МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра Алгоритмической математики

ОТЧЕТ

по лабораторной работе №1
по дисциплине «Машинное обучение»
Тема: Исследование набора данных

Студент гр. 1376 _____ Павлова А.С.

Преподаватель _____ Новикова Е.С.

Санкт-Петербург

2023

Цель работы: в ходе выполнения данного задания выбирается набор данных, который будет использоваться в дальнейшем при исследовании алгоритмов кластеризации и классификации.

Задание:

Задание состоит из последовательного выполнения следующих подзадач:

1. Создать Jupyter Notebook, переименовать его «Lab 1, № Группы, ФИО»
2. Выбор исследуемого датасета.
3. Для каждого датасета представить краткое его описание в вашем Jupyter Notebook:
    a. предметная область, источник данных, характер данных (реальные или имитационные)
    b. какие атрибуты представлены в датасете, их тип (числовой, строковый (категории)), что они обозначают
    c. есть ли описание задачи анализа, если есть - представить
4. Для каждого атрибута нужно определить:
    a. среднее значение, ско
    b. построить гистограмму распределения значений, определить есть ли выбросы
    c. есть ли пропущенные значения, сколько
    d. предложить вариант обработки пропущенных значений
5. Определить корреляцию между параметрами
    a. какие атрибуты высоко коррелированы, определить характер корреляции
    b. какие атрибуты не имеют корреляцию
    c. постройте графики рассеивания (предпочтительнее матрицу графиков рассеивания)
    d. проанализировать полученные результаты

Выполнение работы

1. Выбор датасета.
   Мной был выбран «игрушечный датасет» wine_dataset из библиотеки sklearn.

2. Краткое описание датасета.
   Предметная область: виноделие

   Источник данных: https://scikit-learn.org/stable/datasets/toy_dataset.html

   Характер данных: реальные
   Эти данные являются результатами химического анализа вин, выращенных в одном и том же регионе Италии тремя разными культиваторами. Было проведено тринадцать различных измерений, проведенных для различных компонентов, содержащихся в трех типах вина.

   Атрибуты: 13 числовых атрибутов
   Пропущенные данные: отсутствуют

   Alcohol - алкоголь
   Malic acid - яблочная кислота
   Ash - пепел
   Alcalinity of ash - щелочность золы
   Magnesium - магний
   Total phenols - всего фенолов
   Flavanoids - флаваноиды
   Nonflavanoid phenols - нефлаваноидные фенолы
   Proanthocyanins - проантоцианы
   Color intensity - интенсивность цвета
   Hue - оттенок
   OD280/OD315 of diluted wines - OD280 / OD315 разбавленных вин
   Proline - пролин

3. Расчёт среднего значения, стандартного отклонения.

   Для расчёта величин воспользуемся методами mean() и std() библиотеки pandas. Для лучшей визуализации создадим новый датафрейм для рассчитанных значений.

Ввод [7]:
```
wd_vals = pd.concat([wd.mean(), wd.std()], axis = 1)
wd_vals.columns = ['mean', 'std']
wd_vals
```

Out[7]:

	mean	std
alcohol	13.000618	0.811827
malic_acid	2.336348	1.117146
ash	2.366517	0.274344
alcalinity_of_ash	19.494944	3.339564
magnesium	99.741573	14.282484
total_phenols	2.295112	0.625851
flavanoids	2.029270	0.998859
nonflavanoid_phenols	0.361854	0.124453
proanthocyanins	1.590899	0.572359
color_intensity	5.058090	2.318286
hue	0.957449	0.228572
od280/od315_of_diluted_wines	2.611685	0.709990
proline	746.893258	314.907474

4. Построение гистограмм. Определение выбросов
   Воспользуемся библиотекой seaborn. Для большей компактности объединим полученные графики в матрицу.
   Возможными выбросами могут являться отдельные столбцы, выходящие за пределы основного распределения данных. Идентификация выбросов по гистограмме полностью зависит от визуального анализа.

Ввод [16]:
```
sns.set_style('darkgrid')
sns.set_palette('pastel')
fig, axes = plt.subplots(4, 4, figsize = (15, 15), dpi = 200)
fig.suptitle('Histograms')

for i, column in enumerate(wd.columns):
    sns.histplot(wd[column], ax=axes[i//4, i%4], kde = True, color = 'purple')

plt.tight_layout()
plt.show()
```

Воспользуемся ещё одним методом определения выбросов – методом IQR.

