Московский государственный технический университет им. Н.Э. Баумана
Факультет «Головной учебно-исследовательский и методический центр
профессиональной реабилитации лиц с ограниченными возможностями здоровья (инвалидов)»
Кафедра «Системы обработки информации и управления»

Рубежный контроль

по дисциплине «Методы машинного обучения»

"Методы обработки данных"

СТУДЕНТ:
студент группы ИУ5Ц-21М
Москалик А.А.

ПРЕПОДАВАТЕЛЬ:
Гапанюк Ю.Е.

Москва, 2024

Номер варианта	Номер задачи №1	Номер задачи №2
22	7	33

Задача №7.

Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения медианой.

Задача №33.

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте метод обертывания (wrapper method), алгоритм полного перебора (exhaustive feature selection).

Дополнительные требования по группам:

Для студентов групп ИУ5-21М, ИУ5И-21М, ИУ5Ц-21М - для пары произвольных колонок данных построить график "Диаграмма рассеяния".

Набор данных включает в себя следующие параметры (признаки):

fixed_acidity: Зафиксированная кислотность вина, которая относится к нелетучим кислотам, не испаряющимся легко.
volatile_acidity: Летучая кислотность вина, которая относится к количеству уксусной кислоты в вине, слишком высокие уровни которой могут привести к неприятному вкусу уксуса.
citric_acid: Количество лимонной кислоты в вине, которая может добавлять в вино свежести и аромата.
residual_sugar: Количество сахара, оставшееся после окончания ферментации, измеряется в граммах на литр вина.
quality: Оценка качества вина на основе оценок экспертов (обычно от 0 до 10).

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.feature_selection import RFE
import matplotlib.pyplot as plt

# Загрузка или создание набора данных
data = pd.DataFrame({
    'fixed_acidity': [7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7.5],
    'volatile_acidity': [0.70, 0.88, 0.76, 0.28, 0.70, 0.66, 0.60, 0.65, 0.58, 0.50],
    'citric_acid': [0, 0, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00, 0.02, 0.36],
    'residual_sugar': [1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1],
    'quality': [5, 5, 5, 6, 5, 5, 5, 7, 7, 5]
})

# Добавляем пропуски в данных для имитации реальной ситуации
data.loc[0, 'fixed_acidity'] = None
data.loc[4:6, 'residual_sugar'] = None
```

```python
# Заполнение пропусков медианой для 'fixed_acidity' без использования inplace=True
data['fixed_acidity'] = data['fixed_acidity'].fillna(data['fixed_acidity'].median())
```

```python
# Подготовка данных для отбора признаков
data_for_feature_selection = data.dropna().reset_index(drop=True)
X = data_for_feature_selection.drop('quality', axis=1)
y = data_for_feature_selection['quality']

# Разделение на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Инициализация и обучение RFE
lr = LinearRegression()
rfe = RFE(estimator=lr, n_features_to_select=2)
rfe.fit(X_train, y_train)

# Определение выбранных признаков
selected_features = X_train.columns[rfe.support_]
```

```python
# Построение диаграммы рассеяния для выбранных признаков
plt.figure(figsize=(10, 6))
plt.scatter(data_for_feature_selection[selected_features[0]], data_for_feature_selection[selected_features[1]],
            c=data_for_feature_selection['quality'], cmap='viridis')
plt.title(f'Диаграмма рассеяния для признаков {selected_features[0]} и {selected_features[1]}')
plt.xlabel(selected_features[0])
plt.ylabel(selected_features[1])
plt.colorbar(label='Quality')
plt.show()
```

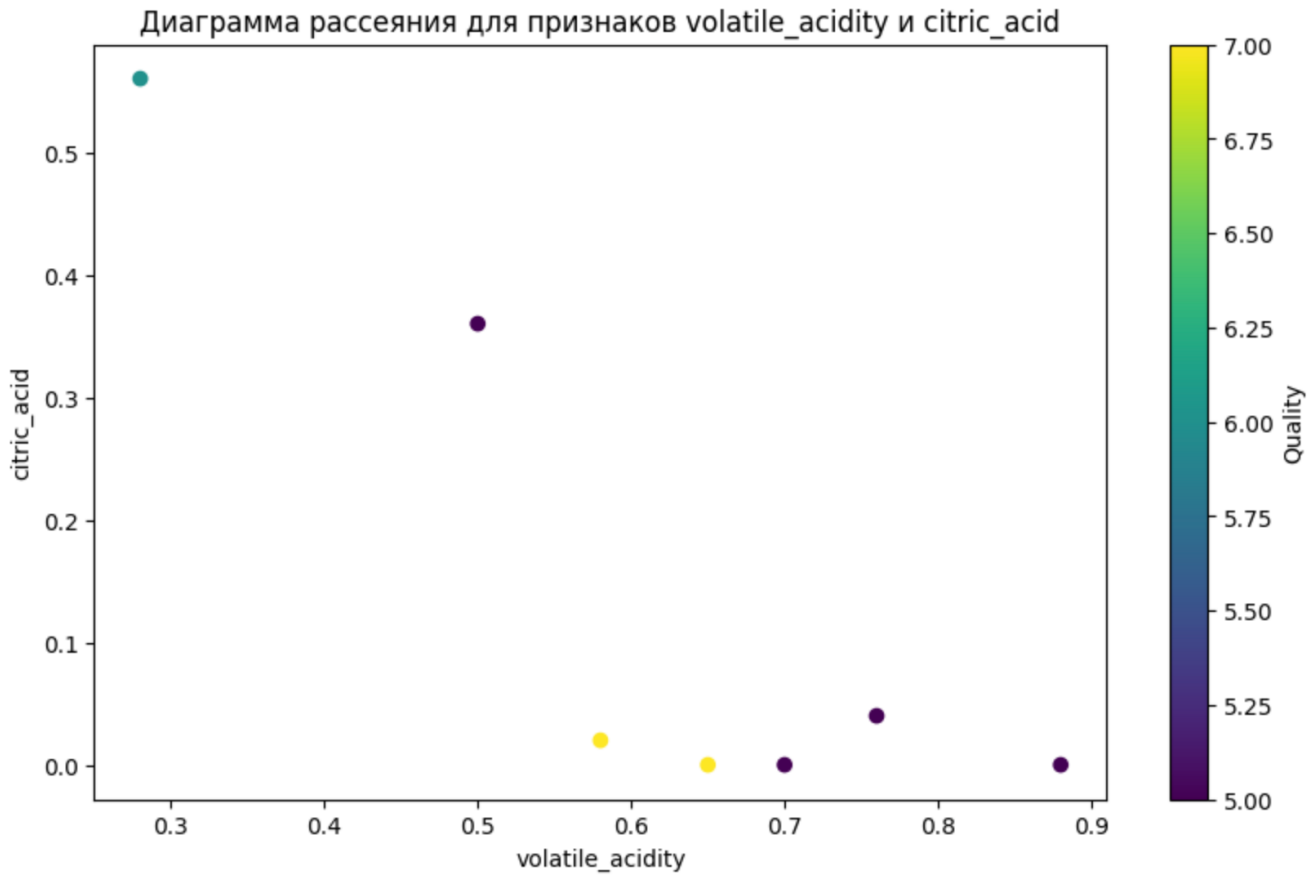
Диаграмма рассеяния для признаков volatile_acidity и citric_acid
citric_acid
volatile_acidity
Quality
0.0
0.1
0.2
0.3
0.4
0.5
0.3
0.4
0.5
0.6
0.7
0.8
0.9
5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75
7.00