SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Predikcia epidemiologických dát v súvislosti s pandémiou COVId-19 metódami strojového učenia a neurónovými sieťami

Tímový projekt

Obsah

[Úvod 1](#_Toc98146768)

[1 Zbierané dáta na Slovensku v súvislosti COVID-19 2](#_Toc98146769)

[2 Microsoft Azure 3](#_Toc98146770)

[2.1 Azure Machine Learning 3](#_Toc98146771)

[2.2 Príprava dát pre trénovanie modelov pomocou AutoML 5](#_Toc98146772)

[3 Modely strojového učenia trénované pomocou AutoML 6](#_Toc98146773)

[3.1 Seasonal Naive 6](#_Toc98146774)

[3.2 ARIMA a ARIMAX 6](#_Toc98146775)

[3.3 Decision Tree 7](#_Toc98146776)

[3.4 Random Forest 8](#_Toc98146777)

[3.5 XGBoost 9](#_Toc98146778)

[3.6 Voting Ensemble 10](#_Toc98146779)

[3.7 Stacking Ensemble 11](#_Toc98146780)

[3.8 Elastic Net 11](#_Toc98146781)

[3.8.1 Support vector machine (SVM) 12](#_Toc98146782)

[3.8.2 Metric learning 13](#_Toc98146783)

[3.8.3 Portfolio optimization 13](#_Toc98146784)

[4 Doterajšie výsledky 14](#_Toc98146785)

[4.1 Predikcia pozitívnych PCR testov 14](#_Toc98146786)

[4.2 Predikcia pozitívnych Ag testov 15](#_Toc98146787)

[4.3 Predikcia úmrtí z viacerých ukazovateľov 16](#_Toc98146788)

[4.4 Predikcia úmrtí z úmrtí 17](#_Toc98146789)

[4.5 Predikcia kompletne zaočkovaných 18](#_Toc98146790)

[4.6 Graf celkových hospitalizácií 19](#_Toc98146791)

[4.7 Graf celkových UPV 20](#_Toc98146792)

[4.8 Tabuľka dosiahnutých výsledkov predikcií 20](#_Toc98146793)

[Zoznam použitej literatúry 21](#_Toc98146794)

Zoznam obrázkov

[Obrázok 1 Ukážka csv formátu datasetu IZA 2](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246850)

[Obrázok 2 Služba Azure Machine Learning (Moreno, 2020) 4](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246851)

[Obrázok 3 Životný cyklus projektu (Microsoft, 2021) 5](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246852)

[Obrázok 4 Vytvorenie modelov ARIMAX a SARIMAX (Date) 7](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246853)

[Obrázok 5 Použitie modelu Decision Tree pri klasifikácií do viacerých tried na vybranej množine dát. Prevzatý obrázok (scikit-learn, 2020) 8](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246854)

[Obrázok 6 Model Random Forest 9](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246855)

[Obrázok 7: Jednoduchá ukážka princípu gradient boosting metódy s využitím rozhodovacích stromov. V každej interacii sa do súboru pripočítava slabý klasifikátor, ktorý dokáže problém riešiť pre obmedzenú podmnožinu vstupov. Kombináciou týchto slabých klasifikátorov dokážeme problém riešiť s vysokou spoľahlivosťou pre ľubovoľný validný vstup. Prevzatý obrázok (Río, 2018). 10](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246856)

[Obrázok 8 Voting Ensemble 11](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246857)

[Obrázok 9: Grafické znázornenie – SVM 12](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246858)

[Obrázok 10: Grafické znázornenie hlbokého - Metric learning 13](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246859)

[Obrázok 11: Grafické znázornenie ukážky - Portgolio optimalization 13](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246860)

[Obrázok 12: Graf časového vývoja počtu pozitívnych PCR testov (modrým) so 14-dňovou predikciou (červeným). 14](#_Toc100246861)

[Obrázok 13 Graf predikcie Ag z PCR, PCR%, Ag, Ag% 16](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246862)

[Obrázok 14 Graf predikcie úmrtí z PCR, PCR%, Ag, Ag% 17](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246863)

[Obrázok 15 Graf predikcie úmrtí z úmrt 18](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246864)

[Obrázok 16 Graf predikcie kompletne zaočkovaných 18](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246865)

[Obrázok 17 Graf celkových hospitalizácií 19](#_Toc100246866)

[Obrázok 18 Graf časového vývoja UPV 20](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc100246867)

[Obrázok 19 Graf predikcie hospitalizácii z parametrov Ag, Ag%, PCR, PCR% 22](#_Toc100246868)

[Obrázok 20 Graf predikcie hospitalizácii z parametrov Ag%, PCR% 22](#_Toc100246869)

[Obrázok 21 Graf predikcie hospitalizácií z parametrov Ag, PCR 23](#_Toc100246870)

[Obrázok 22 Graf predikcie hospitalizácií z parametrov kumulatívne testy, kumulatívne testy% 23](#_Toc100246871)

[Rovnica 1: Rovnica Seasonal Naive 6](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc97399506)

[Rovnica 2 Model ARIMA (PyFlux, 2016) 7](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc97399507)

[Rovnica 3 Model ARIMAX (PyFlux, 2016) 7](file:///C:\Users\Juraj\Desktop\tim_projekt.docx#_Toc97399508)

Zoznam skratiek a značiek

Ag – Antigénový test

CSV – Comma Seperated Value

IZA – Inštitút Zdravotných Analýz

OTP – režim očkovaní, testovaní, prekonaní

PCR - polymerázová reťazová reakciu s reverznou transkripciou

UPV - umelé pľúcne ventilácie

**Úvod**

*(Spracoval: Marek Kačmár, upravil: Filip Frank)*

Pandémia spôsobená koronavírusom COVID-19, ktorý bol identifikovaný v decembri 2019 v Číne, zasiahla takmer všetky aspekty našich každodenných životov. Vlády jednotlivých štátov sveta neustále zavádzajú rôzne opatrenia v snahe zmierniť dopad pandémie na ekonómiu a demografiu štátu.

