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: М. А. Инютин
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 предприимчивого начинающего 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 сплагиаченная работа отразится на репутации По сути в данной лабораторной работе вы выполняете часть работы BI системы.

1 Ход работы

Я выбрал набор данных Smoker Condition [1] для выполнения лабораторной работы. Требуется предсказать, будет ли у человека рак лёгких, основываясь на его генетике и на том, курит ли он.

Признаки в наборе данных:

1. Gene2337 — результат количественного анализа Gene2337.
2. Gene35715 — результат количественного анализа Gene35715.
3. Gene12936 — результат количественного анализа Gene12936.
4. Gene1689 — результат количественного анализа Gene1689.
5. FGFR1 — результат количественного анализа FGFR1 (рецептор фактора роста фибробластов).
6. GATA4 — результат количественного анализа GATA4.
7. type — является ли человек курильщиком. В работе я переименовал признак в «Is smoker?» с числовым значением 0 или 1.
8. Condition — состояние лёгких человека. В работе я переименовал признак в «Has cancer?» с числовым значением 0 или 1.

Перед выявлением зависимостей между признаками следует проверяю целостность набора данных:

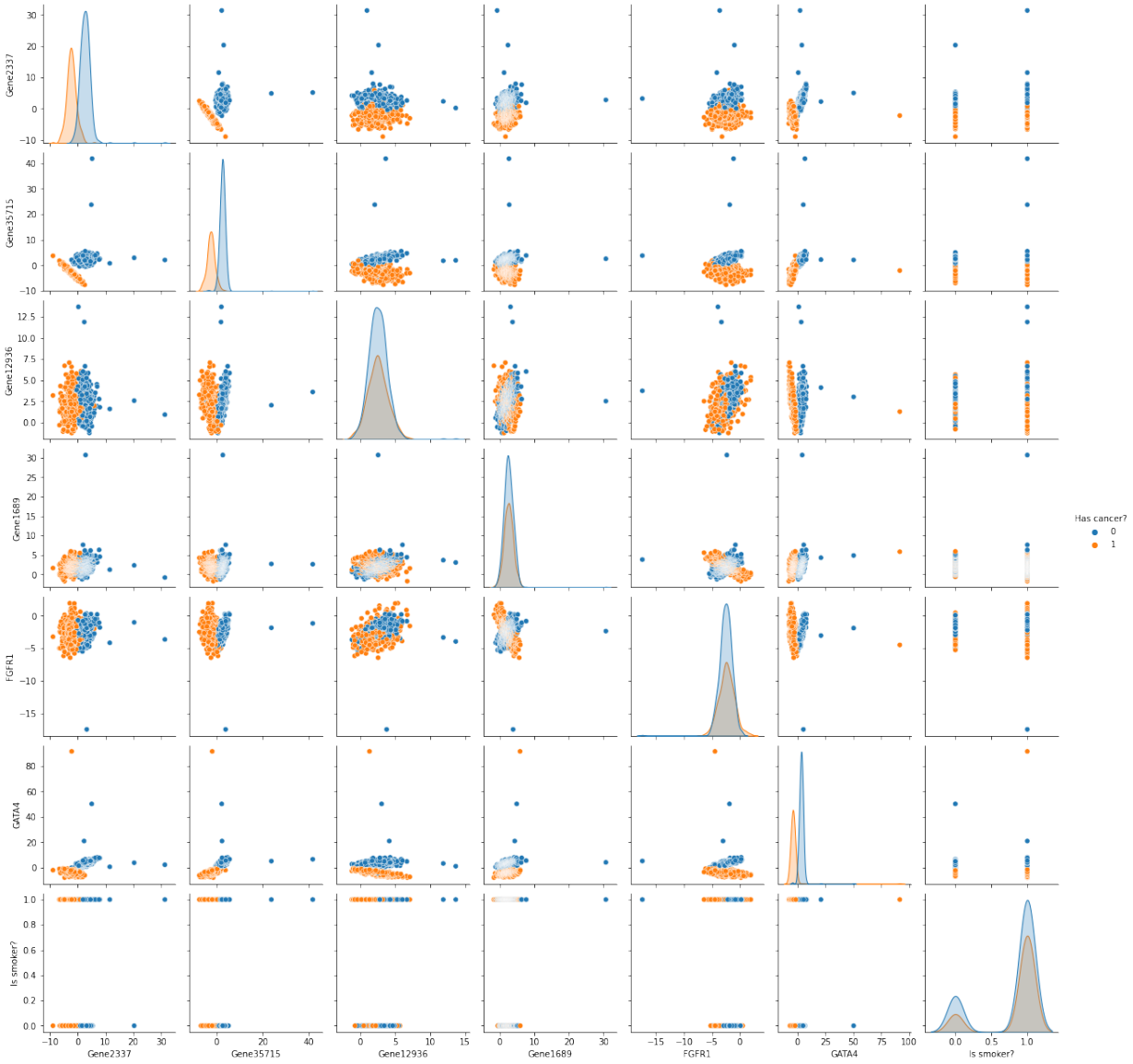
```
RangeIndex: 1023 entries,0 to 1022
Data columns (total 8 columns):
#   Column     Non-Null Count  Dtype
--- ------    -------------- -----
0   Gene2337   1022 non-null   float64
1   Gene35715  1020 non-null   float64
2   Gene12936  1020 non-null   float64
3   Gene1689   1020 non-null   float64
4   FGFR1      1021 non-null   float64
5   GATA4      1021 non-null   float64
6   type       1023 non-null   object
7   Condition  1023 non-null   object
dtypes: float64(6),object(2)
memory usage: 64.1+ KB
```

