ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Fakulta biomedicínského inženýrství
Katedra biomedicínské techniky

Predikce kardiální autonomní neuropatie u pacientů s diabetem

Prediction of cardial autonomy neuropathy of patients with diabetes

Bakalářská práce

Studijní program:	Biomedicínská a klinická technika
Studijní obor:	Biomedicínský technik
Vedoucí práce:	Ing. Jakub Novák

Michaela Benešová

Kladno, květen 2016

České vysoké učení technické v Praze, Fakulta biomedicínského inženýrství

Katedra biomedicínské techniky | Akademický rok: 2015/2016

Zadání bakalářské práce

Student:	**Michaela Benešová**
Obor:	Biomedicínský technik
Téma:	**Predikce kardiální autonomní neuropatie u pacientů s diabetem**
Téma anglicky:	Prediction of cardial autonomy neuropathy of patients with diabetes

Zásady pro vypracování:

Analyzujte a popište naměřená data o pacientech s kardiální autonomní neuropatií adiabetes mellitus s důrazem na příznaky. Pomocí nástrojů pro statistickou analýzu a metod pro předzpracování dat realizujte implementaci pro transformaci dat, filtraci dat a vyvážení skupin. Takto zpracovaná data použijte jako vstup pro metodu "random forest". Pomocí této metody natrénujte prediktor na připravených datech s chybějícími hodnotami. Stejnou metodu aplikujte na data s uměle rekonstruovanými hodnotami a výsledky porovnejte. Zhodnocení dosažených výsledků proveďte na základě kontingenční tabulky, ROC křivky a vypočtených charakteristik jako senzitivity, specificity, pozitivní a negativní prediktivní hodnoty.

Seznam odborné literatury:

[1] Witten I, Frank Eibe, Hall Mark A, Data Mining: Practical Machine Learning Tools and Techniques, ed. 3rd, 2011, Morgan Kaufmann, 978-0123748560

[2] Pyle Dorian, Data preparation for data mining, ed. 1st, 1999, Morgan Kaufmann Publishers, 978-1558605299

[3] Hastie Trevor, Tibshirani Robert, Friedman Jerome, The elements of statistical learning: data mining, inference, and prediction, ed. 3rd, 2001, Springer, 978-0387952840

[4] Tesfaye S1, Boulton AJ, Dyck PJ, Freeman R, Horowitz M, Kempler P, Lauria G, Malik RA, Spallone V, Vinik A, Bernardi L, Valensi P; Toronto Diabetic Neuropathy Expert Group, Diabetic neuropathies: update on definitions, diagnostic criteria, estimation of severity, and treatments, Online, Diabetes Care [online], ed. 2010, [Revidováno 12/2010], [Citováno 2015-09-06], ročník 33, číslo 10, DOI: 10.2337/dc10-1303

zadání platné do:	30.09.2017
Vedoucí:	Ing. Jakub Novák
Konzultant:	doc. MUDr. Lucie Riedlbauchová, Ph.D., FN Motol

.....
vedoucí katedry / pracoviště

.....
děkan

V Kladně dne 22.02.2016

Prohlášení

Prohlašuji, že jsem bakalářskou práci s názvem Predikce kardiální autonomní neuropatie u pacientů s diabetem vypracovala samostatně a použila k tomu úplný výčet citací použitých pramenů, které uvádím v seznamu přiloženém k bakalářské práci.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu §60 Zákona č.121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

V Kladně dne ....

podpis

Poděkování

Ráda bych tímto poděkovala svému vedoucímu Ing. Jakubu Novákovi za vedení mé bakalářské práce, za jeho podnětné rady, připomínky a podporu. Dále bych chtěla poděkovat výzkumné skupině pod vedením doc. Riedlbauchové za poskytnutí použitých dat.

Abstrakt

Práce se zabývá analýzou medicínských dat, jejich předzpracováním a konstrukcí klasifikátoru za účelem predikce kardiální autonomní neuropatie u pacientů s diabetem. Data minigové metody nacházejí často uplatnění v medicíně, kde může klasifikátor sloužit jako pomocný nástroj při diagnostice. Cílem práce je porovnat vlastnosti klasifikátoru random forest v případě, že na vstupu budou nedoplněná data, a v případě, že data na vstupu budou uměle doplněna.

Jsou použita reálná vstupní data. V první řadě je nutné jejich předzpracování, včetně implementace metody na doplnění chybějících hodnot. Byla vybrána metoda k-NN. Pro predikci je využit klasifikátor random forest. Kvalita klasifikace je zhodnocena na základě kontingenční tabulky, ROC křivky (receiver operating characteristic curve), AUC (area under the ROC curve), celkové přesnosti, senzitivity, specificity a pozitivní a negativní prediktivní hodnoty. K porovnání výsledků pro nedoplněná a doplněná data na vstupu byl využit McNemarův test a dvouvýběrový t-test.

Provedené testy nepotvrdily na hladině významnosti 5 % rozdíl v celkové přesnosti klasifikátorů pro nedoplněná a doplněná data, ani ve střední hodnotě AUC. Rozptyl hodnot AUC byl vyšší při klasifikaci doplněných dat.

Z dosažených výsledků plyne, že umělé doplnění chybějících hodnot metodou k-NN nemá vliv na kvalitu klasifikátoru random forest. Aplikace metody k-NN by při použití jiného klasifikátoru, který nemůže mít na vstupu chybějící údaje, umožnila využít i původně nekompletní případy a lze se domnívat, že by zvýšila úspěšnost klasifikátoru.

Cíl práce byl splněn, vliv doplnění hodnot na chybu klasifikace nebyl prokázán.

Klíčová slova

random forest, náhodný les, predikce, diabetes mellitus, kardiální autonomní neuropatie, klasifikace

Abstract

The subject of this thesis is medical data analysis and preprocessing and construction of a classifier able to predict cardial autonomy neuropathy among patients with diabetes. Data mining methods are frequently used in medicine, where classifiers can be used as decision support tools. The aim of the thesis is to compare performance of random forest classifier with input dataset with missing data and with synthetically completed dataset.

Real dataset is used. Initial preprocessing is necessary, including imputation method implementation. Chosen method is k-NN imputation. For prediction purposes, random forest classifier is used. Classifier performance is evaluated using confusion matrix, ROC curve (receiver operating characteristic curve), AUC (area under the ROC curve), accuracy, sensitivity, specificity, positive predictive value and negative predictive value. McNemar's test and two-sample t-test were used for comparison of classifier using incomplete dataset and classifier using completed dataset.

At the 0.05 significance level, performed tests did not confirm a statistically significant difference in accuracy of classifiers for incompleted and completed data, neither in the expected value of AUC. Variance of AUC was higher for classification of imputed dataset.

The results show that synthetic imputation of missing data using k-NN imputation does not influence performance of random forest classifier. Application of k-NN imputation method before using a classifier which is not able to handle incomplete dataset would enable to use initially incomplete samples. It is possible to believe that more samples would increase classification accuracy

The aim of the thesis was accomplished, impact of missing data imputation on classification error was not proved.

Keywords

random forest, prediction, diabetes mellitus, cardial autonomy neuropathy, classification

Obsah

Úvod

Diabetická autonomní neuropatie je častou a významnou komplikací u diabetiků 1. i 2. typu. Její nejzávažnější formou je kardiální autonomní neuropatie (KAN), kdy jsou postiženy autonomní nervy a která způsobuje odchylky v schopnosti adaptace srdeční frekvence na změny míry zátěže srdce.

KAN výrazně zvyšuje pravděpodobnost vzniku vážných kardiovaskulárních komplikací u diabetiků, zejména riziko infarktu myokardu, srdeční slabosti, maligních arytmií a náhlé smrti. Včasné stanovení diagnózy může přispět ke snížení rizika budoucích komplikací. Odhalení autonomní neuropatie v subklinické fázi je však obtížné.

V rámci bakalářské práce je úkolem předzpracovat data o pacientech s diabetem mellitus a kardiální autonomní neuropatií do takové podoby, aby mohla sloužit jako vstup do klasifikátoru random forest. Dalším úkolem je doplnit chybějící data. V následujícím kroku slouží zvlášť nedoplněná a zvlášť doplněná data k natrénování klasifikátoru, který rozděluje případy na pacienty pouze s diabetem a na pacienty s diabetem a zároveň kardiální autonomní neuropatií.

Dalším cílem je porovnat kvalitu klasifikace při natrénování souborem s chybějícími hodnotami a s doplněnými hodnotami na základě kontingenční tabulky, ROC křivky (receiver operating characteristic curve), senzitivity, specificity, pozitivní prediktivní hodnoty a negativní prediktivní hodnoty a zjistit, jaký vliv má umělé doplnění hodnot na tyto parametry.

Pro implementaci algoritmů bylo použito programové prostředí Matlab R2015b.

1 Zpracování lékařských dat

Medicínská data mívají v několika ohledech specifické vlastnosti, které je třeba zohlednit při jejich zpracovávání. Vzhledem k tomu, že data obsahují citlivé osobní údaje, přináší s sebou jejich zpracování a uchovávání i etické a právní závazky [1].

1.1 Lékařská data

Datové soubory získané ve zdravotnictví často obsahují relativně málo případů, ovšem popsané velkým množstvím atributů. Při sběru dat je přítomno několik zdrojů, které mohou do dat vnášet zkreslení, např. subjektivita při sebehodnocení pacientem, interpretace lékaře i nejednotná forma zápisu odborných termínů, zvláště pokud data pochází z více zdrojů [1].

Obezřetně je třeba přistupovat k odlehlým hodnotám. Jako ve všech jiných oblastech mohou být způsobeny chybou při získávání či zápisu, v některých případech jsou však zcela správné a představují vzácné patologické případy, které je nutné do zpracování zahrnout [2].

Vzhledem k velkému počtu atributů je častým krokem při zpracování medicínských dat snižování dimeznionality, což vede k poklesu výpočetní náročnosti dalších metod a také k usnadnění interpretace a vizualizace dat [3].

Problémem společným pro většinu medicínských dat je velké množství chybějících hodnot. Důvody mohou být různé. Údaje mohly být pouze nezaznamenány, nebo vůbec nezměřeny z technických, ekonomických nebo etických důvodů [4]. Pokud je větší množství chybějících dat při zpracování ignorováno, nebo doplněno nevhodně, může to vést k vážnému ovlivnění dat a tím i výsledků jejich zpracování [5].

1.2 Využití klasifikátorů v medicíně

Data mining (dolování dat) je obecné označení pro metody získávání skrytých a potenciálně užitečných informací z datových souborů. V medicíně a zdravotnictví je přítomno velké množství vícerozměrných dat, které při efektivním použití data miningových metod a odhalení skrytých trendů a závislostí mohou být pomocným nástrojem při diagnostice a volbě léčby [6].

Pro popsání závislosti mezi výsledkem a atributy a následnou klasifikaci se v medicíně často využívá lineární regrese, logistická regrese a diskriminační analýza. Tyto metody jsou považovány za „tradiční“. V posledních desetiletích se dostávají do popředí data miningové metody jako například neuronové sítě, support vector machines (SVM), rozhodovací stromy a lesy a naivní Bayes. Tyto nástroje se využívají například v predikci obtížně diagnostikovatelných srdečních chorob [6, 7, 8].

Vzniká množství prací, které srovnávají tradiční přístupy s data miningovými metodami, např. neuronovými sítěmi [7], regresními stromy [9], nebo se věnují porovnání více různých metod [10, 11, 12].

V [13] je pro predikci chronického onemocnění jater využita extrakce části atributů z ultrazvukového obrazu. V [11] bylo při predikci progrese demence a její odlišení od lehké poruchy poznávacích funkcí porovnáno sedm data minigových metod se třemi tradičními klasifikátory. Jednalo se o klasifikaci do dvou tříd, tedy binární klasifikaci. Nejlépe z hlediska celkové přesnosti, senzitivity a specificity obstály v této úloze metody random forest a lineární diskriminační analýza. Random forest byl s úspěchem použit rovněž při predikci diabetu [14].

Výhodou lineární regrese je snadná interpretace modelu, zjevnou nevýhodou je neschopnost vyjádřit nelineární vztahy mezi atributy a výstupem. Předností metody random forest je schopnost pracovat s nelineárními vztahy obsaženými v datech, jak je zmíněno např. v [15]. Další výhodou je možnost zpracování nekompletních vstupních dat použitím metody náhradního dělení [16].

Random forest lze využít i v jiných oblastech, například v bioinformatice při vyhodnocování exprese genů na DNA čipu [17] nebo v ekologii při studiu výskytu druhů [18].

