УНИВЕРЗИТЕТ У БЕОГРАДУ

МАТЕМАТИЧКИ ФАКУЛТЕТ

Милана Грбић

Груписање организама помоћу различитих метода класификације у зависности од генотипских и фенотипских каракеристика

-мастер рад-

Београд, 2016.

Подаци о ментору и члановима комисије

Ментор

др Ненад Митић, ванредни професор, Математички факултет, Универзитет у Београду

Чланови комисије

др Гордана Павловић-Лажетић, редовни професор, Математички факултет, Универзитет у Београду

др Ненад Митић, ванредни професор, Математички факултет, Универзитет у Београду

др Милош Бељански, научни савјетник, Институт за општу и физичку хемију, Београд

Садржај

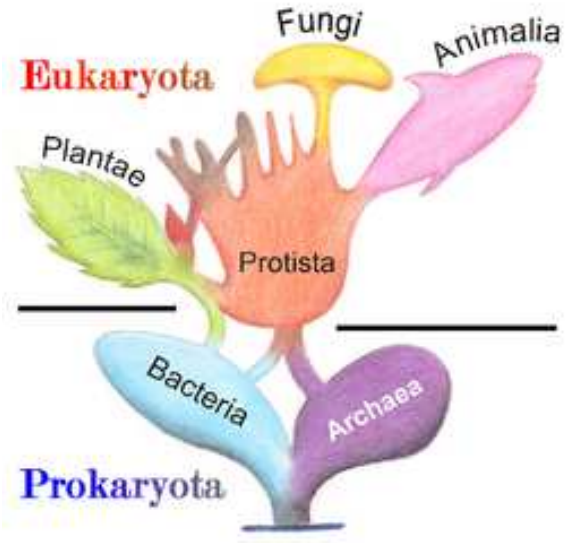
Глава 1

Увод

Количина података који се чувају у разним биоинформатичким базама података широм свијета расте великом брзином. Извлачење закључака из ових података захтијева софистициране рачунарске анализе. Биоинформатика је интердисциплинарна наука тумачења биолошких података помоћу информационих технологија и рачунараских наука. Значај ове науке расте из дана у дан управо због све веће количине података који се свакодневно проналазе и чувају у разним базама података. Посебно активна област истраживања у биоинформатици је примјена и развој техника истраживања података за рјешавање биолошких проблема. Анализирањем великих скупова биолошких података могу се утврдити опште особине или установити специфичности појединих структура [7]. Неки од примјера примјене истраживања података у биоинформатици су: налажење група гена који имају сличне структурално/функционалне особине, класификација ћелија тумора као бенигних или малигних, класификација секундарне структуре протеина и сл.

Генотипска карактеристика организма је заправо генски састав одређеног организма, док фенотипска карактеристика је видљива/уочљива особина која је резултат комбинације гена и утицаја животне средине [3].

У овом раду је представљена примјена методе класификације у циљу груписања организама у зависности од њихових генотипских и фенотипских карактеристика. Рјешавање овог проблема је важно, нарочито у случају класификације појединих потенцијално патогених организама.



Слика 1.1: Прокариотски организми

1.1 Прокариотски организми

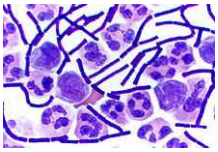
Прокариоти су једноћелијски микроорганизми који немају једро. Постоје двије врсте прокариота - бактерије и археје. Зидови бактеријских ћелија се састоје од пептидогликана муреина, али постоје и бактерије које немају ћелијске зидове. Ћелијски зидови археја не садрже муреин већ су састављени од других полимера.

Већина бактерија и археја су знатно мање од еукариотских ћелија. Живе самостално или у паровима, ланцима и кластерима (гроздовима, групама) у скоро сваком станишту које има довољно влаге. Између осталих станишта, археје се налазе у екстремним окружењима попут сланог језера у Моно Округу у Калифорнији, киселим изворима топле воде у националном парку Јелоустон и у блату, на дну мочваре, у којем нема пуно кисеоника [1].

Бактерије су, поред гљива, једини разлагачи органских материја и имају велику примјену у индустрији.

1.1.1 Бактерије

Бактерије су прокариотски организми и сматра се да су међу најбројнијим организмима на свијету. Бактерије могу живјети и у аеробним и анаеробним условима. Грађу свих бактеријских ћелија чине ћелијска мембрана и цитоплазма, у којој се налазе рибозоми и нуклеоид. Већина бактерија има и ћелијске



Слика 1.2: Бојење по Граму

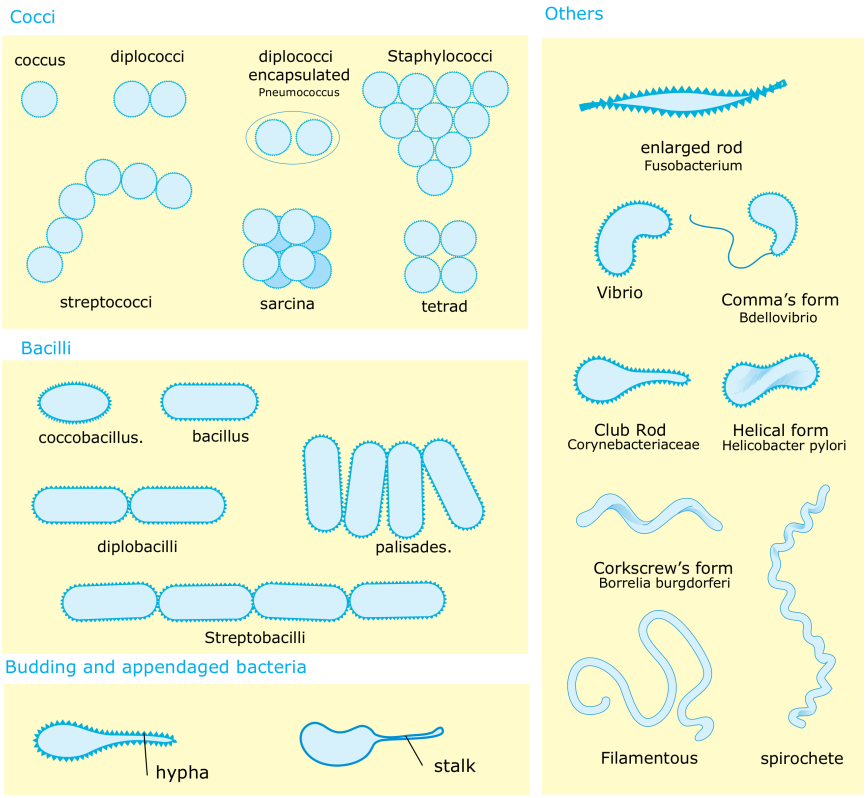
зидове, али постоје и оне које немају (микоплазме и рикеције). Оштећење ћелијског зида доводи до смрти бактерије [4]. Према саставу ћелијског зида и поступку бојења по Граму, бактерије се дијеле на Грам-позитиве и Грам-негативне. Грам-негативне бактерије имају слој липополисахарида који покрива њихов ћелијски зид, док Грам-позитивне немају тај слој, због чега се прве по Граму боје у црвено, а друге у љубичасто. Утврђено је да се због грађе ћелијског зида, Грам-позитивне бактерије лакше уништавају антибиотицима, док су Грам-негативне много отпорније. Поједине врсте бактерија поред наведених дијелова могу да садрже и капсулу, бичеве, фимбрије, тилакоиде и плазмиде. Капсула је слузави, спољашњи омотач који ствара сама бактерија и који штити бактерију од дејства одбрамбеног система организма у којем се налази. Фимбрије су кончићи око тијела бактерије, које ствара сама бактерија и који су протеинске природе, а служе за причвршћивање за подлогу или за међусобно спајање двије јединке при размножавању. Бичеви су дуги, танки израштаји изграђени од протеина флагелина помоћу којих се бактерије крећу. Када изгубе бичеве, бактерије постају непокретне. Тилакоиде посједују бактерије које могу да обављају фотосинтезу - цијанобактерије. Плазмиди су мали прстенасти молекули ДНК који се налазе изван хромозома и дуплирају се независно од њега.

Разликују се три основна облика бактерија:

1. Коке су лоптасте бактерије. Појединачне коке називају се монококе, а удружене су диплококе (две спојене коке), стрептококе (у виду ланца), стафилококе (у облику грозда), тетраде (пакетић од 4 ћелије) и сарцине (пакетић од 8 ћелија).

2. Штапићасте бактерије које образују споре су бацили. Удружени граде диплобациле (два бацила један до другог) и стрептобациле (у виду низа).

3. Спиралне бактерије могу имати облик спирале и онда се називају спирили



Слика 1.3: Различити облици бактерија

(ако имају мањи број благих завоја), спирохете (ако имају већи број оштрих завоја) или, ако су у облику зареза, вибриони.

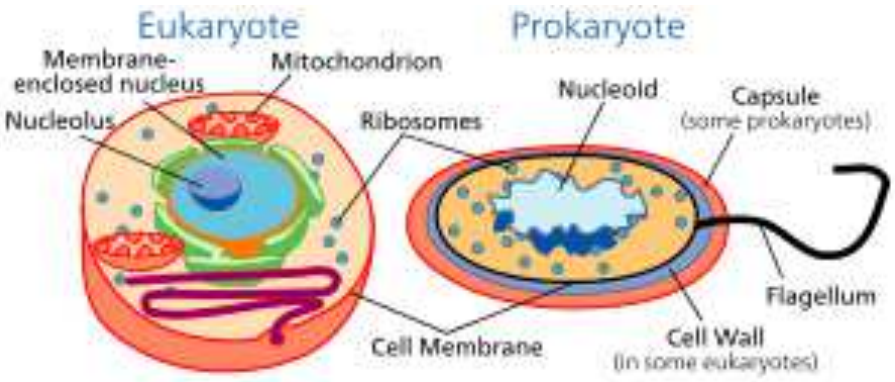
Неки од облика бактерија приказани су на слици 1.3.

Коке су непокретне бактерије, док су штапићасте бактерије покретне јер имају бичеве, издуженог су облика и имају заобљене крајеве.

Патогени организми су они који могу изазвати одређена обољења. Специфични су за посебну врсту домаћина и посебну врсту ткива. Неке врсте бактерија уништавају ћелије свог домаћина. Међутим, највећи број врста бактерија производи токсине (отрове) који наносе штету метаболизму ћелије домаћина [5], [1].

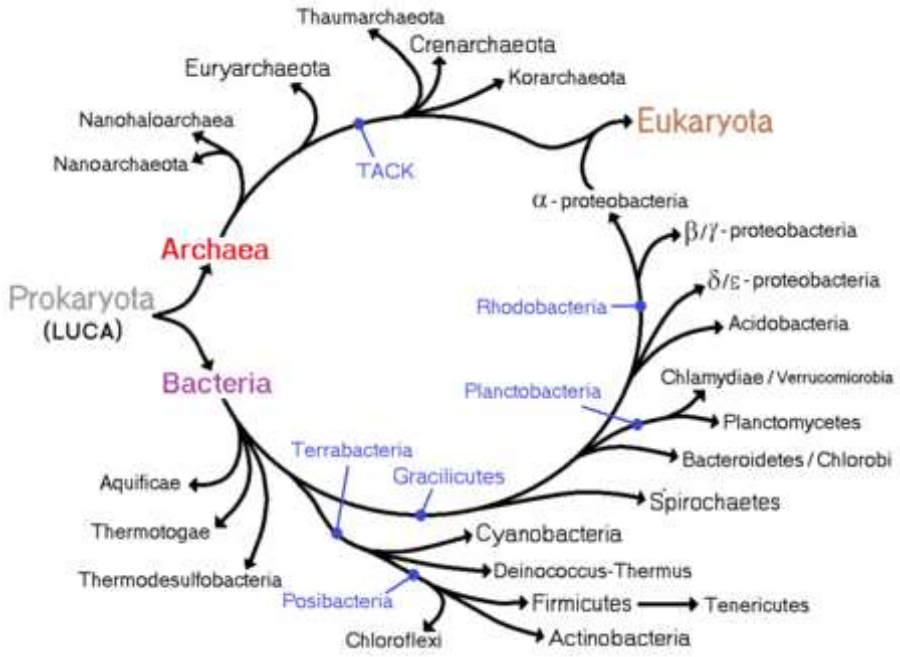
1.1.2 Археје

Археје су прокариотски организми који имају основне цитолошке карактеристике као и бактерије. Разлике између бактерија и археја испољавају се тек на



Слика 1.4: Грађа ћелије прокариотских и еукариотских организама

молекуларном нивоу. Разлике се прије свега огледају у биохемијском саставу ћелијског зида (не садрже пептидогликан муреин) и цитоплазмичне мембране, као и у неким ензимима. Најбољи доказ да су археје филогенетски изоловане је да насељавају термалне воде Пацифика. Првобитно су налажене у екстремним стаништима попут термалних вода, гејзира, веома сланих вода, анаеробних мочвара и подводних вулкана. Карл Воуз (*Carl Woese*) 1977. године је издвојио посебну групу под називом *Archaebacteria*, јер је због екстремних станишта сматрао да су то организми који су старији од бактерија. Међутим, 1990. године је, заједно са Фоксом (*George E. Fox*), установио да је назив неадвекатан и преименовао их у Archaea. Данас су археје пронађене и на многим уобичајним стаништима, а нарочито у водама океана. Различите археје имају различите морфолошке и физиолошке карактеристике. Боје се Грам-позитивно и Грам-негативно. По облику су округле, штапићасте, извијене и полиморфне. Пречник ћелије им је 0,1 до 15μm, а неки кончасти представници могу бити дугачки и до 200μm. Размножавају се диобом, пупљењем и фрагментацијом. У односу на кисеоник могу бити аероби, анаероби и факултативни анаероби. Начин исхране може бити аутотрофан, литотрофан и хетеротрофан. Највећи број ових микроорганизама су хипертермофили, а мали број припада мезофилима. Живе у анаеробним, веома заслањеним и топлим срединама. Они чине око 34% од укупне биомасе прокариота у водама Антартика. За сада нису познате археје које су патогене или које су паразити [5], [1].



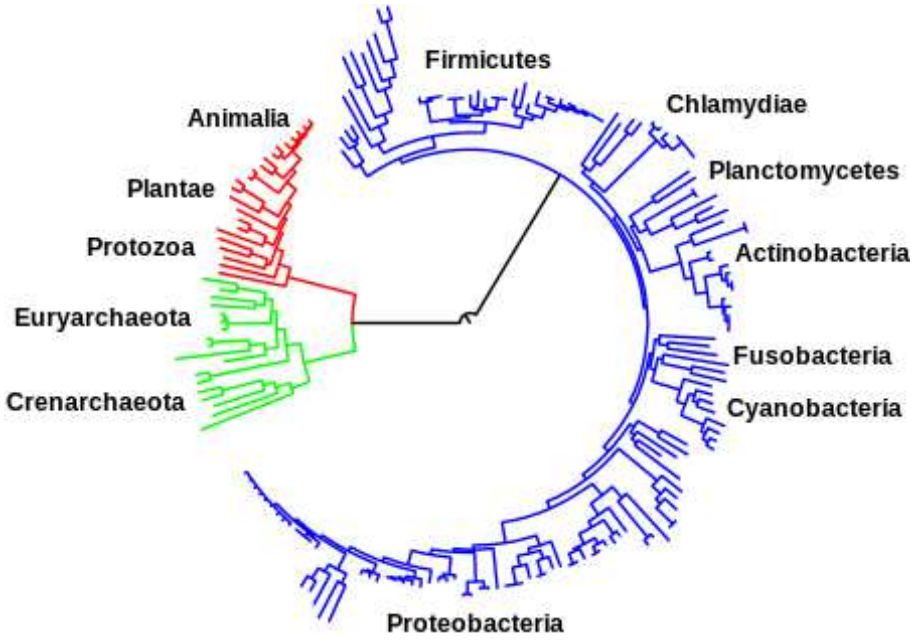
Слика 1.5: Класификација бактерија и археја

1.2 Опис проблема и циљ рада

Циљ рада је да се на основу генотипских и фенотипских карактеристика организама дође до неког новог груписања прокариотских организама. Метода истраживања података која је коришћена у раду је класификација. Подаци на којима је вршена класификација су бактерије и археје. Неке од постојећих класификација ових организама су приказани на сликама 1.5 и 1.6.

Класификација је урађена примјеном више различитих алгоритама. Примјена више алгоритама је потребна због тога што се они различито понашају, односно дају различит квалитет резултата, у зависности од типова података на које се примјењују, величине скупа података и присутности/одсутности података. Класификација је вршена помоћу пакета *IBM InfoSphere Warehouse Intelligent Miner* (http://www.ibm.com/developerworks/data/library/tutorials/iminer/-iminer.html), *Weka* (http://www.cs.waikato.ac.nz/ml/weka/), *Knime* (https://www.-knime.org/) и *IBM SPSS Statistics* (http://www.ibm.com/analytics/us/en/technology/spss/).

У глави 2 су описане методе класификације, док је у глави 3 описана база података над којом је примјењен метод из главе 2, односно описано је значење



Слика 1.6: Класификација бактерија и археја

података који се налазе у бази. У глави 4 су приказани добијени резултати, урађена је упоредна анализа добијених резултата и разматрани су модели који су дали најбоље резултате при урађеним класификацијама. Затим, у глави 5 је предложен најбољи модел и алгоритам за класификацију разматраног скупа организама и приказан даљи план рада. На крају, у додатку у глави 6 налазе се информације које детаљно описују број, типове и могуће вриједности података у бази, као и дио табела које садрже резултате класификације.

Глава 2

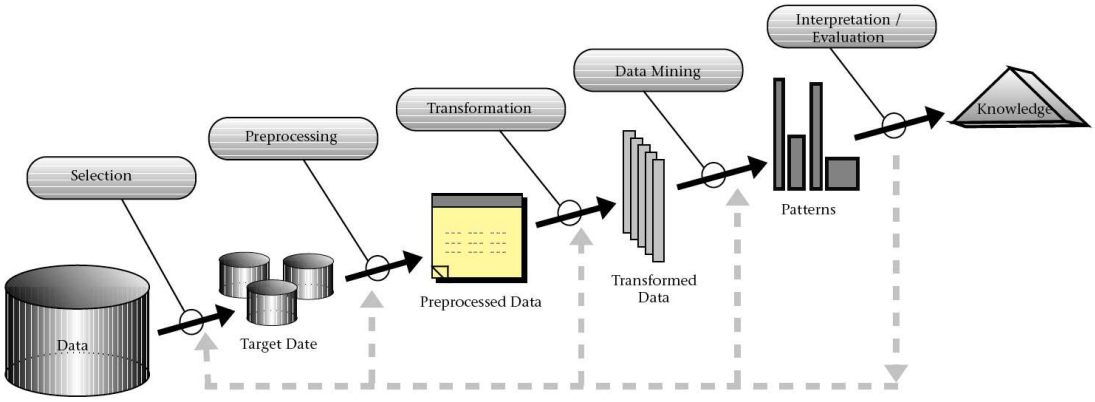
Методе класификације у истраживању података

2.1 Појам истраживања података

Истраживање података се најчешће дефинише као проналажење скривених информација у бази података. Односно, као издвајање претходно непознатих, а потенцијално корисних информација из базе података. Формално, истраживање података је интегрални дио откривања знања у базама података (енгл. *Knowledge Discovery in Databases, KDD*), што је назив за цјелокупни процес претварања равних података у корисне информације.

Често се у базама података налазе "скривене" информације које се не уочавају одмах или које нису лако уочљиве. Аналитичарима је потребно много времена да уоче правилности између података, а традиционалним методама се велики дио података често уопште не анализира, поготово ако су у питању равни подаци. То су само неки од разлога зашто је дошло до развоја области истраживања података.

Такође, бројне су примјене резултата добијених процесом истраживања података. На примјер, у великим пословним кооперацијама из дана у дан расте количина података које је потребно обрадити. Из саме обраде података настоји се добити што квалтетнија информација, која може бити предност у односу на конкуренцију. У науци, медицини и инжињерству се такође свакодневно прикупљају подаци, неке научне симулације генеришу терабајте података који се користе у даљим истраживањима и открићима. Наравно, потребан је начин да



Слика 2.1: Процес откривања знања у базама података

се из тих података открију нека нова знања.

Методе истраживања података се могу подјелити у двије групе:

1. Предиктивне методе
2. Дескриптивне методе

Предиктивне методе предвиђају вриједност циљног атрибута (својство или карактеристика објекта) на основу вриједности осталих атрибута. Односно, предиктивне методе праве модел који је функција осталих атрибута и на основу којег се предвиђа вриједност циљног атрибута. У групу предиктивних метода спадају класификација, регресија, предвиђање и анализа временских серија. Методом класификације предвиђа се вриједност циљног атрибута, који има коначан или пребројиво бесконачан скуп вриједности, тј. циљни атрибут је дискретан. С друге стране, методом регресије се предвиђа вриједност циљног атрибута, чији скуп вриједности је скуп реалних бројева, тј. циљни атрибут је континуалан (непрекидан). Предвиђање је вид класификације којим се прогнозира будуће стање на основу прошлих и садашњих стања. Анализа временских серија истражује промјене вриједности атрибута кроз вријеме.

Дескриптивне методе настоје пронаћи обрасце који описују односе између података. У дескриптивне методе се убрајају кластеровање, сумаризација,

правила придруживања и анализа редослиједа. Кластеровањем се слични подаци (слични у односу на одговарајуће атрибуте) групишу заједно у групе. За разлику од класификације, која је учење под надзором јер су циљне класе унапријед познате, кластеровање је учење без надзора јер број и особине група нису унапријед одређени. Сумаризација пресликава податке у подскупове са придруженим једноставним описима. Правила придруживања откривају обрасце који описују међусобно чврсто повезане особине података. Анализа редослиједа се користи за одређивање образаца у подацима који зависе од редослиједа појављивања.

2.2 Метода класификације

У овом поглављу је детаљно описан метод класификације. Наведени описи су највећим дијелом засновани на [9], а поред тога коришћени су [2] и [8].

Класификација, чији задатак је придруживање једној од неколико унапријед одређених категорија, је распрострањен проблем који се појављује у бројним ситуацијама. На примјер, проблемом класификације можемо сматрати одређивање да ли је пристигло писмо електронском поштом спам или није на основу његовог наслова и садржаја, као и доношење одлуке о томе да ли је ћелија тумора малигна или бенигна на основу MRI скенерског снимка, препознавање галаксија на основу њиховог облика, итд.

