Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky

Soubory klasifikátorů pro predikci finanční výkonnosti regionů

Bc. Lukáš Samek

Diplomová práce
2017

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2016/2017**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Lukáš Samek**
Osobní číslo:	**E15686**
Studijní program:	**N6209 Systémové inženýrství a informatika**
Studijní obor:	**Informatika ve veřejné správě**
Název tématu:	**Soubory klasifikátorů pro predikci finanční výkonnosti regionů**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Cílem práce je charakterizovat přístupy k vytváření souborů klasifikátorů, popsat použitá data, navrhnout model pro predikci finanční výkonnosti regionů, verifikovat jej na předzpracovaných datech a provést statistické porovnání vybraných algoritmů.

Osnova:
- Koncept souborů klasifikátorů a přehled přístupů k jejich vytváření.
- Charakteristika a předzpracování dat.
- Návrh modelu pro predikci finanční výkonnosti regionů.
- Výsledky experimentů a statistické porovnání vybraných algoritmů.

Rozsah grafických prací:

Rozsah pracovní zprávy: **cca 60 stran**

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

HAYKIN, S. Neural Networks: A Comprehensive Foundation. 2nd edition, New Jersey : Prentice Hall, 1999.
KUNCHEVA, L. I. Fuzzy Classifier Design. A Springer Verlag Company, Germany, 2000. ISBN 80-903024-9.
OLEJ V. Modelovanie ekonomických procesov na báze výpočtovej inteligencie. Miloš Vognar - M&V, Hradec Králové, 2003. ISBN 80-903024-9-1.
WITTEN, I.H., FRANK, E., HALL, M.A. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, Amsterdam, 2011.
GAILLARD, N. The Determinants of Moody's Sub-sovereign Ratings. International Research Journal of Finance and Economics, 2009, roč. 31, č. 1, s. 194-209.

Vedoucí diplomové práce: **doc. Ing. Petr Hájek, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **4. září 2016**

Termín odevzdání diplomové práce: **28. dubna 2017**

L.S.

doc. Ing. Romana Provazníková, Ph.D.
děkanka

doc. Ing. Pavel Petr, Ph.D.
vedoucí ústavu

V Pardubicích dne 4. září 2016

PROHLÁŠENÍ

Prohlašuji, že jsem tuto práci vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako Školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně Univerzity Pardubice.

V Pardubicích dne 30.4.2016 Bc. Lukáš Samek

PODĚKOVÁNÍ:

Tímto bych rád poděkoval svému vedoucímu práce doc. Ing. Petru Hájkovi, Ph.D. za jeho odbornou pomoc, cenné rady a poskytnuté materiály, které mi pomohly při zpracování diplomové práce.

ANOTACE

Tato diplomová práce se zabývá predikcí finanční výkonnosti regionů. První část se zabývá tématem ratingu a pojmů souvisejícís ním. Další část práce se zaobírá použitými soubory klasifikátorů strojového učení a jejich nastavením. V závěru práce jsou porovnávány výsledky souborů klasifikátorů a je doporučen klasifikátor použitelný pro predikci.

KLÍČOVÁ SLOVA

rating, strojové učení, finanční výkonnost regionů, meta-učení

TITLE

Files classifiers to predict financial performance of the regions

ANNOTATION

This diploma thesis deals with prediction of financial performance of regions. The first part deals with the topic of evaluation and the concepts related to it. Another part of the thesis deals with used sets of machine learning classifiers and their settings. At the end of the thesis, the results of the classifiers are compared and predictor classifier is recommended.

KEYWORDS

Rating, machine learning, financial performance of regions, meta-learning

OBSAH

SEZNAM TABULEK

SEZNAM OBRÁZKŮ

SEZNAM GRAFŮ

SEZNAM ZKRATEK

BCA	Baseline Credit Asessment
EF	Economic Fundamental
HDP	Gross Domestic Product
IF	Institutional Framework
FP&DP	Financial Performance and Debt Profile
GM	Governance and Management
NN	Neural Nets (neuronová síť)
MLP	Multi Layer Perceptron
SVM	Support Vector Machine
ROC	Receiver Operating Characteristic Curve
ARFF	Attribute-Relation File Format
SMOTE	Synthetic Minority Oversampling Technique
SMO	Sequential Minimal Optimization
WEKA	Waikato Enviroment for Knowledge Analysis

ÚVOD

Predikce finanční výkonnosti regionů a obcí je důležitým aspektem, který může být nápomocen při sestavování finančního plánu, získávání externích zdrojů a rozhodování investorů a bank. Jedná o nástroj, s jehož pomocí investor či banka individuálně posuzuje rizika, která mohou v budoucnu nastat a ohrozit vložené investice. V případě sestavování finančního plánu tato znalost snižuje riziko a zvyšuje důvěryhodnost dané obce nebo regionu, což bude mít za následek možnost získání většího množství externích zdrojů např. z Evropské unie.

Pro hodnocení finanční výkonnosti regionů a obcí (dále jen regionů) je výhodnější použít metodu ratingu, jejíž předností je komplexní hodnocení všech oblastí (příjmů, výdajů, zadluženosti atd.). Jejím nedostatkem je časová a finanční nákladnost na provedení. Proto jsou používané metody strojového učení, jejichž cílem je automaticky klasifikovat regiony do ratingových tříd na základě historických dat poskytovaných ratingovými agenturami. Podobnou studií zabývající se predikcí ratingu je [17], kde bylo využito tzv. probit modelu, jenž je statistický aparát pro odhad ratingových tříd s určitou pravděpodobností. V tomto případě bylo ratingové hodnocení nahrazeno čísly 1 - 20 a následně bylo použito regresního probit modelu pro odhad hodnocení na testovacích datech. V případě této diplomové práce je využito pro predikci finanční výkonnosti souborů klasifikátorů, které simulují rozhodování skupiny expertů provádějících ratingové hodnocení. Dle [34] se tyto soubory klasifikátorů osvědčily již v minulosti při predikci ratingu podniků.

Cílem práce je navrhnout vhodný predikční model založený na souborech klasifikátorů, tento model použít pro predikci finanční výkonnosti (ratingu) regionů a výsledky statisticky porovnat.

Celá diplomová práce je rozdělena do šesti kapitol. Záměrem první kapitoly je vysvětlení pojmu finanční analýzy hospodaření a vymezení hlavních ukazatelů pro stanovení finančního stavu regionů. V rámci druhé kapitoly je definován rating jako jedna z možných metod pro hodnocení finanční výkonnosti regionů . Další část této kapitoly představují ratingové agentury, které se již několik let zabývají stanovováním finanční výkonnosti regionů a obcí. Mezi tyto popisované společnosti patří Moody´s, Standard & Poor´s a Fitch Ratings. V závěru této kapitoly je detailně popsáno výsledné stanovení ratingu regionů a obcí. Důraz bude kladen na postup stanovení základního úvěrového hodnocení (BCA) a pravděpodobnosti mimořádné podpory ze strany jiného subjektu.

Třetí a čtvrtá kapitola je věnována popisu metod strojového učení a meta-učení, které budou následně použity pro klasifikaci. Další část třetí kapitoly je věnována zjištění úspěšnosti souborů klasifikátorů, kde budou popsány možnosti pro přesnost klasifikace modelu jako je matice záměn pro více tříd, nákladová matice a křivka ROC. Pátá kapitola se zabývá charakteristikou a předzpracováním dat. Po dokončení fáze předzpracování dat budou navrhnuty modely pro realizaci klasifikace. Výsledky získané z těchto měření budou následně rozebrány a statisticky porovnány v kapitole šesté. V závěru práce dojde k shrnutí dosažených výsledků na získaných datech a navrhnutí klasifikátoru pro predikci finanční výkonnosti regionů.

1. FINANČNÍ VÝKONNOST REGIONŮ

Při zjišťování finanční situace regionu je prvně nutné vypracovat analýzu jeho hospodaření, kde by měla být zhodnocena minulost, přítomnost a očekávaná budoucnost. V první části kapitoly je vysvětlen pojem finanční analýza a v další části jsou popsány ukazatele, které slouží ke zjištění finančního stavu regionu.

1.1. Finanční analýza

Finanční analýza [21] je standardní nástroj, který se používá k posouzení finanční situace jakéhokoli subjektu. Poskytuje souhrnný i detailní pohled na dosavadní vývoj a na to, jaké předpoklady vytváří pro vývoj budoucí. Dále informuje o existujících trendech a možných slabých místech, která by se mohla stát překážkou zdárnému vývoji. Je tedy důležitým podkladem pro rozhodování představitelů obce, kde jde o rozhodování v širokém slova smyslu, od přípravy rozpočtu až přes úvahu o únosné výši dluhu. Tedy o přijetí takového dluhu, který nenaruší stabilitu obecních financí a výrazně nezúží prostor pro takové výdaje, které jsou nezbytné pro chod obce a poskytování veřejných statků obyvatelům [27].

Zdrojové informace pro finanční analýzu lze členit na externí a interní. V případě interních informací se jedná o data, která se týkají dané obce, vznikají na základě jejich aktivit a bývají evidovány přímo v obci. Tato data poskytuje finanční účetnictví prostřednictvím účetních výkazů. Účetními výkazy jsou rozvaha, výkaz zisků a ztrát, přehled o finančních tocích nebo výkaz o změnách ve vlastním kapitálu. Externí zdroje informací pochází z vnějšího ekonomického prostředí a jedná se o data z Českého statistického úřadu nebo z portálu Ministerstva financí.

Ukazatele finanční analýzy lze dělit na absolutní (extenzivní) a relativní (intenzivní) [21]. Absolutní ukazatele lze dále rozdělit na tokové, které jsou zkoumány za určité časové období (např. výsledovka) a na stavové vypovídající o stavu k určitému datu (např. rozvaha). Nejvýznamnější jsou ukazatele poměrové, které umožňují srovnání v čase a prostoru (jsou založeny na podílech dvou stavových či tokových ukazatelů např. ukazatel likvidity). Přehled základních ukazatelů finanční analýzy je zobrazen v Tabulce č.1.

Na základě rozdělených ukazatelů budou představeny základní metody finanční analýzy. První metodou je analýza stavových ukazatelů, kam patří analýza trendů a procentní analýza. Analýza trendů přebírá data přímo z rozvahy a výkazu zisků a ztrát. Sleduje změny absolutních i relativních hodnot vykazovaných dat v čase. Procentní analýza se zabývá strukturou aktiv a pasiv, které udávají složení hospodářských prostředků. Druhou metodou

je analýza rozdílových a takových ukazatelů, které obsahují analýzu fondu finančních prostředků a analýzu cash-flow. Do poslední analýzy poměrových ukazatelů spadá např. analýza zadluženosti či likvidity [21].

Tabulka 1: Ukazatele finanční analýzy

DÍLČÍ UKAZATELE	
Absolutní (extenzivní)	**Relativní (intenzivní, poměrové a podílové)**
• bazální (základní) • rozdílové • marginální (přírůstkové)	• prostý poměr dvou absolutních ukazatelů • podíl různých hodnot absolutního ukazatele - index bazický - index řetězový • marginální (relativní přírůstkové) • senzitivity (citlivost), tj. poměr relativních marginálních ukazatelů

Zdroj: upraveno podle [29]

1.2. Ukazatele finančního stavu regionu

Níže jsou popsány ekonomické ukazatele, vypovídající o finančním stavu regionu. Obecně by ukazatele měly být srozumitelné, jednoznačné a splňovat významovou funkci pro dosažení zvoleného cíle. Ukazatele se kterými je možné provádět finanční analýzu je celá řada, např. cash-flow nebo ukazatele rentability, ale tyto ukazatele se spíše hodí pro subjekty generující zisky. Jelikož se diplomová práce zabývá tématikou veřejné správy, budou vybrány a popsány ukazatele z oblastí rozpočtového hospodaření. Velmi často bývají využívány ukazatele podílu na obyvatele. Ukazatele jsou systematicky rozděleny dle [16],[39] do několika skupin a to do skupiny příjmy a výdaje, dluh, likvidita a ostatní ukazatele.

1.2.1. Ukazatele na bázi příjmů

Tyto ukazatele patří do skupiny tzv. rozpočtových ukazatelů a jsou zaměřeny na příjmy obce. Pro příjmy v rozpočtu obce jsou srovnávány ukazatele celkových příjmů na obyvatele, daňových příjmů a vlastních příjmů dle [15],[16]

1.2.1.1. Celkové příjmy na obyvatele

Jedná se o jednoduchý základní ukazatel vyjadřující výši zdrojů regionu. Omezením ukazatele celkových příjmů je nezohlednění kapitálových příjmů, investičních transferů

a zaměření na počet obyvatel, nikoli na rozlohu regionu. Výsledky tohoto ukazatele mohou regiony ovlivnit jen omezeně a to pouze malou část daňových příjmů (místní poplatky a svěřené daně), nedaňové a kapitálové příjmy.

1.2.1.2. Ukazatele daňových příjmů

- Daňové příjmy na obyvatele;
- Podíl daňových příjmů na celkových příjmech.

Ukazatel daňových příjmů vyjadřuje výši příjmů z daní, místních, správních a ostatních poplatků. Výše těchto příjmů je v čase stabilní (pokud se nemění příslušná legislativa) a region má minimální možnosti ovlivnit tuto výši příjmů.

1.2.1.3. Ukazatele vlastních příjmů

- Podíl vlastních příjmů na celkových příjmech;
- Vlastní příjmy na obyvatele.

Ukazatele vlastních příjmů patří mezi relativní ukazatele a umožňují porovnat výši vlastních příjmů mezi obcemi i mezi různými roky. Ukazatel podílu vlastních příjmů na celkových příjmech poskytuje informaci o příjmech regionu, které jsou stabilní. Vyjadřuje, do jaké míry je region finančně nezávislý na dotacích neboli čím je procento finanční nezávislosti vyšší, tím více je region nezávislý na dotacích.

V případě ukazatele vlastních příjmů na obyvatele dochází k omezení vyplývající ze zařazení kapitálových příjmů do výpočtu, které v jednotlivých regionech nebývají každoročně ve stabilní výši.

1.2.2. Ukazatele na bázi výdajů

Tyto ukazatele patří do skupiny tzv. rozpočtových ukazatelů a jsou zaměřeny na výdaje regionu. Pro výdaje v rozpočtu regionu jsou srovnávány ukazatele celkových výdajů na obyvatele a běžných i kapitálových výdajů na obyvatele na základě [16],[39].

1.2.2.1. Celkové výdaje na obyvatele

Ukazatel celkových výdajů na obyvatele zahrnuje výši běžných i kapitálových výdajů regionu v přepočtu na jednoho obyvatele. Tento ukazatel umožňuje základní srovnání mezi regiony i mezi roky. Při srovnání mezi regiony se bere v úvahu velikost regionu nebo výše jednorázových výdajů.

1.2.2.2. Běžné a kapitálové výdaje na obyvatele

Poměrový ukazatel celkových výdajů na obyvatele lze dále rozčlenit na běžné a kapitálové výdaje a jejich poměr na jednoho obyvatele regionu. Kapitálové výdaje umožňují srovnání investic regionu v čase i s jinými regiony. Mezi ukazatele běžných výdajů patří výdaje, které se každoročně opakují.

1.2.3. Ukazatele saldo rozpočtu

Popis ukazatelů saldo rozpočtu je založen na studiích [16],[35],[39].

1.2.3.1. Saldo rozpočtu

Vyjadřuje rozdíl mezi rozpočtovými příjmy a výdaji. Pokud jsou naplánované příjmy vyšší než výdaje, je saldo kladné. Kladné saldo informuje, že v rozpočtu je přebytek příjmů. Tyto prostředky mohou být použity na splácení úvěru z minulosti nebo k vytvoření finanční rezervy. Záporné saldo znamená, že v rozpočtu je přebytek výdajů. Chybějící prostředky pocházejí buď z úvěrů (např. na výstavbu) nebo obec využije finanční prostředky uspořené v minulosti. Saldo rozpočtu je jedním z podkladů pro posouzení hospodaření regionů, ale pro analýzu finanční kondice se moc nehodí, jelikož splácení úvěrů je hrazeno výhradně z provozního rozpočtu.

1.2.3.2. Saldo provozního rozpočtu a index provozních úspor

Dalším typem přebytku je saldo provozního rozpočtu vyjádřené jako rozdíl běžných příjmů a běžných výdajů. Provozní přebytek by měl nabývat kladných hodnot, neměl by v čase klesat a výsledek je pro regiony důležitější než saldo rozpočtu. Záporné provozní saldo může vyjadřovat situaci, kdy region nemá dostatek pravidelných příjmů na úhradu samotného provozu. Oproti tomu kladné saldo je použito ve výpočtu indexu provozních úspor. Jedná se o poměr salda provozního rozpočtu k běžným příjmům. Hodnota by se optimálně měla pohybovat okolo 20 %, ale minimální hodnota by neměla klesnou pod 10 % [25].

1.2.4. Ukazatele zadluženosti

Tyto ukazatele nás informují o tom, jakou měrou regiony využívají ke své činnosti dluhy (závazky). V dnešní době již není možné nalézt regiony využívající pouze vlastní kapitál. Ukazatele zadluženosti předchází tomu, aby se regiony neadekvátně zadlužovaly a přicházely tak o potencionální věřitele. Popis ukazatelů zadluženosti je založen na studiích [15],[16],[39].

1.2.4.1. Výše dluhu na obyvatele

Tento ukazatel uvádí, jak velký dluh spadá na jednoho obyvatele regionu a vypočítá se jako podíl celkového dluhu regionu na počtu obyvatel. Vyšší zadluženost na obyvatele nemusí být sama o sobě špatným signálem, pokud region disponuje vyšším provozním saldem na obyvatele a dokáže lépe generovat zdroje na případné splátky.

1.2.4.2. Výše dluhu k dlouhodobému majetku

Ukazatel výše dluhu k dlouhodobému majetku vypovídá o míře krytí dluhů regionu dlouhodobým majetkem a je vhodný pro srovnání v čase a mezi regiony navzájem [15].

1.2.4.3. Výše dluhu k celkovým příjmům

Ukazatel nabízí jednoduché poměření výše zadlužení regionu k velikosti jeho rozpočtu (jeho příjmům). Ukazatel je opět vhodný zejména pro srovnání v čase a mezi regiony navzájem.

1.2.4.4. Výše dluhu k saldu běžného rozpočtu

Ukazatel výše dluhu k saldu běžného rozpočtu vyjadřuje, za kolik let je celkový dluh splatný ze salda běžného rozpočtu regionu. Jelikož je známo zatížení budoucích rozpočtů skrze dluh, je tento ukazatel vhodný zejména pro plánování budoucích výdajů. Kapitálový rozpočet zde není zahrnut z důvodu, že jeho výnosy by měly sloužit spíše na investice.

1.2.5. Ukazatele likvidity

Ukazatele likvidity měří, jak je region schopen uspokojit své krátkodobé závazky v případě vzniku neočekávaných problémů. Jedná se tedy o souhrn všech potencionálních prostředků, které má region k dispozici pro úhradu svých splatných závazků. Likviditu je možné hodnotit dle ukazatele na běžné, pohotové či okamžité likvidity dle [54].

1.2.5.1. Běžná likvidita

"Snižuje krátkodobá aktiva o zásoby, tedy o krátkodobé aktivum, které generuje peníze v případě, že nedojde k jeho prodeji"[1]. Vypočítá se jako podíl krátkodobých závazků na oběžných aktivech. Výpočtem dojde ke zjištění, kolikrát je region schopen uspokojit pohledávky věřitelů v případě, že promění všechna svá oběžná aktiva na peněžní prostředky [54].

[1] převzato z : http://www.faf.cz/Likvidita/Bezna-likvidita.htm

1.2.5.2. Okamžitá likvidita

Okamžitá likvidita vyjadřuje okamžitou schopnost regionu uhradit své krátkodobé závazky a spočítá se jako podíl krátkodobého finančního majetku ke krátkodobým závazkům. Výsledná doporučená hodnota by se měla pohybovat v intervalu <0,2;0,5> dle [54].

2. RATING

Tato kapitola se zabývá popisem jedné z možných metod pro hodnocení finanční výkonnosti regionů. V prvních podkapitolách je vysvětlen obecně pojem rating a v poslední podkapitole je popsán postup stanovení ratingu regionu.

2.1. Pojem rating a jeho využití

Rating je dle [50] nezávislé hodnocení, jehož cílem je zjistit, a to na základě komplexního rozboru veškerých známých rizik hodnoceného subjektu, jak je tento subjekt schopen a ochoten dostát včas a v plné výši všem svým splatným závazkům.

Každý subjekt dostane od agentury hodnocení podle stupnice, která se značí písmeny, kde písmeno A značí dobrou investiční pozici a písmeno C vysoce rizikovou pozici. Při sestavování ratingové stupnice agentury spolupracují s centrální bankou, ministerstvy a vládními agenturami. Ratingové hodnocení slouží především pro investory, investiční banky, makléře a státní instituce. V oblasti investorů se používá toto hodnocení k rozšíření investičních možností a poskytují nezávislé a snadno použitelné hodnocení úvěrového rizika. Investiční banky a makléři používají ratingové hodnocení při výpočtu rizika svého portfolia[2].

Od ratingu se často odvíjí úročení půjčky. Čím horší rating, tím vyšší úroky jsou při poskytování půjčky požadovány. Nižší hodnota ratingu vždy neznamená doporučení pro investici. Kdo má raději vyšší riziko s vyšším výnosem, může zvolit investici do firmy s nižším ratingovým hodnocením.

2.2. Historie ratingu

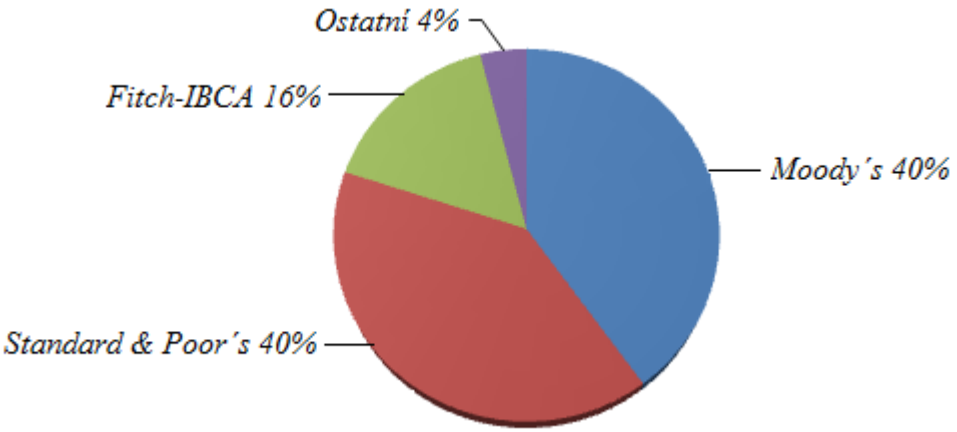
Počátek ratingu je mapován od roku 1909, kdy John Moody poprvé hodnotil obligace železničních společností. Následně se začal tento typ hodnocení používat i pro dluhopisy veřejných podniků a průmyslových společností. Mezi první společnosti, které vydaly své veřejné ratingy, patřily Poor´s Publishing Company a Standart Statistics Company. Toto odvětví se postupem času vyvíjelo a dnes existují tři velké vedoucí agentury zabývající se touto tématikou [32].

