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?

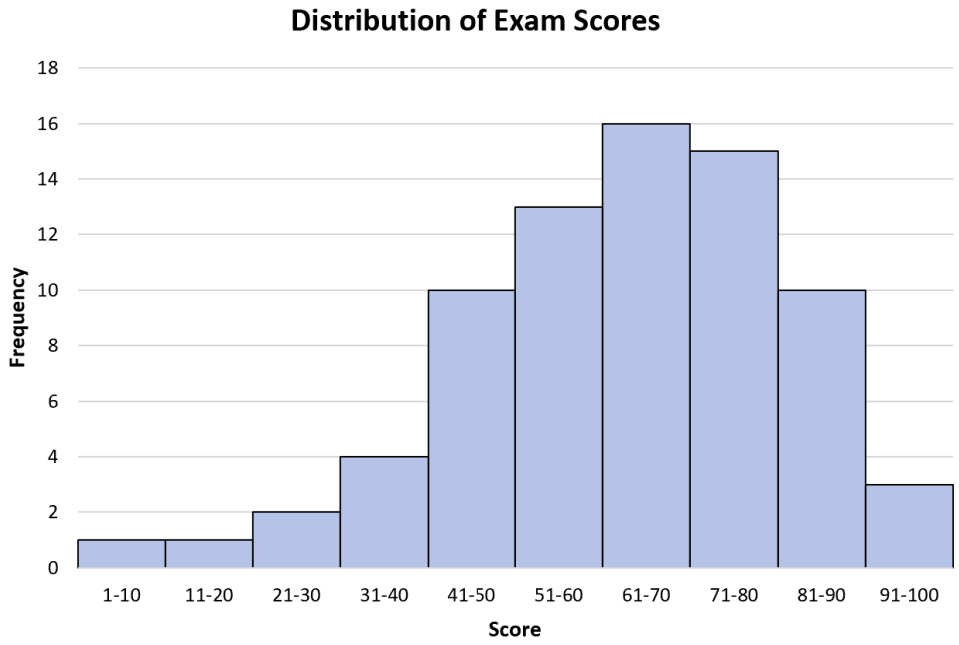
X	Y
Не се сеќавам што имаше тука	Gold
Не се сеќавам што имаше тука	Silver
Не се сеќавам што имаше тука	Gold,Silver

