Биостатистика – Пројектни задатак

Име и презиме: Немања Орељ

Број индекса: 2022/0127

У оквиру Теамс фолдера где се налази овај пројектни задатак пронаћи ћете матрицу *Arcadis.xlsx* над којом је потребно спровести различите статистичке анализе у следећим задацима: 1, 2, 3, 4, и 5. Такође, у задацима 6, 7 и 8 потребно је искористити матрице које се у њима наводе, а које су доступне у истом фолдеру.

Матрица *Arcadis.xlsx* садржи категоријске варијабле:

- *Continent* за коју су дате вредности 1, 2 и 3 које означавају континенте Европу, Северну Америку и Азију, респективно.

НАПОМЕНА: у свим задацима где је потребно обезбедити квази-рандомизацију, потребно је употребити функцију set.seed(123).

Задатак 1 Употребом регуларизационе методе *RIDGE* предвидите туристички индекс *Tourism* на основу варијавли *education, crime, health, affordability, Energy* и *Greenspace*. Затим, одговорите на постављена питања. Претпоставите да су предиктори стандардизовани (стандардизацију обично ради *glmnet* аутоматски, осим ако није другачије задано).

Наведите две главне разлике између *LASSO* (L1) и *Ridge* (L2) регресије.

Одговор:

LASSO (L1) може да постави поједине коефицијенте тачно на нулу, па ради селекцију променљивих,док Ridge (L2) коефицијенте само смањује (shrinkage) и не нулира их, па све променљиве остају у моделу.LASSO користи апсолутне вредности коефицијената L1 као казну, а Ridge L2 казну - квадрате коефицијената**.**

Зашто је потребно стандардизовати променљиве пре *Ridge*/*LASSO* анализа?

Одговор:

Зато што се регуларизација (казнени члан) примењује на величину коефицијената, а величина коефицијената зависи од скале предиктора. Без стандардизације, променљиве са већом скалом/варијансом биле би неправедно више кажњене. Стандардизација доводи све предикторе на исту скалу и омогућава фер поређење њиховог утицаја..

Којим линијама кода се може учитати *Arcadis.xlsx* и направити матрицу *X* са наведеним променљивима и вектор одзива *y* = Tourism?

```
library(readxl)

df <- read_excel("Arcadis.xlsx")

X <- as.matrix(df[, c("education","crime","health","affordability","Energy","Greenspace")])

y <- df$Tourism
```

Примените *Ridge* са *cross-validation* техником, при чему треба водити рачуна да је неопходно обезбедити квази-рандомизацију применом функције set.seed(123). Тада ће параметри модела бити alpha=0,standardize=TRUE, nfolds=10 (type.measure = "mse" је подразумевано),док је минимално **λ**,заокружено на четири децимале, једнако 0.2422**.**

Када је испуњен услов **λ=λ_min** број променљивих које се задржавају у моделу је 6 , а у питању су променљиве: education, crime, health, affordability, Energy, Greenspace.

Одредите вредност *lambda.1se* за *Ridge* модел: 102.4271.

Користећи *Ridge* модел са *lambda.min*, израчунајте предикције за све градове и пронађите 5 градова са највећом апсолутном грешком |*Tourism - pred_Tourism*|. Попуните одговарајућим вредностима следећу табелу: *City, Tourism, pred_Tourism, abs_err*. pred_Tourism представља предвиђену (*fitted*) вредност варијабле Tourism за сваки град, добијену применом Ridge регресионог модела при λ = λ_min.

City	Tourism	pred_Tourism	abs_err
Kuala Lumpur	0.942	0.325	0.617
Paris	0.972	0.363	0.609
Bangkok	0.942	0.334	0.608
Istanbul	0.750	0.182	0.568
Macau	1.000	0.441	0.561

Издвојите само градове који су на континентима *Europe* и *Asia*, и који имају *crime* изнад медијане (медијана се рачуна на целом скупу).

На том подскупу *fit*-ујте Ridge (*cv.glmnet, lambda.min*), израчунајте *pred_Tourism* и *abs_err*, па затим (помоћу *for* петље) за сваки од два континента извуците топ 3 града са највећом abs_err.

