Predikcia teploty ovzdušia

Bc. Tomáš Bujna, Bc. Michal Zeliska

Fakulta informatiky a informačných technológií STU v Bratislave
Ilkovičova 2, 842 16 Bratislava 4

{tomasb199, zeliskamichal53}@gmail.com

Abstrakt. Počasie je na rozdiel od podnebia okamžitým stavom atmosféry na určitom mieste a v určitom čase, zatiaľ čo podnebie popisuje dlhodobý stav počasia na určitom mieste. Tento stav sa opisuje súborom hodnôt meteorologických prvkov, akými sú napríklad teplota, vlhkosť, tlak, úhrn zrážok, oblačnosť, atď. Poveternostné podmienky sa menia nepretržite v závislosti od času, vytvárajú sa spojité rady každého meteorologického údaju a môžu sa použiť na vypracovanie prognostického modelu. Táto práca má za úlohu analyzovať vzorku historicky nameraných meteorologických dát. Následne má za úlohu identifikáciu najrelevantnejších atribútov a ich predspracovanie pre potrebu vytvorenia finálnej dátovej množiny s cieľom predikovania teploty. Na koniec na základe tejto dátovej množiny bude pomocou algoritmov strojového učenia vytvorený model/y na predikciu teploty ovzdušia na vybranom mieste.

Kľúčové slová: počasie, analýza meteorologických dát, regresia, predikcia teploty, lineárna regresia, náhodné lesy, rekurentné neurónové siete

1 Opis problému a motivácia

Presné predpovede počasia v dnešnom svete zohrávajú dôležitú úlohu, keďže rôzne priemyselné a poľnohospodárske sektory sú principiálne závislé od poveternostných podmienok t.j. počasia. Taktiež je možné tieto predpovede využiť na prípadné varovania pred rôznymi prírodnými katastrofami. Predpoveď počasia je možné definovať ako stanovenie správnych hodnôt poveternostných podmienok (viď. Tab. *2*) pričom na základe týchto parametrov je možné ďalej určiť budúce poveternostné podmienky [1]. Aktuálne existuje veľké množstvo nástrojov, ktoré dokážu predpovedať počasie, ale množstvo dát generovaných na predpovedanie počasia je veľké a zväčša neštruktúrované. Z tohto dôvodu sa nejedná o ľahkú úlohu. Táto úloha vyžaduje veľké množstvo parametrov, ktoré sa však môžu rýchlo meniť na základe atmosférických podmienok. Globálna predpoveď počasia predstavuje značne náročný problém, pričom na takúto predikciu počasia je nevyhnutné použitie najnovších technológií s veľkým výpočtovým výkonom [2]. Vzhľadom k veľkej rozsiahlosti problému sme sa rozhodli zúžiť problém na malú oblasť iba s vybranými parametrami počasia.

2 Opis dát

Zvolený dataset [3] obsahuje namerané meteorologické dáta z Írska od začiatku roku 1989 až do konca roku 2017. Merania boli uskutočňované z 25 rôznych meteorologických staníc každú hodinu, pričom tieto stanice sa nachádzajú rôzne rozmiestnené po celej oblasti Írska (Obrázok 1). Keďže tak ako bolo uvedené v úvode dokumentu, počasie je okamžitý stav atmosféry na určitom mieste, rozhodli sme náš dataset rozdeliť do skupín podľa jednotlivých meracích staníc, teda na konkrétne miesta.

Obr. 1 - Vizualizácia polohy meteorologických staníc na mape
(názov stanice: celková priemerná teplota[°C], celkový úhrn zrážok[mm])

Najskôr sme analyzovali počet záznamov v datasete. Nakoľko dataset obsahuje záznamy pre 25 rôznych staníc, vyhodnocovali sme chýbajúce záznamy meraní pre jednotlivé stanice:

Názov stanice	Cork_Airport	Shannon_Airport	Casement	Dublin_Airport	Mullingar	Belmullet	Malin_head	Valentia_Observatory	Claremorris	Roches_Point	Knock_Airport	Finner	Moore_Park	Oak_Park	...	Dunsany	Markree	Gurteen	Mt_Dillon	Athenry	Súčet
Chybné merania	0	0	0	0	0	0	0	1	37	19321	66439	81966	127801	127801	...	166490	166489	167233	167185	197065	**1287828**
Úplnosť dát	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	99,90%	99,90%	92,30%	73,80%	67,70%	49,70%	49,70%	...	34,50%	34,50%	34,20%	34,20%	22,40%	**66,37%**

Tab. 1 - Chýbajúce záznamy pre jednotlivé stanice

V datasete sa nachádza 19 atribútov: 15 numerických (13 – spojitých a 2 - diskrétne) a 4 nominálne. Zoznam a opis jednotlivých atribútov je uvedený v *Tabuľke 2.*

