Министерство науки и высшего образования Российской Федерации  
Пензенский государственный университет  
Кафедра вычислительная техника

**ОТЧЕТ**

по лабораторной работе №5

по дисциплине «Программные средства кибертехнических систем»

на тему «Изучение набора данных»

Выполнили:

студенты группы 22ВВП1

Хоссейни Нежад С.А.С.М.

Амиров И.Р.

Сергунов М. Р.

Приняли:

Зинкин С.А.

Карамышева Н.С.

Пенза 2025

**Цель работы**

Изучить набор данных

**Задание**

1. Для полученного набора данных и сделанного в лабораторной работе 1 описания провести предварительный анализ данных в соответствии с этапами 2-и и 3-й фазы методологии CRISP.

2. Описать имеющиеся данные.

3. Проверить гипотезы о наличии пропусков и выбросов в данных.

4. Выделить наиболее релевантные решаемой задаче признаки.

5. Сформировать предложения о дополнительном сборе данных (если это необходимо и возможно).

6. Оформить отчёт, содержащий информацию о выполнении пунк-тов 1-5 задания.

Вариант 7

Задача определения принадлежности пикселя на фотографии к коже человека или другим объектам (4-й столбец набора данных {1 – кожа, 2 – не кожа}) по его цвету в кодировке RGB:

Blue – первый столбец {0...255}

Green – второй столбец {0...255}

Red – третий столбец {0...255}

Ссылка на репозиторий: https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation

**Ход работы**

Так как данные были взяты с открытого источника они относятся к категории сторонних. Количество экземпляров: 245 057. Частотные характеристики получены с помощью Wavelet Transform.

Таблица содержит 5 столбцов:

1 – Дисперсия вейвлет преобразования изображения.

2 – Асимметрия вейвлет преобразования изображения.

3 – Эксцесс вейвлет преобразования изображения.

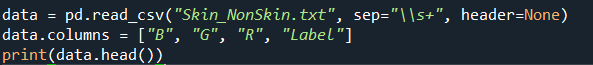
4 – Энтропия изображения.

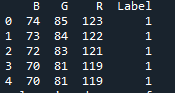
5 – Целевой класс. Подлинная купюра или поддельная.

Нашей задачей будет определить и выявить закономерности позволяющие отличить настоящие банкноты от фальшивых.

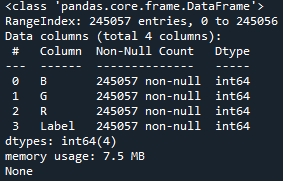
**Анализ данных**

Рассмотрим данные и общие сведения о них.



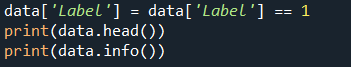


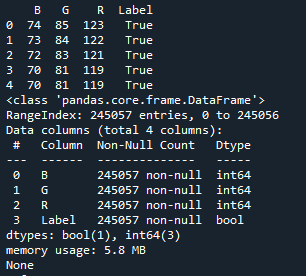




Видим, что количество записей в каждой столбце равно 245 057, а значит пропусков в данных нет. Помимо этого, в таблице нет категориальных признаков.

Приведём целевой параметр к булевой переменной и уменьшим размер датасета.





Мы уже знаем, что пропусков в данных нет, а есть ли дубликаты? Проверим это с помощью duplicated.





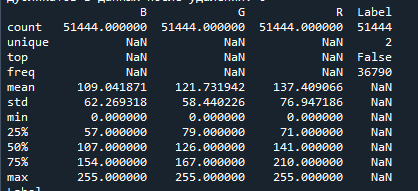
Действительно, 193 613 строки данных оказались дубликатами. Удалим их.





Теперь рассмотрим общие сведения о каждом параметре по отдельности.





Мы видим средние значения, стандартное отклонение, минимальные и максимальные значения для числовых признаков.

По этим данным сложно определить, какие параметры влияют на определение принадлежности класса, поэтому рассмотрим параметры для каждого класса по отдельности.

Сначала определим отношение количества фальшивых банкнот от подлинных.