Continent	City	Tourism	pred_Tourism	abs_err
Europe	Geneva	0.042	0.481	0.439
Europe	London	0.972	0.540	0.432
Europe	Rome	0.828	0.409	0.419
Asia	Macau	1.002	0.406	0.596
Asia	Shenzhen	0.888	0.489	0.399
Asia	Hong Kong	1.002	0.617	0.385

Задатак 2

а) Испитати да ли постоји разлика у вредностима варијабле *Easeofdoingbusiness* ако посматрамо градове у Европи и Азији.

Први корак у анализи је да спроведемо Shapiro-Wilk тест.

Нулта хипотеза Подаци имају нормалну расподелу.

Алтернативна хипотеза Подаци немају нормалну расподелу.

Вредности статистика су 0.95311,0.84781, док су *p* вредности респективно 0.1766, 0.002484.

Закључујемо да Europe има нормалну расподелу док Asia нема нормалну расподелу,

па за испитивање разлика користимо Wilcoxon-Mann-Whitney тест.

На основу добијених резултата тумачимо резултате теста

Нулта хипотеза:Нема разлике у вредностима Easofdoingbusiness између Европе и Азије.

Алтернативна хипотеза: Постоји статистички значајна разлика у вредностима Easeofdoingbussines између Европе и Азије.

Вредност статистике је 586.5, док је *p* вредност 0.0001914.

Нулту хипотезу **ОДБАЦУЈЕМО** ПРИХВАТАМО

ЗАКЉУЧАК Постоји статистички значајна ралике у Easeofdoingbusiness између Европе и Азије.

б) Издвојите само градове из Европе и Азије који имају *Greenspace* изнад медијане (медијана се рачуна на целом скупу).

На том подскупу, за сваку од променљивих *crime*, *health* и *Energy* формирајте табелу са резултатима испитивања разлика између задатих група. У табели приказати (испуњеност критеријума нормалности уз референтну p-value + који тест је коришћен + статистика и p-value коришћеног теста).

Variable	***Shapiro_p_Euro***	***Shapiro_p_Asia***	**Коришћени тест**	***p_value***	***Test statistic***
crime	0.2172	0.0014	Wilcoxon	1.0000	108.0
health	0.0723	0.0442	Wilcoxon	0.9659	109.5
Energy	0.0837	0.0026	Wilcoxon	0.0000	208.0

ц) Формирајте подскуп који садржи само *Europe* и *North America* градове са *affordability* у првом квартилу. Затим, за три прага crime (0.8, 0.9, 0.95) правите додатни подскуп *crime* < праг и за сваки праг тестирајте разлику у *Tourism* између *Europe* и *North America*. Резултате приказати у следећој табели.

Crime **праг**	***n_Europe***	***n_North America***	***Shapiro_p_ Europe***	***Shapiro_p_ NorthAmerica***	**Коришћени тест**	***p_value***
0.8	0	12	NA	0.0078	Mali uzorak(n=0 nije moguce odrediti)	NA
0.9	0	15	NA	0.0010	Mali uzorak(n=0 nije moguce odrediti)	NA
0.95	2	15	NA	0.0010	Mann-Whitney (Wilcoxon)	0.06225

Задатак 3: Испитати да ли постоји разлика у вредностима варијабле *Connectivity* у зависности од континента на коме се град налази.

Први корак у анализи је да спроведемо Shapiro-Wilk тест.

Нулта хипотеза Подаци (Connectivity) у групи имају нормалну расподелу.

Алтернативна хипотеза Подаци (Connectivity) у групи немају нормалну расподелу.

Вредности статистика су 0.97397,0.93577,0.90948, док су *p* вредности респективно 0.6154,0.1618,0.0398.

Закључујемо да једна група није нормално расподељена (Asia),

па за испитивање разлика користимо Kruskal-Wallis тест.

На основу добијених резултата тумачимо резултате теста

Нулта хипотеза: Не постоји разлика у Connectivity између континената.

Алтернативна хипотеза: Постоји разлика у Connectivity између бар два континента.

