Одредување причинско-последични врски меѓу болести

ПРОЕКТ ПО ПОДАТОЧНО РУДАРСТВО 2017/2018

АНДРЕЈ ЈАНЧЕВСКИ 151003
ВИКТОР ДОМАЗЕТОСКИ 151038
СОЊА МИТИЌ 151113

Опис на проблемот

- Дадени се две податочни множества
- Првото множество содржи записи во облик – пациент, возраст, дијагноза и датум на дијагностицирање
- Второто содржи информации за пациент, возраст, тип на терапија и датум на терапијата
- Цел - да се најдат причинско-последични врски помеѓу болестите
- Проблем на каузација

Претпроцесирање на податоците

- Посилен акцент кон множеството на дијагнози заради поголема релевантност и корисност
- Податочното множество за дијагнози содржи 207695 записи
- Множеството содржи вредности кои недостасуваат кои се однесуваат на датумот на дијагностицирање и возраста
- Поради тоа што сочинуваат помалку од 1.5% од множеството беше одлучено истите да бидат отстранети

Претпроцесирање на податоците

- Користени техники за претпроцесирање:
  - Чистење на податоците преку справување со непознати и неконзистентни вредности
  - Анализа на распределбите на атрибутите – хистограм, boxplot како и соодветна примена на Kolmogorov-Smirnov тест и хи-квадрат тест за распределба
  - Min-max нормализација на непрекинатите атрибути на интервалот [0,1]
  - Дополнителна трансформација на дијагнозите – користејќи online ресурси како https://www.icd10data.com/ICD10CM/Codes соодветно беа мапирани кодовите на болестите на повисоко ниво преку замена со кодот на нивната наткатегорија, болестите недостапни на овој портал беа мапирани според други валидни извори

Претпроцесирање на податоците

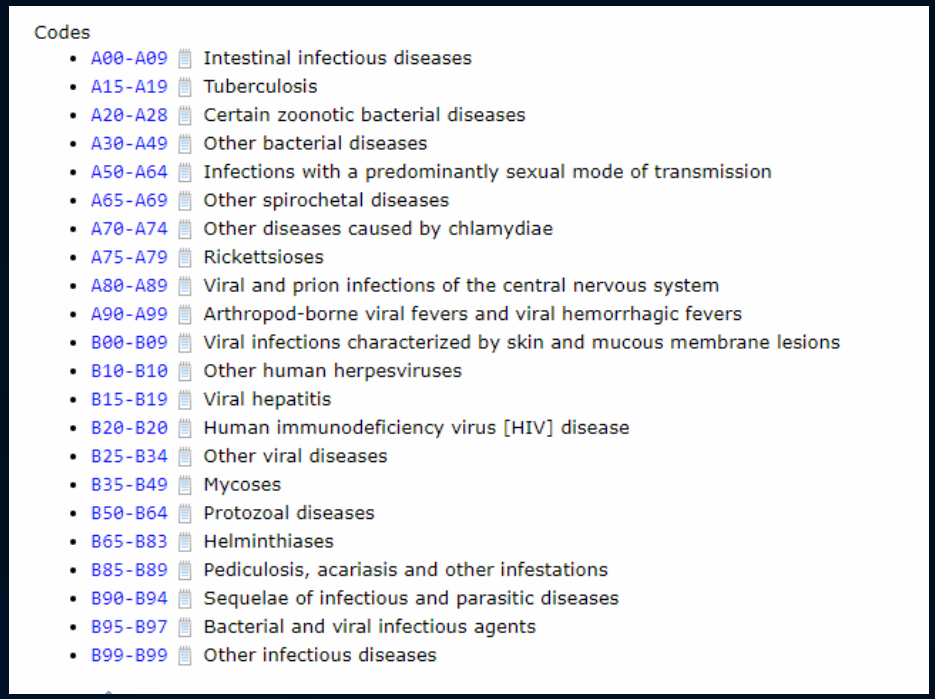
МАПИРАЊЕ ДО НУЛТО НИВО

2018 ICD-10-CM Codes

- A00-B99 Certain infectious and parasitic diseases
- C00-D49 Neoplasms
- D50-D89 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- E00-E89 Endocrine, nutritional and metabolic diseases
- F01-F99 Mental, Behavioral and Neurodevelopmental disorders
- G00-G99 Diseases of the nervous system
- H00-H59 Diseases of the eye and adnexa
- H60-H95 Diseases of the ear and mastoid process
- I00-I99 Diseases of the circulatory system
- J00-J99 Diseases of the respiratory system
- K00-K95 Diseases of the digestive system
- L00-L99 Diseases of the skin and subcutaneous tissue
- M00-M99 Diseases of the musculoskeletal system and connective tissue
- N00-N99 Diseases of the genitourinary system
- O00-O9A Pregnancy, childbirth and the puerperium
- P00-P96 Certain conditions originating in the perinatal period
- Q00-Q99 Congenital malformations, deformations and chromosomal abnormalities
- R00-R99 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- S00-T88 Injury, poisoning and certain other consequences of external causes
- V00-Y99 External causes of morbidity
- Z00-Z99 Factors influencing health status and contact with health services

