УДК 004.9

ОПЕРАЦИЯ РАЗДЕЛЕНИЯ ВЫБОРКИ ДАННЫХ С ПОМОЩЬЮ ИНСТРУМЕНТОВ PYTHON

Моисеенков Д.С., старший оператор научной роты, II курс
Военная академия связи имени Маршала Советского Союза С.М. Будённого, г. Санкт-Петербург

Набор данных для решения задачи: http://archive.ics.uci.edu/ml/datasets/Iris

Датасет является «неувядающей классикой» машинного обучения. Он представляет собой классическую интерпретируемую постановку задачи классификации [1-5].

Набор данных содержит 3 класса по 50 экземпляров в каждом, где каждый класс относится к типу цветка ириса [6]. Один класс линейно отделим от 2 других; последние два линейно НЕ отделимы друг от друга. Прогнозируемый признак: класс ириса [7].

Атрибутивная характеристика:

1. Id (идентификатор записи);
2. Длина чашелистика (единицы измерения – см);
3. Ширина чашелистика (единицы измерения – см);
4. Длина лепестка (единицы измерения – см);
5. Ширина лепестка (единицы измерения – см);
6. Вид ириса (относит цветок к нужному классу).

Первоначально подключим программные библиотеки в проект решения (рисунок 1).

```
import numpy as np
import pandas as pd

from sklearn import model_selection
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report, roc_curve, roc_auc_score
from pandas.plotting import scatter_matrix
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
```

Рисунок 1 – Загрузка необходимых программных библиотек

Для работы с набором данных как массивом информации (датафреймом) понадобятся библиотеки pandas и numpy [8].

Для построения графиков – библиотека matplotlib и входящая в нее pyplot. Также для настройки стилей графиков понадобится библиотека визуализации seaborn [9].

Для построения модели полиномиальной регрессии и оценки ее точности потребуются библиотеки программных методов sklearn [10].

Следующим шагом необходимо загрузить набор данных в переменную датафрейма и представить его обобщенную структуру из первых пяти элементов при помощи метода head():

```
dataset = pd.read_csv("Iris.csv")
dataset.head()
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

Рисунок 2 – Загрузка набора данных в переменную датафрейма

Следующим шагом исследуем зависимость распределения между парами двух взаимосвязанных классов – SepalLengthCm : SepalWidthCm и PetalLengthCm : PetalWidthCm [11].

Для начала создадим исходный программный код для графической отрисовки распределения (рисунок 3).

```
sns.swarmplot(dataset['SepalLengthCm'], dataset['SepalWidthCm'], dataset['Species'])
plt.suptitle('Распределение между длиной чашелистика и шириной чашелистика в зависимости от класса')
plt.show()