2 Teoretický základ

Metody data miningu představují jeden z přístupů při analýze medicínských dat a konstrukci predikčních modelů. Pro vyhodnocení a vyslovení závěru o kvalitě klasifikátoru nestačí znát pouze jeho celkovou přesnost, ale je nutné pohlížet na více charakteristik, např. senzitivitu a specificitu [19].

Při zpracovávání dat je vždy důležité mít povědomí o tom, co data reprezentují a jak byla získána, proto jsou v následující kapitole stručně charakterizována onemocnění, jejichž rozlišením se práce zabývá.

2.1 Diabetes mellitus a kardiální autonomní neuropatie

Diabetes mellitus je chronické onemocnění vznikající v důsledku nedostatečné produkce nebo nedostatečného využití hormonu inzulinu. Tato dysfunkce vede ke zvýšené hladině glukózy v krvi (hyperglykémii). Při fyzické aktivitě bez dostatečného doplnění energie z potravy naopak dochází ke snížení hladiny glukózy v krvi (hypoglykémii). Rozlišuje se diabetes mellitus prvního typu (absolutní nedostatek inzulinu) a druhého typu (relativní nedostatek inzulinu) [20].

Jednou z možných pozdních komplikací je u diabetiků tzv. diabetická neuropatie, což je nezánětlivé poškození funkce a struktury periferních nervů. Příčinou tohoto poškození je dlouhodobá hyperglykémie. Mohou být postiženy nervy motorické, senzitivní i vegetativní (autonomní nervový systém). Podle typu poškozených nervů a postižené oblasti lze rozlišit mnoho druhů neuropatií, nejzávažnější z nich je kardiální (kardiovaskulární) autonomní neuropatie, kdy je v důsledku poškození autonomních nervů negativně ovlivněna funkce srdce a výrazně se zvyšuje riziko výskytu závažných kardiovaskulárních komplikací [20].

Určení prevalence diabetické neuropatie je obtížné vzhledem k tomu, že senzorickomotorická i autonomní neuropatie může dlouho probíhat asymptomaticky. Diagnostika se při prvním kontaktu s nemocným z rizikové skupiny diabetiků provádí formou dotazníku, který se týká přítomnosti symptomů postižení jednotlivých orgánů. Senzitivita i specificita dotazníku je však nízká a tento způsob není vhodný pro včasnou diagnostiku [21].

Standardizované autonomní funkční testy mohou onemocnění odhalit, přestože se ještě neprojevily charakteristické příznaky. Jedná se o vyšetření variability srdeční frekvence pomocí neinvazivních testů z tzv. Ewingovy baterie testů. Patří sem: test hlubokého dýchání, ortostatický test, Valsalvův manévr a isometrický stisk ruky [21].

2.2 Random forest

Metoda random forest (náhodný les) patří do skupiny tzv. souborových metod (*ensemble methods*) [22]. Výsledkem souborových metod je komplexní model složený z dílčích modelů. Kombinování se provádí za účelem zlepšení predikčních vlastností.

Random forest se skládá z binárních rozhodovacích stromů typu CART (Classification and Regression Trees) a lze ho využít pro klasifikační i regresní úlohy. Dokáže zpracovat i data, která v rámci atributů obsahují chybějící hodnoty. Také odstraňuje nepříznivou vlastnost rozhodovacích stromů, kterou je nestabilita [23].

Každý strom souboru je vytvořen s využitím podmnožiny případů a podmnožiny atributů. Trénovací soubor je tzv. bootstrapovým výběrem z datasetu, tedy výběrem s opakováním. Použitím odlišných trénovacích souborů je dosaženo menší korelace mezi stromy. Při trénovaní se u bootstrapového výběru použije m náhodně vybraných prediktorů z celkem M dostupných prediktorů. Pro klasifikaci je doporučená hodnota

$$m = \sqrt{M}, \tag{1}$$

pro regresi je doporučená hodnota

$$m = \log_2 (M + 1). \tag{2}$$

Při klasifikaci případu z testovací množiny je každým stromem provedeno zařazení do klasifikační třídy. Celkový výsledek je potom dán většinovým výsledkem (u regrese je použit průměr výstupů jednotlivých stromů) [22].

Jak bylo zmíněno, při trénování konkrétního stromu nejsou použity všechny případy, je náhodně vybrána jejich podmnožina. Případy mimo tuto podmnožinu se nazývají *out-of-bag* případy, neboli OOB případy. Lze je využít jako testovací množinu již při konstrukci klasifikátoru. Po natrénování n-tého stromu vznikne n-tá

skupina OOB případů. I-tý OOB případ je následně klasifikován každým stromem, při jehož konstrukci nebyl použit. Z poměru správně klasifikovaných případů ku všem provedeným klasifikacím je dosaženo odhadu chyby klasifikace nazývané OOB error [23].

OOB případy lze použít i pro určení důležitosti atributů. Při zjišťování důležitosti atributů se po výpočtu OOB erroru provede náhodná permutace hodnot v rámci m-tého atributu a klasifikace OOB případů se opakuje. Je zaznamenáván vzrůst OOB erroru. Důležitost atributu je tím vyšší, čím je vzrůst OOB erroru výraznější. Naopak pokud daný atribut není pro klasifikaci významný, po jeho permutaci se úspěšnost klasifikace nezmění [22].

Dalším z parametrů metody random forest je počet stromů, ze kterých se skládá. Obecně platí, že čím vyšší počet stromů, tím přesnější výsledek, ovšem při určitém hraničním počtu je již zlepšení nevýznamné a další zvyšování počtu stromů jen zvyšuje čas potřebný k natrénování klasifikátoru [24].

Využití metody random forest je vhodné pro zpracování úloh, kde je ve vstupních datech velké množství atributů [22].

2.3 Testování modelů

K testování modelu se využívá testovacích dat, které nebyly využity při trénování. Časté je dělit data na trénovací a testovací množinu například v poměru 2:1 nebo 4:1 [22, s. 4]. Alternativně lze využít křížovou validaci, která je mj. vhodná v situacích, kdy je k dispozici malé množství případů. Při k-násobné křížové validaci je množina dat rozdělena do k disjunktních podmnožin stejné velikosti. Každá podmnožina je použita právě jednou pro testování modelu vzniklého ze zbylých dat. Trénování a testování se tedy opakuje k-krát. Leave-one-out je extrémní případ křížové validace, kdy se jako trénovací množina použijí všechny případy kromě jednoho, který je využit jako testovací. V tomto případě platí, že počet podmnožin k je rovný p, což je celkový počet případů [22].

2.4 Hodnocení klasifikace

Pro zhodnocení klasifikátoru se standardně využívá několika charakteristik. Jednou z nich je celková přesnost, která je definována jako procento správně klasifikovaných případů.

Přehlednou možností zhodnocení je kontingenční tabulka (*confusion matrix*). V případě binární klasifikace se jedná o tabulku o rozměrech 2×2, kde jsou zaznamenány výsledky klasifikace vzhledem k reálnému stavu, viz tabulka 1.

Tabulka 1: Kontingenční tabulka

		Reálný stav	
		pozitivní (+)	negativní (−)
Výsledek testu	pozitivní (+)	skutečně pozitivní (SP)	falešně pozitivní (FP)
	negativní (−)	falešně negativní (FN)	skutečně negativní (SN)

Mohou nastat čtyři možnosti:

- SP = pozitivní případ je klasifikován jako pozitivní,
- FP = negativní případ je klasifikován jako pozitivní,
- SN = negativní případ je klasifikován jako negativní,
- FN = pozitivní případ je klasifikován jako negativní.

Kontingenční tabulka obsahuje počty případů v jednotlivých kategoriích a může z ní být vypočítáno mnoho parametrů, včetně již zmíněné celkové přesnosti.

Celková přesnost odpovídá součtu správně zařazených případů ku součtu všech klasifikovaných případů

$$celková\,přesnost = \frac{SP + SN}{SP + SN + FP + FN} \tag{3}$$

Senzitivita je pravděpodobnost, že pozitivní případ bude odhalen.

$$senzitivita = \frac{SP}{SP + FN} \tag{4}$$

Specificita vyjadřuje pravděpodobnost, že negativní případ bude správně označen jako negativní.

$$specificita = \frac{SN}{SN + FP} \tag{5}$$

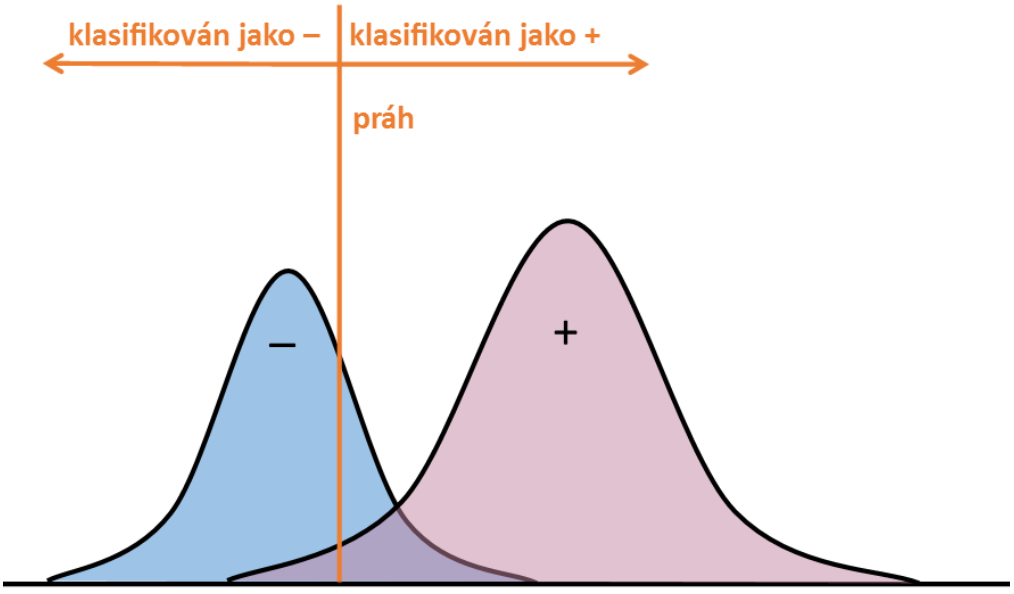
Pozitivní prediktivní hodnota (PPH) udává poměr správně pozitivních výsledků v poměru ke všem pozitivním výsledkům testu.

$$PPH = \frac{SP}{SP + FP} \tag{6}$$

Negativní prediktivní hodnota (NPH) udává poměr správně negativních výsledků v poměru ke všem negativním výsledkům testu [25].

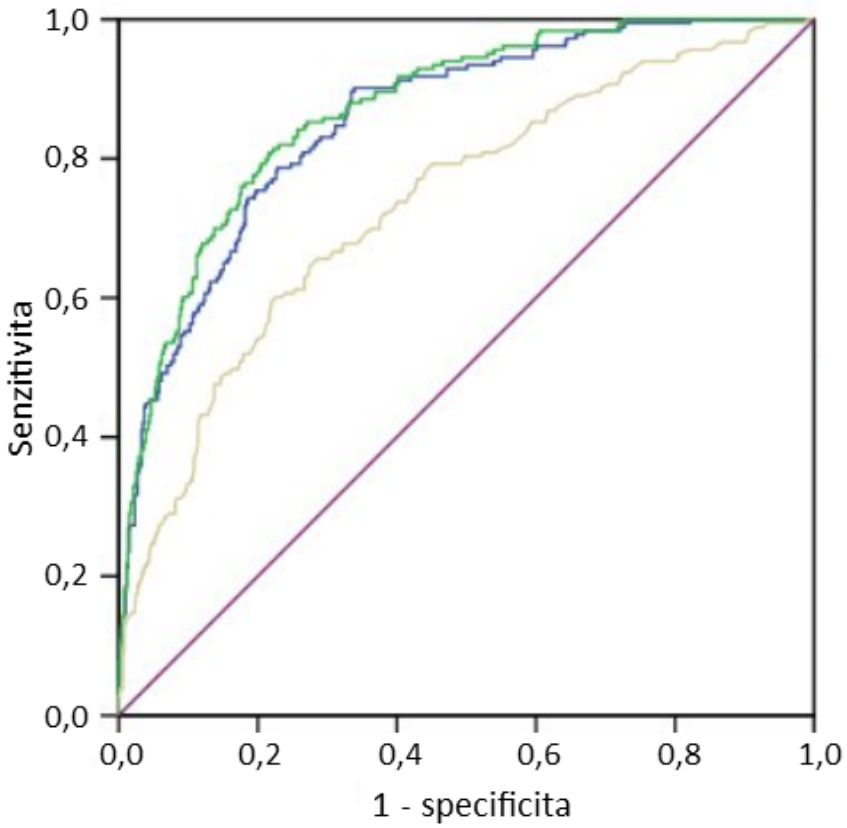
$$NPH = \frac{SN}{SN + FN} \tag{7}$$

ROC křivka je grafické vyjádření závislosti senzitivity a specificity binárního klasifikačního systému na hodnotě prahu. Prahem se myslí mezní hodnota, která rozděluje oblast klasifikovanou jako negativní ($-$) od oblasti klasifikované jako pozitivní ($+$) [26], viz obrázek 1. Posunem prahu se mění počet správně pozitivních a negativních případů i falešně pozitivních a negativních případů a tím i hodnoty senzitivity a specificity.



