Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа №2**

**По курсу «Методы машинного обучения»**

**ИСПОЛНИТЕЛЬ:**

Лю Цзыцзянь

Группа ИУ5И-21М

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ПРОВЕРИЛ:**

Гапанюк Ю.Е.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Лабораторная работа 2

Задание

1) Выбрать набор данных (датасет), содержащий категориальные и числовые 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.) Просьба не использовать датасет, на котором данная задача решалась в лекции.

2) Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:

устранение пропусков в данных;

кодирование категориальных признаков;

нормализация числовых признаков.

**import** numpy **as** np

**import** pandas **as** pd

**import** seaborn **as** sns

**import** matplotlib.pyplot **as** plt

**%matplotlib** inline

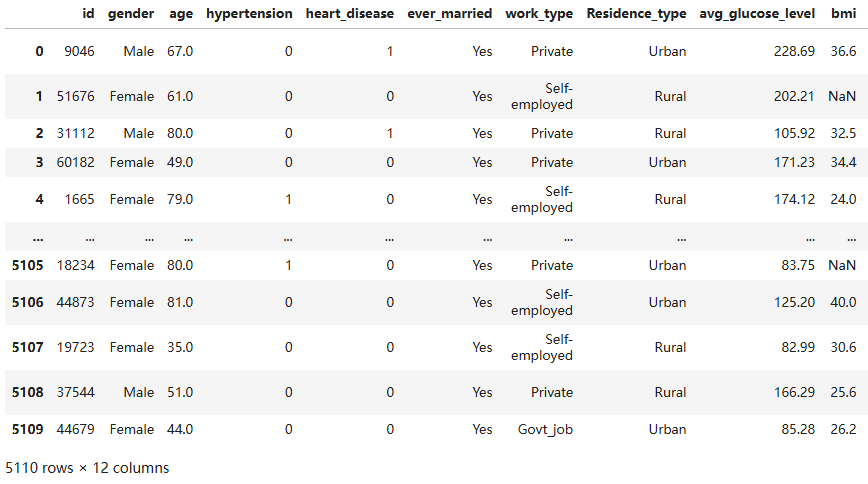
sns**.**set(style**=**"ticks")

**import** plotly.express **as** px

**from** sklearn.impute **import** SimpleImputer

df **=** pd**.**read\_csv('healthcare-dataset-stroke-data.csv')

df



**def** draw\_missing(df):

total **=** df**.**isnull()**.**sum()**.**sort\_values(ascending**=False**)

percent **=** (df**.**isnull()**.**sum()**/**df**.**isnull()**.**count())**.**sort\_values(ascending**=False**)**\***100

missing\_data **=** pd**.**concat([total, percent], axis**=**1, keys**=**['Total', 'Percent'])

**return** missing\_data

draw\_missing(df)**.**round(1)

Out[5]:

|  | **Total** | **Percent** |
| --- | --- | --- |
| **bmi** | 201 | 3.9 |
| **id** | 0 | 0.0 |
| **age** | 0 | 0.0 |
| **gender** | 0 | 0.0 |
| **hypertension** | 0 | 0.0 |
| **heart\_disease** | 0 | 0.0 |
| **work\_type** | 0 | 0.0 |
| **ever\_married** | 0 | 0.0 |
| **Residence\_type** | 0 | 0.0 |
| **avg\_glucose\_level** | 0 | 0.0 |
| **smoking\_status** | 0 | 0.0 |
| **stroke** | 0 | 0.0 |

total\_count **=** df**.**shape[0]

print('Всего строк: {}'**.**format(total\_count))

Всего строк: 5110

Обработка пропусков в числовых данных

*# Выберем числовые колонки с пропущенными значениями*

*# Цикл по колонкам датасета*

num\_cols **=** []

**for** col **in** df**.**columns:

*# Количество пустых значений*

temp\_null\_count **=** df[df[col]**.**isnull()]**.**shape[0]

dt **=** str(df[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 bmi. Тип данных float64. Количество пустых значений 201, 3.93%.