Вредност статистике је 18.8, док је *p* вредност 0.00008272.

Нулту хипотезу **ОДБАЦУЈЕМО** ПРИХВАТАМО

ЗАКЉУЧАК: Постоји статистички значајна разлика у Connectivity између континената.

Издвојите градове који истовремено испуњавају услов: *Greenspace* > медијана (рачуната на целом скупу), па на том подскупу, испитајте да ли постоји разлика у *Tourism* између континената.

Shapiro-Wilk p-value за Европу, Северну Америку и Азију су: 0.08343, 0.09299,0.06468.

Изабрани тест је: ANOVA, а добијена p-value теста 0.388.

Post-hoc тест који можемо спровести је Tukey, са методом (HSD)Honestly Significant Difference.

Овим тестом се показује да постоје разлике између континената

не постоје статистички значајне разлике између континената, јер

су припадајуће p-values (2-1) 0.7634, (3-1) 0.6309, (3-2) 0.3786.

Задатак 4. Коришћењем методе *K-medoids,* потребно је кластеризовати градове према различитим димензијама, а у односу на следеће варијабле: *health, affordability, Energy, Easeofdoingbusiness u Connectivity*. Одаберите k = 3. Након кластеровања, профилишите кластере, израчунајте просечне вредности свих коришћених индикатора у сваком кластеру.

Приликом спровођења *K-medoids* методе, бирају се апстрактне/стварне опсервације које овој методи обезбеђују већу/мању робустност у односу на *K-means* методу, јер су на овај начин кластери више/мање осетљиви на екстремне вредности у подацима.

Приликом спровођења *K-medoids* методе, подразумевана метрика за рачунање одстојања је eucliedan. Ако би било потребно употребити *Manhattan метрику*, онда се прво се мора направити матрица растојања (dist()) а затим се позива функција pam() са одговарајућим параметрима.

Број градова у првом, другом и трећем кластеру је 39, 21, 17, респективно.

Медоиди првог, другог и трећег кластера, респективно, су градови "Munich", "Houston" и "Shenzhen".

Метрика која се користи за интерпретацију ваљаности спроведене кластеризације зове се silhouette width, а у случају спроведеног модела вредност те метрике се чита помоћу функције pamcluster$silinfo$avg.width**,** а вредност је 0.3782.

Кластер са најмањом просечном вредношћу варијабле *health* је кластер чији је медоид град "Houston" а та вредност је 0.523.

Кластер са највећом просечном вредношћу варијабле *Energy* је кластер чији је медоид град "Munich", а та вредност је 0.647,.

Кластер са најмањом просечном вредношћу варијабле *Connectivity*је кластер чији је медоид град "Shenzhen". а та вредност је 0.316**.**

Издвојите само градове који испуњавају услове: *Tourism* > медијана(*Tourism*) на целом скупу и *crime* < медијана(*crime*) на целом скупу. На том подскупу спроведите *pam*() (*k* = 3, *metric* = "*euclidean*") над 5 индикатора одабраних на почетку. Затим, за сваки кластер, користећи *for* петљу, пронађите град који је најудаљенији од свог медоида и попуните следећу табелу:

Кластер	Медоид	Најудаљенији град	Растојање до њега	Број градова у кластеру
1	Antwerp	Bangkok	3.098	7
2	Moscow	Moscow	0.000	1
3	Los Angeles	New York	3.606	6

Издвојите само градове који имају *health* > медијана(*health*) на целом скупу. На том подскупу урадите *pam*() (k = 3, *metric* = "*euclidean*"). Затим, за сваки континент посебно, израчунајте:

- који је доминантан кластер (онај који у себи садржи највише градова тог континента),
- колики је удео доминантног кластера унутар тог континента,
- и укупан број градова у континенту (*n_total*).