Метод межквартильного размаха (IQR) является одним из способов определения выбросов в данных. Для его применения необходимо вычислить медиану и первый и третий квартили в выборке. Затем вычисляется разница между третьим и первым квартилями, которая называется межквартильным размахом (IQR). Далее определяются нижняя и верхняя границы выбросов, которые вычисляются как 1,5*IQR ниже первого квартиля и 1,5*IQR выше третьего квартиля соответственно. Все значения, находящиеся за пределами этих границ, считаются выбросами и могут быть удалены из выборки. Метод IQR более устойчив к наличию выбросов, чем методы, основанные на среднем и стандартном отклонении, поэтому он широко используется для определения выбросов в данных.

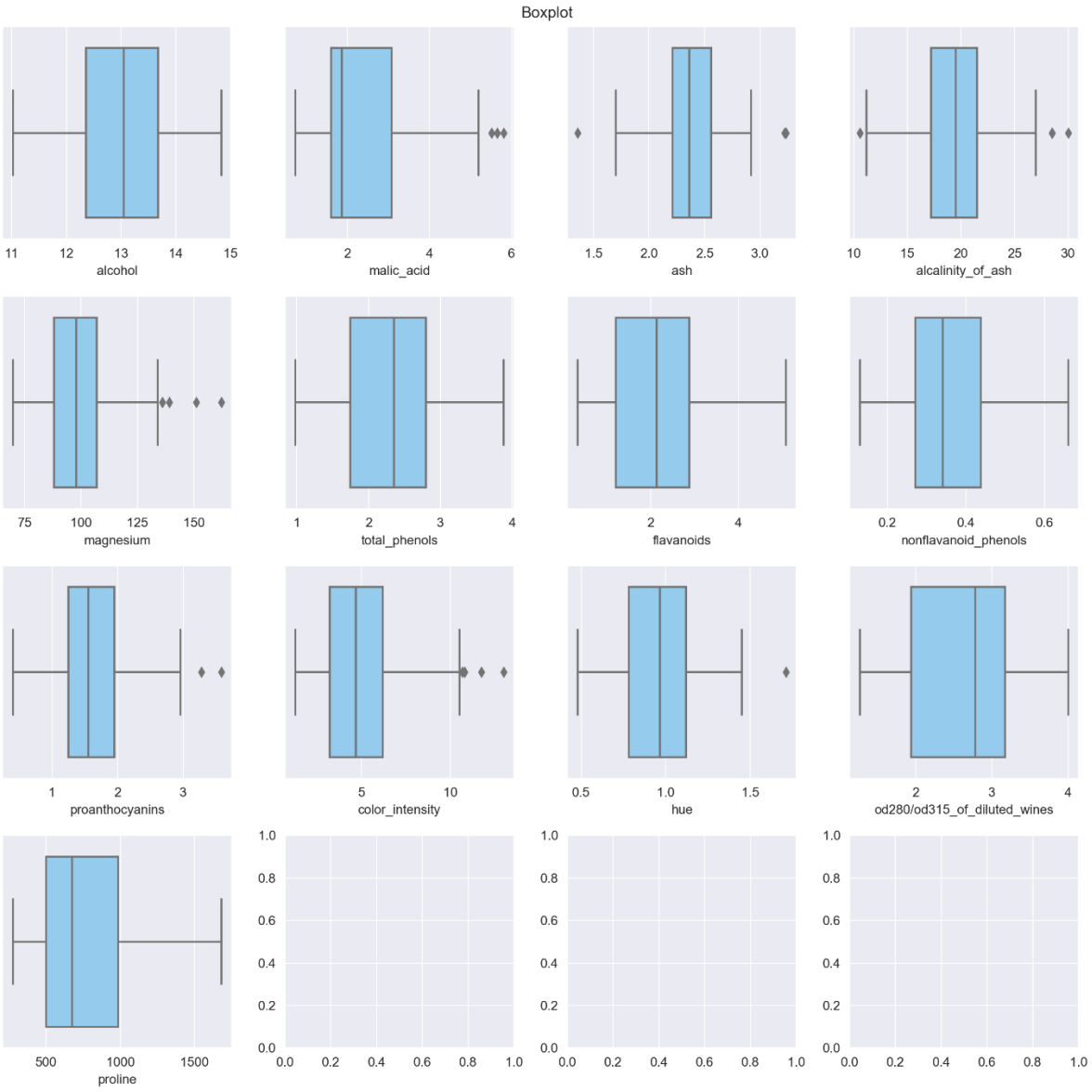
Ввод [9]:
```
outliers = []
for col in wd.columns:
    q1=wd[col].quantile(0.25)
    q3=wd[col].quantile(0.75)
    IQR=q3-q1
    outliers = wd[col][((wd[col]<(q1-1.5*IQR)) | (wd[col]>(q3+1.5*IQR)))]
    print(f'number of outliers in {col} : {str(len(outliers))}', )
    if (len(outliers)) != 0:
        print("value:")
        print(outliers)
    print('-'*50)
```

