

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

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 21

Выполнил:  
Фролов М. К.  
группа ИУ5-64Б

Проверил:  
Гапанюк Ю.Е.

Дата: 06.04.25

Дата:

Подпись:

Подпись:

2025 г.

## Задание:

Номер варианта:

21

Номер задачи:

3. Для заданного набора данных произведите масштабирование данных (для одного признака) и преобразование категориальных признаков в количественные двумя способами (label encoding, one hot encoding) для одного признака. Какие методы Вы использовали для решения задачи и почему?

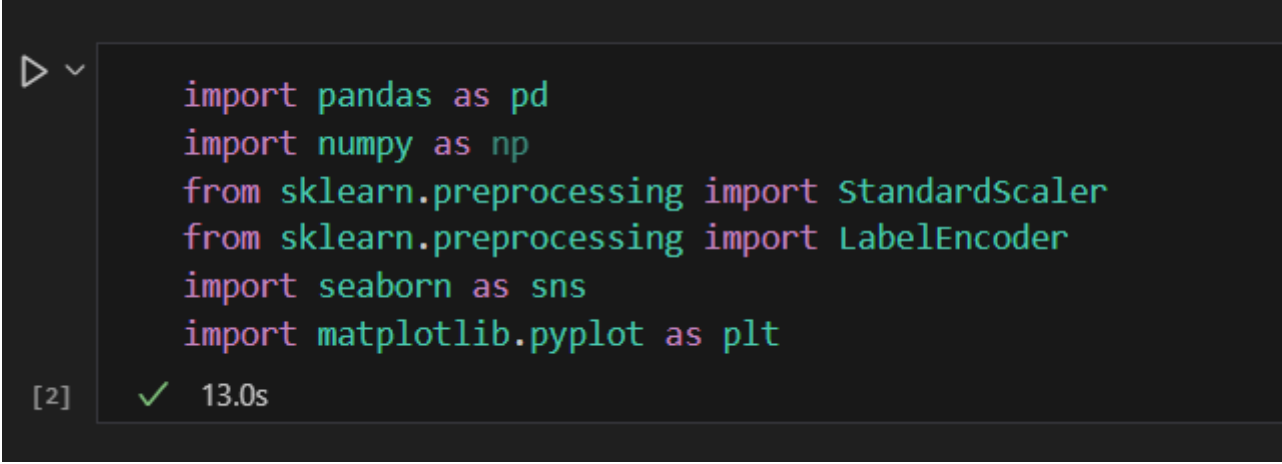
Номер набора данных, указанного в задаче:

5. <https://www.kaggle.com/mohansacharya/graduate-admissions> (файл Admission\_Predict.csv)

Дополнительные требования по группам:

Для студентов группы ИУ5-64Б, ИУ5Ц-84Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

## Ход выполнения:



```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
import seaborn as sns
import matplotlib.pyplot as plt
```

[2] ✓ 13.0s

# Масштабирование данных

Для масштабирования данных я выбрал признак "GRE Score", так как он является числовым и имеет значительный разброс значений

```
data = pd.read_csv('Admission_Predict.csv')

scaler = StandardScaler()
data['GRE Score Scaled'] = scaler.fit_transform(data[['GRE Score']])

data.head()
```

[3] ✓ 0.0s

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit	GRE Score Scaled
0	1	337	118	4	4.5	4.5	9.65	1	0.92	1.762107
1	2	324	107	4	4.0	4.5	8.87	1	0.76	0.627656
2	3	316	104	3	3.0	3.5	8.00	1	0.72	-0.070467
3	4	322	110	3	3.5	2.5	8.67	1	0.80	0.453126
4	5	314	103	2	2.0	3.0	8.21	0	0.65	-0.244998

## Преобразование категориальных признаков

Для преобразования категориальных признаков я выбрал колонку "University Rating", так как она содержит категориальные данные (рейтинг университета от 1 до 5).

```
label_encoder = LabelEncoder()
data['University Rating Label Encoded'] = label_encoder.fit_transform(data['University Rating'])

data.head()
```

[4] ✓ 0.0s

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit	GRE Score Scaled	University Rating Label Encoded
0	1	337	118	4	4.5	4.5	9.65	1	0.92	1.762107	3
1	2	324	107	4	4.0	4.5	8.87	1	0.76	0.627656	3
2	3	316	104	3	3.0	3.5	8.00	1	0.72	-0.070467	2
3	4	322	110	3	3.5	2.5	8.67	1	0.80	0.453126	2
4	5	314	103	2	2.0	3.0	8.21	0	0.65	-0.244998	1

Обоснование выбора Label Encoding:

- Подходит для "University Rating", так как это порядковый признак (рейтинг 1 лучше 2 и т.д.).
- Сохраняет информацию о порядке категорий.

## Скрипичная диаграмма (Violin Plot)

Для визуализации распределения данных я построил скрипичную диаграмму для признака "CGPA". Этот график показывает плотность распределения данных и их статистические характеристики (медиану, квантили).

```
plt.figure(figsize=(8, 6))
sns.violinplot(data=data, y='CGPA', inner='quartile')
plt.title('Violin Plot для CGPA')
plt.ylabel('CGPA')
plt.show()
```

[5] ✓ 0.2s

