

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ  
им. Н.Э. Баумана

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

Рубежный контроль № 1

Выполнил:

студент группы ИУ5И-24М

Аунг Пьюи Нанда

Москва - 2024

## Рубежный контроль (Вариант 16)

### Задача №16.

Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Бокса-Кокса (Box-Cox transformation).

### Решение

#### Загрузка и предобработка данных

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import load_wine
from sklearn.datasets import fetch_california_housing
import scipy.stats as stats
from sklearn.svm import SVR
from sklearn.svm import LinearSVC
from sklearn.feature_selection import SelectFromModel
from sklearn.linear_model import Lasso
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import VarianceThreshold
from sklearn.feature_selection import mutual_info_classif, mutual_info_regression
from sklearn.feature_selection import SelectKBest, SelectPercentile
from IPython.display import Image
%matplotlib inline
sns.set(style="ticks")
```

```
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()
```

```
[3]: # Будем использовать только обучающую выборку
dataset = pd.read_csv("D:\\ИУ-5 2сем\\ММО\\weatherAUS.csv", sep=",")
```

```
[4]: dataset.head(10)
```

```
[4]:
```

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity9am	Humidity3pm	Pressure9am
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	71.0	22.0	1007.7
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	44.0	25.0	1010.6
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	38.0	30.0	1007.6
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	45.0	16.0	1017.6
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	82.0	33.0	1010.8
5	2008-12-06	Albury	14.6	29.7	0.2	NaN	NaN	WNW	56.0	W	...	55.0	23.0	1009.2
6	2008-12-07	Albury	14.3	25.0	0.0	NaN	NaN	W	50.0	SW	...	49.0	19.0	1009.6
7	2008-12-08	Albury	7.7	26.7	0.0	NaN	NaN	W	35.0	SSE	...	48.0	19.0	1013.4
8	2008-12-09	Albury	9.7	31.9	0.0	NaN	NaN	NNW	80.0	SE	...	42.0	9.0	1008.9
9	2008-12-10	Albury	13.1	30.1	1.4	NaN	NaN	W	28.0	S	...	58.0	27.0	1007.0

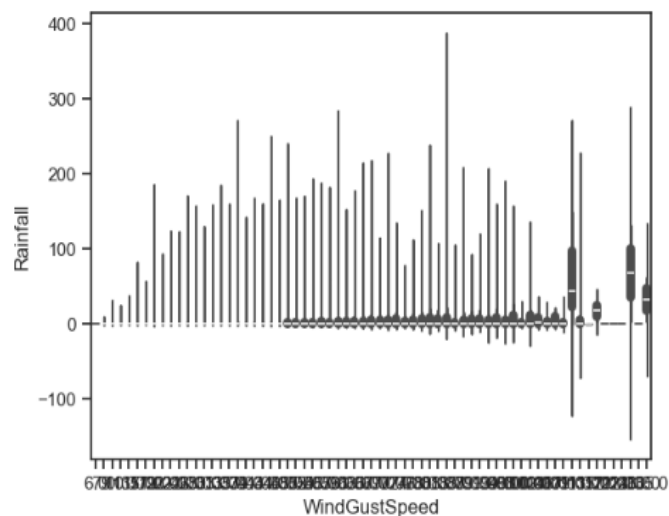
10 rows × 23 columns

```
[5]: data = pd.DataFrame({
    'WindGustSpeed': [44.0, 44.0, 46.0, 24.0, 41.0, 56.0, 50.0, 35.0, 80.0, 28.0],
    'Rainfall': [0.6, 0.0, 0.0, 0.0, 1.0, 0.2, 0.0, 0.0, 0.0, 1.4]
})
```

```
data['WindGustSpeed_boxcox'], param = stats.boxcox(data['WindGustSpeed'])
print('Оптимальное значение  $\lambda$  = {}'.format(param))
sns.violinplot(x='WindGustSpeed', y='Rainfall', data=dataset)
```

Оптимальное значение  $\lambda$  = -0.04037539669852678

<Axes: xlabel='WindGustSpeed', ylabel='Rainfall'>



## Задача №36.

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс `SelectKBest` для 5 лучших признаков, и метод, основанный на взаимной информации.

```
wine = load_wine()
wine_X = wine.data
wine_y = wine.target
wine_feature_names = wine['feature_names']
wine_x_df = pd.DataFrame(data=wine['data'], columns=wine['feature_names'])
```

```
fetch_california_housing = fetch_california_housing()
fetch_california_housing_X = fetch_california_housing.data
fetch_california_housing_y = fetch_california_housing.target
fetch_california_housing_feature_names = fetch_california_housing['feature_names']
fetch_california_housing_x_df = pd.DataFrame(data=fetch_california_housing['data'], columns=fetch_california_housing['feature_names'])
```

```
sel_mi = SelectKBest(mutual_info_regression, k=5).fit(wine_X, wine_y)
```

```
list(zip(fetch_california_housing_feature_names, sel_mi.get_support()))
```

```
[('MedInc', True),
 ('HouseAge', False),
 ('AveRooms', False),
 ('AveBedrms', False),
 ('Population', False),
 ('AveOccup', False),
 ('Latitude', True),
 ('Longitude', False)]
```

```
selected_features_mask = sel_mi.get_support()
selected_feature_indices = [i for i, selected in enumerate(selected_features_mask) if selected]

print("Selected feature indices:", selected_feature_indices)
print("Length of fetch_california_housing_feature_names:", len(fetch_california_housing_feature_names))

selected_feature_names = [fetch_california_housing_feature_names[i] for i in selected_feature_indices if i < len(fetch_california_housing_feature_names)]
```

◀  ▶

```
Selected feature indices: [0, 6, 9, 11, 12]
Length of fetch_california_housing_feature_names: 8
```