Приједлог пројекта из СИАП-а

Овај документ садржи кратак опис онога што је тема пројекта и дефиниција, мотивација за одабрану тему. Након мотивације слиједи преглед владајућих ставова и схватања у литератури, затим скуп података који је укратко описан. Такође, наведен је и софтвер који ће бити коришћен, као и метод евалуације. На самом крају документа налази се план рада на пројекту.

Тема пројекта је “*Speed dating*” – “Брзо упознавање“. На располагању су подаци прикупљени помоћу анкета подјељених учесницима након учествовања у брзом упознавању. Период прикупљања података је од 2002-2004. године, анкетирани су студенти *Columbia Business School* у Њујорку . На располагању је 8378 анкета, на основу којих ће да се врше анализе.

**Дефиниција пројекта**

Циљ пројекта је да основу анализе горе наведених података дођемо да закључка шта је најмање пожељно код партнера, и у чему се разликују мушка у односу на женска становишта о пожељним особинама.

**Мотивација**

Узимајући у обзир брз ритам живота данашњице, проналажење и упознавање нових сродних душа постаје све захтјевније и теже. С тим разлогом се све више људи окреће такозваном брзом упознавању особа које би могле да испуне ту улогу у кратком временском периоду. У овом раду се скреће пажња на особине које би могле да привуку пажњу потенцијалног партнера, како код мушкараца тако и код жена.

**Преглед владајућих ставова и схватања у литератури**

[1] Samuel King, Heinrich Lohner, Rick Radewangen, Ye Wu & Jorn Jansen Schoonhoven (2016) *What matters the most in the first four minutes* [*https://www.kaggle.com/zurfer/what-matters-most*](https://www.kaggle.com/zurfer/what-matters-most)

**Тема рада:** Анализирана је тема брзог упознавања. Циљ је да се обезбједи основно разумјевање људског понашања у оваквим ситуацијама, укључујући атрибуте који су најрелевантнији за успјешно упознавање.

**Подаци:** Скуп података који je истраживан у пројекту назива се "*Speed Dating Experiment*", садржи 8378 анкета, 280 обиљежја која су нумеричка. Генерисан је из серије експерименталних догађаја са брзим упознавањем из 2002. до 2004. године и обухвата податке везане за демографске податке, навике током упознавања, информације о начину живота, упитник за евалуацију атрибута узет када се учесници пријављују и оцјене сваког учесника за друге учеснике током 4 минута интеракције. Анкетирани су студенти *Columbia Business School* у Њујорку .

**Коришћени алгоритми:** Избачене су анкете и обиљежја везане од 6-9 састанка пошто је ту коришћена другачија методологија оцјењивања атрибута у односу на остале састанке. Избацивањем анкета остало их је 6816 и 201 обиљежје. Затим је извршена норамализација атрибута у дистрибуцију од 0 до 100. Експлоративну анализу су вршили помоћу регресије. Линеарна регресија је показала да су најзначајнији атрибути: атрактивност, искреност, интелигенција, духовитост, амбиција и заједничка интересовања. У сљедећем кораку су са нумеричким линеарним регресионим моделом посматрали сваки од горе наведених атрибута независно као и у комбинацији једног од њих са другим атрибутом.

**Остварени резултати:** Створили су неколико линеарних модела регресије како би покушали објаснити факторе атрибута који утичу на број позитивних одговора које су кандидати добијали након одласка на састанак. *K-fold* унакрсна валидација

(детаљније о овој валидацији: [*http://www.ritchieng.com/machine-learning-cross-validation/*](http://www.ritchieng.com/machine-learning-cross-validation/) *)* је кориштена да се одреди најбољи модел линеарне регресије и то је модел који комбинује атрактивност, духовитост и заједничке интересе, тачност је била 66%. Још је било очекивано да је атрактивност најважнија особина за успјешно упознавање, али се испоставило да је особе углавном траже исте особине које и саме посједују.