[2]Pojem portfolio = rozdělení investičních zdrojů s cílem snížení rizika. Používá se ve vztahu k fyzickým osobám jako ukázka skladby cenných papírů ve svém držení.

2.3. Hodnotitelé ratingu

Mezi tři přední ratingové agentury uznávané po celém světě patří *Standard & Poor´s*, *Moody´s a Fitch-IBCA* dle [1],[32]. Jedná se o agentury ze Spojených států amerických, které působí po celém světě a pro které je velmi důležitá nezávislost a důvěryhodnost.

Tržní podíly ratingových agentur jsou znázorněny na grafu viz Obrázek 1.



Obrázek 1: Tržní podíly ratingových agentur

Zdroj: vlastní zpracování podle [50]

2.3.1. Moody´s

Agentura byla založena v roce 1914 Johnem Moodym a mezi ostatními agenturami má přibližně 40 % tržní podíl. Kromě ratingů provádí ekonomické průzkumy a finanční analýzy komerčních i státních subjektů a poskytuje software pro řízení rizik finančních institucí. Společnost má přibližně 4000 zaměstnanců v 27 státech.

2.3.2. Standard & Poor´s

Společnost vznikla v roce 1941 fúzí Standard Statistics Company a Poor's Publishing Company. Zaměřuje se na poskytování mnoha finančních služeb. Kromě ratingů a jiného hodnocení provádí vlastní ekonomické průzkumy, vytváří několik S &P indexů a je jedním z předních světových poskytovatelů nezávislých informací o investicích. Své pobočky má v 23 zemích a na trhu ratingu má přibližně 40 % podílu.

2.3.3. Fitch Ratings

Společnost založil v roce 1913 John Knowles Fitch a dnes je Fitch Ratings jednou ze tří částí finanční společnosti Fitch Group. Fitch Ratings je mezinárodní ratingová agentura, Fitch Solutions je firma poskytující poradenství a služby finančnímu sektoru

a Algorithmics Inc. se zabývá softwarem pro řízení rizik. Fitch Ratings má na trhu menší podíl než předchozí agentury - přibližně 16 %.

2.3.4. Stupnice ratingových agentur

Stupně AAA (nejvyššího ratingu) dosahují subjekty nejvyšší kvality. V případě států a regionů jsou to ty s největší vyspělostí, které nabízejí investorům rostoucí ekonomiku s nízkým zadlužením. Oproti tomu stupeň C (někdy se používá D) je pro země a regiony s vysokou zadlužeností, jejichž ekonomika má závažné problémy a celkovou platební neschopnost viz Tabulka 2. Nejčastěji se v případě tohoto hodnocení hovoří o úvěrovém ratingu, který vyjadřuje důvěryhodnost dlužníka nebo cenného papíru a ten udělují ratingové agentury. Podrobnější popis ratingové stupnice je možné nalézt v [50].

Tabulka 2: Hodnocení jednotlivých ratingových agentur v dlouhém a krátkém období

Moody´s	S &P	Fitch	Hodnocení
Dlouhé období	Dlouhé období	Dlouhé období	
Aaa	AAA	AAA	Nejvyšší kvalita
Aa1	AA+	AA+	Velmi kvalitní
Aa2	AA	AA	
Aa3	AA-	AA-	
A1	A +	A +	Střední kvalita - vyšší
A2	A	A	
A3	A -	A -	
Baa1	BBB+	BBB+	Střední kvalita - nižší
Baa2	BBB	BBB	
Baa3	BBB-	BBB-	
Ba1	BB+	BB+	Spekulativní
Ba2	BB	BB	
Ba3	BB-	BB-	
B1	B +	B +	Vysoce spekulativní
B2	B	B	
B3	B -	B -	
Caa1	CCC+	CCC+	Značná rizika
Caa2	CCC	CCC	Extrémně spekulativní
Caa3	CCC-	CCC-	S velmi nízkou perspektivou
Ca	CC	CC	
C	C	C	Velmi vysoká pravděpodobnost úpadku
	CI	D	
	D		

Zdroj: upraveno podle [4]

2.4. Druhy ratingu

Rating lze členit podle trhu, pro který je určen, podle dluhového instrumentu, se kterým je hodnocení spojené a v neposlední řadě z časového hlediska. V dnešní době je prováděno ratingové hodnocení zemí, regionů, států, bank, penzijních fondů a pojišťoven.

Dělení ratingu z hlediska času dle [50]:

- **Krátkodobý rating** - souvisí obvykle s hodnocením krátkodobých dluhů, tedy závazků se splatností do jednoho roku(např. směnky nebo aktuálně splatné dlouhodobé dluhy).
- **Dlouhodobý rating** - do hodnocení spadají instrumenty s dobou splatnosti delší než jeden rok.

Dělení podle předmětu hodnocení:

- **Rating emise** - vztahuje se ke konkrétnímu cennému papíru, který byl emitován[3].
- **Rating respondenta** - týká se společností nebo států, které cenné papíry vydaly.

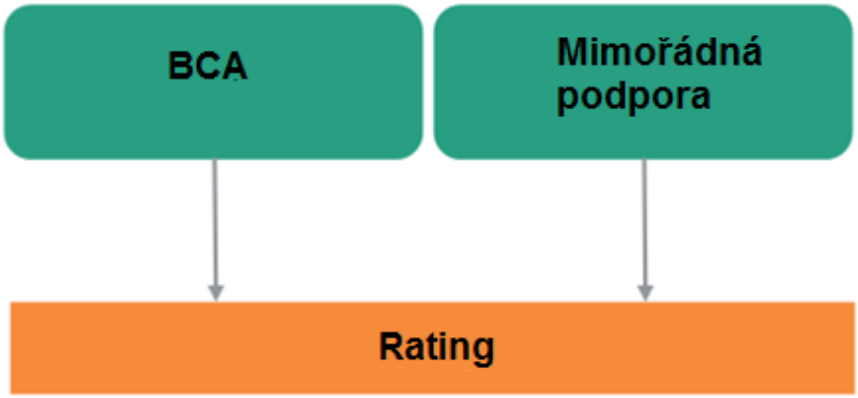
Dělení dle měnového hlediska:

- **Mezinárodní** - týká se závazků emitovaných v zahraniční měně a jedná se o celosvětově porovnatelné hodnocení. Při stanovení známky se používá globální stupnice.
- **Lokální** - vztahuje se pouze na dluhy v lokální měně. Nedá se porovnat s hodnocením zahraničních objektů, jelikož je hodnocení navrženo pouze pro srovnání uvnitř jednotlivých lokálních (národních) ekonomik.

2.5. Rating regionů a měst

Při stanovení ratingu regionů pomocí metodiky Moody´s [44] se vychází z kombinace dvou explicitních faktorů, kdy je nejdříve stanoveno základní úvěrové hodnocení (BCA) a pak se zohledňuje pravděpodobnost mimořádné podpory ze strany jiného subjektu např. vlády viz Obrázek 2.

[3] Emitovat neboli vydávat lze v závislosti na typu společnosti nebo zájmech společnosti nejrůznější druhy cenných papírů, z nichž nejznámější jsou akcie, dluhopisy a podílové listy



Obrázek 2: Postup sestavení výsledného ratingu regionu

Zdroj: upraveno podle [44]

2.5.1. Základní úvěrové hodnocení

Stanovení základního úvěrového hodnocení (Baseline Credit Assessment) dle podrobnější studie [43] je prvním krokem pro určení ratingu regionu. Odpovídá vlastní finanční síle daného regionu bez zohlednění mimořádných dotací nebo transferů od podporující vlády (státu). Smluvní vztahy a jakékoliv pravidelné dotační vztahy se státem jsou zohledňovány v BCA a proto se považují za součást hodnocení vlastní finanční síly daného regionu. Stupně BCA jsou vyjádřeny v alfanumerickém formátu s malými písmeny, které odpovídají alfanumerickým ratingovým stupňům na globální stupnici.

Pro zjištění základního úvěrového hodnocení se využívá bodovací karty individuálního rizika a matice BCA. Bodovací kartu využívá ratingový výbor k hodnocení úvěrové a finanční důvěryhodnosti krajských nebo místních samospráv a obsahuje soubor kvalitativních a kvantitativních ukazatelů, na jejichž základě se stanoví hodnoty základního úvěrového hodnocení. V úvahu jsou brány čtyři faktory, kdy u každého faktoru je bodově hodnoceno několik ukazatelů viz Tabulka 3 a poskytují dobrý statistický odhad pro posouzení síly subjektu jako takového. Všeobecně platí, že u subjektů s nejlepším hodnocením generovaným bodovací kartou se dá očekávat udělení vyššího ratingu.

Prvním hodnoceným faktorem je tzv. ekonomická základna (Economic Fundamentals), která je určena nestabilitou (volatilita) ekonomiky a regionálním HDP na obyvatele/celostátní HDP na obyvatele. Podíl regionálního HDP na obyvatele a celostátní HDP na obyvatele určuje relativní bohatství regionální vlády ve srovnání s celostátním průměrem hrubého domácího produktu dané vlády. Ekonomická volatilita je hodnocena na základě vyhodnocení

hospodářské rozmanitosti regionální vlády. Čím nižší je skóre ekonomické volatility tím vyšší je diverzifikace[4] [44].

Druhým faktorem je tzv. institucionální rámec (Institutional Framework), který se skládá z legislativy a fiskální flexibility. Při stanovení hodnocení pro legislativu se u každého národního institucionálního rámce hodnotí stabilita, předvídatelnost i schopnost reagovat na změny okolností a zda se změny vyskytují předvídatelným a řádným způsobem. U fiskální flexibility je hodnocení založeno na vyhodnocení příjmů z vlastních zdrojů, pružnost výdajů a velikost výpůjček.

Třetím posuzovaným faktorem je finanční výkonnost a profil zadluženosti (Financial Performance and Debt Profile). U tohoto faktoru se posuzuje hrubá provozní rovnováha/provozní příjmy, úrokové platby/provozní příjmy, likvidita, čistý přímý a nepřímý dluh/provozní příjmy a krátkodobý přímý dluh/přímý dluh. Poměr hrubé provozní rovnováhy a provozních příjmů měří schopnost vlády udržovat provozní výdaje pod provozními výnosy a vytvářet přebytky potřebné pro kapitálové výdaje a odpisy dluhů. Čím vyšší je tento poměr, tím nižší je riziko. Čím nižší je poměr úrokové platby a provozních příjmů, který představuje schopnost platby úroků pomocí provozních příjmů a podílů na výnosech, tím nižší je riziko. U likvidity se hodnotí schopnost dostát závazkům prostřednictvím úhrady v peněžních prostředcích. Poměr čistého přímého a nepřímého dluhu a provozních výdajů slouží jako pomocný ukazatel dluhové služby a vyjadřuje dluhovou zátěž. Čím je tento poměr nižší, tím nižší je riziko. Poslední poměr u tohoto faktoru slouží k posouzení refinancování rizik a úrokových rizik větších než jeden rok. Čím nižší je poměr, tím nižší je riziko.

Čtvrtým faktorem ovlivňující výsledek BCA je správa a řízení (Governance and Management), kde se posuzuje správa rizik a finanční řízení, řízení investic a dluhu, transparentnost a úroveň výkaznictví. U správy rizik a finančního řízení dochází k posouzení odbornosti a kvality plánovacích nástrojů vhodných pro finanční řízení. U položky řízení investic a dluhu je posuzována každá investice a dluh. Poslední položka vyjadřuje hodnocení včasnosti, kompletnosti a důvěryhodnosti účetní uzávěrky.

[4] Diverzifikace je ekonomický proces zvyšování rozmanitosti vyráběných výrobků, výrobních prostředků i spotřebních materiálů (http://slovnik-cizich-slov.abz.cz/web.php/slovo/diverzifikace)

Tabulka 3: Základní úvěrové hodnocení pro Moravskoslezský kraj

Základní úvěrové hodnocení Moravskoslezského kraje						
Faktory	**Body**	**Hodnota**	**Váhy dílčích faktorů**	**Dílčí faktory celkem**	**Váhy pro rozvíjející se země**	**Celkem**
Faktor 1 : Ekonomická základna						
Regionální HDP na obyvatele/celostátní HDP na obyvatele	7	86,35	70 %	5 ,2	20 %	1 ,04
Volatilita ekonomiky	1		30 %			
Faktor 2 : Institucionální rámec						
Legislativa	1		50 %	5 ,0	20 %	1 ,00
Fiskální legislativa	9		50 %			
Faktor 3 : Finanční výkonnost a profil zadluženosti						
Hrubý provozní výsledek/provozní příjmy (%)	5	4 ,33	12,5 %	3 ,5	30 %	1 ,05
Úrokové platby/provozní příjmy (%)	1	0 ,23	12,5 %			
Likvidita	1		25 %			
Čistý přímý a nepřímý dluh/provozní příjmy (%)	1	22,90	25 %			
Krátkodobý přímý dluh/přímý dluh (%)	9	44,20	25 %			
Faktor 4 : Správa a řízení kraje						
Správa rizik a finanční řízení	1			1 ,0	30 %	0 ,30
Řízení investic a dluhu	1					
Transparentnost a úroveň výkaznictví	1					
Posouzení individuálního rizika = ∑(bod * váha faktoru)						**3 ,39(3)**
Posouzení systematického rizika						**A1**
Navrhované základní úvěrové hodnocení						**a3**

Zdroj: upraveno podle [43],[44]

2.5.2. Mimořádná podpora

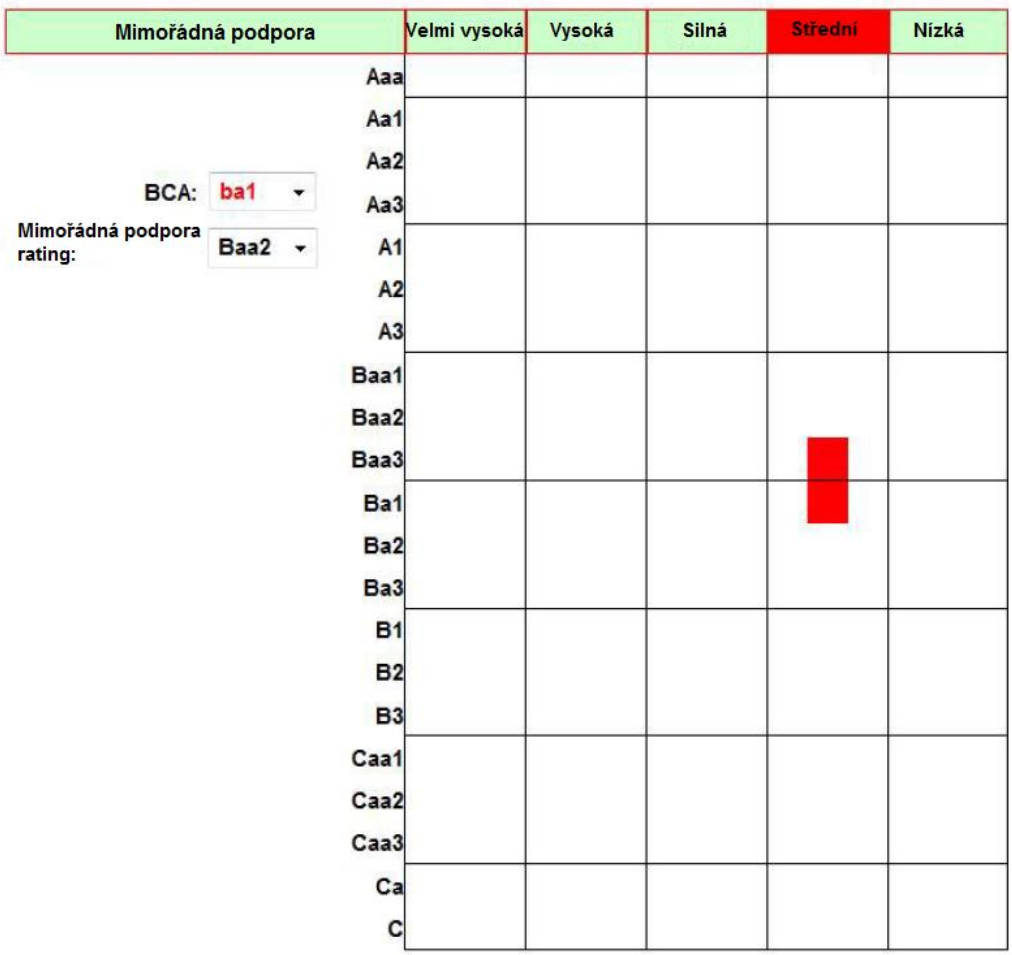
Mimořádná podpora je souhrn opatření, která by přijala podporující vláda, aby zabránila platební neschopnosti místní samosprávy. Tato podpora může mít různé podoby - od formální záruky až po přímou finanční výpomoc či zprostředkování jednání s věřiteli za účelem snadnějšího přístupu k finančním zdrojům. Mimořádná podpora je popisována jako [43]:

- nízká (0 -30 %),
- střední (31-50 %),
- silná (51-70 %),
- vysoká (71-90 %),
- velmi vysoká (91-100 %).

2.5.3. Výsledný rating

Na základě znalosti ratingu základního úvěrového hodnocení a mimořádné vládní podpory, je sestaven interval navrhovaných ratingů a poté je ratingovým výborem stanoven výsledný rating daného regionu. Při rozhodování jsou někdy brány v potaz i jiné nehodnocené faktory a v případě že tyto faktory mají záporný charakter ratingový výbor se s největší pravděpodobností přikloní k nižšímu výslednému ratingu z intervalu [44].

Ukázkový příklad principu stanovení výsledného ratingu je na obrázku níže viz Obrázek 3. Hodnotiteli v tomto případě byla stanovena střední mimořádná vládní podpora, pro základní úvěrové hodnocení byl stanoven rating ba1 (stejné jako označení Ba1) a pro mimořádnou podporu byl rating Baa2. V takovém případě by se ratingový výbor rozhodoval mezi ratingy Baa3, Baa2 a Ba1.



Obrázek 3: Zjištění intervalu pro stanovení výsledného ratingu

Zdroj: upraveno podle [44]

2.6. Rating České republiky

Česká republika získala svůj ratingový stupeň prostřednictvím hodnocení úvěrové důvěryhodnosti České národní banky. Úroveň ratingu ovlivňuje rating regionů v České republice, ale i schopnost centrální banky splácet úroky a jistinu obligací, které vydává. Nejde tedy o hodnocení České republiky jako celku. Rating centrální banky představuje nejvyšší stupeň, jaký může ekonomický subjekt na našem území získat. Například žádná česká banka rating A - v minulosti nevykazovala. Velké banky byly ohodnoceny stupněm BBB++, ostatní české banky měly rating nižší nebo jim nebyl přidělen žádný. Právě to ukazuje na stupeň důvěryhodnosti našeho bankovního sektoru [42],[50].

Česká republika má pro dlouhodobé závazky v domácí měně pro rok 2016 toto ratingové ocenění viz Obrázek 4:

Ratingová agentura	Ocenění (Rating)	Výhled
Moody´s	A1	Stabilní
Standard & Poor´s	AA	Stabilní
Fitch Ratings	A+	Stabilní

Obrázek 4: Rating ČR pro dlouhodobé závazky v domácí měně

Zdroj: [42]

Česká republika má pro dlouhodobé závazky v zahraničních měnách pro rok 2016 ratingové ocenění viz Obrázek 5:

Ratingová agentura	Ocenění (Rating)	Výhled
Moody´s	A1	Stabilní
Standard & Poor´s	AA-	Stabilní
Fitch Ratings	A+	Stabilní

Obrázek 5: Rating ČR pro dlouhodobé závazky v cizí měně

Zdroj: [42]

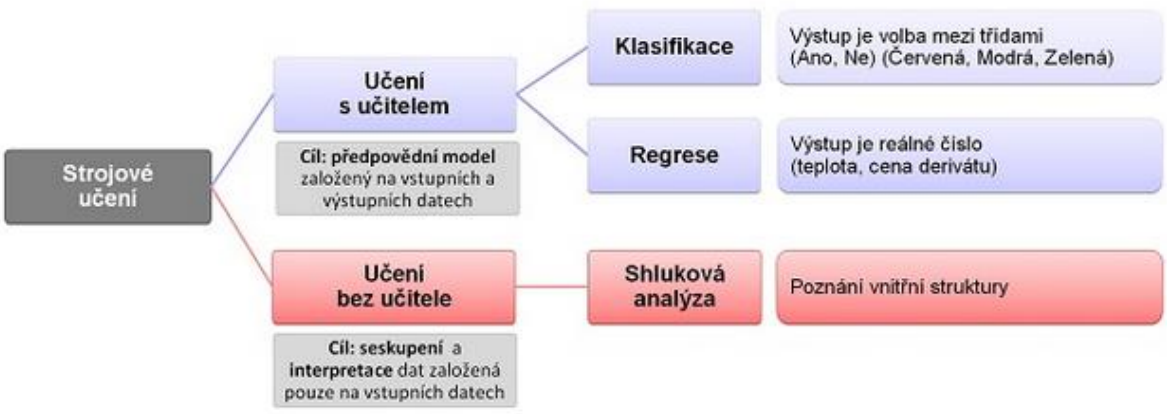
Stabilní výhled ratingové známky ČR vychází z přesvědčivosti účetní bilance vlády a také z dluhových brzd, což odráží připravenost na případné otřesy ze strany hlavních obchodních a investičních partnerů ČR [42],[50].

3. METODY STROJOVÉHO UČENÍ

Vzhledem k vysokým nákladům spojeným s ratingovým hodnocením a omezeným zdrojům regionů je vhodné navrhnout nástroj, který rating přiřadí automaticky. K tomu je vhodné použít metody strojového učení. Ty jsou schopné predikovat ratingy regionů na základě dostupných historických dat a ratingů udělených ratingovými agenturami. V následující kapitole je detailně popsáno strojové učení a jsou vysvětleny obecné principy algoritmů a postup testování. Jednotlivé metody jsou rozděleny podle typů úloh dle [6].

3.1. Strojové učení

Jedná se o vědní disciplínu, která se zabývá algoritmy a technikami, které umožňují počítači učit se ze vstupních dat. Pojem "učení" představuje schopnost zlepšování výkonnosti (přesnosti) modelu vlivem vzrůstající znalosti (nastavení parametrů) na základě zkušeností (dat). Strojové učení pak poskytuje matematický aparát a algoritmy, které jsou používány na řešení úloh v oblasti umělé inteligence (Obrázek 6). Obvyklou oblastí strojového učení v praxi je budoucí spotřeba a vývoj ceny elektrické energie nebo rozpoznávání řeči a obrazu [2].



Obrázek 6: Rozdělení strojového učení podle typu úlohy

Zdroj: upraveno podle [2]

Strojové učení můžeme podle způsobu učení rozdělit do dvou skupin dle [2]:

- **učení s učitelem** (*supervised learning*) - vyžaduje se sada trénovacích dat s výstupním atributem (*Y*). S takto naučeným modelem je poté možné provádět predikci (odhad hodnot) nebo klasifikaci (zařazení do tříd).

