Универзитет у Београду

Математички факултет

Ана Величковић 170/2019

Јелена Митровић 357/2020

Одређивање правила придруживања на основу садржаја базе STUD2020

Семинарски рад

Професор: др Ненад Митић

Универзитет у Београду

Математички факултет

Београд, 2023.

Садржај

1. Увод

Базе података представљају кључну инфраструктуру за складиштење и управљање подацима. Релационе базе података постале су основ модерног информативног друштва, признате и омиљене због своје широке применљивости и ефикасног управљања подацима.

База података је организовани скуп постојаних међусобно повезаних података који се користе од стране система у неком окружењу. Систем прикупља, чува, обрађује и испоручује информације тако да буду доступне и употребљиве кориснику који жели да их користи. Корисници приступају бази података преко упита коришћењем кључних речи. Постоје различите врсте база података.

Релациона база података је врста базе података где су подаци засновани на релационом моделу. Подаци се организују у скуп **релација** између којих су дефинисане одређене везе, база је кориснику приказана у облику табеле. Релација је комплетна табела, а сваки ред у табели јесте торка. Табеле садрже атрибуте, то су колоне у табели.

Релациони упитни језик јесте језик који се користи за комуникацију са релационом базом података, тј. помоћу њега се пишу одређени упити. Језик који је коришћен у овом раду је **SQL**. Више о базама података можете видети у литератури[1].

Правила придруживања користе се за анализу потрошачких корпи, идентификујући релевантне везе између производа у трансакцијама. Рад истражује значајне узорке у скупу података кроз мере квалитета правила придруживања.

2. База података – Stud2020

У овом раду кроз садржај базе података **stud2020**, истражују се фактори који утичу на успех студената, фокусирајући се на време полагања испита, расподелу оцена и корелације између предмета и испитних рокова. Упити над базом података генеришу кључне информације, а анализа података пружа увид у динамику студирања.

База се састоји из следећих табела: dosije, ispit, upisankurs, ispitnirok, upisgodine, kurs, semestar, skolskagodina, studentskistatus, studijskiprogram, predmet, priznatispit, predmetprograma, dosijeext, uslovnipredmet и nivokvalifikacije. У наставку следи опис коришћених табела у овом истраживању:

- **Dosije** – садржи информације о студенту
  - Indeks – индекс студента (PK - примарни кључ)
  - IdPrograma – идентификатор студијског програма који студент студира
  - Ime – име студента
  - Prezime – презиме студента
  - Pol – пол студента {z,m}
  - MestoRodjenja – место где је рођен студент
  - IdStatusa – идентификатор статуса студента
  - DatUpisa – датум уписа на факултет
  - DatDiplomiranja – уколико је дипломирао, када је дипломирао

- **Ispit** – садржи информације о испитима
  - SkGodina – школска година полагања испита (PK)
  - OznakaRoka – ознака испитног рока (PK)
  - Indeks – индекс студента
  - IdPredmeta – идентификатор предмета
  - Semestar – редни број семестра из ког је предмет
  - Status – статус полагања испита одређеног студента, а вредности су:
    - ’p’ - пријављен
    - ’o' - полагао
    - ’n’ – није изашао
    - ’d’ - дисквалификован
    - ’x’ - поништен
    - ’s’ - одустао
  - DatPolaganja – датум полагања испита
  - Poeni – освојен број поена на испиту
  - Ocena – добијена оцена на испиту

- **IspitniRok** – садржи информације о испитним роковима
  - SkGodina – школска година испитног рока (PK)
  - OznakaRoka – ознака испитног рока (PK)
  - Naziv – назив испитног рока
  - DatPocetka – датум почетка испитног рока
  - DatKraja – датум краја испитног рока

- **UpisanKurs** – садржи информације о уписаном курсу
  - Indeks - индекс студента (PK)
  - IdPredmeta - идентификатор предмета (PK)
  - SkGodina - школска година испитног рока (PK)

    - Semestar - редни број семестра из ког је предмет

- **StudijskiProgram** – садржи информације о студијском програму
    - Id – идентификатор студијског програма (PK)
    - Oznaka– ознака студијског програма
    - Naziv – назив студијског програма
    - IdNivoa – идентификатор степена студија
    - ObimESPB – број ЕСПБ студијског програма {30..300}
    - Zvanje – звање након завршетка студијског програма
    - Opis – опис студијског програма

- **PredmetPrograma** – садржи информације о предметима одређеног програма
    - IdPredmeta – идентификатор предмета (PK)
    - IdPrograma – идентификатор студијског програма (PK)
    - Vrsta – врста предмета {obavezan, izborni}
    - Semestar - редни број семестра из ког је предмет

- **Predmet** – садржи информације о предмету
    - Id – идентификатор предмета (PK)
    - Oznaka – ознака предмета
    - Naziv – назив предмета
    - ESPB – број ЕСПБ предмета

3. Опис проблема

Ово истраживање података над базом stud2020, фокусира се на анализу кључних фактора који утичу на успех студената. Циљ рада је идентификовати оптимално време полагања испита, проучити расподелу просечних оцена у различитим испитним роковима, те истражити евентуалне корелације између одређених предмета и испитних рокова. У следећем одељку је приказано решавање овог проблема. Техника која је коришћена у овом истраживању јесте техника правила придруживања. Примењени су алгоритми **apriori** и **FPgrowth**, који су описани у наставку. Коришћени су алати Python-a и IBM SPSS Modeler-a за анализу. Анализе су спроведене на различитим табелама података, а резултати су категоризовани и поређани ради дубљег разумевања међусобних веза између података.

3.1. Правила придруживања

Правила придруживања су истраживачка техника која проналази занимљива правила у огромним скуповима података. То је процес који омогућава проналажење скривених образаца. Проналазе се правила која предвиђају појављивање ставки на основу појављивања одређених других ставки.

Правила придруживања имају широку примену, а један од најчешћих примера јесте коришћење правила придруживања за анализу потрошачке корпе. Сви производи које купац купи јесте једна трансакција која представља један запис тј. слог. Свака трансакција се састоји из скупа ставки.