Претпроцесирање на податоците

МАПИРАЊЕ ДО ПРВО НИВО - ПРИМЕР



Codes

- A00-A09 Intestinal infectious diseases
- A15-A19 Tuberculosis
- A20-A28 Certain zoonotic bacterial diseases
- A30-A49 Other bacterial diseases
- A50-A64 Infections with a predominantly sexual mode of transmission
- A65-A69 Other spirochetal diseases
- A70-A74 Other diseases caused by chlamydiae
- A75-A79 Rickettsioses
- A80-A89 Viral and prion infections of the central nervous system
- A90-A99 Arthropod-borne viral fevers and viral hemorrhagic fevers
- B00-B09 Viral infections characterized by skin and mucous membrane lesions
- B10-B10 Other human herpesviruses
- B15-B19 Viral hepatitis
- B20-B20 Human immunodeficiency virus [HIV] disease
- B25-B34 Other viral diseases
- B35-B49 Mycoses
- B50-B64 Protozoal diseases
- B65-B83 Helminthiases
- B85-B89 Pediculosis, acariasis and other infestations
- B90-B94 Sequelae of infectious and parasitic diseases
- B95-B97 Bacterial and viral infectious agents
- B99-B99 Other infectious diseases

Претпроцесирање на податоците

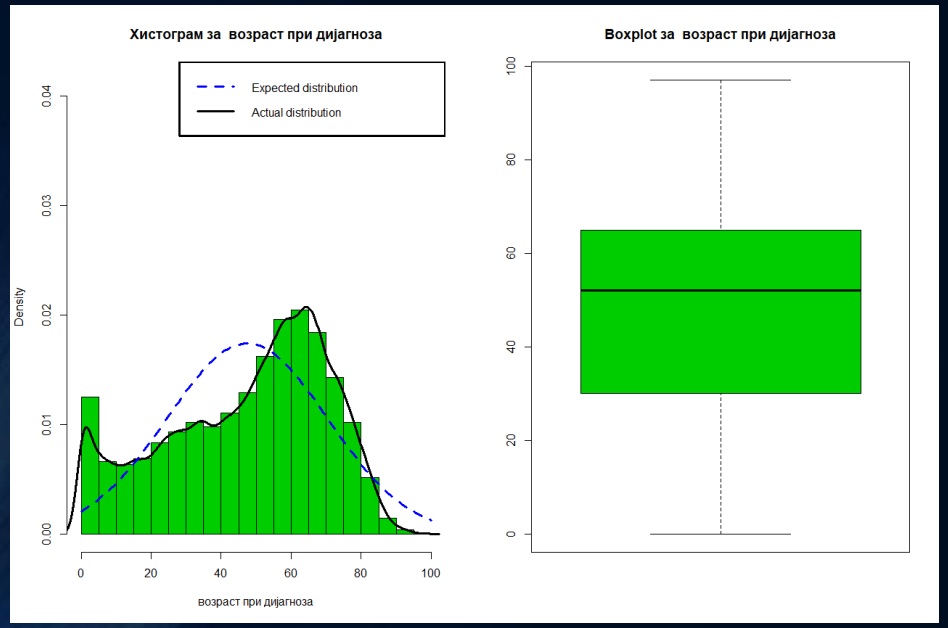
МАПИРАЊЕ ДО ВТОРО НИВО - ПРИМЕР

Codes

- A00 Cholera
- A01 Typhoid and paratyphoid fevers
- A02 Other salmonella infections
- A03 Shigellosis
- A04 Other bacterial intestinal infections
- A05 Other bacterial foodborne intoxications, not elsewhere classified
- A06 Amebiasis
- A07 Other protozoal intestinal diseases
- A08 Viral and other specified intestinal infections
- A09 Infectious gastroenteritis and colitis, unspecified