Резултате прикажите попуњавањем следећих табела:

Табела. Расподела градова по кластерима у односу на припадајуће континенте

Континент	1. кластер	2. кластер	3. кластер
Европа	24	0	0
Северна Америка	1	0	0
Азија	2	8	3

Табела. Преглед доминантних кластера по континентима

Континент	Доминантни кластер	Удео	Укупно градова на континенту
Европа	1	1.000	24
Северна Америка	1	1.000	1
Азија	2	0.615	13

Након спроведене *K-medoids (PAM)* кластеризације са бројем кластера $k = 3$, потребно је упоредити резултате добијене применом **две различите метрике растојања** над истих пет индикатора (*health, affordability, Energy, Easeofdoingbusiness u Connectivity*): еуклидске метрике и *Manhattan* метрике. За еуклидску метрику *PAM* се примењује директно на матрицу података, док се за *Manhattan* метрику најпре конструише матрица растојања помоћу функције *dist(..., method = "manhattan"),* а затим се функција *pam*() позива са аргументом *diss* = *TRUE*. Кроз *for* петљу

потребно је за оба модела израчунати две мере квалитета кластеризације: просечну вредност силуетног индекса (*avg.width*), која показује колико су кластери добро раздвојени, и финалну вредност *objective* функције, која представља укупно растојање унутар кластера. На основу добијених вредности, резултате треба приказати у табели и закључити који модел је бољи, а затим навести и имена градова који су идентификовани као медоиди у том бољем моделу.

Резултати:

Метрика	***avg_sill_score***	***Objective score***
Euclidean	0.3782	1.3501
Manhattan	0.3903	2.3702

Боља метрика је "Manhattan", зато што има већу просечну вредност силуетног индекса.

Медоиди добијени помоћу боље метрике јесу: "Munich","Miami","Macau".

Задатак 5. Коришћењем методе факторске анализе смањити димензионалност дате матрице података и одговорити на следеће захтеве.

Када је потребно спровести чисту редукцију димензије, компресију података или визуализацију без претпоставки о скривеним узроцима, користи се метода **АГК**/ФА. Ако имате велики број корелисаних индикатора и треба вам мањи број синтетичких варијабли за кластеровање, регресију или визуализацију, метода **АГК**/ФА је одговарајућа јер **користи**/не користи све варијансе **променљивих**/фактора и **не захтева**/захтева латентни модел.

Када постоји теоријски разлог да иза корелација стоје неколико латентних конструката (фактора), нпр. у психометрији (мерење особина личности), маркетингу или сродним наукама, користи се метода АГК/**ФА**. Метода АГК/**ФА** издваја **заједничку**/сингуларну варијансу и омогућава интерпретацију променљивих/**фактора** као „узрока" корелација, док **специфичну**/заједничку варијансу третира као шум или мерну грешку.

Израчунајте *p-value* за *Bartlett test of sphericity*: 2.22e-16.

Израчунајте *overall KMO (MSA)*: 0.6529.

Уколико је потребно објаснити барем 70% варијабилитета података, узимајући у обзир кумулативне варијансе компоненти добијених из датог скупа података, потребно је изабрати 2 као број фактора за факторску анализу

У факторској анализи, факторско оптерећење је корелација између променљиве и фактора**.**

У факторској анализи, комуналитет променљиве је**је део варијансе променљиве који објашњавају задржани фактори и рачуна се као збир квадрата факторских оптерећења те променљиве преко свих задржаних фактора.**

Применом *varimax* ротације, за задати број фактора, применом факторске анализе добијају се следећи резултати:

На фактору MR1, највеће апсолутно оптерећење има променљива "Easeofdoingbusiness" са вредношћу 0.961.

Која је променљива најлошије објашњена факторима (има најмању вредност комуналитета) и колика је та вредност? "Energy" , 0.201.

Комуналитет променљиве *health* износи приближно 0.33, што значи да фактори објашњавају 33% њене варијансе.

Ако град има висок MR1 скор, а низак скор на *affordability*, очекује се да буде : **високо развијен**/ниско развијен за бизнис, а приступачан/**скуп**.

Фактор са највећом просечном комуналношћу (*average communality*) је фактор MR1 а та вредност износи 0.4374. Две променљиве са највећим оптерећењем (|*loading*|) на том фактору, из претходног питања, су "Easeofdoingbusiness" и "affordability".

