TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

PRÍPADOVÉ USUDZOVANIE PRI DIAGNOSTIKE KARDIOVASKULÁRNYCH OCHORENÍ

Diplomová práca

2020 **Zuzana Tocimáková, Bc.**

TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

PRÍPADOVÉ USUDZOVANIE PRI DIAGNOSTIKE KARDIOVASKULÁRNYCH OCHORENÍ

Diplomová práca

Študijný program:	Hospodárska informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra kybernetiky a umelej inteligencie (KKUI)
Školiteľ:	prof. Ing. Ján Paralič, PhD.
Konzultant:	Ing. Ľudmila Pusztová

2020 Košice **Zuzana Tocimáková, Bc.**

Abstrakt v SJ

Hlavným cieľom predloženej diplomovej práce je navrhnúť, implementovať a experimentálne overiť systém na podporu rozhodovania kardiológov, ktorý je založený na princípoch prípadového usudzovania. Navrhovaný systém umožňuje používateľovi nájsť najpodobnejšie známe prípady k novému pacientovi a navrhuje najpravdepodobnejší potenciálny výsledok vyšetrenia zvaného koronárna angiografia. Poskytuje aj rôzne užitočné vizualizácie dát prostredníctvom tabuliek a grafov pre kardiológa, ktorý je zodpovedný za finálne rozhodnutie o odporučení nového pacienta na koronárnu angiografiu alebo nie. Implementovaný systém využíva algoritmy strojového učenia ako k-najbližších susedov, rozhodovacie stromy a zhlukovanie.

Kľúčové slova v SJ

systém na podporu rozhodovania, prípadové usudzovanie, k-najbližších susedov, kardiovaskulárne ochorenia, lekárska diagnostika

Abstrakt v AJ

The main purpose of this thesis is to design, implement and experimentally verify a decision support system for cardiologists, which is based on the case-based reasoning principles. The system enables its user to find the most similar historical cases to a new patient and suggests the most probable result of the potential coronary angiography examination. It also provides various useful visualizations via tables and graphs to the cardiologist, who is responsible for the final decision about recommending the coronary angiography for the new patient or not. Implemented system uses machine learning algorithms such as k-Nearest Neighbors, decision trees and clustering.

Kľúčové slova v AJ

Decision Support System, Case-based reasoning, k-Nearest Neighbors, Cardiovascular diseases, Medical diagnostics

57173

TECHNICKÁ UNIVERZITA V KOŠICIACH

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Katedra kybernetiky a umelej inteligencie

ZADANIE DIPLOMOVEJ PRÁCE

Študijný odbor: **Informatika**
Študijný program: **Hospodárska informatika**

Názov práce:

Prípadové usudzovanie pri diagnostike kardiovaskulárnych ochorení

Case-based reasoning in the diagnosis of cardiovascular diseases

Študent: **Bc. Zuzana Tocimáková**
Školiteľ: **prof. Ing. Ján Paralič, PhD.**
Školiace pracovisko: **Katedra kybernetiky a umelej inteligencie**
Konzultant práce: **Ing. Ľudmila Pusztová**
Pracovisko konzultanta: **Katedra kybernetiky a umelej inteligencie**

Pokyny na vypracovanie diplomovej práce:

1. Naštudovať a podať stručný prehľad o prípadovom usudzovaní (CBR - Case Based Reasoning), najmä v kontexte jeho využitia v medicíne.
2. Analyzovať možnosti využitia princípov prípadového usudzovania a metód strojového učenia na podporu rozhodovania kardiológov ohľadom kardiovaskulárneho rizika pacientov.
3. Navrhnúť, implementovať a vhodne overiť aplikáciu pre podporu rozhodovania kardiológov.
4. Realizovať sériu experimentov zameraných na vybrané kľúčové funkcie vytvoreného systému, vyhodnotiť a správne interpretovať získané výsledky.
5. Vypracovať dokumentáciu podľa pokynov katedry a vedúceho práce (hlavná časť práce v rozsahu 50-70 strán, prílohy - používateľská a systémová príručka).

Jazyk, v ktorom sa práca vypracuje: slovenský
Termín pre odovzdanie práce: 04.05.2020
Dátum zadania diplomovej práce: 31.10.2019

prof. Ing. Liberios Vokorokos, PhD.
dekan fakulty

Čestné vyhlásenie

Vyhlasujem, že som celú diplomovú prácu vypracovala samostatne s použitím uvedenej odbornej literatúry.

Košice, 01. mája 2020 .....

vlastnoručný podpis

Poďakovanie

Týmto by som sa chcela poďakovať vedúcemu mojej diplomovej práce, prof. Ing. Jánovi Paraličovi, PhD., za cenné rady a odbornú pomoc pri vypracovaní tejto práce. Moja vďaka taktiež patrí mojej konzultantke, Ing. Ľudmile Pusztovej, za všetky nápady a usmernenia počas písania tejto práce.

Obsah

Zoznam obrázkov 9

Zoznam tabuliek 10

Zoznam symbolov a skratiek 11

Úvod 12

1. Formulácia úlohy a cieľ práce 14

2. Systémy na podporu rozhodovania 15

2.1. Prípadové usudzovanie 16

2.1.1. 1. fáza – RETRIEVE 17

2.1.2. 2. fáza – REUSE/ADAPT 17

2.1.3. 3. fáza – REVISE 17

2.1.4. 4. fáza – RETAIN 18

2.2. Algoritmy strojového učenia 18

2.2.1. K-najbližších susedov (k-NN) 18

2.2.2. Rozhodovacie stromy 20

2.2.3. Zhlukovanie 20

3. Analýza súčasného stavu problematiky 22

3.1. Prípadové a pravidlové usudzovanie 22

3.2. Prípadové usudzovanie, rozhodovací strom a asociačné pravidlá 23

3.3. Prípadové usudzovanie, neurónové siete a iné 24

3.4. Prípadové usudzovanie a rozhodovací strom ID3 25

3.5. Zhrnutie analyzovaných štúdií 26

3.6. Kardiovaskulárne ochorenia 26

3.6.1. Príčiny kardiovaskulárnych ochorení 27

3.6.2. Ischemická choroba srdca 29

3.6.3. Diagnostika Ischemickej choroby srdca 29

3.6.4. Cievna mozgová príhoda 30

3.6.5. Infarkt myokardu 30

3.6.6. Aortálna stenóza 31
3.6.7. Koronarografické vyšetrenie 31
4. Návrh a implementácia vlastného DSS 33
4.1. Návrh DSS 33
4.1.1. Identifikácia používateľských požiadaviek 33
4.1.2. Diagram prípadov použitia 33
4.2. Implementácia navrhovaného DSS 37
4.2.1. Architektúra DSS 37
4.2.2. Funkcie navrhovaného DSS 38
5. Experimentálne vyhodnotenie 44
5.1. Pochopenie problému 44
5.2. Pochopenie dát 44
5.3. Príprava dát 49
5.4. Modelovanie 51
5.4.1. Experimentálna zmena počtu hodnôt cieľového atribútu 53
5.4.2. Experimentálny výber podmnožiny atribútov 57
5.5. Vyhodnotenie 60
Záver 63
Zoznam použitej literatúry 64
Prílohy 67

Zoznam obrázkov

Obr. 1 Hlavné komponenty DSS podľa [2] 15
Obr. 2 Diagram zobrazujúci CBR cyklus 18
Obr. 3 Porovanie Euklidovskej a Manhattanskej vzdialenosti 19
Obr. 4 Štatistika príčin úmrtí na Slovensku za rok 2017 27
Obr. 5 Diagram príadov použitia navrhovaného DSS 34
Obr. 6 Diagram zobrazujúci možnosti implementácie jednotlivých fáz CBR 35
Obr. 7 Architektúra navrhovaného systému 38
Obr. 8 Karta NOVÝ PRÍPAD 39
Obr. 9 Karta DÁTA 40
Obr. 10 Karta SKUPINY PACIENTOV 41
Obr. 11 Karta VIZUALIZÁCIA 42
Obr. 12 Karta PCA - ANALÝZA PRINCIPIÁLNYCH ZLOŽIEK 43
Obr. 13 Koronárne (srdcové) artérie 47
Obr. 14 Graf početností pacientov pre jednotlivé hodnoty atribútu *Nález* 48
Obr. 15 Graf absolútnych a relatívnych početností chýbajúcich hodnôt atribútov 49
Obr. 16 Graf závislosti výšky a váhy s extrémnymi hodnotami 51
Obr. 17 Graf početností pacientov pre jednotlivé hodnoty atribútu Nález: 3 cieľové triedy 54
Obr. 18 Porovnanie odhadovanej kvality metód výpočtu vzdialenosti pre 3 cieľové triedy 54
Obr. 19 Graf početností pacientov pre jednotlivé hodnoty atribútu Nález: 2 cieľové triedy 56
Obr. 20 Porovnanie kvality metód výpočtu vzdialenosti pre 2 cieľové triedy 56
Obr. 21 Porovnanie metrík odhadu kvality metódy Maxmin 57
Obr. 22 Výsledok Forward Stepwise Selection 58
Obr. 23 Presnosť metódy Maxmin pre 2 cieľové triedy na podmnožine podľa FSS 60
Obr. 24 Početnosti *k* susedov pri najvyšších priemerných presnostiach experimentov 62

Zoznam tabuliek

Tab. 1 Zhrnutie analýzy súčasného stavu 26
Tab. 2 Popis atribútov 44
Tab. 3 Konverzia kategorických atribútov na numerické 50
Tab. 4 Najvyššia priemerná presnosť implementovaných metód v percentách 52
Tab. 5 Odvodenie 3 cieľových tried 53
Tab. 6 Odvodenie 2 cieľových tried 55
Tab. 7 Priemerné presnosti metódy Maxmin pre 6 tried na podmnožine podľa FSS 58
Tab. 8 Priemerné presnosti metódy Maxmin pre 3 triedy na podmnožine podľa FSS 59
Tab. 9 Priemerné presnosti metódy Maxmin pre 2 triedy na podmnožine podľa FSS 60
Tab. 10 Porovnanie najvyšších priemerných presností pre všetky dôležité atribúty 61
Tab. 11 Porovnanie najvyšších priemerných presností s výberom atribútov podľa FSS 61

Zoznam symbolov a skratiek

k-NN	k-Nearest Neighbors
CRISP-DM	Cross-Industry Standard Process for Data Mining
DSS	Decision Support System
CDSS	Clinical decision support system
CBR	Case-based reasoning
RBR	Rule-based reasoning
BPN	Back propagation neural network
WRF	Weight ratio functionality
LDL	Low-density lipoproteins
HDL	High-density lipoproteins
ICHS	Ischemická choroba srdca
EKG	Elektrokardiografia
CMP	Cievna mozgová príhoda
KACH	Koronárna artériová choroba
VÚSCH	Východoslovenský ústav srdcových a cievnych chorôb
BMI	Body mass index
SCORE	Systematic coronary risk evaluation
PCA	Principal component analysis
FSS	Forward stepwise selection
AdjRsq	Adjusted R-squared

Úvod

Problematike kardiovaskulárnych ochorení je potrebné venovať zvýšenú pozornosť, pretože dlhodobo obsadzuje vedúcu pozíciu v štatistikách zachytávajúcich príčiny úmrtí, či už svetového, alebo slovenského obyvateľstva. Celosvetový trend podielu úmrtí na tieto ochorenia neustále rastie, a to z 25.59% v roku 1990 na 31.8% v roku 2017 [1]. Z tohto dôvodu sa táto téma stáva stále viac a viac aktuálnejšou.

Jedným zo spôsobov možného zefektívnenia postupov smerujúcich k diagnostike kardiovaskulárnych chorôb je využitie systému na podporu rozhodovania, ktorý by slúžil kardiológom ako podpora pri uskutočňovaní rozhodnutí o vážnosti stavu pacienta. V predloženej diplomovej práci prezentujeme a experimentálne vyhodnocujeme vlastnú implementáciu práve takéhoto typu systému.

Prvá kapitola podáva stručné sformulovanie a vysvetlenie zadania a cieľov práce vychádzajúcich zo zadávacieho listu. Inými slovami, aké konkrétne úlohy sú riešené v rámci tejto diplomovej práce.

Druhá kapitola obsahuje teoretické poznatky o informačných systémoch, konkrétne o systémoch na podporu rozhodovania, a taktiež ako do ich kontextu zapadá metodika prípadového usudzovania, ktorá je tu zároveň podrobne vysvetlená. V nasledujúcej podkapitole je zas v krátkosti objasnená podstata algoritmov strojového učenia (k-NN, rozhodovacie stromy a zhlukovanie), ktorých princípy boli využité aj pri navrhovaní praktického výstupu tejto práce.

Tretia kapitola je venovaná analýze súčasného stavu poznania v oblasti, ktorej sa venuje táto diplomová práca. Najprv sa zaoberá priblížením vybraných existujúcich riešení využívajúcich základne princípy prípadového usudzovania v rôznych medicínskych oblastiach záujmu. Následne pojednáva o kardiovaskulárnych ochoreniach, ich príčinách a existujúcich vyšetreniach slúžiacich na ich diagnostiku.

Štvrtá kapitola obsahuje návrh a následnú implementáciu vlastného systému na podporu rozhodovania kardiológov. Na začiatku sa venuje definovaniu používateľských požiadaviek na navrhovaný systém, a následne možným scenárom jeho použitia. V nasledujúcej časti sa nachádza vývojový diagram znázorňujúci návrh možností realizácie jednotlivých fáz metodiky prípadového usudzovania spolu s jeho podrobným popisom. V časti venujúcej sa implementácii navrhovaného systému je najprv vysvetlená a schematicky zobrazená jeho architektúra, po ktorej nasleduje podrobný popis funkcií systému nachádzajúcich sa v jednotlivých častiach aplikácie.

Obsah piatej kapitoly je rozčlenený podľa jednotlivých fáz metodológie CRISP-DM (Cross-Industry Standard Process for Data Mining). V prvej časti o pochopení riešeného problému sú stanovené ciele práce, za ňou nasleduje podrobný popis využívanej dátovej množiny a úpravy, ktoré v nej boli vykonané s cieľom zvýšiť jej kvalitu a použiteľnosť. V ďalšej fáze zvanej modelovanie bola realizovaná séria experimentov vedúca k odhadnutiu kvality klasifikačných metód implementovaných v systéme na podporu rozhodovania a ich možnému vylepšeniu. Posledná podkapitola je venovaná vyhodnoteniu a interpretácii výsledkov získaných v predošlej fáze modelovania.

1. Formulácia úlohy a cieľ práce

Táto kapitola je venovaná krátkemu objasneniu zadania, cieľov a pokynov pre vypracovanie tejto diplomovej práce.

Prvým bodom zadania práce je podať stručný prehľad o samotnej metodológii prípadového usudzovania (Case Based Reasoning - CBR), a to najmä v kontexte jej využitia v medicíne. Inými slovami je potrebné vhodne zaradiť CBR systémy do topológie informačných systémov a vysvetliť na akých princípoch sú založené. Taktiež je potrebné vymenovať jednotlivé fázy, z ktorých CBR pozostáva a všetky aspoň v krátkosti popísať.

Druhým bodom zadania je analýza možností využitia princípov CBR, ale aj metód strojového učenia na podporu rozhodovania kardiológov ohľadom kardiovaskulárneho rizika pacientov. To znamená, že je potrebné podrobne naštudovať a vysvetliť niekoľko už existujúcich riešení využívajúcich metodiku CBR, prípadne aj v kombinácii s rôznymi metódami strojového učenia, ktoré boli použité v podobnom medicínskom kontexte. Ide o analýzu súčasného stavu poznania v oblasti zamerania tejto diplomovej práce, ktorá je základom pre taký návrh a realizáciu vlastného systému na podporu rozhodovania pre kardiológov, ktorý prinesie pôvodný vlastný prínos k poznaniu v sledovanej problematike.

Treťou úlohou je navrhnúť, implementovať a vhodným spôsobom overiť aplikáciu na podporu rozhodovania kardiológov. Návrh aplikácie by mal byť založený na poznatkoch získaných v predchádzajúcej fáze analýzy súčasného stavu v kombinácii so zapojením vlastných nápadov, ktoré musia byť v súlade s používateľskými požiadavkami. Následne je potrebné realizovať samotnú implementáciu navrhovaného systému vo vhodnom programovacom jazyku. Pre overenie aplikácie je uskutočnené testovanie zamerané na odhad kvality predikčných metód, ktoré daná aplikácia využíva.

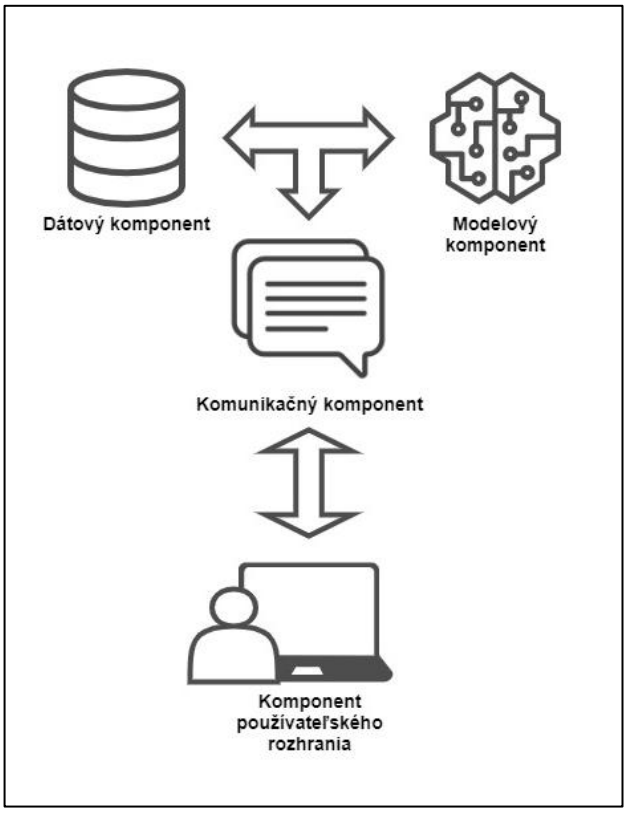
Štvrtou časťou zadania je realizácia série experimentov zameraných na vybrané kľúčové funkcie systému. Po predošlom otestovaní je vhodné nájsť spôsoby, prostredníctvom ktorých by potenciálne bolo možné dosiahnuť lepšie výsledky. Tieto identifikované metódy realizovať a vyhodnotiť získané výsledky.

Posledným bodom zadania je vypracovať dokumentáciu v rozsahu 50 až 70 strán, a zároveň používateľskú a systémovú príručku k implementovanému systému.

2. Systémy na podporu rozhodovania

Systémy na podporu rozhodovania (DSS) sú podľa [2] všeobecne definované ako interaktívne informačné systémy, ktoré pomáhajú ľuďom využívať údaje, dokumenty, znalosti, modely a ich vzájomnú komunikáciu pri riešení problémov a prijímaní rozhodnutí. Sú to iba pomocné systémy, ktorých účelom nie je nahradenie úsudku kvalifikovaných osôb v rozhodovacom procese.

Hlavné komponenty architektúry systémov pre podporu rozhodovania podľa [2] tvorí: používateľské rozhranie, modely a analytické nástroje, databáza a komunikačný komponent (Obr. 1).



Obr. 1 Hlavné komponenty DSS podľa [2]

Typy DSS podľa [3]:

- **Riadené údajmi** - založené na dátach a ich transformácii na informácie,
- **Riadené modelom** - využívajú simulačné a optimalizačné modely,
- **Riadené znalosťami** - využívajú znalostné technológie na splnenie konkrétnych požiadaviek procesu rozhodovania,
- **Riadené dokumentmi** - pomáhajú používateľom získať a spracovať neštruktúrované dokumenty alebo internetové stránky,

- **Riadené komunikáciou** - využívajú komunikačné technológie na podporu kolaborácie medzi používateľskými skupinami.

Systémy na podporu rozhodovania sú aplikované v mnohých oblastiach ľudského pôsobenia, a to od tradičného finančníctva, finančného prognózovania a riadenia, cez agronómiu, metalurgiu, údržbu, logistiku a transport, až po farmáciu a klinickú medicínu [3].

Klinické systémy na podporu rozhodovania (CDSS) sú podľa [4] definované ako informačné systémy určené na ovplyvnenie rozhodovania lekárov o jednotlivých pacientoch v čase uskutočnenia rozhodnutia s cieľom prevencie lekárskych chýb.

CDSS možno podľa [3] rozdeliť do 4 základných skupín, a to na systémy využívajúce:

- **Pravidlové usudzovanie (RBR)** - založené na princípe príčiny a následku (AK -> TAK),
- **Prípadové usudzovanie (CBR)** - založené na princípe analógie s už vyriešenými prípadmi,
- **Bayesovské siete (BBN)** - založené na teórii pravdepodobnosti,
- **Umelé neurónové siete (ANN)** - výpočtový model inšpirovaný štruktúrou a funkcionalitou biologických neurónových sietí.

V tejto diplomovej práci sme sa rozhodli preskúmať možnosti využitia jedného konkrétneho typu CDSS, ktorým je Prípadové usudzovanie.

2.1. Prípadové usudzovanie

Prípadové usudzovanie (CBR) je metodológia využívaná na riešenie problémov, a to na základe predošlých prípadov s už známym riešením. Táto metóda vychádza z princípov ľudského usudzovania, čoho dôkazom je snaha imitovať človeka pri riešení nových problémov, ktorý na to využíva svoje doterajšie skúsenosti a snaží sa ich prispôsobiť aktuálnej situácii [5].

Riešenie problému pomocou CBR zvyčajne pozostáva zo štyroch hlavných fáz [6]:

1. **RETRIEVE** - nájdenie najpodobnejších prípadov k novému problému,
2. **REUSE/ADAPT** - opätovné použitie znalostí z nájdených prípadov na vyriešenie problému,
3. **REVISE** - kontrola navrhovaného riešenia,
4. **RETAIN** - uloženie častí riešenia, ktoré môžu byť užitočné v budúcnosti.

2.1.1. 1. fáza – RETRIEVE

Počiatočná fáza získania najpodobnejších prípadov zo všetkých dostupných prípadov je často považovaná za najdôležitejšiu fázu CBR, pretože tvorí základ celkového fungovania tohto systému. Najčastejšie sa pri realizácii tejto fázy využíva technika k-najbližších susedov [6]. Vo všeobecnosti je možné túto fázu ešte rozdeliť na dva samostatné kroky [7].

Prvým krokom je vyhľadanie takých prípadov, na základe ktorých je potenciálne možné vytvoriť predikciu riešenia nového prípadu. To sa uskutočňuje pomocou atribútov nového problému, ktoré sa použijú ako index v databáze známych prípadov. Vtedy sú nájdené tie záznamy, ktoré sú označené rôznymi podmnožinami hľadaných vlastností [7].

Druhým krokom je selekcia najlepšej množiny prípadov, a to z množiny nájdenej v predchádzajúcom kroku. Cieľom je nájsť jedného alebo hneď niekoľko najrelevantnejších kandidátov. Pre lepšie výsledky porovnania by prípady mali byť porovnávané aj na abstraktnej úrovni reprezentácie, pri ktorej sa berú do úvahy aj odvodené vlastnosti. Napríklad pri posudzovaní podobnosti v medicínskej diagnostike je často prínosnejšie vziať do úvahy hypotetické poruchy alebo choroby, ktoré predstavujú odvodené vlastnosti prípadov, ako iba samotné symptómy vykazované pacientom [7].

2.1.2. 2. fáza – REUSE/ADAPT

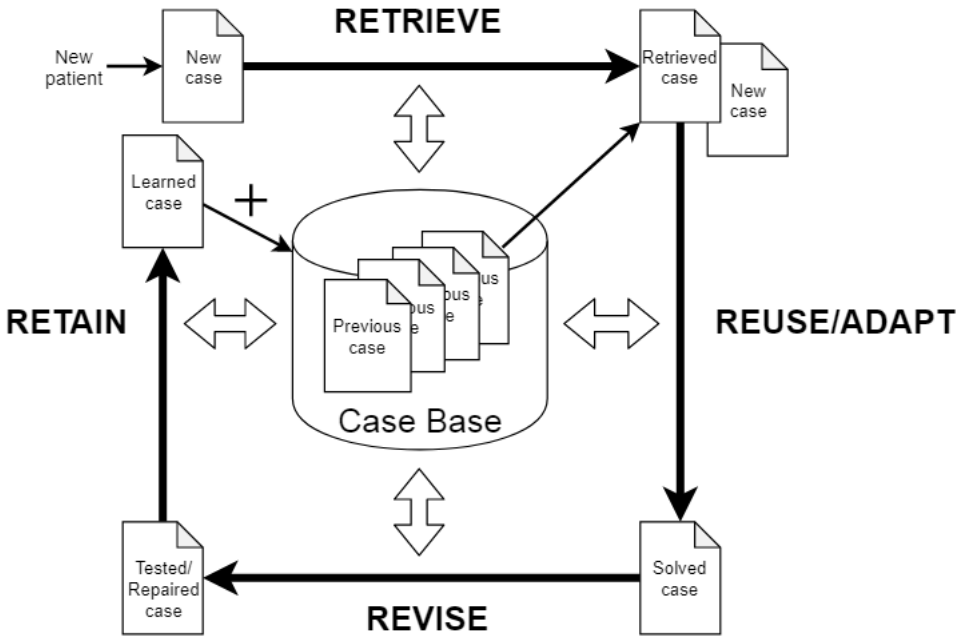
V tejto fáze nastáva formovanie približného riešenia nového problému. Buď je opätovne použité riešenie nájdeného najpodobnejšieho prípadu, alebo nastáva adaptácia za účelom lepšieho prispôsobenia sa aktuálnej situácii. V takomto prípade je potrebné zvážiť čo a akým spôsobom by malo byť adaptované. Jednou zo stratégií je vymazanie sekundárneho komponentu, ak nezohráva žiadnu dôležitú úlohu. Pri uvažovaní o tom, čo je potrebné adaptovať je taktiež potrebné zistiť nekonzistentnosti medzi starými riešeniami a novou situáciou [7].

2.1.3. 3. fáza – REVISE

V tretej fáze je vykonávaná kontrola správnosti navrhovaného riešenia. Validácia môže prebiehať v kontexte predchádzajúcich prípadov, môže byť založená na spätnej väzbe získanej od používateľa systému alebo na mentálnej či reálnej simulácii. Pomocou tejto spätnej väzby je systému umožnené vidieť následky navrhovaného riešenia a nadobudnúť tak nové znalosti. Táto fáza zahŕňa vysvetlenie a odôvodnenie rozdielov medzi prípadmi, projekciu výsledkov a porovnanie alternatívnych možností. Výsledkom posúdenia správnosti môže byť dodatočná adaptácia alebo oprava navrhovaného riešenia [7].