Medzi spomínané opatrenia patrí napríklad prechod škôl na dištančnú formu výučby, nosenie rúšok sa stalo povinným, režim OTP a i.. Vzhľadom na negatívny dopad týchto opatrení je potrebné sledovať ich efektivitu, ktorá sa prejavuje na časovom vývoji pandemických ukazovateľov, akými sú mimo iných počet vykonaných testov, podaných vakcín či obsadenosť lôžok v nemocniciach. Metódy strojového učenia majú potenciál krátkodobo predpovedať vývoj týchto ukazovateľov, čo môže uľahčiť adekvátne rozhodovanie pri riešení pandémie. (Ciotti, 2020)

V tomto tímovom projekte skúmame možnosť využitia cloud computing platformy Microsoft Azure na trénovanie a nasadenie regresných modelov pre predpoveď vývoja ukazovateľov pandémie ochorenia COVID-19 na Slovensku. Naším cieľom je zoznámiť sa s prostredím platformy Microsoft Azure a následne s využitím verejne dostupných dát Inštitútu Zdravotných Analýz natrénovať a nasadiť čo najpresnejšie predikčné modely pre predpoveď vybraných pandemických ukazovateľov.

1. Zbierané dáta na Slovensku v súvislosti COVID-19

*(Spracovali: Branislav Vozár a Adam Šipula, upravil: Filip Frank)*

Hlavným zdrojom dát pre trénovanie našich regresných modelov je Inštitút zdravotníckych analýz, ktorý je analytickým a poradným útvarom Ministerstva zdravotníctva SR. Ide o dáta dokumentujúce časový priebeh pandémie COVID-19 na Slovensku, ktoré sú dostupné na <https://github.com/Institut-Zdravotnych-Analyz/covid19-data>

Dáta sú dostupné v tabuľkovom formáte, sledujeme v nich najmä časový vývoj týchto pandemických ukazovateľov:

* denný počet pozitívnych PCR a antigénových testov
* denné percento pozitívnych PCR a antigénových testov
* denný počet úmrtí
* počet hospitalizovaných pacientov
* Table

  Description automatically generated with low confidencepočet pacientov na umelej pľúcnej ventilácii

Obrázok 1 Ukážka csv formátu datasetu IZA

1. Microsoft Azure

*(Spracoval: Juraj Puszte*r)

Microsoft Azure je takzvaná cloud computing platforma. Poskytuje 4 rôzne formy cloud computingu a to Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS) a tiež serverless (Bigelow, 2020).

Microsoft rozdeľuje Azure cloud služby do množstva kategórií, vrátane: Compute, Mobile, Web, Storage, Analytics, Networking, Media and content delivery network, Integration, Identity, Internet of things, DevOps, Development, Security, Artificial intelligence and machine learning, Containers, Databases, Migration, Management and governance, Mixed reality, Blockchain, Intune (Bigelow, 2020).

Vlastnosti Microsoft Azure:

* Flexible – ľubovoľné presúvanie výpočtových zdrojov
* Open – podporuje takmer akýkoľvek operačný systém, jazyk, nástroj alebo framework
* Reliable – 99.95% dostupnosť SLA a 24x7 technická podpora
* Global – dáta sú uložené v geosynchrónnych dátových centrách
* Economical – platí sa iba za to, čo sa používa (Mccoy)
  1. Azure Machine Learning

*(Spracoval: Juraj Puszter)*

**Azure Machine Learning** je cloudová služba pre urýchlenie a správu životného cyklu projektov strojového učenia. Používa sa pre trénovanie, nasadzovanie modelov a správu MLOps. V Azure Machine Learning je možné vytvoriť model alebo použiť model z open source platformy, ako je napríklad Pytorch, TensorFlow alebo scikit-learn. MLOps nástroje pomáhajú s monitorovaním, pretrénovaním alebo znovu nasadením modelov (Microsoft, 2021).

**Azure Machine Learning Studio** je web portál pre ľudí využívajúcich Azure Machine Learning. Tento web portál uľahčuje a poskytuje prácu s Azure Machine Learning priamo cez internetový prehliadač (Microsoft, 2022).

**Automated Machine Learning (AutoML)** je proces automatizácie časovo náročných, opakujúcich sa úloh vývoja modelu strojového učenia (Microsoft, 2022).

AutoML poskytuje 3 typy trénovania modelov. Sú to Classification, Regression a Time-Series Forecasting.

* + Classification – rozdeľuje dáta do kategórií
  + Regression – predvída numerickú hodnotu dát na základe nezávislých prediktorov
  + Time-Series Forecasting – podobný ako regresia, ale dáta sú závislé na čase (Microsoft, 2022)

Služba Azure Machine Learning podporuje okrem automatizovaného učenia pomocou AutoML aj **code-first** riešenia pre trénovanie modelov, ako aj návrh nasadenie modelov pomocou **drag-n-drop Azure Machine Learning dizajnéra** .

Na obrázku je uvedené, čo nám služba Azure Machine Learning umožňuje riadiť.

Sú to :

* + Škálovateľné výpočty na požiadanie
  + Dátové úložiská a konektivita
  + Riadenie pracovného toku strojového učenia
  + Registrácia a správa modelov
  + Metriky a monitorovanie
  + Nasadenie modelov (Moreno, 2020)

Obrázok 2 Služba Azure Machine Learning (Moreno, 2020)

Životný cyklus každého projektu sa môže líšiť, ale zvyčajne bude vyzerať tak, ako je zobrazené na obrázku. Začína definovaním úlohy a pokračuje prípravou dát, trénovaním a validáciou modelu, nasadením modelu, monitorovaním a riadením životného cyklu. Nakoniec sa znovu dostane do prípravy dát a celý cyklus pokračuje (Microsoft, 2021).