[2] James Hwang, Lucas Cadalzo(2016) *The ugly truth of people decisions in speed dating* [*https://www.kaggle.com/jph84562/the-ugly-truth-of-people-decisions-in-speed-dating*](https://www.kaggle.com/jph84562/the-ugly-truth-of-people-decisions-in-speed-dating)

**Тема рада:** Истраживана је пројекат назван “Speed dating experiment”. Од појединаца је тражено да се изјасне да ли би жељели други састанак с људима које су упознали, и да их оцјене поново након одржаних других састанака. А други аспект који је истраживан је шта учесници мисле да је важно супротном полу.

**Подаци:** У питању исти скуп података који је кориштен и у горе наведеном експерименту. Сам скуп података садржи 8378 анкета.

**Коришћени алгоритми:** Прво је извршена нормализација свих атрибута у дистрибуцију од 0 до 100. Затим су извршили визуализацију података по х и у оси. На х-оси је израчунато колико позитивних одговора особа може добити од особе супротног пола. На у-оси, израчунали су средњу вриједност оцјена за 6 атрибута које је мушкарац примио од 10 жена. Затим је израчунат коефицијент биваријантне корелације за сваки атрибут да би се знало колико је јака корелација између позитивне стопе одговора партнера и оцјене атрибута. Подаци о корелацији су даље подијељени на мушке и женске одлуке, а резултати корелације су поново израчунати, визуализовани помоћу графова и онда су их аутори протумачили.

**Остварени резултати:** За аспект шта је важно супротном полу, испоставило се да мушкарци у женама траже изглед, а не амбицију, уколико не дијеле заједничке интересе. Док је женама јако битна интелигенција. А за аспект о другом састанку, жене су добро погодиле шта је мушкацима најбитније: изглед. И мушкарци су погодили приближно рекавши да је женама најважнији изглед, а у ствари то је интелигенција, па затим изглед и духовитост.

**Скуп података**

За овај пројекат разматраће се скуп податак преузет са следећег линка : [*https://www.kaggle.com/annavictoria/speed-dating-experiment*](https://www.kaggle.com/annavictoria/speed-dating-experiment). Саставили су га професори *Ray Fisman* и *Sheena Iyengar* из *Columbia Business School* у периоду од 2002-2004. године, која је првобитно користила свој рад "*Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment*". Скуп података се састоји од 8378 анкета. Садржи укупно 280 обиљежја, а тај број је толики пошто се одређена обиљежја понављају кроз четири фазе и то из аспекта и мушкараца и жена (та одређена обиљежја дата су у даљем тексту код описа друге фазе). Сва обиљежја су нумеричка, и постоји велики број недостајаћих вриједности код анкета за особе које нису биле најуспјешније.

Подаци су дати у .csv формату као и у Excel. У пројекту ће се највјероватније користити Excel формат. Шест особина које имају велики утицај на резултат су: атрактивност, искреност, интелигенција, духовитост, амбициозност, и заједнички интереси, због тога што је у свакој фази од испитаника било тражено да оцјене партнера по питању тих 6 обиљежја.

Прикупљање података је извршено у четири фазе. У почетну фазу, фазу један, спадају следећи подаци :

* пол(0 – female, 1 - male), број учесника у једном кругу, старост учесника, област студирања , раса учесника (“Black/African American=1, European/Caucasian-American=2, Latino/Hispanic American=3, Asian/Pacific Islander/Asian-American=4, Native American=5, Other=6“), колико је раса битна учеснику (1 - 10), колико је религија битна учеснику (1 - 10), због чега учествују (“Seemed like a fun night out=1, To meet new people=2, To get a date=3, Looking for a serious relationship=4, To say I did it=5, Other=6”), учесталост одлажења на састанке (“Several times a week=1, Twice a week=2, Once a week=3, Twice a month=4, Once a month=5, Several times a year=6, Almost never=7”), колико често излазе (“Several times a week=1, Twice a week=2, Once a week=3, Twice a month=4, Once a month=5, Several times a year=6, Almost never=7”), каријера, заинтересованост за наведене активности (“sports: Playing, sports/ athletics, tvsports: Watching sports, excersice: Body building/exercising, dining: Dining out, museums: Museums/galleries, art: Art, hiking: Hiking/camping, gaming: Gaming, clubbing: Dancing/clubbing, reading: Reading, tv: Watching TV, theater: Theater, movies: Movies, concerts: Going to concerts, music: Music, shopping: Shopping, yoga: Yoga/meditation”), колико учесник очекује да буде срећан са људима које упозна (1 - 10), очекивани број особа којим учесник очекује да се свиди (1 – 20).

