FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Využití metod strojového učení pro predikci technických ztrát v přenosové síti

Bakalářská práce

Vedoucí práce:
Ing. Miloš Fetter

Vypracoval:
Valentin Papazian

Plzeň 2022

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2021/2022

ZADÁNÍ BAKALÁŘSKÉ PRÁCE
(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Valentin PAPAZIAN**
Osobní číslo:	**A19B0310P**
Studijní program:	**B0714A150005 Kybernetika a řídicí technika**
Specializace:	**Automatické řízení a robotika**
Téma práce:	**Využití metod strojového učení pro predikci technických ztrát v přenosové soustavě**
Zadávající katedra:	**Katedra kybernetiky**

Zásady pro vypracování

1. Seznamte se s problematikou technických ztrát v elektroenergetické přenosové soustavě.
2. Analyzujte výsledky případové studie řešící návrh prediktorů technických ztrát v přenosové soustavě ČR s využitím metod strojového učení.
3. Proveďte replikaci ve studii prezentovaných prediktorů pro zvolené horizonty predikce.
4. Navrhněte vlastní, metody strojového učení využívající, prediktory technických ztrát pro zvolené horizonty predikce.

Rozsah bakalářské práce: **30 – 40 stránek A4**
Rozsah grafických prací:
Forma zpracování bakalářské práce: **tištěná**

Seznam doporučené literatury:

Dodají vedoucí a konzultanti práce.

Vedoucí bakalářské práce: **Ing. Miloš Fetter**
Katedra kybernetiky

Datum zadání bakalářské práce: **15. října 2021**
Termín odevzdání bakalářské práce: **23. května 2022**

L.S.

Doc. Ing. Miloš Železný, Ph.D.
děkan

Prof. Ing. Josef Psutka, CSc.
vedoucí katedry

V Plzni dne 15. října 2021

Poděkování

Rád bych poděkoval panu Ing. Miloši Fetterovi za vedení a podporu při vypracování mé bakalářské práce. Dále bych chtěl poděkovat panu Ing. Luboši Šmídlovi, Ph.D. za konzultace a trpělivost při spolupráci. Děkuji také společnosti Česká elektroenergetická přenosová soustava, a. s. za poskytnutí dat. Mé poděkování patří také mé rodině a přátelům za podporu během celého bakalářského studia.

Prohlášení

Předkládám tímto k posouzení a obhajobě bakalářskou práci zpracovanou na závěr studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni. Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí

V Plzni dne 17.6.2022

.....
Valentin Papazian

Abstrakt

Tato bakalářská práce se zabývá tvorbou modelů predikujících ztráty v přenosové síti České republiky. Modely byly vytvářeny v programovacím jazyce Python s využitím knihovny Scikit, která poskytuje efektivní nástroj pro prediktivní analýzu dat. V rámci práce byla zpracovávána data poskytnuta od společnosti Česká elektroenergetická přenosová soustava, a. s.. Data jsou od května 2018 do července 2020 a obsahovala údaje o ztrátách z obchodního měření, teplotě, předpovědi VTE (výkon větrné elektrárny), předpovědi výkonu FVE (fotovoltaické elektrárny) atd.. Pomocí těchto dat bylo stanoveno 102 různých příznaků pro strojové učení. Dále byly testovány a navrženy vhodné trénovací algoritmy. Data byla rozdělena na trénovací a testovací část, kdy trénovací období zahrnovalo prvních 24 měsíců a testovací období poslední 2 měsíce. Získané predikce z trénovací části byly porovnány s testovacími daty. Hodnocení úspěšnosti predikce bylo provedeno výpočtem střední absolutní chyby (*mean absolute error*), střední absolutní procentuální chyby (*mean absolute percentage error*) a největší absolutní chyby (*maximum absolute error*). Nejlepšího výsledku bylo dosaženo při výběru 57 příznaků *Lasso* algoritmem z normalizovaných dat a při výběru *Histogram-based gradient boosting regrese* jako trénovacího algoritmu. Dosažený výsledek: 8,78 střední absolutní chyba, 8,27 střední absolutní procentuální chyba a 50,76 největší absolutní chyba.

Klíčová slova

Energetika, ČEPS, přenosová soustava, strojové učení, prediktivní analýza, umělá inteligence, učení s učitelem

Abstract

This bachelor thesis deals with the creation of models predicting losses in the transmission network of the Czech Republic. The models were created in the Python programming language using the Scikit library, which provides an effective tool for predictive data analysis. As part of the work, data provided by the company Česká energetická prenosová soustava, a. s. were processed. The data are from May 2018 to July 2020 and contained temperature, photovoltaic power plants and wind power plants (complete, understand). Using this data, 102 different neural network training symptoms were determined. Furthermore, suitable training algorithms were tested and designed. The data were divided into training and testing parts. The predictions obtained from the training part were compared with the test data. The prediction success was evaluated by comparing mean absolute error, mean absolute percentage error and maximum absolute error. The best result was achieved by using 57 features with the Lasso algorithm and by using the Histogram-based gradient boosting training algorithm with a result of 8,78 mean absolute error, 8,27 mean absolute percentage error and 50,76 largest absolute error.

Key words

Energetics, Czech Transmission System Operator, Machine learning, Predictive analysis, Artificial intelligence, Supervised learning

Obsah

1 Úvod

V dnešním moderním světě se čím dál víc využívá elektrická energie. Důvodem může být snadný přenos, obnovitelnost, ekologie, výkonnost. S používáním se ale pojí i jedna veliká nevýhoda, a tou je uskladnění elektrické energie. Přestože je vývoj baterií na vzestupu, stále jsou příliš neefektivní a drtivá většina spotřebičů využívá energii přímo ze sítě. Proto je nutné, aby spotřeba a výroba elektrické energie byly co nejvíce v rovnováze. V případě větší spotřeby než produkce, by spotřebiče neměly požadovaný příkon a nemusely by fungovat správně. Naopak kdyby byla větší výroba než spotřeba, energie by nebyla nijak využita a její cenu by žádný zákazník nezaplatil, také by mohlo dojít k poškození sítě a blackoutu - přerušení dodávky elektrické energie. V běžné praxi je vždy výroba větší než spotřeba, aby nedošlo k přerušení dodávky el. energie. Nicméně cílem je dosáhnout co nejnižšího rozdílu výroby a spotřeby. Touto problematikou se zabývá řada společností. Tou hlavní v České republice je Česká elektroenergetická přenosová soustava (ČEPS), která je provozovatelem přenosové soustavy na našem území. Zajišťuje bezpečný a spolehlivý provoz a rozvoj elektroenergetické přenosové soustavy v rámci propojených evropských soustav.

Celková spotřeba elektrické energie závisí na mnoha faktorech. Domácnosti i firmy mají různé nároky na využití elektřiny např. vzhledem k ročnímu období, zda je víkend, pracovní den nebo svátek, o jakou část dne se jedná nebo jaké je počasí. Během dne odebírají elektřinu ve zvýšené míře stroje v továrnách, přístroje na pracovištích, v obchodech a dopravní prostředky (tramvaje, trolejbusy, vlaky). Po západu slunce se zvýší spotřeba energie z důvodu osvětlení. Výrazně nižší je spotřeba elektrické energie od firem během víkendu a státních svátků, jelikož většina zaměstnanců nepracuje a není tedy nutné odebírat energii pro provoz strojů a zařízení, které obsluhují. V zimě jsou vyšší nároky na odběr elektrické energie z důvodu zvýšené nutnosti svícení a kvůli vytápění. Oproti tomu v létě stoupá spotřeba energie při extrémně vysokých teplotách kvůli využití klimatizace. Některé změny v nároku na spotřebu elektrické energie jsou cyklické (střídání dne a noci, všedních dnů a víkendu, ročních období). Jiné změny souvisí se společenskými trendy a socioekonomickým vývojem. Například rostoucí požadavky na odběr ze sítě je i z důvodu nárustu elektromobility. V České republice je poměrně stabilní podnebí na celém území, ale v případě jiných států hraje vliv na spotřebu energie i geologická lokace (oblast u moře, v poušti, na horách apod.).

Celková výroba elektrické energie závisí na výkonu elektráren u nás i v zahraničí. Elektrárny jsou rozlišovány podle využívaných zdrojů na obnovitelné a neobnovitelné. V dnešní době existují elektrárny tepelné, jaderné, vodní, větrné, geotermální, solární a přílivové. Na našem území mají největší zastoupení tepelné a jaderné. V menším rozsahu se využívají i vodní, větrné a solární. Nevýhodou neobnovitelných zdrojů je konečné množství zdrojových materiálů, které jsou nutné k jejich provozu, jako je uhlí nebo zemní plyn. Jejich výkon se ale dá poměrně spolehlivě odhadnout, protože mimo případy technických závad je úměrný dodaným zdrojům. Oproti tomu obnovitelné zdroje mají výhodu v tom, že jejich zdroj není limitovaný, protože využívají přírodní energii jako sluneční svit, vítr, proud vody od řek nebo moře. Tyto zdroje jsou ale nespolehlivé, protože jsou výrazně

ovlivněny ročním obdobím, fázemi dne a počasím.

Ztráta elektrické energie je výsledkem rozdílu celkové výroby a spotřeby. Snahou je dosáhnout minimalizace této hodnoty. Z tohoto důvodu jsou vytvářeny predikce, které pracují se známými měřitelnými parametry. Vzhledem k velkému množství dat je vhodné využít algoritmy strojového učení (*machine learning*)[1]. Největší přesnosti se dosáhne vhodným využitím trénovacích algoritmů. Cílem této práce je na poskytnutých datech nalézt nejvhodnější algoritmus pro předpověď ztrát elektrické energie.

1.1 Cíle práce

1. Prostudovat literaturu zabývající se energetikou v České republice a algoritmy strojového učení.
2. Prostudovat data a kódy poskytnuté od společnosti ČEPS.
3. Seznámit se s prostředím Python 3, porozumět fungování poskytnutých kódů, replikovat a vylepšit výsledky společnosti ČEPS.
4. Nalézt co možná nejlepší kombinace algoritmu předzpracování dat a strojového učení pro získané co nejmenší chybovosti v předpovědi ztrát na přenosové soustavě.

2 Teoretická část

V této kapitole jsou popsány všechny důležité pojmy a vztahy, které jsou využity v praktické části. První dvě podkapitoly se zabývají tématem energetiky v České republice a popisem společnosti Česká elektroenergetická přenosová soustava. Následující tři podkapitoly se věnují oblastem předzpracování dat algoritmům strojového učení. Poslední podkapitola je zaměřena na definování matematických vztahů a vzorců z oblasti statistiky.

2.1 Stav energetiky v České republice

Tato podkapitola představuje téma energetiky v České republice. V prvním oddíle je zaveden pojem elektroenergetika jako obecná vědní disciplína. V dalším oddíle je popsána výroba a spotřeba elektrické energie v ČR. Dále jsou uvedeny nejvýznamější elektrárny a je popsán trh s elektřinou.

2.1.1 Elektroenergetika

Vědní obor, který se zabývá zabezpečením elektrické energie pro lidstvo, se nazývá elektroenergetika. Zkoumá proces výroby elektrické energie, jeho přenos a distribuci, spotřebu, provoz elektrizační soustavy, dispečerské řízení, zabezpečení a řízení rozvoje elektroenergetiky. Elektrizační soustava představuje systém, který zajišťuje výrobu, přenos, distribuci a konečné užití elektrické energie. Mezi základní úkoly elektrizační soustavy patří zajistit, aby elektrické energie bylo dostatečné množství v požadovaném čase, dostatečně kvalitní, spolehlivá dodávka a ekonomičnost.