* 1. Príprava dát pre trénovanie modelov pomocou AutoML

Obrázok 3 Životný cyklus projektu (Microsoft, 2021)

*(Spracoval: Filip Frank)*

Vstupné dáta pre trénovania získavame z repozitára IZA pomocou vlastných jednoduchých skriptov písaných v jazyku Python. Extrahované dáta majú formu CSV súborov a v porovnaní s pôvodnými dátami prešli týmito úpravami:

1. odstránenie nerelevantných stĺpcov
2. odstránenie riadkov obsahujúcich NA hodnoty
3. aplikácia 7-dňového moving average na stĺpce tabuľky (okrem nezávislej premennej dátum).

Príklady vstupných datasetov (každý uvedený dataset obsahuje aj nezávislú premennú dátum):

1. Ag, %Ag
2. PCR, %PCR
3. Ag, %Ag, PCR, %PCR
4. Ag, %Ag, PCR, %PCR, Úmrtia
5. Úmrtia
6. Modely strojového učenia trénované pomocou AutoML
   1. Seasonal Naive

*(Spracoval: Marek Kačmár)*

Seasonal naive je jeden z viacerých modelov, ktoré používa Microsoft Azure pri predpovedi daného pripraveného datasetu. Tento model je veľmi jednoduchý, no napriek tomu efektívny.

Obrázok, na ktorom je text

Automaticky generovaný popisPoužíva sa hlavne pri sezónnych dátach, t. j. každá predpoveď je rovná poslednej pozorovanej hodnote v rovnakej „sezóne“ v danom roku (napr. pri predpovedaní hodnôt z mesiaca Február si model bude brať hodnoty z Februára minulého roku). (T. Baldigara, 2015)

Jeho rovnica vyzerá nasledovne:

Rovnica 1: Rovnica Seasonal Naive

kde m predstavuje hodnotu sezónnej periódy a k je výsledné číslo výpočtu (h-1)/m – počtu kompletných rokov v predpovedi pred stanoveným časom T + h. Model má najlepšiu efektivitu, keď je použitý pri ekonomických alebo finančných predpovediach. (Hyndman, 2018)

* 1. ARIMA a ARIMAX

*(Spracoval: Juraj Puszter)*

ARIMA je skratka pre Auto Regressive Integrated Moving Average. Je to trieda modelov ktorá predpovedá zadané časové rady (time series) na základe svojich vlastných minulých hodnôt. Auto regresive v názve znamená, že je to lineárny regresný model (Kravchuk, 2017).

Model je charakterizovaný 3 výrazmi p, d, q.

* p - Odkazuje na počet oneskorení (lags) Y ktoré sa majú použiť ako predikáty.
* q - Odkazuje na počet oneskorených chýb predpovede (lagged forecast errors), ktoré by mali ísť do ARIMA modelu.
* Obrázok, na ktorom je text, hodinky

  Automaticky generovaný popisd – Minimálny počet rozdielov potrebných na to, aby séria bola stacionárna (Prabhakaran, 2021).

Rovnica 2 Model ARIMA (PyFlux, 2016)

Obrázok, na ktorom je text, hodinky

Automaticky generovaný popisModel ARIMAX je rozšírený model ARIMA. Obsahuje ďalšie nezávislé premenné (prediktory). Model sa tiež uvádza ako vektor ARIMA alebo dynamický regresný model (Kravchuk, 2017).

Rovnica 3 Model ARIMAX (PyFlux, 2016)

Diagram

Description automatically generatedNasledujúci obrázok znázorňuje, ako je kombinovaná lineárna regresia, ARIMA a SARIMA pre vytvorenie modelov ARIMAX a SARIMAX (Date).

Obrázok 4 Vytvorenie modelov ARIMAX a SARIMAX (Date)

* 1. Decision Tree

*(Spracoval: Adam Vozár)*

Rozhodovací strom (Decision Tree) je model na podporu rozhodovania, ktorý sa využíva na klasifikáciu a regresiu. Patrí medzi učenie s učiteľom. Jeho cieľom je vytvorenie modelu, ktorý predpovedá hodnotu premennej pomocou jednoduchých if-else pravidiel. Vďaka jeho jednoduchej interpretácii a možnej vizualizácii je veľmi ľahký na pochopenie. V porovnaní s inými modelmi, vyžaduje iba minimálne zásahy do sledovaných dát. Dáta nie je potrebné normalizovať, pre plnú funkčnosť je však nutné odstrániť chýbajúce hodnoty. (scikit-learn, 2020)



Obrázok 5 Použitie modelu Decision Tree pri klasifikácií do viacerých tried na vybranej množine dát. Prevzatý obrázok (scikit-learn, 2020)

* 1. Random Forest

*(Spracoval: Viet Quoc Le)*

Náhodný les, anglický Random Forest sa skladá z mnohých rozhodovacích stromov (Decision tree) preto sa volá les. Tento model sa dá používať na klasifikáciu aj regresiu. Každý strom spraví svoju predpoveď, môžeme to prirovnať k hlasovaniu. Každý strom má jeden hlas. Pri klasifikácií výsledná trieda lesu je tá, ktorá mala najviac hlasov. Regresia nepredpovedá triedu, ale číselne vyjadrený predpoklad, ktorý nemusí byť celé číslo. Výsledný výsledok lesu pri regresií je priemerná hodnota výsledkov stromov (Donges, 2021).