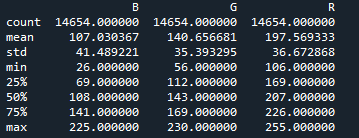


Таким образом, отношение 14654 к 36790 можно приблизительно выразить как 2/5.

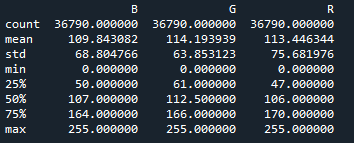
Это приближенное соотношение дает представление о том, что пикселей с кожей в 2 раза меньше, чем пикселей без кожи.

Детальнее исследуем каждый класс по отдельности.









### **Пиксели, относящиеся к коже (Label = True):**

* **Среднее значение:**
  + Красный канал (R): 197.57
  + Зеленый канал (G): 140.66
  + Синий канал (B): 107.03
* **Диапазон значений:** пиксели кожи имеют значения цветов в диапазоне от 26 до 255, что свидетельствует о ярких и насыщенных цветах.
* **Распределение значений:** большинство пикселей имеют высокие значения (75-й процентиль для R — 141, для G — 169, для B — 226), что указывает на преобладание светлых оттенков кожи.

### **Пиксели, не относящиеся к коже (Label = False):**

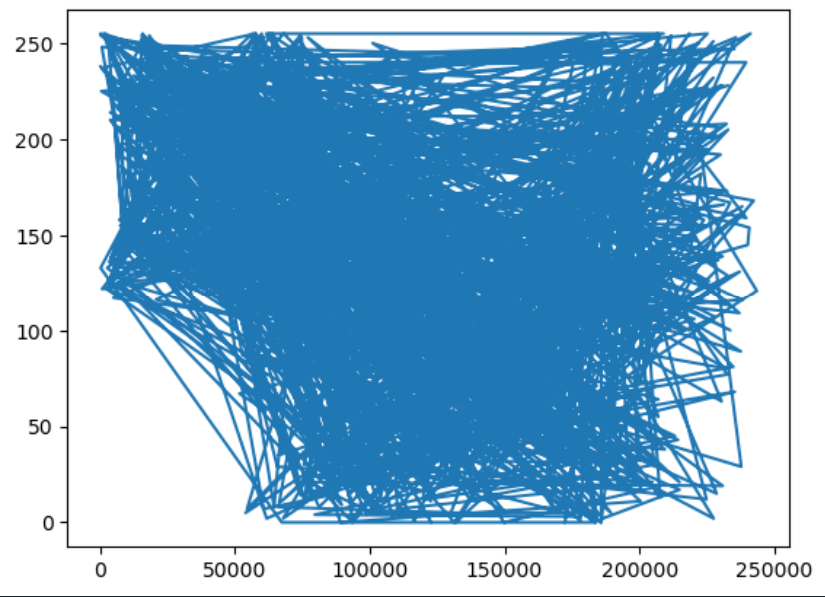
* **Среднее значение:**
  + Красный канал (R): 113.45
  + Зеленый канал (G): 114.19
  + Синий канал (B): 109.84
* **Диапазон значений:** пиксели, не относящиеся к коже, имеют более широкий диапазон значений, от 0 до 255, что включает как темные, так и светлые оттенки.
* **Распределение значений:** пиксели без кожи чаще имеют более низкие значения для всех трех каналов (например, 25-й процентиль для R — 50, для G — 61, для B — 47), что характерно для темных и нейтральных оттенков.

### **Сравнение:**

* Пиксели кожи характеризуются более высокими значениями всех трех цветовых каналов по сравнению с пикселями, не относящимися к коже.
* Пиксели кожи имеют большее количество значений на верхней границе (например, максимум для R — 225, для G — 230, для B — 255), что типично для светлых тонов кожи.
* Пиксели без кожи более разнообразны по цветам, включая более темные оттенки.

Теперь построим различных графики для лучшего понимания картины.

Построим график одного из признаков



Представим признаки на графике в виде точек.