Правила придруживања имају две фазе:

- проналажење честих скупова

- генерисање правила на основу резултата

Деф. Правила придруживања

*Нека су A и B два скупа ставки. Тада се правилно у ознаци A => B назива правило придруживања са минималном подршком **minsup** и минималном поузданошћу **minconf** ако важи:*

- *подршка скупа A U B је ≥ minsup*

- *поузданост правила A => B је ≥ minconf.*

Најважније мере квалитета правила придруживања су:

-Подршка (support) – учесталост појављивања одређеног скупа елемената трансакције у скупу података, рачуна се као количник броја трансакција које садрже А и Б у односу на укупан број трансакција

$$\text{Sup}(A \Rightarrow B) = \frac{\#(A \cup B)}{N}$$

-Поузданост (confidence) – узрочност присутно у правилу, тј. условна вероватноћа да су ставке на десној страни правила присутне ако су пристуне на

левој страни правила, рачуна се као количник броја трансакција које садрже А и Б у односу на број трансакција које садрже А

$$\text{Conf(A=>B)} = \frac{\#(A \cup B)}{\#(A)}$$

Циљ одређивања правила придруживања је наћи сва правила чија је подршка већа или једнака од изабраног минималног прага за подршку као и поузданост већу или једнаку од израбраног минималног прага за поузданост. Често као резултат се добије велики број правила, а пошто је задатак наћи занимљива правила, често се тај праг поставља високо нпр. на 80%. Нека правила која су на тај начин добијена нису интересантна, јер повезују независне ставке или постоје везе између ставки које су већ познате. Зато се често примењује и следећа мера:

- **Лифт** $\text{Lift} = \frac{conf\ (A \rightarrow B)}{\sup(B)}$

  Правило A->B је занимљиво ако је Lift≠1.

У наставку, описане су још неке мере интересантности правила придруживања. У те сврхе, прво је описана **табела контигената** - табела учесталости појављивања ставки у скупу ставки.

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

f_{10} - број трансакција које садрже само А

f_{11} - број трансакција које садрже и А и Б

f_{01} - број трансакција које садрже само Б

f_{00} - број трансакција које не садрже ни А ни Б

f_{1+} - бројач подршке за А

f_{0+} - бројач подршке за $\bar{A}$

f_{+1} - бројач подршке за В

f_{+o} - бројач подршке за $\overline{B}$

Постоје случајеви у којима Лифт мера не пружа адекватну информацију о важности правила придруживања. Из тог разлога, потребно је користити меру која спаја величину и јачину ефекта правила придруживања - **Piatetsky-Shapiro** мера.

$$\text{PS} = s(A,B) - s(A) * s(B) = \frac{f11}{N} - \frac{(f1+*f+1)}{N\text{^}2}$$

Вредност PS(A,B) = 0 указује на међусобну независност A и B, док вредност различита од нуле указује да између A и B постоји корелисаност. У зависност од знака PS(A,B) , колерисаност може бити позитивна или негативна.

Као једна од мера може се користити и **Пирсонов коефицијент корелације**:

$$\rho_{ij} = \frac{sup(i,j) - sup(i) \cdot sup(j)}{\sqrt{sup(i) \cdot sup(j) \cdot (1 - sup(i)) \cdot (1 - sup(j))}}$$

Још једна занимљива мера је мера под називом **Поузданост свих**:

$$all - confidence(X) = \frac{sup(X)}{max_{x \in X}(sup(x)}$$

Поменута мера подржава затворење ка ниже, а њено значење је да сва правила која могу да се изведу из X имају подршку једнаку бар **all - confidence(X)** .

Постоји још много мера које неће бити детаљно обрађене, али су наведене ради потпуности овог истраживања:

- Однос камата
- Deployability
- χ^2 мера
- ИС мера
- Косинусна мера колоне.

Добра мера M мора да задовољава следеће три особине:

- M(A, Б) = 0 ако су A и Б статистички независне,

-M(A, Б) се монотоно повећава са П(A, Б) када П(A) и П(Б) остају непромењене,

- M(A, Б) се монотоно смањује са П(A) [или П(Б)] када П(A, Б) и П(Б) [или П(A)] остају непромењене

Алгоритми

Алгоритми који су коришћени код технике правила придруживања су **apriori** и **FPgrowth**.

Априори алгоритам у фази генерисања честих скупова ставки користи особине подршке како би се смањио број скупова ставки за које је потребно израчунати подршку да би одредио да ли је скуп ставки чест. Алгоритам за смањење кандидатских скупова ставки за које је потребно израчунати подршку користи **априори принцип** и **анти-монотоност.**. .

FPGrowth или алгоритам ФП раста је алгоритам који обилази решетку **'у дубину'**. Посебно је користан за проналажења честих скупова ставки у великим базама података, јер може бити вишеструко ефикаснији од алгоритама који обилазе решетку 'по ширини'. Овај алгоритам се састоји из неколико корака. У првом кораку се прави **ФП дрво** које на почетку не садржи ниједну ставку, а које се допуњује тако што сваки чвор садржи нову ставку и број појављивања у трансакцијама, а то се формира опадајуће по том броју. У другом кораку алгоритам пролази кроз дрво како би идентификовао фреквентне скупове ставки тј. комбинације ставки које се често појављују заједно. У наредном кораку, на темељу фреквентних скупова, алгоритам генерише правила придруживања. То ради рекурзивно где одређује правила за сваки подскуп ставки. Правила се генеришу на темељу подршке и поузданости фрекевентних скупова, а могу и по другим квалитетима мера, као што је нпр. лифт мера. Више о самим алгоритмима, као и правилима придруживања можете видети у литератури[4].

4. Резултати

4.1. Приступ

Извршени су упити над базом података и добијена је помоћна табела **BrojPolaganja**. Табела садржи информације за сваког студента о броју полагања одређеног предмета од стране једног студента. Спојене су следеће табеле BrojPolaganja (BP), Predmet (P), Ispit (I), PredmetPrograma (PP) и StudijskiProgram (SP) и из њих су издвојени следећи атрибути: PP.idPrograma, SP.Naziv, I.idPredmeta, P.Naziv, I.OznakaRoka, I.skGodina, prosecnaOcena, prosecanBrojPolaganja. Резултати који су добијени на овај начин коришћени су у даљој анализи овог истраживачког рада под називом **podaciPoIdProgramu** (више о табели у Додатку 1).

