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

ОТЧЕТ

по лабораторной работе №2

по дисциплине «Машинное обучение»

Тема: Кластеризация

Студентка гр. 1373 _____ Новикова А.С.

Преподаватель _____ Татчина Я.А.

Санкт-Петербург

2023

Цель работы: познакомиться с методом кластеризации K-mean с помощью пакета sklearn.

1. Краткое описание датасета

Для этой работы я взяла другой набор данных, т.к. мне показалось, что датасет из 1 работы не подойдёт для этой.

Физические характеристики крабов, найденных в районе Бостона. Они нужны для того, чтобы предугадать возраст краба.

Все большая часть данных числовые. Есть один строковый тип - пол. В датасете представлены следующие атрибуты:

- Sex - пол краба
- Length - длина в футах
- Diameter - диаметр в футах
- Height - высота в футах
- Weight - вес в унциях
- Shucked Weight - вес без панциря в унциях
- Viscera Weight - вес внутренностей в унциях
- Shell Weight - вес панциря в унциях
- Age – возраст

2. Добавить новый атрибут

Так как почти все характеристики представлены в футах, я добавила атрибут «длина в метрах», для этого надо было умножить столбец Length на 0.305

3. Первичная обработка данных

- Дубликатов в наборе данных не оказалось
- Пропущенных значений в наборе данных не оказалось
- Выбросы в наборе данных есть, они изображены на рис. 1. Было обнаружено 2 аномальных значения. Для остальных выбросов будем считать, что просто есть крабы немного нестандартных размеров.

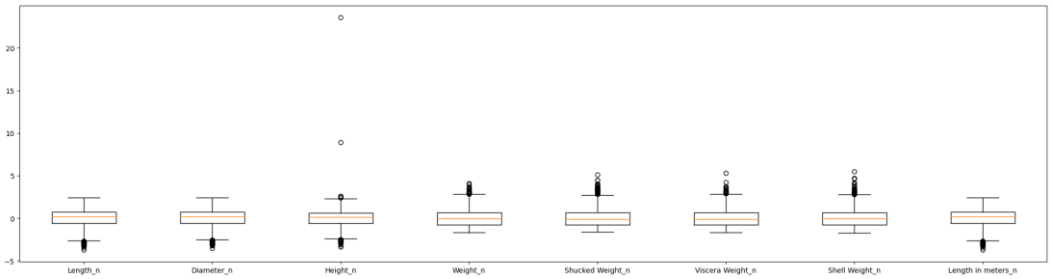

Рисунок 1. Выбросы в наборе данных

Также были найдены нулевые значения. Аномальные и нулевые значения были опустошены, а затем импутированы с помощью k метода ближайших соседей. После этого график выбросов изменился, это можно увидеть на рис. 2.

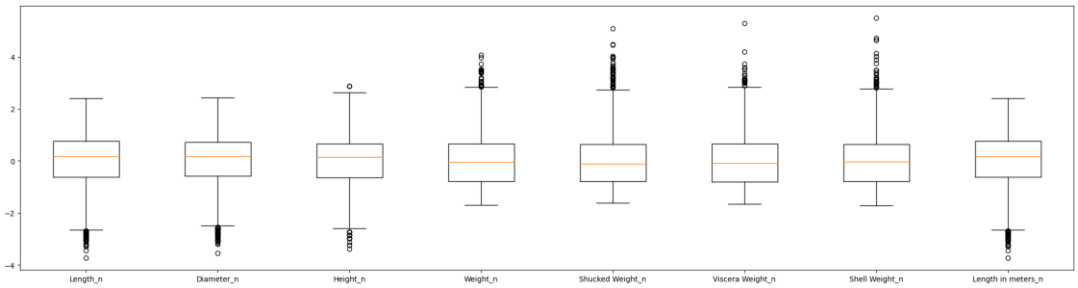

Рисунок 2. Выбросы в наборе данных после применения k метода ближайших соседей.