Полученные данные о выбросах:

```
number of outliers in alcohol : 0
--------------------------------------------------
number of outliers in malic_acid : 3
value:
123    5.80
137    5.51
173    5.65
Name: malic_acid, dtype: float64
--------------------------------------------------
number of outliers in ash : 3
value:
25     3.22
59     1.36
121    3.23
Name: ash, dtype: float64
--------------------------------------------------
number of outliers in alcalinity_of_ash : 4
value:
59     10.6
73     30.0
121    28.5
127    28.5
Name: alcalinity_of_ash, dtype: float64
--------------------------------------------------
number of outliers in magnesium : 4
value:
69    151.0
73    139.0
78    136.0
95    162.0
Name: magnesium, dtype: float64
--------------------------------------------------
number of outliers in total_phenols : 0
--------------------------------------------------
number of outliers in flavanoids : 0
--------------------------------------------------
number of outliers in nonflavanoid_phenols : 0
--------------------------------------------------
number of outliers in proanthocyanins : 2
value:
95     3.28
110    3.58
Name: proanthocyanins, dtype: float64
--------------------------------------------------
number of outliers in color_intensity : 4
value:
151    10.80
158    13.00
159    11.75
166    10.68
Name: color_intensity, dtype: float64
--------------------------------------------------
number of outliers in hue : 1
value:
115    1.71
Name: hue, dtype: float64
--------------------------------------------------
number of outliers in od280/od315_of_diluted_wines : 0
--------------------------------------------------
number of outliers in proline : 0
--------------------------------------------------
```

Дополнительно построим boxplot для наглядности:

```
sns.set_style('darkgrid')
sns.set_palette('pastel')

fig, axes = plt.subplots(4, 4, figsize = (12, 12), dpi = 200)
fig.suptitle('Boxplot')
for i, column in enumerate(wd.columns):
    sns.boxplot(x = wd[column], ax=axes[i//4, i%4], color = 'lightskyblue')

plt.tight_layout()
plt.show()
```



5. Пропущенные значения.
   Для выявления пропущенных значений в датасете воспользуемся методом info() библиотеки pandas.
   Non-Null Count для каждого атрибута одинаков и равен 178, из чего можем сделать вывод, что пропущенных данных нет.

Ввод [20]:
```
wd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   alcohol                       178 non-null    float64
 1   malic_acid                    178 non-null    float64
 2   ash                           178 non-null    float64
 3   alcalinity_of_ash             178 non-null    float64
 4   magnesium                     178 non-null    float64
 5   total_phenols                 178 non-null    float64
 6   flavanoids                    178 non-null    float64
 7   nonflavanoid_phenols          178 non-null    float64
 8   proanthocyanins               178 non-null    float64
 9   color_intensity               178 non-null    float64
 10  hue                           178 non-null    float64
 11  od280/od315_of_diluted_wines  178 non-null    float64
 12  proline                       178 non-null    float64
dtypes: float64(13)
memory usage: 18.2 KB
```

6. Определение корреляции. Построение матрицы рассеивания. Анализ полученных результатов.

   Для определения корреляции воспользуемся методом corr() библиотеки pandas.
   Будем считать высокой корреляцию между параметрами, если по модулю она превышает 0.5.
   Если же корреляция меньше 0.1, будем считать, что она отсутствует.

   Определим сначала высокую корреляцию среди атрибутов, учитывая положительную и отрицательную. С помощью списка pairs и проверки вхождения пар в список, избавимся от повторов.

