Прашања ВНП updated

Колоквиум 1

Квалификациски

1. Дадени се tp, fp, tn, fn, да се најде прецизност. Пример tp=10, fp=25, tn=15, fn=20
precision=___/___

Одговор: 10/35.

2. Кои од наведените се карактеристики за Data Science, за разлика од машинско учење:

- ☑ Пробување на различни параметри и различни модели за решение на одреден проблем.
- ☐ Креирање на модели
- ☐ Докажување на математички својства на модели
- ☑ Разбирање на емпириски својства на моделите

3. Кога користиме одредени податоци за модел во кој е важна далечината, кој енкодер се користи за следните типови на колони:

X	Y
Не се сеќавам што имаше тука	Cat
Не се сеќавам што имаше тука	Dog
Не се сеќавам што имаше тука	Tiger
Не се сеќавам што имаше тука	Fish
Не се сеќавам што имаше тука	Parrot

Dropdown: OneHotEncoder, LabelEncoder, ниедно.

Одговор: OneHotEncoder.

4. Кој енкодер би го користел за следново?

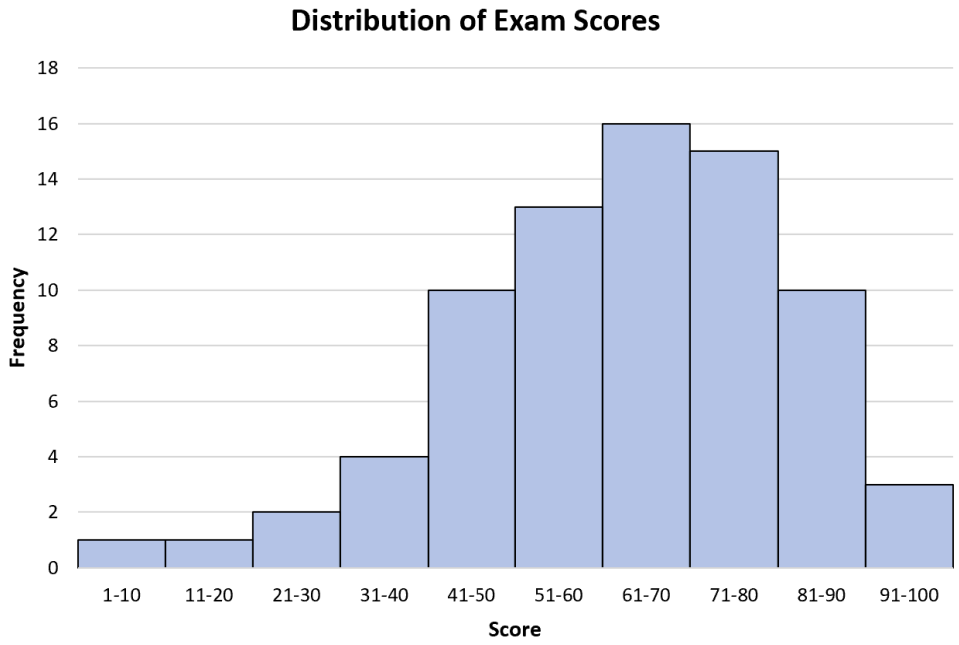
X	Y
Не се сеќавам што имаше тука	Gold
Не се сеќавам што имаше тука	Silver
Не се сеќавам што имаше тука	Gold,Silver

Не се сеќавам што имаше тука	Silver
Не се сеќавам што имаше тука	Gold

a) LabelEncoder
b) OneHotEncoder
c) Ниеден

Одговор: b.

5. Ако има left-skew дистрибуцијата како на сликата, што е точно?



a) Median > mean
b) Median = mean
c) Median < mean
d) Не може да се заклучи

Одговор: a.

6. Дадена е сликата, кога се користи MICE, што од следното е точно:

- ☐ Колоната medv има висока корелација со сите останати колони
- ☑ Помеѓу dis и indus има висока корелација
- ☐ Помеѓу dis и nox нема корелација
- ☑ Помеѓу lstat и medv има висока негативна корелација
- ☐ Доколку расте zn ќе расте и chas
- ☑ Доколку опаѓа dis, age ќе расте

