Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»
(СибГУТИ)

Кафедра прикладной математики и кибернетики

Отчёт
по лабораторной работе № 2 «Решающие деревья»

Выполнил:

студент группы ИП-013

Копытина Татьяна Алексеевна
ФИО студента

Работу проверил: Дементьева Кристина Игоревна
ФИО преподавателя

Новосибирск 2023 г.

Задание

Данная работа носит творческий характер и призвана показать, насколько студент подготовлен к реальному применению полученных знаний на практике. Как известно, в реальной работе никаких вводных данных не предоставляется, тем не менее, мы слегка пренебрегли данным правилом и предоставили теорию и предпочтительный метод для применения.

В приложенном файле (heart_data.csv) располагаются реальные данные по сердечной заболеваемости, собранные различными медицинскими учреждениями. Каждый человек представлен 13-ю характеристиками и полем goal, которое показывает наличие болезни сердца, поле принимает значение 0 или 1 (0 – нет болезни, 1 - есть). Символ ‘?’ в каком-либо поле означает, что для конкретного человека отсутствуют данные в этом поле (либо не производились замеры, либо не записывались в базу).

Требуется имеющиеся данные разбить на обучающую и тестовую выборки в процентном соотношении 70 к 30. После чего по обучающей выборке необходимо построить решающее дерево. Для построения дерева можно пользоваться любыми существующими средствами. Кроме того, для построения дерева необходимо будет решить задачу выделения информативных решающих правил относительно имеющихся числовых признаков.

Разрешается использовать уже реализованные решающие деревья из известных библиотек (например, scikit-learn для Python), либо реализовывать алгоритм построения дерева самостоятельно (все необходимые алгоритмы представлены в теории по ссылке).

В качестве результата работы необходимо сделать не менее 10 случайных разбиений исходных данных на обучающую и тестовую выборки, для каждой построить дерево и протестировать, после чего построить таблицу, в которой указать процент правильно классифицированных данных. Полученную таблицу необходимо включить в отчёт по лабораторной работе.

Код программы

```
import pandas as pd

from sklearn import tree

from sklearn.model_selection import train_test_split

import matplotlib.pyplot as plt

import random


column_names = ["age", "sex", "cp", "trestbps", "chol", "fbs", "restecg", "thalach", "exang", "oldpeak", "slope", "ca", "thal", "target"]

clf = pd.read_csv('heart_data.csv', header=None, names=column_names)

clf = clf.replace("?", None)


clf = clf.apply(pd.to_numeric, errors='ignore')


x = clf.iloc[1:, 0:13]

y = clf.iloc[1:, 13]


clf = tree.DecisionTreeClassifier(random_state=0, max_depth=12, max_leaf_nodes=2)


train_accuracy = []

test_accuracy = []


print('N \t На обучающей ', 'На тестовой ', sep=" ")

for i in range(10):

    random_seed = random.randint(1, 1000)

    X_train, X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.3, random_state=random_seed)

    clf.fit(X_train, Y_train)

    train_acc = clf.score(X_train, Y_train)
```

```
    test_acc = clf.score(X_test, Y_test)
    train_accuracy.append(train_acc)
    test_accuracy.append(test_acc)
    print(f'{(i + 2):2}\t {train_acc:.6f}\t {test_acc:.6f}')

plt.figure(figsize=(8, 6))
plt.plot(range(2, 12), train_accuracy, label='Обучающая')
plt.plot(range(2, 12), test_accuracy, label='Тестовая')
plt.xlabel('Номер итерации')
plt.ylabel('Точность')
plt.legend()
plt.title('Точность и номер итерации')
plt.show()
```

Результаты работы

```
PS C:\MMO> & C:/Users/Татьяна/AppData/Local/Programs/Python/Python310/python.exe c:/MMO/lab2.py
N        На обучающей  На тестовой
 2       0.755396        0.761111
 3       0.760192        0.750000
 4       0.741007        0.794444
 5       0.745803        0.783333
 6       0.755396        0.761111
 7       0.755396        0.761111
 8       0.755396        0.761111
 9       0.769784        0.727778
10       0.736211        0.805556
11       0.748201        0.777778
```

