ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

«СИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТЕЛЕКОММУНИКАЦИЙ И ИНФОРМАТИКИ»

Лабораторная работа №2 по дисциплине

«Методы машинного обучения»

Выполнил: студент 4 курса группы ИП-813

Бурдуковский И.А.

Проверил: ассистент кафедры ПМиК

Морозова К. И.

Новосибирск, 2021 г.

Содержание

[Постановка задачи 3](#_Toc84868729)

[Выполнение 4](#_Toc84868730)

[Код программы 6](#_Toc84868731)

# Постановка задачи

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.

В отчёте следует отразить следующие изменяемые параметры: глубина дерева и количество деревьев для каждого тестируемого случая.

# Выполнение

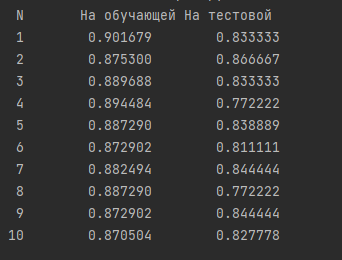
Было реализовано 10 разбиений исходной выборки на обучающую и тестовую.

Для считывания из файла и работой с данными испульзую библиотеку pandas. Заменили все вопросы нулями для одинаковой типизации.

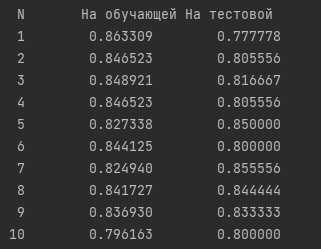
Для постоения дерева испульзовали библиотеку sklearn, а в ней tree. Выбираем х и у для train\_test\_split(x, y) и тестовую выборку 0.3.

Для построения дерева использовали два параметра max\_depth и min\_sample\_leaf. Отвечают за количество свойств в вершине дерева и за длину дерева соответственно. Первыми были 6 и 12

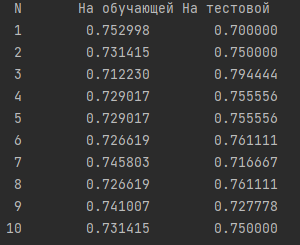
В итоге были получены следующие выводы:



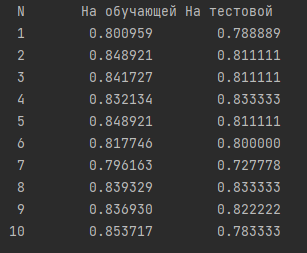
3 и 12



1 и 12



3 и 6



Чем больше мы выбираем высоту дерева, и чем меньше свойств в вершине, тем эффективнее результат программы.

# Код программы

import pandas as pd  
from sklearn import tree  
from sklearn.model\_selection import train\_test\_split  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 clf = pd.read\_csv('heart\_data.csv', header=None)  
 clf = clf.replace("?", None)  
  
 x = clf.loc[1:, 0:12]  
 y = clf.loc[1:, 13]  
 clf = tree.DecisionTreeClassifier(random\_state=0, max\_depth=12, max\_leaf\_nodes=2)  
  
 print(' N \t\t На обучающей', 'На тестовой', sep=" ")  
 for i in range(10):  
 X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(x, y, test\_size=0.3)  
 clf.fit(X\_train, Y\_train)  
 print(f'{(i + 1):2} {clf.score(X\_train, Y\_train) :15.6f} {clf.score(X\_test, Y\_test) :15.6f}')