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

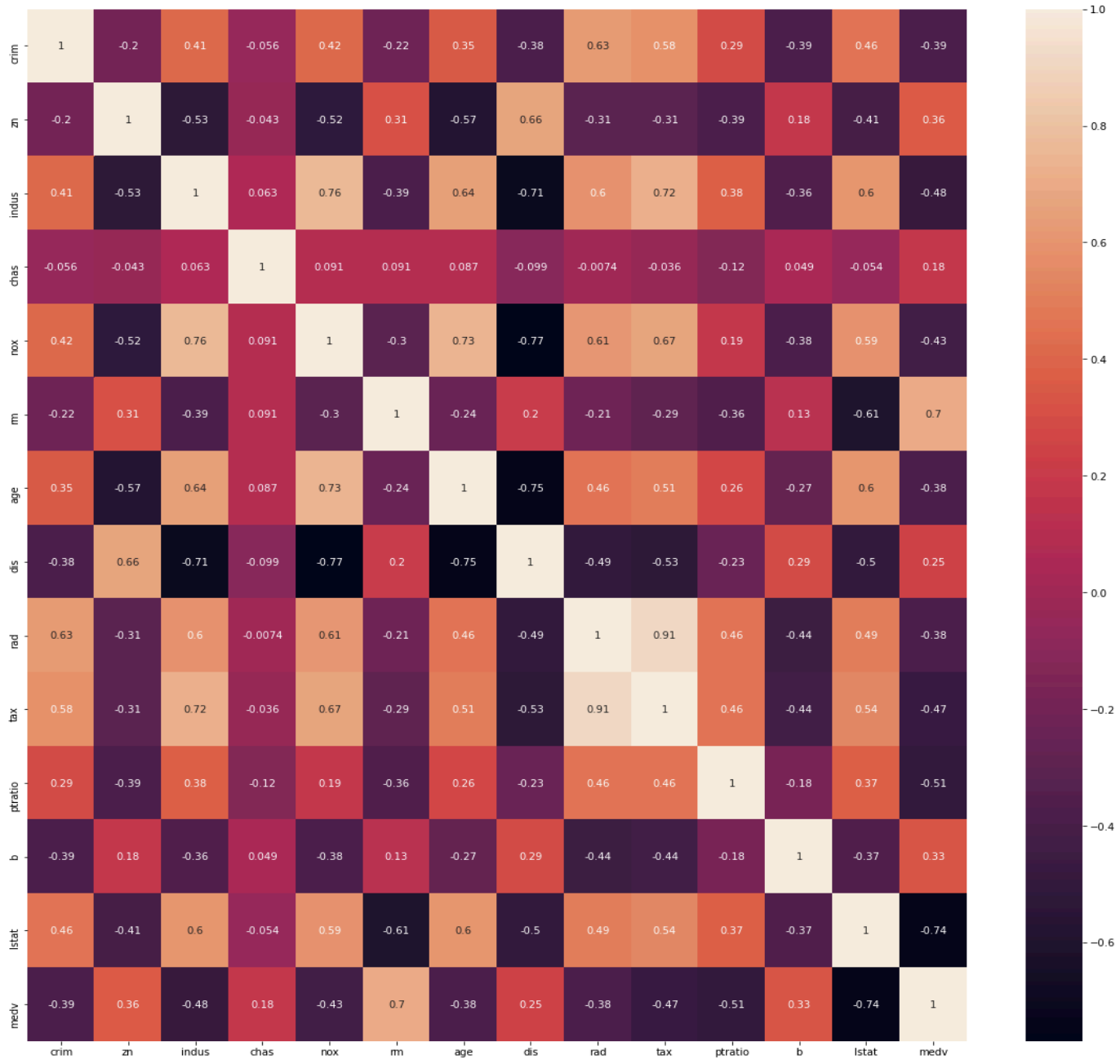
Одговор: a.

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

- ☑ Колоната medv има висока корелација со сите останати колони (дискутабилно)
- ☑ 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



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

a) overfitting

b) underfitting

c) just right

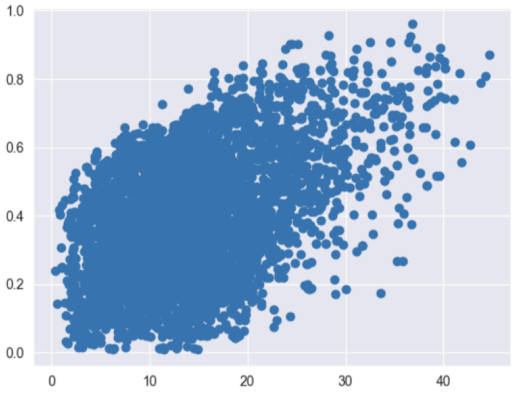
Одговор: a.

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

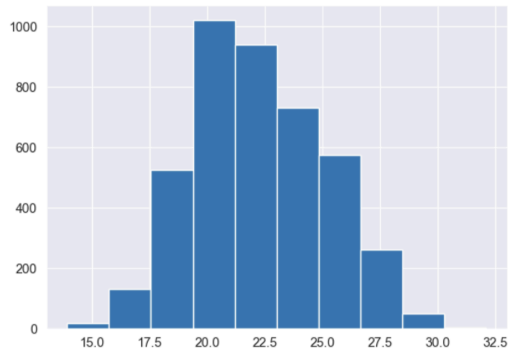
Одговор: a. (А кога тежнее кон 1 податоците се несредени.)