Претпроцесирање на податоците

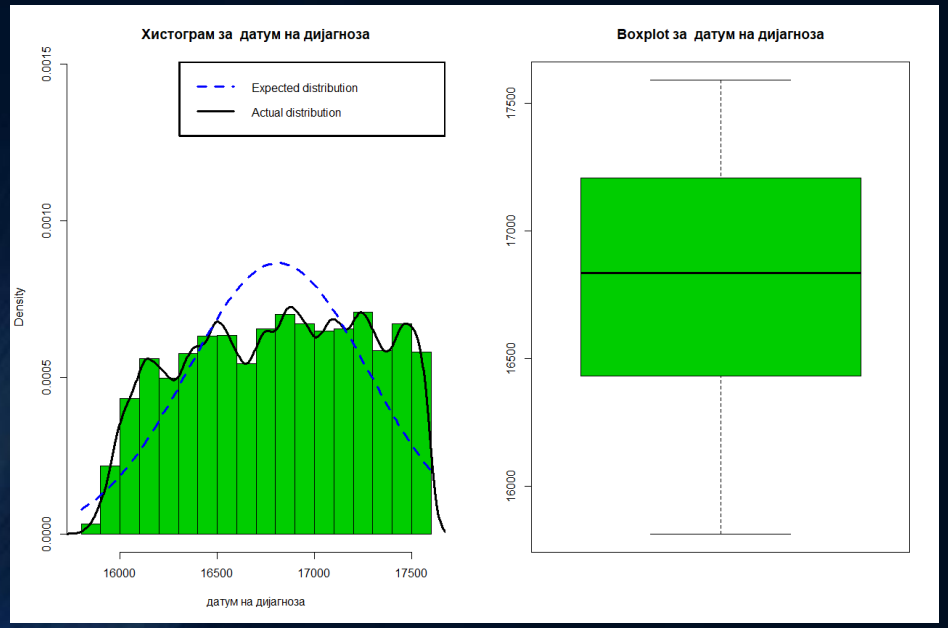
АНАЛИЗА НА РАСПРЕДЕЛБАТА НА ВОЗРАСТА



- KS-тестот за нормална распределба покажува значајно отстапување – доминација на старост од 55 до 65 години
- Возраста содржеше мал број невалидни податоци – негативни вредности поправени користејќи апсолутна вредност

Претпроцесирање на податоците

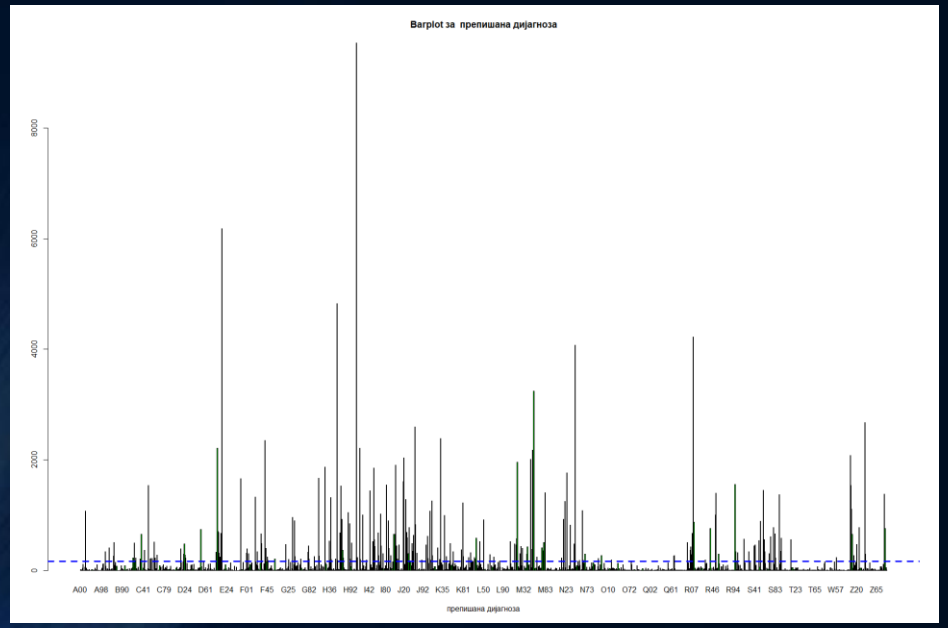
АНАЛИЗА НА РАСПРЕДЕЛБАТА НА ДАТУМОТ НА ПРЕГЛЕД



- KS-тестот за рамномерна распределба не покажува значајност (p=0.47) – промена на временскиот период нема влијание
- Датумот содржеше мал број невалидни податоци – вредности за датум од 20ти век соодветно поправени