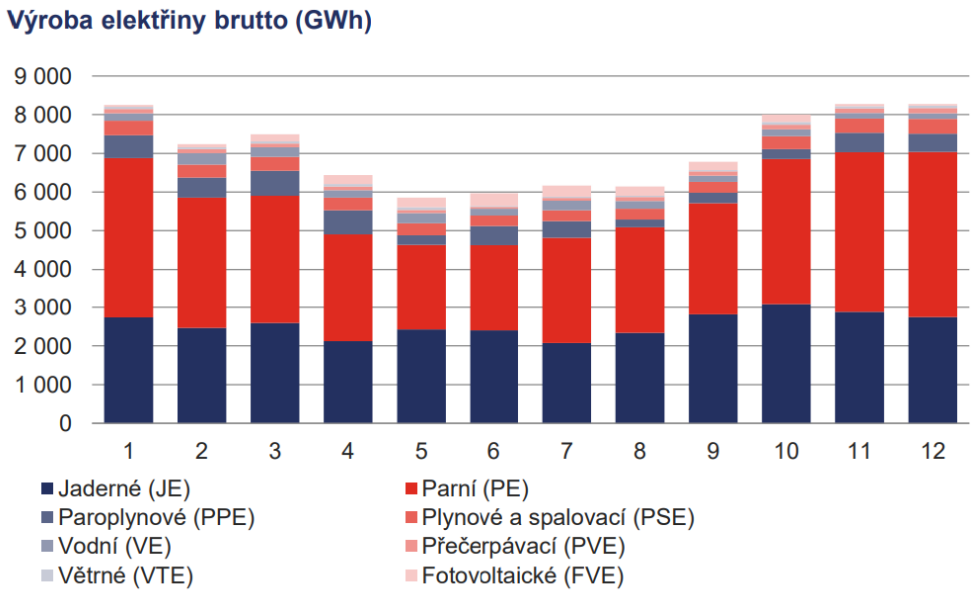
Elektrická energie je obecně získávána přeměnou energie z primárních zdrojů, jako je energie sluneční, jaderná, vodní, větrná a tepelná. K přeměně slouží solární panely, palivové články, elektro-mechanické generátory, termo-elektrické měniče, termo-emisní měniče a MHD měniče. Zdroje se mohou dělit podle toho, jakým způsobem jsou získávány, na prvotní zdroje získávané těžbou, vyrobené zdroje vzniklé zušlechtěním a druhotné zdroje, které vznikají ze ztrát při přeměnách. Další způsob rozdělení zdrojů je na neobnovitelné a obnovitelné. Mezi neobnovitelné zdroje patří fosilní a jaderná paliva. Jedná se o suroviny. kterých je pouze omezené množství. Obnovitelné zdroje jsou takové, které se sami nebo s přispěním člověka jsou schopné částečně nebo úplně obnovit. Patří k nim energie vody, větru, slunečního záření, geotermální energie, biomasy, bioplynu, energie mořských vln, přílivu a odlivu.

2.1.2 Výroba a spotřeba elektrické energie v ČR

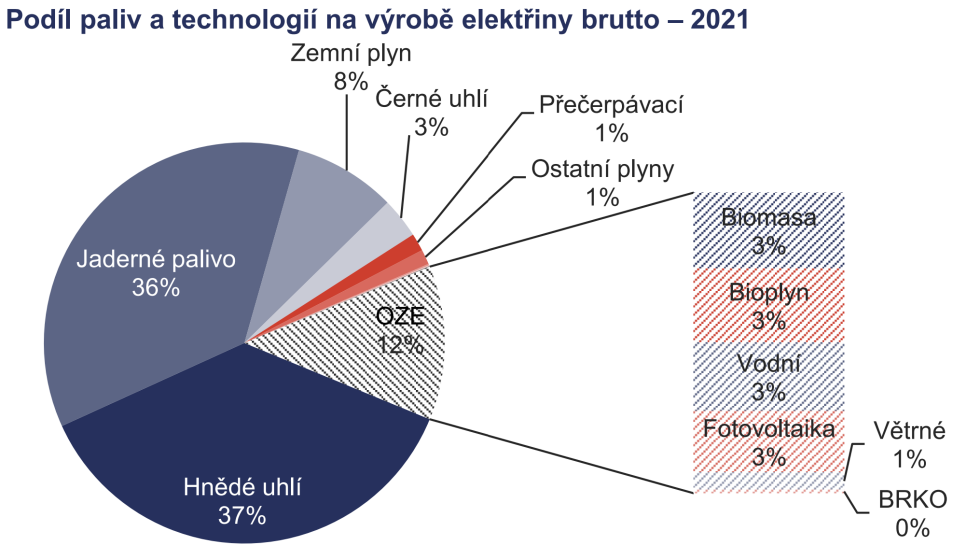
V dokumentu Roční zpráva o provozu elektrizační soustavy České republiky za rok 2021 publikovanou ERU (energetický regulační úřad), se celková výroba brutto elektřiny uvádí 84,9 TWh. Tuzemská brutto spotřeba byla 73,7 TWh. Hodnota výroby elektřiny netto odečítá od brutto hodnotu technologické vlastní spotřeby elektřiny na výrobu elektřiny. Celková výroba elektřiny netto byla 79,3 TWh. Roční maximum zatížení v soustavě bylo

naměřeno 15. února 2021 v 8:45 s hodnotou 12 149 MW. Jedná se o historické maximum zatížení elektrizační soustavy. Roční minimum zatížení soustavy bylo dosaženo 8. srpna 2021 v 5:45. Je evidentní, že na zatížení soustavy má zásadní vliv roční období a část dne.

Největší zastoupení výroby elektrické energie v roce 2021 měly parní energie 44,3 % a jaderné 36,6 %. Nejčastějším zdrojem je tedy hnědé uhlí a jaderné palivo. Další typy byly zastoupeny následovně: paroplynové 6,5 %, plynové a spalovací 4,7 %, vodní 3,0 %, fotovoltaické 2,7 %, přečerpávací 1,5 % a větrné 0,7 %. Obnovitelné zdroje energie jsou na výrobě elektřiny brutto zastoupeny z 12,4 %. Graf vývoje výroby elektřiny brutto v kalendářních měsících roku 2021 podle typu elektrárny jsou uvedeny na obr. 1. Diagram zastoupení paliv a technologií na výrobě elektřiny brutto v roce 2021 je uveden na obr. 2. Zkratka BRKO představuje biologicky rozložitelnou část komunálního odpadu.

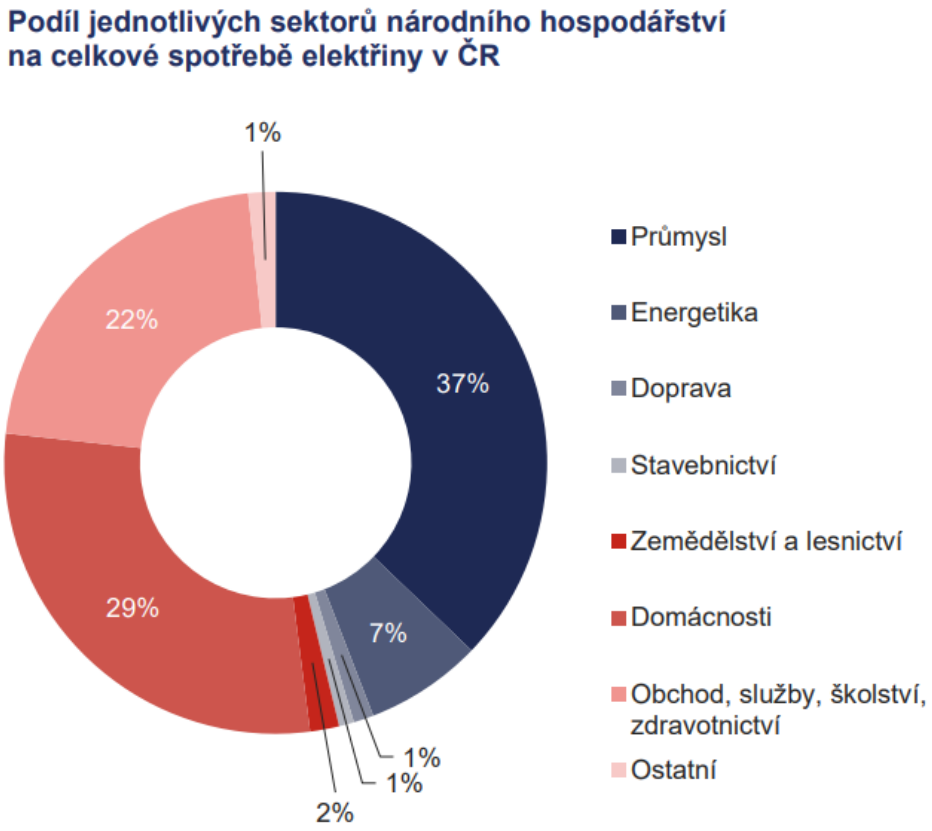


Obrázek 1: Graf množství vyrobené elektřiny v České republice podle typu elektrárny v roce 2021 [2]



Obrázek 2: Podíl paliv a technologií na vyrobené elektřině v České republice v roce 2021 [2]

Elektřina v dnešní době zajišťuje osvětlení, vytápění, transport i fungování rozmanitých zařízení ve fabrikách, podnicích i v domácnostech. Na obr. 3 je znázorněn diagram podílu zastoupení jednotlivých sektorů národního hospodářství na celkové spotřebě elektřiny v České republice za rok 2021. Je zřejmé, že největší zastoupení má s 37 % průmysl, dále s 29 % domácnosti a s 22 % obchody, služby, školství a zdravotnictví. Další sektory jako je energetika, doprava, stavebnictví, zemědělství a lesnictví jsou zastoupeny jen v jednotkách procent [2].



Obrázek 3: Podíl sektorů průmyslu na spotřebě elektřiny [2]

2.1.3 Elektrárny v České republice

Podle dat ERU (Energetický regulační úřad) jsou v České republice zastoupeny následující typy elektráren: jaderné (JE), parní (PE), paroplynové (PPE), plynové a spalovací (PSE), vodní (VE), přečerpávací (PVE), větrné (VTE) a fotovoltaické (FVE). Většinový podíl na výrobě mají jaderné a parní elektrárny. Mezi významné tepelné elektrárny v České republice patří např. Počerady, Tušimice, Poříčí a Chvaletice. Palivem v tepelných elektrárnách je nejčastěji hnědé uhlí, černé uhlí a zemní plyn. Dále se využívá také biomasa, hutní plyn, koksárenský plyn nebo lehký topný olej. Na našem území se také nachází řada již odstavených elektráren, některé už fyzicky ani neexistují. Mezi tyto historické tepelné elektrárny patří např. Tušimice, Prunéřov nebo Mělník. Poslední dvě uvedené byly odstaveny teprve nedávno v průběhu uplynulých dvou let.

V České republice se nacházejí dvě jaderné elektrárny, a to v Dukovanech se 4 bloky a v Temelíně se 2 bloky. Provozovatelem obou je skupina ČEZ. Momentálně je v procesu příprava na výstavbu nového bloku v Dukovanech a diskutuje se i o možnostech budoucího rozšíření elektrárny v Temelíně. Mezi největší vodní elektrárny v Čechách patří Dlouhé Stráně, Orlík, Dalešice, Slapy a Lipno. Vodní elektrárny mohou být průtokové, akumulační

nebo přečerpávací. Akumulační a přečerpávací vodní nádrže fungují v časech velkého vytížení, protože jsou schopny dodat rychle velké množství energie. V nočním režimu, kdy je energie v síti přebytek, akumulují vodu. Provozovateli vodních elektráren v Čechách jsou skupina ČEZ, E.ON Trend, Energo-Pro, Povodí Vltavy, Povodí Ohře a další [2].