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

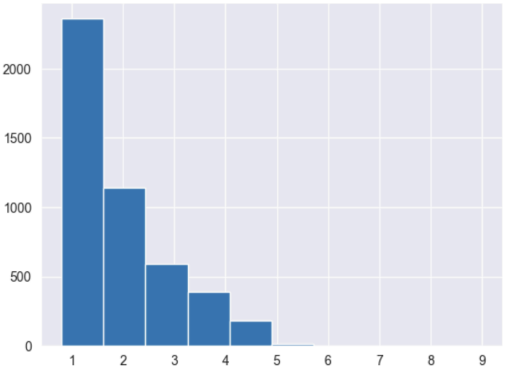
a)

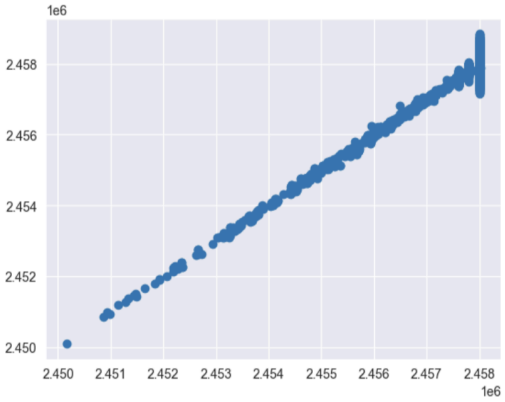


b)



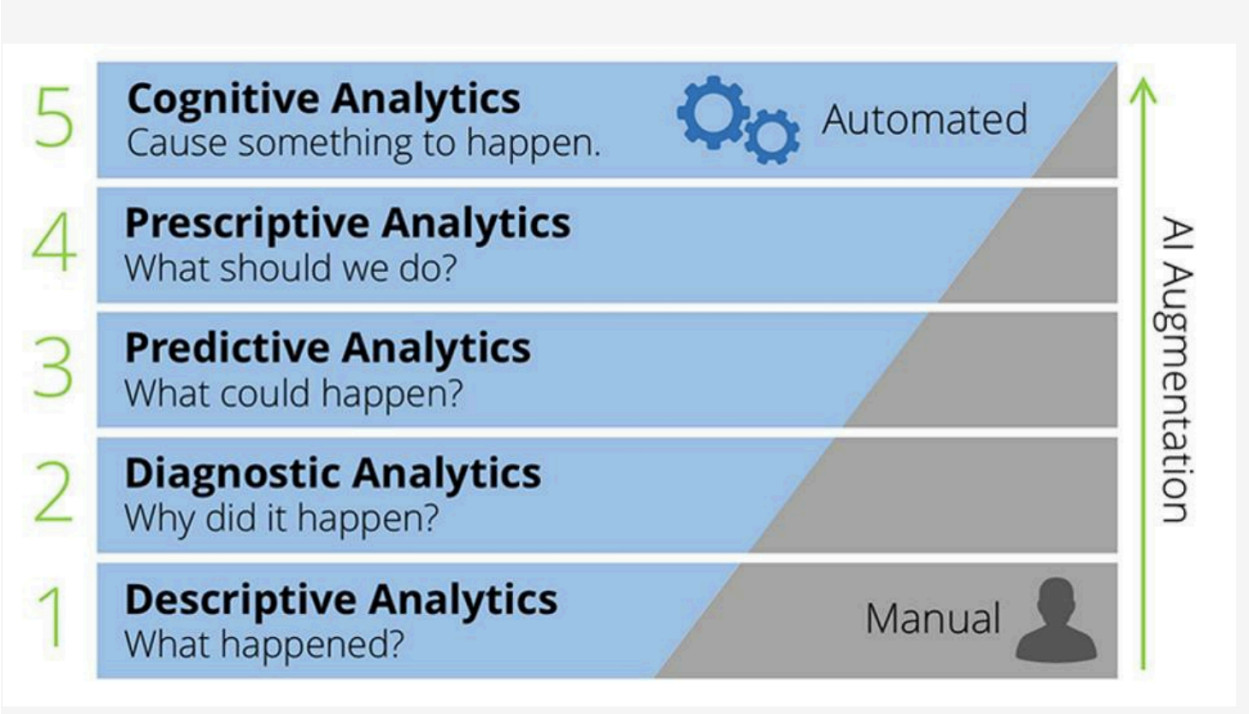
c)



