EKONOMICKÁ UNIVERZITA V BRATISLAVE

FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo : 103004/B/2021/421000214426

STROJOVÉ UČENIE A POROVNANIE SOFTVÉROVÝCH NÁSTROJOV STROJOVÉHO

Bakalárska práca

2021 **Jaroslav GALAĐIK**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

STROJOVÉ UČENIE A POROVNANIE SOFTVÉROVÝCH NÁSTROJOV STROJOVÉHO UČENIA

Bakalárska práca

Študijný program:	Hospodárska informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra aplikovanej informatiky
Školiteľ:	RNDr. Eva Rakovská, PhD
Vedúci katedry:	Ing. Mgr. Peter Schmidt, PhD

Bratislava, 2021 **Jaroslav GALAĐIK**

Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Jaroslav Galadik
Študijný program: hospodárska informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: Bakalárska záverečná práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Strojové učenie a porovnanie softvérových nástrojov strojového učenia

Anotácia: Pojem strojového učenia je pomerne častý, preto práca sa zameriava na prehľadné vysvetlenie charakteristík jednotlivých techník strojového učenia. Tieto charakteristiky následne slúžia v práci na porovnanie vybraných softvérových nástrojov. Práca je podporená komparáciou dvoch nástrojov na konkrétnom príklade.

Vedúci: RNDr. Eva Rakovská, PhD.
Katedra: KAI FHI - Katedra aplikovanej informatiky FHI
Vedúci katedry: Ing. Mgr. Peter Schmidt, PhD.

Dátum zadania: 24.03.2020

Dátum schválenia: 25.03.2020

Ing. Mgr. Peter Schmidt, PhD.
vedúci katedry

POĎAKOVANIE

Chcel by som sa úprimne poďakovať vedúcej práce RNDr. Eva Rakovská, PhD., za trpezlivosť, motiváciu a za poskytnutie cenných a odborných rád pri písaní bakalárskej práce.

ABSTRAKT

GALADIK, Jaroslav: *Strojové učenie a porovnanie softvérových nástrojov strojového učenie*. Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra aplikovanej informatiky. – Vedúci záverečnej práce: RNDr. Eva Rakovská, PhD. – Bratislava: FHI EU, 2021, 49 s.

Cieľom záverečnej práce je objasnenie charakteristík techník strojového učenia a na ich základe porovnať softvérové nástroje pre strojové učenie. Práca je rozdelená do 3 kapitol. Obsahuje 0 grafov, 3 tabuľky a 0 príloh. Prvá kapitola obsahuje teoretický základ strojového učenia, metódy a koncepty strojového učenia, podrobnejší opis dolovania dát a jeho metódy a presnejšie metódy klasifikácie. V ďalšej časti sa charakterizuje cieľ záverečnej práce a metódy skúmania. Záverečná kapitola sa zaoberá opisom a testovaním dvoch voľne prístupných nástrojov strojového učenia, WEKA a GATree. Nástroje boli testované pomocou metódy krížovej kontroly, anglicky cross validation, a merala sa presnosť klasifikácie. Výsledkom riešenia danej problematiky je zhodnotiť výsledky testovania nástrojov WEKA a GATree a určiť. ktorý nástroj je presnejší na implementáciu klasifikácie.

Kľúčové slová: strojové učenie, dolovanie dát, klasifikácia, krížová kontrola, nástroje

ABSTRACT

GALADIK, Jaroslav: *Machine learning and comparison of machine learning software tools*. – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Applied Informatics. – Thesis supervisor: RNDr. Eva Rakovská, PhD. – Bratislava: FHI EU, 2021, 49 p.

The aim of the final thesis is to clarify the characteristics of machine learning techniques and to compare software tools used for machine learning. The work is divided into 3 chapters. It contains 0 graphs, 3 tables and 0 attachments. The first chapter contains the theoretical basis of machine learning, methods and concepts of machine learning, a more detailed description of data mining and its methods and description of classification. In the next part, the objective of the final thesis and examination method is characterized. The final chapter deals with the description and testing of two freely accessible machine learning tools, WEKA and GATREE. Tools were tested using cross-validation methodand we measured classification accuracy. The result of this thesis to evaluate the test results of WEKA and GATREE tools and determine which tool is more accurate for the implantation of classification.

Key words: machine learning, data mining, classification, cross-validation, tools

OBSAH

ZOZNAM OBRÁZKOV

ZOZNAM TABULIEK

Úvod

Každý deň sa takmer každý z nás stretáva s nejakým dôležitým alebo menej dôležitým rozhodnutím. Máme k dispozícii obrovské množstvo údajov, ktoré nám umožňujú ľahšie sa rozhodovať a získavať z nich určité vzory, informácie dôležité pre našu budúcnosť. S veľkými rozhodnutiami sa stretávajú aj mnohé spoločnosti, zdravotníctvo, ekonomika, právnici a pod. Aby sa uľahčila správa tak veľkého množstva dát a uľahčilo sa ich použitie na dosiahnutie určitých želaných rozhodnutí, bolo pred určitým časom vyvinuté strojové učenie a jednotlivé metódy, pomocou ktorých dokážeme vyriešiť zložité problémy v určitých oblastiach. Pri strojovom učení by sa dalo povedať, že stroj naučíme určité algoritmy, pomocou ktorých získame vhodný výsledok s príslušnými údajmi. Existuje aj dolovanie dát, ktoré sa zaoberá hlavne metódami ako dosiahnuť dobrý výsledok.

Dolovanie dát a strojové učenie sú dve rôzne oblasti prieskumu údajov, ktoré sú úzko prepojené. Aby sme mohli pracovať v týchto oblastiach, potrebujeme vhodné nástroje, resp. aplikácie, ktoré nám umožňujú vykonávať metódy strojového učenia. Existuje obrovské množstvo takýchto nástrojov, ktoré sa medzi sebou líšia od vlastností, funkčnosti, ceny a jednoduchosti používania

Cieľom bakalárskej práce je prezentovať koncepciu strojového učenia, dolovania dát a testovanie presnosti klasifikácie dvoch nástrojov strojového učenia, ktoré sú voľne dostupné online. V prvej časti najprv predstavíme teoretickú časť a základné definície strojového učenia a dolovania dát. V druhej časti opíšeme dva voľne prístupné nástroje strojového učenia, Weka a GATree a otestujeme presnosť klasifikácie pri oboch nástrojoch metódou krížovej kontroly. Nakoniec poskytneme výsledky testovania vybraných nástrojov a otestujeme presnosť metódy krížovej kontroly (anglicky cross validation). Nakoniec poskytneme výsledky testovania vybraných nástrojov a porovnáme výsledky presnosti klasifikácie.

1 Súčasný stav problematiky doma a v zahraničí

Strojové učenie (anglicky machine learning, ML) sa v dnešnej dobe prakticky používa všade, napr. pri odosielaní pošty, nakupovaní, v rôznych kasínach, atď. Používajú ho tiež niektoré spoločnosti na zlepšenie obchodných rozhodnutí, zvýšenie produktivity, detekciu chorôb, predpoveď počasia a iné. Potrebujeme nielen najlepšie nástroje, ktoré máme k dispozícii, na pochopenie údajov, ale tiež musíme vedieť, ako tieto údaje pripraviť a porozumieť [2].

O samotnom strojovom učení by sa dalo povedať, že je akýmsi fenoménom zachytávania nových poznatkov [9].

1.1 História strojového učenia

História strojového učenia siaha takmer k úplnému začiatku prvých elektronických počítačov, t. j. od roku 1950. Samotný vývoj strojového učenia je úzko spojený s vývojom umelej inteligencie. Všetko to začalo myšlienkou, že stroje resp. počítače môžu napodobňovať ľudský mozog. V roku 1957 Rosenblatt vyvinul techniku „perceptron“, sochu neskoršej veľmi známej techniky neurónových sietí. Od samého začiatku vývoja strojového učenia sa vyvíjali rôzne smery. Počiatočné smery strojového učenia sú: symbolické učenie sa pravidiel, neurónové siete, podpora učenia, numerické metódy, formálna teória učenia, objavovanie zákonov, konštruktívna indukcia a indukčné programovanie [4]. V smeroch, ktoré sa dnes najčastejšie používajú, nižšie stručne opíšeme ich vývoj a význam.

Symbolické učenie pravidiel sa vyvinulo na začiatku 60. rokov 20. storočia , keď Hunt, Martin a Stone v svojej knihy ktorá sa volá “Experiments in Induction”, predstavili systém CLS (Concept Learning System), na učenie rozhodovacích stromov z príkladov. Stretávali sa s mnohými problémami, ktoré mohli vyriešiť pomocou rôznych systémov strojového učenia, ako napríklad riešenie neznámych a nezmyselných hodnôt, cena na testovanie atribútov, problém statickej spoľahlivosti pravidiel. Pomocou systému CLS bol vyvinutý program ID3 (Iterative Dichotomizer 3), program na automatické vytváranie pravidiel pre veľké súbory údajov, ktorý vyvinul Ross Quinlan v roku 1979 [4].

Pomocou systému ID3 sa zvýšil vývoj systémov na tvorbu rozhodovacích stromov. Veľký vplyv na vývoj a použitie metód strojového učenia priniesol algoritmus AQ11, ktorý sa používal na diagnostiku chorôb. Presnosť tohto algoritmu presiahla presnosť pravidiel odborníkov na diagnostiku chorôb [4]. Patria sem metódy: rozhodovacie stromy, rozhodovacie pravidlá, indukcia logických programov, atď.

Začiatky techniky **neurónových sietí** siahajú do začiatku 40. rokov 20. storočia, keď McCulloch a Pitts vytvorili matematický model nervovej bunky. Samotný model sa časom do dnes nezmenil. Neurónové siete sú oblasťou strojového učenia, ktorá sa zaoberá predovšetkým spracovaním informácií. Táto oblasť je inšpirovaná fungovaním ľudského mozgu, ktoré je plné zložitých sietí a spojení [9]. V tejto oblasti sa nachádzajú viacúrovňové neurónové siete so smerovou spätnou väzbou, neurónové siete Kohonen, neurónové siete Hopfield a niekoľko ďalších metód.

Podpora učenia sa vyvinula v roku 1959, keď vývojár Samuel porovnal dva prístupy k výučbe hry na dámu (angl. checkers). Prvým prístupom bolo memorovanie pozícií s hodnotením kvality situácie. Tieto hodnotenia poskytli výhľad do budúcnosti. Nevýhodou tohto prístupu je veľká spotreba miesta v pamäti. Druhý prístup bol však založený na podpore učenia. Samuel určil svoju váhu parametrami a cieľom naučiť sa nájsť najlepšie nastavenie váh, ktoré sa líšili podľa rozdielu medzi hodnotenou kvalitou a skutočnou kvalitou polohy [4].

Štatistické alebo numerické metódy sú najstarším smerom strojového učenia, najmä kvôli matematickému základu a odvodeniu štatistických metód, ktoré existovali pred použitím počítačov [4]. V tomto smere sa objavujú nasledujúce metódy: k-najbližší susedia, diskriminačná analýza, Bayesovský klasifikátor a pod.

1.2 Koncept strojového učenia

Strojové učenie je odvetvie výskumu umelej inteligencie. Podieľa sa na vývoji techník alebo metód, ktoré umožňujú strojom učiť sa. Ide o získavanie vedomostí na základe nejakých skúseností. Tieto metódy sa používajú na analýzu údajov, dolovanie údajov, konštrukciu numerických a kvalitatívnych modelov a pod. [4].

Dalo by sa povedať, že strojové učenie premieňa dáta na informácie. Nachádza sa na križovatke počítačových vied, štatistiky, inžinierstva a mnohých ďalších vedných disciplín [2]. Súvisí to so štatistikou, ale v porovnaní s ňou sa zaoberá viac návrhom postupov - algoritmami a výpočtovými operáciami. Štatistika však zahŕňa všetko od nastavenia hypotéz až po testovanie. Pre optimálne využitie strojového učenie potrebujeme všetky tieto vlastnosti a funkcie štatistiky. Preto sú tieto dve oblasti tak úzko prepojené [7].