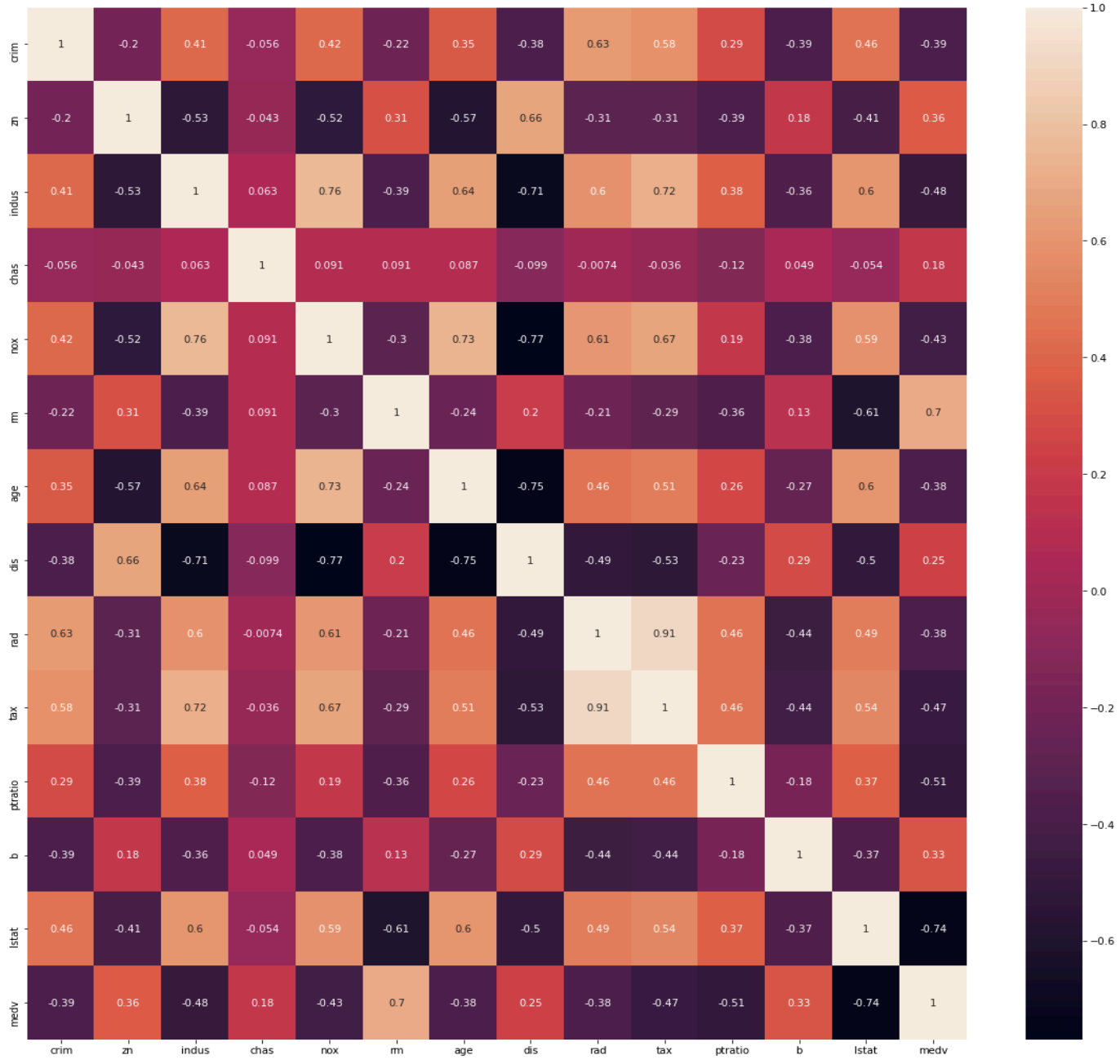
☑ На сликата е прикажана heatmap

***Висока корелација над 0.6/0.7 и -0.6/0.7**

***Нема корелација ако е 0**

***Заедно растат ако е >0, едното расте а другото опаѓа <0**

***Се чита x па y**



7. Доколку имаме KNN класификатор со k=1, дали класификаторот врз податоците ќе биде:

a) overfitting
b) underfitting
c) just right

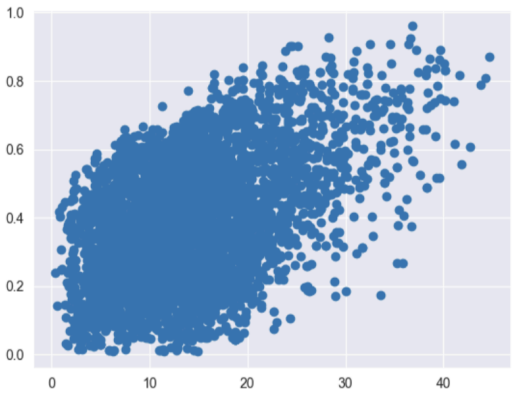
Одговор: a.

8. Што се случува кога ентропијата кај даден датасет тежнее кон нула?
   a) Податоците се добро поделени
   b) Податоците се несредени т.е. Немаат добра поделба

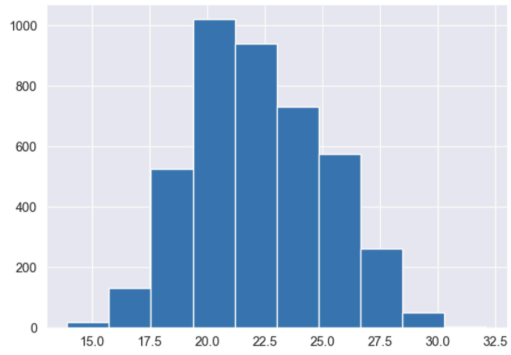
Одговор: a. (А кога тежнее кон 1 податоците се несредени.)