2.2.1 Основни појмови

Улазни податак у класификацију је скуп података. Сваки податак, инстанца или слог, је одређен уређеним паром (X, y), гдје је X скуп атрибута, а y циљни атрибут. Класификацијом ће бити одређена функција која зависи од атрибута из скупа X, а помоћу које се за дати објекат може одредити вриједност циљног атрибута y, тј. може се одредити којој циљној класи припада. У табели 2.1 је приказан скуп атрибута који се користи за класификацију кичмењака у неку од класа: сисари, птице, рибе, гмизавци или водоземци. Скуп атрибута укључује особине кичмењака као што су температура тијела, омотач тијела, начин рађања, способност летења и да ли може да живи у води. Иако су атрибути у табели дискретни, скуп атрибута може да садржи и непрекидне

(континуалне) атрибуте. Међутим, циљни атрибут, односно атрибут који представља циљну класу, мора бити дискретан. Основна разлика између класификације и регресије је то што при регресији циљни атрибут треба да буде непрекидан.

Назив	Темп. тијела	Кожни омотач	Да ли се рађа живо?	Живи у води	Лети	Има ноге	Хиберна-ција	Класа
Људи	топло-крвни	длаке	да	не	не	да	не	сисар
Питон	хладно-крвни	рожни покривач	не	не	не	не	да	гмизавац
Лосос	хладно-крвни	рожни покривач	не	да	не	не	не	риба
Жаба	хладно-крвни	нема	не	да/не	не	да	да	водоземац
Шишмиш	топло-крвни	длаке	да	не	да	да	да	сисар
Голуб	топло-крвни	перје	не	не	да	да	не	птица
Мачка	топло-крвни	крзно	да	не	не	да	не	сисар
Корњача	хладно-крвни	рожни покривач	не	не	не	не	не	гмизавац
Пингвин	топло-крвни	перје	не	да/не	не	да	не	птица
Јегуља	хладно-крвни	рожни покривач	не	да	не	не	не	риба

Табела 2.1: Подаци о кичмењацима

Дефиниција 1. *Класификација је проналажење циљне функције f која сваки скуп атрибута X пресликава у једну од циљних класа y.*

Циљна функција се неформално назива **модел класификације**.

Модел класификације може послужити као објашњење разлика између објеката различитих класа. На примјер, за биологе би било корисно да имају описни модел који ће сумирати податке из табеле 2.1 и објаснити које карактеристике дефинишу кичмењаке као сисаре, рибе, птице, водоземце или гмизавце.

Модел класификације се може користити за предвиђање циљне класе за скуп података за који класа није позната. Нека су у табели 2.2 дате особине бића познатог под именом фламинго.

Можемо користити модел класификације направљен на основу скупа података из табеле 2.1 да одредимо ознаку класе којој припада фламинго.

Назив	Темп. тијела	Кожни омотач	Да ли се рађа живо?	Живи у води	Лети	Има ноге	Хиберна-ција	Класа
Фламинго	топло крвни	крзно	не	не	да	да	не	?

Табела 2.2: Подаци о фламингу

Метод класификације је погодан за предвиђање вриједности или описивање односа података са бинарним и номиналним атрибутима. Мање је ефикасан ако се примјењује на податак чији атрибути су ординални (нпр. класификација особе као члана више, средње или ниже класе), јер не разматра уређеност између категорија. Други облици веза између категорија, као нпр. подкласе или надкласе (нпр. људи и мајмуни су примати, што је подкласа сисара) се такође игноришу.

2.2.2 Процес класификације

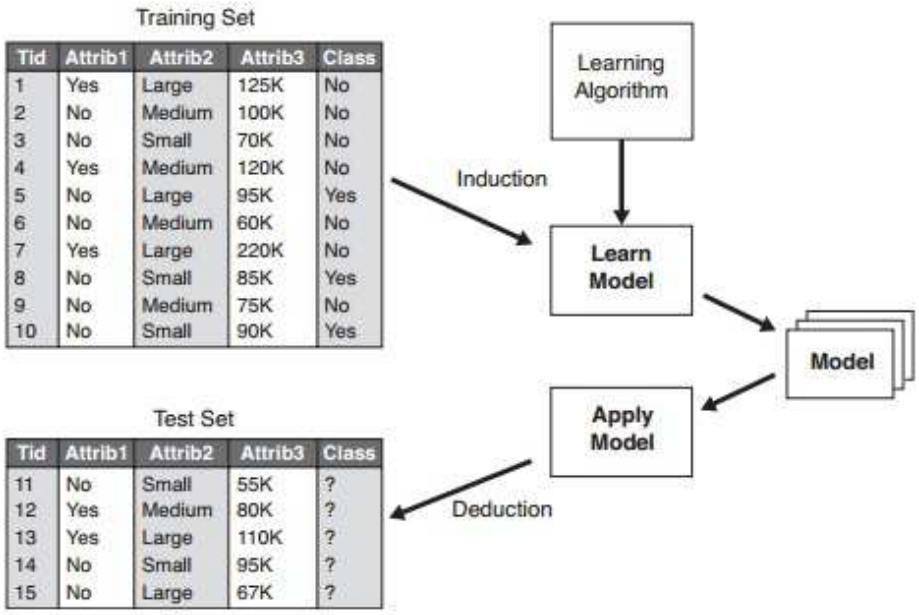
Метод класификације (тј. класификатор) је системски приступ изградње модела класификације на основу улазног скупа података. Неке од најчешће коришћених техника класификације су:

1. Методе засноване на дрветима одлучивања
2. Методе засноване на правилима
3. Неуронске мреже
4. Статистички засноване методе
5. Методе засноване на подржавајућим векторима
6. Наивни Бајесов класификатор

Свака техника користи **алгоритам учења** да одреди модел који најбоље описује везу између атрибута и ознаке класе улазних података. Модел генерисан алгоритмом учења поред тога што треба да коректно класификује улазне податке, треба да што прецизније одређује ознаку класе за њему претходно непознате податке. Дакле, основни циљ алгоритма учења је да генерише класификатор који има способност генерализације, тј. модел који тачно предвиђа ознаку класе за претходно непознате податке.

На слици 2.2 приказан је процес класификације. Улазни подаци се дијеле у два дијела:

1. **Податке за тренинг**, помоћу којих се формира модел
2. **Податке за тестирање**, који се користе за провјеру исправности модела



Слика 2.2: Илустрација процеса класификације

Мјерење перформанси израчунавања модела заснива се на броју коректно и некоректно класификованих тест података тим моделом. Број коректно и некоректно класификованих тест података се представља табелом, која се назива **матрица конфузије**. Табела 2.3 приказује матрицу конфузије за проблем бинарне класификације. Класификација је бинарна ако класификује податке у двије циљне класе. Сваки f_{ij} у табели представља број података класе i за које класификатор предвиђа да су класе j. Нпр. f_{01} је број података класе 0 који се моделом класификације (некоректно) класификује у класу 1. На основу матрице конфузије, можемо одредити број коректно и некоректно класификованих инстанци. Број коректно класификованих инстанци је $f_{00} + f_{11}$, док некоректно класификованих инстанци има $f_{01} + f_{10}$.

Матрица конфузије садржи информације помоћу којих се могу одредити перформансе модела класификације, а затим на основу перформанси је могуће

		Предвиђена класа	
		класа=1	класа=0
Актуелна класа	класа=1	f_{11}	f_{10}
	класа=0	f_{01}	f_{00}

Табела 2.3: Матрица конфузије

поредити различите моделе класификације. Најчешће коришћена метрика за мјерење перформанси система је **тачност**, која се дефинише на сљедећи начин:

$$\text{Тачност} = \frac{\text{број тачно класификованих инстанци}}{\text{укупан број инстанци}} = \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{01} + f_{10}}$$

С друге стране, перформансе модела могу бити изражене и **степеном грешке**, који се дефинише на сљедећи начин:

$$\begin{aligned}\text{Степен грешке} &= \frac{\text{број погрешно класификованих инстанци}}{\text{укупан број инстанци}} \\ &= \frac{f_{10} + f_{01}}{f_{00} + f_{11} + f_{01} + f_{10}}\end{aligned}$$

Многи алгоритми класификације траже модел који постиже што већу тачност, односно што мању грешку на тест подацима.

2.2.3 Процес класификације помоћу дрвета одлучивања

Једна од најчешће коришћених техника класификације је **дрво одлучивања**.

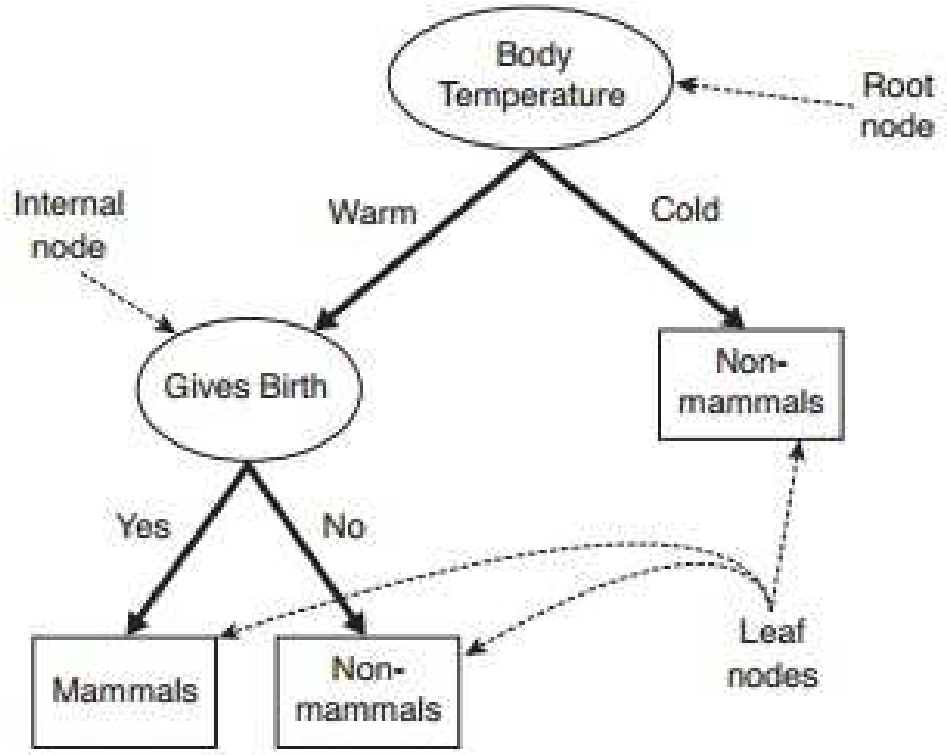
2.2.3.1 Примјена дрвета одлучивања

Да бисмо илустровали како ради дрво одлучивања, посматрајмо једноставан проблем класификације кичмењака из претходног поглавља. Умјесто да кичмењаке класификујемо у пет класа, вршићемо класификацију само у двије класе: сисари и нису-сисари.

Претпоставимо да су научници открили нову врсту. Како ће одлучити да ли је та врста сисар или ипак није сисар? Један од приступа може бити постављање низа питања о карактеристикама те врсте. Прво питање може бити да ли је топлокрвно или хладнокрвно биће. Ако је хладнокрвно, онда сигурно није сисар. У супротном, или је сисар или птица, па постављамо сљедеће питање: Да ли се рађају живи? Ако је одговор "да" онда је сисар, у супротном није

сисар. Сви сисари, осим два изузетка кљунар и спини мравојед, рађају се живи.

Претходни примјер показује како можемо ријешити проблем класификације постављањем низа пажљиво одабраних питања о атрибутима тест података. Након добијеног одговора, слиједи сљедеће питање, све док не закључимо ознаку које класе треба придружити том податку. Скуп питања и њихови могући одговори могу бити организовани у форми дрвета одлучивања, које је хијерархијска структура која са састоји од чворова и грана.



Слика 2.3: Дрво одлучивања за проблем класификације сисара

На слици 2.3 приказано је дрво одлучивања за проблем класификације сисара. Дрво садржи три врсте чворова:

1. **Коријени чвор** (енгл. *root node*) је чвор који нема улазних грана и има нула или више излазних грана.

2. **Унутрашњи чвор** (енгл. *internal node*) је чвор који има тачно једну улазну грану и двије или више излазних грана.

3. **Лист чвор** (енгл. *leaf node*) је чвор који има тачно једну улазну грану

и нема излазних грана. Назива се још и **завршни чвор** (енгл. *terminal node*).

Сваком листу у дрвету одлучивања придружена је ознака неке од циљних класа. Чворови који нису завршни, односно коријен и унутрашњи чворови, садрже услове којима се испитују атрибути и на основу којих се врши раздвајање података који имају различите карактеристике. Нпр. коријени чвор дрвета, које је приказано на слици 2.3 користи атрибут *температура тијела* (енгл. *Body Temperature*) да раздовји топлокрвне и хладнокрвне сисаре. С обзиром да хладнокрвност није особина сисара, лист чвор означен са *Није-сисар* (енгл. *Non-mammlas*) је постављен као десно дијете коријеног чвора. Ако је кичмењак топлокрван, користи се сљедећи атрибут *Да ли се рађа живо?* (енгл. *Gives Birth*) да се раздвоје сисари од осталих топлокрвних кичмењака, углавном птица.

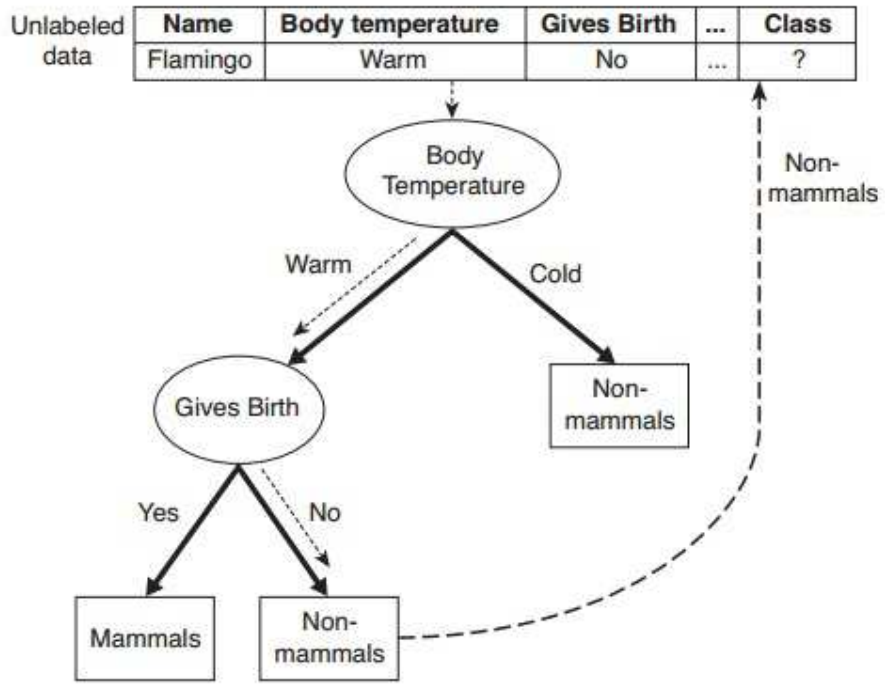
Након формирања дрвета одлучивања, класификација тестних података је праволинијска. Почевши од коријена дрвета, примјењујемо услове теста на податак и пратимо грану која одговара резултату теста. На тај начин долазимо или до сљедећег унутрашњег чвора, за који тестирамо нови услов, или до листа. Ознака класе која се налази у листу се придружује податку. Примјеном дрвета одлучивања на фламинго закључујемо да припада класи *Није сисар*, слика 2.4.

2.2.3.2 Како формирати дрво одлучивања?

За дати скуп атрибута може бити изграђено више дрвета одлучивања. С обзиром да нека дрвета имају већу тачност него остала, претраживање цијелог простора могућих дрвета због величине простора је неизводљиво. Међутим, развијени су ефикасни алгоритми који проналазе дрво одлучивања прихватљиве тачности у разумном временском периоду. Ови алгоритми користе стратегију похлепе (грабљивости) да би подјелили слогове према тестном атрибуту који оптимизује одређени критеријум. Један такав алгоритам је Хантов алгоритам, који се налази у позадини многих алгоритама који индукују дрвета одлучивања. Хантов алгоритам се налази у основи алгоритама ID3, CART и C4.5.

Хантов алгоритам

Хантовим алгоритмом дрво одлучивања расте рекурзивно подјелом тренинг



Слика 2.4: Примјена дрвета одлучивања

података у што "чистије" подскупове. Нека је D_t скуп слогова за тренинг који се налазе у чвору t и нека је $y = \{y_1, y_2, ..., y_c\}$ скуп ознака класа. Рекурзивна дефиниција Хантовог алгоритма је:

- **Корак 1**: Ако сви слогови из скупа D_t припадају истој класи y_t, онда се листу t додјељује ознака класе y_t.
- **Корак 2**: Ако скуп D_t садржи слогове који се налазе у више од једне класе, тада се користи тест атрибут да би се извршила подјела података у мање подскупове. За сваки подскуп формира се дијете чвор, на који се рекурзивно примјењује комплетна процедура.

Због илустрације рада алгоритма, посматрајмо проблем предвиђања да ли ће подносилац захтјева за кредит вратити кредит на вријеме или то неће урадити благовремено. Тренинг скуп за овај проблем може бити формиран на основу података о претходним дужницима. На слици 2.5 су приказани подаци који садрже личне податке о дужнику заједно са ознаком класе да ли је на вријеме вратио кредит.

Почетно дрво одлучивања, које је приказано на слици 2.6(а), садржи само

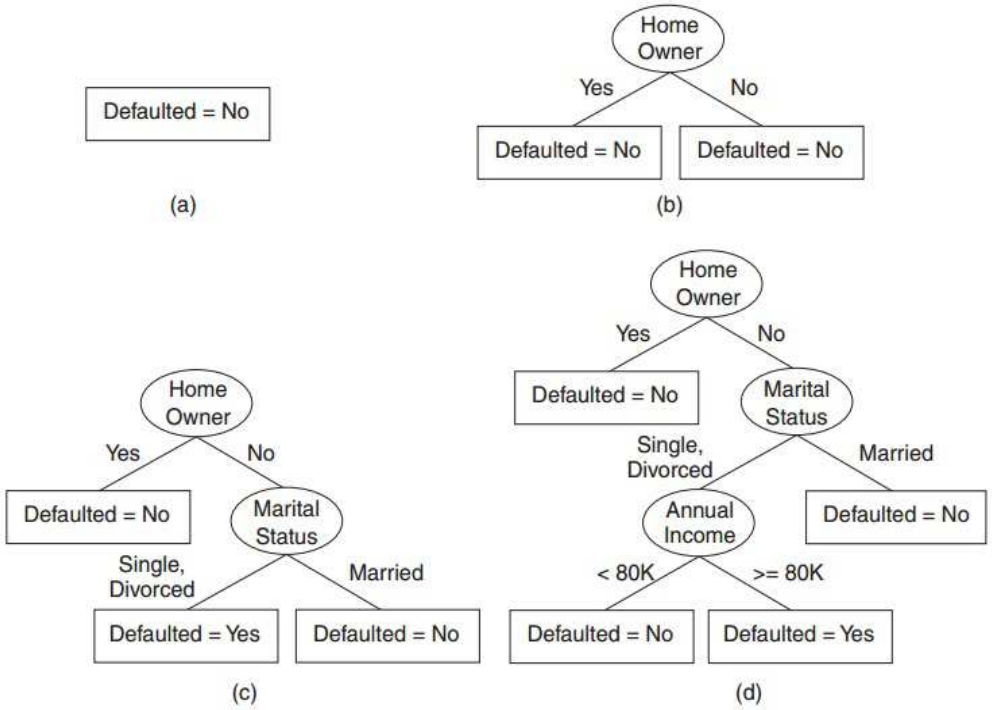
	binary	categorical	continuous	class
Tid	**Home Owner**	**Marital Status**	**Annual Income**	**Defaulted Borrower**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Слика 2.5: Тренинг подаци за налажење дрвета одлучивања о дужницима

један чвор са ознаком класе *Defaulted=No*, што значи да је већина дужника успјешно вратила кредит. Међутим, дрво мора бити редефинисано јер садржи слогове који припадају и једној и другој класи. Затим, подаци су подјељени на два подскупа на основу тест атрибута *Home Owner* (слика 2.6(b)). Зашто је изабран баш овај атрибут биће објашњено касније, за сад претпоставимо да је то најбољи критеријум за подјелу овог чвора. Хантов алгоритам примјењујемо рекурзивно на сваки дијете чвор. Из тренинг скупа са слике 2.5 уочавамо да су сви власници кућа/станова (тј. за које је *Home Owner=Yes*) успјешно вратили кредит, па самим тим ознака класе која се придружује лијевом дијетету коријена је *Defaulted=No* (слика 2.6(b)). За десно дијете настављамо са рекурзивном примјеном Хантовог алгоритма све док не дођемо до скуп чији подаци припадају истој класи. Тако добијена поддрвета су приказана на слици 2.6(c,d)).

Хантов алгоритам ће радити ако је у тренинг скупу присутна свака комбинација атрибута и ако за сваку комбинацију атрибута постоји јединствена ознака класе. Ове претпоставке су сувише јаке да би биле присутне у свим могућим случајевима. У сљедећим случајевима су потребни додатни услови:

1. Могуће је да неки од дијете чворова креираних у кораку 2 буде празан; тј.



Слика 2.6: Хантов алгоритам за извођења дрвета одлучивања

да нема података који су придружени том чвору. Ово се може десити ако ниједан од тренинг података нема комбинацију атрибута која је придружена том чвору. У том случају чвор се декларише као лист са ознаком класе којој припада већина података која је придружена родитељском чвору.

2. Може се десити, у кораку 2, да сви подаци из скупа D_t имају идентичне вриједности атрибута осим ознаке класе, па их је немогуће даље подјелити. У овом случају, чвор се декларише као лист са ознаком класе којој припада већина података придружених том чвору.

Остаје још да се разјасне два питања приликом изградње дрвета одлучивања:

1. **Како подјелити тренинг скуп на два подскупа?** Односно како у сваком рекурзивном кораку изабрати тест атрибут који ће подјелити тренинг скуп на два мања подскупа. Поставља се питање како навести услове за тестирање атрибута и како изабрати најбољу подјелу.

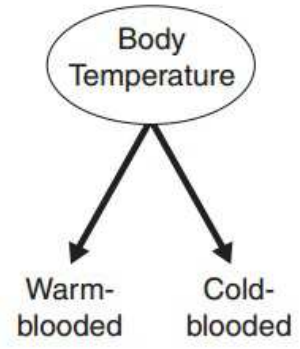
2. **Када стати са подјелом?** Услов за заустављање је неопходан, јер је у

неком моменту потребно стати са формирањем дрвета одлучивања. Једна од могућих стратегија је да се врши подјела све док сви подаци не припадају истој класи или док сви подаци немају исте вриједности атрибута. Иако су оба услова довољна да се процес изградње дрвета заврши, некад се процес може и раније прекинути под неким специјалним условима.