В наборе есть неполные данные, которые я удаляю. Категориальные признаки привожу к числовым, преобразую соответствующие столбцы с данными.

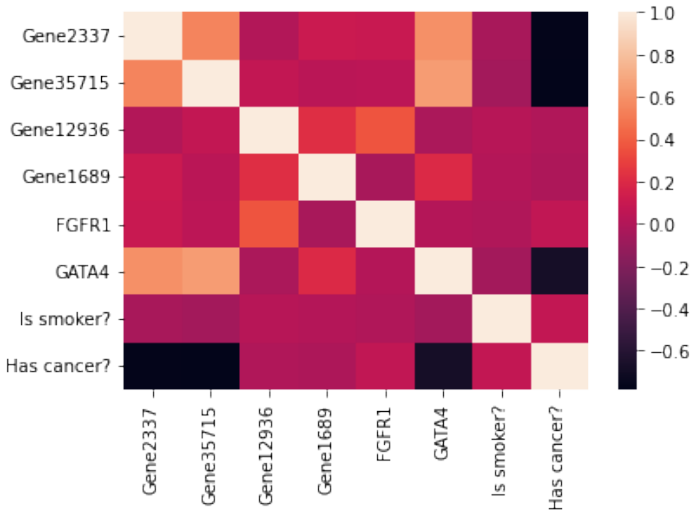
Данные после преобразования имеют следующий вид:

```
Int64Index: 1000 entries,0 to 1019
Data columns (total 8 columns):
#   Column       Non-Null Count  Dtype
--- ------      -------------- -----
0   Gene2337     1000 non-null   float64
1   Gene35715    1000 non-null   float64
2   Gene12936    1000 non-null   float64
3   Gene1689     1000 non-null   float64
4   FGFR1        1000 non-null   float64
5   GATA4        1000 non-null   float64
6   Is smoker?   1000 non-null   int64
7   Has cancer?  1000 non-null   int64
dtypes: float64(6),int64(2)
memory usage: 102.6 KB
```