2.1.4. 4. fáza – RETAIN

Posledná fáza predstavuje uloženie nového vyriešeného prípadu do databázy známych prípadov, aby v budúcnosti mohol byť taktiež využitý na riešenie ďalších problémov. Vtedy je potrebné vybrať si spôsob, pod akým indexom, tzn. označením bude prípad uložený. Indexy by mali byť vybrané tak, aby bol prípad nájdený iba v takej situácii, v ktorej dokáže byť najviac nápomocný. Zároveň je nutné zabezpečiť, aby vždy bolo možné nájsť každý uložený prípad [7].



Obr. 2 Diagram zobrazujúci CBR cyklus

2.2. Algoritmy strojového učenia

Algoritmy strojového učenia môžeme rozdeliť do dvoch hlavných kategórií: kontrolované (supervised) a nekontrolované (unsupervised) učenie. Algoritmy kontrolovaného učenia, ako napr. rozhodovací strom, využívajú informáciu o cieľovom atribúte a podľa jeho hodnoty budujú model, ktorý sa využíva pri prediktívnom dolovaní v dátach. Platí, že ak je cieľový atribút kategorického (diskrétneho) charakteru, jedná sa o klasifikačný model. V opačnom prípade, keď je cieľový atribút numerického (spojitého) typu ide o predikciu. Algoritmy nekontrolovaného učenia akým je napr. zhlukovanie nevyužívajú údaje o cieľovom atribúte, ale snažia sa o vlastnú identifikáciu skupín podobných dátových záznamov [8].

2.2.1. K-najbližších susedov (k-NN)

Algoritmus k-najbližších susedov je univerzálny algoritmus kontrolovaného strojového učenia, ktorý je možné použiť na riešenie problémov vedúcich ku klasifikačným, ale aj regresným úlohám. Podstatou fungovania tohto algoritmu je predpoklad, že záznamy, ktoré sa podľa vybranej metriky

pre výpočet vzdialenosti nachádzajú blízko seba, si sú navzájom podobné [9]. Z tohto dôvodu patrí k najdôležitejším parametrom algoritmu k-NN práve spôsob výpočtu vzdialenosti medzi dátovými bodmi v pomyselnom viacrozmernom priestore.

2.2.1.1. Euklidovská vzdialenosť

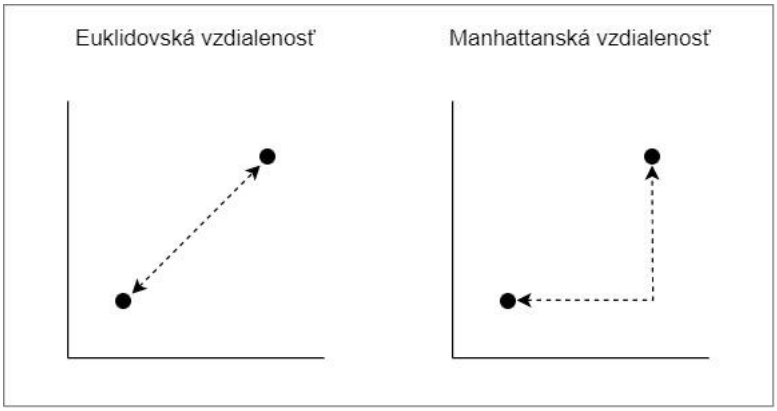
Euklidovská vzdialenosť medzi dátovými bodmi $\bar{X} = (x_1, x_2, \ldots, x_n)$ a $\bar{Y} = (y_1, y_2, \ldots, y_n)$ v n-rozmernom priestore je definovaná vzťahom [10]:

$$Eucl(\bar{X}, \bar{Y}) = \sqrt[2]{\sum_{i=1}^{n} |x_i - y_i|^2}$$

2.2.1.2. Manhattanská vzdialenosť

Manhattanská vzdialenosť medzi dátovými bodmi $\bar{X} = (x_1, x_2, \ldots, x_n)$ a $\bar{Y} = (y_1, y_2, \ldots, y_n)$ v n-rozmernom priestore je definovaná vzťahom [10]:

$$Manh(\bar{X}, \bar{Y}) = \sum_{i=1}^{n} |x_i - y_i|$$



Obr. 3 Porovanie Euklidovskej a Manhattanskej vzdialenosti

Ako môžeme vidieť na Obr. 3, Euklidovská vzdialenosť zodpovedá priamej vzdialenosti medzi bodmi v priestore, zatiaľ čo Manhattanská vzdialenosť kopíruje pravouhlú mriežku (ako vzdialenosť v Manhattane, kde sú ulice na seba kolmé). Obe vyššie spomínané vzdialenosti sú špeciálnymi prípadmi tzv. L_p normy, kde Euklidovská vzdialenosť zodpovedá p = 2 a Manhattanská p = 1. L_p norma medzi dátovými bodmi $\bar{X} = (x_1, x_2, \ldots, x_n)$ a $\bar{Y} = (y_1, y_2, \ldots, y_n)$ v n-rozmernom priestore je definovaná vzťahom [10]:

$$L_p(\bar{X}, \bar{Y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Výsledná hodnota akejkoľvek vzdialenosti vypočítanej pomocou L_p normy berie do úvahy len tie dva záznamy, ktorých vzájomnú vzdialenosť chceme zistiť. Ak by sme chceli pri výpočte vzdialenosti dvoch dátových bodov zohľadniť aj globálne vlastnosti zvyšku dátovej množiny, konkrétne ich celkovú distribúciu, môžeme na to využiť Mahalanobisovu vzdialenosť.

2.2.1.3. Mahalanobisova vzdialenosť

Mahalanobisova vzdialenosť medzi dátovými bodmi $\bar{X} = (x_1, x_2, \ldots, x_n)$ a $\bar{Y} = (y_1, y_2, \ldots, y_n)$ v n-rozmernom priestore, kde C je kovariančná matica dát o rozmeroch n x n, je definovaná vzťahom [10]:

$$Maha(\overline{\mathrm{X}}, \overline{\mathrm{Y}}) = \sqrt[2]{(\bar{X} - \bar{Y})C^{-1}(\bar{X} - \bar{Y})^T}$$

Prvky kovariančnej matice C vyjadrujú kovarianciu medzi jednotlivými atribútmi. Ak je prvok na pozícii *i,j* rovný 0, znamená to, že atribúty *i* a *j* sú nekorelované.

2.2.2. Rozhodovacie stromy

Rozhodovacie stromy [8] predstavujú skupinu klasifikačných algoritmov kontrolovaného učenia. Sú veľmi populárnou formou reprezentácie klasifikátorov, a to hlavne pre jednoduchú interpretáciu ich modelov. Rozhodovací strom sa skladá z nasledovných častí:

- **Medziľahlý uzol** – jeden alebo viac atribútov,
- **Hrana** – test na atribút nachádzajúci sa v nadradenom uzle,
- **Listový uzol** – výsledná cieľová trieda.

Model rozhodovacieho stromu je konštruovaný na základe dátovej množiny slúžiacej na jeho natrénovanie. Následne pri určovaní cieľovej triedy nového prípadu s neznámou hodnotou cieľového atribútu prechádza daný záznam štruktúrou tohto stromu od počiatočného koreňového uzla po jednotlivých hranách podľa podmienok, ktoré spĺňa až pokiaľ neskončí v jednom z listových uzlov nachádzajúcich sa na úplnom konci stromu. Podľa zaradenia príkladu do konkrétneho listového uzla je novému prípadu priradená zodpovedajúca cieľová trieda.

2.2.3. Zhlukovanie

Zhlukovanie je jedna z metód nekontrolovaného učenia, ktorej cieľom je rozdeliť dátové záznamy do skupín tak, aby sa najviac podobné objekty nachádzali v rovnakej skupine (zhluku) a aby si objekty patriace do rôznych zhlukov boli čo najnepodobnejšie. Pre účely výpočtu miery podobnosti jednotlivých príkladov v databáze je možné použiť ľubovoľnú metriku vzdialenosti ako napr. Euklidovská vzdialenosť (kapitola 2.2.1.1) [8].

2.2.3.1. k-medoids

Algoritmus k-medois patrí k do skupiny algoritmov pomocou ktorých je možné realizovať zhlukovanie okolo vybraných centier zhlukov. Typickou vlastnosťou tohto algoritmu je, že za centrum zhluku musí byť vybraný len existujúci dátový bod pochádzajúci z príslušnej dátovej množiny, ktorý je označovaný aj ako „medoid“. Priebeh tohto algoritmu je nasledovný [11]:

1. Náhodná inicializácia *k* objektov, ktoré sa stanú medoidmi (reálne dátové body).
2. Výpočet vzdialeností medzi medoidmi a dátovými objektmi.
3. Priradenie každého dátového bodu k najbližšiemu medoidu na základe vybranej metriky vzdialenosti (napr. Euklidovskej vzdialenosti).
4. Pre každý zhluk je sledované, či nejaký z jeho bodov dokáže znížiť priemernú vzdialenosť v rámci daného zhluku v prípade ak by bol vybraný za medoid.
    - Ak áno:
        - tak je vyberaný bod, ktorý najviac znižuje vzdialenostný koeficient a ten sa stane novým medoidom,
        - ak sa aspoň jeden medoid zmenil, nasleduje opäť krok 3. Ak nie, tak algoritmus končí.
    - Ak nie: algoritmus končí.

3. Analýza súčasného stavu problematiky

Keďže počiatky prípadového usudzovania siahajú až do osemdesiatych rokov minulého storočia, existuje mnoho výskumov využívajúcich práve túto metódu, či už v medicínskej, alebo inej oblasti. Taktiež je tu snaha zo strany výskumníkov o vylepšenie jej výsledkov, a to pomocou kombinácie s ďalšími vhodnými metódami dolovania v dátach.

3.1. Prípadové a pravidlové usudzovanie

Príkladom je pilotná štúdia z roku 2015 [12], v ktorej Saraiva et al. prezentujú využitie hybridného prístupu prípadového (CBR) a pravidlového usudzovania (RBR) na podporu včasnej diagnostiky rôznych druhov rakoviny. Ich cieľom bolo navrhnúť systém na podporu rozhodovania, ktorý by asistoval menej skúseným lekárom pri určení diagnózy, ale zároveň by mohol slúžiť aj ako poradný nástroj pre expertov.

Metodológia RBR reprezentuje znalosti vo forme pravidiel, ktoré sú generované v tvare: „AK <podmienky> => TAK <záver>“. Tieto pravidlá predstavujú nájdené vzory, ktoré by sa mali zhodovať so vzormi v dátach. Najvhodnejšie použitie metódy RBR je počas riešenia jednoduchých problémov, ktoré nevyžadujú veľa pravidiel. V tomto prípade boli pravidlá sformulované na základe vybraných knižných a internetových medicínskych poznatkov.

Ako vstupné dáta výskumníkom slúžili osobné informácie o 48 skutočných onkologických pacientoch z Brazílie, ktoré zahŕňali ich príznaky, pozorované symptómy a taktiež ich diagnózy. Celkovo mali k dispozícii 26 atribútov ako napríklad pohlavie, vek, farba pleti, informácia o fajčení, požívaní alkoholu a o výskyte rakoviny u rodinných príslušníkov. Zaujímavosťou je, že atribút určujúci vek pacienta bol diskretizovaný na 3 intervaly s hodnotami veku do 60 rokov, 61 až 70 rokov a od 71 rokov.

Pre počiatočnú fázu CBR, zvanú RETRIEVE, je v tomto prípade dôležitý tzv. index atribútu. Index je vytvorený pre atribúty vystupujúce v pravidlách, pričom podľa svojej dôležitosti majú stanovenú váhu od 1 do 10, ktorá je zadaná používateľom systému, teda lekárom. Na výpočet podobnosti jednotlivých prípadov využili metódu najbližšieho suseda, ktorá by mala mať v kombinácii s váhami atribútov väčšiu šancu na nájdenie najlepšieho riešenia. Keďže k dispozícii mali iba malú databázu prípadov, výpočet podobnosti uskutočnili pre každý z prípadov. Výstup modelu reprezentoval pravdepodobnosť, že daný pacient má konkrétny typ rakoviny.

Na odhadnutie kvality navrhovaného modelu použili v prípade [12] trojnásobnú krížovú validáciu. Najvyššia dosiahnutá presnosť bola na úrovni 81.25%. V porovnaní s použitím iba

samotnej metódy CBR, validácia hybridnej metódy využívajúcej RBR ukázala zvýšenie presnosti diagnostiky rakoviny v priemere o 22,92%.

3.2. Prípadové usudzovanie, rozhodovací strom a asociačné pravidlá

V taiwanskej štúdii z roku 2006 [13], ktorá je zameraná na prognózu a diagnostiku chronických chorôb, je navrhovaný model „CDPD“ spájajúci metódy dolovania v dátach a prípadového usudzovania. Konkrétne ešte pred aplikovaním metódy CBR sú dáta čistené a predspracované, aby došlo k navýšeniu ich kvality. Následne sú vybrané najpodstatnejšie atribúty a na nich sú trénované algoritmy generujúce rozhodovacie stromy, s cieľom získať platné pravidlá pre výskyt chronických chorôb.

Na dostupné dáta je taktiež aplikovaný aj algoritmus na hľadanie asociačných pravidiel, ktorý slúži na nájdenie implicitných vzťahov v dátach. Všetky tieto pravidlá po ich schválení doktorom používajúcim systém „CDPD“, sú uložené do bázy pravidiel a slúžia na predikciu pravdepodobnosti výskytu konkrétnej chronickej choroby u daného pacienta. Až potom je aplikovaná metóda CBR, ktorej mechanizmus nájde najpodobnejší prípad k aktuálnemu pacientovi, a na základe neho je stanovená následná liečba.

V tejto štúdii bolo po vyčistení dát od anomálií k dispozícii 15 751 skutočných záznamov, pričom každý z nich pozostával z 28 atribútov. Boli to napríklad: HDLC (high-density lipoprotein cholesterol), LDLC (low-density lipoprotein cholesterol), BMI (index telesnej hmotnosti), TG (triglyceridy), pohlavie, vek a váha pacienta, výskyt chronických chorôb u rodinných príslušníkov, ale aj diagnóza a predpísaná liečba.

Keďže navrhovaný systém sa zameriava na viacero chorôb, konkrétne na mŕtvicu, kardiomyopatiu, zvýšený krvný tlak a cukrovku, na nájdenie najpodobnejšieho prípadu vo fáze RETRIEVE sú najprv použité pravidlá nachádzajúce sa v báze pravidiel, ktoré slúžia na zatriedenie nového prípadu do skupiny s podobnými hodnotami atribútov, teda k prípadom týkajúcim sa najpravdepodobnejšej chronickej choroby. Následne je použitá metóda WRF (weight ratio functionality) odvodená od princípu najbližšieho suseda. Táto metóda, využívajúca pomer hodnôt atribútov nového a už známeho prípadu, je obohatená o váhy zohľadňujúce dôležitosť atribútov, pričom najmenej dôležitým vlastnostiam je priradená váha 0.

Vo fáze REUSE/ADAPT použili interpoláciu, priemerné hodnoty atribútu a adaptačné pravidlá. Interpolácia pomocou lineárnej funkcie bola ováhovaná pomerom hodnoty atribútov a celkového rozsahu hodnôt. V prípade, že je potrebné adaptovať parametre násobené váhami, systém

využíva upravenú Shepardovu metódu interpolácie, založenú na relatívnych vzdialenostiach medzi vektormi, ktorá zohľadňuje aj rôzne rozsahy hodnôt.

Ďalší spôsob adaptácie bol realizovaný pomocou priemerných hodnôt. Konkrétne, ak parameter s hodnotou h_1 má byť upravený na hodnotu h_2, systém nájde všetky prípady obsahujúce prvú alebo druhú hodnotu a vypočíta priemerné hodnoty výsledných parametrov v rámci týchto dvoch skupín. Pomer druhého priemeru ku prvému následne vytvorí adaptačné pravidlá. Prípadom nachádzajúcim sa v týchto skupinách sú navyše priradené váhy podľa ich použiteľnosti. Priemerná hodnota použiteľnosti oboch skupín zároveň slúži ako ukazovateľ kvality daného adaptačného pravidla.

Ako posledný spôsob adaptácie využili metódu adaptačných pravidiel. Tieto popisujú zmeny hodnôt, ktoré sú vyjadrené faktormi, rozdielmi alebo kvalitatívnymi zmenami. Faktory sú kombinované násobením, na rozdiel od rozdielov, ktoré sú sčítavané. Ak spojíme oba tieto typy, výsledkom bude kvalitatívna zmena. Pri určovaní liečby je však vždy nutná kontrola správneho použitia adaptačných pravidiel lekárom, čo vlastne predstavuje fázu REVISE. Na konci, v rámci RETAIN fázy, je prípad spolu s riešením uložený do databázy už známych prípadov, a to za účelom zdokonalenia učiaceho sa mechanizmu systému „CDPD“.

3.3. Prípadové usudzovanie, neurónové siete a iné

V štúdii zameranej na návrh podporného systému pre včasnú diagnostiku a liečbu ochorenia pečene [14] bola tiež použitá metóda CBR. Okrem jej samostatného použitia bola vyskúšaná aj jej integrácia s rôznymi technikami dolovania v dátach, konkrétne s metódami: BPN (back-propagation neural network), CART (klasifikačný a regresný strom), logistickou regresiou a diskriminačnou analýzou.

Dáta boli získané od 166 reálnych pacientov medicínskeho strediska v Taiwane, pričom 91 z nich malo diagnostikovanú chorobu pečene. Dátové atribúty zahŕňali napríklad: pohlavie, vek, váhu, krvnú skupinu, stav pečene, a tiež informácie o konzumácii alkoholu, fajčení a prípadnom výskyte žltačky či nádoru.

S cieľom minimalizácie počtu falošne negatívnych prípadov pri diagnostike choroby, na prípady zaradené klasifikátorom do triedy „bez výskytu choroby“ bol použitý hybridný CBR model na potvrdenie tejto diagnózy. Konkrétne bol každý takýto prípad porovnaný s desiatimi najpodobnejšími prípadmi v databáze so zdravými pacientmi a zvlášť aj s najpodobnejšími chorými pacientmi. Ak súčet podobností aktuálneho prípadu so zdravými pacientmi bol väčší ako súčet

podobností s chorými, diagnóza ostala rovnaká ako výsledok určený klasifikátorom. V opačnom prípade bol tento prípad nanovo klasifikovaný do triedy „s výskytom choroby“.

V tomto prípade bola na ohodnotenie kvality modelov použitá 10-násobná krížová validácia, a to za použitia Bernoulliho metódy vzorkovania so zachovaním pomeru chorých a zdravých v každej vzorke. Zo všetkých vytvorených modelov dosiahol najlepšie výsledky model kombinujúci metódy BPN a CBR, a to s presnosťou až na úrovni 95%, senzitivitou 98%, špecificitou 94% a AUC (Area Under the Receiver Operating Characteristics) 96%.

3.4. Prípadové usudzovanie a rozhodovací strom ID3

Metóda CBR môže podľa [15] slúžiť aj ako základ prototypu expertného systému na podporu diagnostiky kardiovaskulárnych chorôb. Navrhovaný systém bol zameraný na ochorenia ako mitrálna stenóza, ľavostranné srdcové zlyhanie, stabilná angína pectoris a esenciálna hypertenzia.

Dáta, s ktorými autori pracovali, obsahovali 110 záznamov o egyptských pacientoch. Každý z nich pozostával z 207 atribútov, ktoré boli buď demografického alebo klinického charakteru. Pomocou odborných znalostí experta a štatistického vyhodnotenia korpusu známych prípadov zisťovali, ktoré z atribútov sú najlepšími prediktormi, a tým bola následne priradená vyššia váha.

Hľadanie najpodobnejších známych prípadov k novému prípadu, teda RETRIEVE fázu CBR cyklu, uskutočňovali dvomi spôsobmi. Najprv prostredníctvom metódy k-najbližších susedov, kde boli do úvahy brané iba prípady s podobnosťou väčšou ako stanovený prah 0.5. *K*, teda počet susedov, bolo vždy nepárne číslo, aby pri hlasovaní o diagnóze nedochádzalo k remíze v počte hlasov za rôzne diagnózy.

Druhý spôsob využíval indukčný algoritmus generujúci rozhodovací strom ID3, ktorý umožňuje jednotlivým prípadom priradiť indexy a pomocou nich aj rýchle nájdenie podobných prípadov. Algoritmus ID3 však nedokáže spracovať prázdne hodnoty, čo sa prejavilo aj na presnosti diagnostiky iba na úrovni 53.8%. Metóda k-najbližších susedov s hlasovaním o najpravdepodobnejšej triede bola značne úspešnejšia, keďže dosiahnutá presnosť bola pri 13 testovacích príkladoch až na úrovni 100%.

3.5. Zhrnutie analyzovaných štúdií

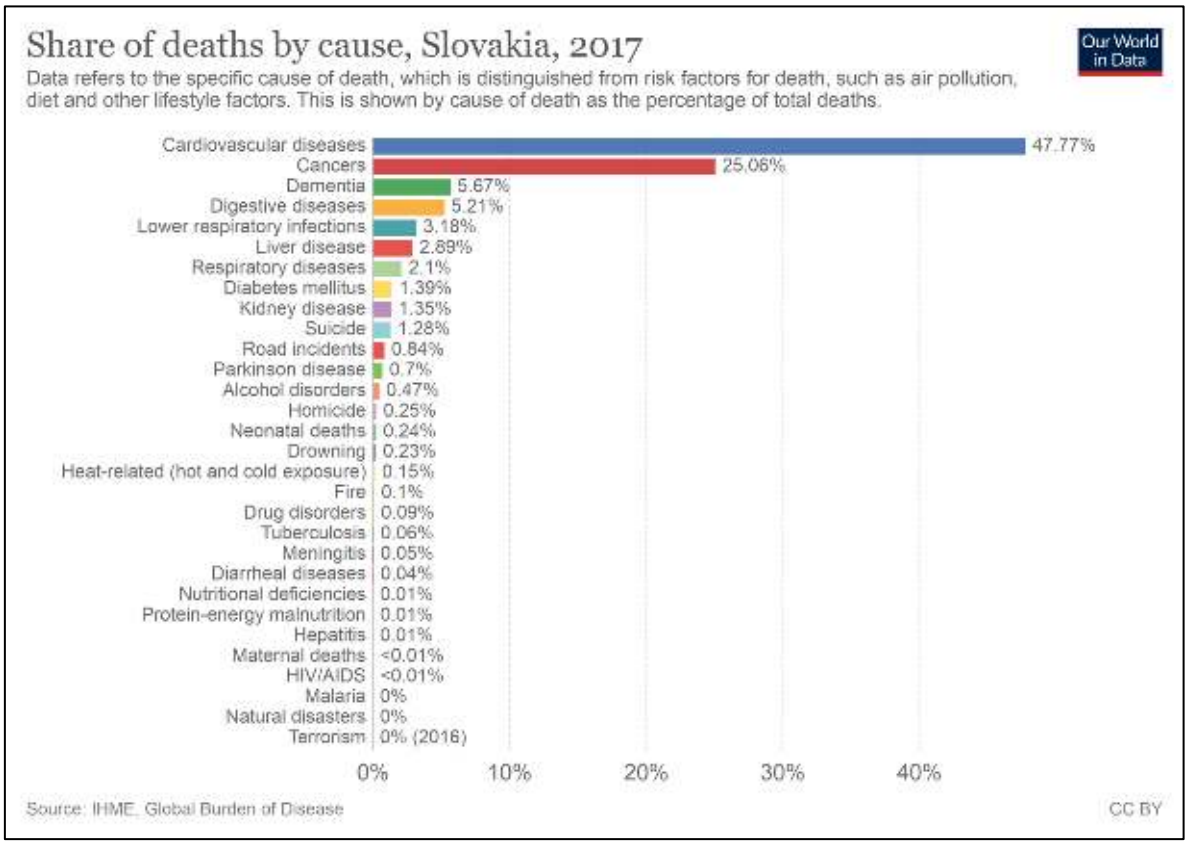
Tab. 1 Zhrnutie analýzy súčasného stavu

Autori	Aplikačná oblasť	Realizácia jednotlivých CBR fáz				
		RETRIEVE/lokálna funkcia podobnosti	REUSE/ADAPT	REVISE	RETAIN	Presnosť
SARAIVA, R. M. et al.	rôzne druhy rakoviny	váha z pravidiel (RBR) * ak $x_i = y_i$ tak: 1 inak: 0 kde: x_i – hodnota atribútu i nového prípadu, y_i – známeho prípadu	–	–	–	81.25%
HUANG, M.-J. et al.	chronické choroby	pravidlá (asociačné, rozhodovací strom) + WRF: ak $x_i < y_i$ tak: $\frac{x_i}{y_i}$ inak: $\frac{y_i}{x_i}$ kde: x_i – h odnota atribútu i nového prípadu, y_i – známeho prípadu	interpolácia, priemerné hodnoty atribútu a adaptačné pravidlá	lekár	✓	–
CHUANG, CH.-L.	choroba pečene	Euklidova vzdialenosť	BPN	10 najbližších susedov	–	95%
SALEM, A.-B. M., ROUSHDY, M., HODHOD, R. A.	kardiovaskulárne choroby	$1 - \frac{\lvert f_i^l - f_i^R \rvert}{\lvert f_{max} - f_{min} \rvert}$ kde: f_i – hodnota atribútu i (ďalej odkazovaná ako Maxmin)	*k*-najbližších susedov kde: *k* – nepárne	kardiológ	–	100%
		ID3	ID3	kardiológ	–	53.8%

3.6. Kardiovaskulárne ochorenia

Keďže oblasťou záujmu tejto diplomovej práce je diagnostika kardiovaskulárnych ochorení, nasledujúca podkapitola je venovaná stručnému objasneniu problematiky srdcovocievnych chorôb, ich príčin, prejavov či vyšetrení na ich odhalenie.

Téme kardiovaskulárnych ochorení je potrebné venovať zvýšenú pozornosť, pretože táto skupina ochorení je najčastejšie sa vyskytujúcou príčinou úmrtí ľudí, či už vo svete, alebo konkrétne v Slovenskej republike, kde každoročne zapríčiňuje skoro polovicu úmrtí (47.7% v 2017) a výrazne vedie nad všetkými ostatnými ochoreniami v tejto štatistike (Obr. 4).



Obr. 4 Štatistika príčin úmrtí na Slovensku za rok 2017[1]

3.6.1. Príčiny kardiovaskulárnych ochorení

Medzi hlavné známe príčiny vzniku srdcovocievnych ochorení, ktoré však nie je možné nijako ovplyvniť patrí [16]:

- **Genetická predispozícia** – výskyt vysokého krvného tlaku u blízkych príbuzných pacienta je predpokladom na prepuknutie choroby aj uňho samotného. Je známe, že pravdepodobnosť zdedenia vysokého krvného tlaku je približne 50%.
- **Vek** – kardiovaskulárna choroba sa u pacienta môže prejaviť v akomkoľvek veku vplyvom rôznych rizikových faktorov. Platí však, že s vyšším vekom pacienta narastá aj riziko vzniku srdcovocievnej choroby.