Poznáme niekoľko delení strojového učenia. Strojové učenie v zásade delíme na induktívne a deduktívne. Induktívne učenie z informácií na najnižšej úrovni všeobecnosti indukuje, t. j.. generuje všeobecnejšiu znalosť. Deduktívne učenie naopak dedukuje znalosť menej všeobecnú a zložitú, ktorá je lepšie prispôsobená na riešenie konkrétneho druhu problémov.

Algoritmy, ktoré sa používajú na riešenie učiacich úloh delíme tiež na inkrementálne a neinkrementálne. Inkrementálne učenie je učenie pomocou algoritmu, ktorý spracováva jeden príklad za druhým a po každom príklade poskytuje riešenie. Z druhej strany, neinkrementálne učenie je učenie pomocou algoritmu, ktorý spracováva množinu príkladov odrazu

Delenie ktoré si podrobnejšie opíšeme je delenie strojového učenia na: kontrolované učenie, nekontrolované učenie a učenie posilňovaním [19].

Kontrolované učenie je proces, v ktorom sú vstupné a výstupné údaje sú vopred určené. Výsledkom je predikčný model [8]. Veľmi dobrým príkladom kontrolovaného učenia je klasifikácia, ktorá obsahuje nasledujúce metódy, ktoré sa v kontrolovanom učení najčastejšie používajú. Jedná sa o rozhodovacie stromy, k-najbližších susedov, diskriminačné funkcie, hybridné algoritmy, umelé neurónové siete a Bayesovské siete [4].

Nekontrolované učenie je algoritmus, ktorý nezávisle rozdeľuje získané vstupné údaje podľa kritérií do niekoľkých rôznych kategórií, z ktorých každá má svoje vlastné charakteristiky. Počet kategórií je určený samotným algoritmom. Samotný algoritmus extrahuje zo vstupných údajov kategórie a charakteristiky údajov. Nie je potrebné zadať žiadny výstup. Nekontrolované učenie je dobrým príkladom regresie vrátane metód: regresné stromy, lineárna regresia, lokálne vážená regresia, metóda podporných vektorov pri regresii a rovnako ako pri klasifikácii aj umelé neurónové siete a hybridné algoritmy [4].

Učenie posilňovaním spočíva v tom, že učený agent koná na základe vstupov z prostredia, ale spätnú väzbu – číselné ohodnotenie (odmenu alebo trest) – môže dostať až oneskorene. Nikdy mu teda nie je priamo prezentované správne konanie v danej situácii, a aj to, či vykonal správnu akciu, môže iba nepriamo odvodiť až z (vo všeobecnosti ľubovoľne) oneskoreného ohodnotenia. Cieľom je maximalizovať toto ohodnotenie [4].

Ako zvoliť správnu metódu? Správnu metódu vyberáme rôznymi spôsobmi, vždy však musíme najskôr zvážiť náš cieľ a definovať čo všetko zahŕňa. Potom sa pozrieme na naše dostupné údaje, z ktorých si môžeme vyberať. Napríklad, ak máme predpovedané cieľové hodnoty, zvolíme kontrolované učenie. Ak nemáme cieľovú hodnotu, potom sa riadime nekontrolovaným učením. Ak sme zvolili kontrolované učenie, pozrieme sa na to, či je naša cieľová hodnota diskrétna hodnota (áno / nie, A / B). Ak je však cieľová hodnota číselná hodnota alebo nemáme cieľové hodnoty, vyberie sa regresia. Metódy sa volia aj mnohými inými možnými spôsobmi [2].

1.3 Kroky strojového učenia

Poznáme všeobecný prístup k používaniu strojového učenia, ktorý obsahuje sedem jednoduchých krokov. Kroky rozvoja strojového učenia sú:

- Zhromažďovanie údajov – najprv musíme zhromaždiť všetky príslušné údaje a vzorky, pomocou ktorých sa môžeme dostať do nášho konečného cieľa
- Príprava vstupných údajov – po zhromaždení všetkých údajov je potrebné ich primerane spracovať a previesť do príslušného formátu.
- Analýza vstupných údajov – v tomto kroku je potrebné tieto údaje správne analyzovať, určiť, či je možné v údajoch identifikovať možné vzorky, či môžu existovať niektoré údaje úplne odlišné od ostatných, atď.
- Ak pracujeme s produkčným systémom a vieme, ako majú vyzerať naše údaje, môžeme tento krok preskočiť. Vo väčšine prípadov nepotrebujeme prítomnosť človeka v tomto kroku. Ak však nemáme automatizovaný systém, potrebujeme prítomnosť osoby, ktorá kontroluje, či sa do systému nedostávajú nesprávne údaje.
- Použitie algoritmu – v piatom kroku prichádza na rad strojové učenie. Algoritmy plníme správnymi a čistými údajmi z prvého a druhého kroku a využívame všetky získané vedomosti a informácie. Tieto získané vedomosti sú uložené vo forme,

ktorá je okamžite k dispozícii na použitie strojového učenia v ďalších dvoch krokoch.

- Testovanie algoritmov – v tomto kroku testujeme výkonnosť algoritmov.
- Využitie vedomostí – v poslednom kroku znova skontrolujeme všetky kroky a použijeme algoritmus pre konkrétny problém [2].

Výsledkom "učenia sa z údajov" môžu byť funkcie, pravidlá, vzťahy, systémy rovníc, ktoré je možné znázorniť rôznymi metódami [4].

1.4 Dolovanie dát – data mining

Ľudia v priebehu času neustále hľadajú nejaké vzorce v dátach. V minulosti to bolo trochu inak a väčšinou s menším počtom dát. Napr. lovci hľadajú vzorky migrácie zvierat, poľnohospodári v oblasti plodín, politici podľa názoru voličov, atď. Dolovanie dát sa líši od každodenného vyhľadávania vzorov v tom, že údaje sú uložené v elektronickej podobe a že vyhľadávanie vzoriek je automatizované t. j. vykonávané pomocou počítačov. Z dôvodu nárastu údajov a množstva zariadení, ktoré umožňujúzber voveľkej kvantite údajov, zvyšujú sa aj príležitosti na dolovanie dát. Dáta sú veľmi cenným zdrojom, pretože nás môžu viesť k novým poznatkom a konkurenčným výhodám v obchodnom prostredí [7].

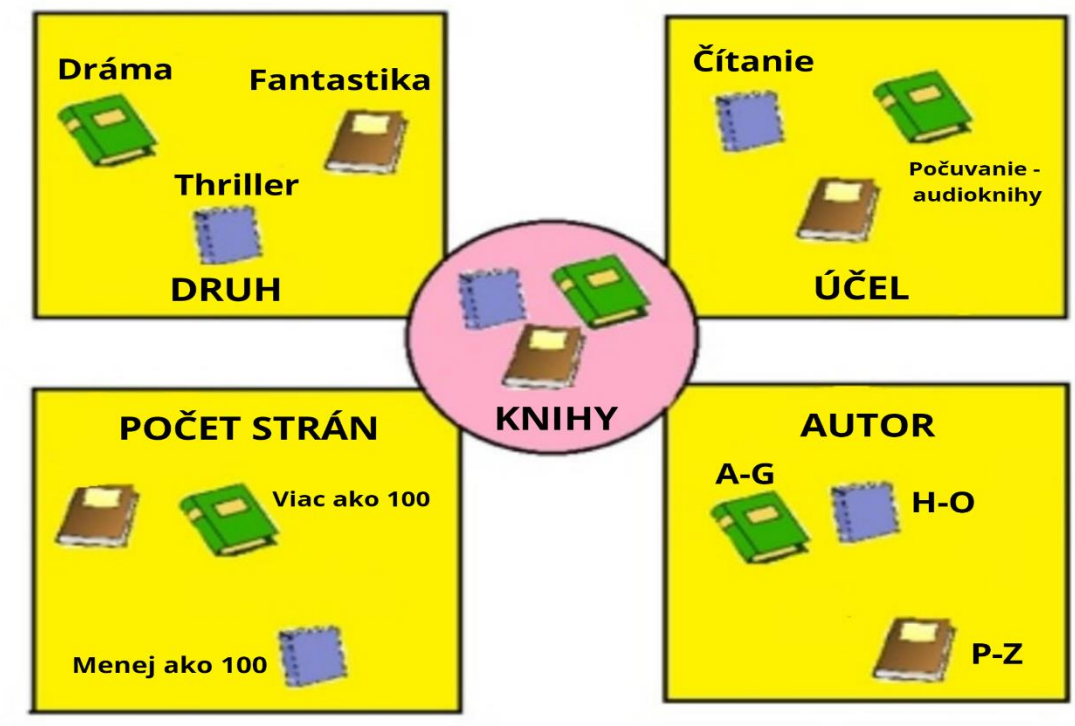
Dolovanie dát a strojové učenie sú dve odlišné, ale zároveň súvisiace oblasti. Dolovanie dát možno definovať ako proces zisťovania vzorov, asociácií, zmien a udalostí v dátach. Objavené vzorce musia mať zmysel a musia viesť k určitým ekonomickým výhodám. Užitočné vzory by nám mali umožniť ľahko predvídať nové údaje [7].

Dolovanie dát je v informačnom priemysle rozšírené najmä kvôli veľkému množstvu údajov a priamej potrebe tieto údaje poskytovať. Získané vzorky je možné použiť pre aplikácie, ktoré nám umožňujú analyzovať, odhaľovať podvody, kontrolovať výrobu a individuálny výskum [3].

Poznáme veľa rôznych techník dolovania dát. V krátkosti popíšeme techniky používané pri úlohách dolovania dát. Klasifikáciu a klasifikačné metódy popíšeme podrobnejšie. Ďalej sa bližšie pozrieme na dôležitú metódu rozhodovacích stromov.

1.5 Segmentácia

Okrem pojmu segmentácie existujú ešte jeden pojem, zhlukovanie (klastrovanie). Zhlukovanie je technika získavania údajov, ktorej hlavným cieľom je rozdeliť veľké množstvo údajov na menšie skupiny, ktoré majú podobné vlastnosti. Príkladom je knižnica. Na policiach v knižnici sú naukladané knihy, ktoré majú určitú spoločnú tému, napr. informatika, geografia, chémia ... Môžu byť usporiadané tiež podľa počtu strán, formátu, autora, názvu [22]. Takáto metóda je znázornená na obrázku nižšie (Obrázok 1). Segmentácia patrí do skupiny nekontrolovaného učenia, a preto nemá žiadne nezávislé premenné [9].



Obrázok 1 Segmentácia[9]

1.6 Vizualizácia (detekcia nezrovnalostí)

Vizualizácia údajov sa najčastejšie používa pri kontrole, interpretácii údajov, úprave údajov, hľadaní dôležitých vzťahov, identifikácii výhod a nevýhod. Dalo by sa povedať, že

transformuje údaje do transparentnejšieho formátu. Pretože máme veľa rôznych údajov, ktoré je potrebné identifikovať, používame rôzne metódy používané pri vizualizácii. Počas celého svojho používania sa ukázala ako spoľahlivá a veľmi jednoduchá metóda, pretože zvyšuje pochopenie a dôležitosť toho, čo sa naučia takíto odborníci, ako aj ďalší ľudia, ktorí sa o túto oblasť zaujímajú. Metódy používané pri vizualizácii sú vizualizácia jedného atribútu, vizualizácia párov atribútov, vizualizácia viacerých atribútov a vizualizácia výsledkov strojového učenia [4].

1.7 Klasifikácia

Klasifikácia je jednou z najbežnejších techník a jednou z najpopulárnejších techník dolovania dát. Na prvý pohľad sa zdá, že je to pre človeka takmer absolútne nevyhnutné, pretože sme čoraz viac konfrontovaní s problémom klasifikácie a kategorizácie. Hlavnou úlohou klasifikácie je preskúmať vlastnosti objektu (problému) a zaradiť ho do jednej z vopred určených tried [1]. V tejto technike máme niekoľko rôznych metód nazývaných klasifikátory. Atribúty sú nezávislé spojité alebo diskrétne premenné, ktoré sa používajú na popis objektov (problémov). Závislá diskrétna premenná je na druhej strane triedou, ktorej priradíme hodnotu vzhľadom na hodnoty nezávislých premenných [4]. Najčastejšie sa klasifikácia používa na odhaľovanie podvodov, pri výrobe, výber obsahu v cielenom marketingu [1] a hľadaní diagnóz v zdravotníctve [4]. Technika klasifikácie sa považuje za techniku kontrolovaného učenia.

