Оцена популарности станице према објектима у околини

Горана Гојић
Факултет техничких наука
Универзитет у Новом Саду
Трг Доситеја Обрадовића 6
21000 Нови Сад
e-пошта: gorana.gojic@uns.ac.rs

Ангелина Вујановић
Факултет техничких наука
Универзитет у Новом Саду
Трг Доситеја Обрадовића 6
21000 Нови Сад
e-пошта: avujanovic@uns.ac.rs

Радован Туровић
Факултет техничких наука
Универзитет у Новом Саду
Трг Доситеја Обрадовића 6
21000 Нови Сад
e-пошта: radovan.turovic@uns.ac.rs

***Сажетак*---У данашње време приметна је експанзија система за изнајмљивање бицикала у великим градовима, у толикој мери да представљају озбиљну алтернативу систему јавног превоза. Са порастом популарности ових система, повећавају се и приноси, али и улагања у њих. Како компаније које изнајмљују бицикле не би пословале са губицима, посебна се пажња посвећује предвиђању локација нових станица за бицикле и уклањању непотребних станица, анализом постојећих станица и фактора које утичу на популарност сваке појединачне станице у систему. Идеја овог пројекта је фокусирање на људима битне објекте који се у одређеној мери налазе у близини станица, како би се близином неког типа објекта објаснила популарност станице и пронашли шаблони који указују на то где би требало поставити нову станицу и/или објашњавају непопуларност неке постојеће станице. У складу са тим спроведене су статистичке анализе које су дале корелације између популарности станица и броја објеката одређеног типа. Развијен је и модел за предвиђање популарности станица америчке компаније Capital BikeShare, коришћењем алгоритма стабла одлучивања. Модел је у већој мери дао резултате који су очекивани. Показало се позитиван утицај на популарност станице имају знаменитости, продавнице, објекти за забаву и релаксацију, туристички објекти, као и објекти за куповину хране и пића. Позитиван утицај објеката јавног превоза није подвргнут моделом стабла одлучивања, већ само статистичким анализама.**

***Кључне речи*---рударење података, стабло одлучивања, корелација, статистика, Capital Bikeshare, станице бицикала, промет, модел система**

1. Увод

Бицикл као превозно средство данас постаје све популарнији због свог позитивног утицаја на здравље човека, могућности јефтиног и брзог транспорта и многих других предности [1], [2]. Са порастом популарности бицикла, почела је да расте и популарност система станица за изнајмљивање бицикала. Популаризација система за изнајмљивање бицикала је постала нарочито изражена у великим градовима у којима је транспорт отежан услед великих гужви у саобраћају. Компаније које су власници ових система нуде својим корисницима могућност да изнајме бицикл на некој од станица за бицикле, користе га за одређену новчану надокнаду и врате на било коју станицу за бицикле из система станица. Са становишта компаније која је власник оваквог система, исплатива станица је она на којој је промет задовољавајућ тј. она на основу које компанија остварује профит. Да би станица остваривала профит, она мора имати добар промет. Станица која има мали промет или уопште нема промет вероватно није довољно значајна за кориснике система. Претпоставка је да се вероватноћа да ће станица имати добар промет смањује, уколико је корисницима потребно више времена да дођу до ње. Због тога је компанији која поседује систем за изнајмљивање бицикала битно да уочи факторе који позитивно, односно негативно утичу на већ постојеће станице бицикала у систему станица. На основу знања о овим факторима, компанија може да предвиди промет на новопостављеној станици. Последично се може предвидети и популарност станице, као и профит који ће новопостављена станица доносити компанији.

Проблем којим се бави овај рад је управо проналажење корелације између популарности постојећих станица за изнајмљивање бицикала и фактора који утичу на популарност станице. Како се станица налази на одређеној локацији која одређује њену популарност, за факторе чији се утицај посматра су

узети објекти који се налазе у околини станице, а битни су за свакодневне људске активности. У сврху решавања проблема су прикупљени подаци о Capital BikeShare систему[1] за изнајмљивање бицикала и подаци о објектима у околини станица за бицикле. За сваку станицу је одређена оцена и типови објеката који се налазе у околини станице. На основу оцене, близине и бројности појединачних типова објеката је формиран модел система. За формирање модела је изабран алгоритам стабла одлучивања [3]. Алгоритам је примењен над подацима о станицама за бицикле, како би се формирао модел. Показало се да формирани модел у одређеној мери може правилно да предвиди популарност станице, уколико су познати типови објеката из околине станице.

У наставку рада ће детаљније бити објашњени различити аспекти решаваног проблема и самог решења. У поглављу 2 је направљен кратак преглед пронађених радова који се баве истом или сличном проблематиком. За сваки рад је детаљније образложена методологија којом је проблем решаван. Поглавље 3 описује процес формирања циљних сетова података коришћених у даљим анализама. У поглављу је детаљније описан и обухват података о станицама за бицикле CBS система и објектима са подручја на којем систем послује. У наставку поглавља је разрађен сам поступак прераде и спајања сакупљених података како би се добили циљни сетови података. Затим следи поглавље 4 које приказује статистичке анализе спроведене над добијеним подацима. Поглавље 5 описује примену алгоритма стабла одлучивања над циљним сетовима података. Приказани су и дискутовани резултати примене алгоритама. На крају, поглавље 6 садржи дискусију на тему резултата добијених на основу формираних модела, сумаризацију рада, као и правце будућег развоја.