За наредну анализу било је потребно одредити успешност за сваки појединачни испит у сваком испитном року. Другим речима, прикупљени су подаци о томе колико је којих оцена постигнуто за сваки специфичан испит у сваком испитном року. За ту сврху, направљено је пет помоћних табела. У једној табели су били представљени подаци о броју оцена 'пет', док је у другој табели био приказан број оцена 'шест', и тако даље. На овај начин су различите табеле одражавале различите оцене. Поред ових информација, свака од табела је садражала информацију о идентификатору предмета, школској години и ознаци рока. Ове помоћне табеле су спојене са табелом Ispit и добијена је табела са следећим атрибутима: IdPredmeta, SkGodina, OznakaRoka, BrojOnihKojiSuPali, BrojSestica, BrojSedmica, BrojOsmica, BrojDevetki и BrojDesetki. Добијени подаци коришћени су у даљем истраживању под називом **podaciOBrojuOcenaZaSvakiIspitUSvakomRoku** (више о табели у Додатку 2).

Наредна табела коришћена за анализу и решавање проблема, садржи информације о успешности појединачног студента у сваком испитном року.Табела садржи информације о индексу, школској години, ознаци рока, броју испита које је студент положио у датом испитном року, просечној оцени коју је студент добио на положеним испитима у том року као и о просечној оцени студента до сада. Подаци добијени на овакав начин, коришћени су за даљу анализу под називом **podaciOStudentuIBrojuPolozenihIspitaUSvakomRoku** (више о табели у Додатку 3).

Алати који су коришћени у истраживању су **Python** и **IBM SPSS Modeler**. У SPSS-у је урађена детаљна анализа над подацима, а у Python-у је извршена провера.

У Пајтону је извршена анализа над табелама података:

- podaciPoIdProgramu, приказано у фајлу association_rules_2.ipynb, који се налази у додатку уз PDF документ.

- podaciOBrojuOcenaZaSvakiIspitUSvakomRoku, приказано у фајлу prvaAnalizaPajton.ipynb, који се налази у додатку уз PDF документ.

- podaciOStudentuIBrojuPolozenihIspitaUSvakomRoku, приказано у фајлу prvaAnalizaPajton.ipynb, који се налази у додатку уз PDF документ.

За анализу у Пајтон коришћен **jupyter notebook** и пакети **pandas** и **mlxtend**. Pandas омогућава да се увезу подаци из различитих типова датотека. Вршено је **препроцесирање** података где су изабрани одређени атрибути који су коришћени за даљу анализу.

Извршене су различите поделе, где су издвојени следећи атрибути:

- idPredmeta, oznakaRoka, prosecnaOcena
- idPredmeta, brojPolaganja
- idPredmeta, oznakaRoka
- idPredmeta, skGodina, OznakaRoka, brojOnihKojiSuPaliPredmet, brojsestica, brojsedmica, brojosmica, brojdevetki, brojdesetki

Коришћењем **TransactionEncoder** објекта и примене методе **fit_transform** врши се трансформација података у формат погодан за алгоритам ФП раст. Формат података који је потребан за овај алгоритам у Python-у јесте у облику листе трансакција, где је свака трансакција листа или скуп података тј. атрибута. Формат је познат као "**sparse transaction format**". Након тога је примењен алгоритам ФП раста за издвајање честих скупова.

Применом алгоритма **association_rules** над честим скуповима ставки, са мером квалитета "поузданост" са минималним прагом 50% добијена су правила која су сортирана по lift мери ради поређења са излазним подацима који су добијени алатом SPSS-ом.

4.2. Анализа правила

Правила придруживања пружају информације о успешности неког студента или предмета у одређеном року, у зависности од студијског програма, степена студија и школске године.

Значајна правила добијена у Пајтону:

За предмете и просечну оцену:

- {IDPREDMETA} -> {PROSECNAOCENA}
  IDPREDMETA = 2211, PROSECNAOCENA=9,3, Лифт = 79,6
  (NazivPredmeta = ‘Uvod u bioinformatiku’)

- {IDPREDMETA, OZNAKAROKA} -> {PROSECNAOCENA}
  IDPREDMETA = 2220, OZNAKAROKA = ‘jan1’, PROSECNAOCENA=8.5,
  Лифт = 51.7 (NazivPredmeta = ‘ Verifikacija softvera’)

- {IDPREDMETA, OZNAKAROKA} -> {PROSECNAOCENA}
  IDPREDMETA = 2213, OZNAKAROKA=’sept1’, PROSECNAOCENA=8.5,
  Лифт = 34.5 (NazivPredmeta = ‘Kriptografija’)

- {IDPREDMETA, OZNAKAROKA} -> {PROSECNAOCENA}
  IDPREDMETA = 2225, OZNAKAROKA=’jun1’, PROSECNAOCENA=7.99,
  Лифт = 15.24 (NazivPredmeta = ‘Upravljanje projektima u industriji i nauc’)

- {IDPREDMETA, OZNAKAROKA} -> {PROSECNAOCENA}
  IDPREDMETA =2226, OZNAKAROKA=’sept2’, PROSECNAOCENA=7.99,
  Лифт = 11.43 (NazivPredmeta = ‘Specijalni kurs - Elementi finansijske matematike’)

- {IDPREDMETA} -> {PROSECNAOCENA}
  IDPREDMETA=2252, PROSECNAOCENA=9.0, Лифт = 9.23
  (NazivPredmeta = ‘Naucno-istrazivacki rad 1’)
- {IDPREDMETA, OZNAKAROKA} -> {PROSECNAOCENA}
  IDPREDMETA = 2025, OZNAKAROKA=’jan1’, PROSECNAOCENA=9.0, Лифт = 7.75 (NazivPredmeta = ‘Racunarska grafika’)

- {IDPREDMETA} -> {PROSECNAOCENA}
  IDPREDMETA = 1983, PROSECNAOCENA=9.98, Лифт = 6.03 (NazivPredmeta = ‘Statisticki softver 4’)

За рокове и предмете:

- {OZNAKAROKA} -> {IDPREDMETA}
  OZNAKAROKA=’kom’ , IDPREDMETA = 2127, Лифт = 20.63
  (NazivPredmeta = ‘Master rad’)

- {IDPREDMETA}->{OZNAKAROKA}
  IDPREDMETA=2371, OZNAKAROKA=’jan1’ Лифт =8.77
  (NazivPredmeta = ‘Odabrana poglavlja analize A’)

- {IDPREDMETA}->{OZNAKAROKA}
  IDPREDMETA=2109, OZNAKAROKA=’janps’ Лифт = 8.44
  (NazivPredmeta = ‘Studijski istrazivacki rad 1’)