- **učení bez učitele** (*unsupervised learning*) - v případě že není k dispozici výstupní atribut (*Y*), pak se model učí na základě hledání podobností mezi vstupními daty. Typickým příkladem spadajícím do této oblasti je shluková analýza.

3.2. Statistické metody

Tyto metody se snaží modelovat vztahy a zákonitosti v datech formou matematických funkcí, vektorů a podmíněných pravděpodobností. Patří sem regresní, diskriminační, bayesovské metody a shluková analýza.

Jedná se o modely založené na pravděpodobnostech, které jsou důležité v případech, kdy není možné nalézt přesné řešení. V takové situaci je nutné určit, jak pravděpodobné jsou jednotlivé hypotézy. Základem metod patřících do bayesovské klasifikace, je Bayesova věta o podmíněných pravděpodobnostech, která určuje pravděpodobnost platnosti hypotézy *H* za předpokladu pozorování hypotézy *E* jako [29]:

$$P(H|E) = \frac{P(E \mid H) * P(H)}{P(E)}, \tag{4.1}$$

kde hypotéza *H* představuje jednotlivé třídy klasifikace a hypotéza *E* vstupní atributy. Pro každou hodnotu vstupního atributu (*E*) je poté hledána cílová třída (*H*) s největší podmíněnou pravděpodobností [30],[33].

Mezi nejznámější a nejpoužívanější bayesovské klasifikátory patří Naivní Bayesův klasifikátor [31] (Naive Bayes), který je založený na práci s podmíněnými pravděpodobnostmi. Pracuje se situací, kdy současně sledujeme více vstupních atributů ($E_1,...,E_k$) a vychází z předpokladu, že vstupní atributy (*E*) jsou podmíněně nezávislé při platnosti cílové třídy (*H*). Podmíněnou pravděpodobnost lze pak určit jako [30]:

$$P(H|E_1,...,E_k) = \frac{P(E_1,...,\ E_k) * P(H)}{P(E_1,...,\ E_k)} = \frac{P(H)}{P(E_1,....,\ E_k)} \prod_{i=1}^{K} P(E_i \mid H) \tag{4.2}$$

Pro tento způsob klasifikace potřebujeme pro každou třídu H_t znát hodnoty pravděpodobností $P(H_t)$ a $P(E_i|Ht)$, které algoritmus získává z předložených trénovacích dat při fázi učení. Požadovaným výsledkem je třída (hypotéza) s největší podmíněnou pravděpodobností [30],[33].

3.3. Symbolické metody umělé inteligence

Symbolické metody umožňují získat informace tzn. nalézt v datech vztahy a strukturu v takové podobě, která je srozumitelná pro běžného uživatele. Do této skupiny patří rozhodovací stromy, které jsou vhodné pro zpracování velkého množství dat.

Rozhodovací stromy (*Decision Tree*) [49] jsou analytické nástroje, které slouží k nalezení pravidel a vztahů v datovém souboru pomocí systematického rozdělování a větvení na nižší úrovně. Hlavním účelem je zjistit atributy, pomocí kterých je možné rozdělit záznamy a tím snížit výslednou nejistotu. Pro tvorbu rozhodovacího stromu se používá algoritmus TDIDT, jenž rozděluje heterogenní množinu na homogennější podmnožiny s nízkou entropií viz Obrázek 7 [36].

Rozhodovací strom se skládá z uzlů, kde uzel na nejvyšší úrovni je označován jako kořenový. Vnitřní uzle představují testy jednotlivých atributů (kořenový uzel je rovněž testem). Větví nazýváme možný výsledek testu. Externí uzly označované jako listy reprezentují jednotlivé třídy [36].

Algoritmus TDIDT

1. Zvol jeden atribut jako kořen dílčího stromu
2. Rozděl data v tomto uzlu na podmnožiny podle hodnot zvoleného atributu a přidej uzel pro každou podmnožinu
3. Existuje-li uzel, pro který nepatří všechna data do téže třídy, pro tento uzel opakuj postup od bodu 1, jinak konec

Obrázek 7: Obecný algoritmus TDIDT pro tvorbu stromu

Zdroj: upraveno podle [36]

Při výběru vhodného atributu k větvení se využívá tzv. entropie a informačního zisku. Entropie popisuje míru neuspořádanosti a je definovaná jako funkce [35]:

$$H = \sum_{t=1}^{T} (p_t * \log_2 p_t), \tag{4.3}$$

kde p_t je pravděpodobnost výskytu třídy t a T je počet tříd. Pro větvení stromu se vybere atribut s nejmenší entropií H [36].

Informační zisk se dle [36] zjistí jako rozdíl entropie pro celá data a pro uvažovaný atribut (měří redukci entropie způsobenou volbou atributu *A*). Výsledkem informačního zisku je atribut s maximální hodnotou:

$$zisk = H(C) - H(A) \tag{4.4}$$

Rozhodovací strom je acyklický graf, kde každý uzel vyjadřuje určitý atribut a větev/e hodnotu tohoto atributu. Mezi běžné algoritmy pro vytváření rozhodovacích stromů patří C5.0 , C &RT, QUEST a CHAID.

Algoritmus C5.0 (nejpoužívanější je jeho implementace J48) je rozšířením algoritmu ID3. Tvorba rozhodovacího stromu z trénovacích dat je shodná s postupem ID3, kdy je v každém uzlu vybrán atribut s nejvyšším informačním ziskem. Algoritmus C5.0 na rozdíl od svého předchůdce umožňuje práci s numerickými daty, chybějícími hodnotami i prořezávání rozhodovacího stromu (*pruning*), které zamezuje přeučení (*overfitting*) [49].

3.4. Subsymbolické metody umělé inteligence

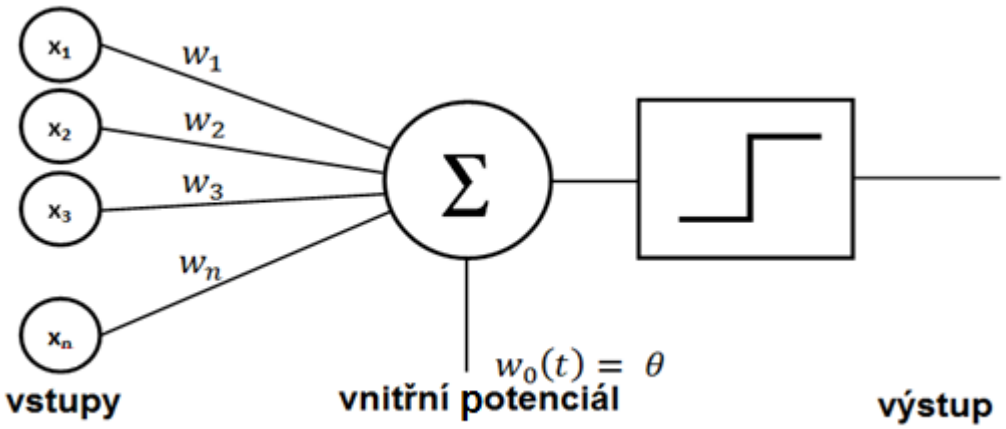
Výstup těchto metod není pro uživatele příliš pochopitelný např. váhy spojů mezi neurony v síti. Do této kategorie se řadí neuronová síť a podpůrné vektorové stroje (Support Vector Machine, SVM).

3.4.1. Neuronové sítě

Neuronová síť (Artificial Neural Network) je velmi populární a výkonná metoda, která se používá k modelování vztahu mezi vícerozměrným vstupním vektorem $\boldsymbol{x}$ a výstupní proměnnou *y*. Historie neuronových sítí sahá do čtyřicátých let 20. století, avšak výpočetní náročnost příslušných algoritmů umožnila jejich reálné využití až s vývojem výkonných počítačů. Neuronové sítě lze považovat za nelineární regresní model, který lze vyjádřit síťovou strukturou. Inspirací pro neuronové sítě byla struktura mozkové tkáně vyšších živočichů, kde je neuron propojen tzv. synapsemi s několika jinými neurony. Synapsemi prochází signál (resp. informace, vzruch), který je každým neuronem zpracován a předán dalšími synapsemi do dalších neuronů [52].

3.4.1.1. Perceptron

Základem umělé neuronové sítě je neuron označovaný jako *perceptron* [3]. Vyjadřuje jednovrstvou neuronovou síť, která je složená ze vstupních a výstupních neuronů. Perceptron dokáže správně klasifikovat pouze lineárně separovatelná data [48].



Obrázek 8: Model perceptronu

Zdroj: upraveno podle [24]

kde:

- x_n je n - tý vstup neuronu,
- w_n je hodnota n - té synaptické váhy,
- w_0 je vnitřní potenciál,
- 0 představuje práh neuronu,
- y je výstup neuronu viz Obrázek 8.

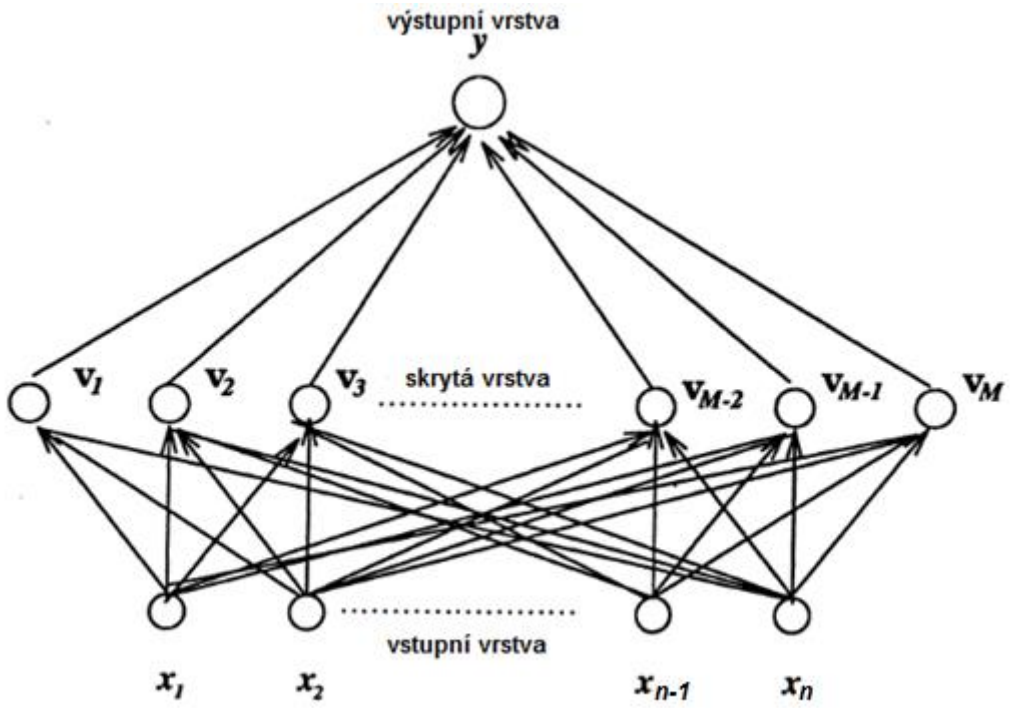
3.4.1.2. Vícevrstvá perceptronová síť

Vícevrstvá perceptronová síť (Multi Layer Perceptron, MLP) patří do kategorie dopředných umělých neuronových sítí se skrytou vrstvou. Struktura MLP s jednou vstupní, skrytou a výstupní vrstvou je znázorněna viz Obrázek 9. Jedná se o nejrozšířenější a nejpoužívanější neuronovou síť [24].

Vlastnosti dle [24]:

- síť je tvořena opakováním základního stavebního prvku – perceptronu,
- počet vstupních neuronů je dán počtem vstupů matematického modelu,
- počet výstupních neuronů je ovlivněn kódováním výstupu (pokud výstup neuronové sítě jde na vstup dalšího programu, lze zde provádět lineární transformaci dat),

- počet neuronů ve skryté vrstvě je volen s ohledem na složitost úlohy, obvykle jako maximum ze vstupů a výstupů,
- učení sítě probíhá za pomoci učení s učitelem,
- za aktivační funkci Φ se obvykle volí funkce lineární nebo sigmoidální,
- učícím algoritmem je nejčastěji algoritmus Back Propagation Error, lze však zvolit i metodu sdružených gradientů nebo jiné metody,
- síť je typická poměrně dlouhým procesem učení.



Obrázek 9: Vícevrstvá perceptronová síť MLP

Zdroj: upraveno podle [24]

Jelikož druhá i třetí vrstva je tvořena neurony klasického typu, pak pro určení hodnot h_k (výstup neuronů ve skryté vrstvě) a y_k (výstup neuronu ve výstupní vrstvě) lze využít vzorce [24]:

$$h_j = \Phi\left(w_{0,j} + \sum_{i=1}^{n} x_i w_{i,j}\right), \tag{4.5}$$

$$y_k = \Phi\left(v_{0,j} + \sum_{j=1}^{H} h_j v_{j,k}\right). \tag{4.6}$$

U sítě MLP dochází k učení metodou zpětného šíření chyby tzv. back-propagation [23], která je založena na tzv. gradientní metodě a efektivně mění synaptické váhy dokud

není dosaženo minimální chyby. Samotný algoritmus obsahuje tři etapy: dopředné šíření vstupního signálu tréninkového vzoru, zpětné šíření chyby a aktualizaci jednotlivých vah. Nastavení vah probíhá postupně, kdy se síti nejdříve předloží vzor, upraví všechny váhy a poté se jí předloží další vzor. Adaptace vah probíhá zpětně od výstupních vrstev ke vstupním a důležitým aspektem pro učení je výběr aktivační funkce, která musí být spojitá, diferencovatelná a monotónně neklesající. Hlavní výhodou této učící metody je, že pro výpočet změny konkrétní váhy je potřeba jen hodnota chyby na jednom konci a hodnota přenášeného signálu na druhém konci [7].

Při dopředném kroku algoritmu obdrží každý neuron ve vstupní vrstvě vstupní signál a provede přenos mezi všechny neurony ve skryté vrstvě. Následně každý neuron ve skryté vrstvě vypočítá svou aktivační funkci, která bude odpovídat skutečnému výstupu na základě předloženého vzoru.

Po provedení dopředného kroku jsou metodou zpětného šíření chyby porovnány vypočtené hodnoty aktivační funkce s výstupními hodnotami pro každý neuron ve výstupní vrstvě a pro každý tréninkový vzor [51]. Na základě odchylky vypočteného výsledku od skutečného je zjištěna velikost chyby (δ_j) dle [23], jež odpovídá části chyby, která se šíří zpětně z daného neuronu ke všem předcházejícím vrstvám majícím s tímto neuronem definované spojení.

$$\delta_j = x_j - y_j , \tag{4.7}$$

kde x_j je vypočtená hodnota a y_j je skutečná hodnota.

Následná úprava váhových hodnot mezi konkrétním neuronem a vyšší vrstvou závisí na odpovídající velikosti chyby a aktivacích neuronů v dané vrstvě [51]. Tento proces učení se opakuje dokud je celková chyba větší než definovaná přesnost.

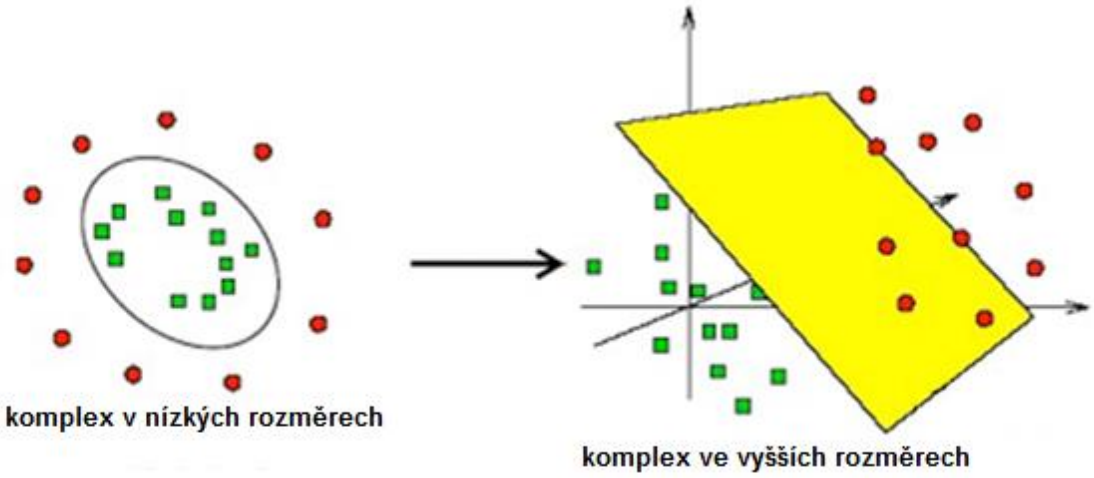
Algoritmus Backpropagation je následující [7]:

- vybere se vzorek,
- neuronová síť provede klasifikaci,
- výsledek klasifikace se porovná s očekávaným výstupem a vypočtou se změny vah podle odchylky,
- a váhy neuronové sítě se aktualizují a vybere se další vzorek podle kroku1.

3.4.2. Metoda podpůrných vektorů

Metoda podpůrných vektorů (Support Vector Machine, SVM) [26] byla vyvinuta Vladimirem Vapnikem roku 1979 a tvoří kategorii tzv. jádrových algoritmů. Snaží se využít

výhody poskytované efektivními algoritmy pro nalezení lineární hranice a zároveň je schopna reprezentovat vysoce složité nelineární funkce. Základním principem je převod původního vstupního prostoru do vícedimenzionálního, kde je již snadné oddělit obě třídy lineárně viz Obrázek 10 [22].

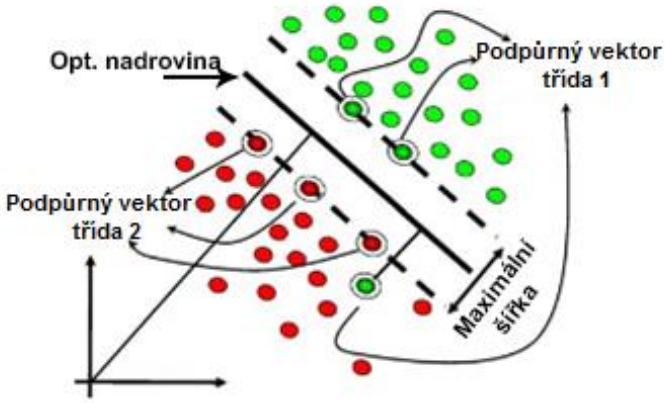


Obrázek 10: Lineární oddělení dvou tříd s nelineárními hranicemi přidáním dimenze

Zdroj: upraveno podle [38],[8]

V původním dvourozměrném prostoru jsou třídy odděleny nelineárně kružnicí, ale přidáním dimenze vzniká možnost prvkům uvnitř kružnice přiřadit další souřadnice, která je posune nahoru, takže pro oddělení obou tříd je možné použít rovinu rovnoběžnou s rovinou danou osami x_1 a x_2.

SVM jak již název napovídá pracuje se sadou podpůrných vektorů, což jsou vstupní vektory, jejichž funkcí je podpora oddělovací nadroviny, která na nich spočívá. Podpůrnými vektory jsou míněny vektory nacházející se nejblíže stanovenému lineárnímu klasifikátoru a jsou znázorněny jako body v kroužku viz Obrázek 11. Ostatní body (vektory) nejsou pro optimální nadrovinu vůbec důležité, protože metoda SVM dokáže z trénovacích dat nalézt pouze ty body (vektory), které jsou pro oddělovač podstatné [22]. Většina učících algoritmů používá totiž všechny poskytnuté trénovací případy, což při velkém množství dat může vést ke snížení efektivity [13].



Obrázek 11: Znázornění podpůrných vektorů a optimální nadroviny

Zdroj: upraveno podle [8]

3.4.3. Algoritmus SMO

Sequential Minimal Optimization (SMO) dle studie [37] je jednoduchý algoritmus pro řešení kvadratického optimalizačního problému (QP), který odpovídá trénování SVM klasifikátoru. Tento algoritmus je použitelný pro řešení větších problémů, dokáže rozložit na nejmenší možné QP podproblémy, které jsou řešeny analyticky s optimalizací v každém kroku.

Výhodou SMO je, že při řešení velkých problémů nemá značné nároky na paměť, protože žádná číselná matice řešená numericky není tak složitá. Tím není tento algoritmus citlivý na chyby přesnosti v numerických metodách.

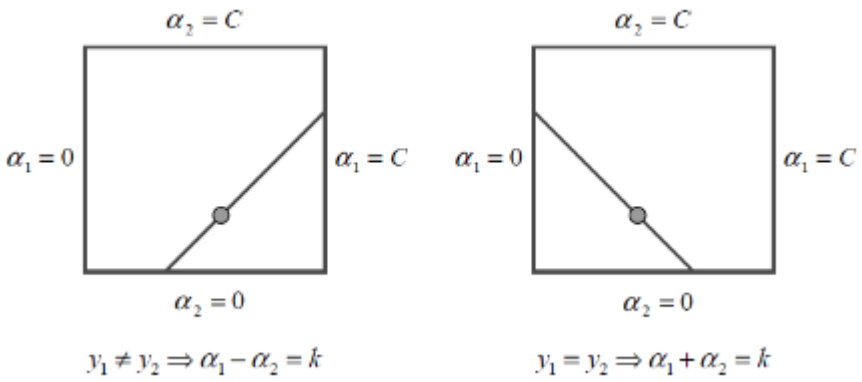
SMO se skládá ze dvou komponent:

- analytické metody pro řešení dvou Lagrangeových násobitelů
- a heuristiky pro výběr těchto násobitelů.

SMO umožňuje upravovat v jednom kroku pouze dva Lagrangeovy násobiteleα a takto zredukovaný problém je možné pak řešit analyticky viz Obrázek 12. Pro tyto dva násobitele α_1 a α_2 musí platit tyto podmínky [36]:

$$0 \leq \alpha_1, \alpha_2 \leq C,$$

$$y_1\alpha_1 + y_2\alpha_2 = K$$



Obrázek 12: Řešení optimalizace dvou násobitelů

Zdroj: [37]

Algoritmus lze zapsat následovně dle [45]:

1 . Najdi Lagrangeovy násobitele α_1, které porušují podmínky pro optimalizační úlohy (podmínky nutné pro hledání optimálního řešení úlohy nelineárního programování)

2 . Vyber si druhého násobitele α_2 a optimalizuj dvojici násobitelů α_1 a α_2

3 . Opakování předchozích dvou kroků, dokud nedojde k optimalizaci

4 . Pokud oba násobitelé splňují podmínku pro optimalizační úlohu, problém je vyřešen. Aby se urychlilo vyhledávání, používá se heuristiky pro výběr páru násobitelů.

3.5. Přesnost klasifikace modelu

3.5.1. Matice záměn pro dvě třídy

Pojmem přesnosti klasifikace modelu se vyjadřuje schopnost klasifikovat neznámá data tzn. data na které nebyl model naučen. Pro měření přesnosti klasifikace je ideální, pokud je k dispozici trénovací (na naučení) i testovací (na otestování úspěšnosti klasifikace) sada.

