

**Московский авиационный институт  
(национальный исследовательский университет)**

**Институт №8 «Информационные технологии и прикладная  
математика»**

**Кафедра 806 «Вычислительная математика и  
программирование»**

**Лабораторная работа №0 по курсу «Искусственный интеллект»**

Студент: Н. П. Ежов  
Преподаватели: Д. В. Сошников  
С. Х. Ахмед  
Группа: М8О-307Б-19  
Дата:  
Оценка:  
Подпись:

**Москва, 2022**

# Лабораторная работа №0

**Задача:** В данной лабораторной работе вы выступаете в роли предпримчивого начинаящего стартапера в области машинного обучения. Вы заинтересовались этим направлением и хотите предложить миру что-то новое и при этом неплохо заработать. От вас требуется определить задачу которую вы хотите решить и найти под нее соответствующие данные. Так как вы не очень богаты, вам предстоит руками проанализировать данные, визуализировать зависимости, построить новые признаки и сказать хватит ли вам этих данных, и если не хватит найти еще. Вы готовитесь представить отчет ваши партнерам и спонсорам, от которых зависит дальнейшая ваша судьба. Поэтому тщательно работайте. И главное, день промедления и вас опередит ваш конкурент, да и сплахищенная работа отразится на репутации По сути в данной лабораторной работе вы выполняете часть работы BI системы.

# 1 Ход работы

Я выбрал набор данных Rice type classification [1] для выполнения лабораторной работы. Требуется предсказать, какой тип риса представлен (Jasmine - 1, Gonen - 0), основываясь на его признаках (таких как площадь, средняя ширина и длина, скругленность и т.д.).

Признаки в наборе данных:

1. Id — номер риса, данный столбец из данных был удалён (очевидно, что номер входных данных признаков не влияет на результат).
2. Area — регион выращивания риса.
3. MajorAxisLength — результат измерения длины риса.
4. MinorAxisLength — результат ширины риса.
5. Eccentricity — степень отклонения контура риса от окружности.
6. ConvexArea — площадь выпуклой поверхности риса.
7. EquivDiameter — диаметр сферы, эквивалентной по объёму рису.
8. Extent — площадь риса.
9. Perimeter — периметр контура риса.
10. Roundness — коэффициент "скругленности" контура риса.
11. AspectRatio — соотношение длины и ширины риса.
12. Class — Тип риса.

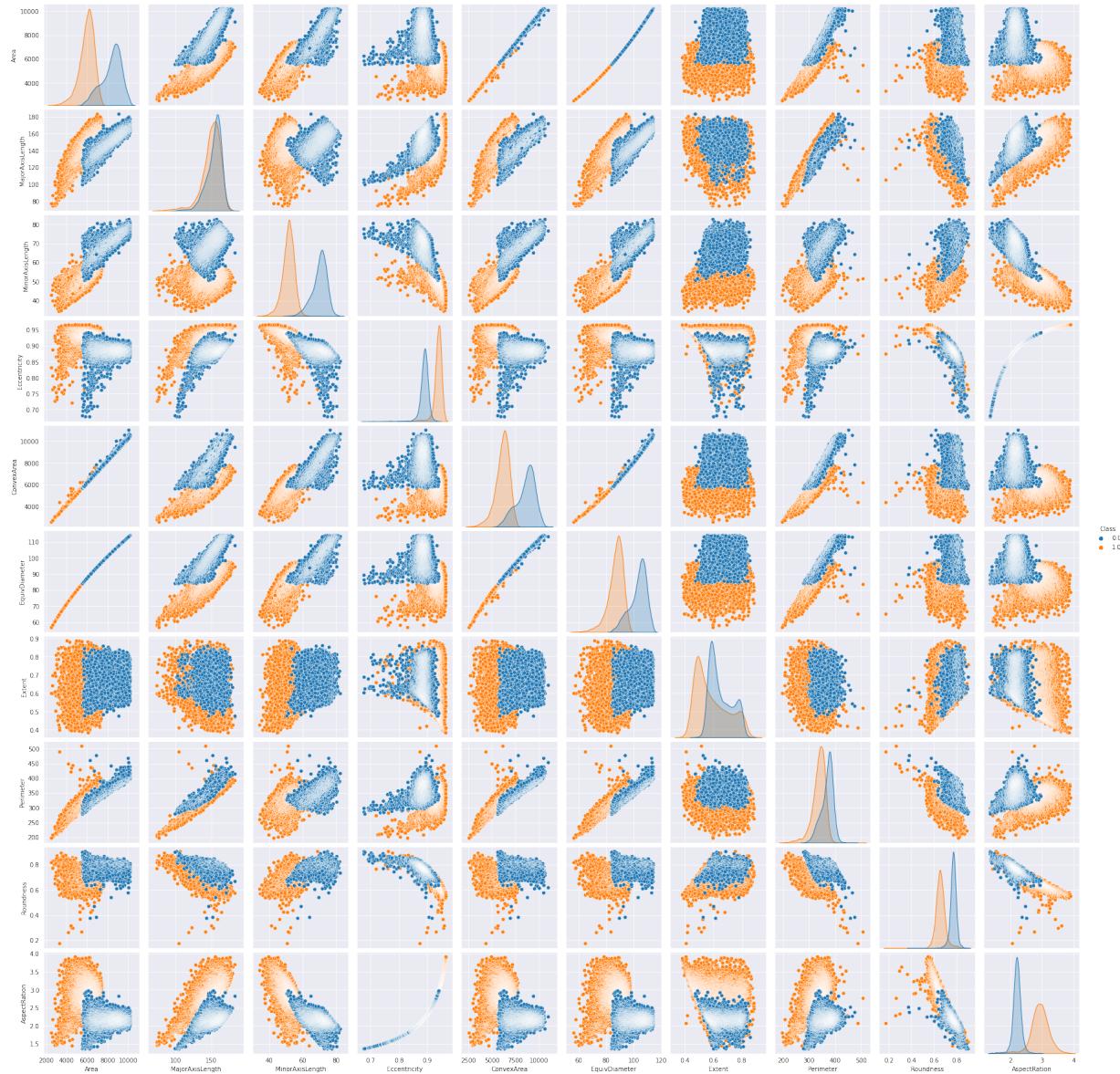
Перед выявлением зависимостей между признаками следует проверять целостность набора данных:

```
RangeIndex: 18185 entries, 0 to 18184
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
 ---  --  
 0   Area             18185 non-null   int64  
 1   MajorAxisLength  18185 non-null   float64 
 2   MinorAxisLength  18185 non-null   float64 
 3   Eccentricity    18185 non-null   float64
```