Obrázek 1: Zobrazení množiny pozitivních a negativních případů a prahu, podle kterého se případy klasifikují na pozitivní a negativní

Příklad ROC křivek je zobrazen na obrázku 2. Pro porovnání více ROC křivek lze využít hodnoty plochy pod křivkou, tzv. AUC (area under the ROC curve). ROC křivka je komplexní charakteristika, která vypovídá nejen o kvalitě klasifikátoru, ale i o charakteru vstupních dat. Jestliže je plocha rovna 1, je klasifikátor nebo test ideální, popřípadě jsou třídy, do nichž probíhá klasifikace, jednoduše oddělitelné, tedy senzitivita i specificita je 100 %.



Obrázek 2: ROC křivky, fialová úsečka představuje ROC křivku náhodné klasifikace, její AUC je 0,5 [27]

2.5 Porovnání klasifikátorů

K otestování statistické významnosti odlišnosti přesnosti klasifikátorů lze použít McNemarův test [28]. K tomu je třeba zaznamenat výsledky klasifikace stejných případů odlišnými klasifikátory, viz tabulka 2.

Tabulka 2: Tabulka výsledků klasifikace dvěma klasifikátory

		Klasifkátor 1	
		třída 1	třída 0
Klasifkátor 2	třída 1	a	b
	třída 0	c	d

Nulová hypotéza říká, že rozdíl v přesnosti klasifikátorů není statisticky významný, alternativní hypotéza říká, že rozdíl v přesnosti je statisticky významný.
Testovací kritérium se vypočítá podle následujícího vzorce

$$\chi^2 = \frac{(b-c)^2}{b+c}, \tag{8}$$

Pokud je součet na diagonále menší než 25, používá se výpočet s Edwardsovou úpravou

$$\chi^2 = \frac{(\mid b-c \mid -1)^2}{b+c}. \tag{9}$$

Za platnosti nulové hypotézy má statistika χ^2 rozdělení s jedním stupněm volnosti. Vypočítané testovací kritérium se následně na zvolené hladině významnosti α porovnává s kritickou hodnotou $\chi^2_{1-\alpha}(1)$. Pokud je χ^2 menší než $\chi^2_{1-\alpha}(1)$, nezamítáme nulovou hypotézu.

Obecně lze pro porovnání dvou nezávislých výběrů použít dvouvýběrový nepárový t-test, který porovnává střední hodnoty obou výběrů. Musí být dodržena podmínka normality obou výběrů [29].

3 Návrh řešení

Data zpracovávaná v bakalářské práci poskytla Fakultní nemocnice v Motole. V první fázi práce byl vytvořen postup pro jejich zpracování, který je popsán v následující kapitole.

3.1 Reálná vstupní data

Získaná data byla multidimenzionální. U celkem 130 vyšetřovaných osob bylo sledováno 130 atributů, které by bylo možné rozdělit do několika skupin:

- základní údaje, např. věk, výška, váha, krevní tlak,
- hodnoty z laboratorních vyšetření, např. koncentrace glykovaného hemoglobinu, cholesterolu, kreatininu, CRP,
- přítomnost/nepřítomnost konkrétního onemocnění, případně provedení/neprovedení konkrétního zákroku, např. ischemická choroba srdeční, stenóza karotid, diabetická noha, kompletní revaskularizace,
- výsledky dotazníků a provedení autonomních funkčních testů, např. test hlubokého dýchání, Valsalvův manévr, ortostatický test,
- výsledky provedení časové a spektrální analýzy variability srdeční frekvence.

Zaznamenávané proměnné byly ve většině případů spojité kvantitativní, v menším zastoupení byly proměnné kvalitativní (nominální i ordinální).

Případy (osoby) byly rozděleny do tří skupin:

- 0 – osoba netrpí diabetem mellitus, ani KAN, jedná se o osobu ze zdravé kontrolní skupiny,
- 1 – osoba má diabetes mellitus prvního či druhého typu a zároveň KAN,
- 2 – osoba má diabetes mellitus prvního či druhého typu, KAN nemá.

Část tabulky ukazuje obrázek 3.

																								SAS dotazník	
Diabetes mellitus (0-ne, 1-typ 1, 2-typ 2)	Rok manifestace diabetu	Thyreopatie (0-ne, 1-hypo-, 2-hyperthyreosa, 3-autoim.thyreoidilitida)	fT4 (pmol/l) době 1.RQA analýzy, ne starší 1 roku	Glykemie nalačno (mmol/l)v den analýzy	HbA1C dle DCCT	MDRD (ml/s)v době 1.RQA analýzy, ne starší 1 roku	Celková bílkovina (g/l) v době 1.RQA analýzy, ne starší 1 roku	Albumin (g/l) v době 1.RQA analýzy, ne starší 1 roku	Albumin/ kreatinin ratio v době 1.RQA analýzy, ne starší 1 roku	Diabetická nefropatie? (0-ne, 1-ano MAU>2,3 / léčená)	Celkový cholesterol (mmol/l) v době 1.RQA analýzy, ne starší 1 roku	TnI (ug/l) v době 1.RQA analýzy, ne starší 1 roku	CRP (mg/l) v době 1.RQA analýzy, ne starší 1 roku	ICHS (0-ne, 1-ano, 2-koronární nález bez ischémie)	Kompletní revaskulárizace? (0-NA, 1-ano, 2-ne)	Fibrilace síní (0-ne, 1-parox. 2-perzist/ chron.)	LVED D (mm)	LVEF (mm)	LA (mm)	NOS (0-11 ženy, 0-12 muži)	TIS (0-55 ženy, 0-60 muži)
0		0	15,18	5,2	2,6258	1,2		45,2		0	5,5					1	49	60	45	2	4
2	2005	0	16,59	13,3	3,367	1,08	62,5	40,2	0,68	0	3,1			0	0	0	54	60	41	5	14
0		0	16,47	6,1	2,7082	1,2		47,1		0	4,5									4	10
0		0	10,9	5,3	2,635	1,2		46,4		0	5			0	0	0	42	60	35	6	22
2	2007	0	16,22	11,2	3,1748	1,2	68,9	38,9	1,13	1	8,5									3	10
0		0	11,42	5	2,6075	1,2				0	5,4			0	0	0	40	60	38	0	0
1	1999	0	13,86	12,3	3,2755	1,2	70,9	45,5	0,94	1	2,9									3	6
0		0	11,64	4,3	2,5435	1,2		45,1		0	6,8			0	0	0	41	60	35	3	6
0		0	14,28	4,9	2,5984	1,2		42,6		0	5,5									2	4
0		0	13,16	4,7	2,5801	1,2		42,3		0	5,6			0	0	0	50	60	26	3	9
0		0	15,09	5,7	2,6716	1,11		42,1		0	4,8			0	0	0	36	60	30	2	6
2	1995	3	22,98	30,3	4,9225	1,2			2,18	0	3,3									6	14
1	1998	3	36,81	8,4	2,9186	1,2	71,3	43,2	0,28	0	3,9	méně 0,010								2	5
1	1985	3	17,31	6,3	2,7265	1,2	73,4	45,5	0,41	0	5,5	méně 0,010								4	11
1	1965	0	14,74	8,3	2,9095	1,2	58,5	32,5	1,67	1	3,1	0,012	14,9							7	11
2	2001	0	16,27	9,1	2,9827	1,2	77,1	43,9	8,91	1	4,5	méně 0,010								2	4
0		0	17,95	5,1	2,6167	1,15		44	0,5	0	5,1									5	11
0		0	14,1	5,2	2,6258	1,2		48,9		0	6,4									3	10
0		0	14,61	5	2,6075	1,2		46,4		0	4,4									1	2
2	2004	0	16,15	7	2,7905	1,2	77,4	45	1,39	0	4,8	méně 0,010	3,6							5	12
1	1992	0	11,47	5,8	2,6807	1,2	77,7	44,6	0,48	0	3,4	méně 0,010	méně 0,5							5	10
2	2000	3	14,88	8,1	2,8912	1,2		73,4	37,1	1	4,1	méně 0,010	21,3							3	10
2	2000	0	17,54	10,1	3,0742	1,2	67,1	45,1	0,85	1	3,7	méně 0,010	2,7							3	9
0		0	12,7	5	2,6075	1,2		44,8		0	7,1									0	0
0		0	13,37	4,2	2,5343	1,17		46,3		0	7,1									4	11
0		0	10,36	5,2	2,6258	1,2		47,1		0	5,8									3	8
0		0	12,14	5	2,6075	1,2	70,9	44,4		0	5,8									1	2
0		0	15,31	6	2,699	1,02		47,1		0	7,2									2	4
1	1977	0	13,15	8	2,882	1,2	65,2		2,08	0	3,6									2	5
0		0	12,99	5,6	2,6624	1,03	71,3			0	4,7									5	11
0		0	16,89	5,4	2,6441	1,2		44,7		0	4,7		méně 0,5							0	0
1	1993	0	14,91	10,1	3,0742	1,2	69,5		0,87	0	4,1	méně 0,010	3,6							5	12

Obrázek 3: Ukázka původní tabulky

3.2 Sjednocení formátu

Vzhledem k tomu, že data byla sbírána na různých odděleních, byl formát zápisu nejednotný. Bylo nutné upravit tabulku do takové podoby, aby bylo možné ji algoritmicky zpracovávat. Úprava musela být provedena ručně po předchozí analýze sledovaných atributů a zahrnovala:

- odstranění sloupců nesoucích stejnou informaci (ponechání pouze jednoho),
- odstranění slovních hodnocení nebo poznámek,
- sjednocení záhlaví,
- sjednocení formátu zápisu,
- sjednocení vyjádření chybějících hodnot.

Výstupem byla tabulka v jednotném formátu. V každém sloupci obsahovala pouze číselné hodnoty (definované v záhlaví) nebo označení pro chybějící (nezjištěnou) hodnotu.

3.3 Definice tříd

Vzhledem k zadanému úkolu bylo třeba pro klasifikaci vybrat pouze pacienty s diabetem. Data byla klasifikována do dvou tříd:

- 0 – pacienti s diabetem libovolného typu bez prokazatelné kardiální autonomní neuropatie
- 1 – pacienti s diabetem libovolného typu s kardiální autonomní neuropatií

Přítomnost či nepřítomnost kardiální autonomní neuropatie byla při sběru dat určena na základě výsledku testu hlubokého dýchání, kdy je při řízené dechové frekvenci 6 dechů/min kontinuálně odečítána srdeční frekvence. V každém dechovém cyklu je určen nejkratší (RRmin) a nejdelší (RRmax) R-R interval a vypočítán jejich průměr přes cykly. Následně je spočítán poměr RRmax/RRmin [21]. Pokud byla vypočítaná hodnota vyšší než stanovená norma pro věk, byl pacient zařazen do skupiny s KAN.

3.4 Předzpracování

Před samotnou klasifikací bylo nutné data upravit (předzpracovat). Předzpracování zahrnovalo odstranění nepoužitelných případů i atributů, ať již z důvodu velkého množství chybějících hodnot, nebo kvůli nevhodnosti z hlediska zadaného úkolu.

Data byla v rámci sloupců normalizována metodou Min-Max na rozsah $(0; 1)$ podle vzorce

$$n_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \tag{10}$$

kde n_i je normalizovaný i-tý prvek ze souboru, x_i je prvek na stejné pozici před normalizací a $\min(x)$ a $\max(x)$ je minimální a maximální hodnota v souboru před normalizací.