2.2.3.3 Начин преказивања тест атрибута

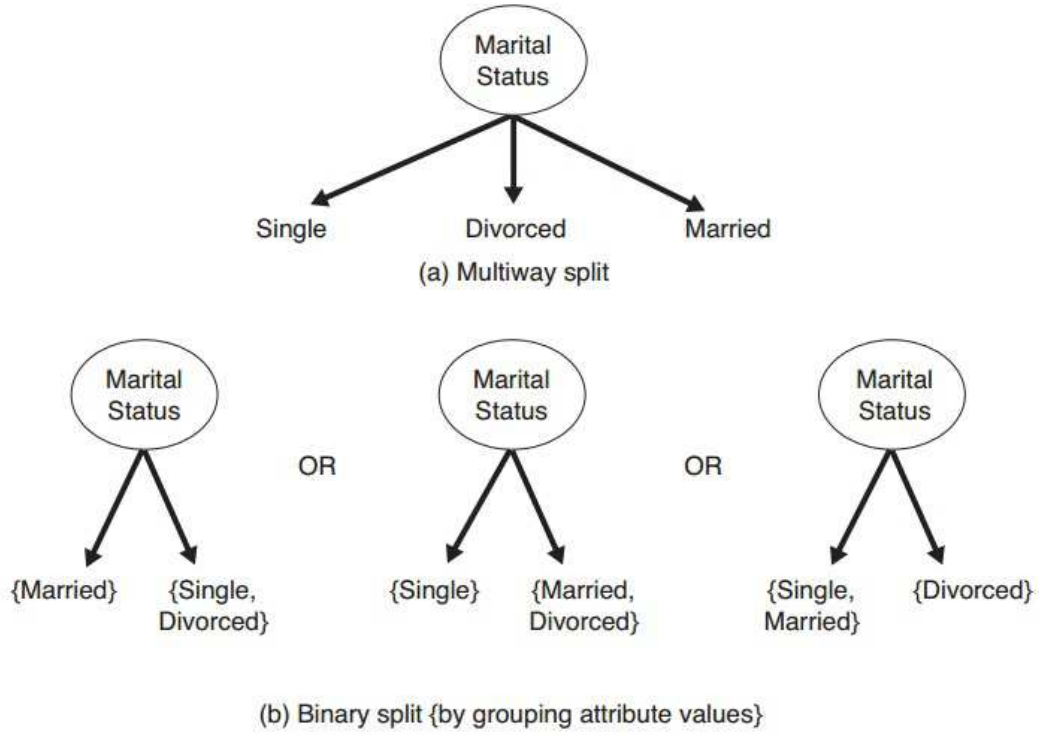
Алгоритми који индукују дрвета одлучивања треба да обезбједе начин приказивања тест атрибута, који ће бити у складу са типовима атрибута.

Бинарни атрибути Ако је тестни атрибут бинарни, онда имамо два могућа резултата (слика 2.7).



Слика 2.7: Бинарни атрибут као тестни

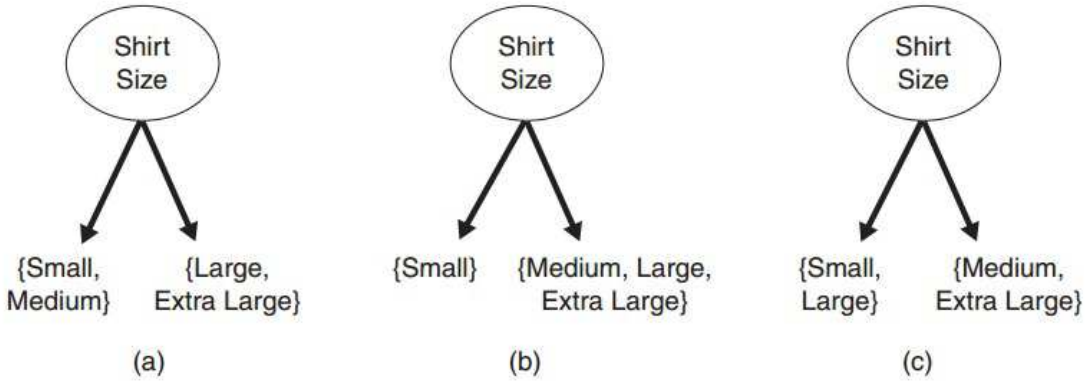
Именски атрибути С обзиром да именски атрибути могу имати више вриједности, тестни услов за њих може бити изражен на два начина као што је приказано на слици 2.8. Ако користимо вишеструку подјелу, као што је приказано на слици 2.8(a), онда је број излазних грана једанак броју различитих вриједности тестног атрибута. Нпр. ако је тестни атрибут брачно стање (енгл. *Marital Status*), који има три могуће вриједности неудата/неожењен (енгл. *Single*), у браку (енгл. *Married*) и разведен (енгл. *Divorced*), он даље доводи до три нове подјеле. С друге стране, неки алгоритми, попут CART-а, праве само бинарне подјеле разматрајући свих $2^{k-1} - 1$ начина добијања бинарних партиција скупа са k атрибута. На слици 2.8(b) приказана су три могућа начина груписања вриједности атрибута брачно стање у два подскупа.



Слика 2.8: Именски атрибут као тестни

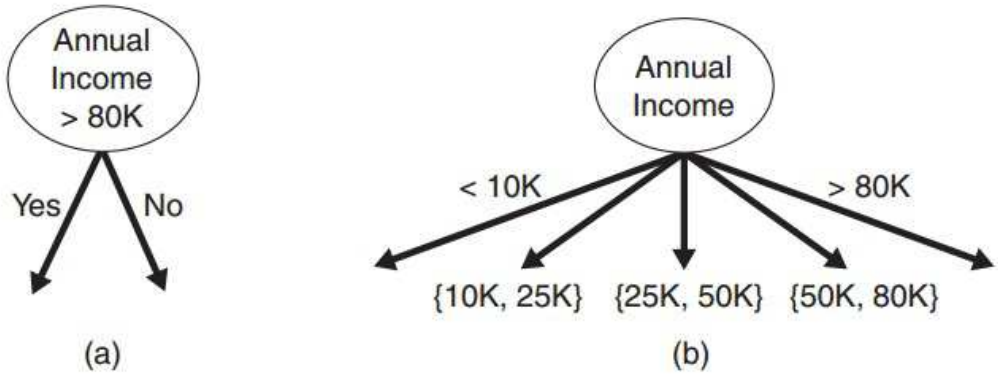
Редни атрибути И за редне атрибуте можемо да користимо бинарну или вишеструку подјелу. Груписање редних атрибута не би требало да нарушава поредак између атрибута. На слици 2.9 су приказани различити начини подјеле тренинг података на основу атрибута величина мајице (енгл. *Shirt Size*). Груписања приказана на слици 2.9(a) и (b) одржавају поредак између редних атрибута, док груписање приказано на слици 2.9(c) нарушава тај поредак јер групише вриједности мало (енгл. *Small*) и велико (енгл. *Large*) у једну партицију, односно средње (енгл. *Medium*) и екстра велико (енг. *Extra Large*) у другу.

Интервални атрибути За интервалне атрибуте тестни услов може бити поређење вриједности $(A < v)$ или $(A \geqslant v)$ са двије излазне гране (бинарна подјела) или подјела вриједности по интервалима $v_i \leqslant A < v_{i+1}$, за $i = 1, 2, ..., k$, са више излазних грана. Разлика између ових приступа приказана је на слици 2.10. За бинарну подјелу, алгоритам који формира дрво одлучивања мора разматрати све могуће подјеле по v и изабрати најбољу међу њима. За вишеструку подјелу, алгоритам треба да разматра све могуће подјеле вриједности



Слика 2.9: Редни атрибут као тестни

тестног атрибута на интервале. Један од приступа којим се ово може ријешити је дискретизација. Након дискретизације, нова вриједност ће бити придружена одговарајућем дискретном интервалу.



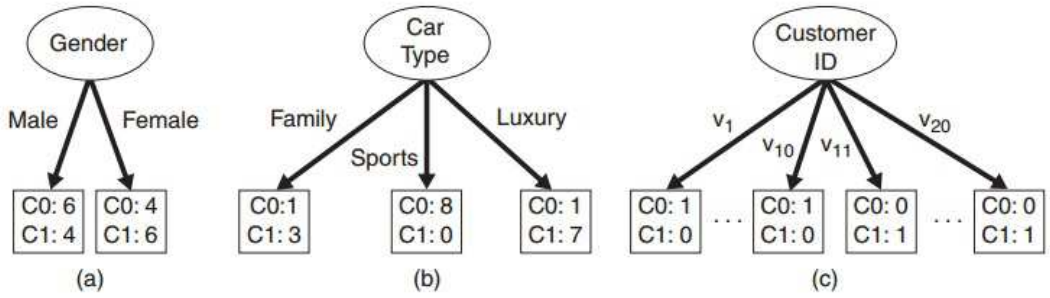
Слика 2.10: Интервални атрибут као тестни

2.2.3.4 Како одредити најбољу подјелу?

Постоје бројне мјере за одређивање најбоље подјеле података. Ове мјере се заснивају на расподјели података по класама прије и послије раздвајања.

Нека је $p(i|t)$ релативна фреквенција података који припадају класи i, а налазе се у чвору t. Понекад се $p(i|t)$ означава са p_i, ако нема забуне на који чвор t се мисли. При бинарној класификацији расподјела по класама за дати чвор се може записати као (p_0, p_1), при чему вриједи $p_1 = 1 - p_0$. Посматрајмо

слику 2.11, јасно је да је расподјела по класама прије подјеле $(0.5, 0.5)$ јер се у свакој класи налази једнак број података. Ако извршимо подјелу по атрибуту пол (енгл. *Gender*), расподјела по класама у добијеним чворовима биће $(0.6, 0.4)$ и $(0.4, 0.6)$, респективно. Очигледно је да нови чворови садрже податке који припадају и једној и другој класи. Подјела по атрибуту тип аута (енгл. *Car Type*), довешће до "чистије" расподјеле.



Слика 2.11: Више верзија бинарне подјеле

Избор атрибута који ће довести до најбоље подјеле се заснива на мјерама нечистоће у новим чворовима. Што је мања нечистоћа, то је подјела боља. На примјер, чвор са расподјелом $(0, 1)$ има степен нечистоће 0, док чвор са расподјелом $(0.5, 0.5)$ има највећи степен нечистоће. Неке од мјера нечистоће су:

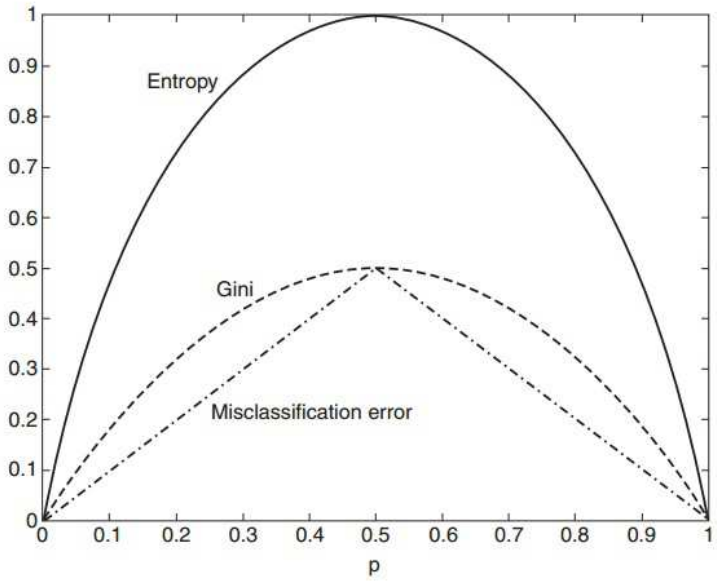
$$Entropija(t) = -\sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t)$$

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

$$Greška\ klasifikacije(t) = 1 - \max_i [p(i|t)]$$

гдје је c број класа, а при израчунавању ентропије узима се да је $0 \cdot \log_2 0 = 0$.

На слици 2.12 је приказано поређење мјера нечистоће за проблем бинарне класификације, при чему p представља дио података који припадају једној од класа. Лако се уочава да све мјере нечистоће достижу максимум за расподјелу $(0.5, 0.5)$, а минумум за расподјеле $(0, 1)$ и $(1, 0)$.



Слика 2.12: Упоређивање мјера нечистоће

2.2.4 Процес класификације наивним Бајесовим класификатором

У овом поглављу је описан метод класификације који не представља експлицитно класификатор, већ користи математичку област теорије вјероватноће да пронађе највјероватнију класификацију. У позадини ове методе класификације налази се Бајесова теорема.

Нека су X и Y случајне варијабле. Заједничка вјероватноћа

$$P(X = x,\ Y = y),$$

заправо представља вјероватноћу да X има вриједност x и Y има вриједност y. Условна вјероватноћа $P(Y = y|X = x)$ представља вјероватноћу да варијабла Y узима вриједност y, ако је познато да варијабла X има вриједност x. Између заједничке и условне вјероватноће постоји сљедећа веза

$$P(X, Y) = P(Y|X) \times P(X) = P(X|Y) \times P(Y).$$

Из претходног слиједи Бајесова формула

$$P(Y|X) = \frac{P(X|Y) \times P(Y)}{P(X)}.$$

2.2.4.1 Примјена Бајесове теореме у класификацији

Прије него што почнемо са процесом класификације, извршимо статистичку формализацију проблема класификације. Нека је X скуп атрибута, а Y ознака класе.

Током процеса тренирања, израчунава се условна вјероватноћа $P(Y|X)$ за сваку комбинацију X и Y из тренинг скупа. Имајући информацију о вриједности ових вјероватноћа, приликом тестирања тест податак X' се сврстава у класу Y', за коју вриједи да је вјероватноћа $P(Y'|X')$ максимална.

Ради илустрације овог приступа посматрајмо податке из табеле са слике 2.5, који ће нам послужити као тренинг скуп. У табели се налазе подаци о томе да ли особа која тражи зајам има сопствену кућу/стан, да ли је у браку и колики јој је годишњи приход. Тражиоци зајма који су благовремено вратили зајам су у класи *No*, док они који нису на вријеме вратили зајам су класификовани као *Yes*. Претпоставимо да имамо тест податак X=*(Home Owner = No, Marital Status = Married, Annual Income = $120K)*. Да бисмо класификовали овај податак потребно је да израчунамо условне вјероватноће $P(Yes|X)$ и $P(No|X)$ на основу података из тренинг скупа. Ако је $P(Yes|X) > P(No|X)$ онда X добија ознаку класе Yes, у супротном ознаку класе No.

Процјена вјероватноће за све могуће комбинације ознаке класе и вриједности атрибута је велики и тежак посао јер то захтјева велики скуп тренинг података. Примјетимо да ако користимо Бајесову формулу за израчунавање вјероватноће да инстанца X припада класи Y да вриједност $P(X)$ можемо занемарити јер је константа.

2.2.4.2 Наивни Бајесов класификатор

Наивни Бајесов класификатор процјењује врјероватноће уз претпоставку да су атрибути и ознака класе међусобно независни. Претпоставка о независности

може се формално исказати на сљедећи начин

$$P(\mathbf{X}|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y),$$

гдје се скуп $\mathbf{X} = (X_1, X_2, ..., X_d)$ састоји од d атрибута.

С обзиром на претпоставку о независности, није потребно да одређујемо вјероватноћу за сваку коминацију атрибута и ознаке класе, већ само вјероватноћу за ознаку класе на основу датих вриједности атрибута. Односно, да би класификовао тестни податак наивни Бајесов класификатор за сваку ознаку класе Y израчунава:

$$P(Y|\mathbf{X}) = \frac{P(Y)\prod_{i=1}^{d} P(X_i)}{P(\mathbf{X})}.$$

Како је вјероватноћа $P(\mathbf{X})$ иста за све ознаке класе Y, довољно је одредити ознаку класе Y за коју је бројилац $P(Y)\prod_{i=1}^{d} P(X_i)$ максималан.

2.2.4.3 Одређивање условне вјероватноће за категоричке атрибуте

За категорички атрибут X_i, условна вјероватноћа $P(X_i = x_i|Y = y)$ се одређује на основу броја инстанци у скупу тренинг података који припадају класи y, а за које посматрани атрибут има вриједност x_i. На примјер, у табели на слици 2.5 троје од седам тражилаца зајма, који су вратили зајам на вријеме, су власници куће/стана. Одакле слиједи да је условна вјероватноћа *P(Home Owner=Yes/No)* једнака $\frac{3}{7}$. Слично, условна вјероватноћа да особа која није вратила зајам на вријеме је неудата/неожењена једнака је

$$P(Marital\,Status = Single|Yes) = \frac{2}{3}.$$

2.2.4.4 Одређивање условне вјероватноће за непрекидне атрибуте

Постоје два начина за одређивање условне вјероватноће за непрекидне атрибуте при класификацији наивним Бајесовим класификатором.

1. Трансформација непрекидних атрибута у категоричке, тј. процес дискретизације који се састоји од двије фазе. У првој фази се одреди број категорија и изврши пресликавање непрекидних атрибута у те категорије. На крају прве фазе, послије сортирања, вриједности непрекидних атрибута се дијеле у n интервала навођењем $(n-1)$ тачке раздвајања. У другој фази вриједности

непрекидних атрибута из истог интервала се пресликавају у исту категоричку вриједност. На овај начин се непрекидни атрибут трансформише у редни атрибут. Условна вјероватноћа $P(X_i|Y = y)$ једнака је броју инстанци тренинг скупа које припадају класи y, а налазе се у интервалу X_i. Колика ће бити грешка при овој процјени зависи од начина дискретизације, као и од броја интервала. Ако је број интервала велики, онда се у сваком интервалу налази мало података за поуздану процјену вјероватноће $P(X_i|Y = y)$. С друге стране, ако је број интервала мали, онда интервали садрже инстанце које припадају различитим класама, па је опет могуће да дође до грешке.

2. Можемо претпоставити да атрибути имају одређену расподјелу и користити тренинг податке за процјену параметара дистрибуције. За непрекидне атрибуте најчешће се користи Гаусова расподјела. Ова расподјела има два параметра, средину μ и варијансу σ^2. За сваку класу y_j, условна вјероватноћа за атрибут X_i рачуна се формулом

$$P(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}.$$

Параметар μ_{ij} се може одредити на основу средње вриједности X_i ($\overline{x}$) за све тренинг податке који припадају класи y_j, док се параметар σ_{ij}^2 одређује на основу стандарне девијације s^2 истих тренинг података. Посматрајмо непрекидни атрибут годишњи приход (енгл. *Annual Income*) из табеле са слике 2.13(а). Средња вриједност и стандардна девијација за овај атрибут у односу на класу No једнаке су

$$\mu = \frac{125 + 100 + 70 + 120 + 60 + 220 + 75}{7} = 110$$

и

$$\sigma^2 = \frac{(125 - 110)^2 + (100 - 110)^2 + ... + (75 - 110)^2}{6} = 2975$$

$$\sigma = \sqrt{2975} = 54.54.$$

Условна вјероватноћа за вриједност атрибута годишњи приход (енгл. *Annual*

$Income$) једнака је

$$P(Annual\ Income = 120|No) = \frac{1}{\sqrt{2\pi}(54.54)} \exp^{-\frac{(120-110)^2}{2 \times 2975}} = 0.0072.$$

2.2.4.5 Примјер примјене наивног Бајесовог класификатора

Посматрајмо скуп података приказаних на слици 2.13(a). На начине описане у 2.2.4.3 и 2.2.4.4, можемо израчунати условне вјероватноће за категоричке и непрекидне атрибуте. Ове вјероватноће су приказане на слици 2.13(b).

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a)

P(Home Owner=Yes|No) = 3/7
P(Home Owner=No|No) = 4/7
P(Home Owner=Yes|Yes) = 0
P(Home Owner=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No) = 1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/3
P(Marital Status=Divorced|Yes) = 1/3
P(Marital Status=Married|Yes) = 0

For Annual Income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

(b)

Слика 2.13: Наивни Бајесов класификатор

Да бисмо одредили ознаку класе за тестни слог $X = (Home\ Owner = No, Marital\ Status = Married, Income = \$120K)$, потребно је да израчунамо вјероватноће $P(Yes|X)$ $P(No|X)$. Из поглавља 2.2.4.2 слиједи да је довољно да израчунамо $P(Y)$ и $\prod_i P(X_i|Y)$. С обзиром да 3 од 10 тренинг података припада класи Yes, онда је $P(Yes) = 0.3$, а како је 7 од 10 тренинг података у класи No, онда је $P(No) = 0.7$. Користећи информације приказане на слици 2.13(b), добијамо:

$$\begin{aligned} P(X|No) &= P(Home\ Owner = No|No) \times P(Marital\ Status = Married|No) \\ &\times P(Income = \$120K|No) \\ &= \frac{4}{7} \times \frac{4}{7} \times 0.0072 \\ &= 0.0024 \end{aligned}$$

$$\begin{aligned} P(X|Yes) &= P(Home\ Owner = No|Yes) \times P(Marital\ Status = Married|Yes) \\ &\times P(Income = \$120K|Yes) \\ &= 1 \times 0 \times 1.2 \times 10^{-9} \\ &= 0 \end{aligned}$$

Коначно, добијамо да је $P(No|X) = \alpha \times \frac{7}{10} \times 0.0024 = 0.0016\alpha$, гдје је $\alpha = \frac{1}{P(X)}$ константа. На исти начин добијамо да је $P(Yes|X) = 0$, јер је $P(X|Yes) = 0$. Како је $P(No|X) > P(Yes|X)$, инстанца X добија ознаку класе No.

2.2.5 Процес класификације примјеном правила

Класификатор заснован на правилима користи правила облика "ако ... онда ..." (енгл. *if...then...*). У оквиру испод приказана су правила која рјешавају проблем класификације кичмењака. Модел класификације се састоји од скупа правила у дисјуктивној нормалној форми $R = (r_1 \vee r_2 \vee ... \vee r_k)$, гдје је R ознака за скуп правила, а r_i ознака појединачних правила за $i \in \overline{1,k}$.

r_1 : (Рађа се живо=не) $\wedge$ (Лети=да) $\rightarrow$ Птице
r_2 : (Рађа се живо=не) $\wedge$ (Живи у води=да) $\rightarrow$ Рибе
r_3 : (Рађа се живо=да) $\wedge$ (Температура тијела=топлокрвни) $\rightarrow$ Сисари
r_4 : (Рађа се живо=не) $\wedge$ (Лети=не) $\rightarrow$ Гмизаваци
r_5 : (Живи у води=да/не) $\rightarrow$ Водоземци

Свако правило класификације може се представити у облику:

$$r_i : (\text{Услов}_i) \rightarrow y_i.$$

Лијева страна правила је (пред)услов и представља конјукцију атрибута, односно облика је

$$\text{Услов}_i = (A_1\ op\ v_1) \wedge (A_2\ op\ v_2) \wedge ... \wedge (A_k\ op\ v_k),$$

при чему је сваки конјукт (A_j, v_j) пар атрибут и његова вриједност, а op је неки од релационих оператора $\{=, \neq, <, \leqslant, >, \geqslant\}$. Десна страна правила је посљедица и садржи ознаку класе y_i.