Skratka	Vysvetlenie	Typ atribútu
station	Meno meteorologickej stanice	nominálny
contry	Krajinná oblasť	nominálny
longitude	Zemepisná dĺžka	numerický-diskrétny
latitude	Zemepisná šírka	numerický-diskrétny
rain	Úhrn zrážok (mm)	numerický-spojitý
temp	Teplota vzduchu (°C)	numerický-spojitý
wetb	Teplota mokrého teplomeru (°C)	numerický-spojitý
dewpt	Teplota rosného bodu (°C)	numerický-spojitý
vappr	Tenzia pary (hPa)	numerický-spojitý
rhum	Relatívna vlhkosť (%)	numerický-spojitý
msl	Priemerný atmosferický tlak na hladine mora (hPa)	numerický-spojitý
wdsp	Priemerná rýchlosť vetra za hodinu (kt)	numerický-spojitý
wddir	Prevažujúci hodinový smer vetra (degrees)	numerický-spojitý
ww	Slovná klasifikácia počasia (sneženie, dážď, slnečno) [0-99]	nominálny
w	Slovná klasifikácia počasia za predchádzajúcu hodinu (sneženie, dážď, slnečno) [0-9]	nominálny
sun	Dĺžka slnečného svitu (hours)	numerický-spojitý
vis	Viditeľnosť (m)	numerický-spojitý
clht	Výška oblačnosti (feet)	numerický-spojitý
clamt	Oblačnosť (okta)	numerický-spojitý

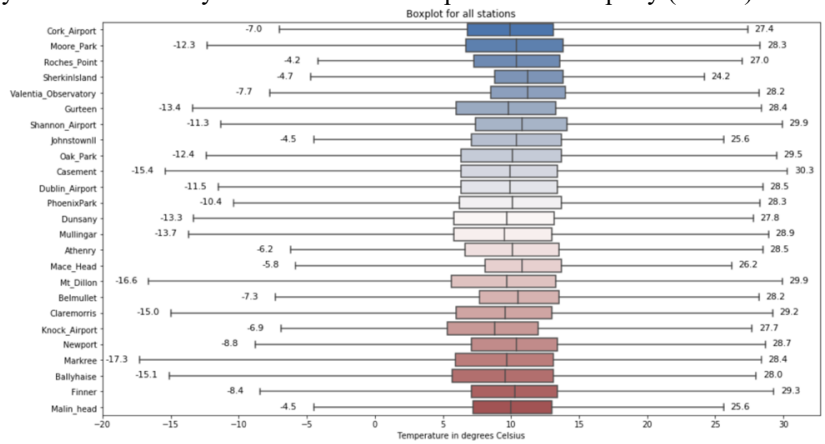
Tab. 2 - Vysvetlenie atribútov v datasete

Následne sme taktiež analyzovali jednotlivé atribúty v záznamoch z pohľadu meracích staníc. Pre niektoré stanice chýbajú kompletné záznamy z meraní v určitom časovom okne alebo taktiež chýbajú niektoré atribúty z meraní. Neúplnosť dát niektorých staníc je spôsobená tým, že boli väčšinou spustené do prevádzky v neskoršom období alebo nevykonávali merania všetkých atribútov, napríklad z dôvodu chýbajúceho technického vybavenia.

Názov stanice	station	rain	temp	vappr	rhum	msl	wdsp	wddir	ww	sun	vis	clht	clamt
Cork_Airport	254208	254208	254208	254206	254206	254208	254208	254208	254208	254208	254192	254208	254208
Shannon_Airport	254208	254208	254208	254208	254208	254208	254207	254207	254208	254208	254208	254208	254208
Casement	254208	254208	254208	254208	254208	254208	254208	254208	254208	254208	254200	254208	254208
Dublin_Airport	254208	254208	254208	254207	254207	254208	254208	254207	254208	254208	254208	254208	254208
Mullingar	254208	254150	254186	252159	252161	233403	254125	253886	0	0	0	0	0

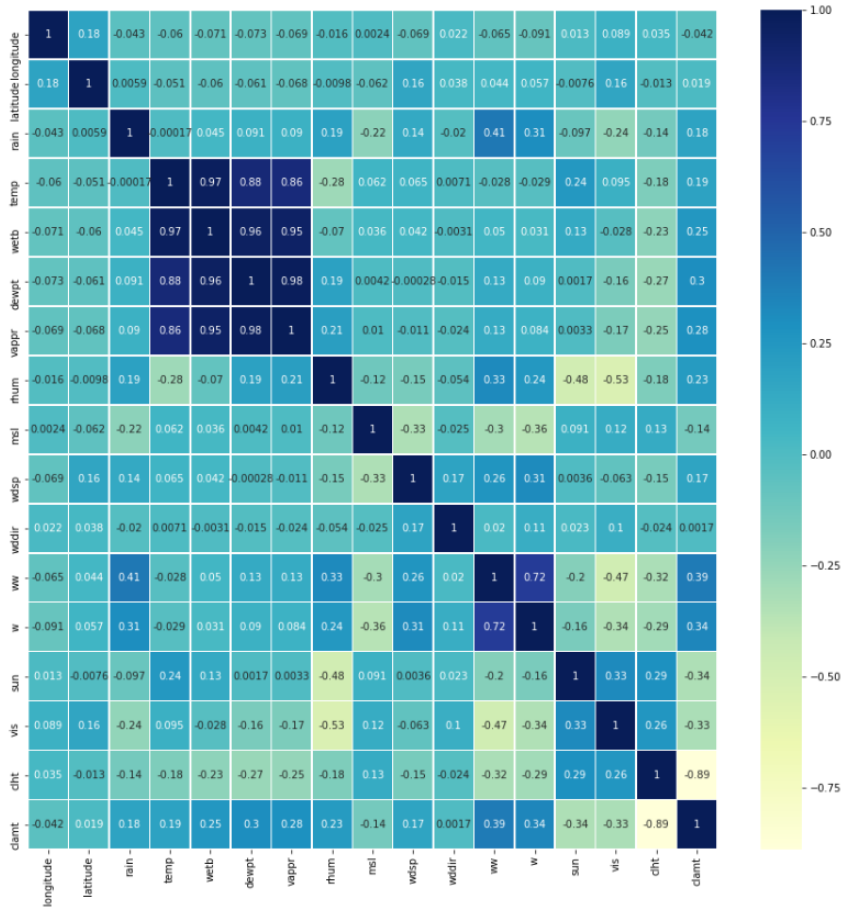
Tab. 3 - Ukážka úplnosti dát niektorých staníc a atribútov(max počet záznamov 254208)

