

Дата - аналітика та прийняття рішень на основі даних

01

Виконав: Оверченко Костягин

Група: КН-20

Керівник: Вельмагіна Н.О.

GENERAL PROVISIONS

Мета

Проведення дослідження серцевих нападів та порівняння інструментів аналізу даних.

Об'єкт дослідження

Набір даних серцевих нападів.

Задачі дослідження

1. Огляд інструментів для аналізу та управління даними.
2. Аналіз набору даних серцевих нападів.
3. Порівняння інструментів R та Python для виконання аналітичних задач.

Предмет дослідження

Інструменти аналізу даних (системи управління базами даних, BI-інструменти, мови програмування Python та R) та їх застосування для аналізу.

02

02

Огляд предметної
області

≡

Застосування аналізу даних у
 медичному секторі

Огляд медичних тем

Теоретичні відомості про етапи
 дослідження

МЕДИЧНІ ТЕРМІНИ

Гострий інфаркт міокарда

Гострий інфаркт міокарда (ГІМ), який зазвичай називають серцевим нападом. ГІМ виникає, коли переривається кровообіг або приплив крові до серцевого м'яза, внаслідок чого серцевий м'яз пошкоджується або гине (стає некротизованим).

ЕКГ

Графічний запис змін електричних потенціалів, які виникають внаслідок збудження серцевого м'яза.

Стенокардія

Напади болю стискаючого характеру, в ділянці серця і за грудиною, які можуть передаватися в ліву руку, лопатку, шию

Тахікардія

Патологічний стан, у якому відбувається зростання частоти скорочень серця (ЧСС) вище 100 ударів на хвилину.

Тиск

Нормальний артеріальний тиск – це тиск в діапазоні 120-129 мм рт. ст. (систолічний)

04

Таласемія

Генетична хвороба крові, коли через мутацію генів, утворюється недостатня кількість гемоглобіну в організмі і відбувається деформація еритроцитів.

ЕТАПИ ДОСЛІДЖЕННЯ

Explore (Дослідження)

Попереднє знайомство з даними, вивчення їхньої структури та змісту. Визначення основних характеристик набору даних: кількість рядків, стовпців, типи даних, наявність пропусків.

Profile (Профілювання)

Детальний аналіз розподілу значень, виявлення викидів, аномалій та закономірностей. Візуалізація даних для кращого розуміння їхніх особливостей.

Clean (Очищення)

Обробка відсутніх значень, видалення або корекція некоректних даних. Стандартизація форматів даних та усунення дублікатів.

Shape (Формування)

Перетворення даних у форму, придатну для подальшого аналізу: агрегація, нормалізація, кодування категоріальних змінних. Створення нових ознак та відбір найбільш інформативних.

Analyse (Аналіз)

Застосування статистичних методів та алгоритмів машинного навчання для виявлення закономірностей. Тестування гіпотез, побудова моделей, інтерпретація результатів та формулювання висновків.

05

АНАЛІТИКА ДАНИХ

- **Хто такий Аналітик**
- **Табличні процесори**
- **Інструменти візуалізації**
- **Інструменти роботи з базами даних**
- **Статистичне програмне забезпечення**

АНАЛІТИКА

Аналітик - це поняття більш широке, ніж просто експерт в якісь галузі знань, його інтелектуальний інструментарій і досвід практичної діяльності набагато ширший і не обмежується однією предметною сферою.

Табличні процесори



Бази даних



Інструменти візуалізації



Статистичні пакети



ПРАКТИЧНА ЧАСТИНА

Формулювання задач

Аналіз даних

Тестування гіпотези

Моделювання

ФОРМУЛУЮВАННЯ ЗАДАЧ



Ідентифікація цільової аудиторії

Компанія прагне отримати уявлення про демографічні характеристики потенційної цільової аудиторії для цього смарт-годинника. Аналізуючи такі змінні, як вік і стать, компанія прагне адаптувати свої маркетингові та рекламні кампанії, щоб ефективно охопити найбільш релевантні сегменти споживачів. Такий цілеспрямований підхід гарантує, що продукт досягне тих, хто може отримати найбільшу користь від його можливостей.