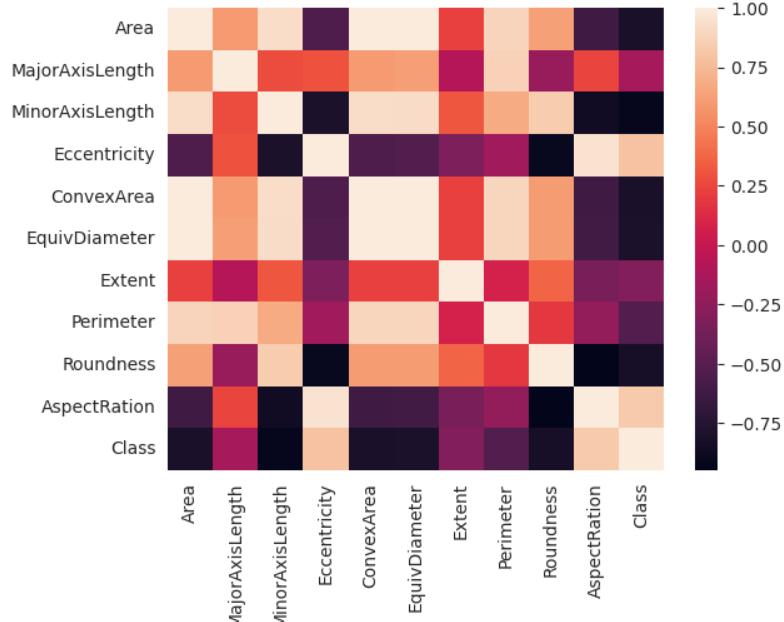
```
4    ConvexArea      18185 non-null  int64
5    EquivDiameter   18185 non-null  float64
6    Extent          18185 non-null  float64
7    Perimeter       18185 non-null  float64
8    Roundness        18185 non-null  float64
9    AspectRation    18185 non-null  float64
10   Class           18185 non-null  int64
dtypes: float64(8),int64(3)
memory usage: 1.5 MB
```

В наборе нет неполных данных, а все признаки - числовые.