*# Фильтр по колонкам с пропущенными значениями*

data\_num **=** df[num\_cols]

data\_num

Out[8]:

|  | **bmi** |
| --- | --- |
| **0** | 36.6 |
| **1** | NaN |
| **2** | 32.5 |
| **3** | 34.4 |
| **4** | 24.0 |
| **...** | ... |
| **5107** | 30.6 |
| **5108** | 25.6 |
| **5109** | 26.2 |

5110 rows × 1 columns

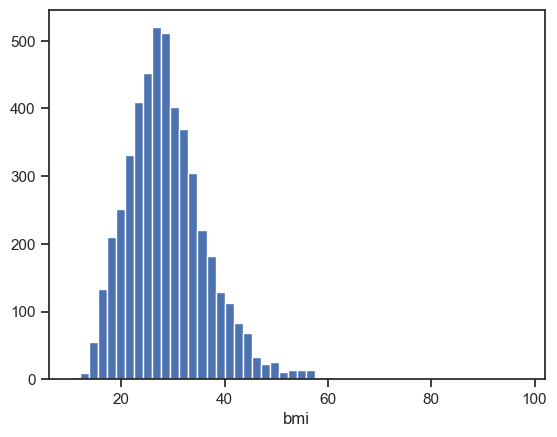
*# Гистограмма по признакам*

**for** col **in** data\_num:

plt**.**hist(df[col], 50)

plt**.**xlabel(col)

plt**.**show()



data\_num\_bmi **=** data\_num[['bmi']]

data\_num\_bmi**.**head()

Out[10]:

|  | **bmi** |
| --- | --- |
| **0** | 36.6 |
| **1** | NaN |
| **2** | 32.5 |
| **3** | 34.4 |
| **4** | 24.0 |

**from** sklearn.impute **import** SimpleImputer

**from** sklearn.impute **import** MissingIndicator

*# Фильтр для проверки заполнения пустых значений*

indicator **=** MissingIndicator()

mask\_missing\_values\_only **=** indicator**.**fit\_transform(data\_num\_bmi)

mask\_missing\_values\_only

Out[12]:

array([[False],

[ True],

[False],

...,

[False],

[False],

[False]])

strategies**=**['mean', 'median', 'most\_frequent']

**def** test\_num\_impute(strategy\_param):

imp\_num **=** SimpleImputer(strategy**=**strategy\_param)

df["bmi"] **=** imp\_num**.**fit\_transform(df[['bmi']])

test\_num\_impute(strategies[1])

Проверим, что пропусков нет

draw\_missing(df)**.**round(1)

Out[16]:

|  | **Total** | **Percent** |
| --- | --- | --- |
| **id** | 0 | 0.0 |
| **gender** | 0 | 0.0 |
| **age** | 0 | 0.0 |
| **hypertension** | 0 | 0.0 |
| **heart\_disease** | 0 | 0.0 |
| **ever\_married** | 0 | 0.0 |
| **work\_type** | 0 | 0.0 |
| **Residence\_type** | 0 | 0.0 |
| **avg\_glucose\_level** | 0 | 0.0 |
| **bmi** | 0 | 0.0 |
| **smoking\_status** | 0 | 0.0 |
| **stroke** | 0 | 0.0 |

Преобразование категориальных признаков в числовые

Использование LabelEncoder

**from** sklearn.preprocessing **import** LabelEncoder

le **=** LabelEncoder()

cat\_enc\_le **=** le**.**fit\_transform(df['work\_type'])

df['work\_type']**.**unique()

array(['Private', 'Self-employed', 'Govt\_job', 'children', 'Never\_worked'],

dtype=object)

np**.**unique(cat\_enc\_le)

array([0, 1, 2, 3, 4])

Кодирование категорий наборами бинарных значений

**from** sklearn.preprocessing **import** OneHotEncoder

cat\_enc**=** df[['work\_type']]

ohe **=** OneHotEncoder()

cat\_enc\_ohe **=** ohe**.**fit\_transform(df[['work\_type']])

cat\_enc\_ohe**.**shape

Out[24]:

(5110, 5)

cat\_enc\_ohe**.**todense()[0:10]