[1] Zdroj: Our World in Data. [Online].
Dostupné na internete: <https://ourworldindata.org/grapher/share-of-deaths-by-cause?country=SVK>

- **Vysoký krvný tlak** – nazývaný aj hypertenzia, označuje hodnoty systolického tlaku prevyšujúce 140 mm/Hg a u diastolického tlaku hodnoty vyššie ako 90 mm/Hg. Vplyvom hypertenzie sa zvyšuje námaha srdca a ciev. Keďže je srdce nútené pracovať v ťažšom režime ako za normálnych podmienok, steny srdca môžu zhrubnúť a srdce sa zväčší, čo spôsobuje komplikácie. U ľudí s vysokým krvným tlakom taktiež nastáva oveľa skôr proces tvrdnutia ciev, ktoré postupne strácajú svoju elasticitu. Riziko výskytu cievnej mozgovej príhody je tak isto u nich vyššie.

- **Pohlavie** – jednotlivé rizikové faktory podnecujúce srdcovocievne ochorenia majú rôznu dôležitosť pri vplyve na mužov a na ženy. U mužov vplýva na vznik kardiovaskulárnych chorôb hlavne celkový cholesterol a LDL (lipoproteíny s nízkou hustotou) cholesterol. Naopak, u žien je dôležité sledovať najmä vysoký krvný tlak, nízku hladinu HDL (lipoproteíny s vysokou hustotou) cholesterolu, cukrovku a zvýšenú hodnotu triglyceridov v krvi. U žien sa zmeny hladín cholesterolu zvyknú prejavovať po dosiahnutí päťdesiatich rokov, u mužov dokonca skôr.

Zároveň existujú aj ovplyvniteľné rizikové faktory, ktoré je možné korigovať napríklad vhodnou liečbou. Medzi tieto rizikové faktory môžeme zaradiť [16]:

- **Cukrovka (diabetes mellitus) 2. typu** – je nevyliečiteľná chronická metabolická choroba. Je celosvetovo rozšírená a počet prípadov neustále rastie. Pri cukrovke nastáva porucha vylučovania inzulínu, teda hormónu nevyhnutného pre umožnenie spracovania cukru v krvi. Bez inzulínu bunky nie sú schopné prijímať glukózu a tá zostáva v krvi. Z tohto dôvodu majú diabetici stále vysokú hladinu cukru v krvi, ak sú neliečení.

- **Cholesterol** – je tuková látka nachádzajúca sa v krvi. Bez neho nedokáže ľudské telo fungovať, pretože je potrebný pre normálny chod buniek. Keďže nie je rozpustný v krvi, prenos medzi bunkami zabezpečujú bielkovinové nosiče zvané lipoproteíny. Za zdravú hladinu cholesterolu sú považované hodnoty menšie ako 5 mmol/l. Ak má pacient dlhodobo zvýšenú jeho hladinu, zvyšuje sa uňho pravdepodobnosť výskytu infarktu myokardu alebo cievnej mozgovej príhody. Vtedy je nutné upraviť životosprávu a pravidelne vykonávať pohybové aktivity, respektíve nasadiť lieky. Najhlavnejšie druhy lipoproteínov sú [17]:

    - *LDL* – jeho úlohou je transport cholesterolu do telových buniek. Všetky bunky ho potrebujú, ale ak ho v krvi koluje príliš veľa, bunky ho nedokážu zužitkovať a hrozí, že sa tuk začne usadzovať na stenách ciev. Tento proces sa nazýva ateroskleróza. Cholesterol, ktorý je naviazaný na LDL je taktiež známy ako „zlý“ cholesterol.

- *HDL* – jeho úlohou je odvádzanie nadbytočného cholesterolu z krvi do pečene. Odtiaľ sa dostáva do žlče a von z ľudského tela. HDL dokáže odstraňovať aj cholesterol, ktorý je usadený na stenách ciev. Z tohto dôvodu je známy aj ako „dobrý“ cholesterol.

- **Fajčenie** – negatívne vplýva na kardiovaskulárny systém, keďže tabakový dym má vysoký obsah škodlivých látok poškodzujúcich srdce a cievy. Nikotín spôsobuje zrýchlenie srdcového rytmu a dočasné zvýšenie krvného tlaku. Môže zapríčiniť zúženie alebo dokonca úplné uzavretie ciev či tepien. Dýchanie cigaretového dymu spôsobuje zníženie množstva kyslíka, ktoré krv transportuje do srdca a celého tela. Z chorôb spôsobených fajčením sa srdcovocievne choroby vyskytujú oveľa častejšie ako rakovina pľúc, pričom na poškodenie ciev stačí aj jedna cigareta za deň [16].

3.6.2. Ischemická choroba srdca

Podstatou ischemickej choroby srdca (ICHS) je zúženie vencovitých tepien, tzn. ciev, ktorých úlohou je zásobovanie srdcového svalu krvou. Krv obsahuje kyslík a živiny, ktoré sú nevyhnutné pre správnu činnosť srdca. Zúženie týchto tepien vyvoláva nerovnováhu medzi množstvom kyslíka, ktoré sa reálne dostane do srdca a množstvom, ktoré srdce v skutočnosti potrebuje [18].

Permanentný nedostatok krvi spôsobuje postupné poškodzovanie srdca. Pacient začína mať ťažkosti, ktoré sú spočiatku iba mierneho charakteru a prejavujú sa hlavne počas vykonávania fyzickej aktivity. Jedná sa konkrétne o bolesť pociťovanú za hrudnou kosťou a predčasné zadýchavanie sa. V pokročilejšom štádiu sa tieto prejavy objavujú aj v pokoji [16].

Okrem všeobecných rizikových faktorov pre srdcovocievne choroby, špecifickými faktormi pre vznik ICHS sú: zvýšená zrážanlivosť krvi, sedavý životný štýl či výbušný typ osobnosti. Riziko vzniku ICHS zvyšuje taktiež nevhodná strava s vysokým obsahom nasýtených tukov a zložitých cukrov [18].

3.6.3. Diagnostika Ischemickej choroby srdca

Medzi neinvazívne metódy vyšetrenia výskytu ischémie srdca, teda bez zásahu do pacientovho tela, patrí [18]:

- **Odber krvi** – zistené odchýlky od normy upozornia lekára na možnú prítomnosť choroby.
- **Elektrokardiografia (EKG)** – sníma a graficky zaznamenáva elektrické potenciály generované v srdcovom svale pomocou elektród umiestnených na telo pacienta. Ak je srdce poškodené v malej miere, výsledky EKG sa môžu javiť normálne. Z tohto dôvodu, ak sa u pacienta vyskytuje podozrenie na ischémiu, je potrebné vykonať EKG vyšetrenie viackrát s niekoľkohodinovým časovým odstupom (zvyčajne šiestich hodín). V takomto

prípade sú sledované zmeny, ktoré sa udiali za tento čas. Ide hlavne o zmeny EKG obrazu a zvýšenie koncentrácie určitých látok v krvi, ktoré značí poškodenie srdca.

- **Ergometria (záťažová EKG)** – u mnohých pacientov trpiacich ICHS sú hodnoty EKG namerané v pokoji normálne. Preto je vhodné realizovať EKG vyšetrenie aj počas vykonávania fyzickej aktivity. Najčastejšie ide o kráčanie na bežiacom páse alebo bicyklovanie. Keďže pri záťaži sa zvyšujú nároky srdca na kyslík, ak je prietok vencových tepien obmedzený, vzniká nepomer medzi množstvom spotrebovaného kyslíka a množstvom dostupného kyslíka, a teda je vyššia šanca prejavu ICHS na EKG zázname.
- **Echokardiografia** – ide o vyšetrenie srdcového svalu prostredníctvom ultrazvukovej sondy. Záznamy sú získavané na základe echa, čiže odrazu ultrazvukových vĺn od tkaniva. Vďaka moderným ultrazvukovým prístrojom je možné nadobudnúť komplexné informácie o anatómii srdcového svalu, a taktiež o štruktúre a funkcii jeho tkanív. Ischémiu srdca najčastejšie naznačuje porucha pohybu stien myokardu.
- **Záťažová echokardiografia** – pre zvýšenie šance odhalenia chronickej (postupnej) ischémie srdca je pacientovi podaná do žily látka zvyšujúca frekvenciu srdca. To spôsobí zvýšenie nárokov srdca na potrebné množstvo kyslíka. Nedostatok krvi v niektorej z častí srdcového svalu zobrazí poruchu pohybu srdca.
- **MR (Magnetická rezonancia) a CT (Výpočtová tomografia) Angiografia srdca** – v rámci týchto vyšetrení je možné detailné znázornenie štruktúry myokardu a ciev pomocou kontrastnej zobrazovacej látky podanej do žily [18].

3.6.4. Cievna mozgová príhoda

Pri zúžení alebo dokonca zablokovaní ciev zabezpečujúcich prívod krvi do mozgu môže dôjsť k cievnej mozgovej príhode (CMP), známej aj ako mozgová mŕtvica alebo porážka. Jedná sa o veľmi nebezpečný stav, kedy je v ohrození život pacienta z dôvodu odumierania mozgu, ktoré začína už po niekoľkých minútach. Medzi príznaky signalizujúce CMP patrí ovisnutie a strata citlivosti v jednej časti tváre, na čo sa dá prísť hlavne pri pokuse o úsmev. Ďalším prejavom môže byť slabosť jednej z horných končatín, kedy je pre pacienta nemožné zdvihnúť naraz obe ruky. Často sa pri CMP vyskytujú aj problémy s rečou ako zlá artikulácia, nezrozumiteľnosť alebo úplná neschopnosť rozprávať. Zriedkavými nie sú ani náhle mdloby a spadnutie na zem [19].

3.6.5. Infarkt myokardu

Infarkt myokardu vzniká za podmienok náhleho a zároveň úplného uzavretia tepny. Jeho najčastejším príznakom, ktorý sa prejavuje až u 80% pacientov, býva silná bolesť v oblasti za hrudnou kosťou, ktorá pretrváva desiatky minút až hodiny. Bolesť v tomto prípade nezávisí od

polohy, v ktorej sa telo nachádza, ani od vykonávaného pohybu. Môže sa prejaviť aj v ramenách, ľavej ruke, krku, dolnej čeľusti či v nadbruší. Často je bolesť sprevádzaná aj nadmerným potením, nutkaním na zvracanie alebo aj samotným zvracaním, a to v kombinácii s psychickým nepokojom vyvolaným strachom o život [18].

3.6.6. Aortálna stenóza

Aortálna stenóza je jedným z kardiovaskulárnych ochorení, ktoré postihuje aortálnu chlopňu nachádzajúcu sa medzi ľavou komorou a aortou. Konkrétne ide o zúženie tejto chlopne, čo zabraňuje adekvátnemu pumpovaniu krvi do aorty, najväčšej tepny v ľudskom tele [20].

3.6.7. Koronarografické vyšetrenie

Koronarografia, nazývaná aj koronárna angiografia, je vyšetrenie vencových (koronárnych) tepien zásobujúcich srdce krvou, a teda kyslíkom a živinami. Cieľom koronarografického vyšetrenia je zistiť potenciálne zúženie vnútorného priemeru srdcových tepien, čo zapríčiňuje zníženie prietoku krvi do niektorých oblastí srdcového svalu. Nedostatočné množstvo krvi a potrebných látok, ktoré sú ňou prenášané do oblasti za zúžením tepny spôsobuje u pacienta bolesť na hrudi. Najskôr iba pri námahe, ale neskôr aj v pokoji. V prípade úplného uzavretia cievy nastáva u pacienta infarkt a odumretie časti srdcovej svaloviny [21].

Koronarografické vyšetrenie je invazívne, a teda vyžaduje zásah do organizmu. Nie je však bolestivé, pretože prebieha pod lokálnym znecitlivením pomocou anestetika v oblasti vpichu. Ak sa počas vyšetrenia nevyskytnú u pacienta žiadne komplikácie, vyšetrenie spolu s prípravou trvá menej ako pol hodiny [21].

Samotné vyšetrenie spočíva v zavedení katétra (tenkej ohybnej trubičky) dovnútra tepny cez pravú slabinu až k srdcu pacienta. Do srdcových tepien sa následne cez katéter vstrekne kontrastná látka (špeciálne farbivo), ktorou sa zviditeľní určitá časť koronárneho riečisko na röntgenovom obraze. Na zachytenie priebehu a výsledku vyšetrenia sa používa záznamové zariadenie (napr. DVD), ktoré umožňuje jeho neskoršie detailné posúdenie. Keďže je koronarografia röntgenové vyšetrenie, počas jeho trvania sa ožaruje okrem samotného pacienta aj lekár a zdravotná sestra. Z tohto dôvodu sa toto vyšetrenie vykonáva iba pri podozrení na zúženie až upchatie srdcových tepien u pacienta, alebo pred každým plánovaným chirurgickým zákrokom na srdci a jeho cievach [21].

Koronarografia dokáže odhaliť prítomnosť koronárnej artériovej choroby (KACH). KACH je poškodenie cievnej steny spôsobené ukladaním tukov a vápnika, ktoré môže viesť k zúženiu alebo až uzavretiu cievy. Výsledkom tohto vyšetrenia je pomoc lekárovi pri výbere najvhodnejšej alternatívy liečby pre daného pacienta, a to takej, ktorá najviac zníži riziko smrti, infarktu

myokardu, zabezpečí normálnu funkciu srdca a zlepšenie kvality života odstránením bolesti na hrudníku [22].

4. Návrh a implementácia vlastného DSS

Táto kapitola je zameraná na návrh DSS, identifikáciu používateľských požiadaviek, prípadov použitia a samotnú implementáciu navrhovaného DSS.

4.1. Návrh DSS

4.1.1. Identifikácia používateľských požiadaviek

Cieľovou skupinou používateľov, pre ktorých je tento DSS navrhovaný sú kardiológovia.

Funkcionálne používateľské požiadavky:

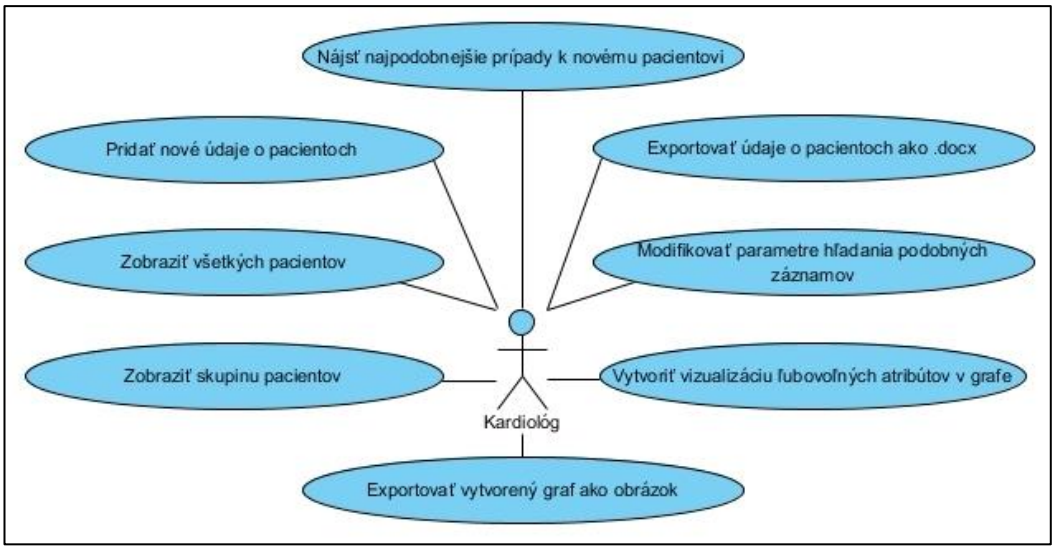
- Zobrazenie dvadsiatich najpodobnejších prípadov k aktuálnemu pacientovi.
- Export dát o aktuálnom pacientovi a najpodobnejších pacientov vo formáte „.docx“.
- Export grafov obsahujúcich vizualizácie dát ako obrázkov.
- Vizualizácia dát o pacientoch roztriedených do špecifikovaných rizikových skupín.

Nefunkcionálne používateľské požiadavky:

- DSS v slovenskom jazyku.
- Jednoduché a intuitívne používanie systému.
- Možnosť potenciálneho rozšírenia dátovej množiny DSS o nové atribúty.

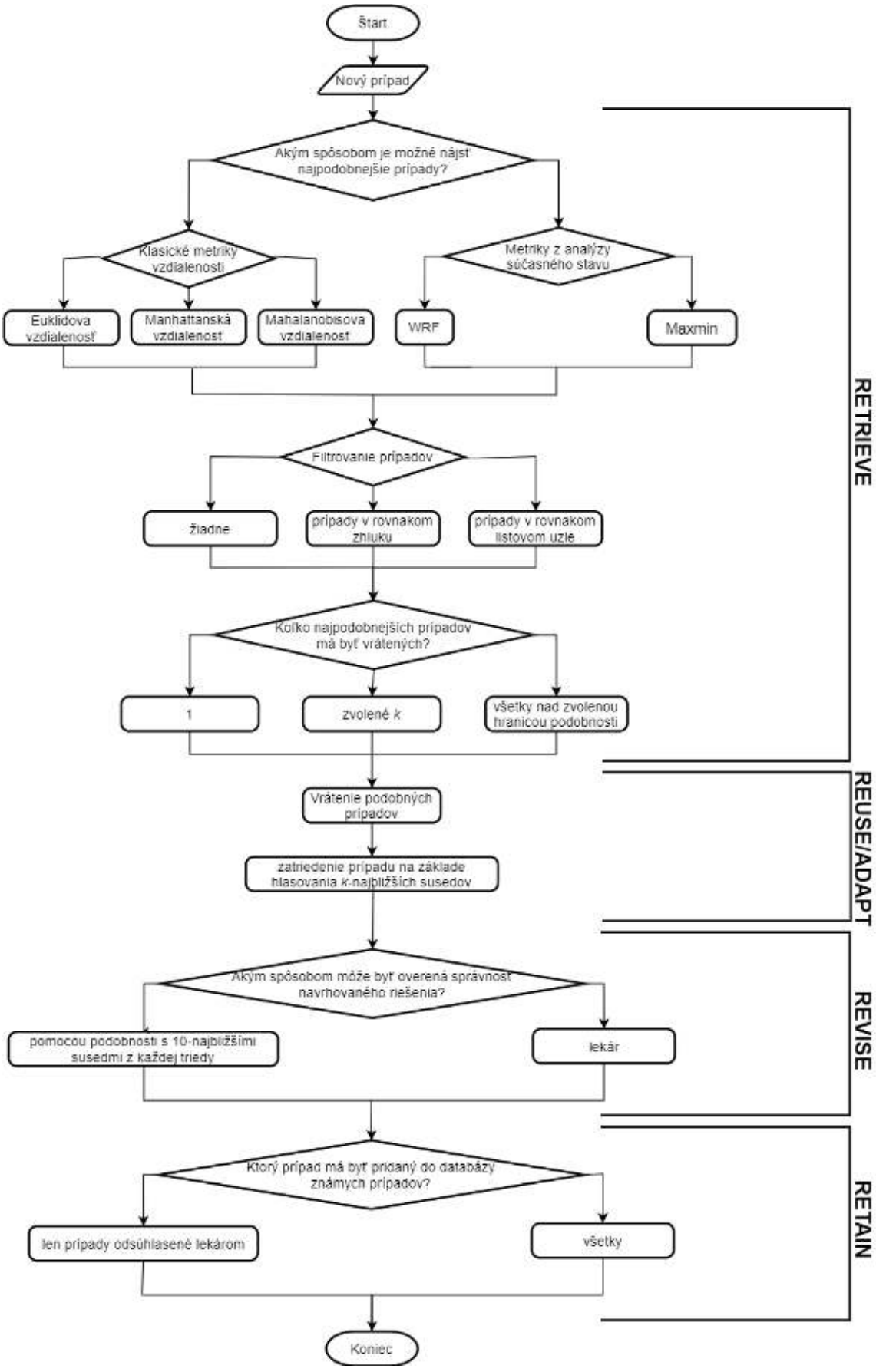
4.1.2. Diagram prípadov použitia

Diagram na Obr. 5 zobrazuje možné prípady použitia navrhovaného systému. Vykresľuje používateľa aplikácie - kardiológa ako jediného aktéra a funkcionality systému, ktoré by mal mať k dispozícii.



Obr. 5 Diagram príadov použitia navrhovaného DSS

Na základe skúmania teoretických znalostí o metóde CBR v kombinácii s poznatkami nadobudnutými prostredníctvom analýzy súčasného stavu bol zhotovený vývojový diagram (Obr. 6), ktorý znázorňuje možnosti implementácie jednotlivých fáz CBR cyklu, ktoré sa javia ako najvhodnejšie pre uplatnenie v nami skúmanej doméne kardiovaskulárnych chorôb.



Obr. 6 Diagram zobrazujúci možnosti implementácie jednotlivých fáz CBR

Navrhovaný systém na podporu diagnostiky kardiovaskulárnych chorôb bude pracovať s dátovou množinou známych prípadov, ktoré sú reprezentované záznamami v relačnej tabuľke. CBR cyklus začína vstupom nového prípadu, v tomto prípade pacienta, ktorý absolvoval potrebné vyšetrenia, ktorých výsledky sú zadané do systému spolu s vybranými osobnými údajmi.

Vtedy môže začať realizácia prvej fázy CBR cyklu RETRIEVE, ktorej úlohou je nájsť najpodobnejšie známe prípady k novému pacientovi. Tento krok môže byť vykonaný prostredníctvom metód založených na štandardných metrikách, akými sú Euklidova, Manhattanská alebo Mahalanobisova vzdialenosť, ale taktiež výpočtami ako napríklad WRF alebo Maxmin (pozri Tab. 1). Ďalšou možnosťou implementácie hľadania podobných prípadov je využitie modelov na dolovanie v dátach, ako napríklad rozhodovacie stromy či zhlukovanie na filtrovanie záznamov. Vtedy sú za najpodobnejšie prípady považované tie záznamy, ktoré sa nachádzajú v rovnakom listovom uzle rozhodovacieho stromu (pozri kapitolu 2.2.2) alebo v tom istom zhluku (pozri kapitolu 2.2.3), do ktorého by bol nový prípad zaradený.

Čo sa týka počtu vrátených najpodobnejších prípadov, môže to byť jeden najviac podobný záznam, vhodne zvolené *k* z intervalu <1;10> alebo všetky prípady podobnejšie ako stanovená hranica podobnosti. V poslednom prípade však môže nastať situácia, kedy sa v databáze nebude vyskytovať žiadny dostatočne podobný prípad a systém nevráti žiadny záznam. Kvôli takejto situácii, keď je nový prípad veľmi odlišný od ostatných, je vhodnejšie zvoliť konkrétny počet vrátených prípadov, ale zároveň používateľovi poskytnúť informáciu o tom, nakoľko sú si prípady podobné, resp. ich vzdialenosť od zadaného prípadu.

Na základe prevažujúcej hodnoty cieľového atribútu u nájdených najpodobnejších záznamov je v rámci fázy REUSE/ADAPT stanovená cieľová trieda nového prípadu. Vo fáze REVISE je overovaná správnosť navrhovaného riešenia. Tento proces je možné uskutočniť pomocou lekára, ktorý na základe svojich znalostí a skúseností môže navrhované riešenie buď odsúhlasiť, alebo zamietnuť. Inou možnou alternatívou, ktorá by mohla byť realizovaná automaticky bez nutnosti akéhokoľvek zásahu zo strany experta, je porovnať súčet podobností desiatich najpodobnejších prípadov z každej triedy voči novému prípadu. Aby sme riešenie mohli považovať za spoľahlivé, najvyššia miera podobnosti by mala vyjsť s tými prípadmi, ktoré patria do rovnakej triedy ako navrhované riešenie. V opačnom prípade by bolo riešenie upravené podľa inej najpodobnejšej triedy.

Podstatou poslednej fázy CBR cyklu zvanej RETAIN je uchovanie správne diagnostikovaného prípadu do databázy známych prípadov, čo umožňuje systému nadobúdať nové znalosti. Tie môže následne využiť pri určovaní diagnózy u budúcich pacientov. Otázkou je, či každý jeden prípad

bude vhodným kandidátom na uloženie do databázy. Prvou alternatívou je pokladať každý nový prípad za prínos pre systém a uložiť ho automaticky. Tento prístup by mohol zapríčiniť, že do databázy známych prípadov sa dostanú prípady s odhadovaným riešením, ktoré sa môže výrazne líšiť od skutočného stavu pacienta a databáza by sa stala nespoľahlivou. Druhou alternatívou je ponechať toto rozhodnutie na experta a znalostnú bázu rozšíriť iba ak to on uzná za vhodné.

4.2. Implementácia navrhovaného DSS

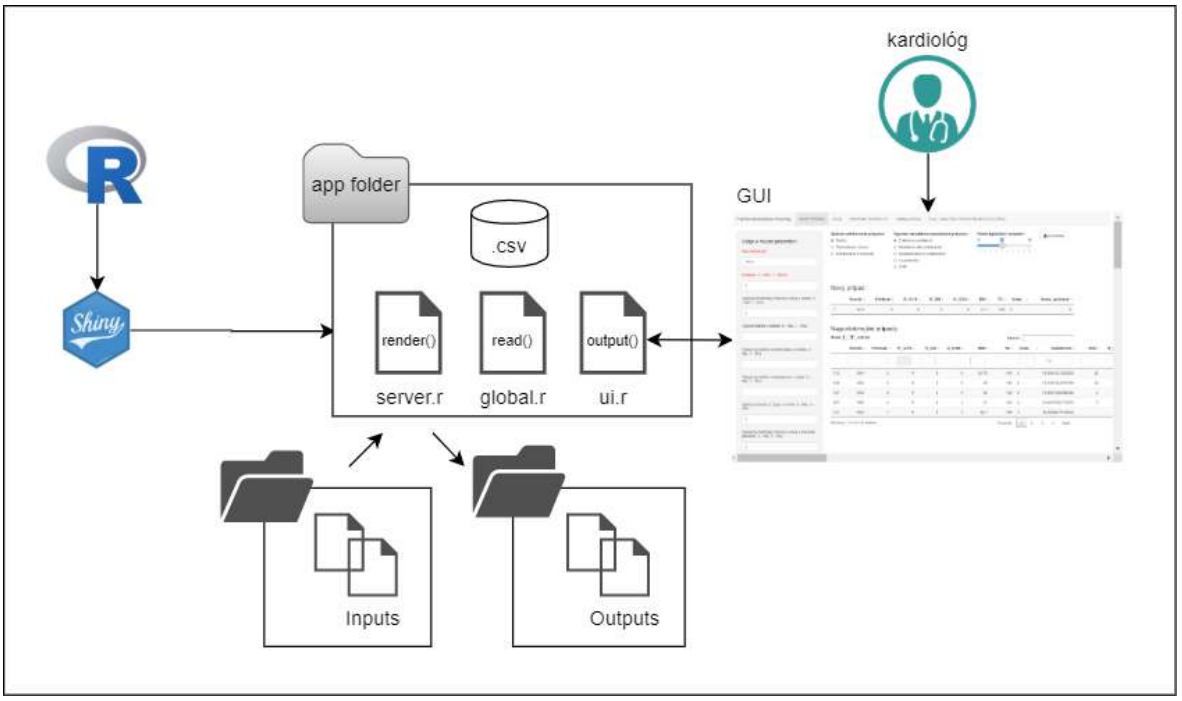
Pre implementáciu navrhovaného systému na podporu rozhodovania bol zvolený open source jazyk R s využitím balíka Shiny, ktorý umožňuje vytvárať interaktívne webové aplikácie.

