VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV INFORMAČNÍCH SYSTÉMŮ
DEPARTMENT OF INFORMATION SYSTEMS

PREDIKCE DEMOGRAFICKÉHO VÝVOJE POPULACE S VYUŽITÍM STROJOVÉHO UČENÍ

PREDICTING POPULATION DEMOGRAPHIC TRENDS USING MACHINE LEARNING

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **ADRIÁN PONECHAL**
AUTHOR

VEDOUCÍ PRÁCE **Ing. RADEK HRANICKÝ, Ph.D.**
SUPERVISOR

BRNO 2025

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání bakalářské práce

161848

Ústav:	Ústav informačních systémů (UIFS)
Student:	**Ponechal Adrián**
Program:	Informační technologie
Název:	**Predikce demografického vývoje populace s využitím strojového učení**
Kategorie:	Data mining
Akademický rok:	2024/25

Zadání:

1. Nastudujte použitelné metody predikce vývoje lidské populace (regresní analýza, simulační techniky, strojové učení).
2. Vytvořte datovou sadu reprezentující demografický vývoj vybrané populace v průběhu let.
3. Po konzultaci s vedoucím navrhněte řešení, které na základě historických dat o populaci dokáže predikovat demografický vývoj vybrané populace (např. rozložení věku, vzdělání, zaměstnání v sektorech apod.).
4. Navržené řešení implementujte.
5. S pomocí vytvořené datové sady experimentálně ověřte funkčnost vašeho řešení.
6. Zhodnoťte dosažené výsledky a diskutujte možná rozšíření.

Literatura:

- Kim, Bumho, Chang-Gyu Lim, Seong-Ho Lee, and Yung-Joon Jung. "A Study on the Population Distribution Prediction in Large City using Agent-Based Simulation." *2021 23rd International Conference on Advanced Communication Technology (ICACT)*, s. 68-71. IEEE, 2021.
- Zhang, Yunong, Jinjin Wang, Qingkai Zeng, Heng Qiu, and Hongzhou Tan. "Near future prediction of European population through Chebyshev-activation WASD neuronet." *2015 Sixth International Conference on Intelligent Control and Information Processing (ICICIP)*, s. 134-139. IEEE, 2015.
- Ying, Jiang Mei, Xiong Li Ran, Hu Zhi Ding, and Niu Le De. "An Analysis of the Change in Population Size of Chinese Wa Nationality and Predictive Based on Census Data and Spectrum Population Prediction Software." *2014 Sixth International Conference on Measuring Technology and Mechatronics Automation*, s. 75-77. IEEE, 2014.
- Guo, Dongsheng, Yunong Zhang, Liangyu He, Keke Zhai, and Hongzhou Tan. "Chebyshev-polynomial neuronet, WASD algorithm and world population prediction from past 10000-year rough data." *The 27th Chinese Control and Decision Conference (2015 CCDC)*, s. 1702-1707. IEEE, 2015.
- Farooq, B., Bierlaire, M., Hurtubia, R. and Flötteröd, G., 2013. Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, s. 243-263.

Při obhajobě semestrální části projektu je požadováno:
Body 1 až 3.

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce:	**Hranický Radek, Ing., Ph.D.**
Konzultant:	Patzelt Petr, Ing.
Vedoucí ústavu:	Kolář Dušan, doc. Dr. Ing.
Datum zadání:	1.11.2024
Termín pro odevzdání:	14.5.2025
Datum schválení:	22.10.2024

Abstrakt

Populačný vývoj začína byť čím ďalej, tým viac skúmanou oblasťou. Narastajúci počet ľudí spôsobuje aj väčšiu spotrebu prírodných alebo štátnych zdrojov. Pri včasnom odhalení neočakávaných trendov je možné prijať opatrenia, ktoré týmto situáciam predídu. Aktuálny vývoj strojového učenia poskytol nástroje, ktoré dokážu takéto trendy odhaliť. V tejto práci boli preskúmané existujúce prístupy, ktoré sú používané na predikciu vývoja populácie, konkrétne štatistické metódy ARIMA a ARIMAX so strojovým učením (XGBoost, LightGBM, Random Forest, RNN, GRU, LSTM). Na základe dátovej sady, ktorá obsahuje dáta o viac ako 200 populáciách, boli navrhnuté a implementované dva hybridné modely: ARIMA + XGBoost a LSTM + XGBoost. Tieto modely vedia zachytiť trend vybranej populácie a predpovedať vývoj parametrov. V najlepších prípadoch modely predpovedali vývoj rozloženia pohlaví s chybou 0,005% (podľa metriky RMSE) a vekových skupín s chybou 0,1% (podľa metriky RMSE). Kombinácia modelov ARIMA + XGBoost je efektívna na presnejšie krátkodobé predikcie, zatiaľčo kombinácia LSTM + XGBoost vie generovať predpovede pre dlhší časový úsek, hoci s menšou presnosťou. Výsledné modely sú využívané na predikciu a úpravu demografických parametrov pri vytváraní populácií digitálnych respondentov, čím sa zaoberá firma Lakmoos AI, s.r.o.

Abstract

Population development is becoming an increasingly studied area. The growing number of people also leads to greater consumption of natural or state resources. By detecting unexpected trends early, it is possible to implement measures to prevent such situations. The recent development of machine learning has provided tools capable of identifying such trends. This work examined existing approaches used to predict population development, specifically the statistical methods ARIMA and ARIMAX with machine learning (XGBoost, LightGBM, Random Forest, RNN, GRU, LSTM). Based on a dataset containing information on more than 200 populations, two hybrid models combining ARIMA + XGBoost and LSTM + XGBoost were designed and implemented. These models can capture the trend of the selected population and predict the development of target parameters. In best cases, the models could predict gender distribution with an error of 0.005% (based on RMSE metric) and development of age group distribution with an error of 0.1% (based on RMSE metric). The ARIMA + XGBoost is effective for the short-term predictions, but the combination LSTM + XGBoost can generate longer predictions, although with less accuracy. The resulting models are used for predicting and adjusting demographic parameters in the creation of digital respondent populations, which is a focus area of the company Lakmoos AI, s.r.o.

Kľúčové slová

Predikcia demografického vývoja, časové rady, populácia, strojové učenie, rekurentné neurónové siete, LSTM, ARIMA, XGBoost, dolovanie dát

Keywords

Demographic development prediction, time series, population, machine learning, recurrent neural networks, LSTM, ARIMA, XGBoost, data mining

Citácia

PONECHAL, Adrián. *Predikce demografického vývoje populace s využitím strojového učení.* Brno, 2025. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Radek Hranický, Ph.D.

Predikce demografického vývoje populace s využitím strojového učení

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Radka Hranického, Ph.D. Ďalšie informácie mi poskytli zamestnanci spoločnosti Lakmoos AI s.r.o. Uviedol som všetky literárne pramene, publikácie a ďalšie zdroje, z ktorých som čerpal. Ako podporný nástroj pre rozdelenie populácii do skupín som použil nástroj ChatGPT od firmy OpenAI.

.....
Adrián Ponechal
13. mája 2025

Poďakovanie

Rád by som poďakoval vedúcemu práce Ing. Radkovi Hranickému, Ph.D za užitočné rady pri tvorení práce, konzultantovi Ing. Petrovi Patzeltovi a Bc. Jánovi Polišenskému za poskytnutie cenných informácii potrebných pre prácu s umelou inteligenciou. Ďalej chcem poďakovať priateľke a rodine za podporu pri písaní tejto práce.

Obsah

Zoznam obrázkov

Kapitola 1

Úvod

Demografické dáta sú veľmi užitočné najmä pri vykonávaní rozhodnutí. Vďaka sledovaniu vývinu populácie možno predpovedať rôzne priaznivé alebo nepriaznivé udalosti. Životná úroveň v populácii je závislá najmä na ľuďoch patriacich do produktívnej skupiny, teda ľudí vo veku od 15–64 rokov.

Predikcia vývoja populácie poskytuje cenné informácie na hodnotenie jej budúcich potrieb spoločnosti. Tieto informácie využívajú štátne orgány k plánovaniu štátneho rozpočtu, zabezpečeniu sociálnych a zdravotníckych služieb. Okrem štátnych úradov sú tieto štatistiky užitočné na regionálnej úrovni na plánovanie mestskej infraštruktúry alebo na firemné účely v oblasti marketingu. Demografické dáta sú užitočné aj pri rekonštrukcii populácie, pretože populačné dáta sú viazané na demografické dáta. Týmto prípadom sa zaoberá firma Lakmoos AI, s.r.o., ktorá predpovedané dáta používa na vytvorenie budúcej populácie.

Cieľom tejto práce je navrhnúť a implementovať prediktívny model na odhad budúceho demografického vývoja na základe historických údajov. Jeho hlavným cieľom je predpovedať vekové rozloženie ľudí v populácii. Model využíva viacstupňový prístup, kde sú najprv predpovedané vplyvné premenné a následne samotné demografické ukazovatele (rozloženie pohlavia a vekových skupín). Použité dáta zahŕňajú časové rady populácií z viac než 200 krajín. To má pomôcť odhaliť spoločné trendy na vývoj akejkoľvek ľudskej populácie. Pomocou naučených vzťahov by potom mal byť model schopný užšej špecifikácie na predikciu konkrétnej populácie pri poskytnutí jej predošlého vývoja.

Táto práca využíva prístup strojového učenia v kombinácii s dátovou sadou, ktorá obsahuje údaje o viacerých populáciách. Cieľom je zistiť, či môžeme vytvoriť univerzálny model schopný predpovedať demografické ukazovatele pre rôzne regióny alebo krajiny na základe historických dát. Okrem toho boli v tejto práci porovnané kombinácie štatistiky a strojového učenia a prístupu založeného výhradne na strojovom učení.

Táto práca pozostáva z 10 kapitol. V kapitole 2 sú zadefinované základné pojmy na lepšie pochopenie zamerania práce. V kapitole 3 je priblížený obraz už o existujúcich metódach, ktoré sa používajú na analýzu a predpoveď demografických trendov vývinu populácie. Najzaujímavejšie z nich sú popísané v kapitole 4. Táto kapitola obsahuje potrebné znalosti na pochopenie použitých metodík. Kapitola 5 sa zaoberá popisom, analýzou a spracovaním dát. Na ich základe je vytvorená dátová sada, ktorá slúži ako vstup pre jednotlivé prediktívne metódy. Návrh riešenia je popísaný v kapitole 6. S použitím vytvorenej dátovej sady sú v kapitole 7 porovnané viaceré metódy používané na predikciu stavu populácie. Najlepšie hodnotené metódy sú zakomponované do výsledného modelu. Kapitola 8 vysvetľuje spôsob implementácie nástroja, spolu s použitými nástrojmi, knižnicami a abstraktným popisom programu. Experimentálne vyhodnotenie modelu je popísané v kapitole 9.

Kapitola 2

Demografia

Demografia je veda zameraná na štúdium ľudskej populácie. Popisuje a zhodnocuje jej aktuálny stav a snaží sa predpovedať jej budúci vývoj [31]. Demografické dáta sú využívané najmä pri tvorbe informovaných rozhodnutí vo vládnucich pozíciách. Študovanie vývoja populácie umožňuje vytvoriť plán tak, aby boli uspokojené ľudské potreby pre život čo najviac členov v populácii - napríklad spotreba jedla a oblečenia [16]. V politických systémoch tieto dáta pomáhajú pri hospodárení so štátnym rozpočtom. Ak systém v danej krajine obsahuje dôchodkový systém, tak na základe predikcie rozdelenia skupín v populácii môžeme očakávať nárast alebo pokles výdavkov, čo môže ovplyvniť daňový systém v krajine.

2.1 Populácia

V kontexte demografie je populácia určitý počet ľudí, ktorý žije na určitom území [31]. Populáciu možno rozdeliť do rôznych skupín podľa určitého parametru, ako je vek, pohlavie, ekonomický status, vzdelanie a podobne. Môže byť lokálna, regionálna, štátna alebo globálna [12]. Demografia skúma tri základné procesy v populácii - plodnosť, úmrtnosť a migráciu. Zmena veľkosti, rozloženia alebo distribúcie v populácii závisí na jednom alebo viacerých z týchto troch procesov [60].

2.2 Modelované skupiny

Každá populácia potrebuje pre svoj vývoj a prežitie určité zdroje, ktoré jedinci v populácii vytvárajú a spotrebúvajú. Rozlišované budú hlavne tieto dve skupiny: skupina producentov a skupina závislých osôb [16].

Producentov možno chápať ako skupinu ľudí, ktorí generujú zdroje. Do tejto skupiny budú patriť pracujúci ľudia, platcovia daní, podnikatelia, živnostníci.

Skupina závislých osôb alebo neproducenti sú súbor ľudí, ktorí zdroje primárne spotrebúvajú. V kontexte tejto práce do tejto skupiny patria deti do 15 rokov a ľudia starší ako 65 rokov.

Na získanie hlavných informácií na vytvorenie obrazu o týchto skupinách je táto práca zameraná na predikciu dvoch konceptov. Prvý sa týka absolútneho počtu ľudí a rýchlosti rastu populácie a distribúcie mužov a žien v populácii. Ak sa rodí veľa detí, tak sa rozširuje skupina neproducentov, po dosiahnutí dospelosti však rozšíria skupinu producentov. Po dosiahnutí veku 65 rokov a vyššie opäť prejdú z produktívnej do neproduktívnej skupiny. Tento koncept priamo súvisí s druhou dôležitou oblasťou v demografii a tou je starnutie.

2.3 Starnutie populácie

Starnutie populácie je fenoménom 21. prvého storočia. Je zapríčinené klesajúcou plodnosťou a rastúcou dĺžkou života, čo má za následok postup veľkých kohort do vyššieho veku, čo spôsobuje rastúci podiel starších ľudí v populácii [9]. Starnutie populácie sa ukázalo ako problém, ktorý sa vyskytuje hlavne vo vyspelých regiónoch [28]. Nezanedbateľným vplyvom starnutia je pohlavné rozloženie obyvateľstva. Medzi staršou skupinou obyvateľstva majú prevahu príslušníci ženského pohlavia a to v prípade rozvinutých aj nerozvinutých krajín [9].

2.4 Dôležité pojmy

V tejto sekcii sú zadefinované a vysvetlené jednotlivé pojmy, ktoré budú použité pri tvorbe dátovej sady. So štruktúrou a starnutím populácie súvisia nasledujúce parametre:

- **Hrubá pôrodnosť** (v angličtine *crude birth rate*) je základna miera vyjadrenia plodnosti. Bežne sa uvádza vzhľadom na 1000 ľudí, kde sa počíta ako pomer pôrodov k celkovému počtu ľudí v populácii krát 1000 [27].
- **Plodnosť alebo miera plodnosti** (v anglictine *fertility rate*) je presnejším ukazovateľom ako hrubá miera pôrodnosti, pretože vyjadruje množstvo narodených detí v prepočte na ženu [48]. Možno ju chápať ako počet pôrodov, ktoré by žena mala, keby porodila deti podľa aktuálnej miery plodnosti podľa veku [27].
- **Migrácia** je definovaná ako zmena zvyčajného miesta pobytu, ktorá zahŕňa prekročenie administratívnej hranice, ako je obec, kraj, provincia, štát alebo medzinárodná hranica [27]. Pre vyjadrenie migrácie v tejto práci bude použitý ukazovateľ čistej migrácie. Čistá migrácia (v angličtine *net migration*) je rozdiel medzi mierou imigrácie a mierou emigrácie [27].
- **Hrubá úmrtnosť** (v angličtine *crude death rate*) sa podobne ako hrubá pôrodnosť uvádza v prepočte na 1000 ľudí. Jedná sa o počet úmrtí na konkrétnom mieste v danom čase za časovú jednotku (typicky rok). Vypočíta sa ako pomer úmrtí k celkovému počtu ľudí v populácii krát 1000 [27].
- **Priemerná dĺžka života** (anglicky *life expectancy at birth*) udáva priemerný počet rokov, ktoré môže osoba očakávať že prežije, na základe súčasnej miery úmrtnosti. Je to ukazovateľ celkového zdravia a kvality života v populácii [48].
- **Pomer vekovej závislosti** (v angličtine *age dependency ratio*) vyjadruje pomer nepracujúcich (mladých a starých) k produktívnemu veku obyvateľov. Tento pomer udáva ekonomické zaťaženie výrobného segmentu obyvateľstvo [48].

Kapitola 3

Súvisiaca práca

Predikcia populácie je komplexná téma a môže prebiehať na rôznych úrovniach abstrakcie (2.1). Na dosiahnutie určitej presnosti je potrebné populáciu rozčleniť na skupiny podľa veku, geografického regiónu a v niektorých prípadoch môže byť vyžadované ďalšie rozčlenenie [12]. V tejto kapitole budú spomenuté rôzne prístupy a metódy, ktoré riešia problém predikcie vývoja populácie. Na predikciu populácie existujú 2 hlavné prístupy. Jeden prístup zahŕňa tvorbu predikcie na základe jedného modelu a druhý je s využitím zloženého modelu [50].

3.1 Jednoduché modely

Za jednoduché budú považované tie modely, ktoré využívajú jednu metódu alebo prístup pre predikciu. Príkladom môžu byť štatistické metódy alebo samostatné modely strojového učenia. V štúdii, ktorú publikovala Australian National University v Canberre, ACT 0200, zameranej na predikciu populácie, rozdelili metódy na extrapolatívne, metódy s očakávaniami a na štrukturálne modelovanie (zložené modely) [12].

3.1.1 Extrapolatívne metódy

Extrapolatívne metódy vychádzajú z predpokladu, že výsledok v budúcnosti je závislý od vývoja v minulosti. Najviac využívanou metódou z tejto kategórie sa pre tento účel používa ARIMA 4.1.1 [12]. Jedná sa o metódu, ktorá sa používa na analýzu časových radov.

Podľa štúdie, ktorá sa zaoberá predikciou demografických indikátorov, ARIMA produkuje validné výsledky aj s nízkym vzorkom dát. Napriek tomu sa veľa vedcov prikláňa k využitiu novodobých riešení s využitím modelov strojového učenia, pretože vykazujú lepšie výsledky pri použití na väčších dátových sadách [25].

3.1.2 Modely strojového učenia

Pred vznikom strojového učenia boli na predikciu používané štatistické metódy. Problém týchto modelov bol ten, že často mali problém odhaliť komplexitu populačnej dynamiky [6].

Štúdia zameraná na vývoj modelu na predpoveď demografických ukazateľov v Azerbajdžane preukázala, že modely strojového učenia preukázali vysokú účinnosť oproti štatistickým metódam, ktoré produkovali skreslené výsledky [25]. Táto konkrétna štúdia porovnávala 4 rôzne metódy s využitím strojového učenia, medzi ktoré patrili lineárna regresia,

rozhodovací strom, Random forest a algoritmus K najbližších susedov (KNN). Z výsledkov vyplýva, že najpresnejšie výsledky dosahovali KNN a random forest. Taktiež je v štúdii uvedené, že je nutné testovať a skúšať iné modely, ktoré sú vhodnejšie na danú problematiku.

S pokračujúcim vývojom boli v mnohých štúdiách použité neurónové siete 4.3.2 a modely hlbokého učenia 4.4. Tieto modely sú náročné na trénovanie, hlavne z pohľadu množstva dát. Toto potvrdzuje študent Maznichenko Lev Vladyslavovych vo svojej bakalárskej práci, kde aplikuje siete LSTM 4.4.2 oproti rôznym variantom modelov ARIMA. Výsledky naznačujú, že siete LSTM neboli vhodným nástrojom pre tento problém a možnou príčinou bol nedostatok trénovacích dát [48].

Naopak druhá štúdia na predikciu populácie v Iraku uvádza, že použitie rekurentných neurónových sietí produkuje lepšie výsledky oproti konvenčným metódam a to dokonca aj pri malom vzorku dát [6].

3.2 Zložené modely

Tento prístup zahŕňa kombináciu viacerých jednoduchých metód a ich spojením do jedného celku. Ukázalo sa, že tento prístup je presnejší ako využitie klasickej jednej metódy. To potvrdila štúdia, ktorá vytvorila kombináciu neurónovej siete a modelu ARIMA. Štúdia uviedla, že tento kombinovaný model zvýšil presnosť predikcie o 9,7% oproti samostatnému využitiu jednotlivých modelov [82].

Podobným spôsobom v štúdii zaoberajúcej sa subpopulačnými národnými prognózami navrhli populačný fúzny transformér (PFT). Tento model je založený na dočasnom fúznom transforméri (TFT) a novom bloku hlbokej hradlovej zvyškovej siete (DGRN). PFT model vykazoval menšiu mieru chybovosti oproti TFT a v štúdii autori uviedli, že kombinácia PFT s ďalšími modelmi prispela k ďalšej redukcii chyby [4].

Zložený model sa môže skladať výhradne zo štatistických prístupov. V roku 2011 využili výskumníci Li Juan a Wei-jie Liu na predikovanie veľkosti populácie v Číne Grey-Markov model 4.1.3. Podľa ich štúdie je tento prístup efektívny na predikciu. Grey model je použitý na predpovedanie sekvencie a následne je použitý Markovov polynóm na revíziu údajov [33].

Kapitola 4

Prediktívne metódy vhodné na predikciu populácie

Táto kapitola podrobnejšie popisuje najviac používané a relevantné metódy používané na generovanie predpovedí vývinu populačných parametrov.

4.1 Štatistické modely

V tejto sekcii sú popísané štatistické modely, ktoré boli najčastejšie používané na predikciu populácie. Tieto modely slúžili často nielen ako základ na porovnávanie výkonu modelov strojového učenia, ale aj ako vylepšovacie prostriedky na zlepšenie presnosti v kombinácii s nimi.

4.1.1 ARIMA

ARIMA(p,d,q) model je jednou z najvyspelejších metód na predikciu časových radov. Využíva minulé a súčasné hodnoty premenných na poskytovanie presných krátkodobých odpovedí. Je vhodná, ak sú údaje v časovom rade na sebe závislé [55]. ARIMA sa skladá z troch zložiek: autoregresia *(AR(p))*, diferenciácia *(I(d))* a kĺzavý priemer *(MA(q))* [6]:

- Autoregresívna zložka *AR(p)* zložka vyjadruje vzťah medzi aktuálnou hodnotou časového radu x_t a jeho predchádzajúcimi hodnotami $(x_{t-1}, x_{t-2}, \ldots, x_{t-p})$. Parameter p určuje počet predchádzajúcich hodnôt zahrnutých v modeli [72].
- Zložka *I(d)* reprezentuje počet potrebných diferenciácií na dosiahnutie stacionarity časového radu. Parameter d udáva počet vykonaných diferenciácií [72].
- Zložká kĺzavého priemeru *MA(q)* predpokladá, že aktuálna hodnota časového radu je lineárnou kombináciou bieleho šumu (náhodných chýb) z predchádzajúcich období. Na rozdiel od autoregresie, model MA(q) vychádza z predchádzajúcich chýb predikcie. Parameter q udáva počet zahrnutých predchádzajúcich chýb [72].

Kombinácia týchto troch zložiek potom tvorí model, ktorého všeobecná forma je vyjadrená nasledovne:

$$y_t = c + AR(p) + MA(q) + \in_t \tag{4.1}$$

kde p a q sú rády autoregresnej zložky a zložky kĺzavého priemeru. d je počet vykonaných diferencií [59].

ARIMA sa dá použiť aj ako štatistická metóda na test stacionarity [18], pretože táto metóda predpokladá stacionaritu a linearitu [6]. Avšak cieľom tejto práce je skúmať vplyv viacerých faktorov na predikovaný parameter, nielen samotná predikcia vybraného parametru. ARIMA bude slúžiť ako jeden zo spôsobov validácie navrhnutého modelu.

4.1.2 ARIMAX

Alternatívou k tejto metóde sú jej dostupné modifikácie ako ARIMAX [57] , ktorá pre predikciu vstupnej sekvencie príjma aj iné časové rady ako vstupné premenné:

$$Y_t = C + \nu(B)X_t + N_t \tag{4.2}$$

kde Y_t je výstupná sekvencia (závislá premenná), X_t je vstupná sekvencia (nezávislá premenná), C je konštanta, N_t je stochastická porucha (séria šumu systému, nezávislá od vstupnej série). $\nu(B)X_t$ je prenosová funkcia, ktorá umožňuje premennej X ovplyvniť premennú Y. B je operátor spätného posunu.

Medzi ďalšie alternatívy prístupy založené na strojovom učení, ktoré môžu byť vhodnejšou alternatívou pre nelineárne vzory alebo zložité závislosti [6].

4.1.3 Grey-Markov model

Grey-Markov model je tvorený kombináciou dvoch modelov - Greyovho modelu a Markovovho reťazca [33].

Grey model

GM(1,1) je model na predikciu časových radov. Jeho výhodou je, že nevyžaduje veľké množstvo dát [40]. Všeobecne GM potrebuje aspoň 4 vzorky [35]. Je vhodný na krátkodobé predikcie sekvencií [33].

Markovov reťazec

Markovov reťazec je dobrý v predpovedaní dlhodobých sekvencií pre náhodné a väčšie nestále údaje, ale prognostický objekt Markovovho reťazca vyžaduje stabilné údaje. Tieto údaje môže poskytnúť GM(1,1) [33].

4.2 Regresná analýza

Regresná analýza sa zaoberá hľadaním a analýzou vzťahov medzi nezávislými a závislými premennými. Na základe týchto vzťahov je možné pomocou vytvoreného regresného modelu vytvárať predikcie. Všeobecný popis regresného modelu je daný rovnicou [68]:

$$y = \alpha + \beta_1 x_1 + e \tag{4.3}$$

α je priesečník s osou y, β je koeficient nezávislej premennej a e je term zobrazujúci chybu rovnice [68, 17].