Poznáme niekoľko rôznych metód resp. klasifikátory, ktoré sú oddelené podľa spôsobu prezentácie funkcie klasifikátora. Najbežnejšie metódy sú rozhodovacie stromy, naivný Bayesov klasifikátor, rozhodovacie pravidlá, technika najbližšieho suseda a umelé neurónové siete [4].

1.7.1 Rozhodovacie stromy

Rozhodovacie stromy sú veľmi populárnou metódou pre klasifikáciu. Sú to symbolické metódy strojového učenia a umožňujú rýchle porozumenie a učenie. Ich štruktúra je v tvare stromu. Využíva sa na rozdelenie veľkého počtu záznamov na menšie množiny (triedy), ktoré na seba nadväzujú v nedefinovanom poradí [1].

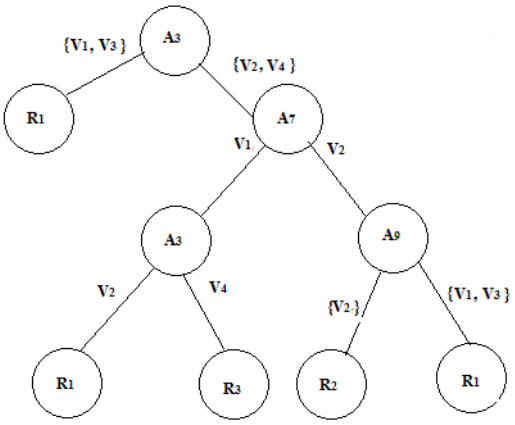
Rozhodovací strom sa skladá z koreňa alebo uzla v hornej časti stromu a jednotlivých vnútorných uzlov, vetví a listov. Interné uzly sú označené názvami atribútov a obsahujú

podmienku, ktorá rozdeľuje učiacu sa množinu na menšie množiny. Vetvy alebo spojenia označené možnými hodnotami pre atribút sú podmnožinou hodnôt atribútov. Listy uzlov spoločného uzla zodpovedajú triedam. Cesta od koreňa k listu zodpovedá jednému rozhodovaciemu pravidlu [4]. Výber atribútov závisí od súboru učebných vzorcov, okolností, od schopnosti vykonávať merania. Rozhodnutie je udalosť, ktorá sa stane v blízkej budúcnosti, ak sa rozhodneme pre ňu. Pomocou rozhodovacích stromov sa môžeme pokúsiť predpovedať udalosť, ktorá sa stane pri určitom rozhodnutí, alebo ho môžeme použiť na nájdenie ďalších najlepších možností na dosiahnutie cieľov. Pomocou nich vytvárame čoraz jednoduchšie pravidlá, ktoré budú mať čo najviac vzorov [9].

Rozhodovací strom je možné zostaviť pomocou výučbovej sady, ktorá sa skladá z niektorých vzoriek. Vlastnosti vzoriek sú popísané množinou atribútov resp. s vlastnosťami a s triedami, do ktorých patria (výsledok). Jedna vzorka učenia patrí do jednej triedy. Pri konštrukcii musíme tiež brať do úvahy, že niekoľko vzoriek opísaných rovnakým vektorom atribútov nesmie mať odlišné rozhodnutia. Ak sú splnené posledné dve podmienky, hovoríme o konzistentných vzoroch učenia. Poznáme však aj nekonzistentné vzorce učenia, ktoré sú znakom toho, že sme v fáze merania atribútov urobili chybu. V učebnej sade máme tiež dve rôzne možnosti atribútov - môžu byť diskrétne, pomocou ktorých môžeme priamo zostaviť strom, alebo máme spojité atribúty, ktoré musíme pred zostavením zmapovať do diskrétnej formy. Proces, ktorým zostavujeme rozhodovacie stromy, sa inak nazýva indukcia [9].

Poznáme niekoľko rôznych druhov rozhodovacích stromov, z ktorých spomenieme najznámejší binárny rozhodovací strom a regresný rozhodovací strom. Binárny rozhodovací strom (Obrázok 2) sa rozpozná, ak má každý interný uzol presne dvoch nástupcov. Akýkoľvek „obyčajný“ rozhodovací strom je možné previesť na binárny formát [4].

Pri regresných stromoch sa používajú spojité aj diskrétne atribúty, v ktorých musí byť vopred známy ich počet. Algoritmus na konštrukciu regresných stromov je veľmi podobný algoritmu na konštrukciu rozhodovacích stromov klasifikácie. Takéto stromy sa dajú použiť na určenie hodnoty závislej premennej nových prípadov. Môžeme tiež skonštruovať binárny regresný strom, ktorý nám umožňuje odhadnúť atribúty odchýlok. S tým dostávame menšie stromy, ktoré nám umožňujú lepšiu presnosť [4].



Obrázok 2 Binárny strom [4]

Rozhodovacie stromy majú jednu zlú vlastnosť, a to, že sa veľmi ťažko upravujú, keď sa zvýši počet atribútov na základe ktorých strom robíme.

Poznáme veľké množstvo rôznych nástrojov, ktoré sa špecializujú na vytváranie rozhodovacích stromov, napríklad Weka a GATree, ktoré si podrobnejšie popíšeme neskôr. Existuje samozrejme veľa ďalších nástrojov.

1.7.2 Bayesov klasifikátor

Bayesov klasifikátor je štatistická klasifikácia. Pomenúva sa podľa anglického matematika Thomasa Bayesa. Pomocou nej môžeme predpovedať pravdepodobnosť, s akou jednotlivá inštancia patrí do konkrétnej triedy. Najznámejšou formou Bayesovho klasifikátora je naivný Bayesovský klasifikátor, ktorý predpokladá nezávislosť vplyvu atribútových hodnôt na triedu [3].

1.7.3 Neurónové siete

Neurónové siete sa začali vyvíjať medzi rokmi 1930 a 1940, ešte predtým, ako digitálny počítač vôbec existoval. V roku 1943 Warren McCulloch a Pitts vytvorili jednoduchý model, ktorý vysvetľuje, ako fungujú biologické neuróny. V roku 1950 boli neurónové siete prvýkrát použité v digitálnom počítači [1]. Počas vývoja sa model neurónovej siete príliš nezmenil. Vývoj zaznamenal veľký pokrok v roku 1986, keď Rumelhart, Hinton a William predstavili algoritmus na prispôsobenie váh vo

viacúrovňových neurónových sieťach. Tem algoritmus nasledovali aj ďalšie algoritmy na učenie neurónových sietí [9].

Neurónové siete, ako aj rozhodovacie stromy, sú veľmi populárnou metódou, pretože sú veľmi spoľahlivé pri mnohých vyhľadávaniach údajov a poskytujú dobrú podporu pre rozhodovacie aplikácie. Neurónové siete sú veľmi podobné biologickým. Ľudský mozog umožňuje ľuďom generalizovať na základe skúseností a počítače pracujú podľa konkrétnych pokynov. Toto môže byť použité pri dolovaní dát, kde nám výskum umožňuje získať nové a lepšie výsledky v budúcnosti [1]. Najdôležitejšie vlastnosti neurónových sietí sú biologická podobnosť, asynchrónne vykonávanie a viacsmerné vykonávanie, flexibilita v reálnom čase, schopnosť učenia, automatické generalizovanie atď. Jedným z nedostatkov je, že nemôžu vysvetliť svoje rozhodnutie [4]. Sieť neurónov má oveľa vyšší limit kapacity ako neuróny. Počet a spôsob spojenia týchto neurónov v sieti sa nazýva topológia neurónovej siete. Niektoré siete majú obmedzenia v spojení medzi úrovňami, tieto sa nazývajú neurónové siete so spätnou väzbou [9].

1.7.4 Metoda K-najbližšieho suseda

Metódu k-najbližších susedov možno tiež stručne nazvať K-NN (k-najbližších susedov). Je to jedna z numerických metód strojového učenia. Umožňuje nám ľahko uložiť všetky učebné prípady, alebo podmnožiny učebných prípadov. Učenie touto metódou takmer neexistuje, takže môžeme povedať, že patrí do kategórie lenivého učenia (angl. lazy learning). Používa sa predovšetkým na riešenie problémov s klasifikáciou a regresiou. Príklad - ak chceme predpovedať triedu knovému príkladu, medzi učebnými príkladmi nájdeme k-najbližšieho suseda, a v klasifikácii predikujeme väčšinovú triedu, t. j. triedu, do ktorej patrí väčšina k-najbližších susedov [4].

1.7.5 Podporné vektory

Poznáme niekoľko rôznych metód podporných vektorov (angl Support Vector Machines). Nepoužívajú sa tak dlho ako iné metódy. Plne sa rozvinuli až v 90. rokoch a považujú sa za najúspešnejšie metódy regresie a klasifikácie. Pracujú na lineárnych a nelineárnych údajoch. Pomocou týchto metód sa snažíme kombinovať atribúty čo najchytrejším spôsobom a zvoliť toľko atribútov, koľko máme k dispozícii, pretože iba tak

metóda extrahuje najdôležitejšie a najžiadanejšie informácie. Transformuje určité atribúty zo základnej formy do vyššej dimenzie a umiestňuje optimálnu nadrovinu. Môžeme mať aj viac nadrovín, ak sú príklady lineárne oddeliteľné. Optimálny je taký, ktorý je rovnaký a najďalej od najbližších príkladov oboch tried [4]. Metódy podporných vektorov sú veľmi presné a dajú sa rýchlo naučiť. Používajú sa tiež na predpovedanie a iné klasifikačné metódy [3]. Metóda podporných vektorov má tiež malú nevýhodu, a to, že interpretácia jednotlivých rozhodnutí je veľmi zložitá [4].

1.7.6 Ostatné metódy klasifikácie

Poznáme tiež ďalšie klasifikačné metódy, ktoré sa nepoužívajú tak často ako tie, ktoré sú opísané vo vyššie uvedených riadkoch, ale sú však tiež dôležité, pretože sa dajú úspešne použiť v určitých aplikáciách, takže je správne spomenúť ich iba stručne.

Genetické algoritmy sú klasifikované ako vyhľadávacie algoritmy. Používame ich na riešenie veľmi náročných problémov s vyhľadávaním a optimalizáciou. Ich základy sú založené na princípe „prežitia najschopnejších“, rovnako ako sa aj prírodný výber riadi týmto princípom, tieto úlohy môžeme úspešne vyriešiť pomocou genetických algoritmov.. Genetické algoritmy pozostávajú z troch genetických operátorov, konkrétne z výberu, pomocou ktorého získame vhodných jednotlivcov. Po výbere vykonáme druhú operáciu, ktorou je kríženie, kde urobíme najlepšie riešenie z dvoch vybraných jednotlivcov. Poslednou operáciou je však mutácia, ktorá predstavuje príležitostnú zmenu jednotlivca, ktorá tiež prispieva k lepšiemu riešeniu [9].

Metódy hrubých množín (anglicky Rough Set) sú založené na základnej teórii množín, ktoré však koexistujú v dvoch oblastiach, a to v matematike a umelej inteligencii. Používajú sa na zisťovanie štrukturálnych vzťahov v súbore údajov, ktoré sú zvyčajne nepresné, zle organizované a predovšetkým neúplné. Sú to práve metódy hrubých množín, ktoré nám umožňujú nové techniky neúplných údajov [9].

Hybridné algoritmy sa kombinujú s inými metódami, aby sa čo najlepšie využili výhody iných individuálnych prístupov. Sú teda kombinované do stále kvalitnejších algoritmov [4]. Podľa spôsobu kombinácií metód poznáme štyri rôzne takzvané hybridy, ktoré sa líšia hlavne spôsobom spojenia, a to: **sekvenčný hybrid** spája dve metódy postupne a obe metódy sú nezávislé; **externý hybrid**, tu sú dve metódy spojené

hierarchicky, pričom prvá metóda je hlavná a hybrid je pomenovaný podľa nej. Funguje to tak, že najskôr prvá metóda spracuje údaje a odošle ich druhej metóde, ktorá po vykonaní svojej úlohy odošle údaje späť k prvej metóde a tá vráti výsledok. V tomto hybride je druhá metóda úplne závislá od prvej; **zabudovaný hybrid**, v tomto hybride sú tieto dve metódy veľmi úzko prepojené, žiadna metóda nemôže fungovať samostatne, tento hybrid sa odporúča používať najviac; **paralelný hybrid**, v ktorom sú obe metódy voľne spojené a fungujú úplne nezávisle jedna od druhej [9].