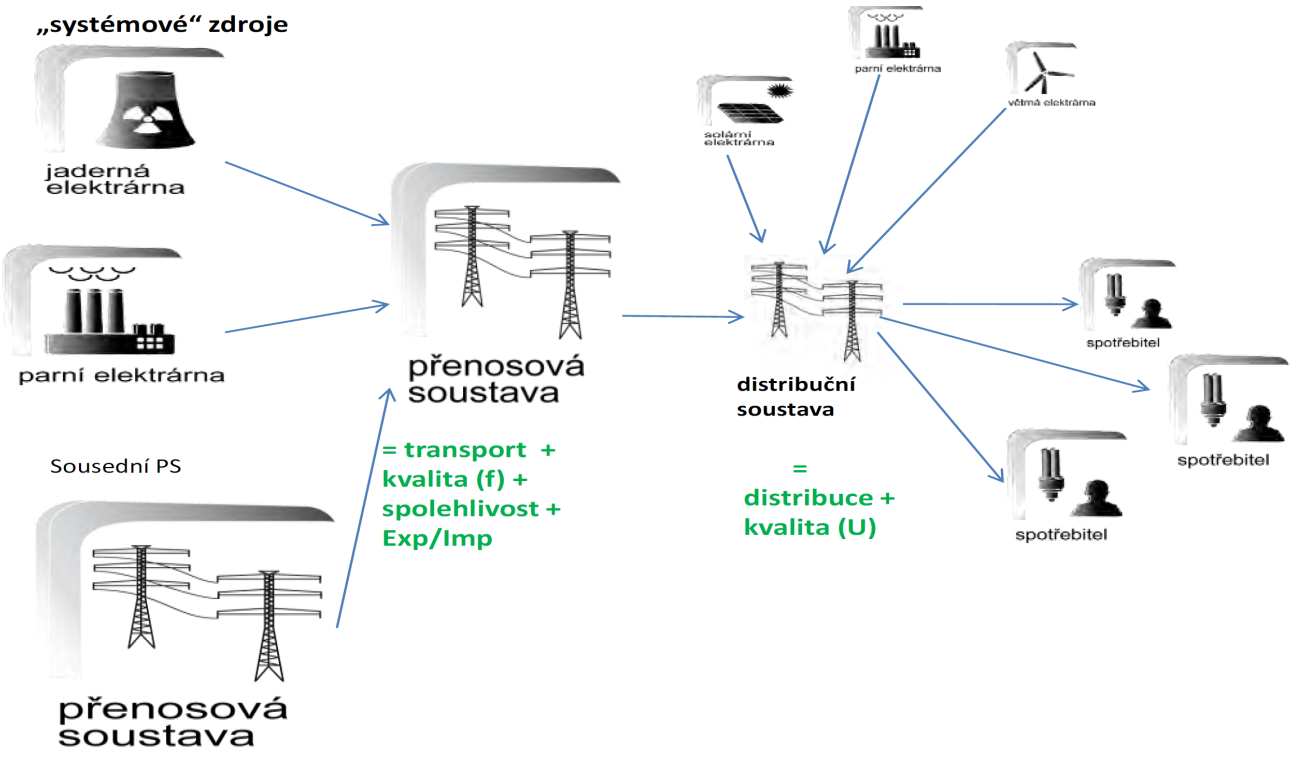
2.1.4 Trh s elektřinou

Elektřina je specifická obchodní položka, jelikož je velmi obtížné ji skladovat. Speciální je i její způsob dopravy přes hromadnou síť, podobně jako plyn nebo voda. U těchto komodit je rozdíl oproti koupi běžného zboží, kdy se předem smluví jeho množství. Síťové komodity spotřebitel čerpá ze sítě téměř neustále a neomezeně a svou objednávku tedy utváří v reálném čase skutečným odběrem. Zároveň musí v síti platit rovnováha mezi dodanou energií a součtu spotřebované energie a ztrát během přepravy. Narušení této rovnosti ovlivňuje nežádoucím způsobem výkon a frekvenci sítě. O rovnováhu se stará dispečink pomocí regulačních energií. V České republice zajišťuje tuto službu společnost ČEPS (viz podkapitola 2.2). Subjekty, které s elektřinou obchodují jsou pokutovány, pokud nedodrží smluvené množství dodané či spotřebované energie. Trh s elektřinou se dělí na velkoobchodní mezi výrobními závody a obchodníky a na maloobchodní mezi koncovými zákazníky a obchodníky. V případě velkoobchodu nesou odpovědnost za odchylky obě strany, v případě maloobchodu pouze prodejci elektřiny. Odpovědnost koncového zákazníka za množství odebrané energie je přenesena na obchodníka [3].

Mezi účastníky trhu s elektřinou patří výrobce, odběratel, obchodník, burza, provozovatel distribuční soustavy, provozovatel přenosové soustavy, operátor trhu s elektřinou a energetický regulační úřad. Pro svou činnost potřebují licenci výrobce, obchodník, provozovatel distribuční soustavy, provozovatel přenosové soustavy a operátor trhu s elektřinou. Všechny licence subjektům vydává Energetický regulační úřad, a to nejméně na 25 let (v případě obchodu s elektřinou nejméně na 5 let) [3]. Trh s elektřinou je možné rozdělit na organizovaný, zajišťovaný např. burzou, a neorganizovaný. Dále se trh s elektřinou dělí na dlouhodobý a krátkodobý, kde se obchoduje v řádech dnů i hodin, a spadají pod něj blokový, denní, vnitrodenní a vyrovnávací trh. Na blokovém trhu se nakupuje na den (*base*), špičku spotřeby, která je mezi 8:00 a 20:00 (*peak*) a období mimo špičku mezi 20:00 a 8:00 (*off-peak*). Na denním trhu se prodává den dopředu formou aukce na každou hodinu dne. Vnitrodenní trh probíhá v daný den a obchoduje se s aktuálními neplánovanými nedostatky nebo přebytky většinou alespoň hodinu dopředu. Na vyrovnávacím trhu nakupují provozovatelé přenosové soustavy 30 minut před začátkem dodávky a mohou díky tomu získat regulační energii [4].

V České republice působí jako burza společnost Power Exchange Centra Europe (PXE a. s.) a Českomoravská komoditní burza Kladno. Dále je pro Česko zásadní burza European Energy Exchange, která se nachází v německém Lipsku, jejíž je PXE dceřiná společnost. Zde dochází k největšímu objemu obchodů ve střední Evropě. Provozovatelem energetické přenosové sítě v Česku je ČEPS, a. s., která se zároveň stará o bilanci energetické soustavy. Největšími provozovateli distribučních sítí elektřiny v Česku jsou ČEZ, Pražská energetika, a. s. (PRE) E.ON a EG.D (*Electricity and Gas Distribution*).

Přenosová soustava má za úkol dálkový přenos elektřiny z míst výroby (velké elektrárny) do míst koncentrované spotřeby. Propojení je vícenásobné, aby byla zajištěna dodávka i v případě výpadku jedné části vedení. V rámci přenosové sítě je regulována rovnováha zdrojů a spotřeby, toky v sítích a zajištěna systémová frekvence a napětí na uzlech. Přenosová síť má páteřní charakter a propojuje hlavní uzly státu a jeho území. Oproti tomu distribuční síť má typicky paprskový charakter, kdy vychází z napojení na přenosovou síť a rozvádí elektřinu ke koncovým odběratelům a také jsou do ní napojeny i menší elektrárny. Velcí průmysloví odběratelé (např. doly, velké ocelárny) jsou napojeni přímo na přenosovou síť, ale většina koncových zákazníků je napojena na distribuční síť. Na úrovni distribuční soustavy se řídí toky v sítích a zajišťují se lokální parametry kvality, zejména napětí. Na obr. 4 je znázorněn zjednodušený model fyzického toku elektřiny přenosovou a distribuční soustavou [3].



Obrázek 4: Schéma přenosové a distribuční sítě [4]

Operátor trhu s elektřinou organizuje od roku 2001 obchodování na denním a krátkodobém trhu a jediným akcionářem je Ministerstvo průmyslu a obchodu. Registruje účastníky obchodu, zpracovává transakce, měří a zúčtovává odchylky. Energetický regulační úřad reguluje část výsledné ceny energie, která se neřídí tržními mechanismy [5].

2.2 Česká elektroenergetická přenosová soustava

Společnost ČEPS a. s. zajišťuje provoz elektroenergetické přenosové soustavy v České republice. Jejím hlavním cílem je zajistit rovnováhu vyrobené a spotřebované elektrické energie v každém okamžiku a tím zajistit spolehlivou a bezpečnou dodávku elektřiny pro všechny uživatele. Tato firma je také součástí trhu s elektřinou v Evropě. Stará se o 44 rozvoden s transformátory, které převádí elektrickou energii z přenosové sítě na distribuční. Dále spravuje elektrická vedení o napětí 400 kV, 220 kV a některé 110 kV. Jednotlivé distribuční sítě jsou spolu propojeny vysokonapěťovým vedením. Do celkové přenosové sítě jsou také připojené velké elektrárny jako například Temelín, Dukovany, Dlouhé stráně, Tisová [6].

2.2.1 Historie a struktura společnosti

Firma byla založena v roce 1998. Nicméně přenosová síť vznikala už o 50 let dříve. V roce 1950 došlo na našem území ke sjednocení dosud samostatných přenosových systémů, které tím mohly začít spolupracovat. Do roku 2003 byl jediným akcionářem společnosti ČEZ. Následně bylo 15 % akcií prodáno Ministerstvu práce a sociálních věcí a 51 % Ministerstvu financí. V roce 2004 odprodal ČEZ, a. s. státu svůj zbývající podíl. V roce 2009 byly akcie společnosti vlastněné Ministerstvem financí převedeny na Ministerstvo průmyslu a obchodu. V roce 2012 byly převedeny akcie Ministerstva práce a sociálních věcí na Ministerstvo průmyslu a obchodu, které se tím stalo stoprocentním držitelem akcií. Jediným akcionářem společnosti ČEPS, a. s. je tedy stát a z jeho pověření vykonává výkon akcionářských práv Ministerstvo průmyslu a obchodu. K 31. 12. 2021 měly akcie celkovou jmenovitou hodnotu přes deset miliard korun českých.

ČEPS, a. s. má na základě svých stanov následující orgány: valná hromada, dozorčí rada, výbor pro audit a představenstvo. Vzhledem k tomu, že ČEPS má jen jediného akcionáře, vykonává působnost valné hromady Ministerstvo průmyslu a obchodu. Dozorčí rada má devět členů, kteří jsou v souladu se stanovami voleni ze dvou třetin valnou hromadou a z jedné třetiny zaměstnanci společnosti. Výbor pro audit má tři členy, kteří jsou ustanoveni na čtyři roky akcionářem společnosti. Představenstvo má pět členů s funkčním obdobím pět let. Členové představenstva nejsou výkonným vedením společnosti, ale určují strategii, řídí obchodní a podnikatelskou činnost a komunikují s akcionářem. ČEPS zaměstnává k 31. 12. 2021 celkově 621 osob [7].

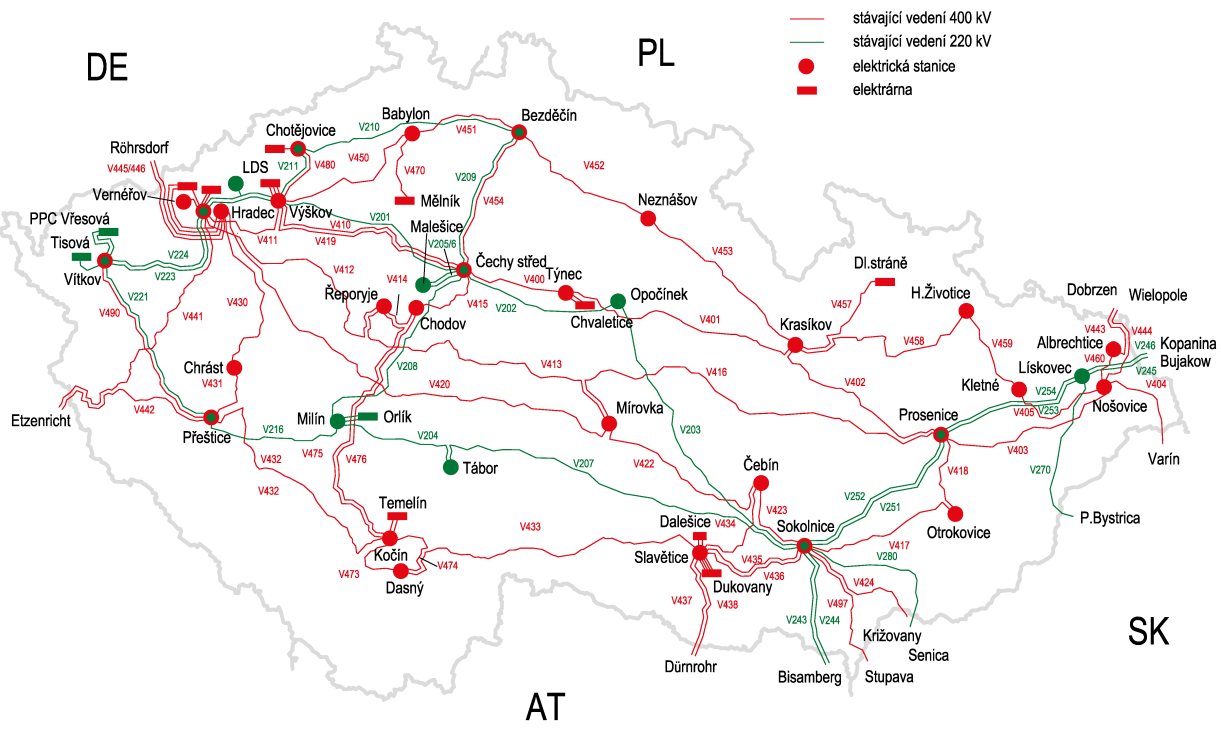
2.2.2 Činnost

Nejdůležitější činnost společnosti ČEPS je provádění dispečerského řízení přenosové soustavy na území Česka v reálném čase. ČEPS je konkrétně zodpovědná za stabilitu výkonu a frekvence, regulace napětí a jalového výkonu. Pro zajišťování stability uvedených parametrů ČEPS nakupuje na trhu potřebné výkonové rezervy. Firma také vlastní monopol na dálkový přenos elektrické energie vysokého napětí.