Прогностичне моделювання для оцінки ризику серцевого нападу

Використовуючи можливості прогностичного моделювання на основі даних, компанія має на меті розробити надійний алгоритм, який може із певною точністю прогнозувати ризик серцевого нападу для людини. Ця прогностична модель буде інтегрована в смарт-годинник, що дозволить здійснювати моніторинг в режимі реального часу і своєчасно сповіщати користувачів, які знаходяться в групі підвищеного ризику.

12

ВИСНОВКИ АНАЛІЗУ ТА EDA

1.

У наборі даних більша частка пацієнтів чоловічої статі порівняно з пацієнтками.

2.

Для обох статей найпоширенішим результатом електрокардіографії є «Норма» (значення 0). Чоловіки мають дещо вищий відсоток аномалій хвилі ST-T (значення 1) порівняно з жінками.

4.

Найпоширенішим типом болю в грудях як для чоловіків, так і для жінок є типова стенокардія (значення 1). Жінки мають дещо вищий відсоток безсимптомних випадків (значення 4) порівняно з чоловіками.

5.

Для обох статей найпоширенішим результатом таласемії є значення 3, яке може представляти певну категорію або діапазон. Чоловіки мають дещо вищу частку значень 2 і 3 порівняно з жінками.

7.

Більшість пацієнтів як чоловічої, так і жіночої статі не страждають від стенокардії, викликаної фізичним навантаженням (значення 0).

3.

Більшість пацієнтів як чоловічої, так і жіночої статі не мають підвищеного рівня цукру в крові натще (значення 0).

6.

Більшість пацієнтів як чоловічої, так і жіночої статі мають 0 або 1 уражену магістральну судину. Чоловіки, як правило, мають дещо вищий відсоток випадків з ураженням 2 або 3 магістральних судин порівняно з жінками.

ВИСНОВКИ КОРЕЛЯЦІЙНОГО АНАЛІЗУ

ПОЗИТИВНА КОРЕЛЯЦІЯ

Такі змінні, як `cp` (тип болю в грудях), `thalachh` (максимальна частота серцевих скорочень), `restecg` (результати електрокардіограми у стані спокою), `s1p` (нахил) та `output` мають позитивну кореляцію.

Позитивна кореляція свідчить про те, що зі збільшенням значення предикторної змінної ймовірність серцевого нападу (вихід = 1) також має тенденцію до зростання.

Наприклад, позитивна кореляція між `cp` і результатом (0,418) вказує на те, що особи з більш вираженими типами болю в грудях (вищі значення `cp`) мають більший ризик серцевого нападу.

НЕГАТИВНА КОРЕЛЯЦІЯ

Такі змінні, як вік, стать, `trtbps` (артеріальний тиск у стані спокою), `chol` (рівень холестерину), `exng` (стенокардія, викликана фізичним навантаженням), `oldpeak` (депресія сегмента ST), `caa` (кількість магістральних судин) і `thall` (талесемія) мають негативну кореляцію з вихідною змінною.

Негативна кореляція свідчить про те, що зі збільшенням значення предикторної змінної ймовірність серцевого нападу (вихід = 1) має тенденцію до зменшення.

Наприклад, більша кількість основних судин, забарвлених під час флюорографії, може свідчити про більш здорову серцево-судинну систему.

ТЕСТУВАННЯ ГІПОТЕЗИ

Тест пропорцій для двох вибірок (two-sample proportion test)

Непараметричний тест, який порівнює пропорції (або ймовірності) бінарного результату між двома незалежними групами.

Нульова гіпотеза H_0

Пропорції одинакові.

$$H_0: pA = pB$$

Альтернативна гіпотеза H_1

Імовірність серцевого нападу у чоловіків та жінок є різною.

$$H_1: pA \neq pB.$$

2-sample test for equality of proportions with continuity correction

```
data: c(sum(females$output), sum(males$output)) out of c(nrow(females), nrow(males))
X-squared = 27.015, df = 1, p-value = 2.019e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 0.2246751 0.4607521
sample estimates:
 prop 1   prop 2 
0.8000000 0.4572864
```

Висновок

Результати тесту свідчать про те, що нульова гіпотеза про рівність пропорцій між чоловіками та жінками може бути відхиlena. Пропорція жінок з ризиком серцевого нападу (0.8) значно вища, ніж пропорція чоловіків (0.45).

