Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Московский государственный технический университет имени Н.Э. Баумана»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ	**«Информатики и систем управления»**
КАФЕДРА	Системы обработки информации и управления

Дисциплина «Технологии машинного обучения»

РУБЕЖНЫЙ КОНТРОЛЬ №1

Вариант 21

Студент	Сахарова Е. К. ИУ5-62Б
Преподаватель	Гапанюк Ю. Е.

1. Задание

Номер варианта	Номер задачи	Номер набора данных, указанного в задаче
21	3	5

Дополнительное требование:

Для студентов групп ИУ5-62Б, ИУ5Ц-82Б – для произвольной колонки данных построить гистограмму.

Задача №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?

Наборы данных:

5. https://www.kaggle.com/noriuk/us-education-datasets-unification-project (файл states_all_extended.csv)

2. Выполнение задания

```
[ ] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    %matplotlib inline
```

```
[ ] data = pd.read_csv('states_all_extended.csv')
```

```
[ ] data.head()
```

	PRIMARY_KEY	STATE	YEAR	ENROLL	TOTAL_REVENUE	FEDERAL_REVENUE	STATE_REVENUE	LOCAL_REVENUE	TOTAL_EXPENDITURE	INSTRUC
0	1992_ALABAMA	ALABAMA	1992	NaN	2678885.0	304177.0	1659028.0	715680.0	2653798.0	
1	1992_ALASKA	ALASKA	1992	NaN	1049591.0	106780.0	720711.0	222100.0	972488.0	
2	1992_ARIZONA	ARIZONA	1992	NaN	3258079.0	297888.0	1369815.0	1590376.0	3401580.0	
3	1992_ARKANSAS	ARKANSAS	1992	NaN	1711959.0	178571.0	958785.0	574603.0	1743022.0	
4	1992_CALIFORNIA	CALIFORNIA	1992	NaN	26260025.0	2072470.0	16546514.0	7641041.0	27138832.0	

5 rows × 266 columns

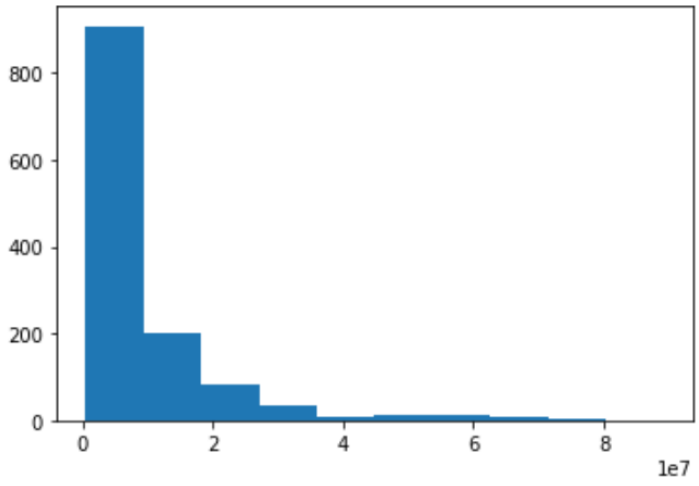
```
[ ] data.shape
```

```
(1715, 266)
```

Масштабирование данных

Для масштабирования данных возьмем колонку TOTAL_REVENUE

```
[ ] plt.hist(data.TOTAL_REVENUE)
    plt.show();
```

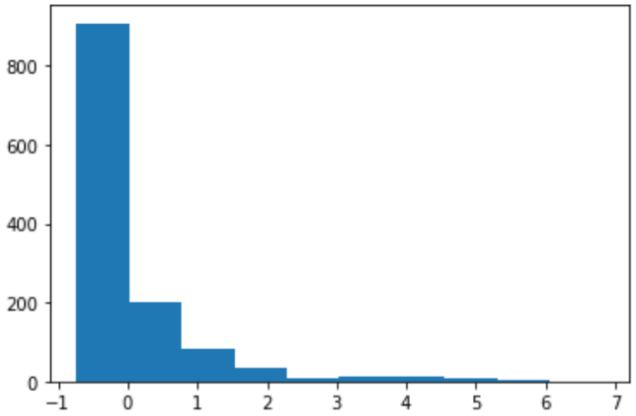


Масштабирование данных входит в подготовку данных. Этот этап подготовки предполагает нормирование или стандартизацию датасета для сдвига значений, что приведет к более удобному изучению данных и повышению качества и скорости обучения моделей. В данном случае используется стандартизация, но для того, чтобы понять, какой метод лучше, нужно прогнать тренировочные данные в модели и сравнить полученные результаты.

```
[ ]  from sklearn.preprocessing import StandardScaler

     scaler = StandardScaler()
     scaled_total_revenue = scaler.fit_transform(data[['TOTAL_REVENUE']])
```

```
[ ]  plt.hist(scaled_total_revenue)
     plt.show();
```