Řízení rovnováhy spotřeby a výroby se provádí nakupováním a prodáváním elektřiny

na burze. To se děje každých 15 minut, ale nakupuje se i na 2 hodiny dopředu a na den dopředu. Je tedy nutné odhadnout co nejpřesněji ztráty (rozdíl výroby a spotřeby), aby bylo možné dokoupit jen nutné množství elektřiny.

Firma také vlastní monopol na dálkový přenos elektrické energie vysokého napětí. Množství elektřiny přenesené přenosovou soustavou na výstupu byl v roce 2021 více než 68 000 GWh. Součástí aktivit společnosti je i údržba komponentů přenosové sítě, jako jsou rozvodny, odpojovače, vypínače, přístrojové transformátory, svodiče přepětí, kotevní a nosné stožáry, kabely a dráty. Údržba zahrnuje jak pravidelné kontroly, opravy a výměny, tak i řešení mimořádných událostí, jako jsou kritické meteorologické jevy. Bouře, silný vítr, povodně nebo dokonce tornádo či zemětřesení mohou v extrémních případech způsobit pád stožárů nebo jiné poškození soustavy. Největší část vedení ČEPS má napěťové úrovně 400 kV s délkou 3 795 km. Dále soustava zahrnuje napěťovou úroveň 220 kV s délkou 1 824 km a 110 kV s délkou 84 km. Přenosová soustava déle obsahuje 44 rozvoden a je napojena na 13 elektráren. Na obrázku 5 je znázorněno umístění elektráren, elektrických stanic (rozvoden) a tras vedení 400 kV a 220 kV na území České republiky.



Obrázek 5: Schéma elektroenergetické sítě ČR

2.2.3 Mezinárodní význam

Společnost ČEPS je součástí Evropské sítě provozovatelů přenosových soustav (*European Network of Transmission System Operators for Electricity*), zkráceně ENTSO-E. Jedná se o uskupení 41 evropských provozovatelů přenosových soustav z 34 zemí Evropy. Provozovatelé přenosových soustav (*Transmission System Operators*), zkráceně TSO, v sousedních státech České republiky jsou: na Slovensku SEPS; v Rakousku Verbund - Austrian Power Grid a VKW-Netz; v Německu EnBW Transportnetze , Tennet TSO, Amprion a 50Hertz Transmission a v Polsku PSE-Operator.

Společnost spolupracuje s těmito přenosovými soustavami sousedních zemí. Z toho především s přenosovou sítí Německa, která obsahuje rozsáhlé větrné elektrárny na severu země. Ty mohou při vhodných podmínkách generovat velké množství elektrické energie, která vlivem propojených evropských soustav nezůstane jen v Německu, ale hledá si cestu nejmenšího odporu a dochází tak k zatěžování i české přenosové soustavy. Proto je u hranice s Německem (Hradec U Kadaně) umístěn speciální transformátor PST, který pomocí regulace fáze usměrňuje tok elektřiny. Příliš velké množství elektřiny by mohlo síť přetížit a došlo by k výpadku. Při výpadku jedné sítě dojde k přetížení dalších linek a mohlo by dojít také k jejich výpadku, nebyla by tedy možná dodávka elektrické energie na větším území. Proto má každá linka (vnitrostátní i mezinárodní) svůj bezpečnostní limit.

ČEPS uzpůsobuje svou činnost tak, aby vedla k postupnému naplnění cílů Zelené dohody pro Evropu (*The European Green Deal*), který zásadním způsobem změní celý energetický sektor i navazující oblasti. Probíhající transformace energetiky zahrnuje snahu o decentralizaci, dekarbonizaci a digitalizaci. V roce 2021 byl společností publikován dokument „11 klíčových podmínek přechodu k nízkoemisní (elektro)energetice v ČR“ [7].

Obrázek 6: Přeshraniční fyzické toky elektřiny [GWh] v roce 2021 [2]

Na obr. 6 jsou znázorněny přeshraniční fyzické toky elektřiny [GWh] mezi Českou republikou a sousedními státy za rok 2021. Celkový export byl 26 228,2 GWh a celkový import 15 153 GWh. Největší export byl do Rakouska, následně do Slovenska a Německa. Největší import je z Polska a následně z Německa. Zbylé hodnoty exportu a importu mají jen minoritní zastoupení.

2.3 Předzpracování dat

Předzpracování dat převádí hrubá data (*raw data*) a signály na reprezentaci dat vhodnou pro aplikaci prostřednictvím sekvence operací. Spočívá v odstranění irelevantních informací a ponechání pouze klíčových vlastností dat. Pokud se systému na vstupu dodají nekvalitní data, nelze očekávat kvalitní výstup, což vyjadřuje anglická fráze *garbage in, garbage out*. Mezi cíle předzpracování dat patří čištění dat, zmenšení dimenze vstupu, normalizace dat a vybrání příznaků.

2.3.1 Čištění dat

Čištění dat je proces oprav, úprav a mazání dat, která jsou nesprávná, neúplná nebo duplikovaná [8]. Může se jednat i o odstraňování odlehlých bodů. Čištění může být jednorázové nebo může být prováděno průběžně. Data je potřeba vyčistit zejména kvůli špatnému vkládání (entry errors), chybějícím datům, špatné dokumentaci, nejednotné metodice při vkládání dat nebo chybám v přenosu. Čištění může probíhat i manuálně.

2.3.2 Zmenšení dimenze vstupu

Protože obtížnost učení roste s počtem dimenzí, je snaha transformovat data z vyšší dimenze do nižší při minimalizaci ztráty jejich informační hodnoty. Jeden z hlavních algoritmů snižujících dimenzi je **PCA** (*Principal Component Analysis* - analýza hlavních komponent) [9].

PCA slouží k dekorelaci dat. Často se používá k snížení dimenze dat s co nejmenší ztrátou informace. V zásadě se jedná o transformaci vstupu do jiné souřadné soustavy. Transformace je lineární, tedy nové příznaky jsou lineární kombinace původních. V novém prostoru je lze přepsat jako osy. První osa vede ve směru největšího rozptylu hodnot. Druhá osa vede ve směru druhého největšího rozptylu hodnot, atd.. Osy jsou ortogonální. Výsledkem je stejný počet os jako původní dimenze, funkce je tedy bezztrátová. Uživatel se ale může rozhodnout nějaké osy nepoužít, tím dojde k redukci celkové dimenze dat. Další algoritmy snižující dimenzi jsou např. ica, svd. Ke snížení dimenze může dojít i na základě expertního posouzení, Lasso algoritmu nebo korelační analýzy.

Lasso algoritmus se může využít jak pro výběr příznaků, tak pro regresi. U regrese se ladí intenzita penalizačního prvku obsaženého v nákladové funkci tak, aby měla nákladová funkce minimální hodnotu. Při výběru příznaků se opět minimalizuje nákladová funkce, veškeré nadbytečné příznaky budou mít váhu nastavenou na 0. Uvažovat se budou tedy pouze příznaky s váhou různou od nuly [10].

Korelace Na základě korelační analýzy se zjistí korelace všech příznaků. Poté jsou příznaky uspořádány podle korelace v absolutní hodnotě. Následně se vybere požadovaný počet příznaků s nejvyšší korelací.

2.3.3 Normalizace dat

Normalizaci dat se doporučuje použít v případech, kdy mají proměnné různý rozsah, ale nemají normální rozdělení či obsahují odlehlé hodnoty (*outliers*). Cílem normalizace je změnit hodnoty číselných sloupců v datové sadě tak, aby používaly společné měřítko, aniž by došlo ke zkreslení rozdílů v rozsazích hodnot nebo ztrátě informace. Sjednocení měřítka vstupních veličin zaručuje zvýšení efektivity algoritmů strojového učení. Dva běžné způsoby normalizace jsou:

1. *min-max* normalizace, změna měřítka sloupce na $\langle 0, 1 \rangle$

$$z = \frac{x - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}. \tag{1}$$

2. *z-score*, v tomto případě se dosáhne normálního rozložení se střední hodnotou 0 a směrodatnou odchylkou 1. Vzorec je

$$z = \frac{x - \mu}{\sigma}, \tag{2}$$

kde x znázorňuje původní sloupec, který chceme normalizovat, μ střední hodnotu a σ směrodatnou odchylku [11].

2.3.4 Rozdělení dat

Dále je za potřebí data rozdělit na trénovací, testovací a validační část. V trénovací části algoritmy hledají souvislosti, čímž se učí. Testovací data by měla být odlišná od trénovacích dat. Slouží k ověření kvality naučeného systému. Systém je dobře naučený v případě, že se stejnou úspěšností vyhodnocuje trénovací i testovací data. Pokud jsou trénovací data vyhodnocena lépe, systém je přeučený (*overfitted*). Validační část se stará o to, aby nedošlo k přeučení. Ke kontrole dochází během trénování. Přesný poměr trénovacích, testovacích a validačních dat není jasný. Záleží na množství dat a typu úlohy. Většinou se data rozdělí v poměru 75:20:5 (trénovací, testovací, validační).

2.3.5 Výběr příznaků

Cílem je nalézt z celkové množiny příznaků takovou podmnožinu příznaků, která maximalizuje schopnost učícího algoritmu. V reálných případech je nemožné nalézt tu nejlepší podmnožinu, protože prostor všech možných podmnožin je příliš velký na to, aby ho šlo celý prohledat. V této úloze se zavádí pojem **relevance příznaku**.

Příznaky se dají rozdělit na silně relevantní příznaky, slabě relevantní příznaky a irelevantní příznaky. Silně relevantní příznak je takový, když jeho vynecháním dojde vždy ke snížení kvality regrese. U slabě relevantního příznaku může dojít při jeho vynechání ke zvýšení ale i snížení kvality regrese. Irelevantní příznaky jsou pak takové, které nespadají ani do jedné kategorie.

Bohužel ale relevantní příznaky nemusí být v optimální podmnožině příznaků. Dokonce i irelevantní příznaky mohou vylepšit regresi. Příznak může být hodnotný pouze s nějakým jiným příznakem, případně kombinací příznaků. Relevantnost příznaků by se měla vyšetřovat v kombinaci s konkrétním trénovacím algoritmem. Běžnou strategií je vytvoření nějaké hodnotící funkce, podle které se příznaky uspořádají od nejrelevantnějších po méně relevantní. Hodnotící funkce pak představuje vhodnou heuristiku.

2.4 Algoritmy strojového učení

Algoritmy strojového učení pomáhají pomocí kódu nalézt a analyzovat význam ve složitých datových sadách. Každý algoritmus obsahuje posloupnost jednoznačných instrukcí, které vedou k požadovanému cíli. Modely strojového učení slouží k vytvoření vzorů, které se dají využít ke kategorizaci informací nebo vytvoření předpovědi [12].

Různé algoritmy analyzují data různými způsoby. Pro zadaná data není většinou předem jisté, jaký algoritmus bude nejlépe fungovat. Proto se musí prověřit větší množství algoritmů, aby bylo možné dosáhnout co nejlepšího výsledku. Všechny algoritmy mají navíc určité parametry, které mohou výrazně ovlivnit kvalitu výstupu, tzv. hyperparametry. Je tedy nutné zjistit vhodný algoritmus a k němu hodnoty parametrů. Základní rozdělení algoritmů strojového učení je podle:

1. **způsobu trénování**
    - učení s učitelem, viz. odstavec 2.4.1.
    - učení bez učitele, viz. odstavec 2.4.2.
    - zpětnovazební učení, viz. odstavec 2.4.3.
2. **výstupu**
    - regrese - produkuje numerickou předpověď na základě vstupu.
    - klasifikace - rozděluje vstupní data do dvou nebo několika tříd.
    - shlukování - zařazuje objekty do skupin (tzv. clustry) s podobnými vlastnostmi, typicky při učení bez učitele.