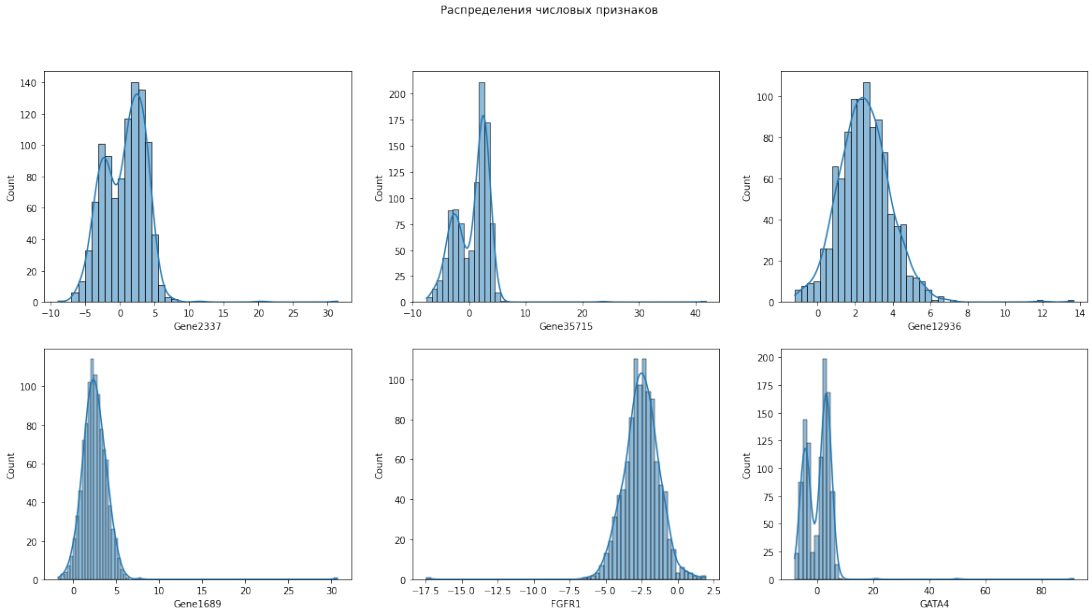
Построю графики для каждой пары признаков. Синим отмечены нормальные лёгкие, оранжевым лёгкие с раковой опухолью:



Построю корреляционную матрицу для признаков:

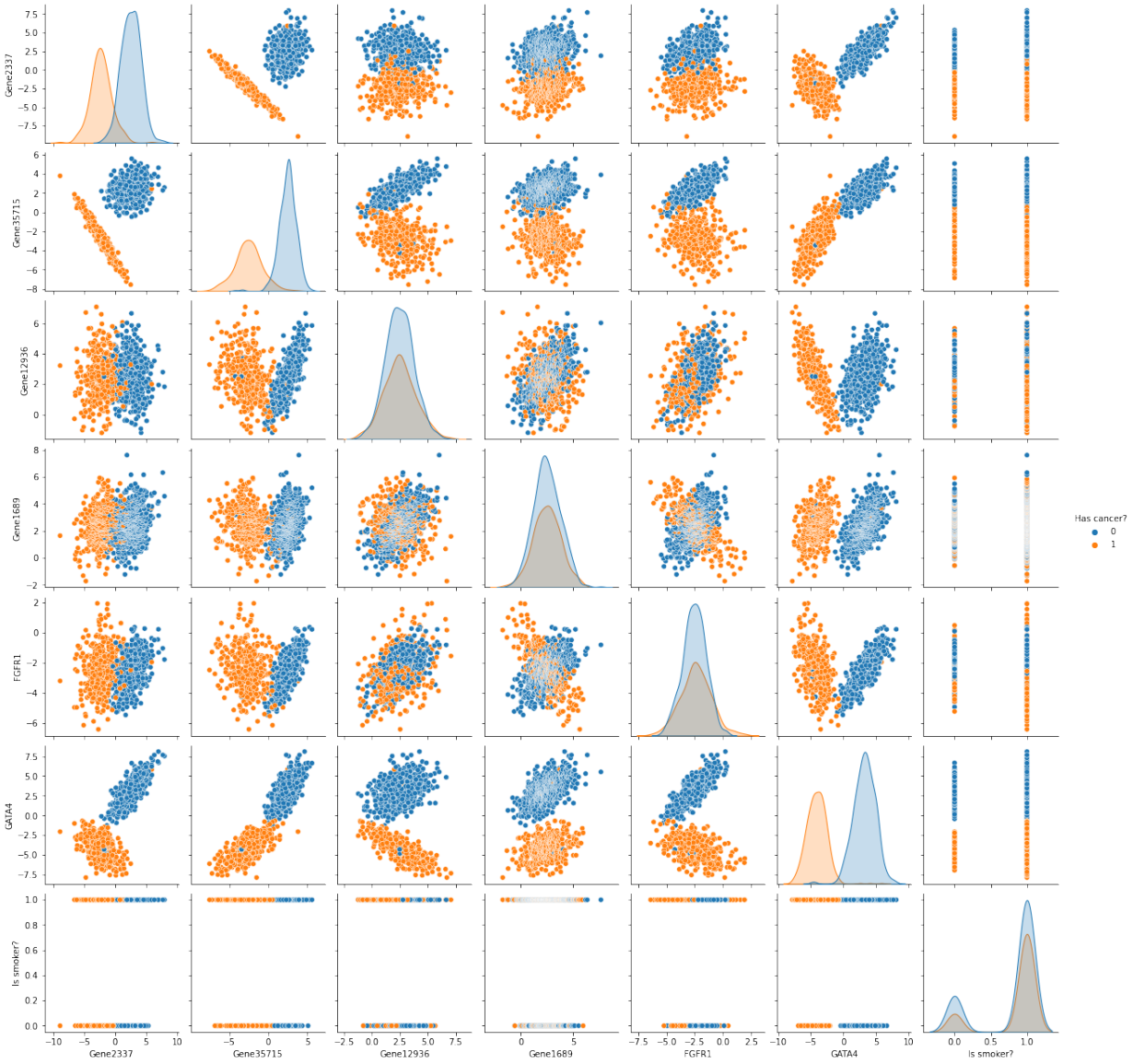


Так же построю гистограммы для числовых признаков:



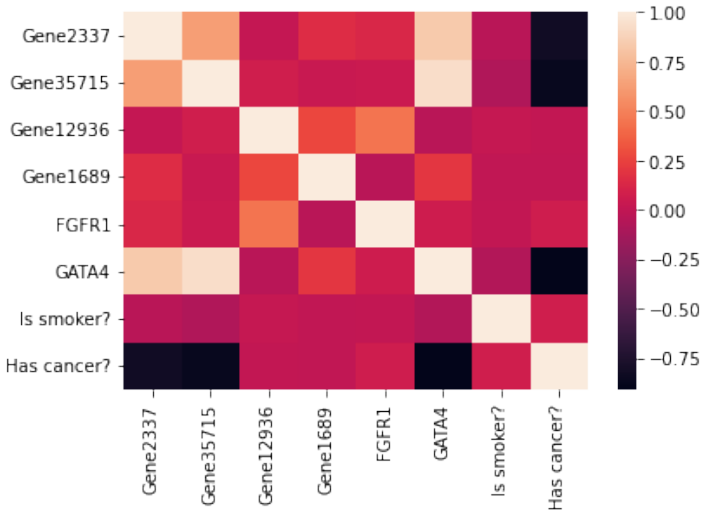
Видно, что в данных присутствуют выбросы, которые могут повлиять на обучение модели. Числовые данные не превосходят по модулю 10, поэтому удалю все элементы, в которых это правило не соблюдается.

