto fáze sme upravili názvy atribútov nasledovne:

- Pridanie merných jednotiek a formátovanie: Názvy atribútov boli upravené tak, aby explicitne obsahovali merné jednotky. Napríklad mernú jednotku metrov za sekundu (m/s) sme namapovali na "inMetersPerSecond". Pre názvy bol zvolený formát camelCase.
- Preklad do anglického jazyka: Prostredníctvom pomocného hárku ("Hlavičky") boli všetky názvy premenných preložené zo slovenského/českého jazyka do angličtiny, čo predstavuje štandard pre medzinárodné programátorské prostredia.

startDatetime	windDire ctionInDe grees	windAvera geSpeedIn MetersPer Second	maximum WindGust DirectionIn Degrees	maximumW indGustSpe edInMeters PerSecond	precipit ationTo tallnMili meters	airTemper atureInDe greesCelsi us	pm10In Microgra msPerCu bicMeter	pm25In Microgr amsPerC ubicMet er	colnMicr ogramsP erCubic Meter	no2InMi crograms PerCubic Meter	noxInMicr ogramsPer CubicMete r	vehicleInTe nsityInVehi clesPerHou r	vehicleAvera geSpeedInKil ometersPerH our
2020-01-01T00:00:00	283	2.9	285	4.6	0	4.7	65	56	407	35.6	54.9		
2020-01-01T01:00:00	268	2.7	259	4.2	0	4.1	35	29	379	33.1	49.5		
2020-01-01T02:00:00	260	1.5	250	2.9	0	3.4	26	21	381	32.5	53.8		

Obrázok 26: Ukážka hodnôt vo finálnom hárku "Dataset2"

Takže atribút "průměrná rychlost" meraný v jednotke kilometre za hodinu (km/h) sme premenovali na "vehicleAverageSpeedInKilometersPerHour". Všimnime si obrázok 26, kde je zobrazená ukážka hárku "Dataset2".

4.2.5 Finálny export

Takto agregovaný, štandardizovaný a preložený dataset sme následne vyexportovali z tabuľkového procesora do formátu .csv (Comma-Separated Values).

Ukážka vyexportovaného súboru je na obrázku 27. Atribúty aj hodnoty sú oddelené bodkočiarkou. Názvy hlavičiek sú zalomené, aby pretekali do ďalšieho riadku, takže

```
startDatetime,windDirectionInDegrees,averageWindSpeedInMetersPerSecond,maxWindDirectionInDegrees,
maxWindSpeedInMetersPerSecond,precipitationInMillimeters,airTemperatureInCelsius,
pm10InMicrogramsPerCubicMeter,pm25InMicrogramsPerCubicMeter,coInMicrogramsPerCubicMeter,
no2InMicrogramsPerCubicMeter,noxInMicrogramsPerCubicMeter,vehicleIntensityInVehiclesPerHour,
vehicleAverageSpeedInKilometersPerHour
2020-01-01T00:00,283,2.9,285,4.6,0.0,4.7,65.0,56.0,487.0,35.6,54.9,,
2020-01-01T01:00,268,2.7,259,4.2,0.0,4.1,35.0,29.0,379.0,33.1,49.5,,
2020-01-01T02:00,260,1.5,250,2.9,0.0,3.4,26.0,21.0,381.0,32.5,53.8,,
```

Obrázok 27: Ukážka vyexportovaného datasetu vo formáte .csv

vidíme každý atribút. Takýto súbor predstavuje primárny, plne štandardizovaný vstupný bod do programovacieho prostredia (jazyka Python), kde môže byť načítaný knižnicami na následnú extrakciu príznakov, analýzu a tréning modelov umelej inteligencie.

Počet riadkov dát v datasete zostal na 8748.

4.3 Predspracovanie dát

Zozbierané a agregované dáta predstavujú len surový vstupný materiál. Pre aplikáciu pokročilých metód strojového učenia, akými sú hlboké neurónové siete, musia dáta spĺňať prísne kritériá na konzistenciu, úplnosť a mierku. Fáza predspracovania dát je preto kritickým krokom, ktorý určuje výslednú presnosť predikčných modelov. Proces transformácie postupuje od surových dát cez čistenie po doplnenie dát.

4.3.1 Analytické zhodnotenie a čistenie anomálií

Ako už bolo podrobne opísané v podkapitole Agregácia a transformácia zozbieraných dát, samotný technický proces odstránenia šumu prebiehal súbežne s integráciou dát (prostredníctvom filtračného atribútu REAL_VALUE). Z metodologického hľadiska je však nevyhnutné zdôvodniť, prečo bola táto operácia pred samotnou analýzou kritická.

Pri prvotnej explorácii zdrojových hárkov bolo zistené, že výpadky meraní (či už pri intenzite dopravy alebo koncentrácii polutantov) neboli v systéme zaznamenávané ako prázdne hodnoty, ale boli do nich natvrdo zapisované nereálne záporné extrémne (napríklad hodnota -5000).

Ponechanie takýchto hodnôt v datasete by modely umelej inteligencie úplne znehodnotilo. Algoritmy by túto chybu senzora interpretovali ako skutočný, extrémne nízky fyzikálny stav (napríklad záporný počet prachových častíc), čo by viedlo ku skresleniu predikcií. Vďaka integrácii čistenia priamo do procesu agregácie boli tieto anomálie úspešne identifikované a nahradené identifikátorom pre chýbajúcu hodnotu (prázdny reťazec), čím sme dataset pripravili na ďalší logický krok - riešenie dier v dátach.

Dôležitým strategickým rozhodnutím počas tohto čistenia bolo neodstraňovať samotné riadky (časové značky), v ktorých sa anomália vyskytla. Odstránenie celých riadkov by porušilo plynulú kontinuitu časového radu, čo by následne znemožnilo použitie

sekvenčných neurónových sietí, ktoré pre správne fungovanie vyžadujú neprerušenu časovú os.

4.3.2 Riešenie chýbajúcich hodnôt

Po nahradení anomálií hodnotami prázdnyimi reťazcami sme zistili, že výpadky dát neboli rozložené rovnomerne. V prvej polovici roka sa nachádzali rozsiahle a kontinuálne bloky chýbajúcich meraní, diery v dátach, kedy boli senzory pravdepodobne dlhodobo nefunkčné. Jednalo sa konkrétne o dáta o doprave.

Pri riešení tohto problému sme zvažovali použitie neurónovej siete. Neurónová sieť by sa naučila priebeh na dátach druhej polovice roka. Tým pádom by sme pomocou naučenej neurónovej siete mohli dopočítať dáta z prvej polovice roka. Tento prístup by však vniesol do datasetu obrovské množstvo syntetického skreslenia. Model, ktorý by sa učil na takýchto dátach by sa v skutočnosti neučil z reality, ale z odhadov iného algoritmu, čo by nemuselo zohľadňovať skutočnosť.

Z tohto dôvodu sme spravili rozhodnutie o skrátení datasetu. Z pôvodného celoročného súboru sme vyextrahovali len súvislý blok dát z druhej polovice roka (približne od 24. júla 2020). Tento blok predstavuje takmer 4000 záznamov s vysokou integritou a kvalitou meraní, čo je mierne postačujúca vzorka na natrénovanie predikčných modelov. Zostalié výpadky, ktoré sa vyskytli v tomto vybranom polroku, sme následne doplnili z okolitých hodnôt - ako to je v tabuľke 2.

Tabuľka 2: Ukážka doplnenia chýbajúcich hodnôt pre atribút oxid uhoľnatý (CO)

Dátum a čas	Hodnota pred doplnením	Hodnota po doplnení
2020-07-31T09:00:00	337	337
2020-07-31T10:00:00	(prázdny reťazec)	394
2020-07-31T11:00:00	(prázdny reťazec)	394
2020-07-31T12:00:00	394	394

Výsledkom tohto kroku je skrátený a naplnený dataset, ktorý budeme používať. Dataset obsahuje 3864 riadkov dát. Kvôli skráteniu datasetu dáta pokrývajú už len druhú polovicu roka od 24. júla po 31. decembra 2020.

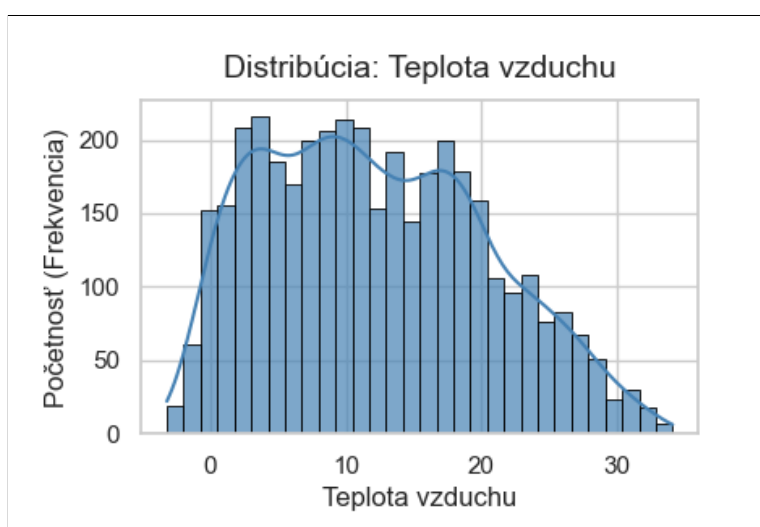
4.4 Exploratívna analýza dát

Po úspešnom predspracovaní a finalizovaní datasetu nasleduje fáza exploratívnej analýzy dát. Jej hlavným cieľom je pochopiť vnútornú štruktúru dát, identifikovať skryté vzorce správania a kvantifikovať vzájomné závislosti medzi dopravou, meteorológiou a kvalitou ovzdušia.

Analýzu sme rozdelili do štyroch oblastí: analýza distribúcie dát a skúmanie extrémov, analýza časových závislostí, korelačná analýza a následne sumarizácia analýz a zistení.

4.4.1 Analýza distribúcie dát a skúmanie extrémov

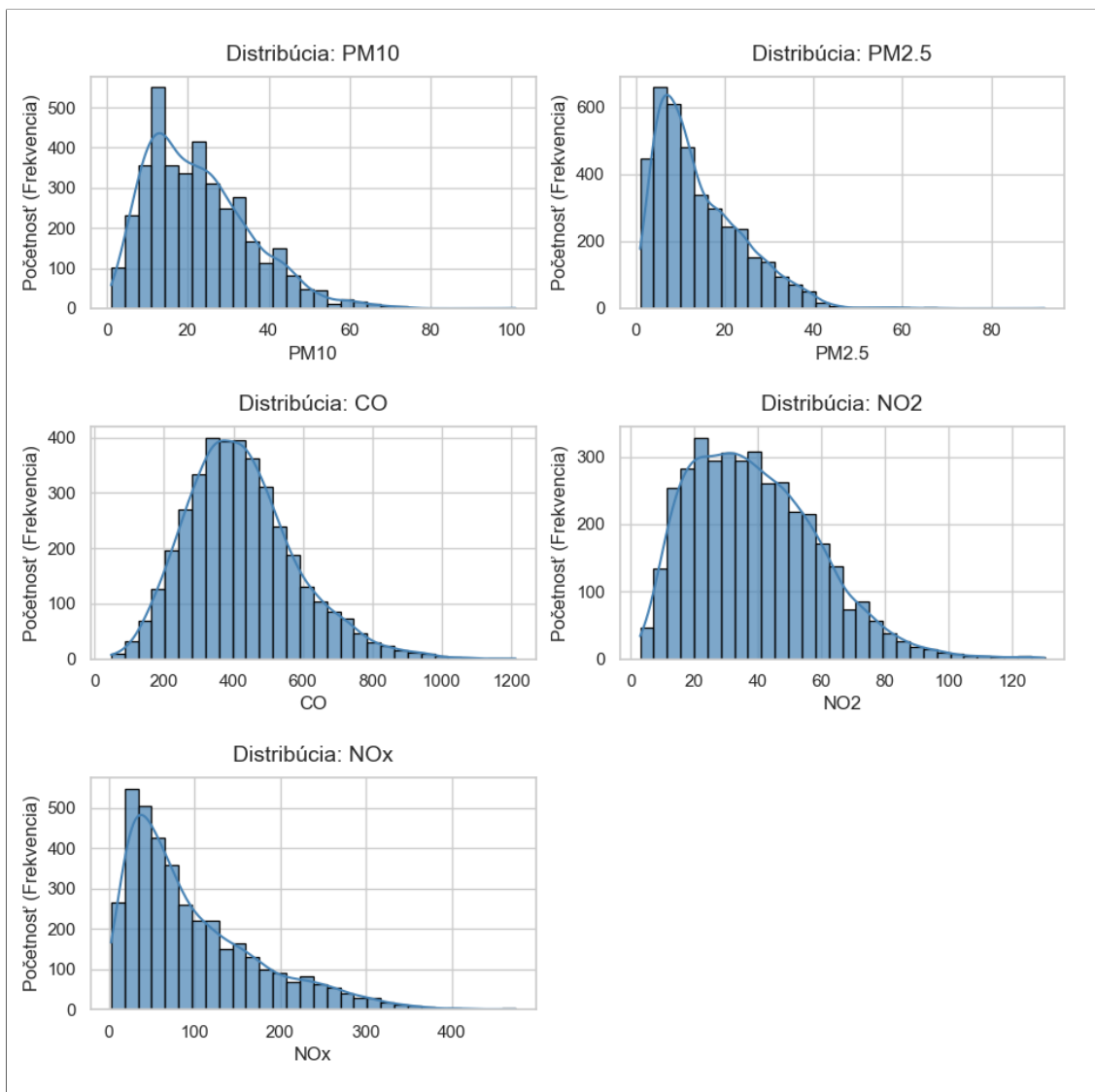
Za účelom pochopenia základných štatistických vlastností a rozloženia analyzovaného súboru dát sme vykonali vizualizáciu prostredníctvom histogramov doplnených o krivku odhadu hustoty pravdepodobnosti, ktorá histogram "vyhladí" a pomôže nám lepšie vizualizovať distribúciu dát. Analýza distribúcie jednotlivých premenných odhalila niekoľko dôležitých charakteristík skúmaného prostredia.



