Лабораторна робота №2

Використання модуля Pandas.

Аналіз даних по серцево-судинних захворюваннях

Мета роботи: набути практичних навичок проведення розвідувального аналізу даних.

Зміст роботи

Завдання 1. Ретельно опрацювати теоретичні відомості:

Лекція 1 і Лекція 2 https - //learn.ztu.edu.ua/course/view.php?id=2303

Intro to Programming - https://www.kaggle.com/learn/intro-to-programming

Python Groupby Tutorial - https://www.kaggle.com/crawford/python-groupby-tutorial

***Завдання* 2.** Провести розвідувальний аналіз даних EDA (Exploratory data analysis) на наборі даних mlbootcamp_train_Soroka.csv

Опис набору даних.

Dataset сформований з реальних даних, і в ньому використовуються ознаки, які можна розбити на 3 групи:

Об'єктивні ознаки:

- Вік (age)
- Зріст (height)
- Вага (weight)
- Пол (gender)

Результати вимірювання:

- Артеріальний тиск верхній і нижній (ap_hi, ap_lo)
- Холестерин (cholesterol)
- Глюкоза (gluc)

Суб'єктивні ознаки (зі слів пацієнта):

- Куріння (smoke)
- Вживання алкоголю (alco)
- Фізична активність (active)

Значення показників холестерину і глюкози представлені одним з трьох класів: норма, вище норми, значно вище норми. Значення суб'єктивних ознак - бінарні.

Всі показники отримані на момент огляду.

Провести первинний аналіз набору даних.

- З бібліотек знадобляться тільки NumPy і Pandas.
- Зчитаємо дані з csv-файлу в об'єкт pandas DataFrame.

```
df = pd.read_csv('mlbootcamp5_train.csv', sep=';',index_col='id')
```

– Подивитися перші 10 записів.

– Використайте метод *describe()* для визначення основних статистичних характеристик.

– Використайте прискорений розвідувальний аналіз даних з використанням бібліотеки *pandas-profiling*.

Для встановлення модуля скористайтеся наступним кодом:

```
! pip install https://github.com/ydataai/ydata-profiling/archive/refs/heads/master.zip
```

Далі потрібно перегрузити ядро. Потім імпортувати бібліотеки:

```
from pandas_profiling import  ProfileReport
import pandas_profiling
```

Виконати профілювання дата сету.

pandas_profiling.ProfileReport(df)

Звіт можна експортувати в інтерактивний HTML файл:

```
profile = pandas_profiling.ProfileReport(df)
profile.to_file(outputfile="AAA data profiling.html")
```

Завдання 3. Отримати відповіді на наступні питання:

Кожну відповідь необхідно проілюструвати фрагментами програмного коду, що відповідають на наступні питання.

Питання 1. Скільки чоловіків і жінок представлено в наборі даних? Не було дано опису ознаки «стать» (якої статі відповідає 1, а якої - 2 в ознаці *gender*) - це можна визначити подивившись на зріст, при розумному припущенні в середньому чоловіки вище.

Питання 2. Хто в середньому рідше вказує, що вживає алкоголь - чоловіки чи жінки?

Питання 3. У скільки разів (округлити, *round*) відсоток курців серед чоловіків більше, ніж відсоток курців серед жинок?

Питання 4. У кого в середньому тиск вище, у жінок чи чоловіків.