4.2.1 Jednoduchá regresia

Jednoduchá (lineárna) regresia je metóda, ktorá sa snaží zredukovať všetky tréningové vzorky do jedného parametru priamky [22]. Ak je tréningová vzorka závislá iba na jednej nezávislej premennej, nazývame ju jednoduchá alebo lineárna regresia:

$$y = \alpha + \beta x \tag{4.4}$$

Parametre a a b je možné získať rôznymi spôsobmi. Najčastejší postup získania týchto parametrov je metóda najmenších štvorcov (OLS) [68]. Niekedy sa táto metóda nazýva aj gradientný zostup. Jej princíp pozostáva z toho, že na začiatku sú hodnoty α a β odhadnuté a následne je hodnota týchto premenných iteratívne aktualizovaná podľa funkcie strednej absolútnej chyby MSE (viď sekciu 7.1.2) [22].

4.2.2 Viacnásobná regresia

Viacnásobná regresia je regresná metóda, ktorá obsahuje viac než jednu nezávislú premennú. Predpis na zahrnutie viac nezávislých premenných [68]:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 ... \beta_n x_n + e \tag{4.5}$$

Princíp trénovania viacnásobnej regresie je rovnaký ako u jednoduchej regresie.

4.2.3 Predpoklady pre použitie regresie

Aby bolo možné dosiahnuť validné výsledky pomocou regresnej analýzy, regresný model musí spĺňať určité predpoklady. Medzi kľúčové predpoklady patrí linearita, nezávislosť, homoskedasticita a normalita [68, 17].

- Linearita hovorí o tom, že vťahy medzi nezávislými a závislými premennými sú lineárne. Lineárne vzťahy môžu byť pozitívne alebo negatívne.
- Nezávislosť sa týka dostupných dátových vzorkov. Každé pozorovanie dátového vzorku musí byť nezávislé od predošlých pozorovaní.
- Predpoklad homoskedasticity vyjadruje, že rozptyl chyby je konštatný. Ak rozptyl konštantý nie je, tak regresná analýza môže byť mylná a výsledky sú potom nepresné.
- Predpoklad normality zaisťuje, že distribúcia chyby podlieha normálnemu rozdeleniu so stredom v nule. Teda najčastejšie chyby sa blížia k nule a platí, že čím väčšia chyba, tým je menšia pravdepodobnosť jej výskytu.

4.3 Prístupy s využitím strojového učenia

Narozdiel od štatistických prístupov majú prístupy strojového učenia schopnosť pracovať s viacdimenzionálnymi dátami. V týchto dátach sú algoritmy strojového učenia schopné identifikovať rôzne vzory, ktoré sa v dátach nachádzajú [17]. Získané vzory môžu byť následne použité na prediktívnu analýzu.

4.3.1 Stromové prístupy

Algoritmy patriace do tejto kategórie vytvárajú štruktúry, ktoré sa nazývajú stromy. Jedná sa o matematické hierarchické štruktúry, ktoré možno popísať ako množinu prepojených uzlov [22]. Pre tieto uzly platí, že každý uzol, okrem koreňového uzla, má práve jedného rodiča a je spojený s dvomi a viac synovskými uzlami. Každý synovský uzol možno považovať za podstrom hlavného stromu. Koncové uzly sa nazývajú listy a obsahujú výslednú hodnotu [17].

Rozhodovací strom

Hlavnou výhodou rozhodovacích stromov je ich malá výpočetná náročnosť, jednoduchosť a jednoduchá interpretabilita [66]. Taktiež sú veľmi univerzálne - každý algoritmus môže byť prevedený do formy rozhodovacieho stromu, sú tak vhodné na klasifikačné aj regresné problémy [22].

Tvorba konštatného regresného rozhodovacieho stromu

Tvorba regresného rozhodovacieho stromu je proces, ktorý zahŕňa rekurzívne rozdelenie (splitting) a viacnásobné regresie. Proces rozdeľovania dátovej sady do podmnožín prebieha rekurzívne, až pokiaľ nie je splnená vopred špecifikovaná konečná podmienka. Každý uzol reprezentuje rozdelenie vstupnej množiny na podmnožiny [66].

Rozdelenie do podmnožín prebieha tak, že pre každú podmnožinu množiny podľa atribútu X je spočítaná miera nečistoty. Následne je zvolená hodnota atribútu tak, že podmnožiny rozdelené podľa tejto hodnoty majú najmenšiu mieru nečistoty. Tá je rovná súčtu štvorcových odchýlok hodnôt vzorkov od priemeru vzoriek v uzle [43].

Po rozdelení dátovej sady do požadovaných podmnožín je pre každý listový uzol vykonaná regresná analýza a vytvorený regresný model, ktorý je platný iba v tom uzle [66].

Random forest

Random forest je algoritmus strojového učenia, ktorý vytvára niekoľko rozhodovacích stromov [17]. Každý rozhodovací strom je generovaný na základe náhodne vybranej sady vstupných atribútov. Konečný výsledok získa kombináciou výsledkov z generovaných stromov [25].

XGBoost

XGBoost je navrhnutý ako algoritmus založený na posilňovaní, tzv. boostingu. [79]. Aplikuje prístup rekurzívneho binárneho rozdelenia (viď 4.3.1) na výber najlepšieho rozdelenia v každom kroku, aby sa dosiahol najlepší model [86]. XGBoost algoritmus v každom kroku vytvorí plytký strom a potom vypočíta zvyšky na predpoveď [79]. Model stromového súboru používa K aditívnych funkcií na predpovedanie výstupu [15]:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{4.6}$$

kde $\mathcal{F}$ je priestor regresných stromov, x_i sú vstupné hodnoty, $\phi(x_i)$ je funkciou modelu a $\hat{y}_i$ je výstupná hodnota.

Pre tréning množiny funkcií použitých v modeli, minimalizujeme nasledujúci regulovaný cieľ [15]:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{4.7}$$

kde $l(\hat{y}_i, y_i)$ je stratová funkcia počítajúca rozdiel medzi predikovanou a aktuálnou hodnotou a Ω je regularizačný parameter ako prevencia proti pretrénovaniu. Regularizačný parameter sa počíta ako [15]:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w_k\|^2 \tag{4.8}$$

kde k je číslo stromu, T je počet listov, $\|w_k\|^2$ je regularizačná norma L2. Parametre γ a λ kontrolujú mieru konzervatívizmu pri prehľadávaní stromu [86].

LigthGBM

LightGBM [38] je podobne ako XGBoost typ rozhodovacieho stromu s posilňovaním gradientu (GBDT). Bol vytvorený za účelom zlepšenia škálovateľnosti a optimalizácie oproti iným stromom zo skupiny GBDT. Na vyriešenie tohto problému boli predstavené 2 nové techniky: Jednostranné vzorkovanie založené na gradiente (*Gradient-based One-Side Sampling*, skrátene GOSS) a exkluzívne zväzky vstupných premenných (*Exclusive Feature Bundling* alebo EFB).

GOSS je technika, urýchľuje trénovanie zameraním sa na dátové inštancie s veľkými gradientmi, ktoré sú informatívnejšie, a náhodným výberom z inštancií s malými gradientmi. Tento prístup zachováva presnosť a zároveň skracuje čas výpočtu.

EFB je technika, ktorá redukuje priestor vzájomne exkluzívnych atribútov ich zväzovaním. Vzájomne exkluzívne atribúty sú také atribúty, ktoré zriedkakedy nadobúdajú súčasne nenulové hodnoty (napríklad atribúty zakódované pomocou kódovania „one-hot"). Tento prístup vedie k efektívnejšiemu tréningu a zníženiu spotreby pamäte.

4.3.2 Neurónové siete

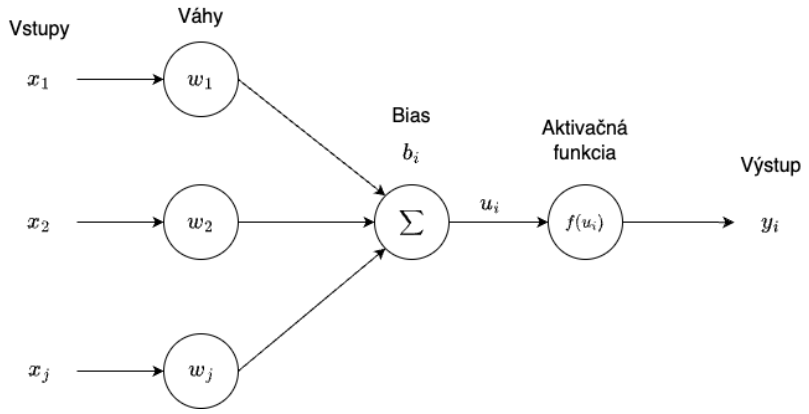
Neurónové siete sú matematické modely inšpirované architektúrou sietí nervových buniek [53]. Architektúra neurónových sietí môže byť definovaná týmito charakteristikami [84]:

1. Definovaním počtu a vrstiev a počtu neurónov v nich
2. Trénovacím mechanizmom aplikovaným na úpravu váh na jednotlivých prepojeniach
3. Aktivačnými funkciami špecifikované pre jednotlivé vrstvy

Neurónová sieť je zložená z umelých neurónov organizovaných do vrstiev [22]. Obrázok 4.2 zobrazuje uloženie neurónov do vrstiev a tiež tri hlavné typy vrstiev. Na príjem vstupných dát slúži vstupná vrstva. Vstupy následne spracujú neuróny v skrytých vrstvách, ktorých výstupy transformuje výstupná vrstva na konečný výstup.

Neurón

Umelý neurón je základnou stavebnou jednotkou neurónových sietí. Tento prvok má tri súbory pravidiel. Sú to sčítanie, násobenie a aktivácia [39]. Každý neurón prijíma niekoľko vstupov a generuje jeden výstup. Každý vstup má svoju určitú váhu [53].



Obr. 4.1: Neurón v neurónovej sieti. Obrázok zobrazuje transformáciu vstupov na výstup pomocou jedného neurónu.

Vstupy vynásobené váhami a sčítané dohromady aj s biasom tvoria dohromady stav neurónu i [10]:

$$u_i = \sum_{j=1}^{N} w(i,j)x(j) + b(i) \tag{4.9}$$

i je číslo neurónu, $x(j)$ je hodnota vstupu a $w(j,i)$ je váha vstupu pre daný neurón i (symbolizuje váhu prepojenia neurónov). Bias term $b(i)$ je skalárna trénovateľná hodnota používaná aj ako prahová hodnota pre aktivačnú funkciu.

Výstupná hodnota neurónu y je rovná stavu neurónu vloženom ako parameter do aktivačnej funkcie f [53]. Proces transformácie vstupných hodnôt na výstupné, ktorú realizuje jeden neurón, je zobrazený na obrázku 4.1.

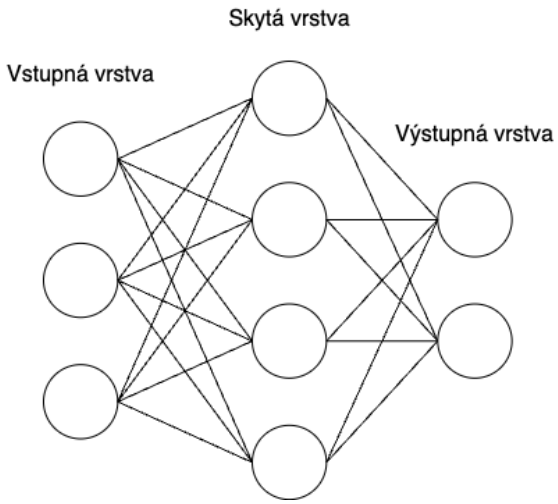
$$y_i = f(u_i) \tag{4.10}$$

Existujú tri typy umelých neurónov - deterministický, pravdepodobnostný a neurón, ktorý produkuje reálnu hodnotu [76]. Deterministický neurón môže nadobudnúť hodnoty 0 alebo 1 (aktivovaný alebo neaktivovaný), pravdepodobnostný v intervale od 0 do 1 a reálny akúkoľvek hodnotu na množine reálnych čísel. Typ neurónu sa odvíja hlavne od toho, aké výstupy produkuje jeho aktivačná funkcia.

Aktivačné funkcie

Neurónová sieť sa bez aktivačnej funkcie správa ako lineárny regresný model [71]. Neexistuje jednoduché vysvetlenie, prečo použitie jednej aktivačnej funkcie v modeli má lepšie výsledky ako použitie druhej. Aktivačné funkcie sú dôležité pri procese učenia [65].

Existuje mnoho druhov aktivačných funkcií, a na rôzne účely sa stále vyvíjajú ďalšie. Táto sekcia popíše základné typy týchto funkcií. Medzi populárne aktivačné funkcie patria skoková funkcia, ReLU, hyperbolický tangens a sigmoidná funkcia [13]. Ich predpisy sú uvedené v tabuľke 4.1. Tieto funkcie sú, s výnimkou skokovej funkcie, nelineárne, čo vedie k možnosti odhaliť aj nelineárne závislosti medzi vstupnými premennými [13].



Obr. 4.2: Základná architektúra neurónovej siete. Zobrazuje vstupnú, skrytú a výstupnú vsrtvu.

Skoková funkcia

Skoková aktivačná je lineárna funkcia, ktorá vracia hodnotu 1, ak je stav neurónu väčší alebo rovný 0, v opačnom prípade vracia hodnotu 0 [76]. Funkcia nie je vhodná na trénovanie modelov trénovaných pomocou spätnej propagácie, pretože táto metóda využíva deriváciu aktivačnej funkcie na úpravu váh a biasov. Prvá derivácia tejto funkcie je rovná 0, čo vedie k neefektívnej úprave parametrov učenia, prípadne k stavu, v ktorom sa model nič nenaučí [17]. Použitím lineárnej funkcie je možné odhaliť len lineárne závislosti, teda nie je vhodná pre komplexné dáta [13].

Sigmoidná funkcia

Problém skokovej aktivačnej funkcie rieši sigmoidná aktivačná funkcia, ktorá je plynulým priblížením skokovej funkcie. Jej hlavným účelom je premena hodnôt z oboru reálnych čísiel do uzavretého intervalu od 0 do 1 [65, 17]. Vďaka tomu je výstup každého neurónu tiež stlačený a to spôsobuje zánik gradientu, hlavne v hlbokých sieťach. To spôsobuje, že od určitého bodu je daná sieť veľmi ťažko optimalizovateľná [65].

Hyperbolický tangens

Podobná funkcia sigmoidnej funkcii je hyperbolický tangens. Hlavný rozdiel medzi týmito funkciami je, že prevádza hodnoty do intervalu od -1 do 1 miesto od 0 do 1 [17]. Hlavnou výhodou oproti sigmoidnej funkcii je tá, že jej prvá derivácia je strmšia oproti prvej derivácii sigmoidnej funkcie, čo znamená, že môže nadobudnúť väčší rozsah hodnôt. Na druhej strane, podobne ako sigmoidná funkcia, má ohraničený výstup, takže podobne ako sigmoidná funkcia má problém so zanikajúcim gradientom [65].

ReLU

Najviac používanou aktivačnou funkciou je ReLU (rektifikovaná lineárna jednotka) [65, 17]. Výhodou tejto funkcie je výpočtová efektivita [17]. ReLU nie je vhodná na použitie ako

aktivačná funkcia do rekurentných neurónových sietí, pretože môže ľahko podporiť problém explodujúceho gradientu.

Ďalším problémom ReLU je možnosť produkcie mŕtvych neurónov. Mŕtve neuróny sú neuróny, ktoré produkujú 0 pre každý výstup, čo sa prejaví ako strata učenia [17]. Tento prípad je špeciálny druh miznúceho gradientu [65]. Je veľmi nepravdepodobné, že sa mŕtvy neurón z tohto stavu zotaví [77, 2].

Aktivačná funkcia	**Predpis**
Skoková funkcia	$f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$
Sigmoidná funkcia	$f(x) = \frac{1}{1+e^{-x}}$
Hyperbolický tangens	$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
ReLU	$f(x) = \max(0, x)$

Tabuľka 4.1: Prehľad základných aktivačných funkcii.

Proces učenia sa neurónových sietí

Učenie sa v kontexte neurónovej siete znamená nájdenie takých kombinácií váh, vďaka ktorým sa daná neurónová sieť začne správať požadovaným spôsobom [69]. Váhy upravujeme na základe toho, že počas procesu tréningu sú neurónovej sieti poskytnuté vzorky dát. Na základe týchto vzoriek sú iteratívne upravované váhy tak, aby bola dosiahnutá čo najmenšia úroveň chyby vo výsledku [13].

Najviac využívaná metóda na tréning neurónových sietí je metóda spätnej propagácie. Princíp je založený na počítaní gradientného zostupu stratovej funkcie vzhľadom na váhy v sieti, a preto tento algoritmus vyžaduje na tréning označené dáta [17].

Pomocou metriky je vypočítaná chyba, ktorú daný model svojou predikciou vykonal. Táto chyba je potom spätne propagovaná do siete, čo zahŕňa výpočet derivácie chyby vzhľadom na každú váhu v sieti [17]. Pri počítaní chýb sa postupuje od výstupnej vrstvy k vstupnej vrstve. Ak sú k dispozícii derivácie jednej vrstvy skrytých neurónov, tak na základe týchto derivácií sú vypočítané derivácie chyby pre vrstvu pod ňou, až pokým nespočítame diferenciu chyby pre všetky vrstvy [13]. Derivácie týchto váh poskytujú informáciu o tom, aký vplyv má určitá zmena váhy na veľkosť chyby. Váhy sú následne upravené v smere, ktorý minimalizuje chybu [17].

4.4 Hlboké učenie

Neurónová sieť, ktorá je tvorená viac ako tromi skrytými vrstvami, sa považuje za hlbokú neurónovú sieť [13].

4.4.1 Rekurentné neurónové siete

Rekurentné neurónové siete (RNN) sú typ neurónových sietí na rozlišovanie vzorov v sekvenciách dát. Na použitie tohto prístupu musia byť dáta v sekvencii na sebe závislé. Na spracovávanie a hľadanie vzorov v sekvenciách využívajú vnútorný stav (pamäť) [17].

Architektúra rekurentných neurónových sietí

Jednoduchá RNN má 3 vrstvy - vstupnú a skryté rekurentné a výstupné vrstvy [67]. Oproti klasickým neurónovým sieťam, majú navyše rekurentné spojenia, ktoré propagujú informáciu medzi neurónmi rovnakej vrstvy. Plne prepojená RNN obsahuje informačný tok z každého do každého ďalšieho neurónu v jeho vrstve vrátane seba [13].

Problémy pri trénovaní rekurentnej neurónovej siete

RNN sa nemôžu učiť dlhodobé časové závislosti ak je na tréning použitý gradientný zostup [8]. Pri spätnej propagácii gradientu v čase gradient exponenciálne klesá, čo zapríčiňuje problém miznúceho gradientu [67].

Okrem klesajúceho gradientu sa môže objaviť aj rastúci gradient. Tento gradient môže naopak pri spätnej propagácii explodovať, čo sa nazýva problém explodujúceho gradientu [67]. Príliš veľký gradient zapríčiňuje nestabilné učenie RNN a môže viesť k veľkej fluktuácii váh v sieti. To môže zhoršiť schopnosť učenia siete a generalizácie nových dát [17].

4.4.2 LSTM

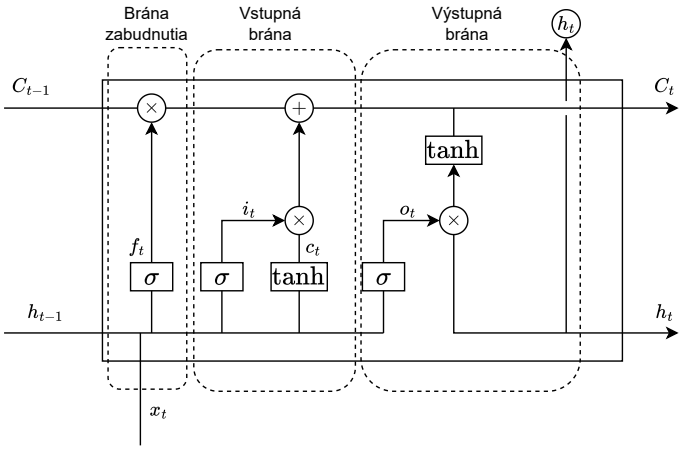
LSTM (dlhá a krátkodobá pamäťová sieť) je druh rekurentnej neurónovej siete, ktorá vďaka svojej štruktúre dokáže popísať sekvenčné dáta s časopriestorovou koreláciou [49]. Architektúra LSTM sietí bola navrhnutá tak, aby uľahčila zapamätanie si informácií počas dlhých časových období, kým nie sú potrebné [24]. Oproti RNN sú tieto siete odolné voči problému miznúceho gradientu [29].

Základnou jednotkou je LSTM bunka , ktorá má 4 brány - vstupnú bránu i_t, bránu bunky c_t, bránu zabudnutia f_t a výstupnú bránu o_t [55]. Brány v bunke a štruktúra samotnej bunky sú zobrazené na obrázku 4.3. Funkcie LSTM bunky sú popísané v nasledujúcich rovniciach [56]:

$$f_t = \sigma(W_{uf}u_t + W_{hf}h_{t-1} + b_f) \tag{4.11}$$
$$i_t = \sigma(W_{ui}u_t + W_{hi}h_{t-1} + b_i) \tag{4.12}$$
$$o_t = \sigma(W_{uo}u_t + W_{ho}h_{t-1} + b_o) \tag{4.13}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{uc}u_t + W_{hc}h_{t-1} + b_c) \tag{4.14}$$
$$h_t = o_t \odot \tanh(c_t), \tag{4.15}$$

kde:

- u_t je vstupný vektor
- i_t, o_t, f_t sú vektory brán
- c_t je vektor stavu bunkovej pamäte
- h_t je stavový výstupný vektor



Obr. 4.3: LSTM bunka s vyznačenými bránami.

- $\sigma(x)$ je sigmoidná funkcia (po prvkoch)
- $x \odot y$ je prvkový súčin
- b_i, b_f, b_o, b_c sú vektory skreslenia (biasov)
- $W_{ui}, W_{hi}, W_{uf}, W_{hf}, W_{uo}, W_{ho}, W_{uc}, W_{hc}$ sú matice lineárnej transformácie

Vstupná brána a brána zabudnutia v pamäťovej bunke určujú jej správanie na základe ich hodnôt. Tie môžu nadobudnúť v uzavretom intervale hodnoty od 0 do 1. Správanie bunky podľa krajných hodnôt týchto brán je zobrazené v tabuľke 4.2 [24].

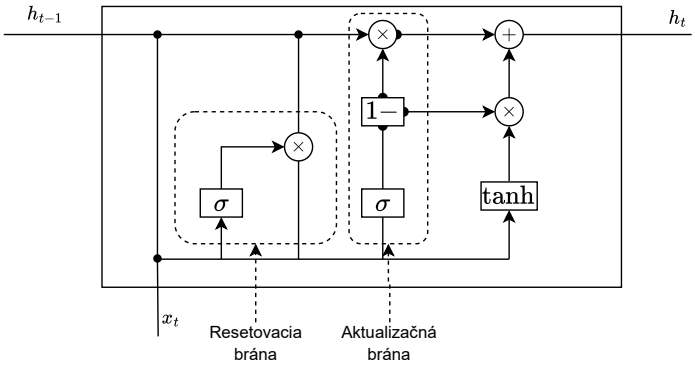
Vstupná brána	**Brána zabudnutia**	**Správanie bunky**
0	1	zachovanie predchádzajúcej hodnoty
1	1	pridanie hodnoty k uchovanej hodnote
0	0	zmazanie uchovanej hodnoty
1	0	prepísanie uchovanej hodnoty

Tabuľka 4.2: Popis správania LSTM bunky na základe hodnôt vstupnej brány a brány zabudnutia.

Množstvo informácií, ktoré má bunková pamäť aktualizovať, zabudnúť a poslať na výstup, určujú brány i_t, o_t, f_t. V závislosti od stavu vektora brány zabudnutia f_t môže byť stav bunky resetovaný alebo obnovený. Podobne fungujú aj vektory i_t a o_t, ktoré sú určené na reguláciu vstupu a výstupu [56].

4.4.3 GRU

GRU (gate recurrent unit) [3] je hradlová jednotka, ktorá moduluje tok informácií v rámci jednotky bez samostatnej pamäťovej bunky [20]. Jednotka GRU má jednoduchšiu základnú štruktúru oproti LSTM. Oproti LSTM má 2 brány: resetovaciu bránu a aktualizačnú bránu. Resetovacia brána určuje mieru straty novej informácie (výpočet zobrazený v rovnici 4.16), zatiaľ čo aktualizačná brána riadi, koľko novej informácie sa zapamätá a koľko sa ponechá



Obr. 4.4: GRU bunka s vyznačenými bránami.

z predchádzajúcej informácie [85]. Výpočet hodnoty aktualizačnej brány je zobrazený v rovnici 4.17 a výpočet vnútorného stavu je zobrazený v rovnici 4.18. GRU jednotka je zobrazená na obrázku 4.4.

