**Министерство науки и высшего образования Российской Федерации**
**Калужский филиал федерального государственного автономного образовательного учреждения высшего образования**
**«Московский государственный технический университет имени Н.Э. Баумана**
**(национальный исследовательский университет)»**
**(КФ МГТУ им. Н.Э. Баумана)**

ФАКУЛЬТЕТ ИУК Информатика и управление

КАФЕДРА ИУК4 Программное обеспечение ЭВМ, информационные технологии

# ДОМАШНЯЯ РАБОТА №1

## «Подготовка данных к решению задач машинного обучения»

### по дисциплине: «Методы машинного обучения»

Выполнил: студент группы ИУК4-72Б _____ Губин Е.В.
(Подпись) (И.О. Фамилия)

Проверил: _____ Семененко М.Г.
(Подпись) (И.О. Фамилия)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:
- Оценка:

Калуга, 2025

**Цели:** разработать и проанализировать модель машинного обучения с использованием методов регрессии и классификации на основе сформированного набора данных.

**Задачи:**

1) Сформировать и предварительно обработать набор данных, включающий медицинские и социальные показатели пациентов.
2) Реализовать вычисление параметров и метрик, необходимых для анализа качества моделей регрессии.
3) Построить и проанализировать примеры применения дерева решений для решения задач классификации и/или регрессии.

## Вариант №4

**Формулировка задания:**

1. Приведите примеры метрик для оценки качества модели линейной регрессии.
2. Что такое дерево решений? Приведите пример его использования в задачах регрессии и/или классификации.
3. Создайте фрейм данных из N = 24 записей со следующими полями: Nrow – номер записи, Name – имя пациента, BirthYear – год рождения, Employ – место работы, Salary – зарплата. Cost – стоимость лечения, Albumin – содержание альбумина в крови, TransFerrin – содержание трансферрина в крови, Ferritin – содержание ферритина в крови. Заполните данный фрейм данными так, что Nrow изменяется от 1 до N.
   Name задается произвольно, BithYear распределен равномерно (случайно) на отрезке [1971,1997], , Cost для пациентов младше 1991 г.р. определяется по формуле Cost = (ln(2013 − BirthY ear) +1) ∗ 11000, для остальных Cost = (log2(2013 − BirthY ear) + 1) ∗ 10000.
   Ранжируйте пациентов по стоимости лечения, начиная с минимальной суммы. Добавьте в таблицу поле, соответствующее общему социальному вычету за лечение (ставка 13%), выплаченному пациенту, если стоимость лечения за каждый год начислялась согласно формулам для Cost, где вместо 2013 следует последовательно подставить каждый год нахождения пациента под наблюдением.

## Теоретическая часть

***Задание №1.*** *Приведите примеры метрик для оценки качества модели линейной регрессии.*

Метрики для оценки качества модели линейно регрессии:

1) Среднеквадратичная ошибка (MSE, чем меньше, тем лучше)
2) Корень из среднеквадратичной ошибки
3) Средняя абсолютная ошибка
4) Коэффициент детерминации
5) Скорректированный $R^2$
6) Средняя абсолютная процентная ошибка

***Задание №2.*** *Что такое дерево решений? Приведите пример его использования в задачах регрессии и/или классификации.*

Дерево решений — это модель машинного обучения, которая принимает решение, последовательно разветвляясь на основе вопросов о признаках данных.

Оно представляет собой структуру в виде дерева:

- узлы — проверка условия (например: “Возраст > 40?”)
- ветви — результат проверки (да/нет)
- листья — итоговое решение (класс или числовое значение)

Модель учится автоматически выбирать признаки и пороги разбиений, чтобы минимизировать ошибку.

***Пример использования дерева решений классификации:***

*Задача:* предсказать, имеет ли пациент диабет.
*Признаки:*

- уровень глюкозы
- ИМТ
- возраст
- давление

*Дерево решений классификации:*

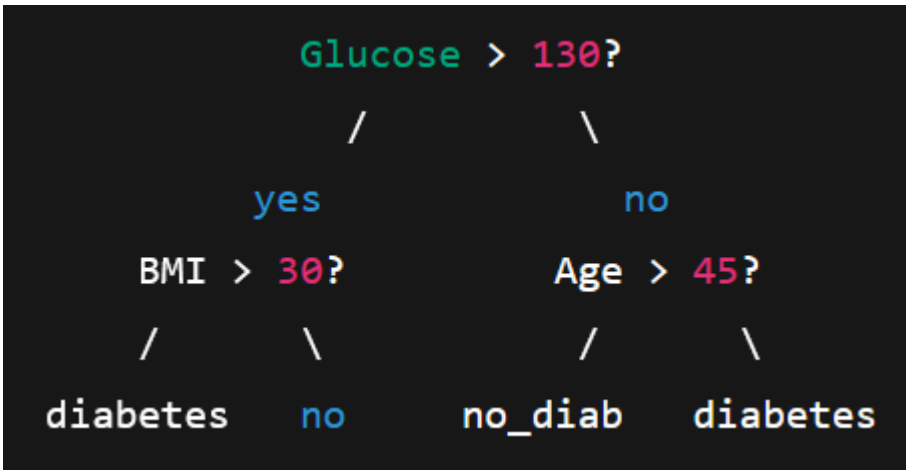


*Рисунок 1 Дерево решений, определяющее наличие диабета*

***Пример использования дерева решений регресии:***

В листе дерева — **числовое значение**, а не класс.

