Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Інститут прикладного системного аналізу
Кафедра математичних методів системного аналізу

ЗВІТ

про виконання лабораторної роботи №1

з дисципліни «Інтелектуальний аналіз даних»

Виконала:

Студентка III курсу

Групи КА-75

Крива А. О.

Перевірила:

Недашківська Н.І.

Київ – 2020

Завдання

Дано масив T, який складається з N рядків, які відповідають прикладам, і m стовпчиків, які відповідають ознакам. Відомо, що ознака x_h приймає значення з множини $\{c_{h1}, c_{h2}, ..., c_{hq_h}\}$. Дано вектор y розмірності N, елементи якого приймають значення з множини $S = \{s_1, s_2, ..., s_v\}$ (мітки класів для прикладів). Знайти ознаку x_h^* та значення цієї ознаки c_{hi}^*:

$$c_{hi}^* = \arg\min_{h,i} Er(h,i),$$

$$Er(h,i) = \frac{p_3(y \neq s_j^* | x_h = c_{hi})}{p_1(x_h = c_{hi})},$$

$$s_j^* = \arg\max_j p_2(y = s_j | x_h = c_{hi}),$$

де $p_1(x_h = c_{hi})$ - кількість прикладів, для яких ознака x_h приймає значення c_{hi}, $p_2(y = s_j | x_h = c_{hi})$ - кількість прикладів, які належать класу s_j і ознака x_h приймає значення c_{hi}, s_j^* - найбільш імовірний клас за умови що ознака x_h приймає значення c_{hi}.

Дані

Number of Instances: 31
Number of Attributes: 4 (all nominal)

Attribute Information:
3 Classes

1 : the patient should be fitted with hard contact lenses,
2 : the patient should be fitted with soft contact lenses,
3 : the patient should not be fitted with contact lenses.

1. age of the patient: (1) young, (2) pre-presbyopic, (3) presbyopic
2. spectacle prescription: (1) myope, (2) hypermetrope
3. astigmatic: (1) no, (2) yes
4. tear production rate: (1) reduced, (2) normal

Текст програми

```
T = pd.read_csv('data.csv', names = ['age', 'prescription', 'astigmatic',
'rate', 'class'])
y = pd.read_csv('data.csv', names=['class'], usecols=[5])

s = np.sort(y['class'].unique())
errors = {}

for f in T.columns[:-1]:
  inner = {}
  for val in np.sort(T[f].unique()):
    p2 = [T[(T[f] == val) & (T['class'] == i)].count()[0] for i in s]
    s_argmax = p2.index(max(p2))
    p3 = sum(p2) - max(p2)
    p1 = T[T[f] == val].count()[f]
    inner[val] = p3 / p1
  min_value, min_key = min((value, key) for key, value in inner.items())
  errors[(f, min_key)] = min_value

min(errors, key=errors.get)
```

Результат роботи програми

```
Out[166]: ('age', 1)
```

Ознака - це вік людини. Значення цієї ознаки - 1 (молода людина).

Висновок

Під час виконання даної лабораторної роботи я вивчила, як працювати з бібліотеками Numpy, Pandas та Matploplib. У роботі було використано булеве маскування, сортування та агрегуючі функції. Розробила декілька варіантів рішення завдання та обрала найоптимальніший. Результат задовольняє умову задачі.