4.2.1. Architektúra DSS

Hlavnými komponentmi Shiny aplikácie sú objekt používateľského rozhrania (UI) a funkcia servera, ktoré sú nevyhnutné na vytvorenie objektu samotnej aplikácie. Jednotlivé komponenty aplikácie boli rozdelené do separátnych súborov vo formáte skriptov v jazyku R. Skript server.r pozostávajúci z render() a reactive() funkcií dokáže v kombinácii so skriptom ui.r obsahujúcim output() funkcie vygenerovať grafické používateľské rozhranie pre koncového používateľa, a teda v našom prípade kardiológa.

V ďalšom skripte s názvom global.r sú umiestnené library() funkcie na načítanie všetkých potrebných balíkov pre chod aplikácie, read() funkcie na načítanie dátovej množiny a algoritmus slúžiaci na predspracovanie načítaných dát. Taktiež sa tu nachádzajú aj definície vlastných funkcií, ktoré tvoria hlavnú funkcionalitu aplikácie, a teda algoritmus hľadania k-najbližších susedov.

Dátový zdroj, ktorý aplikácia využíva na uchovanie záznamov o pacientoch je uložený vo formáte .csv a spolu s ostatnými komponentmi je uložený v priečinku Shiny aplikácie. Interakcie medzi jednotlivými prvkami architektúry sú znázornené na Obr. 7.



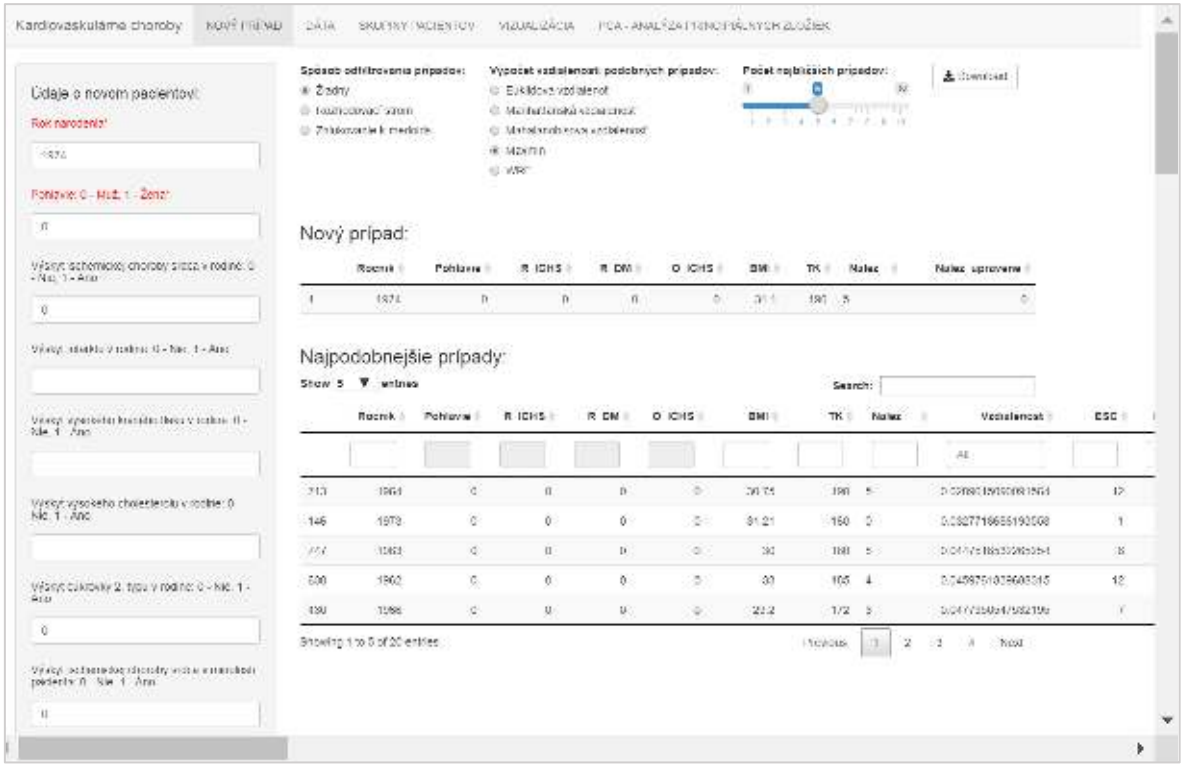
Obr. 7 Architektúra navrhovaného systému

4.2.2. Funkcie navrhovaného DSS

Štruktúru navrhovaného systému na podporu rozhodovania tvorí 5 rôznych kariet. Navigáciu medzi nimi umožňuje hlavné menu v hornej časti obrazovky. Popis funkcií nachádzajúcich sa na jednotlivých kartách aplikácie:

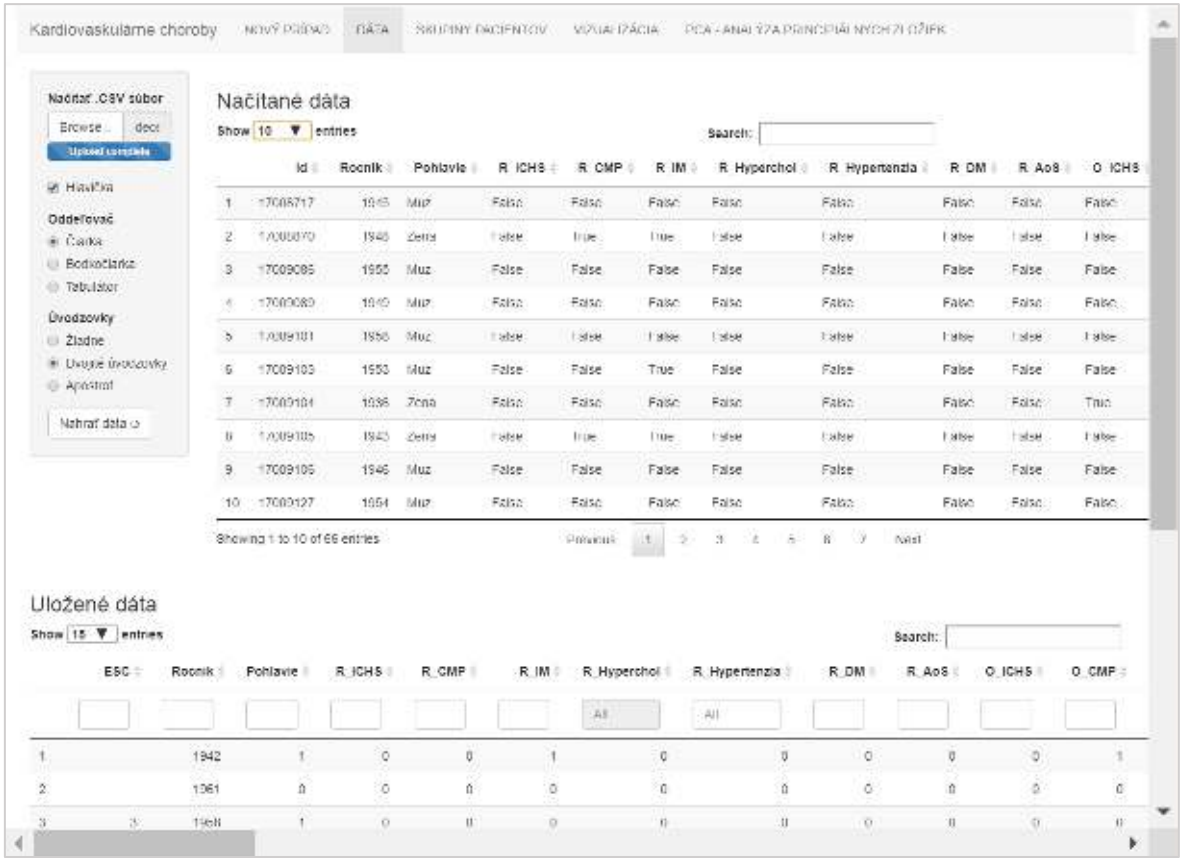
1. Karta **NOVÝ PPRÍPAD**: obsahuje hlavnú funkcionalitu aplikácie. Vľavo (Obr. 8) sa nachádza bočný panel umožňujúci používateľovi zadať základné informácie o novom pacientovi ako sú rok jeho narodenia a pohlavie, ale aj informácie týkajúce sa aktuálneho zdravotného stavu pacienta a histórie výskytu kardiovaskulárnych chorôb u samotného pacienta alebo u jeho rodinných príslušníkov. Množina konkrétnych vstupných atribútov bola vybraná kardiológom podľa toho, či daný atribút je alebo nie je dôležitý pre diagnostiku kardiovaskulárnych ochorení. Po zadaní vstupných informácií sa automaticky spustí algoritmus hľadania najpodobnejších prípadov a v pravej časti sa zobrazia dáta o aktuálnom pacientovi spolu s predikovanou závažnosťou jeho koronarografického nálezu, ale aj dáta o používateľom zvolenom počte (1 až 20) najbližších prípadov podľa aktuálnych nastavení. Ak chce používateľ zmeniť parametre hľadania, môže si vybrať jednu z piatich metód výpočtu vzdialenosti (Euklidovská, Manhattanská a Mahalanobisova vzdialenosť, Maxmmin alebo WRF), ktorú môže, ale nemusí použiť v kombinácii s vybraným algoritmom strojového učenia (rozhodovací strom alebo zhlukovanie typu k-medoids) na odfiltrovanie záznamov.

Ďalším voliteľným parametrom je počet najbližších susedov, na základe ktorých bude predikovaná závažnosť diagnózy pacienta, a to v rámci intervalu 1 až 10. V prípade, že používateľ (kardiológ) chce aktuálny prípad odkonzultovať s jeho kolegami, môže si dáta o týchto prípadoch vyexportovať do reportu vo formáte .docx pomocou tlačidla nachádzajúceho sa v pravom hornom rohu.



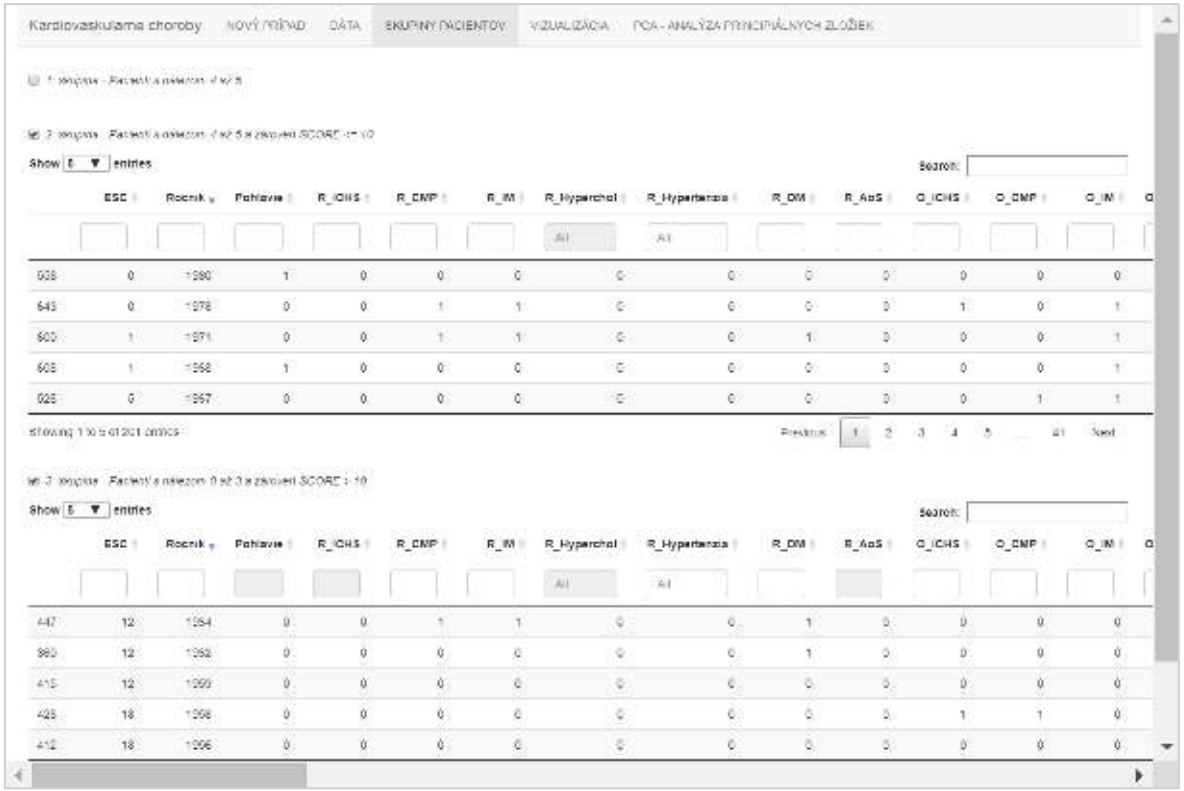
Obr. 8 Karta NOVÝ PRÍPAD

2 . Karta **DÁTA**: umožňuje používateľovi pozrieť si všetky dáta uložené v databáze už známych prípadov, taktiež ich ľubovoľne zoradiť, filtrovať a vyhľadať požadovanú hodnotu. Ďalej umožňuje používateľovi nahrať do aplikácie súbor obsahujúci nové dáta o pacientoch, pričom mu je poskytnutý náhľad aktuálne vybraného súboru a v paneli na ľavej strane sa nachádzajú nastavenia týkajúce sa jeho štruktúry (Obr. 9).



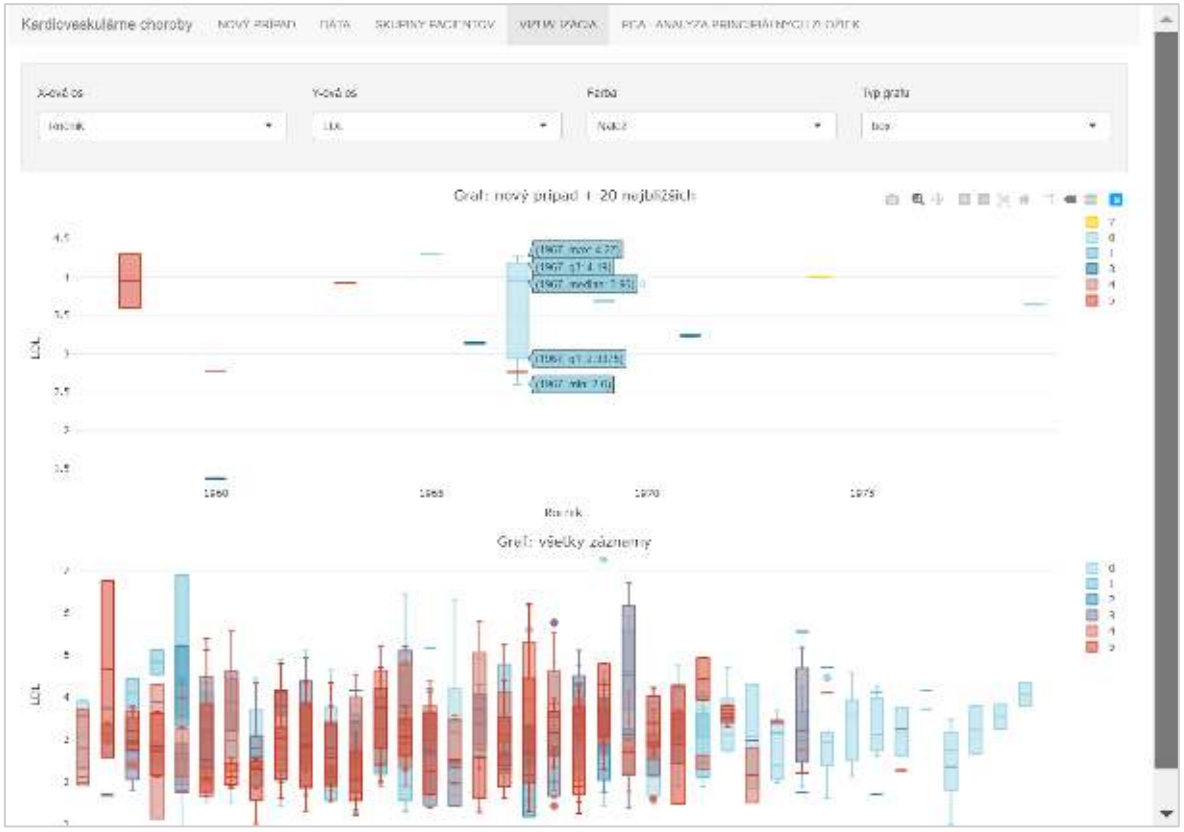
Obr. 9 Karta DÁTA

3 . Karta **SKUPINY PACIENTOV**: poskytuje možnosť zobrazenia a porovnania dát o špecifických skupinách pacientov (Obr. 10), konkrétne o pacientoch s vysokým rizikom výskytu kardiovaskulárnej choroby podľa ich SCORE (pozri kapitolu 5.3), ale zároveň s nízkym stupňom závažnosti nálezu a opačne.



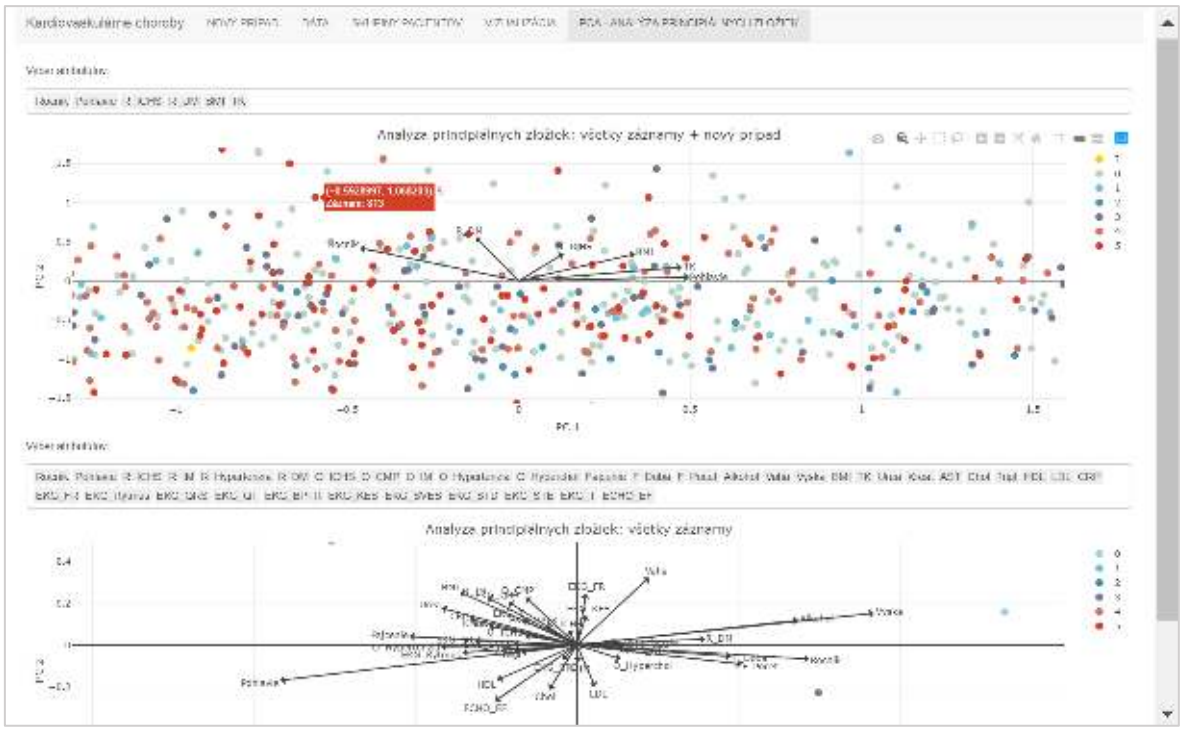
Obr. 10 Karta SKUPINY PACIENTOV

4 . Karta **VIZUALIZÁCIA**: slúži na vykreslenie dát o pacientoch v podobe grafov. Používateľ si môže v hornej časti obrazovky zvoliť, ktorý z atribútov bude zobrazený na osiach x a y, podľa ktorého atribútu budú body rozdelené do farebných skupín, a taktiež aj typ grafu z možností: bodový, stĺpcový alebo krabicový graf (Obr. 11). Na tejto karte sa nachádzajú dva rôzne grafy, pričom v prvom sú vykreslené dáta z nového prípadu zadaného na karte NOVÝ PRÍPAD a 20 jemu najbližších prípadov. V druhom grafe sú zobrazené všetky záznamy z databázy známych prípadov. Oba grafy poskytujú možnosť exportu vizualizácie vo formáte .png.



Obr. 11 Karta VIZUALIZÁCIA

5 . Karta **PCA – ANALÝZA PRINCIPIÁLNYCH ZLOŽIEK**: ponúka vizualizáciu vybranej skupiny atribútov znázornenú prostredníctvom dvoch hlavných komponentov analýzy principiálnych zložiek. Ako v predchádzajúcej karte, aj tu sa nachádzajú dva odlišné grafy, ale nad každým z nich sa nachádza samostatné vstupné pole pre výber atribútov (Obr. 12). Na prvom grafe sú vykreslené všetky záznamy z databázy spolu s údajmi zadanými v karte NOVÝ PRÍPAD, z tohto dôvodu je prednastavený výber skupiny atribútov podľa vyplnených údajov o novom pacientovi. V druhom grafe síce nie je vykreslený nový prípad, ale sú tam všetky ostatné známe prípady, čo umožňuje ľubovoľne meniť skupinu atribútov použitých pre túto vizualizáciu. V každom z týchto grafov je takisto ako v karte VIZUALIZÁCIA možnosť pre ich export ako .png, ale aj približovanie či filtrovanie prípadov pomocou legendy nachádzajúcej sa napravo od daného grafu. Taktiež sa po prejdení kurzorom na akýkoľvek dátový bod zobrazí tooltip s informáciou o jeho presných súradniciach na grafe a o tom, o ktorý záznam v databáze sa jedná spolu so stupňom vážnosti jeho nálezu.



Obr. 12 Karta PCA - ANALÝZA PRINCIPIÁLNYCH ZLOŽIEK

5. Experimentálne vyhodnotenie

Táto kapitola je venovaná podrobnému popisu dátovej množiny, jej následnej analýze a aj experimentom vykonaným na tejto dátovej vzorke. V tejto kapitole bolo postupované podľa metodiky CRISP-DM (Cross-Industry Standard Process for Data Mining), nakoľko úloha má aj silný analytický charakter nad množinou dát extrahovaných z lekárskych záznamov kardiologických pacientov.

5.1. Pochopenie problému

Cieľom tejto diplomovej práce je navrhnúť, implementovať a otestovať systém na podporu rozhodovania s prvkami prípadového usudzovania pre proces diagnostiky kardiovaskulárnych ochorení. Základom takéhoto systému je analýza dostupných dát o pacientoch.

Primárnou úlohou navrhovaného systému je poskytnúť jeho používateľovi, teda kardiológovi dáta o najpodobnejších známych prípadoch k aktuálnemu pacientovi a na základe týchto dát navrhnúť kardiológovi, či je u daného pacienta vysoké riziko výskytu zúženia koronárnych ciev. Tieto informácie môže následne kardiológ využiť v procese rozhodovania o tom, či je nevyhnutné poslať tohto pacienta na koronarografické vyšetrenie (pozri kapitolu 3.6.7) alebo to naopak nie je potrebné.

Cieľom tejto diplomovej práce z pohľadu dolovania v dátach je vytvoriť klasifikátor na báze algoritmu k-najbližších susedov, ktorý by podľa hodnoty cieľovej triedy najbližších prípadov nachádzajúcich sa v okolí zadaného záznamu klasifikoval do ktorej cieľovej skupiny tento záznam patrí, ale zároveň aj zobrazil nájdené najbližšie prípady.

5.2. Pochopenie dát

V tejto diplomovej práci pracujeme s dátovou množinou, ktorá pozostáva z dát získaných od reálnych pacientov z Východoslovenského ústavu srdcových a cievnych chorôb (VÚSCH) v Košiciach. Táto dátová množina obsahuje 66 atribútov a 820 záznamov. Popis jednotlivých atribútov môžete vidieť v Tab. 2. Atribúty, ktorých názov je podčiarknutý, boli označené kardiológom ako dôležité.