4. Построить графики зависимости одной переменной от другой

Были рассмотрены зависимости четырех атрибутов: возраст, вес, рост и длина. Графики можно увидеть на рис. 3.

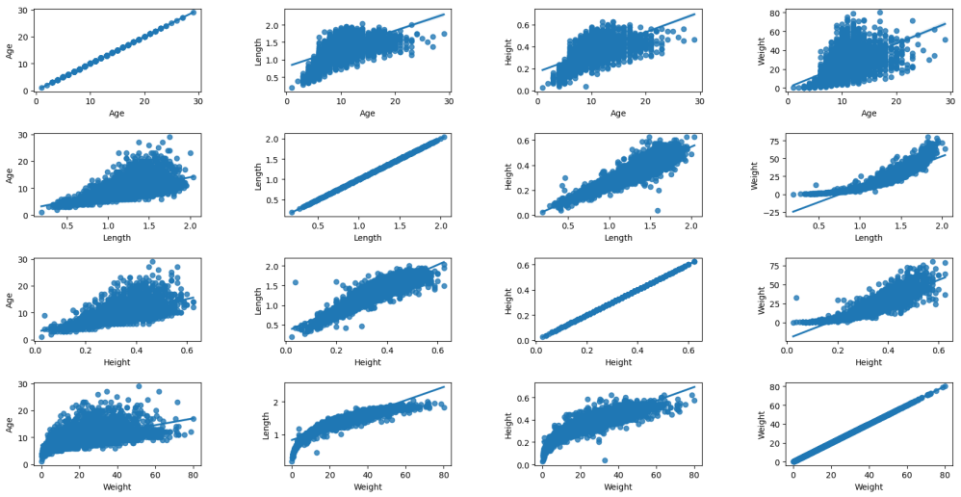

Рисунок 3. Графики зависимостей

Затем были построены те же графики, но в зависимости от пола. Они изображены на рис. 4 – 9.

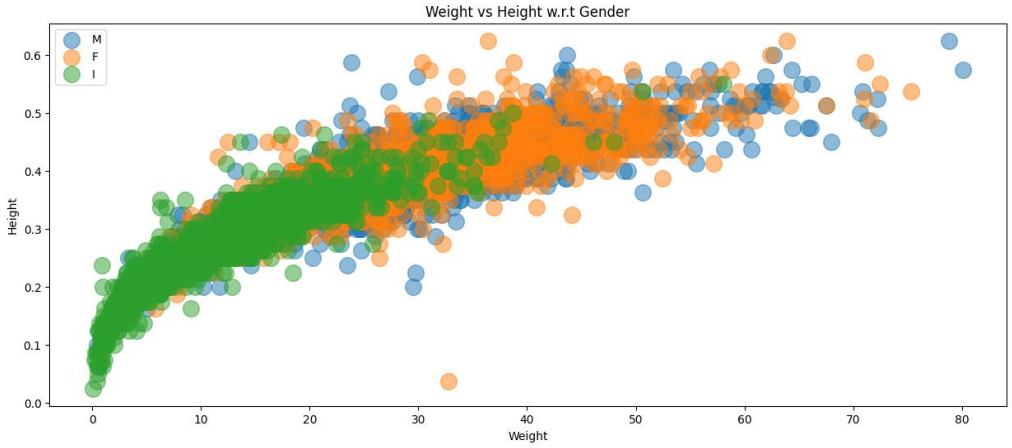

Рисунок 4. График зависимости веса от роста в зависимости от пола.