2 Cieľ práce, metodika práce a metódy skúmania

Cieľ práce je zameraný na objasnenie charakteristík techník strojového učenia a na ich základe porovnať softvérové nástroje pre strojové učenie. Pojem strojového učenia je pomerne častý, preto práca sa zameriava na prehľadné vysvetlenie charakteristík jednotlivých techník strojového učenia. Tieto charakteristiky následne slúžia v práci na porovnanie vybraných softvérových nástrojov. V práci sme porovnávali dva nástroje na konkrétnom príklade.

V prvej časti práce sa zameriavame na základné pojmy strojového učenia. Opísali sme históriu strojového učenia, definovali základné koncepty strojového učenia a definovali kroky strojového učenia. Ďalej sme sa oboznámili s pojmom dolovanie dát a bližšie vysvetlili metódy dolovania dát. Podrobnejšie sme popísali pojem klasifikácia, keďže v praktickej časti budeme merať presnosť klasifikácie strojových nástrojov WEKA a GATree.

V teoretickej časti uvádzame potrebné informácie z odbornej literatúry, ktoré boli potrebné na oboznámenie sa s danou problematikou. Táto časť bude východiskovou pre druhú časť, teda praktickú. Po preskúmaní viacerých nástrojov strojového učenia, ako napríklad Scikit Learn, Accors, Shogun a pod., sme vybrali dva nasledujúce nástroje – Weka a GATree, pretože sú veľmi jednoduché na používanie. Na vysvetlenie funkčnosti nástrojov existujú podrobnejšie návody na rôznych internetových stránkach. Použitie zložitých algoritmov klasifikácie pri vybraných nástrojoch je pomerne jednoduché.

V praktickej časti sa najprv oboznámime s používaním nástrojov WEKA a GATree. Ukážeme si a vysvetlíme určité časti nástrojov, ktoré nám budú potrebné na vykonanie testovania. Nakoniec otestujeme presnosť klasifikácie pri nástrojoch WEKA a GATree použitím metódy krížovej kontroly.

Výsledkom práce je zhodnotiť výsledky testovania nástrojov WEKA a GATree. Výsledky testovania zhrnieme do tabuľky a určíme, ktorý nástroj je presnejší na implementáciu klasifikácie.

3 Výsledky práce

3.1 Nástroje strojového učenia

Existuje mnoho rôznych nástrojov strojového učenia. Tieto nástroje sa líšia najmä ich vlastnosťami, použiteľnosťou, cenou a ďalších funkciách. Žiadny nástroj nie je vhodný pre všetky problémy, ktoré sa vyskytujú v strojovom učení. Určený nástroj strojového učenia je vybraný v závislosti od našich potrieb a vedomostí. Vybrali sme dva nástroje, ktoré poskytujú možnosť porovnávania vybranej metódy – rozhodovacích stromov strojového učenia.

V nasledujúcom texte stručne opíšeme dva rôzne nástroje, ktoré použijeme na tvorbu rozhodovacích stromov a sú voľne prístupné online. Sú to Weka a GATree. Neskôr budeme testovať presnosť klasifikácie daných nástrojov, použitím štatistickej metódou krížovej kontroly klasifikácie, anglicky cross validation, a výsledky nakoniec porovnáme. Budeme sa venovať klasifikačnej úlohe, teda presnosti zaraďovania údajov do vopred určených tried.

Budeme používať rovnaké databázy pre každý nástroj. Pre dané nástroje popíšeme iba základné funkcie a časti programu, ktoré sú potrebné pre tento typ testovania.

3.2 Weka

Weka je veľmi populárny nástroj, ktorý obsahuje veľkú zbierku metód strojového učenia. Je navrhnutý tak, aby rýchlo otestoval existujúce režimy, metódy strojového učenia [6].

Aplikácia Weka (Waikato Environment for Knowledge Analysis) bola vyvinutá v roku 1992 na univerzite na Novom Zélande. Jej názov vyplýva zo slova Waikato, čo znamená životné prostredie na analýzu vedomostí. Logo aplikácie predstavuje vták Meko, ktorý sa nachádza na ostrovoch Nového Zélandu (Obrázok 3). Celý program je napísaný v programovacom jazyku Java. Pracuje na skoro všetkých operačných systémoch, vrátane Linuxu. Najväčšia výhoda aplikácie Weka je, že je voľne prístupná všetkým používateľom online [6].

Obrázok 3 Logo nástroja WEKA

Weka podporuje niekoľko rôznych metód prediktívneho modelovania, medzi ktorými sú:

- Bayesove stromy – anglicky Bayes Tree (AODE, Bayesov klasifikátor, naívny Bayesov klasifikátor, atď.)
- Stromy – anglicky tree (ADTree, Decision Tree, Id3, J48, LMT, M5P, NBTree, RandomForest, RandomTree, REP Tree, atď.)
- Pravidlá – anglicky rules (ConjunctiveRule, DecisioTable, JRip, M5Rules, Nge, OneR Part, Prism, Ridor, ZeroR)
- Lenivé klasifikátory – anglicky lazy clasificators (IB1, IBK, KSTAR, LBR, LWL) [6].

3.2.1 Použitie nástroja WEKA

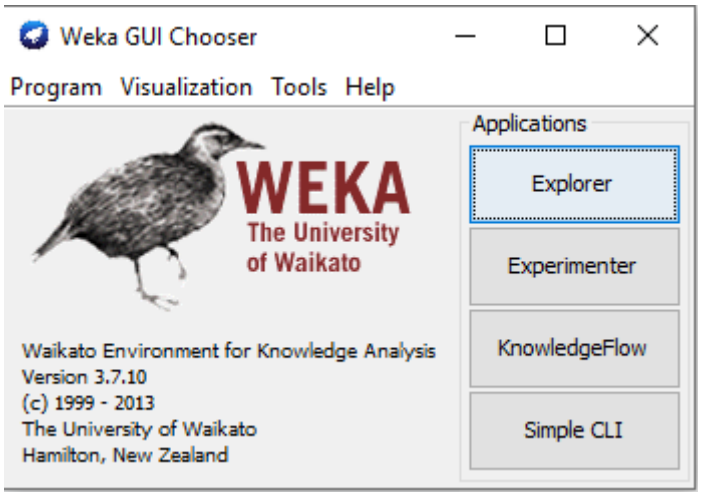
Ako sme opísali vo vyššie uvedených riadkoch, nástroj Weka je veľmi ľahko použiteľný. Má veľmi jednoduché začiatočné menu, ktoré sa zobrazí na začiatku nástroja a ktoré umožňuje štyri rôzne možnosti použitia program. (Obrázok 4)

Ako prvá možnosť je tlačidlo ‘Explorer’ ktorá nám umožňuje preskúmať naše údaje. V tomto prostredí si môžeme vybrať údaje a formátovať ich tak, aby boli pripravené na ďalšie testovanie. Umožňuje nám klasifikovať údaje do určitých skupín, tvorbu rôznych typov stromov a zmenu ich štruktúry. Získané výsledky môžu byť tiež graficky vizualizované [6].

Následne tlačidlo je ‘Experimenter‘, ktoré nám umožňuje vykonávať rozsiahle pokusy, analyzuje štatistické testy a údaje o výkonnosti. Štatistické údaje sa uložia do formátu ARFF (Attribute-Relation File Format), tak, aby mohli byť použité na ďalšie spracovanie [6].

Tlačidlo ‘Knowledge Flow‘ nám umožňuje veľmi podobné funkcie ako prvé tlačidlo ‘Explorer’. Obsahuje rozhranie drag-and-drop. Toto tlačidlo tiež poskytuje, že sa údaje postupne spracúvajú, vizualizácia výkonnosti jednotlivých klasifikátorov sa vykonáva počas spracovania, pridanie nových komponentov pre tok vedomostí a pod [16].

Posledné tlačidlo ‘SimpleCLI‘ nám umožňuje používať metódy dolovania dát pomocou príkazového riadku [9].



Obrázok 4 Začiatočné menu aplikácie WEKA

V nasledujúcom texte popíšeme možnosti používania nástroja WEKA, a ako program prichádza na vhodné výsledky presnosti klasifikácie. Tento proces ďalej budeme opakovať na troch rôznych databázach.

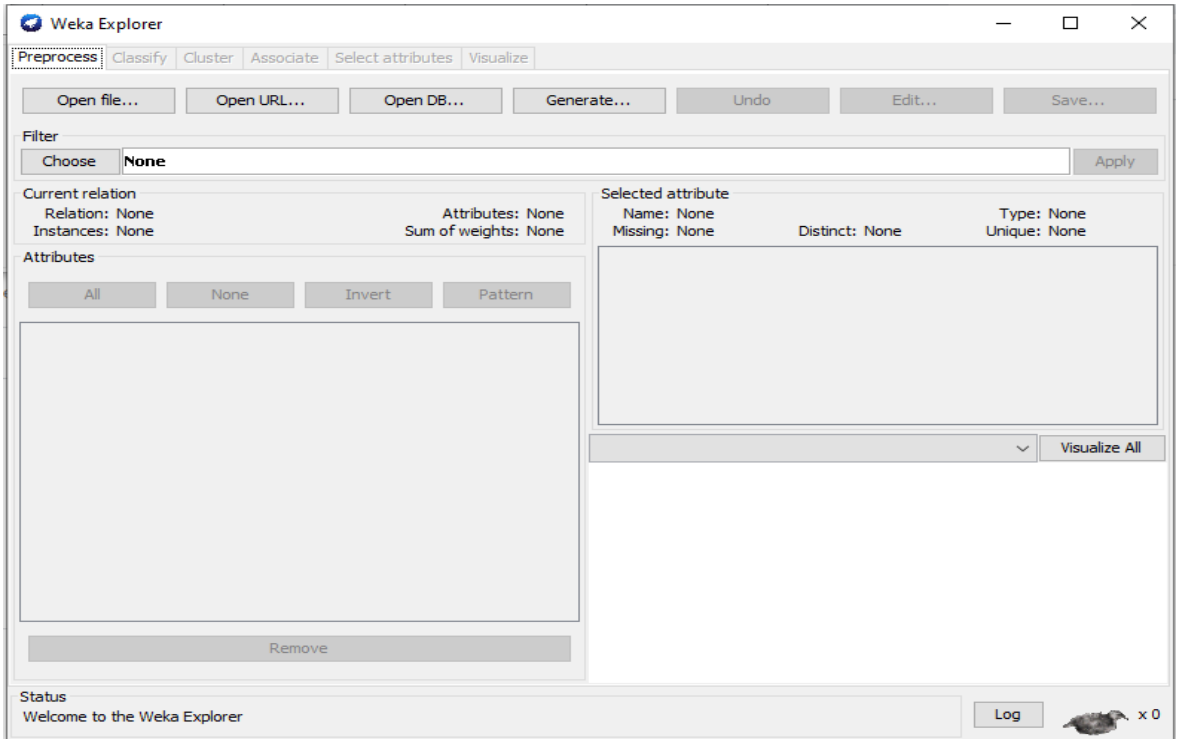
Predtým, ako začneme využívať aplikáciu WEKA, musíme najprv formátovať údaje, ktoré budú zodpovedať formátu ARFF. Atribúty môžu byť číselné, nominálne, krátke reťazce a dátumy. Väčšina súborov ARFF sa skladá z príkladov atribútov, ktoré sú oddelené čiarkami. Weka nám umožňuje previesť CSV (excel súbor) do formátu ARFF.

Na začiatočnom okne aplikácie WEKA, vyberieme prvú možnosť ‘Explorer’. Otvorí sa nám nové okno so záložkami na spracovanie našich údajov (Obrázok 5).

Máme šesť záložiek, ktoré vykonávajú ďalšie funkcie:

- Preprocess – alebo prvé formátovanie údajov, tu sa vyberá súbor údajov (databáza), ktorý je neskôr možné zmeniť
- Classify – alebo triedenie, nám umožňuje klasifikovať a overiť prediktívne modely.
- Clusster – táto možnosť nám umožňuje klasifikovať údaje do skupín

- Associate – kde nájdeme asociatívne pravidlá pre údaje a hodnotíme ich.
- SelectAttributes – kde vyberieme najrelevantnejšie atribúty v našich údajoch, a znovu vykonáme proces navrhovania prediktívnych modelov
- Visualise – umožňuje interakciu v grafickom príkaze [6].