Правило r покрива (обухвата) инстанцу x ако атрибут инстанце задовољава услов правила. Посматрајмо правило r_1 из табеле која је приказана изнад и примјенимо га на податке о соколу (енгл. *hawk*) и медвједу (енгл. *grizzly bear*), који су дати у табели 2.4. Правило r_1 покрива податке о соколу, те се он може класификовати као птица. С друге стране, правило r_1 се не може примјенити на податке о медвједу, јер његови атрибути не задовољавају (пред)услов овог правила.

Назив	Темп. тијела	Кожни омотач	Да ли се рађа живо?	Живи у води	Лети	Има ноге	Хиберна-ција
Соко	топло крвни	перје	не	не	да	да	не
Медвјед	топло крвни	крзно	да	не	не	да	да

Табела 2.4: Подаци о неким кичмењацима

Квалитет класификатора заснованог на правилима може се мјерити одзивом и прецизношћу. Одзив правила је проценат броја слогова који задовољавају лијеву страну правила, док прецизност правила је проценат броја слогова који задовољавају десну страну правила од процента броја слогова који задовољавају лијеву страну правила. Нека је дат скуп података D и правило $r : A \rightarrow y$, онда вриједи

$$\text{Одзив} = \frac{|A|}{|D|}$$

и

$$\text{Прецизност} = \frac{|A \cap y|}{|A|},$$

при чему је $|A|$ број података који задовољавају услов правила, $|A \cap y|$ број података који задовољавају обје стране правила и $|D|$ укупан број података. На примјер, ако је из табеле са слике 2.13 изведено правило $(Marrital\ Status = Single) \rightarrow No$, онда је одзив овог правила $\frac{4}{10} = 40\%$, а тачност $\frac{2}{4} = 50\%$.

2.2.5.1 Начин рада класификатора заснованог на правилима

Да бисмо видјели како ради класификатор заснован на правилима, посматрајмо претходно наведен скуп правила и покушајмо их примјенити на инстанце дате у табели 2.5.

Назив	Темп. тијела	Кожни омотач	Да ли се рађа живо?	Живи у води	Лети	Има ноге	Хиберна-ција
Лемур	топло крвни	крзно	да	не	не	да	да
Корњача	хладно крвни	рожни покривач	не	да/не	не	да	не
Мала ајкула	хладно крвни	рожни покривач	да	да	не	не	не

Табела 2.5: Подаци о неким кичмењацима

- Први организам, лемур, је топлокрван и рађа се жив, па задовољава услов правила r_3 и класификује се као сисар.
- Други кичмењак, корњача, задовољава услове правила r_4 и r_5. С обзиром да ова два правила дају ознаке различитих класа (гмизавци (енгл. *reptiles*) и водоземци (енгл. *amphibians*)), долази до конфликта.
- Мала ајкула не задовољава услове ни једног правила, па му не можемо додјелити ознаку ниједне класе.

Претходни примјер указује на два могућа проблема класификације правилима, када правила нису међусобно искључива и када постоје слогови које не покрива ниједно правило. Пожељно је да класификатор има сљедеће особине

- Класификатор треба да садржи узајмно искључива правила, тј. међусобно независна правила.
- Класификатор треба да посједује потпуно покривање, тј. да садржи правила за све могуће комбинације вриједности атрибута.

Ове двије карактеристике заједно обезбјеђују да је сваки слог покривен бар једним правилом. Нажалост, немају сви класификатори који су засновани на правилима ове двије особине. Ако скуп правила не обезбјеђује потпуно покривање, онда морамо додати предефинисано (*default*) правило

$$r_d : () \rightarrow y_d,$$

које ће покрити инстанце које не задовољавају услове ниједног правила. Предефинисано правило нема (пред)услов, а додјељује ознаку класе којој припада већина тренинг података. Ако правила нису међусобно искључива, онда су могућа два приступа:

- **Уређен скуп правила** Правила се рангирају по приоритету. Када се тестни слог преда класификатору, додјели му се ознака класе по правилу највишег приоритета чији предуслов задовољава.

- **Неуређен скуп правила** Будући да тестни слог може да буде класификован у више различитих класа, након што се преда тестном класификатору и установи којим све класама може да припада, системом гласања се бира класа. Најчешће се додјељује класи која добије највише гласова. Некад се као критеријум при избору користи прецизност правила.

Оба приступа имају предности и недостатке. Неуређена правила су мање подложна погрешној класификацији него уређена правила, због избора уређења међу правилима. Изградња и чување неуређеног скупа правила је јефтиније, јер се не морају чувати у одређеном редослиједу. Међутим, примјена неуређених правила је скупља јер се атрибути тестног слога морају упоредити са (пред)условом сваког правила.

Шеме за одређивања уређења међу правилима могу бити засноване на правилима (тј. правила се рангирају по квалитету) или на класама (правила која припадају истој класи се групишу једно поред другог).

2.2.5.2 Формирање правила класификације

Да бисмо направили класификатор заснован на правилима, потребно је да издвојимо правила која успостављају везу између атрибута података и ознаке класе. Постоје два метода за формирање правила класификације:

1. **Директни метод** - Правила се издвајају директно из тренинг података.

2. **Индиректни метод** - Правила се издвајају из других класификационих модела, као што су дрво одлучивања и неуронске мреже.

Директне методе дијеле скуп атрибута у мање подскупове, такве да се сви подаци који припадају једном подскупу могу класификовати примјеном једног

правила класификације. Индиректне методе заправо дају кратак опис сложенијих метода класификације.

2.2.5.3 Директна метода формирања правила класификације

За издвајање правила директно из података користи се алгоритам секвенцијалног покривања. Овај алгоритам издваја правила редом за сваку класу. Нпр. за проблем класификације кичмењака прво се издвајају правила за птице, па затим редом за сисаре, водоземце, гмизавце и на крају за рибе. Критеријум избора прве класе за коју ће бити генерисана правила зависи од разних фактора, као нпр. може се десити да нека класа преовладава, тј. да већина тренинг инстанци припада тој класи или да се разматра цијена погрешног класификовања у дату класу.

Алгоритам секвенцијалног покривања је приказан на слици 2.14. Почиње од празног скупа правила. Користи функцију *Learn-One-Rule* да издвоји правило за наредну класу. При томе позитивним тренинг подацима сматрају се они који припадају тој класи, а негативним они који не припадају. Добијено правило је пожељно ако покрива већину позитивних инстанци и не покрива или покрива веома мало негативних инстанци. Када се пронађе правило, тренинг подаци који су покривени тим правилом се елиминишу, а ново правило се ставља на врх листе правила R. Понављају се кораци све док се не достигне критеријум заустављања.

Algorithm 5.1 Sequential covering algorithm.

1: Let E be the training records and A be the set of attribute-value pairs, $\{(A_j, v_j)\}$.
2: Let Y_o be an ordered set of classes $\{y_1, y_2, \ldots, y_k\}$.
3: Let $R = \{\ \}$ be the initial rule list.
4: **for** each class $y \in Y_o - \{y_k\}$ **do**
5: **while** stopping condition is not met **do**
6: $r \leftarrow$ Learn-One-Rule (E, A, y).
7: Remove training records from E that are covered by r.
8: Add r to the bottom of the rule list: $R \longrightarrow R \vee r$.
9: **end while**
10: **end for**
11: Insert the default rule, $\{\} \longrightarrow y_k$, to the bottom of the rule list R.

Слика 2.14: Алгоритам секвенцијалног покривања

Функција Learn-One-Rule

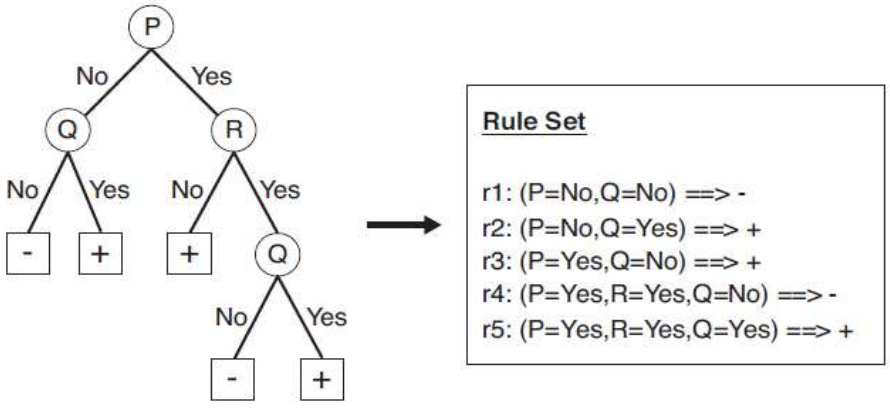
Циљ функције *Learn-One-Rule* је да издвоји правило које покрива већину позитивних инстанци и ниједну (или врло мало) негативних инстанци. Међутим, проналажење оптималног правила је рачунарски захтјеван посао с обзиром да скуп тренинг података може бити јако велики. Функција *Learn-One-Rule* користи стратегију похлепе да ријеши проблем тражења правила. Проналази почетно правило r, које дорађује све док не достигне критеријум заустављања. Након тога врши се поткресивање правила да би се поправила грешка генерализације.

Критеријум за заустављање је израчунавање добити, па ако добит није значајна правило се одбацује.

У директне методе генерисања правила класификације убрајају се RIPPER, CN2 и 1R.

2.2.5.4 Индиректна метода формирања правила класификације

Разматраћемо издвајање правила класификације из дрвета одлучивања. У суштини, сваки пут од коријена до листа се може представити правилом класификације. Тест услови који се налазе на гранама дрвета су конјукти (пред)услова правила, док је ознака класе која се налази у листу посљедица правила. На слици 2.15 приказано је издвајање правила из дрвета одлучивања. Примјетимо да су правила међусобно искључива и да покривају све могуће инстанце.



Слика 2.15: Издвајање правила класификације из дрвета одлучивања

Међутим, нека од правила се могу поједноставити. Размотримо сљедећа три

правила са слике 2.15

$$r_2 : (P = No) \wedge (Q = Yes) \rightarrow +,$$

$$r_3 : (P = Yes) \wedge (R = No) \rightarrow +,$$

$$r_5 : (P = Yes) \wedge (R = Yes) \wedge (Q = Yes) \rightarrow +.$$

Примијетимо да ако је $Q = Yes$, да онда инстанца припада класи $+$, па дата правила можемо поједноставити на сљедећи начин

$$r_2^{'} : (Q = Yes) \rightarrow +,$$

$$r_3 : (P = Yes) \wedge (R = No) \rightarrow +.$$

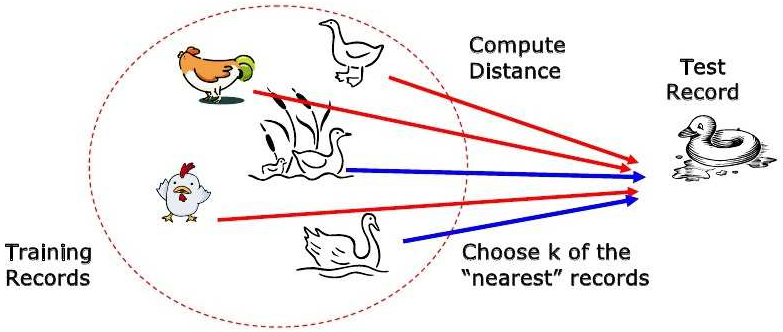
Правило r_3 покрива остале инстанце које припадају класи $+$. Иако, након овога добијена правила нису међусобно искључива, мање су комплексна и лакша су за тумачење.

У индиректне методе генерисања правила класификације убраја се C4.5 rules.

2.2.6 Класификација методом најближег сусједа

Претходно описани методи класификације као што су дрво одлучивања и класификација помоћу правила спадају у *вриједне* класификаторе, јер они одмах након добијања тренинг скупа изграђују модел класификације који ће на основу вриједности атрибута инстанци додјелити ознаку класе. Супротна стратегија је да се изградња модела класификације одлаже све док није потребно класификовати тест податке. Методе које примјењују ову стратегију спадају у *лијене* класификаторе. Примјер лијеног класификатора је *Rote* класификатор који чува цијели тренинг скуп података и примјењује класификацију само на тестне инстанце чије вриједности атрибута се у потпуности поклапају са вриједностима атрибута тренинг инстанци. Недостатак овог приступа је тај што тестне инстанце, чије се вриједности атрибута не поклапају са вриједностима атрибута тренинг инстанци, неће бити класификоване.

Описани приступ се може побољшати ако се пронађу тренинг подаци чији атрибути су слични атрибутима тестних података. Такви тренинг подаци се

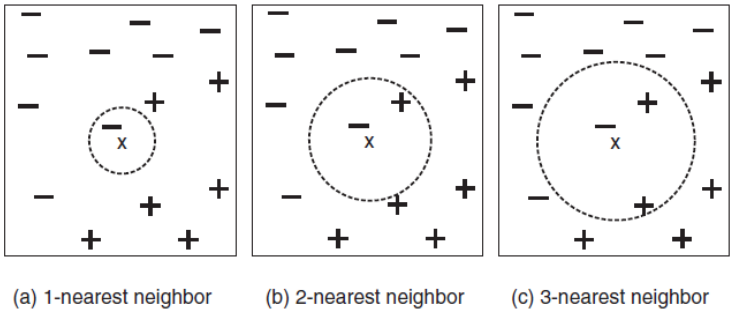


Слика 2.16: Најближи сусјед

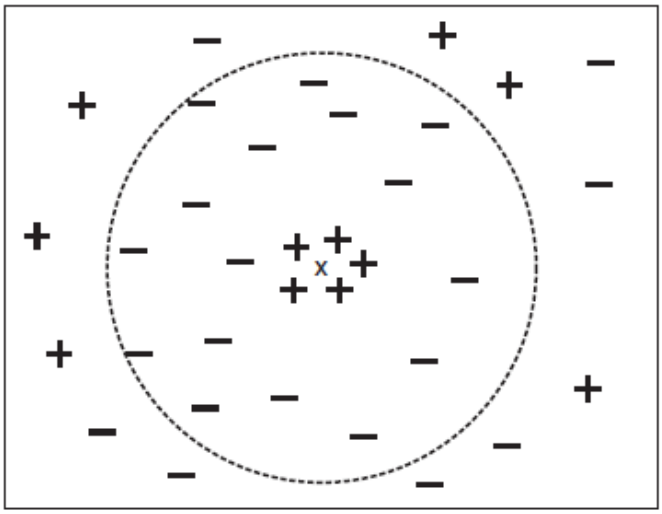
називају **најближи сусједи**. Основа идеја овог приступа се заснива на сљедећем *"Ако шета као патка, кваче као патка, личи на патку онда је вјероватно у питању патка!"* (Слика 2.16). Ако инстанца има d атрибута, онда се при примјени класификатора најближег сусједа представља као тачка у d-димензионалном простору. За дати тестни податак се рачуна блискост са осталим тренинг подацима на основу неке од мјера блискости. Под k најближих сусједа посматране инстанце се подразумјева k тачака које су најближе тачки која представља посматрану инстанцу.

На слици 2.17 приказани су први, други и трећи најближи сусјед центра круга. Тачки се додјељује ознака класе на основу ознака класе њених најближих сусједа. Ако најближи сусједи не припадају истој класи, онда јој се додјељује ознака класе којој припада већина најближих сусједа. На слици 2.17(a) посматра се само један сусјед, који у овом случају има ознаку класе -, па се и центру круга додјељује ознака класе -. С друге стране на слици 2.17(c) се посматрају три најближа сусједа, од којих два припадају класи + а један класи -, па по претходно описаном принципу се центру круга додјељује ознака класе +, јер већина њених најближих сусједа припада тој класи. У ситуацијама попут ове која је приказана на слици 2.17(b) кад једнак број сусједа припада класама + и -, на случајан начин се бира ознака једне од класа.

Из претходног је јасно да је избор броја k најближих сусједа важан. Ако је k сувише мало класификација је осјетљива на шум. С друге стране ако је k сувише велико у сусједе могу да се укључе и тачке из других класа (Слика 2.18).



Слика 2.17: Први, други и трећи најближи сусјед



Слика 2.18: k најближих сусједа за велико k

2.2.6.1 Алгоритам методе најближег сусједа

Алгоритам приказан на слици 2.19 одређује удаљеност (или сличност) између сваког тест податка $z = (x', y')$ и свих тренинг података $(x, y) \in D$ и тако прави листу најближих сусједа D_z. Ако је скуп тренинг података велики, оваква израчунавања могу бити скупа. Међутим коришћењем техника индексирања, може се редуковати број потребних израчунавња да се нађе најближи сусјед за дати тестни примјер.

Након одређивања листе најближих сусједа тестни податак се класификује ознаком класе којој припада већина његових најближих сусједа

$$y' = \arg\max_{v} \sum_{(x_i, y_i) \in D_z} I(v = y_i),$$

Algorithm 5.2 The k-nearest neighbor classification algorithm.

1: Let k be the number of nearest neighbors and D be the set of training examples.
2: **for** each test example $z = (\mathbf{x}', y')$ **do**
3: Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every example, $(\mathbf{x}, y) \in D$.
4: Select $D_z \subseteq D$, the set of k closest training examples to z.
5: $y' = \underset{v}{\text{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
6: **end for**

Слика 2.19: Алгоритам методе најближег сусједа

гдје је v ознака класе, y_i ознака класе неког од најближих сусједа и $I(\cdot)$ функција која враћа 1 ако је вриједност њеног аргумента *true*, у супротном 0.

Претходно описаним приступом, сваки најближи сусјед има једнак утицај на избор ознаке класе датог тестног податка. Управо због тога је претходно описани алгоритам осјетљив на избор вриједности k (као што је и приказано на слици 2.18). Утицај изабраног k се може смањити увођењем тежинске функције за сваког најближег сусједа x_i у односу на његову удаљеност од x' са $\omega_i = \frac{1}{d(x', x_i)^2}$. Тако се постиже да сусједи који су удаљенији од z имају мањи утицај на класификацију у односу на оне који су ближи z. Користећи тежинске функције ознака класе одређује се помоћу

$$y' = \arg\max_{v} \sum_{(x_i, y_i) \in D_z} \omega_i \times I(v = y_i).$$

Глава 3

Материјал

3.1 Опис базе

Подаци који су коришћени при истраживању су преузети са NCBI (*National Center for Biotehnology Information*) сајта, односно листе *lproks summary bct* (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, as of February 9th, 2012). Касније су додате неке карактеристике организама, које су преузете из база *Patrick* (http://patricbrc.org) и *Doe* (http://img.jgi.doe.gov/). При томе нису додати нови организми, већ је скуп карактеристика организама проширен. Тако је направљена табела "карактеристике организама", чији опис се налази у додатку у глави 6. Подаци који се чувају у табели се односе на појединачан организам, који је карактерисан атрибутом који представља идентификацију пројекта (*projectid*). Организми су подјељени у два краљевства Археје (енгл. *Arhaea*) и Бактерије (енгл. *Bacteria*). Археје су подјељене у двије подгрупе Халобактерије (енгл. *Halobacteria*) и Археје без Халобактерија (енгл. *Archaea w/out halobacteria*).

Значење атрибута који се налазе у табели:

1. *proteom_size* - величина протеина (укупна дужина свих протеина у организму);
2. *average_protein_length* - просјечна дужина протеина;
3. *organism_chromosomes* - број хромозома у организму;
4. *organism_plasmides* - број плазмида у организму;

5. *organism_size* - величина организама (у нуклеотидима);
6. *organism_gc_proc* - проценат GC нуклеотида у организму;
7. *gramstain* - Грам позитивне или негативне;
8. *shape* - облик;
9. *arrangement* - уређење;
10. *endospores* - да ли има споре;
11. *motility* - покретљивост;
12. *oxygenreq* - да ли захтјева кисеоник за живот;
13. *habitat* - околина у којој живи (станиште);
14. *temp_range* - температурни опсег у ком живи;
15. *optimal_temp* - оптимална температура на којој живи;
16. *pathogenic* - да ли је патоген;
17. *symbiotic* - да ли живи у симбиози са неким другим организмом;
18. *free_living* - да ли може да живи самостално.

Поред наведених атрибута, у табели се налазе још неке особине протеинске структуре организама. Као резултат великог броја истраживања структуре протеина, уочено је да значајан број протеина не посједује добро дефинисану 3D структуру. Односно, велики број протеина је неуређен, што значи да они немају фиксну 3D структуру или да садрже регионе који не посједују добро дефинисану 3D структуру. Између осталих, један од назива за ову појаву је "неуређеност протеина" (енгл. *disorder proteins*). Протеини могу бити у потпуности неуређени или се састоје од уређених и неуређених региона различитих дужина. Постоји веза између неуређености протеина и његове функције. С обзиром да је експериментално одређивање неуређености протеина компликовано, да би се одредила уређеност/неуређеност протеина организама који се налазе у бази примјењена су три предиктора. Примјењени предиктори *VSL2b* и *IUPred-L* свој рад заснивају на физичко-хемијским својствима аминокиселина у

протеинима [6]. Трећи предиктор који је примјењен је *IsUnstruct*, који је заправо апроксимација математичког модела феромагнетизма статистичке механике и који користи казну за сусједне аминокиселине од којих је једна у уређеном региону, а друга у неуређеном. *IUPred-L* додјељују скор неуређености аминокиселина на основу поравнања размјене енергије.

Раније је поменуто да су Археје које се налазе у табели подјељене у двије групе (Халобактерије и Археје без Халобактерија). Наиме, због специфичности средина у којима организми живе долази до већег степена неуређености протеина појединих органзама. Такве су на примјер Халобактерије, које живе у срединама високе сланости, па су издвојене као посебна подгрупа Археја.

Тако да се у табели налазе и сљедећи подаци:

1. *perc_disorder_aa_1* - проценат аминокиселина у неуређеним регионима протеина организама;

2. *perc_disorder_aa_31* - проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у организму;

3. *perc_prot_dis_1* - проценат протеина који садрже неуређене регионе аминокиселина;

4. *perc_prot_dis_31* - проценат протеина који садрже неуређене регионе аминокиселина дужине бар 31;

5. *chr_perc_disorder_aa_1* - проценат аминокиселина у неуређеним регионима протеина из хромозома организама;

6. *chr_perc_disorder_aa_31* - проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у хромозомима организма;

7. *chr_perc_prot_dis_1* - проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина;

8. *chr_perc_prot_dis_31* - проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина дужине бар 31;

9. *pls_perc_disorder_aa_1* - проценат аминокиселина у неуређеним регионима протеина из плазмида организама;

10. *pls_perc_disorder_aa_31* - проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у плазмидима организма;

11. *pls_perc_prot_dis_1* - проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина;

12. *pls_perc_prot_dis_31* - проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина дужине бар 31.