2.4.1 Učení s učitelem

V tomto případě má algoritmus k dispozici vstupní vektory $\boldsymbol{x}$ a odpovídající korektní výstupní vektory $\boldsymbol{y}$ [13]. Má tedy příklady správné transformace vstupních vektorů na výstupní. Tyto příklady se získávají měřením vstupních a výstupních hodnot systému, který chceme modelovat. Celá množina obsahující vektory $\boldsymbol{x}$ a $\boldsymbol{y}$ znázorňuje známou část chování systému. Tato množina se využívá k naučení algoritmu i ověření jeho funkčnosti.

2.4.2 Učení bez učitele

Tyto algoritmy nemají k dispozici žádné kritérium správnosti transformace vstupních dat. Algoritmy se pokouší uspořádat podobná data do stejných tříd. Počet možných tříd může být předem znám. Do učení nevstupuje žádný arbitr. Všechny dostupné informace jsou obsaženy ve vstupních datech. Tyto algoritmy uspořádávají data do shluků s podobnými vlastnostmi [12].

2.4.3 Zpětnovazební učení

Cílem zpětnovazebního učení je naučit tzv. agenta chovat se tak, aby bylo dosaženo maximálního užitku. Tento typ strojového učení používá algoritmy, které se učí z výsledků a rozhodují, jaký krok udělat dál. Po každém kroku algoritmus obdrží informaci zpětnou vazbou, která určí, jestli byl krok zvolen správně, nesprávně nebo neutrálně. Model zpětnovazebního učení se většinou skládá z množiny stavů, množiny přechodů, pravidel přechodových funkcí, pravidel, která určují bezprostřední odměnu přechodu do jiného stavu a pravidel definující cíle agenta. Úloha se dá tedy také formulovat jako Markovův rozhodovací proces [14].

2.5 Použité algoritmy

Výběr vhodných trénovacích algoritmů je důležitou částí pro vytvoření kvalitního prediktivního modelu. Jelikož neexistuje jeden univerzálně nejlepší algoritmus, je potřeba vyzkoušet více různých algoritmů. Následně jsou popsány algoritmy použité v praktické části.

2.5.1 Lineární regrese

Lineární regrese spočívá v nalezení vah (koeficientů) pro všechny funkční proměnné tak, aby součet kvadrátu odchylek skutečného výsledku a predikovaného byl co nejmenší. Nejprve se sestaví přeurčená soustava rovnic $\mathbf{Ax} = \mathbf{b}$, kde $\mathbf{x}$ je hledaný vektor vah, $\mathbf{A}$ je matice funkcí. Tento vztah se upraví do tvaru $\mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$. Váhy se získají vyřešením této rovnice [15].

2.5.2 Ridge regrese

Vychází z Lineární regrese, je přesnější v situacích, kdy jsou nezávislé proměnné silně korelované. Zmenšuje regresní koeficienty proměnných s nízkou korelací. Toho je dosaženo L2 penalizací, což je součet druhých mocnin koeficientů. Tím jsou koeficienty s nízkou korelací blízko nule. Vzorec pro výpočet je podobný vzorci Lineární regrese. Jeho tvar je $\mathbf{x} = \left(\mathbf{A}^T\mathbf{A} + k\mathbf{I}\right)^{-1}\mathbf{A}^T\mathbf{b}$. Navíc je ve vzorci jednotková matice I a volitelný parametr k.

2.5.3 Bayesovská regrese

V tomto algoritmu se popíše střední hodnota jedné proměnné lineární kombinací jiných proměnných s cílem získat posteriorní pravděpodobnost regresních koeficientů. Bayesovská regrese je vhodná v případě, že jsou data zatížena chybou [16].

2.5.4 Ensemble metody

Algoritmy spadající do této kategorie využívají více modelů, tzv. slabých studentů (*weak learners*), aby vyřešily stejný problém. Kombinace modelů může dosáhnout lepších výsledků než modely samostatně a zároveň lze dosáhnout větší robustnosti. Mimo slabé

studenty je také zadán meta-algoritmus, pomocí kterého vznikne výsledný model. Základní rozdělení Ensemble metod je: Stacking, Boosting a Bagging [17].

- **Stacking**, tento způsob může využívat i úplně odlišné modely (slabé studenty). Každý model má paralelně k dispozici všechna data k trénování. Po natrénování všech modelů (může probíhat najednou) jsou výsledky jednotlivých slabých studentů použity k natrénování meta-modelu.
- **Boosting**, v tomto případě dojde nejprve k natrénování jednoho slabého studenta, který pozmění vstupní data. Tím začne iterativní optimalizační proces. Tato pozměněná data poté slouží k natrénování dalšího slabého studenta, který je opět pozmění. Tímto způsobem se data dostanou až k finálnímu modelu schopného se nyní dobře natrénovat.
- **Bagging**, tento způsob nejprve rozdělí data. Každý slabý student dostane část dat, se kterými dojde k jeho natrénování. Výsledný model vznikne průměrováním výsledků již naučených slabých studentů.

2.5.5 Stacking Ensemble

Tento trénovací algoritmus spadá do kategorie Ensemble method. Používá meta-learning algoritmus, aby se naučil, jak nejlépe kombinovat předpovědi ze dvou nebo více základních algoritmů strojového učení.

2.5.6 Histogram-based gradient boosting

Tento trénovací algoritmus spadá do kategorie Ensemble methods. Na základě statistických informací rozděluje vstupní data do určitých skupin a umožňuje efektivnější výpočet modelu strojového učení. Tyto metody využívají více algoritmů učení k získání lepšího prediktivního výsledku, než jakého by dosáhly samotné jednotlivé algoritmy. Histogram-based gradient boosting využívá algoritmy rozhodovacích stromů. Tato metoda je obvykle lepší než náhodné stromy a je obecně méně citlivá na chybějící údaje.

2.6 Matematické vztahy

V této podkapitole jsou uvedeny matematické pojmy, se kterými se pracuje v praktické části. Tyto pojmy hodnotí úspěšnost různých algoritmů a parametrů. U použitých matematických vztahů jsou uvedeny jejich definice a vzorce.

Střední absolutní chyba (*mean absolute error*)
Střední absolutní chyba vyjadřuje míru chyby mezi dvěma párovými pozorováními, které vyjadřují stejný jev. Ve vzorci značí n počet vzorků, y_i odhadovanou ztrátu a x_i skutečnou ztrátu.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{3}$$

Střední absolutní procentuální chyba (*mean absolute percentage error*)
Střední absolutní procentuální chyba je měřítkem přesnosti prognostické metody ve statistice, kde n je počet vzorků, y_i je odhadovaná ztráta a x_i je skutečná ztráta.

$$MAPE = \frac{100\,\%}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{x_i}\right| \tag{4}$$

Největší absolutní chyba (*maximum absolute error*)
Největší absolutní chyba určuje největší odchylku mezi predikovanou (y_i) a skutečnou (x_i) hodnotou pro n vzorků.

$$i = 1 \to n \qquad \max|x_i - y_i| \tag{5}$$

Korelace
Ve statistice znamená korelace vzájemný lineární vztah mezi dvěma veličinami. Míru korelace udává korelační koeficient, který nabývá hodnot $\langle -1, 1\rangle$. Pokud dosahuje korelační koeficient hodnoty 1, mají mezi sebou porovnávané veličiny zcela přímou úměrnost. Je tedy možné vyjádřit mezi nimi přímou lineární závislost $y = kx$. Při hodnotě korelačního koeficientu -1 platí mezi porovnávanými veličinami úměra nepřímá, tedy y = -kx. V případě, že je korelační koeficient nulový, není mezi veličinami žádná lineární závislost (což ale neznamená, že na sobě veličiny nemohou záviset jinak než lineárně).

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} \tag{6}$$

Vzorec pro korelační koeficient $\rho_{X,Y}$ je uveden v rovnici (6) a určí se převedením kovariance COV(X,Y) na bezrozměrné číslo tím, že se vydělí součinem směrodatných odchylek obou proměnných σ_X a σ_Y.

Vzorce pro kovarianci COV(X,Y) je popsán v rovnici (7), kde proměnná E(X) je střední hodnota.

$$COV(X,Y) = E[(X - E[X])(Y - E[Y])] \tag{7}$$

Směrodatná odchylka σ_X je definována jako odmocnina z rozptylu náhodné veličiny a vypočítá se podle vztahu z rovnice (8), kde var(X) je rozptyl a E(X) je střední hodnota. Směrodatná odchylka vypovídá o tom, jak podobné jsou si průběhy v souboru zkoumaných hodnot. Nízká hodnota vyjadřuje podobnost prvků souboru, naopak vysoká směrodatná odchylka značí velké rozdíly mezi hodnotami proměnné [18].

$$\sigma = \sqrt{var(X)} = \sqrt{E[X - E[X]]^2} \tag{8}$$

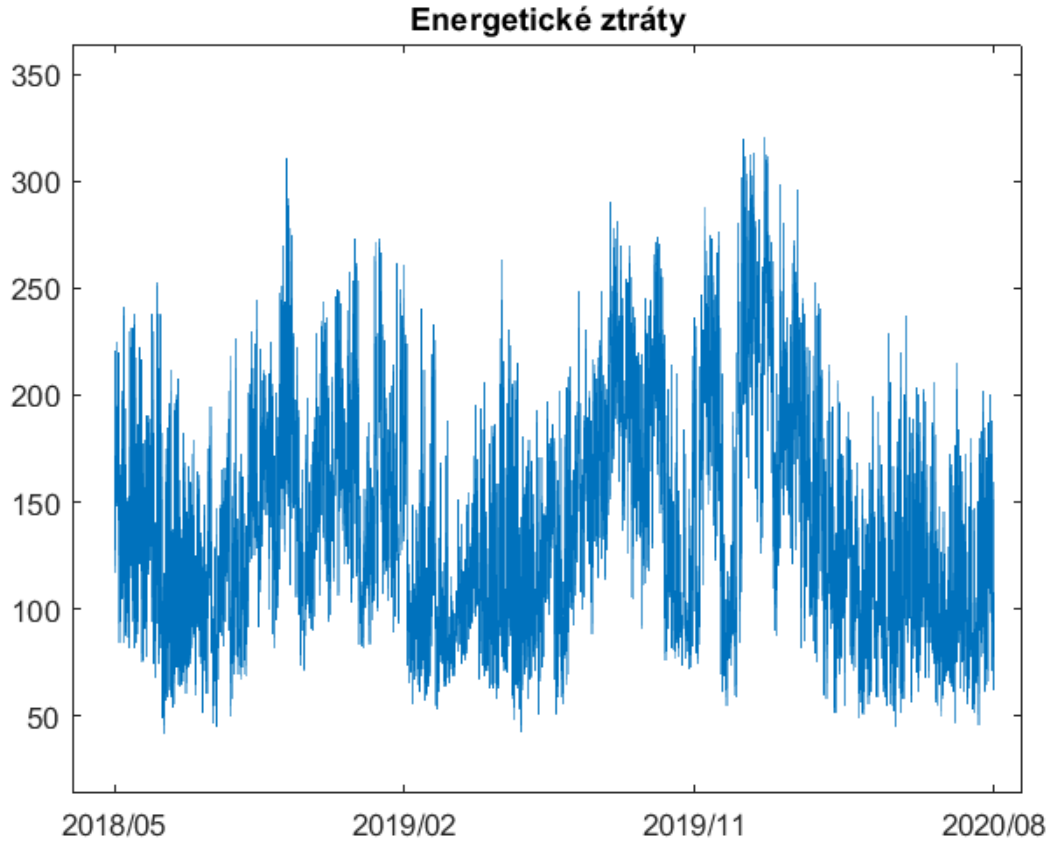
3 Praktická část

Praktická část této bakalářské práce byla provedena v programovacím jazyce Python. Základem byly kódy a data poskytnutá od společnosti ČEPS. Prvním úkolem bylo seznámit se s tím, jaké informace jsou v datech obsaženy a jak fungují algoritmy v kódech. Vstupní data obsahovala např. podrobné informace od větrných a fotovoltaických elektráren z České republiky i z Německa ve formě časových řad. Hrubá data (*raw data*) byla v první fázi zformátovaná a rozčleněna na jednotlivé příznaky. V další fázi s daty pracují algoritmy pro předzpracování dat, které využívají všechny nebo jen vybrané příznaky. Dále je potřeba vybrat a otestovat vhodné algoritmy strojového učení.

Cílem této práce je nalézt optimální řešení pro predikci technických ztrát, a to úpravou dat a vhodně zvoleným algoritmem. Po testování různých algoritmů strojového učení byly vybrány k dalšímu zkoumání následující: Lineární regrese, Ridge regrese, Stacking Ensemble, Bayesovská regrese a Histogram-based gradient boosting. Princip jejich fungování je popsán v kapitole 2.5.