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



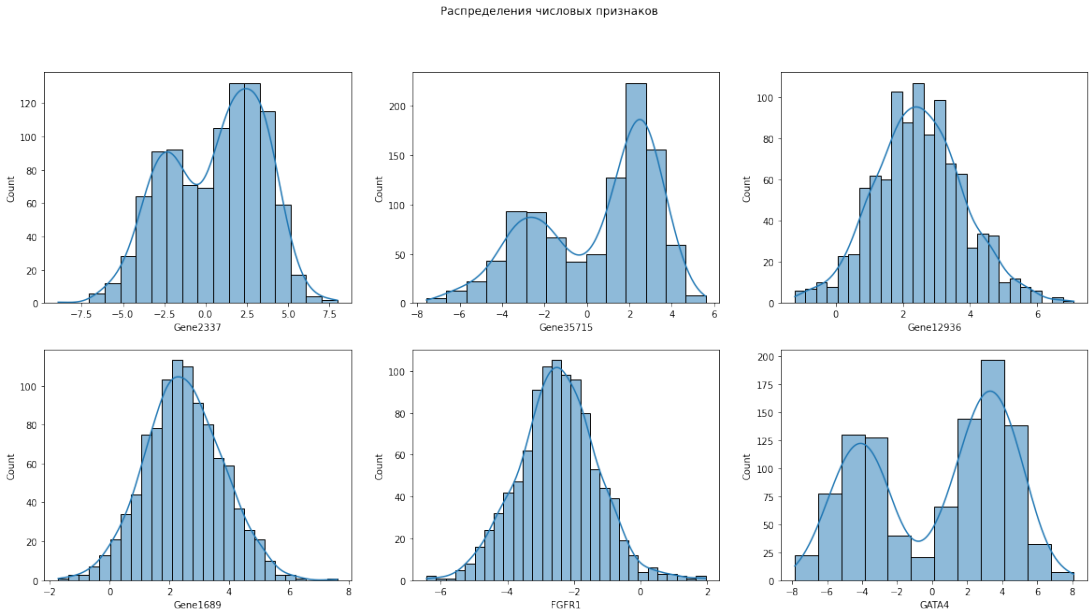
Исходя из парных графиков, можно сделать вывод, что задачу реально решить линейной моделью: на большинстве рисунков можно довольно точно провести прямую, разделяющую синие и оранжевые точки.

Корреляционная матрица после удаления выбросов не изменилась:



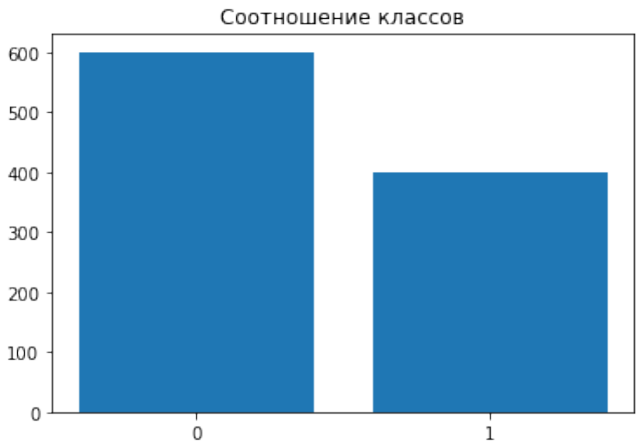
Видно, что больше всего на результат влияют «Gene2337», «Gene35715» и «GATA4», которые ещё и довольно сильно коррелируют между собой. «Gene12936», «Gene1689», «FGFR1» и фактор курения влияют меньше.

Гистограммы 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.

Сперва было довольно трудно выбрать задачу, которую предстоит решать. Многие наборы данных имеют много признаков, из-за этого попарные графики долго отрисовывались и было сложно по ним что-то понять.

Был проанализирован набор данных Smoker Condition [1], результаты получились очень интересные: курение влияет на рак лёгких в меньшей степени, чем генетическая предрасположенность. Понятно, что выборка довольно мала, чтобы делать такой вывод. В работе этого и не требуется.

Список литературы

[1] *Smoker Condition | Kaggle*
URL: `https://www.kaggle.com/datasets/devzohaib/smoker-condition`
(дата обращения: 08.05.2022).

[2] *Exploratory data analysis with Pandas — mlcourse.ai*
URL: `https://mlcourse.ai/book/topic01/topic01_pandas_data_analysis.html`
(дата обращения: 08.05.2022).