Ввод [12]:
```
correlation_matrix = wd.corr()

correlation_mask_high = (correlation_matrix.abs() > 0.5) & (correlation_matrix != 1.0)

pairs = []

for col1 in correlation_matrix.columns:
    for col2 in correlation_matrix.columns:
        if correlation_mask_high.loc[col1, col2] and (correlation_matrix.loc[col1, col2] > 0) and ((col1, col2) not in pairs) and
            pairs.append((col1, col2))
            print(f"Высокая положительная корреляция между {col1} и {col2}: {correlation_matrix.loc[col1, col2]}")
        elif correlation_mask_high.loc[col1, col2] and (correlation_matrix.loc[col1, col2] < 0) and ((col1, col2) not in pairs) a
            pairs.append((col1, col2))
            print(f"Высокая отрицательная корреляция между {col1} и {col2}: {correlation_matrix.loc[col1, col2]}")
```

```
Высокая положительная корреляция между alcohol и color_intensity: 0.546364195083705
Высокая положительная корреляция между alcohol и proline: 0.6437200371782134
Высокая отрицательная корреляция между malic_acid и hue: -0.5612956886649447
Высокая положительная корреляция между total_phenols и flavanoids: 0.8645635000951147
Высокая положительная корреляция между total_phenols и proanthocyanins: 0.6124130837800363
Высокая положительная корреляция между total_phenols и od280/od315_of_diluted_wines: 0.6999493647911861
Высокая отрицательная корреляция между flavanoids и nonflavanoid_phenols: -0.5378996119051984
Высокая положительная корреляция между flavanoids и proanthocyanins: 0.6526917686075153
Высокая положительная корреляция между flavanoids и hue: 0.5434785664899897
Высокая положительная корреляция между flavanoids и od280/od315_of_diluted_wines: 0.787193901866951
Высокая отрицательная корреляция между nonflavanoid_phenols и od280/od315_of_diluted_wines: -0.5032695960789114
Высокая положительная корреляция между proanthocyanins и od280/od315_of_diluted_wines: 0.519067095682523
Высокая отрицательная корреляция между color_intensity и hue: -0.5218131932287572
Высокая положительная корреляция между hue и od280/od315_of_diluted_wines: 0.5654682931826589
```

Аналогично определим отсутствие корреляции.

Ввод [21]:
```
correlation_mask_low = (correlation_matrix.abs() < 0.1)
pairs_low = []

for col1 in correlation_matrix.columns:
    for col2 in correlation_matrix.columns:
        if correlation_mask_low.loc[col1, col2]  and ((col1, col2) not in pairs_low) and ((col2, col1) not in pairs_low):
            pairs_low.append((col1, col2))
            print(f"Корреляция отсутствует между {col1} и {col2}: {correlation_matrix.loc[col1, col2]}")
```

```
Корреляция отсутствует между alcohol и malic_acid: 0.09439694091041516
Корреляция отсутствует между alcohol и hue: -0.0717471973301557
Корреляция отсутствует между alcohol и od280/od315_of_diluted_wines: 0.07234318740052098
Корреляция отсутствует между malic_acid и magnesium: -0.05457509608400031
Корреляция отсутствует между ash и proanthocyanins: 0.009651935152086568
Корреляция отсутствует между ash и hue: -0.07466688903277331
Корреляция отсутствует между ash и od280/od315_of_diluted_wines: 0.0039112306302746085
Корреляция отсутствует между alcalinity_of_ash и magnesium: -0.08333308856795228
Корреляция отсутствует между alcalinity_of_ash и color_intensity: 0.018731980931229433
Корреляция отсутствует между magnesium и hue: 0.05539819560300633
Корреляция отсутствует между magnesium и od280/od315_of_diluted_wines: 0.06600393603204628
Корреляция отсутствует между total_phenols и color_intensity: -0.05513641774236631
Корреляция отсутствует между proanthocyanins и color_intensity: -0.025249930815701583
```