Aby byla zajištěna vyváženost klasifikátoru, bylo třeba dosáhnout stejného počtu případů ve skupině diabetiků bez KAN a ve skupině diabetiků s KAN. Nejjednodušší možnost eliminace nadbytečných případů z početnější skupiny nebyla v tomto případě vhodná, jelikož bylo k dispozici relativně malé množství případů. Nerovnováhu v počtu případů v jednotlivých skupinách lze vyřešit generováním nových (umělých) případů. Jednou z použitelných metod je algoritmus SMOTE (Synthetic Minority Over-sampling Technique) [31]. Nový případ je vytvářen tak, že se ze skupiny dostupných případů, kterou je nutné rozšířit, vybere jeden zástupce. Následně se metodou k-nearest neighbours najde k jeho nejbližších sousedů a náhodně se vybere jeden z nich. Určí se vektor rozdílů původního případu a vybraného souseda. Každý atribut nového případu se vypočítá tak, že se vezme vypočítaný rozdíl vynásobený náhodnou hodnotou z rozsahu $(0; 1)$ a přičte se k hodnotě atributu původního případu. Tedy hodnota atributu bude vždy mezi hodnotou původního případu a vybraného souseda [31].

Byly navrženy čtyři metody doplnění chybějících hodnot – doplnění průměrem, mediánem, metodou k-NN a doplnění na základě odhadu distribuční funkce. Metody byly porovnány tak, že při doplňování záměrně smazaných hodnot byla počítána odchylka od správné hodnoty. Z tohoto srovnání vyšla jako nejvhodnější metoda k-NN. Ta byla použita pro vytvoření finálního datasetu s doplněnými hodnotami, který sloužil jako jeden ze vstupů do klasifikátoru random forest. Alternativním vstupem byl dataset s nedoplněnými hodnotami.

Pro posouzení důležitosti atributů byly použity OOB případy. Při trénování byl počítán OOB error. Následně byly hodnoty v rámci atributu náhodně přeskupeny (permutovány) a byla zaznamenávána změna OOB erroru. Čím více vzroste po permutaci OOB error, tím je atributu přiřazena vyšší důležitost. Atributy byly seřazeny podle důležitosti. Bylo vyzkoušeno pět podvýběrů atributů a na základě AUC byla vybrána nejlepší varianta. Tento proces byl proveden zvlášť pro nedoplněná a zvlášť pro doplněná data.

3.5 Klasifikace

Vstupem do klasifikátoru byla normalizovaná data se stejným počtem případů v obou skupinách a sadou atributů vybraných podle jejich důležitosti. Za použití celého datasetu byly vybrány parametry klasifikátoru.

Důležitým parametrem klasifikátoru random forest je počet rozhodovacích stromů, ze kterých se skládá. Dalšími specifickými volenými parametry jsou velikost bootstrapového výběru, počet atributů použitých pro natrénování jednoho stromu a také další charakteristiky běžně používané u rozhodovacích stromů, například počet případů v terminálním uzlu.

Následně byl klasifikátor natrénován na nedoplněných a doplněných datech.

Úspěšnost klasifikátoru byla zhodnocena pomocí celkové přesnosti, senzitivity, specificity, pozitivní prediktivní hodnoty, negativní prediktivní hodnoty a ROC křivky s vypočtenou AUC.

4 Realizace

Na základě návrhu byly metody v pořadí uvedeném v předchozí kapitole aplikovány na reálná data. Po předzpracování následovala fáze klasifikace a zhodnocení dosažených výsledků. V následující kapitole jsou obsaženy detaily prováděných kroků a jejich dílčí výsledky.

4.1 Filtrace dat

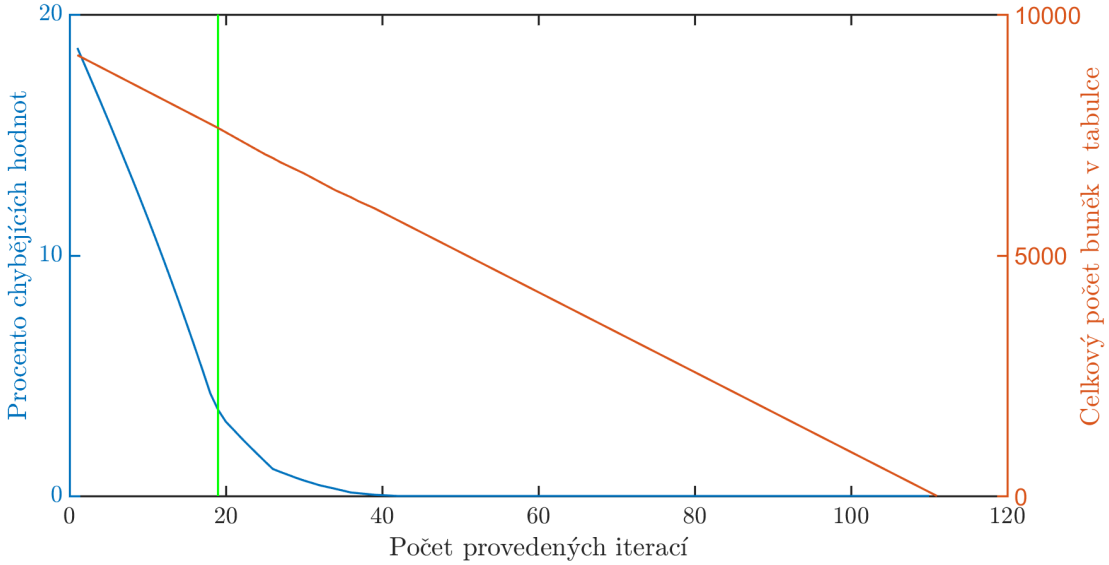
V první řadě bylo ze souboru vyřazeno 46 případů ze zdravé kontrolní skupiny, protože odlišení zdravých jedinců od diabetiků s KAN nebylo v souladu se zadáním práce. Vstupní data potom obsahovala 84 případů (45 diabetiků a 39 diabetiků s KAN).

Sloupec splňuje/nesplňuje normu pro věk u testu hlubokého dýchání musel být ze vstupního datasetu vyloučen, protože korelace s výsledkem byla velmi vysoká a k zařazení do správné třídy by stačil pouze tento jeden atribut. Protože stejná informace by se dala odvodit i ze zbývajících dvou sloupců testu hlubokého dýchání (změřená hodnota a norma pro věk), byly odstraněny i tyto sloupce.

Ve vstupních datech bylo velké množství chybějících hodnot. Jejich rozmístění nebylo náhodné. Velké množství chybějících hodnot vznikalo v případech, kdy se pravděpodobně upustilo od sledování jistého atributu a ten byl zaznamenán jen u dříve vyšetřených pacientů. Podobně tomu bylo u atributů, které byly přidány později a u předešlých pacientů dosud nebyly doplněny. Chybějící data v rámci řádků vznikala například pokud se pacient nedostavil na nějaké vyšetření, tedy celá sada souvisejících atributů nemohla být změřena a ohodnocena.

Cílem byla filtrace co největšího množství chybějících hodnot při zachování co nejvyššího počtu buněk tabulky.

Při filtraci bylo vypočítáno procento chybějících hodnot v rámci sloupců a řádků. Hodnoty byly seřazeny sestupně a postupně odstraňovány sloupce/řádky s nejvyšším procentem chybějících údajů. Po každém kroku byly zaznamenán celkový počet buněk tabulky, viz obrázek 4.



Obrázek 4: Závislost procenta chybějících hodnot (modře) a celkového rozměru tabulky (červeně) na počtu smazaných sloupců/řádků, zeleně je vyznačen výsledný stav po filtraci

Vyšší procento hodnot chybělo nejdříve v rámci sloupců, v několika případech to bylo téměř 100 %. Každý smazaný sloupec znamenal úbytek o stejný počet buněk, proto celkový rozměr tabulky klesal lineárně. Po smazání sloupců, které obsahovaly méně než 10 % dat, byl pokles procenta chybějících hodnot pozvolnější. V této fázi, po 19 iteracích, byla vytvořená tabulka uložena. Z grafu je patrné, že po provedení 42 iterací již byly odstraněny všechny chybějící hodnoty.

4.2 Normalizace

Data byla metodou Min-Max normalizována na rozsah $(0; 1)$. Důvodem normalizace bylo následné zamýšlené použití metody k-NN, kde se využívá euklidovská vzdálenost, a při různých rozsazích atributů by byl výsledek vychýlený. Random forest je k normalizaci invariantní, protože v algoritmu se nikdy vzájemně neporovnávají hodnoty různých atributů.

4.3 Doplňování

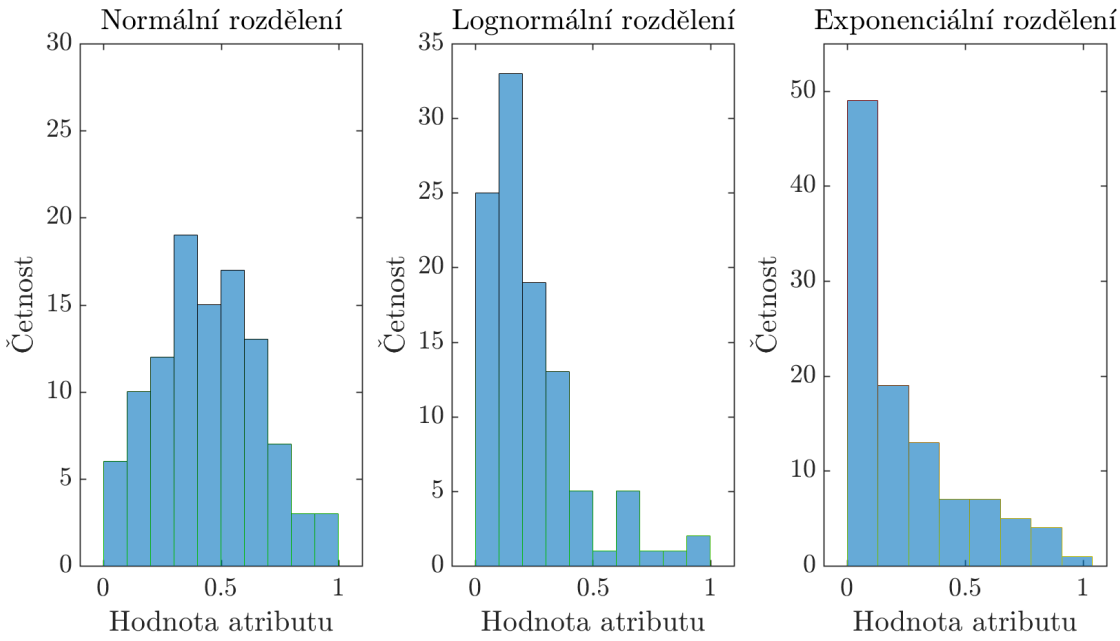
Zprvu bylo zásadním problémem datasetu velké množství chybějících hodnot. Ve výchozím stavu to bylo téměř 20 % z celkového počtu buněk. Nicméně při filtraci bylo

zjištěno, že většina pochází z atributů, které jsou vyplněny jen z pěti až deseti procent. Vzhledem k tomu, že dalším krokem při předzpracování bylo doplňování dat, musely být tyto atributy vyloučeny, protože jejich účinné doplnění by nebylo možné. Po provedení filtrace tabulka obsahovala celkem 280 chybějících hodnot, což představovalo méně než 4 % z celkového počtu buněk.

Byly vyzkoušeny a porovnány následující čtyři metody doplnění chybějících hodnot.