Глава 4

Резултати

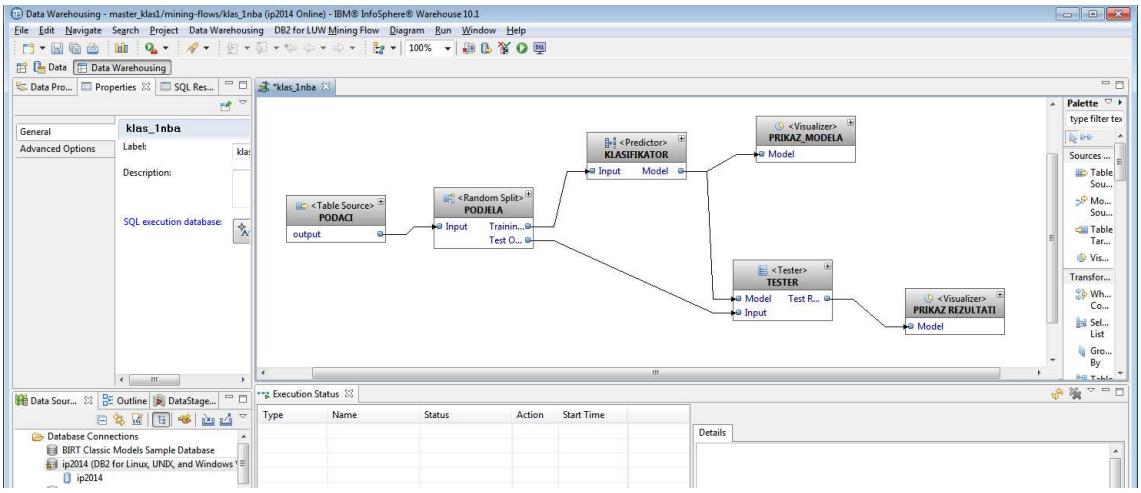
Класификација датих података, у односу на различите атрибуте, обављена је са четири различита алгоритма

- Дрво одлучивања,
- Наивни Бајесов алгоритам,
- Класификација правилима,
- Алгоритам најближег сусједа,

који су описани у глави 2. При томе тестирана су два алгоримтма дрвета одлучивања, од којих је један из пакета *InfoSphere Warehouse Intelligent Miner* (у наставку *IM*), а други из пакета *IBM SPSS Statistics 23* (у наставку *SPSS*). Такође, тестирана су два наивна Бајесова алгоритма, један из *IM* а други из пакета *WEKA*. Алгоритам за класификацију правилима је из пакета *WEKA*, а алгоритам најближег сусједа из *KNIME*-а.

InfoSphere Warehouse је пакет производа који користе *DB2* сервер. Приликом употребе алгоритама за класификацију са ове платформе прављени су ткз. "токови истраживања" (енгл. *mining flow*), један од њих је приказан на слици 4.1. Сваки од токова истраживања садржи извор података (у *IM*-у *Table Source*), у који се учитавају подаци који ће се користити при анализи. Затим се учитани подаци даље шаљу на подјелу на тренинг и тест податке помоћу дијела тока који се зове случајна подјела (у *IM*-у *Random Split*). Након извршене подјеле добијамо два скупа података, односно тренинг и тест податке. Тренинг подаци се прослијеђују као улазни подаци класификатору у IM-у (у *IM*-у *Predictor*),

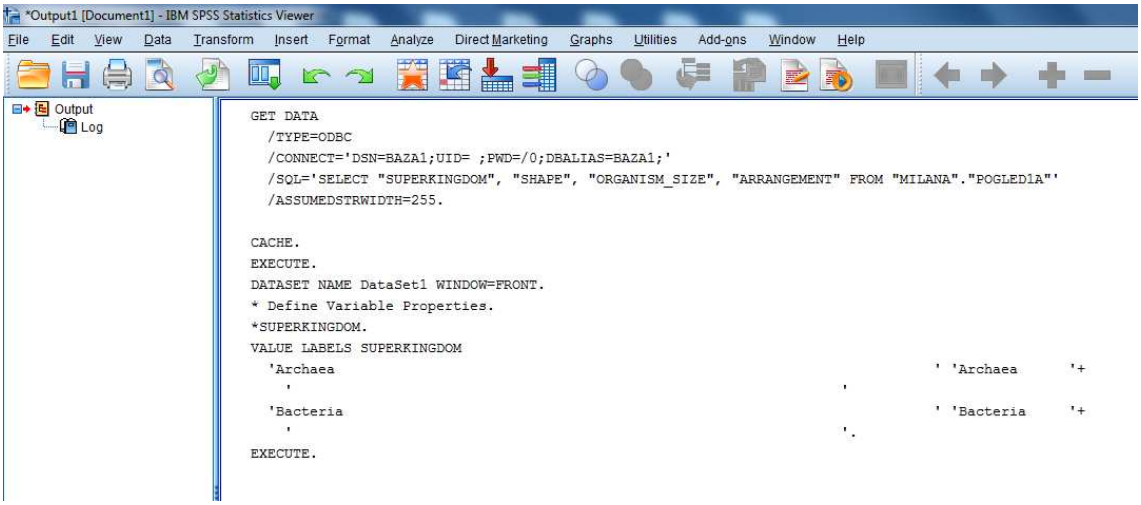
који на основу изабраног алгоритма прави модел класификације. Из овог пакета су коришћени алгоритми дрво одлучивања (*Sprinter*) и наивни Бајесов (*Naive Bayes*). Затим се добијени модел из класификатора и тестни скуп података прослијеђују као улазни подаци тестеру (у *IM*-у *Tester*), који враћа информацију о примјени модела на тест подацима. Ако се као алгоритам користи дрво одлучивања онда су доступне информације о броју (проценту) коректно/некоректно класификованих тренинг података, о броју (проценту) коректно/некоректно класификованих тест података, као и квалитет модела на тренинг и квалитет модела на тест подацима. С друге стране, ако се за изградњу модела користи наивни Бајесов алгоритам онда се као резултат добијају подаци о квалитету модела на тренинг подацима и квалитету модела на тест подацима. Добијени подаци се графички представљају помоћу програма за приказивање (у *IM*-у *Visualizer*).



Слика 4.1: Ток истраживања у InfoSphere Wahrehouse Intelligent Miner

IBM SPSS Statistics је софтверски пакет који се првобитно користио за статистичку анализу и истраживање података, док данас има примјену и у другим областима као што су маркетинг и здравствене науке. Дрво одлучивања које се користило за израду модела класификације формирано је алгоритмом *CHAID*. Као резултат класификације података овим пакетом добија се проценат коректно/некоректно класификованих тренинг података, као и проценат коректно/некоректно класификованих тест података.

WEKA је систем који се користи за истраживање података и развијен је на универзитету Ваикато на Новом Зеланду (*University of Waikato, New Zealand*).



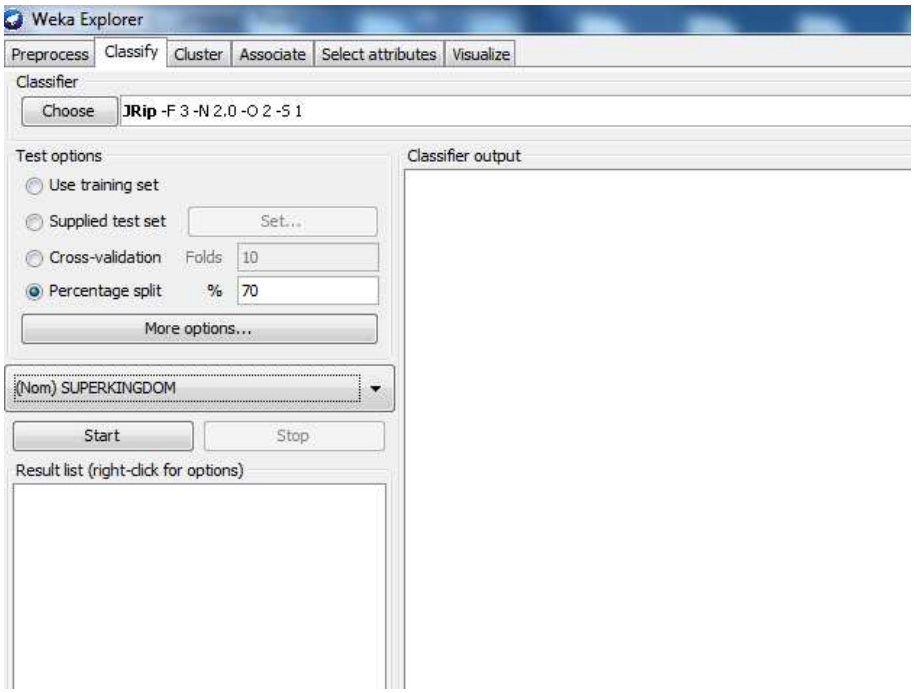
Слика 4.2: Радно окружење у SPSS-у

Заправо, *WEKA* представља колекцију алгоритама машинског учења који се углавном користе за истраживање података. Приликом класификације података наивним Бајесовим алгоритмом коришћен је алгоритам *Naive Bayes Simple*, а при класификацији правилима *Jrip* алгоритам. При употреби ових алгоритама из *WEKA* пакета као резултат добија се проценат коректно/некоректно класификованих тест података.

KNIME (*Konstanz Information Miner*) је јавно доступан пакет који као и *WEKA* садржи скуп алгоритама машинског учења који се користе при истраживању података. Користи се за моделирање и анализу података. Из овог пакета је тестиран алгоритам најближег сусједа (*K-Nearest Neighbour*). Разматрано је $k = 3$ најближих сусједа, а као резултат добијени су подаци о проценту коректно/некоректно класификованих тест података. Имплемнтација алгоритма најближег сусједа која је коришћена при овом истраживању изградњу модела класификације заснива само на атрибутима нумеричког типа.

Приликом примјене свих наведених алгоритама, подаци су дјељени на тренинг и тест податке у односу 70 : 30.

С обзиром да нису сви алгоримти из истог пакета, не враћају сви исте врсте резултата. Тако да је при упоредној анализи вршено упоређивање процента коректно/некоректно класификованих тест података за све алгоритме осим за наивни Бајесов алгоритам из *IM*. Квалитет модела на тренинг и тест подацима које враћа наивни Бајесов алгоритам из *IM* упоређени су са квалитетом модела



Слика 4.3: Радно окружење у WEKA-и након учитавња података који се класификују

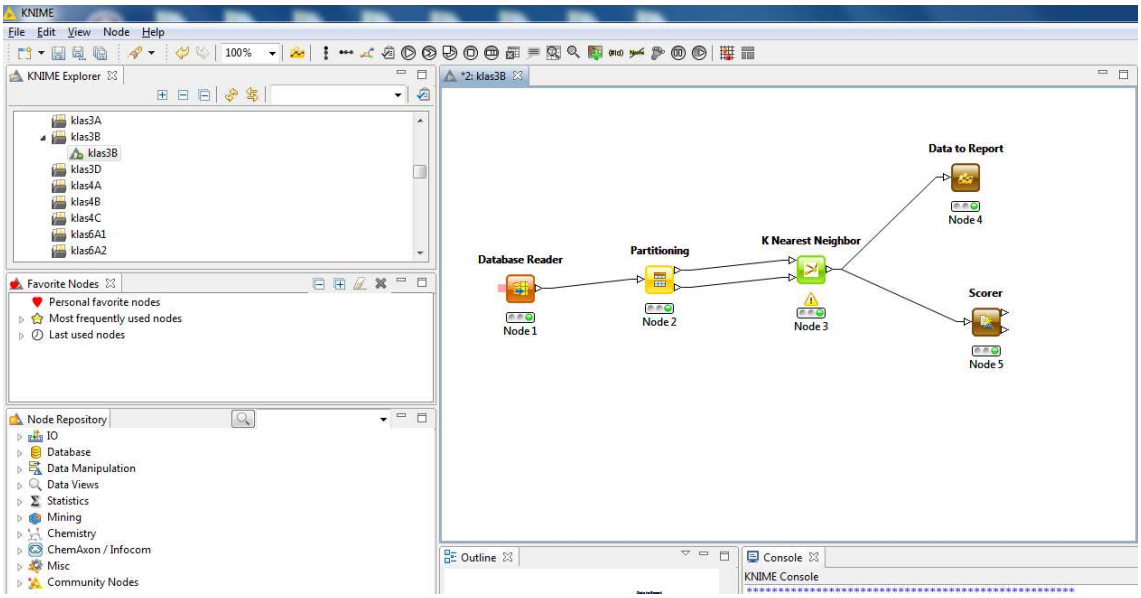
на тренинг и тест подацима који изгради алгоритам дрвета одлучивања из *IM*. Поред тога, упоређени су проценти коректно/некоректно класификованих тренинг података алгоритмом дрвета одлучивања из *IM*-а и *SPSS*-а. Све наведене анализе приказане су у поглављу 4.2.

Модели класификације у електронској форми се налазе у додатку овог рада.

4.1 Резултати класификације

У овом поглављу су приказани и разматрани добијени резулати класификација чији су модели формирани помоћу претходно наведених алгоритама.

1. Разматране су фенотипске карактеристике прокариота, односно њихов облик (*shape*), величина организма (*organism_ size*) и уређеност (*arrangement*), па на основу њих су организми класификовани у једну од класа Археја или Бактерија. Резултати те класификације су приказани у табели 4.1. Поред тога, на основу истих атрибута обављена је класификација организама у раздјеле (*phylum*) и добијени резултати су приказани у табели 4.2.



Слика 4.4: Процес класификације у пакету KNIME

При изградњи модела за класификацију у Археје и Бактерије, алгоритам дрвета одлучивања из *IM*-а највише користи атрибут величина организма (55.09%), док наивни Бајесов алгоритам из *IM*-а највише користи атрибут облик (53.62%), а оба алгоритма најмање користе атрибут уређеност (дрво 17.30%, Бајес 11.59%).

Алгоритам најближег сусједа при изградњи модела наведених класификација не користи ненумеричке атрибуте, односно облик и уређеност, па модели који су добијени овим алгоритмом су формирани само на основу атрибута величина организма.

Табела 4.2 не садржи информације о резултатима класификације наивним Бајесовим алгоритмом из *WEKA*-е и из *IM*-а. Наиме, верзија овог алгоритма из *WEKA*-е не формира модел јер атрибут величина организма нема двије различите вриједности за један од раздјела, док у *IM*-у направи модел чији је квалитет на тренинг подацима 0.53, али не може га тестирати на тест подацима јер класа раздјел има 35 различитих вриједности и при подјели материјала у тестним подацима се налази нека од вриједности које нема у тренинг подацима. Приликом примјене модела на тест податке наилази се на организам који припада управом том раздјелу којег нема у тренинг подацима и долази до прекида програма. Рјешење овог проблема је да се повећа проценат тренинг података са 70% на 80% и тако смањи проценат тестних података на 20%. Тако се добија

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	94.06%	5.94%	93.75%	6.25%	0.75	0.805
Дрво одлучивања-SPSS	94%	6%	92.7%	7.3%	-	-
Наивни Бајесов-WEKA	-	-	92.555%	7.455%	-	-
Наивни Бајесов-IM	-	-	-	-	0.782	0.787
Класификација правилима	-	-	93.7394%	6.2606%	-	-
Најближи сусјед	-	-	90.372%	9.628%	-	-

Табела 4.1: Класификација у Археје или у Бактерије у односу на облик, уређеност и величину организма

модел чији је квалитет на тренинг подацима 0.528, али са истим проблемом приликом примјене на тестне податке. Повећавајући проценат тренинг података на 90% формира се модел чији је квалитет на тренинг подацима 0.52, а на тестним подацима 0.5888. У прилогу овог рада сачувана је верзија модела са подјелом 90 : 10. С обзиром да су остали модели формирани при подјели података у односу 70 : 30 при поређењу резултата ради конзистентности коришћен је податак о квалитету модела на тренинг подацима који је добијен при овој подјели, а квалитет на тестним подацима из наведених разлога није упоређиван.

Из табеле 4.1 видимо да сви алгоритми коректно класификују око 93% тестних података, осим алгоритма најближег сусједа који коректно класификује око 90% тестних података. На тренинг подацима добијају се слични резултати дрветом одлучивања из *IM*-а и из *SPSS*-а. Бољи квалитет на тест подацима има модел добијен дрветом одлучивања из *IM*-а, док бољи квалитет на тренинг подацима има модел добијен наивним Бајесовим алгоритмом из *IM*-а.

Посматрајући резултате приказане у табели 4.2 закључујемо да најбољи проценат коректно класификованих тест података у раздјеле има алгоритам заснован на правилима, док алгоритам дрвета одлучивања из *IM*-а боље калсификује тренинг податке него алгоритам дрвета одлучивања из *SPSS*-а.

2. Како облик организма зависи од спора, а с друге стране облик организма утиче на његову могућност кретања, тестирана је веза између облика

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	58%	42%	46.4%	53.6%	0.683	0.628
Дрво одлучивања-SPSS	52.4%	47.6%	51.1%	48.9%	-	-
Класификација правилима	-	-	53.9763%	46.0237%	-	-
Најближи сусјед	-	-	49.155%	50.845%	-	-

Табела 4.2: Класификација у раздјеле (phylum) у односу на облик, уређеност и величину организма

(*shape*), покретљивости (*motility*) и да ли организам има споре (*endospores*). Организми су класификовани по наведеним атрибутима у Археје или Бактерије и добијени резултати су приказани у табели 4.3. На основу истих атрибута дати прокариоти су класификовани у раздјеле (*phylum*) и резултати се налазе у табели 4.4.

За изградњу модела класификације у Археје и Бактерије дрветом одлучивања из *IM*-а највише се користи атрибут покретљивост (56.13%), док наивни Бајесов алгоритам из истог пакета овај атрибут користи најмање, односно само 8.33%. С друге стране, алгоритам дрвета одлучивања из *SPSS*-а при изградњи овог модела не користи атрибут споре.

Алгоритам најближег сусједа не може примјенити при овим класификацијама јер ниједан од атрибута није нумеричког типа.

Наивни Бајесов алгоритам из *IM*-а не направи модел при класификацији у раздјеле из истог разлога који је наведен при класификацији у раздјеле са атрибутима облик, величина организма и уређеност. Повећањем процента тренинг података на 80% добија се модел квалитета 0.549 на тренинг и 0.455 на тестним подацима. Исти је сачуван у електронској верзији рада али због већ наведених разлога није коришћен при обради резултата.

Из табеле 4.3 видимо да алгоритам заснован на правилима најбоље класификује тестне податке, а да алгоритам дрвета одлучивања из *IM*-а има најмањи проценат коректно класификованих тест података (само 21.07%), иако коректно класификује 93.26% тренинг података. Ипак, алгоритам дрвета одлучивања из *SPSS*-а боље класификује тренинг податке него алгоритам дрвета одлучивања

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	93.26%	6.74%	21.07%	78.93%	0.794	0.099
Дрво одлучивања-SPSS	94.7%	5.3%	93.3%	6.7%	-	-
Наивни Бајесов-WEKA	-	-	92.7242%	7.2758%	-	-
Наивни Бајесов-IM	-	-	-	-	0.736	0.841
Класификација правилима	-	-	94.247%	5.753%	-	-

Табела 4.3: Класификација у Археје или у Бактерије у односу на облик, покретљивост и споре

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	46%	54%	1%	99%	0.609	0.276
Дрво одлучивања-SPSS	57.2%	42.8%	56.5%	43.5%	-	-
Наивни Бајесов-WEKA	-	-	53.1303%	46.8697%	-	-
Класификација правилима	-	-	54.4839%	45.5161%	-	-

Табела 4.4: Класификација у раздјеле у односу на облик, покретљивост и споре

из *IM*-а. Слично, иако има бољи квалитет модела на тренинг подацима него наивни Бајесов алгоритам из *IM*-а алгоритам дрвета одлучивања из *IM*-а има лошији квалитет модела на тест подацима (0.099). Модел добијен дрветом одлучивања у *IM*-у све тестне податке који су Археје класификује као Бактерије, док 78% тестних података који су Бактерије класификује као Археје, па је тачност овог модела на тестним подацима који су Археје 0, а који су Бактерије 0.213.

Разматрајући резултате приказане у табели 4.4 закључујемо да алгоритам дрвета одлучивања из *SPSS*-а најбоље класификује и тренинг и тест податке. Квалитети модела добијених алгоритмима дрвета одлучивања и наивним Бајесовим алгоритмом из *IM*-а се не могу упоредити јер наивни Бајесов алгоритам не изгради модел због већ наведених разлога. Низак проценат од 1% коректно класификованих тестних података моделом заснованим на алгоритму дрвета одучивања из *IM*-а је посљедица тога да се једино 0.9% организма раздјела *Euryarchaeota* коректо класификује овим моделом, све остале инстанце

се класификују погрешно.

3. На основу еколошких карактеристика организма, односно на основу станишта (*habitat*), температурног опсега на којем живе (*temp_ range*) и оптималне температуре на којој живе (*optimal_ temp*) дати прокариоти су класификовани у Археје или Бактерије и резултати те класификације су приказани у табели 4.5. На основу истих атрибута организми су класификовани у раздјеле (*phylum*) и добијени резултати су приказани у табели 4.6. Такође, на основу температурног опсега на којем живи (*temp_ range*) и оптималне температуре на којој живи (*optimal_ temp*), извршена је класификација организама по стаништима (*habitat*) и резултати су приказани у табели 4.7.

За изградњу модела класификације у класе Археја или Бактерија, алгоритам дрвета одлучивања из *IM*-а не користи атрибут станиште, док исти алгоритам из *SPSS*-а ни за овај модел ни за модел класификације у раздјеле не користи атрибут оптимална температура. Алгоритам најближег сусједа формира модел за обе ове класификације, али са поруком да је атрибуте станиште и температурни опсег није користио јер нису нумеричког типа.

У табели 4.6 нема резултата класификације наивним Бајесовим алгоритмом из *WEKA*-е. Наиме, за класу *Fibrobacteras* атрибут оптимална температура нема двије различите вриједности, па се наивни Бајесов алгоритам не може формирати модел класификације. У истој табели нема резултата ни за наивни Бајесов алгоритам из *IM*-а јер као и раније при класификацијама у раздјеле због великог броја могућих вриједности ове класе долази до немогућности изградње модела или до његове примјене на тестне податке. Тек подјелом на тренинг и тест податке у односу 95 : 5 добијен је модел квалитета 0.398 на тренинг подацима и 0.281 на тест подацима, који је сачуван у прилогу.

За табелу 4.7 недостаје информација о резултатима модела који је добијен наивним Бајесовим алгоритмом из *IM*-а. Модел формиран за однос 70 : 30 тренинг и тест података има квалтитет на тренинг подацима 0.564, али приликом примјене на тестне податке добија се порука да модел не враћа информације о квалитету. А ако се направи подјела 50 : 50 добија се модел квалитета 0.735 на тренинг и 0.083 на тест подацима.