РОЗПОДІЛ БЮДЖЕТУ ДЛЯ ЗАПУСКУ РЕКЛАМНОЇ КАМПАНІЇ

Розподіл бюджету

Чоловіки: $(prop2) / (prop1 + prop2) * Total Budget = 0.45/(0.8+0.45)* Total Budget = 0.36 * Total Budget$

Жінки: $(prop1) / (prop1 + prop2) * Total Budget = 0.8/(0.8+0.45)* Total Budget = 0.64 * Total Budget$

Чоловіки: $0.36 * 10\,000 = 3600$

Жінки: $0.64 * 10\,000 = 6400$

Налаштування таргетингу

Вікову категорію обираємо найширшу, спираючись на гістограму розподілення. Таким чином початкові налаштування таргетингу для аудиторії можна виставити від 40 до 65 років.



МОДЕЛЮВАННЯ

Фінальна модель

```
logit = 6.610980 - 0.016553 * age - 1.568152 * sex + 0.866726 * cp - 0.018966 *  
trtbps - 0.009528 * chol - 0.578602 * fbs + 0.382247 * restecg + 0.016547 *  
thalachh - 1.035652 * exng - 0.442631 * oldpeak + 0.621918 * slp - 0.776689 *  
caa - 1.193826 * thall
```

Чутливість

R: 0,8

Python: 0,96

Точність

R: 0,82

Python: 0,8

Специфічність

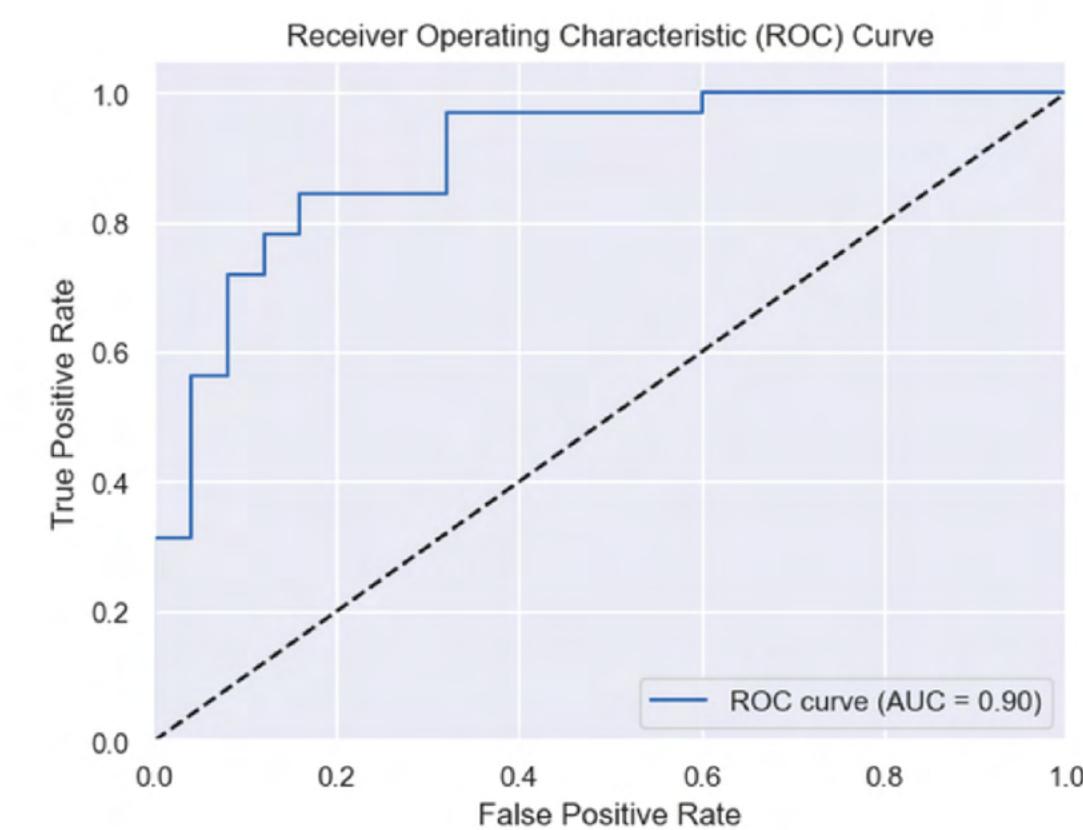