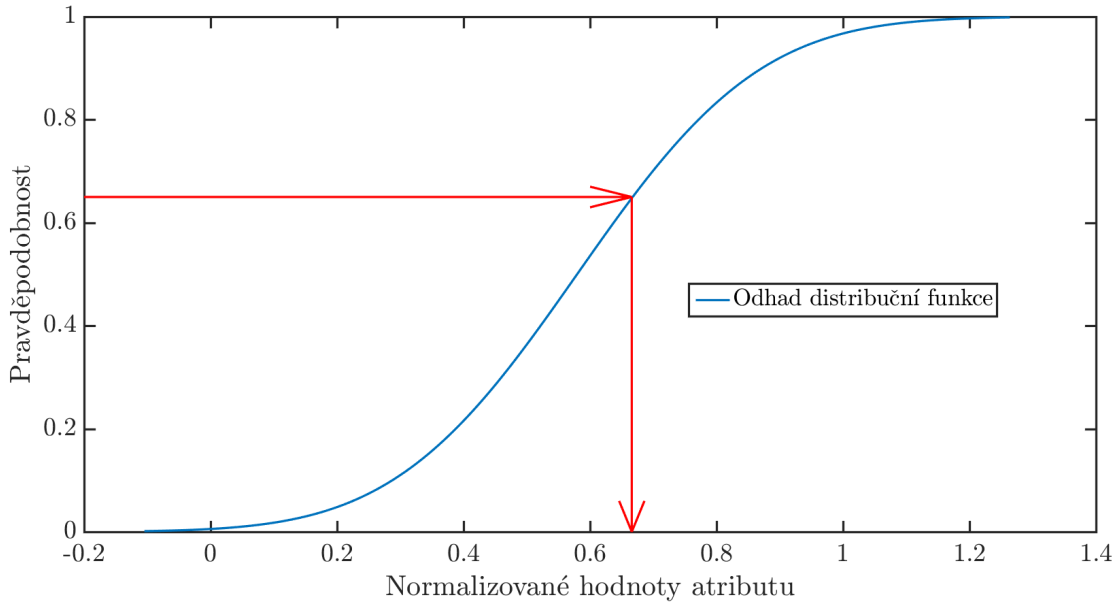
Nejdříve byla implementována a aplikována metoda pro doplnění hodnot aritmetickým průměrem vypočítaným z dostupných hodnot v rámci každého atributu. Další použitou metodou bylo doplňování mediánem určeným z dostupných hodnot [30].

Další metodou bylo doplnění na základě odhadu distribuční funkce. Distribuční funkce vystihuje rozložení dostupných hodnot. Chybějící elementy jsou doplněny tak, aby vyhovovaly vypočítanému rozložení. Za tímto účelem byly kvantitativní atributy rozděleny na tři skupiny podle předpokládaného rozdělení odhadnutého z histogramu – na veličiny s normálním rozdělením, lognormálním rozdělením a exponenciálním rozdělením, viz obrázek 5.



Obrázek 5: Ukázka histogramů normalizovaných hodnot atributů zařazených podle rozdělení do jedné ze tří skupin

Následně byly z existujících hodnot určeny parametry rozdělení (střední hodnota a případně rozptyl) a odpovídající distribuční funkce. Doplňovaná hodnota byla určena tak, že pro náhodně vygenerované číslo z rovnoměrného rozdělení z rozsahu $(0; 1)$ byla pomocí distribuční funkce odečtena příslušná hodnota na ose x (viz obrázek 6).



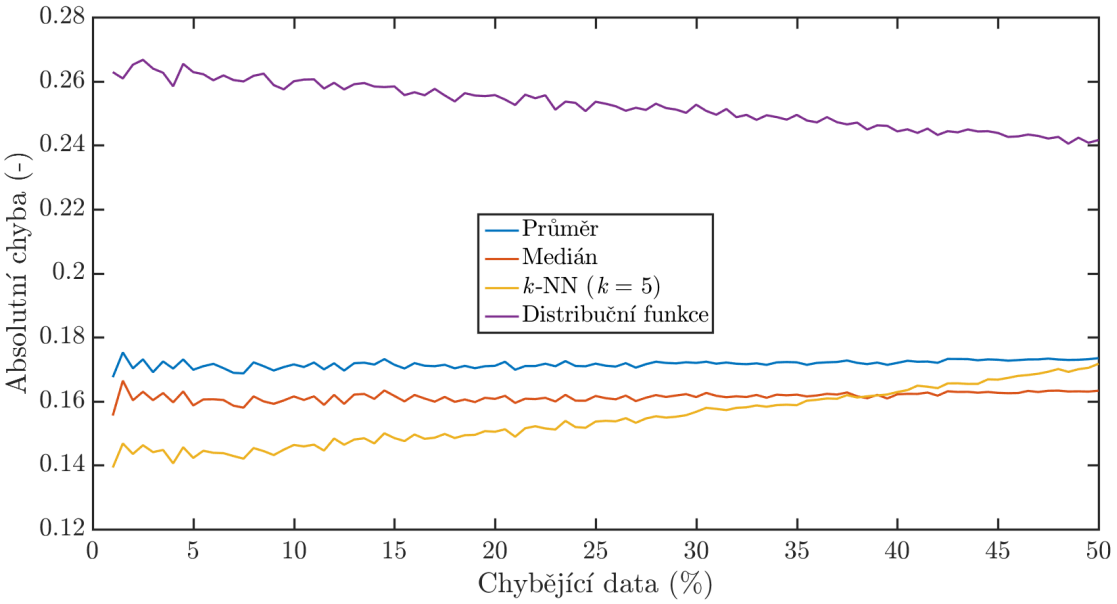
Obrázek 6: Doplnění chybějící hodnoty dle distribuční funkce odhadnuté z dostupných hodnot

Dalším implementovaným postupem bylo doplňování chybějících hodnot metodou k-NN (k-nearest neighbours) [30]. Metoda pracuje s jednotlivými případy (řádky). Pokud případ obsahuje chybějící hodnotu, vypočítá se jeho euklidovská vzdálenost od všech ostatních případů (chybějící hodnoty v obou případech jsou ignorovány). Vybere se k nejbližších sousedů a chybějící hodnota se doplní váženým průměrem hodnoty tohoto atributu k sousedů. Váha je upravena tak, že hodnoty s nižší vzdáleností mají větší váhu. Váha konkrétního prvku je nepřímo úměrná vzdálenosti sloupce, ve kterém se nachází, od sloupce, do nějž je doplňováno.

4.3.1 Srovnání použitých metod

Úspěšnost metod byla testována následovně. Byla vytvořena tabulka normalizovaných kompletních dat, z níž bylo určité procento hodnot náhodně odstraněno. Na tuto nekompletní tabulku byly opakovaně nezávisle použity čtyři metody pro

doplnění chybějících hodnot a byla počítána absolutní chyba. Stejný postup byl opakován pro různé množství odstraněných hodnot od 1 % do 50 % s krokem 0,5 % (viz obrázek 7). U metody k-NN bylo počítáno s 5 nejbližšími sousedy.

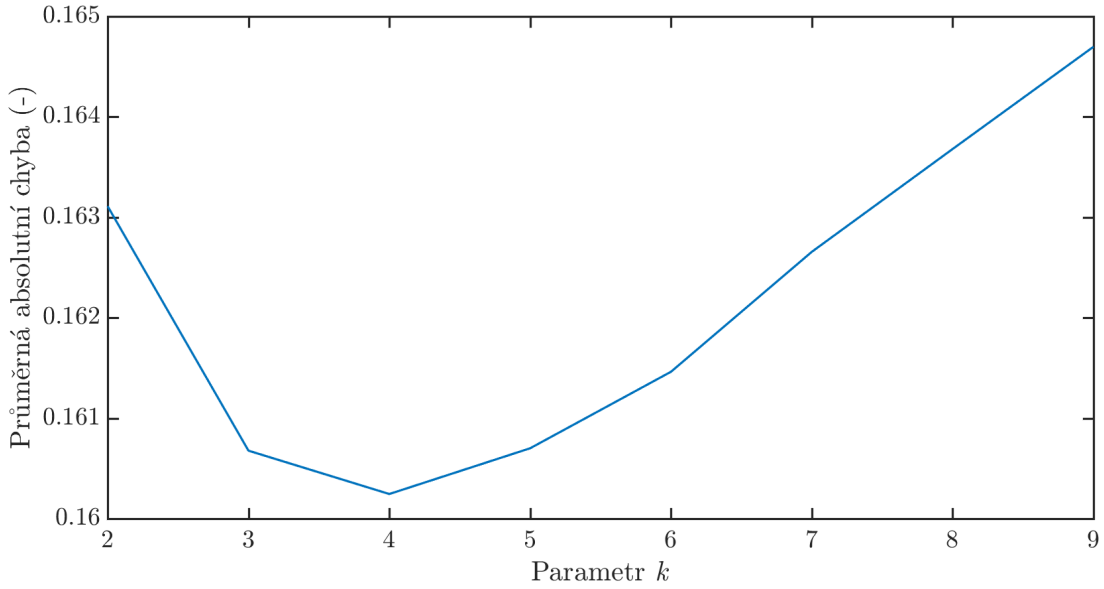


Obrázek 7: Průměr absolutní chyby ze sta doplnění jednotlivými metodami v závislosti na procentu odstraněných hodnot

4.3.2 Volba parametru metody *k*-NN

Z implementovaných metod dosáhla nejlepších výsledků metoda k-NN, proto byla zvolena pro doplnění dat použitých jako vstup do klasifikátoru.

U této metody je klíčovým parametrem zvolený počet nejbližších sousedů k. Při volbě parametru k byla cílem co nejnižší odchylka od správné hodnoty. Byla použita kompletní tabulka vytvořená za účelem srovnání metod pro doplňování hodnot. Z tabulky bylo náhodně smazáno 3,42 % hodnot, což odpovídá procentu chybějících hodnot v reálných datech. Tato tabulka byla tisíckrát doplňována metodou k-NN s různými hodnotami parametru k a byla počítána chyba. Parametr k pro doplnění reálně chybějících hodnot byl vybrán podle těchto výpočtů jako ten s minimální odchylkou (tzn. $k = 4$), viz obrázek 8.



Obrázek 8: Závislost průměrné absolutní chyby z tisíce doplnění na parametru k v metodě k-NN

4.4 Vyvážení skupin

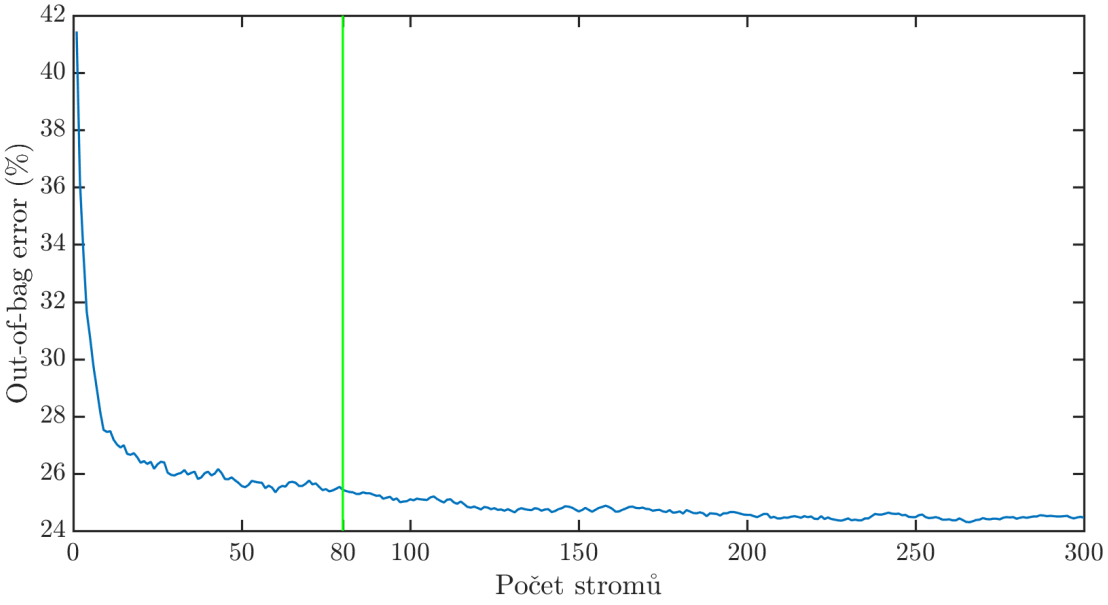
K vyvážení skupin byl implementován a použit algoritmus SMOTE, počet použitých nejbližších sousedů byl 5. Počet byl zvolen podle autorů metody [31]. Bylo třeba vytvořit šest nových případů ze skupiny diabetiků s KAN. Nejprve byl náhodně vybrán vzorek šesti případů, který byl vstupem do algoritmu SMOTE. Na základě každého vybraného případu a jednoho z pěti jeho nejbližších sousedů byl vygenerován jeden nový případ.

4.5 Optimalizace klasifikátoru

Po předzpracování byla provedena klasifikace pomocí metody random forest. V první řadě bylo nutné definovat vhodné parametry klasifikátoru.

Pro výběr optimálního počtu stromů byl na všech datech trénován les s počtem 300 stromů. Při volbě nejvhodnějšího počtu byl zohledněn odhad chyby klasifikace OOB error. V průběhu trénování byl po každém dalším vytvořeném stromu počítán OOB error. Výsledky zprůměrované ze sta opakování ukazuje obrázek 9.