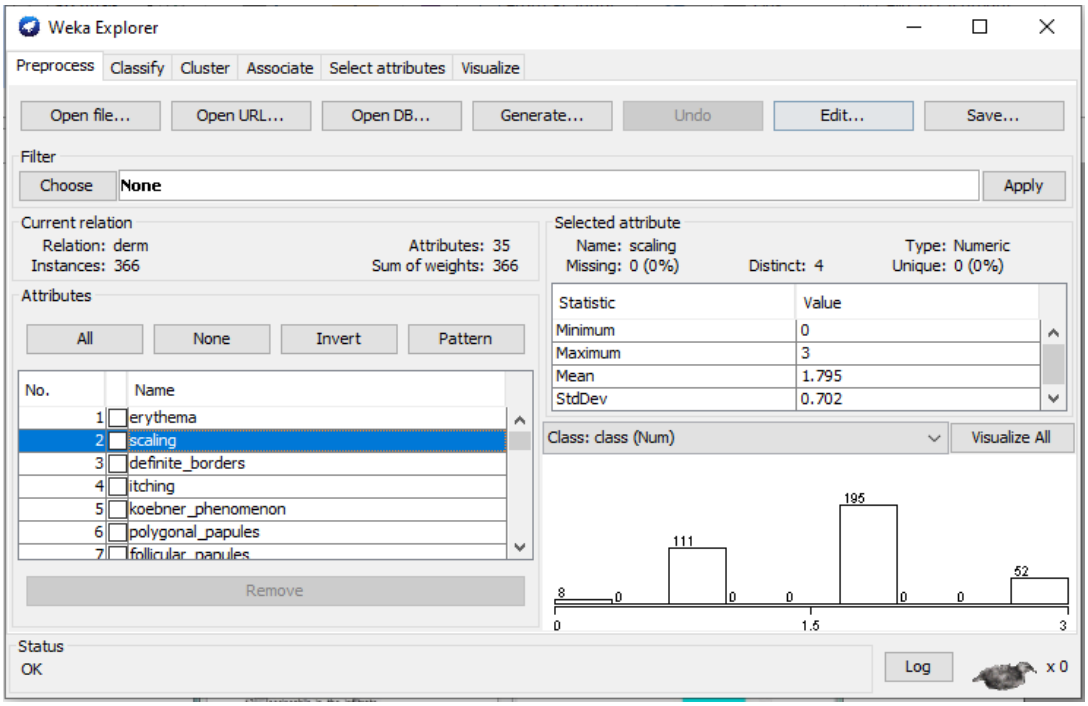
Obrázok 5 Okno Explorer v aplikácií WEKA

Ak chceme použiť tieto okná, musíme najprv na okno Preprocess (Obrázok 6) importovať súboru vo formáte ARFF. Na otváranie súboru máme k dispozícii viac možností, môžeme priamo otvoriť súbor s údajmi z počítača (Open file), alebo vyberieme webovú adresu, kde sú údaje uložené (Open URL). Máme možnosť aj čítať údaje z databázy (Open DB). WEKA nám tiež umožňuje vytvorenie umelých údajov (Generate).

My sme si vybrali prvú možnosť (Open file). Keď sa údaje importujú, okno bude doplnené informáciami o súbore údajov. Kliknutím na tlačidlo "Choose" si môžete vybrať filter pre vymazanie určitých vlastností zo sady údajov, alebo inými slovami, ručne vyberieme nastavenia atribútov [6].

V sekcii "Atributes" vidíme všetky atribúty nášho dátového súboru. Kliknutím na konkrétne pole alebo tlačidlo, môžeme pridať alebo odstrániť určitý atribút. Stlačením na akýkoľvek atribút, sa v okne "Selected attribute" vypíše štatistika daného atribútu. V

pravom dolnom rohu je nakreslený histogram pre konkrétny atribút. Kliknutím na tlačidlo "Visual All", môžeme v rovnakom čase vidieť grafy všetkých atribútov.



Obrázok 6 Okno preprocess v aplikácií WEKA

Ďalej si podrobnejšie opíšeme záložku "Classify" (Obrázok 7), ktorá nám pomáha triediť a overovať prediktívne modely. V ľavej hornej časti máme tlačidlo "Choose". Kliknutím naň sa otvorí hierarchické menu, kde si vyberieme typ modelu. V našom prípade si vyberieme typ rozhodovacieho stromu J48 [6].

J48 je jedným z populárnych algoritmov klasifikácie, ktorý vytvára rozhodovací strom. J48 je open source java implementácia algoritmu C4.5 v nástroji Weka, a bol vytvorený projektovým tímom Weka.Je to jeden z najužitočnejších prístupov k tvorby rozhodovacích stromov pre problémy s klasifikáciou [20].

Každý aspekt informácií je potrebné rozdeliť na menšie podskupiny na základe rozhodnutia. J48 ukazuje na štandardizovaný zisk údajov, ktorý skutočne rozdelí informácie výberom určitého atribútu. Menšie podmnožiny sú vrátené algoritmom. Rozdelené stratégie sa zastavia, ak podmnožina má miesto s podobnou triedou vo všetkých inštanciách. J48 rozvíja uzol rozhodovania s využitím očakávaných odhadov triedy [14].

V ľavej časti "Test options" máme štyri rôzne testovacie režimy:

- Use training set – kde testujeme údaje, na ktorých bol model postavený. Hodnotenie sa získa rýchlo, je to veľmi optimistické, ale napriek tomu môže byť aj

užitočné, pretože všeobecne predstavuje strop na vykonávanie modelu v nových dátach.

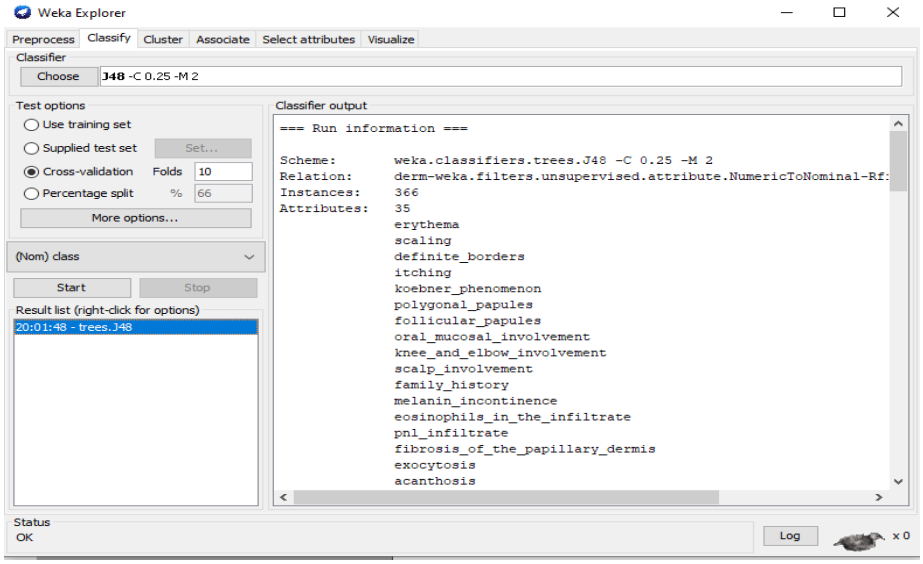
- Supplied test set – používa sa tam kde je prediktívny model testovaný na dátach načítaných z testovacieho súboru.
- Cross-validation – používa sa tam, kde je prediktívny model testovaný krížnou kontrolou, t.j. k-krát krížovou kontrolou. Úroveň krížovej kontroly je určená v pravom poli vedľa možnosti „Cross-validation“
- Percentage split – klasifikátor sa odhaduje podľa určitého percentuálneho podielu vstupných údajov [6].

Keď určíme model, parametre a spôsob overenia, môžeme spustiť proces učenia kliknutím na tlačidlo “Start”. Tento proces môžeme hocikedy zastaviť.

Po ukončení procesu sa výsledky overovania a testovania zobrazujú priamo v časti “Classifier output”. Výsledky sú uvedené v piatich častiach, a to:

- Run information – v tejto časti vidíme, ktorý model bol vybraný, názov dátového súboru, počet údajov, číslo a názvy atribútov a spôsob testovania
- Classifier mode – tu vidíme textové zobrazenie rozhodovacieho stromu nášho prediktívneho modelu
- Summary – dáva nám štatistické zhrnutie prediktívneho modelu a testovania
- Detailed accuracy by class – ukazuje na presnosť zobrazenia podľa určitých tried
- Confusion matrix –ukazuje klasifikáciu prípadov z každej triedy.

Výsledky môžu byť uložené v akomkoľvek formáte, budú dočasne uložené v ľavej časti “Result list”, kde môžeme vidieť aj ďalšie výsledky, ktoré sme urobili.



Obrázok 7 Okno classify v aplikácií WEKA

Nástroj Weka nemá žiadne zvláštne nevýhody, môžeme poukázať len na to, že môžeme pracovať s údajmi len vtedy, ak sú vo formáte, ktorý nástroj podporuje. Ako slabú stránku by sme mohli zdôrazniť skutočnosť, že program je napísaný v programovacom jazyku Java, ktorý je o niečo pomalší ako iné programovacie jazyky.

3.3 GATree

GATREE (Obrázok 8) je tiež nástrojom, s ktorým riešime problémy pomocou rozhodovacích stromov. Môžeme tiež formulovať genetické algoritmy, ktoré priamo rozvíjajú binárne rozhodovacie stromy. GATREE umožňuje rozvíjanie rozhodovacích stromov tak dlho, ako je to potrebné. To znamená, že počet vetiev a listov je ľubovoľný. Rozvíjanie stromu sa môže zastaviť, keď sú výsledky uspokojivé. Poskytuje tvorbu rôznych stromov a všetky tieto stromy môžu byť použité striedavo [12].

Obrázok 8 Logo aplikácie GATree

3.3.1 Použitie apilkácie GATree

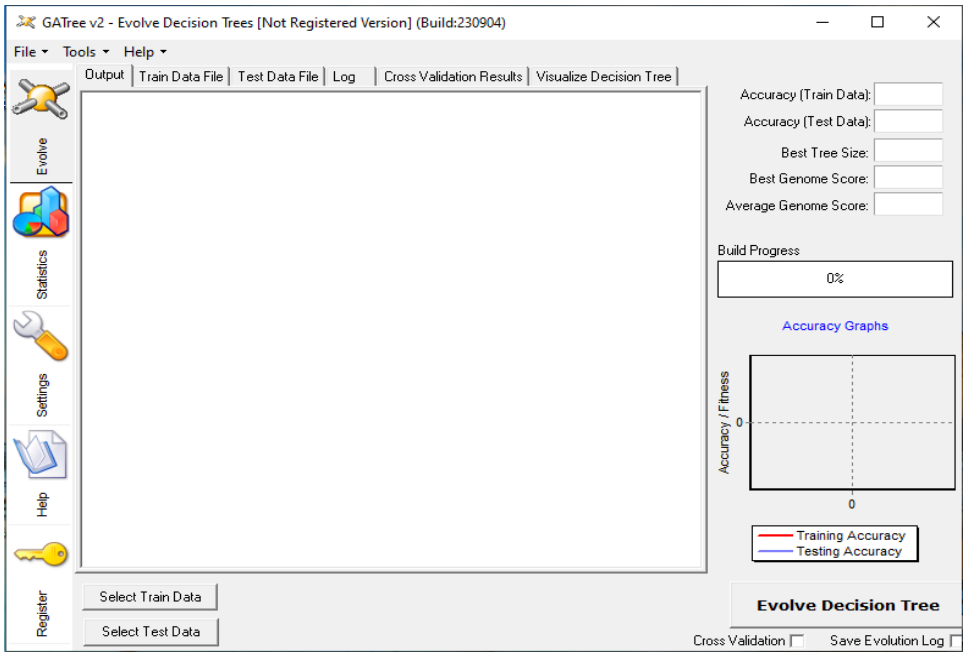
Pri spustení aplikácie GATree sa otvorí počiatočné okno (Obrázok 9), v ktorom, rovnako ako pri predchádzajúcom programe, musíme najprv vybrať aktívny výber údajov na vytvorenie rozhodovacieho stromu. Dátový súbor musí byť v ARFF formáte, rovnako ako s nástrojom Weka.