sns.swarmplot(dataset['PetalLengthCm'], dataset['PetalWidthCm'], dataset['Species'])
plt.suptitle('Распределение между длиной лепестка и шириной лепестка в зависимости от класса')
plt.show()
```

Рисунок 3 – Создание программного кода отрисовки графической зависимости. Непосредственно отображение зависимости представлено на рисунках 4 – 5.

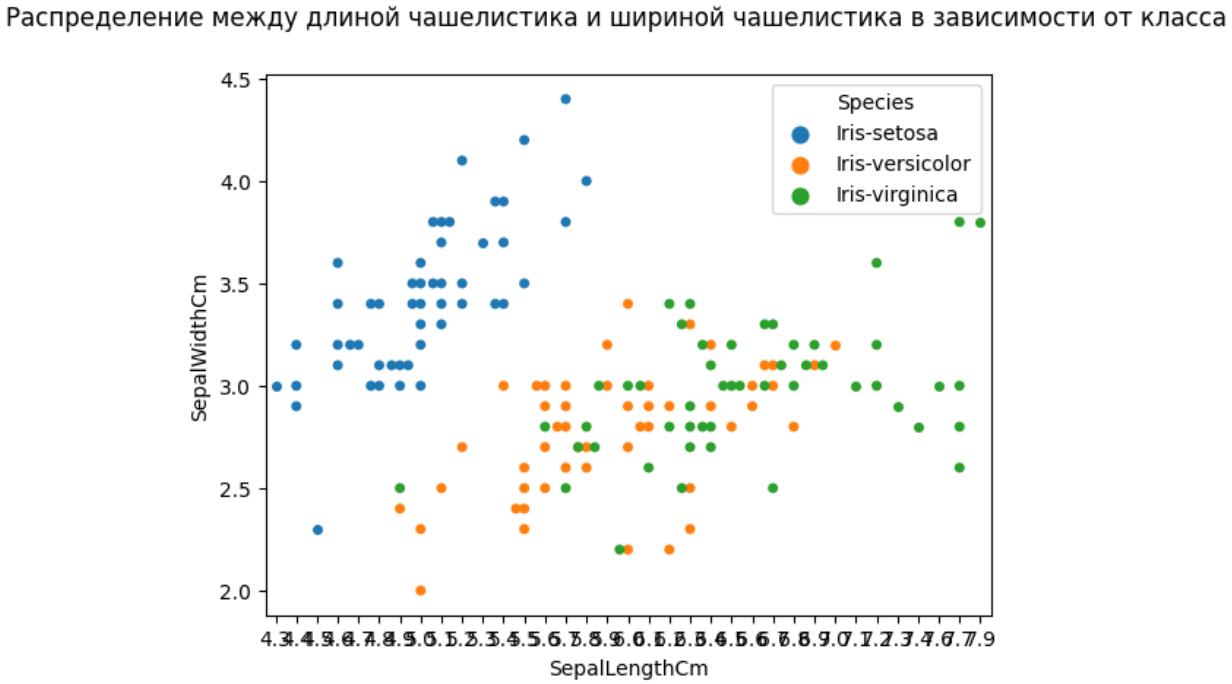

Рисунок 4 – Зависимость распределения между длиной чашелистика [SepalLengthCm] и его шириной [SepalWidthCm].

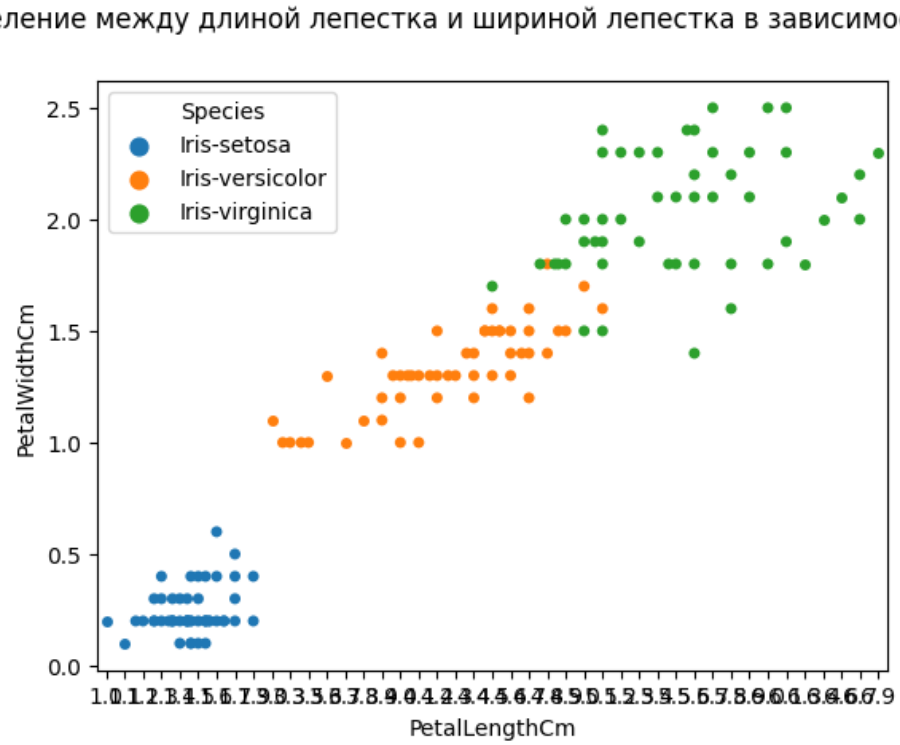

Рисунок 5 – Зависимость распределения между длиной лепестка [PetalLengthCm] и его шириной [PetalWidthCm]. Из рисунков 4 – 5 следует, что пространство распределения данных можно обобщить аппроксимирующей функцией полиномиального характера [1].

Приступим к подготовке модели полиномиальной регрессии. Для начала разделим входные данные на набор признаков и целевой прогнозируемый класс (рисунок 6).

```
array = dataset.values
X = array[:, 0:4]
Y = array[:, 4]
```

Рисунок 6 – Разделение выборки на набор прецедентов и целевую переменную

Таким образом, была реализована операция разделения выборки данных с помощью инструментов Python.

Список литературы:

1. Свидетельство о государственной регистрации программы для ЭВМ № 2023680124 Российская Федерация. BrainPower : № 2023669010 : заявл. 16.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QXBJIM.
2. Математические и программные методы построения моделей глубокого обучения : Учебное пособие / А. В. Протодьяконов и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 176 с. – ISBN 978-5-9729-1484-5. – EDN PZLUAH.
3. Свидетельство о государственной регистрации программы для ЭВМ № 2023680335 Российская Федерация. Maitak Intelligence Natural Language Processing Module : № 2023669704 : заявл. 27.09.2023 : опубл. 28.09.2023 / Р. В. Майтак.
4. Методы восстановления непараметрической регрессии в условиях несбалансированных данных / А. Д. Салычева и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2024. – 192 с. – ISBN 978-5-9729-1856-0. – EDN AAJATW.
5. Свидетельство о государственной регистрации программы для ЭВМ № 2023684619 Российская Федерация. Efficient Network: № 2023684038: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
6. Свидетельство о государственной регистрации программы для ЭВМ № 2023680070 Российская Федерация. Модернизированная модель DBSCAN для определения скрытых взаимосвязей : № 2023668841 : заявл. 13.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN KQUUKF.
7. Асимптотический анализ поведения прикладных моделей машинного обучения : Учебное пособие / А. В. Протодьяконов и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 144 с. – ISBN 978-5-9729-1455-5. – EDN APHQME.

8. Свидетельство о государственной регистрации программы для ЭВМ № 2023684621 Российская Федерация. Destructed Deep Random Forest: № 2023684050: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
9. Свидетельство о государственной регистрации программы для ЭВМ № 2023684622 Российская Федерация. Mask Made AI: № 2023684042: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
10. Свидетельство о государственной регистрации программы для ЭВМ № 2023680103 Российская Федерация. Cognitive Solution : № 2023669189 : заявл. 19.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QEMFJA.
11. Свидетельство о государственной регистрации программы для ЭВМ № 2023684624 Российская Федерация. Программа автоматического распознавания лиц в видеопотоке: № 2023684236: заявл. 15.11.2023: опубл. 16.11.2023 / П. А. Пылов.