2. Претходна решења

Приликом тражења других решења за проблем одређивања популарности станица за бицикле, преферирана су решења добијена на основу већ постојећих, већих система за изнајмљивање бицикала са подручја Сједињених Америчких Држава. Међу пронађеним решењима издвајамо *„Predicting the Popularity of Bicycle Sharing Stations: An Accessibility-Based Approach Using Linear Regression and Random Forests“* [4] као најсличније решаваном проблему. Наведени рад решава проблем одређивања популарности станица за изнајмљивање бицикала заснован на доступности станица корисницима, који изнајмљене бицикле користе за путовања до посла и назад. Доступност је изражена временом које је потребно кориснику да дође од неке локације која представља радно место или место становања, до неке од станица за изнајмљивање бицикала. Рад је у обзир узео седам параметара: радна места на 30 или 60 минута вожње од било које станице, стамбене објекте на 30 или 60 минута вожње удаљене од било које станице, радна места и стамбене објекте на 10 минута пешачења од било које станице и све станице за изнајмљивање бицикала до којих се са дате станице стиже за мање од 30 минута вожње бицикла. Подаци су изведени помоћу OpenStreetMaps и OpenTripPlanner софтвера. Модел је формиран на основу података о CBS система уз коришћење линеарне регресије и Random Forest алгоритма. Формирани модел се добро показао у предвиђању популарности станица унутар CBS система над којим је формиран. Затим је модел испробан над подацима сакупљеним о друга два система за изнајмљивање бицикала на територији Сједињених Америчких Држава. Формирани модел се није добро показао над преостала два система.

Осим пословних, главни разлози за изнајмљивање бицикала су још и путовања која су приватног карактера, нпр. забава, одлазак у куповину и друго [1]. За разлику од наведеног рада који у обзир узима само стамбене и пословне објекте, овај рад се заснива на претпоставци да почетна и крајња тачка путовања могу бити било која два објекта. Због тога су у обзир узети и други типови објеката који могу утицати на локацију станице за изнајмљивање бицикала као што су, на пример, ресторани или продавнице. Доступност станица за бицикле се дефинише другачије у односу на претходно наведени рад. Уведени су степени доступности, такозвани рангови. Сви објекти који се налазе у одређеном рангу у односу на неку станицу, имају исту доступност станици. Рангови су представљени концентричним прстеновима око станице који имају површину. У претходно наведеном раду је за сваки објекат била позната тачна доступност, док приступ са ранговима може довести до проблема да два различита објекта унутар ранга имају исту доступност, а да је време потребно да се стигне до станице различито за сваки објекат. Овај проблем је нарочито изражен уколико је ширина прстена који представља ранг велика. Проблем може бити ублажен смањењем ширине прстенова, чиме се умањује разлика у времену потребном да се стигне до станице, за објекте на различитим полупречницима у односу на станицу.

3. Формирање сетова података

Ово поглавље детаљно описује поступак настанка циљних сетова података, потребноих за статистичке анализе и даље формирање модела система. Циљни сетови захтевају да за сваку станицу за бицикле буде позната оцена која представља њену популарност, број изнајмљених и враћених бицикала на годишњем нивоу, као и број свих типова објеката у

1. Више информација о Capital BikeShare систему на званичном веб сајту. У наставку рада ће се уместо Capital BikeShare користити само CBS.

свим ранговима у околини станице. Да би један овакав сет био формиран, било је неопходно доћи до података о станицама и објектима у околини станица. Први део овог поглавља се бави обухватом података о станицама за бицикле и промету на станицама, а затим и обухватом података о објектима у околини станица. Други део поглавља се фокусира на трансформације које је било потребно применити да би се од прикупљених података добили цильни сетови података.

3.1. Обухват података

На сајту компаније CBS дати су сетови података који се односе на историју изнајмљивања/враћања бицикала за претходних пет година. За овај пројекат преузети су сетови који се односе на 2015. годину[2]. Сетови су организовани у CSV датотеке по кварталима и садрже податке о трајању путовања, времену изнајмљивања/враћања бицикла (на нивоу сата), називу и терминалу станице (идентификатор станице) са које је изнајмљен, односно на коју је враћен бицикл. Уз претходно наведене податке, за свако изнајмљивање се бележе и идентификатор бицикла и тип чланарине члана који је користио бицикл. Преузета су укупно четири сета података о изнајмљивањима бицикала који су спојени у један сет. Добијени сет ће у наставку бити референиран још и као *сет изнајмљивања*. Подаци о типу чланарине и идентификатору бицикла нису коришћени приликом формирања цильног сета података.
Подаци о станицама – назив, географске координате, број заузетих/слободних места, терминал, време постављања станице, време уклањања станице, да ли је станица привремена, да ли је станица закључана и идентификатор станице, такође су преузети са сајта CBS система[3]. У наставку, овај сет ће бити референциран као *сет станица*.
Подаци о објектима у околини станица преузети су помоћу OverPass API-ja [5]. Задате су координате правоугаоника који обухвата све станице за бицикле компаније CBS и њихову околину. На основу упита са координатама - југ, запад, север, исток: 38.6678492, -77.3598547, 39.250558, -76.7899497 добијени су сви постојећи чворови на мапи у оквиру задатих координата у XML формату. Сваки добијени чвор представља један објекат означен на некој локацији обухваћеној претходно поменутим правоугаоником. Појединачни објекат је описан идентификатором, географским координатама и опционо тагом. Таг представља скуп парова кључ-вредност који детальније описују сам објекат. У наставку, пар кључ-вредност ће се звати атрибут, где је кључ назив атрибута, а вредност вредност атрибута. На овај начин је формиран иницијални *сет објеката* који је даље редукован и прилагођен решавању проблематике овог рада. Подаци о семантици атрибута добављани су помоћу OSM TagFinder [6] претраживача кључева за OpenStreetMaps [7] мапе.

