Рубежный контроль №1

Никулин Данила ИУ5-61Б Вариант 10

Для студентов групп ИУ5-61Б, ИУ5Ц-81Б - для пары произвольных колонок данных построить график "Диаграмма рассеяния".

Задача №2.

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Набор данных №2

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine

```
from sklearn.datasets import load_wine
import pandas as pd
import numpy as np

data = load_wine()

df = pd.DataFrame(data.data, columns=data.feature_names)

df.head()
```


```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 178,\n  \"fields\": [\n    {\n      \"column\": \"alcohol\",\n      \"properties\": {\n\"dtype\": \"number\",\n        \"std\": 0.8118265380058575,\n\"min\": 11.03,\n        \"max\": 14.83,\n\"num_unique_values\": 126,\n        \"samples\": [\n          11.62,\n          13.64,\n          13.69\n        ],\n\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"malic_acid\",\n\"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 1.1171460976144627,\n        \"min\": 0.74,\n        \"max\": 5.8,\n\"num_unique_values\": 133,\n        \"samples\": [\n          1.21,\n2.83,\n          1.8\n        ],\n        \"semantic_type\": \"\",\n\"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"ash\",\n      \"properties\": {\n        \"dtype\": \"number\",\n\"std\": 0.27434400906081485,\n        \"min\": 1.36,\n\"max\": 3.23,\n        \"num_unique_values\": 79,\n\"samples\": [\n          2.31,\n          2.43,\n          2.52\n],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n}\n    },\n    {\n      \"column\": \"alcalinity_of_ash\",\n
```

\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
3.339563767173505,\n   \"min\": 10.6,\n   \"max\": 30.0,\n
\"num_unique_values\": 63,\n   \"samples\": [\n   25.5,\n
28.5,\n   15.6\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"magnesium\",\n   \"properties\": {\n   \"dtype\":
\"number\",\n   \"std\": 14.282483515295665,\n   \"min\":
70.0,\n   \"max\": 162.0,\n   \"num_unique_values\": 53,\n
\"samples\": [\n   126.0,\n   85.0,\n   162.0\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"total_phenols\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
0.6258510488339893,\n   \"min\": 0.98,\n   \"max\": 3.88,\n
\"num_unique_values\": 97,\n   \"samples\": [\n   1.68,\n
2.11,\n   1.35\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"flavanoids\",\n   \"properties\": {\n   \"dtype\":
\"number\",\n   \"std\": 0.9988586850169467,\n   \"min\":
0.34,\n   \"max\": 5.08,\n   \"num_unique_values\": 132,\n
\"samples\": [\n   3.18,\n   2.5,\n   3.17\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"nonflavanoid_phenols\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
0.12445334029667937,\n   \"min\": 0.13,\n   \"max\": 0.66,\n
\"num_unique_values\": 39,\n   \"samples\": [\n   0.58,\n
0.41,\n   0.39\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"proanthocyanins\",\n   \"properties\": {\n   \"dtype\":
\"number\",\n   \"std\": 0.5723588626747613,\n   \"min\":
0.41,\n   \"max\": 3.58,\n   \"num_unique_values\": 101,\n
\"samples\": [\n   0.75,\n   1.77,\n   1.42\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"color_intensity\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
2.318285871822413,\n   \"min\": 1.28,\n   \"max\": 13.0,\n
\"num_unique_values\": 132,\n   \"samples\": [\n   2.95,\n
3.3,\n   5.1\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"hue\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 0.22857156582982338,\n   \"min\": 0.48,\n
\"max\": 1.71,\n   \"num_unique_values\": 78,\n
\"samples\": [\n   1.22,\n   1.04,\n   1.45\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"od280/od315_of_diluted_wines\",\
n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 0.7099904287650504,\n   \"min\": 1.27,\n   \"max\":
4.0,\n   \"num_unique_values\": 122,\n   \"samples\": [\n
4.0,\n   1.82,\n   1.59\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\

n   },\n   {\n   \"column\": \"proline\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 314.9074742768491,\n   \"min\": 278.0,\n   \"max\": 1680.0,\n   \"num_unique_values\": 121,\n   \"samples\": [\n   1375.0,\n   1270.0,\n   735.0\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n]\n}","type":"dataframe","variable_name":"df"}

```
df.dtypes
```

```
alcohol                         float64
malic_acid                      float64
ash                             float64
alcalinity_of_ash               float64
magnesium                       float64
total_phenols                   float64
flavanoids                      float64
nonflavanoid_phenols            float64
proanthocyanins                 float64
color_intensity                 float64
hue                             float64
od280/od315_of_diluted_wines    float64
proline                         float64
dtype: object
```

```
df['target'] = data.target

# Создадим категориальный признак на основе числового
df['category'] = pd.cut(df['alcohol'], bins=3, labels=['low',
'medium', 'high'])
```

```
df.isnull().sum()
```

```
alcohol                         0
malic_acid                      0
ash                             0
alcalinity_of_ash               0
magnesium                       0
total_phenols                   0
flavanoids                      0
nonflavanoid_phenols            0
proanthocyanins                 0
color_intensity                 0
hue                             0
od280/od315_of_diluted_wines    0
proline                         0
target                          0
category                        0
dtype: int64
```

```
# Добавим пропуски
df.loc[df.sample(frac=0.1).index, 'alcohol'] = np.nan
df.loc[df.sample(frac=0.1).index, 'category'] = np.nan

# Заменим пропуски в количественном признаке медианным значением
df['alcohol'].fillna(df['alcohol'].median(), inplace=True)

# Заменим пропуски в категориальном признаке наиболее часто
встречающимся значением
df['category'].fillna(df['category'].mode()[0], inplace=True)

df.isnull().sum()
```