Гледајући сваки задржани фактор, пронађите градове који имају највеће вредности за сваки фактор, па у следећу табелу упишите те градове и њихове одговарајуће вредности:

Фактор	**Град**	**Скор**
MR1	*New York*	1.3511
MR2	*Tokyo*	1.6623
MR3	*Moscow*	2.6479
MR4	*Lisbon*	1.2933

Спроведите факторску анализу са *nfactors* = 4 и *rotate* = "*varimax*" (обавезно израчунајте *factor scores*). Затим издвојите само градове из Европе који су у горњем квартилу по *Tourism* (односно задовољавају услов: ≥ *quantile(..., 0.75*), рачунато унутар Европе). На том подскупу, уз помоћ *for* петље израчунајте *cor()* између *Tourism* и сваког фактора (*MR1, MR2, MR3, MR4*), па наведите фактор са највећом |*cor*|. Дакле, резултати су редом:

Вредност горњег квартила, на основу које се врши селекција података, износи: 0.5505.

Вредности коефицијента корелације између варијабле *Tourism* и сваког од фактора су:

-0.0743,0.2964,0.4837,0.2918.

Фактор који има највећи коефицијент корелације јесте MR3 и та вредност је 0.4837.

Посматрајте само променљиве *health, affordability, Energy, Easeofdoingbusiness u Connectivity*, и за сваки континент посебно спроведите анализу погодности и факторску анализу на следећи начин: најпре израчунајте *p-value Bartlett*-овог теста сферичности, затим одредите укупну *KMO (MSA)* меру адекватности узорка, након чега на основу сопствених вредности (*eigenvalues* > 1) корелационе матрице изаберите број фактора према *Kaiser*-овом критеријуму; потом спроведите факторску анализу са *varimax* ротацијом и за изабрани број фактора идентификујте променљиву са највећим комуналитетом. Добијене резултате (по континентима) уписати у табелу испод.

Континент	***n***	***Bartlett_p***	***KMO MSA value***	***n_factors***	***top_communality_variable***	***Comm_value***
Европа	32	4e-05	0.6254	2	Connectivity	0.9410
Северна Америка	22	0e+00	0.4664	2	Easeofdoingbusiness	0.9785
Азија	23	0e+00	0.8965	1	Easeofdoingbusiness	0.9150

Потребно је да се најпре спроведе факторска анализа над нумеричким променљивама, уз фиксиран број фактора *nfactors = 4*, примену *varimax* ротације и израчунавање факторских скорова методом *scores = "regression"*. Затим је неопходно издвојити подскуп градова који се налазе у горњем квартилу по променљивој *Connectivity* и истовремено у доњем квартилу по променљивој *affordability*, при чему се квартилни прагови рачунају на целом скупу података. На том подскупу треба, коришћењем *for* петље, израчунати просечан факторски скор за сваки фактор и идентификовати фактор код кога је апсолутна вредност тог просека највећа, јер тај фактор најјаче карактерише издвојени подскуп градова. Након тога, за тај изабрани фактор потребно је рангирати градове према вредности њиховог факторског скора и издвојити пет градова са највећим позитивним вредностима фактора ако је просек позитиван, односно пет градова са најмањим (најнегативнијим) вредностима фактора ако је просек негативан, јер су то градови који највише одступају од нуле у истом смеру као и просечан факторски скор и тиме најјаче испољавају карактеристике тог фактора. Резултати се уписују у табелу која садржи следеће колоне: Град (назив града), Континент (континент на ком се град налази), *Tourism* (вредност туристичког индекса), Одабрани фактор (ознака изабраног фактора, нпр. *MR1*), и *Score* (факторски скор тог града за наведени фактор).

Просечне вредности сваког од фактора (*MR1,MR2, MR3, MR4*) су: 0.7092, 0.8744,0.4628,

-1.0553.

Од свих фактора бира се MR4, јер има највећу апсолутну вредност просечног фактора на издвојеном скупу.