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \tag{4.16}$$

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \tag{4.17}$$

$$h_t = \tanh(W_t + r_t \oplus h_{t-1}) \tag{4.18}$$

kde:

- x_t je vstupný vektor
- W, U sú váhové matice
- r_t je hodnota restovacej brány
- z_t je hodnota aktualizačnej brány
- h_t je skrytý stav bunky

Sekvenčné modelovanie

LSTM siete sa využívajú na predikciu sekvencii zo sekvencii. Tento proces sa nazýva sekvenčné modelovanie. Pri sekvenčnom modelovaní dostane model na vstupe viacrozmerný časový rad $\mathbf{Y} \in \mathbb{Y}^{m \times T}$ konštantnej dĺžky T z množiny obsahujúcej m časových radov rovnakej dĺžky. $\mathbf{Y}_t(i)$ je i-tá časová rada v čase t. Cieľom modelu je potom predikovať časovú radu $\mathbf{Y}_{T+1}(\cdot)$ [46]. V praxi sa na predikciu budú používať hodnoty časových radov na základe sekvencie posledných n časových radov. Ak je dĺžka celej časovej rady m, tak predikovaná hodnota bude predpovedaná zo sekvencie $Y_n^m = (Y_n(\cdot), Y_{n+1}(\cdot), \ldots, Y_m(\cdot))$.

Kapitola 5

Zber dát

Na trénovanie modelov je potrebné mať čo najväčšie množstvo kvalitných dát. Platí koncept GIGO (Garbage in, garbage out), ktorý popisuje, že z nekvalitných vstupných dát môže vyplynúť len nekvalitný výstup [13]. Záznamy o vývoji populácie sú vo forme časových radov. Časová rada je typ dát, ktoré sa zhromažďujú v pravidelných časových intervaloch [73]. Táto kapitola je zameraná na popis, analýzu a metódy predspracovania získaných dát.

5.1 Zdroj dát

Dáta boli preberané z oficiálneho a voľne prístupného zdroja `data.worldbank.org`[1]. Jedná sa o spoľahlivý zdroj obsahujúci dáta z nielen demografických, ale aj ekonomických a iných oblastí pre rôzne krajiny. Konkrétne dátové sady použité pre túto prácu sú uvedené aj so stručným popisom v tabuľke 5.2. Originálne dátové sady sú poskytnuté v súborovom formáte `csv`. Majú uniformnú štruktúru - každá dátová sada obsahuje rovnaké stĺpce. Formát dátových sád je popísaný na obrázku 5.1.

Uniformná štruktúra dátových sád						
Meno krajiny	Kód krajiny	Meno indikátoru	Kód indikátoru	1960	...	2023
<meno1>	<kód1>	<indikator>	<kód_i>	<i_1_1960>	...	<i_1_2023>
<meno2>	<kód2>	<indikator>	<kód_i>	<i_2_1960>	...	<i_2_2023>

Obr. 5.1: Štruktúra dát jednotlivej dátovej sady. Zvýraznená časť zodpovedá názvom stĺpcov a premenné sú zobrazené vo formáte `<meno_premennej>`. Premenná znázorňujúca hodnotu indikátora danej krajiny v danom roku je zapísaná v tvare `<i_{číslo_krajiny}_{rok}>`. Stĺpec pre kód krajiny a kód indikátoru je v dátovej sade vynechaný.

5.1.1 Populačné dátové sady

Vytvorenú dátovú sadu možno rozdeliť do menších celkov podľa vybranej populácie. Tieto dátové sady sú užitočné najmä na extrapoláciu chýbajúcich dát (viď sekciu 5.2.2). Môžu

[1] https://data.worldbank.org/

Dátová sada	**Popis indikátora**
Fertility rate, total (births per woman)	Počet pôrodov na ženu (viď sekciu 2.4)
Population, total	Celkový počet ľudí v populácii
Net migration	Rozdiel medzi počtom imigrantov a emigrantov (viď sekciu 2.4)
Arable land (% of land area)	Percentuálna časť ornej pôdy vzhľadom na celkovú plochu krajiny
Birth rate, crude (per 1,000 people)	Počet pôrodov na 1 000 obyvateľov (viď sekciu 2.4)
GDP growth (annual %)	Ročný rast hrubého domáceho produktu (HDP) v percentách
Death rate, crude (per 1,000 people)	Počet úmrtí na 1 000 obyvateľov (viď sekciu 2.4)
Population ages 15-64 (% of total population)	Percento obyvateľstva vo veku 15-64 rokov v porovnaní s celkovým počtom obyvateľov
Population ages 0-14 (% of total population)	Percento obyvateľstva vo veku 0-14 rokov v porovnaní s celkovým počtom obyvateľov
Agricultural land (% of land area)	Percentuálna časť poľnohospodárskej pôdy vzhľadom na celkovú plochu krajiny
Population ages 65 and above (% of total population)	Percento obyvateľstva vo veku 65 rokov a viac v porovnaní s celkovým počtom obyvateľov
Rural population (% of total population)	Percento vidieckeho obyvateľstva v porovnaní s celkovým počtom obyvateľov
Rural population growth (annual %)	Ročný rast vidieckeho obyvateľstva v percentách
Age dependency ratio (% of working-age population)	Pomer závislosti obyvateľov mimo produktívneho veku (deti a starší ľudia) k obyvateľom v produktívnom veku (viď sekciu 2.4)
Urban population (% of total population)	Percento mestského obyvateľstva v porovnaní s celkovým počtom obyvateľov
Population growth (annual %)	Ročný rast populácie v percentách
Adolescent fertility rate (births per 1,000 women ages 15-19)	Počet pôrodov na 1 000 žien vo veku 15-19 rokov
Life expectancy at birth, total (years)	Očakávaná dĺžka života pri narodení v rokoch (viď sekciu 2.4)
Population, female	Podiel žien v populácii vyjadrený v percentách
Population, male	Podiel mužov v populácii vyjadrený v percentách

Tabuľka 5.2: Konkrétne použité dátové sady pre tvorbu dátovej sady. Dáta obsiahnuté v jednotlivej dátovej sade obsahujú záznamy daného indikátoru pre viacero populácii od roku 1960 - 2023. Dáta obsahujú chýbajúce hodnoty.

byť použité aj na predikciu vývoja danej konkrétnej populácie alebo na dolaďovanie modelu na získanie presnejších predikcií na zvolenú populáciu.

5.1.2 Tréningová a validačná dátová sada

Na vyhodnotenie presnosti metódy je potrebné rozdeliť dátovú sadu na 2 časti. Prvú časť tvoria dáta určené na tréning modelu a druhá časť je použitá na ohodnotenie jeho výkonu. Modelu budú poskytnuté len dáta potrebné na predikciu a následne budú výsledky modelu porovnané s referenčnými hodnotami.

Keďže sa jedná o časovú radu, tak tréningové a validačné dáta budú rozdelené podľa časových období. Každý vývoj populácie je rozdelený na 2 sekvencie zvlášť. Trénovacia dátová sada obsahuje údaje od najstaršieho vývoja populácie až po deliaci rok. Validačná dátová sada obsahuje záznamy od deliaceho roku po najnovšie záznamy každej populácie.

5.2 Tvorba dátovej sady

Dátová sada je tvorená spojením niekoľkých dátových sad. Každá čiastková dátová sada obsahuje údaje ako krajina, meno ukazateľa a roky v maximálnom rozmedzí 1960 - 2023. Každý rok predstavuje jeden stĺpec z hodnotou indikátora v danom roku. Proces tvorby dátovej sady je zobrazený na obrázku 5.3. Zobrazuje postupnosť operácií aplikovaných na získané dáta.

5.2.1 Reštrukturalizácia dát

Na vytvorenie záznamov závislých od času je nutné transformovať roky z viacerých stĺpcov do jedného. To je vykonané prevodom z tzv. širokého formátu do dlhého formátu. Tieto dva formáty sa odlišujú tým, že dlhý formát obsahuje viac riadkov ako široký formát, ale za to menej stĺpcov ako široký formát [34].

Naopak bolo nutné rozšíriť hodnotu indikátorov z dlhého do širokého formátu. Každá čiastková dátová sada obsahovala práve jeden indikátor, takže transformácia prebehla tak, že hodnoty v stĺpci pre daný indikátor boli agregované do samostatného stĺpca. Tento stĺpec potom obsahoval hodnoty v daných rokoch. Prevod zo širokého do dlhého formátu je zobrazený na obrázku 5.2.

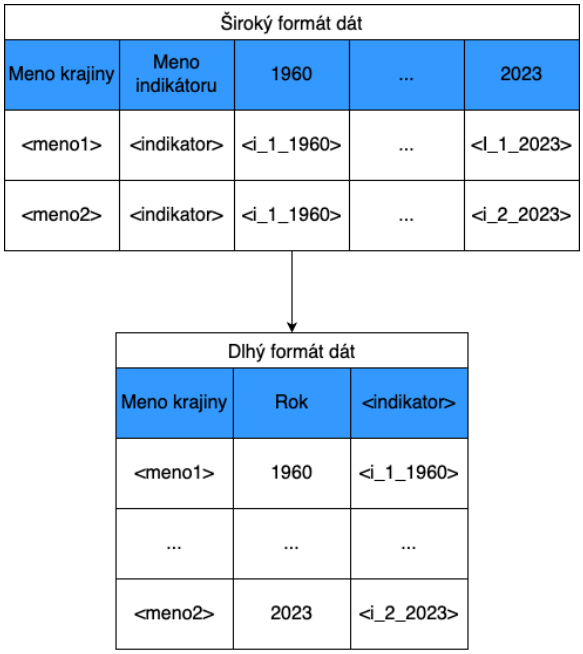
5.2.2 Chýbajúce hodnoty

Po zjednotení všetkých potrebných čiastkových údajov vznikla dátová sada, ktorá obsahovala chýbajúce hodnoty. Na získanie čo najväčšieho počtu údajov bola dátová sada rozdelená podľa jednotlivých populácií. V stĺpcoch populácie, kde chýbalo menej ako 30 percent dát, boli dogenerované pomocou spline interpolácie tretieho rádu (kubická interpolácia). Interpolácia je metóda predikcie neznámych dát na základe známych dát [52].

5.2.3 Čistenie dát

Po doplnení niektorých chýbajúcich hodnôt stále ku každej populácii neboli prístupné všetky potrebné indikátory pre všetky roky. Väčšina chýbajúcich dát bola odstránená pri spájaní čiastkových dátových sád do jednej. Záznamy, kde chýbali nejaké potrebné údaje, boli zahodené.

Dátová sada obsahovala pravidelné aj nepravidelné sekvencie údajov pre rôzne populácie. Aby bola zachovaná pravidelnosť sekvencií, tak populácie, ktorých počet chýbajúcich rokov bol menší alebo rovný 20 percentám, boli dopočítané opäť pomocou interpolácie.



Obr. 5.2: Prevod zo širokého do dlhého formátu jednej základnej dátovej sady pre jeden indikátor. Zvýraznená časť zodpovedá názvom stĺpcov a premenné sú zobrazené vo formáte `<meno_premennej>`. Premenná hodnoty indikátora danej krajiny v danom roku je zapísaná v tvare `<i_{číslo_krajiny}_{rok}>`.

Týmto spôsobom bol počet záznamov zredukovaný z pôvodných 17 024 na 12 400 záznamov.

5.3 Predspracovanie dát

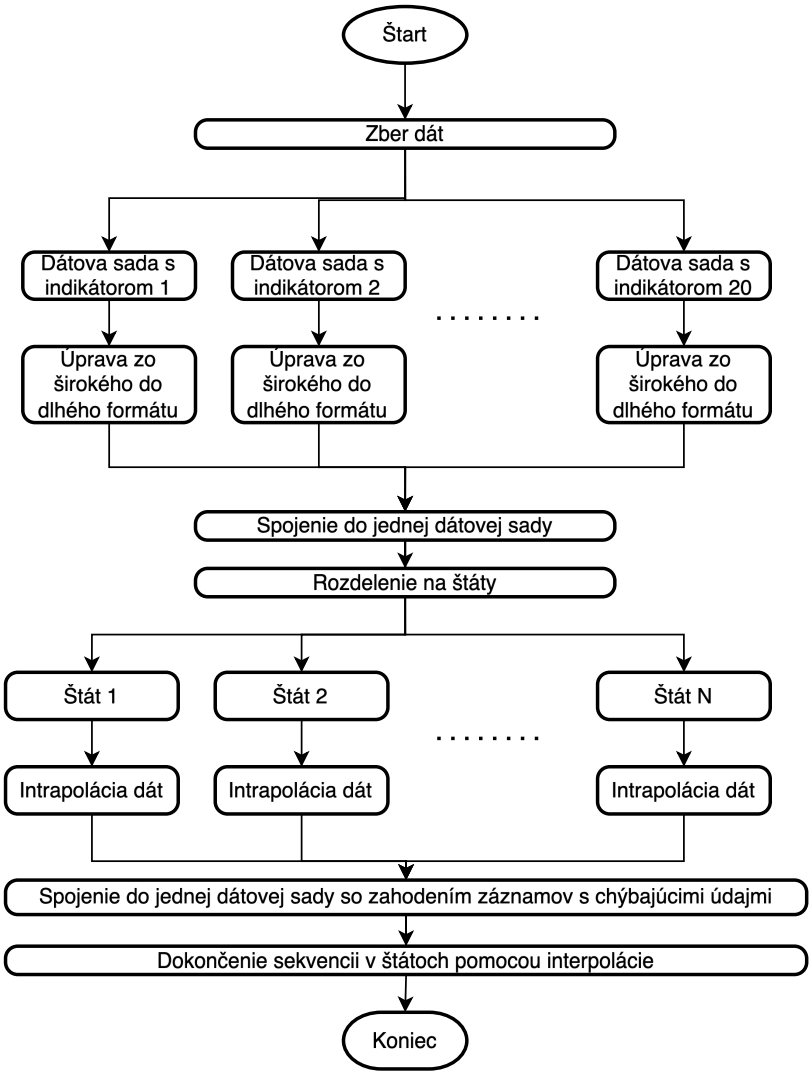
Pred použitím dát budú vybrané dáta podliehať predspracovaniu. V tomto prípade to zahŕňa rozdelenie dátovej sady na tréningovú a validačnú, transformáciu formátu dát a ich škálovanie. Ako boli dáta predspracované, je stručne zobrazené na obrázku 5.4.

5.3.1 Transformácia dát

Vstupné dáta boli rozdelené do niekoľkých kategórií: kategorické premenné, absolútne hodnoty numerických premenných (premenné, s nízkym rozptylom hodnôt), percentuálne premenné a premenné s vysokým rozsahom hodnôt. Kategorické dáta sú prevádzané na numerické použitím kódovania štítkov (anglicky *label encoding*). Absolútne numerické premenné ostali bezo zmeny a percentuálne ukazatele boli normalizované, teda prevedené z intervalu $(0, 100)$ do intervalu $(0, 1)$.

Medzi premenné s vysokým rozsahom hodnôt patria ukazatele čistá migrácia, celkový počet ľudí v populácii a hodnota gdp. Pre údaje s veľkými aj malými zložkami sa používa logaritmická transformácia. Konkrétne bola použitá bi-symetrická logaritmická transformácia [80]:

$$y = sgn(x) \cdot log_{10}(1 + |(x/C|) \tag{5.1}$$



Obr. 5.3: Proces tvorby dátovej sady.

kde x je pôvodná hodnota, y je transformovaná hodnota a C je konštanta, ktorá upravuje sklon prenosovej funkcie blízko začiatku. Potom inverzná transformácia zodpovedá rovnici:

$$x = sgn(x) \cdot C \cdot (-1 + 10^{|y|}) \tag{5.2}$$

5.3.2 Škálovanie dát

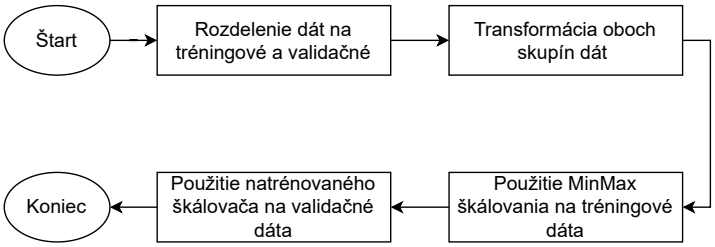
Škálovanie premenných potrebných na predikciu je nevyhnutným krokom pri vytváraní efektívnych prediktívnych modelov [45]. Najviac používanými metódami na škálovanie dát sú normalizácia a štandardizácia. Jedná sa o procesy, ktoré upravujú rozpätie dát.

Normalizácia prevádza dáta na hodnoty, ktoré patria do intervalu 0 až 1 alebo -1 až 1. Je vhodné ju použiť, ak existujú veľké rozdiely v rozsahoch premenných použitých na predikciu [5, 45]. Na druhej strane štandardizácia upraví príslušné dáta tak, aby ich priemer bol rovný 0 a smerodajná odchýlka rovná 1 [17, 45]. Štandardizácia je vhodná na použitie najmä ak hodnoty dát podliehajú gaussovému rozloženiu.

Normalizácia dát sa používa na zníženie variability a urýchlenie konvergencie modelu počas trénovania. Štúdia z roku 2025, ktorá sa zaoberá predikciou cien akcií, ukazuje, že škálovanie dát pomocou Min-Max škálovača produkuje presnejšie výsledky ako použitie štandardného škálovača [74]. Z tohto dôvodu bola v tomto prípade na úpravu dát použitá normalizácia na miesto štatandardizácie. Min-Max škálovač prevádza dáta do intervalu $(0, 1)$ nasledovne:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5.3}$$

kde X' je výsledok škálovanej hodnoty, X je škálovaná hodnota a X_{min} a X_{max} zodpovedajú minimálnej a maximálnej poznanej hodnote premennej. Škálovač je natrénovaný na trénovacích dátach a potom použitý aj na validačné dáta.



Obr. 5.4: Postup použitého predspracovania dát.

5.4 Spracovanie a analýza získaných dát

Získané dáta obsahujú nezávislé premenné (vlastnosti, príznaky) a cieľové (závislé) premenné, ktoré bude pomocou nich predpovedať.

5.4.1 Predpovedané parametre poppulácie

Predpovedané parametre populácie boli rozdelené do 3 skupín:

1. Starnutie populácie - pomer ľudí v určitých vekových skupinách.
2. Pomer počtu mužov a žien v populácii.
3. Počet ľudí v populácii.

Konkrétne premenné aj s rozdelením do skupín sú zobrazené a popísané v tabuľke 5.3.

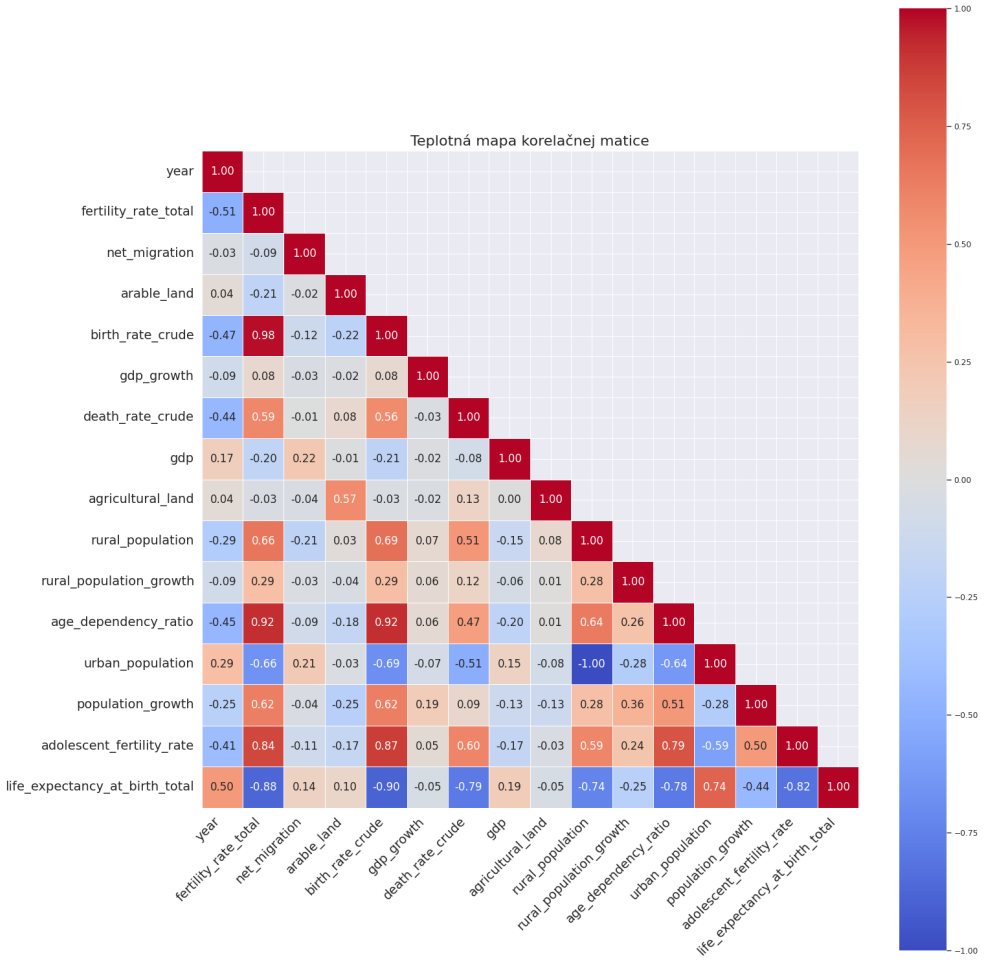
5.4.2 Výber vlastností pre predikciu

Nie všetky zozbierané indikátory sú vhodné na predikciu cieľových premenných. Dobrá podmnožina nezávislých premenných je taká, ktorá vysoko koreluje s cieľovou premennou, ale zároveň medzi sebou nekorelujú [26]. Na vyfiltrovanie redundantných vstupných premenných bola použitá korelačná matica (viď obrázok 5.5):

Na obrázku 5.5 možno vidieť, že niektoré vstupné premenné majú medzi sebou vysokú, buď kladnú, alebo zápornú koreláciu. Všetky dvojce (nezahrňujúce cieľové premenné - viď

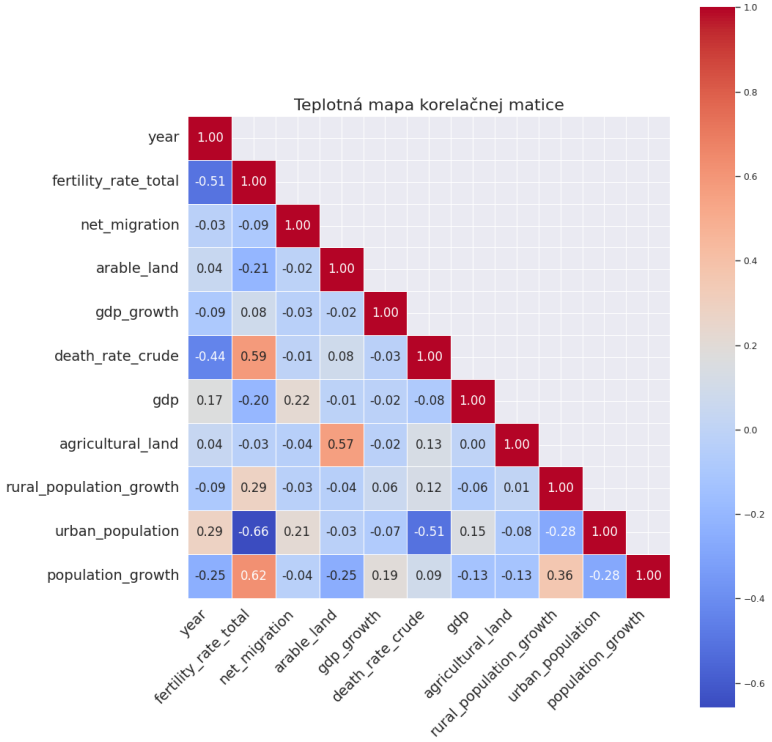
Skupina	Názov premennej	Popis
Starnutie populácie	population_ages_0-14	Podiel ľudí vo veku 0-14 rokov
	population_ages_15-64	Poduel ľudí vo veku 15-64 rokov
	population_ages_65_and_above	Podiel ľudí vo veku 65+ rokov
Rozloženi pohlavia	population_female	Podiel žien v populácii
	population_male	Podiel mužov v populácii
Veľkosť populácie	population_total	Absolútny počet ľudí v populácii

Tabuľka 5.3: Cieľové premenné.



Obr. 5.5: Korelačná matica so všetkými vstupnými indikátromi v dátovej sade.

tabuľka 5.3), ktorých absolútna hodnota korelačného koeficientu bola vyššia ako 0.8, boli vyfiltrované. Obrázok 5.6 znázorňuje korelačnú maticu s vybranými vstupnými premennými a mieru ich korelácie medzi sebou. Obrázok 5.7 zase znázorňuje mieru korelácie medzi jednotlivými vybranými vstupnými premennými a na obrázku 5.7 je znázornená miera korelácie medzi vybranými vstupnými a výstupnými premennými.



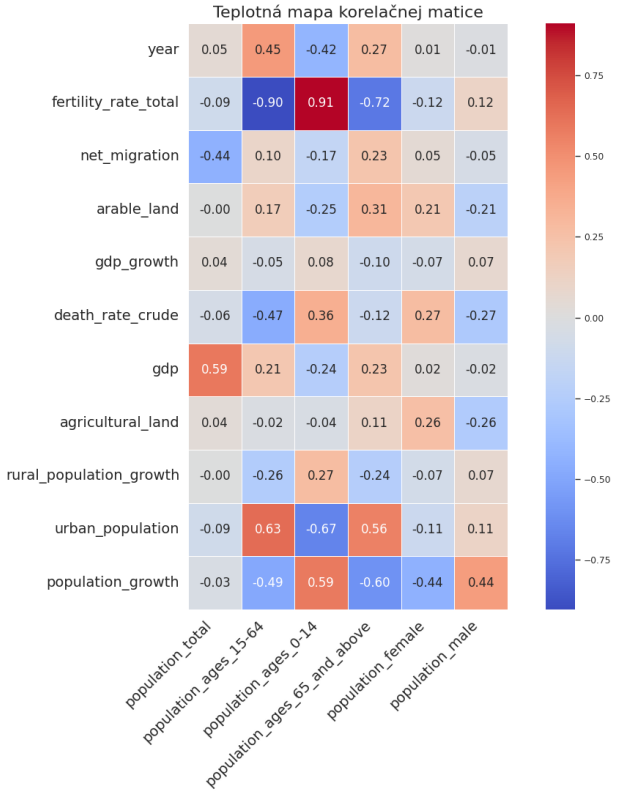
Obr. 5.6: Korelačná matica vybraných vstupných premenných.