У фази два и у свакој другој фази извршено је испитивање следећих особина :

* ниво свих шест особина које учесник тражи код партнера (збир свих шест = 100), оцјена свих шест особина које учесник тражи код партнера (1 - 10), која од шест особина највише утиче на одлуку учесника (збир свих = 100), шта учесник мисли да траже у супротном полу (1 – 10), оцјена сопствених особина (1 – 10).

Фази три припадају следећи подаци :

* број састанка с одговарајућим партнером, да ли су особе на састанку нашле заједнички језик, да ли су исте расе, старост партнера, раса партнера , одлука партнера, општи утисак (1 – 10).

Фаза четири : Након завршених састанака (након 3-4 седмице), учесницима је послато да попуне још једну анкету о особама са којима се најбоље подударају. Следећи подаци су везани за то :

* колико је учесника датог учесника позвало на састанак, колико је учесник позвао на састанак, да ли је учесник био на барем на једном састанку, са колико је партнера био на састанку, колико пута.

Опис скупа података се може наћи на истом мјесту гдје и сам скуп података.

**Методологија**

Прво ће бити извршен *data wrangling* кроз учитавање података из Excel табеле, разумјевање података, затим чишћење података - обиљежја која имају највише недостајућих вриједности, ће бити избачена из разматрања. Затим, у неколико анализа учесници су добили различита упутства. На примјер, у већини догађаја, учесницима је дат 100 бодова за дистрибуцију до 6 особина како би показали њихов приоритет. Међутим, за неке фазе су тражиле да рангирају значај атрибута на скали 1-10. У оба случаја, подаци ће бити скларини да би се формирала дистрибуција од 0-10.

Биће извршена експлоративна анализа за одређивање обиљежја која су најзначајнија, уз визуализацију добијених резултата. То приликом биће одређен коефицијент корелације између промјенљивих. Редукција димензионалност и обиљежја биће вршена ПЦА алгоритмом.

Што се тиче алгоритма за анализу података, користићемо машинске алгоритме и то у првом реду логистичку регресију, затим стабло одлучивања (примјенити тако што ћемо вјероватно посматрати 6 основних атрибута у 4 фазе) и XGBoost алгортима .

**Софтвер**

*У RapidMiner* софтверу за обраду, анализу и складиштење података, биће извршена и експлоративна анализа скупа података, као и *data wrangling*. Апликација ће бити израђена у *Python* програмском језику, и помоћу овог језика биће имплементирани алгоритми машинског учења и додата библиотека за XGBoost.

**Метод евалуацијe**

Обучавајући скуп података биће формиран што ће бити случајно одабрано 70% очишћених података, а валидациони скуп чиниће остатак. Тачност модела биће евалуирана валидационим скупом. Биће коришћена прецизност и F1 мјера као метрика.

**План**

Плана рада на овом пројекту обухвата следеће тачке:

• прикупљање података,

• експлоративна анализа,

• трансформација података,

• креирање модела,

• провјера модела,

• визуелизација добијених резултата помоћу ***bar/line* и *radar/spider* дијаграма**.

**Тим**

Тим чине : Милица Говедарица (E2 49/2017), Десанка Тодић (E2 56/2017) .