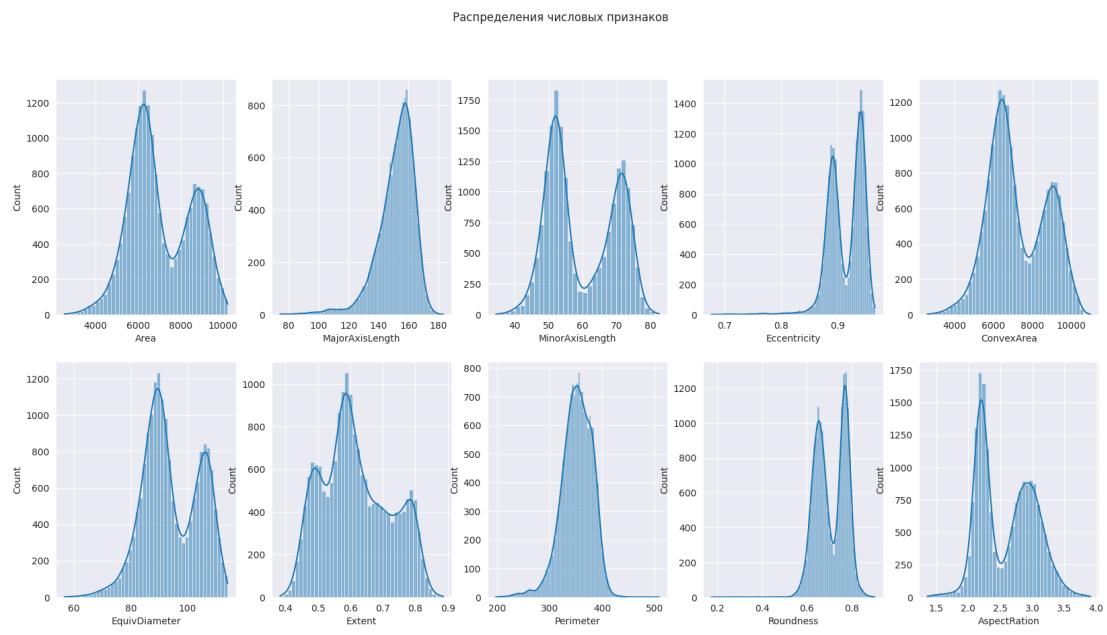
Построю графики для каждой пары признаков. Синим отмечен рис типа Jasmin, оранжевым - Gonen:



Построю корреляционную матрицу для признаков:

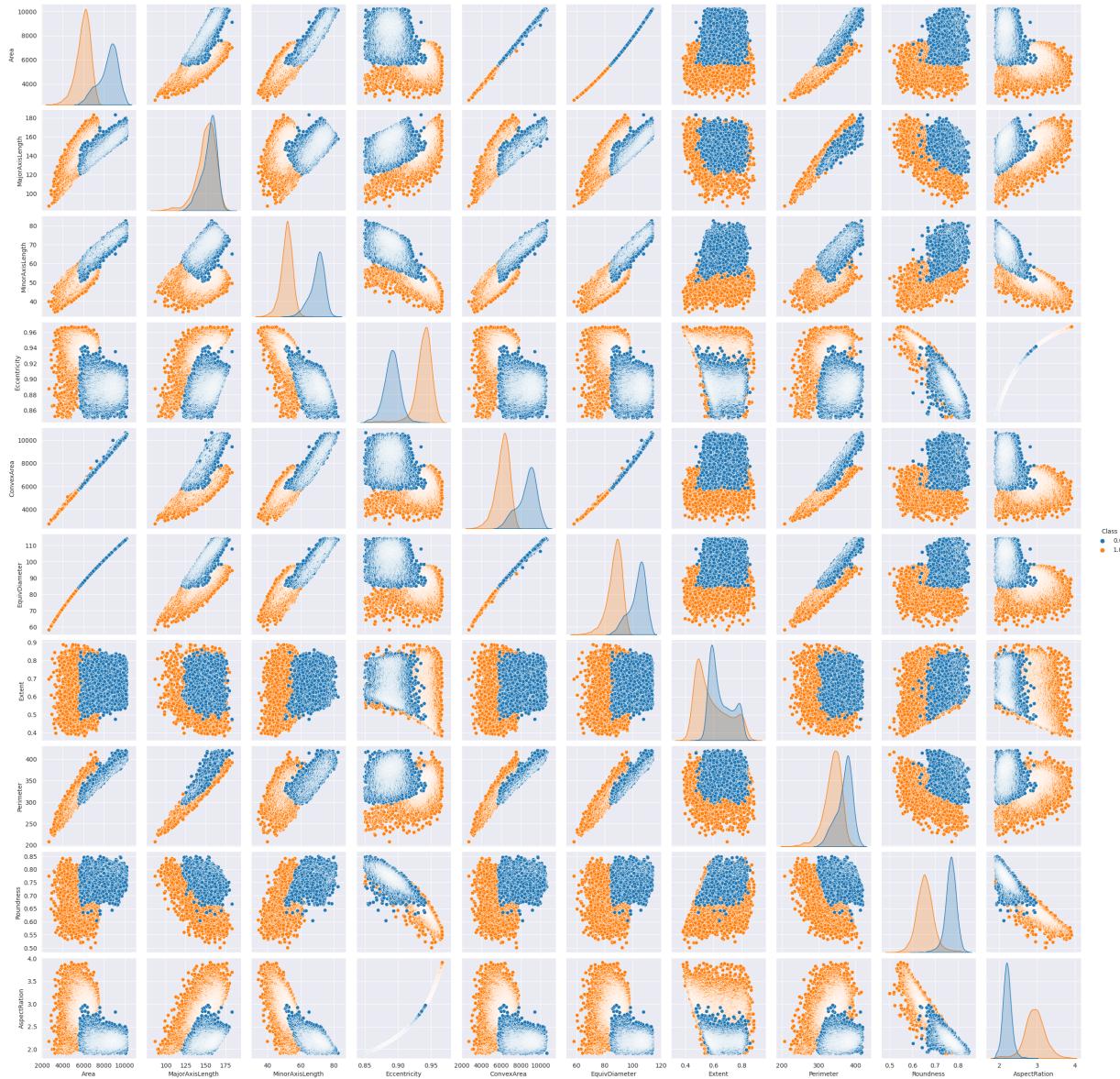


Так же построю гистограммы для числовых признаков:



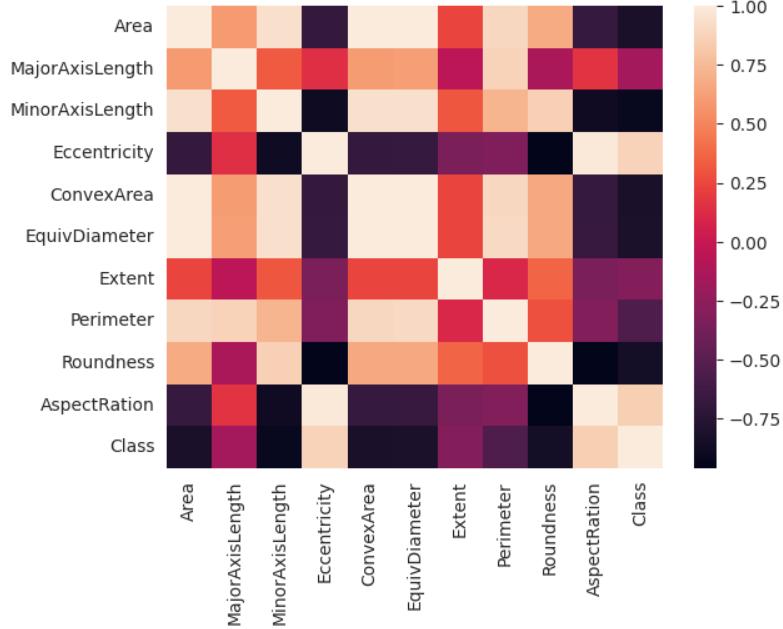
Имеется небольшое количество выбросов у некоторых признаков. Удалию все данные, где Roundness < 0.5, Perimeter > 420 и Eccentricity < 0.85

Теперь ничего не мешает анализу. Построю те же графики для обработанного набора данных:



Исходя из парных графиков, можно сделать вывод, что задачу реально решить линейной моделью. Однако, на некоторой части графиков не представляется возможным провести точную прямую линию, которая бы разделяла два класса, а признак MajorAxisLength вообще практически не коррелирует с целевым классом, что отчётливо заметно на одном из диагональных графиков.