5.4.3 Ukážka vytvorenej dátovej sady

Dátová sada bude uložená dvomi spôsobmi: celá dátová sada obsahujúca všetky zaznamenané populácie a dáta jednotlivých populácií samostatne. Je to kvôli uľahčeniu práce s časovými radmi jednotlivých populácií. Na celkové uloženie dát je využitý formát `csv`.

Vytvorená dáta obsahuje viac ako 12-tisíc záznamov 243 populácií. Neobsahuje žiadne chýbajúce hodnoty. Časový úsek zaznamenaných vývojov populácií sa medzi nimi môže líšiť. Priemerný rok záznamov sa pohybuje okolo roku 1993. Minimálny zaznamenaný rok je 1961 a maximálny je 2021.

Výsledná dátová sada obsahuje indikátory z každej čiastkovej dátovej sady popísanej v tabuľke 5.2. Každý indikátor slúži ako stĺpec. Okrem nich sa v dátovej sade nachádza rok a meno krajiny. Ukážka dátovej sady je zobrazená v tabuľkách 5.4 a 5.5. Dátová sada je zoradená vzostupne podľa mena krajiny a následne podľa roku záznamu.



Obr. 5.7: Korelačná matica vybraných vstupných s cieľovými premennými.

Meno krajiny	**Rok**	**Miera plodnosti, celková**	**Počet obyvateľov, spolu**	**Čistá migrácia**	**...**
Afghanistan	2001	7,446	19688632,0	-192286,0	...
Afghanistan	2002	7,339	21000256,0	1327074,0	...
Afghanistan	2003	7,220	22645130,0	388632,0	...
Afghanistan	2004	7,069	23553551,0	-248616,0	...
Afghanistan	2005	6,905	24411191,0	252185,0	...

Tabuľka 5.4: Dátová sada - prvá časť. V tabuľke je zobrazených prvých 5 záznanov dátovej sady spolu s prvými tromi indikátormi.

...	**Rast populácie**	**Miera plodnosti dospievajúcich**	**Priemerná dĺžka života pri narodení, spolu**
...	0,742517	150,863	55,798
...	6,449321	148,141	56,454
...	7,541019	143,370	57,344
...	3,933178	138,827	57,944
...	3,576508	133,071	58,361

Tabuľka 5.5: Druhá časť dátovej sady. Obsahuje prvých 5 záznanov dátovej sady spolu s poslednými tromi indikátormi.

Kapitola 6

Návrh modelu

Cieľom tejto práce je vytvoriť model, ktorý rieši problém predikcie demografického vývoja populácie - budúce rozloženie vekových skupín v populácii a pohlavia v populácii, celkový počet ľudí v populácii. Na predikciu cieľových premenných závislých od času sú potrebné dva typy dát:

1. Dáta, ktoré zaznamenávajú predošlý vývoj vybranej cieľovej premmennej.
2. Hodnoty premmenných, ktoré majú vplyv na hodnotu cieľovej premennej.

6.1 Architektúra modelu

Keďže ku cieľovej predikcii sú potrebné dva druhy dát, ktorých budúci vývoj treba poznať, tak navrhovaný model sa bude skladať z dvoch samostatných modelov. Jeden model bude určený k predikcii hodnoty výstupných vlastností na základe ich predošlého vývoja a hodnôt premenných vplývajúcich na hodnotu vlastností. Druhý model je potrebný na predikciu budúcich hodnôt vlastností vplývajúcich na cieľovú charakteristiku populácie.

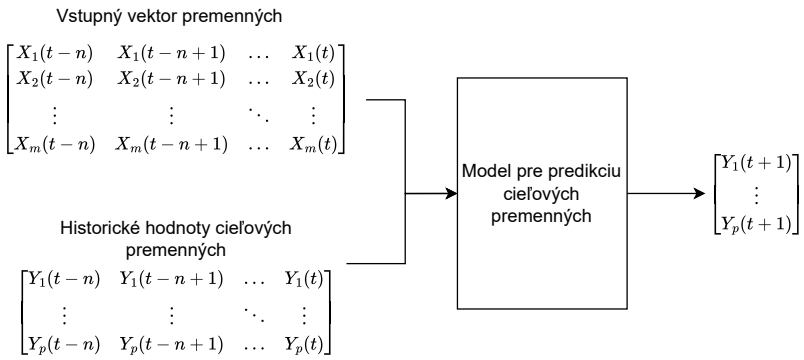
6.1.1 Model pre predikciu cieľovej premennej

Ako vstup príjma tento model historické hodnoty cieľovej premennej $Y = Y(t), Y(t-1), ..., Y(t-n)$, kde n je dĺžka zaznamenanej sekvencie minulej sekvencie a hodnoty nezávislých premenných $X(t)$ v čase t. V prípade použitia metód, ktoré umožňujú zadať aj históriu vývoja nezávislých premenných, bude namiesto vstupných hodnôt $X(t)$ zadaný vektor $X = X(t) = X(t), X(t-1), ..., X(t-n)$. Na základe vstupných údajov predpovie hodnotu cieľovej premenej v čase $Y(t+1)$. Predikcia cieľovej vlastnosti je zobrazená na obrázku 6.1. Takto vytvorený model dokáže predpovedať budúcu hodnotu parametru $Y(t)$ len do ďalšej časovej jednotky.

6.1.2 Model pre predikciu vplývajúcich vlastností na cieľovú premennú

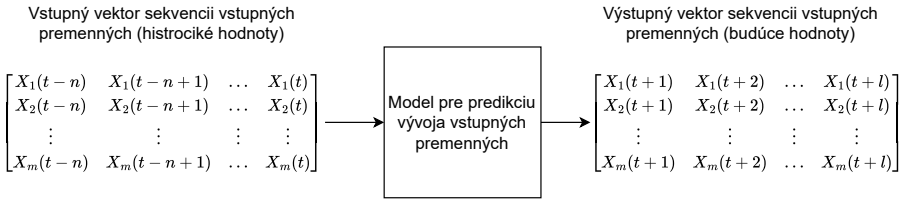
Vytvorený model by mal byť schopný predpovedať vývoj populácie od posledného do zvoleného roku. Z tohto dôvodu je nutné poznať aj vývoj premenných vplývajúcich na skúmané charakteristiky stavu populácie.

Z tohto dôvodu je navrhnutý model, ktorého úlohou je z vektora predchádzajúcich sekvencií vstupných premenných (nezahŕňajúcich predchádzajúci vývoj cieľových premenných)



Obr. 6.1: Vstupy a výstupy modelu pre predikciu vektoru cieľových premenných

$X = X(t), X(t-1), \dots, X(t-n)$, kde n je špecifikovaná dĺžka sekvencie, určiť ich hodnoty v intervale $(X(t), X(t+l))$ (viď sekcia 6.2).



Obr. 6.2: Vstupy a výstupy modelu pre predikciu vývoja vstupných premenných

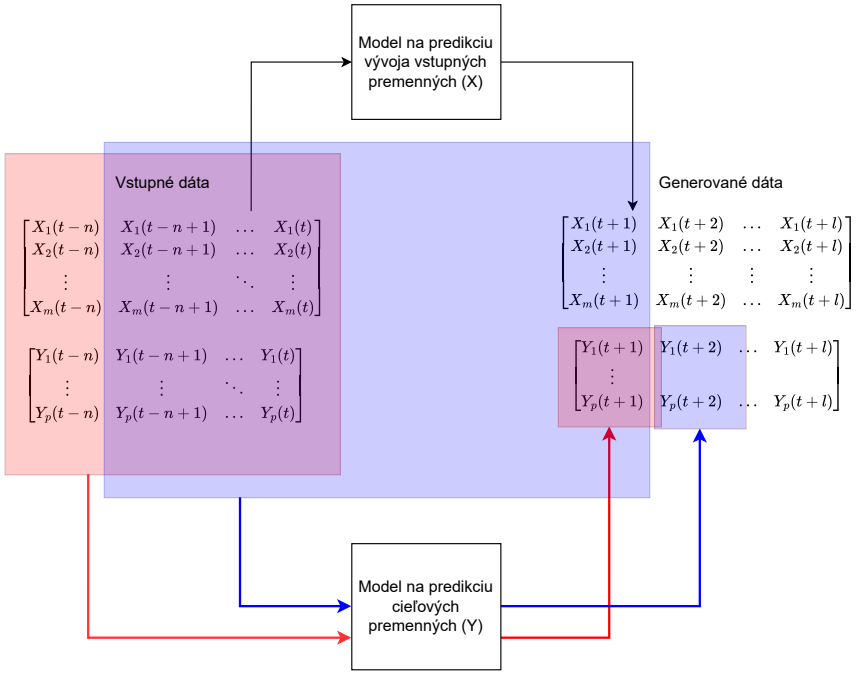
6.2 Predikcia do špecifikovaného roku

Pri návrhu modelu treba zohľadniť spôsob predikcie viacerých rokov, tzn. nie len jeden krok dopredu. Existuje niekoľko spôsobov na predikciu viacerých bodov, v tejto práci budú zohľadnené tieto 3:

1. S využitím rekurzívneho prístupu (viď sekciu 6.2.1).
2. Natrénovaním modelu pre predikciu fixnej dĺžky do budúcnosti (MIMO, viď sekciu 6.2.2).
3. S využitím rekurzívneho prístupu v kombinácii s MIMO, (viď sekcia 6.2.3).

6.2.1 Rekurzívna stratégia

V tejto stratégii je model natrénovaný na predpovedanie jedného kroku dopredu [75], teda hodnotu cieľových premenných v nasledujúcom roku. Ak je posledný rok R a požadovaný rok predikcie T, pričom platí $R < T$, aby bolo predpovedať všetky hodnoty pre roky intervalu $(R, R+T)$, tak model používa hodnoty generované v predchádzajúcich krokoch, čo zároveň spôsobuje hromadenie chýb. V závislosti od šumu prítomného v časových radoch a horizontu



Obr. 6.3: Spôsob zreťazenia modelov za sebou. Na ľavej strane obrázku sú zobrazené vstupné dáta a na pravej sú zobrazené generované dáta.

predpovede môže rekurzívna stratégia trpieť nízkym výkonom v úlohách viackrokového predpovedania [75].

6.2.2 MIMO

MIMO (Multi-Input Multi-Output) [11] je stratégia založená na predpovedaní viac krokov dopredu. Počet krokov je pevne definovaný predom. Oproti rekurzívnej stratégii sa táto stratégia vyhýba akumulácii chýb. Medzi nevýhody tohto prístupu je rovnaká štruktúra modelu pre všetky horizonty (fixná dĺžka predpovedí) [75] .

6.2.3 RECMO

RECMO [32] (rekurzívny MIMO) je kombinácia rekurzívnej stratégie a MIMO. Princíp spočíva v tom, že je vytvorený model predikujúci s krokov do budúcna. Následne sú tieto kroky zaradené do vstupu pre predikciu ďalších s krokov, až do konca generácie celej prognózy.

6.3 Zreťazenie modelov

Modely budú použité postupne za sebou: najprv bude použitý model, ktorý predpovedá vývoj premenných a výstup z neho využije druhý model na generáciu predikcií cieľových parametrov.

Na obrázku 6.3 je zobrazený spôsob generácie predikcií s použitím oboch modelov. Model na predikciu vstupných premenných X najprv vygeneruje prognózu vývoja vstupných premenných. Ak to bu

Druhý model potom pomocou vstupných dát vygenerovaných premenných predpovedá vývoj cieľových premenných, ktoré chceme predpovedať. Na obrázku 6.3 sú červenou farbou označené vstupné a výstupné dáta prvého cyklu predikcie a modrou farbou druhý cyklus modelu na predikciu cieľových parametrov.

6.4 Návrh API

Pre nasadenie a použitie modelu bude vytvorené jednoduché aplikačné programové rozhranie, prostredníctvom ktorého je možné zistiť budúci vývoj vybranej populácie do požadovaného roku.

6.4.1 Koncové body

Pre generáciu predikcií sú navrhnuté dva koncové body. Obidva príjmajú na vstupe dáta predošlého vývoja vybranej populácie, meno populácie / štátu a cieľový rok, do ktorého sa má uskutočniť predikcia. Dáta musia obsahovať údaje o rokoch. V prípade, ak dáta o vybranej populácii nie sú špecifikované, sú použité dáta, ktoré sú obsiahnuté v dátovej sade. Prehľad koncových bodov je zobrazený v tabuľke 6.1.

Metóda	**Cesta**	**Účel**
GET	/info	Zobrazí dostupné modely.
POST	/predict	Predikcia vývoja populácie do vybraného roku zadaným modelom.
POST	/lakmoos-predict	Predpoveď populácie vhodný ako vstupné dáta pre generovanie populácie.

Tabuľka 6.1: Popis koncovych bodov aplikačného programového rozhrania. a ich účel.

6.4.2 Kompatiblita s Lakmoos AI API

Lakmoos AI, s.r.o je spoločnosť zaoberajúca sa vytváraním prieskumov trhu pomocou digitálnych respondentov. Títo digitálni respondenti sú združovaní do populácií. Populácie sú vytvorené pomocou generátorov, ktoré generujú respondentov podľa rozloženia pravdepodobnosti, ktoré slúži ako vstup do týchto generátorov. Pre umožnenie tvorby budúcich populácií je vytvorený koncový bod "/lakmoos-predict"(viď 6.1), ktorý poskytuje demografické dáta vo formáte pravdepodobnosti kompatibilných s týmito generátormi.

Formát vstupu do demografického generátoru

Všeobecne generátor príjma na vstupe dáta vo forme pravdepodobnosti výskytu danej veličiny. Navrhnutý model bude kompatibilný hlavne s generátorom zodpovedným za generáciu veku a pohlavia pre jednotlivého virtuálneho respondenta[1]. Ukážka formátu je zobrazená v tabuľke 6.2. Prevod prenosu dát z formátu výstupu modelu do formátu uvedeného kompatibilného pre generátor je popísaný v sekcii 8.3.3

[1]Virtuálny respondent predstavuje simulovanú osobu so zvolenými demografickými charakteristikami (napr. vek, pohlavie), ktorú používa model na predikciu (generovanie) odpovedí.

Vek	**Celkom**	**Žena**	**Muž**
0	0,001	0,00048	0,00052
1	0,0012	0,000576	0,000624
2	0,0013	0,000624	0,000676
...	...	...	...

Tabuľka 6.2: Príklad vstupných dát do generátora základnej demografie. Zobrazené dáta nie sú reálne. Celková pravdepodobnosť predstavuje pravdepodobnosť generovania virtuálneho respondenta v danom veku a je rovná súčtu pravdepodobností výskytu muža a ženy.

Kapitola 7

Výber metód pre navrhovaný model

Ešte pred návrhom modelu je potrebné zistiť, ktoré metódy sú pre tento prípad použitia vhodné. Táto kapitola je zameraná na porovnanie zmienených metód (kapitola 4) a vyhodnotenie ich presnosti pri predikcii rozloženia vekových skupín v populácii a ich schopnosti zachytenia trendov v populácii. Metódy budú porovnané podľa rovnakých metrík.

7.1 Použité metriky

Na hodnotenie výkonu modelov v tejto práci boli použité nasledujúce metriky, ktoré patrí medzi najčastejšie používané pri regresných úlohách:

- MAE (stredná absolútna chyba).
- MSE (stredná štvorcová chyba).
- RMSE (odmocnina zo strednej štvorcovej chyby).
- MAPE (priemerná absolútna percentuálna chyba).
- Skóre R^2.

Medzi najpopulárnejšie spôsoby vyjadrenia miery presnosti modelu na predikciu časových radov sú metriky RMSE a MAPE [14]. Hoci je MSE dobrým meradlom celkovej chyby predikcie, nie je tak intuitívne a ľahko interpretovateľné ako napríklad MAE. Preto sa často uprednostňuje RMSE pred MSE, pretože MSE je vyjadrené v inej mierke ako sú dáta (namiesto od RMSE). RMSE a MSE sú citlivejšie na odľahlé hodnoty ako napríklad MAE [30].

Na rozdiel od RMSE, MAPE nezvýrazňuje extrémne chyby a je preferovaná pre svoju jednoduchú interpretáciu v percentách [1]. MAE poskytuje ľahko pochopiteľné meradlo priemernej chyby a podobne ako MAPE nezvýrazňuje extrémne odchýlky.

Metrika R^2 označuje pomer rozptylu vysvetleného regresnými modelmi vzhľadom na rozptyl y. Keďže časové rady majú vo všeobecnosti vysoký stupeň autokorelácie, aktuálnu hodnotu y možno do značnej miery vysvetliť najnovšou nameranou hodnotou y. V dôsledku toho má hodnota R^2 pre regresné modely zostrojené s použitím informácií o najnovšej predpovedanej hodnote y tendenciu byť vysoká [36]. Z tohto dôvodu je R^2 použitá len ako doplnková metrika.

Ako hlavné metriky boli v tejto práci zvolené MAE, MAPE a RMSE, a to najmä kvôli ich ľahkej interpretovateľnosti. Táto kombinácia zároveň umožňuje posúdiť veľkosť absolútnej chyby predikcie a identifikovať modely s nižšou robustnosťou voči odľahlým hodnotám.

7.1.1 MAE

Stredná absolútna chyba (MAE) vyjadruje rozdiel medzi predikovanými hodnotami a aktuálnymi hodnotami. Je počítaná ako priemer absolútnych rozdielov medzi predikovanými hodnotami a aktuálnymi hodnotami [17].

7.1.2 MSE

Stredná štvorcová chyba (MSE) meria priemer kvadratických diferencií medzi predpovedanými hodnotami a aktuálnymi hodnotami. Výpočet prebieha tak, že spočíta sumu kvadratickej odchýlky pre každý dátový bod a súčet vydelí celkovým počtom pozorovaných vzoriek. Táto metrika zvýrazňuje odľahlé hodnoty v dátach, čo môže viesť k zarovnaniu modelu k týmto hodnotám, čo môže viesť k zlému výkonu v niektorých prípadoch [17].

7.1.3 RMSE

Odmocnina zo strednej štvorcovej chyby (RMSE) je výhodná metrika týkajúca sa interpretability, pretože sa dá interpretovať priamo vzhľadom k predikovaným hodnotám. Počíta sa ako odmocnina z MSE. Podobne ako MSE je to výhodná metrika najmä vtedy, ak dáta podliehajú normálnemu rozloženiu [17].

7.1.4 MAPE

MAPE (priemerná absolútna percentuálna chyba) [19]. Často sa používa kvôli jednoduchej interpretácii. Vyjadruje relatívnu chybu k cieľovej predikcii:

$$MAPE = \frac{|g(x) - y|}{|y|} \tag{7.1}$$

$MAPE$ je hodnota chyby, g je testovaný model, $g(x)$ sú hodnoty jeho predikcií y sú aktuálne hodnoty.

7.1.5 Skóre R^2

Skóre R^2 alebo koeficient determinácie je štatistická metrika, ktorá predstavuje podiel rozptylu pre závislú premennú vysvetlenú nezávislými premennými v regresnom modeli. Nadobúda hodnoty od $-\inf$ do 1, pričom 1 indikuje najlepší výkon [17].

7.2 Výber metódy na predikciu zvolených demogofrafických parametrov

V tejto sekcii budú porovnané zvolené metódy popísané v kapitole 4 na predikciu cieľových skupín (viď tabuľku 5.3). Metódy boli porovnané pre každú skupinu zvlášť.

7.2.1 Spôsob porovnania

Modely budú na základe rovnakých vstupných premenných predikovať výstupné premenné. Porovnanie modelov bude vykonané pre každú skupinu výstupných premenných zvlášť. Vybrané modely budú predpovedať vždy výstup jednej časovej jednotky dopredu $Y(t+1)$ na základe poznaných vstupov predošlého kroku. Vstupné premenné sú zobrazené v tabuľke 7.1.

Názov premennej	**Popis**
year	Rok.
fertility_rate_total	Celková miera plodnosti.
net_migration	Čistá migrácia.
arable_land	Podeil ornej pôda.
gdp_growth	Rast HDP.
gdp	Hodnota HDP v amerických dolároch.
death_rate_crude	Hrubá miera úmrtnosti.
agricultural_land	Poľnohospodárska pôda.
rural_population_growth	Rast vidieckej populácie.
urban_population	Podiel mestskej populácie.
population_growth	Rast populácie.

Tabuľka 7.1: Vstupné premenné pre porovnanie modelov.

Okrem vstupných premenných bude vstup do modelu obsahovať aj historický vývoj cieľových premenných, ktorý bude obsahovať posledných 10 historických hodnôt premenných vybranej cieľovej skupiny (viď 5.3).

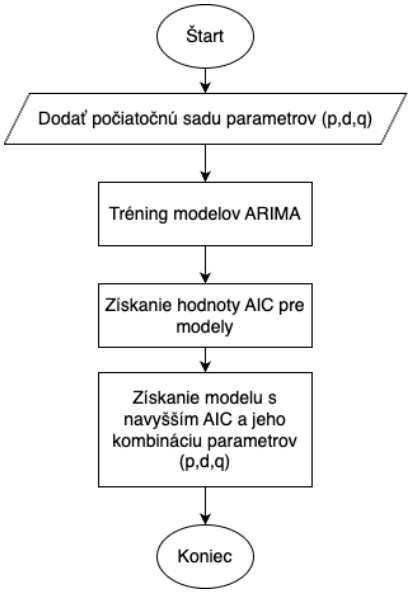
7.2.2 Porovnávané metódy

Na predikciu starnutia boli zvolené metódy vhodné na predikciu dát časových radov. Zo štatistických metód bola zvolená ARIMA (viď 4.1.1), ktorá sa veľmi často používa pre tento účel. Z prístupov strojového učenia boli zvolené stromové prístupy (viď 4.3.1) a rekurentné neurónové siete (viď 4.4.1 a jej varianty GRU (viď 4.4.3) a LSTM (viď 4.4.2). V tejto sekcii sú popísané konkrétne parametre týchto metód.

Štatistické metódy

Zo štatistických metód boli pre porovnanie vybrané 2 metódy - ARIMA a ARIMAX (viď sekciu 4.1.1). Tieto metódy vyžadujú, aby dáta boli stacionárne. Pre zaistenie stacionarity boli premenné s vysokým rozsahom hodnôt (ako sú ukazovatele celkovej migrácie, počet ľudí v populácii alebo ukazateľ hodnoty HDP) transformované pomocou logaritmickej transformácie (viď 5.3.1).

Pre každý štát a každú predikovanú premennú bol natrénovaný samostatný model ARIMA alebo ARIMAX. Základné porovnávané verzie týchto modelov sú ARIMA(1,1,1) a ARIMAX(1,1,1). Pre výber modelu bol použitý zjednodušený algoritmus na selekciu modelu ARIMA [7]. Algoritmus je zobrazený na obrázku 7.1.



Obr. 7.1: Algoritmus na výber modelu ARIMA.

Rekurentné neurónové siete

Na porovnanie boli zvolené 3 typy rekurentných neurónových sietí: klasická neurónová sieť, LSTM sieť a GRU sieť. Všetky siete obsahujú vstupnú vrstvu a výstupnú vrstvu, skryté vrstvy obsahujú len bunky typické pre daný typ siete.

Avšak všetky štandardné architektúry neurónových sietí sú nachýlné na pretrénovanie [23]. Z tohto dôvodu bol pri trénovacom experimente použitý tzv. *early stopping* (skoré zastavenie) založené na validácii [61]. Pri tréningu sa zaznamenáva trénovacia a validačná strata. Ak sa validačná strata nezlepší o definovanú úroveň, tréning sa zastaví po definovanom počte epoch, a použije sa model s najnižšou validačnou stratou.

Na výber parametrov pre rôzne neurónové siete bol použitý algoritmus, ktorý natrénuje jednotlivé neurónové typy RNN, GRU a LSTM pre rôzny počet vrstiev a rôzny počet neurónov vo vrstve. Kedže vytvorená dátová sada je pomerne malá, tak je predpokladané, že lepší výkon dosiahnu hlavne menšie modely. Z toho dôvodu boli skúšané kombinácie parametrov počtu vrstiev v sieti 1, 2, 3 v kombinácii s počtom neurónov vo vrstve 32, 64, 128, 256 a 512. Porovnávané boli na základe metriky MSE. Výsledky sú zobrazené v tabuľke 7.2.

Typ siete	Počet vrstiev	Počet neurónov vo vrstve
RNN	1	128
GRU	2	512
LSTM	2	256

Tabuľka 7.2: Najlepšie architektúry modelov z porovnávaných kombinácii.

Názov parametru	Hodnota
Koeficient rýchlosti učenia	0,0001
Počet epoch	30
Dĺžka sekvencie spracovávanej sekvencie	10
Veľkosť dávky vstupov	16

Tabuľka 7.3: Spoločné trénovacie hyerparametere pre rekurentné neurónové siete.