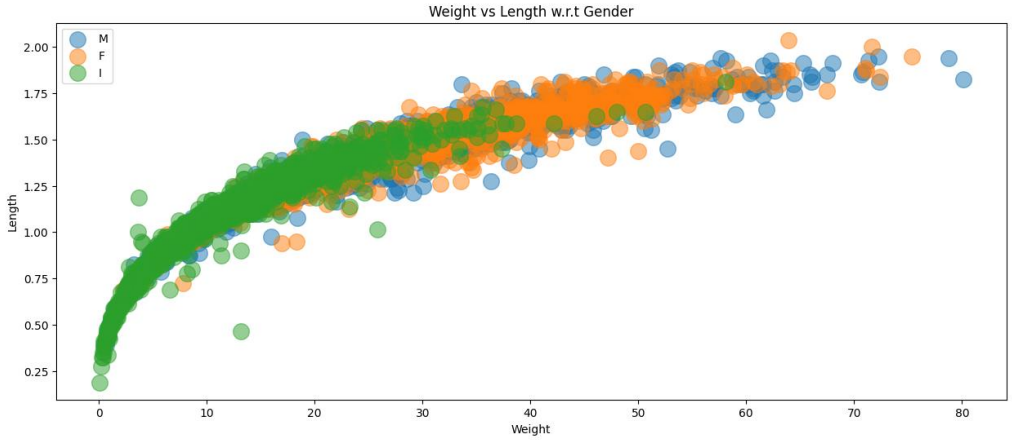

Рисунок 5. График зависимости веса от длины в зависимости от пола.

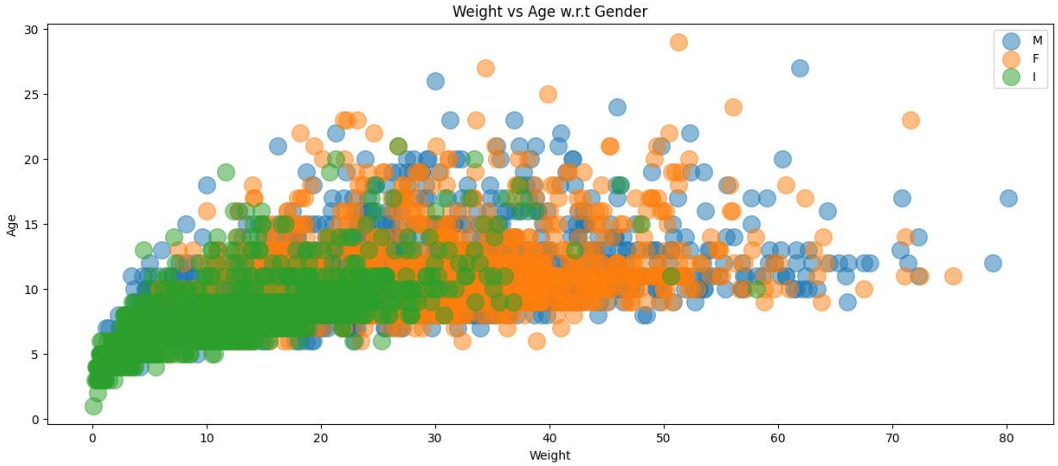

Рисунок 6. График зависимости веса от возраста в зависимости от пола.