- {OZNAKAROKA} -> {IDPREDMETA}
  OZNAKAROKA=’sept2’, IDPREDMETA=2221, Лифт = 6.86
  (NazivPredmeta = ‘Obrada digitalnih slika’)

- {OZNAKAROKA} -> {IDPREDMETA}
  OZNAKAROKA=’sept2’, IDPREDMETA=1833, Лифт = 6.86
  (NazivPredmeta = ‘Odabrana poglavlja vangalakticke astronomije’)

- {IDPREDMETA}->{OZNAKAROKA}
  IDPREDMETA=2127, OZNAKAROKA=’janps’ Лифт = 5.63
  (NazivPredmeta = ‘Master rad’)

- {IDPREDMETA}->{OZNAKAROKA}
  IDPREDMETA=2405, OZNAKAROKA=’jan1’, Лифт = 5.26
  (NazivPredmeta = ‘Teorija aproksimacija’)

Сва правила добијена анализом су у прилогу Pravila.txt.

За предмете и број полагања:

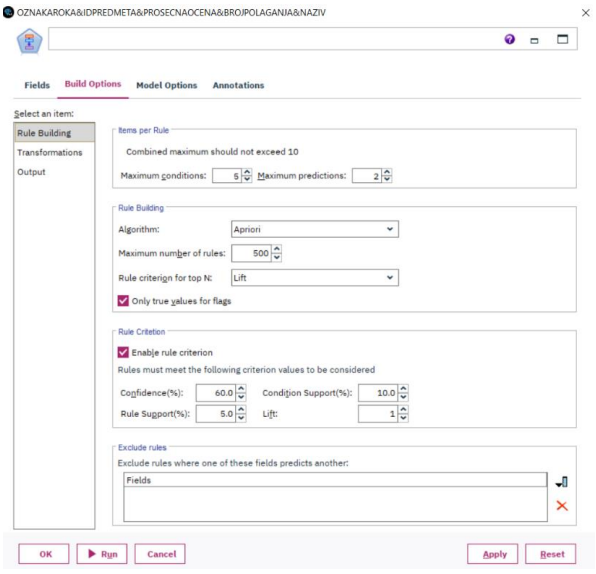
{IDPREDMETA} -> {BROJPOLAGANJA}

- IDPREDMETA=2329, BROJPOLAGANJA=4, Лифт = 28.65
  (NazivPredmeta = ‘Verovatnoca’)
- IDPREDMETA=1979, BROJPOLAGANJA=4, Лифт = 28.65
  (NazivPredmeta = ‘Teorija uzoraka’)
- IDPREDMETA=2417, BROJPOLAGANJA=4, Лифт = 28.65
  (NazivPredmeta = ‘Instrumenti i tehnike astrofizickih posmatranja’)

- IDPREDMETA=1941, BROJPOLAGANJA=4, Лифт = 28.65
  (NazivPredmeta = ‘Kompilacija programskih jezika’)
- IDPREDMETA=2199, BROJPOLAGANJA=1, Лифт = 4.5
  (NazivPredmeta = ‘Medjuzvezdana materija’)
- IDPREDMETA=2207, BROJPOLAGANJA=1, Лифт = 4.5
  (NazivPredmeta = ‘Metodologija strucnog i naucnog rada’)
- IDPREDMETA=2251, BROJPOLAGANJA=1, Лифт = 4.5
  (NazivPredmeta = ‘Specijalni kurs 1’)
- IDPREDMETA=2093, BROJPOLAGANJA=1, Лифт = 4.5
  (NazivPredmeta = ‘Odabrana poglavlja globalne analize’)
- IDPREDMETA=2392, BROJPOLAGANJA=3, Лифт = 2.75
  (NazivPredmeta = ‘Obrada astronomskih posmatranja 1’)
- IDPREDMETA=2388, BROJPOLAGANJA=3, Лифт = 2.75
  (NazivPredmeta = ‘Uvod u nebesku mehaniku’)
- IDPREDMETA=2457, BROJPOLAGANJA=3, Лифт = 2.75
  (NazivPredmeta = ‘Masinsko ucenje’)
- IDPREDMETA=1908, BROJPOLAGANJA=2, Лифт = 2.69
  (NazivPredmeta = ‘Uvod u teorijsku mehaniku’)
- IDPREDMETA=2366, BROJPOLAGANJA=2, Лифт = 2.69
  (NazivPredmeta = ‘Metodika nastave racunarstva A’)

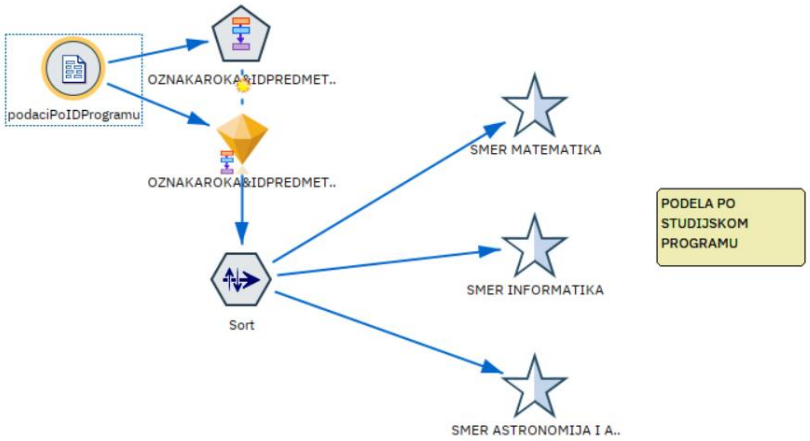
Сва правила добијена анализом у пајтону над овим подацима можете видети у прилогу под називом pravilaPrvaAnaliza.txt.

У SPSS-у је рађена анализа над табелом podaciPoIdProgramu приказано у стримму IP2podaciPoSmeru.str, који се налази у додатку уз PDF документ. Употребом чвора **Association Rules-a** из дела **Association** добијена су одређена правила придруживања. Следећи атрибути су имали двоструку улогу (**Both**) тј. служили су или као услов (**condition**) или као последица (**prediciton**) ових правила: OznakaRoka, IdPredmeta, ProsecnaOcena,BrojPolaganja, Naziv. Коришћен је алгоритам априори, постављен је услов да има максималан број атрибута у услову на 5, а максималан број у последици јесте на 2. Критеријуми по ком су правила извршена јесте минимална поузданост 60%, минимална условна подршка 10% и минимална лифт мера је 1. Дозвољено је да буде изабрано највише 500 правила која су сортирана по лифт мери (слика 1).