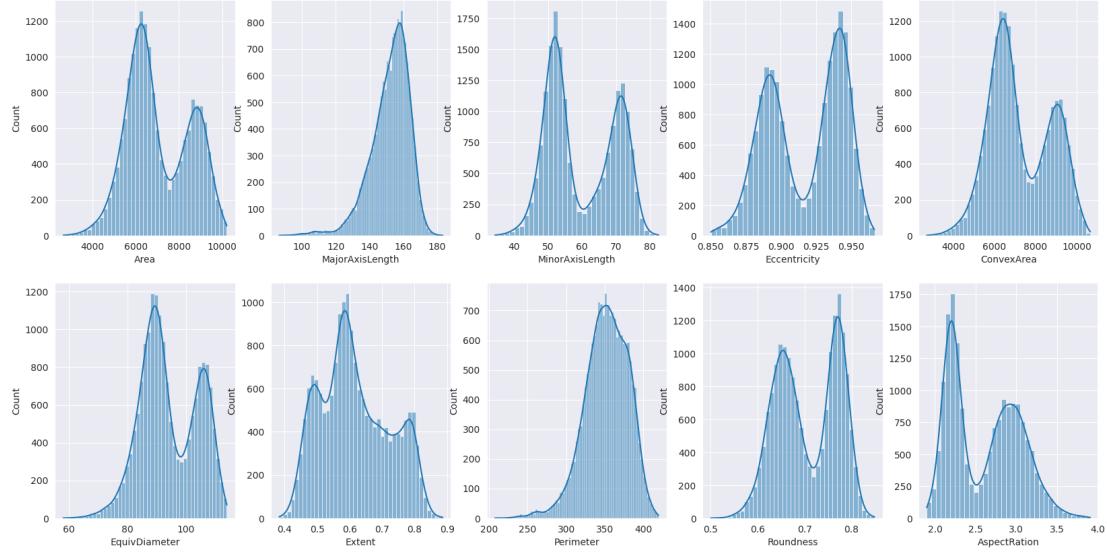
Корреляционная матрица после удаления выбросов не изменилась:



Видно, что почти все признаки, кроме «MajorAxisLength», «Extent» и «Perimeter», очень сильно влияют на конечный результат. Также некоторые признаки довольно сильно коррелируют между собой, что может быть объяснено тем, что всё это, по сути, геометрические параметры, которые связаны между собой, т.к. контур риса обоих сортов напоминает эллипс.

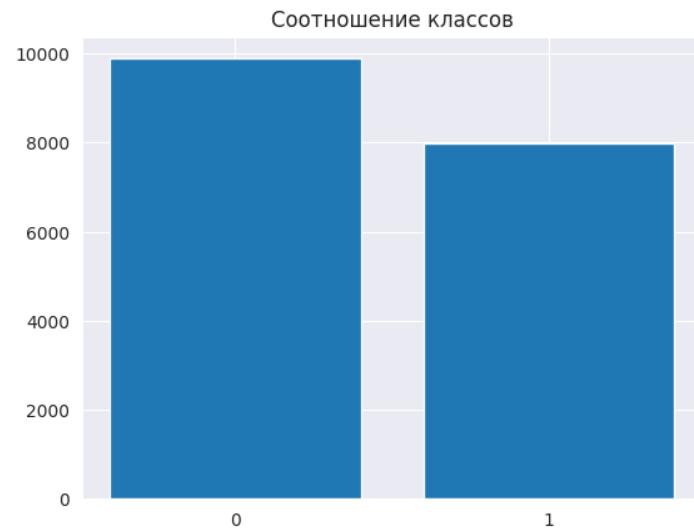
Гистограммы распределения числовых признаков:

Распределения числовых признаков



Полученных данных достаточно для построения модели, об этом говорят попарные графики и корреляционная матрица. Добавление новых признаков не требуется.

Соотношение классов объектов:



Объектов разных классов примерно одинаковое количество, oversampling не требуется. Данные готовы к обучению.

## 2 Выводы

В ходе выполнения лабораторной работы я освежил в памяти курс математической статистики: гистограмму, корелляцию и корреляционную матрицу для наборов данных. Так же я изучил библиотеку Pandas, она оказалась очень удобной для анализа данных.

Трудно было найти подходящий набор данных, который подходил бы под параметры для обучения линейных моделей. В ходе своих поисков я также пробовал провести анализ и обучение на датасете для выявление диабета, но, по парным графикам все значения были «перемешаны» и корреляция почти всех признаков была  $\leq 10\%$ .

Был проанализирован набор данных Rice type [1], результаты получились законо-мерные: тип риса напрямую зависит от геометрических параметров и региона выращивания. Но, исходя из корреляционной матрицы можно заметить, что регион выращивания почти полностью определяет геометрические параметры риса.

## Список литературы

- [1] *Rice type / Kaggle*  
URL: <https://www.kaggle.com/datasets/mssmartypants/rice-type-classification>  
(дата обращения: 30.05.2022).
- [2] *Exploratory data analysis with Pandas – mlcourse.ai*  
URL: [https://mlcourse.ai/book/topic01/topic01\\_pandas\\_data\\_analysis.html](https://mlcourse.ai/book/topic01/topic01_pandas_data_analysis.html)  
(дата обращения: 30.05.2022).