Dôležité pre náhodný lesy je zabezpečiť variáciu a nezávislosť. Toto dosiahneme použitím Bagging metódy. Bagging rozdelí trénovací dataset na menšie náhodne datasety, ktoré sa pridelia stromom, každý strom preto bude mať náhody datasetDiagram

Description automatically generated (Patro, 2017).

Obrázok 6 Model Random Forest

* 1. XGBoost

*(Spracoval: Filip Frank)*

XGBoost (skratka pre Extreme Gradient Boosting) je implementácia gradient boosting metódy súborového strojového učenia s učiteľom vytvorená pre efektívne riešenie klasifikačných a regresných problémov. Gradient boosting využíva ako slabý klasifikátor (regresor) rozhodovacie stromy, ktoré sú jeden za druhým kombinované do silného klasifikátora (regresora) za účelom minimalizácie chybovej (loss) funkcie. Aby sa zabránilo pretrénovaniu silného regresora, využívajú sa metódy regularizácie pre udržanie rozhodovacích stromov na úrovni slabých klasifikátorov (stromy majú obmedzenú hĺbku, počet listov, využívajú sa L1/L2 regularizácie pre váhy listov, atď.). Hlavnou výhodou XGBoost implementácie je jej škálovateľnosť: XGBoost vylepšuje gradient boosting metódu o optimalizačné algoritmy, ktoré zabezpečujú komparatívne rýchlejší priebeh trénovania a minimalizujú nároky na potrebné výpočtové zdroje.  (Tianqi Chen, 2016)

Obrázok 7: Jednoduchá ukážka princípu gradient boosting metódy s využitím rozhodovacích stromov. V každej interacii sa do súboru pripočítava slabý klasifikátor, ktorý dokáže problém riešiť pre obmedzenú podmnožinu vstupov. Kombináciou týchto slabých klasifikátorov dokážeme problém riešiť s vysokou spoľahlivosťou pre ľubovoľný validný vstup. Prevzatý obrázok (Río, 2018).

* 1. Voting Ensemble

*(Spracoval: Tomáš Singhofer)*

Model Voting Ensemble je model, ktorý predikuje odpoveď na základe viacerých predpovedí z iných modelov. Môže byť použitý ako na regresiu tak aj na klasifikáciu

V prípade regresie model vytvorí predpoveď, ktorá je priemerom viacerých iných regresných modelov.

Klasifikácia sa delí na dve hlavné metódy a to hard a soft voting. Pri metóde hard voting model sčíta hlasy z viacerých modelov a vyberie sa trieda s najväčším počtom hlasov. Pri soft votingu sa sčítajú pravdepodobnosti a vyberie sa trieda s najväčšou pravdepodobnosťou. (Brownlee, 2020)



Obrázok 8 Voting Ensemble

* 1. Stacking Ensemble

*(Spracoval: Filip Frank)*

Stacking je technika súborového strojového učenia, pre ktorú sú typické dve úrovne trénovania modelu (Brownlee, 2018):

1. Trénovanie jednotlivých podmodelov (môžu to byť slabé i silné klasifikátory/regresory)
2. Trénovanie druhej úrovne, v ktorej sa model učí predikcie týchto podmodelov nakombinovať tak, aby poskytol optimálnu odpoveď
   1. Elastic Net

*(Spracoval: Branislav Šipula)*

ElasticNet je penalizačný regresný model rozšírením lineárnej regresie, ktorá pridáva regulačné sankcie k funkcii, straty počas tréningu. Ako vyhodnotiť model elastickej siete a použiť konečný model na predpovedanie nových údajov. Ako nakonfigurovať model ElasticNet pre nový súbor údajov pomocou vyhľadávania v mriežke a automaticky.

Lineárna regresia označuje model, ktorý predpokladá lineárny vzťah medzi vstupnými premennými a cieľovou premennou. Pri jedinej vstupnej premennej je tento vzťah čiarou a pri vyšších rozmeroch si tento vzťah možno predstaviť ako tzv. „hyper-rovinu“, ktorá spája vstupné premenné s cieľovou premennou. Jedným z populárnych trestov je penalizácia modelu na základe súčtu hodnôt štvorcových koeficientov. Toto sa nazýva trest L2. Pokuta L2 minimalizuje veľkosť všetkých koeficientov, aj keď zabraňuje odstráneniu akýchkoľvek koeficientov z modelu.

Ďalšou populárnou penalizáciou je penalizácia modelu na základe súčtu hodnôt absolútnych koeficientov. Toto sa nazýva trest L1. Pokuta L1 minimalizuje veľkosť všetkých koeficientov a umožňuje, aby sa niektoré koeficienty minimalizovali na hodnotu nula, čím sa z modelu odstráni predikujúci ukazovateľ (ukazovateľ úspešnosti alebo neúspešnosti).

Táto technika kombinuje LASSO (operátor najmenšieho absolútneho zmrštenia/výberu) a využitie penalizačných funkcií (L1 a L2). Príkladov použitia ElasticNet je hneď niekoľko, uvedieme si niektoré v nasledujúcich podkapitolách. (Jason Brownle, 2020)

* + 1. Support vector machine (SVM)

*(Spracoval: Branislav Šipula)*

SVM v preklade znamená *metóda podporných vektorov* a používa sa lineárnu klasifikáciu alebo regresnú analýzu problému. Jej cieľom je nájsť tzv. „hyper-rovinu“ teda optimálne rozdelenie priestoru do dvoch tried. (sk.education-wiki, 2022)