Претпроцесирање на податоците

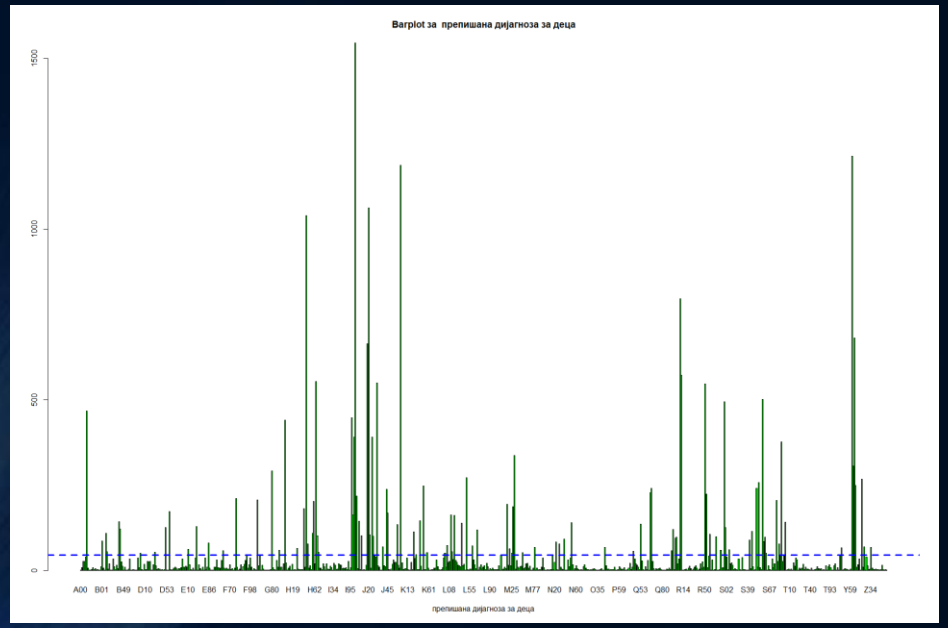
АНАЛИЗА НА РАСПРЕДЕЛБАТА НА ДИЈАГНОЗИТЕ

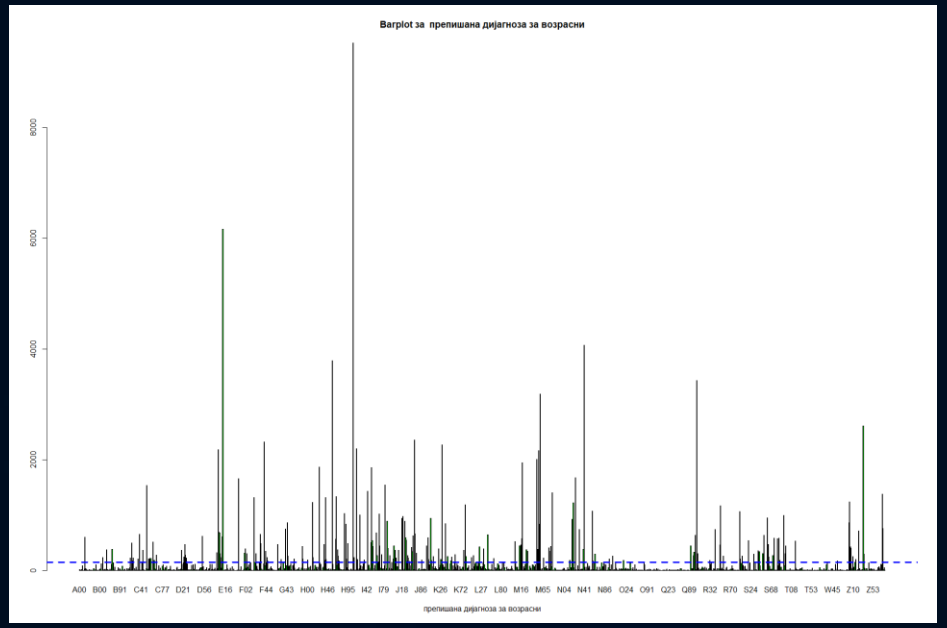


- Дискретен номинален атрибут
- Хи-квадрат тестот за рамномерна распределба покажува значајно отстапување – некои болести имаат видливо поголема честота