Výsledkem klasifikace je dle [47] tzv. matice záměn "confusion matrix" viz Tabulka 4, která ukazuje:

- kolik výsledků je skutečně pozitivních TP (true positive),
- kolik výsledků je chybně negativních FN (false negative),
- kolik výsledků je skutečně negativních TN (true negative),
- a kolik výsledků je chybně pozitivních FP (false positive).

Tabulka 4: Znázornění matice záměn pro dvě třídy

	Skutečnost	
Predikce	Ano (+)	Ne (-)
Ano (+)	**TP**	**FN**
Ne (-)	**FP**	**TN**

Zdroj: vlastní zpracování podle [46]

Z takto naměřených hodnot lze spočítat základní míry hodnocení úspěšnosti klasifikace testovacích dat dle [46]:

$$\text{přesnost } (\textit{accuracy}) = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4.8}$$

$$\text{chyba } (\textit{error}) = \frac{FP + FN}{TP + TN + FP + FN}, \tag{4.9}$$

$$\text{senzitivita } (\textit{sensitivity}) = \frac{TP}{TP + FN}, \tag{4.10}$$

$$\text{specificita } (\textit{specificity}) = \frac{TN}{TP + FP}. \tag{4.11}$$

Přesnost (4.8) udává podíl správně klasifikovaných objektů ke všem testovacím záznamům. Chyba (4.9) naopak pracuje s podílem chybně klasifikovaných objektů ke všem testovacím záznamům. Senzitivita (4.10) je podíl správně klasifikovaných k součtu skutečně pozitivních a chybně negativních (např. v mamologii procento výsledků, které jsou správně pozitivní v přítomnosti rakoviny prsu). Specificita (4.11) je podíl skutečně negativních k součtu skutečně pozitivních a chybně pozitivních.

3.5.2. Matice záměn pro více tříd

Matice záměn pro více tříd využívá stejné vzorce pro výpočet celkové správnosti, chyby, senzitivity a specifikace jako matice záměn pro dvě třídy (Tabulka 5). Rozdíl je v dosazování hodnot (kromě skutečně pozitivního výsledku), kdy v tomto případě se jednotlivé chybně negativní výsledky, skutečně negativní výsledky a chybně pozitivní výsledky musí sečíst před dosazením do vzorce.

Tabulka 5: Matice záměn pro výpočet úspěšnosti klasifikace pro třídu 1

		Skutečná třída								
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Predikovaná třída	**1**	TP	FN	FN	FN	FN	FN	FN	FN
	2	FP	TN	FN	FN	FN	FN	FN	FN
	3	FP	FN	TN	FN	FN	FN	FN	FN
	4	FP	FN	FN	TN	FN	FN	FN	FN
	5	FP	FN	FN	FN	TN	FN	FN	FN
	6	FP	FN	FN	FN	FN	TN	FN	FN
	7	FP	FN	FN	FN	FN	FN	TN	FN
	8	FP	FN	FN	FN	FN	FN	FN	TN

Zdroj: vlastní zpracování

3.5.3. Nákladová matice

Nákladová matice je rozvinutí chybné klasifikace a je výstižně charakterizována již svým názvem. Nese informaci o nákladech spojených s chybnou klasifikací Tabulka 6. Odlišením závažnosti chyby lze pak ovlivnit i to, jak bude vytvářený model naučen. Číselné ohodnocení závažnosti chyby však nemusí být triviální problém. Je zpravidla podmíněno znalostí modelované problematiky a k jejímu stanovení může být nutné provést analýzu výstupní veličiny [47].

Tabulka 6: Ilustrativní nákladová matice

		Skutečná třída								
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Predikovaná třída	**1**	0	1	2	3	4	5	6	7
	2	1	0	1	2	3	4	5	6
	3	2	1	0	1	2	3	4	5
	4	3	2	1	0	1	2	3	4
	5	4	3	2	1	0	1	2	3
	6	5	4	3	2	1	0	1	2
	7	6	5	4	3	2	1	0	1
	8	7	6	5	4	3	2	1	0

Zdroj: vlastní zpracování

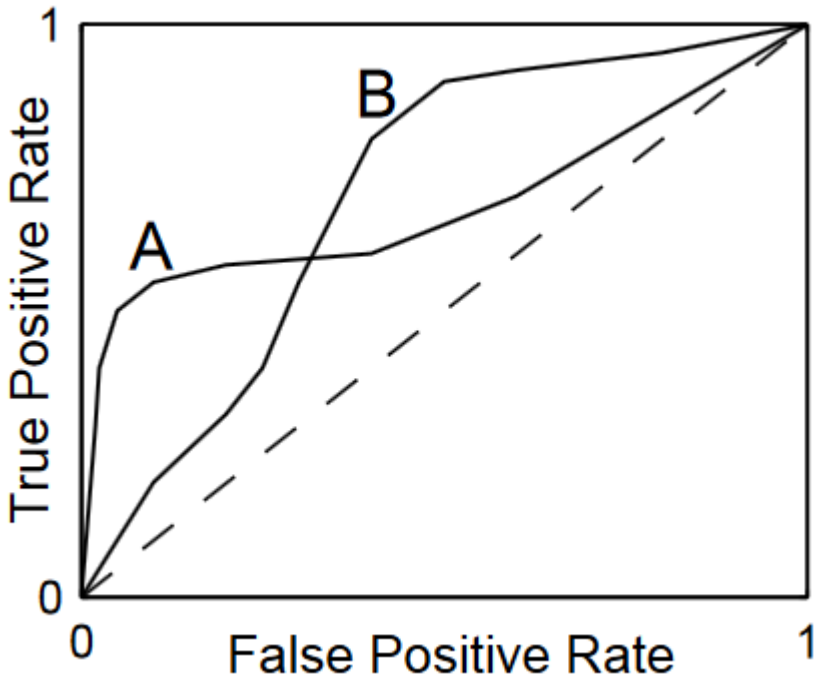
Nákladových matic může být několik typů, ale v této práci se bude využívat matice, kde se při chybné klasifikaci náklady zvyšují, čím více se skutečná třída liší od cílové

ratingové třídy viz Tabulka 6. V praxi se však náklady mohou lišit podle uživatele modelu, např. samosprávy, investora, banky, apod.

3.5.4. Křivka ROC

ROC křivka (Receiver Operating Characteristic Curve - graf prahové operační charakteristiky) [47] v dnešní době nachází široké uplatnění v oblasti medicíny (při lékařském rozhodování) a hlavně ve strojovém učení při vyhodnocování úspěšnosti klasifikátorů při klasifikaci do dvou tříd. Jedná se o grafické hodnocení úspěšnosti klasifikátoru, ale i o číselné hodnocení (plocha pod ROC křivkou, AUC). Dále je vhodným měřítkem kvality klasifikace než přesnost v případě nevybalancovaných dat (s různým počtem případů v predikovaných třídách).

V ideálním případě by měl model mít senzitivitu i specificitu rovnu jedné, ale v praxi tomu často není, proto se využívá křivka ROC, která zobrazuje závislost sensitivity na opačné hodnotě specificity (1 -specificita). Lze srovnávat více křivek predikčních modelů. Čím více je křivka blíže bodu jedna tím dochází k lepšímu oddělení tříd viz Obrázek 13. Pro zjištění velikosti míry oddělení složek se využívá plochy pod křivkou tzv. AUC (Area Under Curve), která by se měla co nejvíce blížit hodnotě jedna [20].



Obrázek 13: ROC křivka pro dva odlišné klasifikátory

Zdroj: [47]

4. META-UČENÍ

Meta-učící metody jsou někdy označované také jako *ensemble* metody nebo soubory klasifikátorů. Mezi tyto metody patří *Boosting, Bagging a Stacking*, které mají za cíl naučit více modelů pomocí jednoho nebo různých algoritmů a následně je zkombinovat. Ve většině případů použití souborů těchto metod (algoritmů) vede k dosažení vyšší přesnosti.

4.1. Boosting

Hlavní myšlenkou metody Boosting [5] je zlepšení klasifikační přesnosti libovolného algoritmu strojového učení. Snaží se posílit slabé učící algoritmy na algoritmy silné tím, že použije více modelů stejného typu se slabým učícím algoritmem, jejichž přesnost je lepší než náhodná (více než 50 %). Na začátku je každému záznamu v trénovací sadě přiřazena váha $w_i(x)=1$. Jakmile dojde k vytvoření prvního modelu, jsou trénovací záznamy klasifikovány a váha úspěšně klasifikovaných je oslabena (budou se méně podílet na nastavení), oproti tomu váha neúspěšně klasifikovaných je posílena (zvětšena) [40],[47].

Zástupcem této metody je algoritmus AdaBoost, který vezme jednoduchý (slabý) klasifikátor, který má alespoň 50 % přesnost a poté je mezi sebou zkombinuje, aby dosáhl požadované přesnosti. Touto kombinací vzniká silný klasifikátor vytvořený z několika slabých klasifikátorů viz Obrázek 14. Tím dochází k exponenciálnímu zmenšení trénovací chyby v závislosti na počtu klasifikátorů a pokud každý klasifikátor dává lepší výsledky než náhodné je zaručen pokles trénovací chyby [47]:

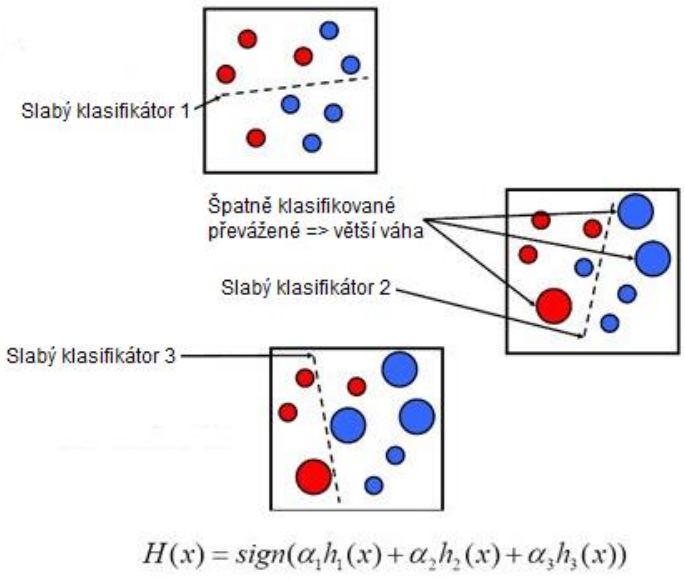
$$H(x) = sign \sum_{i=1}^{N} a_i h_i(x_i), \tag{5.1}$$

kde a_i je váha, kterou je potřeba nastavit učením.

Algoritmus AdaBoost.M.1lze zapsat následujícím pseudokódem:

Vstup:	Datová sada $D = \{(x_1,y_1),(x_2,y_2), \ldots ,(x_m,y_m)\}$; $x \in X$, $y \in Y$
	Slabý algoritmus učení K
	Parametr T specifikující počet iterací
Inicializace:	$D_1(i) = 1/m$ distribuce vah pro všechny trénovací vzorky
Proces:	Pro $t = 1 , \ldots , T$:
1 .	Vyber z D trénovací množinu D_t podle rozdělení D_t; použij algoritmus učení K a vytvoř hypotézu (klasifikátor) $h_t = K(D, D_t)$;
2 .	Vypočti chybu ε_t a vyber hypotézu (slabý klasifikátor) h_t s nejmenší chybou $\epsilon_t = Pr_{i\sim Di}[h_t(x_i \neq y_i)]$;
3 .	Výpočet váhy hypotézy h_p (klasifikátoru) $\alpha_t = \frac{1}{2} ln \frac{1-\epsilon_t}{\epsilon_t}$
4 .	Aktualizace vah vzorků v trénovací množině $D_{t+1}(i) = \frac{D_t(i) exp(-\alpha_t y_i ht(x_i))}{Z_t}$, kde Z_t je normalizační faktor, zvolený tak, aby D_{t+1} zůstala pravděpodobnostím rozdělením.
Výstup:	**Výsledný klasifikátor sestavený ze slabých klasifikátorů** $H(x) = sign \sum_{t=1}^{T} \alpha_t h_t(x)$

Zdroj: upraveno podle [53]



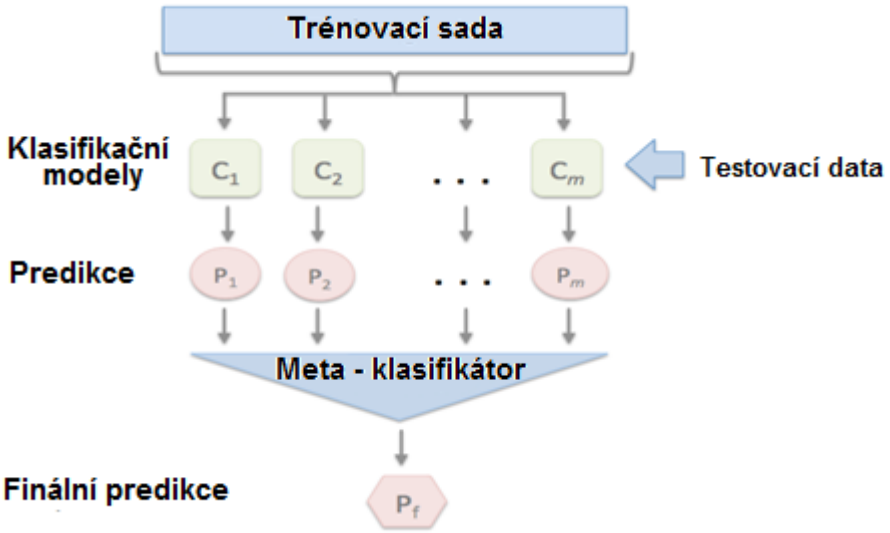
Obrázek 14: Postupné získávání silného klasifikátoru ze slabých klasifikátorů

Zdroj: upraveno podle [14]

Výhodou algoritmu AdaBoost je použití rychlých klasifikátorů bez nutnosti obětovat přesnost.

4.2. Bagging

Bagging [5] patří k nejjednodušším postupům z kategorie meta-learningových metod a poprvé ji představil Loe Breiman. Název metody je odvozen od *boostrap aggregation*, kde se využívá více trénovacích množin, kdy každá nová množina je tvořena výběrem záznamů ze základní trénovací množiny. Jde tedy o metodu generování několika verzí modelů z jedné trénovací sady, kdy je vybrána množina *n* vzorků (vzorky se v množinách mohou opakovat) a na každé takto vytvořené množině je natrénován model viz Obrázek 15. Obvykle jsou modely *M* stejného typu, nejčastěji to jsou rozhodovací stromy nebo neuronové sítě. Tímto principem je dosaženo toho, že každý model je odlišný, protože je pro něj použita jiná množina vzorků z trénovací sady. Jakmile dojde k naučení všech modelů, dochází každým modelem k určení třídy, do které má být prvek klasifikován. Výsledná třída je poté dána hlasováním (nejčastěji predikována) jednotlivých modelů a třída s nejvíce hlasy vyhraje [47]. Tato metoda se používá v případě nedostatku dat, zvyšuje stabilitu a snižuje riziko přeučení. [40],[47]



Obrázek 15: Metoda Boostrap Aggregation

Zdroj: upraveno podle[14]

Metoda Random Forest dle studie [11] patřící do této kategorie je použitelná pro klasifikaci i predikci. Hlavní myšlenkou je natrénovat každý strom nezávisle na ostatních, tak že se budou od sebe lišit. Vychází z metody Bagging, kdy nelze zadat všem modelům

stejná trénovací data, protože by výsledky byly totožné. Proto se u této metody využívá tzv. boostrapového výběru z trénovací sady.

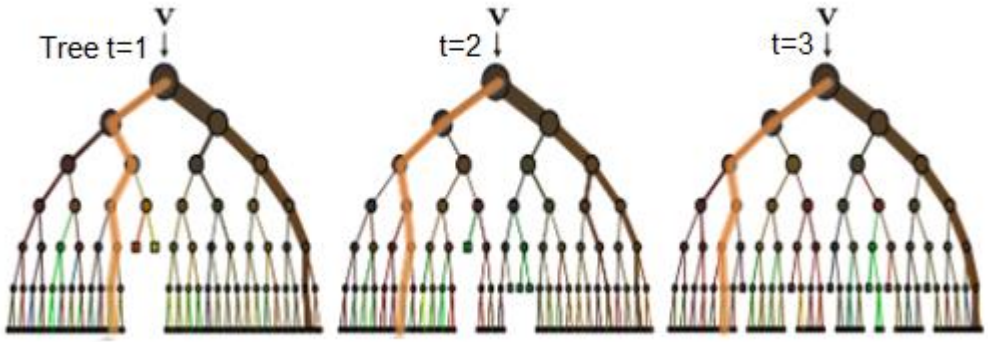
Při použití náhodného lesa pro klasifikaci je výsledkem většinové hlasování

$$C_{\mathrm{RF}} = vets\ \_\ hlas\ \{C_i(x)\}_i^N\ , \tag{5.2}$$

Při použití pro predikci je výsledkem průměr ze všech stromů

$$f_{\mathrm{RF}}(x) = \frac{1}{N}\left(\sum_{i=1}^{N} T_i(x)\right)\ , \tag{5.3}$$

Konstrukce rozhodovacího stromu probíhá hledáním nejvhodnějšího atributu o maximálním informačním zisku. Naproti tomu u konstrukce náhodného rozhodovacího stromu je nejvhodnější atribut pro dělení stromu vybrán náhodně z náhodné trénovací podmnožiny podle maximálního informačního zisku. Vzhledem k náhodnému výběru a omezenému počtu záznamů z trénovací sady nedostaneme samotným rozhodovacím stromem nejlepší možný výsledek, ale při konstrukci většího množství náhodných stromů naučených na náhodných podmnožinách, lze získat velmi silný a spolehlivý klasifikátor, viz Obrázek 16. Nevýhodou je paměťová náročnost z důvodu vytvoření náhodného rozhodovacího stromu pro každou náhodnou trénovací podmnožinu. Příkladem použití náhodného lesa je rychlá klasifikace částí těla a gest pro Microsoft Kinect.



Obrázek 16: Ukázka konstrukce tří náhodných stromů

Zdroj: upraveno podle [12]

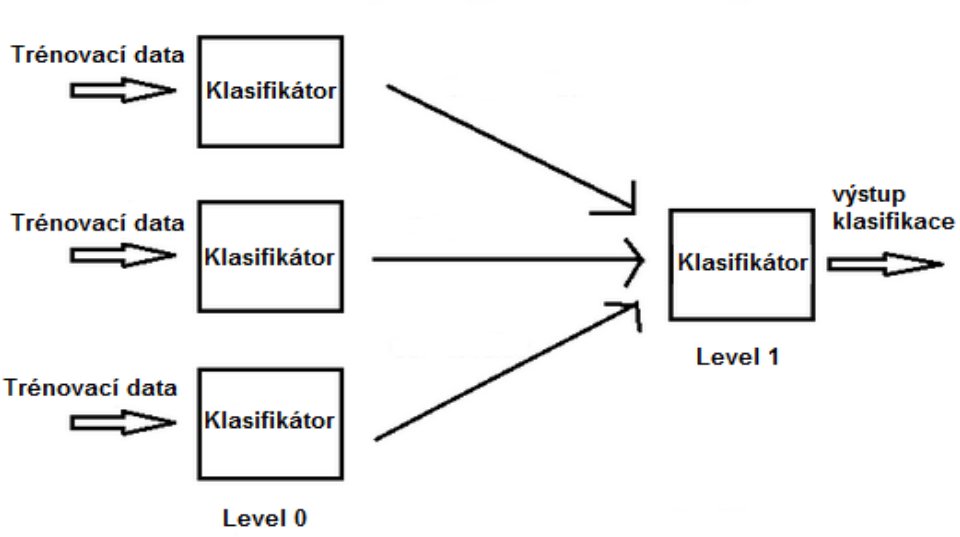
Algoritmus Random Forest lze zapsat pseudokódem:

Vstup:	Datová sada $D = \{(x_1,y_1),(x_2,y_2), \ldots ,(x_m,y_m)\}$; $x \in X$, $y \in Y$
	Funkce F
	Počet stromů B
	Parametr T specifikující počet iterací
Funkce RANDOMFOREST (D ,F): Pro t = 1 , ..., T :	
1 .	Vyber z D náhodnou trénovací podmnožinu $D^{(t)}$
2 .	Volání funkce RANDOMIZEDTREELEARNS($D^{(t)}$,F) a výstup uložen do proměnné h_t
3 .	Výstup h_t z funkce RANDOMFOREST uložen do proměnné $H = H \cup \{h_t\}$
4 .	Funkce RANDOMIZEDTREELEARNS (D ,F) vrací náhodný strom
Výstup:	**Klasifikátor tvořený několika rozhodovacími stromy**

Zdroj: upraveno podle [41]

4.3. Stacking

Tato metoda kombinuje více modelů různých typů(několik základních klasifikátorů), které jsou rozděleny na úroveň 1 a 0. K učení modelů dochází postupně (nejprve se učí modely na úrovni 0). Modely na úrovni 0 mohou být libovolného druhu a mají určenou váhu (míru důvěryhodnosti predikce), ale modely na úrovni 1 by měly mít jednoduchou strukturu jako jsou rozhodovací stromy viz Obrázek 17 [10].



Obrázek 17: Princip metody Stacking

Zdroj: upraveno podle [14]

Pro aplikaci metody stacking jsou data rozdělena na trénovací a testovací. Na trénovacích datech jsou naučeny modely úrovně 0. Jejich výstupy budou vstupy do úrovně 1. Trénovací data pro metamodel jsou použita odložená testovací data. Tato data musí projít modely úrovně 0, tím se získají trénovací data pro model úrovně 1. Po jeho nastavení jsou použita pro přenastavení všech modelů úrovně 0 všechna trénovací data, avšak již bez zásahu do metamodelu úrovně 1. Tímto je dosaženo zlepšené celkové přesnosti [10],[14].

Algoritmus Stacking lze zapsat pseudokódem:

Vstup:		Datová sada $D = \{(x_1,y_1),(x_2,y_2), \ldots ,(x_m,y_m)\}; x \in X , y \in Y$
		První stupeň algoritmu učení $K_1, \ldots ,K_T;$
		Druhý stupeň algoritmu učení K
		Parametr T specifikující počet iterací
Proces:		Pro t =1 , ..., T :
	1 .	Natrénování prvního stupně klasifikátorů na originální datové sadě D, $h_t = K_t(D);$
		end;
	2 .	Deklarace nové proměnné D' pro novou datovou sadu $D' = 0 ;$
		Pro i = 1 , ... ,m :
		Pro t = 1 , ... ,T :
	3 .	Uložení výstupů z klasifikátorů prvního stupně $z_{it} = h_t(x_i);$ $D' = D' \cup \{((z_{i1},z_{i2},\ldots,z_{iT}),y_i)\};$
		end;
	4 .	Natrénování druhého stupně klasifikátorů na nové datové sadě D', $h' = K (D');$
Výstup:		**Výsledný klasifikátor** $H (x) = h'(h_1(x), \ldots ,h_T(x))$

Zdroj: upraveno podle [53]

5. DATA A NÁVRH MODELU

Datová sada s veškerými statistickými údaji o regionech sledované ratingovou agenturou Moody´s poskytuje údaje o determinantech finanční výkonnosti 254 regionů, 4 regionů České republiky (Praha, Brno, Ostrava a Moravskoslezský kraj). Tento soubor dat, který byl poskytnut ratingovou agenturou Moody´s [5] pro účely diplomové práce obsahuje 36 atributů s výstupním atributem o stavu finanční výkonnosti (ratingu). Atributy HDP na obyvatele, nezaměstnanost, paritu kupní síly atd. představují hodnoty pro predikci ratingu v následujícím roce, protože delší časová predikce vzhledem ke stále se měnícím ekonomickým podmínkám by byla méně spolehlivá. Výstupní atribut je klasifikován za rok 2016 do osmi tříd a vyjadřuje stav finanční výkonnosti v rozmezí (1-velmi dobrý až 8-velmi špatný). Stanovené statistické údaje (vstupní atributy) jednotlivých regionů jsou průměry z let 2011až 2015. Důvodem je, že pětileté časové období je pro hodnocení požíváno ratingovými agenturami. Je tak omezen vliv vychýlených hodnot na výsledný rating.