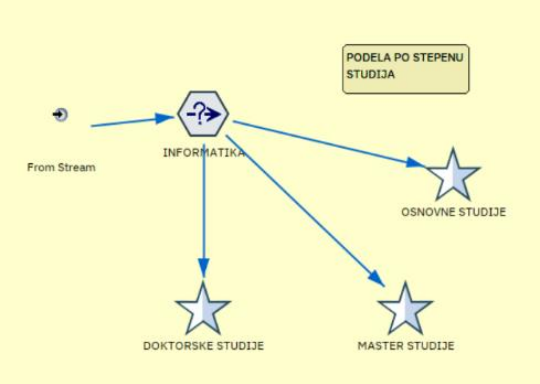
Слика 1. Критеријуми за правила придруживања (IP2podaciPoSmeru.str)

Правила су сортирана опадајуће по лифт мери, како би се извршила детаљна анализа успешности, подаци су подељени по студијском програму: Matematika, Informatika и Astronomija i astrofizika (слика 2).

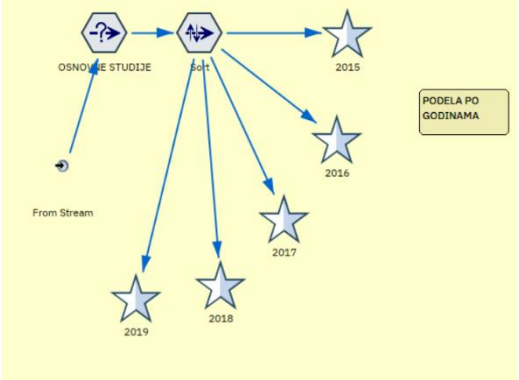


Слика 2. Подела по студијском програму (IP2podaciPoSmeru.str)

Сваки студијски програм је подељен по степену студија на основне, мастер и докторске студије (слика 3).

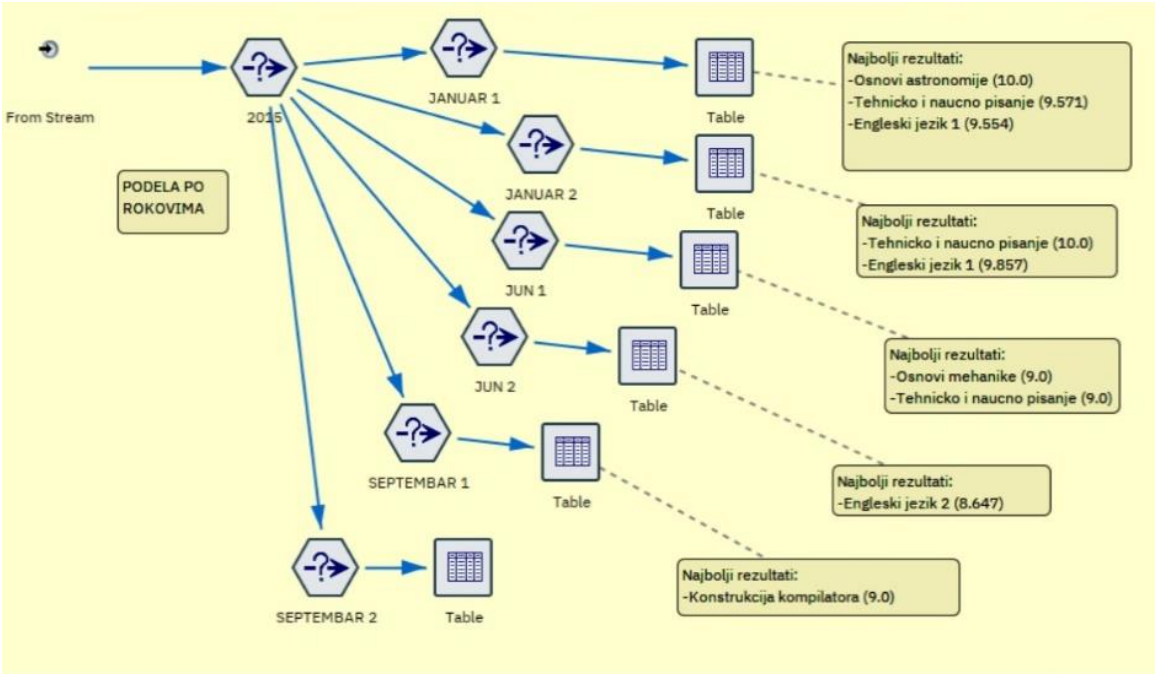


Слика 3. Подела по степену студија (IP2podaciPoSmeru.str)



Слика 4. Подела по школским годинама (IP2podaciPoSmeru.str)

За сваки студијски програм је испитана успешност по годинама: 2015, 2016, 2017, 2018, 2019, (слика 4) а за сваку годину, у зависности од тога који су испитни рокови одржани у тој години су разврставани по ознаци рока (слика 5), а испитни рокови су: јануар 1, јануар 2, јануар ПС, јун 1, јун 2, септембар 1, септембар 2, септембар 3, септембар 4 и ком. За сваки рок су добијени предмети и просечна оцена за те предмете, као и просечан број полагања који представља у просеку из ког пута је студент положио тај предмет.