Stromové prístupy

Na porovnanie boli zvolené 3 typy stromových prístupov - Random Forest (náhodný les - viď 4.3.1), XGBoost (viď 4.3.1) a LightGBM (viď 4.3.1). Maximálny počet vytvorených stromov pre stromové prístupy je nastavený na 100 stromov.

7.2.3 Porovnanie metód na predikciu starnutia v populácii

Výsledok porovnania metód pre skupinu starnutia v populácii sú zhrnuté v tabuľke 7.4.

Model	MAE	MSE	RMSE	MAPE	R2
XGBoost	0,469907	0,773085	0,594448	0,023742	-3,162772
LightGBM	0,474313	0,730706	0,571284	0,023544	-3,477708
ARIMA	0,756170	1,665824	0,931751	0,037795	-7,856792
Random forest	0,748777	1,724899	0,879608	0,034142	-11,892065
ARIMAX	0,969567	3,015138	1,168805	0,044859	-49,691072
RNN	1,302317	3,052082	1,407487	0,095964	-68,266003
LSTM	1,421084	4,666500	1,546407	0,071161	-64,632170
GRU	1,606564	4,749886	1,715321	0,104869	-64,959586

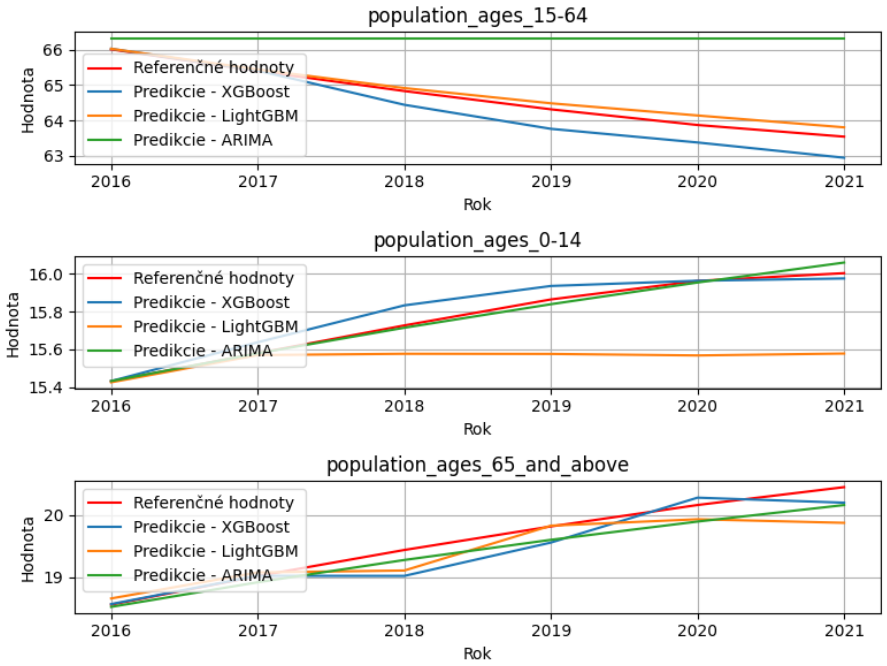
Tabuľka 7.4: Porovnanie modelov na predikciu starnutia v populácii.

Z tabuľky 7.4 vyplýva, že najlepšie navrhnutou metódou na predikciu vývoja starnutia v rôznych populáciách je XGBoost. Všetky modely majú zápornú metriku R^2, čo znamená, že všetky modely majú celkovo horší výkon pri predpovedaní nasledujúceho kroku ako predpovedanie priemeru. Celková metrika sa počíta ako priemer metrík pre všetky dostupné populácie v dátovej sade. Podľa výsledkov sú stromové prístupy vhodnejšie na predikciu ďalšej hodnoty cieľovej premennej. XGBoost a LightGBM prekonali všetky ostatné metódy.

7.2.4 Porovnanie metód na predikciu roloženia pohlaví v populácii

Výsledok porovnania metód pre premenné rozloženie pohlaví vo všetkých populáciách. Porovnanie podľa celkových metrík je zobrazené v tabuľke 7.5. Ukážka predpovedí troch modelov s najlepším výkonom na predpovedanie vývoja rozloženia pohlaví českej populácie 7.3.

Výsledky ukazujú podobnú situáciu ako pri výbere modelu starnutia. Opäť sa na prvých priečkach objavili stromové prístupy XGBoost a LightGBM, čo sa týka presnosti na základe MAE. Podľa metriky R^2 ARIMA lepšie zachytila rozptyl hodnôt cieľových premenných ako spomínané stromové prístupy.



Obr. 7.2: Najlepšie 3 modely na predikciu vývoja vekových skupín ľudí v populácii Česka.

Model	**MAE**	**MSE**	**RMSE**	**MAPE**	**R2**
XGBoost	0,152435	0,102863	0,188660	0,003103	-9,786589
LightGBM	0,164130	0,135818	0,193887	0,003360	-10,233497
ARIMA	0,194362	0,274776	0,232729	0,003949	-13,581815
Random forest	0,220588	0,459863	0,249659	0,004634	-15,561229
ARIMAX	0,272545	1,383705	0,323066	0,005545	-1425,358241
LSTM	0,690887	1,513531	0,717819	0,013913	-925,760349
GRU	0,844248	2,546971	0,883037	0,016798	-1673,681444
RNN	0,984102	2,138364	1,027802	0,020173	-1828,977594

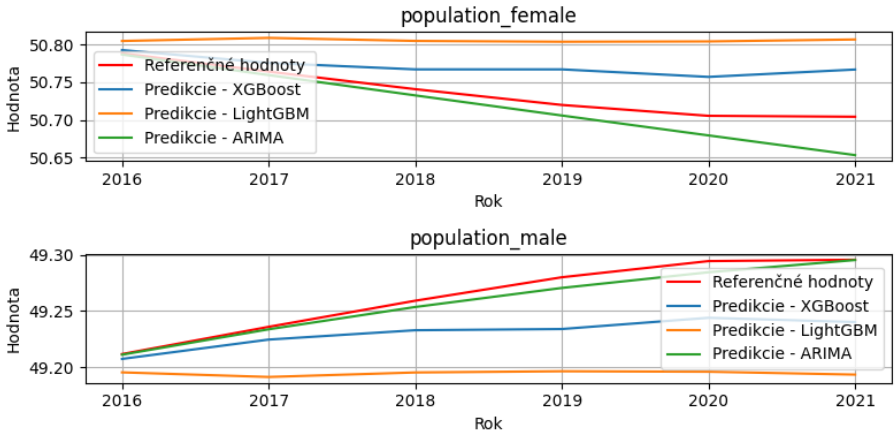
Tabuľka 7.5: Porovnanie modelov na predikciu rozloženia pohlaví – celkové hodnotenie pre všetky štáty.

7.2.5 Porovnanie metód na predikciu celkového počtu ľudí v populácii

Celkové hodnotenie metód na validačnej sade podľa zvolených metrík je zobrazené v tabuľke 7.6. Na obrázku 7.2 sú zobrazené príklady predikcií na populácii Česka. Z výsledkov vyplýva, že najlepší prístup pri aktuálnych vstupných premenných je použiť model ARIMA. Môže to byť z dôvodu chýbajúcich vstupných premenných ovplyvňujúcich celkový počet ľudí v populácii 5.7.

7.3 Výber metódy na predikciu vplyvných premenných

Účelom tohto modelu je predpovedať budúce hodnoty všetkých premenných, ktoré ovplyvňujú cieľové premenné.



Obr. 7.3: Predikcie najlepších 3 modelov na predikciu rozloženia pohlaví v Česku.

Model	**MAE**	**MSE**	**RMSE**	**MAPE**	**R2**
ARIMA	$2{,}07 \times 10^{6}$	$3{,}66 \times 10^{13}$	$2{,}62 \times 10^{6}$	0,0203	-1,8044
ARIMAX	$8{,}12 \times 10^{6}$	$1{,}15 \times 10^{15}$	$1{,}00 \times 10^{7}$	0,0394	-13,4359
XGBoost	$2{,}27 \times 10^{7}$	$9{,}60 \times 10^{15}$	$2{,}71 \times 10^{7}$	0,1208	$-1{,}6463 \times 10^{4}$
Random forest	$2{,}96 \times 10^{7}$	$2{,}06 \times 10^{16}$	$3{,}39 \times 10^{7}$	0,0620	-99,6803
LightGBM	$2{,}28 \times 10^{7}$	$1{,}21 \times 10^{16}$	$2{,}61 \times 10^{7}$	0,2252	$-7{,}8499 \times 10^{4}$
GRU	$1{,}79 \times 10^{8}$	$3{,}28 \times 10^{17}$	$1{,}84 \times 10^{8}$	0,3033	-3713,2135
LSTM	$2{,}27 \times 10^{8}$	$4{,}97 \times 10^{17}$	$2{,}31 \times 10^{8}$	0,3696	-1370,6115
RNN	$2{,}39 \times 10^{8}$	$5{,}68 \times 10^{17}$	$2{,}45 \times 10^{8}$	0,4718	-1781,7058

Tabuľka 7.6: Celkové hodnotenie výkonu modelov pri predpovedaní celkového počtu ľudí v populácii.

7.3.1 Porovnávané metódy

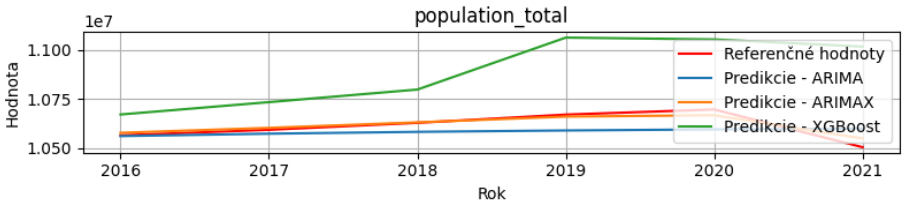
V tejto sekcii budú porovnané metódy vhodné na predikciu časových radov - ARIMA a rekurentné neurónové siete a kombinácia rekurentnej neurónovej siete s plne prepojenou vrstvou. Pri kombinovaných sieťach bola posledná rekurentná vrstva nahradená plne prepojenou vrstvou so 64 neurónmi. Parametre porovnávaných rekurentných čistých neurónových sietí sú rovnaké ako pri predošlom experimente (viď 7.2). Podobne prebiehal aj výber modelu ARIMA (viď 7.1).

Zo štúdie zameranej na porovnanie štatistických metód na analýzu časových radov a modelmi strojového učenia ako sú Random forest, majú štatistické metódy tendenciu prekonať tieto modely strojového učenia [47]. Z toho dôvodu stromový typ modelov nebol v tejto sekcii porovnaný.

Porovnávané boli 3 typy neurónových sietí: čisto rekurentné, rekurentné s jednou vrstvou pridanou plne prepojenou vrstvou a univariantné neurónové siete. Konkrétne hodnoty parametrov sú zobrazené v tabuľke 7.3.

7.3.2 Výsledky porovnania

V tabuľke 7.7 sú zobrazené výsledky porovnania modelov pre predpoveď budúceho vývoja vplyvných parametrov. Najlepšie hodnotenie podľa celkových metrík dosiahol model



Obr. 7.4: Najlepšie 3 modely na predikciu vývoja počtu ľudí v populácii v Česku.

ARIMA. Dosiahol lepšie výsledky ako rekurentné neurónové siete. Môže to byť spôsobené slabou kvalitou dostupných demografických dát alebo neseúvisiacimi vstupnými premennými. Rekurentné neurónové siete vyžadujú pre svoj tréning veľké množstvo trénovacích dát, čo môže znamenať, že daná dátová vzorka nemusí byť dostatočne veľká [41].

Model	**MAE**	**MSE**	**RMSE**	**MAPE**	**R2**
ARIMA	$7{,}10 \times 10^{10}$	$6{,}06 \times 10^{23}$	$8{,}01 \times 10^{10}$	$8{,}66 \times 10^{15}$	$-2{,}52 \times 10^{27}$
LSTM	$1{,}01 \times 10^{11}$	$1{,}07 \times 10^{24}$	$1{,}20 \times 10^{11}$	$7{,}57 \times 10^{15}$	$-1{,}91 \times 10^{29}$
RNN	$1{,}37 \times 10^{11}$	$2{,}40 \times 10^{24}$	$1{,}43 \times 10^{11}$	$8{,}62 \times 10^{14}$	$-8{,}29 \times 10^{29}$
GRU + NN	$1{,}29 \times 10^{11}$	$2{,}53 \times 10^{24}$	$1{,}42 \times 10^{11}$	$4{,}17 \times 10^{17}$	$-7{,}69 \times 10^{30}$
RNN + NN	$1{,}52 \times 10^{11}$	$3{,}08 \times 10^{24}$	$1{,}61 \times 10^{11}$	$1{,}55 \times 10^{16}$	$-4{,}12 \times 10^{28}$
GRU	$1{,}52 \times 10^{11}$	$3{,}05 \times 10^{24}$	$1{,}94 \times 10^{11}$	$2{,}19 \times 10^{15}$	$-1{,}16 \times 10^{30}$
LSTM + NN	$1{,}73 \times 10^{11}$	$4{,}07 \times 10^{24}$	$1{,}80 \times 10^{11}$	$3{,}08 \times 10^{17}$	$-7{,}33 \times 10^{29}$
Uni GRU	$1{,}87 \times 10^{11}$	$7{,}60 \times 10^{24}$	$2{,}21 \times 10^{11}$	$2{,}58 \times 10^{17}$	$-7{,}16 \times 10^{29}$
Uni LSTM	$2{,}19 \times 10^{11}$	$1{,}09 \times 10^{25}$	$2{,}60 \times 10^{11}$	$5{,}18 \times 10^{16}$	$-2{,}81 \times 10^{29}$
Uni RNN	$6{,}73 \times 10^{11}$	$2{,}18 \times 10^{26}$	$8{,}88 \times 10^{11}$	$4{,}25 \times 10^{15}$	$-1{,}01 \times 10^{29}$

Tabuľka 7.7: Porovnanie modelov na predpoveď vývoja vplyvných premenných na celej dátovej sade pre všetky premenné. Kombinované siete sú vo forme napríklad GRU + NN. Skratka *Uni* znamená, že model tvorí súbor modelov, z ktorých každý predpovedá univariantnú (samostatnú) časovú radu parametra.

Názov premennej	**Model**
agricultural_land	ARIMA
arable_land	ARIMA
death_rate_crude	ARIMA
fertility_rate_total	ARIMA
gdp	ARIMA
gdp_growth	Uni LSTM
net_migration	Uni LSTM
population_growth	LSTM + NN
rural_population_growth	ARIMA
urban_population	ARIMA

Tabuľka 7.8: Porovnanie celkové hodnotenie modelov na predikciu jednotlivých premenných samostatne.

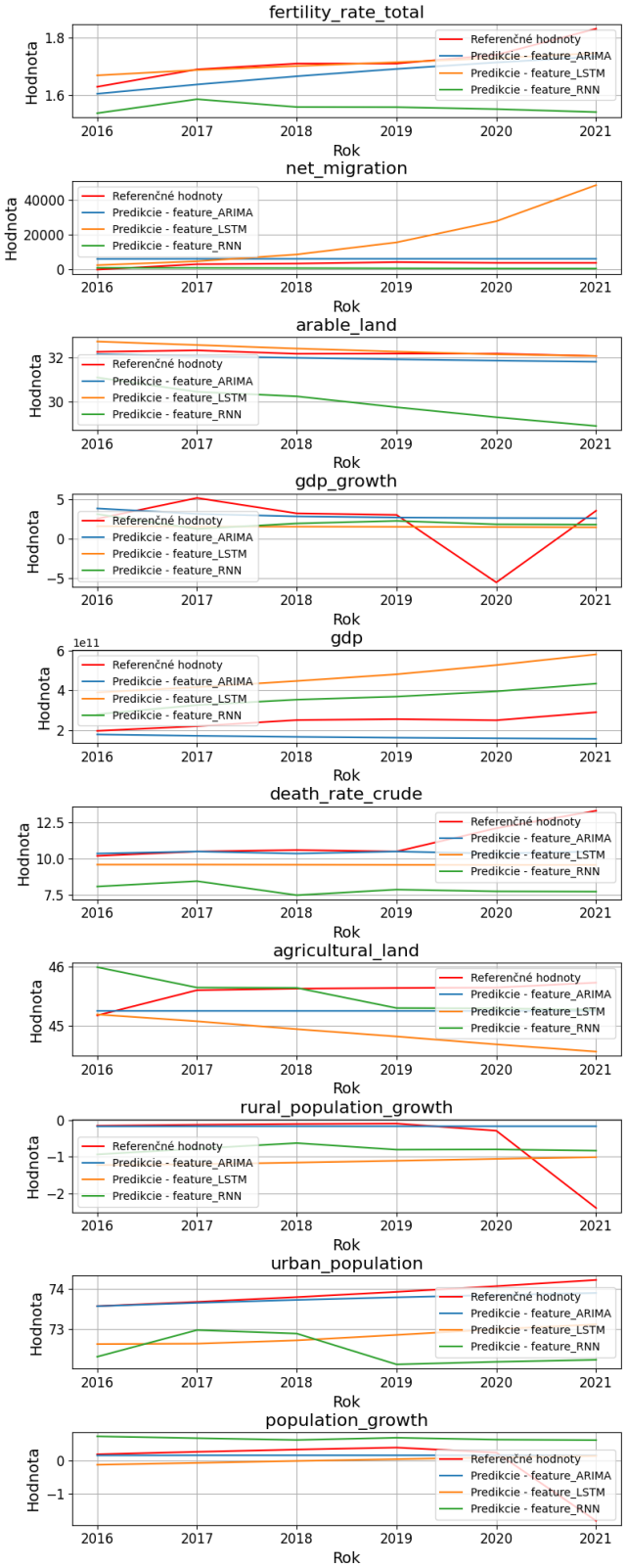
Z výsledkov v tabuľke vyplýva, že ARIMA je všeobecne zo zvolených modelov najlepšou metódou na predikciu vývoja zvolených parametrov. Avšak výsledky môžu byť zavá-

dzajúce - vysoká miera chyby môže spôsobiť, že parameter s rádovo vyššími hodnotami môže ovplyvniť celkové poradie modelu. Preto boli v tomto prípade modely porovnané podľa presnosti pre jednotlivé premenné. Výsledky sú zobrazené v tabuľke 7.8. Okrem parametrov rastu HDP (gdp_growth), čistej migrácie (net_migration) a miery rastu populácie (population_growth) je ARIMA najvhodnejšou metódou pre celkovú predpoveď vybraných parametrov.

7.4 Zhodnotenie výsledkov

Ako model na predikciu cieľových parametrov vekového rozloženia a podielu mužov a žien v populácii bol zvolený XGBoost. Cieľom tejto práce je vytvoriť univerzálny model a na danej dátovej sade dosiahol najlepšie výsledky podľa zvolených metrík. Na predikciu celkového počtu ľudí v populácii dosiahla najlepšie výsledky ARIMA, ktorá nevyužíva ostatné sekvencie, preto ju na tento účel možno použiť samostatne.

Na všeobecnú predikciu vstupných premenných do modelu na predikciu cieľových premenných bola zvolená ARIMA. ARIMA sa ukázala ako univerzálna metóda použiteľná samostatne na predikciu rôznych sekvencií. LSTM mala o niečo horšie výsledky, ale v niektorých prípadoch jej varianty dokázali predpovedať lepšie vývoj niektorých premenných (viď tabuľku 7.8), napríklad rast HDP (gdp_growth) alebo mieru čistej migrácie (net_migration). Preto bude na zhotovenie finálneho modelu použitý aj prístup s použitím modelu LSTM ako modelu na predikciu vstupných premenných.



Obr. 7.5: Príklady predikcii jednotlivých vplyvných sekvencii (vstupných premenných) pomocou 3 modelov s najlepším výkonom.

Kapitola 8

Implementácia

V tejto kapitole sú popísané implementačné detaily, abstraktný popis navrhnutého modelu, použité nástroje a knižnice.

8.1 Použité nástroje

Na implementáciu je použitý jazyk Python [63], vďaka svojej ľahko čitateľnej syntaxi, kompatibilite medzi platformami a širokej škále dostupných knižníc pre strojové učenie a dátovú analýzu [37]. Bola použitá verzia Pythonu 3.11, na zaistenie kompatibility so systémami firmy Lakmoos AI.

8.1.1 Správa balíčkov a verzovanie

Na správu balíčkov bol použitý nástroj Poetry [62] kvôli jednoduchej správe a inštalácii rôznych balíčkov. Automaticky rieši kompatibilitu verzií balíčkov a umožňuje použiť vlastné rozhranie príkazového riadku. Na verzovanie kódu bol použitý nástroj Git. Celý nástroj je dostupný na portále GitHub pod MIT licenciou[1].

8.1.2 Knižnice

Na implementáciu neurónových sietí bola použitá knižnica PyTorch [64], pretože je široko používaná a dobre podporovaná knižnica pre oblasť hlbokého strojového učenia. Umožňuje rýchlejšie trénovanie neurónových sietí pomocou paralelizácie výpočtov s využitím GPU. Pre klasické algoritmy umelej inteligencie bola použitá knižnica scikit-learn [70]. Scikit-learn ponúka okrem implementácie základných modelov umelej inteligencie aj rôzne nástroje na normalizáciu a škálovanie dát.

Na prácu s dátami bola použitá knižnica Pandas [54]. Táto knižnica je vhodná na spracovanie tabuľkových dát. Bola použitá pri tvorbe dátovej sady a pri manipulácii s dátami (napríklad pri predspracovaní dát). Pandas je kompatibilná s knižnicou NumPy [51] a PyTorch na úpravu dát (napríklad do tenzorov) na efektívnejšie spracovanie modelom. Na prevod dát do distribúcie pravdepodobností bola použitá knižnica SciPy [78].

Osobitné knižnice LightGBM [42] a XGBoost [81] boli použité na implementáciu gradient boosting modelov XGBoost a LightGBM.

Na ladenie modelu bola použitá knižnica SHAP[44]. SHAP (SHapley Additive exPlanations) je teoretický prístup založený na teórii hier na vysvetlenie výstupu akéhokoľvek

[1].https://github.com/andrewp-dot/demography-predictor

modelu. Táto knižnica používa hodnoty SHAP ako metriku na vysvetlenie vplyvov vstupných premenných na hodnotu predikcií.

Navrhnuté aplikačné programové rozhranie (API) bolo implementované pomocou knižnice FastAPI. FastAPI je širokopoužívaná a dobre zdokumentovaná knižnica na tvorbu aplikačných programových rozhraní.

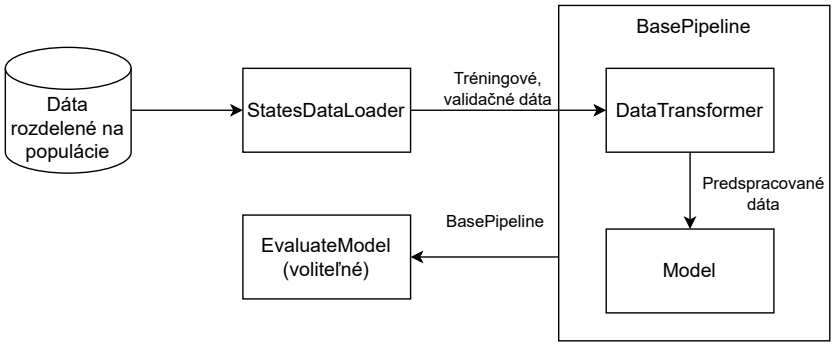
8.2 Implementácia modelu

V tejto sekcii sú popísané implementačné detaily nástroja na predikciu tvorby prediktorov experimentov popísaných v kapitole 7. Prediktor modelu je tvorený dvomi modelmi: 1. modelom na predikciu vývoja sekvencií a 2. modelom, ktorého účel je zachytiť na vstupných sekvenciách a historických údajoch parametrov predpovedať ich hodnoty do budúcna. Celý demografický prediktor je zapuzdrený do triedy PredictorPipeline. Trieda prijíma na vstupe surové dáta a na výstupe dáta v pôvodnom formáte. Celá trieda je zodpovedná za predspracovanie dát, ich spracovanie a vrátenie do pôvodného formátu.

8.2.1 Trénovanie modelov

Načítanie a rozdelenie dát na trénovacie a validačné zabezpečuje trieda `StatesDataloader`. Celá dátová sada je rozdelená na skupiny sekvencií pre jednotlivé populácie (štáty). Trénovacie a validačné dáta sú rozdelené spôsobom popísaným v sekcii 5.1.2.

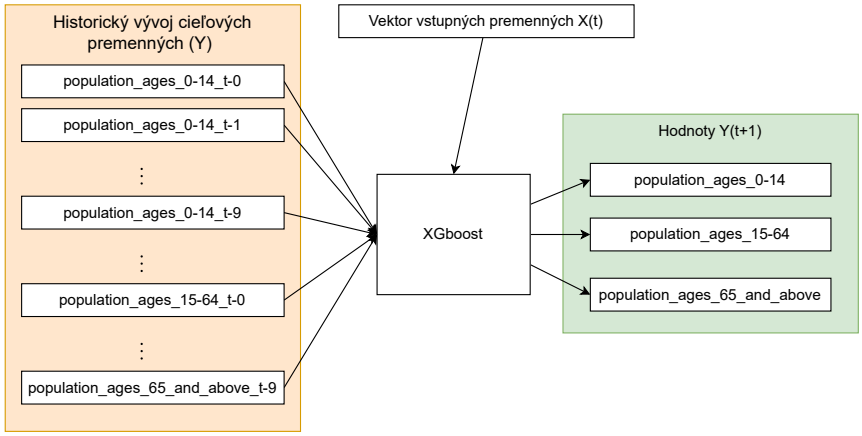
Načítané dáta sú potom predspracované pomocou triedy `DataTransformer`. Táto trieda zabezpečuje predspracovanie dát, ako je ich transformácia a škálovanie (viď 5.3.1). Jeho súčasťou je natrénovaný škálovač, ktorý sa potom uloží spoločne s modelom do jedného celku na zreťazené spracovanie. Základná trieda na tvorbu zreťazeného spracovania je `BasePipeline`. Tvorí základné rozhranie pre každú vytvorenú pipeline. Zreťazené modely slúžia potom ako vstup na vyhodnotenie presnosti modelu. Celý proces je graficky zobrazený na obrázku 8.1.