Из табеле 4.5 видимо да дрво одлучивања из *SPSS*-а има најбољи проценат

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	97.55%	2.45%	83.79%	16.21%	0.876	0.381
Дрво одлучивања-SPSS	95.2%	4.8%	96.6%	3.4%	-	-
Наивни Бајесов-WEKA	-	-	94.247%	5.753 %	-	-
Наивни Бајесов-IM	-	-	-	-	0.851	0.568
Класификација правилима	-	-	95.7699%	4.2301%	-	-
Најближи сусјед	-	-	92.513%	7.487%	-	-

Табела 4.5: Класификација у Археје или у Бактерије у односу на станиште, температурни опсег и оптималну температуру на којој живи

коректно класификованих тест података, док у односу на њега алгоритам дрвета одлучивања из *IM*-а боље класификује тренинг податке. Међутим, алгоритам дрвета одлучивања из *IM*-а има најлошији проценат коректно класификованих тест података (83.79%). Модел изграђен дрветом одлучивања у *IM*-а има бољи квалитет на тренинг подацима у односу на модел формиран наивним Бајесовим алгоритмом у *IM*-а, али лошији квалитет модела на тест подацима.

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	19%	81%	4%	96%	0.385	0.444
Дрво одлучивања-SPSS	48%	52%	47.5%	52.5%	-	-
Класификација правилима	-	-	43.824%	56.176 %	-	-
Најближи сусјед	-	-	52.023%	47.977%	-	-

Табела 4.6: Класификација у раздјеле у односу на станиште, температурни опсег и оптималну температуру на којој живи

При класификацији у раздјеле на основу ових атрибута из табеле 4.6 уочавамо да алгоритам најближег сусједа најбоље класификује тестне податке. Тренинг податке боље класификује дрво одлучивања из *SPSS*-а.

Иако дрво одлучивања из *IM* при класификацији, чији су резултати приказани у табели 4.7, боље класификује тренинг податке, има најлошији проценат коректно класификованих тест података (8%). Наиме, класа станиште има 5 различитих вриједности а модел на тест подацима све организме који се налазе на неком од три станишта погрешно класификује. Алгоритам најближег сусједа

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	65%	35%	8%	92%	0.616	0.171
Дрво одлучивања-SPSS	49.5%	50.5%	52.2%	47.8%	-	-
Наивни Бајесов-WEKA	-	-	49.0672%	50.9328%	-	-
Класификација правилима	-	-	47.2015%	52.7985 %	-	-
Најближи сусјед	-	-	55.346%	44.654%	-	-

Табела 4.7: Класификација у станишта у односу на температурни опсег и оптималну температуру на којој живи

најбоље класификује тест податке.

4. Познато је да су патогени организми углавном факултативни анаероби и да највећи број болести изазивају анаеробне бактерије. Поред тога, већина патогених организама живи на температури на којој живи и организам домаћина. Због наведеног су организми класификовани у Археје или Бактерије на основу атрибута патогеност (*pathogenic*), да ли захтјева кисеоник за живот (*oxygenreq*) и оптимална температура на којој живи (*optimal_ temp*) и добијени резултати су приказани у табели 4.8. На основу истих атрибута организми су класификовани у раздјеле (*phylum*) и резултати су приказани у табели 4.9. Дати прокариоти су класификовани као патогени или непатогени на основу атрибута да ли захтјева кисеоник за живот (*oxygenreq*) и оптимална температура на којој живи (*optimal_ temp*) и резултати су приказани у табели 4.10.

Алгоритми дрвета одлучивања из *IM*-а и *SPSS*-а за формирање модела класификације у Археје и Бактерије не користе атрибут патогеност, а за модел класификације у класе патогено и непатогено не користе атрибут оптимална температура. Модели у све три наведене класификације који су изграђени алгоритмом најближег сусједа формирани су само на основу атрибута оптимална температура јер преостала два атрибута нису нумеричког типа.

Наивни Бајесов алгоритам из *WEKA*-е не направи модел за класификацију у раздјеле, јер за класу *Fibrobacteras* атрибут оптимална температура на којој живи нема двије различите вриједности. Проблем се из истих разлога као и раније јавља при формирању модела за исту класификацију наивним Бајесовим

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	97.52%	2.48%	87.23%	12.77%	0.884	0.363
Дрво одлучивања-SPSS	93.5%	6.5%	94%	6%	-	-
Наивни Бајесов-WEKA	-	-	93.0626%	6.9374%	-	-
Наивни Бајесов-IM	-	-	-	-	0.883	0.331
Класификација правилима	-	-	94.5854%	5.4146%	-	-
Најближи сусјед	-	-	91.573%	8.427%	-	-

Табела 4.8: Класификација у Археје или у Бактерије у односу на патогеност, захтјев кисеоника за живот и оптималне температуре на којој живи

алгоритмом из *IM*-а који при подјели података на тренинг и тест у односу 70 : 30 не формира модел, при подјели 80 : 20 формира модел чији квалитет на тренинг подацима је 0.42 али при покушају да га примијени на тест податке добија се порука у којој стоји да модел на тест подацима не враћа информације о квалитету. Тек подјелом 90 : 10 добија се модел квалитета 0.357 на тренинг и 0.216 на тест подацима, који је сачуван у прилогу али није разматран при упоређивању резултата.

Из резултата класификације приказаних у табели 4.8 уочавамо да алгоритам заснован на правилима најбоље класификује тестне податке, док тренинг податке боље класификује алгоритам дрвета одлучивања из *IM*-а. Модели изграђени дрветом одлучивања и наивним Бајесовим алгоритмом у *IM*-у су скоро истог квалитета на тренинг подацима, али на тест подацима бољи квалитет има дрво одлучивања.

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	19%	81%	9%	91%	0.431	0.488
Дрво одлучивања-SPSS	47.3%	52.7%	44.3%	55.7%	-	-
Класификација правилима	-	-	43.3164 %	56.6836%	-	-
Најближи сусјед	-	-	50.909%	49.091%	-	-

Табела 4.9: Класификација у раздјеле у односу на патогеност, захтјев кисеоника за живот и оптималне температуре на којој живи

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	86.88%	13.12%	48.34%	51.66%	0.803	0.275
Дрво одлучивања-SPSS	80.8%	19.2%	83.6%	16.4%	-	-
Наивни Бајесов-WEKA	-	-	64.467%	35.533%	-	-
Наивни Бајесов-IM	-	-		-	0.642	0.328
Класификација правилима	-	-	82.2335%	17.7665%	-	-
Најближи сусјед	-	-	76.647%	23.353%	-	-

Табела 4.10: Класификација у патогено или непатогено у односу на захтјев кисеоника за живот и оптималне температуре на којој живи

Из резултата класификације у раздјеле, приказаних у табели 4.9, јасно је да алгоритам дрвета одлучивања из *IM*-а најлошије класификује и тренинг и тест податке. Тренинг податке боље класификује алгоритам дрвета одлучивања из *SPSS*-а, а тест податке најбоље класификује алгоритам најближег сусједа.

Из резултата у табели 4.10 слиједи да тренинг податке боље класификује алгоритам дрвета одлучивања из *IM*-а, док тест податке најбоље класификује дрво одлучивања из *SPSS*-а. Модел изграђен дрветом одлучивања у *IM*-у има бољи квалитет на тренинг подацима, а на тест подацима бољи квалитет има модел формиран наивним Бајесовим алгоритмом из *IM*-а.

5. С обзиром да гени који се налазе на плазмидима узоркују инфективност, разматрана је веза између патогености и плазмида. Односно, вршена је класификација организама као патогених или непатогених у односу на атрибут број плазмида у организму (*organism_plasmides*). Добијени резултати су приказани у табели 4.11. При томе, наивни Бајесов алгоритам из *IM* формира модел чији је квалитет на тренинг подацима 0.878, али при примјени на тестне податке добија се порука да модел на тест подацима не враћа информације о квалитету. Повећањем процента тренинг података на 80% добија се модел квалитета 0.319 на тренинг, односно 0.117 на тест подацима. У електронском прилогу овог рада сачуван је посљедњи модел, а при разматрању резултата је узет у обзир квалитет модела на тренинг подацима који је добијен при подјели 70 : 30.

Из резултата приказаних у табели 4.11 видимо да дрво одлучивања из *IM*-а боље класификује тренинг податке него дрво одлучивања из *SPSS*-а. Алгоритам најближег сусједа најбоље класификује тестне податке.

Такође, тестирана је веза између патогености организма и неуређености протеина који се налазе у плазмидима организма. Резултати класификације у класу патогено или непатогено у односу на проценат аминокиселина у неуређеним регионима протеина из плазмида организма (*pls_perc_disorder_aa_1*), проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у плазмидима организма (*pls_perc_disorder_aa_31*), проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина (*pls_perc_prot_dis_1*) и проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина дужине бар 31 (*pls_perc_prot_dis_31*), добијених са сва три предиктора приказани су у табелама од 6.13 до 6.24 у додатку 6.3. У 11 од ових 12 класификација наивни Бајесов алгоритам при подјели на тренинг и тест податке у односу 70 : 30 формира модел чији је квалитет на тренинг подацима висок, односно између 0.808 (у табели 6.16) и 0.849 (у табели 6.17), али тестни модел јавља већ поменуту поруку да нема информације о квалитету модела на тест подацима. Даље, смањујући проценат тестних података на 20%, 10% или 5% добијају се модели нешто мањег квалитета на тренинг подацима нпр. редом 0.804, 0.778 и 0.77 али са истим проблемом са моделом на тестним подацима. Тек при подјели 96 : 4 (негдје и 97 : 3) на тренинг и тест податке добија се модел знатно нижег квалитета нпр. 0.302 на тренинг подацима и 0.235 на тест подацима. Само при класификацији приказаној у табели 6.19 се при подјели 70 : 30 добија модел који одмах враћа информацију и на тренинг и на тест подацима. Као и у претходним сличним случајевима, у крајњем разматрању у обзир су узети само резултати добијени при подјели 70 : 30.

При томе када се као атрибут користи проценат протеина из плазмида организма који садрже неуређене регионе аминокиселине (*pls_perc_prot_dis_1*) према програму *IsUnstruct* (табела 6.23) дрво одлучивања из *IM*-а и *WEKA*-е не формирају алгоритам. Наивни Бајесов алгоритам из *WEKA*-е не формира модел јер се све инстанце налазе у класи *No*, тј. нису патогене. Дрво одлучивања формира модел при подјели на тренинг и тест податке у односу 80 : 20, али тај модел све тестне инстанце погрешно класификује.

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	88.99%	11.01%	15.97%	84.03%	0.886	0.097
Дрво одлучивања-SPSS	51.7%	48.3%	47.5%	52.5%	-	-
Наивни Бајесов-WEKA	-	-	51.9459%	48.0541%	-	-
Класификација правилима	-	-	50.423%	49.577%	-	-
Најближи сусјед	-	-	52.703%	42.297%	-	-

Табела 4.11: Класификација у патогено или није патогено у односу на број плазмида у организму

Модели дрвета одлучивања за све наведене класификације формирани у *SPSS*-у имају само један чвор, па све инстанце класификују као патогене.

Анализирајући резултате класификације који се налазе у прилогу, увиђамо да у свих 11 класификација за које се могу упоређивати резултати алгоритам дрвета одлучивања из *IM*-а има бољи проценат коректно класификованих тренинг података него алгоритам дрвета одлучивања из *SPSS*-а. Међутим, алгоритам дрвета одлучивања из *IM*-а има најмањи проценат коректно класификованих тестних података, док алгоритам дрвета одлучивања из *SPSS*-а у 9 класификација има најбољи проценат коректно класификованих тестних података. У три класификације најбољи проценат коректно класификованих тестних података има алгоритам најближег сусједа.

Поредећи квалитете модела формираних у пакету *IM* алгоритмима дрво одлучивања и наивни Бајесов при подјели на тренинг и тест податке у односу $70:30$, уочавамо да у 11 класификација наивни Бајесов алгоритам има бољи квалитет модела на тренинг подацима. Али, при томе треба узети у обзир претходно описану анализу ових резултата, тј. да модели формирани наивним Бајесовим алгоритмом или не дају резултате на тест подацима или производе резултате сумњивог квалитета.

6. Вршена је класификација у Бактерије или Археје без Халобактерија у односу на проценат *GC* нуклеотида у организму (*organism_GC_proc*) и проценат аминокиселина у неуређеним регионима протеина из хромозома организма (*chr_perc_disorder_aa_1*) добијен програмом *IUPred-L*, па се резултати

ове класификације налазе у табели 4.12. Такође, организми су класификовани у класе Бактерије или Археје без Халобактерија у односу на исте атрибуте при чему су информације о неуређености добијене са преостала два програма и добијени резултати се налазе у додатку 6.3 у табелама 6.25 и 6.26.

Даље, вршена је класификација у Бактерије или Археје без Халобактерија у односу на проценат *GC* нуклеотида у организму (*organism_ GC_ proc*) и проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у хромозомима организма (*chr_ perc_ disorder_ aa_ 31*) / проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина (*chr_ perc_ prot_ dis_ 1*) / проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина дужине бар 31 (*chr_ perc_ prot_ dis_ 31*) добијених са сва три програма и резултати класификације приказани у табелама од 6.27 до 6.35 у додатку 6.3.

При формирању модела класификација чији су резултати приказани у табелама 6.25, 6.28, 6.30 и 6.34 алгоритам дрвета одлучивања из *SPSS*-а не користи атрибут проценат GC нуклеотида у организму, док за формирање модела класификације чији су резултати приказани у табелама 6.31 и 6.32 нису коришћени атрибути о неуређености. За формирање модела дрветом одлучивања из *IM*-а коришћена су оба атрибута у девет класификација, док за три класификације модел не враћа информације о проценту коришћења атрибута при изградњи модела.

У класификацији у којој се као атрибут користи проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина (*chr_ perc_ prot_ dis_ 1*) добијен програмом *IsUnstruct* (табела 6.32) наивни Бајесов алгоритам из *WEKA*-е не формира модел класификације јер стандардна девијација атрибута проценат протеина који садрже неуређене аминокиселине при чему се протеини налазе у хромозомима организма за ознаку класе Археје без Халобактерија једнака је нули.

При класификацији која користи податке о неуређеним регионима аминокиселина су изостављени организми који припадају Халобактеријама јер је код њих уочен висок степен неуређености па могу навести на погрешне резултате.

Разматрањем резултата ових класификација уочава се да тестне податке у шест

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	95.84%	4.16%	95.7%	4.3%	0.854	0.847
Дрво одлучивања-SPSS	95.3%	4.7%	92.6%	7.4%	-	-
Наивни Бајесов-WEKA	-	-	94.5392%	5.4608%	-	-
Наивни Бајесов-IM	-	-	-	-	0.827	0.819
Класификација правилима	-	-	95.7338%	4.2662%	-	-
Најближи сусјед	-	-	96.246%	3.754%	-	-

Табела 4.12: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат аминокиселина у неуређеним регионима протеина из хромозома организма према програму IUPred-L

класификација најбоље класификује алгоритам заснован на правилима, у три алгоритам најближег сусједа, у једној дрво одлучивања из *IM*-а и у једној дрво одлучивања из *SPSS*-а. При томе у једној класификацији једнак проценат коректно класификованих тестних података имају алгоритам заснован на правилима и наивни Бајесов алгоритам из *WEKA*-е. Алгоритам дрвета одлучивања из *IM*-а боље класификује тренинг податке него алгоритам дрвета одлучивања из *SPSS*-а. Модел формиран наивним Бајесовим алгоритмом у *IM*-у има бољи квалитет на тренинг подацима, док модел формиран алгоритмом дрвета одлучивања у *IM*-у има бољи квалитет на тестним подацима.

7. Организми су класификовани у Бактерије и Археје без Халобактерија на основу својих генотипских особина, односно на основу атрибута величина протеома (*proteom_size*), просјечна дужина протеина (*average_protein_length*) и проценат аминокиселина у неуређеним регионима протеина организма (*perc_disorder_aa_1*) добијен програмом *IUPred-L*, па су резултати ове класификације приказани у табели 4.13.

У додатку 6.3, тачније у табелама од 6.36 до 6.46 се налазе резулатати класификација у Бактерије и Археје без Халобактерија на основу атрибута величина протеома (*proteom_size*), просјечна дужина протеина (*average_protein_length*) и проценат аминокиселина у неуређеним регионима протеина организма (*perc_disorder_aa_1*) (добијен са преостала два програма)/ проценат аминокисели-

на у неуређеним регионима протеина дужине бар 31 у организму (*perc_disorder-_aa_31*)/ проценат протеина који садрже неуређене регионе аминокиселина (*perc_prot_dis_1*)/ проценат протеина који садрже неуређене регионе аминокиселина дужине бар 31 (*perc_prot_dis_31*), добијени са сва три програма. У класификацији у којој се као атрибут користи проценат протеина који садрже неуређене регионе аминокиселина (*perc_prot_dis_1*) добијен програмом *IsUnstruct* (табела 6.43) наивни Бајесов алгоритам из *WEKA*-е не формира модел класификације јер стандардна девијација атрибута проценат протеина који садрже неуређене регионе аминокиселина за организме из класе Археје без Халобактерија једнака је нули.

Модел класификације изграђен дрветом одлучивања из *IM*-у у неким случајевима не користе атрибут величина протеома (нпр. у класификацијама чији су резултати приказани у табелама 4.13, 6.38, 6.40, 6.44 и 6.46), док при изградњи модела класификације из табеле 6.43 не користи атрибут о неуређености. С друге стране, модел класификације чији су резултати приказани у табели 4.13 формиран је дрветом одлучивања из *SPSS*-а само на основу атрибута о неуређености, док друга два атрибута нису коришћена. Дрво одлучивања из *SPSS*-а за изградњу модела класификације из табела 6.36, 6.45 и 6.46 не користи атрибут величина протеома, а за моделе каласификација из табела 6.42 и 6.43 не користи информације о неуређености.

При класификацији која користи податке о неуређености аминокиселина су изостављени организми који припадају Халобактеријама јер је код њих уочен висок степен неуређености па могу навести на погрешне резултате.

Анализом добијених резултата класификација уочавамо да алгоритам дрвета одлучивања из *IM*-а има најбољи проценат коректно класификованих тестних података у 4 класификације, у истом броју класификација као најбољи се показује алгоритам заснован на правилима. Тренинг податке у 8 класификација боље класификује алгоритам дрвета одлучивања из *IM*-а, а у преосталих 4 алгоритам дрвета одлучивања из *SPSS*-а. У већини класификација модел формиран наивним Бајесовим алгоритмом у *IM*-у има бољи квалитет и на тренинг и на тест подацима.

8. Организми су класификовани у Бактерије или у Археје на основу стани-

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	97.47%	2.53%	96.88%	3.12%	0.789	0.807
Дрво одлучивања-SPSS	94.8%	5.2%	93.8%	6.2%	-	-
Наивни Бајесов-WEKA	-	-	93.0034%	6.9966%	-	-
Наивни Бајесов-IM	-	-	-	-	0.857	0.895
Класификација правилима	-	-	96.587%	3.413%	-	-
Најближи сусјед	-	-	96.622%	7.338%	-	-

Табела 4.13: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат аминокиселина у неуређеним регионима протеина организма према програму IUPred-L

шта на коме живе (*habitat*) и покретљивости (*motility*), па су резултати те класификације приказани у табели 4.14. Вршена је и класификација организама у њихове раздјеле (*phylum*) на основу истих атрибута и резултати су приказани у табели 4.15.

Алгоритам дрвета одлучивања из *SPSS*-а формира модел за класификацију у раздјеле на основу атрибута станиште, док исти алгоритам из *IM*-а за изградњу тог модела користи оба атрибута. У овим класификацијама алгоритам најближег сусједа се не може примјенити јер ниједан од атрибута није нумеричког типа.

У табели 4.15 недостају резултати класификације наивним Бајесовим алгоритмом из *IM*-а, јер наведени алгоритам не формира модел при подјелама 70 : 30 и 80 : 20 на тренинг и тест податке, док при подјелама 90 : 10, 95 : 5, 96 : 4 и 97 : 3 формира моделе чији су квалитети на тренинг подацима редом 0.486, 0.48, 0.478 и 0.465, али за сваки од њих на тест подацима добија се порука да нема информација о квалитету.

При класификацији, чији су резултати приказани у табели 4.14, најбољи проценат коректно класификованих и тренинг и тест података има алгоритам дрвета одлучивања из *SPSS*-а. Квалитет модела на тренинг и тест подацима бољи је код модела који је формиран на основу наивног Бајесовог алгоритма у *IM*-а, него код модела који је формиран на основу дрвета одлучивања у *IM*-а.

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	90.94%	9.06%	55.83%	44.17%	0.826	0.256
Дрво одлучивања-SPSS	94%	6%	92.7%	7.3%	-	-
Наивни Бајесов-WEKA	-	-	92.2166%	7.7834 %	-	-
Наивни Бајесов-IM	-	-	-	-	0.843	0.496
Класификација правилима	-	-	92.2166%	7.7834%	-	-

Табела 4.14: Класификација у Археје или у Бактерије на основу станишта и покретљивости

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	38%	62%	0%	100%	0.492	0.31
Дрво одлучивања-SPSS	46.5%	53.5%	47.3%	52.7%	-	-
Наивни Бајесов-WEKA	-	-	48.9002%	51.0998 %	-	-
Класификација правилима	-	-	46.0237%	53.9763%	-	-

Табела 4.15: Класификација у раздјеле на основу станишта и покретљивости

Алгоритам	Број класификација
Дрво одлучивања-IM	6
Дрво одлучивања-SPSS	17
Наивни Бајесов-WEKA	3 или 2
Класификација правилима	13 или 14
Најближи сусјед	10
Укупно	49

Табела 4.16: Упоређивање резултата на тест подацима

При класификацији у раздјеле тренинг податке боље класификује алгоритам дрвета одлучивања из *SPSS*-а. Тестне податке најбоље класификује наивни Бајесов алгоритам из *WEKA*-е.

4.2 Анализа резултата

У овом поглављу је урађена упоредна анализа резултата који су приказани у поглављу 4.1.

У истраживању је урађено 49 класификација, са сваким од шест алгоритама. Као што је раније наведено не враћају сви тестирани алгоритми исте врсте резултата, па је пет од шест алгоритама поређено по проценту коректно класификованих тестних података и у табели 4.16 је за сваки од тих пет алгоритама приказан податак у колико класификација је имао најбољи проценат коректно класификованих тест података. У једној класификацији алгоритам заснован на правилима и наивни Бајесов из *WEKA*-е имају исти проценат коректно класификованих тестних података, због тога у табели стоји 3 или 2 за наивни Бајесов из *WEKA*-е и 13 или 14 за алгоритам заснован на правилима.

Модели формирани помоћу дрвета одлучивања из *IM*-а и *SPSS*-а су као информацију вратили и број (проценат) коректно/некоректно класификованих тренинг података, па је у табели 4.17 приказан податак у колико класификација који од ова два алгоритма је имао бољи проценат коректно класификованих тренинг података. Као што је објашњено раније при једној класификацији алгоритам дрвета одлучивања из *IM*-а при подјели 70 : 30 не формира модел, па је укупан број класификација које су упоређене је 48.