Obrázok 9: Grafické znázornenie – SVM

* + 1. Metric learning

*(Spracoval: Branislav Šipula)*

Jedná sa o učenie pomocou funkcie vzdialenosti medzi objektami, detekcia bodov. Využitie má v k-NN klasifikácií, klasifikácií obrázkov, zhlukovaní, získavaní informácií a na riešenie problémov ako „učenie pod dohľadom“ alebo „učenie pod slabým dohľadom“. (CJ Carey, Yuan Tang, William de Vazelhes, Aurélien Bellet and Nathalie Vauquier, 2015)

****

Obrázok 10: Grafické znázornenie hlbokého - Metric learning

* + 1. Portfolio optimization

*(Spracoval: Branislav Šipula)*

Ide o proces výberu najlepšieho portfólia, distribúcia aktív, zo súboru všetkých zvažovaných portfólií podľa určitého cieľa. Cieľ zvyčajne maximalizuje faktory, ako je očakávaný výnos, a minimalizuje náklady, ako je finančné riziko. Faktory, ktoré sa berú do úvahy, sa môžu pohybovať od hmotných, ako sú aktíva, pasíva, zisky alebo iné fundamenty až po nehmotné, ako je selektívny predaj. (Bradford Lynch Levy, 2015)



Obrázok 11: Grafické znázornenie ukážky - Portgolio optimalization

1. Doterajšie výsledky
   1. Predikcia pozitívnych PCR testov

*(Spracoval: Filip Frank)*

Stĺpce datasetu (boli použité pre trénovanie predikčných modelov):

1. dátum (10.10.2020 – 8.12.2021)

2. denný počet pozitívnych PCR testov

3. denné percento pozitívnych PCR testov

4. denný počet pozitívnych Ag testov

5. denné percento pozitívnych Ag testov

Víťazný predikčný model: Arimax



Obrázok 12: Graf časového vývoja počtu pozitívnych PCR testov (modrým) so 14-dňovou predikciou (červeným).

Predikcia bola vykonaná na 14 nasledujúcich dní (9.12-22.12.2021), na pôvodných dátach bol aplikovaný algoritmus 7-dňového moving average pre vyhladenie vysvetliteľne nekonzistentných dát (napr. periodické významné poklesy denného počtu testov v dôsledku nízkej miery testovania cez víkendy).

Na grafe vidíme červenou predpoveď denného počtu pozitívnych PCR testov, reálny počet pozitívnych testov reprezentuje modrý graf. Medzi predpovedaným a reálne zisteným počtom PCR testov je výrazný rozdiel, pretože práve v čase, kedy začíname s predikciou (9.12.2021), nastal prudký zlom v počte zistených pozitívnych prípadov. Natrénovaný model nie je tak dokonalý, aby dokázal tento pokles predpovedať, namiesto toho pokračuje v predpovedi podľa trendu predchádzajúcich dní. Výrazný rozdiel pozorujeme už v prvom dni predikcie, kde model predpovedá počet PCR testov 8361.53, pričom reálna hodnota je 7337.57 (pozn.: hodnota po aplikácii 7-dňového moving average). V poslednom dni predikcie model predpovedá 8892.61, reálne však prírastok zodpovedá číslu 3618.14. Celkovú chybovosť modelu pri tejto predikcii vyjadruje chybová metrika RMSE (root-mean-square error), ktorá dosiahla hodnotu 4500.426. Táto hodnota je vzhľadom na predpovedané hodnoty vysoká a zodpovedá neschopnosti modelu predpovedať zmenu trendu vývoja počtu pozitívnych testov, ktorá nastala bezprostredne v deň začiatku predikcie.

* 1. Predikcia pozitívnych Ag testov

(*Spracoval: Juraj Puszter)*

Model neurónovej siete sme trénovali na základe:

* pozitívnych antigénových testov
* pozitívnych antigénových testov v percentách
* počte pozitívnych PCR testov
* počte pozitívnych PCR testov v percentách

Pri predpovedaní sme dáta pozitívnych antigénových testov nepoužili, nakoľko tie sme predpovedali.

Predpoveď a trénovanie modelu neurónovej siete sa uskutočnilo na dátach zozbieraných z datasetu poskytnutým Inštitútom Zdravotných Analýz, ktorý sa nachádza na githube. V datasete s názvom daily stats sa nachádzajú rôzne údaje, vrátane údajov o celkovom a tiež pozitívnom počte antigénových a PCR testov. Z týchto dát sme vyjadrili počet pozitívnych testov v percentách. Z datasetu sme využili dáta od dňa 11. Októbra 2020 do 30. Novembra 2021. Predikovali sme na 7 dní, od 1. do 7. Decembra 2021. Dáta boli pred predikciou upravené pomocou 7 dňového pohyblivého priemeru. Predikciu sme neskôr porovnali s reálnymi dátami a zobrazili na grafe.

Chart, line chart

Description automatically generated

Obrázok 13 Graf predikcie Ag z PCR, PCR%, Ag, Ag%

Modrá čiara na grafe znázorňuje počet pozitívnych antigénových testov. Červená čiara znázorňuje našu predikciu a zelená reálne hodnoty. Ako na grafe môžeme vidieť, naša predikcia sa od reálnych hodnôt značne líši. Predpovedaný počet antigénových testov má 7. Decembra hodnotu 2,490.341 a reálna hodnota je 1,817.71. Pri porovnaní predikovaných a reálnych hodnôt pomocou RMSE nám vyšlo 378.21833995587417. Vieme vyčítať, že reálne hodnoty klesli značne viac, ako naše predikované.

