**Московский государственный технический университет им. Н. Э. Баумана**

Курс «Технологии машинного обучения»

Отчёт по лабораторной работе №2

Выполнил:
Здобняков Ф. А.
группа ИУ5-64Б

Проверил:
Гапанюк Ю.Е.

Дата: 07.04.25

Подпись:

Дата:

Подпись:

Москва, 2025 г.

**Цель лабораторной работы:** изучение способов предварительной обработки данных для дальнейшего формирования моделей.

**Задание:**

1. Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:
    a. обработку пропусков в данных;
    b. кодирование категориальных признаков;
    c. масштабирование данных.

**Ход выполнения:**

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
%matplotlib inline
sns.set(style="ticks")
```

```
data = pd.read_csv('student_admission_record_dirty.csv', sep=",")
```

```
data.head()
```

| | Name | Age | Gender | Admission Test Score | High School Percentage | City | Admission Status |
|---|---|---|---|---|---|---|---|
| 0 | Shehroz | 24.0 | Female | 50.0 | 68.90 | Quetta | Rejected |
| 1 | Waqar | 21.0 | Female | 99.0 | 60.73 | Karachi | NaN |
| 2 | Bushra | 17.0 | Male | 89.0 | NaN | Islamabad | Accepted |
| 3 | Aliya | 17.0 | Male | 55.0 | 85.29 | Karachi | Rejected |
| 4 | Bilal | 20.0 | Male | 65.0 | 61.13 | Lahore | NaN |

```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 157
```

```
[11]
    data.columns
```

```
Index(['Name', 'Age', 'Gender', 'Admission Test Score',
       'High School Percentage', 'City', 'Admission Status'],
      dtype='object')
```

```
[13]
    print(data.isnull().sum())
```

```
Name                        10
Age                         10
Gender                      10
Admission Test Score        11
High School Percentage      11
City                        10
Admission Status            10
dtype: int64
```

1.

```
[20]
    data_cleaned = data.dropna()
```

```
[21]
    print(data_cleaned.isnull().sum())
```

```
Name                        0
Age                         0
Gender                      0
Admission Test Score        0
High School Percentage      0
City                        0
Admission Status            0
dtype: int64
```

2. модой возраст

+ Code + Markdown

```
[22]
data['Age'] = data['Age'].fillna(data['Gender'].mode()[0])  # Заполнение модой
```

3)категориальные другие значения

```
[23]
data['Gender'] = data['Gender'].fillna('Неизвестно')
```

```
[24]
# Использование get_dummies для преобразования категориальных признаков
data = pd.get_dummies(data, columns=['City'], drop_first=True)

print(data.head())
```

```
...       Name   Age  Gender  Admission Test Score  High School Percentage  \
0  Shehroz  24.0  Female                  50.0                   68.90
1    Waqar  21.0  Female                  99.0                   60.73
2   Bushra  17.0    Male                  89.0                     NaN
3    Aliya  17.0    Male                  55.0                   85.29
4    Bilal  20.0    Male                  65.0                   61.13

  Admission Status  City_Karachi  City_Lahore  City_Multan  City_Peshawar  \
0         Rejected         False        False        False          False
1              NaN          True        False        False          False
2         Accepted         False        False        False          False
3         Rejected          True        False        False          False
4              NaN         False         True        False          False

   City_Quetta  City_Rawalpindi
0         True            False
1        False            False
2        False            False
```

```
scaler = MinMaxScaler()

data_cleaned[['Age', 'High School Percentage']] = scaler.fit_transform(data_cleaned[['Age', 'High School Percentage']])
```

```
print(data_cleaned.head())
```

```
       Name   Age  Gender  Admission Test Score  High School Percentage  \
0   Shehroz  1.00  Female                  50.0                0.654772
3     Aliya  0.72    Male                  55.0                0.790788
7     Rabia  0.84  Female                  82.0                0.544979
9    Kamran  0.76    Male                  53.0                0.904398
10   Shafiq  0.72    Male                  78.0                0.000000

       City Admission Status
0    Quetta         Rejected
3   Karachi         Rejected
7    Lahore         Accepted
9    Multan         Rejected
10   Quetta         Rejected
```