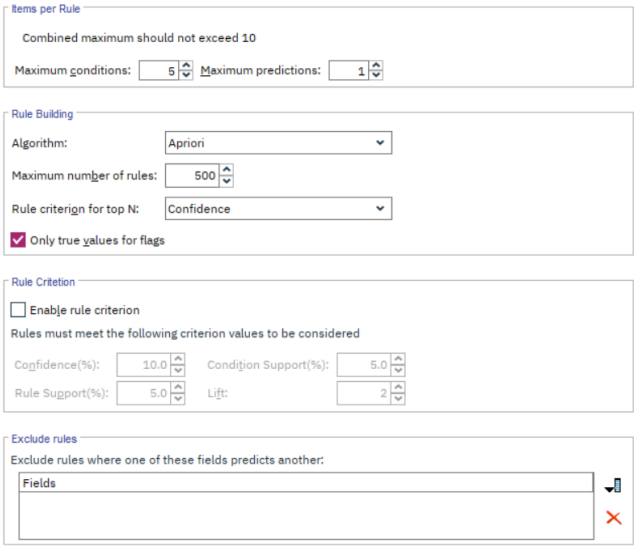


Слика 5. Подела по испитним роковима (IP2podaciPoSmeru.str)

Следећа анализа у SPSS-у рађена је над табелом под називом podaciOBrojuOcenaZaSvakiIspitUSvakomRoku, приказано у стримму IP2BrojOcenaPoRokovimaZaSvakiPredmet.str, који се налази у додатку уз PDF документ. За добијење правила придруживања над овим скупом података, искоришћен је поменут чвор Association Rules. Скоро сви атрибути из ове табеле су коришћени или као услов (condiction) или као последица (prediction), сем атрибута који немају улогу, а то су idPredmeta и skGodina, док се за предвиђање циља користе атрибути brojDesetki, brojDevetki, brojOsmica, brojSedmica, brojSestica, brojOnihKojiSuPali, а као циљни атрибут је издвојен атрибут oznakaRoka. Искоришћен је алгоритам априори, уз следеће услове:

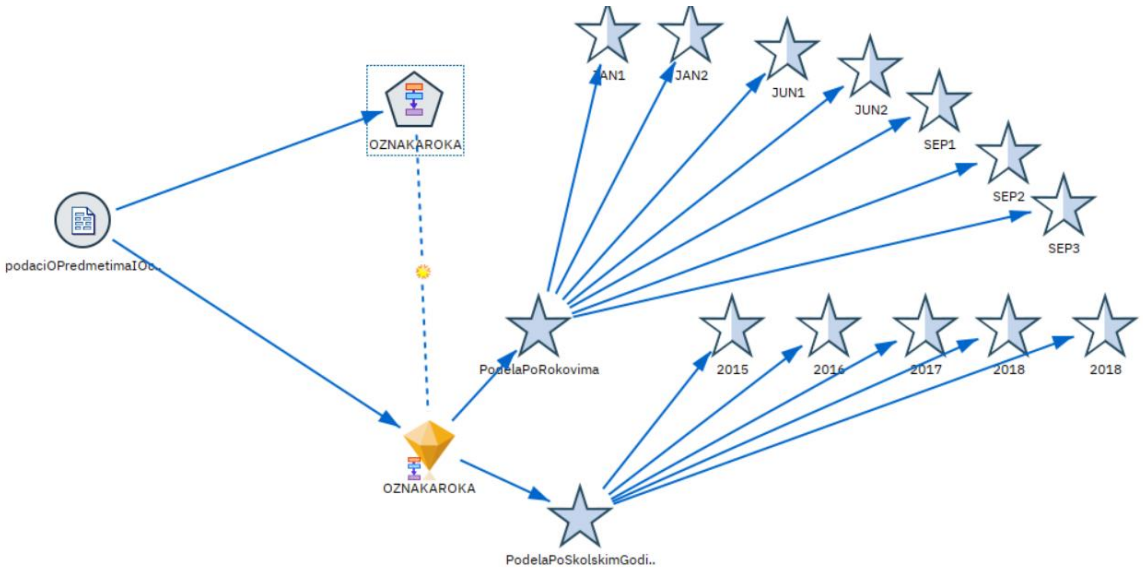
- максималан број атрибута у услову је 5,
- максималан број у последици је 3,
- минимална поузданост правила је 60%,
- минимална условна подршка правила је 10%,
- минимална лифт мера је 1.

Дозвољено је да буде изабрано највише 1000 правила која су сортирана по лифт мери (слика 6).



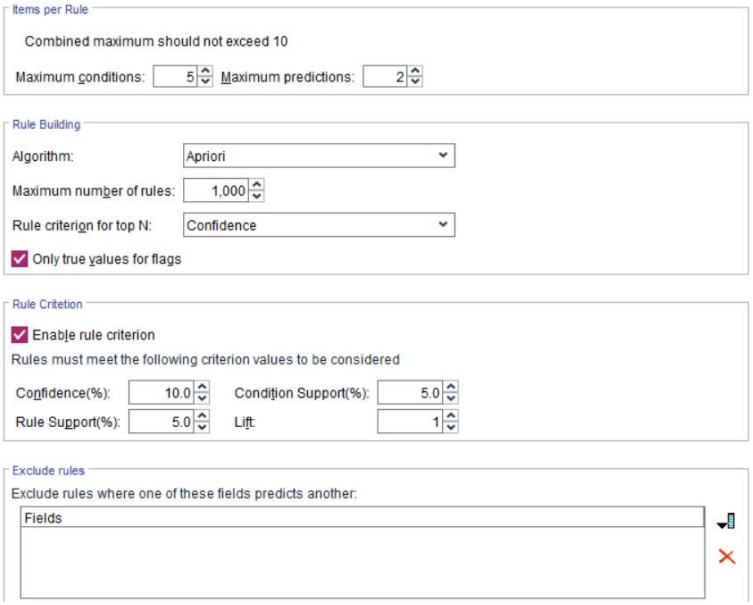
Слика 6. Критеријуми за правила придруживања (IP2BrojOcenaPoRokovimaZaSvakiPredmet.str)

Након сортирања правила по лифт мери опадајуће, извршена је подела по испитним роковима,као и по школској години, ради детаљније анализе (слика 7).



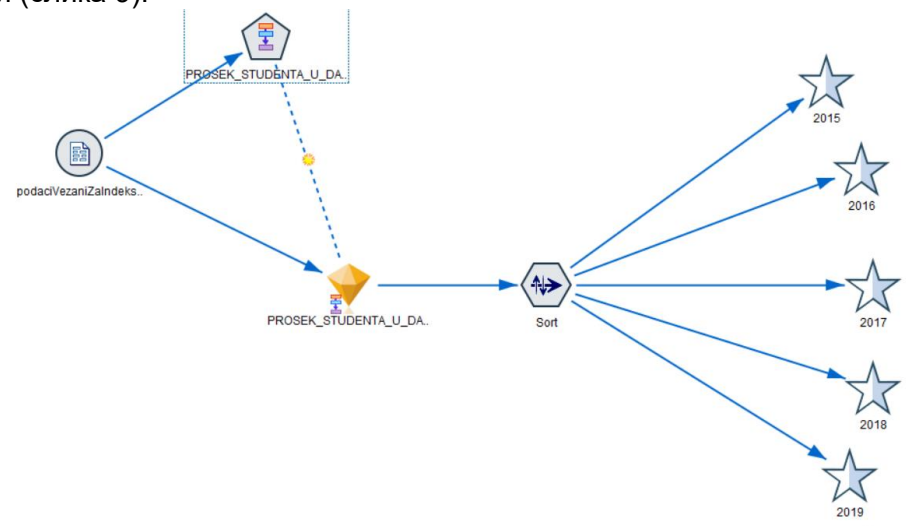
Слика 7. Подела по испитним роковима и школским годинама (IP2BrojOcenaPoRokovimaZaSvakiPredmet.str)