*Задача:* предсказать стоимость лечения пациента.
*Признаки:*

- возраст
- зарплата
- показатель альбумина
- наличие хронических болезней

*Дерево решений регрессии:*

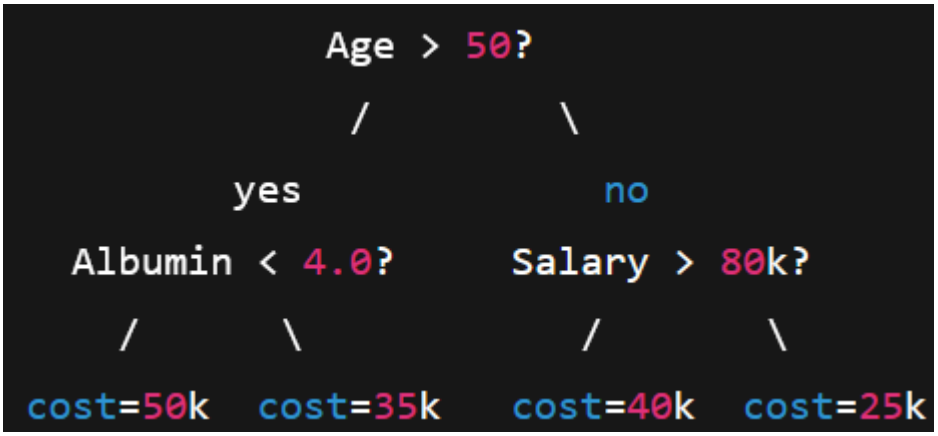


*Рисунок 2 Дерево классификации для решения регрессии*

## Практическая часть

### Результаты выполнения программы:

```
=== Исходный датафрейм ===
    Nrow        Name  BirthYear    Employ  ...   Albumin  TransFerrin    Ferritin  SocialDeduction
0      1   Patient_1       1977   Factory  ...  4.300170   239.743136  158.262767     25971.281173
1      2   Patient_2       1990        IT  ...  4.366776   201.104423  166.365192     23261.511674
2      3   Patient_3       1985  Hospital  ...  3.462136   363.092286  139.711485     24460.169647
3      4   Patient_4       1981    Office  ...  5.423962   341.371469   27.117355     25265.549336
4      5   Patient_5       1978        IT  ...  4.937832   345.801434   50.209600     25802.851432
5      6   Patient_6       1991   Factory  ...  5.348747   354.254069   28.800172     27848.086045
6      7   Patient_7       1977  Hospital  ...  5.237068   214.808930  198.194915     25971.281173
7      8   Patient_8       1996        IT  ...  4.494750   271.693146  108.019675     25742.220867
8      9   Patient_9       1989    Office  ...  5.304686   223.173812  162.399794     23522.231500
9     10  Patient_10       1993    Office  ...  3.221231   372.620685  274.118613     27075.416486
10    11  Patient_11       1981     Store  ...  3.489957   324.659625   89.801824     25265.549336
11    12  Patient_12       1981     Store  ...  3.113068   266.179605  134.907218     25265.549336
12    13  Patient_13       1994   Factory  ...  3.813326   212.711670  231.554319     26656.909963
13    14  Patient_14       1991        IT  ...  3.971693   262.196464   84.063486     27848.086045
14    15  Patient_15       1974        IT  ...  3.678373   265.036664   41.554375     26448.686059
15    16  Patient_16       1978     Store  ...  5.071844   345.921236  101.130407     25802.851432
16    17  Patient_17       1994        IT  ...  3.891883   327.511494   65.141960     26656.909963
17    18  Patient_18       1973        IT  ...  3.702336   377.442549  280.315343     26599.351597
18    19  Patient_19       1992    Office  ...  4.356740   294.442985  246.273706     27471.727442
19    20  Patient_20       1991     Store  ...  3.352311   223.918849  197.353052     27848.086045
20    21  Patient_21       1972  Hospital  ...  5.005492   342.648957  264.008965     26746.147037
21    22  Patient_22       1994     Store  ...  3.186377   352.157010  245.028182     26656.909963
22    23  Patient_23       1982  Hospital  ...  5.467217   312.255440   72.239616     25074.599445
23    24  Patient_24       1976    Office  ...  4.930612   354.193436  269.916520     26134.887975

[24 rows x 10 columns]
```