Keďže našim hlavným cieľom je predikcia teploty, rozhodli sme sa taktiež bližšie analyzovať aké hodnoty nadobúda v datasete práve atribút teploty (Obr. *2*).



Obr. 2 - Boxplot atribútu teploty pre jednotlivé stanice

Taktiež sme sa rozhodli v rámci procesu analýzy a spracovania dát overiť vzájomné súvislosti, vzťahy a prepojenia jednotlivých atribútov datasetu, aby bolo možné dopredu identifikovať dôležité atribúty, ktoré môžu veľmi vplývať na výsledok predikcie (Obrázok 3).



Obr. 3 - Korelačná matica

3 Definovanie úlohy objavovania znalosti

Cieľom nášho projektu je predikovanie atribútu teploty na základe historických dát a taktiež predikcia teploty v susednom meste na základe aktuálnych dát. Úloha ktorú riešime, teda predikcia teploty na základe iných atribútov, spadá pod regresiu.
V prvom prípade budeme predikciu teploty robiť vždy pre jednu konkrétnu meraciu stanicu. Predikcia bude uskutočňovaná na základe meraní niekoľkých predošlých dní a predikovať budeme určitý časový úsek do budúcnosti napríklad 24 hodín.

V druhom prípade budeme predikciu robiť vždy na základe vybraných atribútov počasia pre jednu meraciu stanicu napr. *Dublin Airport* a na základe aktuálnych parametrov počasia v tejto meracej stanici budeme predpovedať atribút teploty napr. pre susednú stanicu *Casament*.

3.1 Predpokladaný scenár riešenia (problémy)

Predpokladaným scenárom pri predspracovaní dát je výber meracej stanice, pre ktorú bude predikovaná teplota. V tomto prípade sa potykáme s problémami akými sú veľká dimenzionalita dát, chýbajúce merania alebo hodnoty niektorých atribútov. Dimenzionalitu dát je možné vyriešiť extrakciu vybraných čŕt (z angl. *fe*ature *selection*), chýbajúce merania vieme eliminovať výberom vhodnej meracej stanice bez chýbajúcich meraní a prípadné chýbajúce hodnoty atribútov vieme nahradiť napr. interpoláciou alebo opakovaním poslednej známej hodnoty. Po vytvorení trénovanej a testovacej množiny je potreba dáta vhodne normalizovať, nakoľko niektoré atribúty dosahujú hodnoty v tisíckach a niektoré sú v stotinách. Ďalším problémom môže byť výber vhodných vyhodnocovacích funkcií potrebných na ohodnotenie presnosti modelov.

4 Opis prác iných autorov

V tejto kapitole uvádzame existujúce práce, ktoré sa zaoberajú problematikou predikovania počasia. Pri ich popise sme sa zamerali hlavne na použité DM metódy a taktiež na dosiahnuté výsledky.

Názov článku: Application of Data Mining Techniques in Weather Prediction and Climate Change Studies [4] (2012)

Použité metódy: Neurónové siete a rozhodovacie stromy.

Výsledky práce: Autori sa vo svojej práci zamerali na predpovedanie maximálnej teploty, zrážok, odparovania a rýchlosti vetra. Použili meteorologické údaje zhromaždené medzi 2000 a 2009 z mesta Ibadan, Nigéria. Na vyhodnocovanie ich výstupov použili korelačný koeficient (z angl. *Correlation Coefficient, CC*), druhú mocninu strednej chyby (z angl. *Mean Squared Error,* MSE) a druhú odmocninu strednej chyby (z angl. *Root Mean Square Error*, RMSE). Vzhľadom na ich typ úlohy sa im podarilo pre testovacie dáta dosiahnuť výsledky MSE - 0.2028 a RMSE - 0.3290. Vo výsledku práce dospeli k tomu, že pre lepší výsledok je potrebný väčší súbor údajov a to najlepšie z priebehu mnohých desaťročí.

Názov článku: Implementation of Data Mining Techniques for Meteorological Data Analysis [5] (2011)

Použité metódy: Neurónové siete a lineárna regresia.