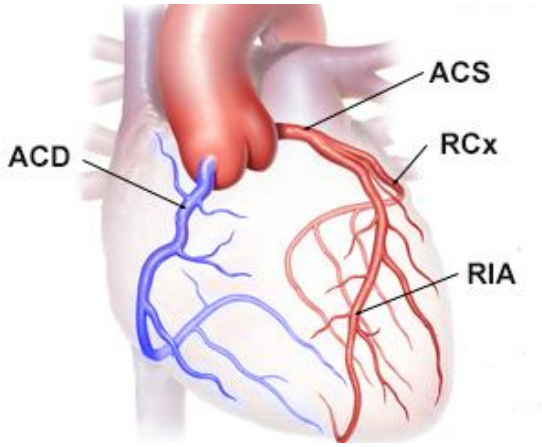
Tab. 2 Popis atribútov

Názov atribútu	Typ atribútu	Popis atribútu
Id	Celé číslo	Jedinečný identifikátor pre každý záznam
RC	Reťazec	Rodné číslo pacienta (zašifrované)
Rocnik	Celé číslo	Rok narodenia pacienta
Pohlavie	Kategorický (Muz, Zena)	Pohlavie pacienta
R_ICHS	Binárny	Výskyt ischemickej choroby srdca u rodinného

		príslušníka pacienta
R_CMP	Binárny	Výskyt cievnej mozgovej príhody u rodinného príslušníka pacienta
R_IM	Binárny	Výskyt infarktu u rodinného príslušníka pacienta
R_Hypertenzia	Binárny	Výskyt vysokého krvného tlaku u rodinného príslušníka pacienta
R_Hyperchol	Binárny	Výskyt vysokého cholesterolu u rodinného príslušníka pacienta
R_DM	Binárny	Výskyt cukrovky 2. typu u rodinného príslušníka pacienta
R_AoS	Binárny	Výskyt aortálnej stenózy (zúženia aorty) u rodinného príslušníka pacienta
O_ICHS	Binárny	Výskyt ischemickej choroby srdca v minulosti u pacienta
O_CMP	Binárny	Výskyt cievnej mozgovej príhody v minulosti u pacienta
O_IM	Binárny	Výskyt infarktu v minulosti u pacienta
O_Hypertenzia	Binárny	Výskyt vysokého krvného tlaku v minulosti u pacienta
O_Hyperchol	Binárny	Výskyt vysokého cholesterolu v minulosti u pacienta
O_DM	Binárny	Výskyt cukrovky 2. typu v minulosti u pacienta
O_AoS	Binárny	Výskyt aortálnej stenózy (zúženia aorty) v minulosti u pacienta
Fajcenie	Kategorický (1, 2, 3)	Výskyt fajčenia u pacienta (1 - fajčiar, 2 - exfajčiar, 3 - nefajčiar)
F_Doba	Číslo	Počet rokov fajčenia, prípadne nefajčenia u pacienta
F_Pocet	Číslo	Počet vyfajčených cigariet za deň pacientom
Alkohol	Binárny	Užívanie alkoholu pacientom
Vaha	Celé číslo	Váha pacienta (kg)
Vyska	Celé číslo	Výška pacienta (cm)
BMI	Číslo	BMI (Index telesnej hmotnosti) pacienta
TK	Celé číslo	Systolický krvný tlak pacienta (tzv. „horný“)
Urea	Číslo	Hladina močoviny v krvi pacienta
Kreat	Číslo	Hladina kreatinínu v krvi pacienta
AST	Číslo	Hladina enzýmu vylučovaného pri poškodení pečene (Asparate transaminase) pacienta
Na	Číslo	Hladina sodíka v krvi pacienta
K	Číslo	Hladina draslíka v krvi pacienta
Chol	Číslo	Hladina cholesterolu u pacienta
Trigl	Číslo	Hladina triacylglycerolov u pacienta
HDL	Číslo	Hladina lipoproteínov s vysokou hustotou u pacienta (tzv. „dobrý cholesterol“, ktorý transportuje prebytočný cholesterol z tkanív)
LDL	Číslo	Hladina lipoproteínov s nízkou hustotou u pacienta (tzv. „zlý cholesterol“, ktorý transportuje cholesterol do tkanív)
CRP	Číslo	Hladina C-reaktívneho proteínu u pacienta (indikuje akútny zápal v tele pacienta)
CL	Číslo	Hladina chloridov v krvi pacienta
FBG	Číslo	Hladina fibrinogénov u pacienta
HIV	Binárny	Výsledok testu na vírus HIV u pacienta

HBs	Binárny	Prítomnosť antigénu značiaceho prítomnosť vírusu žltačky typu B u pacienta
EKG_FR	Celé číslo	Frekvencia úderov srdca za minútu u pacienta
EKG_Rytmus	Kategorický (SR, Fib)	Typ srdcového rytmu u pacienta (SR - pravidelný rytmus, Fib - fibrilácia)
EKG_PQ	Celé číslo	Dĺžka intervalu od začiatku P vlny (frekvencie sťahov komôr) do začiatku komorového komplexu v milisekundách
EKG_QRS	Celé číslo	Doba depolarizácie komôr srdca v milisekundách
EKG_QT	Celé číslo	Doba od začiatku QRS po koniec T vlny v milisekundách
EKG_BLTR	Binárny	Prítomnosť blokády ľavého Tawarového ramienka u pacienta (porucha vedenia vzruchu myokardom)
EKG_BPTR	Binárny	Prítomnosť blokády pravého Tawarového ramienka u pacienta
EKG_KES	Binárny	Prítomnosť komorových extrasystol (predčasných sťahov) u pacienta
EKG_SVES	Binárny	Prítomnosť supraventrikulárnych (predsieňových) extrasystol u pacienta
EKG_STD	Binárny	Prítomnosť depresií (poklesov) v ST úseku - medzi vlnou P a počiatkom QRS (normálny výsledok je izoelektrická rovina)
EKG_STE	Binárny	Prítomnosť elevácií (výstupov) v ST úseku u pacienta
EKG_T	Binárny	Nález pri meraní T vlny u pacienta - reprezentuje repolarizáciu komorového myokardu (normálny výsledok je negatívna vlna vo zvodoch *V1* a *aVR*)
ECHO_EF	Celé číslo	Ejekčná frakcia ľavej komory, percentuálny podiel vytlačenej krvi pri každom sťahu srdca u pacienta
ECHO_PH	Kategorický (0, 1, 2, 3)	Stupeň pľúcnej hypertenzie (zvýšeného tlaku) u pacienta (0 - bez PH, 1 - ľahká PH, 2 - mierna PH, 3 - závažná PH)
SV_mostik	Binárny	Prítomnosť svalového myokardiálneho mostíka v jednej z vetiev u pacienta
ACS	Číslo	Percentuálne zúženie vetvy ACS
RIA	Číslo	Percentuálne zúženie vetvy RIA
RD1	Číslo	Percentuálne zúženie vetvy RD1, súčasťou RIA
RD2	Číslo	Percentuálne zúženie vetvy RD2, súčasťou RIA
RCX	Číslo	Percentuálne zúženie vetvy RCX
RIM	Číslo	Percentuálne zúženie vetvy RIM, súčasťou RCX
RMS1	Číslo	Percentuálne zúženie vetvy RMS1, súčasťou RCX
RMS2	Číslo	Percentuálne zúženie vetvy RMS2, súčasťou RCX
ACD	Číslo	Percentuálne zúženie vetvy ACD
RIP	Číslo	Percentuálne zúženie vetvy RIP
Nalez	**Kategorický (0 až 5)**	**Cieľový atribút označujúci stupeň závažnosti nálezu: 0 - žiadny nález 1 - aspoň 1 vetva obsahuje 10 % zúženie 2 - akákoľvek vetva, okrem RIA má 20 - 50% zúženie 3 - vetvy RIA majú 20 - 50%, ostatné 50 - 70% zúženie 4 - vetvy RIA majú maximálne 70% zúženie (50% - 70%), ostatné 70% - 100% zúženie 5 - aspoň jedna z vetiev RIA má viac ako 70% zúženie**

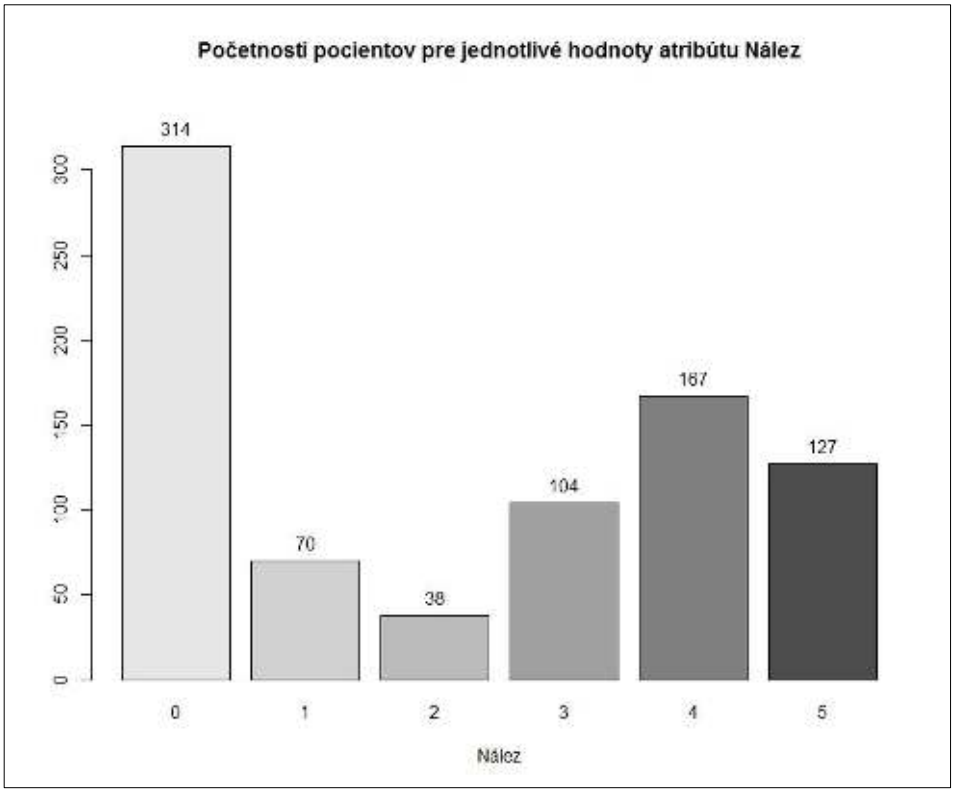
Cieľový atribút *Nalez* vznikol odvodením z nasledovných atribútov: ACS, RIA, RD1, RD2, RCX, RIM, RMS1, RMS2, ACD, RIP. Tieto atribúty hovoria o percentuálnom zúžení jednotlivých vetiev koronárnych artérií u daného pacienta. Ich hodnoty boli zistené prostredníctvom koronárnej angiografie (kapitola 3.6.7). Pre znázorenie je na Obr. 13 zobrazená pravá (ACD) a ľavá (ACS) koronárna artéria, ktoré sa postupne ďalej rozvetvujú.



Obr. 13 Koronárne (srdcové) artérie[2]

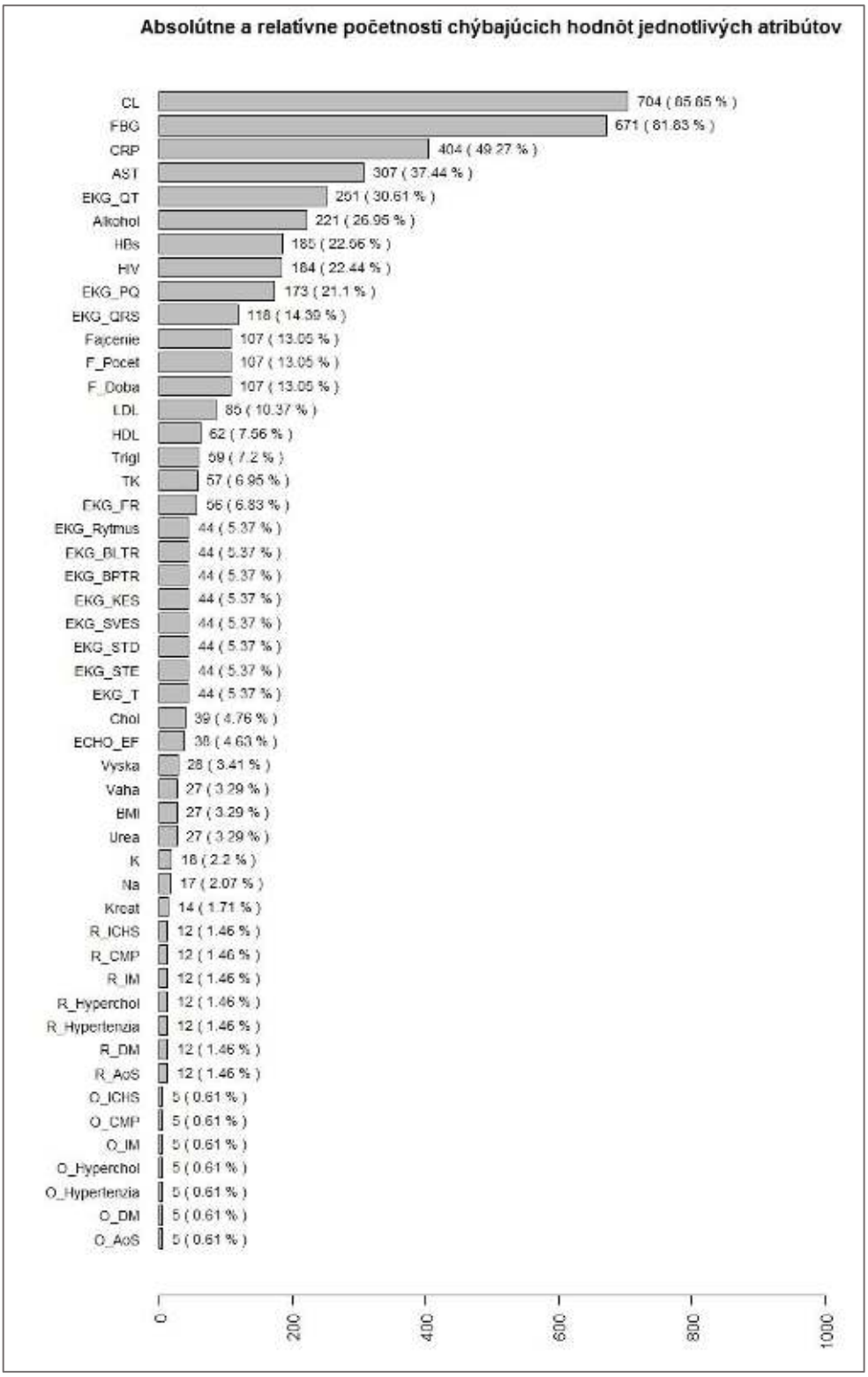
Na nasledovnom grafe (Obr. 14) je možné pozorovať koľko záznamov z dátovej množiny spadá do jednotlivých kategórií cieľového atribútu *Nalez*. Zo všetkých šiestich kategórií spadá najviac záznamov do triedy 0 (38,29%), ktorá prestavuje zdravých ľudí. Naopak najmenej prípadov patrí do triedy 2 (4,36%). Na základe týchto skutočností je zrejmé, že triedy cieľového atribútu sú nevyvážené.

[2] Zdroj: BLAHÚT, P. TECHmED: Cievne zásobenie srdca. [Online]. Dostupné na internete: <https://www.techmed.sk/ekg-kniha/obr/002/coronary-arteries.png>



Obr. 14 Graf početností pacientov pre jednotlivé hodnoty atribútu *Nález*

V tejto dátovej množine sa nachádza aj množstvo prázdnych hodnôt. Ich absolútne aj relatívne početnosti sú vyobrazené na Obr. 15, zoradené zostupne pre každý atribút obsahujúci chýbajúce hodnoty. Na grafe je možné pozorovať, že najvyššie hodnoty (nad 80%), dosiahli atribúty CL a FBG. Naopak, atribúty týkajúce sa diagnóz rôznych chorôb v minulosti pacienta (O_ICHS, O_CMP, O_IM, O_Hypertenzia, O_Hyperchol, O_DM, O_AoS) alebo jeho rodinných príslušníkov (R_ICHS, R_CMP, R_IM, R_Hypertenzia, R_Hyperchol, R_DM, R_AoS) obsahovali iba zanedbateľné percento chýbajúcich hodnôt (do 2%).



Obr. 15 Graf absolútnych a relatívnych početností chýbajúcich hodnôt atribútov

5.3. Príprava dát

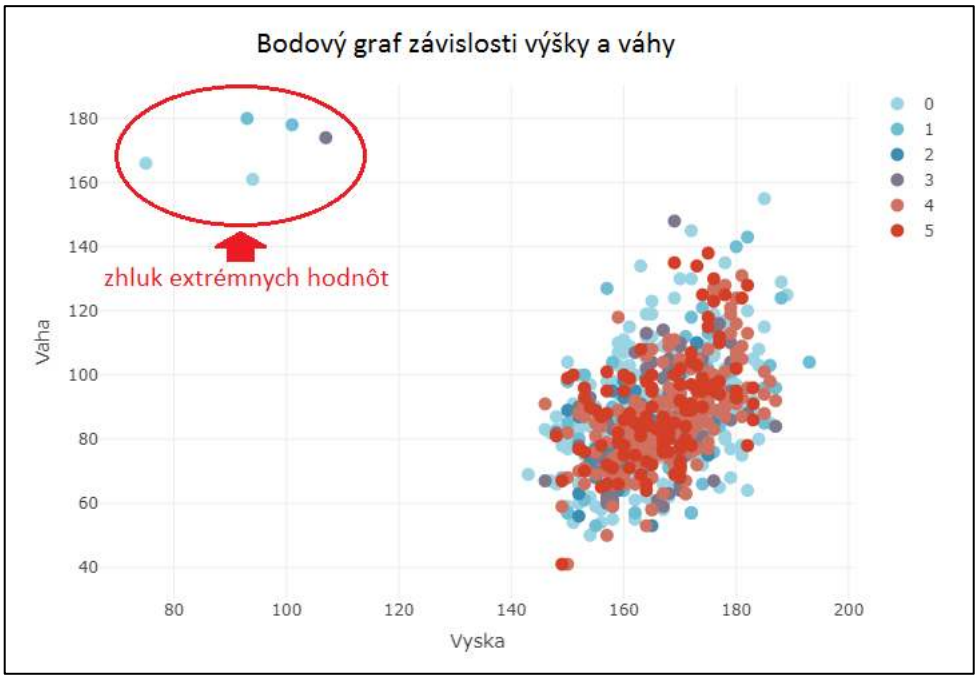
V rámci prípravy dát boli z dátovej množiny odstránené atribúty: Id a RC (šifrované rodné číslo). Takisto bola odstránená aj skupina atribútov, na základe ktorej bol odvodený cieľový atribút *Nalez*, konkrétne: ACS, RIA, RD1, RD2, RCX, RIM, RMS1, RMS2, ACD, RIP.

Pre ďalšiu prácu s dátovou množinou bolo nutné konvertovať všetky kategorické atribúty obsahujúce textové hodnoty na numerické hodnoty. Konverzie boli vykonané týmto spôsobom (Tab. 3):

Tab. 3 Konverzia kategorických atribútov na numerické

Názov atribútu	Pôvodná hodnota	Nová hodnota
Pohlavie	„Muz“	0
	„Zena“	1
EKG_Rytmus	„SR“	0
	„Fib“	1

V rámci prípravy dát bolo uskutočnené aj čistenie dát v tých prípadoch, kde sa vyskytovali extrémne hodnoty, ktoré boli pravdepodobne chybné. V situácii vykreslenej na Obr. 16 je zobrazený bodový graf závislosti výšky a váhy pacientov, kde je možné pozorovať zhluk piatich extrémnych hodnôt. Po bližšom preskúmaní týchto hodnôt vzniklo podozrenie, že hodnoty výšky by v skutočnosti mali predstavovať váhu a hodnoty váhy by mali byť zas výška. Ako kontrola slúžil atribút BMI, ktorý je vypočítavaný na základe výšky a váhy. Preto mohol nastať prípad, kedy bola aj hodnota BMI extrémna, keďže bola vypočítaná z chybných hodnôt výšky a váhy. Z tohto dôvodu boli pri čistení tieto hodnoty navzájom vymenené a následne bola vypočítaná nová hodnota BMI. Druhým prípadom boli nevychyľujúce sa hodnoty BMI, pretože boli vypočítané z opačných hodnôt. To znamenalo, že je potrebné vymeniť hodnoty výšky a váhy, ale ponechať pôvodnú hodnotu BMI.



Obr. 16 Graf závislosti výšky a váhy s extrémnymi hodnotami

Pri uvažovaní o tom, aký spôsob by bol najvhodnejší pre vysporiadanie sa s početnými chýbajúcimi hodnotami atribútov, bolo brané do úvahy, že dáta s ktorými pracujeme pochádzajú od reálnych pacientov. Na to, aby ich mohol brať kardiológ skutočne do úvahy pri uskutočňovaní svojich rozhodnutí, musia byť tieto dáta správne a musí sa na nich dať v plnej miere spoľahnúť. V prípade, ak by sa v systéme miešali reálne aj umelo generované dáta o známych prípadoch, mohlo by to znížiť dôveryhodnosť získaných výsledkov a potenciálne by to tak mohlo viesť k reluktancii pri používaní navrhovaného systému na podporu rozhodovania. Z tohto dôvodu neboli chýbajúce hodnoty nahradzované, ale výpočty prebiehajúce v systéme boli uspôsobené na využitie iba tých záznamov, ktoré majú vyplnené všetky aktuálne potrebné hodnoty atribútov.

S cieľom rozdeliť pacientov do rôznych rizikových skupín podľa zaužívaných smerníc bol do dátovej množiny pridaný atribút „ESC“, ktorý predstavuje SCORE (Systematic COronary Risk Evaluation) pacienta [23]. Atribút bol odvodený na základe pohlavia, veku, fajčenia, hladiny cholesterolu a systolického krvného tlaku pacienta. Tento atribút môže nadobúdať hodnoty z intervalu 0 až 47 a vyjadruje percentuálny odhad rizika výskytu smrteľnej kardiovaskulárnej choroby v najbližších desiatich rokoch u daného pacienta.

5.4. Modelovanie

Keďže sa táto diplomová práca zameriava na využitie princípov prípadového usudzovania v rozhodovacom procese, ako hlavný typ modelu pre klasifikáciu pacientov do cieľových skupín bol využívaný algoritmus k-najbližších susedov (kapitola 2.2.1). Ako doplnkové algoritmy

strojového učenia, ktoré boli použité v kombinácii s hlavným klasifikátorom k-NN na filtrovanie dátovej množiny, sme využili algoritmy rozhodovacieho stromu (kapitola 2.2.2) a k-medoids (kapitola 2.2.3.1).

Na odhad kvality implementovaných metód hľadania najbližších prípadov v rámci navrhovaného DSS bola použitá 10-násobná krížová validácia. Tá bola realizovaná pre rôzny počet najbližších susedov, a to z intervalu 1 až 10. Atribúty použité na testovanie pochádzali z množiny atribútov, ktorá bola označená kardiológom ako „Dôležité atribúty“, pričom daný atribút sa bral do úvahy iba vtedy, keď jeho hodnota bola už zadaná v novom prípade. Atribúty, ktorých hodnota v novom prípade chýbala neboli zahrnuté do výpočtu vzdialenosti. Najvyššie priemerné hodnoty presnosti, ktoré vznikli výpočtom aritmetického priemeru presností v rámci desiatich iterácií krížovej validácie pre konkrétny počet susedov a jednotlivé metódy výpočtu sú zaznamenané v Tab. 4:

Tab. 4 Najvyššia priemerná presnosť implementovaných metód v percentách

Metóda výpočtu vzdialenosti	Spôsob filtrovania prípadov					
	Žiadny		Rozhodovací strom		Zhlukovanie	
	Najvyššia priemerná presnosť	K susedov	Najvyššia priemerná presnosť	K susedov	Najvyššia priemerná presnosť	K susedov
Euklidovská	34.02%	5	29.39%	3	33.54%	10
Manhattanská	35.98%	4	29.39%	10	32.32%	9
Mahalanobisova	39.27%	9	**30.12%**	**9**	33.41%	5
Maxmin	39.39%	10	29.88%	4	**33.66%**	**10**
WRF	**39.88%**	**6**	29.76%	10	33.41%	10

Na základe výsledkov v Tab. 4 môžeme konštatovať, že pre každú z využitých metód výpočtu vzdialenosti bola najvyššia priemerná presnosť dosiahnutá v prípade, keď do výpočtu nebola zahrnutá možnosť filtrovania záznamov. Najvyššie priemerné presnosti, a to nad 39% dosiahli metódy Mahalanobisova vzdialenosť, Maxmin a WRF, pričom Manhattanská vzdialenosť za nimi zaostávala o zhruba 3% a Euklidovská až o 5%. V prípade testovania za využitia kombinovaných metód boli rozdiely v priemerných presnostiach medzi metódami iba minimálne, do 1%. Z týchto dôvodov boli nasledujúce experimenty realizované len na vybraných metódach, teda tých ktoré dosiahli najvyššie priemerné presnosti, ale bez použitia filtrácie záznamov.

Ak uvažujeme o základnom (baseline) modeli, ktorý zjednodušene reprezentuje súčasné rozhodovanie kardiológov ohľadom odporučenia pacienta na koronarografické vyšetrenie, predstavuje ho pravdepodobne model klasifikujúci všetkých pacientov, u ktorých sa vyskytujú symptómy kardiovaskulárnych ochorení do triedy 5. Presnosť takéhoto modelu vypočítaná na základe nami používanej dátovej množiny je na úrovni 15.49%. Najvyššie priemerné presnosti

všetkých implementovaných metód boli v tomto prípade vyššie ako presnosť baseline modelu a ich rozdiely dosahovali od 13.9% až do 24.39%.

5.4.1. Experimentálna zmena počtu hodnôt cieľového atribútu

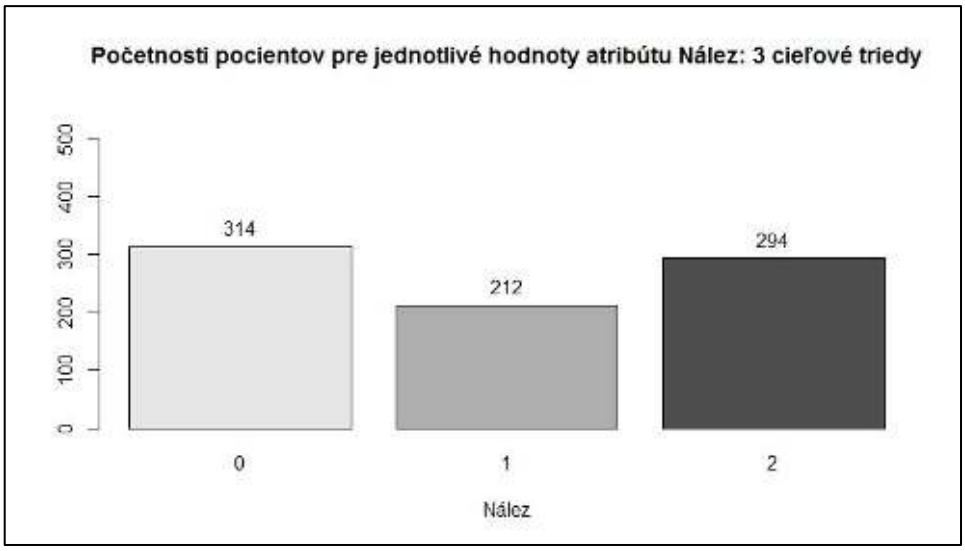
Cieľový atribút našej dátovej množiny *Nalez* pozostáva zo šiestich rôznych hodnôt: 0, 1, 2, 3, 4 a 5. Keďže najvyššia priemerná odhadovaná presnosť predikcie hodnôt atribútu *Nalez* bola iba na úrovni necelých 40%, táto podkapitola je venovaná experimentom so zmenou počtu cieľových tried so zámerom zvýšiť odhadovanú presnosť predikcie. Takáto možnosť zjednodušenia definície úlohy bola zároveň podrobne odkonzultovaná s expertom – kardiológom, ktorý vidí praktické opodstatnenie aj využitie vo vytvorení skupín podľa žiadneho, nezávažného a závažného nálezu.