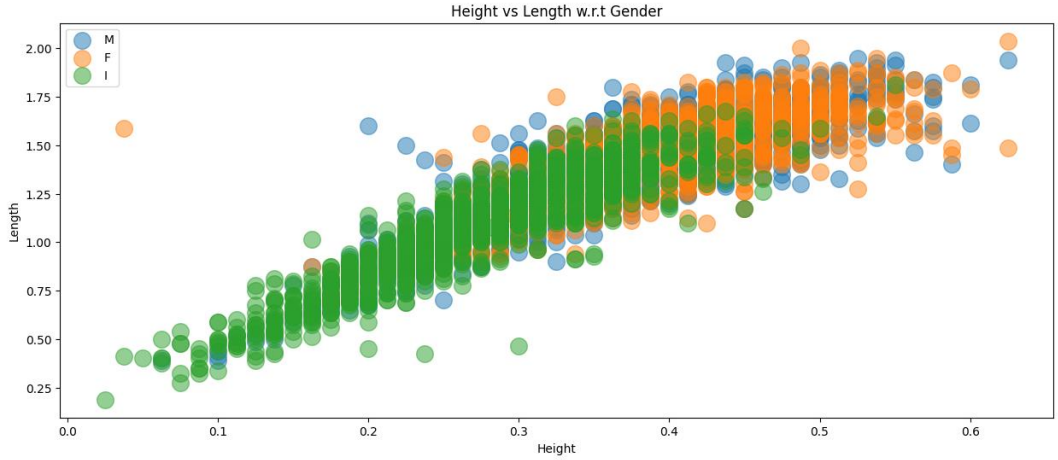

Рисунок 7. График зависимости роста от длины в зависимости от пола.

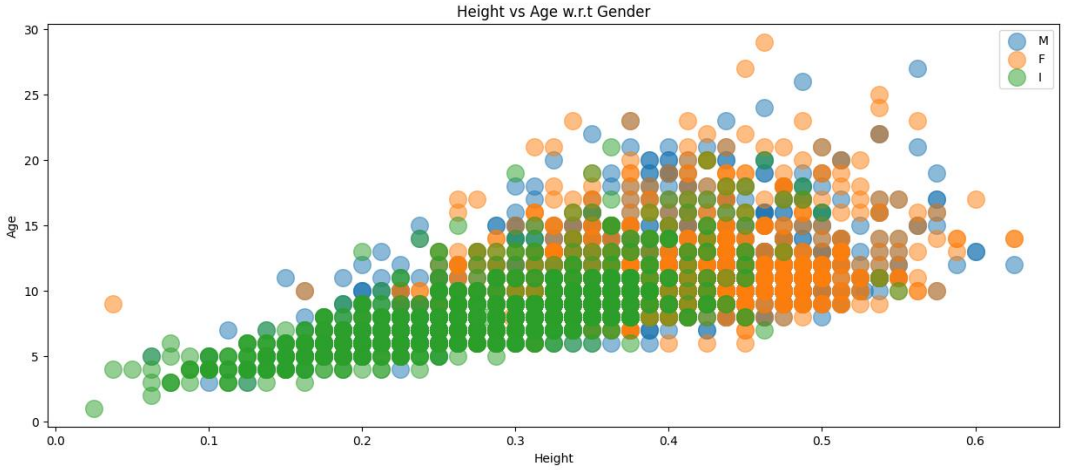

Рисунок 8. График зависимости роста от возраста в зависимости от пола.

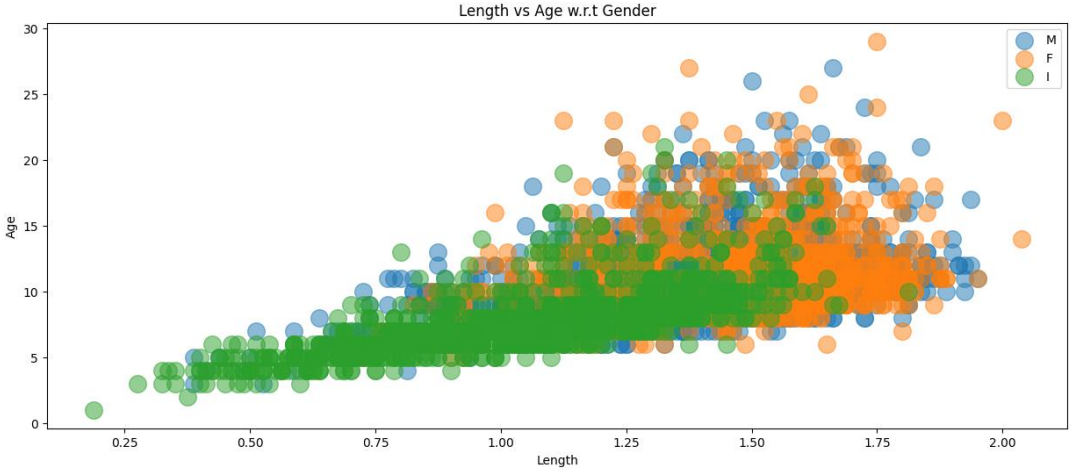

Рисунок 9. График зависимости длины от возраста в зависимости от пола.