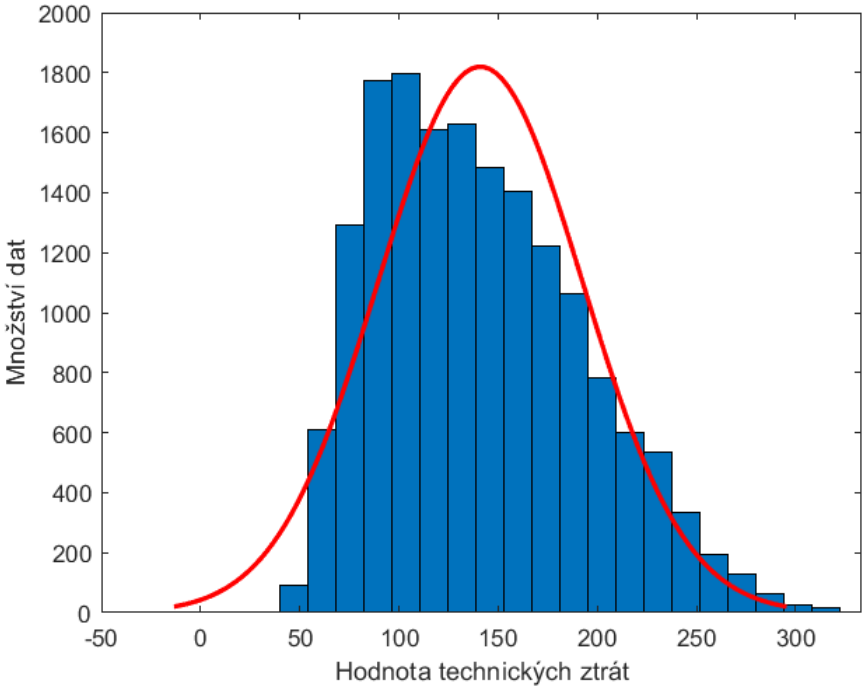
3.1 Klasifikace úlohy

Pro správné použití dat a dosažení co nejlepších výsledků byla provedena datová analýza. Jelikož neexistuje přesné analytické řešení a je k dispozici dostatečné množství dat, je možné využít algoritmy strojového učení. Vzhledem k existenci výstupního vektoru je vhodné použít algoritmy učení s učitelem. Dále je požadováno, aby výstupem byla numerická předpověď. Budou se tedy využívat regresní modely. Na obrázku 7 jsou zobrazeny hodnoty technických ztrát od května 2018 do srpna 2020. Jelikož nejsou data zatížena výraznými výkyvy nad rámec střídání ročních období, lze všechna data použít k trénování.



Obrázek 7: Technické ztráty od května 2018 do srpna 2020

Dále je na obrázku 8 vykresleno rozložení technických ztrát proložené normálním rozdělením, aby se dalo lépe odhadnout, jakých hodnot, s jakou pravděpodobností mohou technické ztráty nabývat. Je vidět, že technické ztráty nabývají téměř normálního rozdělení se střední hodnotou 141,04 a směrodatnou odchylkou 51,48.



Obrázek 8: Histogram zastoupení hodnot technických ztrát a jejich proložení normálním rozdělením

3.2 Replikace výsledků

V prvním kroku bylo potřeba ověřit nejlepší výsledek z případové studie. Poskytnutý kód proběhl s dodanými daty. V případové studii je zobrazena tabulka nejlepších dosažených výsledků. Ty byly vyhodnoceny pomocí MAE, MAPE a největší odchylky. Výsledky uvedené ve studii a dodané kódy s daty ale nepřinesly stejný výsledek. To je ukázáno v tabulce 1. U každého algoritmu byla odchylka MAE mezi výsledkem replikace a případové studie mezi 0,5 % a 11 %. K této odchylce došlo i u algoritmů s jednoznačným řešením, jako je např. Lineární regrese. Nelze tedy tuto chybu přičítat změně počátečních podmínek apod.. Důvodem odchylky je pravděpodobně jiné rozdělení trénovacích a testovacích dat nebo trénování na jiných datech, než je uvedeno v případové studii.

Algoritmus	MAE studie	MAE replikace	Odchylka
Lineární r.	10,41	9,73	6,53 %
Lasso r.	10,86	9,83	9,48 %
Ridge r.	10,07	9,52	5.46 %
ElasticNetCV	17,11	16,95	0.93 %
Stacking ens.1	9,97	9,46	5,11 %
Stacking ens.2	10,15	9,83	3.15 %
Extra stromy	10,44	9,44	9,57 %
Náhodné lesy	10,61	9,48	10,65 %
SVM(rbf)	11,17	10,7	4,20 %
SVM(poly.)	15,92	15,67	1,57 %
Rozhodovací s.	15,68	14,06	10,33 %
KNN(3)	18,16	17,90	1,43 %
KNN(8)	16,25	16,15	0,61 %
XGBoost r.	11,64	10,78	7,38 %
SGD regrese	10,65	10,01	6,00 %

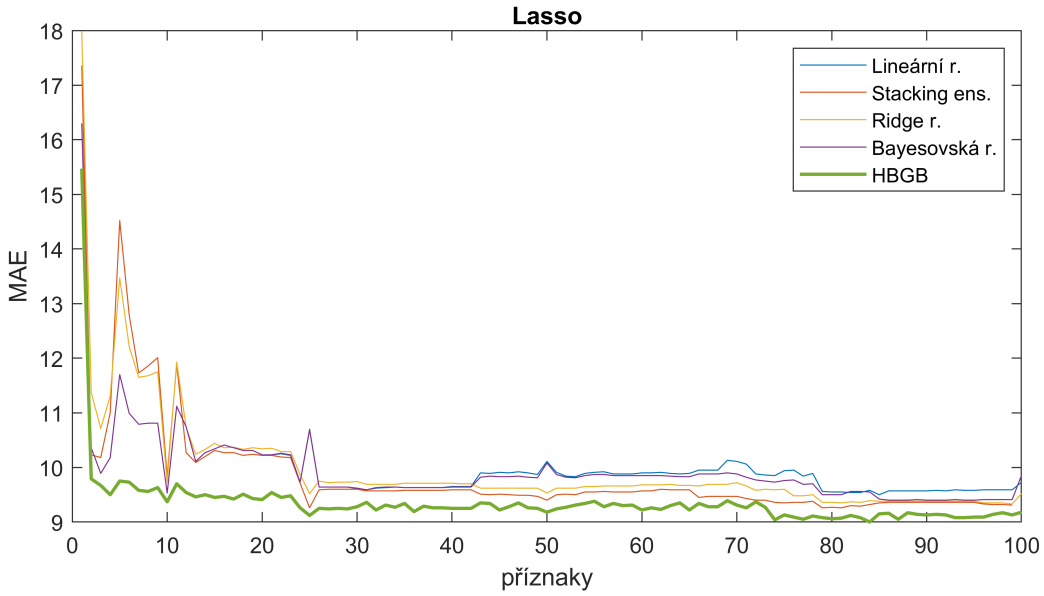
Tabulka 1: Porovnání MAE z případové studie a replikace

3.3 Algoritmy předzpracování dat

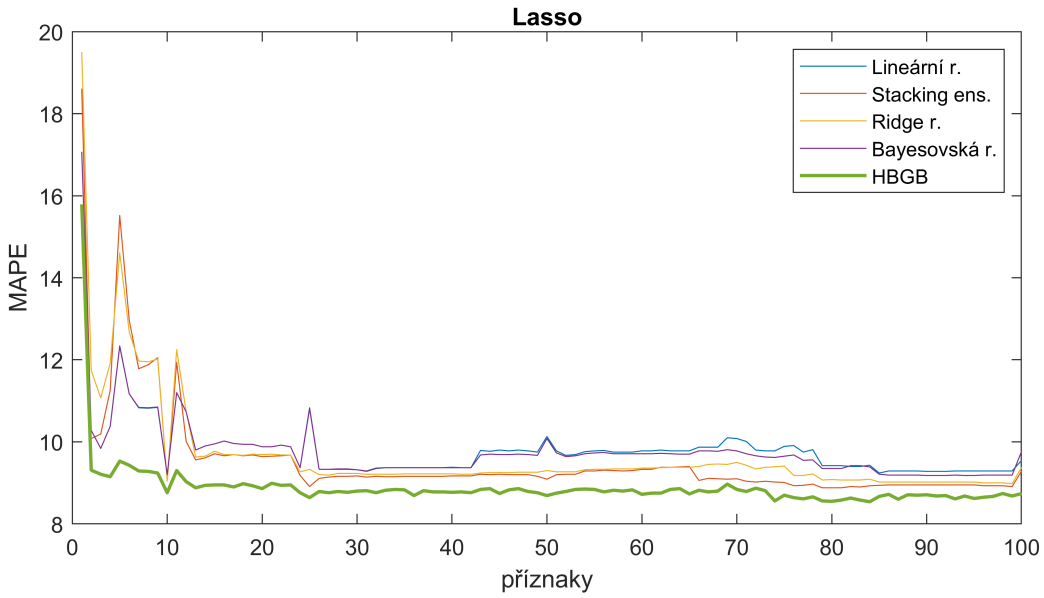
Během předzpracování dat bylo na základě korelace zjištěno, že z původních 107 příznaků, obsahuje 5 příznaků stejnou korelaci s výstupním vektorem. Kontrolou bylo zjištěno, že obsahují i stejné hodnoty. Protože obsahují redundantní informaci, nebyly dále uvažovány. Celkem je tedy k dispozici 102 různých příznaků. Po získání tohoto vektoru stavu je zapotřebí vhodně upravit data tak, aby je mohly algoritmy strojového učení dobře využít. Experimentálně bylo zjištěno, že nejlepšího výsledku se dosáhne při *z-score* normalizaci vstupních i výstupních dat. Dále byly testovány dva způsoby výběru příznaků, **Lasso algoritmem** a podle **nejvyšší korelace**.

3.3.1 Lasso

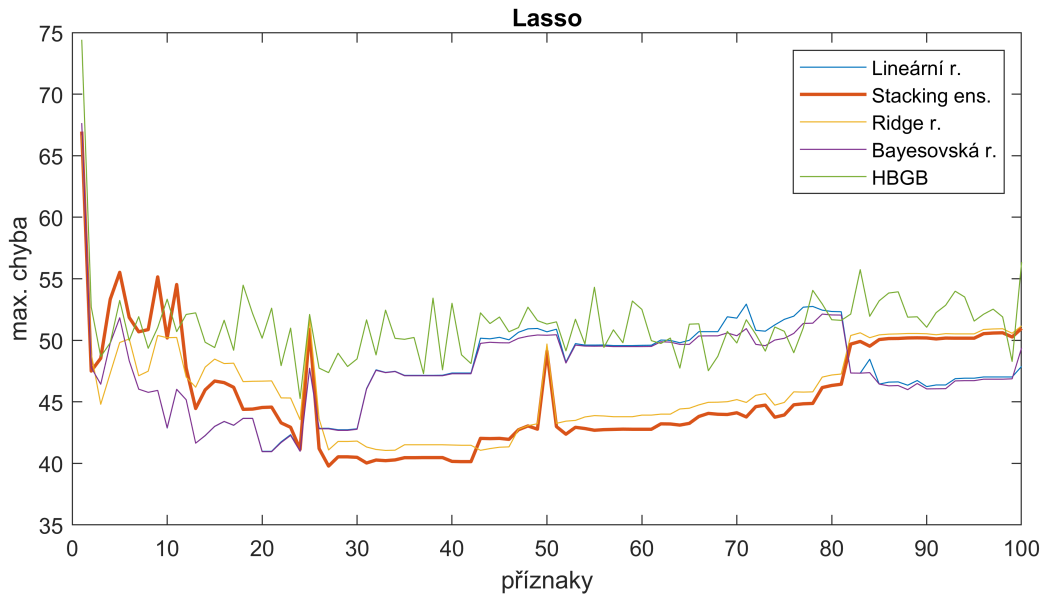
Lasso algoritmus umožňuje vybrání konkrétního počtu příznaků. Jelikož není dopředu jasné, s jakým počtem příznaků budou algoritmy strojového učení dosahovat nejlepších výsledků, je testováno celé rozpětí příznaků. Pro všechny podmnožiny počtu příznaků byly vyzkoušeny všechny zmiňované algoritmy strojového učení: Lineární regrese, Stacking Ensemble, Ridge regrese, Bayesovská regrese a Histogram-based gradient boosting regrese (HBGB).