3.2. Припрема података о станицама и изнајмљивањима

За складиштење и обраду података из сета изнајмљивања и сета станица, коришћен је софтвер Microsoft SQL Server 2008 R2 Developer Edition. Коришћењем пропратног софтвера Microsoft SQL Servera, за сваку станицу је срачунат укупан број изнајмљених и враћених бицикала на нивоу године. Укупан број изнајмљених тј. враћених бицикала је добијен проналажењем свих уноса у сету изнајмљивања који одговарају називу сваке појединачне станице. Затим је за сваку станицу на основу пронађених уноса сумиран број изнајмљених и враћених бицикала. Приликом рачунања уочено је да се у сету података о изнајмљивањима називи тринаест станица не поклапају са називима који се налазе у сету станица. Проблем је делимично решен тако што је претходно описани алгоритам поново примењен над проблематичним станицама, али уз коришћење терминала уместо назива станице. За неке од преосталих станица је уочено да је редослед речи у називу обрнут. Након корекције редоследа речи у називу, било је могуће срачунати број изнајмљених и враћених бицикала и за те станице. Проблем је остао нерешен код укупно две станице, које у наставку формирања цильног сета нису узете у обзир.

3.3. Припрема података о објектима

Један од највећих изазова креирања цильног сета је била трансформација иницијално прикупљеног сета објеката у сет објеката који је могуће искористити за даље анализе. Иницијални сет објеката је бројао око деведесет хиљада чворова којима су представљени објекти на територији рада CBS система. Установљено је да корисници система користе бицикле приликом путовања на посао, али и личних ствари попут забаве, куповине, одлазака у ресторане и за друге социјалне потребе [1]. Из претходног следи да би објекти од интереса биле станице других видова транспорта (аутобус, воз), паркинзи, паркови, објекти за куповину хране, разне продавнице, радна места и други објекти потребни за свакодневне послове. Међутим, међу прикупљеним подацима се нашло много објеката чији је значај за кориснике система минималан, нпр. плочници, канте за отпатке, барикаде, трафостанице и други. Због претходно наведеног је било потребно прво редуковати иницијални сет података о објектима.
При прегледању података о објектима, уочено је да постоји велики број чворова описаних само географским координатама. Овакви чворови су одбачени

2. Подаци о изнајмљивањима су преузети са следеће везе
3. Подаци о станицама су преузети са следеће везе

одмах на почетку пречишћавања података. Површине на мапи се описују полигонима чија темена су чворови на мапи без атрибута. Преостали подаци су филтрирани на основу атрибута којима су описани. Анализом атрибута објеката је уочено да постоје атрибути који нису информативни са становишта функције објекта. Пример таквих атрибута су ауторски и датумски атрибути. Неинформативни атрибути су брисани из свих објеката у којима се појављују. Након брисања неинформативног атрибута су брисани сви објекти који су остали без иједног атрибута. Поступак је итеративно примењиван докле год нису остали објекти који садрже само информативне атрибуте. Остатак података о објектима је ручно прегледан и обрисани су они за које је утврђено да немају смисла.
Са становишта формулације проблема, два издвојена објекта који имају исту функцију (нпр. две продавнице) није потребно разликовати. Довољно је знати да су оба објекта продавнице како би се могао срачунати утицај тих објеката на станице за бицикле. Због тога је након филтрирања података о објектима, сваком објекту из редукованог сета објеката додељен тип. Приликом прављења типова, у обзир су узете прикупљене информације о коришћењу изнајмљених бицикала [1]. Други фактор који је утицао на одређивање типова је потреба да се што више искористе прикупљени подаци о објектима и идентификују и неки нови типови објеката који би могли да утичу на популарност станица. Формирани су следећи типови објеката:

- *food_drink* тип који представља објекте у којима је могуће купити храну и пиће,
- *shop* тип који представља све објекте у којима је могуће нешто купити, као што су супермаркети, продавнице воћа и поврћа, трафике итд.,
- *business* тип који представља канцеларијска, занатска и услужна радна места,
- *recreation* тип који представља објекте за рекреацију попут паркова, базена, спортских објеката итд.,
- *entertainment* тип који представља објекте забавног карактера, попут разних ноћних клубова,
- *sight* тип који представља знаменитости попут, статуа, скулптура и сличног,
- *health* тип који представља објекте за здравствену негу људи и животиња,
- *religion* тип који представља објекте верског карактера,
- *tourism* тип који представља објекте занимљиве туристима. На пример, галерије, музеји, камп места, разне атракције, хотеле, мотеле итд.,
- *historic* тип који представља грађевине од историјског значаја,
- *education* тип који представља образовне институције, попут школе, колеџа и универзитета,
- *culture* тип који представља објекте од културног значаја, попут библиотека, биоскопа, књижара и других и
- *better_place* тип који представља гробља.

Од наведених типова објеката, првих пет типова има директну везу са претходно потврђеним узроцима изнајмљивања бицикала. Остали типови су експерименталног карактера.

