МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГАОУ ВО «СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ ИНФОРМАЦИОННЫХ

ТЕХНОЛОГИЙ И ТЕЛЕКОММУНИКАЦИЙ

КАФЕДРА ПРИКЛАДНОЙ ИНФОРМАТИКИ

**Лабораторная работа №1**

по дисциплине

«Искусственный интеллект и машинное обучение»

**Выполнил:**

Сердюков Никита Анатольевич

Студент 2 курса группы \_ПИН-б-о-22-1

Направления подготовки

09.03.03 Прикладная информатика

очной формы обучения

Ставрополь, 2023 г.

Тема: Инструментальные средства машинного обучения.

Цель работы: изучение программных средств для организации рабочего места специалиста по анализу данных и машинному обучению.

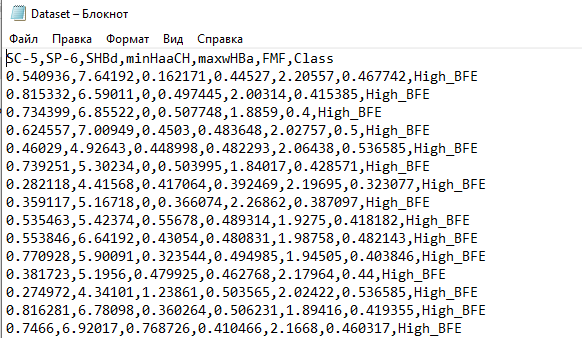
Выполнение работы:

Датасет: Sirtuin6 малые молекулы

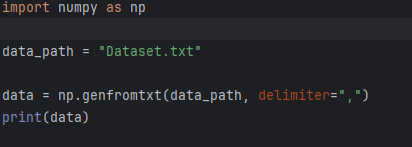
Описание датасета: Набор данных включает 100 молекул с 6 наиболее подходящими дескрипторами для определения потенциальных ингибиторов целевого белка Sirtuin6. Молекулы сгруппированы в зависимости от их низкого и высокого BFE. Признаки белка: SC-5, SP6, SHBd, minHaaCH, maxwHBa, FMF

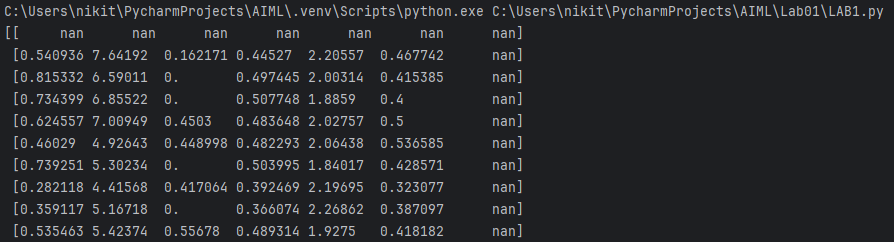
Модель которую можно построить на данном наборе данных: построение модели машинного обучения для предсказания активности молекул в качестве ингибиторов Sirtuin6, т .е является ли молекула дескриптором к белку или нет.

1. Загрузим набор данных Sirtuin6 Small Molecules, запишем в удобном формате в текстовый документ:

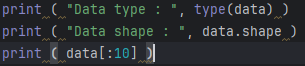


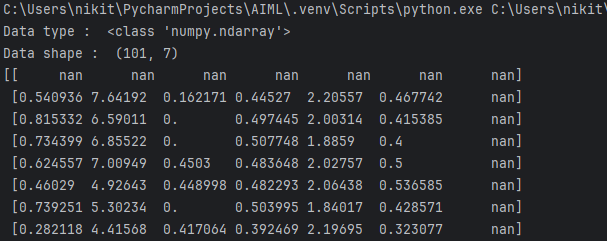
1. С помощью метода genfromtxt преобразуем датасет в ndarray, и выведем его