Obrázek 9: MAE vzhledem k příznakům získaných Lasso algoritmem



Obrázek 10: MAPE vzhledem k příznakům získaných Lasso algoritmem



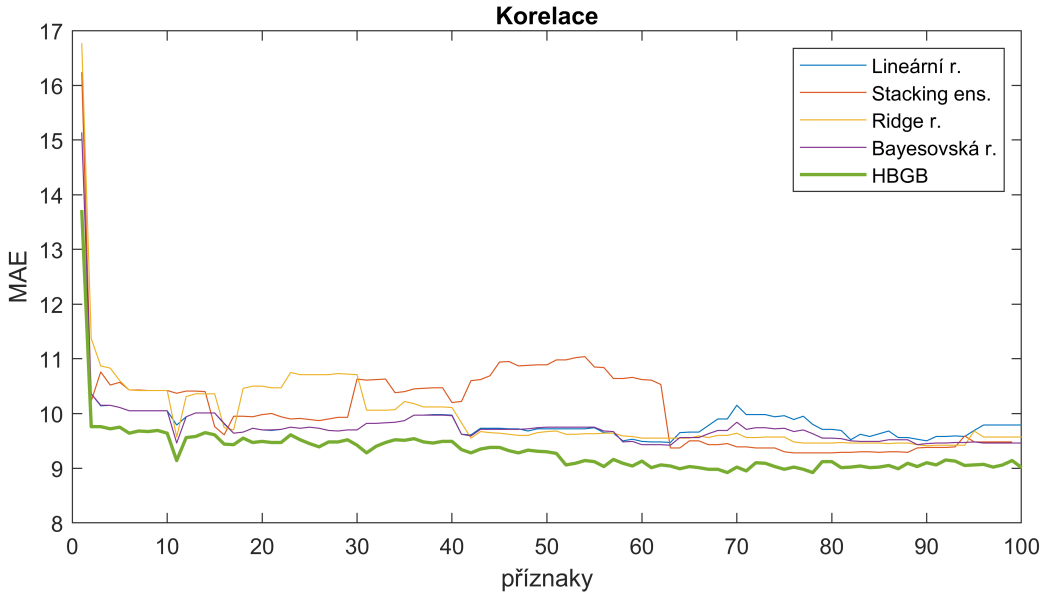
Obrázek 11: Maximální odchylka vzhledem k příznakům získaných Lasso algoritmem

Na obrázku 9 je zobrazena MAE pro všechny algoritmy s vybraným počtem příznaků. Je tedy vidět, že u všech skupin vybraných příznaků dosáhl algoritmus Histogram-based gradient boosting regrese nejlepšího výsledku. Dále se dá pozorovat, že nejmenší odchylky dosáhl algoritmus v rozmezí 75-85 vybraných příznaků. Nejmenší zaznamenaná hodnota je 9,00 u 84 vybraných příznaků. Na obrázku 10 je vidět graf MAPE opět pro všechny algoritmy. Tento graf má stejný průběh jako obrázek 9. Nejnižší zaznamenaná hodnota je 8,54 % znovu u 84 vybraných příznaků. Posledním měřeným kritériem kvality predikce je maximální odchylka predikovaných a skutečných ztrát. Nejnižší maximální odchylka byla dosažena algoritmem Stacking Ensemble s hodnotou 39,78 u 27 vybraných příznaků.

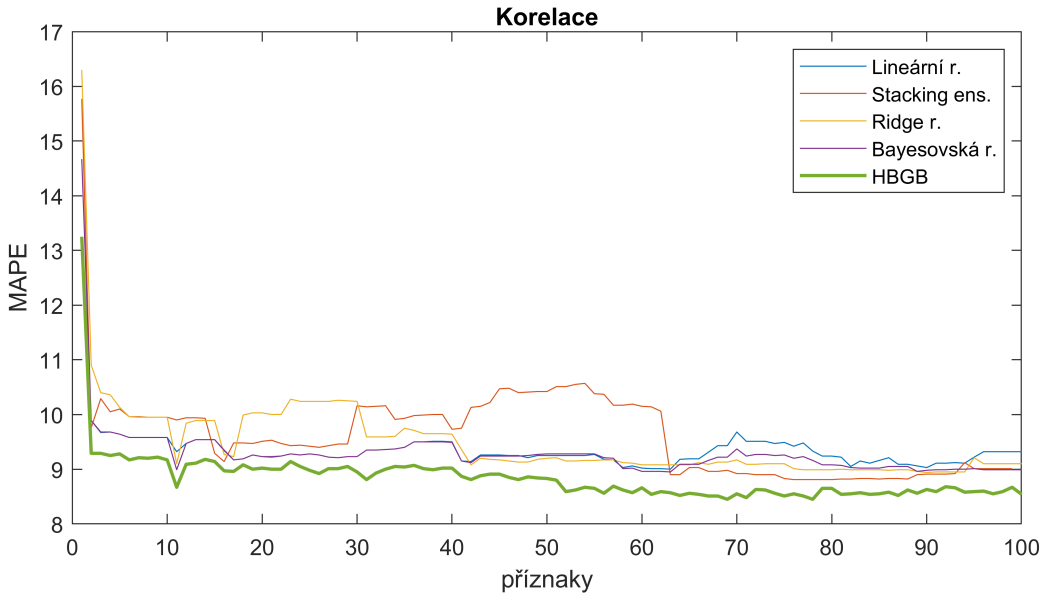
Na základě dosažených výsledků je pokládán ze nejlepší model získaný pomocí algoritmu je Histogram-based gradient boosting regrese, přestože nedosahuje nejnižší maximální odchylky ze všech testovaných algoritmů.

3.3.2 Nejlepší příznaky

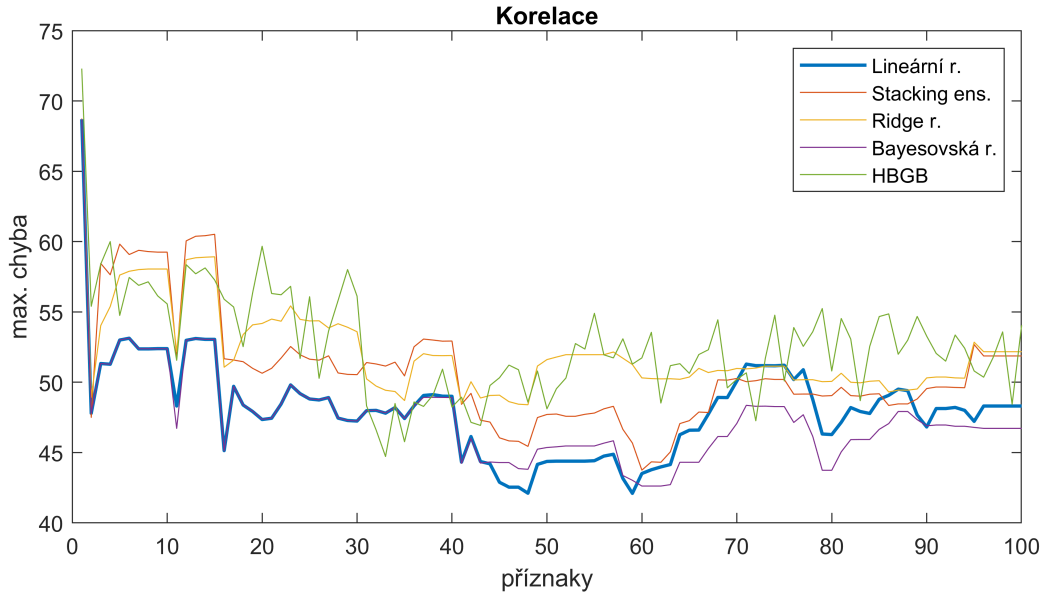
Další výběr příznaků proběhl na základě korelační analýzy. Nejprve se seřadily všechny příznaky dle výše korelace v absolutní hodnotě. Následně byly postupně vybírány ty, u kterých byla zjištěna nejvyšší hodnota. Korelace u 2 příznaků přesáhla hodnotu 0.8. Protože opět není předem jisté, jaký počet příznaků s nejvyšší korelací bude nejlepší, jsou znovu testovány všechny podmnožiny příznaků.



Obrázek 12: MAE vzhledem k příznakům získaných korelační analýzou



Obrázek 13: MAPE vzhledem k příznakům získaných korelační analýzou



Obrázek 14: Maximální odchylka vzhledem k příznakům získaných korelační analýzou

Na obrázku 12 je zobrazena metrika MAE pro všechny použité algoritmy. Je vidět, že algoritmus Histogram-based gradient boosting opět poskytuje nejlepší výsledky na celé šířce příznakových množin. Nejmenší odchylka je 8,92 u 69 a 78 vybraných příznaků. Na následujícím obrázku 13 je vidět graf MAPE znovu pro všechny algoritmy. Graf má znovu stejný průběh jako obrázek 12, pouze má posunutou osu y. Opět tedy algoritmus Histogram-based gradient boosting dosáhl nejlepšího výsledku. Už od dvou vybraných příznaků dosahuje tento algoritmus velmi dobrých výsledků. Nejnižší zaznamenaná hodnota je 8,45 % u 69 a 78 vybraných příznaků. Na obrázku 14 je zobrazena maximální odchylka použitých algoritmů vzhledem k vybraným příznakům korelační analýzou. V tomto případě dosáhl nejlepšího výsledku algoritmus Lineární regrese s hodnotou 41,1 u 59 vybraných příznaků.

Obě použité metody výběru příznaků, Lasso i podle nejvyšší korelace dosáhly s testovanými algoritmy podobných výsledků. Výběr podle nejvyšší korelace byl ale o 0,09 % úspěšnější.

3.4 Výpočetní algoritmy

Pro získání co nejlepších výsledků bylo otestováno několik různých trénovacích algoritmů. Více se v této práci zkoumají algoritmy: Lineární regrese, Ridge regrese, Stacking Ensemble, Bayesovká regrese a Histogram-based gradient boosting regrese.

3.4.1 Lineární regrese, Ridge regrese, Bayesovská regrese

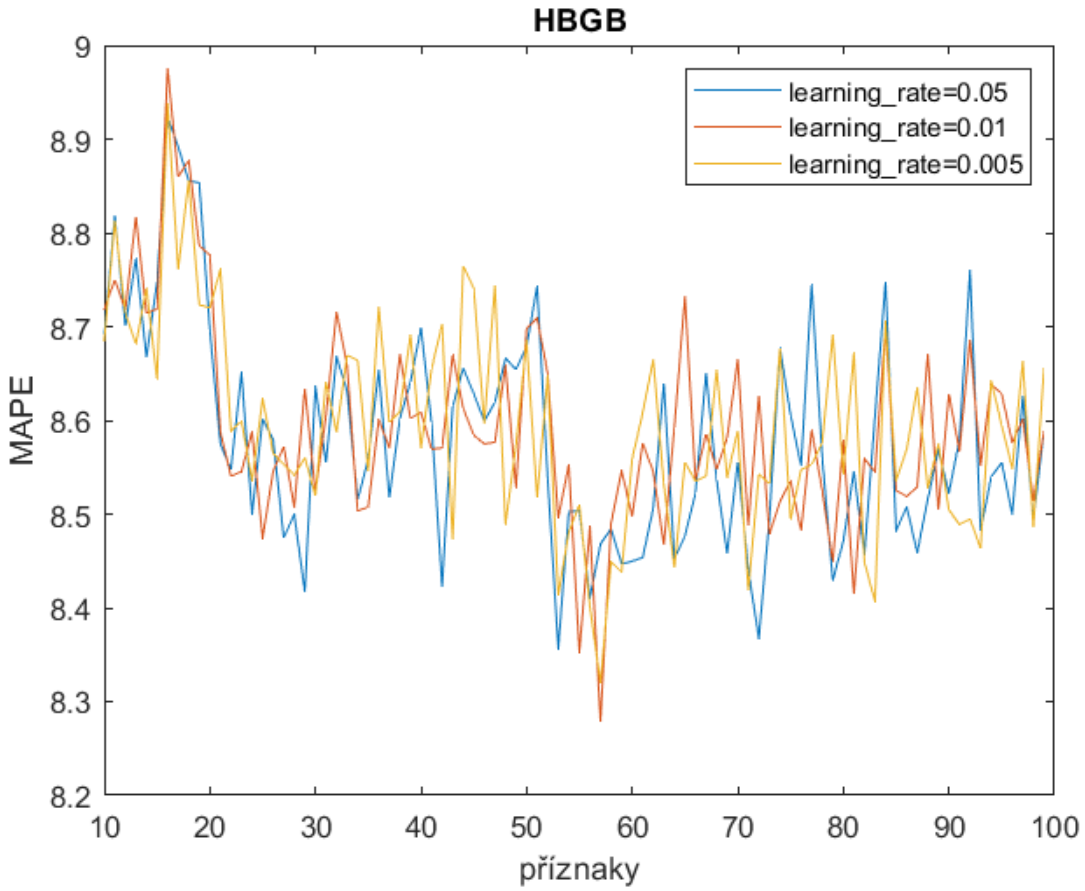
Tyto algoritmy vynikají efektivitou výpočtu a jednoduchostí interpretace modelu. Trénování jednotlivých algoritmů se všemi příznaky proběhlo do 2 sekund. I přes výše zmíněné výhody se nebudou tyto algoritmy dále zkoumat kvůli nedostatečné přesnosti dosažené v rámci této úlohy.