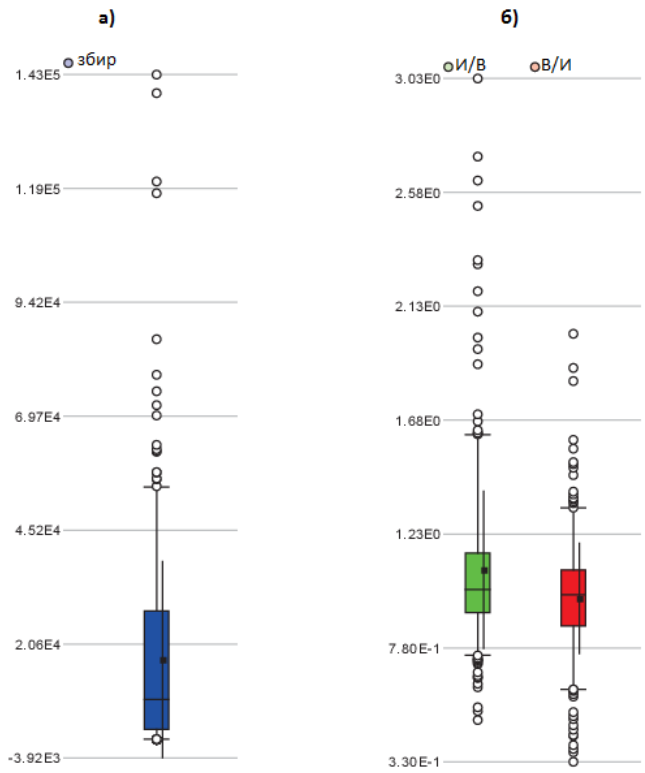
3.4. Спајање података о станицама и објектима

Након што су подаци о објектима прочишћени и сваком објекту додељен тип, израчунато је растојање (у метрима) између сваке станице и сваког објекта. У наставку формирања циљног сета избачени су уноси станица - објекат код којих је растојање веће од 4 км. Ово растојање је изабрано као граница узимајући у обзир да просечна брзина човековог хода износи 5 км/ч и да је максимално време које би човек утрошио пешачећи до станице један сат. Како се дата брзина односи на праволинијски ход ваздушном линијом, апроксимативно је узето 4 км као граница. Овај временски период је експериментално изабран и не постоји доказ да је то његова тачна вредност (за потребе овог пројекта није спроведена анкета међу корисницима CBS која би дала тачну вредност). Поред овог сета, креиран је и сет где су узети у обзир уноси станица - објекат чије је растојање мање од 1.6 км. Растојање је добијено узимајући у обзир, као и за претходни сет, просечну брзину човековог хода, али је у овом сету за максимално време узет временски период од 20 мин. Временски период је, као и за претходни случај, експериментално изабран.
У даљем поступку формирања циљних сетова било је потребно одредити који објекти се налазе у околини којих станица. Околина станица је дефинисана помоћу рангова. Рангови су представљени концентричним прстеновима око станице који имају површину. На тај начин сваки објекат у зависности од растојања у односу на станицу припада одговарајућем рангу. Приликом одређивања рангова коришћена су два приступа: рангови са једнаким површинама и рангови са једнаким полупречницима.
Код формирања сета са једнаким површинама, централни круг и сви прстенови су исте површине. Најпре је израчуната укупна површина околине, подељена је бројем рангова да би се добила површина сваког ранга, а потом су на основу тих површина израчунати одговарајући полупречници. За ову категорију су формирана укупно четири ранга над укупном површином од 4 километра. У поглављу 4 ће се сет са два ранга референцирати као *први сет*. За даље објашњавање је одабран баш овај сет јер се он најбоље показао приликом формирања модела стабла одлучивања из категорије сетова са једнаким

површинама. Детаљи ће бити објашњени у поглављу 5.
Из категорије рангова са једнаким полупречницима су такође направљена четири сета. Сет са најмањим бројем рангова има укупно три ранга. Први ранг за 0-5, други за 5-10 и трећи за 10-15 минута праволинијске шетње од станице. Уз овај су направљена још три сета где сваки има по један ранг више у односу на претходни. У поглављу 4 ће се сет са четири ранга референцирати као *други сет*. За даље објашњавање је одабран баш овај сет јер се он најбоље показао приликом формирања модела стабла одлучивања из категорије сетова са једнаким полупречницима. Детаљи ће бити објашњени у поглављу 5.

3.5. Формирање оцене популарности станице

На основу претходно изведених података о станицама и изнајмљивањима, изведене су по три оцене за сваку станицу. Прва оцена је оцена промета и добијена је сабирањем укупног броја изнајмљених и враћених бицикала за сваку станицу на нивоу годину дана. Мотивација за формирање овакве оцене потиче из претпоставке да су станице са већим прометом значајније за функционисање компаније која поседује систем за изнајмљивање бицикала. Друга оцена представља однос броја изнајмљених и враћених бицикала за сваку станицу за период од годину дана. Трећа оцена је редундантна и представља реципрочну вредност друге оцене, односно однос броја враћених и изнајмљених бицикала за сваку станицу за период од годину дана. Мотивација за другу и трећу оцену проистиче из чињенице да је за добру снабдевеност станица за бицикле потребно вршити ребалансирање броја бицикала. Ребалансирање је превожење бицикала са оптерећених станица на мање оптерећене станице које обавља компанија, како би бицикли увек били доступни на станицама на којима су потребни. Стога су са аспекта компаније власника система за изнајмљивање бицикала исплативије станице на које није потребно утрошити додатни новац због ребаланса. Подаци о промету на станицама CBS система су представљени помоћу нормалне расподеле. Подаци су приказани на слици 1 помоћу дијаграма квартила. На основу дијаграма је формирано укупно пет оцена. Сваком оценом је представљен интервал укупног броја изнајмљених и враћених бицикала на станици. Оцена нула је третирана посебно и додељена је станицама које уопште нису имале промет током 2015. године. Сматра се да су то станице које су вероватно биле нефункционалне током године и због тога не могу бити равноправне са станицама које су макар неко време биле функционалне. У табели 1 су приказане горње и доње границе интервала које одговарају свакој од оцена. Подаци о односу изнајмљених и враћених бицикала, као и подаци о промету враћених и изнајмљених