Výstupní třída v datovém souboru obsahuje stupně ratingu, které jsou stanoveny z hodnocení dlouhodobé národní stupnice podle agentury Moody´s. Hodnotící stupnice jsou uvedeny s modifikátorem "nn", který značí příslušnou zemi např. pro Českou republiku je použito označení AAA.cz viz Tabulka 7.

Tabulka 7: Popis použité dlouhodobé národní rating stupnice

Rating podle agentury Moody´s	Stupeň ratingu	Popis
Aaa.nn	1	Nejvyšší úvěrová způsobilost
Aa.nn	2	Velmi vysoká úvěrová způsobilost
A .nn	3	Nadprůměrná úvěrová způsobilost
Baa.nn	4	Průměrná úvěrová způsobilost
Ba.nn	5	Podprůměrná úvěrová způsobilost
B .nn	6	Slabá úvěrová způsobilost
Caa.nn	7	Velmi slabá úvěrová způsobilost
Ca.nn	8	Extrémně slabá úvěrová způsobilost

Zdroj: [46]

5.1. Datový slovník

Vytvořený datový slovník (Tabulka 8) obsahuje důležité informace o všech vstupních atributech pro lepší pochopení významu a ulehčení práce s daty. Poskytuje následující informace:

[5]https://www.moodys.com/

- název atributu (český název proměnných),
- typ (datový typ proměnných - všechny vstupní proměnné jsou datového typu Real),
- zkratka (zkrácený název pro práci s proměnnými v programu Weka),
- a rozsah atributu (udává maximální a minimální hodnotu dané proměnné).

Tabulka 8: Přehled vstupních atributů

Název atributu	**Typ**	**Zkratka**	**Rozsah atributu**
Populace	Real	Population	[20;43117]
HDP	Real	GDP	[1799;798750]
HDP na obyvatele	Real	GDPPerCapita	[38;654]
HDP/národní průměr	Real	GDP/NationalAve	[4022;107305]
HDP na obyvatele v paritě kupní síly	Real	GDP_PPP	[372;112473]
Reálné HDP	Real	RealGDP	[-4;10]
Míra nezaměstnanost	Real	Unemployment	[1;37]
Národní míra nezaměstnanosti	Real	NationalUnemployment	[3;30]
Celkový přímý a nepřímý dluh	Real	TotalDirectIndirectDebt	[0;273327]
Čistý přímý a nepřímý dluh	Real	NetDirectIndirectDebt	[0;257758]
Čistý přímý a nepřímý dluh na obyvatele	Real	NetDebtPerCapita	[0;29310]
Čistý přímý a nepřímý dluh/HDP	Real	Debt/GDP	[0;84]
Čistý přímý a nepřímý dluh/provozní výnosy	Real	Debt/OperatingRevenue	[0;385]
Čistý přímý a nepřímý dluh/celkový příjem	Real	NetDirectIndirectDebt/TR	[0;346]
Přímý dluh v cizí měně(před výměnou)/přímý dluh	Real	FXDirectDebt(beforeswap)	[0;94]
Přímý dluh v cizí měně(po výměně)/přímý dluh	Real	FXDirectDebt(afterswap)	[0;94]
Krátkodobý přímý dluh/přímý dluh	Real	ShortDebt/Debt	[0;85]
Krátkodobý variabilní poměr dlouhodobého přímého dluhu/přímý dluh	Real	LongDebt/Debt	[0;100]
Vážený průměr splatnosti přímého dluhu	Real	MaturityDebt	[0;22]
Vlastní zdroje příjmů/provozní výnosy	Real	OwnRev/OperRev	[1;122]
Mezivládní převody/provozní příjmy	Real	GovTransf/OperRe	[-22;99]
Vyčleněné příjmy/provozní výnosy	Real	EarRev/OperRev	[0;95]
Úrokové platby/ provozní výnosy	Real	Inter/OperRev	[0;12]
Dluhová služba/celkové příjmy	Real	DebtSer/TR	[0;61]

Akruální finanční přebytek/celkový příjem	Real	AccuaFinancingSurplus/TR	[-41;14]
Peněžní přebytek/celkové příjmy	Real	CashSurp/TR	[-31;20]
Hrubá výpůjční potřeba/celkový příjem	Real	BorrowNeed/TR	[-11;111]
Celkové výdaje na obyvatele	Real	TEPerCapita	[10;46865]
Celkové výdaje na obyvatele/HDP	Real	TE/GDP	[0;80]
Primární operační bilance/provozní výnosy	Real	OperBalance/OR	[-14;82]
Hrubé provozní bilance/provozní výnosy	Real	GrossOB/OR	[-23;76]
Čistá provozní bilance/provozní výnos	Real	NetOperatingBalance/OR	[-128;76]
Poměr samofinancování	Real	SelfFinRatio	[-1;4]
Kapitálové výdaje/celkové výdaje	Real	CapitalSpend	[2;67]
5ti leté celkové tržby - 5ti leté celkové výdaje	Real	TR-TE	[-7;17]
Čistý pracovní kapitál/celkové výdaje	Real	NWC/TE	[-61;218]

Zdroj: vlastní zpracování

Detailní popis atributů, které byly použity v této diplomové práci, je založen na studii [46].

Populace (Population) - populace nemusí být nutně ratingovým faktorem, ale demografické trendy jako je populační růst a věkové rozdělení pomáhají podmínit fiskální prostředí vládního sektoru. Hodnoty populace jsou použity v níže zmíněných atributech pro přepočet na jednoho obyvatele.

HDP (GDP) - je standardní mezinárodní měřítko velikosti ekonomik, které kvantifikuje hodnotu konečného zboží a služeb na běžných cenách přepočtených na běžný americký dolar. HDP je pro mnoho velkých regionálních vlád snadno dostupné a poskytuje srovnání na mezinárodní úrovni.

HDP na obyvatele (GDPPerCapita) - jedná se o přepočet HDP na jednoho obyvatele (sumu celkového hrubého domácího produktu vydělíme počtem obyvatele daného kraje). Pomocí HDP na hlavu můžeme porovnávat kraje (země) s různě velkým hospodářským prostorem.

HDP/Národní průměr (GDP/NationalAverage) - vyjadřuje HDP na obyvatele jako procento z národního HDP na obyvatele. Slouží pro odhadnutí relativní ekonomické výkonnosti a posouzení dopadu na fiskální výkonnost. Hodnoty přesahující 100 % udávají, že úroveň bohatství země je nad národní normy.

HDP na obyvatele v paritě kupní síly (GDPPPP) - HDP a příjmy domácností jsou nejprve vypočítány v národních měnách a poté jsou přepočítány v paritě kupní síly, které berou v potaz rozdílné cenové úrovně v jednotlivých členských státech a umožňují

přesnější porovnání. Pomocí PPP se tyto ukazatele převedou na uměle vytvořenou společnou měnu zvanou PPS (standard kupní síly). V jednoduchých verzích PPP představuje poměr cen v národních měnách za stejné výrobky a služby v různých zemích.

Real GDP (RealGDP) - měří celkovou hodnotu produkce za dané období ve stálých cenách, kde pod pojmem stálá cena rozumíme ceny určitého zvoleného období. Jedná se o nominální HDP očištěný od inflace a podává skutečný obraz o výkonu ekonomiky, skutečném fyzickém přírůstku nebo úbytku produkce.

Míra nezaměstnanosti (Unemployment) - patří mezi nejvíce sledované ukazatele v makroekonomii a počítá se jako podíl počtu nezaměstnaných k pracovní síle:

$$u = \frac{N}{EA} * 100 = \left(\frac{N}{Z + N}\right) * 100 \quad , \tag{6.1}$$

kde N je počet nezaměstnaných a Z počet zaměstnaných. Udává průměrný údaj pro celou zemi nebo pro regiony dané země. Informuje o stavu, kolik procent z ekonomického aktivního obyvatelstva nemá ve sledovaném období zaměstnání.

Národní míra nezaměstnanosti (NationalUnemployment) - slouží jako srovnávací bod pro ukazatel míry nezaměstnanosti a poskytuje informace o stavu na trhu práce v průmyslových zemích.

Celkový přímý a nepřímý dluh (TotalDirectIndirectDebt) - je měřítkem stavu dluhu zahrnující všechny krátkodobé a dlouhodobé dluhy vlády, dlužné závazky vydané vládou a také sem patří dluhové nástroje, kterými jsou kapitálové pronájmy nebo partnerství veřejného a soukromého sektoru.

Čistý přímý a nepřímý dluh (NetDirectIndirectDebt) - vypočítá se odečtením hrubého dluhu z vládních finančních aktivit.

Čistý přímý a nepřímý dluh na obyvatele (NetDebtPerCapita) - poskytuje možnost mezinárodního srovnání úrovní dluhu jednotlivých zemí. Z důvodu rozdílů mezi úrovní bohatství průmyslově vyspělých a rozvíjejících se ekonomik je doporučeno mezinárodní srovnání zadluženosti provádět mezi zeměmi podobného nebo stejného ekonomického vzrůstu.

Čistý přímý a nepřímý dluh/HDP (Debt/GDP) - poskytuje procentuální vyjádření čistého přímého a nepřímého dluhu vzhledem k HDP dané země. Například v České republice je dluh způsoben deficitním rozpočtem vlády, uvádět dluh vůči HDP je velmi nepřesné vzhledem k tomu, že se jedná o nepodložený odhad.

Čistý přímý a nepřímý dluh/Provozní výnosy (Debt/OperatingRevenue) - poskytuje alternativu pro vyjádření přesnější analýzy dluhové zátěže dané země. Použití provozních výnosů ve jmenovateli odráží rozpočtovou strukturu mnoha vlád, které rozlišují mezi investiční a provozní činností. U některých zemí je část provozních výnosů vyčleněna pro jiné účely, než je zadluženost země.

Čistý přímý a nepřímý dluh/Celkový příjem (NetDirectIndirectDebt/TR) - tento atribut vychází z předchozího poměru, jen je zde přidán do jmenovatele kapitálový příjem. Příjmy nejsou omezeny a tím poskytují ucelenější obraz příjmových toků.

Přímý dluh v cizí měně (po výměně)/Přímý dluh (FXDirectDebt(afterswap)) - tento poměr měří devizový dluh po výměně měny jako procento přímého dluhu, který nezahrnuje nepřímý dluh.

Krátkodobý přímý dluh/Přímý dluh (ShortDebt/Debt) - zahrnuje položky se splatností kratší než jeden rok jako jsou např. obchodní cenné papíry. K financování státního krátkodobého dluhu jsou vydávány krátkodobé dluhopisy se splatností do 1 roku prodávané na peněžním trhu.

Krátkodobý a variabilní poměr dlouhodobého přímého dluhu/Přímý dluh (LongDebt/Debt) - tento poměr oproti výše zmíněnému pracuje navíc s dlouhodobou variabilní sazbou přímého dluhu a tím zachycuje celkovou výši úrokového rizika v krátkodobém a dlouhodobém horizontu. Pomocí těchto hodnot je možné hodnotit dopad změn úrokových sazeb na dluh a refinancování nákladů.

Vážený průměr splatnosti přímého dluhu (MaturityDebt) - kromě rizik zachycených ve výše uvedených proměnných je nutné také zkoumat změny úrokových sazeb,a to u rozvíjejících se zemí vzhledem ke splatnosti dluhu. Jedná se o schopnost vlády získávat příjmy z občanů dané země a to prostřednictvím daní nebo poplatků, aby byly dodrženy závazky vůči dluhům. Všechna data jsou převzata z auditovaných účetních závěrek.

Vlastní zdroje příjmů/Provozní výnosy (OwnRev/Operating Revenue) - položka vlastní zdroje příjmů bere v úvahu veškeré příjmy jako jsou daně, poplatky a jiné zdroje, které jsou kontrolovány vládou dané země. Z důvodu zachycení rozdílů na dostupných zdrojích příjmů a vládní schopnost rozpoznat sazby při niž jsou vybírány daně a poplatky. Proto se je analytici snaží izolovat jako sadu provozních výnosů, tržeb a zdrojů, které jsou přímo řízeny vládou.

Mezivládní převody/Provozní příjmy (Intergovernmental Transfers/Operating Revenue) - mezivládní převody vyjadřují všechny provozní výnosy převedené z vyšších úrovní vlády jako je daňová podpora nebo sdílení nákladů na jednotlivé kategorie výdajů.

Vyčleněné příjmy/Provozní výnosy (Earmarked Revenue/Operating Revenue) - v některých případech vlády jde o mezivládní přijímání nebo vybírání výnosů, které mohou být použity pouze k pokrytí specifických výdajů. Vyčleněné příjmy, jedná se o příjmy, které jsou určeny pro specifické účely jiné než týkající se oblasti dluhu.

Úrokové platby/Provozní výnosy (Interest Payments/Operating Revenue) - vyjadřují schopnost provádět platby úroků pomocí provozních příjmů a podílů na provozních výnosech. Pokud dojde ke zvýšení úrokových sazeb, musí se zvýšit objem příjmů při, kterém analytici hodnotí dopad na dluhové závazky.

Dluhová služba/Celkové příjmy (DebtService/Total Revenue) -do čitatele se přidají splátky jistiny a jmenovatel zahrnuje všechny kapitálové výnosy, převody ze státních podniků a jiných subjektů. Ve jmenovateli je tedy zachycován celkový dopad závazků dluhu na vládní příjmy.

Aktuální finanční přebytek/Celkový příjem (AccrualFinancingSurplus/Total Revenue) - analytici Moody´s vypočítají přebytek na aktuální financování, které měří celkové výnosy (provozní i kapitálové) snížené o celkové náklady. Aktuální finanční přebytek v dlouhém období by se měl projevit snížením emisí dluhu a snížením úrovně dluhu.

Peněžní přebytek/Celkové příjmy (Cash Financing Surplus/Total Revenue) - u tohoto poměru představuje peněžní přebytek hotovost potřebnou pro provozní a investiční aktivity. Tento přebytek by měl vést ke snížení úrovně dluhu, zatímco velké nároky na peněžní prostředky mohou znamenat fiskální nerovnováhu vedoucí ke zvýšení úrovně dluhu.

Hrubá výpůjční potřeba/Celkový příjem (Gross Borrowing Need/Total Revenue) - hrubá výpůjční potřeba se skládá z čisté výpůjční potřeby vlády a rozšířená o splátky a odkupy státních dluhopisů v daném roce, splátky půjček EIB, zpětné odkupy a výměny státních dluhopisů splatných v dalších letech a přecenění rezerv financování.

Celkové výdaje na obyvatele (Total Expenditures per capita) - ačkoli úroveň fiskálního bohatství se liší mezi subjekty, vyjádření celkových výdajů na jednoho obyvatele nabízí srovnatelnou velikost vlády. Mezinárodní srovnání probíhá většinou přednostně mezi podobnými hospodářskými a finančními profily.

Celkové výdaje/HDP (Total Expenditures/GDP) - výsledek vyjadřuje, zda rozpočtová politika byla rozšiřující nebo zužující.

Primární operační bilance/Provozní výnosy (Primary Operating Balance/Operating Revenue) - slouží k měření celkových provozních výnosů po odečtení celkových provozních výdajů. Výhodná pro subjekty, které byly amortizovány hmotným kapitálovým aktivem, kde je vyloučena částka odpovídající amortizaci. Poměr vyjadřuje schopnost vlády generovat přebytky z běžných a opakujících se operací, které jsou poté k dispozici k pokrytí dluhu.

Hrubé provozní bilance/Provozní výnosy (GrossOperating Balance/Operating Revenue) - tato hrubá provozní bilance navazuje na předchozí poměr tím, že zahrnuje úrokové platby.

Čistá provozní bilance/provozní výnos (NetOperatingBalance/Operating Revenue) - čistá provozní bilance vychází z předchozího poměru zahrnující platby jistiny. Vlády, které rozlišují mezi provozními a kapitálovými rozpočty, pravidelně refinancují splatnost dluhu a mají kladné čisté provozní zůstatky, naznačují silnou schopnost samofinancování.

Poměr samofinancování (Self-financing Ratio) - označuje schopnost financovat plánované investice z vlastních zdrojů. Patří do ukazatelů založených na cash-flow. Poměr samofinancování se vypočítá z hrubé provozní bilance + kapitálové příjmy a to celé děleno celkovými kapitálovými výdaji. Poměr samofinancování s hodnotou menší než jedna vyjadřuje, že je třeba si vypůjčit, aby byly splněny požadavky kapitálového rozpočtu. Pro vlády, které rozlišují mezi provozními a investičními rozpočty se měří pomocí tohoto poměru míra tlaku na celkovou finanční způsobilost.

Kapitálové výdaje/Celkové výdaje (Capital Spending/Total Expenditures) - v tomto případě patří do kapitálových výdajů libovolné provozní výdaje. Akorát je zde limitována vláda, která nemůže odložit kapitálové výdaje, protože by to mělo negativní vliv na hospodářský růst. Tento poměr slouží k indikaci kapitálových výdajových trendů.

5ti leté celkové tržby (5 -year Total Revenue - 5 -year Total Expenditures) - jedná se o rozdíl celkových příjmů a celkových výdajů. Kladná hodnota říká, že růst tržeb předběhl růst výdajů a fiskální rovnováha byla zachována. Záporná hodnota naznačuje, že růst výdajů předběhl růst tržeb a vzniká fiskální nerovnováha.

Čistý pracovní kapitál/Celkové výdaje (Net Working Capital/Total Expenditures) - pojem čistý pracovní kapitál je definován jako krátkodobá aktiva včetně hotovosti, peněžních ekvivalentů a krátkodobých obchodovatelných cenných papírů mínus krátkodobé závazky

včetně krátkodobého dluhu a splatných závazků. Tento poměr poskytuje informaci o potřebě přístupu na trh v krátkém období.

5.2. Předzpracování dat a statistická analýza

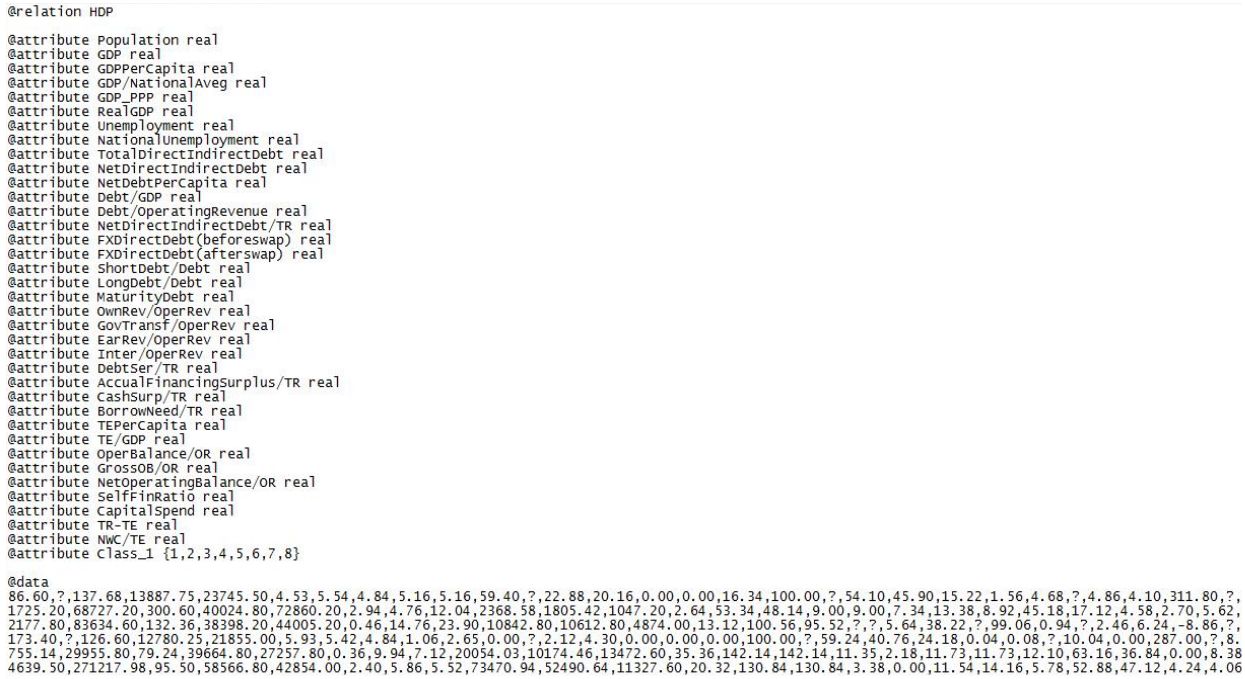
Na datech před použitím došlo k ošetření chybějících hodnot u většiny atributů. Pro nahrazení chybějících hodnot bylo použito symbolu "?". V příloze A je možnost nahlédnutí na již ošetřené vstupní hodnoty s chybějícími hodnotami.

Statická analýza tohoto datového souboru byla řešena v prostředí softwaru IBM SPSS Modeler. U spojitých proměnných došlo k výpočtu aritmetického a geometrického průměru, součtu hodnot, minima, maxima, rozpětí, mezikvartilového rozpětí a mediánu. V příloze B je možnost nahlédnutí do statistické analýzy získané z IBM SPSS Modeler.

5.3. Práce s daty

Po ukončení fáze předzpracování dat byla datová sada pro metody klasifikace rozdělena na trénovací a testovací množinu. Data byla pro statistické porovnání a získání nezkreslených výsledků 10 - krát náhodně zamíchána a rozdělena v poměru 70:30. Trénovací množiny obsahují 179 instancí, na kterých se daný klasifikační algoritmus učila testovací množina obsahuje 77 instancí pro testování přesnosti daného modelu.