Ранг	**Град**	**Континент**	***Вредност варијабле Tourism***	**Одабрани фактор**	***Score***
1	Zurich	Evropa	0.354	MR4	-1.6065
2	New York	Severna Amerika	0.726	MR4	-1.5848
3	Beijing	Azija	0.252	MR4	-1.5023
4	Tokyo	Azija	0.330	MR4	-1.4809
5	Shanghai	Azija	0.336	MR4	-1.4103

Задатак 6. Потребно је учитати библиотеке за рад са методом анализе преживљавања. Успоставите радно окружење тако што ћете инсталирати и учитати *survival, survminer u readxl*. Базу података *plucaSurvival.xlsx* учитајте, преко функције *read_excel*. Дефинишите догађај (*event*): Креирајте колону *event* као индикатор смрти – вредност 1 када је *status* == 2, иначе 0. Ово омогућава правилну интерпретацију у моделирању.

У бази *plucaSurvival.xlsx* дата је категоричка варијабла пол (1 – мушкарац, 2 – жена). Спровести анализу преживљавања и одговорити на следеће захтеве

У скупу анализираних података, постоји 228 опсервација. Број одиграних догађаја (хазардних догађаја) је 149.

Медијално време преживљавања је 344.0637 временских јединица.

Да би се статистички проверила разлика између полова, спроводи се log-rank тест. Статистика овог теста износи 0.9122, док је одговарајућа *p* вредност једнака 0.3395. Закључује се да не постоји статистички значајна разлика у преживљавању између мушкараца и жена.

Зарад потребе провере статистички значајне потврде пола као варијабле са прогностичким фактором преживљавања спроводи се Cox PH регресија. Одговарајућа *p* вредност једнака је 0.3410. Закључује се да пол није статистички значајан прогностички фактор преживљавања.

Из спроведеног модела могло би да се претпостави да **жене**/мушкарци имају **14.82** % **мањи**/већи ризик од хазардног догађаја у односу на жене/**мушкарце**.

Креирајте мултиваријантни Cox PH модел који укључује променљиве *sex*, *age* и *ph.ecog*. Допуните:

Вредности exp(coef) (hazard ratio) за пол је 0.7903; за узраст је 1.0006; а за ph.ecog је 1.2550. *P*-вредности су 0.1680 (пол), 0.9493 (узраст) и 0.0082 (ph.ecog), а то значи да је само ph.ecog статистички значајан предиктор преживљавања док пол и узраст нису.

Повећање ph.ecog за једну јединицу повећава ризик од хазардног догађаја за 25.50%, јер је HR = 1.2550 .

Жене имају HR**:** 0.7903, што значи да имају 20.97% мањи ризик од мушкараца. Та разлика јесте/**није** статистички значајна јер је *p-вредност* једнака 0.1680.

Над већ учитаном и сређеном матрицом података, формирајте три старосне групе *AgeGroup* на основу терцила променљиве *age* (групе *Low*, *Medium*, *High*). Спроведите *Kaplan–Meier* анализу по групама и урадите *log-rank* тест да проверите да ли постоји разлика у преживљавању између ове три групе (прикажите *chisq u p-value*), а затим за сваку групу израчунајте и прикажите у табели: број пацијената *n*, број догађаја *events*, медијално време преживљавања *median_survival* (у данима) и вероватноћу преживљавања на *t = 365 (S_365)*; након тога, „најжилавију" групу дефинишете као ону са највећом медијаном преживљавања, и за њу израчунајте колико је дана „дужа" њена медијана у односу на преостале две групе (*delta_days = median_best - median_other*), као и колико то износи додатних дана по сваком месецу трајања друге групе, где месец рачунате као 30 дана, тј. *extra_days_per_month = delta_days / (median_other/30*) (ове разлике прикажете у посебној табели за поређења најжилавије групе са преостале две).