Obrázok 28: Histogram distribúcie teploty vzduchu

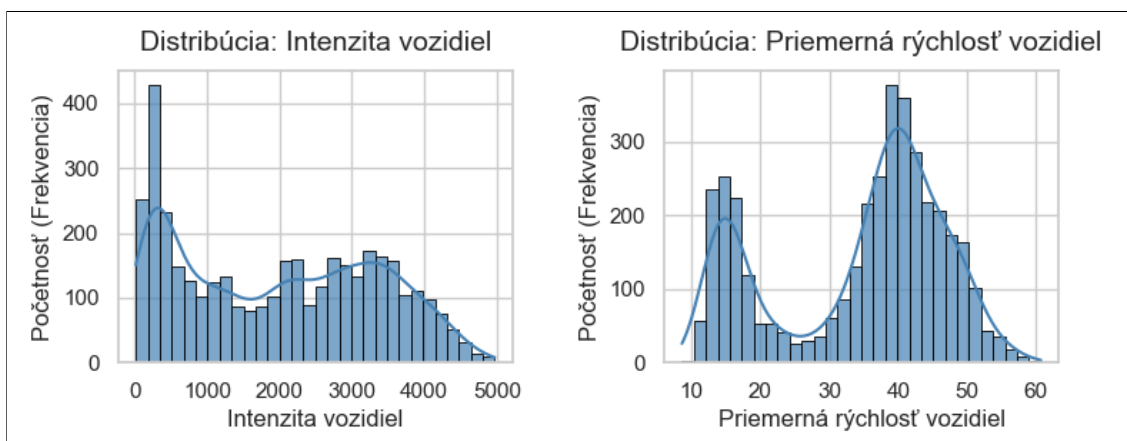
Pri meteorologickom ukazovateli teplota vzduchu na obrázku 28, sa blíži ku štandardnému normálnemu rozdeleniu. Tento jav je logicky odôvodniteľný rozsahom samotného datasetu. Keďže analyzované dáta pokrývajú výhradne druhú polovicu roka (od 24. júla do 31. decembra), v datasete chýbajú neskoré zimné (január, február) a jarné mesiace (marec, apríl, máj, jún a viac než polovica júla), čo prirodzene narúša symetriu ročného teplotného cyklu a mierne deformuje tvar Gaussovej krivky. Histogram pre úhrn zrážok tu ani nezobrazujeme, keďže väčšina hodnôt boli rovné 0.

Výraznou črtou analyzovaných dát je prítomnosť pozitívneho zošikmenia. Tento asymetrický tvar distribúcie s dlhým chvostom na pravej strane sme identifikovali pri priemernej rýchlosti vetra a pri všetkých sledovaných látkach znečisťujúcich ovzdušie (PM10, PM2.5, CO, NO₂ a NO_x), všimnime si obrázok 29. Z hľadiska kvality ovzdušia tento tvar krivky indikuje, že počas sledovaného obdobia prevažovali dni s relatívne priaznivými a stabilnými podmienkami (hodnoty sa koncentrujú v ľavej časti grafu bližšie k nule). Dlhý chvost smerom k vysokým hodnotám však reprezentuje výskyt občasných, ale o to výraznejších epizód s vysokou koncentráciou znečistenia, ktoré sú



Obrázok 29: Histogram distribúcie kvality ovzdušia

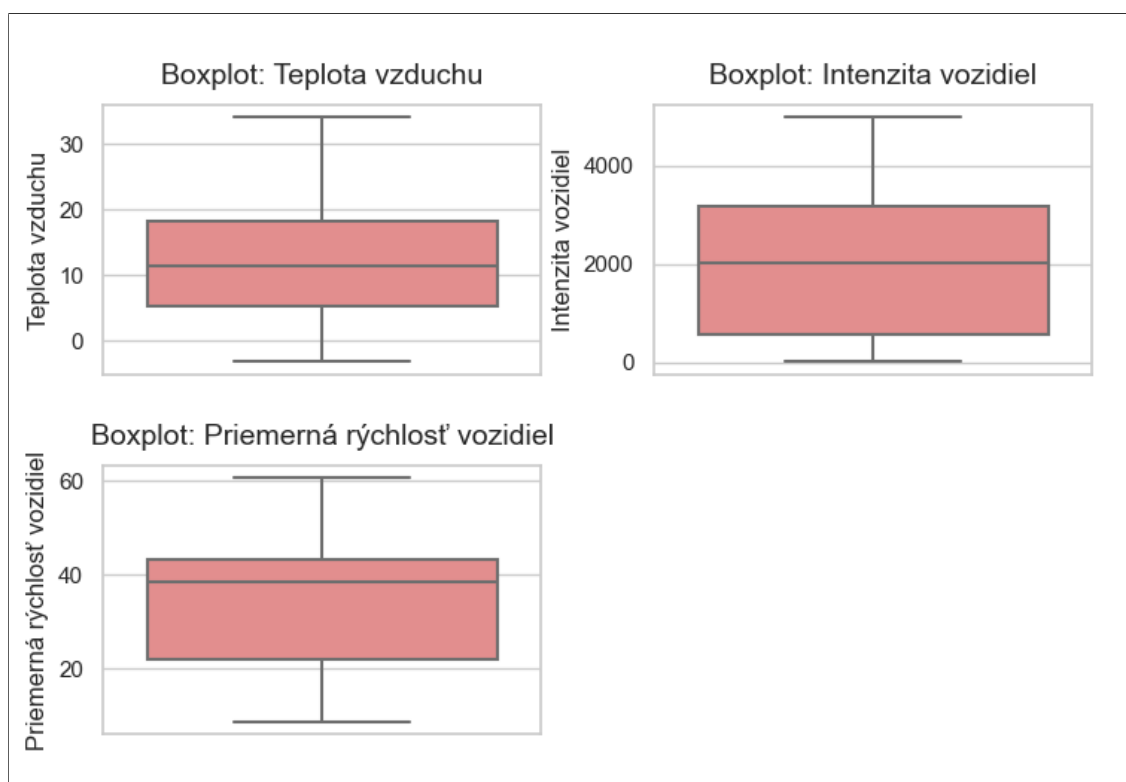
typické napríklad pre zimné smogové situácie alebo zhoršené rozptylové podmienky.



Obrázok 30: Histogram distribúcie dopravy

Zaujímavým zistením z hľadiska dopravných charakteristík je rozloženie priemernej rýchlosti vozidiel a intenzity vozidiel, ktoré obe vykazujú znaky tzv. bimodálnej distribúcie (rozdelenie s dvoma lokálnymi vrcholmi), všimnime si obrázok 30. Tento jav odzrkadľuje prirodzený cyklus a dynamiku dopravy na sledovanom úseku. Pri intenzite dopravy jeden vrchol predstavuje nízky počet vozidiel počas nočných hodín a sedla, zatiaľ čo druhý vrchol reprezentuje vysoký objem dopravy počas denných špičiek. Tento cyklus priamo súvisí s bimodalitou pri priemernej rýchlosti, ktorú možno interpretovať dvoma dominantnými stavmi: prvým, ľavým vrcholom odzrkadľujúcim spomalenú premávku počas dopravných špičiek alebo dopravným zápchem (nižšia rýchlosť) a druhým, pravým vrcholom reprezentujúcim plynulú premávku vo voľných hodinách (vyššia rýchlosť).

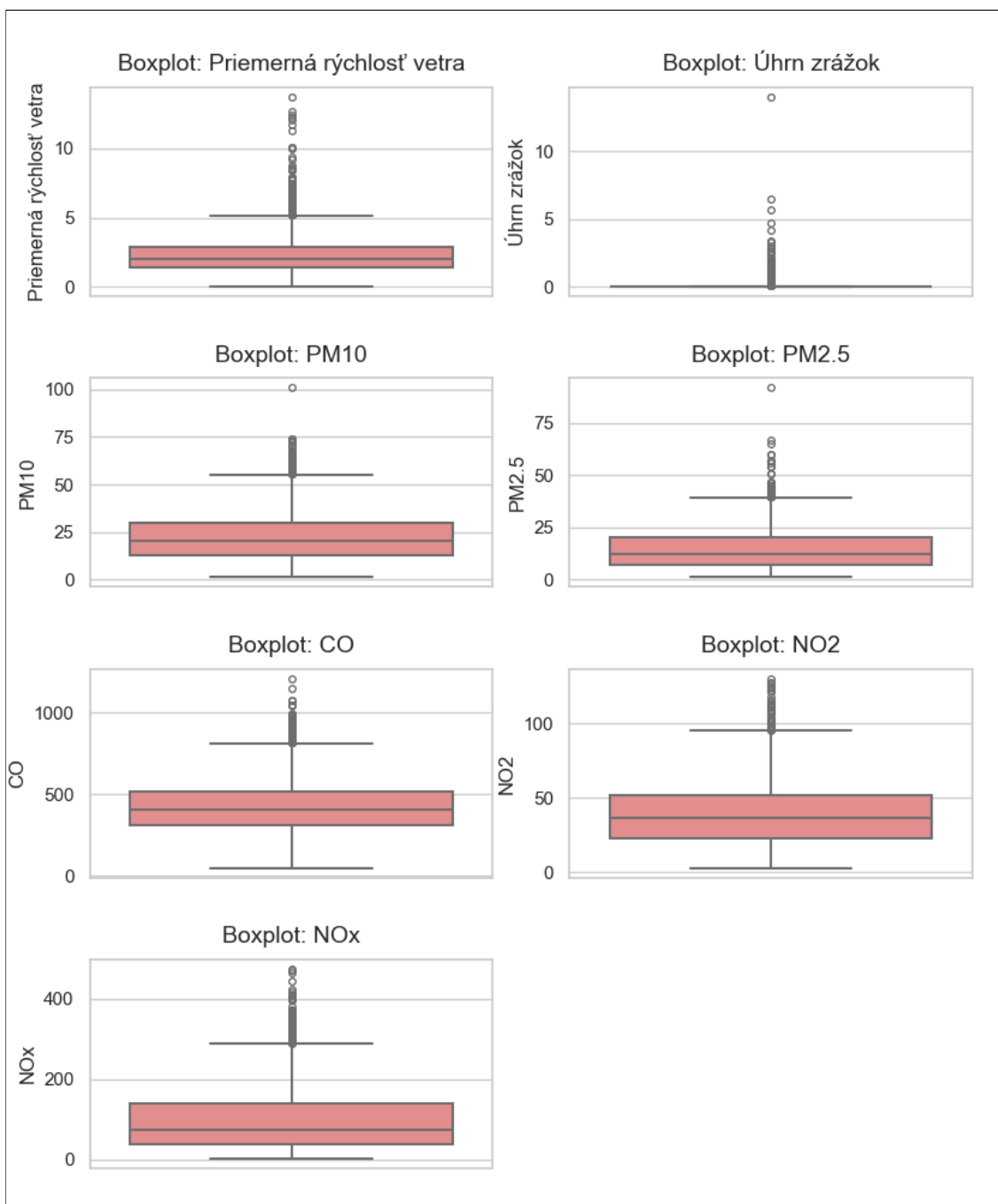
Na hlbšie pochopenie štruktúry dát a identifikáciu extrémnych výkyvov v meraniach sme využili vizualizáciu prostredníctvom boxplotov. Táto metóda umožňuje prehľadne zobrazovať medián, medzikvartilové rozpätie a predovšetkým odľahlé pozorovania (outliery, extrémny), ktoré sú v grafoch znázornené ako samostatné body nachádzajúce sa mimo štandardného rozsahu (nad tzv. fúzmi grafu).



Obrázok 31: Boxplot bez odľahlých hodnôt

Analýza ukázala jasné rozdelenie premenných do dvoch skupín. Prvú skupinu tvoria premenné, pri ktorých sa extrémne hodnoty nevyskytujú, konkrétne teplota vzduchu, intenzita vozidiel a priemerná rýchlosť vozidiel, všimnime si obrázok 31. Absencia outlierov pri teplote je prirodzená, keďže zmeny teplôt v prírode prebiehajú postupne a bez náhlych, extrémnych skokov mimo bežného sezónneho rámca - teplota nezvykne

podskočiť len tak o 20°C. V prípade dopravných charakteristík je absencia extrémov daná fyzikálnymi obmedzeniami samotnej infraštruktúry - intenzita vozidiel nemôže prekročiť maximálnu priepustnosť komunikácie a rýchlosť je limitovaná dopravnými predpismi a hustotou premávky. Tieto premenné teda oscilujú v predvídateľných a uzavretých hraniciach, takže grafy nemajú odľahlé hodnoty.