Takto předpřipravená data byla v programu MS Excel převedena do požadovaného textového datového formátu ARFF (Attribute-Relation File Format), se kterým umí program Weka pracovat. Ukázka formátu ARFF pro program Weka se kterým je pracováno v predikčních modelech viz Obrázek 18.



```
@relation HDP

@attribute Population real
@attribute GDP real
@attribute GDPPerCapita real
@attribute GDP/NationalAveg real
@attribute GDP_PPP real
@attribute RealGDP real
@attribute Unemployment real
@attribute NationalUnemployment real
@attribute TotalDirectIndirectDebt real
@attribute NetDirectIndirectDebt real
@attribute NetDebtPerCapita real
@attribute Debt/GDP real
@attribute Debt/OperatingRevenue real
@attribute NetDirectIndirectDebt/TR real
@attribute FXDirectDebt(beforeswap) real
@attribute FXDirectDebt(afterswap) real
@attribute ShortDebt/Debt real
@attribute LongDebt/Debt real
@attribute MaturityDebt real
@attribute OwnRev/OperRev real
@attribute GovTransf/OperRev real
@attribute EarRev/OperRev real
@attribute Inter/OperRev real
@attribute DebtSer/TR real
@attribute AccualFinancingSurplus/TR real
@attribute CashSurp/TR real
@attribute BorrowNeed/TR real
@attribute TEPerCapita real
@attribute TE/GDP real
@attribute OperBalance/OR real
@attribute GrossOB/OR real
@attribute NetOperatingBalance/OR real
@attribute SelfFinRatio real
@attribute CapitalSpend real
@attribute TR-TE real
@attribute NWC/TE real
@attribute Class_1 {1,2,3,4,5,6,7,8}

@data
86.60,?,137.68,13887.75,23745.50,4.53,5.54,4.84,5.16,5.16,59.40,?,22.88,20.16,0.00,0.00,16.34,100.00,?,54.10,45.90,15.22,1.56,4.68,?,4.86,4.10,311.80,?,
1725.20,68727.20,300.60,40024.80,72860.20,2.94,4.76,12.04,2368.58,1805.42,1047.20,2.64,53.34,48.14,9.00,9.00,7.34,13.38,8.92,45.18,17.12,4.58,2.70,5.62,
2177.80,83634.60,132.36,38398.20,44005.20,0.46,14.76,23.90,10842.80,10612.80,4874.00,13.12,100.56,95.52,?,?,5.64,38.22,?,99.06,0.94,?,2.46,6.24,-8.86,?,
173.40,?,126.60,12780.25,21855.00,5.93,5.42,4.84,1.06,2.65,0.00,?,2.12,4.30,0.00,0.00,0.00,100.00,?,59.24,40.76,24.18,0.04,0.08,?,10.04,0.00,287.00,?,8.
755.14,29955.80,79.24,39664.80,27257.80,0.36,9.94,7.12,20054.03,10174.46,13472.60,35.36,142.14,142.14,11.35,2.18,11.73,11.73,12.10,63.16,36.84,0.00,8.38
4639.50,271217.98,95.50,58566.80,42854.00,2.40,5.86,5.52,73470.94,52490.64,11327.60,20.32,130.84,130.84,3.38,0.00,11.54,14.16,5.78,52.88,47.12,4.24,4.06
```

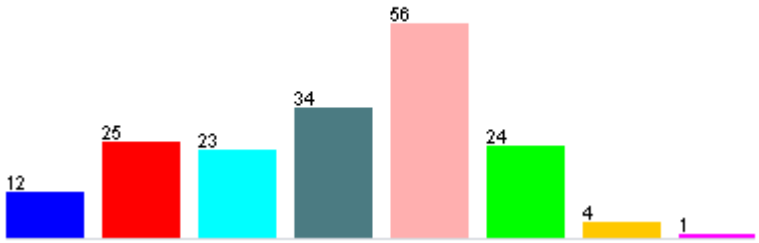
Obrázek 18: Ukázka souboru ARFF

Zdroj: vlastní zpracování

Z důvodu nerovnováhy zastoupení některých tříd musely být všechny trénovací soubory v programovém prostředí WEKA vyrovnány pomocí převzorkovací techniky SMOTE (**S**ynthetic **M**inority **O**versampling **TE**chnique). Jedná se o metodu odběru vzorků, kdy vyrovnáním výstupních tříd je dosaženo vyšší přesnosti klasifikace. Obecně je trénování nad nevyváženými daty velmi zkoumané téma, jelikož takových případů se v reálném světě nachází mnoho [28]. Vyrovnání vzorků je třeba udělat pouze na trénovacích datech, aby nedošlo ke zkreslení přesnosti predikčního modelu.

Trénovací data byla vyrovnána pomocí SMOTE na hodnotu 56 podle třídy 5 s pomocí nastavování parametrů:

- vyrovnávaná třída (index třídy, na kterou má být SMOTE aplikováno) → 1-8,
- nebližší sousedé (počet nejbližších sousedů, které mají být použity) → pro třídy 1-6 byl nastaven počet nejbližších sousedů na hodnotu 5 a pro třídu 7 (obsahuje jen 4 případy) na hodnotu 3. V případě třídy 8 nemohlo být SMOTE použito, jelikož obsahuje pouze jeden případ,
- procento (procento instancí, které chceme vytvořit) → 0-100 %,
- a náhodné vzorkování → 1.



Obrázek 19: Ukázka nevyrovnaných trénovacích dat bez SMOTE

Zdroj: vlastní zpracování

5.4. Návrh modelu

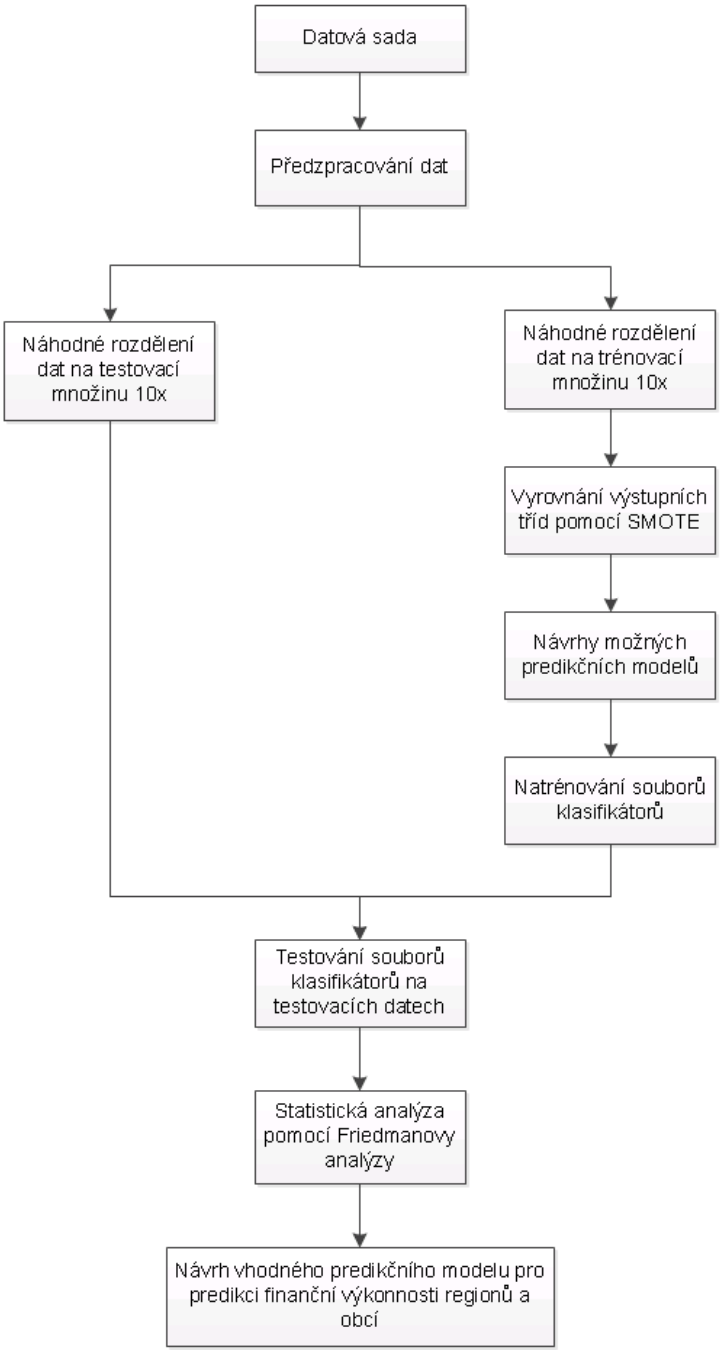
Pro účely klasifikace bylo použito celkem 4 základních klasifikátorů:

- NaiveBayes,
- metoda Podpůrných vektorových strojů (SMO),
- neuronová síť typ Multi Layer Perceptron (MLP),
- rozhodovací strom J48

a 4 meta klasifikátorů (použito stejné nastavení základních klasifikátorů):

- metoda AdaBoost,
- metoda Bagging,
- Random Forest,
- metoda Stacking.

Koncept modelu pro predikci je znázorněn na Obrázek 20.



Obrázek 20: Koncept výběru modelu pro predikci finanční výkonnosti obcí a regionů

Zdroj: vlastní zpracování

Nastavení pro klasifikátor SMO:

- jádrová funkce → Puk,
- parametr komplexnosti *c*→ 8,
- parametr tolerance → 0,001.

Nastavení neuronové sítě MLP:

- neuronů ve skryté vrstvě→ 50,
- trénovací čas → 1000,
- rychlost učení → 0,35,
- momentum → 0,25.

Nastavení rozhodovacího stromu J48:

- minimální počet instancí → 5,
- prořezávání → ano,
- faktor spolehlivosti → 0,25.

Nastavení AdaBoost:

- klasifikátor → SMO,
- počet iterací → 15.

Nastavení Bagging:

- klasifikátor → SMO,
- velikost trénovací sady → 110 %,
- počet stromů → 8,
- počet iterací → 15.

Nastavení stromu Random Forest:

- maximální hloubka stromu → bez omezení,
- počet iterací → 500.

Nastavení meta algoritmu Stacking:

- klasifikátory → NaiveBayes, SMO, MLP a J48,
- meta klasifikátor → Random Forest.

Výše popsané parametry nastavení klasifikačních modelů byly vybrány jako nejlepší, při jejich změně se průměrné náklady na klasifikaci zvyšovaly. Dále bylo vyzkoušeno nastavení i ostatních parametrů, které daný klasifikátor nabízel, ale tato změna nastavení

snižovala průměrné náklady. Pro metody AdaBoost a Bagging bylo vyzkoušeno dalších několik základních klasifikátorů, ale nejlépe při vyhodnocení z nich dopadl algoritmus SMO.

5.5. Použitý software

Pro měření klasifikace bylo využito programového prostředí softwaru Weka ve verzi 3.6 a 3.8 dle studie [8],[18] (**W**aikato **E**nviroment for **K**nowledge **A**nalysis), který byl vyvinut na univerzitě Waikato v Hamiltonu na Novém Zélandu. Celý software je sepsán v programovacím jazyce Java a distribuován pod licencí GNU Public licence. Jedná se o nástroj sloužící pro tzv. "dolování dat" (datamining), podporující předzpracování dat, vizualizaci, výběr atributů, klasifikaci, predikci, shlukování a tvorbu asociačních pravidel [19].

Pro porovnání výsledků klasifikace bylo použito Friedmanovy ANOVY a Wilcoxonova párového testu v programovém prostředí Statistica 12.

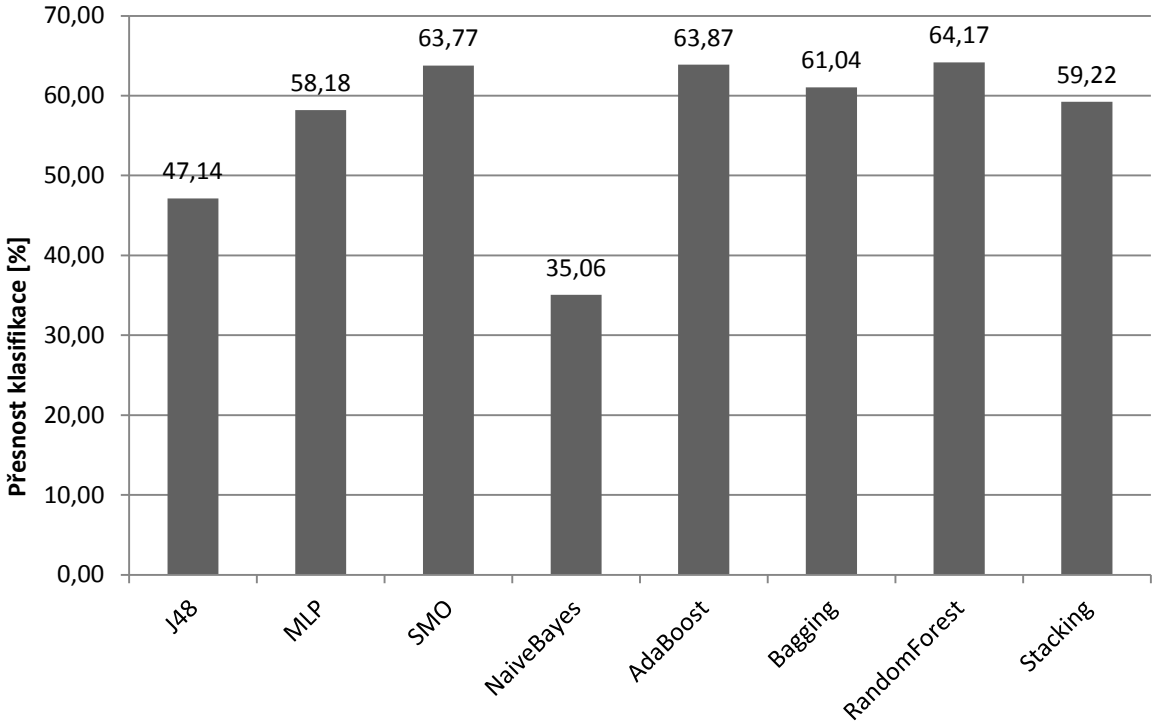
6. VÝSLEDKY PREDIKCE

Výsledky klasifikace provedené v programu WEKA pro základní a meta algoritmy jsou uvedeny v příloze C. Tyto výsledné tabulky odpovídají naměřeným hodnotám na testovací množině dat. Pro samotné vyhodnocení výsledků bylo využito grafického aparátu, kde je porovnáváno:

- procento správně klasifikovaných instancí (přesnost klasifikace),
- průměrné náklady na klasifikaci,
- plocha pod křivkou ROC.

6.1. Přesnost klasifikace

Procento správně klasifikovaných instancí udává, kolik případů v přepočtu na procenta klasifikátor správně zařadit do výstupní třídy na testovací množině dat. Nejlepší procentuální úspěšností pro základní algoritmy bylo dosaženo u metody podpůrných vektorových strojů SMO (SVM) s průměrnou přesností klasifikace 63,77 % a pro meta klasifikátory u Random Forest s přesností 64,17 %. Nejhoršího přesnosti bylo dosaženo u klasifikátoru NaiveBayes s úspěšností 35,06 % viz Graf 1.

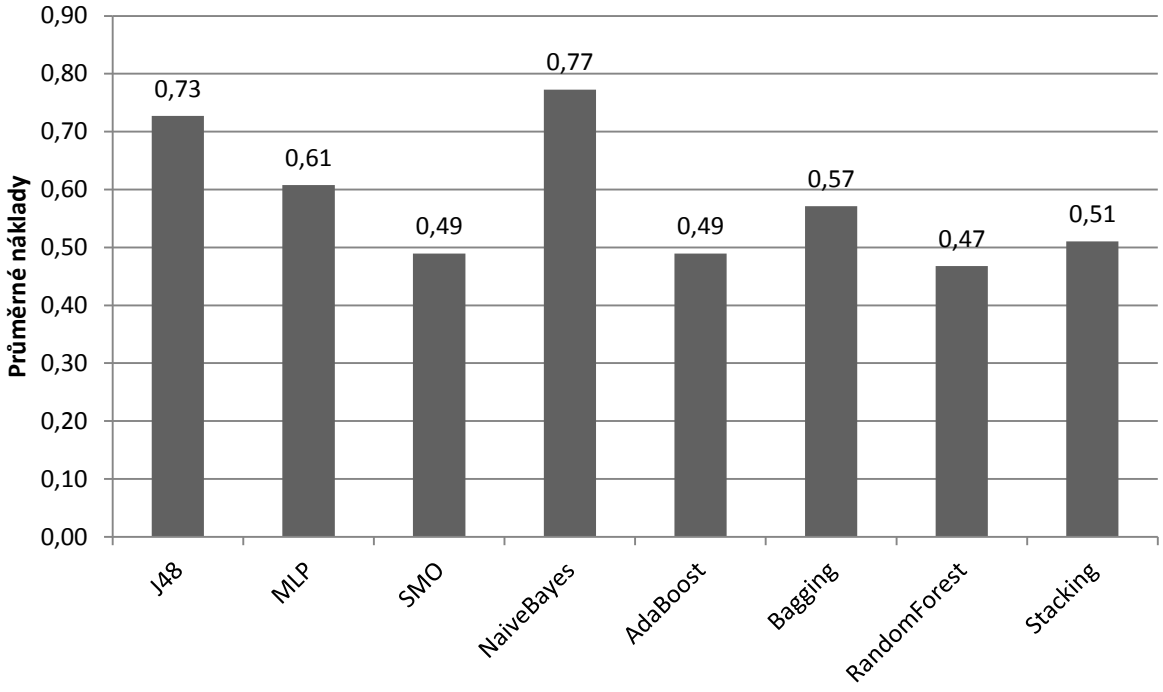


Graf 1: Průměrná přesnost klasifikace

Zdroj: vlastní zpracování

Průměrné náklady na klasifikaci jsou v tomto případě měření lepším porovnávacím znakem přesnosti klasifikace než procento správně klasifikovaných instancí. Z průměrných nákladů na klasifikaci můžeme říct o kolik se skutečná třída liší od cílové třídy. Procentuální přesnost klasifikace je vhodným ukazatelem spíše pro výstupní dvě třídy.

Čím jsou náklady na klasifikaci menší, tím je daný klasifikátor úspěšnější. Nejnižších průměrných nákladů pro základní algoritmy bylo dosaženo u SMO (průměrné náklady 0,49) a pro meta algoritmy u Random Forest (průměrné náklady 0,47). Nejlepším klasifikátorem ze všech použitých se stal Random Forest a nejhorším NaiveBayes (průměrné náklady 0,77) viz Graf 2.

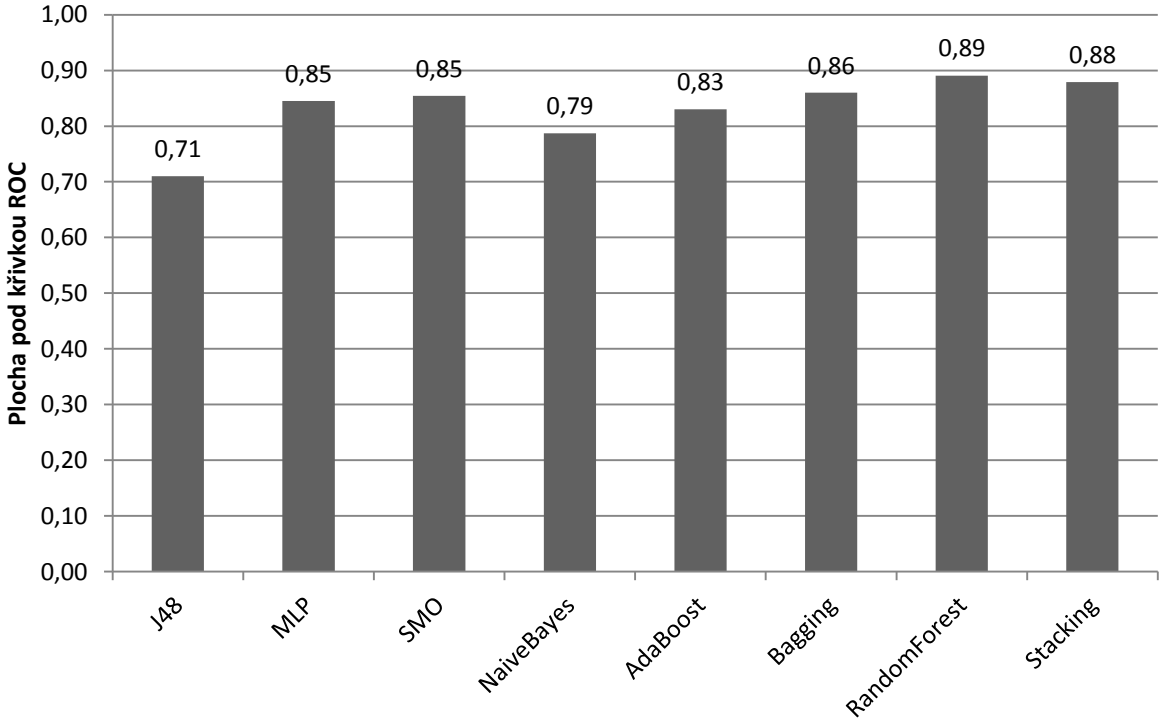


Graf 2: Průměrné náklady na klasifikaci

Zdroj: vlastní zpracování

Velikost plochy pod křivkou ROC (AUC) udává účinnost klasifikace na testovacích datech, přičemž hodnota AUC se může pohybovat v hodnotách 0,5 - 1. Čím vyšší je tato hodnota, tím lepší je účinnost klasifikátoru.

Nejvyšší hodnoty plochy pod křivkou u základních algoritmů bylo naměřeno u klasifikátorů MLP a SMO se stejnou velikostí plochy 0,85 a u meta algoritmů u klasifikátoru Random Forest s velikostí plochy pod křivkou ROC 0,89 viz Graf 3. Nejhoršího naměřeného výsledky plochy pod křivkou bylo dosaženo u základního algoritmu J48.



Graf 3: Plocha pod křivkou ROC

Zdroj: vlastní zpracování

6.2. Statistické porovnání výsledků

Pro zjištění rozdílů mezi klasifikačními algoritmy bylo využito neparametrického testu (na reálných datech není obvykle možné předpokládat normální rozdělení u naměřených výsledků) pro dvoufaktorovou analýzu rozptylu tzv. Friedmanovy ANOVY. Na základě tohoto rozhodnutí byla stanovena hypotéza *H1: Předpokládáme, že výsledky přesnosti klasifikace & průměrných nákladů& plochy pod křivkou ROC se pro dané algoritmy liší*.

***H1*$_0$**: Předpokládáme, že mezi klasifikačními algoritmy neexistují statisticky významné rozdíly.

***H1*$_1$**: Předpokládáme, že mezi klasifikačními algoritmy existují statisticky významné rozdíly.

Tabulka 9: Porovnání všech algoritmů pomocí Friedmanovy ANOVY

Algoritmy	**Průměrné pořadí pro přesnost klasifikace** (Chi2 = 46.35, *p* = 0.000)	**Průměrné pořadí pro průměrné náklady** (Chi2 = 46.79, *p* = 0.000)	**Průměrné pořadí pro ROC plochu** (Chi2 = 52.18, *p* = 0.000)
J48	7,10	6,70	7,80
MLP	4,65	5,35	4,60
SMO	2,80	2,85	3,90
NaiveBayes	7,70	7,60	7,10
AdaBoost	2,65	2,95	4,50
Bagging	4,05	5,10	4,10
Random Forest	2,60	2,10	1,70
Stacking	4,45	3,35	2,30

Zdroj: vlastní zpracování

Hypotéza byla testována na hladině významnosti 0,000 a kritickou hodnotu testovacího kritéria (*Chi2*) 46,79. V tomto případě zamítáme nulovou hypotézu $H1_0$ a přijímáme alternativní hypotézu $H1_1$. Sloupce průměrného pořadí pro zkoumané oblasti Tabulka 9 určují průměrné pořadí naměřených hodnot a zobrazují nejmenší a největší odlišnosti mezi klasifikátory. Nejmenšího statisticky významného rozdílu tzn. největší přesnosti bylo zjištěno u klasifikátoru Random Forest s nejnižším průměrným pořadím u všech měřítek. Naproti tomu největšího rozdílu tzn. nejmenší přesnosti bylo zjištěno u klasifikátoru NaiveBayes viz Tabulka 9.