Po počátečním strmém poklesu se hodnota OOB erroru snižuje velmi pozvolna. Jako vhodná volba se jeví hodnota bezprostředně po největším poklesu – 10 stromů. Ovšem při použití jiných dat by mohl být klasifikátor s tímto počtem stromů ne-



Obrázek 9: Závislost OOB erroru na počtu stromů, zeleně je vyznačena zvolená hodnota počtu stromů

stabilní. Je patrné, že přesnost klasifikátoru by byla tím vyšší, čím by byl vyšší počet stromů. Nicméně s rostoucím počtem stromů roste i čas nutný k natrénování klasifikátoru a příspěvek dalšího navyšování počtu stromů není tak výrazný. Počet stromů byl stanoven na 80.

Dalším voleným parametrem byla velikost bootstrapového výběru pro trénování. Podle hodnoty popisované v článku autora metody random forest [32] byla definována jako dvě třetiny z celkového počtu případů.

Počet náhodně vybraných prediktorů m byl ponechán na doporučené hodnotě pro klasifikační úlohy, tedy odmocnina z celkového počtu prediktorů.

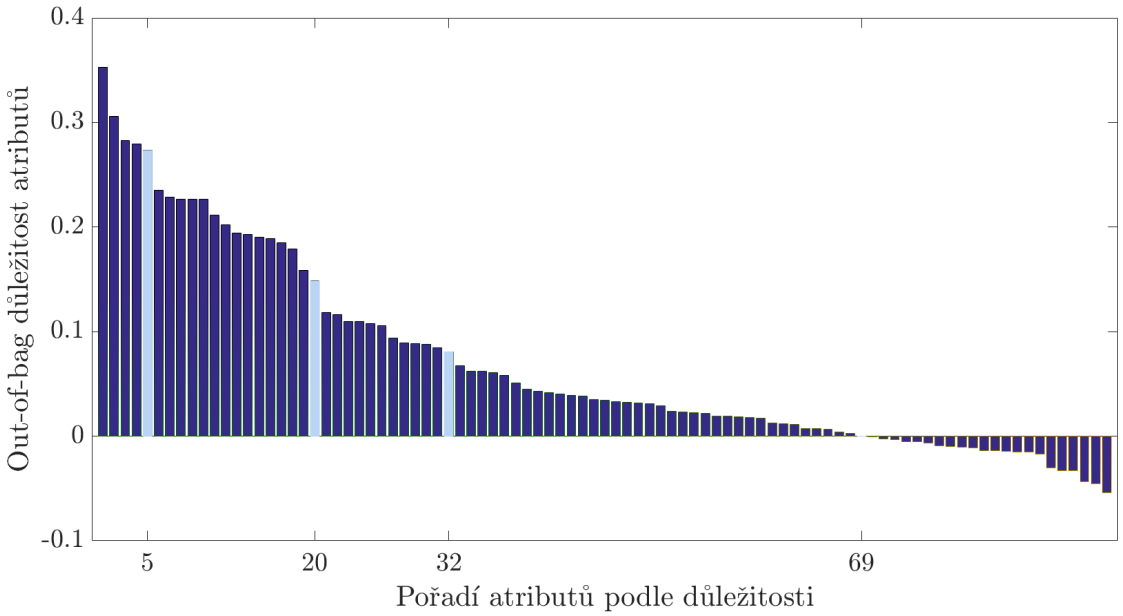
Klasifikátor se stejnými parametry byl použit i pro výpočet důležitosti atributů a srovnání kvality klasifikace vstupů s různým počtem atributů.

4.6 Důležitost atributů

Datasety po dosavadních úpravách (filtrace, normalizace, vyvážení skupin, případně doplnění chybějících hodnot) obsahovaly 90 případů a 91 atributů. Výpočet důležitosti atributů byl proveden za účelem redukce dimenzionality. Předpoklad byl takový, že některé atributy mají na zařazení do výsledné třídy větší vliv než jiné. Cílem bylo vybrat pouze ty atributy, jejichž absence by způsobovala největší chybu klasifikace.

Při vysokém počtu atributů má klasifikátor tendenci nacházet i velmi jemné vztahy mezi atributy a na tento „šum“ se adaptovat.

K určení důležitosti atributů byly využity OOB případy. Při trénování klasifikátoru byl počítán OOB error a byl sledován vzrůst OOB erroru při náhodné permutaci prvků konkrétního atributu. Z tohoto vzrůstu (případně poklesu) byla vypočítána důležitost atributu. Postup byl opakován stokrát, čímž bylo pro každý atribut získáno 100 hodnocení důležitosti. Výsledky pro jednotlivé atributy byly zprůměrovány a atributy byly seřazeny podle vypočítané důležitosti sestupně. Výsledky pro původní soubor s chybějícími hodnotami ukazuje obrázek 10.

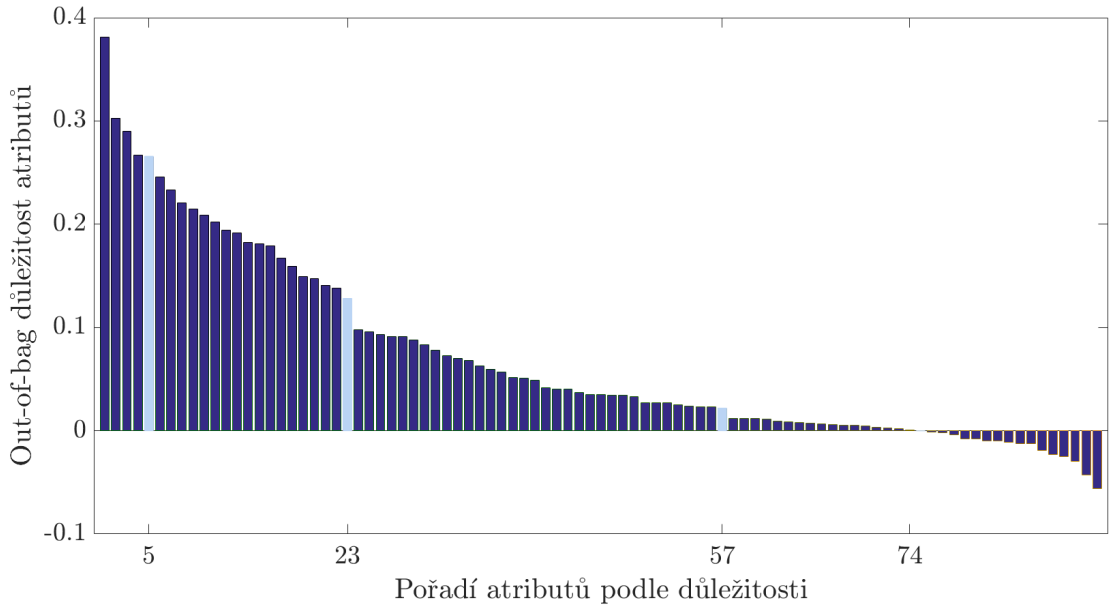


Obrázek 10: OOB důležitost atributů pro soubor s nedoplněnými hodnotami, světle modře jsou označeny mezní sloupce podvýběrů

Byly vytvořeny čtyři nové podmnožiny atributů. Hodnoty byly zvoleny na základě grafu v místech předcházejících prudšímu poklesu. První podmnožina obsahovala pouze atributy s vypočítanou OOB důležitostí větší než nula, jejich počet byl 69. Důležitost menší než nula značí, že při permutaci prvků přesnost klasifikace vzrostla. Další výběry obsahovaly pouze 5, 20 a 32 atributů s nejvyšší dosaženou důležitostí.

Obrázek 11 ukazuje výsledky pro doplněný soubor. Postup byl stejný jako u nedoplněných dat. Rozdíl v důležitosti proti nedoplněným datům byl velmi malý, pořadí se ovšem mírně lišilo, proto byl každý dataset vyhodnocen zvlášť. U doplněných dat

byly z grafu stejným způsobem jako k předchozím případě zvoleny mezní hodnoty počtu atributů 5, 23, 57 a 74.



Obrázek 11: OOB důležitost atributů pro soubor s doplněnými hodnotami, světle modře jsou označeny mezní sloupce podvýběrů

4.7 Klasifikace

Vytvořené datasety sloužily jako vstup do klasifikátoru random forest. K vyhodnocení byla použita leave-one-out křížová validace. Byly porovnány hodnoty plochy pod ROC křivkou (AUC) pro jednotlivé vstupy. Výsledky shrnuje tabulka 3.

Pro závěrečné srovnání byly vybrány vstupy s nejvyšší hodnotou AUC. Pro nedoplněná data to byl dataset s 32 nejdůležitějšími atributy, pro doplněná data soubor s 57 atributy.

Tabulka 3: Vypočítané hodnoty AUC pro klasifikaci nedoplněných a doplněných dat s různým počtem atributů, tučně jsou vyznačeny nejvyšší dosažené výsledky a odpovídající počty atributů

	Nedoplněná data				
Počet atributů	5	20	**32**	69	91
AUC	0,7638	0,8097	**0,8274**	0,8197	0,8079
	Doplněná data				
Počet atributů	5	23	**57**	74	91
AUC	0,7603	0,8152	**0,8314**	0,8234	0,8199

4.8 Hodnocení kvality klasifikace

Vzhledem k poměrně nízkému počtu případů byla k vyhodnocení byla použita křížová validace. Pro porovnání byla provedena desetinásobná křížová validace a leave-one-out křížová validace. Informace o klasifikaci prvků z testovací množiny byly uchovávány po dobu jednoho průběhu validace. Z výsledných zařazení testovacích případů v rámci jedné validace byla vytvořena kontingenční tabulka a vypočítány ostatní parametry, včetně ROC křivky. Celý proces byl opakován desetkrát a výsledky vypočítaných parametrů jsou prezentovány ve formě průměr ± směrodatná odchylka, viz tabulka 6. V kontingenční tabulce byly výsledky pouze průměrovány, viz tabulky 4 a 5.

5 Výsledky

U klasifikátorů byly zaznamenávány a vyhodnocovány ROC křivky, kontingenční tabulky a z nich vypočítané parametry přesnost, senzitivita, specificita, pozitivní a negativní prediktivní hodnota.

Byly využity dvě možnosti validace – desetinásobná křížová validace a leave-one-out křížová validace. Nižší směrodatně odchylky od průměrů vypočítaných parametrů byly získány z leave-one-out křížové validace, průměrné hodnoty byly zpravidla vyšší. Výsledky obou metod jsou pro srovnání uvedeny v příloze A a B.

5.1 Kontingenční tabulky

Kontingenční tabulky obsahují počet případů zařazených do jednotlivých tříd vzhledem k reálné třídě, do které případy náleží.

Tabulka 4: Zprůměrovaná kontingenční tabulka pro klasifikaci nedoplněných dat (metoda leave-one-out křížová validace)

		Reálný stav	
		třída 1	třída 0
Výsledek klasifikace	třída 1	36,2	8,8
	třída 0	13,5	31,5

Tabulka 5: Zprůměrovaná kontingenční tabulka pro klasifikaci doplněných dat (metoda leave-one-out křížová validace)

		Reálný stav	
		třída 1	třída 0
Výsledek klasifikace	třída 1	37,6	7,4
	třída 0	14,3	30,7

5.2 Odvozené parametry

Následující charakteristiky byly vypočítány z kontingenční tabulky pro každé provedení leave-one-out křížové validace, na závěr byly výsledky zprůměrovány.

Pro nedoplněná data byla průměrná celková přesnost 75,22 %, pro doplněná vyšla přesnost 75,89 %. V tabulce 6 je uveden průměr a směrodatná odchylka tohoto i ostatních vypočítaných parametrů.

Tabulka 6: Odvozené parametry klasifikátoru pro nedoplněná a doplněná data (metoda leave-one-out křížová validace)

	Nedoplněná data		Doplněná data	
	průměr	směrodatná odchylka	průměr	směrodatná odchylka
AUC	0,83	0,01	0,84	0,01
celková přesnost (%)	75,22	1,58	75,89	1,58
senzitivita (%)	72,24	1,21	72,44	1,16
specificita (%)	78,24	2,54	80,65	2,62
pozitivní prediktivní hodnota (%)	80,44	2,93	83,56	2,81
negativní prediktivní hodnota (%)	70,00	1,57	68,22	1,50

Srovnání

K porovnání přesností klasifikátorů byl využit McNemarův test s Edwardsovou úpravou. Za tím účelem byla provedena leave-one-out křížová validace a byly zaznamenávány informace o klasifikaci testovacího případu klasifikátorem natrénovaným na nedoplněných datech (klasifikátor N) a na doplněných datech (klasifikátor D), viz tabulka 7.