3.4.2 Stacking Ensemble

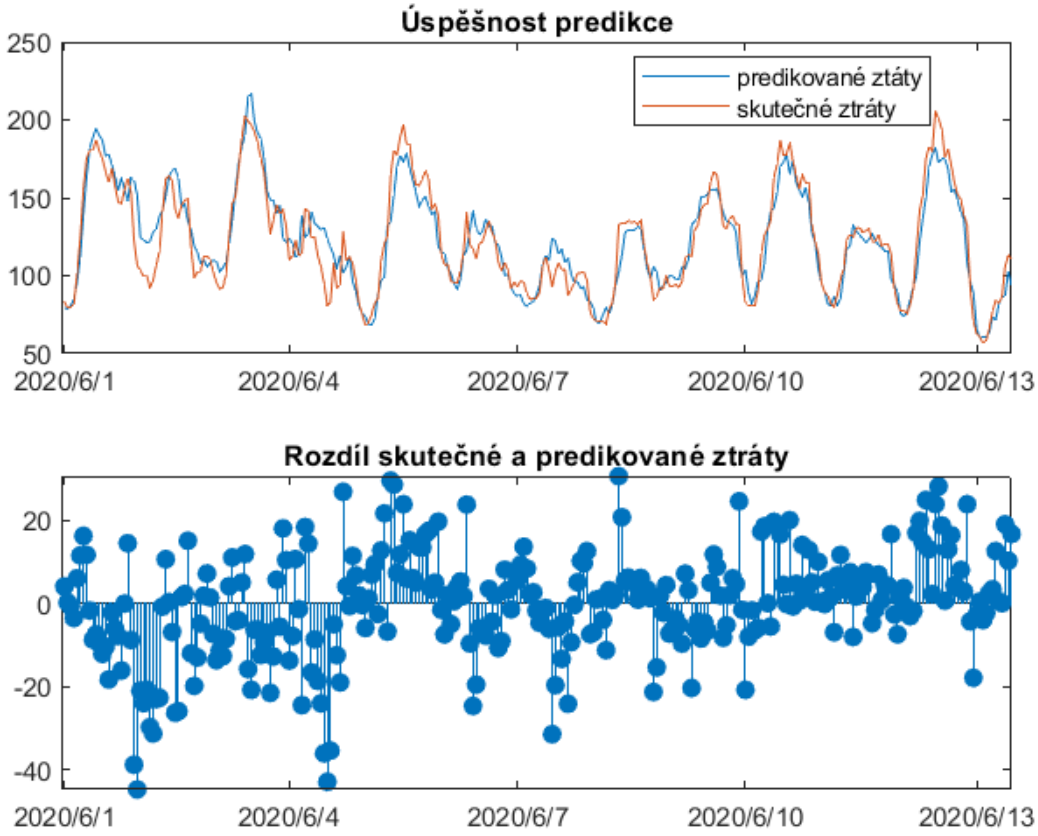
Tento algoritmus využívá více základních algoritmů pro trénování. Experimentálně bylo zjištěno, že nejlepších výsledků dosahuje kombinace Lineární regrese a Ridge regrese. Jako meta-algoritmus byla zvolena Lineární regrese. Ve srovnání s ostatními modely poskytoval tento algoritmus poměrně přesné předpovědi. Vzhledem k jeho složitosti ale trvá relativně dlouho jeho natrénování. Doba trénování se všemi příznaky přesahuje 15 sekund.

3.4.3 Histogram-based gradient boosting regrese

Použití tohoto algoritmu přineslo nejlepší výsledky, proto u něj proběhne optimalizace hyperparametrů. Nejdříve byla zvolena ztrátová funkce. Z možností *squared_error*, *least_squares*, *absolute_error*, *least_absolute_deviation* a *poisson* dopadla nejlépe *squared_error*. Dále bylo zjištěno, že pro tento algoritmus s upravenými parametry je lepší použít pro výběr příznaků lasso. Vzhledem k tomu, že počet vzorků je větší než 10 000, tak je defaultně povoleno předčasné zastavení trénování. Dále byl optimalizován parametr *learning_rate*, ten byl nakonec zvolen na hodnotu 0,01. Tento parametr určuje rychlost učení. Jako další parametr byl zkoumán *min_samples_leaf*. Jeho konečná hodnota byla nastavena na 20. Proces optimalizace hodnoty *learning_rate* je vykreslen na obrázku 15. Tento algoritmus dával ze všech testovaných nejlepší výsledky. Jeho natrénovaní se všemi příznaky trvalo kolem 4 sekund.



Obrázek 15: Porovnání výsledku HBGB algoritmu pro různé nastavení parametru *learning_rate*



Obrázek 16: Porovnání skutečné a predikované ztráty

Porovnání predikce tohoto modelu a skutečné ztráty je zobrazena na obrázku 16. Je vidět, že predikovaná ztráta dobře kopíruje trendy skutečné hodnoty. Oproti výsledkům uvedených v případové studii se podařilo přesnost predikce vylepšit o více než 1,5 %. Výsledky této práce jsou možné využít k přesnějšímu nakupování na burze, a tedy snížení nákladů spojených s obchodováním s elektřinou.

Pro další výzkum v této oblasti by se dalo využít složitějších modelů, které mají potenciál dosáhnout ještě lepších výsledků, či předefinování úlohy pro predikci ztrát v přenosové soustavě pro delší časový úsek.

4 Závěr

Tato bakalářská práce se zabývá tématem Využití metod strojového učení pro predikci technických ztrát v přenosové soustavě. Nejprve došlo k seznámení se s problematikou technických ztrát v elektroenergetické přenosové soustavě. Získané teoretické poznatky byly popsány v kapitole 2. Podkapitola 2.1 popisuje stav energetiky v České republice a přináší souhrnný přehled trhu s elektřinou a výroby a spotřeby elektrické energie na našem území. V podkapitole 2.2 byla představena společnost Česká elektroenergetická přenosová soustava, jejíž data byla využita v praktické části práce. Blíže byla popsána její historie, struktura, činnost a mezinárodní význam. Podkapitola 2.3 uvádí výčet metod pro předzpracování dat. V podkapitole jsou uvedeny popisy základních typů algoritmů strojového učení, kterými jsou učení s učitelem, učení bez učitele a zpětnovazební učení. V podkapitole 2.5 je přehled konkrétních algoritmů použitých v praktické části.

Kapitola 3 obsahuje praktickou část, ve které byla nejprve analyzována případová studie řešící návrh prediktorů technických ztrát v přenosové soustavě ČR s využitím metod strojového učení. Následně byla provedena replikace výsledků prezentovaných ve studii s použitím stejných prediktorů a zvoleného horizontu predikce v programovém prostředí Python 3 s využitím knihovny Scikit. Jelikož se jednalo pravděpodobně o jiný časový horizont testovaných dat, než byl použit ve studii, nebyly získané výsledky totožné.

V poslední podkapitole praktické části byly navrženy další metody strojového učení, a to tak, aby výsledný rozdíl predikovaných a reálných hodnot byl pro zvolený časový horizont co nejnižší. Během práce bylo vyzkoušeno několik variant algoritmů předzpracování dat a algoritmů strojového učení. Nejlepšího výsledku bylo dosaženo při výběru 57 příznaků Lasso algoritmem z normalizovaných dat a při výběru trénovacího algoritmu Histogram-based gradient boosting. Touto metodou bylo dosaženo výsledku: 8,78 ve střední absolutní chybě, 8,27 % ve střední absolutní procentuální chybě a 50,76 v největší absolutní chybě. Jedná se o výsledky, které jsou lepší, než jakých bylo dosaženo v případové studii společnosti ČEPS, a. s.. Vzhledem k vysokým cenám energií může i malé zlepšení v prediktivním modelu přinést velkou finanční úsporu, jelikož lepší odhad vývoje energetických ztrát umožní výhodnější včasný nákup energie.

5 Sezam použité literatury

[1] Mitchell, T., Machine Learning. 1997, New York: McGraw Hill

[2] Roční zpráva o provozu elektrizační soustavy České republiky za rok 2021, 2022 [online] Energetický regulační úřad, [cit. 10. 8. 2022]. Dostupné z: https://www.eru.cz/rocni-zprava-o-provozu-es-cr-pro-rok-2021

[3] Flášar, P., Fousek, J., Jícha, T. a další, Trh s elektřinou: úvod do liberalizované energetiky, 2016, Asociace energetických manažerů

[4] Salavec, J., Trh s elektřinou – specifika, účastníci trhu a rozdělení, 2017 [online], O energetice [cit. 11. 8. 2022]. Dostupné z: https://oenergetice.cz/trh-s-elektrinou/trh-s-elektrinou

[5] Očenášková, A., Trh s energiemi přehledně, 2021, [online] Aktuálně.cz, [cit. 11. 8. 2022]. Dostupné z: https://zpravy.aktualne.cz/ekonomika/trh-s-energiemi/r c1498712334411ecb02dac1f6b220ee8/

[6] ČEPS, a. s.. [online]. Dostupné z: https://www.ceps.cz/cs/

[7] Výroční zpráva ČEPS, a. s., 2022 [online] ČEPS, a. s., [cit. 6. 8. 2022]. Dostupné z: https://or.justice.cz/ias/ui/vypis-sl-detail?dokument=72525099subjektId=70456spis=77860

[8] Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data [online], Tableau, [cit. 14. 8. 2022]. Dostupné z: https://www.tableau.com/learn/articles/what-is-data-cleaning

[9] van der Maaten, L., Postma, E., van den Herik, J., Dimensionality Reduction: A Comparative Review, 2009 [online], [cit. 13.8.2022]. Dostupné z: https://members.loria.fr/moberger/Enseignement/AVR/Exposes/TR$_D$*imensiereductie.pdf*

[10] Gianluca M., Feature selection in machine learning using lasso regression, 2021 [online], [cit. 14.8.2022]. Dostupné z: https://towardsdatascience.com/feature-selection-in-machine-learning-using-lasso-regression-7809c7c2771a

[11] Normalization | Codecademy. Learn to Code - for Free | Codecademy [online], [cit. 15.08.2022]. Dostupné z: https://www.codecademy.com/article/normalization

[12] Koďousková, B., Co je strojové učení a jak souvisí s umělou inteligencí?, [online], [cit. 13.08.2022]. Dostupné z: https://www.rascasone.com/cs/blog/strojove-uceni-ml-metody-klasifikace

[13] Russell S., Norvig P., Artificial Intelligence: A Modern Approach, Prentice Hall, 2010, ISBN 9780136042594.

[14] van Otterlo, M., Wiering, M., Reinforcement Learning, 2012, ISBN 978-3-642-27644-6.

[15] Yan X., Linear Regression Analysis: Theory and Computing, 2009, [online], [cit. 14. 8. 2022]. Dostupné z: https://books.google.cz/books?id=MjNv6rGv8NICpg=PA1redir$_esc$ = $yv = onepageqf = false$

[16] Koehrsen, W., Introduction to Bayesian Linear Regression, 2018, [online], Towards Data Science, [cit. 10. 8. 2022]. Dostupné z: https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7

[17] Rocca, J., Ensemble methods: bagging, boosting and stacking, 2019, [online], Towards Data Science, [cit. 14.8.2022]. Dostupné z: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

[18] Dowdy, S. and Wearden, S., Statistics for Research, 1983, ISBN 0-471-08602-9