Obr. 8.1: Schéma typického trénovania modelov. Predspracované tréningové dáta idú do modelu, ktorý obsahuje funkciu na jeho tréning. Vytvorenú pipeline možno potom pomocou triedy `EvaluateModel` ohodnotiť na základe validačných dát.

8.2.2 Model na predikckiu cieľových premenných

Ako model na predikciu cieľových premenných pre starnutie a rozloženie pohlaví v populácii bol na základe experimentov vybraný XGBoost. Model s objektom triedy `DataTransformer`



Obr. 8.2: Shéma úpravý historických vstupných premennných na vstup do modelu XGBoost.

je zoskupený do triedy `TargetPredictorPipeline`. Tento typ pipeline okrem predspracovania dát ich aj prevedie do korektného formátu na vstup do modelu XGBoost. Na implementáciu bola vytvorená trieda `TargetModelTree`.

XGBoost na spracovávanie časových radov potrebuje poznať predchádzajúce hodnoty predpovedanej veličiny. Vstupný vektor obsahuje vstupné premenné a historické hodnoty v tvare oneskorených príznakov. Počet historických hodnôt cieľových premenných definuje dĺžka sekvencie. Ak je dĺžka sekvencie 10, tak vstupný vektor bude obsahovať 10 posledných historických hodnôt každej cieľovej premennej (viď obrázok 8.2. Na optimalizáciu parametrov bol využitý parameter tuning s použitím triedy `GridSearchCV` dostupnej v knižnici Scikit-learn (viď použité nástroje v sekcii 8.1.2).

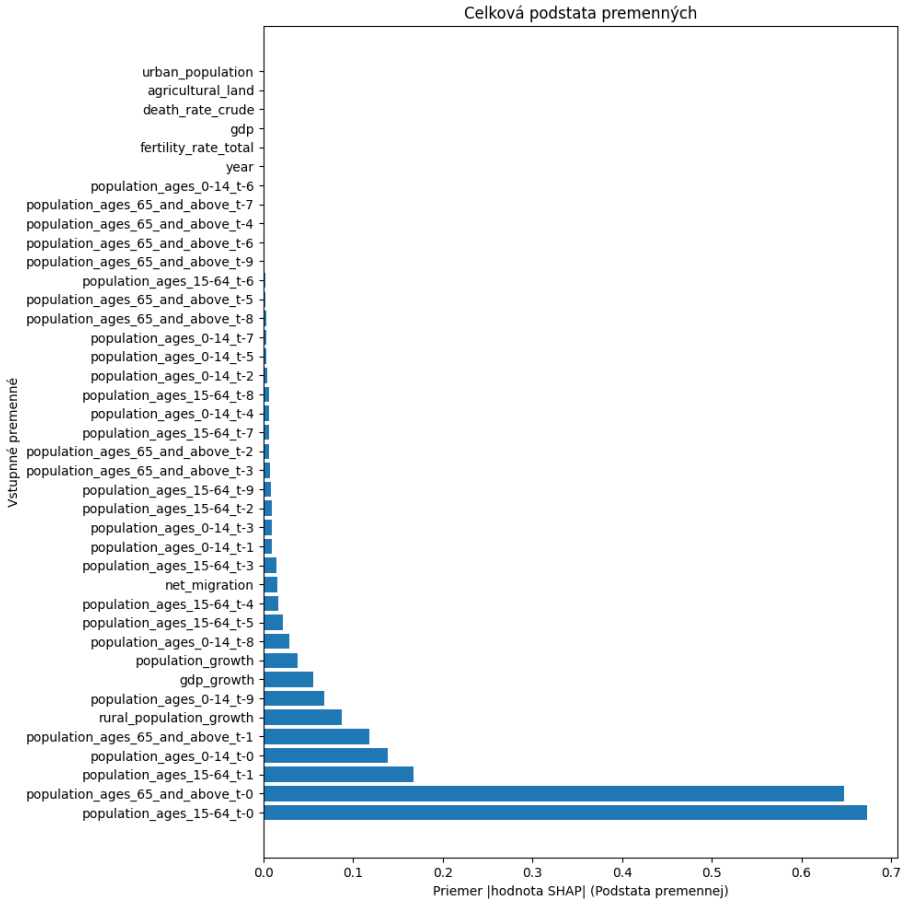
Na zobrazenie vstupných premenných a ich vplyv na vstup bol použitý stĺpcový graf z knižnice SHAP (viď 8.1.2). Tento graf zobrazuje priemernú absolútnu hodnotu SHAP pre každý prvok vo všetkých predikciách a kvantifikuje veľkosť vplyvu na predikciu modelu [58]. Na obrázku 8.3 je zobrazený stĺpcový graf pri predikcii českej populácie.

8.2.3 Model na predikciu vplyvných sekvencii

Na základe experimentov bola na tento účel zvolená metóda ARIMA, pretože dosiahla najlepšie výsledky podľa experimentu (viď tabuľku 7.7).

Implementácia modelu ARIMA

ARIMA model je potrebné natrénovať pre každú sekvenciu zvlášť. Na tento účel bola vytvorená trieda `PureEnsembleModel`, ktorá obsahuje skupinu modelov ARIMA predpovedajúca každá 1 parameter 1 sekvencie. Model je na kompatibilitu rozhrania zapuzdrený v triede `CustomARIMA`. V tejto triede je implementovaný algoritmus na hľadanie optimálnych parametrov pre danú sekvenciu. Na predpovedanie viacerých populácii je vytvorená trieda `StatisticalMultistateWrapper` zoskupujúca viac inštancií triedy `PureEnsembleModel`, každá na predikciu vývoja parametrov pre jeden štát. Tento model bude slúžiť ako základný model,

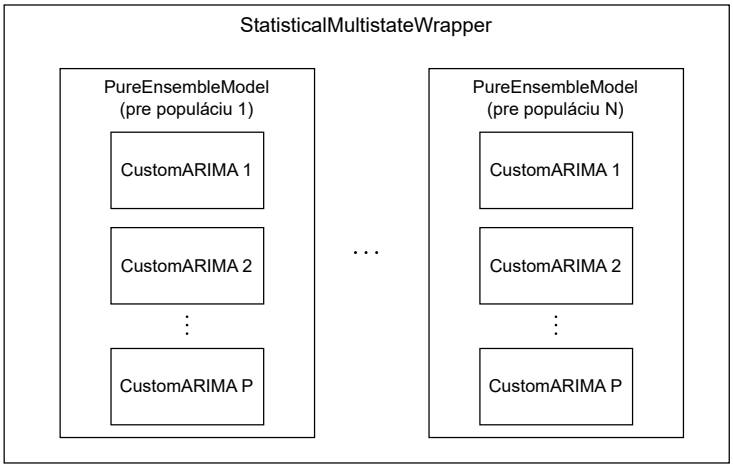


Obr. 8.3: Vstupné premenné do modelu a ich priemerná hodnota na predikciu Česka.

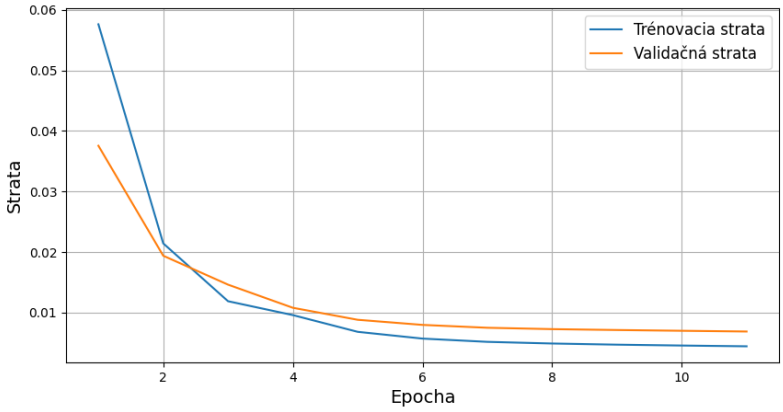
Implementácia modelu LSTM

Na špecifikáciu rekurentnej neurónovej siete je vytvorená trieda `RNNHyperparameters`. Táto trieda obsahuje všetky nastaviteľné parametre (tzv. hyperparametre) potrebné pre definíciu architektúry (počet vrstiev, počet neurónov vo vrstve, dĺžka spracovávanej sekvencie, dimenzia vstupu a výstupu) vrátane parametrov, ako sú miera učenia (anglicky *learning rate*), počet epoch alebo počet naraz spracovaných sekvencií pri trénovaní (anglicky *batch size*). Parametre natrénovaného modelu sú rovnaké ako model použitý v experimentoch na výber metódy (viď tabuľku 7.2). Definíciu architektúry, tréning a predikciu pomocou modelu LSTM zabezpečuje trieda `BaseRNN`. Ako prevencia proti pretrénovaniu bola použitá metóda skorého zastavenia (anglicky *early stopping*, viď 7.2.2).

Graf 8.5 zobrazuje konkrétne hodnoty tréningovej a validačnej straty neurónovej siete počas trénovania. Počas prvých dvoch epôch je validačná strata nižšia ako tréningová. To môže byť spôsobené tým, že v dôsledku malého množstva trénovacích dát počiatočné validačné sekvencie obsahujú hodnoty, ktoré model videl už pri trénovaní. Tréning bol zastavený po uplynutí 11 epôch následkom skorého zastavenia.



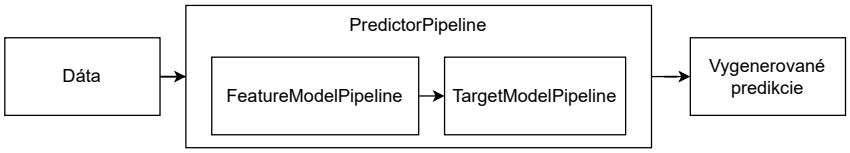
Obr. 8.4: Zložený ARIMA model na predikciu vývojov parametrov všetkých dostupných populácii.



Obr. 8.5: Hodnoty tréningovej a validačnej funkcie straty počas trénovania LSTM siete.

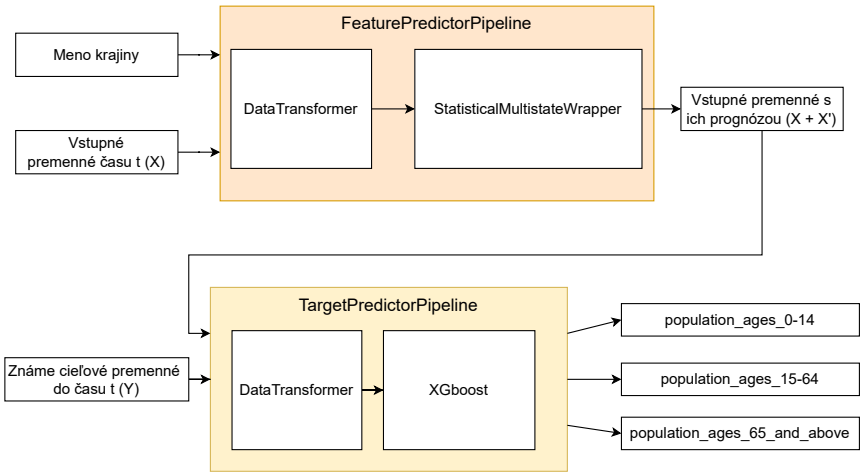
8.2.4 Demografický prediktor

Demografický prediktor je implementovaný pomocou tried `FeatureModelPipeline` na predikciu vývoja vstupných premenných a `TargetModelPipeline` na predikciu cieľových premenných. Každá z nich obsahuje vlastný dátový transformér na predspracovanie a spätnú transformáciu dát do originálneho formátu. To umožňuje to, že každý model môže použiť inú formu predspracovania dát. Abstraktná schéma je zobrazená na obrázku 8.6.



Obr. 8.6: Abstraktná schéma implementácie prediktoru.

Konkrétnu podobu prediktora na predikciu starnutia je zobrazená na obrázku 8.7. Prediktor na predikciu rozloženia pohlaví je implementovaný rovnakým spôsobom, s rozdielom výberu cieľových premenných Y. Konkrétne premenné sú zobrazené v tabuľke 5.3.



Obr. 8.7: Predikcia pomocou demografického prediktora na predikciu starnutia. Na obrázku je je zobrazený postup predikcie pomocou prediktora.

8.3 Implementácia API

Na jednoduché a rýchle nasadenie modelu bolo navrhnuté aplikačné programové rozhranie. Jeho návrh je popísaný v sekcii 6.4.

8.3.1 Informačný koncový bod

Jeho výstupom je zoznam dostupných modelov a zoznam dostupných populácií, ktoré je možné pomocou vybraného modelu predpovedať.

- **URL:** `/info`
- **Metóda:** `GET`
- **Odozva:** Objekt typu `JSON`, ktorý obsahuje:
  - `models` – zoznam dostupných modelov (napr. `["modelA_cz", "modelB"]`)
  - `populations` – zoznam populácií, ktoré možno predikovať (napr. `["Czechia", "United States"]`)
- **Stavový kód:** `200 OK` pri úspechu

8.3.2 Prediktívny koncový bod

Tento koncový bod je zodpovedný za generovanie predikcií. Jeho výstupom sú nové predikcie vybranej populácie do požadovaného roku.

- **URL:** `/predict`
- **Metóda:** `POST`
- **Vstup:** Objekt typu `JSON`, ktorý obsahuje:
  - `model_key` – názov modelu, ktorý sa má použiť
  - `state` - názov predikovanej populácie
  - `target_year` - konečný rok predikcie populácie
- **Odozva:** Objekt typu `JSON`, ktorý obsahuje:
  - `state` - názov predikovanej populácie.
  - `predictions` – predikované hodnoty.
- **Stavové kódy:**
  - `200 OK` pri úspechu.
  - `404 NOT FOUND` pri nenájdenej zadanej populácii alebo modeli.
  - `500 INTERNAL SEVER ERROR` chyba na strane servera.

8.3.3 Koncový bod kompatibliný s Lakmoos AI API

Tento koncový bod vznikol v spolupráci s firmou Lakmoos AI, s.r.o na podporu generácie vstupných dát do generátora (viď 6.4.2). Výstupom tohto koncového bodu sú okrem predikcií aj rozloženie pravdepodobnosti pre vekové alebo pohlavné rozloženie v populácii.

- **URL:** `/lakmoos-predict`
- **Metóda:** `POST`
- **Vstup:** Objekt typu `JSON`, ktorý obsahuje:
  - `model_key` – názov modelu, ktorý sa má použiť
  - `state` - názov predikovanej populácie
  - `target_year` - konečný rok predikcie populácie
  - `max_age` - maximálny predpokladaný vek virtuálneho respondenta.
- **Odozva:** Objekt typu `JSON`, ktorý obsahuje:
  - `state` - názov predikovanej populácie
  - `predictions` – predikované hodnoty.
  - `distribution` – distribúcie pre pohlavie a vek vo formáte popísanom v tabuľke 6.4.2
- **Stavové kódy:**
  - `200 OK` pri úspechu
  - `404 NOT FOUND` pri nenájdenej zadanej populácii alebo modeli

– `500 INTERNAL SEVER ERROR` chyba na strane servera

Predpovedané dáta sú prevedené na hodnoty pravdepodobnosti. Týka sa to predikcie rozloženia pohlaví v populácii a vekových skupín v populácii. Prevod podielu vekových skupín na pravdepodobnosť je implementovaný pomocou šikmo-normálneho rozloženia. Na implementáciu šikmej normálnej distribúcie bola použitá implementácia z knižnice Pythonu SciPy. Na výpočet tejto distribúcie bolo potrebné dopočítať parametre, ako vekový priemer a smerodajnú odchýľku (pozri rovnice 8.1 a 8.2) [83]. Ďalej bolo potrebné dopočítať ešte koeficient šikmosti (pozri rovnicu 8.3). Ako koeficient šikmosti bol použitý Pearsonov druhý koeficient šikmosti [21].

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i m_i}{n} \tag{8.1}$$

$$\sigma = \left[\frac{\sum_{i=1}^{k} f_i m_i^2 - \left(\frac{\sum_{i=1}^{k} f_i m_i}{n} \right)^2 \cdot n}{n-1} \right]^{1/2} \tag{8.2}$$

kde $\bar{X}$ je priemerný vek skupín a σ je smerodajná odchýľka. f_i je percentuálne zastúpenie skupiny (frekvencia výskytu), m_i je stred i-tej skupiny, k je počet skupín a n je celkový počet observácií. Výpočet koeficientu šikmosti zahŕňa hodnotu μ, ktorá predstavuje predpokladanú strednú hodnotu.

$$\text{Šikmosť}_{\text{Pearson 2}} = \frac{3(\bar{X} - \mu)}{\sigma} \tag{8.3}$$

Vypočítané pravdepodobnosti výskytu ľudí pre jednotlivú skupinu (každý rok tvorí jednu skupinu) na základe získanej funkcie hustoty pravdepodobnosti.

Kapitola 9

Evaluácia metódy

Táto kapitola je zameraná na experimentálnu evaluáciu vytvorených modelov. Prediktory sa líšia modelom na predikciu vývoja vplyvných premenných (prvým modelom).

9.1 Spôsob evaluácie

Ako vstupné dáta na evaluáciu modelov budú slúžiť posledné známe dáta jednotlivých štátov. Model potom predpovie dáta do posledného známeho roku validačných dát. Výsledok sa porovná s referenčnými hodnotami skúmaných premenných.

9.2 Evaluácia modelov pre všetky populácie

Na evaluáciu celého modelu sú použité všetky dostupné dáta. Cieľom experimentu je zistiť priemernú celkovú presnosť vytvorených modelov pre obidva ciele predikcie.

1. Model	Cieľ modelu	MAE	MSE	RMSE	MAPE	R2
ARIMA	Rozloženie pohlaví	0,1993	0,2267	0,2284	0,0041	-6,4635
	Vekové skupiny	0,5840	0,9201	0,7107	0,0293	-4,7180
LSTM	Rozloženie pohlaví	0,2501	0,3636	0,2798	0,0051	-7,4312
	Vekové skupiny	0,7112	1,4454	0,8406	0,0328	-11,7351

Tabuľka 9.1: Výkonnosť modelov na predikciu rozloženia pohlavia a veku. Metriky sú platné pre celú dátovú sadu. Hodnoty sú zaokrúhlené na 4 desatinné miesta.

Celkové hodnotenie oboch modelov je zobrazené podľa metrík v tabuľke 9.1. Metriky sú počítané váhovaným priemerom, kde evaluácia pre populácie s dlhšími sekvenciami má väčšiu váhu. Do hodnotenia sa rátajú chyby zo všetkých cieľových premenných. Zaujímavé sú hodnoty metriky R^2. Hodnoty R^2 sú v tomto prípade záporné, čo znamená, že model celkové hodnotenie predpovedí modelov sú horšie, ako predikcie priemernej hodnoty. To môže byť spôsobené akumulovanou chybou, pretože hodnota y_t môže byť do veľkej miery vysvetlená hodnotou y_{t-1} (viď 7.1).

9.2.1 Najlepšie a najhoršie predpovedané štáty

Evaluácia prebiehala aj pre každý štát zvlášť na overenie možného rozptylu chyby.

Modely rozloženia pohlaví

Tabuľky 9.2 a 9.3 zobrazujú najlepšie a najhoršie hodnotené štáty pre model predpovedajúci pohlavné rozloženie v jednotlivých štátoch.

1. Model	Meno populácie	MAE	MSE	RMSE	MAPE	R2
ARIMA	Serbia	0,0041	0,0000	0,0045	0,0001	-0,0962
	Albania	0,0076	0,0001	0,0101	0,0002	-0,1010
	Madagascar	0,0098	0,0001	0,0108	0,0002	-5,1217
LSTM	Somalia	0,0081	0,0001	0,0099	0,0002	-0,0922
	World	0,0113	0,0002	0,0131	0,0002	-0,0290
	Albania	0,0127	0,0002	0,0141	0,0003	-1,1651

Tabuľka 9.2: Najlepšie predikované štáty podľa metrík – modely na predikciu rozloženia pohlaví. Zoradené podľa metriky RMSE.

1. Model	Meno populácie	MAE	MSE	RMSE	MAPE	R2
ARIMA	Central African Republic	1,1365	1,3637	1,1678	0,0227	-17,6005
	Maldives	2,3440	8,1355	2,8523	0,0492	-1,7065
	Oman	4,5254	24,2321	4,9226	0,0985	-4,3572
LSTM	Qatar	2,2043	6,1545	2,4808	0,0580	-3,6679
	Maldives	3,4268	14,8085	3,8482	0,0717	-3,9265
	Oman	5,1762	31,3752	5,6014	0,1126	-5,9363

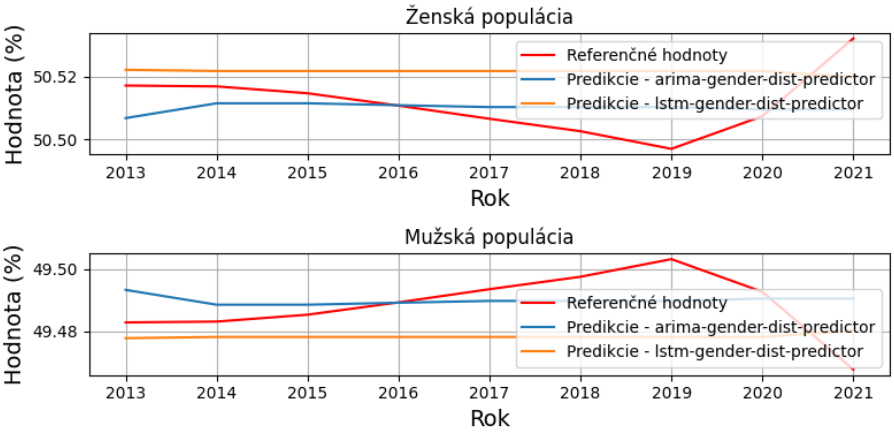
Tabuľka 9.3: Najhoršie predikované štáty podľa metrík – modely na predikciu rozloženia pohlaví. Zoradené podľa metriky RMSE.

Obidva modely vedia s pomerne malou chybou predpovedať vývoj rozloženia pohlaví v Albánsku (viď obrázok 9.1) a majú horší výkon pri predpovedaní rozloženia pohlaví v Ománe (viď obrázok 9.2). Z obidvoch obrázkov vyplýva, že modely majú problém so zachytením sezónnych trendov. Populácie so stabilnými trendami a malou varianciou vedia modely predpovedať celkom presne. Slabé zachytenie sezónnych trendov môžu spôsobiť zle zvolené atribúty pre tento trend a nedostatočné množstvo známych dát, na základe ktorých by bolo možné určiť sezónne trendy.

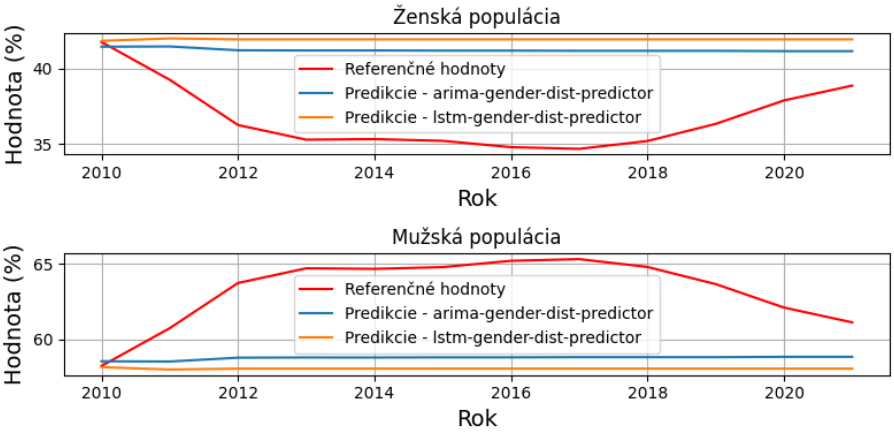
Modely na predikciu veku

V tabuľkách 9.4 a 9.5 sú zobrazené hodnotenia výkonu pre vybrané najlepšie a najhoršie hodnotené štáty podľa RMSE.