Obrázok 32: Boxplot s odľahlými hodnotami

Druhú, z hľadiska výskumu podstatne zaujímavejšiu skupinu, tvoria premenné s výrazným výskytom odľahlých pozorovaní. Patria sem meteorologické javy (priemerná rýchlosť vetra, úhrn zrážok) a všetky sledované ovzdušie znečisťujúce látky (PM10,

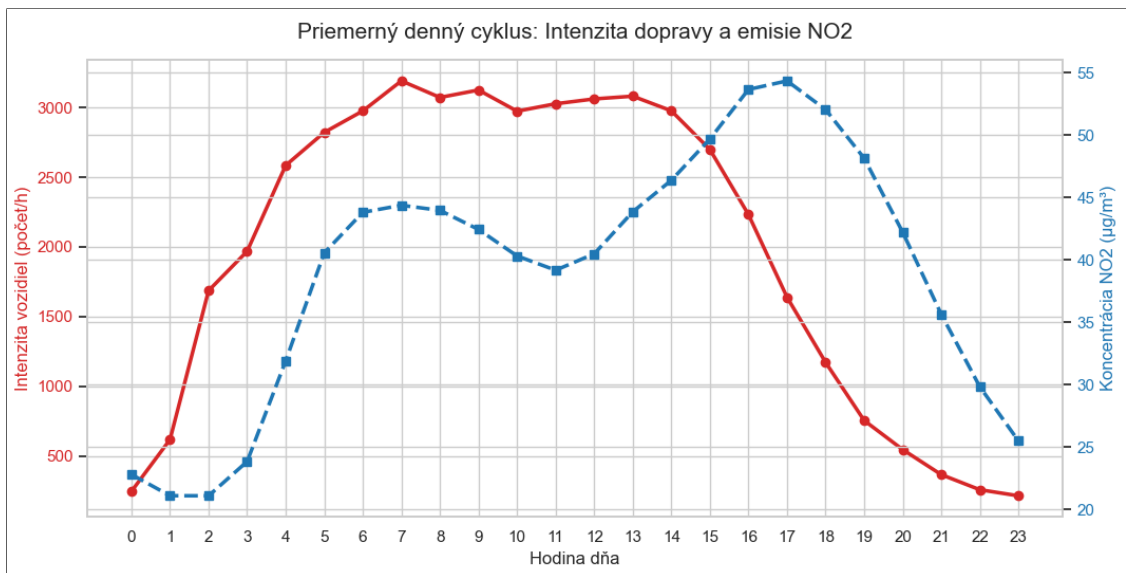
PM2.5, CO, NO₂ a NO_x), všimnime si obrázok 32. Vzhľadom na povahu týchto dát nemusí ísť o chyby meracích prístrojov, ale o reálne a mimoriadne dôležité environmentálne epizódy.

Pri vetre a zrážkach tieto extrémny reprezentujú ojedinelé búrkové udalosti alebo nárazové vetry - extrémne počasie. V prípade emisií (CO, NO₂, NO_x) a prachových častíc (PM₁₀, PM_{2.5}) body nad fúzmi boxplotov vizualizujú kritické smogové situácie. Tieto extrémny typicky vznikajú počas zimných mesiacov v dôsledku zhoršených rozptylových podmienok a teplotných inverzií, kedy nedochádza k dostatočnému premiešavaniu vrstiev ovzdušia a znečistenie z dopravy a lokálneho vykurovania sa kumuluje pri zemi - zimný smog [11].

Z metodologického hľadiska je dôležité zdôrazniť, že identifikované odľahlé pozorovania sme z datasetu neodstránili. V kontexte analýzy kvality ovzdušia tieto extrémne hodnoty nesú najvyššiu informačnú hodnotu, keďže reprezentujú obdobia s najväčším negatívnym dopadom na ľudské zdravie a životné prostredie.

4.4.2 Analýza časových závislostí a sezónnosti

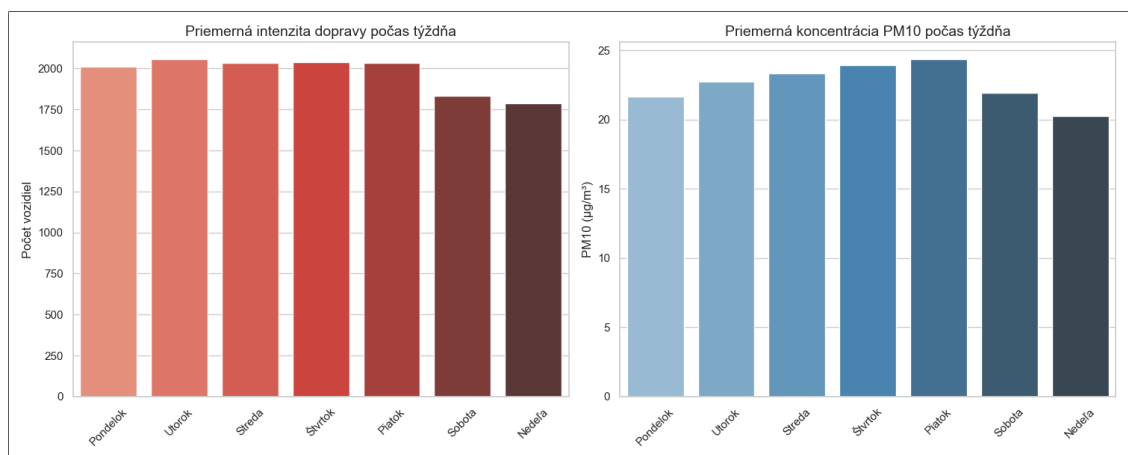
Pochopenie vplyvu dopravy na kvalitu ovzdušia si vyžaduje analýzu dát v rôznych časových mierkach. Hodnotenie bolo preto rozdelené na analýzu denného cyklu, týždenného chodu a dlhodobého mesačného trendu s ohľadom na nástup zimnej sezóny.



Obrázok 33: Priemerný denný cyklus intenzity dopravy a emisie NO₂

Priemerný denný cyklus odhalil špecifické správanie oboch skúmaných fenoménov. Krivka intenzity dopravy nenadobúda klasický tvar s dvoma oddelenými a ostrými dopravnými špičkami (rannou a poobednou), ale vykazuje skôr tvar lichobežníka. Zvýšená intenzita dopravy začína približne o 6:00 hodine rannej a pretrváva na relatívne stabilnej a vysokej úrovni kontinuálne až do 14:00, všimnime si obrázok 33.

Na druhej strane, denný chod koncentrácií oxidu dusičitého (NO_2) vykazuje zreteľnejšiu bimodalitu (dva lokálne vrcholy). Hoci emisie NO_2 primárne pochádzajú zo spaľovacích motorov vozidiel, ich koncentrácia v ovzduší nekopíruje intenzitu vozidiel absolútne presne. Medzi ranným a poobedným maximom NO_2 je pozorovateľný pokles, napriek tomu, že intenzita dopravy zostáva vysoká. Toto fázové posunutie a pokles koncentrácií je možné pripísať lepším rozptylovým podmienkam v poludňajších hodinách vplyvom denného ohrievania vzduchových hmôt (termická konvekcia), čo zabezpečuje lepšie odvetrávanie znečistenia.



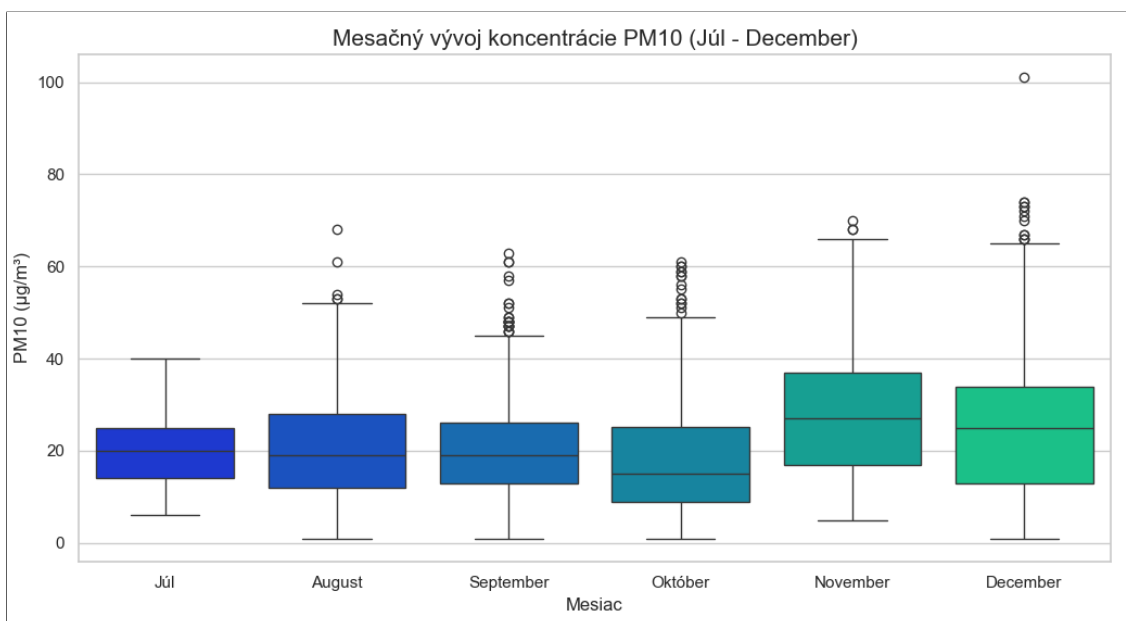
Obrázok 34: Priemerná intenzita dopravy počas týždňa a priemerná koncentrácia PM10 počas týždňa

Analýza týždenného cyklu, podľa obrázku 34, potvrdila pokles intenzity dopravy počas víkendu (sobota, nedeľa) v porovnaní s pracovnými dňami (pondelok až piatok). Tento pokles je vizuálne detekovateľný, avšak nie je výrazný, čo naznačuje, že sledovaný úsek je kontinuálne zaťažovaný a plní dôležitú funkciu aj mimo pracovného týždňa.

Zaujímavý trend vykazujú koncentrácie prachových častíc PM10. Kým v nedeľu dosahujú svoje lokálne minimum, hodnoty namerané v sobotu sú prekvapivo porovnateľné s pondelkom. Zvyšné pracovné dni vykazujú stabilne vyššie koncentrácie oproti začiatku týždňa. Tento jav pravdepodobne súvisí s kumuláciou znečistenia v ovzduší v priebehu pracovného týždňa a s rozličnými zvykmi obyvateľstva cez víkend (napríklad sobotné nákupy a presuny ale pokojná nedeľa).

Mesačný vývoj (júl až december 2020) zreteľne vizualizuje vplyv sezónnosti a nástupu chladnejších mesiacov na kvalitu ovzdušia, všimnime si obrázok 35. Prvé náznaky zhoršených rozptylových podmienok sa na boxplotoch prejavujú výskytom extrémov už v priebehu septembra a októbra.

V novembri je síce frekvencia extrémnych výkyvov nepatrne nižšia, avšak je viditeľný signifikantný nárast celkovej priemernej úrovne znečistenia oproti predchádzajúcim mesiacom. Najkritickejšim mesiacom je jednoznačne december, pre ktorý je charakteristický vysoký nárast extrémnych hodnôt. Tento exponenciálny nárast



Obrázok 35: Mesačný vývoj koncentrácie PM10

smogových epizód na konci roka priamo súvisí s plným nástupom vykurovacej sezóny, častejším výskytom teplotných inverzií a bezvetrím, ktoré v kombinácii s dopravnými emisiami vedú k výraznej degradácii kvality ovzdušia.