V hlavnom menu môžeme neustále monitorovať náš rozhodovací strom. V samotnom okne nástroja máme na ľavej strane hlavné tlačidlá na používanie programu. Sú to:

- Evolve – rozvíjanie rozhodovacieho stromu, v tejto časti môžeme sledovať, na aký spôsob je náš strom vytvorený. Individuálne záložky nám umožňujú viac príležitostí na preskúmanie nášho rozhodovacieho stromu.
- Statistics – štatistika, tu program umožňuje zobraziť niekoľko grafov, ktoré ukazujú, čo sa stane s naším stromom v procese vývoja. Tieto grafy nám umožňujú sledovať proces v reálnom čase a rýchlejšie môžeme nájsť potenciálne problémy a trendy.
- Settings – nastavenia, tu program poskytuje kontrolu všetkých aspektov v procese tvorby stromu. Existujú dva typy nastavení: základné nastavenia a rozšírené nastavenia, v závislosti od použiteľnosti a zložitosti.
- Help – pomoc [13].

V hlavnom menu máme uvedených šesť okien, ktoré nám ponúkajú nasledujúce možnosti:

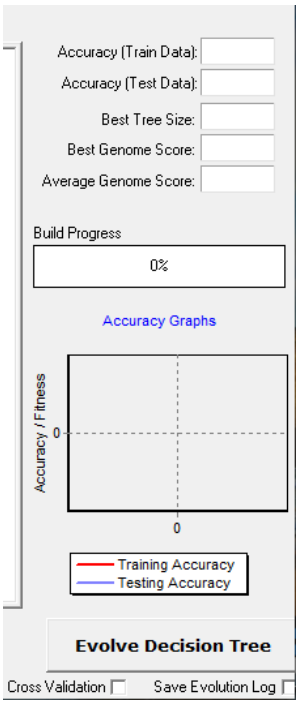
- Output – tu môžeme vidieť nás rozhodovací strom v reálnom čase
- Train Data File / Test Data File – tu si môžeme prezrieť obsah nášho súboru (obrázok 9)
- Log – tu môžeme vidieť najlepšie rozhodovacie stromy za každých 5 generácií v procese, program ich automaticky ukladá
- Cross Validation Results- tu vidíme výsledky merania presnosti klasifikácie
- Visualize Decision Tree – tu môžeme vidieť náš vizualizovaný rozhodovací strom [13].



Obrázok 9 Začiatočné menu v aplikácii GATree

V pravom hornom rohu (Obrázok 10) v hlavnom menu máme možnosť vidieť nasledovné výsledky:

- Accuracy (Train Data/Test Data) – tu môžeme vidieť presnosť rozhodovacieho stromu
- Best TreeSize – najlepšia veľkosť rozhodovacieho stromu
- Best GenomeScore / Average GenomeScore – kontrola počtu nesprávnych stromov, ktoré nahradíme novými, a priemer nesprávnych stromov
- BuildProgress – v tomto poli vidíme, koľko percent nášho rozhodovacieho stromu je už vytvorených
- AccuracyGraphs – v tejto časti je nakreslený graf presnosti



Obrázok 10 Výpis výsledkov v aplikácií GATree

Ako jediné dva nedostatky v nástroji GATREE sú, že má zlé pokrytie štatistík pre tvorbu rozhodovacích stromov a dlhý čas testovania [12].

3.4 Testovanie aplikácií

V tejto kapitole budeme testovať presnosť klasifikácie nástrojov Weka a GATree, ktoré sme stručne opísali v predchádzajúcich častiach. Otestujeme presnosť klasifikácie pri oboch nástrojoch, pomocou metódy krížovej kontroly, ktorú si aj popíšeme (anglicky cross validation). Pre klasifikačné problémy sa zvyčajne používa k-krát krížová kontrola. Je to populárna metóda, jednoduché je ju pochopiť. Všeobecne vedie k menej optimistickému odhadu modelovej zručnosti ako iné metódy, ako napríklad metóda train/test rozdelenie.

Pri testovaní budeme používať rovnaké databázy pre oba nástroje, pretože tak môžeme najpresnejšie identifikovať, ktorý nástroj je najlepší na používanie klasifikácie. Riešime úlohu klasifikácie, t.j. porovnanie presnosti klasifikácie v nástrojoch Weka a GATree, na troch dátových súborov, teda na konkrétnej doménovej oblasti ktorá je popísaná dátami v určitej databáze. V oboch nástrojoch si vytvoríme rozhodovacie stromy pre každú databázu, a potom použitím krížovej kontroly uvidíme presnosť zaraďovania údajov do vopred určených tried, teda aká je presnosť klasifikácie pri našich databázach

v oboch nástrojoch. Inými slovami, určíme v percentách koľko údajov je nesprávne kategorizovaných.

3.4.1 Krížová kontrola (cross validation)

Krížová kontrola je štatistická metóda oceňovania a porovnávania učiacich sa algoritmov, ktoré rozdeľujú údaje na dva segmenty. Prvý segment sa používa na učenie alebo tréningový model, zatiaľ čo druhý na kontrolu modelu. Najtypickejším príkladom krížovej kontroly je verifikácia a certifikácia, ktorá sa neustále opakuje v cykloch po sebe idúcich, aby bolo možné overiť každý dátový bod. Základnou formou je k-krát krížová kontrola.

Procedúra k-krát krížovej kontroly má jeden parameter s názvom K, ktorý sa vzťahuje na počet segmentov, na ktoré sa daná vzorka údajov má rozdeliť. Najčastejšie používaná hodnota K je 10. V takom prípade táto forma sa volá10-krát krížová kontrola [21].

Pri metóde k-krát krížová kontrola sa overovanie údajov rozdelí na k-rovnaké alebo približne rovnaké segmenty alebo kusy. Prvé segmenty sa používajú na učenie, ktoré potom predpovedia posledný segment. Proces sa opakuje k-krát, zatiaľ čo vždy predpovedáme ďalší segment údajov a jedna časť údajov je ponechaná na testovanie. Po dokončení testovania k-násobku sa dosiahne výsledok priemernej presnosti modelov. Výsledok, ktorý sa získa, nám hovorí presnosť prediktívneho modelu na údajoch, ktoré neboli použité na tvorbu stromu [10].

V nasledovnom texte použijeme štatistickú metódu krížovej kontroly K-10 (10-krát) na meranie presnosti klasifikácie nástrojov strojového učenia. Vytvoríme rozhodovacie stromy, skontrolujeme presnosť klasifikácie a uvedieme percento chyby.

3.4.2 Databázy

Pri kontrole presnosti klasifikácie v nástrojoch strojového učenia Weka a GATree, budeme používať rovnaké databázy v formáte ARFF. Databázy, ktoré budeme používať v bakalárskej práci, sú voľne dostupné na internete. Databázy majú rozdielny počet údajov, pretože to uľahčí pochopenie toho, ako nástroj hodnotí presnosť klasifikácie podľa rozdielneho množstva údajov.

Prvá databáza ktorú použijeme, sa týka témy dermatológie. Obsahuje 366 údajov, t. j. vzoriek a 34 rôznych atribútov (Obrázok 11), z ktorých 33 sú lineárne a jeden nominálny. Údaje sú pravdivé, obsahujú informácie o rôznych kožných ochoreniach. Databáza je číselného typu. Atribúty opisujú príznaky určitej choroby a čísla od 0-3 predstavujú stupeň intenzity symptómov ochorenia. Atribút pod sériovým číslom 11 je vždy označený 0 alebo 1, čo nám hovorí, či je prítomná rodinná história tohto ochorenia. Atribút číslo 34 označuje vek pacienta. Hodnoty chýbajúcich atribútov sú označené znakom "?" [17].

No.	Name
1	erythema
2	scaling
3	definite_borders
4	itching
5	koebner_phenomenon
6	polygonal_papules
7	follicular_papules
8	oral_mucosal_involvement
9	knee_and_elbow_involvement
10	scalp_involvement
11	family_history

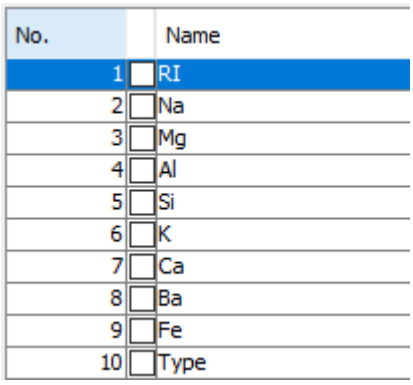
Obrázok 11 Záznamy z databázy Dermatology

Druhá databáza sa vzťahuje na tému skla. Obsahuje 214 údajov, t. j. vzoriek a 9 rôznych atribútov (Obrázok 12). Typy atribútov sú tiež numerické. Prvým atribútom je ID počet skiel, všetky ostatné atribúty opisujú zloženie skla. Atribút 4 až 9 opisuje percento prítomnosti konkrétnej látky. Posledný atribút opisuje, pre ktorú vec sa používa konkrétna kompozícia [15].

Atribúty v databázy Glass:

1. Index lomu
2. Na - sodík
3. Mg - horčík
4. Al - hliník
5. Si - kremník
6. K - draslík
7. Ca - vápnik
8. Ba - bárium
9. Fe – železo
10. Typ skla

- 1 stavebné okná
- 2 obytné okná
- 3 prepracované okná vozidiel
- 4 neprepracované okná vozidiel
- 5 kontajnery
- 6 riad
- 7 svetlomet



No.	Name
1	RI
2	Na
3	Mg
4	Al
5	Si
6	K
7	Ca
8	Ba
9	Fe
10	Type

Obrázok 12 Záznamy z databázy Glass

Tretia databáza je najmenšia, v porovnaní so zvyškom dvoch a obsahuje údaje o rastline (Obrázok 13). Obsahuje tri triedy, z ktorých každá obsahuje 50 údajov, t. j. vzoriek, to znamená len 150 údajov. Prvá trieda je lineárne oddelená od ostatných dvoch. Údaje sú reálne a numerické. Databáza obsahuje štyri rôzne atribúty [11].

Atribúty v databázy Iris:

- Kvet – dĺžka v cm
- Kvet – šírka v cm
- List – dĺžka v cm
- List – šírka v cm
- Trieda
    1. Iris Setos
    2. Iris Versicolur
    3. Iris Virginica

No.	Name
1	Id
2	SepalLengthCm
3	SepalWidthCm
4	PetalLengthCm
5	PetalWidthCm
6	Species

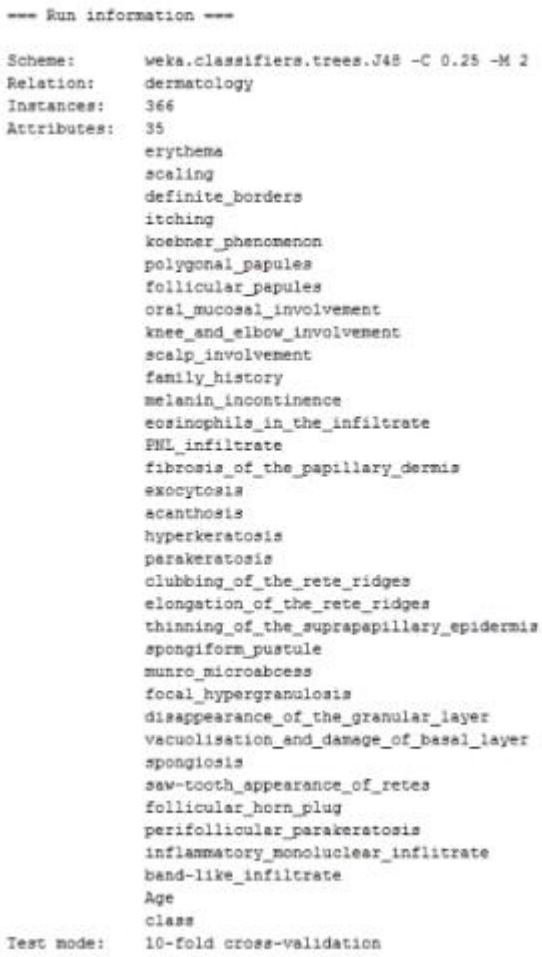
Obrázok 13 Záznamy z databázy Iris

Dané databázy použijeme pre oba nástroje a riešime úlohu klasifikácie, t.j. porovnanie presnosti klasifikácie v nástrojoch Weka a GATree. Vytvoríme rozhodovacie stromy pre každú databázu použitím 10-krát krížovej kontroly, a uvidíme ako presne sú údaje zaradené do vopred určených tried.