Мне кажется, что набор данных можно разделить на группы по полу, так как на графиках отчётливо видно, что маленькие физические показатели у крабов с неопределённым полом, со средними показателями у крабов женского пола, а большие – у крабов мужского пола.

Также я думаю, что на графиках, изображенных на рис. 4, 5 и 6, можно увидеть, что набор данных можно разделить на группы в зависимости от веса. Например, группа с маленьким весом от 0 до 20, со средним весом от 20 до 40 и с большим весом от 40.

5. Применить к датасету метод KMeans

Графики изображены на рис. 10 – 15.

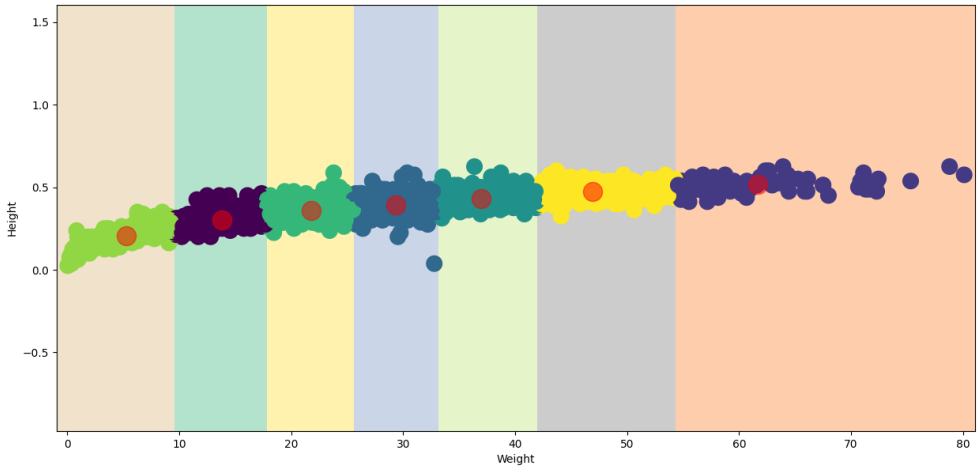