V dalším kroku byl použit Wilcoxonův párový test pro porovnání jednotlivých klasifikátorů navzájem. Jedná se o neparametrický test, který porovnává párové hodnoty výběrového souboru. V tomto případě byly porovnávány mezi sebou meta klasifikátory v závislosti na výsledku přesnosti klasifikace, průměrných nákladů a plochy pod křivkou ROC na různých hladinách významnosti viz Tabulka 10. Hodnota "Z" vyjadřuje velikost rozdílu mezi dvěma klasifikačními algoritmy. V případě algoritmu Random Forest došlo ke statisticky lepším výsledkům než u ostatních algoritmů v případě plochy pod ROC křivkou. U průměrných nákladů nebyl výsledek statisticky významný u SMO a AdaBoost. To znamená, že tyto algoritmy nebyly statisticky významně překonány. Podobně pak u přesnosti navíc Bagging také nebyl metodou Random Forest významně překonán. I další výsledky ukazují, že meta algoritmy většinou dokázaly jednoduché klasifikátory překonat.

Tabulka 10: Test významnosti přesnosti klasifikace, průměrných nákladů a plochy pod křivkou ROC pomocí Wilcoxonova párového testu

	Přesnost klasifikace			Průměrné náklady			Plocha pod křivkou ROC		
Algoritmy	***Z* - hodnota**	***p* - hodnota**		***Z* - hodnota**	***p* - hodnota**		***Z* - hodnota**	***p* - hodnota**	
RF vs. J48	2 ,803	0 ,005	***	2 ,803	0 ,005	***	2 ,803	0 ,005	***
RF vs. MLP	1 ,784	0 ,074	*	2 ,701	0 ,007	***	2 ,701	0 ,007	***
RF vs. SMO	0 ,415	0 ,678		0 ,866	0 ,386		2 ,497	0 ,013	**
RF vs. NB	2 ,803	0 ,005	***	2 ,803	0 ,005	***	2 ,803	0 ,005	***
RF vs. AB	0 ,474	0 ,636		0 ,866	0 ,386		2 ,701	0 ,007	***
RF vs. Bag	0 ,764	0 ,445		2 ,497	0 ,013	**	2 ,650	0 ,008	***
RF vs. Stack	2 ,521	0 ,012	**	2 ,293	0 ,022	**	2 ,497	0 ,013	**
Bag vs. J48	2 ,803	0 ,005	***	2 ,395	0 ,017	**	2 ,803	0 ,005	***
Bag vs. MLP	1 ,185	0 ,236		1 ,070	0 ,285		1 ,784	0 ,074	
Bag vs. SMO	2 ,310	0 ,021	**	2 ,803	0 ,005	***	1 ,274	0 ,203	
Bag vs. NB	2 ,803	0 ,005	***	2 ,803	0 ,005	***	2 ,803	0 ,005	***
Bag vs. AB	2 ,369	0 ,018	**	2 ,803	0 ,005	***	1 ,784	0 ,074	
Bag vs. RF	0 ,764	0 ,445		2 ,497	0 ,013	**	2 ,650	0 ,008	***
Bag vs. Stack	0 ,770	0 ,441		2 ,039	0 ,041	**	2 ,242	0 ,025	**
AB vs. J48	2 ,803	0 ,005	***	2 ,803	0 ,005	***	2 ,803	0 ,005	***
AB vs. MLP	1 ,988	0 ,047	**	2 ,395	0 ,017	**	0 ,652	0 ,515	
AB vs. SMO	-	1 ,000		-	1 ,000		-	1 ,000	
AB vs. NB	2 ,803	0 ,005	***	2 ,803	0 ,005	***	1 ,784	0 ,074	
AB vs. Bag	2 ,369	0 ,018	**	2 ,803	0 ,005	***	1 ,784	0 ,074	
AB vs. RF	0 ,474	0 ,636		0 ,866	0 ,386		2 ,701	0 ,007	***
AB vs. Stack	1 ,580	0 ,114		0 ,459	0 ,646		2 ,497	0 ,013	**
Stack vs. J48	2 ,803	0 ,005	***	2 ,701	0 ,007	***	2 ,803	0 ,005	***
Stack vs. MLP	0 ,561	0 ,575		2 ,192	0 ,028	**	2 ,599	0 ,009	***
Stack vs. SMO	1 ,580	0 ,114		0 ,459	0 ,646		2 ,191	0 ,028	**
Stack vs. NB	2 ,803	0 ,005	***	2 ,803	0 ,005	***	2 ,803	0 ,005	***
Stack vs. AB	1 ,580	0 ,114		0 ,459	0 ,646		2 ,497	0 ,013	**
Stack vs. Bag	0 ,770	0 ,441		2 ,039	0 ,041	**	2 ,242	0 ,025	**
Stack vs. RF	2 ,521	0 ,012	**	2 ,293	0 ,022	**	2 ,497	0 ,013	**

* na hladině významnosti p <0 .1 , ** p <0 .05, *** p <0 .01

Zdroj: vlastní zpracování

ZÁVĚR

Tato diplomová práce se zabývá hodnocením finanční výkonnosti regionů a jeho predikcí. Cílem práce bylo navrhnout model použitelný pro predikci na základě dostupné datové sady. Pro trénování a testování souborů klasifikátorů bylo použito vstupních atributů za roky 2011-2015 a výstupní rating z roku 2016. Pro přesnější a spolehlivější měření výsledků byl soubor náhodně 10 - krát rozdělen v poměru 70:30 na trénovací a testovací sadu. Následně z důvodu nevyrovnaných instancí ve výstupní třídě byly počty instancí v třídách vyrovnány na hodnotu 56 pomocí vzorkovacího filtru SMOTE. Výsledek predikce byl hodnocen podle přesnosti, průměrných nákladů klasifikace a plochou pod ROC křivkou. Porovnání klasifikátorů ukázalo, že Random Forest poskytl nejpřesnější predikci ratingu regionů. Použití souborů klasifikátorů tedy vedlo ke snížení rozptylu a systémové chyby oproti použitým základním klasifikátorům.

Analýza výsledků klasifikačních modelů zahrnovala procentuální schopnost modelů klasifikovat objekty, průměrné náklady na klasifikaci a plochu pod křivkou ROC. V případě základních klasifikátorů bylo ve všech těchto výsledných statistikách nejlepších výsledků dosaženo u základního klasifikátoru SMO(SVM) s hodnotou přesnosti klasifikace 63,77 %, průměrnými náklady 0,49 a plochou pod křivkou ROC 0,85. U meta klasifikátorů bylo nejlepšího výsledku dosaženo klasifikátorem Random Forest s hodnotou přesnosti klasifikace 64,17 %, průměrnými náklady 0,47 a plochou pod křivkou ROC 0,89. Při statistickém srovnání pomocí Friedmanovy ANOVY bylo zjištěno, že nejmenšího statisticky významného rozdílu tzn. významné přesnosti se dosahuje u meta klasifikátoru Random Forest.

Vhodným modelem použitelným pro predikci finanční výkonnosti regionů je meta klasifikátor Random Forest, který dosáhl nejlepších výsledků při sledovaných kritériích týkající se přesnosti klasifikace, průměrných nákladů a plochy pod křivkou ROC. Návrhem predikčního modelu, jeho ověřením na reálných datech a doporučením metody Random Forest pro predikci finanční výkonnosti regionů došlo k naplnění cíle této práce. Hlavní nevýhodou navrženého modelu je, že byl otestován pouze na ratingu z jednoho roku. Ačkoliv je toto hodnocení relativně stabilní (nedochází k častým změnám hodnocení), je třeba výsledky v budoucnu ověřit na dalších datech. Další možnou nevýhodou se může jevit generování souboru mnoha rozhodovacích stromů, což může být výpočetně náročné. Další nevýhodou může být nižší srozumitelnost modelu oproti základním klasifikátorům, např. rozhodovacím stromům, kde je zřejmé, jak se algoritmus na základě vstupních atributů rozhoduje o klasifikaci do ratingových tříd. U metody Random Forest je těchto stromů příliš

mnoho, než aby bylo možné její postup snadno interpretovat. Výhodou tohoto modelu je možnost použití i pro jiné ratingové agentury než je Moody´s, ratingové agentury se liší pouze ve výsledném označení ratingu a navíc se jejich hodnocení většinou shoduje. Mezi možné výhody patří schopnost modelu vytřídit důležité a nedůležité atributy vhodné pro predikci či možnost pracovat se spojitými i kategorickými atributy. Tato schopnost výběru atributů je již vnořena v rozhodovacích stromech (používají se atributy s nejlepší schopností klasifikovat data). Ke zvýšení přesnosti predikce by také byla vhodná datová sada bez chybějících hodnot u některých atributů.

Tento model pro predikci finanční výkonnosti regionů a obcí je možné použít jako spolehlivý nástroj na podporu při rozhodování o investicích, případně je možné pomocí něho odhalit finanční potíže, které by mohly nastat a následně těmto potížím včas předejít. Model je možné využít i při rozhodování o přidělování externích zdrojů regionům. Vybraný predikční model Random Forest šetří finanční i časové náklady vlivem simulace rozhodnutí expertů provádějících ratingové hodnocení.

POUŽITÁ LITERATURA

[1] *A short history of Standard & Poor's: Q&A* [online]. [cit. 2017-06-16]. Dostupné z: http://www.telegraph.co.uk/finance/financialcrisis/8937653/A-short-history-of-Standard-and-Poors-QandA.html.

[2] *Analytics-driven embedded systems, part 2 – Developing analytics and prescriptive controls. Embedded Computing Design* [online]. 14.03.2016 [cit. 2017-06-16]. Dostupné z: http://embedded-computing.com/articles/analytics-driven-embedded-systems-part-2-developing-analytics-and-prescriptive-controls.

[3] ATKINSON, Peter M.; TATNALL, A. R. L. Introduction neural networks in remote sensing. *International Journal of Remote Sensing*, 1997, roč. 18, č. 4, s. 699-709.

[4] BARESA, Suzana; BOGDAN, Sinisa; IVANOVIC, Sasa. Role, interests and critics of credit rating agencies. *UTMS Journal of Economics*, 2012, roč. 3, č. 1, s. 71-82.

[5] BAUER, Eric; KOHAVI, Ron. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 1999,roč. 36, č. 1, s. 105-139.

[6] BERKA, Petr. *Dobývání znalostí z databází*. Vyd. 1. Praha: Academia, 2003. 366s. ISBN 80-200-1062-9.

[7] *Biologické algoritmy (6) - Neuronové sítě* [online]. [cit. 2017-06-17]. Dostupné z: https://www.root.cz/clanky/biologicke-algoritmy-6-neuronove-site.

[8] BOUCKAERT, Remco R., et al. *WEKA manual for version 3-7-3*. Waikato: *The University of Waikato, New Zealand*, 2010, s. 588-595.

[9] *Classification Algorithm*. *Television Commercial Detection* [online]. John H. Marcoux [cit. 2017-06-17]. Dostupné z: http://www-personal.umich.edu/~johnhugo/commercial/design.htm.

[10] COHEN, William W. *Stacked sequential learning*. Pittsburgh: Carnegie-Mellon University 2005.

[11] CRIMINISI, A.; SHOTTON, J.; KONUKOGLU, E. Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, 2011, roč. 5, č. 6, s. 12.

[12] CRIMINISI, Antonio; SHOTTON, Jamie (ed.). *Decision forests for computer vision and medical image analysis*. Berlin : Springer Science & Business Media, 2013.

[13] DUMAIS, Susan; CHEN, Hao. Hierarchical classification of Web content. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2000. s. 256-263.

[14] *Ensemble Learning* [online]. [cit. 2017-06-27]. Dostupné z: https://kgpdag.wordpress.com.

[15] *Finanční kondice vybrané obce* [online]. Brno, 2016 [cit. 2017-06-16]. Dostupné z: https://is.muni.cz/th/370259/esf_m/DP_-_Martincova_Tereza.pdf.

[16] *Finanční zdraví obcí a jeho regionální diferenciace* [online]. Brno, 2012 [cit.2017-06-16]. Dostupné z: https://is.muni.cz/th/42351/esf_d /DisP_Oplustilova_final.pdf. Disertační práce. MASARYKOVA UNIVERZITA.

[17] GAILLARD, Norbert. The determinants of Moody's sub-sovereign ratings. *International Research Journal of Finance and Economics*, 2009, roč. 31, č. 1, s. 194-209.

[18] HALL, Mark, et al. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 2009, roč. 11, č. 1, s. 10-18.

[19] HALL, Mark, et al.:*Weka 3: Data mining software in java*. Waikato: The University of Waikato, 2009. Dostupné z http://www. cs. waikato. ac. nz/~ ml/weka.

[20] HANLEY, James A.; MCNEIL, Barbara J. *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. *Radiology*, 1982, roč. 143, č. 1, s. 29-36.

[21] HOLEČKOVÁ, J. *Finanční analýza firmy*. 1. vyd. Praha: ASPI, 2008, 208 s. ISBN 978-80-7357-392-8.

[22] HSU, Chih-Wei, et al. A practical guide to support vector classification. Taiwan: National Taiwan University, 2003.

[23] CHAUVIN, Yves; RUMELHART, David E. (ed.). *Backpropagation: theory, architectures, and applications*. Hillsday: Psychology Press, 1995.

[24] CHENG, Bing; TITTERINGTON, D. Michael. Neural networks: A review from a statistical perspective. *Statistical Science*, 1994, roč. 9, č. 1, s. 2-30.

[25] Jak připravovat rozpočet obce. *Moderní obec* [online]. 2012, **18**(9) [cit. 2017-06-16]. Dostupné z: http://moderniobec.cz/jak-pripravovat-rozpocet-obce.

[26] JOACHIMS, Thorsten. *Making large-scale SVM learning practical*. Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.

[27] KAMENÍČKOVÁ, Věra. *Je finanční analýza nutná?* Deník veřejné správy [online]. 1999, 1999(5) [cit. 2017-06-16]. SSN 1213-6336. Dostupné z: http://www.dvs.cz/clanek.asp?id=23356.

[28] KATORE, Lokesh S; UMALE, J. S. Comparative study of recommendation algorithms and systems using WEKA. *International Journal of Computer Applications*, 2015, roč. 110, č. 3, s. 14-17.

[29] KRAFTOVÁ, Ivana. Finanční analýza municipální firmy. Praha: C.H. Beck, 2002. C.H. Beck pro praxi. ISBN 80-717-9778-2.

[30] LEUNG, K. Ming. Naive bayesian classifier. New York: *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007.

[31] MCCALLUM, Andrew, et al. A comparison of event models for naive bayes text classification. In: *AAAI-98 workshop on learning for text categorization*. 1998. s. 41-48.

[32] *Moody's Corporation. Moody's History*. [online]. [cit. 2017-06-16]. Dostupné z: https://www.moodys.com:443/Pages/atc001.aspx.

[33] MURPHY, Kevin P. *Naive bayes classifiers*. Vancouver: University of British Columbia, 2006.

[34] NANNI, Loris; LUMINI, Alessandra. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 2009, roč. 36, č. 2, s. 3028-3033.

[35] PAVLÍK, Marek. Jak úspěšně řídit obec a region: cíle, nástroje, trendy, zahraniční zkušenosti. 1. vyd. Praha: Grada, 2014. ISBN 978-80-247-5256-3.

[36] PETR, Pavel. *Metody data miningu*. Vyd. 1. Pardubice: Univerzita Pardubice, 2014. ISBN 978-80-7395-872-5.

[37] PLATT, John. *Sequential minimal optimization: A fast algorithm for training support vector machines*. Technical Report MSR-TR98-14, 1998.

[38] *Playing with Support Vector Machines (SVM)*. LinkedIn [online]. [cit. 2017-06-16]. Dostupné z: https://www.linkedin.com/pulse/playing-support-vector-machines-svm-s-hiring-data-scientists.

[39] PROVAZNÍKOVÁ, Romana. *Financování měst, obcí a regionů: teorie a praxe*. 2. aktualiz. a rozš. vyd. Praha: Grada, 2009, 304 s. ISBN 9788024727899.

[40] QUINLAN, J. Ross, et al. Bagging, boosting, and C4. 5. In: *AAAI/IAAI, Vol. 1*. 1996. s. 725-730.

[41] *Random Forest* [online]. 1-2 [cit.2017-06-16]. Dostupné z: http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf.

[42] *Rating*. Ministerstvo financí ČR [online]. Praha 1, 2005 [cit. 2017-06-16]. Dostupné z: http://www.mfcr.cz/cs/verejny-sektor/rizeni-statniho-dluhu/zakladni-informace/rating.

[43] *Ratingová zpráva Moravskoslezský kraj*.[online]. [cit. 2017-06-16]. Dostupné z: http://www.msk.cz/assets/verejna_sprava/msk_ratingova-zprava_cerven-2016_publikovana-verze.pdf.

[44] *Regional and Local Governments* [online]. 1-39 [cit. 2017-06-16]. Dostupné z: https://www.moodys.com/researchdocumentcontentpage.aspx?docid=PBC_147779.

[45] RIFKIN, Ryan Michael. *Everything old is new again: a fresh look at historical approaches in machine learning*. [online]. [cit. 2017-06-16], 2002. PhD Thesis. Cambridge: Massachusetts Institute of Technology.

[46] RUBINOFF, David. HANZLOVA, Katerina. *Sources and Uses of Statistical Data in Moody's Analysis of Regional and Local Governments* [online]. [cit. 2017-06-16]. Dostupné z: https://www.moodys.com/researchdocumentcontentpage.aspx?docid=PBC_147779.

[47] SUN, Yanmin, et al. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 2007, roč. 40, č. 12, s. 3358-3378.

[48] ŠÍMA, Jiří a Roman NERUDA. *Teoretické otázky neuronových sítí*. Vyd. 1. Praha: Matfyzpress, 1996. 390 s. ISBN 80-85863-18-9.

[49] TURNEY, Peter D. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 1995, roč. 2, s. 369-409.

[50] VINŠ, Petr a Václav LIŠKA. *Rating*. Praha: C.H. Beck, 2005. C.H. Beck pro praxi. ISBN 80-717-9807-X.

[51] VOLNÁ, Eva. *Neuronové sítě 1*. Ostrava, 2002. Učební texty. Ostravská Universita, Přírodovědecká fakulta.

[52] ZHANG, Guoqiang Peter. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2000, roč. 30, č. 4, s. 451-462.

[53] ZHOU, Zhi-Hua. Ensemble learning. *Encyclopedia of biometrics*, 2015, s. 411-416.

[54] ZIKMUND, Martin. *Ukazatelé likvidity: Běžná likvidita*. [online]. [cit. 2017-06-16]. Dostupné z: http://www.businessvize.cz/financni-analyza/ukazatele-likvidity.

SEZNAM PŘÍLOH

Příloha A - Ukázka dat

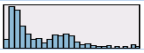
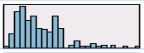
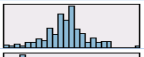
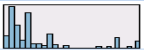
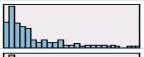
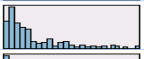
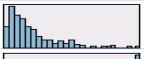
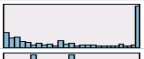
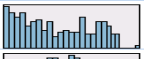
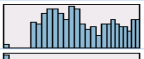
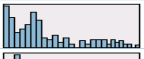
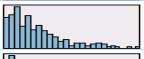
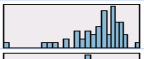
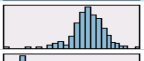
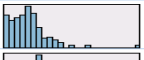
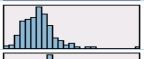
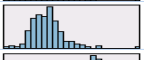
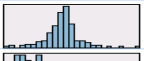
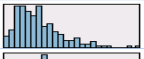
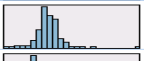
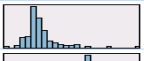
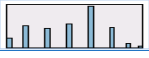
Příloha B – Popisná statistika atributů