Алгоритми који су из пакета *IBM InfoSphere Intelligent Miner*, односно алго-

Алгоритам	Број класификација
Дрво одлучивања-IM	35
Дрво одлучивања-SPSS	13
Укупно	48

Табела 4.17: Упоређивање резултата на тренинг подацима

Алгоритам	Модел тренинг	Модел тест
Дрво одлучивања-IM	14	11
Наивни Бајесов-IM	30	20
Укупно	44	31

Табела 4.18: Упоређивање квалитета модела

ритам дрвета одлучивања и наивни Бајесов алгоритам из *IM*-а као резултат враћају податак о квалитету модела на тренинг подацима и квалитету модела на тест подацима. Због структуре података за пет класификација наивни Бајесов алгоритам не формира модел при подјели 70 : 30, док алгоритам дрвета одлучивања не направи модел за једну класификацију (за исту ту не направи модел ни наивни Бајесов алгоритам). За 13 класификација наивни Бајесов алгоритам не враћа информације о квалитету модела на тестним подацима, па уз оних 5 за које не формира модел укупан број класификација при којима информација о квалитету модела на тестним подацима није доступна је 18. Због наведеног је укупан број упоређених класификација мањи од 49. У табели 4.18 је приказан податак у колико класификација је који од ова два алгоритма имао бољи квалитет модела на тренинг подацима и квалитет модела на тест подацима.

Даље, вршено је упоређивање резултата алгоритама у зависности од типова атрибута. Од 49 класификација у њих 37 су сви атрибути нумеричког типа, односно у 13 класификација користи се један атрибут нумеричког типа, у 12 два атрибута нумеричког типа и у 12 три атрибута нумеричког типа. У табели 4.19 су приказани подаци о томе у колико класификација је који алгоритам имао најбољи проценат коректно класификованих тестних података, ако је број атрибута један, два или три и сви су нумеричког типа. Подаци о томе који од два алгоритма дрвета одлучивања има бољи проценат коректно класификованих тренинг података ако је број атрибута један, два или три и сви су нумеричког типа приказани су у табели 4.20. При посљедењем упоређивању треба имати у виду да као што је раније објашњено при једној класификацији алгоритам

Алгоритам	Један атрибут	Два атрибута	Три атрибута	Укупно
Дрво одлучивања-IM	0	1	4	5
Дрво одлучивања-SPSS	9	1	3	13
Наивни Бајесов-WEKA	0	0-1	1	1-2
Класификација правилима	0	7-6	4	11-10
Најближи сусјед	4	3	0	7

Табела 4.19: Упоређивање резултата на тест подацима при чему су сви атрибути нумеричког типа

Алгоритам	Један атрибут	Два атрибута	Три атрибута	Укупно
Дрво одлучивања-IM	12	10	7	29
Дрво одлучивања-SPSS	0	2	5	7

Табела 4.20: Упоређивање резултата на тренинг подацима при чему су сви атрибути нумеричког типа

дрвета одлучивања из *IM*-а не формира модел. Упоређивање квалитета модела на тренинг подацима које формирају алгоритми дрвета одлучивања и наивни Бајесов из *IM*-а дати су у табели 4.21. Међутим, за једну класификацију над једним нумеричким атрибутом наивни Бајесов алгоритам и алгоритам дрвета одлучивања не изграде модел. Упоређивање квалитета модела на тест подацима које формирају алгоритми дрвета одлучивања и наивни Бајесов из *IM*-а дати су у табели 4.22. При томе, наивни Бајесов алгоритам за 11 класификација над једним нумеричким атрибутом не враћа информацију о квалитету.

У 4 класификације од разматраних 49 су сви атрибутни текстуалног типа. У двије класификације од ове четири су коришћена два текстуална атрибута, а у друге двије три текстуална атрибута. При томе, најбољи проценат коректно класификованих тест података у класификација по два атрибута дају алго-

Алгоритам	Модел тренинг над једним атрибутом	Модел тренинг над два атрибута	Модел тренинг над три атрибута	Укупно
Дрво одлучивања-IM	2	3	3	8
Наивни Бајесов-IM	10	9	9	28

Табела 4.21: Упоређивање квалитета модела над тренинг подацима при чему су сви атрибути нумерички

Алгоритам	Модел тест над једним атрибутом	Модел тест над два атрибута	Модел тест над три атрибута	Укупно
Дрво одлучивања-IM	1	7	1	9
Наивни Бајесов-IM	0	5	11	16

Табела 4.22: Упоређивање квалитета модела над тест подацима при чему су сви атрибути нумерички

ритми дрвета одлучивања из *SPSS*-а и наивни Бајесов алгоритам из *WEKA*-е, док по три атрибута најбоље резултате дају алгоритам заснован на правилима и дрво одлучивања из *SPSS*-а. Упоређивањем процената коректно класификованих тренинг података у све четири разматране класификације уочава се да се најбољи резултати добијају употребом алгоритма дрвета одлучивања из *SPSS*-а. За двије класификације од ове четири наивни Бајесов алгоритам из *IM*-а не успијева да направи модел класификације за подјелу 70 : 30. У два преостала модела бољи квалитет над тестним подацима има наивни Бајесов алгоритам из *IM*-а, док квалитет модела над тренинг подацима у случају класификације са три атрибута бољи је код алгоритма дрвета одлучивања, а у случају два податка код наивног Бајесовог алгоритма.

Од 49 класификација у њих шест су два атрибута текстуалног типа, а један нумеричког типа. Од тих шест класификација за двије је најбољи проценат коректно класификованих тестних података добијен алгоритмом заснованим на правилима, за двије алгоритмом најближег сусједа, за једну алгоритмом дрвета одлучивања из *IM*-а и за једну алгоритмом дрвета одлучивања из *SPSS*-а. У четири класификације бољи проценат коректно класификованих тренинг инстанци добијен је алгоритмом дрвета одлучивања из *IM*-а, а за преостале двије алгоритмом дрвета одлучивања из *SPSS*-а. Интересантно је то да за класификације које имају најбољи проценат коректно класификованих тестних података алгоритмом заснованим на правилима, дрветом одлучивања из *SPSS*-а или дрветом одлучивања из *IM*-а, као најбољи алгоритам за класификацију њихових тренинг података се показао алгоритам дрвета одлучивања из *IM*-а. С друге стране, за оне класификације код којих се за тестне податке као најбољи показао алгоритам најближег сусједа као најбољи алгоритам за класификацију тренинг инстанци добија се алгоритам дрвета одлучивања из *SPSS*-а. Наивни Бајесов алгоритам из *IM*-а за двије класификације не направи модел при подјели

података у односу 70 : 30, а за једну не обезбјеђује информације о тестним подацима. Тако да је квалитет модела на тренинг подацима бољи код модела направљеног дрветом одлучивања него код модела направљеног наивним Бајесовим алгоритмом у односу 3 : 1. За модел на тестним подацима је тај однос 2 : 1 за алгоритам дрвета одлучивања.

У преостале двије класификације коришћени су један нумерички и један текстуални атрибут. При једној од тих класификација најбољи проценат коректно класификованих тест података има алгоритам најближег сусједа, а при другој алгоритам дрвета одлучивања из *SPSS*-а. У обе класификације бољи проценат коректно класификованих тренинг података има алгоритам дрвета одлучивања из *IM*-а. Бољи квалитет модела на тренинг подацима у оба случаја има алгоритам дрвета одлучивања, док модел који формира наивни Бајесов алгоритам из *IM*-а у једном случају не враћа информацију о квалитету модела на тестним подацима, а у другом има бољи квалитет на тест подацима него модел формиран дрветом одлучивања.

Као циљне класе, при овом истраживању, коришћене су:

1. Археје и Бактерије,
2. Археје без Халобактерија и Бактерије,
3. патогено и није патогено,
4. раздјели (*phylum*),
5. станишта.

У табели 4.23 је дат преглед у колико класификација, у односу на циљну класу, који алгоритам је имао најбољи проценат коректно класификованих тест података. Из приказаних табела се види да ако је циљна класа Археје без Халобактерија или Бактерије да се као најбољи издваја алгоритам класификације правилима, док при класификацији организама као патогених или непатогених најбољи проценат коректно класификованих тестних података има алгоритам дрвета одлучивања из *SPSS*-а.

У табели 4.24 је дат преглед који од алгоритама дрвета одлучивања, у односу на циљну класу, је имао бољи проценат коректно класификованих тренинг података. При класификацији тренинг података за све циљне класе, осим ако

Класа / Алгоритам	Археје и Бактерије	Археје без Халобактерија и Бактерије	патогено и није патогено	раздјел	станиште	Укупно
Дрво одлучивања-IM	1	5	0	0	0	6
Дрво одлучивања-SPSS	2	4	10	1	0	17
Наивни Бајесов-WEKA	0	1 или 2	0	1	0	3 или 2
Класификација правилима	2	11 или 10	0	1	0	13 или 14
Најближи сусјед	0	3	4	2	1	10

Табела 4.23: Упоређивање алгоритама према циљним класама у односу на проценат коректно класификованих тест података

Класа / Алгоритам	Археје и Бактерије	Археје без Халобактерија и Бактерије	патогено и није патогено	раздјел	станиште	Укупно
Дрво одлучивања-IM	3	17	13	1	1	35
Дрво одлучивања-SPSS	2	7	0	4	0	13

Табела 4.24: Упоређивање алгоритама према циљним класама у односу на проценат коректно класификованих тренинг података

је циљна класа раздјел, бољи проценат коректно класификованих података има алгоритам дрвета одлучивања из *IM*-а.

У табели 4.25 је дат преглед у колико класификација који од алгоритама дрвета одлучивања, у односу на циљну класу, је имао бољи квалитет модела на тренинг подацима, а у табели 4.26 упоређени су квалитети модела на тест подацима. При томе нису поређени квалитети модела на тест подацима у случају да су циљне класе раздјел или станиште, а да су подаци подјељени на тренинг и тест податке у односу 70 : 30. Наиме, у случају да је циљна класа станиште модел формиран наивним Бајесовим алгоритмом не враћа информацију о квалитету, а кад је циљна класа раздјел од 5 модела формира један, али ни тај један не обезбјеђује информацију о квалитету на тестним подацима.

Класа / Алгоритам	Археје и Бактерије	Археје без Халобактерија и Бактерије	патогено и непатогено	раздјел	станиште	Укупно
Дрво одлучивања-IM	3	6	3	1	1	14
Наивни Бајесов-IM	2	18	10	0	0	30

Табела 4.25: Упоређивање алгоритама према циљним класама у односу на квалитет модела на тренинг подацима

Класа / Алгоритам	Археје и Бактерије	Археје без Халобактерија и Бактерије	патогено и није патогено	Укупно
Дрво одлучивања-IM	2	8	1	11
Наивни Бајесов-IM	3	16	1	20

Табела 4.26: Упоређивање алгоритама према циљним класама у односу на квалитет модела на тест подацима

Ако су циљне класе Археје/Бактерије, раздјел или станишта бољи квалитет на тренинг подацима има модел формиран алгоритмом дрвета одлучивања из *IM*-а, док у осталим случајевима бољи је квалитет модела формираног наивним Бајесовим алгоритмом из *IM*-а.

Модел формиран алгоритмом дрвета одлучивања из *SPSS*-а на основу атрибута станиште, температурни опсег на којем организам живи и оптимална температура на којој живи, односно на основу еколошких карактеристика организма, има најбољи проценат од 96.6% коректно класификованих тестних података у класе Археје и Бактерије. С друге стране, најбољи проценат од 97.83% коректно класификованих тестних података у класе Археје без Халобактерија и Бактерије има модел формиран алгоритмом дрвета одлучивања из *IM*-а на основу атрибута величина протеома, просјечна дужина протеина и проценат аминокиселина у неуређеним регионима протеина дужине бар 31, који је је добијен програмом *IsUnstruct*.

Најбољи модел за класификацију организама у раздјеле је добијен алгоритмом дрвета одлучивања из *SPSS*-а на основу атрибута облик, покретљивост и споре и има 56.5% коректно класификованих тестних података. Модел, који је формиран дрветом одлучивања из *SPSS*-а и користи атрибуте да ли организам захтјева кисеоник за живот и оптимална температура на којој живи, има најбољи проценат од 83.6% коректно класификованих тестних организама као патогених или непатогених.

Глава 5

Закључак

5.1 Закључак

Из претходно разматраних резултата закључујемо да најбољи проценат коректно класификованих тестних инстанци има алгоритам дрвета одлучивања из *SPSS*-а. Истим алгоритмом добијени су најбољи резултати и при класификацијама које се заснивају само на нумеричким атрибутима и при класификацијама које користе само текстуалне податаке. Ако су неки од атрибута нумеричког, а неки текстуалног типа у 37.5% случајева најбољи резултат на тест подацима добијен је алгоритмом најближег сусједа, у 25% случајева алгоритмом дрвета одлучивања из *SPSS*-а, у 25% случајева алгоритмом заснованим на правилима и 12.5% случајева алгоритмом дрвета одлучивања из *IM*-а. При томе треба имати у виду да коришћена имаплементација алгоритма најближег сусједа при изградњи модела не користи текстуалне атрибуте.

При поређењу процената коректно класификованих тренинг података алгоритмима дрвета одлучивања из *IM*-а и *SPSS*-а, уочавамо да, осим ако су сви подаци текстуалног типа, бољи резултати се добијају примјеном алгортима дрвета одлучивања из *IM*-а.

У већини разматраних случајева, бољи квалитет модела на тренинг подацима и модела на тест подацима добија се примјеном наивног Бајесовог алгоритма из *IM*-а. Међутим, овдје треба узети у обзир да се при класификацији овим алгоритмом јавило много сумњивих резултата (нпр. при класификацији организама као патогених или непатогених на основу атрибута о неуређености протеина) или да чак у неким случајевима не даје резултате, нарочито на тестним

подацима. Употреба наивног Бајесовог алгоритма из *IM*-а се показала као неадекватна при изградњи модела чија циљна класа има много различитих вриједности, нпр. у овом истраживању при класификацији у раздјеле. У случајевима гдје је коришћен овај алгоритам за класификацију у раздјеле је долазило до ситуације да се у тестним подацима налази неки од раздјела којег нема у тренинг подацима и да онда формирани модел не може да се примијени. Поред тога, при анализи је примијећено да алгоритам дрвета одлучивања из *IM*-а има висок проценат коректно класификованих тренинг података, а слаб квалитет модела на тест подацима нарочито ако су циљне класе патогено и непатогено (нпр. табеле од 6.14 до 6.24). Наиме, овдје сви формирани модели све тестне инстанце класификују као патогене.

Према овом истраживању модел заснован на алгоритму дрвета одлучивања из *SPSS*-а је најпогоднији за класификацију посматраних организама без обзира да ли је груписање организама по фенотипским, генотипским или еколошким карактеристикама или по некој од комбинација ових особина. Такође, треба имати у виду и чињеницу да три од четири модела који су предложени као најбољи у претходном поглављу су формирани алгоритмом дрвета одлучивања из *SPSS*-а, па се долази до закључка да за дати скуп организама овај алгоритам даје најбоље резултате класификације.

5.2 Даљи рад

Примјењујући наведене алгоритме на још неку комбинацију атрибута могу се проширити добијени резултати. Такође, могуће је на исте скупове атрибута примјенити друге алгоритме класификације и тиме добити још резултата који би се упоредили да постојећим, а можда и наметнули и неко боље рјешење од предложеног. Поред тога се на основу добијених резултата може се анализирати да ли при класификацији, која као атрибут користи неки од података који се односе на уређеност/неуређеност региона аминокиселина у протеинима, проценат коректно/некоректно класификованих података зависи од предиктора (*IUPred-L, VSL2b* и *IsUnstruct*) којим је добијен податак о неуређености.

Глава 6

Додатак

6.1 Табела карактеристике организама

У овом додатку дат је опис табеле карактеристике организама над којом је вршена класификација. Табела је формирана кодом

create table karakteristike_organizama(
superkingdom varchar(26) not null,
phylum varchar(45),
ordo varchar(38),
projectid integer not null,
proteom_size integer,
average_protein_length decimal(5,2),
organism_chromosomes smallint,
organism_plasmides smallint,
organism_size integer,
organism_GC_proc decimal(5,2),
gramstain char(1),
shape varchar(30),
arrangement varchar(43),
endospores varchar(3),
motility varchar(12),
oxygenreq varchar(15),
habitat varchar(15),
temp_range varchar(17),

optimal_temp decimal(5,2),
pathogenic char(3),
symbiotic char(3),
free_living char(3),
disorder_prediktor character(10) not null,
perc_disorder_aa_1 decimal(5,2),
perc_disorder_aa_31 decimal(5,2),
perc_prot_dis_1 decimal(5,2),
perc_prot_dis_31 decimal(5,2),
chr_perc_disorder_aa_1 decimal(5,2),
chr_perc_disorder_aa_31 decimal(5,2),
chr_perc_prot_dis_1 decimal(5,2),
chr_perc_prot_dis_31 decimal(5,2),
pls_perc_disorder_aa_1 decimal(5,2),
pls_perc_disorder_aa_31 decimal(5,2),
pls_perc_prot_dis_1 decimal(5,2),
pls_perc_prot_dis_31 decimal(5,2),
primary key
(superkingdom,projectid,disorder_prediktor)
) not logged initially;.

6.2 Детаљи о подацима из табеле

У табели се налази укупно 6290 инстанци, од којих је 1971 различита инстанца.

• Могуће вриједности атрибута *superkingdom* су: *Bacteria, Archaea, Halobacteria, Archaea w/out halobacteria.* Података чија је вриједност атрибута *superkingdom Bacteria* има 1845, док података чија је вриједност атрибута *superkingdom Archaea* има 126. С обзиром да су *Archaea*, због већ наведених разлога, подјељене на *Halobacteria* и *Archaea w/out halobacteria*, првих има 18, а других 108.

• Атрибут *phylum* може да узима неку од сљедећих 35 вриједности: *Acidobacteria* (8), *Actinobacteria* (206), *Aquificae* (10), *Bacteriodetes* (78), *Caldiserica* (1), *Chlamydiae* (41), *Chlorobi* (11), *Chloroflexi* (16), *Chrysiogentes* (1), *Crenar-*

chaeota (43), *Cyanobacteria* (43), *Deferribacteres* (4), *Deinococcus-Thermus* (17), *Dictyoglomi* (2), *Elusimicrobia* (1), *Euryarchaeota* (79), *Fibrobacteres* (2), *Firmicutes* (406), *Fusobacteria* (5), *Gemmatimonadetes* (1), *Ignavibacteria* (1), *Korarchaeota* (1), *Nanoarchaeota* (1), *Nitrospirae* (3), *Planctomycetes* (5), *Proteobacteria* (862), *Spirochates* (47), *Synergistetes* (4), *Tenericutes* (48), *Thaumarchaeota* (1), *Thermobaclum* (1), *Thermodesulfobacteria* (2), *Thermotogae* (15), *Verrucomicrobia* (4). У загради је наведен број инстанци које имају одговарајућу вриједност атрибута *phylum*. Једна инстанца нема вриједност за овај атрибут.

- Атрибут *ordo* може да узима неку од 72 вриједности.
- Атрибут *proteom_ size* има 1970 различитих вриједности из интервала $[29853, 3762377]$.
- Просјечна дужина протеина, односно атрибут *average_ protein_ length* узима 1772 различите вриједности из интервала $[232.07, 416.35]$.
- Организми, који се налазе у табели, имају 1, 2 или 3 хромозома (атрибут *organism_ chromosomes*). Један хромозом има 1970 организама, два 97, а три 15.
- Број плазмида у организму варира од 0 до 21 и атрибут *organism_ plasmides* узима 17 различитих вриједности.
- Величина протеина (*organism_ size*) има вриједности из интервала $[138927, 13033779]$. Различитих вриједности овог атрибута има 1969.
- Проценат GC нуклеотида (*organism_ GC_ proc*) у организму се креће од 13.53 до 74.90 и различитих вриједности има 1450.
- Грам позитивних организама (*gramstain*=+) има 576, Грам негативних (*gramstain*=-) има 1377, док за 18 организама нема податка о томе да ли су Грам позитивни или Грам негативни.
- Облик (*shape*) организама може бити: *Pleomorphic* (55), *Rod bacillus* (1161), *Rod curved* (146), *Sphere coccus* (329) или *Other* (10). У загради је наведен број инстанци које имају одговарајући облик. За 270 инстанци овај податак недостаје.
- Уређење (*arrangement*) организама може бити: *Chains, filaments, hyphae* (94), *Clusters, aggregates* (11), *Multiple forms* (417), *Pairs* (27) или *Single* (457). У загради је наведен број инстанци које имају одговарајуће уређење. За 965 инстанци овај податак недостаје.

• Споре има 177 организама из табеле (*endospores=Yes*), 1102 организма нема споре (*endospores=No*), док за 692 инстанце нема информације о томе да ли имају споре.

• 864 организама из табеле је покретљиво (*motility=Yes*), 760 није (*motility=No*), док за 347 организма нема података о томе.

• Атрибут *oxygenreq* узима неку од вриједности: *Aerobic* (394), *Anaerobic* (283), *Facultative* (394) или *Microaerophilic* (35). У загради је наведен број инстанци које имају одговарајућу вриједност атрибута *oxygenreq*. За 865 инстанци овај податак недостаје.

• Станиште (*habitat*) организма може бити: *Aquatic* (269), *Host associated* (744), *Multiple* (435), *Specialized* (204) или *Terrestrial* (147). У загради је наведен број организама који живе на одговарајућем станишту. За 172 организма нема података о станишту.

• Температурни опсег у којем организам живи (*temp_range*) може бити: *Hyperthermophilic* (78), *Mesophilic* (1548), *Psychrophilic* (26) или *Thermophilic* (138). У загради је наведен број организама који живе у одговарајућем температурном окружењу. За 181 организам нема података о температурном окружењу.

• Организама који су патогени (*pathogenic=Yes*) има 977, док оних који нису (*pathogenic=No*) има 994.

• Организама који могу да живе у симбиози са другим организмима (*symbiotic=Yes*) има 182, а оних који не могу да живе у симбиози са другима (*symbiotic=No*) има 1789.

• Самостално може да живи (*free_living=Yes*) 905 организама, а не може (*free_living=No*) 1066 организама.

• Проценат аминокиселина у неуређеним регионима протеина организма (*perc_disoreder_aa_1*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.1.

• Проценат аминокиселина у неуређеним регионима дужине бар 31 протеина организма (*perc_disoreder_aa_31*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.2.

• Проценат протеина који садрже неуређене регионе аминокиселина (*perc-*

_prot_dis_1) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.3.

• Проценат протеина који садрже неуређене регионе аминокиселина дужине бар 31 (*perc_prot_dis_31*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.4.