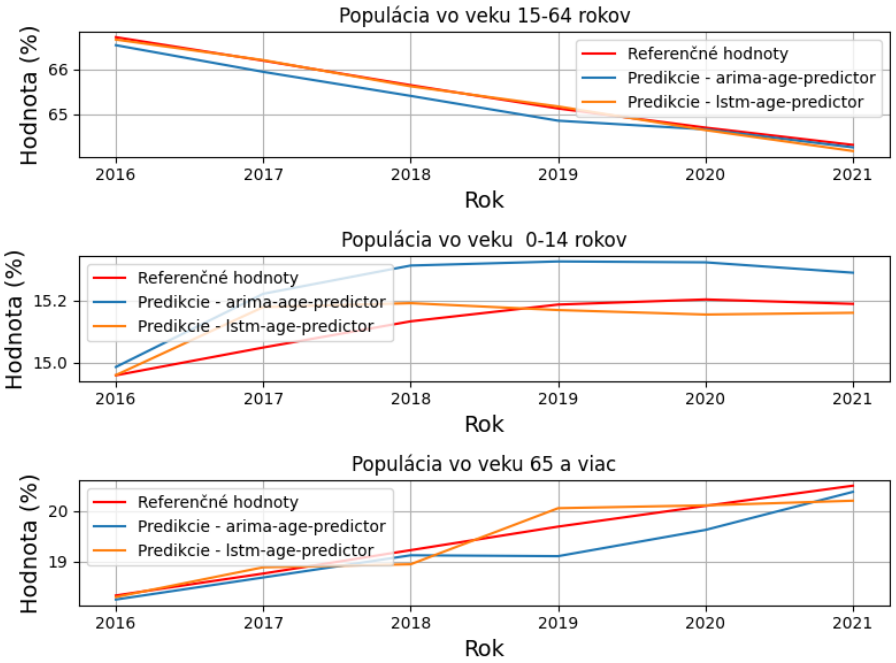
Na obrázkoch 9.3 a 9.4 môžeme vidieť, že modelu sa napriek chybám podarilo zachytiť rastúci alebo klesajúci trend vybranej vekovej skupiny. Zaujímavé je, že predikcie rozloženia vekových skupín dosiahli kladné hodnoty R^2, čo naznačuje, že model aspoň čiastočne vystihol variabilitu referenčných dát. V prípadoch, kde bol trend populácie monotónny, dokázal model predikovať vývoj pomerne presne. Naopak, nedokážu dobre predpovedať sezónny vývoj - napríklad pri predpovedi percenta ľudí od 15–64 rokov v rokoch 2012–2020 na obrázku 9.4. Čím monotónnejší je trend populácie, tým jednoduchšie ho vedia modely predpovedať.



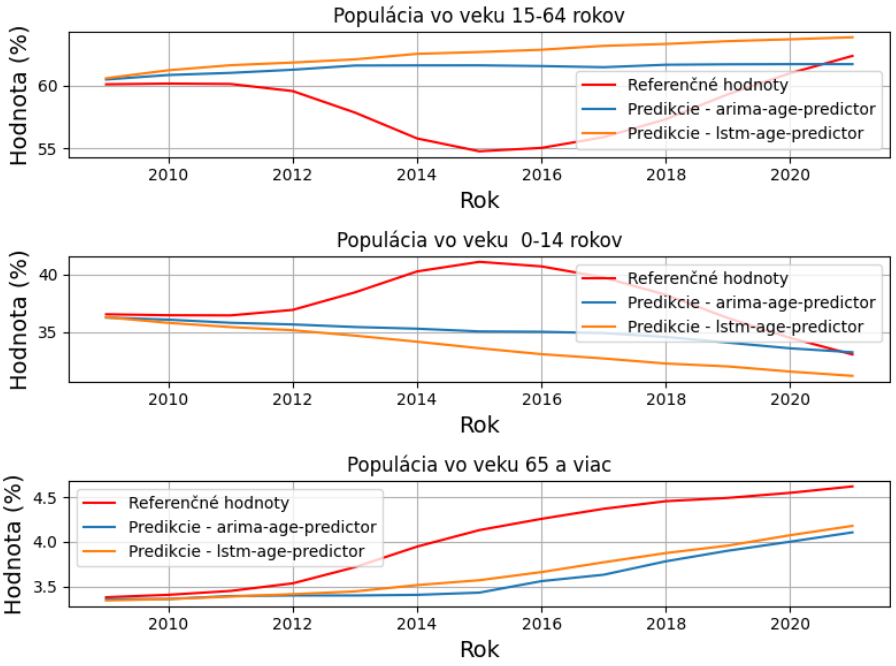
Obr. 9.1: Predikcia rozloženia pohlavia v Albánsku.



Obr. 9.2: Predikcia rozloženia pohlavia v Ománe.



Obr. 9.3: Predikcia rozloženia obyvateľstva do vekových skupín v Slovinsku.



Obr. 9.4: Predikcia rozloženia obyvateľstva do vekových skupín v Sýrskej arabskej republike.

1. Model	Meno populácie	MAE	MSE	RMSE	MAPE	R2
ARIMA	Virgin Islands (U.S.)	0,0870	0,0148	0,1004	0,0024	0,7239
	Angola	0,0871	0,0167	0,1086	0,0050	-0,4944
	Congo, Dem. Rep.	0,1071	0,0179	0,1291	0,0163	-4,1657
LSTM	Afghanistan	0,0938	0,0113	0,1058	0,0130	-55,9173
	Slovenia	0,0955	0,0204	0,1200	0,0045	0,7984
	Yemen, Rep.	0,1172	0,0179	0,1303	0,0115	-7,8439

Tabuľka 9.4: Najlepšie predikované štáty podľa metrík – modely na predikciu vekových skupín. Zoradené podľa metriky RMSE.

1. Model	Meno populácie	MAE	MSE	RMSE	MAPE	R2
ARIMA	Japan	1,9784	7,7303	2,3362	0,0780	-2,3700
	Saudi Arabia	1,8325	8,8607	2,4430	0,1688	-38,5822
	Syrian Arab Republic	2,0471	8,9241	2,5899	0,0743	-1,0072
LSTM	Qatar	2,6645	18,8226	2,8535	0,1278	-36,6061
	Syrian Arab Republic	2,7984	15,3710	3,3365	0,0867	-2,0600
	Japan	3,7821	22,4317	4,2050	0,1433	-8,7932

Tabuľka 9.5: Najhoršie predikované štáty podľa metrík – modely na predikciu roloženia vekových skupín. Zoradené podľa metriky RMSE.

9.2.2 Zhrnutie

Z experimentu vyplýva, že obidva modely majú problém so zachytením sezónnych hodnôt. Modely dokážu dobre predpovedať iba monotónne trendy. Výsledky evaluácie umožnili teda identifikovať populácie so stabilným a nestabilným trendom.

9.3 Evaluácia modelu pre skupiny populácie

Tento experiment je zameraný na zhodnotenie štátov s určitou charakteristikou, ako je miera bohatstva alebo geolokácia populácií. Cieľom tohto experimentu je zistiť, či model vie predpovedať lepšie určité skupiny štátov ako iné, alebo či je model skutočne univerzálny.

9.3.1 Rozdelené populácie podľa bohatstva

Populácie boli rozdelené podľa úrovne bohatstva na populácie s vysokým príjmom, stredným vyšším príjmom, stredným nižším príjmom a nízkym príjmom.

Na základe tabuliek 9.6 a 9.7 možno zhodnotiť, že obidva modely majú problém predpovedať populácie, ktoré patria do skupiny s vysokým príjmom. Priemerná chyba v predpovediach populácií so stredným alebo nízkym príjmom je podľa metriky RMSE približne rovnaká. Jedným z možných dôvodov je, že populácie s vyšším príjmom majú komplexnejší vývoj ako populácie s nižšími príjmami. Ďalšou možnou príčinou sú nezahrnuté faktory v dátovej sade, ktoré ovplyvňujú vývoj populácií s vyšším príjmom. Môže ísť aj o nevyvážené trénovacie dáta alebo bias, prípadne samotná metóda nemusí tieto skupiny dobre zachytiť.

Skupina	1. Model	MAE	MSE	RMSE	MAPE	R2
Nízky príjem	ARIMA	0,1055	0,0478	0,1177	0,0021	-3,9230
	LSTM	0,1353	0,0600	0,1480	0,0027	-17,5901
Nižší stredný príjem	ARIMA	0,1468	0,0648	0,1680	0,0029	-11,4737
	LSTM	0,1792	0,0971	0,2008	0,0036	-4,7752
Vyšší stredný príjem	ARIMA	0,1921	0,1854	0,2226	0,0039	-8,9102
	LSTM	0,2333	0,3091	0,2620	0,0047	-6,0309
Vysoký príjem	ARIMA	0,3606	0,6572	0,4127	0,0075	-3,0064
	LSTM	0,4773	1,0734	0,5324	0,0101	-4,9205

Tabuľka 9.6: Porovnanie výkonnosti modelov podľa príjmových skupín krajín - predikcia distribúcie pohlaví.

Skupina	1. Model	MAE	MSE	RMSE	MAPE	R2
Nízky príjem	ARIMA	0,5124	0,7917	0,6262	0,0242	-3,3319
	LSTM	0,6248	1,3089	0,7384	0,0280	-8,9190
Nižší stredný príjem	ARIMA	0,4496	0,4447	0,5482	0,0250	-7,5597
	LSTM	0,6636	1,0681	0,7809	0,0304	-25,4443
Vyšší stredný príjem	ARIMA	0,5730	0,9167	0,6973	0,0264	-2,5573
	LSTM	0,7102	1,2390	0,8528	0,0327	-7,8049
Vysoký príjem	ARIMA	0,8286	1,6767	1,0030	0,0428	-5,7705
	LSTM	0,8872	2,3899	1,0323	0,0413	-5,6577

Tabuľka 9.7: Porovnanie výkonnosti modelov ARIMA a LSTM podľa príjmových skupín – predikcia distribúcie vekových skupín.

9.3.2 Rozdelené populácie podľa geolokácie

Populácie boli rozdelené podľa úrovne kontinentov a oblastí, ako sú Európa, Ázia, Severná Amerika, Južná Amerika, Afrika a Oceánia.

Región	1. Model	MAE	MSE	RMSE	MAPE	R2
Afrika	ARIMA	0,1502	0,0777	0,1682	0,0030	-4,0282
	LSTM	0,1994	0,1046	0,2203	0,0040	-15,7081
Ázia	ARIMA	0,4253	0,9924	0,4846	0,0090	-12,5419
	LSTM	0,6194	1,7345	0,6795	0,0132	-7,1387
Európa	ARIMA	0,2415	0,1385	0,2800	0,0048	-13,4857
	LSTM	0,2391	0,1301	0,2760	0,0048	-4,9536
Severná Amerika	ARIMA	0,1325	0,0481	0,1556	0,0027	-1,7403
	LSTM	0,1404	0,0504	0,1620	0,0028	-2,1330
Oceánia	ARIMA	0,2688	0,1334	0,3152	0,0054	-12,3667
	LSTM	0,3095	0,1885	0,3614	0,0062	-8,6550
Južná Amerika	ARIMA	0,1371	0,0502	0,1616	0,0027	-2,2659
	LSTM	0,1629	0,0645	0,1866	0,0033	-5,3024

Tabuľka 9.8: Výkonnosť modelov podľa svetových regiónov – predikcia rozloženia pohlaví.

Podľa tabulky 9.8 majú obidva modely na predikciu rozloženia pohlaví najlepší výkon pri predpovedaní štátov Afriky, Severnej a Južnej Ameriky. Najväčší problém tvoria prognózy

pre ázijské štáty. To môže byť spôsobené tým, že tieto štáty majú veľmi rôznorodý vývoj, čo môže modelu sťažovať schopnosť správne generalizovať.

Región	1. Model	MAE	MSE	RMSE	MAPE	R2
Afrika	ARIMA	0,5107	0,6453	0,6062	0,0267	-5,1249
	LSTM	0,7915	1,6006	0,9196	0,0378	-15,2748
Ázia	ARIMA	0,8334	2,0726	1,0168	0,0452	-5,0723
	LSTM	1,0651	3,5255	1,2379	0,0508	-7,4047
Európa	ARIMA	0,6188	0,8544	0,7417	0,0270	-3,1185
	LSTM	0,6030	0,7662	0,7143	0,0259	-4,9553
Severná Amerika	ARIMA	0,6634	1,1161	0,8132	0,0319	-0,7860
	LSTM	0,6551	1,1070	0,7810	0,0311	-2,7771
Oceánia	ARIMA	0,4687	0,4645	0,5669	0,0257	-31,0361
	LSTM	0,6090	0,8710	0,7278	0,0317	-99,2455
Južná Amerika	ARIMA	0,5287	0,6683	0,6660	0,0243	-2,6938
	LSTM	0,5825	0,7882	0,7149	0,0275	-1,6285

Tabuľka 9.9: Výkonnosť modelov podľa svetových regiónov – predikcia vekových skupín.

Presnosť modelov pri predikciách rozloženia vekových skupín v rôznych regiónoch sa pre jednotlivé regióny s výnimkou Ázie veľmi nelíši (viď tabuľku 9.9). Podobne ako pri predikciách rozloženia pohlaví, aj tieto modely majú najväčší problém s predpovedaním ázijského regiónu. Aj v tomto prípade to môže súvisieť s vyššou komplexnosťou a rôznorodým vývojom jednotlivých krajín z tohto regiónu, ktoré model zrejme nedokáže dostatočne zachytiť.

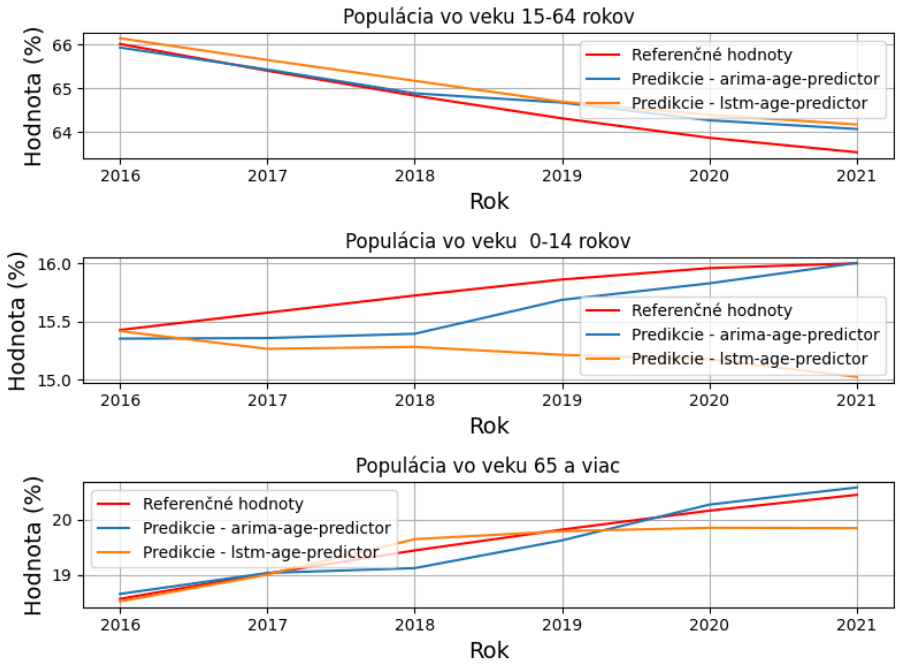
9.4 Zhodnotenie konvergencie

Existuje predpoklad, že predikcie tejto metódy začnú po čase konvergovať. Základom predpokladu je graf zobrazujúci vplyvné premenné do druhého modelu (viď 8.3). Z grafu vyplýva, že najdôležitejšie premenné ovplyvňujúce výslednú predikciu sú posledné hodnoty cieľových premenných.

Ako populácia pre tento experiment bola zvolená populácia Česka, pretože je to relatívne mladý štát. K dispozícii má malé množstvo známych historických záznamov o jeho vývoji, preto sa predpokladá, že model by mohol začať v tomto prípade konvergovať rýchlejšie. Navyše obidva modely ukázali dobrú schopnosť zachytenia trendu pri predpovedi evaluačných dát (viď obrázok 9.5). Jeho vývoj bol predpovedaný do roku 2050.

Na obrázku 9.6 sú porovnané predikcie modelov na predikciu rozloženia vekových skupín. Zaujímavé je, že model, ktorý používa ako prvý model ARIMA, začne konvergovať pre každú predpovedanú skupinu v iný rok. Najzreteľnejšia je konvergencia pre populáciu vo veku 0–14 rokov, kde začne model predpovedať od roku 2035 rovnakú hodnotu. Model s prvým modelom LSTM v tomto prípade začne konvergovať neskôr. Napriek tomu v prípade predpovede populácie vo veku 15–64 rokov začne konvergovať skôr ako model, ktorý ako prvý model používa model ARIMA.

Pri predpovedaní rozloženia pohlaví obidva modely začali konvergovať oveľa skôr (viď obrázok 9.7). Model s využitím ARIMA modelu začal konvergovať už krátko po predpovedaní rokov 2025. Model, ktorý využíva ako prvý LSTM model, začal, na rozdiel od predošlého modelu, plne konvergovať až po roku 2040.



Obr. 9.5: Porovnanie predpovedí oboch modelov na predikciu veku pre referenčné dáta Česka.

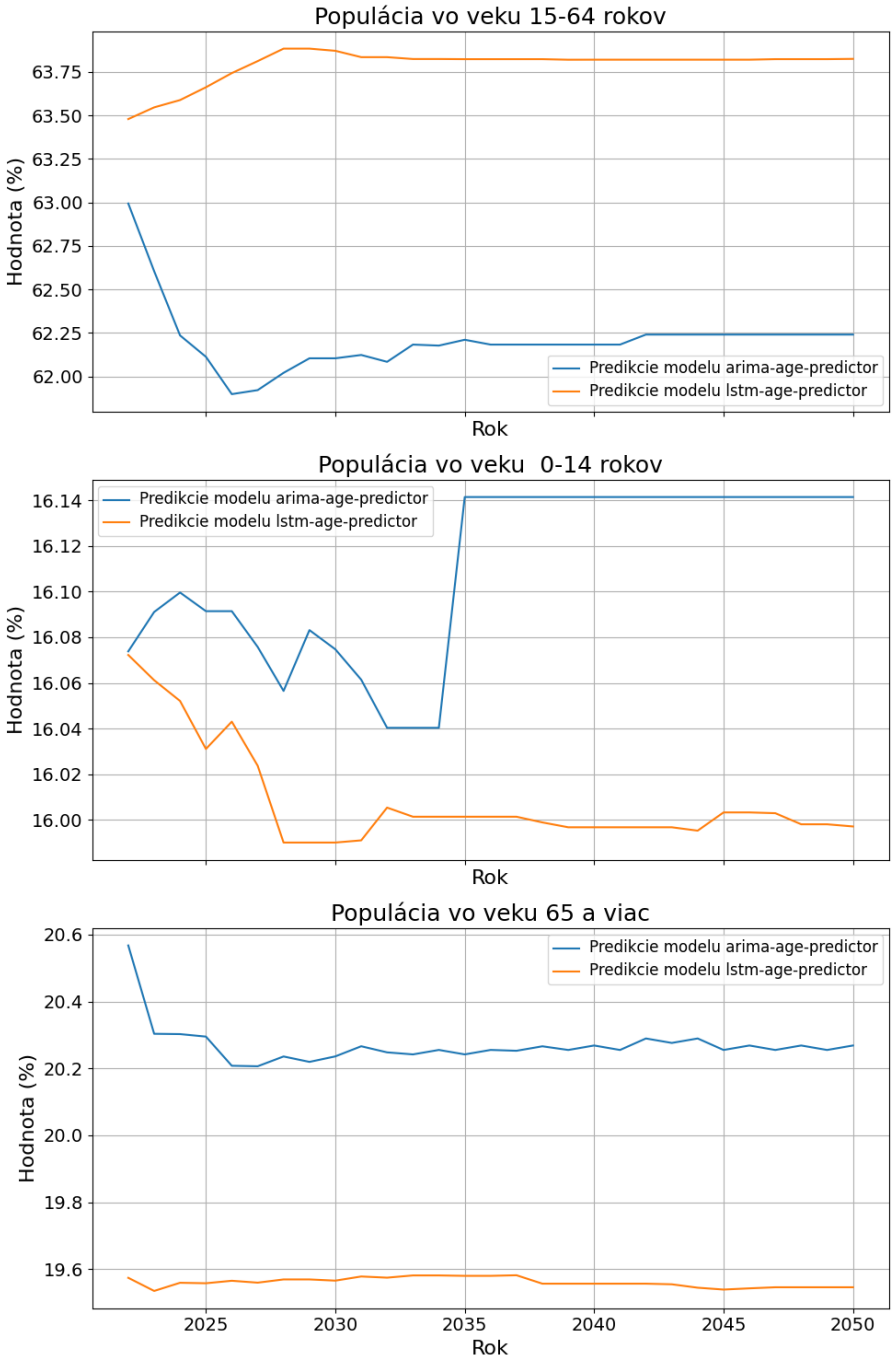
Modely na predpovedanie veku začnú konvergovať výrazne neskôr ako modely na predpoveď rozloženia pohlavia. Výsledky ukázali, že LSTM model dosahuje nižšiu presnosť ako model ARIMA, avšak na rozdiel od neho model LSTM umožňuje predpovedať vývoj viac časových krokov dopredu ako model ARIMA.

9.5 Diskusia

Táto sekcia sumarizuje výsledky vykonaných experimentov, hodnotí výkonnosť použitých prediktorov a načrtáva možné praktické využitia navrhnutej metódy. Zároveň naznačuje možné smerovanie budúceho výskumu.

9.5.1 Zhrnutie experimentov

Experimenty ukázali, že zvolená metóda je vhodná na predikciu krátkodobých a pomerne stabilných trendov. Nie je schopná na základe zozbieraných dát spoľahlivo predpovedať sezónne trendy. To môže byť spôsobené tým, že známe sekvencie pre populácie sú pomerne krátke. Metóda sa ukázala ako univerzálna, s výnimkou predpovedí pre krajiny s vysokým príjmom a krajiny patriace do regiónu Ázie. Tento typ populácií môže mať komplexnejší vývoj, ktorý daná metóda nedokáže zachytiť. V dátovej sade nie sú zahrnuté faktory, ktoré môžu ovplyvniť vývoj týchto populácií (napríklad politická situácia, úroveň vzdelania, úroveň zdravotníctva, technologický pokrok). Tieto parametre neboli zahrnuté, pretože kvalita dát, ktoré sú vhodné na tento účel, je pomerne nízka. Nájdené dáta obsahujú pre niektoré populácie veľa chýbajúcich hodnôt, sú nekonzistentné.



Obr. 9.6: Predikcia populácie Česka - model ma predikciu vekových skupín.

9.5.2 Zhodnotenie výsledkov a limity

Výsledky ukazujú, že modely sú schopné zachytiť hlavne dlhodobé trendy, vďaka čomu môžu poskytnúť dobrý obraz o vývoji populácie. Kvôli zvolenému rekurzívnemu prístupu predpovede, zostrojené modely začnú po istom predpovedanom období konvergovať. Myslím si, že je to opäť kvôli nedostatku dostupných dát. Prediktory sú teda vhodné na použitie krátkodobých predikcií.

Obr. 9.7: Predikcia populácie Česka - model ma predikciu rozloženia pohlaví.

Prediktor, ktorý využíva na predikciu vývoja vplyvných sekvencií model ARIMA, dosahuje lepšiu presnosť pre krátkodobé predikcie (približne do roku 2035). Pri predpovedaní rozloženia pohlaví je táto hranica nižšia - približne od roku 2027. Prediktor kombinujúci model LSTM s modelom XGBoost dosiahol horšie výsledky pri predpovediach rozloženia vekových skupín (v priemere o 0.13% podľa metriky RMSE), aj pri predpovediach rozloženia pohlaví (v priemere o 0.05% podľa metriky RMSE).

9.5.3 Budúca práca a odporúčania

Na základe získaných výsledkov a zistených limitácií existuje niekoľko smerov, ktorými sa nasledujúci výskum môže uberať.

Ako prvý spôsob na zlepšenie modelov je zahrnutie do dátovej sady faktory, ktoré do väčšej miery ovplyvňujú predpovedané veličiny. Vzhľadom na malé množstvo dostupných historických dát môže byť kľúčom k lepšiemu zachyteniu sezónnych trendov a komplexných závislostí pridanie atribútov a vytvorenie kvalitnejšej dátovej sady. Presnejší obraz o distribúcii veku v populácii je možné získať rozdelením cieľových premenných na viac menších častí.

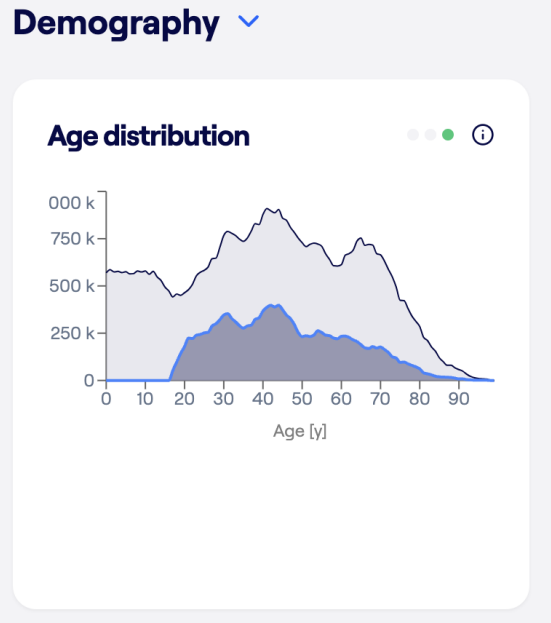
Ďalším smerom môže byť užšia špecifikácia modelu na základe spoločných rysov populácií. K tomuto účelu môžu byť využité známe fakty, ako znalosť ich polohy alebo úrovne

ekonomiky. Pokročilejší prístup môže zahŕňať rozdelenie štátov pomocou strojového učenia bez učiteľa, napríklad vytvorením zhlukov. Populácie s podobnými rysmi zoskupiť a natrénovať model, ktorý predpovedá vývoj populácií s podobnými rysmi. Ďalšou možnosťou užšej špecifikácie (na štát alebo na skupinu štátov) je doladenie (anglicky *finetuning*) modelov pomocou vytrénovanej skupiny populácií.