Příloha C - Podrobné výsledky experimentů

Příloha A

GDP_PPP	P_change	hployment	ment Rate	irect Debt	irect Debt	per_capita	debt/GDP	g_revenue	l Revenue	irect Debt	irect Debt	debt/debt	debt/debt	urity_debt	/oper_rev	/oper_rev	/oper_rev	/oper_rev	/total_rev	l Revenue	surplus/TR	g_need/TR	per_capita	TE/GDP	alance/OR	alance/OR	g Revenue	cing_ratio	ending/TE	TR-TE(%)	NWC/TE	redikce_8_trid
26093,75	-1,75	10,43	11,00	2141,25	1794,25	1349,25	4,48	45,28	40,80	0,00	0,00	10,45	?	10,95	6,98	64,60	79,50	2,35	5,53	0,80	13,10	-9,83	3348,75	8,35	3,18	0,83	-2,93	?	9,88	1,13	-13,63	4
67428,25	3,60	5,08	7,14	25539,12	10848,16	2681,40	3,74	28,02	28,02	3,60	3,60	1,88	1,94	13,92	86,58	13,42	0,00	1,38	2,36	-4,22	-12,82	39,93	11016,80	14,22	3,48	2,10	?	?	8,90	-1,80	63,74	2
24957,00	0,30	33,48	23,90	30815,80	30195,80	3592,00	17,12	104,00	97,10	0,00	?	15,12	40,03	?	60,60	39,40	?	3,06	9,92	-12,70	?	19,76	4322,40	19,82	-4,02	-7,06	-14,62	0,06	11,06	-1,72	?	5
32866,00	-3,82	23,92	24,26	208,34	208,34	308,80	0,18	38,18	36,78	0,00	0,00	9,26	?	?	60,20	39,80	?	0,96	4,00	?	3,60	-0,36	812,40	0,48	4,46	3,46	0,26	1,94	3,68	-2,00	?	7
35460,20	2,76	7,38	5,94	4406,58	3085,58	2047,00	5,04	143,80	126,04	?	?	18,18	49,62	5,83	90,58	9,42	?	11,58	38,78	?	-23,70	?	1580,40	3,84	?	11,08	?	?	44,02	?	-24,76	2
23911,00	-2,00	15,70	14,36	2274,20	2274,18	9244,40	49,08	249,72	189,02	0,15	?	51,40	?	?	76,20	23,80	?	2,02	6,78	?	-3,10	8,32	5185,20	26,82	12,82	10,76	3,40	0,90	34,62	2,92	?	5
?	2,00	3,28	5,20	85069,80	74084,00	6983,40	13,58	149,42	144,62	?	?	19,64	?	?	?	19,10	?	4,26	25,64	?	-0,08	22,90	4936,80	9,52	10,76	6,50	-15,68	0,98	9,54	1,02	?	1
54759,60	0,74	15,94	23,90	1586,20	1359,20	843,40	1,78	45,64	43,12	?	?	15,66	54,74	4,26	55,52	44,48	?	1,28	7,06	5,76	?	1,72	1879,60	3,96	22,08	20,78	14,64	1,32	20,52	0,10	?	4
20999,25	-1,40	14,10	11,00	369,28	369,28	637,50	2,65	18,73	16,23	0,00	0,00	5,53	74,23	7,00	5,20	65,90	76,70	0,70	1,53	0,43	4,53	-3,63	4001,50	16,53	9,55	8,90	7,78	?	20,48	2,15	-0,68	4
44005,20	0,46	14,76	23,90	10842,80	10612,80	4874,00	13,12	100,56	95,52	?	?	5,64	38,22	?	99,06	0,94	?	2,46	6,24	-8,86	?	12,70	5721,40	14,88	1,00	-1,52	-5,50	0,30	11,56	-0,12	?	4
?	2,16	3,04	5,20	47898,80	45627,00	3634,40	6,98	78,12	74,84	?	?	13,48	?	?	?	13,26	?	2,08	8,82	?	3,44	4,18	4809,00	9,18	13,26	11,20	4,06	1,32	11,80	1,30	?	1
?	1,74	10,32	5,20	########	99235,40	29310,00	67,80	361,64	346,28	?	?	13,36	?	?	?	37,26	?	8,84	44,06	?	1,14	33,56	8579,00	19,78	12,86	4,04	-33,12	1,20	7,00	3,28	?	2
?	0,98	3,22	3,84	2762,58	1056,92	7637,80	?	99,08	95,98	?	?	26,50	?	?	45,32	43,80	4,00	5,64	22,06	0,82	?	15,74	7888,80	?	10,86	5,20	-12,18	1,74	7,14	-0,88	?	2
41979,60	0,32	16,16	23,90	3022,60	2257,00	1955,60	5,60	26,86	26,72	?	?	12,84	60,30	?	96,86	3,14	?	0,64	2,92	1,88	?	0,66	7435,80	20,34	7,06	6,42	4,10	1,38	5,18	0,46	?	4
43187,25	0,28	4,05	11,00	1687,48	369,33	718,00	1,43	5,93	5,85	0,00	0,00	1,65	?	?	?	10,45	?	0,08	0,50	1,25	2,80	-2,35	12392,50	24,85	29,43	29,33	28,88	?	29,65	0,73	3,45	3
?	1,30	7,34	5,20	23434,60	23414,40	9538,60	30,10	191,00	177,64	?	?	23,08	?	?	?	34,02	?	5,04	41,74	?	2,50	30,32	5349,00	16,74	14,60	9,56	-30,26	1,20	13,64	1,62	?	2
39640,25	1,54	6,60	7,14	53818,66	34448,08	7513,60	16,76	86,98	86,98	14,94	0,44	8,25	23,52	12,45	83,06	16,94	0,00	3,82	10,78	-0,40	-6,26	13,22	8993,80	19,30	10,12	4,36	?	?	7,86	-0,60	-8,40	1
18919,25	-1,38	19,50	11,00	11306,45	10893,18	1866,50	8,60	65,30	60,65	11,80	0,00	2,98	?	?	?	67,10	81,65	1,80	2,93	-11,65	-3,73	6,23	3522,25	16,15	-3,18	-4,93	-6,35	?	12,90	-0,33	-2,25	5
?	?	5,38	5,24	3920,30	3920,30	7048,80	10,58	152,80	146,86	7,70	7,70	2,90	?	?	?	39,26	?	1,78	4,42	?	-4,44	2,10	5129,60	7,48	3,82	2,04	-0,72	0,60	9,94	-0,30	?	6
32127,80	0,34	19,58	23,90	10408,20	10335,80	4115,60	15,22	115,52	110,76	1,10	?	9,94	40,74	?	69,34	30,66	?	3,58	10,16	-12,80	?	19,48	4328,00	15,48	0,48	-3,10	-10,08	0,08	12,42	-1,18	?	4
26209,40	0,06	27,44	23,90	13792,20	13791,80	6604,00	29,78	202,06	189,94	?	?	21,20	46,50	?	67,80	32,20	?	6,26	21,12	-23,50	?	38,74	4521,80	19,62	-12,36	-18,64	-34,96	-0,90	9,58	-2,10	?	5
39291,40	0,74	20,86	23,90	80277,60	74549,60	9971,00	29,98	269,76	257,64	?	?	27,62	46,64	4,90	82,58	17,42	?	7,34	33,72	-25,48	?	52,20	4993,00	14,56	-9,36	-16,70	-44,70	-0,84	11,20	-1,48	?	5
65497,40	0,88	9,04	10,00	298,16	247,66	1247,80	2,34	157,54	133,14	?	?	7,90	?	?	60,68	39,32	?	4,00	12,26	-6,06	?	14,96	1029,40	1,52	25,48	21,52	10,98	0,90	37,30	-0,26	?	3
34044,60	-0,80	?	11,18	84,52	84,52	1601,60	4,68	103,78	98,64	?	?	5,60	0,46	?	73,82	26,18	26,18	5,14	11,78	4,78	-5,30	12,44	1617,60	0,02	13,90	8,74	1,50	2,66	8,34	1,52	0,28	5
40194,60	0,84	?	8,06	6932,88	6932,88	1500,00	4,56	59,96	59,94	?	?	18,44	?	?	2,72	97,26	?	1,92	5,60	-3,42	?	7,10	2611,80	6,22	0,02	-1,88	-5,56	-0,55	1,90	-3,86	?	2
19688,00	2,60	5,08	6,78	1174,80	1174,80	2177,20	?	66,10	57,42	?	?	9,80	?	?	41,42	53,86	?	2,98	30,26	?	6,08	21,58	3568,00	?	8,32	5,32	-26,46	?	12,40	1,40	8,26	2
34735,60	0,88	10,26	10,00	649,46	373,12	628,00	?	58,74	55,42	?	?	9,28	?	?	36,84	30,98	?	0,88	5,50	0,20	?	4,20	1135,00	3,10	14,40	13,52	8,94	1,04	18,30	0,54	?	2
41005,50	1,98	7,65	7,18	326,00	325,93	503,25	?	31,53	28,35	0,00	0,00	6,45	6,45	?	79,48	20,53	20,53	0,80	2,28	?	10,48	0,00	1556,00	?	16,13	15,35	13,60	1,63	20,13	1,38	142,23	1
23268,40	0,26	30,62	23,90	3527,80	3507,60	3235,20	16,58	74,88	67,22	?	?	15,38	55,37	?	51,90	48,10	?	2,76	7,78	-10,40	?	15,68	5507,00	27,06	-4,60	-7,34	-13,24	0,18	12,20	-1,40	?	4
36813,60	4,56	4,56	5,62	1276,20	1101,40	22698,00	44,86	106,50	104,86	0,00	0,00	30,46	30,46	0,00	86,46	13,54	0,00	2,34	17,90	-3,94	?	19,54	22568,80	44,62	0,00	-0,26	-13,22	0,18	5,26	-0,12	?	2
44023,80	0,98	4,86	8,06	23954,58	18288,66	2847,40	7,94	66,70	66,68	?	?	34,36	?	?	31,66	68,34	?	0,54	3,40	2,12	?	2,26	5701,00	12,44	12,54	12,02	9,14	1,26	10,06	-4,26	?	2
42156,40	1,23	4,90	3,98	25960,34	24377,82	16702,80	37,90	227,02	217,16	?	?	5,70	11,10	?	?	26,22	22,84	3,58	18,42	?	5,50	12,20	7565,80	16,50	16,18	12,58	4,78	?	21,64	?	?	3
36027,40	0,93	4,92	3,98	39400,14	30175,18	13308,60	20,30	269,14	257,06	?	?	4,16	5,70	?	?	36,94	16,16	3,50	15,40	?	-4,20	18,82	3027,40	8,26	10,54	6,84	-5,28	?	14,98	?	?	3
29867,20	0,42	20,04	23,90	11197,80	11000,00	3984,80	15,76	121,86	111,28	?	?	13,78	41,40	?	70,44	29,56	?	3,96	15,24	-9,52	?	21,16	4031,60	15,46	3,90	-0,02	-12,74	0,48	16,68	-1,02	?	4
43718,20	5,24	8,24	7,92	6014,80	494,60	11278,60	1,64	10,18	10,18	0,00	0,00	15,88	?	?	95,02	5,62	0,00	3,54	13,08	?	4,30	7,84	8978,80	15,94	7,64	4,10	-2,22	?	?	?	-4,26	1
53642,25	5,38	15,04	7,10	190,18	198,35	5623,00	9,30	10,85	10,85	0,00	0,00	12,80	12,80	8,38	18,42	81,58	0,00	0,42	1,22	6,06	1,86	1,88	46864,80	79,65	11,26	10,82	?	?	10,84	?	21,13	2
39558,00	2,60	4,30	6,78	323,20	323,20	2293,60	?	175,22	161,52	?	?	4,48	?	?	58,72	41,24	?	3,62	26,50	?	5,80	19,16	1346,80	?	18,82	15,20	-10,40	?	16,78	1,56	51,90	2
41005,50	2,04	7,50	7,14	461,78	459,02	886,80	?	66,80	52,62	0,00	0,00	6,46	6,46	10,58	82,46	17,54	17,54	1,48	4,30	?	8,48	0,00	1241,60	?	24,70	23,16	19,20	1,32	40,26	3,56	217,64	1
35919,80	?	3,12	3,98	5012,62	5005,68	6214,80	16,56	120,54	113,80	?	?	9,58	9,58	?	?	27,68	22,22	2,24	13,24	?	6,50	9,02	5302,60	13,58	16,78	14,54	6,40	?	18,02	?	?	3
38491,80	1,53	3,36	3,98	24985,22	22570,96	7878,00	20,98	286,76	271,60	?	?	5,12	5,36	?	?	38,04	14,44	4,60	21,48	?	-1,68	19,78	3087,80	7,86	11,20	6,60	-8,36	?	16,32	?	?	3
41344,60	1,13	4,02	4,00	23058,30	25669,10	18216,00	42,08	269,80	258,70	?	?	5,56	5,94	?	?	30,96	24,88	4,28	18,56	?	0,54	14,86	7226,00	16,08	11,06	6,76	-4,76	?	10,74	?	?	3
26197,20	2,60	7,02	6,78	1864,40	1864,40	1583,60	?	58,08	54,48	?	?	44,22	?	?	34,00	54,74	?	2,52	60,72	?	0,92	57,42	2876,00	?	4,38	1,84	-60,35	?	7,06	-1,10	-9,92	2
37916,50	-1,00	11,20	11,40	20767,60	19975,40	3442,20	8,60	114,70	112,44	1,00	0,00	3,40	9,60	?	7,00	41,40	72,60	4,00	7,40	-11,40	-10,80	14,60	3537,00	6,60	-1,80	-5,80	-9,60	?	7,00	1,40	-32,40	5
32833,60	-1,02	9,16	11,34	1284,90	972,96	613,00	1,60	19,78	18,12	0,00	0,00	4,30	40,16	?	5,38	60,76	71,52	1,06	1,82	-3,62	-1,12	1,94	3619,60	9,58	0,92	0,56	-1,04	?	9,52	0,24	4,66	4
38570,50	-0,38	7,38	11,00	2980,03	2765,30	278,25	0,63	9,73	9,33	13,30	0,00	5,75	6,00	12,30	?	42,80	80,20	0,40	0,98	0,25	1,20	-0,73	3045,25	6,98	4,35	3,93	3,30	?	7,63	0,33	-6,63	4
41005,50	2,04	7,50	7,14	348,70	348,70	935,00	?	33,28	35,54	0,00	0,00	11,88	11,88	5,12	81,54	18,46	18,46	0,98	6,50	?	5,98	1,18	2385,80	?	30,72	29,78	24,64	1,36	19,08	-0,32	53,90	1
?	?	4,88	5,24	10187,36	10187,36	6288,60	16,08	110,18	103,14	26,87	26,65	17,16	?	?	?	25,98	?	1,58	9,04	?	0,18	1,50	6215,40	15,50	6,34	4,78	-3,38	1,04	10,46	1,28	?	2
30568,00	-3,33	15,80	14,36	7921,28	5073,92	20462,80	84,30	385,32	344,22	?	?	9,86	?	?	79,12	20,88	?	10,56	21,84	?	-31,44	43,96	7954,00	32,26	4,72	-5,86	-20,12	0,26	25,64	-7,44	?	6
44189,40	0,56	18,78	23,90	27203,00	27411,60	4263,60	11,40	153,52	150,74	?	?	16,60	32,48	?	121,76	-21,76	?	4,94	21,04	-18,26	?	34,44	3467,80	9,02	-3,22	-8,18	-24,68	-0,48	10,16	-2,50	?	4
38191,75	2,28	5,42	7,14	30379,60	18386,80	14538,60	33,70	144,66	144,66	24,33	6,08	11,40	22,30	13,88	72,08	27,92	0,00	5,86	25,86	-5,00	-10,50	30,52	10983,40	24,50	4,68	-1,16	?	?	8,94	-1,40	-29,28	2

Příloha B

Field	Graph	Measurement	Min	Max	Sum	Mean	Std. Dev
Population		Continuous	20.300	43116.780	1308723.060	2556.100	4167.253
GDP		Continuous	1798.640	798749.800	32464327.960	86803.016	126281.178
GDP_per_capita		Continuous	37.660	654.280	54021.800	105.511	56.604
GDP/national_average		Continuous	1.000	107305.200	13349735.540	25140.745	21205.177
GDP_PPP		Continuous	372.800	112473.500	13655637.360	27755.360	17091.966
real_GDP_change		Continuous	-3.820	9.660	972.940	1.954	1.845
unemployment		Continuous	1.440	37.080	4459.100	8.778	6.750
National Unemployment Rate		Continuous	3.040	29.800	4550.020	8.887	6.655
Total Direct and Indirect Debt		Continuous	0.320	273327.200	4978780.320	9724.180	30158.702
Net Direct and Indirect Debt		Continuous	0.500	257758.200	4367386.000	8530.051	27730.315
net_debt_per_capita		Continuous	0.000	29310.000	1435431.000	2803.576	4908.822
debt/GDP		Continuous	0.000	84.300	3748.100	10.709	13.987
debt/operating_revenue		Continuous	0.600	385.320	36140.120	70.586	73.303
Net Direct and Indirect Debt/Total Revenue		Continuous	0.600	346.280	33641.320	65.706	69.423
FX Direct Debt(before swaps)/Direct Debt		Continuous	0.000	93.700	2763.380	7.159	16.560
FX Direct Debt(after swaps)/Direct Debt		Continuous	0.000	93.700	2496.540	6.535	16.791
short_debt/debt		Continuous	0.000	84.620	7637.880	14.918	13.476
long_debt/debt		Continuous	0.000	100.000	21723.600	51.970	40.111
maturity_debt		Continuous	0.000	21.600	1497.640	8.509	4.647
own_rev/oper_rev		Continuous	1.360	121.760	20070.580	44.208	29.808
gov_transf/oper_rev		Continuous	-21.760	98.600	23159.600	45.234	28.586
earmarked_rev/oper_rev		Continuous	0.000	94.760	10364.060	26.305	23.900
interest/oper_rev		Continuous	0.000	11.580	1281.120	2.502	2.158
debt_ser/total_rev		Continuous	0.000	60.720	4401.320	8.596	9.231
Accrual Financing Surplus(Requirement)/Total Revenue		Continuous	-40.680	13.600	-543.460	-4.383	9.026
cash_surplus/TR		Continuous	-31.440	20.320	-618.400	-1.316	5.951
borrowing_need/TR		Continuous	-11.220	111.500	5202.060	10.240	14.533
TE_per_capita		Continuous	10.000	46864.800	1582953.700	3091.706	4807.046
TE/GDP		Continuous	0.000	79.650	4465.160	12.613	9.427
oper_balance/OR		Continuous	-14.180	82.360	4417.120	8.661	10.248
gross_oper_balance/OR		Continuous	-23.040	76.180	3149.720	6.152	10.294
Net Operating Balance/Operating Revenue		Continuous	-128.460	76.180	-549.560	-1.159	18.329
self_financing_ratio		Continuous	-1.140	4.100	337.420	0.953	0.624
capital_spending/TE		Continuous	1.900	67.580	8803.220	17.261	10.110
TR-TE(%)		Continuous	-7.440	17.480	64.940	0.138	2.161
NWC/TE		Continuous	-60.920	217.640	2514.760	8.327	30.706
Predikce_8_trid		Continuous	1.000	8.000	2062.000	4.027	1.586

Příloha C

Základní klasifikátory

Rozhodovací strom J48

Přesnost klasifikace %	Přesnost	Průměrné náklady	Senzitivita	Specificita	Plocha pod ROC křivkou
54,54 %	0,55	0,64	0,55	0,59	0,78
46,75 %	0,47	0,70	0,47	0,50	0,75
41,56 %	0,42	0,74	0,42	0,50	0,74
51,95 %	0,52	0,68	0,52	0,52	0,80
54,54 %	0,55	0,57	0,55	0,54	0,79
29,87 %	0,30	1,04	0,30	0,30	0,63
50,65 %	0,51	0,58	0,51	0,53	0,75
40,26 %	0,40	0,82	0,40	0,40	0,72
50,65 %	0,51	0,73	0,51	0,50	0,41
50,65 %	0,51	0,78	0,51	0,59	0,74
47,14 %		0,73			0,71

Neuronová síť MLP

Přesnost klasifikace %	Přesnost	Průměrné náklady	Senzitivita	Specificita	Plocha pod ROC křivkou
62,34 %	0,62	0,61	0,62	0,65	0,86
58,44 %	0,58	0,58	0,58	0,62	0,83
64,93 %	0,65	0,49	0,65	0,64	0,86
55,84 %	0,56	0,73	0,56	0,60	0,87
49,35 %	0,49	0,70	0,49	0,55	0,83
59,74 %	0,60	0,48	0,60	0,59	0,87
64,93 %	0,65	0,56	0,65	0,65	0,87
49,35 %	0,49	0,75	0,49	0,47	0,82
58,44 %	0,58	0,58	0,58	0,60	0,79
58,44 %	0,58	0,58	0,58	0,60	0,85
58,18 %		0,61			0,85

SMO (SVM)

Přesnost klasifikace %	Přesnost	Průměrné náklady	Senzitivita	Specificita	Plocha pod ROC křivkou
63,64 %	0,64	0,56	0,64	0,65	0,85
57,14 %	0,57	0,52	0,57	0,57	0,84
61,04 %	0,61	0,56	0,61	0,62	0,85
63,64 %	0,64	0,40	0,64	0,66	0,88
63,64 %	0,64	0,52	0,64	0,64	0,85
71,43 %	0,71	0,32	0,71	0,75	0,90
70,13 %	0,70	0,40	0,70	0,69	0,87
64,94 %	0,65	0,51	0,65	0,62	0,84
57,14 %	0,57	0,61	0,57	0,64	0,82
64,94 %	0,65	0,48	0,65	0,66	0,85
63,77 %		0,49			0,85

NaiveBayes

Přesnost klasifikace %	Přesnost	Průměrné náklady	Senzitivita	Specificita	Plocha pod ROC křivkou
36,36 %	0,36	0,87	0,36	0,44	0,79
44,15 %	0,44	0,78	0,44	0,56	0,81
44,16 %	0,44	0,77	0,44	0,52	0,74
50,65 %	0,51	0,64	0,51	0,51	0,79
54,54 %	0,55	0,65	0,55	0,55	0,83
46,75 %	0,47	0,71	0,47	0,55	0,86
49,35 %	0,49	0,66	0,49	0,57	0,80
35,06 %	0,35	0,91	0,35	0,36	0,76
38,96 %	0,39	0,82	0,39	0,45	0,76
32,47 %	0,33	0,92	0,33	0,27	0,74
43,25 %		0,77			0,79

Soubory klasifikátorů (meta klasifikátory)

AdaBoost

Přesnost klasifikace %	Přesnost	Průměrné náklady	Senzitivita	Specificita	Plocha pod ROC křivkou
63,64 %	0,64	0,56	0,64	0,65	0,85
57,14 %	0,57	0,52	0,57	0,57	0,84
61,04 %	0,61	0,56	0,61	0,62	0,85
64,64 %	0,64	0,40	0,64	0,66	0,88
63,64 %	0,64	0,52	0,64	0,64	0,85
71,43 %	0,71	0,32	0,71	0,75	0,66
70,13 %	0,70	0,40	0,70	0,69	0,87
64,94 %	0,65	0,52	0,65	0,62	0,84
57,14 %	0,57	0,61	0,57	0,64	0,82
64,94 %	0,65	0,48	0,65	0,66	0,85
63,87 %		0,49			0,83

Bagging

Přesnost klasifikace %	Přesnost	Průměrné náklady	Senzitivita	Specificita	Plocha pod ROC křivkou
59,74 %	0,60	0,69	0,60	0,62	0,86
57,14 %	0,57	0,53	0,57	0,60	0,87
55,84 %	0,56	0,60	0,56	0,56	0,85
63,64 %	0,64	0,52	0,64	0,68	0,89
61,04 %	0,61	0,60	0,61	0,59	0,87
67,53 %	0,68	0,42	0,68	0,67	0,90
66,23 %	0,66	0,55	0,66	0,65	0,87
57,14 %	0,57	0,69	0,57	0,55	0,83
58,44 %	0,58	0,62	0,58	0,57	0,83
63,64 %	0,64	0,51	0,64	0,63	0,84
61,04 %		0,57			0,86

Random Forest

Přesnost klasifikace %	Přesnost	Průměrné náklady	Senzitivita	Specificita	Plocha pod ROC křivkou
68,83 %	0,69	0,43	0,69	0,69	0,91
65,93 %	0,65	0,47	0,65	0,64	0,89
62,34 %	0,62	0,48	0,62	0,65	0,85
67,53 %	0,68	0,35	0,68	0,67	0,91
65,64 %	0,64	0,47	0,64	0,65	0,89
62,34 %	0,62	0,43	0,62	0,63	0,90
61,55 %	0,57	0,56	0,57	0,60	0,90
58,44 %	0,58	0,56	0,58	0,55	0,89
60,25 %	0,53	0,55	0,53	0,53	0,87
68,83 %	0,69	0,38	0,69	0,69	0,90
64,17 %		0,47			0,89

Stacking

Přesnost klasifikace %	Přesnost	Průměrné náklady	Senzitivita	Specificita	Plocha pod ROC křivkou
68,83 %	0,69	0,40	0,69	0,69	0,90
62,34 %	0,62	0,48	0,62	0,63	0,88
55,84 %	0,56	0,52	0,56	0,57	0,83
59,74 %	0,60	0,48	0,60	0,63	0,90
62,34 %	0,62	0,49	0,62	0,61	0,90
61,04 %	0,61	0,43	0,61	0,61	0,90
51,95 %	0,52	0,62	0,52	0,56	0,87
55,84 %	0,56	0,60	0,56	0,54	0,85
53,25 %	0,53	0,58	0,53	0,57	0,86
61,04 %	0,61	0,49	0,61	0,62	0,89
59,22 %		0,51			0,88