Претпроцесирање на податоците

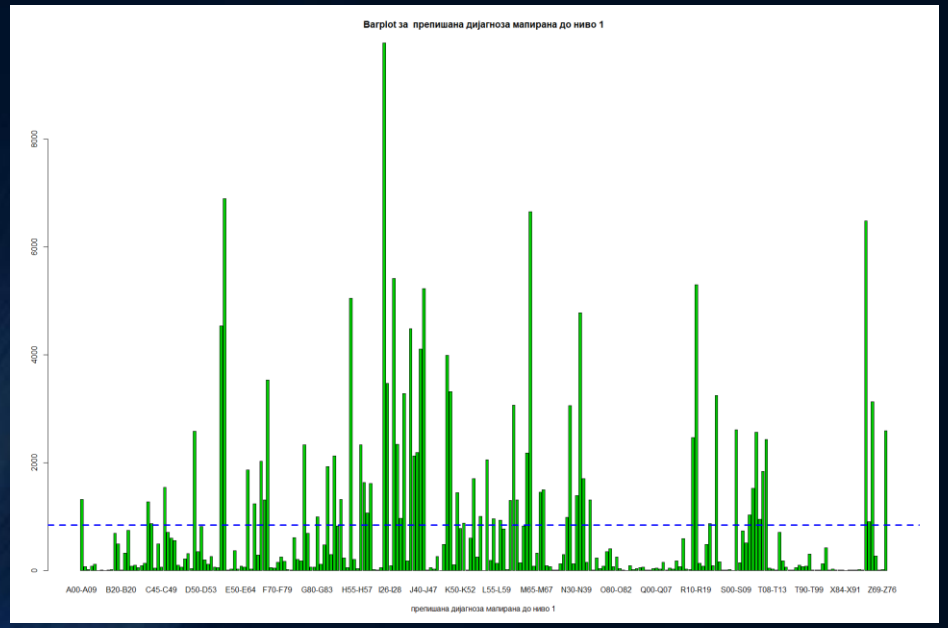
СПОРЕДБА НА ДИЈАГНОЗИТЕ КАЈ ДЕЦА И ВОЗРАСНИ

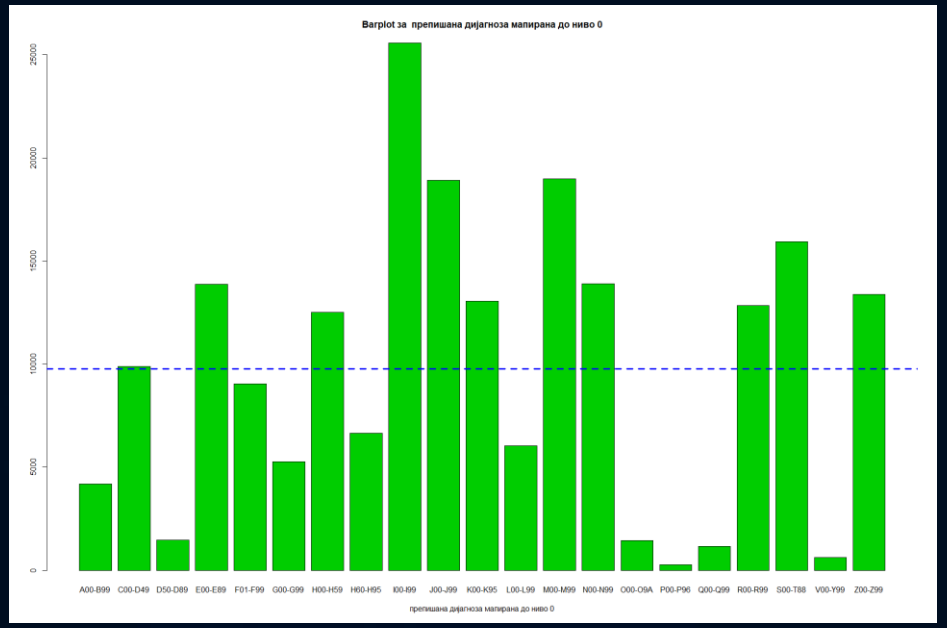




Претпроцесирање на податоците

СПОРЕДБА НА ДИЈАГНОЗИТЕ КАЈ ПОВИСОКИТЕ НИВОА НА МАПИРАЊЕ





Пристапи кон проблемот

1. Анализа на асоцијации користејќи го Apriori алгоритамот
2. Тренирање на класификатор за специфична болест (група од болести)

Анализа на асоцијации

- Прв начин за откривање на врски меѓу болестите – обид за премин од асоцијација до корелација преку следните чекори:
  1. Податочното множество се сортира прво по бројот на пациент па потоа по датумот на дијагнозата
  2. За секој пациент се изминуваат дијагнозите по сортираниот редослед, почнувајќи од празна листа во која се додаваат уникатните дијагнози
  3. Во секој чекор моменталната состојба на листата претставува трансакција која се зачувува во множество на трансакции, притоа ресетирајќи ја листата за следниот пациент
  4. За секое ниво на мапирање на болестите се креира различно множество на трансакции – вкупно 3
  5. Над секое од нив се применува алгоритамот Apriori со 0.005 како минимална поддршка и 0.6 минимална доверба за правилата