Наредна анализа у SPSS-у извршена је над табелом података podaciOStudentuIBrojuPolozenihIspitaUSvakomRoku, приказано у стримму ip2PodaciOIndeksimaBrojuPolozenihIspitaPoRokuIProseku.str, који се налази у додатку уз PDF документ. За ову аланизу је, такође, искоришћен алгоритам априори из чвора Association Rules. Сви атрибути се користе или као услов, или као последица, осим атрибута indeks, који се није користио при налажењу правила придруживања. За предвиђање се користе skGodina, oznakaRoka, prosek_studenta, а циљни атрибути су broj_polozenih_predmeta_u_datom_roku и prosek_studenata_u_datom_roku_za_polozene_predmete.На слици 8. можемо видети који су критеријуми коршћени за одређивање правила придруживања у овом скупу података.



Слика 8. Критеријуми за правила придруживања (ip2PodaciOIndeksimaBrojuPolozenihIspitaPoRokuIProseku.str)

Након сортирања правила опадајуће по лифт мери, извршена је подела по школској години (слика 9).



Слика 9. Подела по школским годинама (ip2PodaciOIndeksimaBrojuPolozenihIspitaPoRokuIProseku)

Значајна правила:

Из табеле **podaciPoIdProgramu:**

- {NAZIV = 'Matematika'} => {BROJPOLAGANJA >= 2 && BROJPOLAGANJA <= 3},{NAZIV = 'Informatika'}=> {BROJPOLAGANJA >= 1 && BROJPOLAGANJA <= 2}

Предмети на математичком смеру се најчешће полажу из 2. или 3. пута (Лифт = 1,89), а предмети на информатичком смеру се најчешће полажу из 1. или 2. пута (Лифт = 1.48)

- {IDPROGRAMA < 200 && SKGODINA = 2019} => {BROJPOLAGANJA >= 2 && BROJPOLAGANJA <= 3 && PROSECNAOCENA < 6.667}

Предмете на основним студијама 2019. године најчешће су студенти полагали из 2. или 3. пута са просечном оценом мањом од 6.667 (Лифт = 1.49)

- {IDPROGRAMA < 200 && SKGODINA <= 2016} => {PROSECNAOCENA < 6.667}

Најчешћа просечна оцена по предмету 2015. и 2016. године на основним студијама била је испод 6.667 (лифт = 1.49)

- {IDPROGRAMA < 200 && OZNAKAROKA = sept1} => {PROSECNAOCENA <= 6.667}

Просечна оцена у року ‘sept1’ на основним студијама је мања или једнака 6.667

-Студенти на основним студијама највише полажу испите у јунским роковима, а нешто мање у ‘sept1’

- {IDPROGRAMA >= 200 && IDPROGRAMA <= 300} => {BROJPOLAGANJA >= 1 && BROJPOLAGANJA <= 2}

Студенти на мастер студијама углавном положе испит из 1. или 2. пута (Лифт = 1.46)

- {IDPROGRAMA >= 200 && IDPROGRAMA <= 300} => {PROSECNAOCENA >= 8.33}

Велики број испита на мастер студијама је успешно положен код великог броја студената, па је просечна оцена изнад 8.33 (Лифт = 1.42)

-Просечна оцена по предмету на Математичком факултету је најчешће мања од 6.667 (45.08% од укупног броја), између 6.667 и 8.33 је 27.49%, док је изнад 8.33 скоро исти проценат, а то је 27.43%

-Студенти на математичком факултету предмете најчешће полажу испите у року септембру 2 (21.36%), а мање у јуну 2 (12.06%), јануару 1 (11.32%) и јануару 2 (11.19%), док у осталим роковима још мање

Још нека правила која можемо извући, а да нисмо већ поменули у претходним јесу из табеле **podaciOBrojuOcenaZaSvakiIspitUSvakomRoku:**

-Студент најчешће по испитном року полаже један или два испита (чак 85.20% од укупног броја студената)

-Просек студената у највећем броју јесте између 6.8 и 8.4

И још нека из **podaciOStudentuIBrojuPolozenihIspitaUSvakomRoku:**

-Број деветки за један предмет у једном року је мањи од 15 (95%), број осмица је мањи од 12, број шестица мањи од 9, а број људи који су пали један предмет у једном року је у просеку мањи од 8

- {OZNAKAROKA = 'jan2'} =>{BROJOSMICA > 10 && BROJOSMICA < 20}

У јануару 2 у просеку број оцена 8 је између 11 и 19

- {OZNAKAROKA = 'jun1'} => {BROJONIHKOJISUPALI < 8}

У јуну 1 у просеку број људи који падне одређени испит је испод 8

Додатно о правилима може да се пронађе у фолдеру достављеном уз пдф, као и више информација о самим стримовима.

Вршене су анализе у Пајтону и SPSS, следи поређење резултата:

-Бољи резултати су добијени алгоритмом ФП раста

-У Пајтону смо добили информације за тачне индексе предмета у ком испитном року је најбоље полагати поједине предмете, као и из ког пута студенти положе дати предмет

-Резултати у SPSS нам дају бољи увид у то како студенти напредују из године у годину, на којем степену студија су најуспешнији, као и на ком смеру

-Резултати који су добијени и једним и другим алатом су:

*__Најуспешнији рок је септ2, после тога јун1, затим јануарски рокови__

(Студенти на математичком факултету предмете најчешће полажу испите у року септембру 2 (21.36%), а мање у јуну 2 (12.06%), јануару 1 (11.32%) и јануару 2 (11.19%), док у осталим роковима још мање)

*__Предмети на математичком смеру се углавном полажу из 3. или 4.пута, а на информатичком из 1. или 2__.