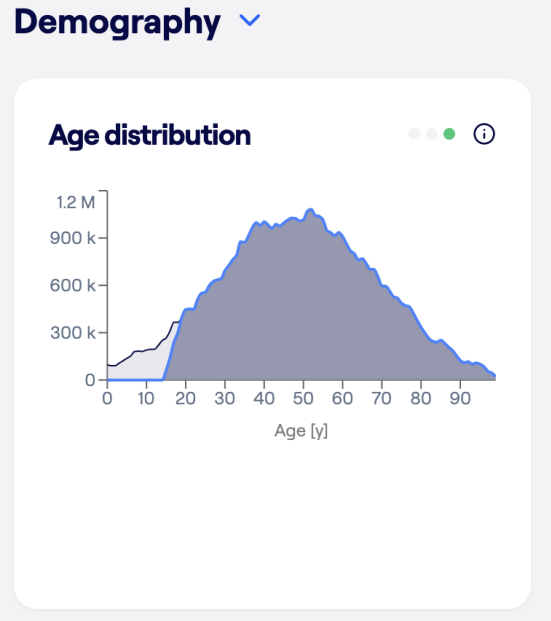
9.6 Praktické využitie

Modely sú nasadené prostredníctvom navrhnutého a implementovaného aplikačného rozhrania (API). Modely sa využívajú na predikciu a úpravu demografie pri generovaní populácie na základe predpovedaného rozloženia transformovaného do pravdepodobnostnej krivky (viď 8.3.3).

Na obrázku 9.8 je zobrazená aktuálna generovaná populácia na základe aktuálnych dát. Obrázok 9.9 ukazuje vekové rozloženie digitálnych respondentov vo vytvorenom vzorku českej populácie v roku 2035. Obrázky sú vytvorené ako snímka obrazovky z aplikácie, kde boli tieto distribúcie použité.



Obr. 9.8: Generovaná vzorka digitálnych respondentov českej populácie vytvorená na základe aktuálnych dát.



Obr. 9.9: Generovaná vzorka digitálnych respondentov českej populácie vytvorená na základe predpovedaných rozložení českej populácie v roku 2035.

Kapitola 10

Záver

Na základe porovnaných metód boli navrhnuté a implementované dva modely schopné predpovedať vývoj vybranej populácie na základe jej historických dát. Cieľom obidvoch typov modelov je predpovedať vekové a pohlavné rozloženie pre ľubovoľnú populáciu obsiahnutú v dátovej sade.

Na analýzu predošlého vývoja bola navrhnutá dátová sada spojením niekoľkých ukazateľov pre viac ako 200 populácií. Dáta boli preberané z portálu World Bank Group.

Na vytvorenej dátovej sade boli porovnané najčastejšie používané prediktívne metódy strojového učenia podľa 5 metrík — MAE, RMSE, MAPE, MSE a R^2. Na tento účel bola dátová sada rozdelená na tréningovú a validačnú. Súčasne bola rozdelená na individuálne populácie, pričom každá populácia predstavuje jeden časový rad.

Na základe výberu metódy bol navrhnutý hybridný model, ktorý kombinuje modely ARIMA a model XGBoost. Všeobecne stromové prístupy preukázali najlepší výkon spomedzi všetkých metód na predikciu distribúcie pohlavia a rozloženia vekových skupín.

Aby bolo možné predpovedať vývoj populácie na dlhšiu dobu, tak bolo potrebné predikovať aj vývoj vstupných premenných pre model XGBoost. Na tento účel boli porovnávané rôzne architektúry rekurentných neurónových sietí. Porovnávané boli jednoduché rekurentné neurónové siete, GRU a LSTM siete a kombinácie rekurentných neurónových sietí s plne prepojenou vrstvou. ARIMA prekonala všetky testované architektúry. Dôvodom môže byť dátová sada obsahujúca krátke sekvencie vývoja a pomerne malý počet trénovacích dát.

Testované boli kombinácie modelov ARIMA + XGBoost a LSTM + XGBoost. Na základe experimentov bolo zistené, že obidve kombinácie vedia celkom dobre zachytiť jednoduché a monotónne trendy. Kombinácia štatistického prístupu s prístupom strojového učenia (ARIMA + XGBoost) dokáže s pomerne uspokojivou presnosťou predpovedať hodnoty cieľových premenných pre jednotlivé štáty. Prediktor LSTM + XGBoost dosiahol horšie výsledky, čo sa týka presnosti predpovedí. Napriek tomu táto kombinácia modelov umožňuje generovanie dlhších predpovedí. Prediktor nezahŕňa vplyv nečakaných udalostí, ako sú vojny, pandémie a pod. Napriek tomu obidva zostrojené modely dokážu v pomerne dobrej miere zachytiť a predpovedať dlhodobé demografické trendy.

Výsledky práce ukazujú, že je možné vytvoriť univerzálny model na predikciu vývoja ľubovoľnej populácie, hoci s určitými obmedzeniami. V budúcnosti by bolo vhodné rozšíriť model pre lepšie zachytenie aj sezónnych trendov alebo navrhnúť niekoľko modelov s užším zameraním na populácie s podobnými rysmi.

Literatúra

[1] Adhikari, R. a Agrawal, R. K. An introductory study on time series modeling and forecasting. *ArXiv preprint arXiv:1302.6613*, 2013.

[2] Agarap, A. Deep learning using rectified linear units (relu). *ArXiv preprint arXiv:1803.08375*, 2018.

[3] Alfredo, C. S.; Adytia, D. A. et al. Time series forecasting of significant wave height using GRU, CNN-GRU, and LSTM. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 2022, zv. 6, č. 5, s. 776–781.

[4] Alghanmi, N.; Alotaibi, R.; Alshammari, S. a Mahmood, A. Population Fusion Transformer for Subnational Population Forecasting. *International Journal of Computational Intelligence Systems*. Springer, 2024, zv. 17, č. 1, s. 26.

[5] Ali, P. J. M.; Faraj, R. H.; Koya, E.; Ali, P. J. M. a Faraj, R. H. Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, 2014, zv. 1, č. 1, s. 1–6.

[6] Ashour, M. A. H.; Al Dahhan, I. A. a Sherzad, S. S. Utilizing Artificial Intelligence for Accurate Population Predictions for Iraq Through 2030, 2023, s. 17–21.

[7] Awe, O.; Okeyinka, A. a Fatokun, J. O. An Alternative Algorithm for ARIMA Model Selection. In: *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)*. 2020, s. 1–4.

[8] Bengio, Y.; Simard, P. a Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*. IEEE, 1994, zv. 5, č. 2, s. 157–166.

[9] Bloom, D. E. a Luca, D. L. The global demography of aging: facts, explanations, future. In: *Handbook of the economics of population aging*. Elsevier, 2016, sv. 1, s. 3–56.

[10] Bodiansky, E. a Rudenko, O. Artificial neural networks: architectures, training, applications. *TELETECH, Kharkov*, 2004.

[11] Bontempi, G. Long term time series prediction with multi-input multi-output local learning. *Proc. 2nd ESTSP*, 2008, s. 145–154.

[12] Booth, H. Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 2006, zv. 22, č. 3, s. 547–581. ISSN 0169-2070. Dostupné z:

https://www.sciencedirect.com/science/article/pii/S016920700600046X. Twenty five years of forecasting.

[13] Buduma, N. *Fundamentals of deep learning : designing next-generation machine intelligence algorithms.* Second edition. Sebastopol: O'Reilly, 2022. ISBN 978-1-492-08218-7.

[14] Chen, C.; Twycross, J. a Garibaldi, J. M. A new accuracy measure based on bounded relative error for time series forecasting. *PloS one*. Public Library of Science San Francisco, CA USA, 2017, zv. 12, č. 3, s. e0174202.

[15] Chen, T. a Guestrin, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, s. 785–794.

[16] Cox, P. *Demography*. Cambridge University Press, 1976. ISBN 9780521096126. Dostupné z: https://books.google.cz/books?id=naU5AAAAIAAJ.

[17] Cuantum. *Machine learning with Python : Keras, PyTorch, and TensorFlow*. First edition. Dallas: Cuantum Technologies, 2023. ISBN 9798397584081.

[18] Dai, J. a Chen, S. The application of ARIMA model in forecasting population data. In: IOP Publishing. *Journal of Physics: Conference Series*. 2019, sv. 1324, č. 1, s. 012100.

[19] de Myttenaere, A.; Golden, B.; Le Grand, B. a Rossi, F. Mean Absolute Percentage Error for regression models. *Neurocomputing*, 2016, zv. 192, s. 38–48. ISSN 0925-2312. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0925231216003325. Advances in artificial neural networks, machine learning and computational intelligence.

[20] Dey, R. a Salem, F. M. Gate-variants of Gated Recurrent Unit (GRU) Neural Networks. In: *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*. September 2017, s. 1597–1600.

[21] Doane, D. P. a and, L. E. S. Measuring Skewness: A Forgotten Statistic? *Journal of Statistics Education*, 2011, zv. 19, č. 2.

[22] Friedland, G. *Information-driven machine learning : data science as an engineering discipline*. Cham: Springer, 2024. ISBN 978-3-031-39476-8.

[23] Geman, S.; Bienenstock, E. a Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 1992, zv. 4, č. 1, s. 1–58.

[24] Grosse, R. Lecture 15: Exploding and vanishing gradients. *University of Toronto Computer Science*, 2017.

[25] Hajirahimova, M. S. a SAliyeva, A. Development of a Prediction Model on Demographic Indicators based on Machine Learning Methods: Azerbaijan Example, 2023.

[26] Hall, M. A. *Correlation-based Feature Selection for Machine Learning*. Hamilton, New Zealand, 1999. Dizertačná práca. The University of Waikato. Dostupné z: https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf.

[27] Hamilton, E. R. Measuring Populations: Mortality, Fertility, and Migration. *Global Population and Reproductive Health*. Jones & Bartlett Publishers, 2014, s. 53.

[28] Hayward, M. D. a Zhang, Z. Demography of aging. *Handbook of aging and the social sciences*. Academic Press San Diego, 2001, zv. 5, s. 69–85.

[29] Hochreiter, S. Long Short-term Memory. *Neural Computation MIT-Press*, 1997.

[30] Hyndman, R. J. a Koehler, A. B. Another look at measures of forecast accuracy. *International journal of forecasting*. Elsevier, 2006, zv. 22, č. 4, s. 679–688.

[31] Jeřábek, B. M. *Analýza vývoje počtu obyvatel v ČR*. 2011. Diplomová práce. Univerzita Pardubice Fakulta ekonomicko-správní.

[32] Ji, Y.-j.; Gao, L.-p.; Chen, X.-s. a Guo, W.-h. Strategies for multi-step-ahead available parking spaces forecasting based on wavelet transform. *Journal of Central South University*. Springer, 2017, zv. 24, č. 6, s. 1503–1512.

[33] Juan, L. a Liu, W.-j. Population Forecasting in China Based on the Grey-Markov Model. In: *2011 International Conference on Information Management, Innovation Management and Industrial Engineering*. 2011, sv. 3, s. 133–136.

[34] Julia Kim Walther, B. N. a Zitzmann, S. To Be Long or To Be Wide: How Data Format Influences Convergence and Estimation Accuracy in Multilevel Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*. Routledge, 2024, zv. 31, č. 5, s. 759–774. Dostupné z: https://doi.org/10.1080/10705511.2024.2320050.

[35] Julong, D. et al. Introduction to grey system theory. *The Journal of grey system*. Citeseer, 1989, zv. 1, č. 1, s. 1–24.

[36] Kaneko, H. Beware of r2 even for test datasets: Using the latest measured y-values (r2LM) in time series data analysis. *Journal of Chemometrics*. Wiley Online Library, 2019, zv. 33, č. 2, s. e3093.

[37] Karl, T. *6 Reasons Why Is Python Used for Machine Learning*. 2023. Dostupné z: https://www.newhorizons.com/resources/blog/why-is-python-used-for-machine-learning.

[38] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, s. 3146–3154. Dostupné z: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[39] Krenker, A.; Bešter, J. a Kos, A. Introduction to the artificial neural networks. *Artificial Neural Networks: Methodological Advances and Biomedical Applications. InTech*, 2011, s. 1–18.

[40] Lee, Y.-S. a Tong, L.-I. Forecasting energy consumption using a grey model improved by incorporating genetic programming. *Energy conversion and Management*. Elsevier, 2011, zv. 52, č. 1, s. 147–152.

[41] Li, O.; Zhao, W.; Huang, X.; Chen, Y.; Gan, L. et al. Scaling the Training of Recurrent Neural Networks on Sunway TaihuLight Supercomputer. In: Rodrigues, J. M. F.; Cardoso, P. J. S.; Monteiro, J.; Lam, R.; Krzhizhanovskaya, V. V. et al., ed. *Computational Science – ICCS 2019*. Cham: Springer International Publishing, 2019, s. 427–440. ISBN 978-3-030-22734-0.

[42] LightGBM Developers. *LightGBM Documentation*. 2025. Dostupné z: https://lightgbm.readthedocs.io/en/latest/index.html.

[43] Loh, W.-Y. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*. Wiley Online Library, 2011, zv. 1, č. 1, s. 14–23.

[44] Lundberg, S. M. a Lee, S.-I. *SHAP: Explaining the output of machine learning models* https://shap.readthedocs.io/en/latest/index.html. 2017. [cit. 2025-05-01].

[45] Mahmud Sujon, K.; Binti Hassan, R.; Tusnia Towshi, Z.; Othman, M. A.; Abdus Samad, M. et al. When to Use Standardization and Normalization: Empirical Evidence From Machine Learning Models and XAI. *IEEE Access*, 2024, zv. 12, s. 135300–135314.

[46] Mariet, Z. a Kuznetsov, V. Foundations of sequence-to-sequence modeling for time series. In: PMLR. *The 22nd international conference on artificial intelligence and statistics*. 2019, s. 408–417.

[47] Marzak, Z.; Benabbou, R.; Mouatassim, S. a Benhra, J. Forecasting Multivariate Time Series with Trend and Seasonality: A Random Forest Approach. In: Dassisti, M.; Madani, K. a Panetto, H., ed. *Innovative Intelligent Industrial Production and Logistics*. Cham: Springer Nature Switzerland, 2025, s. 128–144. ISBN 978-3-031-80775-6.

[48] Maznichenko, L. V. Time series analysis and forecasting of demographics of developed and developing countries. Igor Sikorsky Kyiv Polytechnic Institute, 2024.

[49] Minghui, G. et al. Population Prediction of China Based on ARIMA-LSTM Combined Model. *Journal of Sociology and Ethnology*. Clausius Scientific Press, 2023, zv. 5, č. 3, s. 19–25.

[50] Montero Manso, P.; Athanasopoulos, G.; Hyndman, R. J. a Talagala, T. S. FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, 2020, zv. 36, č. 1, s. 86–92. ISSN 0169-2070. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0169207019300895. M4 Competition.

[51] NumPy Developers. *NumPy Documentation*. 2025. Dostupné z: https://numpy.org/doc/. Citované dňa: 30.4.2025.

[52] Oh, C.; Han, S. a Jeong, J. Time-Series Data Augmentation based on Interpolation. *Procedia Computer Science*, 2020, zv. 175, s. 64–71. ISSN 1877-0509. Dostupné z: https://www.sciencedirect.com/science/article/pii/S1877050920316914. The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC),The 15th International Conference on Future Networks and Communications (FNC),The 10th International Conference on Sustainable Energy Information Technology.

[53] Ovchynnikova, O.; Belovsky, C. a Khan, O. Neural Network Forecasting of International Population Migration. In: *2021 11th International Conference on Advanced Computer Information Technologies (ACIT)*. 2021, s. 147–152.

[54] pandas development team. *Pandas Documentation*. 2025. Dostupné z: https://pandas.pydata.org/docs/.

[55] Pane, S. F.; Adiwijaya; Sulistiyo, M. D. a Gozali, A. A. LSTM and ARIMA for Forecasting COVID-19 Positive and Mortality Cases in DKI Jakarta and West Java. In: *2022 Seventh International Conference on Informatics and Computing (ICIC)*. 2022, s. 1–6.

[56] Park, S. H.; Kim, B.; Kang, C. M.; Chung, C. C. a Choi, J. W. Sequence-to-Sequence Prediction of Vehicle Trajectory via LSTM Encoder-Decoder Architecture. In: *2018 IEEE Intelligent Vehicles Symposium (IV)*. 2018, s. 1672–1678.

[57] Peter, Ď. a Silvia, P. ARIMA vs. ARIMAX–which approach is better to analyze and forecast macroeconomic time series. In: *Proceedings of 30th international conference mathematical methods in economics*. 2012, sv. 2, s. 136–140.

[58] Ponce Bobadilla, A. V.; Schmitt, V.; Maier, C. S.; Mensing, S. a Stodtmann, S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*. Wiley Online Library, 2024, zv. 17, č. 11, s. e70056.

[59] Poole, R. W. The statistical prediction of population fluctuations. *Annual Review of Ecology and Systematics*. JSTOR, 1978, zv. 9, s. 427–448.

[60] Poston Jr, D. L. a Bouvier, L. F. *Population and society: An introduction to demography*. Cambridge University Press, 2010.

[61] Prechelt, L. Early stopping-but when? In: *Neural Networks: Tricks of the trade*. Springer, 2002, s. 55–69.

[62] Python Poetry Contributors. *Poetry: Python dependency management and packaging tool* https://python-poetry.org/docs/. 2025. Dostupné z: https://python-poetry.org/docs/. [cit. 2025-04-28].

[63] Python Software Foundation. *Python Official Website* https://www.python.org/. 2024. [cit. 2025-04-28].

[64] PyTorch Team. *PyTorch Documentation*. 2025. Dostupné z: https://pytorch.org/docs/stable/index.html.

[65] Rasamoelina, A. D.; Adjailia, F. a Sinčák, P. A Review of Activation Function for Artificial Neural Network. In: *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. 2020, s. 281–286.

[66] Rodriguez Galiano, V.; Sanchez Castillo, M.; Chica Olmo, M. a Chica Rivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*. Elsevier, 2015, zv. 71, s. 804–818.

[67] Salehinejad, H.; Sankar, S.; Barfett, J.; Colak, E. a Valaee, S. Recent advances in recurrent neural networks. *ArXiv preprint arXiv:1801.01078*, 2017.

[68] Sarstedt, M. a Mooi, E. Regression Analysis. In: *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, s. 209–256. ISBN 978-3-662-56707-4. Dostupné z: https://doi.org/10.1007/978-3-662-56707-4_7.

[69] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, 2015, zv. 61, s. 85–117. ISSN 0893-6080. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0893608014002135.

[70] Scikit-learn Developers. *Scikit-learn: Machine Learning in Python*. 2025. Dostupné z: https://scikit-learn.org/stable/index.html.

[71] Sharma, S.; Sharma, S. a Athaiya, A. Activation functions in neural networks. *Towards Data Sci*, 2017, zv. 6, č. 12, s. 310–316.

[72] Shumway, R. H.; Stoffer, D. S.; Shumway, R. H. a Stoffer, D. S. ARIMA models. *Time series analysis and its applications: with R examples*. Springer, 2017, s. 75–163.

[73] Sitaram, D.; Dalwani, A.; Narang, A.; Das, M. a Auradkar, P. A Measure of Similarity of Time Series Containing Missing Data Using the Mahalanobis Distance. In: *2015 Second International Conference on Advances in Computing and Communication Engineering*. 2015, s. 622–627.

[74] Switrayana, I. N.; Hammad, R.; Irfan, P.; Sujaka, T. T. a Nasri, M. H. Comparative Analysis of Stock Price Prediction Using Deep Learning with Data Scaling Method. *JTIM: Jurnal Teknologi Informasi dan Multimedia*, 2025, zv. 7, č. 1, s. 78–90.

[75] Taieb, S. B.; Bontempi, G.; Atiya, A. F. a Sorjamaa, A. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert systems with applications*. Elsevier, 2012, zv. 39, č. 8, s. 7067–7083.

[76] Taylor, J. G. Neural Networks: An Overview. In: Landau, L. J. a Taylor, J. G., ed. *Concepts for Neural Networks: A Survey*. London: Springer London, 1998, s. 1–12. ISBN 978-1-4471-3427-5. Dostupné z: https://doi.org/10.1007/978-1-4471-3427-5_1.

[77] Trottier, L.; Giguere, P. a Chaib Draa, B. Parametric exponential linear unit for deep convolutional neural networks. In: IEEE. *2017 16th IEEE international conference on machine learning and applications (ICMLA)*. 2017, s. 207–214.

[78] Virtanen, P.; Gommers, R.; Oliphant, T. E. et al. *SciPy 1.6.3 Reference Guide*. 2020. Dostupné z: https://docs.scipy.org/doc/scipy/.

[79] Wang, C.-Y. a Lee, S.-J. Regional population forecast and analysis based on machine learning strategy. *Entropy*. MDPI, 2021, zv. 23, č. 6, s. 656.

[80] Webber, J. B. W. A bi-symmetric log transformation for wide-range data. *Measurement Science and Technology*. IOP Publishing, dec 2012, zv. 24, č. 2, s. 027001. Dostupné z: https://dx.doi.org/10.1088/0957-0233/24/2/027001.

[81] XGBoost Developers. *XGBoost Python API Documentation*. 2025. Dostupné z: https://xgboost.readthedocs.io/en/stable/python/python_intro.html.

[82] Xie, X. a Ma, H. Application and Optimization of ARIMA Model and BP Neural Combined Model in the Prediction of Population Aging. In: *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*. 2022, s. 1686–1689.

[83] Yilmaz, A. E. a Aktaş, S. Mean and Standard Deviation for Open-Ended Grouped Data. *Politeknik Dergisi*. Gazi University, 2022, zv. 25, č. 4, s. 1603–1611.

[84] Zakaria, M.; Mabrouka, A. a Sarhan, S. Artificial neural network: a brief overview. *Neural networks*, 2014, zv. 1, s. 2.

[85] Zaman, L.; Sumpeno, S. a Hariadi, M. Analisis Kinerja LSTM dan GRU sebagai Model Generatif untuk Tari Remo. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, May 2019, zv. 8, č. 2, s. 142.

[86] Zhang, X.; Yan, C.; Gao, C.; Malin, B. A. a Chen, Y. Predicting missing values in medical data via XGBoost regression. *Journal of healthcare informatics research*. Springer, 2020, zv. 4, s. 383–394.

Zoznam príloh

Príloha A

Inštalácia

Táto príloha stručne opisuje inštaláciu implementovaného nástroja. Pre viac informácií navštívte stránky GitHub https://github.com/andrewp-dot/demography-predictor.

A.1 Požiadavky systému

Pre úspešnú inštaláciu nástroja sú v prostredí potrebné nainštalované nasledujúce programy:

- 1. Python (verzia 3.11).
- 2. Poetry (minimálna verzia 2.1.3).

A.2 Inštalácia závislostí

Pre vytvorenie virtuálneho prostredia a inštaláciu závislostí stačí spustiť `poetry install`.

A.3 Nastavenie prostredia

Pre správne použitie programu, ako je nastavenie API a použitie rozhrania príkazového riadku (CLI) je nutné vytvoriť v koreňovom adresári konfiguračný súbor s názvom `.env`. Príklad súboru `.env` s názvom `.env.example` sa nachádza rovnako v koreňovom adresári.

Príloha B

Obsah priloženého média

Priložené dátové médium obsahuje:

- Výslednú správu vo formáte pdf.
- Zložku `demography-predictor` so zdrojovými kódmi vytvoreného nástroja.
- Zložku `dokumentacia` obsahujúca zdrojové kódy na vytvorenie tejto správy (LaTeX).

Zdrojové kódy pre implementovaný nástroj:

```
.
+-- .env
+-- .env.example
+-- .git
+-- .gitattributes
+-- .gitignore
+-- LICENSE
+-- README.md
+-- client
|   +-- client.py
|   +-- client_requests.py
+-- config.py
+-- data_science
|   +-- README.md
|   +-- archives
|   +-- data
|   +-- datasets
|   +-- feature_engineering
|   +-- preprocessors
|   +-- visualizations
+-- explanations
|   +-- arima-age-predictor
|   +-- arima-gender-dist-predictor
|   +-- lstm-age-predictor
|   +-- lstm-gender-dist-predictor
+-- imgs
```

```
+-- loggers_config.yaml
+-- logs
+-- model_experiments
|   +-- README.md
|   +-- __init__.py
|   +-- base_experiment.py
|   +-- experiment_results
|   +-- experiments
|   +-- main.py
|   +-- model_selection
+-- poetry.lock
+-- poetry.toml
+-- pyproject.toml
+-- src
|   +-- .gitattributes
|   +-- README.md
|   +-- api
|   +-- base.py
|   +-- cli
|   +-- compare_models
|   +-- evaluation.py
|   +-- feature_model
|   +-- pipeline.py
|   +-- preprocessors
|   +-- shap_explainer
|   +-- state_groups.py
|   +-- statistical_models
|   +-- target_model
|   +-- train_scripts
|   +-- utils
+-- trained_models
|   +-- arima-age-predictor
|   +-- arima-gender-dist-predictor
|   +-- lstm-age-predictor
|   +-- lstm-gender-dist-predictor
```

Zložka obsahuje súbor `.env` pre nastavenie prostredia. Hlavné zdrojové kódy sa nachádzajú v zložke `src`. Podrobnejší popis obsahu jednotlivých zložiek v zložke `src` je možné nájsť v súbore `src/README.md`. Súbor `README.md` v koreňovom adresári popisuje spôsob inštalácie a použitia daného nástroja. V zložke `data_science` sú súbory a zdrojové kódy potrebné pre vytvorenie dátovej sady použitej v práci. Vytvorené modely sú uložené v zložke `trained_models`.