Výsledky práce: Autori vo svojom článku predpovedali dennú priemernú teplotu na základe atribútov: deň, mesiac, priemerné teploty 3 predchádzajúcich dní, vlhkosť a rýchlosť vetra. Na trénovanie použili 70% dát ako spojitý rad a zvyšok bol použitý na testovanie. Použitá NN je obyčajná dopredná neurónová sieť so 6 skrytými vrstvami. V prípade použitej lineárnej regresie sa jedná o *Least Median Squares Linear Regression*. Na vyhodnocovanie ich výstupov použili CC a RMSE. V tomto prípade sa im podarilo dosiahnuť výsledky RMSE – 1,691 (LR) a 1,726 (NN) a pre CC – 0,924 (LR) a 0,933 (NN). Zvyšok ich práce sa venoval klasifikácií teploty do určitých intervalov na základe teplotných rozmedzí. Ich dataset obsahoval 9 ročný denný interval priemerných denných pozorovaní.

Názov článku: Efficient Weather Forecasting using Artificial Neural Network as Function Approximator. [6] (2014)

Použité metódy: Neurónové siete.

Výsledky práce: Autori na predpoveď využili neurónové siete typu Radial basic function network (RBFN). Predpovedali aktuálnu teplotu na základe aktuálnych atribútov ako: tlak, vlhkosť, teplota rosného bodu, rýchlosť vetra a smer vetra. Na trénovanie použili 1 rok dát a na testovanie použili ďalší rok dát. Záznamy boli merané každé 3 hodiny. Ako vyhodnocovaciu metriku použili iba MSE. Autori experimentovali s rôznym počtom neurónov v skrytej vrstve (2-60), pričom najlepšie výsledky boli dosiahnuté pri 49 neurónoch – 0,174462 MSE.

5 Použité DM metódy

Na realizáciu našej úlohy sme sa rozhodli využiť nasledovné DM metódy:

- **Lineárna regresia** (z angl. *Linear regression, LR*) - Lineárna regresia je sada techník pre odhad vzťahov medzi dvoma alebo viacerými premennými. Lineárna regresia teda hľadá riešenie na otázku, ako premenné navzájom súvisia. Výsledkom lineárnej regresie je regresná rovnica, ktorá sa môže použiť na predpovede údajov. Vzťah jednej vysvetľujúcej premennej sa nazýva jednoduchá lineárna regresia a vzťah pre viac ako jednu vysvetľujúcu premennú sa nazýva viacnásobná lineárna regresia.

- **Náhodné lesy** (z angl. Random Forests, RF) - Náhodné lesy sú jedným z najúčinnejších modelov učenia s učiteľom používaných na klasifikáciu a regresiu. Celok učiacich modelov združuje viac modelov strojového učenia, čo umožňuje celkovo lepší výkon. Logika použitia RF je taká, že každý z použitých modelov je slabý, ak je použitý samostatne, ale je silný, ak je zostavený do väčšieho celku. V prípade náhodných lesov sa používa veľké množstvo rozhodovacích stromov, ktoré pôsobia ako „slabé“ faktory a ich výstupy sa agregujú, pričom výsledok predstavuje „silný“ celok. Náhodné lesy sa v porovnaní s rozhodovacími stromami vyznačujú

nižšou variáciou. Variácia je miera v akej by sa naša predikcia zmenila, ak by sme ju trénovali na iných dátach. Vysoká variácia zvyčajne znamená, že sa prispôsobujeme našim trénovací dátam, teda nachádzame vzory, ktoré sú viac náhodné a nie sú schopné zovšeobecniť nové údaje.

- **Neurónové siete** (z angl. *Neural Network*, NN) - Umelé neurónové siete sú založené na množstve prepojených jednotiek alebo uzlov nazvaných umelé neuróny, ktoré napodobňujú neuróny v biologickom mozgu. Každé spojenie, rovnako ako synapsie v biologickom mozgu, môžu prenášať signál do iných neurónov. Umelý neurón, ktorý prijíma signál ho potom spracuje a môže signalizovať neuróny, ktoré sú s ním spojené. Existuje množstvo architektúr umelých neurónových sietí, pričom každá vyniká v rôznych oblastiach pri riešení rôznych problémov. Z nášho pohľadu predpovede počasia sú zaujímavé hlavne rekurentné neurónové siete (RNN), kde spojenia medzi neurónmi tvoria orientovaný graf pozdĺž dočasnej postupnosti, čo im umožňuje vykazovať dočasné dynamické správanie. RNN používajú svoj vnútorný stav (pamäť) na spracovanie sekvencií vstupov s premenlivou dĺžkou, vďaka čomu sú silným nástrojom pri spracovávaní dát ako sú rôzne časové rady, zvuk, reč, text, video a pod. RNN teda majú v porovnaní s inými technikami oveľa lepšie "chápanie" danej postupnosti a jej súvislostiach. Pri každom rozhodnutí RNN berie do úvahy okrem aktuálnych vstupných informácií aj informácie, ktoré už sieťou prešli predtým. Rozšírením RNN sú Long Short Term Memory (LSTM) siete, ktoré riešia problém strácajúceho sa gradientu, čím umožňujú sieti pamätať si vstupy po dlhšiu dobu a spracovávať dlhšie sekvencie.