• Проценат аминокиселина у неуређеним регионима протеина из хромозома организма (*chr_perc_disorder_aa_1*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.5.

• Проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у хромозомима организма (*chr_perc_disorder_aa_31*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.6.

• Проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина (*chr_perc_prot_dis_1*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.7.

• Проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина дужине бар 31 (*chr_perc_prot_dis_31*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.8.

• Проценат аминокиселина у неуређеним регионима протеина из плазмида организма (*pls_perc_disorder_aa_1*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.9.

• Проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у плазмидима организма (*pls_perc_disorder_aa_31*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.10.

• Проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина (*pls_perc_prot_dis_1*) је одређен са три различита програма

(*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.11.

• Проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина дужине бар 31 (*pls_perc_prot_dis_31*) је одређен са три различита програма (*VSL2b, IUPred-L, IsUnstruct*). Информације о броју различитих вриједности, најмањој и највећој вриједности дате су у табели 6.12.

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	974	13.68	50.23
IUPred-L	961	1.13	27.94
IsUnstruct	1025	7.90	36.48

Табела 6.1: Информације о *perc_ disorder_ aa_ 1*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	874	2.82	34.00
IUPred-L	496	0.08	11.79
IsUnstruct	788	1.18	19.56

Табела 6.2: Информације о *perc_ disorder_ aa_ 31*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	43	99.03	100.00
IUPred-L	1640	19.61	93.84
IsUnstruct	1	100.00	100.00

Табела 6.3: Информације о *perc_ prot_ dis_ 1*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	1405	10.85	74.14
IUPred-L	1037	0.47	36.58
IsUnstruct	1396	4.71	58.09

Табела 6.4: Информације о *perc_ prot_ dis_ 31*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	974	13.68	50.23
IUPred-L	961	1.13	27.94
IsUnstruct	1025	7.90	36.48

Табела 6.5: Информације о *chr_ perc_ disorder_ aa_ 1*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	874	2.82	34.00
IUPred-L	496	0.08	11.79
IsUnstruct	788	1.18	19.56

Табела 6.6: Информације о *chr_perc_ disorder_ aa_ 31*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	43	99.03	100.00
IUPred-L	1640	19.61	93.84
IsUnstruct	1	100.00	100.00

Табела 6.7: Информације о *chr_perc_prot_dis_1*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	1405	10.85	74.14
IUPred-L	1037	0.47	36.58
IsUnstruct	1396	4.71	58.09

Табела 6.8: Информације о *chr_perc_prot_dis_31*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	574	13.47	69.27
IUPred-L	469	2.35	49.26
IsUnstruct	553	8.42	62.14

Табела 6.9: Информације о *pls_perc_disorder_aa_1*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	546	0.96	52.35
IUPred-L	390	0.14	36.76
IsUnstruct	532	0.62	43.92

Табела 6.10: Информације о *pls_perc_disorder_aa_31*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	41	97.67	100.00
IUPred-L	476	20.00	100.00
IsUnstruct	1	100.00	100.00

Табела 6.11: Информације о *pls_perc_prot_dis_1*

Програм	Број различитих	Најмања вриједност	Највећа вриједност
VSL2b	494	11.11	100.00
IUPred-L	430	1.11	71.42
IsUnstruct	477	7.01	100.00

Табела 6.12: Информације о *pls_perc_prot_dis_31*

6.3 Резултати класификације - табеле

Неки од резултата класификације налазе се у табелама у овом поглављу.

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	83.78%	16.22%	10.8%	89.2%	0.82	0.097
Дрво одлучивања-SPSS	52.6%	47.4%	50%	50%	-	-
Наивни Бајесов-WEKA	-	-	47.099%	52.901%	-	-
Наивни Бајесов-IM	-	-	-	-	0.843	-
Класификација правилима	-	-	47.4403%	52.5597%	-	-
Најближи сусјед	-	-	53.409%	46.591%	-	-

Табела 6.13: Класификација у патогено или непатогено у односу на проценат аминокиселина у неуређеним регионима протеина из плазмида организма према програму IUPred-L

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	83.71%	16.29%	8.86%	91.4%	0.818	0.089
Дрво одлучивања-SPSS	51.9%	48.1%	51.9%	48.1%	-	-
Наивни Бајесов-WEKA	-	-	48.4642%	51.5358%	-	-
Наивни Бајесов-IM	-	-	-	-	0.827	-
Класификација правилима	-	-	48.6348%	51.3652%	-	-
Најближи сусјед	-	-	53.416%	46.584%	-	-

Табела 6.14: Класификација у патогено или непатогено у односу на проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у плазмидима организма према програму IUPred-L

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	82.77%	17.23%	9.62%	90.38%	0.808	0.096
Дрво одлучивања-SPSS	51.6%	48.4%	52.7%	47.3%	-	-
Наивни Бајесов-WEKA	-	-	47.2696%	52.7304%	-	-
Наивни Бајесов-IM	-	-	-	-	0.835	-
Класификација правилима	-	-	47.099%	52.901%	-	-
Најближи сусјед	-	-	50.943%	49.057%	-	-

Табела 6.15: Класификација у патогено или непатогено у односу на проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина према програму IUPred-L

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	85.44%	14.56%	10.3%	89.7%	0.836	0.088
Дрво одлучивања-SPSS	51.4%	48.6%	53.1%	46.9%	-	-
Наивни Бајесов-WEKA	-	-	47.7816%	52.2184%	-	-
Наивни Бајесов-IM	-	-	-	-	0.849	-
Класификација правилима	-	-	48.1229%	51.8771%	-	-
Најближи сусјед	-	-	45.732%	54.268%	-	-

Табела 6.16: Класификација у патогено или непатогено у односу на проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина дужине бар 31 према програму IUPred-L

Алгориатм	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	80.44%	19.56%	11.33%	88.67%	0.778	0.111
Дрво одлучивања-SPSS	51.1%	48.9%	53.3%	46.7%	-	-
Наивни Бајесов-WEKA	-	-	45.3925%	54.6075%	-	-
Наивни Бајесов-IM	-	-	-	-	0.808	-
Класификација правилима	-	-	45.9044%	54.0956%	-	-
Најближи сусјед	-	-	50.829%	49.171%	-	-

Табела 6.17: Класификација у патогено или непатогено у односу проценат аминокиселина у неуређеним регионима протеина из плазмида организма према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	81.19%	18.81%	13.87%	86.13%	0.789	0.119
Дрво одлучивања-SPSS	50.1%	49.9%	55.3%	44.7%	-	-
Наивни Бајесов-WEKA	-	-	46.4164%	53.5836%	-	-
Наивни Бајесов-IM	-	-	-	-	0.813	-
Класификација правилима	-	-	46.4164%	53.5836%	-	-
Најближи сусјед	-	-	46.154%	53.846%	-	-

Табела 6.18: Класификација у патогено или непатогено у односу на проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у плазмидима организма према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	98.89%	1.11%	0.62%	99.38%	0.989	0.006
Дрво одлучивања-SPSS	51.5%	48.5%	52%	48%	-	-
Наивни Бајесов-WEKA	-	-	46.587%	53.413%	-	-
Наивни Бајесов-IM	-	-	-	-	0.491	0.005
Класификација правилима	-	-	46.4164%	53.5836%	-	-
Најближи сусјед	-	-	50%	50%	-	-

Табела 6.19: Класификација у патогено или непатогено у односу на проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	81.96%	18.04%	11.26%	88.74%	0.796	0.105
Дрво одлучивања-SPSS	50.8%	49.2%	53.6%	46.4%	-	-
Наивни Бајесов-WEKA	-	-	45.5631%	54.4369%	-	-
Наивни Бајесов-IM	-	-	-	-	0.816	-
Класификација правилима	-	-	47.4403%	52.5597%	-	-
Најближи сусјед	-	-	50.256%	49.744%	-	-

Табела 6.20: Класификација у патогено или непатогено у односу на проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина дужине бар 31 према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	82.12%	17.88%	10.21%	89.79%	0.797	0.1
Дрво одлучивања-SPSS	51.7%	48.3%	53.1%	46.9%	-	-
Наивни Бајесов-WEKA	-	-	45.7338%	54.2662%	-	-
Наивни Бајесов-IM	-	-	-	-	0.834	-
Класификација правилима	-	-	45.2218%	54.7782%	-	-
Најближи сусјед	-	-	52.308%	47.692%	-	-

Табела 6.21: Класификација у патогено или непатогено у односу на проценат аминокиселина у неуређеним регионима протеина из плазмида организма према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	82.02%	17.98%	10.56%	89.44%	0.797	0.102
Дрво одлучивања-SPSS	50.9%	49.1%	54.6%	45.4%	-	-
Наивни Бајесов-WEKA	-	-	45.3925 %	54.6075%	-	-
Наивни Бајесов-IM	-	-	-	-	0.825	-
Класификација правилима	-	-	44.7099%	55.2901%	-	-
Најближи сусјед	-	-	51.064%	48.936%	-	-

Табела 6.22: Класификација у патогено или непатогено у односу на проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у плазмидима организма према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-SPSS	53.8%	46.2%	48.5%	51.5%	-	-
Класификација правилима	-	-	45.3925%	54.6075%	-	-
Најближи сусјед	-	-	53.5%	46.5%	-	-

Табела 6.23: Класификација у патогено или непатогено у односу на проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	82.72%	19.28%	11.02%	88.98%	0.782	0.109
Дрво одлучивања-SPSS	50%	50%	56.9%	43.1%	-	-
Наивни Бајесов-WEKA	-	-	45.7338%	54.2662%	-	-
Наивни Бајесов-IM	-	-	-	-	0.812	-
Класификација правилима	-	-	45.3925%	54.6075%	-	-
Најближи сусјед	-	-	55.738%	44.262%	-	-

Табела 6.24: Класификација у патогено или непатогено у односу на проценат протеина из плазмида организма који садрже неуређене регионе аминокиселина дужине бар 31 према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	95.2%	4.8%	95.52%	4.48%	0.607	0.775
Дрво одлучивања-SPSS	94.6%	5.4%	94.1%	5.9%	-	-
Наивни Бајесов-WEKA	-	-	95.0512%	4.9488%	-	-
Наивни Бајесов-IM	-	-	-	-	0.748	0.755
Класификација правилима	-	-	95.9044%	4.0956%	-	-
Најближи сусјед	-	-	93.857%	6.143%	-	-

Табела 6.25: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат аминокиселина у неуређеним регионима протеина из хромозома организма према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	95.88%	4.12%	94.56%	5.44%	0.824	0.764
Дрво одлучивања-SPSS	94.7%	5.3%	94%	6%	-	-
Наивни Бајесов-WEKA	-	-	95.0512%	4.9488%	-	-
Наивни Бајесов-IM	-	-	-	-	0.826	0.762
Класификација правилима	-	-	95.7338%	4.2662%	-	-
Најближи сусјед	-	-	95.392%	4.608%	-	-

Табела 6.26: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат аминокиселина у неуређеним регионима протеина из хромозома организма према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	95.64%	4.36%	96%	4%	0.848	0.888
Дрво одлучивања-SPSS	94.4%	5.6%	94.7%	5.3%	-	-
Наивни Бајесов-WEKA	-	-	94.5392%	5.4608%	-	-
Наивни Бајесов-IM	-	-	-	-	0.84	0.835
Класификација правилима	-	-	95.0512%	4.9488%	-	-
Најближи сусјед	-	-	95.222%	4.778%	-	-

Табела 6.27: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у хромозомима организма према програму IUPred-L

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	95.59%	4.41%	91.42%	8.58%	0.775	0.558
Дрво одлучивања-SPSS	94.5%	5.5%	94.3%	5.7%	-	-
Наивни Бајесов-WEKA	-	-	95.0512%	4.9488%	-	-
Наивни Бајесов-IM	-	-	-	-	0.806	0.654
Класификација правилима	-	-	96.0751%	3.9249%	-	-
Најближи сусјед	-	-	94.027%	5.973%	-	-

Табела 6.28: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у хромозомима организма према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	95.26%	4.74%	96.03%	3.97%	0.805	0.881
Дрво одлучивања-SPSS	95.3%	4.7%	95%	5%	-	-
Наивни Бајесов-WEKA	-	-	95.0512%	4.9488%	-	-
Наивни Бајесов-IM	-	-	-	-	0.82	0.897
Класификација правилима	-	-	97.6109%	2.3891%	-	-
Најближи сусјед	-	-	96.416%	3.584%	-	-

Табела 6.29: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у хромозомима организма према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	96.3%	3.7%	95.99%	4.01%	0.817	0.823
Дрво одлучивања-SPSS	94%	6%	95.5%	4.5%	-	-
Наивни Бајесов-WEKA	-	-	94.7099%	5.2901%	-	-
Наивни Бајесов-IM	-	-	-	-	0.82	0.735
Класификација правилима	-	-	96.2457 %	3.7543 %	-	-
Најближи сусјед	-	-	97.44%	2.56%	-	-

Табела 6.30: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина према програму IUPred-L

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	94.73%	5.27%	93.89%	6.11%	0.474	0.601
Дрво одлучивања-SPSS	95%	5%	93.3%	6.7%	-	-
Наивни Бајесов-WEKA	-	-	95.0512%	4.9488%	-	-
Наивни Бајесов-IM	-	-	-	-	0.697	0.641
Класификација правилима	-	-	95.0512%	4.9488%	-	-
Најближи сусјед	-	-	93.345%	6.655%	-	-

Табела 6.31: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	94.96%	5.04%	93.39%	6.61%	0.475	0.566
Дрво одлучивања-SPSS	94.2%	5.8%	95.2%	4.8%	-	-
Наивни Бајесов-IM	-	-	-	-	0.687	0.668
Класификација правилима	-	-	95.0512%	4.9488%	-	-
Најближи сусјед	-	-	93.857%	6.143%	-	-

Табела 6.32: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	95.6%	4.4%	95.41%	4.59%	0.852	0.864
Дрво одлучивања-SPSS	95.2%	4.8%	92.9%	7.1%	-	-
Наивни Бајесов-WEKA	-	-	94.5392%	5.4608%	-	-
Наивни Бајесов-IM	-	-	-	-	0.851	0.794
Класификација правилима	-	-	95.0512%	4.9488%	-	-
Најближи сусјед	-	-	96.075%	3.925%	-	-

Табела 6.33: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина дужине бар 31 према програму IUPred-L

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	94.59%	5.41%	94.41%	5.59%	0.473	0.637
Дрво одлучивања-SPSS	94.5%	5.5%	94.4%	5.6%	-	-
Наивни Бајесов-WEKA	-	-	95.0512%	4.9488%	-	-
Наивни Бајесов-IM	-	-	-	-	0.634	0.718
Класификација правилима	-	-	96.4164%	3.5836%	-	-
Најближи сусјед	-	-	96.075%	3.925%	-	-

Табела 6.34: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина дужине бар 31 према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	96.45%	3.55%	97.5%	2.5%	0.809	0.919
Дрво одлучивања-SPSS	95%	5%	95.7%	4.3%	-	-
Наивни Бајесов-WEKA	-	-	95.0512%	4.9488 %	-	-
Наивни Бајесов-IM	-	-	-	-	0.826	0.897
Класификација правилима	-	-	97.6109%	2.3891%	-	-
Најближи сусјед	-	-	96.246%	3.754%	-	-

Табела 6.35: Класификација у Археје без Халобактерија или у Бактерије у односу на проценат GC нукеотида у организму и проценат протеина из хромозома организма који садрже неуређене регионе аминокиселина дужине бар 31 према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	96.92%	3.08%	94.44%	5.56%	0.854	0.616
Дрво одлучивања-SPSS	94.2%	5.8%	95.2%	4.8%	-	-
Наивни Бајесов-WEKA	-	-	96.4164%	3.5836%	-	-
Наивни Бајесов-IM	-	-	-	-	0.891	0.785
Класификација правилима	-	-	95.9044%	4.0956%	-	-
Најближи сусјед	-	-	93.003%	6.997%	-	-

Табела 6.36: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат аминокиселина у неуређеним регионима протеина организма према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	97.53%	2.47%	96.7%	3.3%	0.939	0.829
Дрво одлучивања-SPSS	97.5%	2.5%	95.4%	4.6%	-	-
Наивни Бајесов-WEKA	-	-	96.9283%	3.0717%	-	-
Наивни Бајесов-IM	-	-	-	-	0.868	0.864
Класификација правилима	-	-	97.2696%	2.7304%	-	-
Најближи сусјед	-	-	94.027%	5.973%	-	-

Табела 6.37: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат аминокиселина у неуређеним регионима протеина организма према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	96.6%	3.4%	96.34%	3.66%	0.858	0.91
Дрво одлучивања-SPSS	97.3%	2.7%	97.3%	2.7%	-	-
Наивни Бајесов-WEKA	-	-	93.3447%	6.6553%	-	-
Наивни Бајесов-IM	-	-	-	-	0.893	0.911
Класификација правилима	-	-	96.9283%	3.0717%	-	-
Најближи сусјед	-	-	93.345%	6.655%	-	-

Табела 6.38: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у организму према програму IUPred-L

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	95.51%	4.49%	93.36%	6.64%	0.902	0.769
Дрво одлучивања-SPSS	95.5%	4.5%	96.2%	3.8%	-	-
Наивни Бајесов-WEKA	-	-	93.5154%	6.4846%	-	-
Наивни Бајесов-IM	-	-	-	-	0.896	0.842
Класификација правилима	-	-	96.0751%	3.9249%	-	-
Најближи сусјед	-	-	93.857%	6.143%	-	-

Табела 6.39: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у организму према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	97.49%	2.51%	97.83%	2.17%	0.863	0.875
Дрво одлучивања-SPSS	96.5%	3.5%	96.6%	3.4%	-	-
Наивни Бајесов-WEKA	-	-	96.4164 %	3.5836%	-	-
Наивни Бајесов-IM	-	-	-	-	0.893	0.891
Класификација правилима	-	-	97.4403%	2.5597%	-	-
Најближи сусјед	-	-	95.734%	4.266%	-	-

Табела 6.40: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат аминокиселина у неуређеним регионима протеина дужине бар 31 у организму према програму IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	96.84%	3.16%	95.95%	4.05%	0.831	0.746
Дрво одлучивања-SPSS	94.9%	5.1%	93.4%	6.6%	-	-
Наивни Бајесов-WEKA	-	-	92.6621 %	7.3379 %	-	-
Наивни Бајесов-IM	-	-	-	-	0.871	0.814
Класификација правилима	-	-	95.0512%	4.9488%	-	-
Најближи сусјед	-	-	93.857%	6.143%	-	-

Табела 6.41: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат протеина који садрже неуређене регионе аминокиселина према програмом IUPred-L

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	93.59%	6.41%	94.96%	5.04%	0.801	0.881
Дрво одлучивања-SPSS	94.4%	5.6%	94.7%	5.3%	-	-
Наивни Бајесов-WEKA	-	-	94.5392%	5.4608%	-	-
Наивни Бајесов-IM	-	-	-	-	0.814	0.844
Класификација правилима	-	-	93.3447%	6.6553%	-	-
Најближи сусјед	-	-	93.003%	6.997%	-	-

Табела 6.42: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат протеина који садрже неуређене регионе аминокиселина према програмом VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	93.29%	6.71%	92.8%	7.2%	0.772	0.799
Дрво одлучивања-SPSS	94.8%	5.2%	93.8%	6.2%	-	-
Наивни Бајесов-IM	-	-	-	-	0.845	0.844
Класификација правилима	-	-	94.7099%	5.2901%	-	-
Најближи сусјед	-	-	94.369%	5.631%	-	-

Табела 6.43: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат протеина који садрже неуређене регионе аминокиселина према програмом IsUnstruct

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	97.38%	2.62%	96.01%	3.99%	0.87	0.768
Дрво одлучивања-SPSS	97.7%	2.3%	97.6%	2.4%	-	-
Наивни Бајесов-WEKA	-	-	92.1502%	7.8498%	-	-
Наивни Бајесов-IM	-	-	-	-	0.889	0.849
Класификација правилима	-	-	96.587%	3.413%	-	-
Најближи сусјед	-	-	93.515%	6.485%	-	-

Табела 6.44: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат протеина који садрже неуређене регионе аминокиселина дужине бар 31 према програму IUPred-L

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	95.68%	4.32%	93.09%	6.91%	0.899	0.688
Дрво одлучивања-SPSS	94.7%	5.3%	94%	6%	-	-
Наивни Бајесов-WEKA	-	-	92.8328%	7.1672%	-	-
Наивни Бајесов-IM	-	-	-	-	0.89	0.863
Класификација правилима	-	-	95.7338%	4.2662%	-	-
Најближи сусјед	-	-	91.638%	8.632%	-	-

Табела 6.45: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат протеина који садрже неуређене регионе аминокиселина дужине бар 31 према програму VSL2b

Алгоритам	Коректно класиф. тренинг	Некоректно класиф. тренинг	Коректно класиф. тест	Некоректно класиф. тест	Квалитет модела тренинг	Квалитет модела тест
Дрво одлучивања-IM	97.24%	2.76%	97.06%	2.94%	0.808	0.894
Дрво одлучивања-SPSS	97.6%	2.4%	95.8%	4.2%	-	-
Наивни Бајесов-WEKA	-	-	95.3925%	4.6075%	-	-
Наивни Бајесов-IM	-	-	-	-	0.871	0.912
Класификација правилима	-	-	97.099%	2.901%	-	-
Најближи сусјед	-	-	92.833%	7.167%	-	-

Табела 6.46: Класификација у Археје без Халобактерија или у Бактерије у односу на величину протеома, просјечну дужину протеина и проценат протеина који садрже неуређене регионе аминокиселина дужине бар 31 према програму IsUnstruct

Литература

[1] R. W. Bauman, E. Machunis-Masuoka, and I. Tizard. *Microbiology*. Pearson, 2004.

[2] M. Bramer. *Principles of Data Mining*. Springer, 2013.

[3] C. Frederick. "Wilhelm Johannsen and the Genotype Concept". *Journal of the History of Biology* 7 (1974), pp. 5–30.

[4] Z. Gitai. "The new bacterial cell biology: moving parts and subcellular architecture". *PubMed* (2005).

[5] M. Jarak and M. Govedarica. *Mikrobiologija*. Poljoprivredni fakultet Novi Sad, 2003.

[6] G. Pavlović-Lažetić, N. Mitić, J. Kovačević, Z. Obradović, S. Malkov, and M. Beljanski. "Bioinformatics analysis of disordered proteins in prokaryotes". *BMC Bioinformatics* 12 (2011), pp. 1–22.

[7] K. Raza. "Application of data mining in bioinformatics". *Indian Journal of Computer Science and Engineering* 1(2) (2010), pp. 114–118.

[8] L. Rokach and O. Maimon. *Data Mining with Decision Trees*. World Scientific Publishing, 2008.

[9] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson, 2006.