9. Кој график ни укажува дека ако расте Open ќе расте и Closed ако Open е на x-оската, а Closed на y-оската :

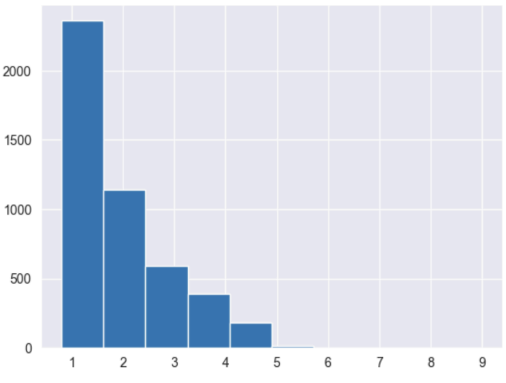
a)

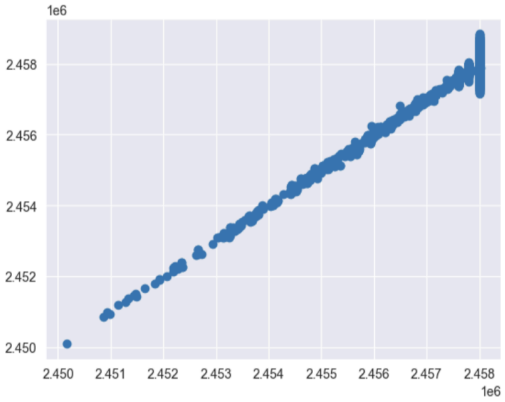


b)



c)