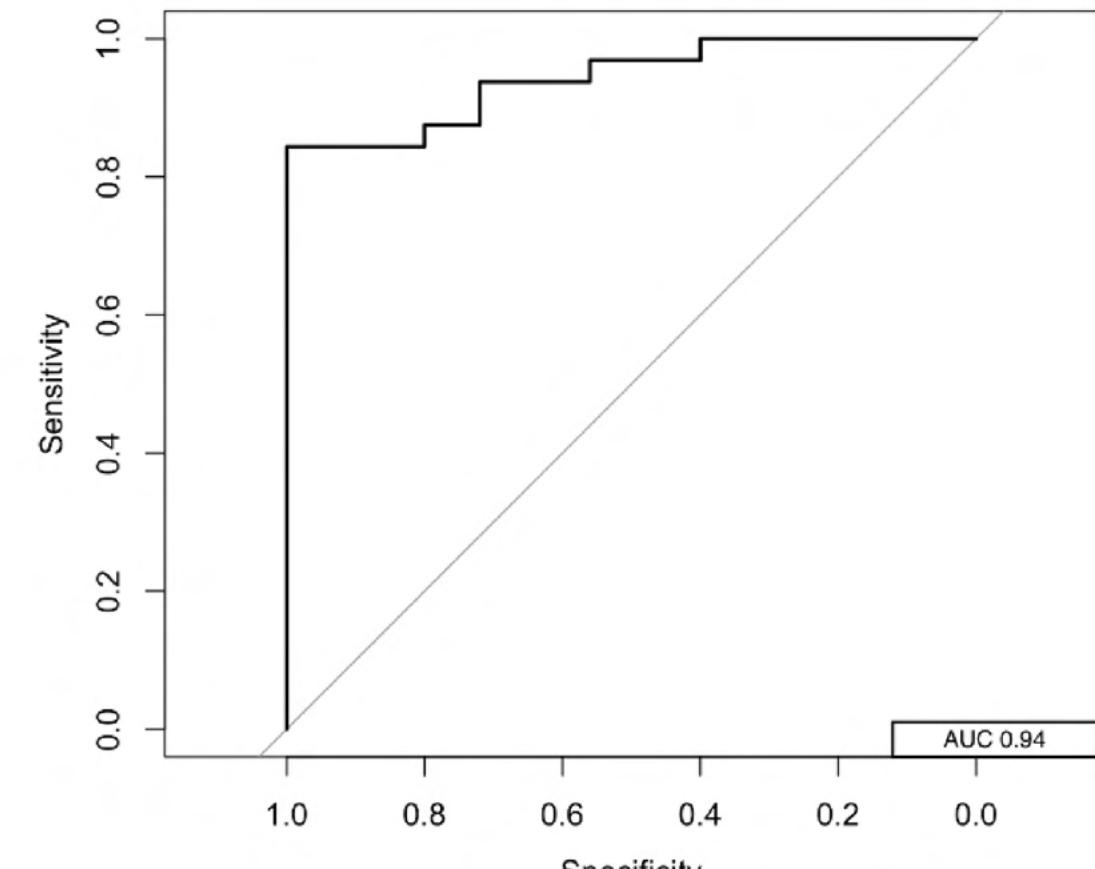
R: 0,84

Python: 0,6

AUC

R: 0,94

Python: 0,9



Висновки

1.

Аналітика великих даних корисна для прогнозування серцевих нападів та управління лікуванням серцево-судинних захворювань.

2.

Дослідження визначило потенційну цільову аудиторію та розробило модель прогнозування ризику серцевого нападу для медичної компанії, що планує створити смартгодинник.

3.

Результати показали вищу пропорцію жінок з ризиком серцевого нападу та визначили цільову вікову категорію 40-65 років.

4.

Розроблена модель машинного навчання демонструє високу точність (82,46%) та відмінну здатність розрізняти пацієнтів за ймовірністю серцевого нападу (AUC 0,9388).

5.

Проект продемонстрував можливості аналітичних інструментів R та Python. Обидва інструменти мають як переваги, так і недоліки. Проте R підходить для наукових досліджень, а Python більш універсальний інструмент для застосування та інтеграції з іншими середовищами.

Дякую за увагу!