Према резултату *log-rank теста* (p = 0.8453), да ли постоји статистички значајна разлика у преживљавању између *AgeGroup* група? Одговор: НЕ (ДА/НЕ)

На нивоу значајности α = 0.05, нулту хипотезу „криве преживљавања су исте у све три групе": НЕ ОДБАЦУЈЕМО (ОДБАЦУЈЕМО / НЕ ОДБАЦУЈЕМО)

Табела 1: Резиме резултата према групама старости AgeGroup

AgeGroup	n	events	median_survival	S_365
Medium	76	52	301.6766	NA
Low	76	48	353.0000	NA
High	76	49	361.1070	NA

„Најжилавија група" је група: **high.**

Табела 2: Резиме резултата анализе преживљавања између „најжилавије групе" и других група

Најжилавија Група	**Преостала група**	**Median_days Најжилавија група**	**Median_days Преостала група**	**Delta_days**	**Extra_days_ per_month**
High	Low	361.1070	353.0000	8.1070	0.6890
High	Medium	361.1070	301.6776	59.4304	5.9100

Посматрајте само пацијенте који су *sex = "Жена"* и имају *ph.ecog ≤ 1*, а затим унутар тог подскупа дефинишите кандидате за старосни праг (*cutoff*) као квантилне вредности променљиве *age* добијене за нивое *0.20, 0.30, 0.40, 0.50, 0.60, 0.70 и 0.80*. За сваки такав *cutoff* формирајте бинарну групну променљиву *AgeBin* са нивоима *age < cutoff* и *age ≥ cutoff*, али у даљу анализу укључите само оне *cutoff* вредности за које обе групе садрже најмање *n_min = 5* пацијената и најмање *events_min = 5* догађаја (смрти). За сваки важећи *cutoff* спроведите *log-rank* тест поређења кривих преживљавања и израчунајте вредности χ^2 статистике (*chisq*) и одговарајуће *p_value*, након чега као „најбољи" изаберите онај *cutoff* који даје највећу *chisq* вредност. На крају, за изабрани *cutoff*, прикажите за обе старосне групе (*AgeBin < cutoff и AgeBin ≥ cutoff*) следеће параметре: број пацијената (n), број догађаја (events), медијално време преживљавања (median_survival, у данима) и вероватноћу преживљавања у времену t = 365 дана (S(365)).

Вредност χ^2 статистике *log-rank* теста за изабрани *cutoff* износи 0.8641.

Одговарајућа *p-value* за *log-rank* тест износи 0.3526.

На основу *log-rank* теста за оптимални *cutoff* = 65, да ли постоји статистички значајна разлика у преживљавању између две старосне групе (AgeBin < cutoff и AgeBin ≥ cutoff)?

Одговор: НЕ ПОСТОЈИ (ПОСТОЈИ / НЕ ПОСТОЈИ)

Табеларно упоредно прикажите разлике у *AgeBin* групама за оптимални *cutoff* и то према колонама: n, број догађаја, медијално време преживљавања и S(365).

AgeBin	n	events	median_survival	S(365)
age < 65	39	22	433.0000	NA
Age >= 65	21	16	349.9066	NA

Задатак 7. Потребно је учитати библиотеке за рад са методом *biclustering*. Потребно је учитати податке из предефинисане базе података *Food.sav,* која је дата на дељеном диску. Приликом спровођења анализе максималан број бикластера које алгоритам може да формира ограничити на 25. Параметар delta = 1.5 и alpha = 1 и method = BCCC().

Средња вредност променљиве V17 у другом бикластеру је 34.6.

Највећа вредност променљиве V13 у четвртом бикластеру је 69.3.

Медијана променљиве V7 у петом бикластеру износи 31.2.

Минимална вредност променљиве V2 у трећем бикластеру је 100.

У неуређеном Bicluster membership графику променљива V5 се појављује у 5 бикластера.

У сортираном Bicluster membership графику променљива V11 се појављује у 2 бикластера.

Променљива која се појављује у највише бикластера је V3 и појављује се 16 пута.

Практична вредност употребе ове методе огледа се и у проналажењу најзаступљенијих парова заједничких појављивања две варијабле у оквиру различитих бикластера. Стога, у датој матрици података, највећи број заједничког појављивања две варијабле у различитим бикластерима је

5.

Варијабле које се заједно појављују највећи број пута су: V3-V4 и V3-V6 .