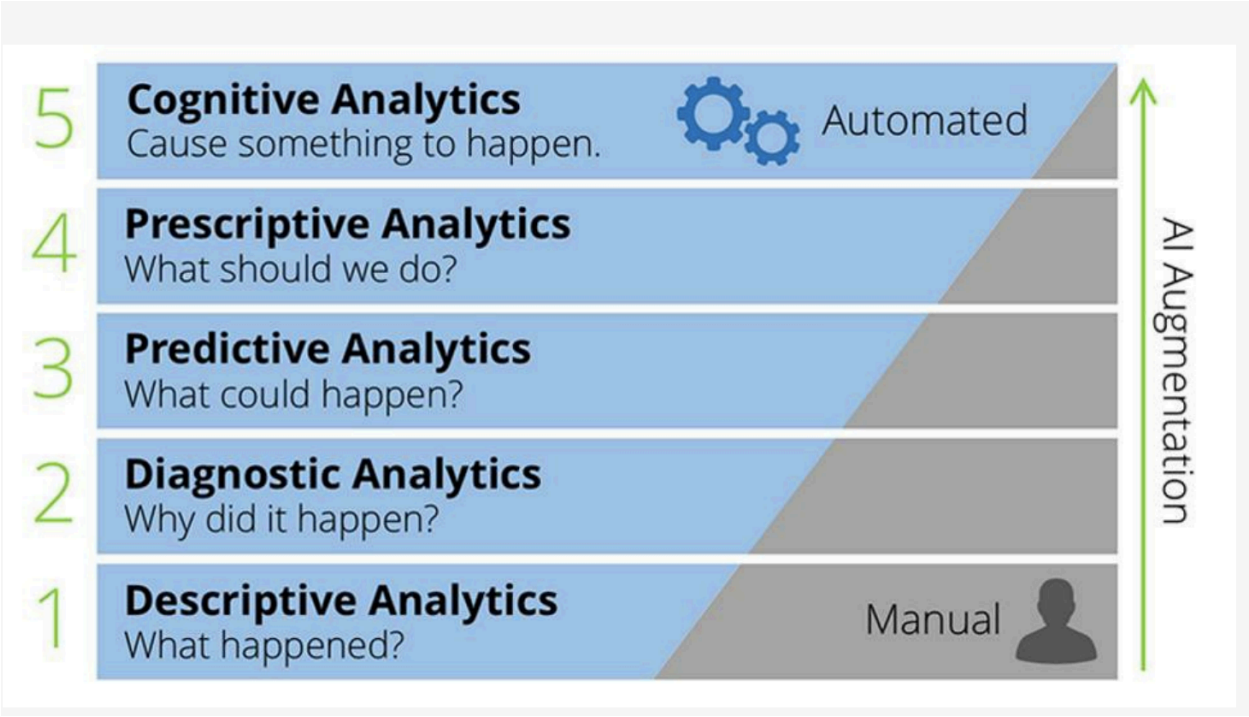




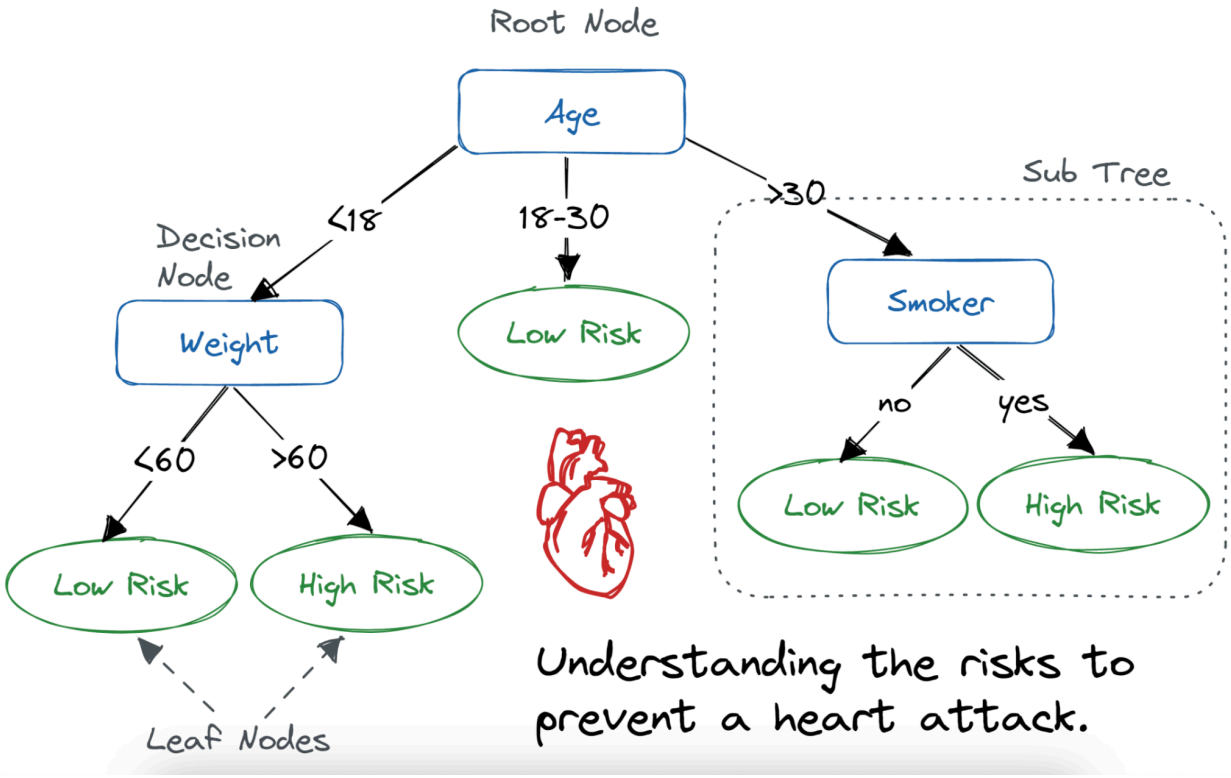
Одговор: d

11.Да се подредат видовите на Analytics од 1 до 5, од долу нагоре, според AI Augmentation:
Descriptive Analytics, Cognitive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics.

Одговор:



12. Дадени влезни податоци и слика од дрво на одлука. Да се предвиди според дрвото која класа ќе ја имаат податоците.



https://images.datacamp.com/image/upload/v1677504957/decision_tree_for_heart_attack_prevention_2140bd762d.png

На пример оваа слика дадена и податоци: Age: 25, Weight 25, Smoker: Yes и да се одреди дали ќе се додели класата Low Risk или High Risk.

14. Имаме 2 колони A и B. A содржи одредени податоци, B други. И во двете колони фалат податоци во одредени редици. Задачата беше да се импутираат вредностите за двете колони. За A да се импутираат со KNN импутација со k=2, а на другата колона да се импутираат со мода. Потоа дадена ваква слична табела и да се пополни (Вредностите за A_miss и B_miss се пополнуваат со 1 ако во таа редица фали податок):

A	B	A_new	B_new	A _miss	B_miss

1	null	//////////////////////	Imputacija 1	0	1
5	1	//////////////////////	//////////////////	0	0
null	0	Imputacija (5+4)/2	//////////////////	1	0
4	null	//////////////////////	Imputacija 1	0	1
3	1	//////////////////////	//////////////////////	0	0

Дополнително - теорија

1. Податочно множество од 50 редици и 5 колони (имаше и 100 редици и 10 колони) и да искоментираш за min_samples_leaf=10 колку би можела да биде вредноста на max_depth.

2. Опиши го R2 score.

3. Разлика и примена помеѓу Precision и Recall

4. Дадени редици со null вредности за променливата A или B - A е непрекината, B е категориска(енкодирана ваљда), и треба ако се користи KNNImputer/SimpleImputer со мода да се напише кои вредности ќе се стават на местото на nullовите, и од страна имаш дополнителни две колони A_miss & B_miss, ако во соодветната редица вредностите за A и B ти се, на пример, 15 и 0 - тогаш A_miss и B_miss се 0 и 0 (ги имаш и двата податока), а ако се на пример Null и 1, би биле 1 и 0 бидејќи A е missing.

5. Доколку имаме dataset во кој нема некоја силна линеарна врска помеѓу влезните податоци и таргет колоната, кој тип на регресија може да се искористи и зошто.

6. Небалансирани податоци во Supervised learning.

Дополнително - задачи

Треба да се одговорат прашања во однос на задачите кои всушност ќе нѐ водат што да правиме, во кој редослед.