Рис1. Вывод работы в числовом соотношении

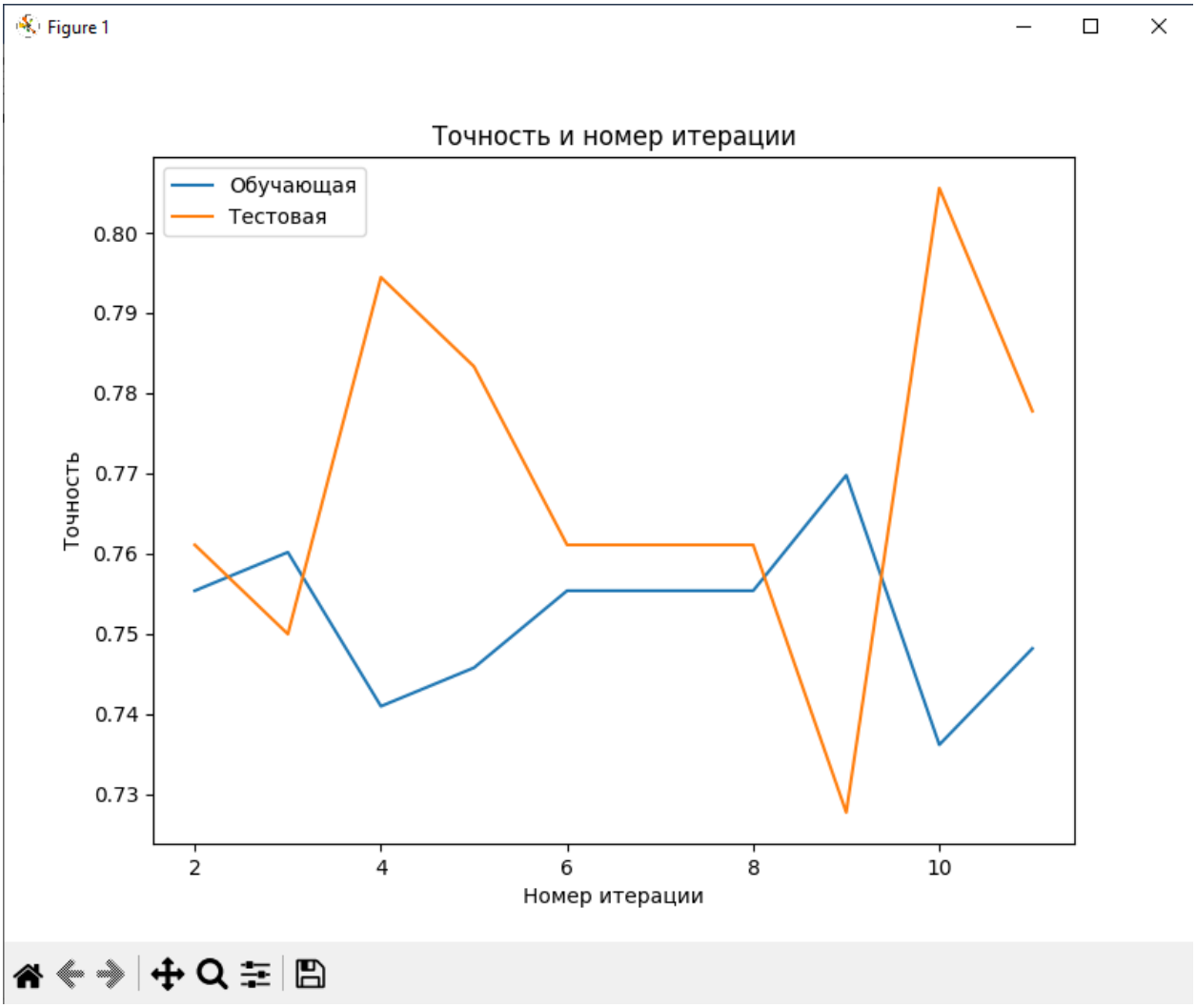

Рис2. Графическое отображение работы программы.

Выводы

В ходе работы я изучила и применила на практике метод решающих деревьев, используя библиотеку для машинного обучения sklearn.