5.1 Predspracovanie a výber atribútov

Na výber atribútov sme použili *SelectKBest* algoritmus, ktorý na základe zvolenej vyhodnocovacej funkcie vyberie *k* najlepších atribútov. V našom prípade sme ako vyhodnocovaciu funkciu použili *mutual_info_regression*, ktorá odhaduje vzájomnú informáciu pre spojitý cieľový atribút, čiže teplotu a taktiež funkciu *f_regression*, ktorá berie do úvahy koreláciu medzi každým regresorom a cieľovým atribútom (teplotou). Výsledky aplikovaného *SelectKBest* algoritmu je možné vidieť nižšie v Tab. *4*.

Poradie	Atribút
1	temp
2	day
3	hour
4	vappr
5	rhum

Tab. 4 - Výsledok SelectKBest

Vzhľadom k tomu, že väčšina atribútov nadobúdala s atribútom teploty nízku koreláciu, rozhodli sme sa pridať ešte ďalšie atribúty a to kĺzavý priemer (z angl. *moving average,* MA) teplôt, pričom sme skúsili vytvoriť MA s rôznou dĺžkou – 6, 12, 24, 48 a 720 hodín. Čím bola dĺžka menšia, tým väčšiu koreláciu daný atribút nadobúdal.
Predspracovanie dát je ďalej možné rozdeliť do dvoch kategórií na základe typu úlohy, ktorý riešime. V prvom prípade, ak predpovedáme budúcu teplotu pre konkrétnu

stanicu, sme najskôr dáta rozdelili do viacerých dataframov, pričom každý dataframe predstavuje záznamy pre jednu meraciu stanicu. Následne sme zo všetkých staníc vybrali meraciu stanicu, ktorá mala najmenší počet chýbajúcich záznamov a atribútov (*Dublin Airport*). Ďalej sme pre túto meraciu stanicu vytvorili trénovaciu a testovaciu množinu v pomere 80:20. Vzhľadom k tomu, že jednotlivé záznamy meraní sú uskutočňované postupne za sebou v čase, je nevyhnutné dbať na tento fakt aj pri ich rozdelení do týchto množín. To znamená, že dáta do týchto množín nemôžu byť rozdelené náhodne, ale je potrebné, aby dáta boli usporiadané v časovej následnosti. Následne sme dáta normalizovali pomocou *MinMax Scalera* a pre takto predpripravené dáta sme vytvorili časové okná na predikciu. Vytvorenie časových okien prebieha dynamicky na základe dvoch parametrov a to počet predchádzajúcich hodín a počet predpovedaných hodín. Tvar dát je teda [x,y,z], kde x je počet časových okien, y počet riadkov (záznamov) v jednom časovom okne a z je počet atribútov časového okna.

Príklad sekvencie teplôt: [1, 2, 3, 4, 5, 6, 7, 8]
Dĺžka predchádzajúcej sekvencie: 4
Dĺžka predikovanej sekvencie: 2

Predchádzajúce hodnoty	Predikované hodnoty
[1, 2, 3, 4]	[5, 6]
[2, 3, 4, 5]	[6, 7]
[3, 4, 5, 6]	[7, 8]

Tab. 5 - Ukážka vytvárania časových okien (iba s teplotou)

V druhom prípade ak predpovedáme aktuálnu teplotu v susednom meste si taktiež dáta rozdelíme do viacerých dataframov na základe meracích staníc. Následne vyberieme 2 stanice. Jednu nazveme hlavná (napr. *Dublin Airport*), teda tá pre ktorú poznáme reálne parametre počasia a druhá bude susedná (napr. *Casament*), teda tá pre ktorú sa budeme snažiť predpovedať aktuálnu teplotu/zrážky. Následne vytvoríme nový dataset, ktorý bude obsahovať všetky atribúty hlavnej stanice a atribút teploty susednej stanice. Ďalej tento dataset rozdelíme na trénovaciu a testovaciu množinu v pomere 80:20. V tomto prípade však nie je potrebné brať do úvahy ich časovú následnosť. Na takto predpripravené dáta následne aplikujeme LR, RF a NN.

5.2 Spôsob vyhodnocovania

Ako spôsoby vyhodnocovania sme si zvolili 3 najpoužívanejšie metriky štandardne používané pri predpovediach a analýzach časových radov na vyhodnotenie presnosti spojitých atribútov, a to:
MSE (z angl. *mean squared error*) – vyjadruje presnosť odhadov pomocou strednej hodnoty druhých mocnín rozdielov medzi každou predikovanou hodnotou a jej zodpovedajúcou správnou hodnotou.
RMSE (z angl. *root mean squared error*) – druhá odmocnina z MSE.

MAE (z angl. *mean absolute error*) – vyjadruje priemerný počet chýb v predpovediach. Je to vlastne priemer absolútnych rozdielov medzi predikovanými hodnotami a skutočnými hodnotami, kde všetky individuálne rozdiely majú rovnakú váhu.
Pri všetkých týchto metrikách platí, čím menšia hodnota bližšie k nule, tým lepší výsledok.

