МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра «Измерительно-вычислительные комплексы»

«Методы искусственного интеллекта»

Классификация. Метод k ближайших соседей (k-nn)

Отчёт по лабораторной работе №4

Вариант №12

Выполнила:

студентка группы ИСТбд-42

Кучина Анна

Проверил:

доцент кафедры ИВК, к.т.н.

Шишкин В. В.

Ульяновск

2022

Задание:

1. Создать симулированный набор данных и записать его на диск в виде csv файла со следующими параметрами:

- продукт;

- сладость;

- хруст;

-класс.

|  |  |  |  |
| --- | --- | --- | --- |
| продукт | сладость | хруст | класс |
| Яблоко | 7 | 7 | Фрукт |
| салат | 2 | 5 | Овощ |
| бекон | 1 | 2 | Протеин |
| банан | 9 | 1 | Фрукт |
| орехи | 1 | 5 | Протеин |
| рыба | 1 | 1 | Протеин |
| сыр | 1 | 1 | Протеин |
| виноград | 8 | 1 | Фрукт |
| морковь | 2 | 8 | Овощ |
| апельсин | 6 | 1 | Фрукт |

Подготовить для классификации несколько примеров в соответствии с рисунком

Хруст

Фрукты

Овощи

Протеин

Сладость

2. Запрограммировать метрический классификатор по методу k-NN. Для проверки решить ту же задачу методом k-NN библиотеки sklearn.

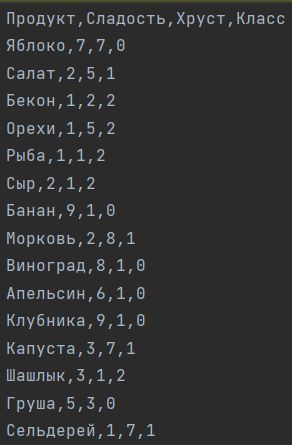
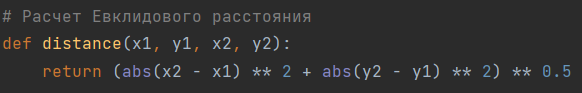
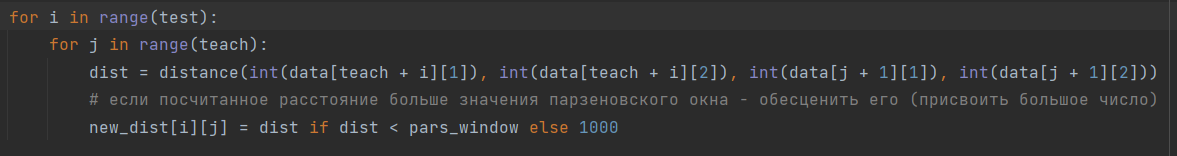
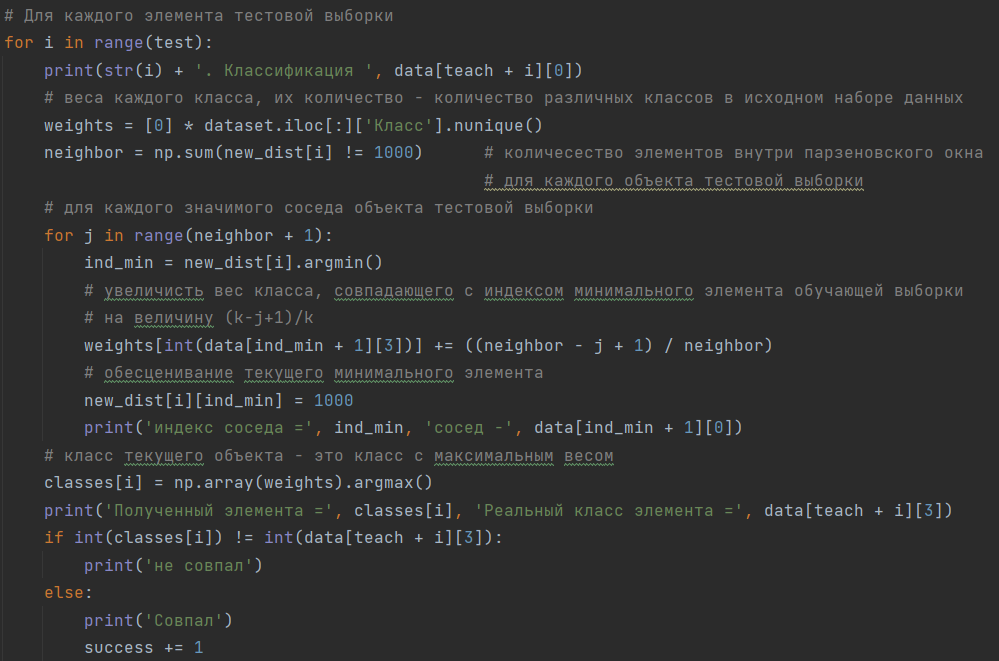
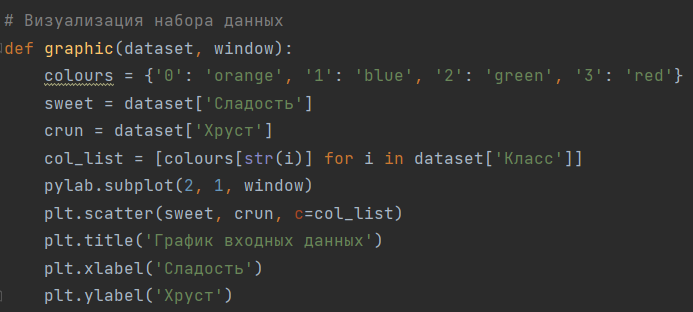
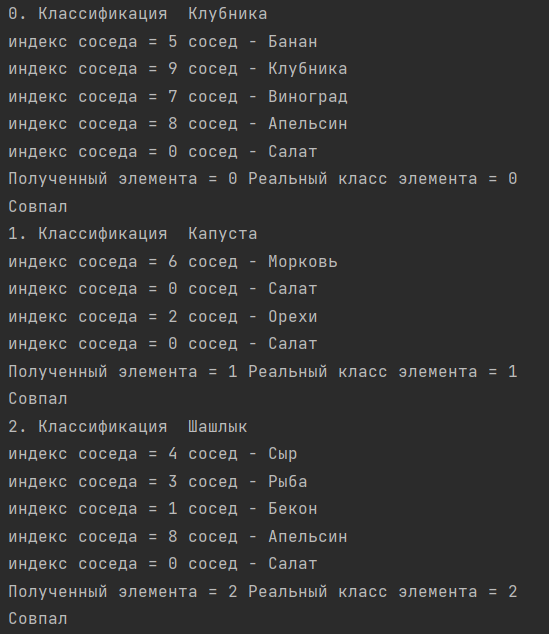
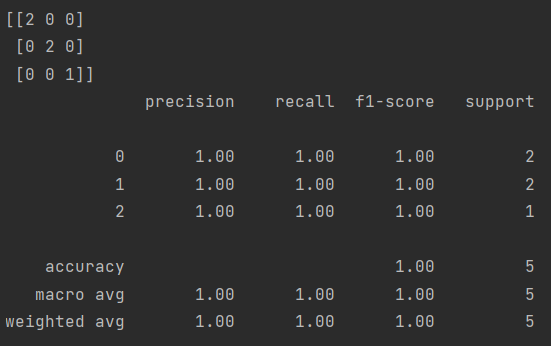
3. Прочитать сгенерированный набор данных. Настроить классификатор. Провести эксперимент по классификации с контролем для подготовленных примеров.