Слика 1. Квартилни дијаграми за: а) збир изнајмљених и враћених бицикала, б) однос изнајмљених и враћених бицикала

Табела 1. Границе за оцену промета

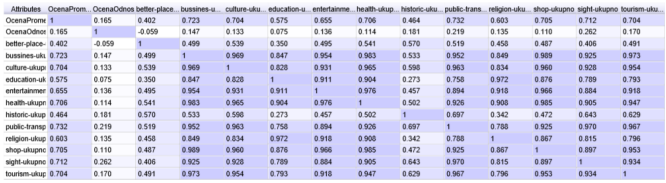
оцена	доња граница	горња граница
0	0	0
1	0+	376
2	376+	2566
3	2566+	28646
4	28646+	54607
5	54607+	∞

бицикала су такође представљени нормалном расподелом. Као и подаци о промету, и ови подаци су представљени дијаграмом квартила приказаног на слици 1. На дијаграму је са *В/И* представљен однос враћених и изнајмљених, а са *И/В* однос изнајмљених и враћених бицикала по станици. Примећено је да однос враћених и изнајмљених има финију расподелу у односу на однос изнајмљених и враћених. Због овога се из даљег разматрања избацује оцена која представља однос изнајмљених и враћених бицикала по станици. Као и у случају оцене промета и овде ће свака оцена представљати неки интервал. Интервали који одговарају оценама, као и границе интервала, дате су у табели 2. У наставку ће се под појмом *оцена односа* подразумевати оцена која описује однос враћених и изнајмљених бицикала.
Након формирања оцене популарности сваке поје-

Табела 2. Границе за оцену односа

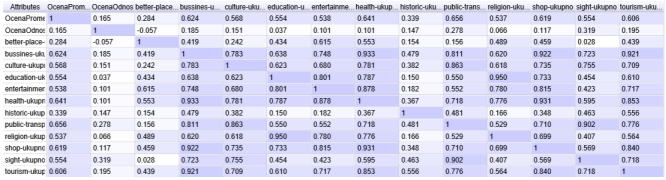
оцена	доња г 1	горња г 1	доња г 2	горња г 2
0	0	0	-	-
1	0+	~0.866	~1.332+	∞
2	~0.866+	~0.616	~1.087+	~1.332
3	~0.616+	~1.087	-	-

диначне станице извршено је пребројавање свих типова објеката у свим ранговима у околини станице. На тај начин добијени су сетови података где је за сваку станицу позната оцена која представља њену популарност, број изнајмљених и враћених бицикала на годишњем нивоу, као и број свих типова објеката у свим ранговима у околини станице.



Attributes	OcenaProm...	OcenaOdnos	better-place...	bussines-u...	culture-uku...	education-u...	entertainme...	health-ukup...	historic-uku...	public-trans...	religion-uku...	shop-ukupno	sight-ukupno	tourism-uku...
OcenaProme	1	0.165	0.402	0.723	0.704	0.575	0.655	0.706	0.464	0.732	0.603	0.705	0.712	0.704
OcenaOdnos	0.165	1	-0.059	0.147	0.133	0.075	0.136	0.114	0.181	0.219	0.135	0.110	0.262	0.170
better-place-	0.402	-0.059	1	0.499	0.539	0.350	0.495	0.541	0.570	0.519	0.458	0.487	0.406	0.491
bussines-uku	0.723	0.147	0.499	1	0.969	0.847	0.954	0.983	0.533	0.952	0.849	0.989	0.925	0.973
culture-ukup	0.704	0.133	0.539	0.969	1	0.828	0.931	0.965	0.598	0.963	0.834	0.960	0.928	0.954
education-uk	0.575	0.075	0.350	0.847	0.828	1	0.911	0.904	0.273	0.758	0.972	0.876	0.789	0.793
entertainmer	0.655	0.136	0.495	0.954	0.931	0.911	1	0.976	0.457	0.894	0.918	0.966	0.884	0.918
health-ukup	0.706	0.114	0.541	0.983	0.965	0.904	0.976	1	0.502	0.926	0.908	0.985	0.905	0.947
historic-uku	0.464	0.181	0.570	0.533	0.598	0.273	0.457	0.502	1	0.697	0.342	0.472	0.643	0.629
public-transp	0.732	0.219	0.519	0.952	0.963	0.758	0.894	0.926	0.697	1	0.788	0.925	0.970	0.967
religion-ukup	0.603	0.135	0.458	0.849	0.834	0.972	0.918	0.908	0.342	0.788	1	0.867	0.815	0.796
shop-ukupno	0.705	0.110	0.487	0.989	0.960	0.876	0.966	0.985	0.472	0.925	0.867	1	0.897	0.953
sight-ukupno	0.712	0.262	0.406	0.925	0.928	0.789	0.884	0.905	0.643	0.970	0.815	0.897	1	0.934
tourism-ukup	0.704	0.170	0.491	0.973	0.954	0.793	0.918	0.947	0.629	0.967	0.796	0.953	0.934	1

Слика 2. Матрица корелације за први сет без н.в.