Афинитет бикластера дефинише се као однос између средње вредности свих елемената унутар конкретног бикластера и средње вредности истих одабраних варијабли у целој матрици. Бикластер са највећим афинитетом је кластер број 19, а вредност афинитета је 1.42692.

За сваки бикластер израчунајте *within_sd = sd() свих елемената унутар бикластера*. Задржите само бикластере са n_rows >= 6 и n_cols >= 5 и пронађите бикластер са најмањом *within_sd*.

Према задатим критеријумима, набољи бикластер је 3. Он има 9 редова и 5 колона. Вредност за његов параметар *within_sd* износи: 2.04868.

Задатак 8. Применом методе логистичке регресије потребно је детаљније истражити матрицу *Arcadis.xlsx*. Прво је потребно од променљиве *Employment* направити категоричку променљиву *EmploymentCategory* са ознакама „*low*“ за вредности од минималне до 0,3 , затим „*medium*“ од 0,3 до 0,6 и онда „high“ за све веће од 0,6. Друго, потребно је од променљиве *Education* направити категоричку променљиву *EducationCategory* са ознакама „*low*“ за вредности од минималне до 0,4 , затим „*medium*“ од 0,4 до 0,6 и онда „high“ за све веће од 0,6.

Након тога, креирати варијаблу *HighEmployment,* тако да уколико је *EmploymentCategory* =“high“ онда је вредност 1, а у супротном 0.

Поделити скуп података на тренинг и тест сет у односу 70:30.

Креирати и евалуирати успешност модела логистичке регресије који предвиђа варијаблу *HighEmployment* помоћу варијабли, *crime, Drinkingwaterandsanitation, Airpollution, Energy, Greenspace, EducationCategory*. Није потребно делити матрицу података на тренинг и тест податке.

Код варијабле *EmploymentCategory* број инстанци категорија од најмање до највеће категорије, респективно је 10 , 47 , 20.

Код варијабле *EducationCategory* број инстанци категорија од најмање до највеће категорије, респективно је 16, 24 , 37.

Када се креира варијабла *HighEmployment*, она броји 57 инстанци за вредност 0, а 20 инстанци за вредност 1.

Статистички значајни предиктори су варијабле:

crime, Energy, Greenspace, EducationCategoryMedium.

Оцењена вредност коефицијента за варијаблу *EducationCategoryMedium* је 3.0260 и она јесте статистички значајна. Ако се ова варијабла повећа за 1 мерну јединицу, то значи да се ако град пређе из low у medium образовање,шансе за висок ниво запослености расту око 20 пута, уз фиксне остале варијабле.

Додатно, шансе да варијабла *HighEmployment*=1 су 20.6 пута **веће**/мање за оне градове са *Medium* образовањем у односу на *Low*, када су све остале варијабле фиксне.

Након извршења предикције бинарне категорије задате варијабле *HighEmployment* (узети граничну вредност одлуке „>0,5“), добијају се вредности евалуационих метрика:

Accuracy = *0.792* *Precision* = *0.667* *Recall* = *0.400* *F1* = *0.500*

Урадите 10 различитих подела на train set / test set (70:30) за *seed* вредности од 123 до132. За сваки *seed*: тренирајте исти *glm* модел (као што је задат у задатку) и израчунајте на *test set*-у *Accuracy, Precision, Recall, F1* за *threshold = 0.5*. На основу добијених резултата, попуните следећа празна поља:

Seed	Accuracy	Precision	Recall	F1
123	0.750	1.000	0.143	0.250
124	0.667	0.500	0.125	0.200
125	0.625	0.500	0.333	0.400
126	0.792	0.667	0.333	0.444
127	0.750	0.400	0.400	0.400
128	0.833	1.000	0.429	0.600
129	0.792	0.500	0.200	0.286
130	0.667	0.429	0.429	0.429
131	0.750	1.000	0.250	0.400
132	0.792	0.500	0.400	0.444

Због уједначеног сагледавања врлина сваког модела, метрика која се користи за оцену, од приказаних, је F1. На основу ње, најбољи модел је онај чија *seed* вредност износи 128.