Московский государственный технический университет им. Н. Э. Баумана

Курс «Технологии машинного обучения»

Отчёт по лабораторной работе №2

Выполнил:
Горенков А.А.
группа ИУ5-63Б

Проверил:
Гапанюк Ю.Е.

Дата: 28.02.25

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:

Загрузка и первичный анализ

```
[ ] import numpy as np
    import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
    %matplotlib inline
    sns.set(style="ticks")
```

```
[ ] df0 = pd.read_csv("/dataset2.csv")
    df0.info()
    df0.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4998 entries, 0 to 4997
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Ids                 4998 non-null   int64
 1   Employer            4998 non-null   object
 2   Name                4998 non-null   object
 3   Salary              4998 non-null   bool
 4   From                2183 non-null   float64
 5   To                  1374 non-null   float64
 6   Experience          4998 non-null   object
 7   Schedule            4998 non-null   object
 8   Keys                4998 non-null   object
 9   Description         4998 non-null   object
 10  Area                4998 non-null   object
 11  Professional roles  4998 non-null   object
 12  Specializations     4998 non-null   object
 13  Profarea names      4998 non-null   object
 14  Published at        4998 non-null   object
dtypes: bool(1), float64(2), int64(1), object(11)
memory usage: 551.7+ KB
```

Ids	Employer	Name	Salary	From	To	Experience	Schedule	Keys	Description	Area	Professional roles	Specializat

Обработка пропусков: удаление

```
[ ] df1 = df0.copy()
```

```
[ ] # Удаление колонок, содержащих пустые значения
    df1_1 = df1.dropna(axis=1, how='any')
    (df1.shape, df1_1.shape)
```

```
((4998, 14), (4998, 12))
```

```
[ ] # Удаление строк, содержащих пустые значения
    df1_2 = df1.dropna(axis=0, how='any')
    (df1.shape, df1_2.shape)
```

```
((4998, 14), (1116, 14))
```

```
[ ] df1.head()
```

	Ids	Employer	Name	Salary	From	To	Experience	Schedule	Keys	Area	Professional roles	Specializations	Profarea
0	49313809	Space307	Golang Developer (Кипр)	True	251322.0	NaN	От 3 до 6 лет	Полный день	['Docker', 'Golang', 'Redis', 'Английский язык...	Санкт-Петербург	['Программист, разработчик']	['Программирование, Разработка']	['Информаци техно инт тел
1	48813842	Монополия	E-mail маркетолог	True	60900.0	NaN	От 1 года до 3 лет	Полный день	['Грамотность', 'Написание текстов', 'Грамотна...	Санкт-Петербург	['Менеджер по маркетингу и рекламе']	['Маркетинг']	['Информаци техно инт тел
2	49413720	Eden Springs	Оператор call-центра (удаленно)	False	NaN	NaN	От 1 года до 3 лет	Удаленная работа	['Клиентоориентированность', 'Ориентация на ре...	Санкт-Петербург	['Оператор call-центра, специалист контактного...	['Маркетинг', 'Продажи по телефону, Телемаркет...	['Про 'Информаци техно ин
3	46460892	Импорт Хоум	Ведущий SMM специалист	True	60000.0	80000.0	От 1 года до 3 лет	Полный день	['Продвижение бренда', 'Креативность', 'Adobe ...	Санкт-Петербург	['SMM-менеджер, контент-менеджер']	['Управление маркетингом', 'PR, Маркетинговые ...	['Информаци техно инт теле
4	49555567	Pride Games	UX/UI Designer	False	NaN	NaN	От 1 года до	Полный	['UI', 'UX', 'gamedev', 'game	Санкт-	['Дизайнер,	['Игровое ПО', 'Программирование,	['Марк реклам

Обработка пропусков: импьютация

```
[ ] df2 = df0.copy()
```

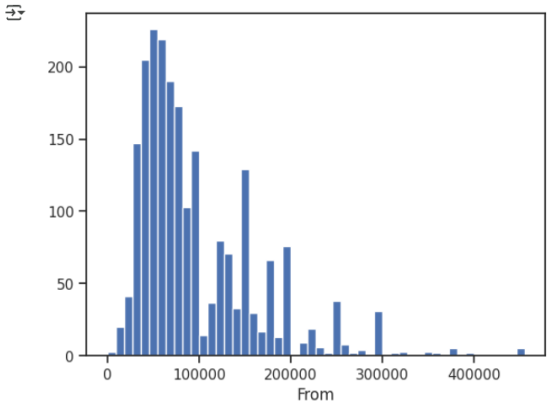
```
[ ] # Выберем числовые колонки с пропущенными значениями
    num_cols = []
    total_count = df2.shape[0]

    for col in df2.columns:
        # Количество пустых значений
        temp_null_count = df2[df2[col].isnull()].shape[0]
        dt = str(df2[col].dtype)
        if temp_null_count>0 and (dt=='float64' or dt=='int64'):
            num_cols.append(col)
            temp_perc = round((temp_null_count / total_count) * 100.0, 2)
            print('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format(col, dt, temp_null_count, temp_perc))
```

```
Колонка From. Тип данных float64. Количество пустых значений 2815, 56.32%.
Колонка To. Тип данных float64. Количество пустых значений 3624, 72.51%.
```

```
[ ] df2_num = df2[num_cols]
    df2_num
```

	From	To
0	251322.0	NaN
1	60900.0	NaN
2	NaN	NaN
3	60000.0	80000.0
4	NaN	NaN
...	...	...
4993	60000.0	NaN
4994	NaN	147900.0
4995	NaN	NaN
4996	NaN	NaN

```
[ ] # Гистограмма по признакам
    for col in df2_num:
        plt.hist(df2[col], 50)
        plt.xlabel(col)
        plt.show()
```