3.4.3 Testovanie aplikácie WEKA

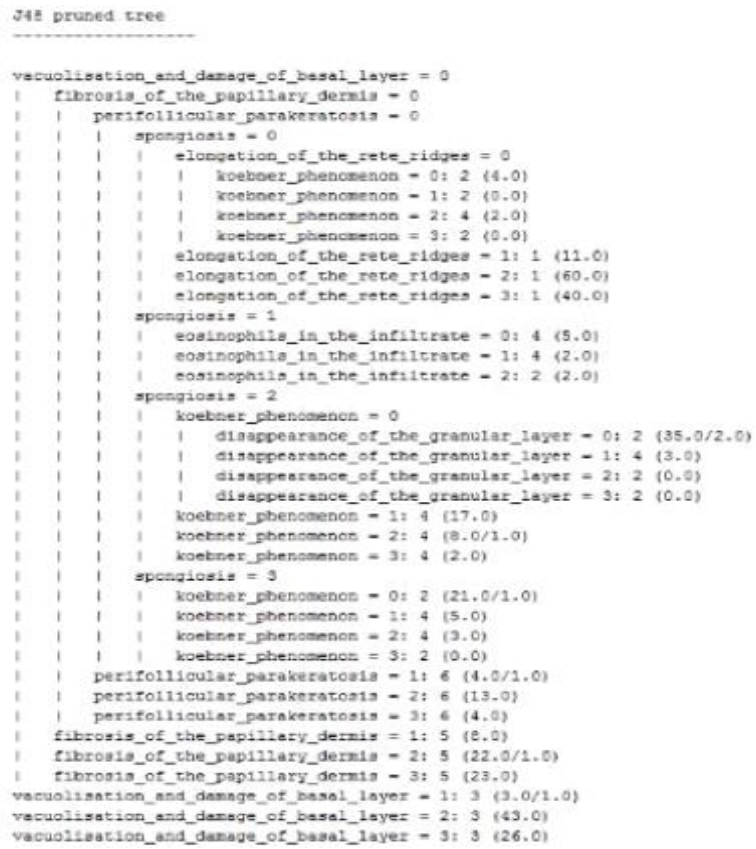
Pri testovaní nástroja Weka boli v tvorbe rozhodovacieho stromu nastavené iba základné nastavenia, ako bolo minule opísané. Použili sme metódu k-krát krížovej kontroly (k=10) a vytvorili najlepší štandardný rozhodovací strom J48 ktorý sme už popísali. Rozhodovací strom J48 je implementácia algoritmu ID3, vyvinutý projektovým tímom nástroja Weka [18].

Najprv sme testovali prvú databázu Dermatology. Po nastavení programu na testovanie, program Weka vypísal informácie z rozhodovacieho stromu. Z výpisu (Obrázok 14) môžeme zistiť, že bol vybraný typ rozhodovacieho stromu J48, môžeme vidieť že databáza obsahuje 34 atribúty, a že sme pre testovanie použili metódu 10-k krížovej kontroly.



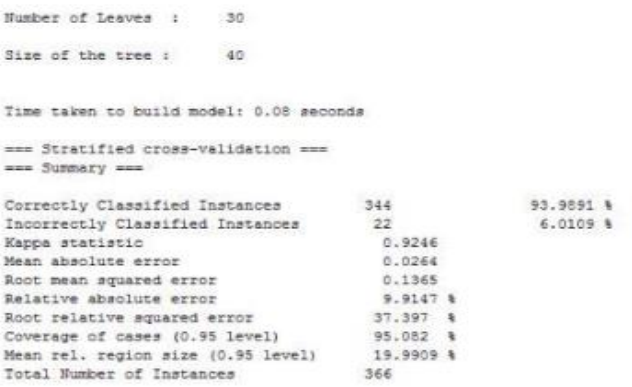
Obrázok 14 Výpis údajov o databázy v aplikácií WEKA (dermatology)

V pokračovaní na obrázku nižšie (Obrázok 15) vidíme vytvorený rozhodovací strom pre databázu Dermatology.



Obrázok 15 Výpis rozhodovacieho stromu Dermatology v aplikácií WEKA

V nižšie uvedenom obrázku, kde je výsledok krížovej kontroly, vidíme (Obrázok 16), že rozhodovací strom obsahuje 30 listov a že veľkosť stromu (počet uzlov a listov) je 40. Tiež je zobrazený čas budovania stromu, a v našom prípade je to 0,8 sekundy. Ďalej je vypísaná presnosť vybudovaného modelu. V našom prípade je presnosť stromu 93,99%, čo znamená, že správne zoradených údajov je 344 a počet chybne klasifikovaných údajov Weka odhadla na 6,01%, čiže 22 údaje.



Obrázok 16 Výsledky krížovej kontroly v aplikácií WEKA (Dermatology)

Potom sme testovali databázu Glass (Obrázok 17), ktorá obsahuje údaje o skle použitých na rôznych výrobkoch. Nastavenie nastavíme rovnakým spôsobom ako pri prvej databáze.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Glass
Instances:    214
Attributes:   10
              RI
              Na
              Mg
              Al
              Si
              K
              Ca
              Ba
              Fe
              Type
Test mode:    10-fold cross-validation
```

Obrázok 17 Výpis údajov o databázy v aplikácií WEKA (Glass)

Zo dolného výpisu (Obrázok 18) môžeme pochopiť, že rozhodovací strom obsahuje 30 listov a že jeho veľkosť je 59. Aplikácia Weka potrebovala 0,13 sekundy na vytvorenie stromu s 214 údajmi a 10 atribútmi.

Z obrázku (Obrázok 18) môžeme vidieť presnosť vybudovaného modelu. V našom prípade je presnosť stromu 66,82%, čo znamená, že správne klasifikovaných údajov je 143 a počet chybne klasifikovaných údajov Weka odhadla na 33,18%, čiže 71 záznamov.

```
Number of Leaves  :     30

Size of the tree :      59


Time taken to build model: 0.13 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         143               66.8224 %
Incorrectly Classified Instances        71               33.1776 %
Kappa statistic                          0.55
Mean absolute error                      0.1026
Root mean squared error                  0.2897
Relative absolute error                 48.4507 %
Root relative squared error             89.2727 %
Coverage of cases (0.95 level)          78.972  %
Mean rel. region size (0.95 level)      21.4286 %
Total Number of Instances              214
```

Obrázok 18 Výsledky krížovej kontrole v aplikácií WEKA (Glass)

Ako poslednú databázu sme použili databázu Iris, ktorá obsahuje informácie o kvete iris, ktorá obsahuje iba 150 údajov a 5 atribútov. Tiež sme si vybrali krížovú kontrolu presnosti klasifikácie, ktorú môžeme vidieť na uvedenom obrázku nižšie (Obrázok 19).

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Iris
Instances:    150
Attributes:   5
              sepal_length
              sepal_width
              petal_length
              petal_width
              class
Test mode:    10-fold cross-validation
```

Obrázok 19 Výpis údajov o databázy v aplikácií WEKA (Iris)

Z Obrázku 20 vidíme ďalšie údaje. Rozhodovací strom obsahuje 5 listov, a že jeho veľkosť je 9 uzlov. Čas tvorby rozhodovacieho stromu je 0,07 sekundy. Presnosť stromu je 96%, čo znamená, že správne klasifikovaných údajov je 144 a počet chybne klasifikovaných údajov Weka odhadla na 4%, čiže 6 záznamov.

```
Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Coverage of cases (0.95 level)          96.6667 %
Mean rel. region size (0.95 level)      33.7778 %
Total Number of Instances              150
```

Obrázok 20 Výsledky krížovej kontrole v aplikácií WEKA (Iris)

Po testovaní aplikácie Weka boli výsledky zhrnuté v tabuľke nižšie (tabuľka 1). Dostali sme výsledky, ktoré hovoria že tento nástroj je najpresnejší v klasifikácii malých databáz. V našom prípade to bola databáza Iris, ktorá obsahovala 150 údajov a 4 atribúty, kde bola presnosť klasifikácie odhadovaná na 96%. Veľmi dobré výsledky sú tiež aj pri testovaní databáze Dermatology, ktorá obsahuje 366 údajov a 34 atribúty, kde je presnosť

klasifikácie 93,99%. Najhoršie odhadovaná presnosť klasifikácie je určená pri databáze Glass, ktorá obsahuje 214 údajov a 10 atribútov, kde presnosť klasifikácie bola iba 66,82%.

Tabuľka 1 Výsledky testovania aplikácie Weka

Aplikácia - Databáza	Dermatology	Glass	Iris
Weka	93.99%	66.82%	96%

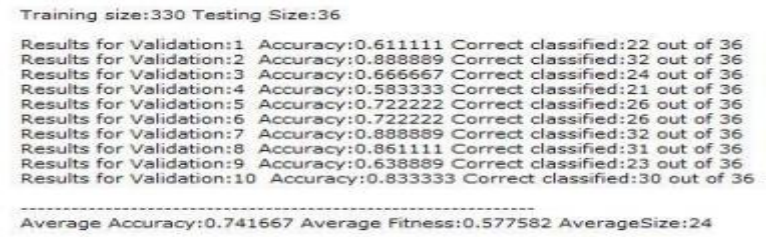
3.4.4 Testovanie aplikácie GATree

Pri testovaní nástroja GATree sme použili tie isté databázy, ktoré sme použili pri testovaní nástroja Weka. Pri tvorbe rozhodovacích stromoch boli nastavené tie základné nastavenia a krížová kontrola klasifikácie bola nastavená na K-10.

V nástroji GATREE, po konštrukcii stromu databázy Dermatology, v pravej časti sú uvedené nasledujúce výsledky o našom rozhodovacom strome (Obrázok 21). V prvých dvoch častiach vidíme najlepšiu presnosť stromu pri krížovej kontrole klasifikácie, v ďalšom okne vidíme, aká je naša najlepšia veľkosť stromu, čo je v našom prípade 25 uzlov. Najpresnejší výsledok krížového overovania prvej databázy je 86,06%.

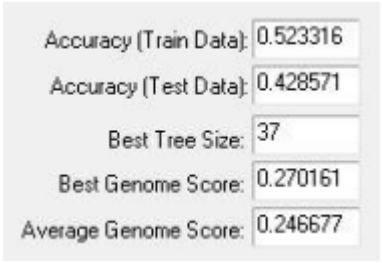
Obrázok 21 Výpis o vytvorení stromu v nástroji GATRee (Dermatology)

Z nasledujúceho obrázku (Obrázok 22) môžeme vidieť, ako nástroj GATree distribuuje údaje na desať približne rovnakých častí. Program vypočíta priemernú presnosť stromu, ktorá je v našom prípade 74,17%.



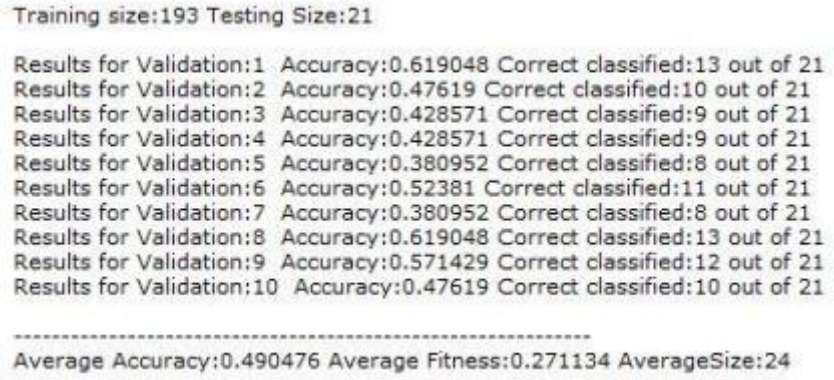
Obrázok 22 Výsledok krížovej kontrole v nástroji GATree (Dermatology)

Ďalšia databáza, podľa ktorej sme testovali presnosť klasifikácie v nástroji GATree, je, rovnako ako aj pri nástroji WEKA, databáza Glass. Najlepšia veľkosť stromu so všetkými uzlami je 37. Z dolného Obrázku 23 môžeme vidieť najlepšiu presnosť stromu podľa krížovej kontroly klasifikácie, ktorá je 52,33%.