4. По возможности результаты визуализировать.

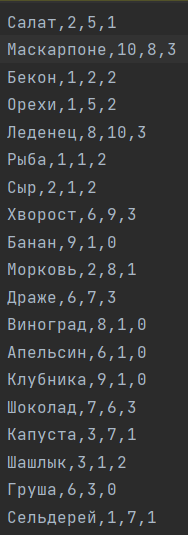
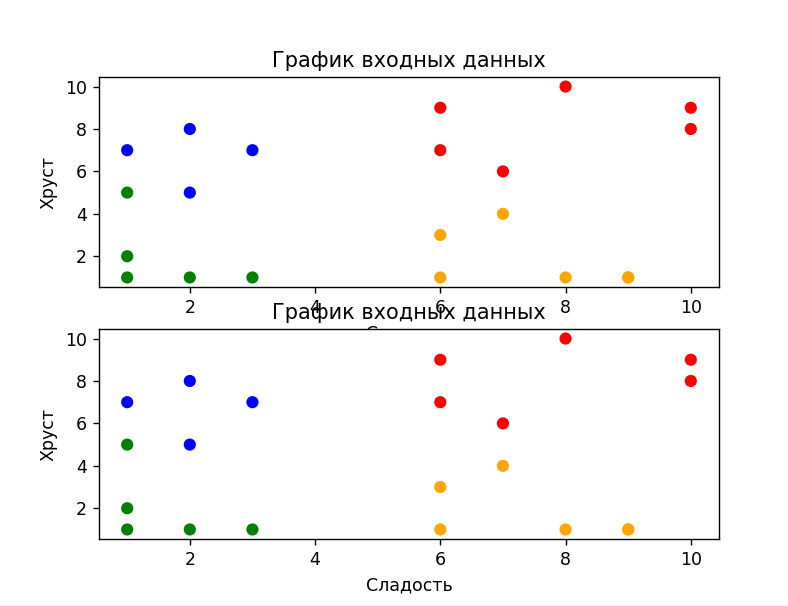
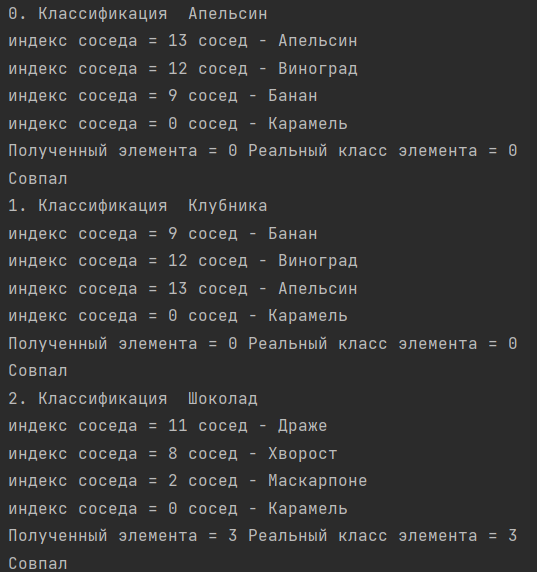
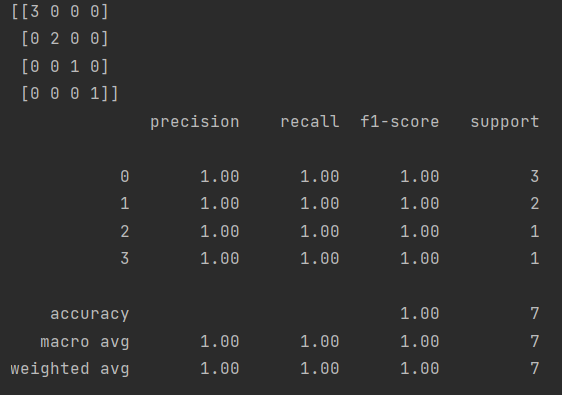
5. Ввести в набор данных и примеры продукты еще одного класса (возможно изменив набор параметров) и повторить эксперимент.

6. По результатам эксперимента подготовить отчет по лабораторной работе.

**Часть 1**

1. CSV файл с симулированным набором данных для трех классов, созданный по схеме из задания:  
     
   Первые десять элементов относятся к обучающей выборке, остальные пять – к тестовой. У них также был указан класс, чтобы после применения метода классификации сравнить его с исходными данными.
2. Написание алгоритма классификации K-NN, а точнее его частного вида с применением Парзеновского окна и весов расстояний:
   1. Функция для расчета Евклидового расстояния между двумя точками на плоскости по их координатам:  
      
   2. Вычисление расстояний от каждого элемента тестовой выборки до каждого элемента обучающей выборки и заполнение матрицы расстояний. Применение Парзеновского окна и обесценивание всех элементов, которые оказались дальше указанного радиуса окна:  
      
   3. Реализация алгоритма классификации k-nn для элементов тестовой выборки:  
      
3. Визуализация результатов:  
     
     
     
   Графики исходной выборки (из CSV файла) и данные с применением запрограммированного вручную классификатора:
4. Первые три объекта тестовой выборки с их ближайшими соседями, полученным и реальным классом, а также классы всех объектов тестовой выборки, полученные с помощью классификатора k-nn:  
     
     
   
5. Классификация объектов тестовой выборки с помощью библиотеки sklearn и визуализация результатов в виде матриц классификации и статистики ошибок:  
     
   Из матрицы ошибок видно, что среди классифицируемых объектов есть два объекта, которые были отнесены к первому классу («0»), два объекта – ко второму («1»), и один – к третьему («2»). В действительности мы имеем такое же количество объектов каждого класса. Можно сказать, что эксперимент прошел успешно.

**Часть 2**

1. Набор данных для второго эксперимента. Были добавлены несколько новых продуктов, и, соответственно, новый класс для них. В наборе 18 элементов, из которых 11 относятся к обучающей выборке, а 7 – к тестовой. Набор данных из CSV файла для второго эксперимента:  
   
2. Визуализация полученных результатов классификации для второго эксперимента:  
   
3. Первые несколько элементов с информацией о ближайших соседях, исходных и полученных классах, а также список полученных классов и количество совпавших:  
     
   
4. Классификация этих же данных с помощью библиотеки sklearn и визуализация результатов в виде матрицы классификации и матрицы статистики:  
   

В ходе проведенных экспериментов было выявлено, что оптимальным размером Парзеновского окна при таком количестве данных обучающей выборки является число 4 в единицах системы координат.

Все эксперименты прошли успешно, ошибок классификации нет.

Одной из самых больших трудностей стал принцип выбора веса расстояния. Неизвестно, насколько точна данная величина, однако в проведенных экспериментах она не дала ошибки.