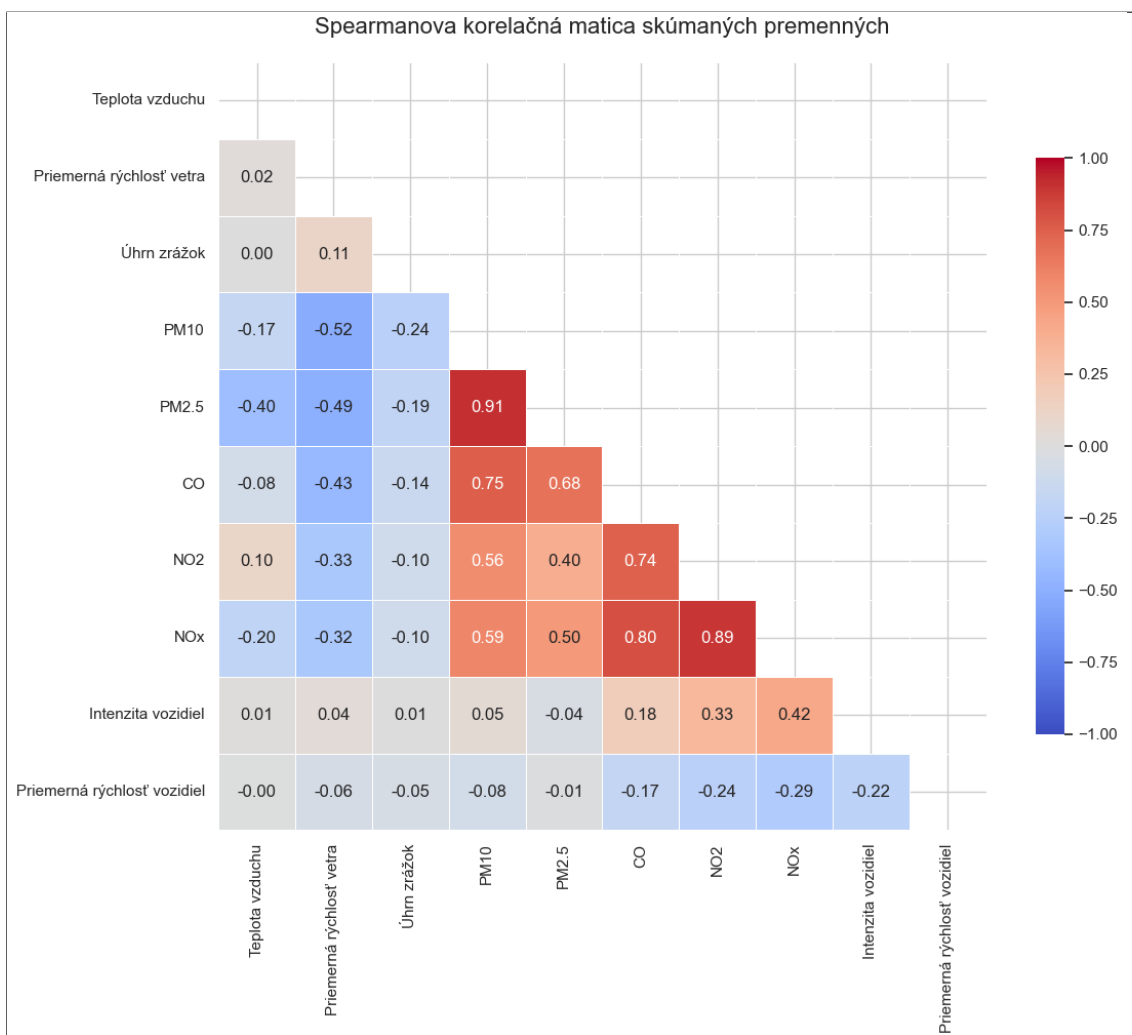
4.4.3 Korelačná analýza

Na kvantifikáciu sily a smeru vzájomných závislostí medzi meteorologickými podmienkami, charakteristikami dopravného prúdu a koncentráciou znečisťujúcich látok sme vykonali korelačnú analýzu. Vzhľadom na prítomnosť odľahlých pozorovaní (extrémov) a asymetrické rozdelenie väčšiny premenných bol na výpočet zvolený neparametrický Spearmanov koeficient poradovej korelácie [29], ktorý je voči týmto vlastnostiam dát robustnejší. Spearmanov koeficient budeme označovať r_s a vypočítame ho rovnicou (14):

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (14)$$

kde n je celkový počet pozorovaní, d_i predstavuje rozdiel medzi poradím i -tej dvojice hodnôt ($d_i = \text{Poradie}(x_i) - \text{Poradie}(y_i)$) pre $i = 1, 2, \dots, n$. Koeficient nadobúda hodnoty z intervalu $\langle -1, 1 \rangle$, pričom hodnota 1 indikuje priamu poradovú závislosť, hodnota -1 dokonalú nepriamu závislosť a hodnota blízka 0 naznačuje absenciu poradovej korelácie. Ak je poradie skúmanej hodnoty rovnaké pre viaceré hodnoty, tak sa z nich vypočíta aritmetický priemer.

Výsledky analýzy potvrdili viaceré dôležité predpoklady o správaní sa skúmaného systému.

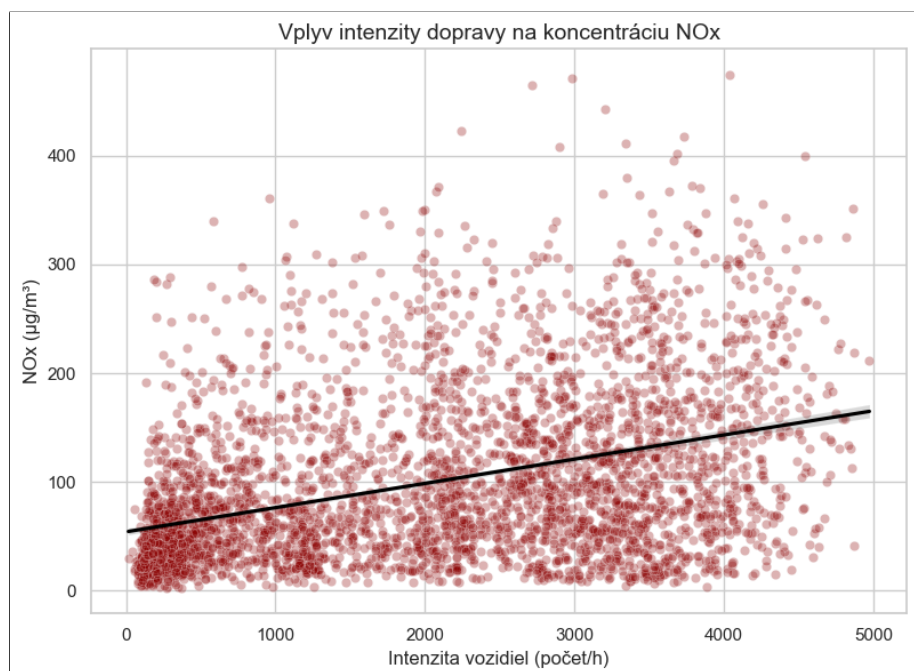


Obrázok 36: Spearmanova korelačná matica

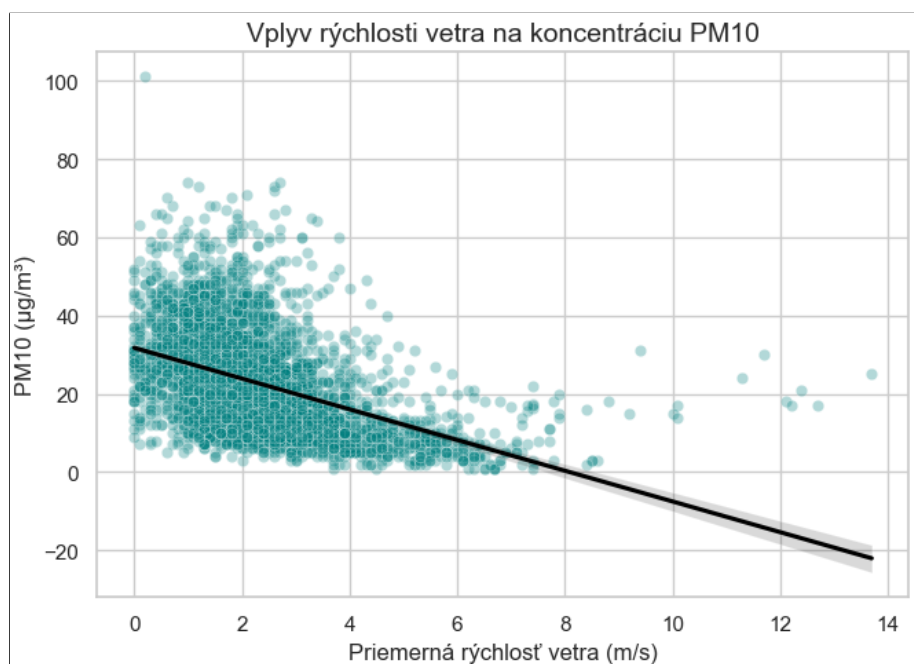
Intenzita vozidiel preukázala jednoznačnú pozitívnu koreláciu s primárnymi emisiami pochádzajúcimi zo spaľovacích motorov. Najsilnejší vzťah bol zaznamenaný pri oxidoch dusíka ($r_s = 0,42$ pre NOx a $r_s = 0,33$ pre NO2), všimnime si obrázok 36. Tieto hodnoty v kontexte komplexných environmentálnych dát potvrdzujú, že cestná doprava je na sledovanom úseku dominantným zdrojom znečistenia oxidmi dusíka.

Detailnejšie zobrazenie vplyvu intenzity dopravy na koncentráciu oxidov dusíka (NOx) sú zobrazené na obrázku 37. Slabšia pozitívna korelácia sa preukázala aj pri oxide uhoľnatom (CO, $r_s = 0,18$), čo rovnako korešponduje s produkciou výfukových plynov.

Priemerná rýchlosť vetra sa ukázala ako najdôležitejší meteorologický faktor ovplyvňujúci čistotu ovzdušia. Bola zaznamenaná stredne silná negatívna korelácia (všetky hodnoty pod úrovňou $-0,32$) so všetkými sledovanými znečisťujúcimi látkami (PM10, PM2.5, CO, NO2 a NOx). Na obrázku 38 je zobrazený detailnejší pohľad na vplyv rýchlosti vetra na koncentráciu prachových častíc PM10. Tento inverzný vzťah dokazuje, že so stúpajúcou rýchlosťou vetra dochádza k lepšiemu premiešavaniu vzduchových hmôt, rozptylu emisií a tým k rýchlemu poklesu ich koncentrácie pri zemi. Inak povedané - vietor



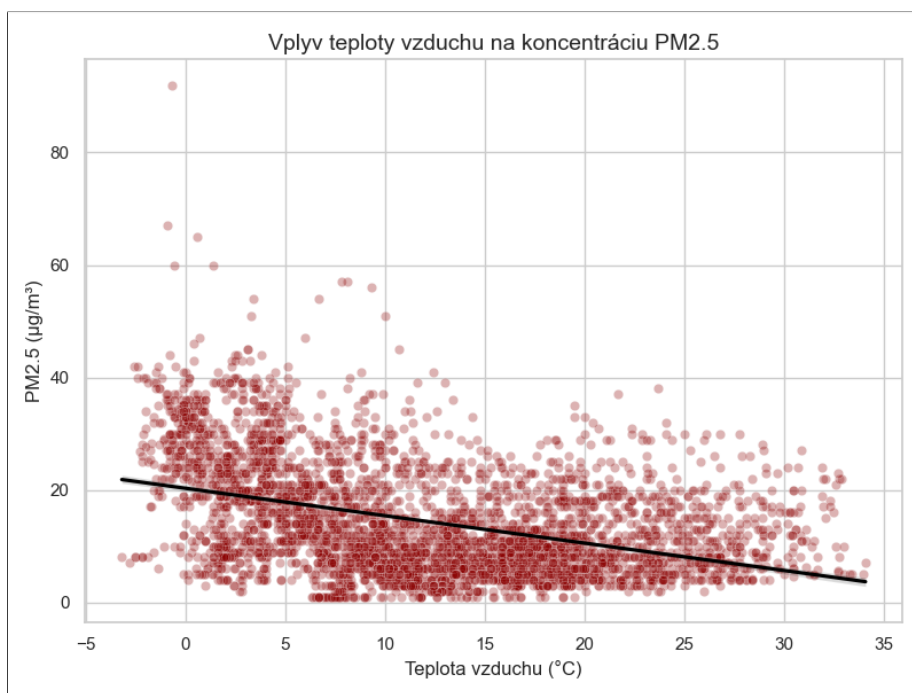
Obrázok 37: Vplyv intenzity dopravy na koncentráciu oxidov dusíka (NOx)



Obrázok 38: Vplyv rýchlosti vetra na koncentráciu prachových častíc PM10

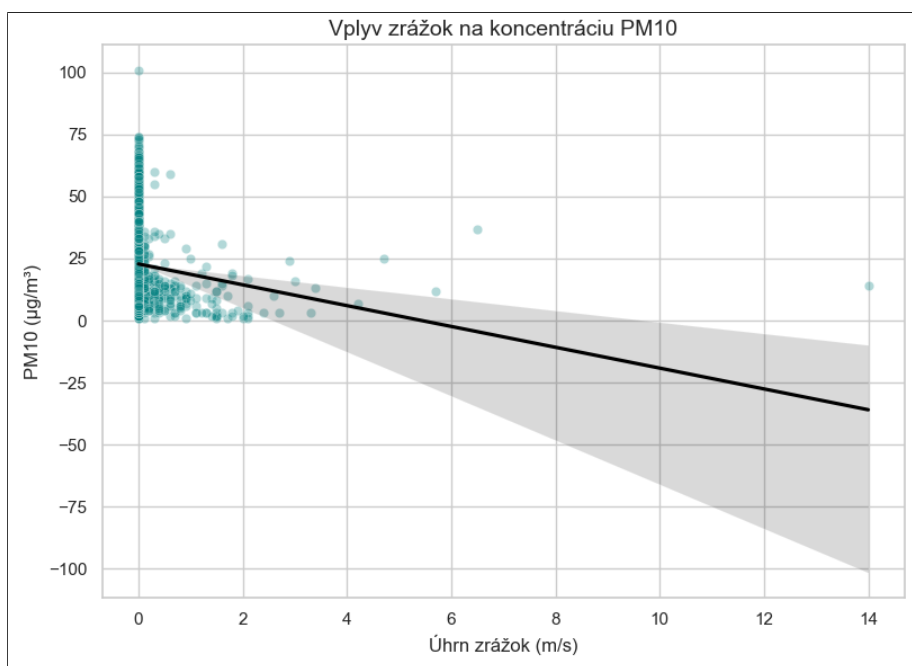
znečistenia odfúkne, rozptýli.