Рисунок 10. Кластеризация с использованием роста и веса

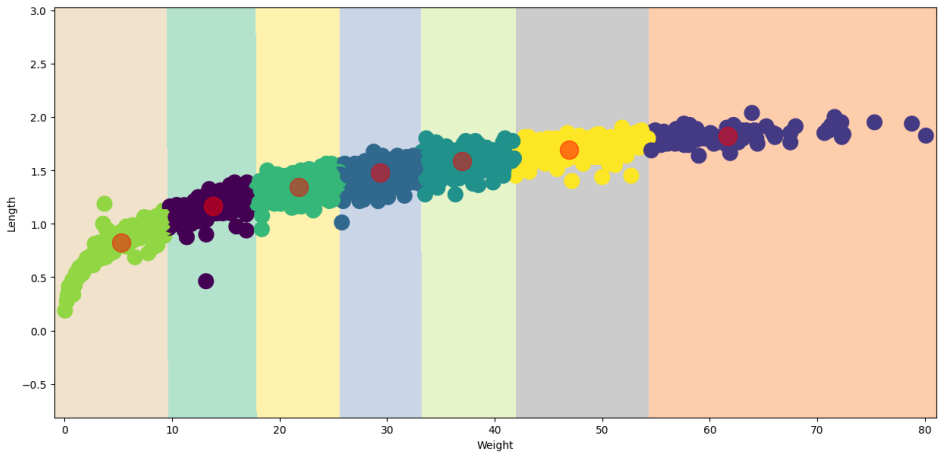

Рисунок 11. Кластеризация с использованием длины и веса

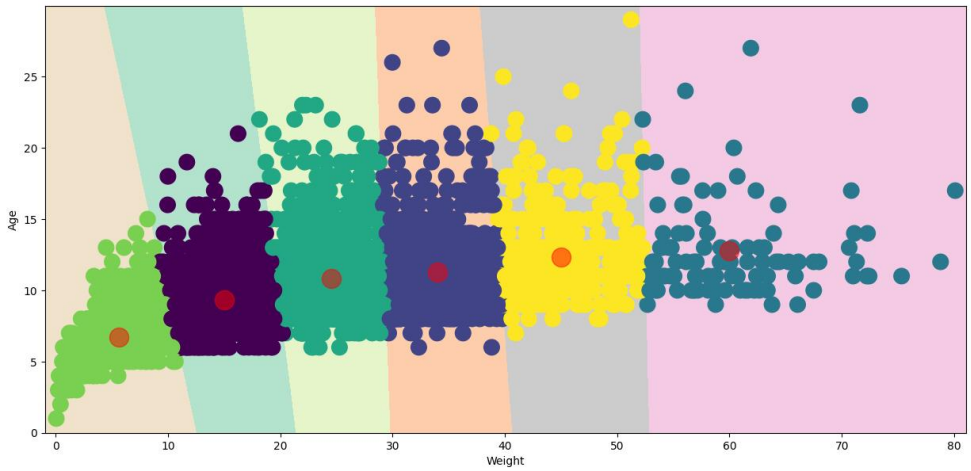

Рисунок 12. Кластеризация с использованием возраста и веса

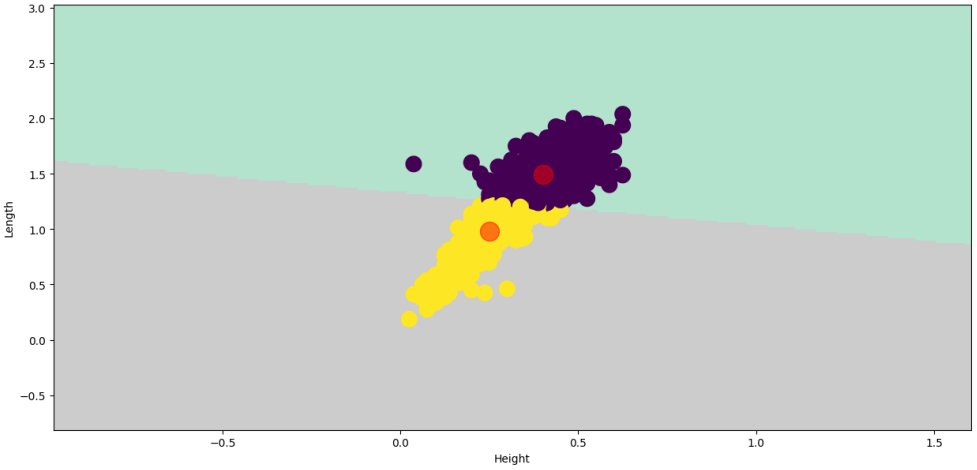

Рисунок 13. Кластеризация с использованием длины и роста

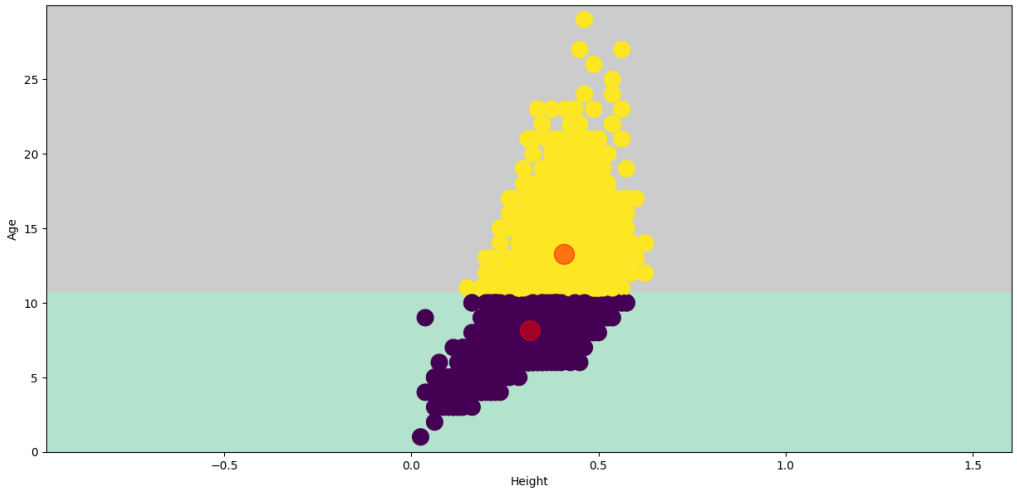

Рисунок 14. Кластеризация с использованием возраста и роста

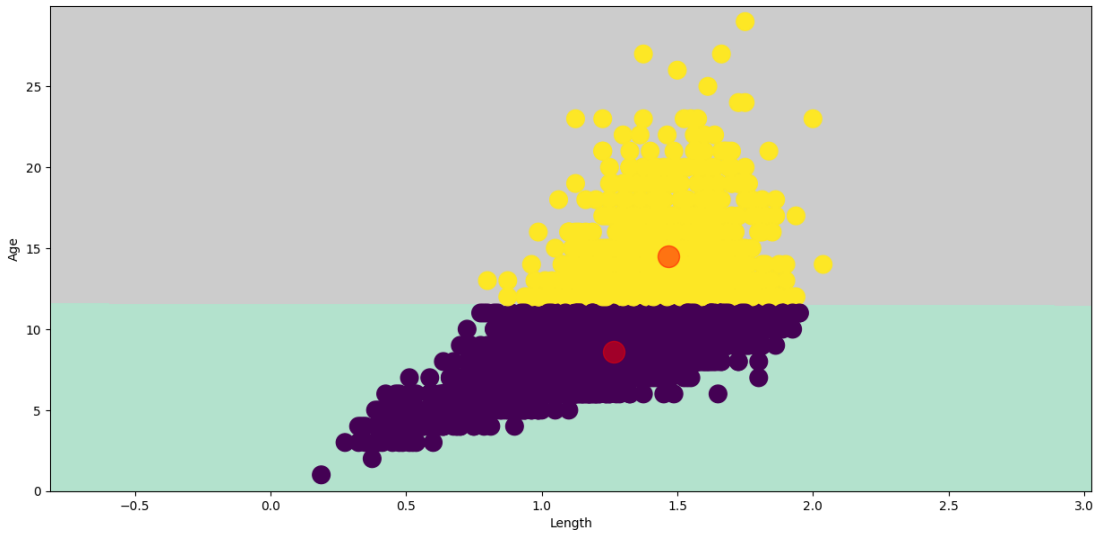

Рисунок 15. Кластеризация с использованием длины и возраста

6. Выводы

В ходе этой работы была проведена первичная обработка набора данных: были удалены аномальные и нулевые значения и импутированы новые с помощью k метода ближайших соседей. Пропущенных значений и дубликатов в наборе данных не оказалось.

Также были построены графики зависимостей в зависимости от пола, по которым можно сделать вывод, что если физические показатели краба довольно малы, то определить его пол ещё нельзя. Чем больше показатели, тем больше вероятность, что краб мужского пола. Краб со средними физическими показателями, скорее всего, будет женского пола.

Был изучен метод кластеризации KMeans, с помощью которого мы попробовали разделить набор данных на кластеры.