5.3 Experimentovanie

V prvom prípade predpovedáme budúcu teplotu na základe hodnôt predchádzajúcich teplôt v jednom meste. Experimentami budeme sledovať presnosť modelov pri zvyšujúcom sa počte predchádzajúcich meraní, ktoré budú mať modely k dispozícií a taktiež počtu hodín, ktoré budú musieť modely predpovedať. V prípade neurónovej siete môžeme skúsiť, ako bude viac atribútov ovplyvňovať výsledok predikcie. Taktiež V druhom prípade ak predpovedáme aktuálnu teplotu v susednom meste sme si tento scenár vybrali práve kvôli experimentom, o ktoré v takomto prípade nebude núdza. Hlavný experiment, ktorý týmto scenárom sledujeme, je porovnanie ako vplýva vzdialenosť hlavnej stanice od susedenej na výsledok predikcie. Ďalej je taktiež v tomto prípade možné experimentovať s parametrami poveternostných podmienok, na základe ktorých budeme robiť predikciu. Taktiež je možné experimentovať aj s tým, ktorý parameter (teda nemusí to byť výhradne teplota) je najlepšie predpovedať z hľadiska presnosti a ktoré parametre poveternostných podmienok je v takom prípade vhodné použiť.

6 Experimenty

V tejto kapitole uvádzame popis našich experimentov, ktoré sme definovali už vyššie v kapitole *5.3 Experimentovanie*. Vzhľadom k tomu, že naše experimentovanie vychádza z dvoch samostatných prípadov, budeme sa každému prípadu venovať v samostatnej podkapitole.

6.1 Predikcia budúcej teploty v jednej meracej stanici

Na predikciu budúcej teploty sme využili 3 rôzne typy DM metód: lineárnu regresiu (LR), náhodné lesy (RF) a neurónové siete (NN), aby bolo možné vzájomne vyhodnotiť ich úspešnosť. Ako vyhodnocovacie metriky sme zvolili už vyššie spomínané RMSE, MSE a MAE. V našom experimente sme taktiež vyhodnocovali ako na predikciu vplýva dĺžka sekvencie dát, na základe ktorej je vytváraná predikcia a taktiež sme vyhodnocovali ako vplýva na výsledok predikcie dĺžka časového obdobia, na ktoré predikciu vytvárame. V prípade LR a RF sme využili v podstate štandardnú implementáciu bez výraznejšieho ladenia a táto predikcia je vytváraná len na základe atribútu teploty. V prípade NN sme sa rozhodli teplotu predikovať aj na základe viacerých atribútov, nakoľko prvotné experimenty ukázali, že bez využitia viacerých atribútov predikcia pomocou NN dosahovala zväčša najhoršie výsledky z použitých metód, okrem predikcie na dlhšie časové obdobie (v našom prípade 12h), kde NN

dosahovala aj bez využitia ďalších atribútov približne o 1°C lepšie výsledky. Vzhľadom k tomuto potenciálu sme sa rozhodli rozšíriť atribúty pre NN. Na výber atribútov pre NN sme využili výsledky zo *SelectKBest* algoritmu a taktiež sme pridali aj kĺzavé priemery. Na základe našich testov sa ukázalo najvhodnejšie použitie práve MA_12 a MA_24, ktoré sme následne aj ďalej používali v rámci našej predikcie. Nižšie v Tab. *6* uvádzame podrobný popis NN, ktorá bola využitá na predikciu.

Architektúra NN	
LSTM(75 n.)→Dropout(0.2)→LSTM(75 n.)→Dropout(0.2) →Dense(dĺžka predpovede n.)	
Atribúty	temp, day, hour, vappr, rhum, MA_12 a MA_24
Early stopping	max. 30 epochov a zastavenie po 3 zhoršeniach validačnej MSE
Batch size	32
Použité dáta	50 000 (40 000 trénovacie a 10 000 testovacie)

Tab. 6 - Popis NN pre predikciu budúcej teploty

Nižšie na Obr. *4* následne uvádzame dosiahnuté výsledky v prípade všetkých 3 DM metód. Ako je možné vidieť z nášho testovania, najlepšie výsledky sa podarilo dosiahnuť odladenej NN s využitím viacerých atribútov.

V prípade NN, LR a RF iba s atribútom teploty môžeme pozorovať vzrastajúcu presnosť predikcie so zvyšujúcou sa dĺžkou sekvencie dát, na základe ktorých je vytváraná sekvencia. V prípade zvyšujúcej sa dĺžky predikcie na 12h dopredu, môžeme pozorovať vzrastajúcu dominanciu NN nakoľko v tomto prípade dosahuje približne o 1/3 lepšie výsledky ako LR alebo RF a to aj bez využitia dodatočných atribútov.