5.4.1.1. Tri cieľové triedy

Pri prvom experimente boli zo šiestich pôvodných tried odvodené 3 nové cieľové triedy. Pri tomto rozdelení bolo prihliadané na početnosť príkladov patriacich do jednotlivých tried (str. 48 Obr. 14). V snahe vyrovnať ich nevyváženosť bolo rozdelenie realizované spôsobom uvedeným v Tab. 5. Početnosti novovytvorených tried môžeme vidieť na Obr. 17:

Tab. 5 Odvodenie 3 cieľových tried

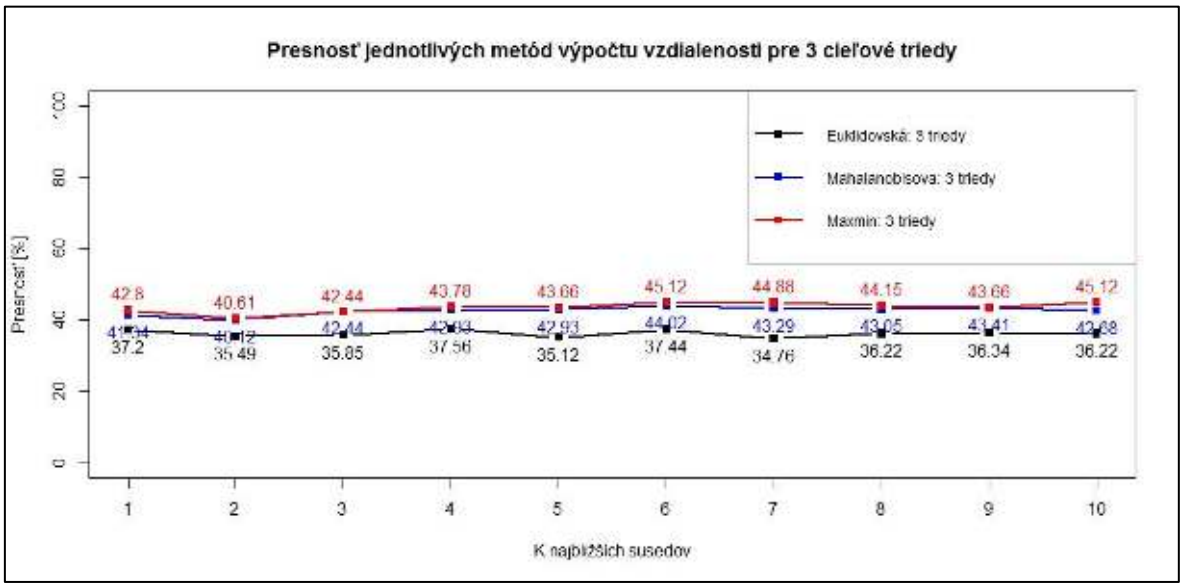
Názov atribútu	Pôvodná hodnota	Nová hodnota
Nalez	0	0
	1	1
	2	
	3	
	4	2
	5	



Obr. 17 Graf početností pacientov pre jednotlivé hodnoty atribútu Nález: 3 cieľové triedy

Toto nové rozdelenie pacientov do troch tried korešponduje taktiež so závažnosťou ich koronarografického nálezu, keďže u pacientov v prvej skupine nebol zistený žiadny nález a sú považovaní za zdravých z kardiovaskulárneho hľadiska. U pacientov v druhej skupine bol síce zaznamenaný pozitívny nález, ale iba nezávažného charakteru. Naopak, do tretej skupiny pacientov sú zaradení tí, ktorých koronarografický nález je už považovaný za vážny.

Pri novom rozdelení cieľového atribútu boli opätovne otestované vybrané metódy predikcie. Každá z týchto metód dosiahla vyššiu odhadovanú presnosť v porovnaní s pôvodným rozdelením cieľového atribútu v priemere o 4,7%.



Obr. 18 Porovnanie odhadovanej kvality metód výpočtu vzdialenosti pre 3 cieľové triedy

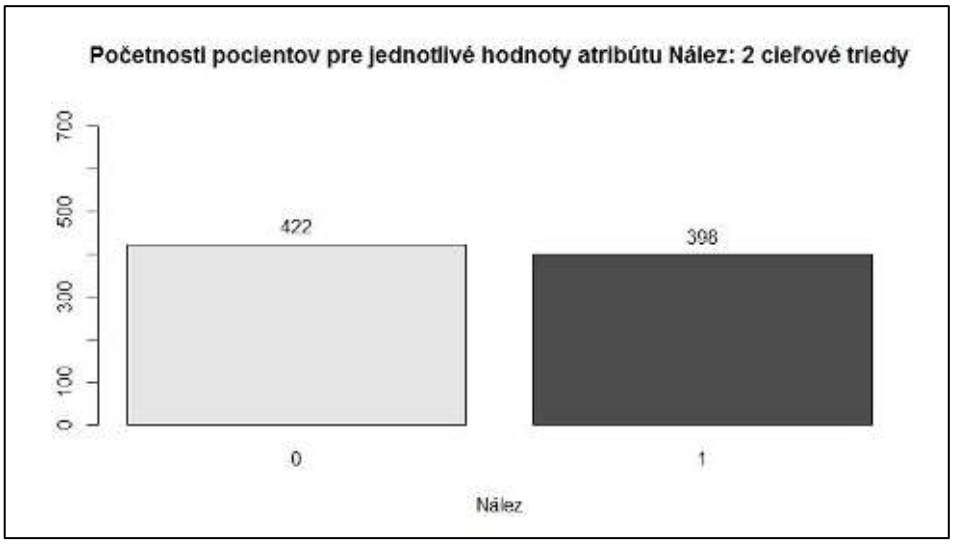
Ak by sme chceli dosiahnuté výsledky znova porovnať s baseline modelom, tento by klasifikoval všetkých symptomatických pacientov do triedy 2. To znamená, že za týchto podmienok by bola jeho presnosť 35.86%. Pri porovnaní s najvyššími priemernými presnosťami navrhovaných modelov na základe výsledkov na Obr. 18, jeho presnosť tentokrát zaostáva o 1.7% až 9.26%.

5.4.1.2. Dve cieľové triedy

Pri ďalšom experimente boli zo šiestich pôvodných tried odvodené 2 nové cieľové triedy, a to spôsobom uvedeným v Tab. 6. Z hľadiska početností jednotlivých cieľových tried je toto rozdelenie najvyrovnanejšie (Obr. 19) v porovnaní s tromi alebo pôvodnými šiestimi triedami. Trieda 0 v tomto prípade reprezentuje všetkých pacientov, u ktorých bol koronarografický nález negatívny, respektíve pozitívny, avšak nie závažný. Teda tých, u ktorých momentálne nie je nevyhnutné vykonať koronarografické vyšetrenie. Trieda 1 naopak združuje tie prípady, u ktorých sa vyskytol závažnejší nález, a teda je u nich nutné vykonať koronárnu angiografiu. Vzhľadom na odporúčací charakter tohto systému je rozdelenie na dve triedy úplne postačujúce a dáva kardiológom relevantné vstupné informácie pre ich finálne rozhodnutie ohľadom indikácie spomínaného vyšetrenia.

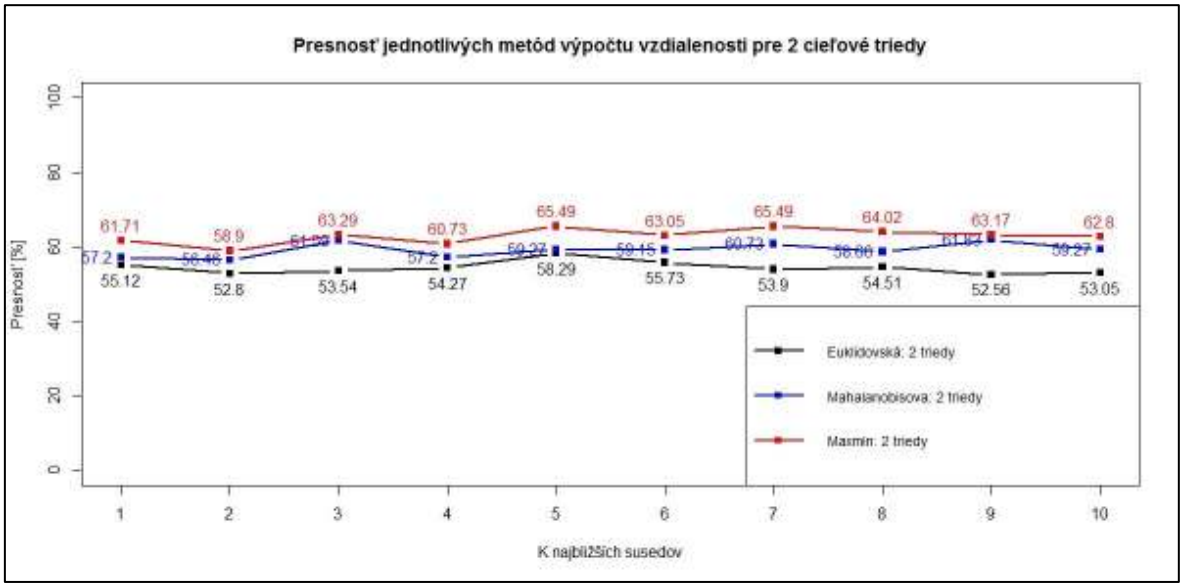
Tab. 6 Odvodenie 2 cieľových tried

Názov atribútu	Pôvodná hodnota	Nová hodnota
Nalez	0	0
	1	
	2	
	3	1
	4	
	5	

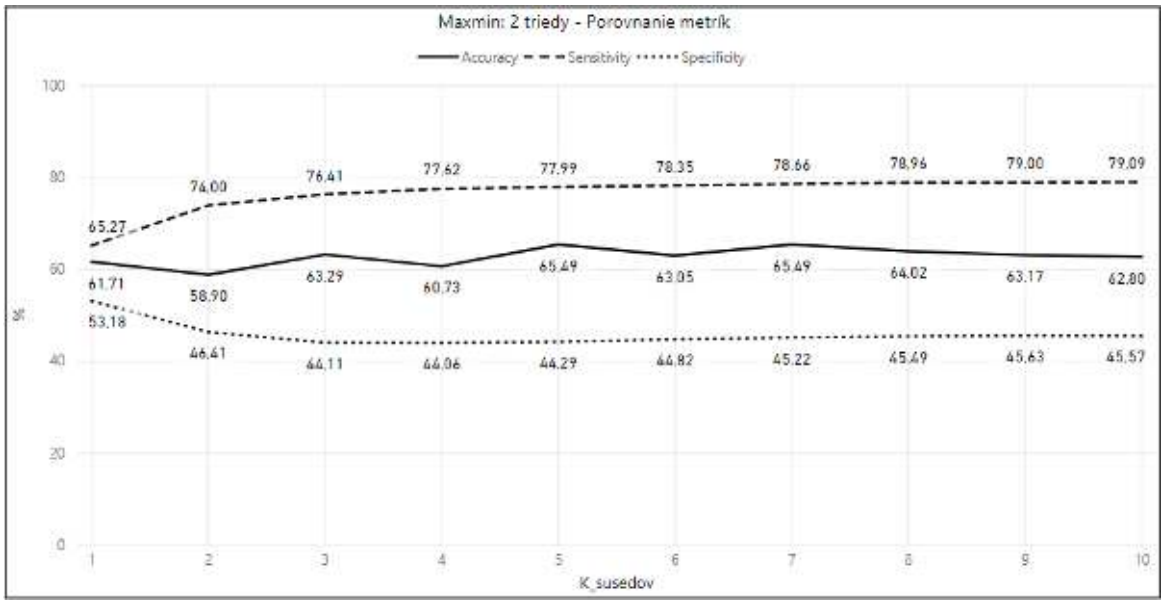


Obr. 19 Graf početností pacientov pre jednotlivé hodnoty atribútu Nález: 2 cieľové triedy

Výsledky tohto experimentu sú zobrazené na Obr. 20. Najvyššiu priemernú odhadovanú presnosť pri predikcii cieľovej triedy dosiahla metóda Maxmin, a to na úrovni 65.49% pri počte najbližších susedov 5 a 7. Pre túto metódu boli vypočítané aj ďalšie metriky na odhad kvality klasifikátora ako senzitivita a špecificita, ktorých výsledky sú vyobrazené na Obr. 21. Najvyššia priemerná dosiahnutá senzitivita dosiahla úroveň 79.09% pri 10 najbližších susedoch a najvyššia priemerná špecificita dosahovala hodnotu 53.18% pri iba jednom susedovi.



Obr. 20 Porovnanie kvality metód výpočtu vzdialenosti pre 2 cieľové triedy



Obr. 21 Porovnanie metrík odhadu kvality metódy Maxmin

Ak by sme opäť chceli porovnať dosiahnuté výsledky s baseline modelom, v tomto prípade by klasifikoval všetkých pacientov vykazujúcich symptómy do triedy 1. Za takýchto okolností by dosiahnutá presnosť bola na úrovni 48.54%. To značí, že najvyššie priemerné presnosti navrhovaných modelov podľa výsledkov na Obr. 20 sú vyššie o 9.75% až 16.95%.

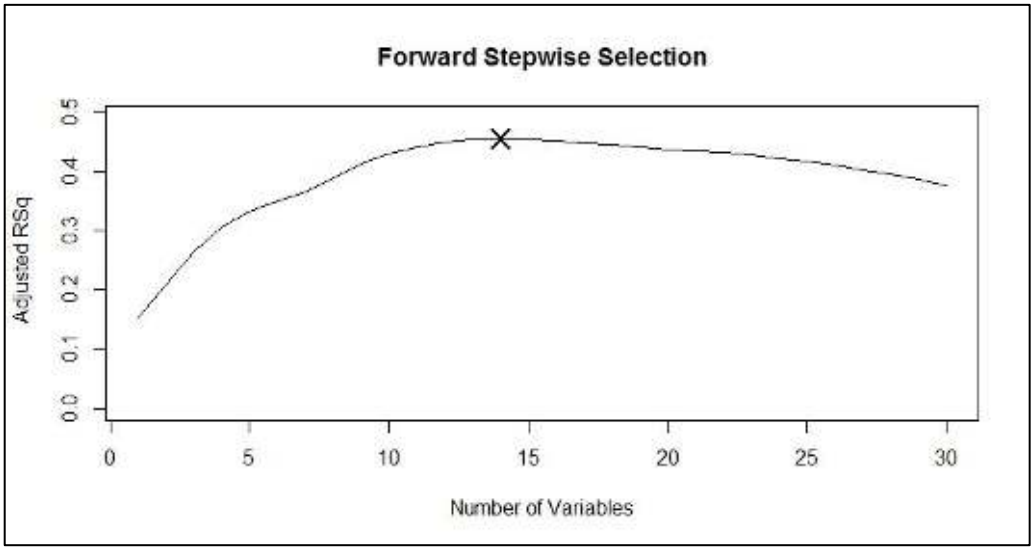
5.4.2. Experimentálny výber podmnožiny atribútov

Výber podmnožiny atribútov bol realizovaný na základe algoritmu Forward Stepwise Selection [24]. Tento algoritmus má na začiatku model obsahujúci nula prediktorov a postupne doň pridáva v každej iterácii jeden nový prediktor. Svoj výber uskutočňuje podľa toho, ktorý z ostávajúcich atribútov najviac prispeje k zníženiu RSS (Reziduálneho súčtu štvorcov), a teda k vylepšeniu budovaného modelu.

Optimálny počet atribútov v modeli bol stanovený podľa ukazovateľa Adjusted R^2 (korigovaného koeficientu determinácie). Jeho hodnota sa znižuje, ak sú do modelu zahrnuté nepotrebné atribúty. Z tohto dôvodu bol vybraný model s najvyšším AdjRSq na úrovni 0.4544771 (Obr. 22), ktorý obsahoval nasledovných 14 atribútov:

1. Rocnik
2. Pohlavie
3. R_ICHS
4. R_IM
5. R_Hypertenzia
6. O_IM
7. O_Hypertenzia
8. Chol
9. HDL

10. EKG_FR
11. EKG_QRS
12. EKG_SVES
13. EKG_STE
14. ECHO_EF



Obr. 22 Výsledok Forward Stepwise Selection

Na tejto podmnožine 14 atribútov bolo opäť vykonané testovanie pre jednu z najúspešnejších metód výpočtu vzdialenosti – Maxmin. V Tab. 7 je zobrazené porovnanie priemerných presností 10-násobnej krížovej validácie pre jednotlivé počty najbližších susedov spolu s ich percentuálnym rozdielom oproti testovaniu s využitím všetkých dôležitých atribútov. Najvyšší percentuálny rozdiel bol síce iba na úrovni 1.47%, ale keďže v 7 z 10 prípadov bol zaznamenaný nárast presnosti, budeme tento experiment realizovať aj pre ostatné rozdelenia cieľového atribútu.

Tab. 7 Priemerné presnosti metódy Maxmin pre 6 tried na podmnožine podľa FSS

K susedov	Priemerná presnosť pre 6 cieľových tried v %		Rozdiel v %
	bez výberu atribútov	iba vybrané atribúty	
1	32.07	31.83	-0.24
2	35.12	35.85	0.73
3	35.00	35.85	0.85
4	35.98	36.34	0.36
5	37.44	37.20	-0.24
6	37.56	36.46	-1.10
7	37.93	38.54	0.61
8	38.90	39.76	0.86
9	39.02	**40.49**	**1.47**
10	**39.39**	39.63	0.24

5.4.2.1. Tri cieľové triedy

V tomto experimente bola opäť porovnávaná presnosť metódy Maxmin na vybranej podmnožine 14 atribútov s presnosťou na množine dôležitých atribútov, avšak cieľový atribút bol rozdelený do troch skupín rovnakým spôsobom ako v kapitole 5.4.1.1. V tomto prípade boli všetky z porovnávaných presností vyššie. Rozdiely medzi jednotlivými presnosťami boli tentokrát výraznejšie, pričom najvyšší rozdiel dosahoval až 7% (Tab. 8).

Tab. 8 Priemerné presnosti metódy Maxmin pre 3 triedy na podmnožine podľa FSS

K susedov	Priemerná presnosť pre 3 cieľové triedy v %		Rozdiel v %
	bez výberu atribútov	iba vybrané atribúty	
1	42.80	43.78	0.98
2	40.61	44.27	3.66
3	42.44	47.44	5.00
4	43.78	47.07	3.29
5	43.66	48.54	4.88
6	**45.12**	48.41	3.29
7	44.88	49.76	4.88
8	44.15	49.27	5.12
9	43.66	**50.73**	**7.07**
10	**45.12**	50.61	5.49

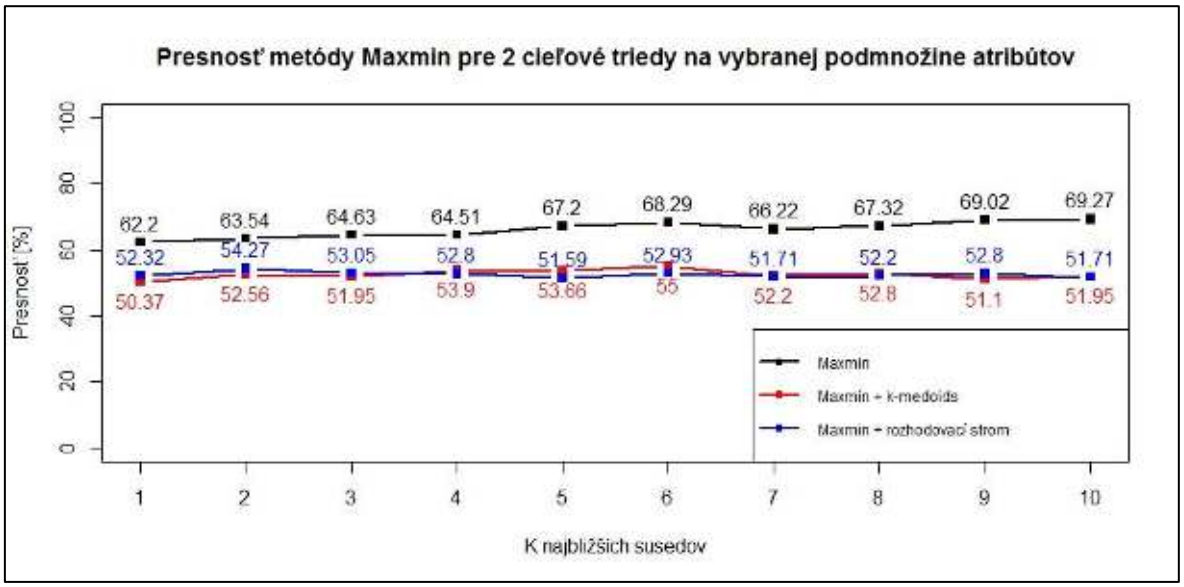
5.4.2.2. Dve cieľové triedy

V tomto prípade boli porovnávané presnosti metódy Maxmin pri rozdelení cieľového atribútu na dve triedy podľa kapitoly 5.4.1.2, pričom opäť boli porovnávané presnosti bez výberu atribútov a presnosti s výberom atribútov podľa Forward Stepwise Selection. Všetky rozdiely medzi presnosťami boli znova kladné ako pri rozdelení do troch tried a najvyšší rozdiel dosahoval výšku 6.47%. Pôvodná najvyššia priemerná presnosť sa teda zvýšila z 65.49% dosahovaných pri 5 a 7 susedoch na 69.27% pri 10 najbližších susedoch (Tab. 9).

Tab. 9 Priemerné presnosti metódy Maxmin pre 2 triedy na podmnožine podľa FSS

K susedov	Priemerná presnosť pre 2 cieľové triedy v %		Rozdiel v %
	bez výberu atribútov	iba vybrané atribúty	
1	61.71	62.20	0.49
2	58.90	63.54	4.64
3	63.29	64.63	1.34
4	60.73	64.51	3.78
5	**65.49**	67.20	1.71
6	63.05	68.29	5.24
7	**65.49**	66.22	0.73
8	64.02	67.32	3.30
9	63.17	69.02	5.85
10	62.80	**69.27**	**6.47**

Pre overenie predpokladu, že metódy výpočtu vzdialenosti dosahujú vyššie hodnoty presnosti bez dodatočnej filtrácie záznamov bola metóda Maxmin opätovne otestovaná aj v kombinácii s rozhodovacím stromom a zhlukovaním k-medoids. Tento predpoklad bol potvrdený, keďže najvyššia presnosť pri testovaní metódy Maxmin na vybranej množine atribútov (69.7%) bola dosiahnutá pri nastavení 10 najbližších susedov bez dodatočnej filtrácie záznamov (Obr. 23).



Obr. 23 Presnosť metódy Maxmin pre 2 cieľové triedy na podmnožine podľa FSS

5.5. Vyhodnotenie

Pri porovnávaní výsledkov získaných vo fáze modelovania sa ukázalo, že experimenty ohľadom redukcie počtu cieľových tried vedú v každom prípade (Tab. 10 a Tab. 11) k zvýšeniu najvyššej priemernej presnosti u všetkých testovaných metód.

Tab. 10 Porovnanie najvyšších priemerných presností pre všetky dôležité atribúty

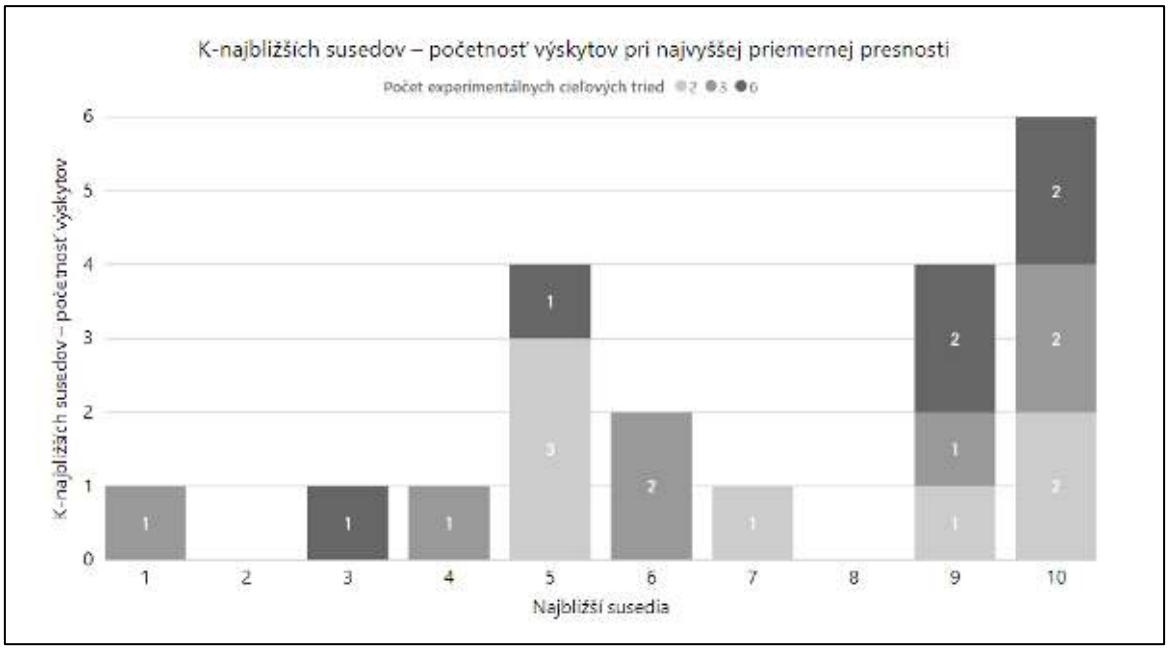
Metóda výpočtu vzdialenosti	6 cieľových tried		3 cieľové triedy		2 cieľové triedy	
	Najvyššia priemerná presnosť	K susedov	Najvyššia priemerná presnosť	K susedov	Najvyššia priemerná presnosť	K susedov
Euklidovská	34.02%	5	37.56%	4	58.29%	5
Mahalanobisova	39.27%	9	44.02%	6	61.83%	9
Maxmin	**39.39%**	10	**45.12%**	6 a 10	**65.49%**	5 a 7

Po porovnaní výsledkov rovnakých metód výpočtu vzdialenosti pre množinu všetkých dôležitých atribútov (Tab. 10) a iba jej podmnožinu vybranú pomocou FSS (Tab. 11) sa ukázalo, že metódy Mahalanobisova vzdialenosť a Maxmin dosiahli vždy vyššiu najvyššiu priemernú presnosť pri použití na vybranej podmnožine atribútov prostredníctvom FSS. Tento výsledok je pravdepodobne z časti zapríčinený aj tým, že do výpočtu zohľadňujúceho menší počet atribútov je zahrnutých viac kompletných záznamov (bez chýbajúcich hodnôt na potrebných miestach). Tieto dve metódy dosahovali veľmi podobné výsledky za použitia pri 6 a 3 cieľových triedach, badateľnejší rozdiel nastal až pri 2 cieľových triedach, kedy bol výsledok metódy Maxmin vyšší o viac ako 2% pri hodnote 69.27%.