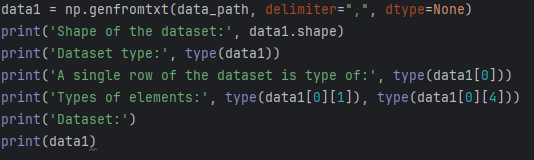


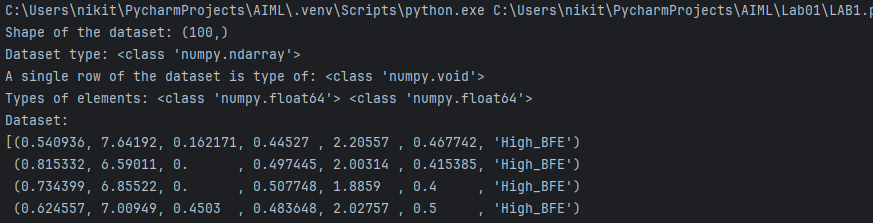
1. Создадим и выведем подвыборку данных для поверхностного обзора исследуемой выборки



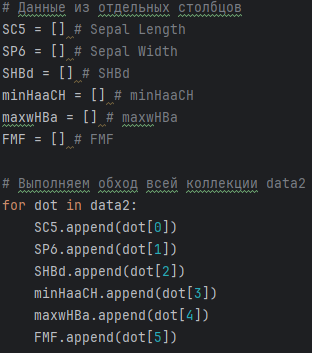


1. Доработаем вывод подвыборки данных. Сделаем вывод для типа строка

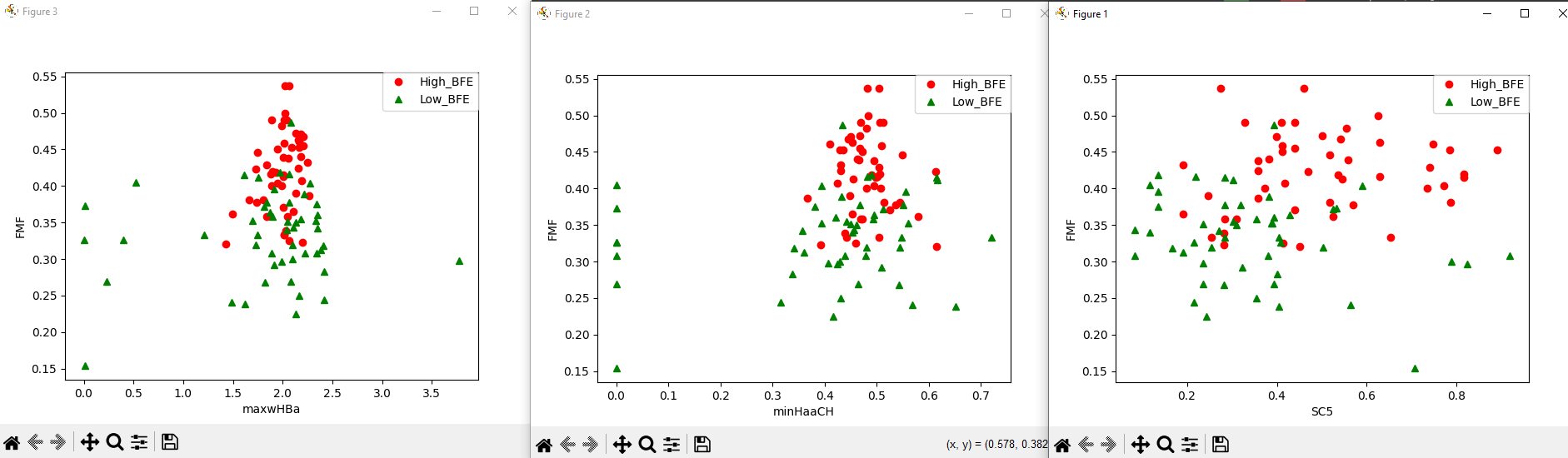




1. Сделаем визуализацию данных







Листинг:

import numpy as np  
import matplotlib as mpl  
import matplotlib.pyplot as plt  
  
data\_path = "Dataset.data"  
  
dt = np.dtype("f8, f8, f8, f8, f8, f8, U30")  
data2 = np.genfromtxt(data\_path, delimiter=",", dtype=dt)  
print('Shape of the dataset:', data2.shape)  
print('Dataset type:', type(data2))  
print('A single row of the dataset is type of:', type(data2[0]))  
print('Types of elements:', type(data2[0][1]), type(data2[0][6]))  
print('Dataset slice:')  
print(data2[:10])  
  