Анализа на асоцијации

ПРИМЕР ТРАНСАКЦИИ

```
M00-M99,,,,,,,,,,,,,
M00-M99,I00-I99,,,,,,,,,,,,
M00-M99,I00-I99,H00-H59,,,,,,,,,,,
M00-M99,I00-I99,H00-H59,R00-R99,,,,,,,,,,
M00-M99,I00-I99,H00-H59,R00-R99,J00-J99,,,,,,,,,
M00-M99,I00-I99,H00-H59,R00-R99,J00-J99,S00-T88,,,,,,,,
K00-K95,,,,,,,,,,,,,
K00-K95,M00-M99,,,,,,,,,,,,
K00-K95,M00-M99,G00-G99,,,,,,,,,,,
K00-K95,M00-M99,G00-G99,Z00-Z99,,,,,,,,,,
K00-K95,M00-M99,G00-G99,Z00-Z99,E00-E89,,,,,,,,,
K00-K95,M00-M99,G00-G99,Z00-Z99,E00-E89,R00-R99,,,,,,,,
K00-K95,M00-M99,G00-G99,Z00-Z99,E00-E89,R00-R99,I00-I99,,,,,,,
K00-K95,M00-M99,G00-G99,Z00-Z99,E00-E89,R00-R99,I00-I99,S00-T88,,,,,,
N00-N99,,,,,,,,,,,,,
N00-N99,F01-F99,,,,,,,,,,,,
N00-N99,F01-F99,I00-I99,,,,,,,,,,,
N00-N99,F01-F99,I00-I99,R00-R99,,,,,,,,,,
N00-N99,F01-F99,I00-I99,R00-R99,E00-E89,,,,,,,,,
N00-N99,F01-F99,I00-I99,R00-R99,E00-E89,M00-M99,,,,,,,,
N00-N99,F01-F99,I00-I99,R00-R99,E00-E89,M00-M99,H60-H95,,,,,,,
N00-N99,F01-F99,I00-I99,R00-R99,E00-E89,M00-M99,H60-H95,K00-K95,,,,,,
N00-N99,F01-F99,I00-I99,R00-R99,E00-E89,M00-M99,H60-H95,K00-K95,Z00-Z99,,,,,
N00-N99,F01-F99,I00-I99,R00-R99,E00-E89,M00-M99,H60-H95,K00-K95,Z00-Z99,G00-G99,,,,
Z00-Z99,,,,,,,,,,,,,
Z00-Z99,J00-J99,,,,,,,,,,,,
```

```
M17,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,H04,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,H04,H52,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,H04,H52,R51,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,H04,H52,R51,M50,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,H04,H52,R51,M50,H10,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,H04,H52,R51,M50,H10,I69,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,H04,H52,R51,M50,H10,I69,J41,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,H04,H52,R51,M50,H10,I69,J41,S80,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
M17,I10,H00,H04,H52,R51,M50,H10,I69,J41,S80,S90,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,G54,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,G54,Z45,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,G54,Z45,Z95,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,G54,Z45,Z95,E11,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,G54,Z45,Z95,E11,R42,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,G54,Z45,Z95,E11,R42,I10,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,G54,Z45,Z95,E11,R42,I10,M77,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,G54,Z45,Z95,E11,R42,I10,M77,I73,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
K29,K43,K58,K51,K80,M54,M79,G54,Z45,Z95,E11,R42,I10,M77,I73,S52,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

Анализа на асоцијации

- Пример добиени правила од нулто ниво:
  - {Pregnancy, childbirth and the puerperium; Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified} -> {Factors influencing health status and contact with health services} confidence=0.816 support=0.006