1. Дадена е една колона од податочните множества која треба да се предвиди и може да биде која било од колоните, зависно од групата која ќе се падне на студентот. Пример прашања се:
    - Колку колони имаат missing values
    - Колку колони ќе употребите за тренирање на моделот
    - Каков модел ќе користите за предвидување на оваа колона
    - Подредете по редослед што ќе правите за да се справите со missing values.

    https://colab.research.google.com/drive/1UCY49__CRn0l4enW_dvNH6gqAUHBLE6X

2. Треба да се најдат најдобрите хиперпараметри за дрво на одлука и да се употреби KFold со 5 поделби. На пример параметрите што треба да се најдат се: criterion, max_depth, min_samples_split. Да се употребат најмногу 3 можни вредности за секој параметар од кои ќе се избере 1 (за да се не се извршува програмата предолго).
    Пример прашања се:
    - Колку пати ќе се тренира со една поделба од KFold
    - Колку пати ќе се тестира врз една поделба од KFold
    - Дали е target колоната балансирана или не

    https://colab.research.google.com/drive/1x3d2fZFdpLP-7wd3rfZrwCmWO3s-rm3l

- **Vtor del**

Даден е сегмент од некое податочно множество во кој имаме две влезни карактеристики A и B, при што има податоци што недостасуваат. Ваша задача е да ги пополните овие вредности, при што за карактеристиката A ќе користите метод на импутација со KNN со k=2, а додека за карактеристиката B ќе ја користите модата како импутациски метод.

***мода е најповторуван податок**

A	B

10	0
5	null
null	0
21	1
15	null
null	0
16	null
7	0

Резултат: Во следната табела пополнете ги вредностите што недостасуваат, при што треба да ги пополните и соодветните missingness колони за влезните карактеристики A и B.
НАПОМЕНА: Резултатите да се заокружат на 1 децимала, со користење на децимална точка!

A	B	A_{new}	B_{new}	A_{miss}	B_{miss}
10	0	10	0	Answer 0	Answer 0
5	null	5	Answer 0	Answer 0	Answer 1
null	0	Answer (5+21)/2	0	Answer 1	Answer 0
21	1	21	1	Answer 0	Answer 0
15	null	15	Answer 0	Answer 0	Answer 1

null	0	Answer (15+16)/2	0	Answer 1	Answer 0
16	null	16	Answer 0	Answer 0	Answer 1
7	0	7	0	Answer 0	Answer 0

2.Дадено е податочно множество кое се состои од 5 колони и 50 редици. Потребно е да се изгради дрво на одлука кое има дефиниран параметар min_leaf_samples=10. Колку би можела да биде максималната длабочина на дрвото кое би се добило во овој случај. Образложете го одговорот.

3.

НАПОМЕНА:
На истата страница со ова прашање подолу ви се наоѓа есејско прашање каде треба да прикачите решение на задачата.
Затоа што испитот е open book се воведени следните правила:
- Во случај да одговорите точно на прашањата поставени подолу, а тоа да го нема во кодот ќе освоите нула поени на прашањата.

Дадено е податочното множество кое вклучува резултати од три тестови на ученици во државно училиште и разновидни лични и социо-економски фактори кои може да имаат ефекти врз нив.

Уренек за задачата е даден на следниот линк.

Според колоните кои датасетот ги содржи, ваша задача е да ја предвидите етничката група (EthnicGroup).

Колку колони од понудените ќе ги искористите за input колони во моделот?
Answer

Колку колони имаат missing values?
Answer

За колоната TestPrep кои non null вредности ги содржи:
Answer
'None', 'complited', 'none', ‘nan’, 'completed', само 'complited'
Ако во колоната TestPrep има вредности кои недостигаат, изберете ги чекорите кои се најсоодветни за пополнување на null вредностите ако користите *KNNImputer*:

1.
Answer
поставување на np.nan како вредност за соодветната енкодирана лабела
нема потреба од ништо, колоната е веќе средена
справување со вредности кои недостигаат директно со fit_transform функцијата
енкодирање на текстуалните податоци

2.
Answer
енкодирање на текстуалните податоци
поставување на np.nan како вредност за соодветната енкодирана лабела
справување со вредности кои недостигаат директно со fit_transform функцијата
нема потреба од ништо, колоната е веќе средена

3.
Answer
енкодирање на текстуалните податоци
нема потреба од ништо, колоната е веќе средена
поставување на np.nan како вредност за соодветната енкодирана лабела
справување со вредности кои недостигаат директно со fit_transform функцијата

Дали има значителна разлика во MathScore-от ако родителот има завршено магистерски или само средно образование (ПОСТАВИ ВИЗУЕЛИЗАЦИЈА ВО ТЕТРАТКАТА - без неа одговорот нема да е земен во предвид):
Не
Да

Каква дистрибуција има MathScore колоната (ПОСТАВИ ВИЗУЕЛИЗАЦИЈА ВО ТЕТРАТКАТА - без неа одговорот нема да е земен во предвид):
Answer
гама распределба
нормална распределба

експоненцијална распределба

Кој моделот ќе го искористите за предвидување на таргет колоната:
Answer
KNN моделлинеарна регресијалогистичка регресија

Кои метрики ќе ги искористите?
F1 Score
Recall Score
Mean Squered Error
R2 Score

4.
Потребно е во прилог да прикачите .ipynb / .py file, (На Colab: File-->Download-->Download .ipynb) од задачата која претходно ја решававте со насловена: Task1_{index}.ipynb / Task1_{index}.py на местото на {index} го поставувате вашиот индекс

НАПОМЕНА: Документите кои не се именувани според правилото нема да бидат прегледани. Shareable линкови до вашиот Colab notebook исто така не се прегледуваат.