NN - Stateful = False (s atribútmi)												
sequence_length	**24h**			**48h**			**72h**			**168h**		
prediction_length	**rmse**	**mse**	**mae**	**rmse**	**mse**	**mae**	**rmse**	**mse**	**mae**	**rmse**	**mse**	**mae**
1	0.648	0.421	0.467	0.662	0.438	0.481	0.686	0.471	0.5	0.645	0.416	0.464
2	0.8	0.641	0.581	0.807	0.651	0.578	0.81	0.657	0.58	0.822	0.675	0.595
3	0.956	0.913	0.704	0.917	0.841	0.67	0.937	0.877	0.683	0.931	0.867	0.679
12	1.458	2.125	1.091	1.434	2.055	1.072	1.453	2.112	1.098	1.491	2.224	1.131
NN - Stateful = False (iba teplota)												
1	0.682	0.465	0.486	0.670	0.449	0.481	0.688	0.473	0.496	0.669	0.448	0.479
2	0.872	0.761	0.630	0.859	0.738	0.620	0.852	0.726	0.613	0.851	0.723	0.607
3	1.051	1.105	0.765	1.017	1.033	0.745	0.988	0.975	0.722	0.976	0.953	0.724
12	1.673	2.798	1.255	1.585	2.511	1.203	1.554	2.415	1.173	1.529	2.337	1.157
LR (polynomialita=1) (iba teplota)												
1	0.680	0.462	0.484	0.667	0.446	0.477	0.664	0.440	0.474	0.658	0.432	0.471
2	0.859	0.737	0.614	0.838	0.702	0.603	0.830	0.689	0.598	0.818	0.669	0.591
3	1.018	1.037	0.724	0.989	0.979	0.709	0.975	0.951	0.702	0.961	0.923	0.692
12	1.998	3.991	1.497	1.880	3.534	1.419	1.842	3.395	1.398	1.807	3.266	1.377
RF (100 stromov) (iba teplota)												
1	0.684	0.468	0.490	0.671	0.450	0.482	0.668	0.446	0.479	0.666	0.444	0.477
2	0.861	0.742	0.616	0.833	0.694	0.600	0.832	0.693	0.599	0.826	0.682	0.595
3	1.015	1.031	0.726	0.980	0.961	0.703	0.978	0.957	0.702	0.975	0.951	0.702
12	1.927	3.714	1.426	1.834	3.363	1.386	1.812	3.282	1.375	1.773	3.144	1.349

Obr. 4 - Predikcia budúcej teploty

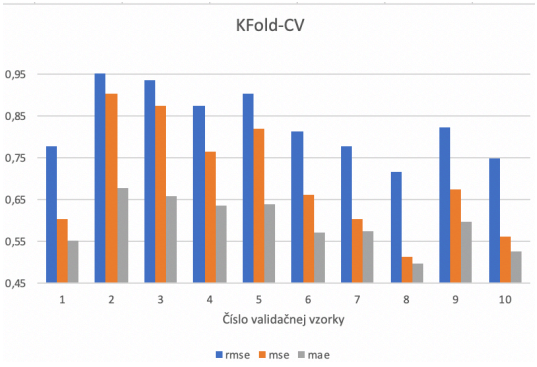
6.2 Predikcia aktuálnej teploty v susednej meracej stanici

Na predikciu aktuálnej teploty v susednom meste sme taktiež využili všetky 3 DM metódy, tak ako aj v predchádzajúcom prípade. Tu sme však použité metódy odladili. V prípade LR sme zväčšovali polynomialitu, pričom najideálnejší počet polynómov bol 3. V prípade RF sme ladili počet využitých stromov, pričom najideálnejší počet stromov, pri koľkých už nemalo význam pridávať ďalšie bol 100, keďže dosiahnutý výsledok predikcie sa od 50 stromov zlepšoval iba minimálne. V prípade NN sme využili ladenie hyperparametrov pomocou techniky *Grid Search*, kde sme skúšali rôzny *batch_size* (8, 16, 32), rôzny počet *epoch* (10, 20, 30) a rôzny počet neurónov (*units*) na 2. vrstve (32, 48, 64). Hyperparameter tuning ukázal, že najlepšia kombinácia parametrov je: *batch_size* – 8, počet *epoch* – 20 a na druhej vrstve – 48 neurónov(*units*). Zvyšné parametre ako *optimizer*, *aktivačnú funkciu*, *dropout* a *learning_rate* sme odladili ručne. V Tab. *7* uvádzame podrobný popis NN, ktorá bola využitá na predikciu.

Architektúra NN	
Dense (64 n.)->Dropout(0,1)-> Dense (48 n.)->Dropout(0,1) ->Dense(1)	
Atribúty	Použité všetky atribúty
Early stopping	max 20 epochov a zastavenie po 5 zhoršeniach validovanej MSE
Aktivačná f.	sigmoid
Optimizer	Nadam
Learning rate	0,01
Batche size	8
Použité dáta	50 000 (40 000 trénovacie a 10 000 testovacie)

Tab. 7 - Popis NN pre predikciu pre druhý scenár

Na vyhodnotenie nášho modelu sme využili K-Fold krížovú validáciu, na základe ktorej sme dostali nasledujúce výsledky, ktoré sú uvedené na Obr. *5*. Z tohto obrázka môžeme vidieť, že v rámci nášho modelu nedochádza k žiadnemu výraznejšiemu skresleniu.