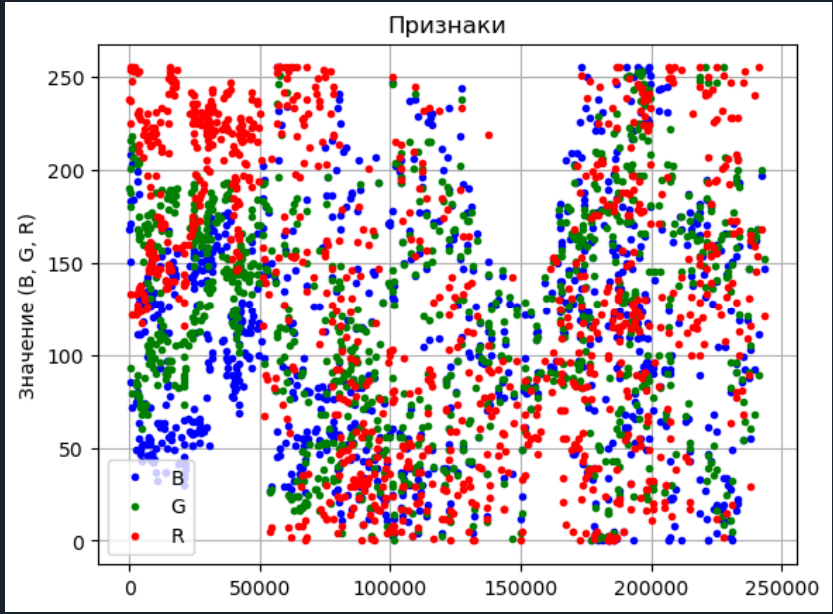
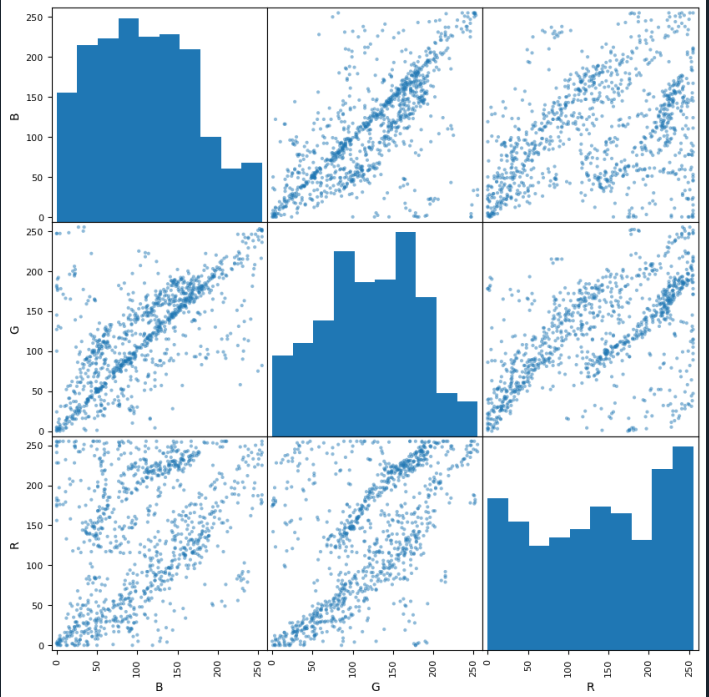


Диаграмма рассеяния, или точечная диаграмма:



Точечная диаграмма, на которой каждый класс объектов раскрашен в свой цвет.