5
НАПОМЕНА:
На истата страница со ова прашање подолу ви се наоѓа есејско прашање каде треба да прикачите решение на задачата.
Затоа што испитот е open book се воведени следните правила:
- Во случај да одговорите точно на прашањата поставени подолу, а тоа да го нема во кодот ќе освоите нула поени на прашањата.

Дадено е податочно множество кое содржи податоци од био-сигнали за пациенти кои се пушачи или непушачи. Целта е да се предвиди дали одреден пациент е пушач или непушач (колона smoking). Стартер кодот е даден на следниот линк.

Изградете дрво на одлука (DecisionTree) како модел, при тоа изберете ги најдобрите вредности на следните хиперпараметри за моделот и дадените податоци: criterion, max_depth и min_samples_split. При избор на параметрите користете cross-validation на целото множество со 5 поделби. За најдобар модел изберете го оној што има оптимална вредност на најсоодветната метрика која за таргет променливата

(земете го предвид типот на проблем класификација/регресија и балансираност/небалансираност).

НАПОМЕНА:

ПРОЦЕСОТ НА ИЗБИРАЊЕ НА НАЈДОБРИ ХИПЕР-ПАРАМЕТРИ МОРА ДА ГО ИМАТЕ ВО КОД (НЕ САМО РАЧНО ДА МЕНУВАТЕ ПАРАМЕТРИ И НЕКОЛКУ ПАТИ ДА ЈА ИЗВРШИТЕ КЕЛИЈАТА). МОРА ДА ИСПРОБАТЕ БАРЕМ 3 РАЗЛИЧНИ ВРЕДНОСТИ ЗА СЕКОЈ ОД ХИПЕР-ПАРАМЕТРИТЕ.

Откако ќе го изберете најдобриот модел, направете предвидувања со него на целото множество со cross-validation со 5 поделби, и направете евалуација користејќи ги сите метрики соодветни за множеството и моделот.

Одговорете ги следните прашања за кодот што сте го напишале.

1. Дали податочното множество е балансирано? (ВАЖНО: МОРА ДА ИМАТЕ ДЕЛ ОД КОДОТ ВО ПРИКАЧЕНАТА ТЕТРАТКА КОЈ ГО ИЛУСТРИРА ОВА ПРАШАЊЕ ЗА ДА ВИ СЕ ПРИФАТИ ОДГОВОРОТ)

Да
Не
Прашањето не е валидно за типот на таргет променливата
1. Изберете која метрика е валидна при правење на евалуацијата за најдобар модел според дадениот таргет?

mean_squared_error r2_score f1_score accuracy error_rate

2. Кои вредности може да ги има параметарот criterion за моделот и таргетот кој е даден?

squared_error absolute_error entropy error_rate log_loss f1_score

3. Кои од следните вредности се валидни за параметарот max_depth за дадените податоци?

1000 -1 5 10000 1 10

4. При правење на cross-validation, за секоја комбинација на хипер-параметри моделот се тренира
Answer
5 3 2 7 6 4 пати, и во една итерација се користи/ат
Answer
7 4 6 1 3 2 5 дел/ови од множеството за тренирање на моделот, а
Answer
4 3 1 5 6 2 7 дел/ови за тестирање.

5. Во кои од наведените случаи е корисно да се прави cross-validation?
кога имаме многу податоци
кога имаме малку податоци
кога сакаме да ги процениме перформансите на моделот на целото множество
кога сакаме да определиме најдобри хипер-параметри за моделот
кога имаме зависности или групирање во податоците

6. За ваков тип на таргет променлива соодветно би било да се истренира и некој од следниве модели наместо DecisionTree:
Линеарна регресија
Логистичка регресија
KNN
Ridge регресија

6
Потребно е во прилог да прикачите .ipynb / .py file, (На Colab: File-->Download-->Download .ipynb) од задачата која претходно ја решававте со насловена: Task2_{index}.ipynb / Task2_{index}.py на местото на {index} го поставувате вашиот индекс

НАПОМЕНА: Документите кои не се именувани според правилото нема да бидат прегледани. Shareable линкови до вашиот Colab notebook исто така не се прегледуваат.

НАПОМЕНА:
На истата страница со ова прашање подолу ви се наоѓа есејско прашање каде треба да прикачите решение на задачата.
Затоа што испитот е open book се воведени следните правила:
- Во случај да одговорите точно на прашањата поставени подолу, а тоа да го нема во кодот ќе освоите нула поени на прашањата.

Дадено е податочното множество кое вклучува резултати од три тестови на ученици во државно училиште и разновидни лични и социо-економски фактори кои може да имаат ефекти врз нив.

Уренек за задачата е даден на следниот линк.

Според колоните кои датасетот ги содржи, ваша задача е да ја предвидите етничката група (EthnicGroup).