Attributes	OcenaProm...	OcenaOdnos	better-place...	bussines-u...	culture-uku...	education-u...	entertainme...	health-ukup...	historic-uku...	public-trans...	religion-uku...	shop-ukupno	sight-ukupno	tourism-uku...
OcenaProme	1	0.165	0.284	0.624	0.568	0.554	0.538	0.641	0.339	0.656	0.537	0.619	0.554	0.606
OcenaOdnos	0.165	1	-0.057	0.185	0.151	0.037	0.101	0.101	0.147	0.278	0.066	0.117	0.319	0.195
better-place-	0.284	-0.057	1	0.419	0.242	0.434	0.615	0.553	0.154	0.156	0.489	0.459	0.028	0.439
bussines-uku	0.624	0.185	0.419	1	0.783	0.638	0.748	0.933	0.479	0.811	0.620	0.922	0.723	0.921
culture-ukup	0.568	0.151	0.242	0.783	1	0.623	0.680	0.781	0.382	0.863	0.618	0.735	0.755	0.709
education-uk	0.554	0.037	0.434	0.638	0.623	1	0.801	0.787	0.150	0.550	0.950	0.733	0.454	0.610
entertainmer	0.538	0.101	0.615	0.748	0.680	0.801	1	0.878	0.182	0.552	0.780	0.815	0.423	0.717
health-ukup	0.641	0.101	0.553	0.933	0.781	0.787	0.878	1	0.367	0.718	0.776	0.931	0.595	0.853
historic-ukup	0.339	0.147	0.154	0.479	0.382	0.150	0.182	0.367	1	0.481	0.166	0.348	0.463	0.556
public-transp	0.656	0.278	0.156	0.811	0.863	0.550	0.552	0.718	0.481	1	0.529	0.710	0.902	0.776
religion-ukup	0.537	0.066	0.489	0.620	0.618	0.950	0.780	0.776	0.166	0.529	1	0.699	0.407	0.564
shop-ukupno	0.619	0.117	0.459	0.922	0.735	0.733	0.815	0.931	0.348	0.710	0.699	1	0.569	0.840
sight-ukupno	0.554	0.319	0.028	0.723	0.755	0.454	0.423	0.595	0.463	0.902	0.407	0.569	1	0.718
tourism-ukup	0.606	0.195	0.439	0.921	0.709	0.610	0.717	0.853	0.556	0.776	0.564	0.840	0.718	1

Слика 3. Матрица корелације за други сет без н.в

4. Статистика

Ово поглавље се бави статистичким анализама које су спроведене над претходно направљеним сетом података. Циљ ових анализа је увидети да ли постоји повезаност између оцене популарности станице и броја објеката одређеног типа у околини те станице. За спровођење статистичких анализа коришћен је оператор Correlation Matrix у оквиру Rapid Miner алата. На овај начин добијене су корелације између свих атрибута и оне нам говоре да ли су и колико јако парови атрибута у вези. Графичка репрезентација корелација је дата у виду матрице. Корелација је представљена бројем између -1 и +1. Позитивна вредност корелације подразумева и позитивну везу, тачније вредности атрибута се пропорционално повећавају. Негативна вредност корелације подразумева негативну или инверзну везу. У том случају вредности атрибута се обрнуто пропорционално повећавају.

За потребе пројекта статистичка анализа спроведена је над првим и другим сетом, са и без станица на којима је промет био нула (13 станица). Уочено је да су боље вредности корелација за сетове без станица на којима је промет нула. Разлике у вредностима корелација сетова са и без станица на којима је промет нула (даље у тексту *први сет без н.в.* и *други сет без н.в.*) износе око 10-15%. Једино су за корелацију између оцене и броја објеката типа *historic* вредности боље у сетовима са станицама на којима је промет био нула. За први сет без н.в. матрица корелације дата је на слици 2, док је за други сет без н.в. матрица приказана на слици 3. За први сет без н.в. уочено је да је најмања вредност корелације између оцене промета и броја објеката типа *better_place*. Исто је уочено и за други сет без н.в. Највећа вредност корелације у првом сету без н.в. уочена је између оцене промета и броја објеката типа *public_transport*. Исто је уочено и за други сет без н.в. То наводи на закључак да на промет неке станице има утицај број објеката типа *public_transport*, тачније што је њихов број већи, већи је и промет на станици. Поред броја објеката овог типа, знатан утицај има и број објеката типа *business*, *health* и *shop*.

За оцену односа је уочено да је вредност корелације близу нуле са бројем објеката типа *better_place* у оба сета без н.в., односно корелација не постоји. Највећа вредност корелације у оба сета без н.в. за оцену односа уочена је са бројем објеката типа *sight*. Као и за оцену промета, велики је утицај и броја објеката типа *public_transport* на оцену односа.

5. Модели система

Модели су формирани са циљем тестирања следећих хипотеза:

- *"The top bikeshare trip purposes overall were for personal/non-work trips."*[4] Из овога следи да објекти типа *shop*, *entertainment*, *food_drink*, *recreation* и *culture* позитивно утичу на промет на станици,
- *"A large share of members used bikeshare for their trip to work."*[4] Из овог следи да би требало да објекти типа *buisiness* такође позитивно утичу на промет на станици и
- *"Capital Bikeshare also served as a feeder service to reach transit stops."*[5] На основу ове тврдње следи да и објекти типа *public_transport* имају позитиван утицај на промет на станици.

Уколико се модели покажу добро за потврђивање горе наведених хипотеза, могуће је са већом поузданошћу утврдити утицај других типова објеката на популарност станица тј. промет на њима.

Модел система је направљен коришћењем алгоритма стабла одлучивања у имплементацији RapidMiner алата. Алгоритам је примењен над свим формираним сетовима података из категорије сетова са истим површинама и сетова са истим полупречницима. Након формирања стабла одлучивања за сваки

4. [1], страна 3
5. [1], страна 4

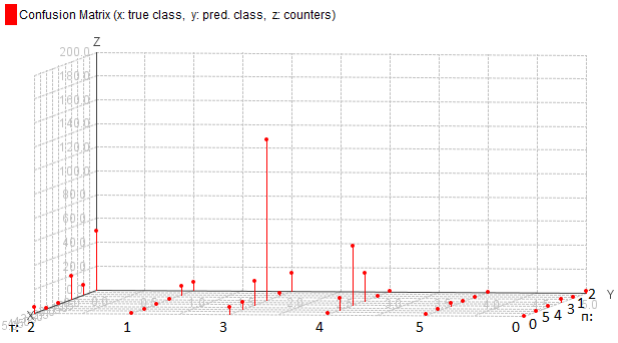
сет података, одабран је по један репрезентативни сет података из категорије сетова са једнаким површинама и једнаким полупречницима. Као мера за евалуацију решења је узета тачност решења, па је одабрано по једно најтачније решење из обе категорије сетова.