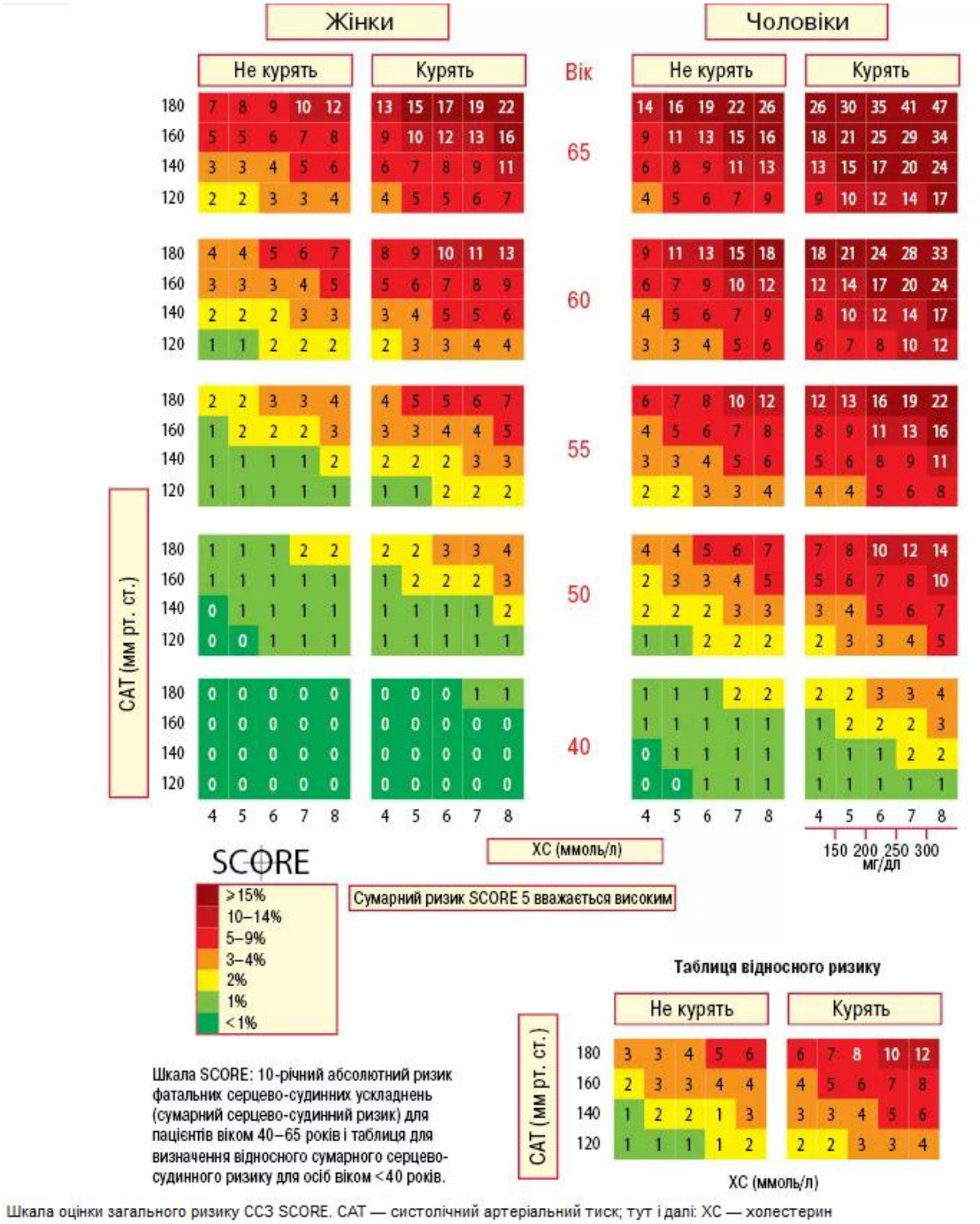
Питання 5. В чому вимірюється вік? На скільки місяців (приблизно) відрізняються медіанне значення віку курців і тих хто не курить?

Питання 6. У статті Wikipedia про серцево-судинний ризик показана шкала SCORE для розрахунку ризику смерті від серцево-судинного захворювання в найближчі 10 років. На рис.1. представлена оцінка ризику серцево-судинних захворювань.

SCORE - це абревіатура англійських слів «систематична оцінка коронарного ризику», тобто ризику захворювань серця і судин. Ця шкала була запропонована групою експертів Європейського товариства кардіологів

у 2003 р. і розроблена на підставі результатів досліджень, проведених в 12 європейських країнах із загальною кількістю пацієнтів понад 205 тисяч.

Шкала - це система квадратів, в якій застосовано принцип світлофора – три основні кольори: зелений – це низький ризик, що відповідає 1% або менше, жовтий колір – увага! ризик помірний і коливається в межах 2–4%, червоний колір – небезпека! 5% і більше. Для більшої диференціації застосовані відповідні відтінки цих трьох основних кольорів.