# Данные из отдельных столбцов  
SC5 = [] # SC5  
SP6 = [] # SP6  
SHBd = [] # SHBd  
minHaaCH = [] # minHaaCH  
maxwHBa = [] # maxwHBa  
FMF = [] # FMF  
  
# Выполняем обход всей коллекции data2  
for dot in data2:  
 SC5.append(dot[0])  
 SP6.append(dot[1])  
 SHBd.append(dot[2])  
 minHaaCH.append(dot[3])  
 maxwHBa.append(dot[4])  
 FMF.append(dot[5])  
  
# Строим графики по проекциям данных  
# Учитываем, что каждые 50 типов идут последовательно  
plt.figure(1)  
High\_BFE, = plt.plot(SC5[:50], FMF[:50], 'ro', label='High\_BFE')  
Low\_BFE, = plt.plot(SC5[50:100], FMF[50:100], 'g^', label='Low\_BFE')  
plt.legend(bbox\_to\_anchor=(1, 1), loc=1, borderaxespad=0.)  
plt.xlabel('SC5')  
plt.ylabel('FMF')  
  
plt.figure(2)  
High\_BFE, = plt.plot(minHaaCH[:50], FMF[:50], 'ro', label='High\_BFE')  
Low\_BFE, = plt.plot(minHaaCH[50:100], FMF[50:100], 'g^', label='Low\_BFE')  
plt.legend(bbox\_to\_anchor=(1, 1), loc=1, borderaxespad=0.)  
plt.xlabel('minHaaCH')  
plt.ylabel('FMF')  
  
plt.figure(3)  
High\_BFE, = plt.plot(maxwHBa[:50], FMF[:50], 'ro', label='High\_BFE')  
Low\_BFE, = plt.plot(maxwHBa[50:100], FMF[50:100], 'g^', label='Low\_BFE')  
plt.legend(bbox\_to\_anchor=(1, 1), loc=1, borderaxespad=0.)  
plt.xlabel('maxwHBa')  
plt.ylabel('FMF')

plt.show()

Контрольные вопросы:

1. Для организации рабочего места специалиста Data Science используются инструменты для работы с данными, такие как Jupyter Notebook, IDE (например, PyCharm), системы контроля версий (например, Git), базы данных (например, SQL), библиотеки Python для анализа данных и машинного обучения.

2. Некоторые популярные библиотеки Python для работы в области машинного обучения:

- NumPy: Библиотека для работы с массивами и матрицами, предоставляющая высокопроизводительные математические функции.

- Pandas: Инструмент для обработки и анализа данных, предоставляющий удобные структуры данных и функции для работы с таблицами.

- Scikit-learn: Библиотека с открытым исходным кодом для машинного обучения, предоставляющая инструменты для классификации, регрессии, кластеризации и пр.

3. Библиотеки Python получили широкое распространение при реализации систем машинного обучения из-за следующих причин:

- Обширное сообщество разработчиков: Python имеет активное сообщество разработчиков, которые создают и поддерживают множество библиотек для машинного обучения.

- Простота использования: Python предлагает простой и понятный синтаксис, что делает его доступным для начинающих и опытных специалистов.

- Мощные библиотеки: В Python есть множество высококачественных библиотек для работы с данными и машинным обучением, так что разработчики могут быстро создавать и тестировать модели.

Вывод: изучил программные средства для организации рабочего места специалиста по анализу данных и машинному обучению.