  - {Diseases of the nervous system; Endocrine, nutritional and metabolic diseases; Diseases of the musculoskeletal system and connective tissue} -> {Diseases of the circulatory system} confidence=0.740 support=0.007
- Пример добиени правила од прво ниво:
  - {Complications predominantly related to the puerperium} -> {Persons encountering health services in circumstances related to reproduction} confidence=0.916 support=0.006
  - {Symptoms and signs involving the digestive system and abdomen; Spondylopathies} -> {Other dorsopathies} confidence=0.657 support=0.009
  - {Ischemic heart diseases; Episodic and paroxysmal disorders} -> {Hypertensive diseases} confidence=0.696 support=0.005
- Пример добиени правила од второ ниво:
  - {Type 1 diabetes mellitus; Essential (primary) hypertension} -> {Type 2 diabetes mellitus} confidence=0.919 support=0.007
  - {Angina pectoris; Cardiomyopathy} -> {Essential (primary) hypertension} confidence=0.703 support=0.005
  - {Thoracic, thoracolumbar, and lumbosacral intervertebral disc disorders; Abdominal and pelvic pain} -> {Dorsalgia} confidence=0.559 support=0.007

Класификација според специфична болест или група на болести

- Втор начин за откривање на врски меѓу болестите – корелација преку анализа на фреквенција на појавување на болести и возраст во следните чекори:
  1. Податочното множество повторно се сортира прво по пациент па потоа по датум
  2. За секој пациент се креира вектор на фреквенции на појавување на секоја од болестите освен целната болест за која се прави моделот – броењето на болестите се врши само до првото дијагностицирање на целната болест и на овој вектор се додава тогашната возраст на пациентот
  3. Од ова множество со непрекинати атрибути се добива множество на дискретни атрибути преку дискретизација
  4. Во зависност од типот на множеството се тренира различна група на модели

Креирање на множествата

- Множествата се креирани во однос на специфична болест (или група на болести) – избрани според релевантност и задоволувајќи со доволен број на позитивни примероци
- Секој пациент станува запис во множеството
- Секоја болест (доколку се работи на ниво 2) или група болести (доколку се работи на поопштите нивоа) станува атрибут на множеството
- Класата е бинарен атрибут кој кажува дали пациентот бил или не бил дијагностициран со болеста (или групата болести)

Дискретни наспроти континуирани множества

- Во дискретните множества, болестите преставуваат бинарен атрибут кој ни кажуваат дали пациентот до првата дијагноза на болеста избрана како класа ја имал таа болест
- Во непрекинатите множества болестите претставуваат атрибути кои ни кажуваат колку пати пациентот бил дијагностициран со таа болест до првото појавување на болеста избрана како класа