Tabulka 7: Zprůměrovaná tabulka výsledků klasifikace dvěma klasifikátory

		klasifkátor D	
		třída 1	třída 0
klasifkátor N	třída 1	64,0	2,8
	třída 0	4,2	18,9

Nulová hypotéza je formulována tak, že přesnost klasifikátorů *acc* není statisticky významně odlišná, tedy že umělé doplnění chybějících dat neovlivní celkovou přesnost klasifikátoru random forest. Alternativní hypotéza je, že přesnost je po doplnění statisticky významně odlišná.

$$H_0 : acc_N = acc_D$$
$$H_1 : acc_N \neq acc_D$$

Výpočet testovacího kritéria:

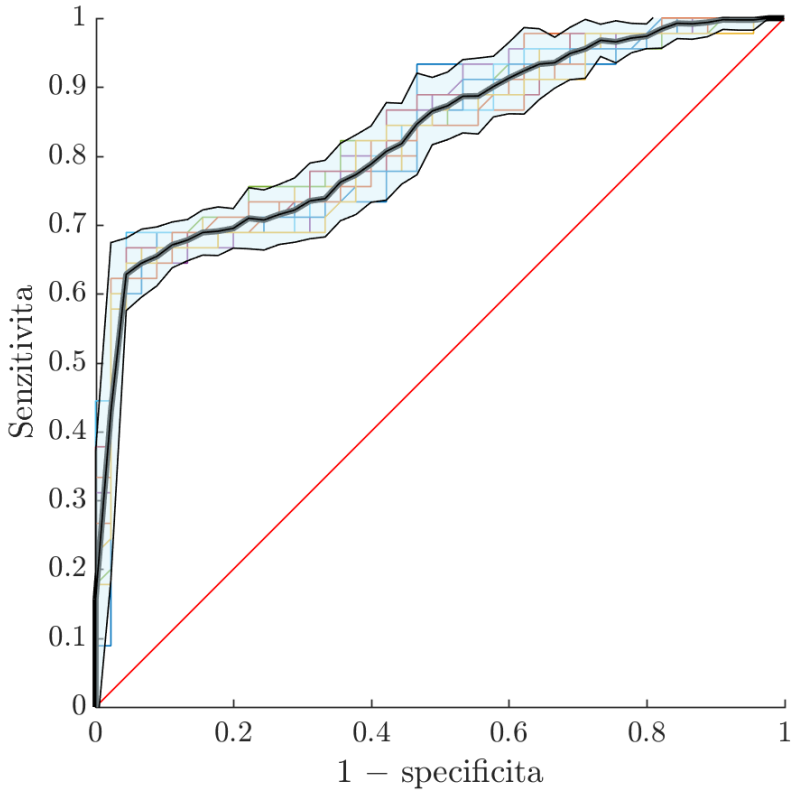
$$\begin{aligned} \chi^2 &= \frac{(\mid 2,8 - 4,2 \mid -1)^2}{2,8 + 4,2} \\ \chi^2 &= 0,057 \end{aligned}$$

Výsledek byl porovnán s kritickou hodnotou χ^2 rozdělení se zvolenou hladinou významnosti 0,05 a jedním stupněm volnosti, $\chi^2_{0,95}(1) = 3,84$.

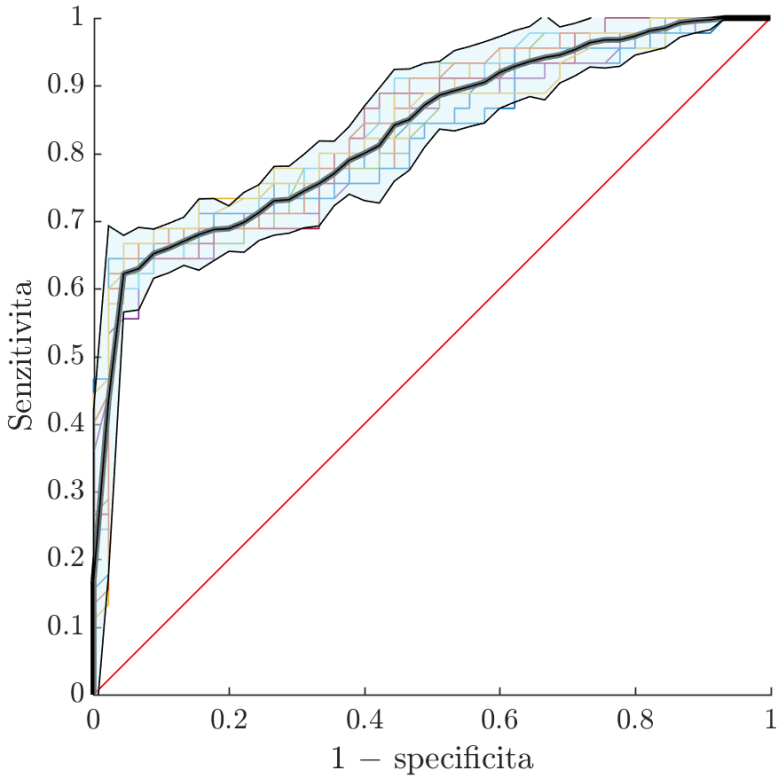
Jelikož je testovací kritérium menší než kritická hodnota, nezamítáme nulovou hypotézu, lze tedy i nadále předpokládat, že doplnění hodnot neovlivňuje klasifikaci.

5.3 ROC křivky

V rámci vyhodnocení bylo získáno deset ROC křivek pro deset průběhů leave-one-out křížové validace. Hodnoty ROC křivek byly průměrovány a je zobrazována průměrná hodnota ROC křivek spolu s pásem ± 2 směrodatné odchylky. Pás pokrývá oblast, kde se křivka nachází s 95% pravděpodobností, viz obrázky 12 a 13.



Obrázek 12: ROC křivka klasifikace souboru s nedoplněnými hodnotami



Obrázek 13: ROC křivka klasifikace souboru s doplněnými hodnotami

Srovnání

Cílem bylo vyšetřit odlišnost dosažených ROC křivek. K porovnání byly použity vypočtené plochy pod těmito křivkami (AUC). Předpoklad byl takový, že střední hodnoty ani směrodatné odchylky AUC nebudou při opakovaném sestrojení ROC křivky z klasifikace nedoplněných a doplněných dat odlišné.

Za tímto účelem byla s nedoplněnými daty stokrát provedena leave-one-out křížová validace, výstupem bylo 100 vypočítaných AUC, označených jako AUC_N. Stejný postup byl opakován s doplněnými daty a byl získán soubor AUC_D.

K testování normality souborů AUC_N a AUC_D byl použit χ^2 test dobré shody. Testováno bylo na hladině významnosti 5 %. Pro AUC_N byla p-value 0,64, pro AUC_D 0,89. Normalita obou souborů byla potvrzena.

Pro porovnání rozptylů byl použit dvouvýběrový F-test. Nulová hypotéza říká, že rozptyl obou výběrů (σ_N^2 a σ_D^2) se neliší, alternativní hypotéza popisuje opačnou situaci.

$$H_0 : \sigma_N^2 = \sigma_D^2$$
$$H_1 : \sigma_N^2 \neq \sigma_D^2$$

Rozptyl AUC_N byl $0,68 \cdot 10^{-5}$, pro AUC_D byl $1,17 \cdot 10^{-5}$, počet prvků v souboru n_N a n_D byl v obou případech 100.

Výpočet testovacího kritéria:

$$\begin{aligned} F &= \frac{\sigma_D^2}{\sigma_N^2} \\ F &= \frac{1,17 \cdot 10^{-5}}{0,69 \cdot 10^{-5}} \\ F &= 1,70 \end{aligned}$$

Výsledek byl porovnán s kritickou hodnotou F-rozdělení se zvolenou hladinou významnosti 0,05 a stupni volnosti $n_N - 1$ a $n_D - 1$, $F_{0,95}(99, 99) = 1,39$.

Jelikož je testovací kritérium vyšší než kritická hodnota, zamítáme nulovou hypotézu, rozptyly jsou na dané hladině významnosti odlišné.

Pro porovnání středních hodnot výběrů byl použit dvouvýběrový t-test. Nulová hypotéza říká, že střední hodnoty obou výběrů (μ_N a μ_D) se neliší, alternativní hypotéza říká, že střední hodnoty se liší.

$$H_0 : \mu_N = \mu_D$$
$$H_1 : \mu_N \neq \mu_D$$

Výpočet testovacího kritéria:

$$\begin{aligned} t &= \frac{\mid \mu_N - \mu_D \mid}{\sqrt{\frac{\sigma_N^2}{n_N} + \frac{\sigma_D^2}{n_D}}} \\ t &= \frac{\mid 0,8313 - 0,8295 \mid}{\sqrt{\frac{0,69 \cdot 10^{-5}}{100} + \frac{1,17 \cdot 10^{-5}}{100}}} \\ t &= 1,31 \end{aligned}$$

Výsledek byl porovnán s kritickou hodnotou t-rozdělení se zvolenou hladinou významnosti 0,05 a vypočítaným stupněm volnosti vzhledem k odlišnosti rozptylů, který byl 185. (Při stejných rozptylech by byl počet stupňů volnosti $n_N - n_D - 2$.) Kritická hodnota je $t_{0,95}(185) = 1,97$.

Jelikož je testovací kritérium nižší než kritická hodnota, nezamítáme nulovou hypotézu, shoda středních hodnot byla potvrzena.

6 Diskuze

Při filtraci byly ze zpracování úplně vyloučeny případy osob, u kterých nebylo diagnostikováno ani jedno ze sledovaných onemocnění. Takový klasifikátor by v praxi nepřinášel žádnou výhodu, protože k rozlišení skupin stačí pouze přítomnost diabetu a diagnostika tohoto onemocnění není nijak komplikovaná. Ovšem zdravé případy by mohly být užitečné ve fázi předzpracování, například při doplňování hodnot.

Po eliminaci sloupců s více než 90 % chybějících hodnot zůstalo v datech 280 chybějících hodnot, což představovalo 3,42 % z celkového počtu buněk v tabulce. Nebylo ověřeno, že dosažené výsledky by byly obdobné při vyšším procentu chybějících hodnot.

Normalizace byla aplikována z důvodu metod používajících výpočet vzdálenosti (SMOTE, k-NN), z pohledu vybraného klasifikátoru nebyla nutná. Bylo by možné použít jiné metody normalizace, např. Z-score nebo Soft-max normalizaci.

Byly implementovány čtyři metody pro doplnění chybějících hodnot – průměrem, mediánem, doplnění metodou k-NN a doplnění podle odhadu distribuční funkce. Pro výsledné doplnění byla zvolena metoda k-NN s počtem nejbližších sousedů 4. Doplňování pomocí odhadu distribuční funkce vykazuje proti očekávání poměrně velkou chybu. Pravděpodobnou příčinou je nízký počet případů. Při zvýšení počtu případů by bylo možné lépe aproximovat rozdělení daného atributu, což by zlepšilo přesnost metody. Perspektivně by bylo možné přidat další metody doplnění chybějících hodnot, případně vybrat více metod a porovnat jejich vhodnost podle chyby klasifikace při jejich uplatnění.

Za účelem vyvážení skupin byl použit algoritmus SMOTE. Namísto syntetického vytváření nových případů by bylo možné zvolit váhování, kdy případy v minoritní skupině mají vyšší váhu a jejich nesprávná klasifikace je výrazněji penalizována.

Důležitosti atributů byla vypočítána pomocí OOB případů v průběhu trénování klasifikátoru random forest. Ke srovnání klasifikace různých sad atributů byla použita plocha pod ROC křivkou. Alternativně by bylo možné vyzkoušet jinou možnost výběru atributů a porovnat výsledky.

McNemarův test prokázal, že celková přesnost obou klasifikátorů není na hladině významnosti 5 % odlišná. Použití testu bylo problematické vzhledem k počtu případů, které byly jednotlivými klasifikátory zařazeny do odlišných tříd. Tyto údaje

se používají k výpočtu testovacího kritéria. Ke spolehlivějšímu výpočtu by bylo nutné mít k dispozici více případů.

Testování rozdílnosti ROC křivek proběhlo na základě parametru AUC. Byl porovnáván soubor 100 hodnot AUC z klasifikace nedoplněných dat a 100 hodnot AUC z klasifikace doplněných dat. Na hladině významnosti 5 % nebyla F-testem potvrzena rovnost rozptylů výběru, t-test však na stejné hladině významnosti potvrdil shodu středních hodnot.