5.1. Модел система за оцену промета - сетови једнаких површина

Из категорије сетова са једнаким површинама је изабрано стабло одлучивања формирано над сетом података са два ранга, у радијусу 4 км у околини станице. Ово стабло одлучивања је дало већу тачност у односу на сва остала формирана стабла одлучивања (65.41%). У табели 3 је приказан вектор перформанси за поменуто стабло. Из табеле је могуће видети да се оцене 0 и 5 предвиђају са најмањом прецизношћу (0% и 25% респективно), што је и очекивано пошто обучавајући скуп има мали број уноса о оваквим станицама. Оцена која се предвиђа са највећом прецизношћу је оцена три (75%). На слици 4 је дат графички приказ података из табеле 3 помоћу матрице конфузије [8].

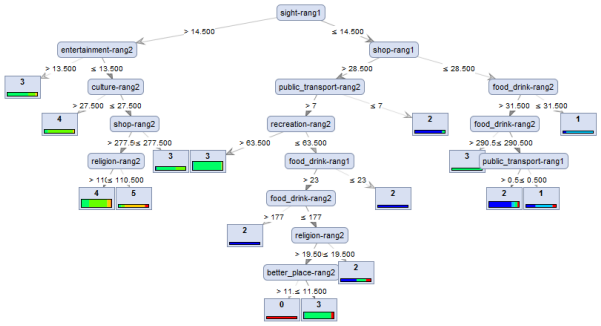
Табела 3. Приказ вектора перформанси (**п** - претпоставка, **и** - истина, **прец.** - прецизоност)

	и 2	и 1	и 3	и 4	и 5	и 0	прец.
п 2	49	8	19	0	0	5	60%
п 1	7	7	0	0	0	0	50%
п 3	15	2	135	20	6	6	73%
п 4	0	0	23	50	10	1	60%
п 5	0	0	0	2	1	1	25%
п 0	1	0	2	0	0	0	0%
одзив	68%	41%	75%	69%	6%	0%	



Слика 4. Матрица конфузије за репрезентативни сет из категорије једнаких површина

На слици 5 је приказано стабло одлучивања за претходно описано стабло.



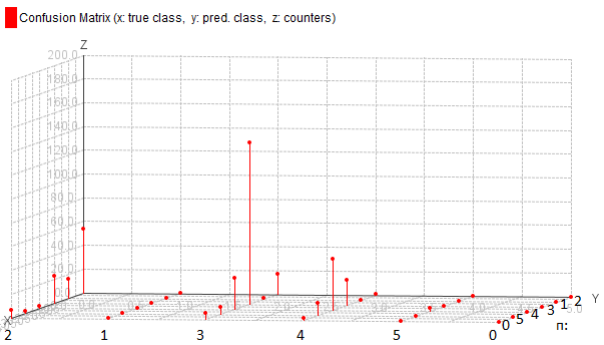
Слика 5. Стабло одлучивања за репрезентативни сет из категорије једнаких површина

5.2. Модел система за оцену промета - сетови једнаких полупречника

Из категорије сетова са једнаким полупречницима је изабрано стабло одлучивања формирано над сетом података са четири ранга, где је сваки ранг додатних 5 минута праволинијског ходања у односу на претходни. Тачност овог стабла одлучивања је 62.97%. У табели 4 је приказан вектор перформанси за поменуто стабло. Из табеле је могуће видети да на основу овог модела није могуће предвидети оцене 0, 1 и 5. Оцена са најбољом прецизношћу је оцена 3 и могуће ју је предвидети са прецизношћу од 71.20%. На слици 6 је дат графички приказ података из табеле 4 помоћу матрице конфузије.

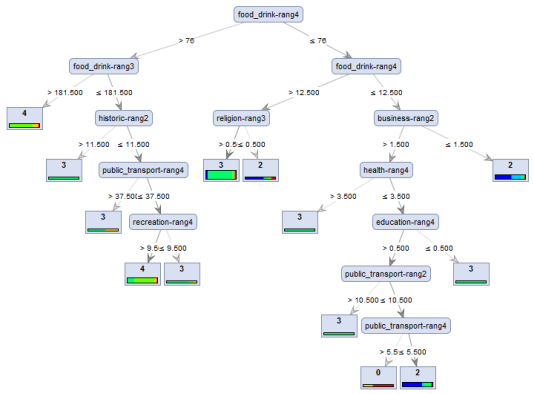
Табела 4. Приказ вектора перформанси (**п** - претпоставка, **и** - истина, **прец.** - прецизност)

	и 2	и 1	и 3	и 4	и 5	и 0	пред.
п 2	47	15	21	1	1	3	53%
п 1	0	0	0	0	0	0	0%
п 3	24	2	129	27	5	8	66%
п 4	1	0	29	44	11	2	50%
п 5	0	0	0	0	0	0	0%
п 0	0	0	0	0	0	0	0%
одзив	65%	0%	72%	61%	0%	0%	



Слика 6. Матрица конфузије за репрезентативни сет из категорије једнаких полупречника

На слици 7 је приказано стабло одлучивања за претходно описано стабло.