Колку колони од понудените ќе ги искористите за input колони во моделот? 2

Колку колони имаат missing values? 1

За колоната **TestPrep** кои non null вредности ги содржи: 'none','complited'

Ако во колоната **TestPrep** има вредности кои недостигаат, изберете ги чекорите кои се најсоодветни за пополнување на null вредностите ако користите ***KNNImputer***:

1. поставување на np.nan како вредност за соодветната енкодирана лабела
2. енкодирање на текстуалните податоци
3. поставување на np.nan како вредност за соодветната енкодирана лабела

Дали има значителна разлика во MathScore-от ако родителот има завршено магистерски или само средно образование (**ПОСТАВИ ВИЗУЕЛИЗАЦИЈА ВО ТЕТРАТКАТА - без неа одговорот нема да е земен во предвид**):

- ○ Не
- ○ Да

Каква дистрибуција има MathScore колоната (**ПОСТАВИ ВИЗУЕЛИЗАЦИЈА ВО ТЕТРАТКАТА - без неа одговорот нема да е земен во предвид**):

Кој моделот ќе го искористите за предвидување на таргет колоната:

Кои метрики ќе ја искористите?

- ☐ F1 Score
- ☐ Recall Score
- ☐ Mean Squered Error
- ☐ R2 Score

2

Time left 0:50:1

НАПОМЕНА:

На истата страница со ова прашање подолу ви се наоѓа есејско прашање каде треба да прикачите решение на задачата.

Затоа што испитот е open book се воведени следните правила:

- Во случај да одговорите точно на прашањата поставени подолу, а тоа да го нема во кодот ќе освоите нула поени на прашањата.

Дадено е податочно множество кое содржи податоци од био-сигнали за пациенти кои се пушачи или непушачи. Целта е да се предвиди дали одреден пациент е пушач или непушач (колона **smoking**). Стартер кодот е даден на следниот линк.

Изградете дрво на одлука (**DecisionTree**) како модел, при тоа изберете ги најдобрите вредности на следните хиперпараметри за моделот и дадените податоци: **criterion**, **max_depth** и **min_samples_split**. При избор на параметрите користете **cross-validation** на целото множество со **5** поделби. За најдобар модел изберете го оној што има оптимална вредност на најсоодветната метрика која за таргет променливата (земете го предвид типот на проблем класификација/регресија и балансираност/небалансираност).

НАПОМЕНА:

ПРОЦЕСОТ НА ИЗБИРАЊЕ НА НАЈДОБРИ ХИПЕР-ПАРАМЕТРИ МОРА ДА ГО ИМАТЕ ВО КОД (НЕ САМО РАЧНО ДА МЕНУВАТЕ ПАРАМЕТРИ И НЕКОЛКУ ПАТИ ДА ЈА ИЗВРШИТЕ КЕЛИЈАТА). МОРА ДА ИСПРОБАТЕ БАРЕМ 3 РАЗЛИЧНИ ВРЕДНОСТИ ЗА СЕКОЈ ОД ХИПЕР-ПАРАМЕТРИТЕ.

Откако ќе го изберете најдобриот модел, направете предвидувања со него на целото множество со cross-validation со 5 поделби, и направете евалуација користејќи ги сите метрики соодветни за множеството и моделот.

Одговорете ги следните прашања за кодот што сте го напишале.

1. Дали податочното множество е балансирано? **(ВАЖНО: МОРА ДА ИМАТЕ ДЕЛ ОД КОДОТ ВО ПРИКАЧЕНАТА ТЕТРАТКА КОЈ ГО ИЛУСТРИРА ОВА ПРАШАЊЕ ЗА ДА ВИ СЕ ПРИФАТИ ОДГОВОРОТ)**

- ○ Да
- ○ Не
- ○ Прашањето не е валидно за типот на таргет променливата

1. Изберете која метрика е валидна при правење на евалуацијата за најдобар модел според дадениот таргет?

☐ mean_squared_error ☐ r2_score ☐ f1_score ☐ accuracy ☐ error_rate

2. Кои вредности може да ги има параметарот **criterion** за моделот и таргетот кој е даден?

☐ squared_error ☐ absolute_error ☐ entropy ☐ error_rate ☐ log_loss ☐ f1_score

3. Кои од следните вредности се валидни за параметарот **max_depth** за **дадените податоци**?

☐ 1000 ☐ -1 ☐ 5 ☐ 10000 ☐ 1 ☐ 10

4. При правење на cross-validation, за секоја комбинација на хипер-параметри моделот се тренира [ ] пати, и во една итерација се користи/ат [ ] дел/ови од множеството за тренирање на моделот, а [ ] дел/ови за тестирање.

5. Во кои од наведените случаи е корисно да се прави cross-validation?