Zaujímavé zistenia priniesla analýza vplyvu teploty vzduchu. Najvýraznejšia negatívna korelácia ($r_s = -0,40$) bola zaznamenaná medzi teplotou a jemnými prachovými časticami PM2.5 na obrázku 39. Tento jav priamo indikuje sezónny charakter tohto znečistenia - s klesajúcou teplotou v jesenných a zimných mesiacoch rýchle stúpa koncentrácia PM2.5. Hlavnou príčinou je súbeh dvoch faktorov: začiatok vykurovacej sezóny (lokálne kúreniská produkujúce jemný prach) a častejší výskyt teplotných inverzií,



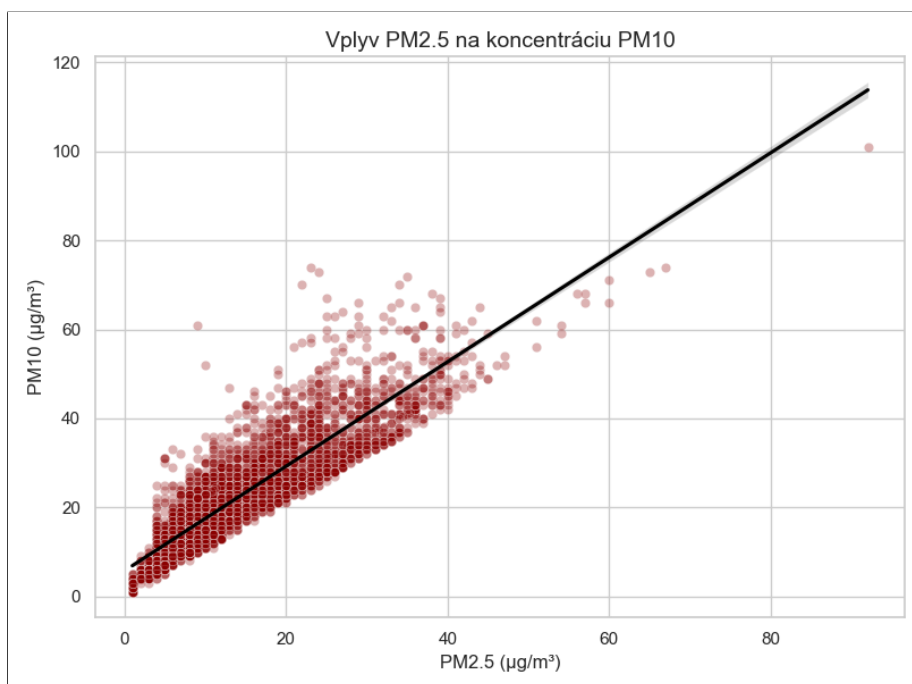
Obrázok 39: Vplyv teploty vzduchu na koncentráciu PM2.5

ktoré zamedzujú rozptylu znečistenia.



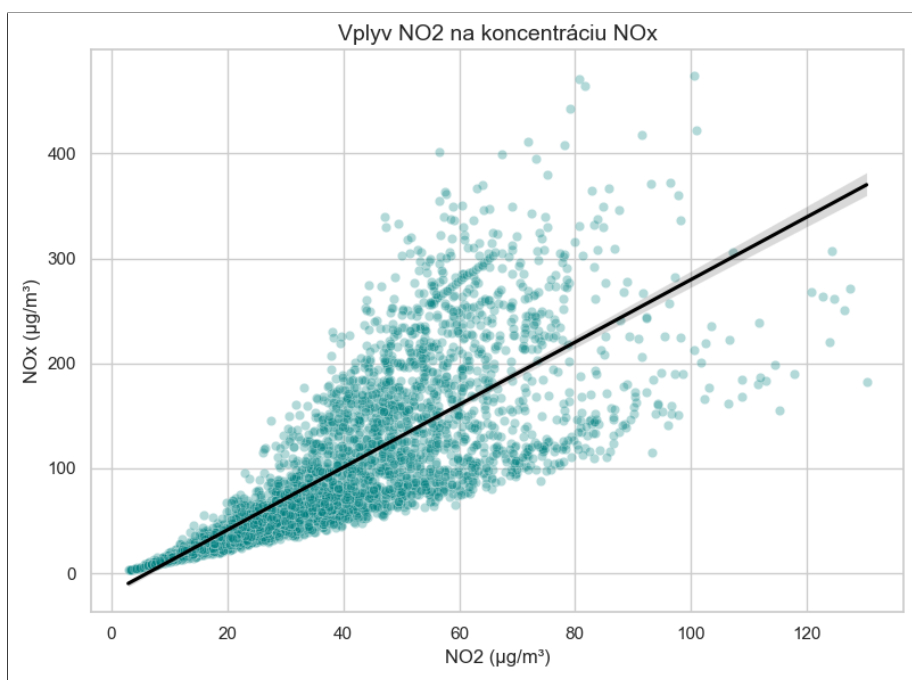
Obrázok 40: Vplyv zrážok na koncentráciu PM10

Úhrn zrážok vykazuje predpokladaný očisťujúci efekt, a to najmä vo vzťahu k prachovým časticiam (PM10, $r_s = -0,24$). Všimnime si obrázok 40. Tento proces, známy ako mokrá depozícia (vymývanie atmosféry), viaže pevné častice na kvapky vody a sťahuje ich na zemský povrch, čím znižuje ich koncentráciu vo vzduchu. Mierny pokles bol zaznamenaný aj pri NOx ($r_s = -0,10$).



Obrázok 41: Vplyv PM2.5 na koncentráciu PM10

Medzi jednotlivými znečisťujúcimi látkami sa potvrdila stredná až extrémne vysoká miera pozitívnej korelácie. Ide o dôsledok spoločných zdrojov znečistenia (najmä dopravy) a podobného správania sa v atmosfére. Extrémne silná korelácia bola zaznamenaná medzi časticami PM10 a PM2.5 ($r_s = 0,91$), čo je fyzikálne dané tým, že frakcia PM2.5 je hmotnostnou podmnožinou častíc PM10, vidíme to na obrázku 41.



Obrázok 42: Vplyv NO2 na koncentráciu NOx

Podobne silná väzba ($r_s = 0,89$) existuje medzi NO₂ a celkovými NO_x, keďže oxid dusičitý tvorí podstatnú časť všetkých oxidov dusíka v ovzduší, obrázok 42. Najnižšia z tejto skupiny, hoci stále významná ($r_s = 0,40$), bola korelácia medzi PM_{2.5} a NO₂, čo poukazuje na skutočnosť, že hoci majú čiastočne spoločný zdroj v doprave, dynamika ich šírenia a sekundárne zdroje (napríklad kúrenie pre PM_{2.5}) sa mierne líšia.

4.4.4 Sumarizácia analýz a kľúčové zistenia

Prevedená analýza potvrdila, že mestské prostredie funguje ako komplexný a dynamický systém, v ktorom sú doprava, počasie a kvalita ovzdušia úzko prepojené. Hlavné zistenia možno zhrnúť do nasledujúcich bodov:

- **Dominancia dopravy ako zdroja emisií:** Korelačná analýza exaktne potvrdila priamu závislosť medzi intenzitou dopravy a koncentráciami oxidov dusíka (NO_x a NO₂). V obdobiach ranných a poobedných špičiek, kedy intenzita dopravy dosahovala kritické hodnoty z hľadiska priepustnosti, dochádzalo k pravidelnému prekračovaniu bezpečného 24-hodinového limitu $25 \mu\text{g}/\text{m}^3$ pre NO₂, čo priamo znižuje kvalitu života v sledovanej lokalite.
- **Čistiaci efekt meteorologických podmienok:** Rýchlosť vetra sa ukázala ako najvýznamnejší negatívny korelát voči všetkým sledovaným polutantom. Analýza potvrdila doménové kritérium, že pri rýchlostiach nad 3 m/s dochádza k efektívnemu rozptylu emisií. Naopak, stavy bezvetria a nízkych rýchlostí vetra (pod 1 m/s) sú v kombinácii s vysokou intenzitou dopravy primárnou príčinou akumulácie znečistenia na úrovni ulice.
- **Sezónnosť a vplyv teplotných inverzií:** Distribúcia prachových častíc (PM₁₀ a PM_{2.5}) vykazuje silnú závislosť od ročného obdobia. Kým v letných mesiacoch boli koncentrácie relatívne stabilné, v decembri s nástupom vykurovacej sezóny a nižších teplôt došlo k nárastu extrémnych hodnôt presahujúcich kritickú hranicu $100 \mu\text{g}/\text{m}^3$. Tento jav potvrdzuje vplyv teplotných inverzií, kedy sa znečistenie uzatvára v prízemnej vrstve atmosféry.
- **Bimodálny charakter dopravného prúdu:** Histogramy intenzity dopravy odhalili dve výrazné maximá zodpovedajúce denným špičkám, čo potvrdzuje potrebu modelovať dopravu ako periodický dej. Priemerná rýchlosť vozidiel vykazovala poklesy práve v časoch najvyššej intenzity, čo korešponduje so vznikom kongescií a následným nárastom lokálneho znečistenia.

Tieto zistenia definujú kľúčové interakcie v systéme a potvrdzujú, že pre úspešnú predikciu nie je možné uvažovať o doprave izolovane bez meteorologického kontextu. Získané znalosti o koreláciách a distribučných vlastnostiach dát tvoria základ pre proces inžinierstva príznakov a následný návrh architektúry neurónovej siete.

5 Modelovanie na báze umelej inteligencie

V tejto kapitole sa zameriame na proces návrhu a implementácie predikčného modelu, ktorý využíva metódy umelej inteligencie na predikciu dopravnej situácie. Implementácia logiky vychádza z návodu na implementáciu siete GRU, pričom bola prispôbená pre potreby nášho konkrétneho prípadu [30]. Cieľom je podrobne opísať celý pracovný postup od prípravy dát pre model až po vyhodnotenie úspešnosti vytvoreného modelu.

Prvým krokom je transformácia údajov do formátu vhodného pre rekurentnú neurónovú sieť typu GRU. Tento proces zahŕňa import potrebných programových knižníc, načítanie datasetu, škálovanie hodnôt do jednotného intervalu a tvorbu sekvencií pomocou metódy posuvného okna. Dôležitou súčasťou je aj chronologické rozdelenie dát na tréningovú a testovaciu množinu tak, aby sa zachovala časová následnosť dejov.

V časti návrh a architektúra modelu opíšeme štruktúru modelu. Podrobne rozoberieme funkciu jednotlivých vrstiev - od vstupnej a rekurentnej vrstvy (GRU) cez regularizačnú vrstvu (Dropout) až po plne prepojené vrstvy určené na generovanie finálnej predikcie.

Časť tréning modelu opisuje parametre učiaceho procesu, výber optimalizátora a definovanie stratovej funkcie. Pozornosť budeme venovať aj technikám monitorovania tréningu, konkrétne mechanizmu EarlyStopping, ktorý slúži ako ochrana pred preučeníím modelu.

Záver kapitoly je venovaný overeniu schopností modelu na neznámych dátach. Úspešnosť predikcie budeme posudzovať pomocou kvantitatívnych štatistických metrík doplnených o vizuálnu analýzu distribúcie chýb.

5.1 Príprava dát pre modelovanie

Pred samotným spustením učiaceho procesu rekurentnej neurónovej siete GRU sme museli pripraviť dáta do vhodného formátu. Tento proces bol naprogramovaný v jazyku Python a pozostával z nadväzujúcich krokov: importovanie knižníc, načítanie dát, určenie vstupov a výstupov, škálovanie hodnôt vstupov a výstupov, tvorba sekvencií a chronologické rozdelenie datasetu.