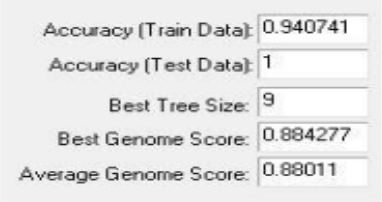
Obrázok 23 Výpis o vytvorení stromu v nástroji GATRee (Glass)

Priemerná presnosť krížovej kontroly všetkých desiatich častí sa odhaduje na 49,04% (Obrázok 24)



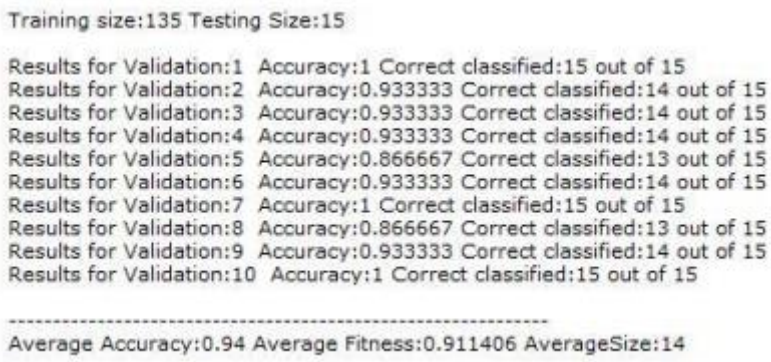
Obrázok 24 Výsledok krížovej kontrole v nástroji GATree (Glass)

Posledná databáza, podľa ktorej sme testovali nástroj GATree, je, rovnako ako aj pri nástroji WEKA, databáza Iris. Najlepšia veľkosť stromu je 9 uzlov. Najlepšia presnosť stromu podľa krížovej kontroly klasifikácie je 94.07% (Obrázok 25).



Obrázok 25 Výpis o vytvorení stromu v nástroji GATRee (Iris)

Z nižšieho obrázku (Obrázok 26) môžeme vidieť, že program odhaduje priemernú presnosť krížovej kontroly na 94%.



Obrázok 26 Výsledok krížovej kontrole v nástroji GATree (Iris)

Po testovaní nástroja GATree tiež kladieme výsledky do jednej tabuľky (tabuľka 2). Použili sme výsledky priemernej krížovej kontroly na desiatich častiach a nie najlepšie výsledky krížovej kontroly. Výsledky sú horšie ako v nástroji WEKA. Najlepší výsledok bol získaný pri testovaní najmenšej databázy Iris, s 150 údajmi, kde sme dostali presnosť 94%. V databáze Dermatology sa získal mierne horší výsledok, kde je presnosť krížovej kontroly 74.17%. Najhorší výsledok sme dostali pri databáze Glass, rovnako ako pri nástroji WEKA. Pri tejto databáze, ktorá obsahuje 214 dát a 10 atribútov sme získali presnosť krížovej kontroly 49,04%.

Tabuľka 2 Výsledky testovania aplikácie GATree

Aplikácia - Databáza	Dermatology	Glass	Iris
GATree	74.17%	49.04%	94%

3.4.5 Porovnanie nástrojov WEKA a GATree

Po všetkých popisoch testovania nástrojov sme všetky výsledky zhrnuli do jednej menšej tabuľky (Tabuľka 3), v ktorej môžeme vidieť výsledky oboch nástrojov s konkrétnymi databázami.

Rovnako ako pri menšom množstve údajov, aj pri väčšom množstve údajov môžeme pochopiť, že pri kontrole presnosti klasifikácie je presnejším nástrojom Weka. Vo všetkých prípadoch databáz bol menej presný nástroj GATree.

Z nižšie uvedenej tabuľky (Tabuľka 2), môžeme tiež pochopiť, že podľa všetkých databáz je nástroj Weka veľmi presný pri kontrole presnosti klasifikácie pre databázy s akýmkoľvek množstvom údajov. Percento presnosti vo všetkých troch databázach je vysoká. Výsledky presnosti klasifikácie sú ovplyvnené nie len objemom databázy, ale aj obsahom a typmi údajov v databáze.

Tiež sme vypočítali priemer všetkých výsledkov krížovej kontroly klasifikácie. Priemer bol vypočítaný podľa štandardného vzorca pre výpočet priemeru. Pri výpočte priemeru presnosti zo všetkých troch databáz sme dostáli nasledujúce výsledky: Nástroj Weka má priemernú hodnotu presnosti krížovej kontroly 85,60%, a nástroj GATree má menšiu priemernú hodnotu presnosti krížovej kontroly, a to len 72,40%. Priemer výsledkov konkrétneho nástroja nám hovorí, ktorý nástroj bol presnejší, a ktorý nástroj by bol najvhodnejší na používanie pre klasifikáciu databáz rôznych veľkostí. V našom prípade, vyššiu priemernú presnosť klasifikácie má nástroj WEKA. Oveľa horšiu presnosť klasifikácie (o 13.20%) má nástroj GATree.

Tabuľka3 Výsledky testovanie aplikácie všetkých nástrojov

Aplikácia - Databáza	Dermatology	Glass	Iris	PRIEMER
Weka	93.99%	66.82%	96%	85.60%
GATree	74.17%	49.04%	94%	72.40%

Záver

Strojové učenie je veľmi zaujímavou a rozsiahlou oblasťou. Pokrýva obrovské znalosti a skúsenosti z iných oblastí ako sú štatistika, učenie z dát, databázy, atď. Metódy strojového učenia môžu byť použité v mnohých oblastiach v závislosti od problému, ktorý je potrebné vyriešiť.

Na začiatku bakalárskej práce sme stanovili viac častí. Jednou z najdôležitejších bola, že v teoretickej časti sme predstavili metódy a koncepcie strojového učenia. Konkrétnejšie sme popísali metódy ťažby dát, segmentáciu, klasifikáciu a vizualizáciu. Ďalej sme podrobnejšie popísali koncepciu klasifikácie a techniky rozhodovacích stromov.

V praktickej časti sme sa venovali testovaniu dvoch voľne prístupných nástrojov strojového učenia, Weka a GATree na spustenie rovnakého klasifikačného algoritmu. Vybraté nástroje sme najprv opísali a potom sme otestovali presnosť klasifikácie použitím metódy krížovej kontroly.

Na testovanie uvedených nástrojov, ktoré podporujú používanie klasifikácie, sme použili tri identické voľne prístupné databázy (Glass, Dermatology a Iris). Databázy sa navzájom líšia v množstve údajov, počte atribútov a typoch údajov. Dostali sme výsledky, podľa ktorých môžeme povedať, že nástroj Weka je presnejší ako nástroj GATree, keď sa jedná o presnosť krížovej kontroly klasifikácie.

Tvorba bakalárskej práce na tému strojového učenia bola veľmi poučná. Najviac problémov sme mali pri získavaní literatúry pre konkrétny nástroj a nastavenie nástroja podľa určitých parametrov na účel testovania presnosti klasifikácie určitého nástroja.

Zoznam použitej literatúry

Knižné zdroje:

[1] BERRY, Michael J.A. – LINOFF, Gordon S. – *Data Mining Techniques: For Marketing Sales, and Customer Relationship Management.* 2. vyd. Wiley Publishing, 2004, 672 s. ISBN 9780471470649

[2] HARRINGTON, Peter – *Machine Learning in Action*. 1. vyd. Shelter Island: Manning Publications, 2012, 384 s. ISBN978-1617290183.

[3] HAN, Jiawei – KAMBER, Micheline – *Data Mining: Concepts and Techniques.* 2. vyd. San Francisco: Morgan Kaufmann, 2011, 744 s. ISBN 0123814790

[4] KONONENKO, Igor – *Strojno učenje.* Ljubljana: Faculty of Computer and Information Science, 2005, 450 s ISBN 9789616209526

[5] SUTTON, Richard S - BARTO, Andrew G. – BACH, Francis – *Reinforcement Learning: An Introduction*. Cambridge MA: MITPress, 1998, 322 s.

[6] WITTEN, Ian H. – EIBE, Frank – HALL, Mark A. – *Data mining: Pracital Machine Learning Tools and Techniques*. 2.vyd. San Francisco: Morgan Kaufmann, 2005, 560 s. ISBN 978-0120884070

[7] WITTEN, Ian H. – EIBE, Frank – HALL, Mark A. – *Data mining: Pracital Machine Learning Tools and Techniques*. 3.vyd. San Francisco: Morgan Kaufmann, 2011, 664 s. ISBN 978-0123748560

[8] ZHU, Xiaojin – GOLDBERG, Andrew B. – *Introduction to Semi-Supervised Learning*.Synthesis Lectures on Artificial Intelligence and Machine Learning. 2009, 130 s. ISBN 978-1598295474

[9] ZORMAN, Milan et al – *Inteligentni sistemi in profesionalni vsakdan.* Maribor: Univerza, 2003, 220 s. ISBN 9788643505731

Elektronické zdroje:

[10] BROWNLEE, Jason - A Gentle Introduction to k-fold Cross-Validation [online]. 2018. [25.04.2021]. Dostupné na: https://machinelearningmastery.com/k-fold-cross-validation/

[11] FISHER, Ronald A. - *Iris plants Database*.[online] 1988[04.04.2021]. Dostupné na: https://www.kaggle.com/uciml/iris

[12]GATree. *GATree Home*. 2002. [04.04.2021]. Dostupné na: http://www.gatree.com/
[13] GATree . *Usage tutorial GATree*. 2002.[04.04.2021]. Dostupné na: http://www.gatree.com/?page_id=6 [04.04.2021]
[14] GAYATHRI, Vaidyanathan - SARAVAN, N - Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm. [online] . In: International Journal of Computational Intelligence and Informatics, Vol. 7. 2018. 11s. [25.04.2021]. Dostupné na:
[15] GERMAN, B. *Glass Identification Database*.[online] 1987.[04.04.2021] Dostupné na: https://www.kaggle.com/uciml/glass
[16] HALL, Mark – REUTEMANN, Peter. *Weka KnowledgeFlow Tutorial for Version 3-5-8*.[online] 2008 [03.04.2021]. 15s. Dostupné na: http://software.ucv.ro/~eganea/AIR/KnowledgeFlowTutorial-3-5-8.pdf
[17] ILTER, Nilsel – ALTAY, Guvenir H. - *Dermatology Database*. [online] *1998* [04.04.2021]. Dostupné na: https://www.kaggle.com/syslogg/dermatology-dataset
[18]J48 Decision Tree. [25.04.2021]. Dostupné na: http://data-mining.business-intelligence.uoc.edu/home/j48-decision-tree
[19] MACHOVÁ, Kristína – *Strojové učenie, princípy a algoritmy*. [online] 2002 [26.02.2021]. 125 s. Dostupné na: http://people.tuke.sk/kristina.machova/pdf/SU4.pdf
[20] MAULANA, Mohamad F. – DEFRIANI, Meriska - *Logistic Model Tree and Decision Tree J48. Algorithms for Predicting the Length of Study Period*.[online] 2020. [25.04.2021]. ISSN 2620-3553. Dostupné na: https://www.researchgate.net/publication/340504096_Logistic_Model_Tree_and_Decision_Tree_J48_Algorithms_for_Predicting_the_Length_of_Study_Period/link/5e8dc023a6fdcca789fe069a/download
[21] REFAEILZADEH, Payam – TANG, Li – LIU, Huan - *Cross-Validation, Synonyms Rotation Estimation.*[online] [26.02.2021]. 6s. Dostupné na: https://docplayer.net/2885209-Cross-validation-synonyms-rotation-estimation.html
[22]Zentut. Data Mining Techniques [2021-04-01]. Dostupné na: http://www.zentut.com/data-mining/data-mining-techniques/