- ☐ кога имаме многу податоци
- ☐ кога имаме малку податоци
- ☐ кога сакаме да ги процениме перформансите на моделот на целото множество
- ☐ кога сакаме да определиме најдобри хипер-параметри за моделот
- ☐ кога имаме зависности или групирање во податоците

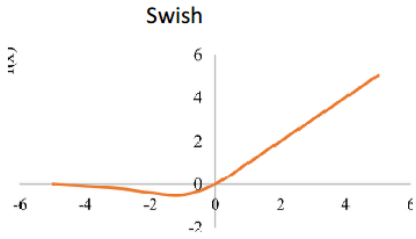
6. За каков тип на таргет променлива соодветно би било да се истренира и некој од следниве модели наместо DecisionTree:

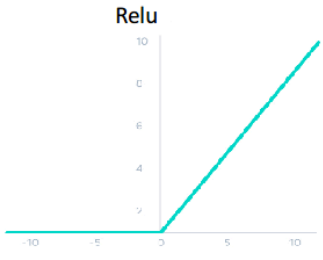
- ☐ Линеарна регресија
- ☐ Логистичка регресија
- ☐ KNN
- ☐ Ridge регресија

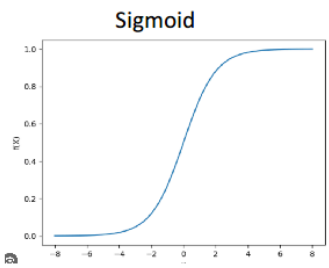
Колоквиум 2

Квалификациски

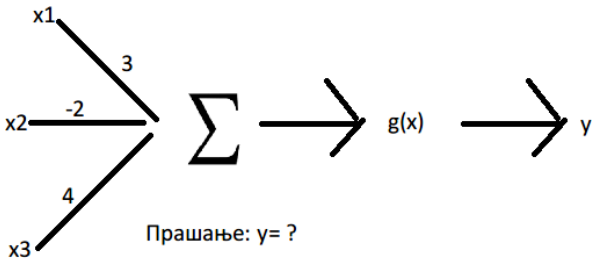
8. Drag and drop





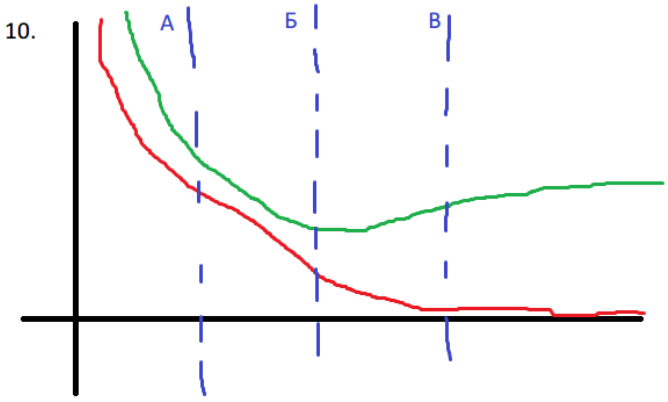


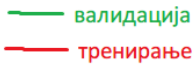
9.



Прашање: y= ?

10.





-Во која точка тренирањето на моделот треба да престане?
-Во која точка има overfitting?
-Во која точка има underfitting?

1.

2. Нека е дадена реченицата: “The government debt problems turned into banking crises as happened in 2009.” Skip-gram со големина на прозорец три за зборот banking е:

O: problems turn into crises as happened

3. Во кој случај би било најдобро да се употреби Sigmoid како излезно ниво кај невронските мрежи?

О: бинарна класификација

4. На кои од наведените модели за кластирање потребно е да се наведе бројот на кластери?
О: K-means clustering

5. Ако имаме случај со купувачи на Зара, кои се од различен тип, но имаат исти патерни кој вид на кластерирање е најсоодветен? - DBScan | HDBScan | KNN

6. Positional embedding за прв пат се појавува кај (LSTM | XGBoost | Transformer), тие служат за означување позиции на (податоци независни од тип | временски серии | зборови)

7. Еден од најдобрите јазични модели BERT се потпира на трансформер архитектура. Кој дел од трансформер архитектура се користи во BERT?

8. Избери што е точно за bagging и boosting
-bagging прави усреднување при класификација
-bagging прави усреднување при регресија
-кај bagging се градат многу независни модели врз делови од множеството
-XGBoost користи gradient descent
-Gradient boosting алгоритмите конвергираат кога ламбда=1
-кај boosting секој наредени модел ги користи грешките на претходниот

Дополнително - теорија

1. Имате задача да определите дали цената на акциите ќе расте, дадени се колоните Open, Close, Volume, High, Low. За каков проблем станува збор? Дали ќе користите дрво на одлука или случајна шума како алгоритам за градење на моделот? Објаснете го изборот на алгоритмот што го направивте!

2. Дадено е податочно множество за 50000 различни апликации за кредит. За секоја кредитна апликација се чуваат вредности за 34 карактеристики Одговорете на следните прашања: За каков проблем станува збор? Дали ќе користите дрво на одлука или случајна шума како алгоритам за градење на моделот? Објаснете го изборот на алгоритмот што го направивте!

3. Правите модел за класификација на вино според квалитетот. За секое вино се чуваат вредности за 13 карактеристики. Одговорете на следните прашања:За каков проблем станува збор?Дали ќе користите дрво на одлука или случајна шума како алгоритам за градење на моделот? Објаснете го изборот на алгоритмот што го направивте!

4. Имате задача да направите класификатор на вести што се однесуваат на различни медицински области. Дали во овој случај би ги користеле стандардните Word2Vec embeddings или не? Објаснете зошто! Како пример во објаснувањето може да ја користите следната реченица The patient exhibited symptoms of dyspnea and tachycardia.

Дополнително - задачи

1.