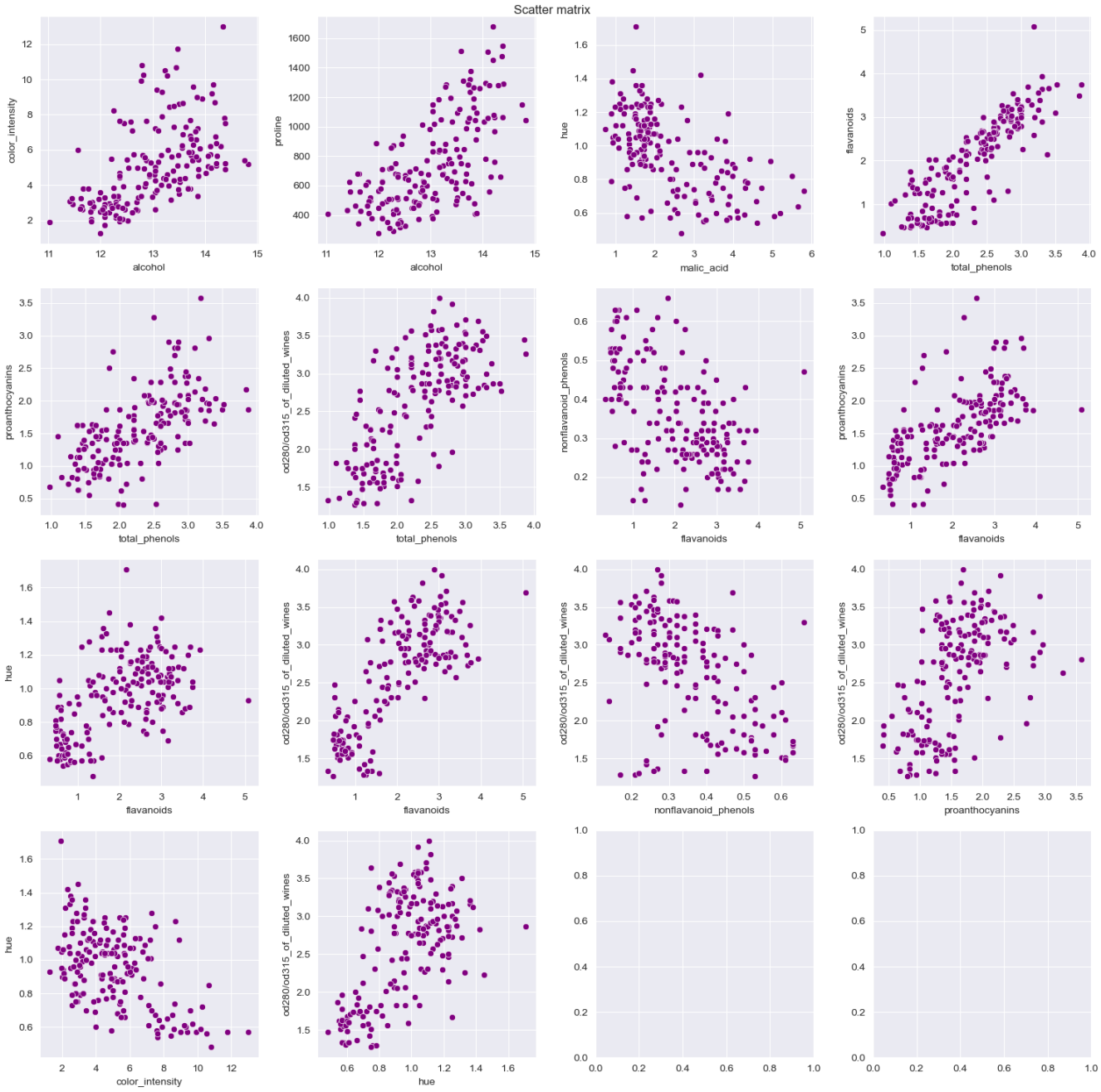
Построение матрицы рассеивания.

Создадим датафрейм, используя список pairs из предыдущего пункта.

Воспользуемся библиотекой seaborn для построения матрицы.

Ввод [19]:
```
subframe_pairs = pd.DataFrame(pairs, columns = ['Attribute1', 'Attribute2'])

print(subframe_pairs)

fig, axes = plt.subplots(4, 4, figsize = (15, 15))
fig.suptitle('Scatter matrix')


for i in range(len(subframe_pairs)):
    sns.scatterplot(x = wd[subframe_pairs.loc[i][0]], y = wd[subframe_pairs.loc[i][1]], ax=axes[i//4, i%4], color = 'purple')

plt.tight_layout()
plt.show()
```

```
              Attribute1                  Attribute2
0                alcohol              color_intensity
1                alcohol                     proline
2             malic_acid                         hue
3          total_phenols                  flavanoids
4          total_phenols              proanthocyanins
5          total_phenols  od280/od315_of_diluted_wines
6             flavanoids         nonflavanoid_phenols
7             flavanoids              proanthocyanins
8             flavanoids                         hue
9             flavanoids  od280/od315_of_diluted_wines
10  nonflavanoid_phenols  od280/od315_of_diluted_wines
11       proanthocyanins  od280/od315_of_diluted_wines
12       color_intensity                         hue
13                   hue  od280/od315_of_diluted_wines
```



Полученные коэффициенты корреляции совпадают с матрицей рассеивания.

Качество вина не зависит от одного параметра напрямую, каждый признак влияет на качество и состав вина.

Например, атрибуты "Alcohol" и "Proline" имеют положительную корреляцию, что может означать, что чем выше содержание алкоголя в вине, тем выше содержание пролина. Также было выяснено, что атрибуты "Ash" и "Magnesium" имеют отрицательную корреляцию, что может означать, что чем выше содержание золы в вине, тем ниже содержание магния. Анализ корреляций между атрибутами позволяет понять взаимосвязь между ними и определить наиболее важные атрибуты для предсказания целевой переменной. Также это может помочь в выборе наиболее эффективных методов анализа данных и построения моделей.