* 1. Predikcia úmrtí z viacerých ukazovateľov

*(Spracoval: Viet Quoc Le)*

Semestrálnou úlohou bolo spraviť predikciu počtu úmrtí na 14 dni, z denných štatistík na Slovensku, ktoré sú dostupné v repozitári IZA na githube v priečinku DailyStats. Dáta sa začali zbierať od 6.3.2020, nakoľko dáta boli na začiatku nekonzistentné až po 10.10.2020, kedy sa začalo plošné testovanie začali byť dáta konzistentné, som použil dáta od 10.10.2020. Použité príznaky na trénovanie boli denne pozitívne Ag, pozitívne Ag v percentách, pozitívne PCR, pozitívne PCR v percentách. Pripravený súbor s údajmi bol vložený do Azure na trénovanie a vyhodnotenie najlepšieho modelu. Najlepší model bol neskôr použitý na predikciu.

Graphical user interface, chart, line chart

Description automatically generated

Obrázok 14 Graf predikcie úmrtí z PCR, PCR%, Ag, Ag%

Predpovede modelu spočiatku sú výrazne lepšie ako neskoršie predpovede, po 7 dní sa vzdialenosť medzi predpoveďou a skutočnými údajmi zväčšuje. Predpovedaný počet úmrtí na konci, teda dňa 23.12.2021 je 16 618, skutočný počet úmrtí je 16 124.

* 1. Predikcia úmrtí z úmrtí

*(Spracoval: Tomáš Singhofer)*

Mojím zadaním bolo predikovať celkový počet úmrtí. Sieť sa učila pomocou zozbieraných dát z minulosti z githubu inštitútu zdravotných analýz a na základe toho predikovala vývoj situácie na ďalších 7 dní (9.12. – 15.12.). Sieť predikovala počet úmrtí na deň 15.12. niečo menej ako 15 500. V skutočnosti bol počet úmrtí v tento deň až 15 730. Pri mojom trénovaní mal najlepšie výsledky model Voting Ensemble.



Obrázok 15 Graf predikcie úmrtí z úmrt

* 1. Predikcia kompletne zaočkovaných

*(Spracoval: Marek Kačmár)*

Chart, line chart

Description automatically generatedFinálnou úlohou semestra bolo predikovať vývoj počtu pribúdajúcich plne zaočkovaných jedincov, ktorým bola podaná 2. dávka. Prvým krokom bolo pripraviť si dáta, ktoré budem predikovať. Z githubu Inštitútu Zdravotných Analýz som stiahol verejne dostupný csv súbor s dátami udávajúcimi o stave zaočkovanosti populácie Slovenska. Po úprave datasetu sa mi podarilo zoradiť dáta podľa jednotlivých týždňov a počtu druhých dávok, ktoré indikovali o plne zaočkovanom jedincovi. Na pripravené dáta som aplikoval moving average so 7-dňovým oknom, následne som z Azure použil model, ktorý najlepšie predikoval môj dataset, a ten som využil vo svojej vizualizácii predpovede vývoja počtu plne zaočkovaných na ďalších 7 týždňov.

Obrázok 16 Graf predikcie kompletne zaočkovaných

Na grafe môžeme pozorovať porovnanie predikovaného vývoja počtu plne očkovaných a skutočného vývoja počtu plne zaočkovaných jedincov v jednotlivých týždňoch. Predikcia bola vytvorená ešte v polovici decembra a samotné porovnanie bolo vykonané 22. januára. Ako môžeme vidieť na grafe, počet pribúdajúcich plne zaočkovaných ľudí v jednotlivých týždňoch klesal postupne k nule, od čoho sa odvíjala aj jeho predpoveď, avšak vplyvom rôznych činiteľov, ktoré v predpovedi neboli zahrnuté, ako napr. zvýšenie počtu nakazených a hospitalizovaných (strach z nakazenia/úmrtia), politická podpora zaočkovanosti a i. tento počet zvýšili, čo je možné vidieť na skutočnom počte plne zaočkovaných v prislúchajúcich týždňoch.

* 1. Graf celkových hospitalizácií

*(Spracoval: Branislav Šipula)*

Graf znázorňuje stav/vývoj počtu hospitalizovaných pacientov v nemocniciach za obdobie od 30.04.2020 do 9.12.2021. Z grafu je vidieť, že najvyšší nárast hospitalizovaných pacientov bol v mesiaci marec a druhy najvyšší nárast bol v mesiaci december 2021. Najnižší pokles bol medzi týmito extrémami bolo v letných mesiacoch a z toho najmenší v mesiaci august 2021.



Obrázok 17 Graf celkových hospitalizácií

* 1. Graf celkových UPV

*(Spracoval: Adam Vozár)*

Záverečná úloha pozostávala z vytvorenia grafu obsadenosti miest na umelej pľúcnej ventilácii. K vytvoreniu grafu bolo potrebné použiť dáta z Inštitútu Zdravotných Analýz, konkrétne súbor Hospital\_UPV\_AdmissionDischarge.csv, ktorý v prvom stĺpci obsahuje dátum a v druhom stĺpci číslo, ktoré predstavuje o koľko sa počet miest na UPV zvýšil resp. znížil. Dataset som doplnil o ďalší stĺpec suma, v ktorom sa počet obsadených miest na UPV kumulatívne sčituje ku každému dátumu. Z grafu môžeme vidieť viditeľný nárast UPV od októbra 2020 ktorý dosiahol svoje maximum dňa 20. marca 2021 s počtom 422 UPV.