Tab. 11 Porovnanie najvyšších priemerných presností s výberom atribútov podľa FSS

Metóda výpočtu vzdialenosti	6 cieľových tried		3 cieľové triedy		2 cieľové triedy	
	Najvyššia priemerná presnosť	K susedov	Najvyššia priemerná presnosť	K susedov	Najvyššia priemerná presnosť	K susedov
Euklidovská	32.56%	3	41.31%	1	56.83%	5
Mahalanobisova	**42.20%**	10	50.12%	10	66.95%	10
Maxmin	40.49%	9	**50.73%**	9	**69.27%**	10

Pri hľadaní najvyššej priemernej presnosti rôznych metód výpočtu *k*-najbližších susedov bol sledovaný aj ich počet, a to z dôvodu nájdenia optimálnej hodnoty parametra *k*. Na nasledovnom obrázku (Obr. 24) sú zobrazené početnosti výskytov jednotlivých *k* pri najvyššej priemernej presnosti Euklidovskej vzdialenosti, Mahalanobisovej vzdialenosti a metódy Maxmin z dvoch predošlých tabuliek (Tab. 10 a Tab. 11). Na základe týchto experimentov sa ukazuje, že najčastejšie vyšiel najvyšší priemerný výsledok pri počte susedov 10. Druhý najpočetnejší výskyt bol zaznamenaný pri počtoch susedov 5 a 9. Výskyt ostatných hodnôt *k* bol skôr ojedinelý. Z toho vyplýva, že ak si používateľ v aplikácii zvolí parameter *k* v hodnote 10, 9 alebo 5, má najväčšiu šancu na dosiahnutie najpresnejšieho výsledku predikcie koronarografického nálezu.



Obr. 24 Početnosti *k* susedov pri najvyšších priemerných presnostiach experimentov

Zhrnutie záverov vyplývajúcich z predošlých experimentov, ktoré vedú k dosiahnutiu najvyššej presnosti pri predikcii závažnosti koronarografického nálezu:

1. Metóda výpočtu vzdialenosti: Maxmin.
2. Metóda filtrácie: Žiadna.
3. Počet najbližších susedov: 10, prípadne 9 alebo 5.
4. Počet cieľových tried: 2.
5. Množina atribútov: výber podľa FSS.

Záver

Zámerom predloženej diplomovej práce bolo navrhnúť a implementovať systém na podporu rozhodovania, ktorý by mohol slúžiť kardiológom ako pomoc pri rozhodovaní o vážnosti stavu upchatia srdcových tepien pacienta, a teda nutnosti jeho odporučenia na podrobenie sa koronarografickému vyšetreniu.

Systém na podporu rozhodovania bol navrhnutý na základe identifikovaných používateľských požiadaviek zakomponovaných do vlastnej aplikácie využívajúcej princípy metodiky prípadového usudzovania, ktorá pre svoje fungovanie primárne využíva algoritmus strojového učenia zvaný k-najbližších susedov. Ďalšou možnosťou bola aj jeho kombinácia s ďalšími algoritmami strojového učenia akými sú rozhodovacie stromy alebo zhlukovanie.

Implementácia systému bola realizovaná prostredníctvom jazyka R, a to za využitia balíka Shiny. Hlavným prínosom aplikácie je zobrazenie najpodobnejších známych prípadov k ľubovoľnému novému pacientovi, o ktorom sú do aplikácie zadané aspoň základné informácie. Používateľ si taktiež môže meniť parametre tohto hľadania, a to počet hľadaných prípadov a metódu výpočtu na ich nájdenie. K vedľajším funkcionalitám patria rôzne interaktívne vizualizácie dát v podobe tabuliek a grafov, ako aj ich možný export. Z dôvodu práce s citlivými údajmi o reálnych kardiologických pacientoch táto aplikácia nie je voľne dostupná pre verejnosť. Článok popisujúci návrh a implementáciu tohto systému bude publikovaný v zborníku z medzinárodnej konferencie MIE 2020 (30th Medical Informatics Europe conference) [25].

Na základe vykonaných experimentov bolo zistené, že systém dosahuje najvyššiu presnosť klasifikácie (69%) pomocou metódy Maxmin pri rozdelení cieľového atribútu na 2 cieľové triedy za využitia vybraných atribútov pomocou FSS, ku ktorým patrili: vek, pohlavie, hladina cholesterolu, výskyt hypertenzie alebo ischemickej choroby u pacienta alebo jeho rodinných príslušníkov a výsledky EKG vyšetrenia.

V budúcnosti, za predpokladu vyzbierania ďalších dát pochádzajúcich od nových pacientov a ich následného nahratia do vyvinutej aplikácie, by mohol tento systém na podporu rozhodovania prinášať presnejšie výsledky a stať sa tak spoľahlivým pomocným nástrojom pre kardiológov.

Zoznam použitej literatúry

[1]. Our World in Data. 2017. [citované 21.4.2020]. [Online]. Dostupné na internete: <https://ourworldindata.org/grapher/share-of-deaths-by-cause?time=..2017>.

[2]. POWER, D. J. 2002. Decision Support Systems: Concepts and Resources for Managers. Quorum Books. 251s. ISBN 9781567204971.

[3]. JAO, C. 2011. Efficient Decision Support Systems - Practice and Challenges From Current to Future. IntechOpen. 558s. ISBN 978-953-307-326-2.

[4]. BERNER, E. S. 2016. Clinical Decision Support Systems Theory and Practice. Springer International Publishing. 313s. ISBN 978-3-319-31911-7.

[5]. BEGUM, S. et al. 2011. Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments. In: Ieee Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, roč. 41, č.4, s. 421-434.

[6]. CHOUDHURY, N., BEGUM, S. 2016. A Survey on Case-based Reasoning in Medicine. In: International Journal of Advanced Computer Science and Applications, roč. 7, č. 8, s. 136-144.

[7]. KOLODNER, J. L. 1992. An Introduction to Case-Based Reasoning. In: Artificial Intelligence Review, roč. 6, s. 3-34.

[8]. PARALIČ, J. 2003. Objavovanie znalostí v databázach. 1. vydanie. Elfa s.r.o.. 80s. [Online]. Dostupné na internete: <http://people.tuke.sk/jan.paralic/knihy/ObjavovanieZnalostivDB.pdf>

[9]. HARRISON, O. 2018. Machine Learning Basics with the K-Nearest Neighbors Algorithm. [Online]. Publikované 10.9.2018. [citované 26.2.2020]. Dostupné na internete: <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>.

[10]. SHARMA, N. 2019. Importance of Distance Metrics in Machine Learning Modelling. [Online]. Publikované 13.1.2019. [citované 14.3.2020]. Dostupné na internete: <https://towardsdatascience.com/importance-of-distance-metrics-in-machine-learning-modelling-e51395ffe60d>.

[11]. KASSAMBARA, A. 2018. DataNovia: K-Medoids in R: Algorithm and Practical Examples, 2018. [citované 20.3.2020]. [Online]. Dostupné na internete: <https://www.datanovia.com/en/lessons/k-medoids-in-r-algorithm-and-practical-examples/>.

[12]. SARAIVA, R. M. et al. 2015. A Hybrid Approach Using Case-Based Reasoning and Rule-Based Reasoning to Support Cancer Diagnosis: A Pilot Study. In: Studies in Health Technology and Informatics, č. 216, s. 862–866.

[13]. HUANG, M.-J. et al. 2007. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. In: Expert Systems with Applications, roč. 32, č.3, s. 856-867 .

[14]. CHUANG, CH.-L. 2011. Case-based reasoning support for liver disease diagnosis In: Artificial Intelligence in Medicine, roč. 53, č. 1, s. 15-23.

[15]. SALEM, A.-B. M., ROUSHDY, M., HODHOD, R. A. 2005. A Case Based Expert System for Supporting Diagnosis of Heart Diseases. In: ICGST International Journal of Artificial Intelligence and Machine Learning AIML, roč. 5, č. 1, s. 33-39.

[16]. STAŠKO, P. 2017. Národný portál zdravia: Najčastejšie zomierame na srdcovocievne choroby. Koho ohrozujú najviac? [Online]. Publikované 5.5.2016. Aktualizované 20.1.2017. [citované 3.6.2019]. Dostupné na internete: <https://www.npz.sk/sites/npz/Stranky/NpzArticles/2013_06/Prevencia_srdcovocievnych_ochoreni.aspx?did=3&sdid=25&tuid=0&page=full&>.

[17]. Slovenská nadácia srdca. Cholesterol. [Online]. [citované 4.6.2019]. Dostupné na internete: <http://www.tvojesrdce.sk/article/195--cholesterol>.

[18]. VARGOVÁ, V., FEDAČKO, J. 2016. Ischemická choroba srdca. [Online]. Publikované 5.2.2016. Aktualizované 2.6.2016. [citované 5.6.2019]. Dostupné na internete: <https://www.npz.sk/sites/npz/Stranky/NpzArticles/2013_06/Ischemicka_choroba_srdca.aspx?did=4&sdid=31&tuid=0&page=full&>.

[19]. Únia pre zdravšie srdce. 2013. Srdcovocievne ochorenia: Cievna mozgová príhoda. [Online]. [citované 5.6.2019]. Dostupné na internete: <https://www.presrdce.eu/srdcovocievne-ochorenia/cievna-mozgova-prihoda>.

[20]. KLINOVSKÝ, J. 2005. Aortálna stenóza. [Online]. Publikované 12.5.2005. [citované 5.6.2019]. Dostupné na internete: <https://www.zdravie.sk/choroba/23221/aortalna-stenoza>.

[21]. Stredoslovenský ústav srdcových a cievnych chorôb, a.s.. Koronarografické vyšetrenie. [Online]. [citované 10.3.2020]. Dostupné na internete: <https://www.suscch.eu/page.php?63>.

[22]. Národný ústav srdcových a cievnych chorôb, a. s.. Informácie o koronarografii. [Online]. [citované 10.3.2020]. Dostupné na internete: <https://www.nusch.sk/sk/informacie-o-koronarografii>.

[23]. European Society of Cardiology. SCORE Risk Charts. [Online]. [citované 15.2.2020]. Dostupné na internete: <https://www.escardio.org/Education/Practice-Tools/CVD-prevention-toolbox/SCORE-Risk-Charts>.

[24]. JAMES, G. et al. 2017. An Introduction to Statistical Learning. 7. vydanie. Springer. 440s. ISBN 978-1-4614-7138-7.

[25]. The European Federation of Medical Informatics. 2020. [Online]. [citované 27.4.2020]. Dostupné na internete: <https://mie2020.org/en/>

Prílohy

Príloha A: CD médium

Príloha B: Používateľská príručka

Príloha C: Systémová príručka

TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

PRÍPADOVÉ USUDZOVANIE PRI DIAGNOSTIKE KARDIOVASKULÁRNYCH OCHORENÍ
Používateľská príručka

Študijný program:	Hospodárska informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra kybernetiky a umelej inteligencie (KKUI)
Školiteľ:	prof. Ing. Ján Paralič, PhD.
Konzultant:	Ing. Ľudmila Pusztová

2020 Košice **Zuzana Tocimáková, Bc.**

Obsah

Zoznam obrázkov 70

1. Návod na spustenie a použitie aplikácie 71

1.1. Spustenie aplikácie v prostredí RStudio 71

1.2. Návod na použitie aplikácie 72

1.2.1. Karta NOVÝ PRÍPAD 72

1.2.2. Karta DÁTA 73

1.2.3. Karta SKUPINY PACIENTOV 74

1.2.4. Karta VIZUALIZÁCIA 75

1.2.5. Karta PCA – ANALÝZA PRINCIPIÁLNYCH ZLOŽIEK 76

Zoznam obrázkov

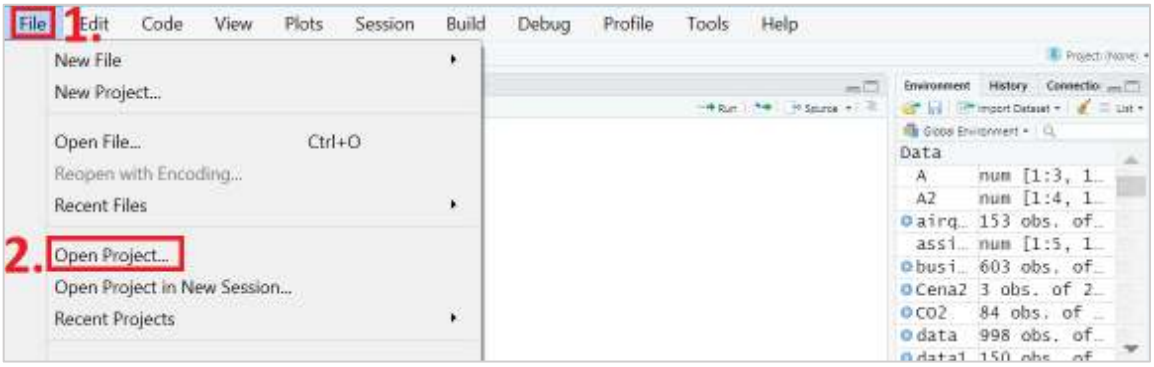
Obr. 1 Otvorenie projektu 71
Obr. 2 Inštalácia potrebných R balíkov a spustenie aplikácie 71
Obr. 3 Karta NOVÝ PRÍPAD 73
Obr. 4 Karta DÁTA 74
Obr. 5 Karta SKUPINY PACIENTOV 75
Obr. 6 Karta VIZUALIZÁCIA 76
Obr. 7 Karta PCA - ANALÝZA PRINCIPIÁLNYCH ZLOŽIEK 77

1. Návod na spustenie a použitie aplikácie

Praktickým výstupom tejto diplomovej práce je RShiny aplikácia vytvorená pomocou programovacieho jazyka R verzie 3.6.1 vo vývojovom prostredí RStudio verzie 1.2.1335. Jednotlivé časti aplikácie sú uložené v skriptoch s príponou .r v priečinku samotnej RShiny aplikácie. V nasledujúcej časti sa nachádza návod ako v tomto prostredí môžeme spustiť a používať túto aplikáciu, ktorá predstavuje systém na podporu rozhodovania kardiológov.

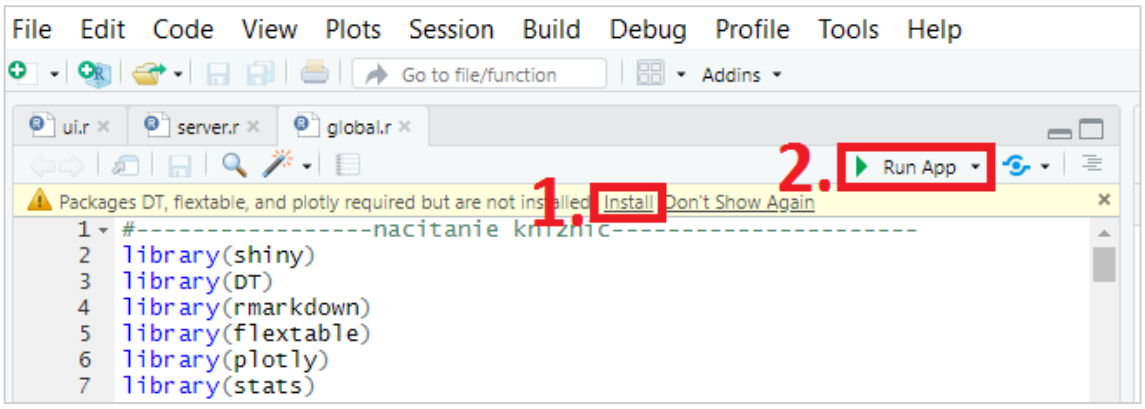
1.1. Spustenie aplikácie v prostredí RStudio

Otvoríme si prostredie RStudio a klikneme v ľavom hornom rohu na položku *File* (pozri 1 Obr. 1). Zobrazí sa nám zoznam, v ktorom klikneme na položku *Open Project...*(pozri 2 Obr. 1). Následne vyberieme projekt s názvom „System_na_podporu_rozhodovania_kardiologov“, v ktorom je uložená táto diplomová práca.



Obr. 1 Otvorenie projektu

Pri prvom otvorení aplikácie je nutné nainštalovať všetky R balíky, ktoré sú potrebné na spustenie aplikácie. Urobíme tak kliknutím na tlačidlo *Install* (pozri 1 Obr. 2). Po dokončení inštalácie je potrebné kliknúť na tlačidlo *Run App* (pozri 2 Obr. 2), čím sa aplikácia spustí.



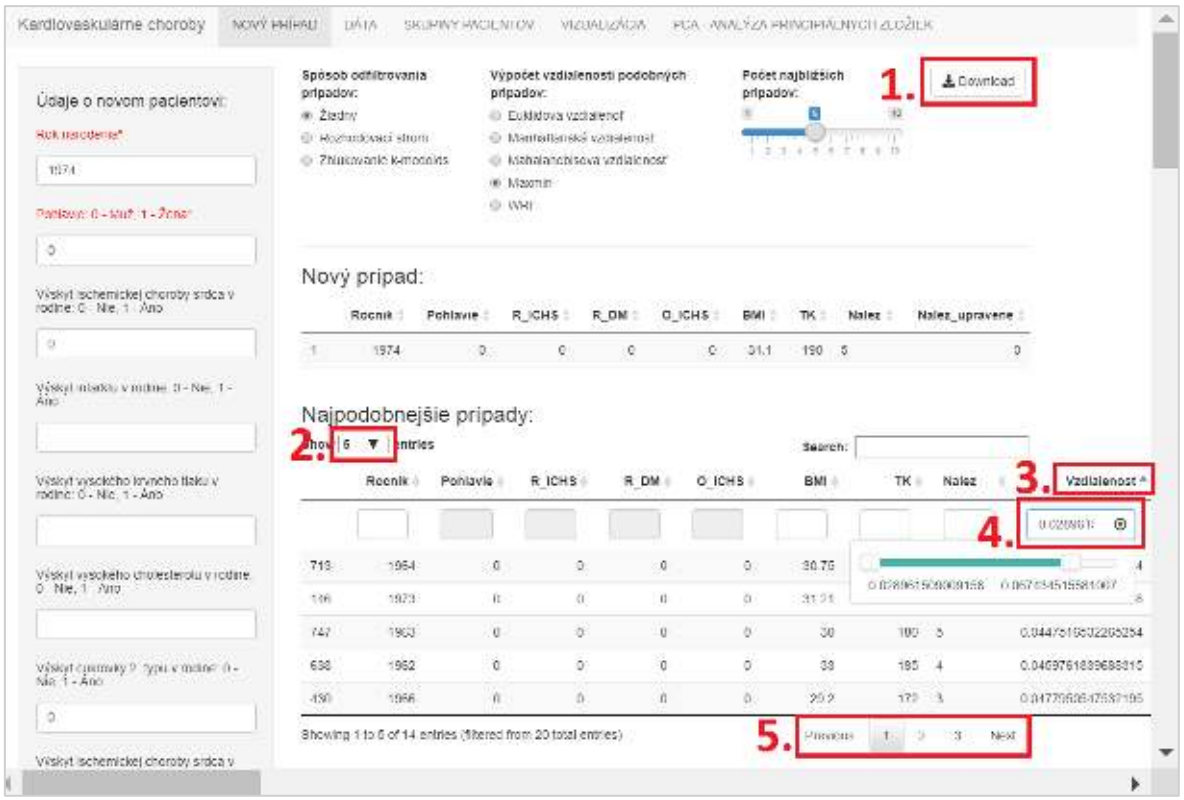
Obr. 2 Inštalácia potrebných R balíkov a spustenie aplikácie

1.2. Návod na použitie aplikácie

Na základe tohto návodu by mal používateľ zistiť na čo slúži táto aplikácia a mal by byť schopný sa v nej orientovať. Aplikácia obsahuje 5 rôznych kariet a pre navigáciu medzi jednotlivými kartami aplikácie slúži menu v hornej časti obrazovky (Obr. 3).

1.2.1. Karta NOVÝ PRÍPAD

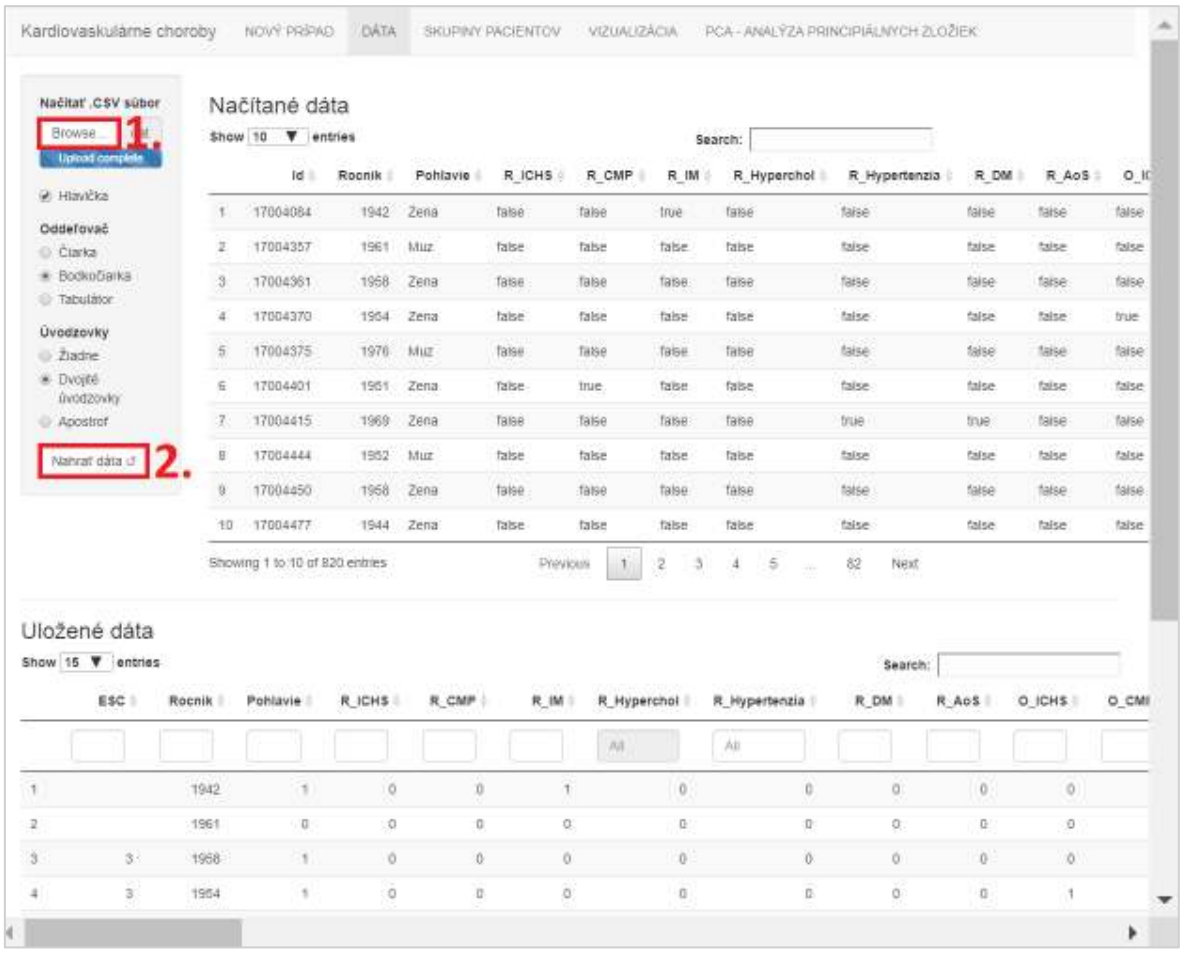
Po spustení aplikácie sa zobrazí prvá karta s názvom NOVÝ PRÍPAD. Na ľavej strane obrazovky sa v bočnom paneli nachádzajú polia pre zadávanie vstupných hodnôt jednotlivých atribútov o novom pacientovi (Obr. 3). Po vyplnení povinných atribútov, ktorými sú *Rocnik* a *Pohlavie*, je automaticky spustený algoritmus hľadania podobných prípadov. V hornej časti obrazovky je možné nastaviť rôzne parametre tohto algoritmu. Prvým možným nastavením je *Spôsob odfiltrovania prípadov s možnosťami*: žiadny, rozhodovací strom a zhlukovanie k-medoids. Ďalším nastaviteľným parametrom je *Výpočet vzdialenosti podobných prípadov*, kde je k dispozícii 5 možností: Euklidova vzdialenosť, Manhattanská vzdialenosť, Mahalanobisova vzdialenosť, Maxmin a WRF. Posledným modifikovateľným parametrom je *Počet najbližších prípadov*, kde sú na výber čísla od 1 do 10. V pravom hornom rohu obrazovky (pozri 1 Obr. 3) sa nachádza tlačidlo *Download*, ktoré slúži na exportovanie údajov o aktuálnom pacientovi spolu s jemu najbližšími známymi prípadmi vo formáte .docx. V časti *Nový prípad* sú zobrazené aktuálne zadané hodnoty atribútov spolu s predikovanou hodnotou Nálezu. V sekcii *Najpodobnejšie prípady* sú zobrazené údaje o zvolenom počte najpodobnejších známych prípadov. Ak chceme zobraziť viac alebo menej prípadov, v rolovacom zozname (pozri 2 Obr. 3) máme na výber z možností 5, 10 alebo 20 záznamov. Nájdené záznamy sú zoradené vzostupne podľa ich vzdialenosti k novému prípadu. V prípade potreby je možné záznamy usporiadať podľa ľubovoľného atribútu kliknutím na názov atribútu. Po prvom kliknutí sa zoradia záznamy vzostupne, po druhom zas zostupne. Smer usporiadania označuje šípka nachádzajúca sa vedľa názvu atribútu (pozri 3). V prípade, že chceme obmedziť hodnoty atribútov v tabuľke iba na konkrétny interval, je potrebné kliknúť na pole pod názvom daného atribútu (pozri 4 Obr. 3) a nastaviť ľubovoľný interval hodnôt. Pre prechod na nasledujúcu alebo predošlú stránku dátovej tabuľky slúžia tlačidlá *Previous* a *Next* nachádzajúce sa pod touto tabuľkou (pozri 5 Obr. 3).