Слика 7. Стабло одлучивања за репрезентативни сет из категорије једнаких полупречника

5.3. Верификација

За верификацију свих добијених модела је коришћена унакрсна верификација. Верификација је извршена коришћењем оператора X-Validation [9] из RapidMiner алата. За потребе унакрсне верификације, скуп података је издељен на 10 подсетова који су даље коришћени у процесу верификације.

6. Закључак

У овом поглављу су кроз дискусију образложени резултати добијени формирањем различитих модела система. У наставку се налази сумаризација рада, као и правци будућег развоја.

6.1. Дискусија

Анализирајући добијене резултате алгоритма над свим подацима, уочено је да су сви модели приближно исте тачности, која се креће од 54% до 65%. Генерално, оцене промета 2, 3 и 4 се предвиђају са највећом прецизношћу од 50% до 75%. Слична ситуација је и са одзивом одговарајућих оцена, што се може и видети у табелама 3 и 4. Такође, оно што је интересантно јесте да околина станица са оценом промета 5 има највећи број објеката *business*, *shop* и *food_drink* типа. У поглављу 5 су приказани само репрезентативни случајеви стабала. Осим њих, формиран је и одређени број стабала која нису детаљније разматрана у овом раду, јер њихова тачност није била максимална. Међутим, код свих стабала одлучивања су уочене одређене сличности. На пример, уочено је да је на основу сваког стабла одлучивања могуће предвидети оцену 3 са прецизношћу која је минимално 50%, а врло често и преко 70%. Осим претходног, примећено је да висока фреквенција појављивања неких типова објеката у околинама станица повећава промет на станици, тј. има позитиван утицај на популарност станице. Са друге стране, уочено је да постоје и типови објеката који негативно утичу на популарност станице.
Типичан пример неповољног утицаја показују објекти типа *better_place* чија повећана фреквенција појављивања у околини је повезана са смањеним прометом на станици. *Public_transport* је тип објеката за који се не може са сигурношћу рећи да ли позитивно или негативно утиче на промет бицикала на станицама. Типови објеката чији велик број појављивања у близини станице најчешће поспешује промет на њој су: *food_drink*, *shop*, *recreation*, *tourism*, *business*, *entertainment*. Узимајући у обзир резултате и саму њихову тачност можемо рећи да су прве две хипотезе потврђене, док трећа хипотеза није потврђена на основу модела, док корелација говори да постоји веза између бољег промета на станицама и јавног превоза.

6.2. Сумаризујте рад

Проблем којим се бави овај рад је предвиђање популарности станице за бицикле, где је популарност представљена оценом промета на станици. Овај рад настоји да искористи велику количину неструктуираних података о објектима, како би се након одређеног структуирања ти подаци искористили за стицање знања о популарности станица за бицикле. Први корак ка решењу проблема је био формирање сетова података над којим су касније спроведене даље анализе. Добијени сетови су прво обрађени статистички, а онда и уз коришћење алгоритма стабла одлучивања. Потврђен је већи део постављених хипотеза о утицају одређених типова објеката на станице за бицикле. На основу овога је оправдано веровати да је направљене моделе могуће искористити да би се из њих добило знање о утицају других типова објеката, за које се тренутно не зна тачан утицај на популарност станице.

6.3. Предложите правце будућег рада

Како су се добијени резултати показали као задовољавајући, наставак рада на овом пројекту би требало базирати на проналажењу везе између односа броја враћених и изнајмљених бицикала на станици са близином објеката одређеног типа. Оцена односа броја враћених и изнајмљених бицикала на свакој станици је у току рада на овом пројекту израчуната, али није коришћена. Један од разлога за потребу проналаска корелација између објеката одређеног типа са овом оценом лежи у чињеници да промет не говори да ли се на некој станици више изнајмљују, односно враћају бицикли. Самим тим ни компанија само на основу промета не може вршити ре-

баланс, тачније ефикасније распоређивање бицикала по станицама. Боље распоређивање бицикала по станицама утиче како на финансијски аспект, тако и на задовољство корисника.
Такође, постоји потреба и за унапређењем поступка разврставања објеката по категоријама. Мана код приступа који је примењен у овом раду је та што је разврставање по категоријама у већој мери субјективно. Један од даљих праваца рада је и покушати разврставање учинити што је више могуће објективним.

Библиографија

[1] ``2014 capital bikeshare member survey report,'' Capital Bikeshare System, Tech. Rep., 2015. [Online]. Available: http://www.capitalbikeshare.com/assets/pdf/cabi-2014surveyreport.pdf

[2] B. Alberts, J. Palumbo, and E. Pierce, ``Vehicle 4 change: Health implications of the capital bikeshare program,'' The George Washington University, Tech. Rep., 2012. [Online]. Available: http://www.capitalbikeshare.com/assets/pdf/v4c_capstone_report_final.pdf

[3] [Online]. Available: http://docs.rapidminer.com/studio/operators/modeling/predictive/trees/parallel_decision_tree.html

[4] M. W. Conway, ``Predicting the popularity of bicycle sharing stations: An accessibility-based approach using linear regression and random forests,'' 2014. [Online]. Available: http://www.indicatrix.org/publications/2014/Conway-Bikeshare-Accessibility.pdf

[5] [Online]. Available: http://wiki.openstreetmap.org/wiki/Overpass_API

[6] [Online]. Available: http://tagfinder.herokuapp.com/

[7] [Online]. Available: http://www.openstreetmap.org/#map=5/51.500/-0.100

[8] [Online]. Available: https://en.wikipedia.org/wiki/Confusion_matrix

[9] [Online]. Available: http://docs.rapidminer.com/studio/operators/validation/x_validation.html