Obrázok 18 Graf časového vývoja UPV

* 1. Tabuľka dosiahnutých výsledkov predikcií

*(Spracoval: Juraj Puszter*)

V tabuľke sa nachádza súhrn dosiahnutých výsledkov predikcií z kapitoly 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predikcia** | **Počet dní** | **Hodnota predikcie** | **Reálna hodnota** | **RMSE** |
| Pozitívne PCR testy | 14 | 8892.61 | 3618.14 | 4500.426 |
| Pozitívne Ag testy | 7 | 2,490.341 | 1,817.71 | 378.218 |
| Úmrtia z viacerých ukazovateľov | 14 | 16 618 | 16 124 | 40.676 |
| Úmrtia z úmrtí | 7 | 15 500 | 15 730 | 260.38 |
| Kompletne zaočkovaní | 49 | 2 327 | 33 493 | 24441.36 |

1. Predikcie hospitalizácií
   1. Predikcie z dát od konca delty a začiatku omikronu

*(Spracoval: Juraj Puszter, Adam Vozár, Branislav Šipula, Marek Kačmár)*

V tejto kapitole sa nachádzajú výsledky predikcií hospitalizácií z dát od konca delty a začiatku omikronu. Modely sme trénovali na dátach od 14.1.2022 do 10.3.2022. Predikovali sme na 10 dní, od 11.3.2022 do 20.3.2022. Dáta sme pred predikciou upravili pomocou 7 dňového pohyblivého priemeru. Predikcie sme porovnali s reálnymi dátami a zobrazili na grafoch.

Trénovanie a predikciu sme vykonávali na základe štyroch rôznych kombinácií parametrov a to:

1. Ag pozitívne, Ag pozitívne v %, PCR pozitívne, PCR pozitívne v %
2. PCR pozitívne v %, Ag pozitívne v %
3. PCR pozitívne, Ag pozitívne
4. Kumulatívne testy, Kumulatívne testy v %

Dáta o Ag a PCR pozitívnych testoch sme získali priamo z datasetu, zatiaľ čo dáta o Ag a PCR pozitívnych testoch v % sme vyjadrili pomocou celkových počtoch Ag a PCR testov. Kumulatívne testy sme vypočítali pomocou vzorca kumulatívne testy = Ag pozitívne testy \* 2.68 + PCR pozitívne testy. Potom sme vypočítali pomocný výpočet pre celkový počet testov pomocou vzorca celkový počet = Ag celkových testov \* 2.68 + PCR celkových testov. Tento pomocný výpočet sme využili a vypočítali kumulatívne testy v % vzorcom kumulatívne testy v % = kumulatívne testy / celkový počet.

Pre všetky 4 trénovania nám vyšiel najlepší model VotingEnsemble.

Na nasledujúcich 4-och obrázkoch sa nachádzajú grafy predikcií hospitalizácií vyššie uvedenými parametrami. Modrá čiara znázorňuje počet hospitalizácií. Červená čiara znázorňuje našu predikciu a zelená reálne hodnoty.



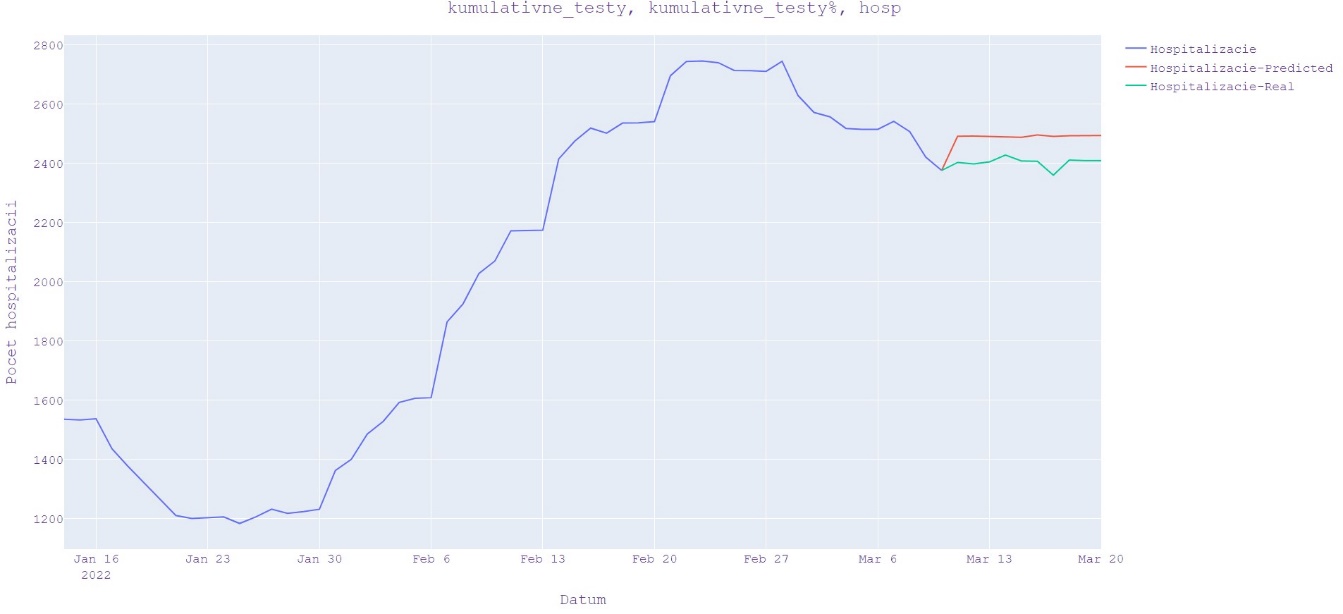
Obrázok 19 Graf predikcie hospitalizácii z parametrov Ag, Ag%, PCR, PCR%



Obrázok 20 Graf predikcie hospitalizácii z parametrov Ag%, PCR%



Obrázok 21 Graf predikcie hospitalizácií z parametrov Ag, PCR



Obrázok 22 Graf predikcie hospitalizácií z parametrov kumulatívne testy, kumulatívne testy%

Na grafoch vidíme, že naše predikcie sú vyššie ako reálne hodnoty. Rozdiel medzi nimi je približne 80. Nasledujúca tabuľka obsahuje súhrn informácií o našich predikciách.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parametre** | **Hodnota predikcie** | **Reálna hodnota** | **RMSE** |
| Ag, Ag%, PCR, PCR% | 2487.26 | 2409 | 87.793 |
| Ag%, PCR% | 2484.44 | 87.364 |
| Ag, PCR | 2488.24 | 86.863 |
| Kumulatívne, kumulatívne% | 2494.08 | 89.773 |

**Zoznam použitej literatúry**

**—.** *Decision Trees.* [Online] [Dátum: 13. 3 2022.] https://scikit-learn.org/stable/modules/tree.html.

**Bigelow, Stephen J. 2020.** TechTarget. [Online] 4 2020. [Dátum: 5. 3 2022.] https://www.techtarget.com/searchcloudcomputing/definition/Windows-Azure.