Преобразование категориальных признаков

```
[ ]  data.YEAR.unique()
```

```
array([1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002,
       2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013,
       2014, 2015, 2016, 1986, 1987, 1988, 1989, 1990, 1991, 2017, 2019])
```

```
[ ]  data.STATE.unique()
```

```
array(['ALABAMA', 'ALASKA', 'ARIZONA', 'ARKANSAS', 'CALIFORNIA',
       'COLORADO', 'CONNECTICUT', 'DELAWARE', 'DISTRICT_OF_COLUMBIA',
       'FLORIDA', 'GEORGIA', 'HAWAII', 'IDAHO', 'ILLINOIS', 'INDIANA',
       'IOWA', 'KANSAS', 'KENTUCKY', 'LOUISIANA', 'MAINE', 'MARYLAND',
       'MASSACHUSETTS', 'MICHIGAN', 'MINNESOTA', 'MISSISSIPPI',
       'MISSOURI', 'MONTANA', 'NEBRASKA', 'NEVADA', 'NEW_HAMPSHIRE',
       'NEW_JERSEY', 'NEW_MEXICO', 'NEW_YORK', 'NORTH_CAROLINA',
       'NORTH_DAKOTA', 'OHIO', 'OKLAHOMA', 'OREGON', 'PENNSYLVANIA',
       'RHODE_ISLAND', 'SOUTH_CAROLINA', 'SOUTH_DAKOTA', 'TENNESSEE',
       'TEXAS', 'UTAH', 'VERMONT', 'VIRGINIA', 'WASHINGTON',
       'WEST_VIRGINIA', 'WISCONSIN', 'WYOMING', 'DODEA', 'NATIONAL'],
      dtype=object)
```

Произведем категориальное преобразование для колонок YEAR (one hot encoding) и STATE (label encoding).

Так как штатов больше, чем годов, для удобства используем эти два способа именно таким образом.

Для метода One-Hote Encoding можно использовать методы библиотеки scklearn или pandas. В данном случае используется метод get_dummies() второй библиотеки. Различие между ними состоит в том, что get_dummies() по умолчанию, в отличие от OneHotEncoder, конвертирует строковые значения в One-Hot представление данных, для OneHotEncoder их нужно преобразовать в целочисленный тип данных.

С помощью метода get_dummies() библиотеки Pandas преобразуем колонку YEAR

```
[ ] data_year_ohe = pd.get_dummies(data=data.YEAR)
    data_year_ohe.head()
```

	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	19
0	0	0	0	0	0	0	1	0	0	0	0	0	
1	0	0	0	0	0	0	1	0	0	0	0	0	
2	0	0	0	0	0	0	1	0	0	0	0	0	
3	0	0	0	0	0	0	1	0	0	0	0	0	
4	0	0	0	0	0	0	1	0	0	0	0	0	

```
[ ] data_year_ohe.shape
```

```
(1715, 33)
```

Из-за этого способа данные могут намного увеличиться (в данном случае вместо одного столбца в датасете стало 33), что в будущем может повлиять на скорость обучения модели.

LabelEncoder намного экономнее One-Hot Encoding, так как итоговый результат кодирует различные типы только в одной колонке:

Теперь используем интсрумент LabelEncoder для кодирования штатов и получим итоговую таблицу

```
[ ] from sklearn.preprocessing import LabelEncoder

    le = LabelEncoder()
    data.STATE = le.fit_transform(data.STATE)
    data.head()
```

	PRIMARY_KEY	STATE	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2
0	1992_ALABAMA	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
1	1992_ALASKA	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
2	1992_ARIZONA	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
3	1992_ARKANSAS	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
4	1992_CALIFORNIA	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	

5 rows × 298 columns

Дополнительное требование

Построим гистограмму для колонки TOTAL_EXPENDITURE

```
[ ] plt.hist(data.TOTAL_EXPENDITURE)
    plt.show();
```