Obr. 3 Karta NOVÝ PRÍPAD

1.2.2. Karta DÁTA

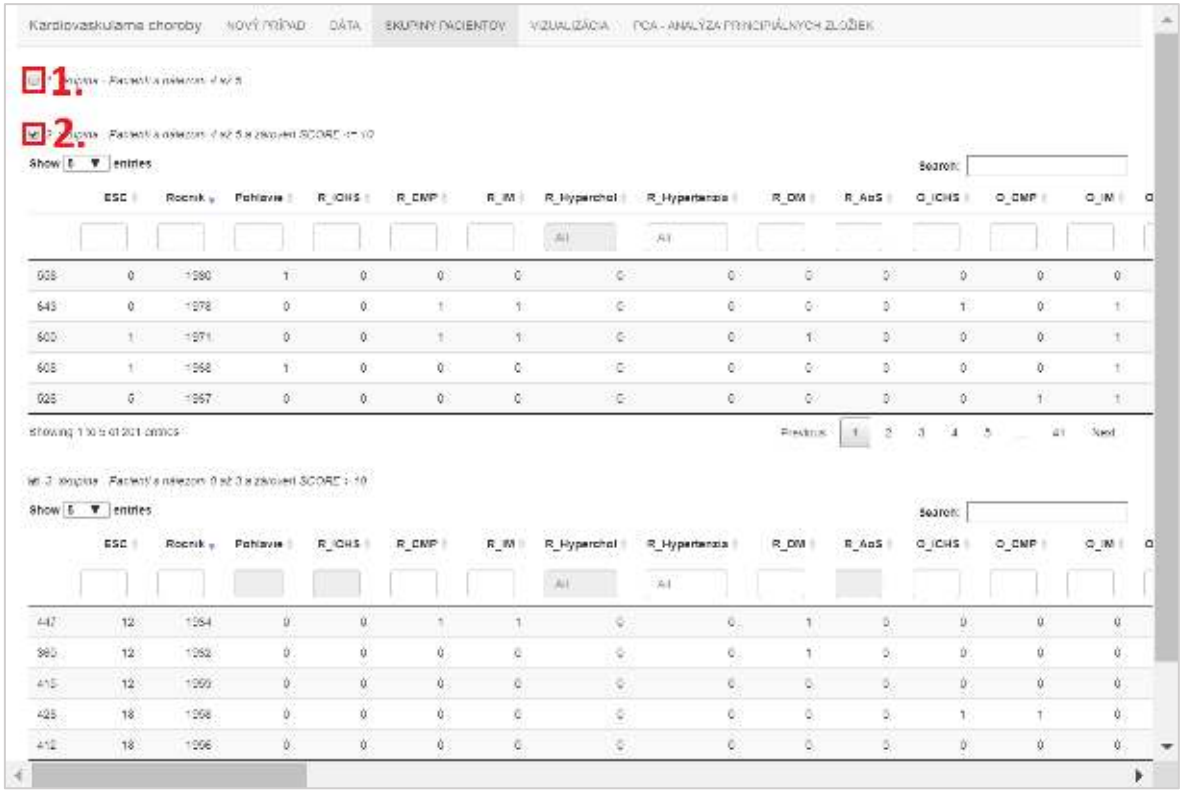
V spodnej časti nasledujúcej karty DÁTA je možné si prezrieť všetky dáta uložené v databáze (Obr. 4). Tak ako v predchádzajúcej karte, aj tu je možné dáta ľubovoľne zoradiť alebo filtrovať. Tentokrát je možné nastaviť počet zobrazených záznamov na 15, 20 alebo 50 riadkov. V ľavom hornom rohu sa nachádza panel, ktorý používateľovi umožňuje načítať do aplikácie súbor formátu .csv obsahujúci dáta o nových pacientoch po stlačení tlačidla *Browse* (pozri 1 Obr. 4). V sekcii *Načítané dáta* je poskytovaný náhľad dát nachádzajúcich sa v aktuálne vybranom súbore. V ľavom paneli sa taktiež nachádzajú nastavenia týkajúce sa štruktúry načítaného súboru, a to či obsahuje hlavičku, výber oddeľovača s možnosťami: čiarka, bodkočiarka a tabulátor, alebo nastavenie prítomnosti alebo typu úvodzoviek s možnosťami: žiadne, dvojité úvodzovky a apostrof. V prípade, že načítané dáta chceme definitívne nahrať, a teda pridať ich k už uloženým záznamom, stačí stlačiť tlačidlo *Nahrať dáta* (pozri 2 Obr. 4). Následne sa aplikácia automaticky obnoví a spustí odznova na prvej karte, pričom už obsahuje nové dáta.



Obr. 4 Karta DÁTA

1.2.3. Karta SKUPINY PACIENTOV

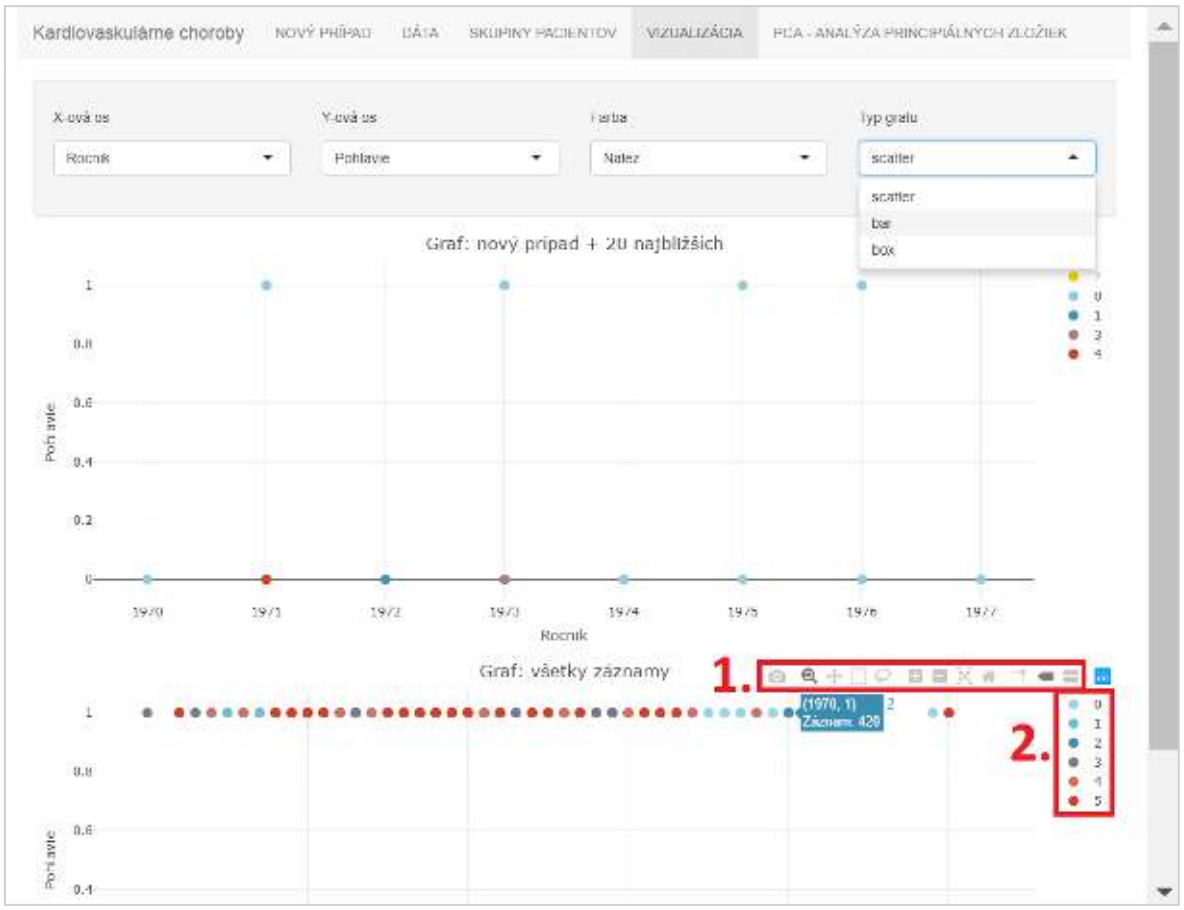
Táto karta poskytuje možnosť zobrazenia (pozri 1 Obr. 5) alebo skrytia (pozri 2 Obr. 5) dát o vybranej skupine pacientov. K dispozícii sú 4 rôzne skupiny pacientov: Pacienti s nálezom 4 až 5, Pacienti s nálezom 4 až 5 a zároveň SCORE<=10, Pacienti s nálezom 0 až 3 a zároveň SCORE > 10 a Pacienti s nálezom 0 až 3. Tieto skupiny slúžia hlavne na nájdenie a porovnanie pacientov s vysokým rizikom výskytu kardiovaskulárnej choroby podľa ich SCORE, ale zároveň s nízkym stupňom závažnosti ich koronarografického nálezu a opačne. Zobrazenie dátovej podmnožiny je opäť možné upravovať pomocou filtrovania a triedenia hodnôt atribútov alebo zmeny počtu zobrazených záznamov s možnosťami 5, 10 a 20.



Obr. 5 Karta SKUPINY PACIENTOV

1.2.4. Karta VIZUALIZÁCIA

V hornej časti karty zvanej VIZUALIZÁCIA (Obr. 6) sa nachádza panel, v ktorom je možné zvoliť nastavenia oboch grafov nachádzajúcich sa na tejto karte, a teda ktorý z atribútov bude zobrazený na osiach x a y, podľa ktorého atribútu budú dátové body rozdelené do farebných skupín, a taktiež aj typ grafu z možností: bodový, stĺpcový alebo krabicový graf. Po presune kurzora myši na plochu ktoréhokoľvek grafu sa zobrazí menu (pozri 1 Obr. 6) s možnosťami exportu grafu vo formáte .png, priblížením ľubovoľnej časti grafu alebo resetovaním na pôvodné nastavenia osí. Vedľa grafu sa nachádza aj legenda (pozri 2 Obr. 6), ktorá pri kliknutí na konkrétnu kategóriu skryje dáta patriace do nej. Opätovným kliknutím na rovnakú kategóriu sa skryté dáta znovu zobrazia. Naopak, zobrazenie iba vybranej kategórie dát je možné prostredníctvom dvojkliku na danú kategóriu v legende. V prípade, že chceme bližšie preskúmať konkrétny dátový bod na grafe, stačí naň premiestniť kurzor myši a pomocou tooltipu sa zobrazia jeho presné súradnice, identifikátor a kategória, do ktorej patrí.



Obr. 6 Karta VIZUALIZÁCIA

1.2.5. Karta PCA – ANALÝZA PRINCIPIÁLNYCH ZLOŽIEK

Posledná karta ponúka vizualizáciu vybranej skupiny atribútov znázornenú prostredníctvom dvoch hlavných komponentov analýzy principiálnych zložiek. Na prvom grafe (Obr. 7) sú vykreslené všetky záznamy z databázy spolu s údajmi zadanými v karte NOVÝ PRÍPAD. Z tohto dôvodu je prednastavený výber skupiny atribútov podľa vyplnených údajov o novom pacientovi. V druhom grafe nie je síce vykreslený nový prípad, ale sú tam všetky ostatné známe prípady, čo umožňuje ľubovoľne meniť skupinu atribútov použitých pre túto vizualizáciu. Po kliknutí na konkrétny atribút (pozri 1 Obr. 7) a následnom stlačení klávesu *backspace* alebo *delete* sa atribút z výberu odstráni. Pre opätovné pridanie odstráneného atribútu je potrebné naň opäť kliknúť v rolovacom zozname (pozri 2 Obr. 7). Pre obidva grafy platia rovnaké možnosti zobrazenia dát a exportu grafu ako aj v predchádzajúcej karte VIZUALIZÁCIA (kapitola 1.2.4).

Obr. 7 Karta PCA - ANALÝZA PRINCIPIÁLNYCH ZLOŽIEK

TECHNICKÁ UNIVERZITA V KOŠICIACH
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

PRÍPADOVÉ USUDZOVANIE PRI DIAGNOSTIKE KARDIOVASKULÁRNYCH OCHORENÍ
Systémová príručka

Študijný program:	Hospodárska informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra kybernetiky a umelej inteligencie (KKUI)
Školiteľ:	prof. Ing. Ján Paralič, PhD.
Konzultant:	Ing. Ľudmila Pusztová

2020 Košice **Zuzana Tocimáková, Bc.**

Obsah

Zoznam obrázkov 80

1. O Aplikácii 81

2. Popis hlavných funkcií 82

2.1. CBR (Case Based Reasoning) cyklus 82

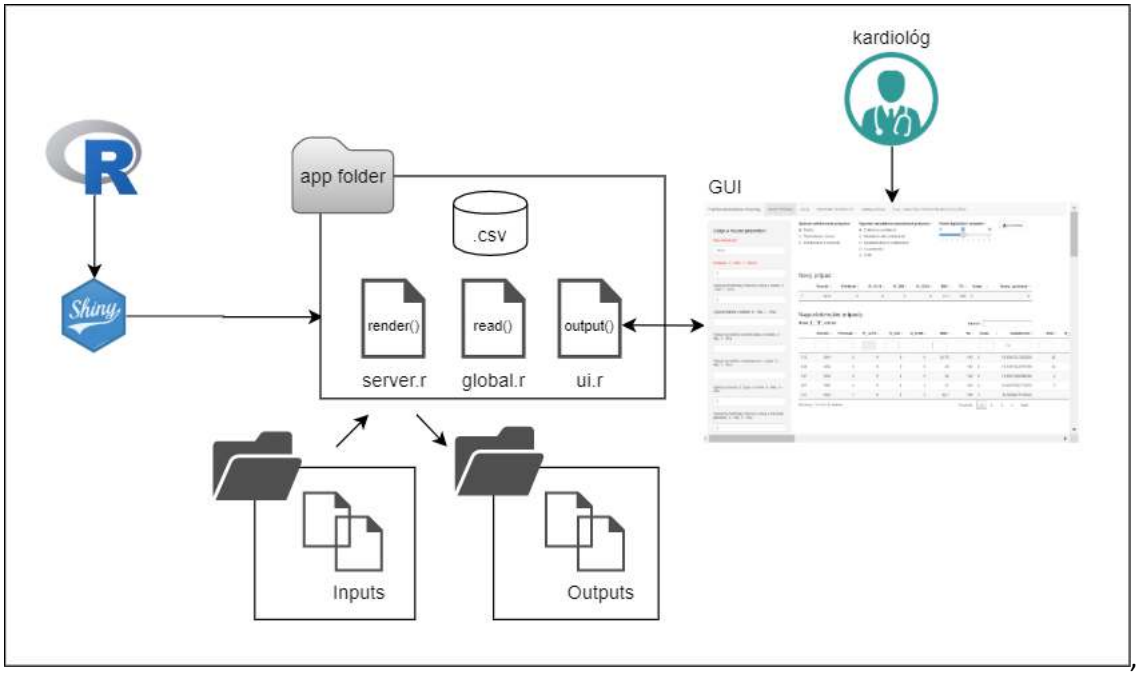
2.2. Graf PCA (Analýzy principiálnych zložiek) 86

Zoznam obrázkov

Obr. 1 Architektúra aplikácie 81

1. O Aplikácii

Praktickým výstupom tejto diplomovej práce je RShiny aplikácia predstavujúca systém na podporu rozhodovania kardiológov. Bola implementovaná prostredníctvom programovacieho jazyka R verzie 3.6.1 vo vývojovom prostredí RStudio verzie 1.2.1335. Jednotlivé časti aplikácie sú uložené v skriptoch s príponou .r v priečinku samotnej RShiny aplikácie. Spolu so skriptami sa v priečinku aplikácie nachádza aj súbor formátu .csv obsahujúci známe dáta o predošlých pacientoch. Schéma architektúry aplikácie (Obr. 1):



Obr. 1 Architektúra aplikácie

2. Popis hlavných funkcií

Táto kapitola je venovaná popisu implementácie vybraných funkcií aplikácie a zároveň aj okomentovaným ukážkam z ich zdrojového kódu.

2.1. CBR (Case Based Reasoning) cyklus

Na realizáciu CRB cyklu bola implementovaná vlastná funkcia *knn*, ktorá využíva predovšetkým princíp algoritmu k-najbližších susedov. Jej povinnými parametrami sú: *data* - dátová množina známych prípadov, *ncase* – dáta o novom pacientovi, *imp_col* – číslo stĺpca cieľového atribútu, *train_cols* – čísla stĺpcov obsahujúcich atribúty z nového prípadu. Nepovinnými parametrami s prednastavenými hodnotami sú: *k = 5* – počet najbližších susedov použitých na predikciu cieľovej triedy, *max_k = 20* – maximálny počet vrátených najbližších susedov, *revise_k = 10* - počet najbližších susedov z každej cieľovej triedy pre fázu REVISE, *method = „euclidean“* – metóda výpočtu vzdialenosti medzi známymi prípadmi a novým pacientom, *clustering_method = NULL* – metóda použitá na filtrovane prípadov. Zdrojový kód funkcie *knn*:

```
   knn <- function(data, ncase, imp_col, train_cols, k=5, max_k=20,
revise_k=10, method = "euclidean", clustering_method = NULL) {

  # dataframe obsahujúci iba záznamy bez chýbajúcich hodnôt v rámci množiny
atribútov train_cols
  data_complete <-
data[complete.cases(data[,c(train_cols,imp_col)]),c(train_cols,imp_col)]
  # dataframe obsahujúci nový záznam s neznámym Nalezom
  data_incomplete <- cbind(ncase,"Nalez" = NA)
  # počet kompletných známych prípadov
  ncomplete <- length(data_complete[,1])
  nincomplete <- length(data_incomplete[,1])
  # počet stĺpcov kompletných známych prípadov
  ncols <- ncol(data_complete)
  sum_class <- NULL

  for(j in 1:nincomplete) { # cyklus pre prechádzanie novými záznamami
    d <- numeric(ncomplete) # inicializácia vektora vzdialeností známych
záznamov
    i <- 1
    while(i <= (ncols-1)){ # cyklus pre prechádzanie cez všetky stĺpce okrem
cieľového

      if(method=="euclidean"){ # výpočet Euklidovej vzdialenosti
        d <- d + (data_complete[,i] - data_incomplete[j,i])^2
        if(i==(ncols-1)){
          d <- sqrt(d)
        }
      }
      else if (method=="manhattan") { # výpočet Manhattanskej vzdialenosti
        d <- d + abs(data_complete[,i] - data_incomplete[j,i])
      }

      else if (method=="vzdialenost") { # výpočet vzdialenosti metódou "Maxmin"
```

```
        if((max(data_complete[,i])-min(data_complete[,i]))>0){
        d <<- d + abs(data_complete[,i]-
data_incomplete[j,i])/(max(data_complete[,i])-min(data_complete[,i]))
        }
        else{
        d <<- d + abs(data_complete[,i]-data_incomplete[j,i])
        }
        if(i==(ncols-1)){
          d <<- d/i
        }
      }


      # výpočet vzdialenosti metódou WRF
      else if (method=="wrf") {

        for(m in 1:nrow(data_complete)){

          if(data_complete[m,i] >= data_incomplete[j,i]){
            if(data_complete[m,i] == 0){
              d[m] <<- d[m] + 0
            }
          else{
           d[m] <<- d[m] +  1 - (data_incomplete[j,i]/data_complete[m,i])
           }
          }
          else{
           d[m] <<- d[m] + 1 - (data_complete[m,i]/data_incomplete[j,i])
          }
        }
      }

      #odstránenie stĺpcov s variačným rozpätím 0 z výpočtu vzdialenosti
      else if (method=="mahalanobis") {
        if((max(rbind(data_complete[,i],data_incomplete[,i]))-
min(rbind(data_complete[,i],data_incomplete[,i])))<=0){
          data_complete[,i]<<-NULL
          data_incomplete[,i]<<-NULL
          i <<- i-1
          ncols <<- ncol(data_complete)
        }
      }
      i <<- i+1 # zvýšenie i o 1 pre ďalšiu iteráciu cyklu while
    }
    # koniec cyklu pre jednotlivé stĺpce

    if(method=="mahalanobis"){ #výpočet Mahalanobisovej vzdialenosti
      ncols <<- ncol(data_complete)
      d <<- mahalanobis.dist(data.x=data_complete[,1:(ncols-1)],data.y =
data_incomplete[,1:(ncols-1)])
    }

    if(clustering_method=="tree"){ # vytvorenie rozhodovacieho stromu
      tree_model <<- tree(Nalez ~ .,data = data_complete)

      tree_prediction <<- predict(tree_model, data_incomplete, type="where")

      node_neighbors <<- tree_model$where[tree_model$where == tree_prediction]
```

```
      #obmedzenie celej množiny na prípady z cieľového uzla
      data_complete <-
data_complete[complete.cases(data_complete[names(node_neighbors),]),]

      node_length <- nrow(data_complete)

      d <- d[as.integer(names(node_neighbors))]

      #ošetrenie prípadu keď je k väčšie ako počet záznamov v konečnom
stromovom uzle
      if(node_length < k){
        k <- node_length # zmena k na počet záznamov v uzle
        revise_k <- node_length
        max_k <- node_length
      }
      else if(node_length < revise_k){
        revise_k <- node_length
        max_k <- node_length
      }
      else if(node_length < max_k){
        max_k <- node_length
      }

    }

    #zhlukovanie k-medoids
    if(clustering_method == "k-medoids"){

      if(method == "manhattan"){
        clust <- Cluster_Medoids(data =
rbind(data_complete[,1:(ncol(data_complete)-1)],

data_incomplete[,1:(ncol(data_incomplete)-1)]),
                                 clusters = 6, distance_metric = "manhattan")
      }
      else if(method == "mahalanobis"){
        clust <- Cluster_Medoids(data =
rbind(data_complete[,1:(ncol(data_complete)-1)],

data_incomplete[,1:(ncol(data_incomplete)-1)]),
                                 clusters = 6, distance_metric = "mahalanobis")
      }

      else{
        clust <- Cluster_Medoids(data =
rbind(data_complete[,1:(ncol(data_complete)-1)],

data_incomplete[,1:(ncol(data_incomplete)-1)]), clusters = 6)
      }

      cluster_prediction <-
clust$clusters[nrow(data_complete)+nrow(data_incomplete)]

      #obmedzenie na prípady z rovnakého zhluku ako nový prípad
      data_complete <- data_complete[(clust$clusters[1:length(clust$clusters)-
1]) == cluster_prediction, ]

      cluster_length <- nrow(data_complete)
```

```
      d <<- d[as.integer(rownames(data_complete))]

      #ošetrenie k
      if(cluster_length < k){
        k <<- cluster_length
        revise_k <<- cluster_length
        max_k <<- cluster_length
      }
      else if(cluster_length < revise_k){
        revise_k <<- cluster_length
        max_k <<- cluster_length
      }
      else if(cluster_length < max_k){
        max_k <<- cluster_length
      }

    }

    # výber k-najbližších susedov
    # uloženie indexov k-NN nájdených vo fáze RETRIEVE
    knn_index <<- head(sort(d,index.return=T)$ix,k)
    knn_index_more <<- head(sort(d,index.return=T)$ix,max_k)
    # zoradenie cieľových tried podľa ich početností zostupne
    nn <<- sort(table(data_complete[knn_index,ncols]), T)
    # priradenie najčastejšej cieľovej triedy k aktuálnemu prípadu
    # fáza REUSE
    data_incomplete[j,ncols]<<- names(nn)[1]
    d <<-data.frame(d)

    # vyfiltrovanie prípadov z triedy i
    # fáza REVISE
    for(i in 0:5){
      #iba keď sú v data_complete prípady z triedy i
      if(nrow(data_complete[data_complete$Nalez == i,]) > 0){

      data_one_class <<- cbind(data_complete[data_complete$Nalez == i,],Vzd =
d[data_complete$Nalez == i,1])

      if(revise_k > nrow(data_one_class)){
      #ak by bolo príliš málo záznamov z danej triedy
      knn_index2 <<-
head(sort(data_one_class[,"Vzd"],index.return=T)$ix,nrow(data_one_class))
      }
      else{
      knn_index2 <<-
head(sort(data_one_class[,"Vzd"],index.return=T)$ix,revise_k)
      }
      data_one_class_neigh <<- data_one_class[knn_index2,]
      # výpočet priemernej vzdialenosti prípadov v rámci i-tej triedy
      sum_class[i+1] <<- sum(data_one_class_neigh[,"Vzd"])/nrow(data_one_class)
      }
    }

    revised_class <<- (sort(sum_class,index.return=T)$ix[1])-1

    data_incomplete[j,ncols+1]<<- revised_class # priradenie upravenej triedy
novému prípadu
    colnames(data_incomplete) <<- c(colnames(data_incomplete[,-
(ncols+1)]),"Nalez_upravene")
```

```
  }
  # funkcia vracia objekt typu zoznam s 3 položkami
  return(list(
    # 1. dáta z nového prípadu doplnené o Nalez a upravený Nalez
    class = data_incomplete,
    # 2. vybraný počet najbližších susedov spolu s ich vzdialenosťami od nového
prípadu
    neighbours = cbind(data_complete[knn_index,], Vzdialenost = d[knn_index,]),
    # 3. 20 najbližších susedov spolu s ich vzdialenosťami od nového prípadu
    neighbours_more = cbind(data_complete[knn_index_more,],Vzdialenost =
d[knn_index_more,])
   ))
}
```

2.2. Graf PCA (Analýzy principiálnych zložiek)

Na vykreslenie všetkých grafov v aplikácii bol použitý balík *plotly*, ktorý umožňuje vytvárať interaktívne vizualizácie s rôznymi ovládacími prvkami ako priblíženie, filtrovanie dát a export grafu v .png formáte. Ukážka zdrojového kódu slúžiaceho na vykreslenie grafu dvoch hlavných komponentov PCA pre všetky uložené záznamy:

```
   output$plot_pca <- renderPlotly({
     # graf zobrazujúci 2 hlavné komponenty PCA
     p <- plot_ly(data_pca_df(), # dáta po aplikovaní PCA
                  x = data_pca_df()[,1], # 1. zložka PCA
                  y = data_pca_df()[,2], # 2. zložka PCA
                  text =  paste("Záznam:", rownames(data_pca_df())), # text
tooltipu
                  # farebné kategórie podľa Nalezu
                  color = data_pca_df()$Nalez,
                  type = "scatter", marker = list(size = 10),
                  # farebná škála podľa stupňa Nalezu
                  colors = c("#9bd4e4","#64b8d1","#14719f","#cf7c70","#d43f28")
                  )
     # pridanie Názvov do grafu
     p <- layout(p, title = "Analýza principiálnych zložiek: všetky záznamy",
                 xaxis = list(title = "PC 1"), # názov x-vej osi
                 yaxis = list(title = "PC 2")) # názov y-vej osi
     p %>%
     # funkcia pre pridanie šípok do grafu
       add_annotations(
         x= rep(0,ncol(data_pca()$x)),
         y= rep(0,ncol(data_pca()$x)),
         axref = "x",
         ayref = "y",
         ax = ~data_pca()$rotation[,1], # 1. zložka PCA
         ay = ~data_pca()$rotation[,2], # 2. zložka PCA
         text = rownames(data_pca()$rotation), # názov atribútu
         arrowside = "start",
         showarrow = T
       )
   })
```