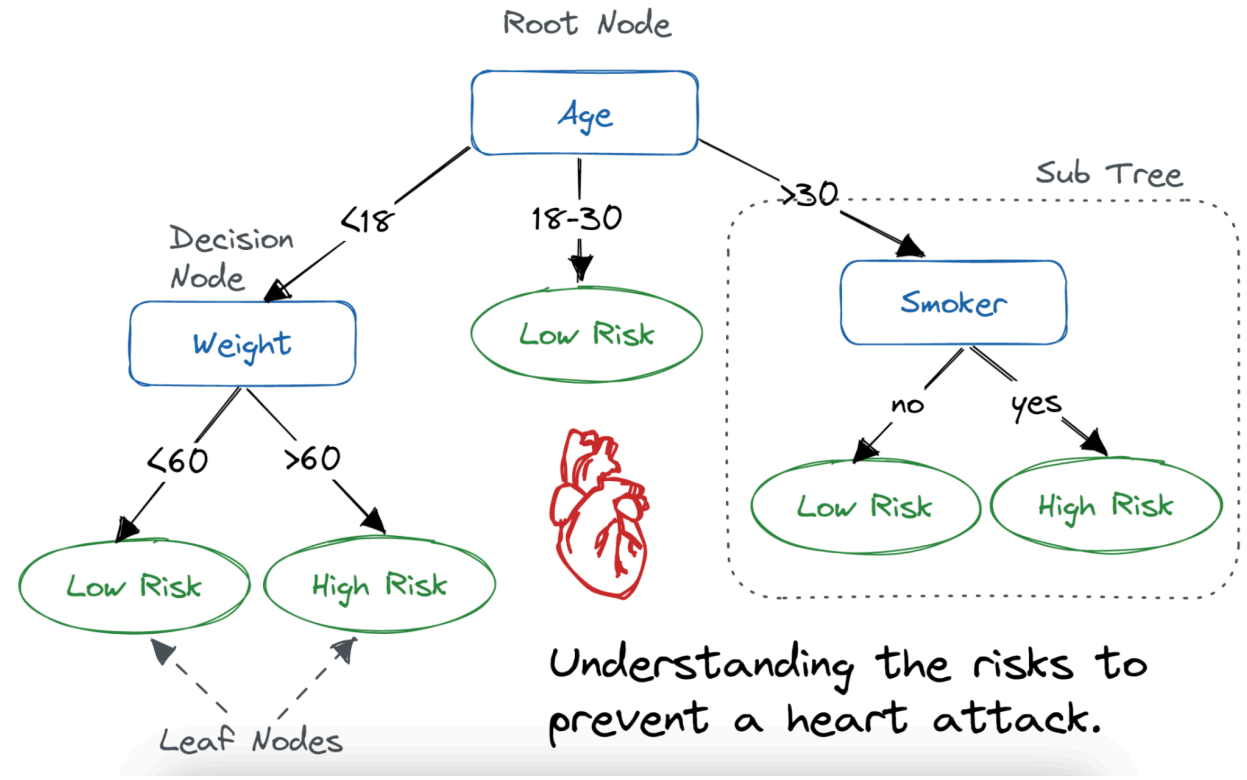


11.Да се подредат видовите на Analytics од 1 до 5, од долу нагоре, според AI Augmentation:
Descriptive Analytics, Cognitive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics.

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	0	1

5	1	//////////////////	//////////////////	0	0
null	0	imputacija	//////////////////	1	0
4	null	//////////////////	imputacija	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. 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.

4. 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.

5. 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