Шкала оцінки загального ризику ССЗ SCORE. САТ — систолічний артеріальний тиск; тут і далі: ХС — холестерин

Рис.1. Оцінка ризику серцево-судинних захворювань

Правий верхній прямокутник відображає сегмент чоловіків, які палять у віці від 60 до 64 років включно. (Неочевидно, але тут для віку і тиску цифри означають верхню межу, і вона не включається).

Бачимо 9-ку в лівому нижньому кутку цього прямокутника і 47 - в правому верхньому. Тобто якщо при цьому систолічний (тобто верхній) артеріальний тиск - менше 120 мм рт.ст., а рівень холестерину - 4 ммоль/л, то ризик ССЗ оцінюється приблизно в 5 разів нижче, ніж якби тиск знаходився в інтервалі [160, 180), а холестерину було б 8 ммоль/л.

Порахуємо аналогічне значення на заданих даних.

Уточнення:

– Створіть нову ознаку *age_years* - вік в роках, округливши до цілих (*round*). Для даного прикладу відберіть чоловіків від 60 до 64 років включно з кращими показниками.

– Категорії рівня холестерину на малюнку і в даних відрізняються. Відображення значень на зображенні в значення ознаки *cholesterol* наступне: 4 ммоль/л →→ 1, 5-7 ммоль/л →→ 2, 8 ммоль/л →→ 3.

– Цікавлять 2 підгрупи чоловіків, які палять вік від 60 до 64 років включно: перша з верхнім артеріальним тиском строго менше 120 мм рт.ст. і концентрацією холестерину - 4 ммоль/л, а друга - з верхнім артеріальним тиском від 160 (включно) до 180 мм рт.ст. (Не включно) і концентрацією холестерину - 8 ммоль/л.

У скільки разів (*round*) відрізняються частки хворих людей (відповідно до цільової ознаки, *cardio*) в цих двох підвибірках?

Питання 7. Побудуйте нову ознаку - BMI (*Body Mass Index*). Для цього треба вагу у кілограмах поділити на квадрат зросту в метрах. Нормальними вважаються значення BMI від 18.5 до 25. Виберіть вірні твердження.

Твердження:

– Медіанний BMI по вибірці перевищує норму.
– У жінок в середньому BMI нижче, ніж у чоловіків.
– У здорових в середньому BMI вище, ніж у хворих.
– У хворих в середньому BMI вище, ніж у здорових.

Питання 8. Можна помітити, що дані не чисті, багато в них «бруду» і неточностей. Краще це можна побачити на візуалізації даних.

Відфільтруйте наступні сегменти пацієнтів (вважаємо помилками в даних):

– вказане нижнє значення артеріального тиску строго вище верхнього;

– зріст строго менше 2.5% - перцентілі або строго більше 97.5% - перцентілі (використовуйте *pd.Series.quantile*)

– вага строго менше 2.5% - перцентілі або строго більше 97.5% - перцентілі

Це не вся чистка даних, яку можна було виконати, але поки можна зупинимося на цьому.

Скільки відсотків даних (round) було видкінуто?

Питання 9. Скільки чоловіків і жінок мають зайву вагу?

BMI в межах [25:29.9]- передожиріння, в межах [30:34.9] – 1 ступінь ожиріння, в межах [35:39.9] – 2 ступінь ожиріння, >40 – 3 ступінь ожиріння,

Методичні рекомендації

Python - відмінна мова для аналізу даних і в першу чергу завдяки фантастичній екосистемі пакетів, орієнтованих на дані. Pandas є одним з таких пакетів і значно спрощує імпорт і аналіз даних.

Функція Pandas *dataframe.groupby()* використовується для поділу даних на групи на основі деяких критеріїв.

Приклад: Дано дві марки авто і споживання CO2 для кожної марки автомобіля.

```
data = {
  'co2': [95, 90, 99, 104, 105, 94, 99, 104],
  'model': ['Citigo', 'Fabia', 'Fiesta', 'Rapid', 'Focus', 'Mondeo', 'Octavia
', 'B-Max'],
  'car': ['Skoda', 'Skoda', 'Ford', 'Skoda', 'Ford', 'Ford', 'Skoda', 'Ford']
}

df = pd.DataFrame(data)
df
```

Знайти середнє споживання CO2 для кожної марки автомобіля:

```
print(df.groupby(["car"]).mean())
```

Визначити кількість авто представлених в масиві:

```
print(df.groupby(["car"]).count())
```