1. **Importovanie knižníc:** Pre implementáciu modelu boli využité štandardizované open-source knižnice. Knižnica Pandas slúžila na manipuláciu s tabuľkovými dátami, NumPy pre prácu s poľami, joblib pre ukladanie a načítanie výstupov, Scikit-learn (sklearn) poskytla optimalizované algoritmy pre normalizáciu a samotná architektúra neurónovej siete bola postavená nad vysokoúrovňovým rozhraním Keras (s backendom TensorFlow). Bližšie informácie o verziách knižníc sa nachádzajú v prílohe Systémová dokumentácia.

```

1 import pandas as pd
2 import numpy as np
3 import joblib
4 from sklearn.preprocessing import MinMaxScaler
5 from keras.models import Sequential
6 from keras.layers import GRU, Dense, Dropout
7 from keras.callbacks import EarlyStopping

```

Zdrojový kód 1: Importovanie knižníc

2. **Načítanie dát:** Vstupom bol finálny dataset "final_dataset.csv". Keďže pracujeme s časovými radmi, je dôležité zachovať ich chronologickú postupnosť. Preto bol atribút určujúci čas merania "startDatetime" konvertovaný na natívny časový formát (datetime) a následne nastavený ako hlavný index celého dátového rámca (DataFrame, df). Takže ak by sa aj stalo, že by sme spravili chybu - zamenili jeden riadok za druhý - tak to týmto krokom opravíme.

```

1 df = pd.read_csv(r"final_dataset.csv")
2 df["startDatetime"] = pd.to_datetime(df["startDatetime"])
3 df.set_index("startDatetime", inplace=True)

```

Zdrojový kód 2: Načítanie dát

3. **Určenie vstupov a výstupov modelu:** Určili sme si vstupy a výstupy modelu. Na vstupe boli všetky atribúty nášho datasetu (počasie, kvalita ovzdušia a doprava), na výstupe boli údaje o doprave. Rozhodli sme sa tak, kvôli schopnosti identifikovať vzťahy modelu GRU.

```

1 feature_columns = df.columns.tolist()
2 target_columns = [
3     "vehicleIntensityInVehiclesPerHour",
4     "vehicleAverageSpeedInKilometersPerHour",
5 ]

```

Zdrojový kód 3: Určenie vstupov a výstupov modelu

4. **Škálovanie hodnôt:** Tento algoritmus lineárne transformoval všetky atribúty do zjednoteného intervalu $\langle 0, 1 \rangle$. Používa minimálnu a maximálnu hodnotu zo všetkých hodnôt atribútu. Ak je minimálna hodnota 0 a maximálna hodnota 100 a chceme škálovať hodnotu 50, tak dostaneme 0,5. Z dôvodu zachovania inverznej transformácie pre finálne vyhodnotenie boli vstupné premenné (features) a výstupné premenné (targets) škálované oddelene.

```

1 feature_scaler = MinMaxScaler()
2 target_scaler = MinMaxScaler()
3 scaled_features = feature_scaler.fit_transform(
4     df[feature_columns]
5 )
6 scaled_targets = target_scaler.fit_transform(
7     df[target_columns]
8 )

```

Zdrojový kód 4: Škálovanie dát

5. **Tvorba časových okien:** Klasické dopredné siete prijímajú dáta vo forme jednorozmerného vektora. Model GRU pre svoju činnosť vyžaduje 3D tenzor so štruktúrou [počet vzoriek, počet časových krokov, počet príznakov]. V našom prípade to je [3840 vzoriek, 24 časových krokov, 13 príznakov]. Počet vzoriek sa skrátil o 24 hodín, z 3864, čo je v poriadku, keďže potrebujeme 24 hodín na predpoveď 25-tej hodiny - nemôžeme predpovedať z prázdnych údajov. Na prevod údajov z tabuľky do požadovaného formátu sme použili funkciu `create_sequences`, ktorá implementuje metódu posuvného okna. Algoritmus iteruje naprieč datasetom a pre každý časový krok t vytvorí pole historických hodnôt s dĺžkou 24 hodín. Vďaka tomuto kroku si model pri predikcii každej novej hodnoty "pamätá" kontext celého predchádzajúceho dňa (24 hodín).

```

1 def create_sequences(features, targets, sequence_length):
2     inputs = []
3     outputs = []
4     for i in range(len(features) - sequence_length):
5         inputs.append(features[i : (i + sequence_length)])
6         outputs.append(targets[i + sequence_length])
7     return np.array(inputs), np.array(outputs)
8
9
10 input_data, output_data = create_sequences(
11     scaled_features, scaled_targets, 24
12 )

```

Zdrojový kód 5: Tvorba časových okien

6. **Rozdelenie dát:** Záverečnou fázou prípravy bolo rozdelenie dát na tréningovú a testovaciu množinu v zaužívanom pomere 80:20, aj kvôli menšiemu počtu dát. Pri časových radoch nesmie dôjsť k náhodnému premiešaniu vzoriek, pretože by sa tým zničila časová následnosť dejov - model by sa nemal naučiť tie hodnoty, ktoré bude predikovať. Kvôli tomu bolo rozdelenie vykonané chronologicky - prvých 80 % historických dát slúžilo na tréningovanie, kým posledných 20 % z konca roka predstavovalo neznámu budúcnosť, na ktorej sme následne model testovali. Po

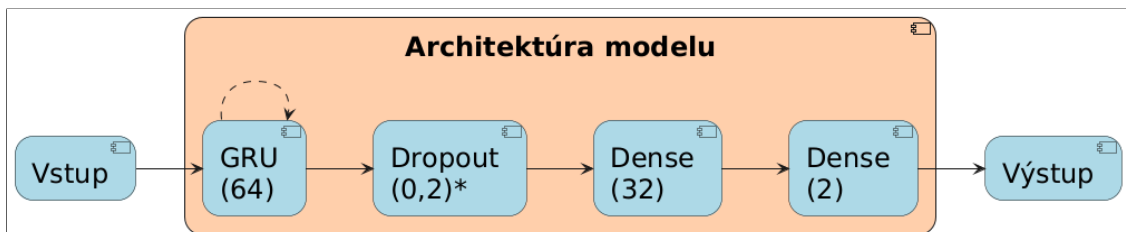
rozdelení dát vyzerá 3D tenzor pre trénováciu množinu nasledovne: [3072 vzoriek, 24 časových krokov, 13 príznakov].

```
1 def split(data, index: int):
2     return data[:index], data[index:]
3
4
5 split_index = int(len(input_data) * 0.8)
6 input_train_data, _ = split(
7     input_data, split_index
8 )
9 output_train_data, _ = split(
10     output_data, split_index
11 )
```

Zdrojový kód 6: Rozdelenie dát

5.2 Návrh a architektúra modelu

Po ukončení fázy prípravy a transformácie dát sme pristúpili k samotnému návrhu architektúry modelu. Pre implementáciu predikčného modelu bola zvolená knižnica Keras, ktorá beží nad knižnicou TensorFlow. Model je vytvorený ako Sequential, kde výstup každej vrstvy slúži ako priamy vstup pre nasledujúcu vrstvu. Hlavným cieľom návrhu bolo vytvoriť architektúru, ktorá dokáže predikovať dáta pre nasledujúcu hodinu.



Obrázok 43: Architektúra modelu

Na obrázku 43 je zobrazená architektúra modelu. Rozpísaná je nižšie.

1. **Vstupná vrstva:** Rozmer tejto vrstvy je definovaný štruktúrou pripravených sekvencií, čo v našom prípade predstavuje 24 časových krokov (historické okno 24 hodín) a 13 príznakov. Trénovacích vzoriek je 3072.
2. **Rekurentná vrstva (GRU):** Základom je vrstva GRU (Gated Recurrent Unit) s veľkosťou výstupu 64 neurónov. Vstupný rozmer tejto vrstvy (input_shape) je definovaný štruktúrou dát - to je popísané vo vstupnej vrstve. Parameter return_sequences bol ponechaný na predvolenej hodnote False. Znamená to, že vrstva spracuje celú 24-hodinovú sekvenciu a do ďalšej časti siete

odovzdá len finálny kontextový vektor reprezentujúci súhrnnú informáciu z celého predchádzajúceho dňa.

3. **Regularizačná vrstva (Dropout):** Na zvýšenie robustnosti modelu a elimináciu rizika preučenia (overfitting) sme za rekurentnú vrstvu zaradili vrstvu Dropout s koeficientom 0,2. Tento mechanizmus počas trénovacieho procesu náhodne deaktivuje 20 % neurónov, čím núti sieť učiť sa redundantné a všeobecnejšie reprezentácie dát. V našom prípade deaktivuje 20 % zo 64, takže približne 13 neurónov. Zostáva aktívnych 51 neurónov. Vzhľadom na to, že náš dataset obsahuje reálne senzorické údaje, ktoré môžu byť zaťažené šumom, je táto forma regularizácie nevyhnutná pre zachovanie schopnosti modelu zovšeobecňovať na neznámych testovacích dátach.
4. **Plne prepojená skrytá vrstva (Dense):** Pre ďalšiu nelineárnu transformáciu príznakov vyextrahovaných GRU vrstvou bola použitá vrstva Dense s 32 neurónmi. Ako aktivačná funkcia bola zvolená ReLU (Rectified Linear Unit).
5. **Výstupná vrstva (Dense):** Finálnu časť architektúry tvorí výstupná vrstva Dense s 2 neurónmi. Každý z týchto neurónov zodpovedá jednej z našich cieľových premenných hodinová intenzita dopravy a priemerná rýchlosť vozidiel. Keďže ide o regresnú úlohu, v tejto vrstve je použitá lineárna aktivácia, čo umožňuje sieti generovať hodnoty v kontinuálnom rozsahu potrebnom pre predikciu fyzikálnych veličín.

```
1 model_layers = [  
2     GRU(  
3         64,  
4         input_shape=(  
5             input_train_data.shape[1],  
6             input_train_data.shape[2]  
7         ),  
8         return_sequences=False,  
9     ),  
10    Dropout(0.2),  
11    Dense(32, activation="relu"),  
12    Dense(len(target_columns)),  
13 ]  
14 model = Sequential(model_layers)  
15 model.compile(optimizer="adam", loss="mse")
```

Zdrojový kód 7: Architektúra modelu

6. **Kompilácia modelu a nastavenie optimalizátora:** Model je kompilovaný s využitím optimalizačného algoritmu Adam (Adaptive Moment Estimation), najmä kvôli jeho efektívnosti pri riešení problémov hlbokého učenia [2][30]. Ako stratovú

funkcia (loss function) sme vybrali stredná kvadratická chyba (MSE), ktorá je citlivá na väčšie odchýlky a núti model minimalizovať rozdiely medzi predikovanou a skutočnou hodnotou intenzity a rýchlosti dopravy.

5.3 Tréning modelu

Po definovaní architektúry a kompilácii modelu nasleduje samotná fáza trénovania modelu. Tento proces sme realizovali na predpripravenej trénovacej množine, pričom proces učenia sa bol monitorovaný a optimalizovaný s využitím techník hlbokého učenia.

```
1 early_stopping = EarlyStopping(  
2     monitor="val_loss", patience=10, restore_best_weights=True  
3 )  
4  
5 history = model.fit(  
6     input_train_data,  
7     output_train_data,  
8     epochs=100,  
9     batch_size=32,  
10    validation_split=0.1,  
11    callbacks=[early_stopping],  
12    shuffle=False,  
13 )
```

Zdrojový kód 8: Trénovanie modelu

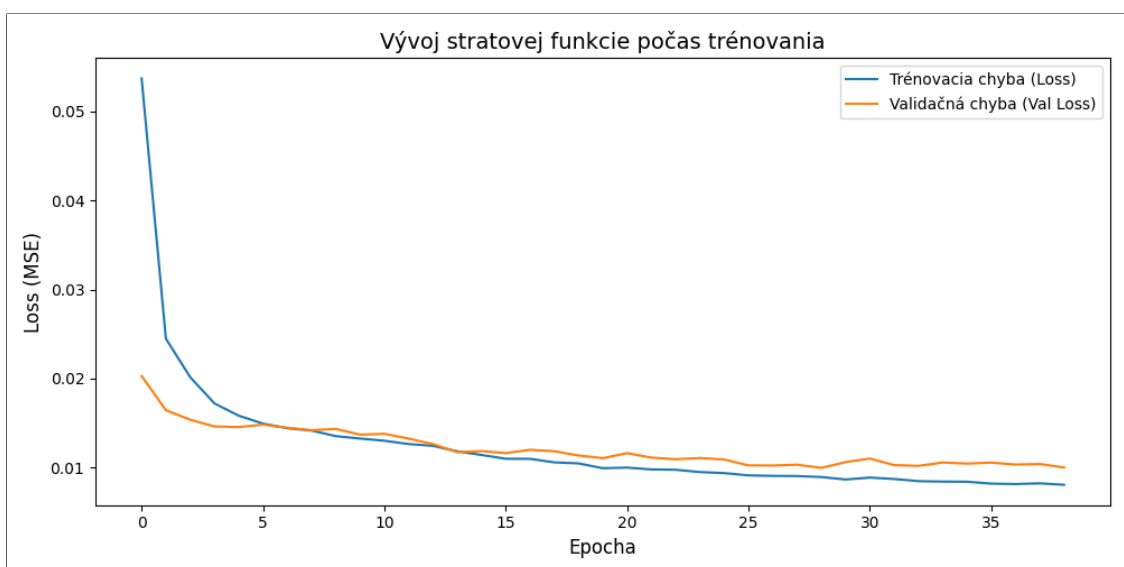
```
1 model.save("gru_model.keras")  
2 joblib.dump(feature_scaler, "gru_scaler_inputs.joblib")  
3 joblib.dump(target_scaler, "gru_scaler_outputs.joblib")  
4 joblib.dump(history, "gru_model_history.joblib")
```

Zdrojový kód 9: Uloženie modelu, škálovačov a histórie

1. **Parametre učiaceho procesu:** Proces učenia sa bol spustený na maximálny počet 100 epoch. Veľkosť dávky (parameter `batch_size`), ktorá určuje počet vzoriek spracovaných sieťou pred jednou aktualizáciou váh, bola nastavená na hodnotu 32. Pre potreby priebežného overovania schopnosti modelu generalizovať bol parameter `validation_split` nastavený na hodnotu 0,1. Znamená to, že z vyčlenenej trénovacej množiny sieť počas každej epochy použila 90 % dát na samotnú úpravu váh a zvyšných 10 % slúžilo ako nezávislá validačná vzorka na výpočet validačnej chyby.
2. **Ochrana pred preučeníím:** Jedným z najväčších rizík pri trénovaní hlbokých neurónových sietí na časových radoch je preučenie (overfitting), kedy sa model "naučí naspamäť" trénovacie dáta, no stratí schopnosť predikovať nové situácie.

