Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №2
по дисциплине

«Методы машинного обучения»

на тему

«Обработка признаков (часть 2)»

Выполнил:
студент группы ИУ5и-24М
Аунг Пьио Нанда

Москва — 2024 г.

1.Цель лабораторной работы

Изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

2.Задание

1. Выбрать один или несколько наборов данных (датасетов) для решения следующих задач. Каждая задача может быть решена на отдельном датасете, или несколько задач могут быть решены на одном датасете. Просьба не использовать датасет, на котором данная задача решалась в лекции.
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   i. масштабирование признаков (не менее чем тремя способами);
   ii. обработку выбросов для числовых признаков (по одному способу для удаления выбросов и для замены выбросов);
   iii. обработку по крайней мере одного нестандартного признака (который не является числовым или категориальным);
   iv. отбор признаков:
      - один метод из группы методов фильтрации (filter methods);
      - один метод из группы методов обертывания (wrapper methods);
      - один метод из группы методов вложений (embedded methods).

3. Ход выполнения работы

3.1) Загрузка и первичный анализ данных

На этом шаге мы загрузим данные и ознакомимся с ними.

[41]:
```
import pandas as pd
data = pd.read_csv("D:\\ИУ-5 2сем\\ММО\\Wine_quality.csv")
data.head()
```

[41]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

3.2) Масштабирование признаков

[45]:
```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
data_scaled_minmax = data.copy()
data_scaled_minmax[data.columns] = scaler.fit_transform(data)
```

3.2.1- Min-Max Scaling

[46]:
```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data_scaled_standard = data.copy()
data_scaled_standard[data.columns] = scaler.fit_transform(data)
```

3.2.2- Standard Scaling (Z-score Normalization)

47]:
```
from sklearn.preprocessing import RobustScaler
scaler = RobustScaler()
data_scaled_robust = data.copy()
data_scaled_robust[data.columns] = scaler.fit_transform(data)
```

3.2.3- Robust Scaling

3.3) Обработка выбросов для числовых признаков

```
[53]: Q1 = data['chlorides'].quantile(0.25)
      Q3 = data['chlorides'].quantile(0.75)
      IQR = Q3 - Q1
      data_no_outliers = data[~((data['chlorides'] < (Q1 - 1.5 * IQR)) | (data['chlorides'] > (Q3 + 1.5 * IQR)))]
```

3.3.1- Удаление выбросов с помощью межквартильного размаха

```
[54]: import numpy as np
      # Определение квартилей и межквартильного размаха
      Q1 = data['chlorides'].quantile(0.25)
      Q3 = data['chlorides'].quantile(0.75)
      IQR = Q3 - Q1
      median = data['chlorides'].median()
      data['chlorides'] = np.where((data['chlorides'] < Q1 - 1.5 * IQR) | (data['chlorides'] > Q3 + 1.5 * IQR), median, data['chlorides'])
```

3.3.2- Замена выбросов медианным значением

3.4) Обработка нестандартного признака

Для обработки нестандартного признака, например, текстового, можно использовать методы кодирования, такие как TF-IDF или Word2Vec.

3.5) Отбор признаков

```
]: import pandas as pd
   data = pd.read_csv("D:\\ИУ-5 2сем\\ММО\\Wine_quality.csv")
   df = pd.DataFrame(data)
   # Calculate the correlation between features and the target variable
   correlation = df.corrwith(df['quality']).sort_values(ascending=False)
   # Print the correlation results
   print("Correlation with quality:")
   print(correlation)
```

```
Correlation with quality:
quality                1.000000
alcohol                0.476166
sulphates              0.251397
citric acid            0.226373
fixed acidity          0.124052
residual sugar         0.013732
free sulfur dioxide   -0.050656
pH                    -0.057731
chlorides             -0.128907
density               -0.174919
total sulfur dioxide  -0.185100
volatile acidity      -0.390558
dtype: float64
```

3.5.1- Filter Methods

[63]:
```
from sklearn.feature_selection import RFE
from sklearn.ensemble import RandomForestClassifier
# Определение модели
model = RandomForestClassifier()
# Создание объекта RFE
rfe = RFE(model, n_features_to_select=5)
# Применение RFE к данным
rfe.fit(df.drop('quality', axis=1), df['quality'])
# Вывод результатов
print("Selected features by RFE:")
print(df.drop('quality', axis=1).columns[rfe.support_])
```

```
Selected features by RFE:
Index(['volatile acidity', 'total sulfur dioxide', 'density', 'sulphates',
       'alcohol'],
      dtype='object')
```

3.5.2- Wrapper Methods

[64]:
```
# Обучение модели случайного леса
model_rf = RandomForestClassifier()
model_rf.fit(df.drop('quality', axis=1), df['quality'])
# Важность признаков
feature_importance = pd.Series(model_rf.feature_importances_, index=df.drop('quality', axis=1).columns).sort_values(ascending=False)
# Вывод результатов
print("Feature importance from RandomForestClassifier:")
print(feature_importance)
```

```
Feature importance from RandomForestClassifier:
alcohol                0.152516
sulphates              0.109814
total sulfur dioxide   0.104936
volatile acidity       0.104158
density                0.089171
chlorides              0.077535
pH                     0.076569
fixed acidity          0.074279
citric acid            0.073418
residual sugar         0.071532
free sulfur dioxide    0.066072
dtype: float64
```

3.5.3- Embedded Methods

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2024. — Режим доступа: https://github.com/ugapanyuk/courses_current/wiki/LAB_MMO__FEATURES.