*Рисунок 3 Исходный DataFrame*

```
=== Датафрейм, отсортированный по стоимости лечения (Cost) ===
    Nrow        Name  BirthYear    Employ  ...   Albumin  TransFerrin    Ferritin  SocialDeduction
1      2   Patient_2       1990        IT  ...  4.366776   201.104423  166.365192     23261.511674
8      9   Patient_9       1989    Office  ...  5.304686   223.173812  162.399794     23522.231500
2      3   Patient_3       1985  Hospital  ...  3.462136   363.092286  139.711485     24460.169647
22    23  Patient_23       1982  Hospital  ...  5.467217   312.255440   72.239616     25074.599445
10    11  Patient_11       1981     Store  ...  3.489957   324.659625   89.801824     25265.549336
11    12  Patient_12       1981     Store  ...  3.113068   266.179605  134.907218     25265.549336
3      4   Patient_4       1981    Office  ...  5.423962   341.371469   27.117355     25265.549336
15    16  Patient_16       1978     Store  ...  5.071844   345.921236  101.130407     25802.851432
4      5   Patient_5       1978        IT  ...  4.937832   345.801434   50.209600     25802.851432
6      7   Patient_7       1977  Hospital  ...  5.237068   214.808930  198.194915     25971.281173
0      1   Patient_1       1977   Factory  ...  4.300170   239.743136  158.262767     25971.281173
23    24  Patient_24       1976    Office  ...  4.930612   354.193436  269.916520     26134.887975
7      8   Patient_8       1996        IT  ...  4.494750   271.693146  108.019675     25742.220867
14    15  Patient_15       1974        IT  ...  3.678373   265.036664   41.554375     26448.686059
17    18  Patient_18       1973        IT  ...  3.702336   377.442549  280.315343     26599.351597
20    21  Patient_21       1972  Hospital  ...  5.005492   342.648957  264.008965     26746.147037
16    17  Patient_17       1994        IT  ...  3.891883   327.511494   65.141960     26656.909963
21    22  Patient_22       1994     Store  ...  3.186377   352.157010  245.028182     26656.909963
12    13  Patient_13       1994   Factory  ...  3.813326   212.711670  231.554319     26656.909963
9     10  Patient_10       1993    Office  ...  3.221231   372.620685  274.118613     27075.416486
18    19  Patient_19       1992    Office  ...  4.356740   294.442985  246.273706     27471.727442
5      6   Patient_6       1991   Factory  ...  5.348747   354.254069   28.800172     27848.086045
13    14  Patient_14       1991        IT  ...  3.971693   262.196464   84.063486     27848.086045
19    20  Patient_20       1991     Store  ...  3.352311   223.918849  197.353052     27848.086045

[24 rows x 10 columns]
```

*Рисунок 4 DataFrame, отсортированный по стоимости лечения*

```
=== Минимальная стоимость лечения ===
Nrow                          2
Name                  Patient_2
BirthYear                  1990
Employ                       IT
Salary                   117313
Cost               45490.436375
Albumin                4.366776
TransFerrin          201.104423
Ferritin             166.365192
SocialDeduction    23261.511674
Name: 1, dtype: object
```

*Рисунок 5 Пациент с минимальной стоимость лечения*

```
=== Максимальная стоимость лечения ===
Nrow                         20
Name                 Patient_20
BirthYear                  1991
Employ                    Store
Salary                    97121
Cost               54594.316186
Albumin                3.352311
TransFerrin          223.918849
Ferritin             197.353052
SocialDeduction    27848.086045
Name: 19, dtype: object
```

*Рисунок 6 Пациент с максимальной стоимостью решения*

**Листинг программы:**

```
import pandas as pd
import numpy as np
import math

N = 24
np.random.seed(42)

names = [f"Patient_{i}" for i in range(1, N+1)]
birth_years = np.random.randint(1971, 1998, N)
employ = np.random.choice(["Office", "Factory", "Store", "IT", "Hospital"], N)
salary = np.random.randint(30000, 150000, N)
albumin = np.random.uniform(3.0, 5.5, N)
transferrin = np.random.uniform(200, 400, N)
ferritin = np.random.uniform(20, 300, N)

def calc_cost(by):
    if by < 1991:
        return (math.log(2013 - by) + 1) * 11000
    else:
        return (math.log2(2013 - by) + 1) * 10000

cost = [calc_cost(by) for by in birth_years]

years_obs = list(range(2010, 2014))

def total_social_deduction(by):
    total_cost = 0
    for y in years_obs:
        if by < 1991:
            total_cost += (math.log(y - by) + 1) * 11000
        else:
            total_cost += (math.log2(y - by) + 1) * 10000
    return total_cost * 0.13

social_deduction = [total_social_deduction(by) for by in birth_years]

df = pd.DataFrame({
    "Nrow": range(1, N+1),
    "Name": names,
    "BirthYear": birth_years,
    "Employ": employ,
    "Salary": salary,
    "Cost": cost,
    "Albumin": albumin,
    "TransFerrin": transferrin,
    "Ferritin": ferritin,
    "SocialDeduction": social_deduction
})

df_sorted = df.sort_values(by="Cost")

print("=== Исходный датафрейм ===")
print(df)

print("\n=== Датафрейм, отсортированный по стоимости лечения (Cost) ===")
print(df_sorted)

print("\n=== Минимальная стоимость лечения ===")
print(df_sorted.iloc[0])

print("\n=== Максимальная стоимость лечения ===")
```

```
print(df_sorted.iloc[-1])
```

**Вывод:** в ходе выполнения работы я получил практические навыки формирования данных, расчёта метрик и применения моделей машинного обучения.