Aby sme tomuto javu predišli, implementovali sme mechanizmus EarlyStopping. Tento mechanizmus bol nastavený tak, aby monitoroval hodnotu validačnej straty (val_loss). Parameter patience (slovensky trpezlivosť) bol nastavený na hodnotu 10. To znamená, že ak sa hodnota validačnej chyby nezlepšila počas desiatich po sebe nasledujúcich epoch, trénovací proces bol predčasne ukončený, aby sa zabránilo degradácii modelu. Zároveň bol aktivovaný parameter restore_best_weights=True, ktorý po zastavení trénovania automaticky obnovil konfiguráciu váh z tej epochy, v ktorej model dosiahol najnižšiu validačnú chybu.

3. **Analýza chybovej krivky:** Priebeh a úspešnosť optimalizácie váh počas trénovania vizualizuje chybová krivka.



Obrázok 44: Vývoj chybovej krivky počas trénovania

Graf na obrázku 44 zobrazuje vývoj strednej kvadratickej chyby (MSE) pre trénovaciu množinu (Trénovacia chyba / Loss) a validačnú množinu (Validačná chyba / Val Loss). Z priebehu kriviek môžeme vyvodit' niekoľko záverov:

- **Rýchla konvergencia:** V prvých epochách obe chybové krivky strmo klesajú, čo dokazuje, že model si rýchlo začal osvojovať základné vzorce a závislosti medzi doménovými príznakmi.
- **Stabilita modelu:** Počiatočný strmý pokles plynule prechádza do stabilizačnej fázy, kde sa hodnoty chýb ustália na minime.
- **Absencia preučenia:** Kľúčovým zistením je, že validačná krivka tesne kopíruje trénovaciu krivku a dochádza len k miernemu nárastu v neskorších fázach učenia sa.

Z týchto záverov môžeme potvrdiť, že kombinácia architektúry GRU, regularizácie Dropout a mechanizmu EarlyStopping úspešne ochránila model pred preučeníím sa.

Vďaka tejto optimalizácii sme získali predikčný model, ktorého reálnu úspešnosť môžeme následne kvantifikovať na testovacej množine, ktorú model zatiaľ nevidel.

5.4 Vyhodnotenie natrénovaného modelu

Po úspešnej optimalizácii váh bol model podrobený finálnej validácii na izolovanej testovacej množine (20 % dát z konca sledovaného obdobia), ktorú sieť počas tréningu nikdy nevidela. Aby bolo možné výsledky kvantifikovať a vizualizovať v reálnych fyzikálnych jednotkách, predikované aj skutočné hodnoty boli najskôr spätne transformované z normalizovaného intervalu $\langle 0, 1 \rangle$ pomocou škálovača.

```
1 output_predicted_scaled = model.predict(input_test_data)
2 output_predicted = target_scaler.inverse_transform(
3     output_predicted_scaled
4 )
5
6 output_test_real_data = target_scaler.inverse_transform(
7     output_test_data
8 )
```

Zdrojový kód 10: Spätne škálovanie hodnôt

Vyhodnotenie úspešnosti modelu sme vykonali dvoma spôsobmi: kvantitatívneho (pomocou štatistických metrík) a vizuálneho (grafickou analýzou distribúcie chýb).

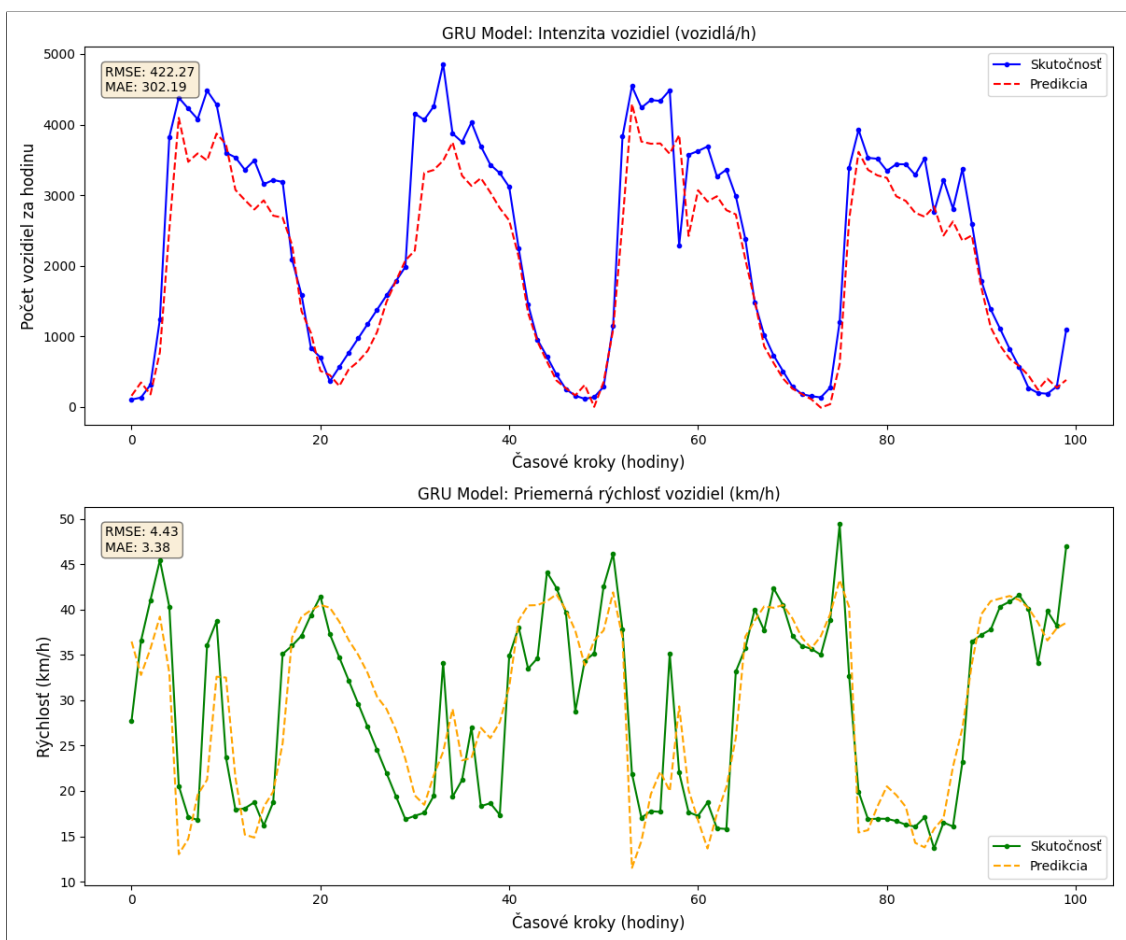
5.4.1 Kvantitatívne vyhodnotenie

Pre objektívne posúdenie presnosti sme pre obe výstupné premenné vypočítali dve kľúčové metriky: MAE podľa vzorca (2) a RMSE podľa vzorca (1).

- **Predikcia intenzity vozidiel:** Pri predikcii hodinovej intenzity vozidiel dosiahla hodnotu MAE 302 vozidiel za hodinu a hodnotu RMSE 422. Vzhľadom na prípustnosť a celkové objemy dopravy v analyzovanom úseku predstavuje táto chyba vysokú mieru presnosti.
- **Predikcia priemernej rýchlosti vozidiel:** V prípade predikcie rýchlosti dopravného prúdu dosiahol model MAE na úrovni 3 km/h a hodnotu RMSE 4. Takto nízka odchýlka potvrdzuje, že model je schopný s vysokou istotou predikovať plynulosť premávky a varovať pred vznikom dopravných zápch.

5.4.2 Vizuálna analýza predikcie

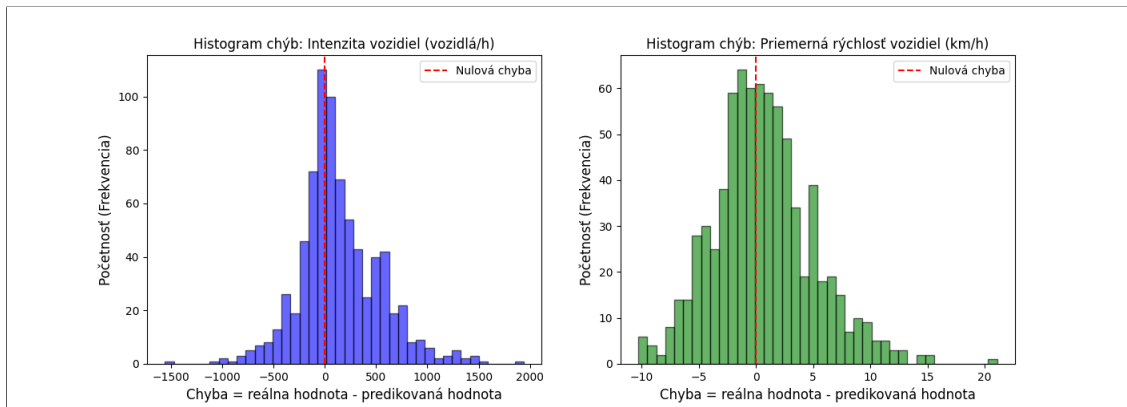
Kvantitatívne metriky poskytujú celkový obraz o presnosti, no pre detailné pochopenie správania sa modelu je nevyhnutná vizualizácia.



Obrázok 45: Porovnanie predikcie a skutočnosti intezity dopravy a priemernej rýchlosti vozidiel

Na obrázku 45, ktorý porovnáva skutočný (čierna krivka) a predikovaný (oranžová krivka) vývoj, je evidentné, že architektúra GRU excelentne zachytáva dynamiku časového radu. Model s presnosťou reaguje na zmeny a spoľahlivo predikuje vznik ranných aj poobedných dopravných špičiek. Mierne odchýlky sa vyskytujú iba pri extrémnych a neštandardných výkyvoch, čo je prirodzeným dôsledkom faktu, že model hľadá zovšeobecnený vzorec správania a nie presné anomálie.

Histogram na obrázku 46 vizualizuje početnosť jednotlivých odchýlok predikcie od reálnych hodnôt. Hodnoty vľavo od nulovej chyby znamenajú, že model predikuje hodnoty nad reálnu hodnotu a napravo, že model predikuje hodnoty pod reálnu hodnotu. Graf potvrdzuje, že chyby modelu majú rozdelenie blízke normálnemu rozdeleniu, pričom absolútne maximum početnosti je sústredené tesne okolo nuly. Výskyt výraznejších chýb smerom k okrajom grafu je menej podstatný.



Obrázok 46: Histogram distribúcie chýb modelu

Tieto výsledky potvrdzujú, že implementácia integrovaného GRU modelu v prostredí inteligentného mesta predstavuje spoľahlivý nástroj pre predikciu dopravy a získavanie znalostí pre včasné obmedzenie dopravy.

Záver

Hlavným cieľom predkladanej záverečnej práce bola analýza rozsiahlych údajov v doprave a modelovanie rizík v inteligentných mestách pomocou metód umelej inteligencie. Táto problematika je mimoriadne aktuálna, keďže rastúca urbanizácia vytvára obrovský tlak na infraštruktúru a životné prostredie, čo si vyžaduje prechod na proaktívne rozhodovanie založené na dátach.

V teoretickej časti sme špecifikovali kľúčové úlohy Smart Cities, ktoré úzko prepájajú mestskú dopravu, znečistenie ovzdušia a meteorologické javy. Pre úspešné riešenie tohto problému sme navrhli kritériá kvality, ktoré sme rozdelili do troch kategórií: doménové kritériá (časový predstih predikcie, výpočtová efektivita, nasaditeľnosť), kritériá kvality života (plynulosť dopravy a limity znečistenia ovzdušia stanovené WHO) a vyhodnocovacie metriky modelu (RMSE a MAE).

Praktická časť práce sa opierala o rozsiahly dataset z Legerovej ulice v Prahe, ktorý v sebe integroval údaje o doprave, počasi a kvalite ovzdušia za rok 2020. Surové dáta prešli dôkladným procesom čistenia a agregácie, pričom pre zachovanie kontinuálnej časovej osi, ktorá je nevyhnutná pre sekvenčné neurónové siete, bol dataset skrátený na druhú polovicu roka.

Následná exploratívna a korelačná analýza úspešne zosumarizovala zistenia a vzájomné korelácie v skúmanom systéme. Potvrdili sme priamu pozitívnu závislosť medzi intenzitou dopravy a emisiami oxidov dusíka (NO_x). Zároveň sa preukázal silný čistiaci efekt vetra na koncentráciu všetkých polutantov a negatívny vplyv klesajúcej teploty a teplotných inverzií v zimných mesiacoch, ktoré spôsobujú kritické hromadenie prachových častíc.

Následovala implementácia predikčného modelu s využitím umelej inteligencie, konkrétne rekurentnej neurónovej siete architektúry GRU (Gated Recurrent Unit). Model bol navrhnutý tak, aby na základe 24-hodinového historického okna dokázal predikovať dopravnú situáciu. Architektúra využívala vrstvu Dropout a mechanizmus EarlyStopping na ochranu pred preučeníím.

Vyhodnotenie natrénovaného modelu na nezávislej testovacej množine preukázalo jeho vysokú úspešnosť. Pri predikcii intenzity vozidiel model dosiahol MAE 302 vozidiel za hodinu a RMSE 422. Pre predikciu priemernej rýchlosti vozidiel bola dosiahnutá MAE na úrovni 3 km/h a RMSE 4. Vizuálna analýza potvrdila, že model s vysokou presnosťou zachytáva dynamiku časového radu a spoľahlivo predikuje vznik ranných aj poobedných dopravných špičiek.