Кодирование категорий целочисленными значениями (label encoding)

```
[ ] from sklearn.preprocessing import LabelEncoder
```

```
[ ] df3 = df0.copy()
```

```
[ ] df3.head()
```

	Ids	Employer	Name	Salary	From	To	Experience	Schedule	Keys	Description	Area	Professional roles	Specializat
0	49313809	Space307	Golang Developer (Кипр)	True	251322.0	NaN	От 3 до 6 лет	Полный день	['Docker', 'Golang', 'Redis', 'Английский язык...	Мы в Space307 разрабатываем международную торг...	Санкт-Петербург	['Программист, разработчик']	['Программиров Разраб
1	48813842	Монополия	E-mail маркетолог	True	60900.0	NaN	От 1 года до 3 лет	Полный день	['Грамотность', 'Написание текстов', 'Грамотна...	С 2015 года наш IT блок меняет рынок автотранс...	Санкт-Петербург	['Менеджер по маркетингу и рекламе']	['Марке
2	49413720	Eden Springs	Оператор call-центра (удаленно)	False	NaN	NaN	От 1 года до 3 лет	Удаленная работа	['Клиентоориентированность', 'Ориентация на ре...	Что нужно будет делать: Принимать входящие зв...	Санкт-Петербург	['Оператор call-центра, специалист контактного...	['Марке 'Прода телес Телемар
3	46460892	Импорт Хоум	Ведущий SMM специалист	True	60000.0	80000.0	От 1 года до 3 лет	Полный день	['Продвижение бренда', 'Креативность', 'Adobe ...	В данный момент мы ищем в нашу команду самого ...	Санкт-Петербург	['SMM-менеджер, контент-менеджер']	['Управл маркетингом' Маркетингов
4	49555567	Pride Games Studio	UX/UI Designer	False	NaN	NaN	От 1 года до 3 лет	Полный день	['UI', 'UX', 'gamedev', 'game design', 'проект...	Pride Games Studio — это команда единомышленни...	Санкт-Петербург	['Дизайнер, художник']	['Игровое 'Программиров Разработ

```
[ ] encoder = LabelEncoder()

    df3['AreaEncoded'] = encoder.fit_transform(df3['Area'])
```

```
[ ] unique_rows = df3.drop_duplicates(subset='Area')
```

Кодирование шкал порядка

```
[ ] df4 = df0.copy()
```

```
[ ] unique_rows = df4.drop_duplicates(subset='Experience')
    unique_rows
```

	Ids	Employer	Name	Salary	From	To	Experience	Schedule	Keys	Description	Area	Professional roles	Specializations	I
0	49313809	Space307	Golang Developer (Кипр)	True	251322.0	NaN	От 3 до 6 лет	Полный день	['Docker', 'Golang', 'Redis', 'Английский язык...	Мы в Space307 разрабатываем международную торг...	Санкт-Петербург	['Программист, разработчик']	['Программирование, Разработка']	['\
1	48813842	Монополия	E-mail маркетолог	True	60900.0	NaN	От 1 года до 3 лет	Полный день	['Грамотность', 'Написание текстов', 'Грамотна...	С 2015 года наш IT блок меняет рынок автотранс...	Санкт-Петербург	['Менеджер по маркетингу и рекламе']	['Маркетинг']	['\
6	45573632	Global Ports Management	Специалист технического сопровождения	True	47850.0	56550.0	Нет опыта	Сменный график	['Грамотная речь', 'Грамотность', 'Работа с бо...	Группа компаний «Глобал Портс» – международная...	Санкт-Петербург	['Специалист технической поддержки']	['Морские/Речные перевозки', 'Контейнерные пер...	['\
70	49293222	Lumio Technologies Ltd	Senior/Lead PHP Engineer	True	324801.0	389762.0	Более 6 лет	Полный день	['Git', 'Работа в команде', 'Symfony', 'PHPUni...	Location: Remote or Relocation (London, United...	Санкт-Петербург	['Программист, разработчик']	['Банковское ПО', 'Программирование, Разработк...	['\

```
[ ] sizes = ['Нет опыта', 'От 1 года до 3 лет', 'От 3 до 6 лет', 'Более 6 лет']
    pd_sizes = pd.DataFrame(data={'sizes':sizes})
    pd_sizes
```

	sizes
0	Нет опыта
1	От 1 года до 3 лет
2	От 3 до 6 лет
3	Более 6 лет

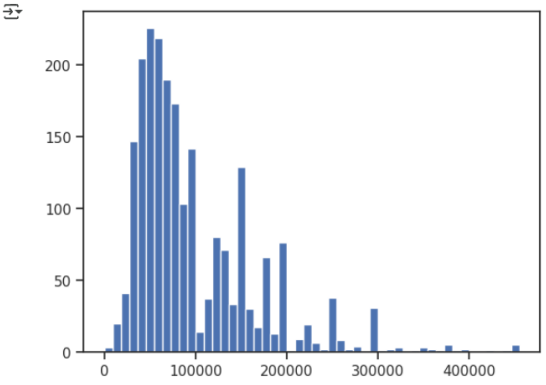
MinMax масштабирование

```
[ ] from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normalizer
```

```
[ ] df5 = df0.copy()
```

```
[ ] sc1 = MinMaxScaler()
    sc1_data = sc1.fit_transform(df5[['From']])
```

```
[ ] plt.hist(df5['From'], 50)
    plt.show()
```




```
[ ] plt.hist(sc1_data, 50)
    plt.show()
```