({NAZIV = 'Matematika'} => {BROJPOLAGANJA >= 2 && BROJPOLAGANJA <= 3},{NAZIV = 'Informatika'}=> {BROJPOLAGANJA >= 1 && BROJPOLAGANJA <= 2}

Предмети на математичком смеру се најчешће полажу из 2. или 3. пута (Лифт = 1,89), а предмети на информатичком смеру се најчешће полажу из 1. или 2. пута (Лифт = 1.48))

*__Студенти на мастер студијама углавном полажу предмете из 1. или 2.пута__

({IDPROGRAMA >= 200 && IDPROGRAMA <= 300} => {BROJPOLAGANJA >= 1 && BROJPOLAGANJA <= 2}

Студенти на мастер студијама углавном положе испит из 1. или 2. пута (Лифт = 1.46))

*__Просечна оцена на мастер студијама је изнад 8.0__

({IDPROGRAMA >= 200 && IDPROGRAMA <= 300} => {PROSECNAOCENA >= 8.33}

Велики број испита на мастер студијама је успешно положен код великог броја студената, па је просечна оцена изнад 8.33 (Лифт = 1.42))

5. Закључак

Овај семинарски рад истражује релевантне узорке у студентској бази података stud2020, кроз пажљиву анализу табела попут испита, досијеа студената итд. откривене су важне безе и информације које пружају увид у факторе који утичу на успех студената. Коришћена су два алата: Пајтон и SPSS ради поређења резултата. Закључак јесте да су боља правила добијена алгоритмом ФП раста. Између резултата добијених у Пајтону и SPSS има разлика, које су наведене у последњем одељку, али су те разлике битне за анализу, али има и доста сличности. Коришћењем напредних техника правила придруживања, имплементираних помоћу наведених алата, идентификоване су кључне тачке фокуса за унапређење квалитета образовања. На темељу анализе података и примене алгоритама придруживања, ово истраживање је идентификовало значајне обрасце и тенденције у вези са успехом студената. Узорци успеха по предметима, програмима, годинама и испитним роковима пружају корисне смернице за прилагођавање студената. Пажљиво примењени алгоритми омогућили су дубље разумевање података. Неки од закључака јесте да су студенти како пролазе године све успешнији, тј. успешнији су у 2019. у односу на године пре. Студенти много више излазе на испите у септембарским роковима, па и у јунским, него у јануарски. Разлог томе је што су студенти вероватно опуштенији на почетку године, док на крају године студенти имају мотивацију као што је на пример да ухвате буџет, заврше студије, заузму место у студентском дому, конкурса за студентски кредит/стипендију или нешто друго. Информатичари су успешнији од математичара по подацима, разлог томе може бити што су бољи студенти, а с друге стране у корист математичара јесте што су можда поједини предмети на том смеру тежи. Студенти који се одлуче за мастер студије, углавном имају бољи просек од студената који су на основним студијама. Ова анализа је помогла око бољег увида у рад студената на Математичком факултету, али и о тежини појединих предмета. За детаљнију анализу, о појединим роковима, школским годинама, студентима погледати фајлове који долазе уз овај семинарски рад.

Литература

[1] C.J.Date, SQL and Relational Theory: How to Write Accurate SQL Code, O'Reilly Media, Inc, 2nd Edition, 2011

[2] C.J.Date: Database Design and Relational Theory: Normal Forms and All That Jazz, O'Reilly, 2012.

[3] C.J.Date: An Introduction to Database Systems, VIII ed, Addison Wesley Inc, 2004.

[4] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar: Introduction to Data Mining, 2nd ed, Pearson Education, 2019.

[5] Xindong Wu, Vipin Kumar (eds.): The Top Ten Algorithms in Data Mining, CRC Press, 2009.

[6] Charu C. Aggarwal: Data Mining The Textbook, Springer, 2015.

[7] Презентације са сајта професора др. Ненада Митића из предмета Истраживање података 1

[8] Презентације са сајта професора др. Ненада Митића из предмета Релационе базе података

[9] Истраживање података, Никола Ајзенхамер, Ања Букуров, Војислав Станковић (2017)

[10] An Overview of Association Rule Mining Algorithms, Trupti A. Kumbhare, Prof. Santosh V. Chobe

[11] Association Rule Mining: A Survey, Qiankun Zhao, Sourav S. Bhowmick

Додатак 1

podaciPoIdPrograma – садржи информације о предметима и просечним оценама

- IDPROGRAMA - идентификатор студијског програма који студент студира
- NAZIV - назив студијског програма
- IDPREDMETA - идентификатор предмета
- NAZIVPREDMETA – назив предмета
- OZNAKAROKA - ознака испитног рока
- SKGODINA - школска година полагања испита
- PROSECNAOCENA – просечна оцена појединог предмета на одређеном студијском програму
- BROJPOLAGANJA - колико пута су у просеку студенти који су изашли на испит полагали тај предмет

Додатак 2

podaciOBrojuOcenaZaSvakiIspitUSvakomRoku- садржи информације о броју оцена за сваки предмет полаган у свим испитним роковима

- IDPREDMETA - идентификатор предмета
- SKGODINA - школска година полагања испита
- OZNAKAROKA - ознака испитног рока
- BROJONIHKOJISUPALI - број студената који су пали одређен предмет у датом року
- BROJSESTICA - број добијених оцена шест на испиту из одређеног предмета у датом року
- BROJSEDMICA - број добијених оцена седам на испиту из одређеног предмета у датом року
- BROJOSMICA - број добијених оцена осам на испиту из одређеног предмета у датом року
- BROJDEVETKI - број добијених оцена девет на испиту из одређеног предмета у датом року
- BROJDESETKI - број добијених оцена десет на испиту из одређеног предмета у датом року

Додатак 3

podaciOStudentuIBrojuPolozenihIspitaUSvakomRoku - садржи податке о броју испита које је сваки студент положио у сваком року, као и просечну оцену

- INDEKS - индекс студента

- SKGODINA - школска година полагања испита

- OZNAKAROKA - ознака испитног рока

- PROSEK_STUDENTA - просечна оцена студента

- BROJ_POLOZENIH_PREDMETA_U_DATOM_ROKU- број испита које је студент успешно положио у датом испитном року

- PROSEK_STUDENTA_U_DATOM_ROKU_ZA_POLOZENE_ISPITE - просечна оцена коју је студент остварио успешним полагањем испита у датом испитном року