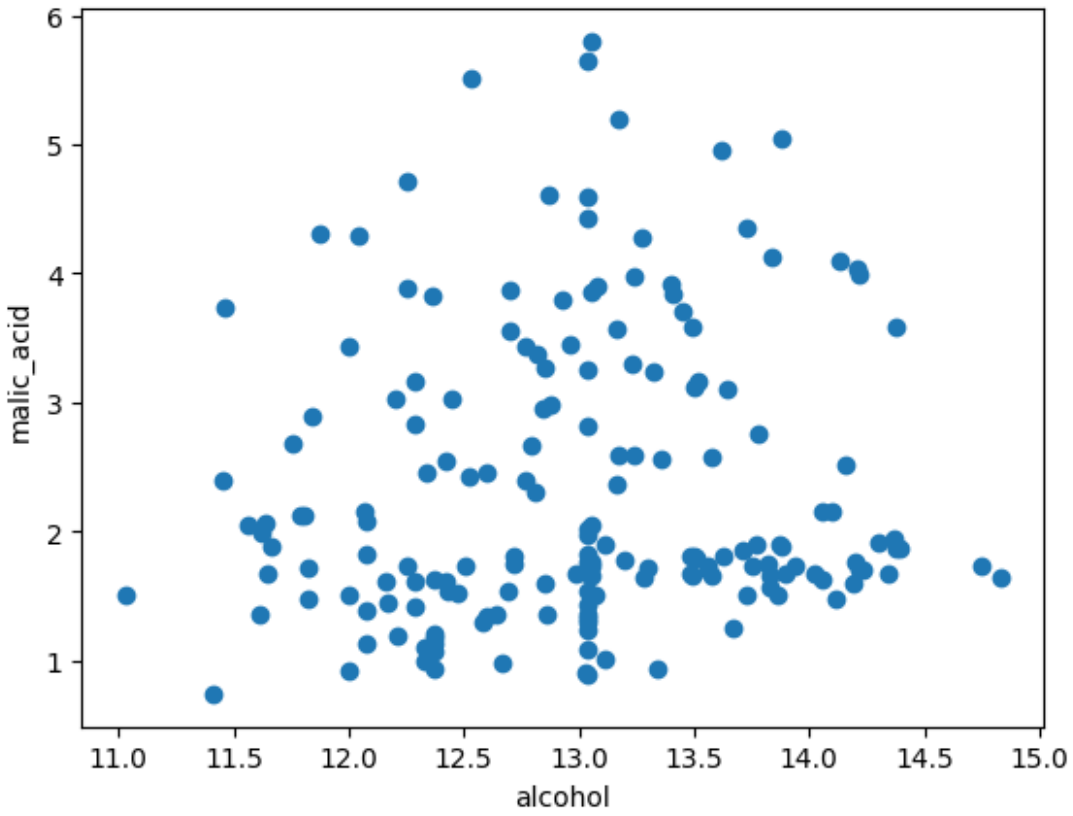
```
alcohol                         0
malic_acid                      0
ash                             0
alcalinity_of_ash               0
magnesium                       0
total_phenols                   0
flavanoids                      0
nonflavanoid_phenols            0
proanthocyanins                 0
color_intensity                 0
hue                             0
od280/od315_of_diluted_wines    0
proline                         0
target                          0
category                        0
dtype: int64
```

Диаграмма рассеивания

```
import matplotlib.pyplot as plt

# Выберем две колонки
col1 = 'alcohol'
col2 = 'malic_acid'

# Построим диаграмму рассеяния
plt.scatter(df[col1], df[col2])
plt.xlabel(col1)
plt.ylabel(col2)
plt.show()
```



Вывод:

Для количественного признака 'alcohol' я использовал метод замены пропусков медианным значением. Этот метод является одним из наиболее распространенных для обработки пропусков в числовых данных, так как медиана минимизирует влияние выбросов и сохраняет распределение данных.

Для категориального признака 'category' была использована замена пропусков наиболее часто встречающимся значением, так как он сохраняет структуру данных и не вносит искусственных значений, которые могут исказить результаты анализа.

Для дальнейшего построения моделей машинного обучения можно использовать целевую переменную 'target' и созданный категориальный признак 'category', а также другие количественные признаки, такие как 'alcohol' и 'malic_acid', которые, имеют влияние на качество вина. Эти признаки могут быть использованы для создания моделей, способных предсказывать классификацию вин на основе их химических свойств.