Out[25]:

matrix([[0., 0., 1., 0., 0.],

[0., 0., 0., 1., 0.],

[0., 0., 1., 0., 0.],

[0., 0., 1., 0., 0.],

[0., 0., 0., 1., 0.],

[0., 0., 1., 0., 0.],

[0., 0., 1., 0., 0.],

[0., 0., 1., 0., 0.],

[0., 0., 1., 0., 0.],

[0., 0., 1., 0., 0.]])

cat\_enc**.**head(10)

Out[26]:

|  | **work\_type** |
| --- | --- |
| **0** | Private |
| **1** | Self-employed |
| **2** | Private |
| **3** | Private |
| **4** | Self-employed |
| **5** | Private |
| **6** | Private |
| **7** | Private |
| **8** | Private |
| **9** | Private |

pd**.**get\_dummies(cat\_enc)**.**head()

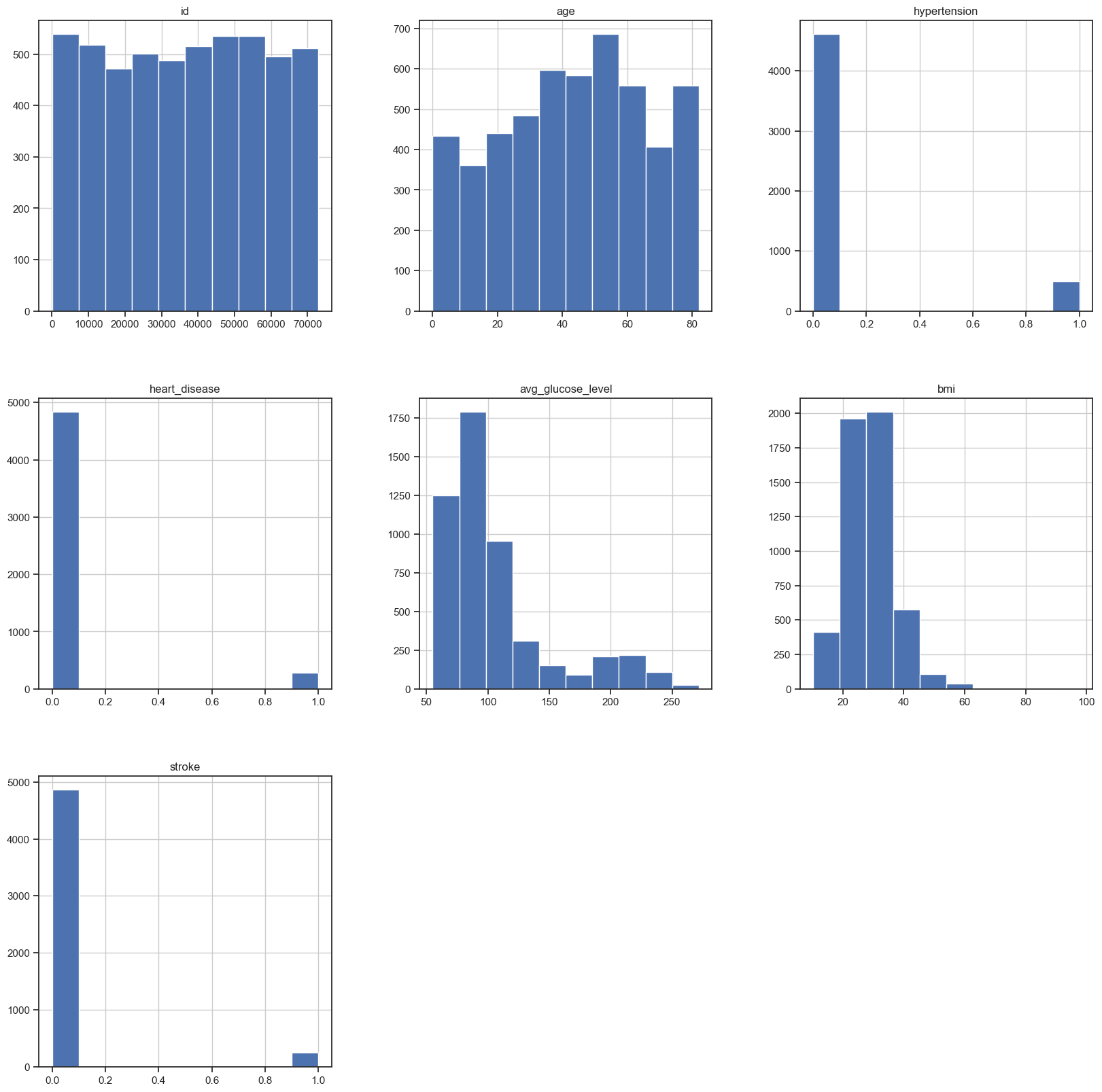
Out[27]:

|  | **work\_type\_Govt\_job** | **work\_type\_Never\_worked** | **work\_type\_Private** | **work\_type\_Self-employed** | **work\_type\_children** |
| --- | --- | --- | --- | --- | --- |
| **0** | False | False | True | False | False |
| **1** | False | False | False | True | False |
| **2** | False | False | True | False | False |
| **3** | False | False | True | False | False |
| **4** | False | False | False | True | False |

# Нормализация числовых признаков

df**.**hist(figsize**=**(20,20))

plt**.**show()



**import** scipy.stats **as** stats

**def** diagnostic\_plots(df, variable):

plt**.**figure(figsize**=**(15,6))

*# гистограмма*

plt**.**subplot(1, 2, 1)

df[variable]**.**hist(bins**=**30)

*## Q-Q plot*

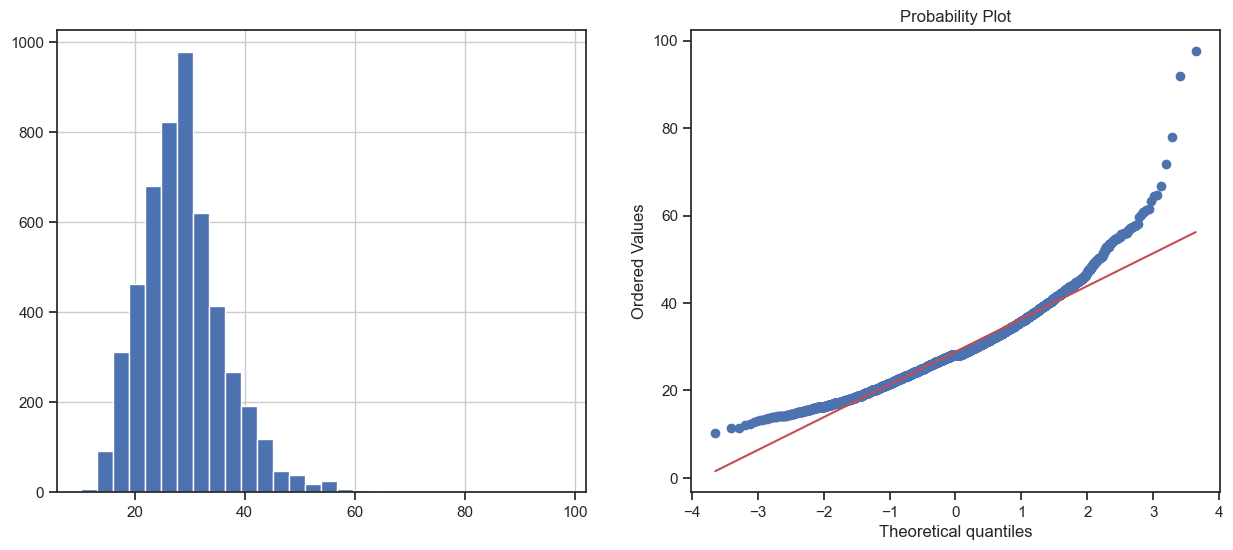
plt**.**subplot(1, 2, 2)

stats**.**probplot(df[variable], dist**=**"norm", plot**=**plt)

plt**.**show()

In [31]:

diagnostic\_plots(df, 'bmi')



In [32]:

df['bmi\_boxcox'], param **=** stats**.**boxcox(df['bmi'])

print('Оптимальное значение λ = {}'**.**format(param))

diagnostic\_plots(df, 'bmi\_boxcox')

Оптимальное значение λ = 0.0024460113998946375