- Дополнително, во множествата има атрибут возраст, кој ја кажува возраста на која пациентот првпат ја добил дијагнозата за болеста избрана како класа (доколку класниот атрибут е 1)
- Инаку возраста на последно-дијагностицираната болест (доколку класниот атрибут е 0)

Класификација според специфична болест или група на болести

- За непрекинатите множества се тренираат следните типови на модели:
  - Логистичка регресија
  - K најблиски соседи – Евклидово растојание, K=5, тежинско-базирана класификација во однос на растојанието
  - Невронска мрежа – трослојна мрежа со 15 скриени неврони
  - SVM – RBF кернел функција
- Параметрите за секој од моделите беа поставени по извршена валидација со случаен стратификуван примерок

Класификација според специфична болест или група на болести

- За дискретните множества се тренираат следните типови на модели:
  - Наивен Баесов класификатор
  - CN2 правила на одлука – минимум покриеност: 2
  - Дрво на одлука – минимум инстанци во лист: 5
  - Random Forest – 20 дрва
- Параметрите за секој од моделите беа поставени по извршена валидација со случаен стратификуван примерок

Резултати

НЕПРЕКИНАТИ МНОЖЕСТВА

		Класификатор			
		Логистичка регресија	К најблиски соседи	Невронска мрежа	SVM
Податочно множество	A00-B99	0,802	0,859	0,859	0,802
	E00-E89	0,749	0,836	0,825	0,740
	G00-G99	0,874	0,894	0,879	0,874
	I00-I99	0,799	0,811	0,812	0,813
	K00-K95	0,599	0,729	0,736	0,596
	B00-B09	0,957	0,962	0,956	0,957
	C15-C26	0,989	0,989	0,989	0,989
	E08-E13	0,904	0,927	0,933	0,904
	I10-I16	0,803	0,834	0,837	0,834
	K70-K77	0,977	0,980	0,976	0,977
	I10	0,810	0,826	0,825	0,754
Просек		0,842	0,877	0,875	0,840

ДИСКРЕТНИ МНОЖЕСТВА

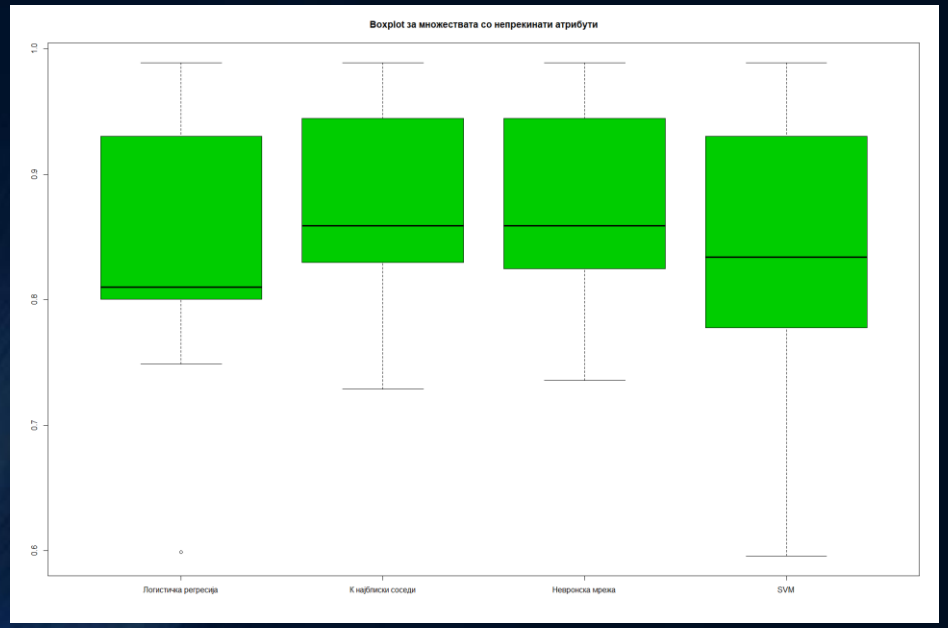
		Класификатор			
		Наивен Баесов	CN2 правила	Дрво на одлука	Random Forest
Податочно множество	A00-B99	0,807	0,849	0,859	0,863
	E00-E89	0,817	0,825	0,833	0,837
	G00-G99	0,874	0,901	0,874	0,905
	I00-I99	0,800	0,796	0,798	0,805
	K00-K95	0,739	0,727	0,733	0,739
	B00-B09	0,954	0,961	0,957	0,957
	C15-C26	0,972	0,988	0,989	0,989
	E08-E13	0,913	0,926	0,904	0,907
	I10-I16	0,800	0,815	0,818	0,826
	K70-K77	0,970	0,977	0,977	0,977
	I10	0,795	0,821	0,819	0,733
Просек		0,858	0,871	0,869	0,867

Евалуација на моделите

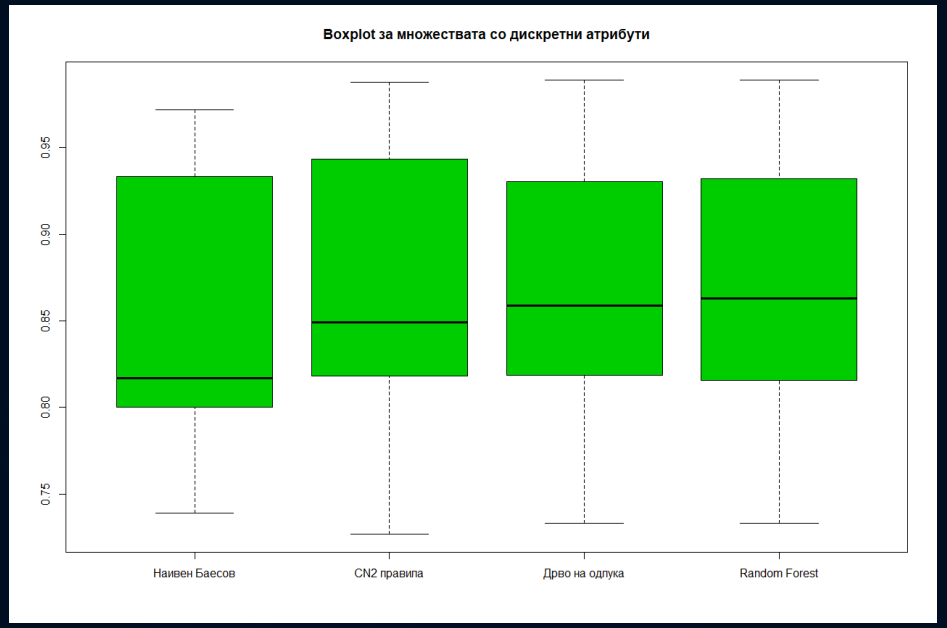
- Со цел да се спореди перформансот на моделите F1 мерката беше користена како главен критериум
- Применувајќи t-тест за споредба на просечниот перформанс на невронската мрежа и KNN алгоритамот, кои покажаа најдобри резултати кај непрекинатите множества, не беше откриена значајност
- Применувајќи t-тест за споредба на просечниот перформанс на Random Forest и CN2 алгоритамот, кои покажаа најдобри резултати кај дискретните множества, исто така не беше откриена значајност

Евалуација на моделите

НЕПРЕКИНАТИ МНОЖЕСТВА



ДИСКРЕТНИ МНОЖЕСТВА



Прашања?

Ви благодариме за вниманието :D