**Brownlee, Jason. 2020.** How to Develop Voting Ensembles With Python. [Online] 2020. [Dátum: 14. 12 2021.] https://machinelearningmastery.com/voting-ensembles-with-python/.

**—. 2018.** Stacking Ensemble for Deep Learning Neural Networks in Python. [Online] 2018. [Dátum: 10. 1 2022.] https://machinelearningmastery.com/stacking-ensemble-for-deep-learning-neural-networks/.

**Ciotti, Ciccozzi, Terrinoni, Jiang, Wang, Bernardini. 2020.** *The COVID-19 pandemic.* [Online] Taylor & Francis, 2020. [Dátum: 6. 3 2022.] https://doi.org/10.1080/10408363.2020.1783198.

**Date, Sachin.** Time Series Analysis, Regression and Forecasting. [Online] [Dátum: 5. 3 2022.] https://timeseriesreasoning.com/contents/regression-with-arima-errors-model/.

**Donges, Niklas. 2021.** builtin. [Online] 16. 9 2021. [Dátum: 3. 1 2022.] https://builtin.com/data-science/random-forest-algorithm.

**Hyndman, Athanasopoulos. 2018.** Forecasting: principles and practice. [Online] 2018. [Dátum: 6. 3 2022.] https://otexts.com/fpp2/simple-methods.html.

**Kravchuk, Kostiantyn. 2017.** r-bloggers. [Online] 5. 5 2017. [Dátum: 5. 3 2022.] https://www.r-bloggers.com/2017/05/forecasting-arimax-model-exercises-part-5/.

**Mccoy, Logan.** ccb technology. [Online] [Dátum: 5. 3 2022.] https://ccbtechnology.com/what-microsoft-azure-is-and-why-it-matters/.

**Microsoft. 2021.** Microsoft. [Online] 15. 12 2021. [Dátum: 5. 3 2022.] https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-machine-learning.

**—. 2022.** Microsoft. [Online] 4. 2 2022. [Dátum: 5. 3 2022.] https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-machine-learning-studio.

**—. 2022.** Microsoft. [Online] 26. 2 2022. [Dátum: 5. 3 2022.] https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml#when-to-use-automl-classification-regression-forecasting--computer-vision.

**Moreno, Caio. 2020.** Medium. [Online] 8. 5 2020. [Dátum: 5. 3 2022.] https://caiomsouza.medium.com/aml-azure-machine-learning-introduction-13505d3393b9.

**Patro, Tomáš. 2017.** BudFIT. [Online] 15. 10 2017. [Dátum: 3. 1 2022.] https://casopis.fit.cvut.cz/technologie/umela-inteligencia-ako-funguje-preco-dobre-sa-nu-zaujimat/.

**Prabhakaran, Selva. 2021.** machine learning plus. [Online] 22. 8 2021. [Dátum: 5. 3 2022.] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/.

**PyFlux. 2016.** PyFlux. [Online] 2016. [Dátum: 5. 3 2022.] https://pyflux.readthedocs.io/en/latest/arima.html.

**—. 2016.** PyFlux. [Online] 2016. [Dátum: 5. 3 2022.] https://pyflux.readthedocs.io/en/latest/arimax.html.

**Río, Rafael del. 2018.** A simple example of visualizing gradient boosting. [Online] 2018. [Dátum: 14. 12 2021.] https://www.researchgate.net/figure/A-simple-example-of-visualizing-gradient-boosting\_fig5\_326379229.

**T. Baldigara, M. Mamula. 2015.** Modelling International Tourism Demand Using Seasonal Arima Models. [Online] 2015. [Dátum: 6. 3 2022.] https://hrcak.srce.hr/file/206819. 21.

**Tianqi Chen, Carlos Guestrin. 2016.** XGBoost: A Scalable Tree Boosting System. [Online] 2016. [Dátum: 14. 12 2021.] https://arxiv.org/abs/1603.02754.

**Jason Brownle. 2020.** How to Develop Elastic Net Regression Models in Python. [Online] 2020-2022 [Dátum: 13. 3. 2022] https://machinelearningmastery.com/elastic-net-regression-in-python/

**2022 Algoritmus SVM.** [Dátum: 13. 3. 2022] https://sk.education-wiki.com/4245111-svm-algorithm

**CJ Carey, Yuan Tang, William de Vazelhes, Aurélien Bellet and Nathalie Vauquier.** **2015-2020**. metric-learn [Online] 2015 [Dátum: 13. 3. 2022] http://contrib.scikit-learn.org/metric-learn/introduction.html

**Wikipedia.** **2021.** Portfolio optimization[Online] [Dátum: 13. 3. 2022] https://en.wikipedia.org/wiki/Portfolio\_optimization

**Bradford (Lynch) Levy. 2015.** Investment Portfolio Optimization. [Online] 2013 - 2022. [Dátum: 13. 3. 2022] http://www.bradfordlynch.com/blog/2015/12/04/InvestmentPortfolioOptimization.html