Záverom možno konštatovať, že všetky body zadania záverečnej práce boli splnené. Navrhnutá metodika a vytvorený model umelej inteligencie predstavujú spoľahlivý nástroj pre predikciu vývoja dopravy, na základe ktorého je možné včasne indikovať rizikové situácie, akými sú dopravné zápchy s následným zhoršením kvality ovzdušia.

Zoznam bibliografických odkazov

- [1] Smart Cities Infrastructure. *Adtell Integration* [online]. [b. r.] [cit. 2026-04-23]. Dostupné na: <https://adtellintegration.com/smart-cities-infrastructure/>.
- [2] MANDAL, Komal. *Working of RNN, LSTM, GRU and Bidirectional Models with Practical Python Implementation* [online]. 2025 [cit. 2026-04-28]. Dostupné na: <https://medium.com/@komalmandal1907/working-of-rnn-lstm-gru-and-bidirectional-models-with-practical-python-implementation-d7f2abe2419e>.
- [3] SEZNAM.CZ, a.s., 2026. *Výrez mapy: Okolie Prahy* [mapa]. [cit. 2026-04-24]. Dostupné na: <https://mapy.com/s/judukusogo>.
- [4] Inteligentné mestá a obce. *Európska komisia* [online]. 2025 [cit. 2026-04-02]. Dostupné na: <https://digital-strategy.ec.europa.eu/sk/policies/smart-cities-and-communities>.
- [5] Definícia Smart Cities. *Smart City* [online]. 2022 [cit. 2026-04-02]. Dostupné na: <https://www.smartcity.gov.sk/definicia-smart-cities/index.html>.
- [6] Ranking of European medium-sized cities. *Rudolf Giffinger* [online]. 2007 [cit. 2026-04-02]. Dostupné na: https://www.smart-cities.eu/download/smart_cities_final_report.pdf.
- [7] CARAGLIU, A. et al. Smart cities in europe. *VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics, Serie Research Memoranda* [online]. 2009 [cit. 2026-04-03]. Dostupné na: <https://doi.org/10.1080/10630732.2011.601117>.
- [8] ZANELLA, A. et al. Internet of things for smart cities. 1. *Internet of Things Journal, IEEE* [online]. 2012 [cit. 2026-04-03]. Dostupné na: <https://doi.org/10.1109/JIOT.2014.2306328>.
- [9] Výpočet kapacít pozemných komunikácií. *Ministerstvo dopravy SR* [online]. 2024 [cit. 2026-04-23]. Dostupné na: https://www.ssc.sk/files/documents/technicke-predpisy/tp/tp_102-2024.pdf.
- [10] Transport and environment report 2022. *EUROPEAN ENVIRONMENT AGENCY (EEA)* [online]. 2023 [cit. 2026-04-03]. Dostupné na: <https://www.eea.europa.eu/en/analysis/publications/transport-and-environment-report-2022>.
- [11] TRNKOVÁ, Katarína, Drímal, Marek. Vybrané kapitoly zo zdravia a životného prostredia. *Belianum. Vydavateľstvo Univerzity Mateja Bela v Banskej Bystrici*

- [online]. 2023 [cit. 2026-04-03]. Dostupné na: <https://doi.org/10.24040/2023.9788055720807>.
- [12] LOOMIS, D. et al. The carcinogenicity of outdoor air pollution. *The Lancet Oncology* [online]. 2013 [cit. 2026-04-03]. Dostupné na: [https://doi.org/10.1016/S1470-2045\(13\)70487-X](https://doi.org/10.1016/S1470-2045(13)70487-X).
- [13] WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. *World Health Organization* [online]. 2021 [cit. 2026-04-03]. Dostupné na: <https://www.who.int/publications/i/item/9789240034228>.
- [14] Program ochrony powietrza dla województwa małopolskiego. *Kraków: Urząd Marszałkowski Województwa Małopolskiego* [online]. 2023 [cit. 2026-04-04]. Dostupné na: <https://powietrze.malopolska.pl/program-ochrony-powietrza/>.
- [15] SCULLEY, D. et al. Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems* [online]. 2015 [cit. 2026-04-20]. Dostupné na: https://www.researchgate.net/publication/319769912_Hidden_Technical_Debt_in_Machine_Learning_Systems.
- [16] Kvalita ovzduší v České republice v roce 2020. *ČHMÚ* [online]. 2021 [cit. 2026-04-18]. Dostupné na: <https://info.chmi.cz/rocnka/ko2020/>.
- [17] CHAI, T., Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* [online]. 2014 [cit. 2026-04-05]. Dostupné na: <https://gmd.copernicus.org/articles/7/1247/2014/gmd-7-1247-2014.pdf>.
- [18] CHO, K. et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation* [online]. 2014 [cit. 2026-04-28]. Dostupné na: <https://doi.org/10.3115/v1/D14-1179>.
- [19] KORETSKYI, Max. *How neural networks learn: deep dive into backpropagation and gradient descent* [online]. 2026 [cit. 2026-04-28]. Dostupné na: <https://indepth.dev/posts/1001/en/how-neural-networks-learn-backpropagation-gradient-descent/#the-vanishing-gradient-problem>.
- [20] ALAKE, Richmond. *A Data Scientist's Guide to Gradient Descent and Backpropagation Algorithms* [online]. 2022 [cit. 2026-04-28]. Dostupné na: <https://developer.nvidia.com/blog/a-data-scientists-guide-to-gradient-descent-and-backpropagation-algorithms/>.

- [21] LEO, Cristian. *The Math behind Adam Optimizer* [online]. 2024 [cit. 2026-04-28]. Dostupné na: <https://towardsdatascience.com/the-math-behind-adam-optimizer-c41407efe59b/>.
- [22] SUPERANNOTATE. *Activation functions in neural networks* [online]. 2024 [cit. 2026-04-28]. Dostupné na: <https://www.superannotate.com/blog/activation-functions-in-neural-networks>.
- [23] BAHETI, Pragati. *Activation Functions in Neural Networks [12 Types & Use Cases]* [online]. 2021 [cit. 2026-04-28]. Dostupné na: <https://www.v7labs.com/blog/neural-networks-activation-functions>.
- [24] CHAUDHARY, Mukesh. *Activation Functions: Sigmoid, Tanh, ReLU, Leaky ReLU, Softmax* [online]. 2020 [cit. 2026-04-28]. Dostupné na: <https://medium.com/@cmukesh8688/activation-functions-sigmoid-tanh-relu-leaky-relu-softmax-50d3778dcea5>.
- [25] LANG, Niklas. *Activation Functions in Neural Networks: How to Choose the Right One* [online]. 2024 [cit. 2026-04-27]. Dostupné na: <https://towardsdatascience.com/activation-functions-in-neural-networks-how-to-choose-the-right-one-cb20414c04e5/>.
- [26] BIRD, George. *Backpropagation Through Time For Networks With Long-Term Dependencies* [online]. 2021 [cit. 2026-04-28]. Dostupné na: <https://arxiv.org/html/2103.15589v3>.
- [27] UNIVERZITA MATEJA BELA V BANSKEJ BYSTRICI. *Súbor údajov o doprave, počasi a kvalite ovzdušia (UDAJE_Praha_Legerova ulica.xlsx)* [online]. Interný systém Moodle. 2023 [cit. 2025-11-03]. Dostupné na: <https://lms.umb.sk/mod/resource/view.php?id=237402>.
- [28] OPENDATA CHMI. *Súbor údajov o kvalite ovzdušia (OpenISKO_1969-2024.zip)* [online]. 2025 [cit. 2025-11-04]. Dostupné na: https://opendata.chmi.cz/air_quality/historical/OpenISKO_1969-2024.zip.
- [29] MARKECHOVÁ, Dagmar, Beáta STEHLÍKOVÁ a Anna TIRPÁKOVÁ. *Základy štatistiky pre pedagógov. Základy štatistiky pre pedagógov* [online]. Nitra: UKF, 2011 [cit. 2026-04-25]. ISBN 978-80-8094-899-3. Dostupné na: https://fzsp.truni.sk/sites/default/files/dokumenty/e-kniznica/e-ucebnice/Anal%C3%BDza-epidemiologick%C3%BDch-d%C3%A1t_frameset%20-%20EPSTRUNI%20%28HTML%29/data/media/aed/20120130_90405__1.pdf.

- [30] GEEKSFORGEEKS. *Gated Recurrent Unit Networks* [online]. 2026 [cit. 2026-04-27]. Dostupné na: <https://www.geeksforgeeks.org/machine-learning/gated-recurrent-unit-networks/>.

Prílohy

Príloha 1 Používateľská príručka

Príloha 2 Systémová dokumentácia

Príloha 1 Používateľská príručka

**UNIVERZITA MATEJA BELA V BANSKEJ BYSTRICI
FAKULTA PRÍRODNÝCH VIED**

**ANALÝZA ROZSIAHLYCH ÚDAJOV V DOPRAVE
A MODELOVANIE RIZÍK V INTELIGENTNÝCH
MESTÁCH POMOCOU METÓD UMELEJ
INTELIGENCIE**

Používateľská príručka

Banská Bystrica 2026

Bc. Adam Valašťan

Základné ustanovenia

Táto používateľská príručka má slúžiť ako návod na základné používanie vytrénovaného modelu. Náš výstup/model nemá používateľské rozhranie, keďže sme vytrénovali len model na predikciu. Budeme teda používať výpis z konzole na zobrazenie výsledkov.

Predikovanie hodnôt s vytrénovaným modelom

Pred pokračovaním sa uistíme, že sme si najprv prečítali systémovú dokumentáciu. Na predikovanie hodnôt použijeme skript s názvom "predict_on_sample.py", ktorý sa nachádza priamo v repozitári, ktorý máme stiahnutý. Skript obsahuje komentáre pre ľahké orientovanie sa v kóde.

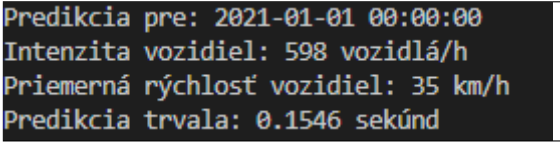
1. Spustenie skriptu:

```
1 python predict_on_sample.py
```

Ak nám vyhlási chybu, že sa Pytho nenašiel, tak môžeme skúsiť príkaz:

```
1 python3 predict_on_sample.py
```

2. Výstup skriptu:



```
Predikcia pre: 2021-01-01 00:00:00  
Intenzita vozidiel: 598 vozidlá/h  
Priemerná rýchlosť vozidiel: 35 km/h  
Predikcia trvala: 0.1546 sekúnd
```

Obrázok 47: Výstup skriptu v konzole

Model nám predikoval pre dátum 1.1.2021 a čas 00:00, že bude intenzita vozidiel 598 vozidiel za hodinu a priemerná rýchlosť vozidiel 35 km/h. Predikcia trvala 0,15 sekúnd.

Pre iné výsledky môžeme zmeniť údaje v .csv súbore "dataset/sample.csv".

Príloha 2 Systémová dokumentácia

**UNIVERZITA MATEJA BELA V BANSKEJ BYSTRICI
FAKULTA PRÍRODNÝCH VIED**

**ANALÝZA ROZSIAHLYCH ÚDAJOV V DOPRAVE
A MODELOVANIE RIZÍK V INTELIGENTNÝCH
MESTÁCH POMOCOU METÓD UMELEJ
INTELIGENCIE**

Systémová dokumentácia

Banská Bystrica 2026

Bc. Adam Valašťan

Zdrojové kódy

Zdrojové kódy, datasets a výstupy práce sú dostupné v repozitári autora záverečnej práce na platforme GitLab Univerzity Mateja Bela v Banskej Bystrici:

- <https://gitlab.umb.sk/avalastan/masters-thesis>

Systémové požiadavky

Uvedené systémové požiadavky boli použité pri vývoji a môžu sa líšiť od minimálnych systémových požiadaviek. Klientom je hrubý klient, t. j. beh programu sa nespolieha na vzdialený server.

- Procesor: Intel(R) Core(TM) i5-14500 2,6 GHz
- Grafická karta: Nvidia GeForce RTX 5060 Ti 16 GB
- Operačná pamäť: 48 GB
- Minimálny voľný úložný priestor na disku: 8 GB
- Stabilné internetové pripojenie
- Programovací jazyk Python: verzia 3.12.10, pip verzia 26.0.1

Inštalácia závislostí

Prv sa uistíme, že náš systém spĺňa systémové požiadavky. Na spustenie ukázkových skriptov je potrebné mať nainštalovaný programovací jazyk Python.

- Stiahneme si celý repozitár z adresy ako súbor zip.
- Po stiahnutí repozitára ako súbor .zip extrahujeme všetko do adresára.
- Adresár otvoríme v editore kódu, napríklad Visual Studio Code.
- Nainštalujeme závislosti pomocou príkazu:

```
1 pip install -r requirements.txt
```

Týmto by sme mali mať závislosti úspešne nainštalované.

Námety na vylepšenie

1. Vytvorenie používateľského rozhrania:

Pre vyskúšanie modelu je momentálne potrebné, aby používateľ mal k dispozícii počítač s nainštalovanými závislosťami. Navrhujeme vytvoriť webovú aplikáciu s jednoduchým používateľským rozhraním. Napríklad pomocou knižnice FastAPI, ktorá predvolene implementuje OpenAPI a aj jednoduché používateľské rozhranie Swagger.