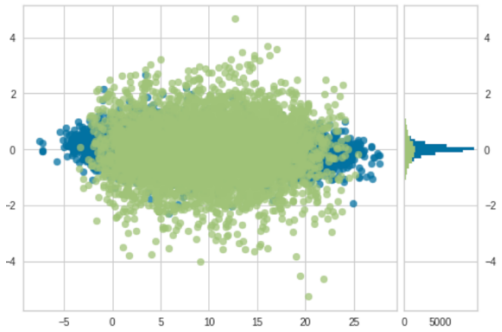


Obr. 5 - K-Fold CV

Prípad II		NN			LR (polynomialita=3)			RF(100 stromov)		
Vzájomný vzťah	Predikovaný Atribút	rmse	mse	mae	rmse	mse	mae	rmse	mse	mae
Bízko seba(19km, 24m.n.m.)	teplota	0.834	0.696	0.585	0.778	0.605	0.565	0.766	0.587	0.541
Ďalej od seba (73km, 22m.n.m)	teplota	1.179	1.389	0.870	1.098	1.205	0.807	1.099	1.208	0.802
Bízko seba(19km, 24m.n.m.)	dážď	0.354	0.126	0.095	0.359	0.129	0.098	0.346	0.119	0.086
Ďalej od seba (73km, 22m.n.m)	dážď	0.404	0.163	0.134	0.403	0.163	0.161	0.413	0.170	0.155

Tab. 8 - Predikcia aktuálnej teploty a dažďa v susedných mestách

Z Tab. *8* môžeme vidieť, že v tomto prípade bola suverénne najlepšia predikcia za pomoci RF, ktorá dosahovala vo väčšine prípadov značne lepšie výsledky. Taktiež môžeme vidieť, že v prípade zväčšujúcej sa vzdialenosti medzi mestami dochádza k výraznému zhoršeniu dosahovaných výsledkov ako v prípade teploty, tak aj v prípade dažďa. Vidíme, že relatívne presnú predpoveď aktuálnej teploty alebo dažďa v susednom meste, teda s priemernou absolútnou odchýlkou menšou ako 1, je možné vykonávať len na určité vzdialenosti, pričom z našich testov je možné vidieť, že by to mohlo byť približne 75km. V budúcnosti ešte však bude treba vykonať dôkladnejšie testovania a taktiež bude potrebné overiť, či na zhoršenie predikcie nebude mať vplyv aj rozdiel miest v nadmorských výškach, nakoľko v našom prípade bol rozdiel miest v nadmorských výškach len minimálny.



Obr. 6 - Residual plot pre RF (blízko seba, teplota)

7 Zhodnotenie

Nami navrhnuté riešenie umožňuje predikciu teploty na niekoľko ďalších hodín dopredu a taktiež predikciu aktuálnej teploty a dažďa v susednom meste. Oba tieto prípady použitia boli implementované pomocou 3 rôznych DM metód, aby bolo možné pre každý prípad vybrať tú najvhodnejšiu. V prípade predikcie do budúcna najlepšie hodnoty nadobúdala NN s využitím viacerých parametrov. V prípade predikcie v susednom meste najlepšie hodnoty nadobúdala metóda RF. Vzhľadom k tomu, že ide o veľmi špecifické prípady použitia, je obtiažne nájsť podobné riešenia, s ktorými by sme vedeli naše riešenie priamo porovnať. V porovnaní s uvedenými riešeniami najväčší rozdiel spočíva v tom, že tieto riešenia nepracujú s hodinovými záznamami, ale používajú spriemerované dlhšie časové úseky a iné meteorologické údaje. Napriek týmto rozdielom je však možné tvrdiť, že nami dosiahnuté výsledky sa približujú k existujúcim riešeniam na porovnateľne dobrej úrovni. Medzi ďalšie pokračovanie

práce môžeme zahrnúť otestovanie iných DM metód, komplexnejšie odladenie existujúcich modelov a použitie zložitejšej architektúry NN, prípadne pridanie ďalších atribútov, ktoré by mohli pomôcť k presnejšej predikcií.

Referencie

[1] D. N. Fente and D. Kumar Singh, „Weather Forecasting Using Artificial Neural Network," rev. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore, 2018, pp. 1757-1761.

[2] A. K. Pandey, C. P. Agrawal and M. Agrawal, „A hadoop based weather prediction model for classification of weather data," rev. *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, 2017, pp. 1-5.

[3] Conor Rothwell, „Irish Weather (hourly data)," 2019. [Online]. Available: https://www.kaggle.com/conorrot/irish-weather-hourly-data.

[4] Folorunsho, Olaiya & Adeyemo, Adesesan, „Application of Data Mining Techniques in Weather Prediction and Climate Change Studies," rev. *International Journal of Information Engineering and Electronic Business*, 2012.

[5] Kohail, Sarah and Alaa El-Halees, „Implementation of Data Mining Techniques for Meteorological Data Analysis," rev. *A case study for Gaza Strip*, 2011.

[6] El-Feghi, I. & Zubi, Zakaria & Abozgaya, S., „Efficient Weather Forecasting using Artificial Neural Network as Function Approximator," rev. *INTERNATIONAL JOURNAL of NEURAL NETWORKS and ADVANCED APPLICATIONS*, 2014, pp. 49-55.