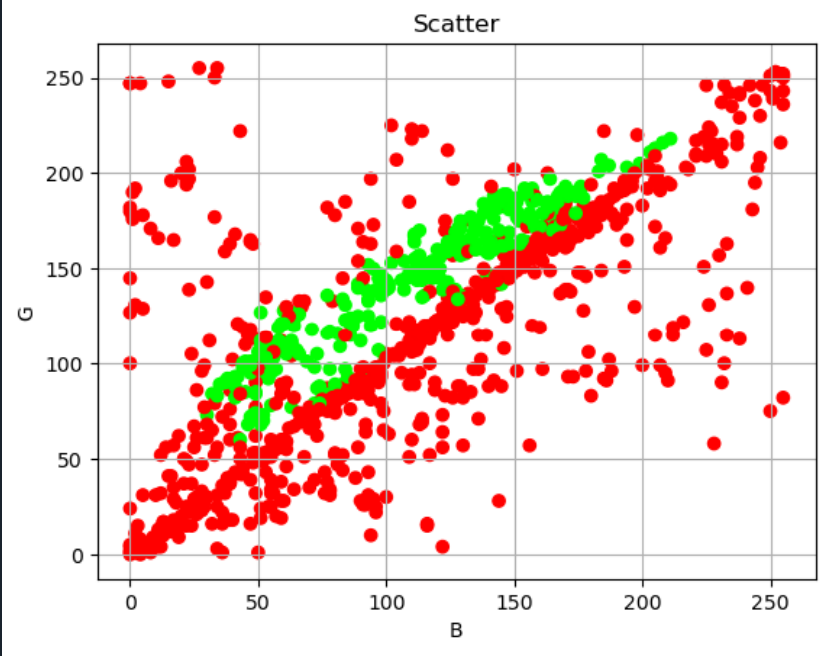
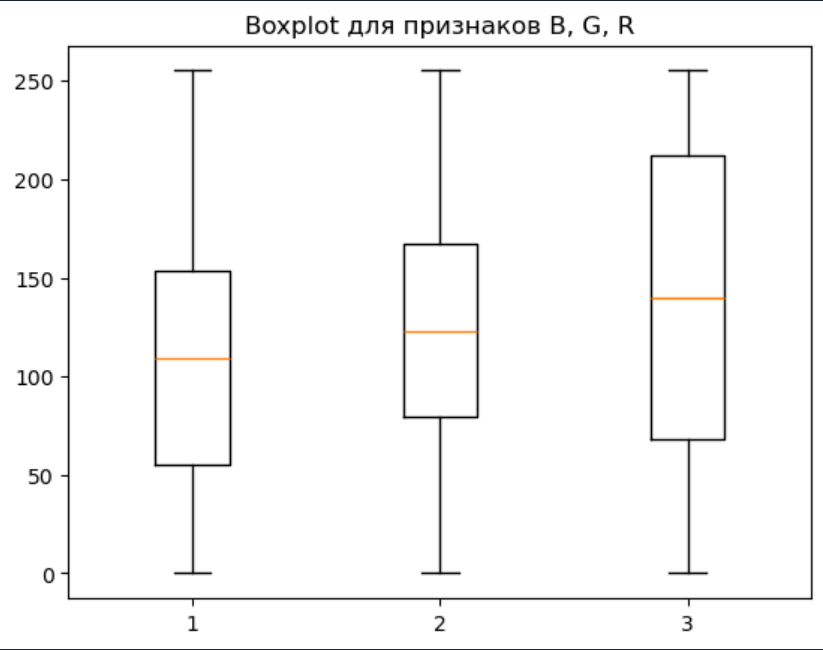


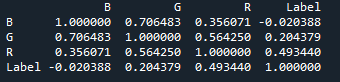
Диаграмма размаха



Теперь на графике просматривается медиана (красная ли ния), 25-й и 75-й процентили распределения (верхняя и нижняя стороны прямоугольника), максимальное и минимальное значе ния (границы усов) и выбросы (кружки). На графике нет отдельных точек (кружков), что может означать отсутствие выбросов в данных.

Напоследок рассмотрим корреляцию между признаками.





Корреляция между цветами (особенно между синим и зеленым каналами) показывает, что цветовые каналы не независимы, а имеют некоторое взаимное влияние.

Более значительное влияние на классификацию пикселей оказывает красный канал (R), что указывает на его важность при определении, является ли пиксель кожей или не кожей. В то время как синий канал имеет слабую связь с меткой, зеленый канал оказывает умеренное влияние.

Подводя итог, можно сделать вывод, что данные пригодны для использования и построение модели. В них отсутствуют пропуски, но присутствуют дубликаты, которые могут быть легко удалены. Выбросы отсутствуют.

**Вывод**

В результате выполнения лабораторной работы изучили набор данных