Z dosažených výsledků plyne, že syntetické doplnění chybějících hodnot metodou k-NN významně neovlivní kvalitu klasifikátoru random forest z pohledu celkové přesnosti ani AUC. Použitá metoda má vliv na rozptyl hodnot AUC při opakovaném vyhodnocení klasifikace.

Závěr

Práce se zabývala problematikou klasifikace medicínských dat.

Na reálných datech bylo provedeno předzpracování zahrnující filtraci, normalizaci, vyvážení skupin a doplnění chybějících hodnot. Byl vytvořen klasifikátor a zhodnocena klasifikace. Pro doplnění chybějících hodnot byla zvolena metoda k-NN, použitým klasifikátorem byl random forest. K vyhodnocení byla použita desetinásobná křížová validace a leave-one-out křížová validace. Hodnocenými parametry byla ROC křivka a charakteristiky vypočítané z kontingenční tabulky – celková přesnost, senzitivita, specificita, pozitivní a negativní prediktivní hodnota.

Práce obsahuje popis všech aplikovaných metod a porovnání výsledků klasifikace pro doplněná a nedoplněná data.

Výstupem byly dva klasifikátory, klasifikátor natrénovaný na nedoplněných datech dosahuje celkové přesnosti (75,22 ± 1,58) % a klasifikátor natrénovaný na doplněných datech má přesnost (75,89 ± 1,58) %.

Bylo potvrzeno, že hodnocené parametry nejsou pro původní data s chybějícími hodnotami a pro data s uměle doplněnými hodnotami na hladině významnosti 5 % odlišné, tedy že umělé doplnění chybějících hodnot metodou k-NN nemá vliv na kvalitu klasifikátoru random forest.

Seznam použitých zdrojů

[1] ASHWINKUMAR, U. M. a K. R. ANANDAKUMUR. Ethical and Legal Issues for Medical Data Mining. *International Journal of Computer Applications*. 2010, **1**(28), s. 7-11.

[2] LHOTSKÁ, Lenka, Miroslav BURŠA, Michal HUPTYCH a Matěj HRACHOVINA. Big data versus vzácné případy. In: *MEDSOFT*. Praha: Tech-market, 1996, s. 117-125. ISSN 1803-8115.

[3] CIOS, Krzysztof a George William MOORE. Uniqueness of Medical Data Mining. *Artificial Intelligence in Medicine*. 2002, **26**(1-2), s. 1-24. DOI: 10.1016/S0933-3657(02)00049-0. ISSN 09333657. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/S0933365702000490

[4] CISMONDI, Federico, André S. FIALHO a Susana M. VIEIRA et al. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine*. 2013, **58**(1), s. 63-72. DOI: 10.1016/j.artmed.2013.01.003. ISSN 09333657. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/S0933365713000055

[5] UENAL, Hatice, Benjamin MAYER a Jean-Baptist DU PREL. Choosing Appropriate Methods for Missing Data in Medical Research: A Decision Algorithm on Methods for Missing Data. *Journal of Applied Quantitative Methods*. 2014, **9**(4), s. 10-21.

[6] SONI, Jyoti, Ujma ANSARI, Sunita SONI a Dipesh SHARMA. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*. 2011, **17**(8), s. 43-48.

[7] GREEN, Michael, Jonas BJÖRK, Jakob FORBERG et al. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. *Artificial Intelligence in Medicine* [online]. 2006, **38**(3), s. 305-318 [cit. 2016-04-27]. DOI: 10.1016/j.artmed.2006.07.006. ISSN 09333657. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0933365706001059

[8] PETER, T. John a Kumar SOMASUNDARAM. An empirical study on prediction of heart disease using classification data mining techniques. In: *2012 International Conference on Advances in Engineering, Science and Management (ICAESM - 2012)*. Nagapattinam: IEEE, 2012, s. 514-518. ISBN 978-81-909042-2-3.

[9] AUSTIN, Peter C. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine*. 2007, **26**(15), s. 2937-2957. DOI: 10.1002/sim.2770. ISSN 02776715. Dostupné také z: http://doi.wiley.com/10.1002/sim.2770

[10] KURT, Imran, Mevlut TURE a A. Turhan KURUM. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*. 2008, **34**(1), s. 366-374. DOI: 10.1016/j.eswa.2006.09.004. ISSN 09574174. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/S0957417406002855

[11] MAROCO, João et al. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes* [online]. 2011, **4**(1), s. 299-312 [cit. 2016-04-27]. DOI: 10.1186/1756-0500-4-299. ISSN 1756-0500. Dostupné z: http://www.biomedcentral.com/1756-0500/4/299

[12] LEHMANN, Christoph, Thomas KOENIG, Vesna JELIC et al. Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *Journal of Neuroscience Methods*. 2007, **161**(2), s. 342-350. DOI: 10.1016/j.jneumeth.2006.10.023. ISSN 01650270. Dostupné také z: http://linkinghub.elsevier.com/retrieve/pii/S0165027006005425

[13] RIBEIRO, Ricardo T., Rui Tato MARINHO a J. Miguel SANCHES. Classification and Staging of Chronic Liver Disease From Multimodal Data. *IEEE Transactions on Biomedical Engineering*. 2013, **60**(5), s. 1336-1344.

[14] KUMAR, Sharddha. A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier. *International Journal of Computer Applications*. 2015, **120**(8), s. 36-39.

[15] HASSAN, Syeda et al. Bioprocess data mining using regularized regression and random forests. *BMC Systems Biology* [online]. 2013, **7**(1), s. 5-11 [cit. 2016-04-27]. DOI: 10.1186/1752-0509-7-S1-S5. ISSN 1752-0509. Dostupné z: http://www.biomedcentral.com/1752-0509/7/S1/S5

[16] CEVALLOS VALDIVIEZO, Holger a Stefan VAN AELST. Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences* [online]. 2015, 311, s. 163-181 [cit. 2016-05-14]. DOI: 10.1016/j.ins.2015.03.018. ISSN 00200255. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0020025515001838

[17] DÍAZ-URIARTE, Ramón a Sara ALVAREZ DE ANDRÉS. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* [online]. **7**(1), s. 3-15 [cit. 2016-04-27]. DOI: 10.1186/1471-2105-7-3. ISSN 14712105. Dostupné z: http://www.biomedcentral.com/1471-2105/7/3

[18] CUTLER, D. Richard, Thomas C. EDWARDS, Jr. a Karen H. BEARD et al. Random forests for classification in ecology. *Ecology*. 2007, **88**(11), s. 2783-2792.

[19] BELLAZZI, Riccardo a Blaz ZUPAN. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics* [online]. 2008, **77**(2), s. 81-97 [cit. 2016-05-09]. DOI: 10.1016/j.ijmedinf.2006.11.006. ISSN 13865056. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S1386505606002747

[20] American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*. 2004, **27**(1), s. 5-10.

[21] LACIGOVÁ, Silvie, Alexandra JIRKOVSKÁ a Zdeněk RUŠAVÝ. Standardy diagnostiky a léčby diabetické neuropatie. *Diabetes, metabolismus, endokrinologie a výživa (DMEV)*. 2012, **15**(1), s. 36-40.

[22] KOMPRDOVÁ, Klára. *Rozhodovací stromy a lesy*. Vyd. 1. Brno: Akademické nakladatelství CERM, 2012. ISBN 978-80-7204-785-7.

[23] Random forests – classification description. *Statistics at UC Berkeley: Department of Statistics*. [online]. © 2014 [cit. 2016-05-14]. Dostupné z: https://www.stat.berkeley.edu/ breiman/RandomForestscc_home.htm

[24] OSHIRO, Thais Mayumi, Pedro Santoro PEREZ a José Augusto BARANAUSKAS. How Many Trees in a Random Forest? [online]. s. 154-168 [cit. 2016-05-10]. DOI: 10.1007/978-3-642-31537-4_13. Dostupné z: http://link.springer.com/10.1007/978-3-642-31537-4_13

[25] Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value . *STAT 507*. [online]. © 2016 [cit. 2016-04-29]. Dostupné z: https://onlinecourses.science.psu.edu/stat507/node/71

[26] Principy imunoanalytických metod: ROC křivka. *Creative Education – Vital Application*. [online]. © 2016 [cit. 2016-04-29]. Dostupné z: http://www.ceva-edu.cz/mod/book/view.php?id=3088&chapterid=1642

[27] ROC Curve. *IBM Knowledge Center*. [online]. © 2013 [cit. 2016-04-29]. Dostupné z: http://www.ibm.com/support/knowledgecenter/SSLVMB_22.0.0/com.ibm.spss.statistics.cs/spss/tutorials/roc_curve_bankloan_01.htm

[28] WESTFALL, Peter H., James F. TROENDLE a Gene PENNELLO. Multiple McNemar Tests. *Biometrics* [online]. 2010, **66**(4), s. 1185-1191 [cit. 2016-05-06]. DOI: 10.1111/j.1541-0420.2010.01408.x. ISSN 0006341x. Dostupné z: http://doi.wiley.com/10.1111/j.1541-0420.2010.01408.x

[29] Two-Sample t-Test for Equal Means. *e-Handbook of Statistical Methods*. [online]. 30.10.2013 [cit. 2016-05-10]. Dostupné z: http://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm

[30] BATISTA, Gustavo E.A.P.A. a Maria Carolina MONARD. A Study of K-Nearest Neighbour as an Imputation Method. In: *Second International Conference on Hybrid Intelligent Systems, Santiago, Chile*, s. 251–260. IOS Press, Amsterdam (2002)

[31] CHAWLA, Nitesh, Kevin BOWYER, Lawrence HALL a Philip KEGELMEYER. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligenece Research*. 2002, 16: s. 321-357.

[32] BREIMAN, Leo. Random Forests. *Machine Learning*. 2001, **45**(1), s. 5-32. DOI: 10.1023/A:1010933404324.

Seznam obrázků

Seznam tabulek

Seznam příloh

Příloha A: Kontingenční tabulky

Tabulka 1: Zprůměrované kontingenční tabulky pro klasifikaci nedoplněných a doplněných dat (metoda desetinásobné křížové validace)

		Reálný stav			
		Nedoplněná data		Doplněná data	
		třída 1	třída 0	třída 1	třída 0
Výsledek klasifikace	třída 1	34,4	10,6	37,6	7,4
	třída 0	13,7	31,3	13,8	31,2

Tabulka 2: Zprůměrované kontingenční tabulky pro klasifikaci nedoplněných a doplněných dat (metoda leave-one-out křížové validace)

		Reálný stav			
		Nedoplněná data		Doplněná data	
		třída 1	třída 0	třída 1	třída 0
Výsledek klasifikace	třída 1	36,2	8,8	37,9	7,1
	třída 0	13,5	31,5	14,4	30,6

Příloha B: Odvozené parametry klasifikace

Tabulka 3: Odvozené parametry klasifikátoru pro nedoplněná a doplněná data (metoda desetinásobné křížová validace)

	Nedoplněná data		Doplněná data	
	průměr	směrodatná odchylka	průměr	směrodatná odchylka
AUC	0,83	0,01	0,83	0,01
celková přesnost (%)	73,00	1,58	76,44	1,72
senzitivita (%)	71,52	1,06	73,16	1,41
specificita (%)	74,83	2,87	80,93	2,95
pozitivní prediktivní hodnota (%)	76,44	3,81	83,56	3,18
negativní prediktivní hodnota (%)	69,56	1,83	69,33	2,04

Tabulka 4: Odvozené parametry klasifikátoru pro nedoplněná a doplněná data (metoda leave-one-out křížová validace)

	Nedoplněná data		Doplněná data	
	průměr	směrodatná odchylka	průměr	směrodatná odchylka
AUC	0,83	0,01	0,84	0,01
celková přesnost (%)	75,22	1,58	75,89	1,58
senzitivita (%)	72,24	1,21	72,44	1,16
specificita (%)	78,24	2,54	80,65	2,62
pozitivní prediktivní hodnota (%)	80,44	2,93	83,56	2,81
negativní prediktivní hodnota (%)	70,00	1,57	68,22	1,50

Příloha C: Obsah CD

- Klíčová slova.pdf
- Abstrakt (česky).pdf
- Abstrakt (anglicky).pdf
- Zadání.pdf
- Bakalářská práce.pdf
- Data
  - Tato složka obsahuje výchozí data a dílčí vstupy a výstupy jednotlivých metod.
- Zdrojové kódy
  - Tato složka obsahuje zdrojové kódy k použitým metodám.