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: В. Ю. Юревич
Преподаватели: Д. В. Сошников
С. Х. Ахмед
Группа: М8О-407Б-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 Diabetes classification for beginner [1] для выполнения лабораторной работы. В описании датасета указано, что он состоит из реальных диагностических данных и предлагается создать модель, которая сможет предсказать, болен человек или нет на основании результатов анализов.

Признаки в наборе данных:

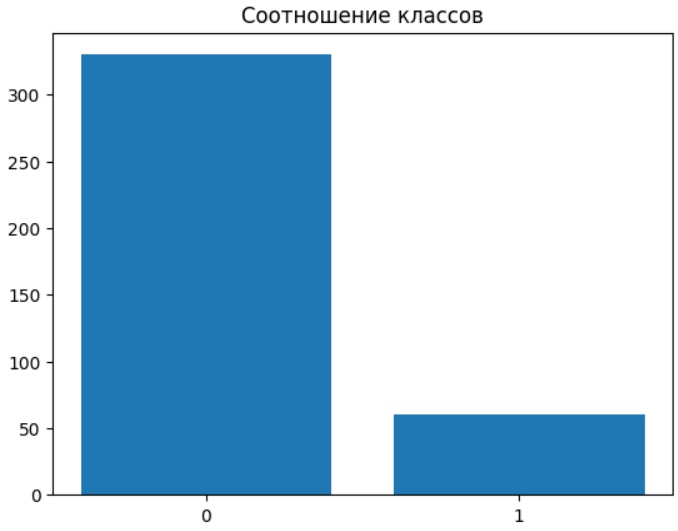
1. cholesterol — числовое выражение количества общего холестерина в крови человека
2. glucose — числовое выражение количества глюкозы в крови человека
3. hdl_chol — числовое выражение количества холестерина-ЛПВП (HDL) в крови человека
4. chl_hdl_ratio — соотношение количества общего холестерина и холестерина-ЛПВП в крови человека (Общий/HDL)
5. age — возраст пациента
6. gender — пол пациента
7. height — рост пациента, выраженный в дюймах
8. weight — вес пациента, выраженный в фунтах
9. bmi — индекс массы тела
10. diabetes — диагноз: имеет челове диабет или нет.

Перед выявлением зависимостей между признаками следует проверяю целостность набора данных:

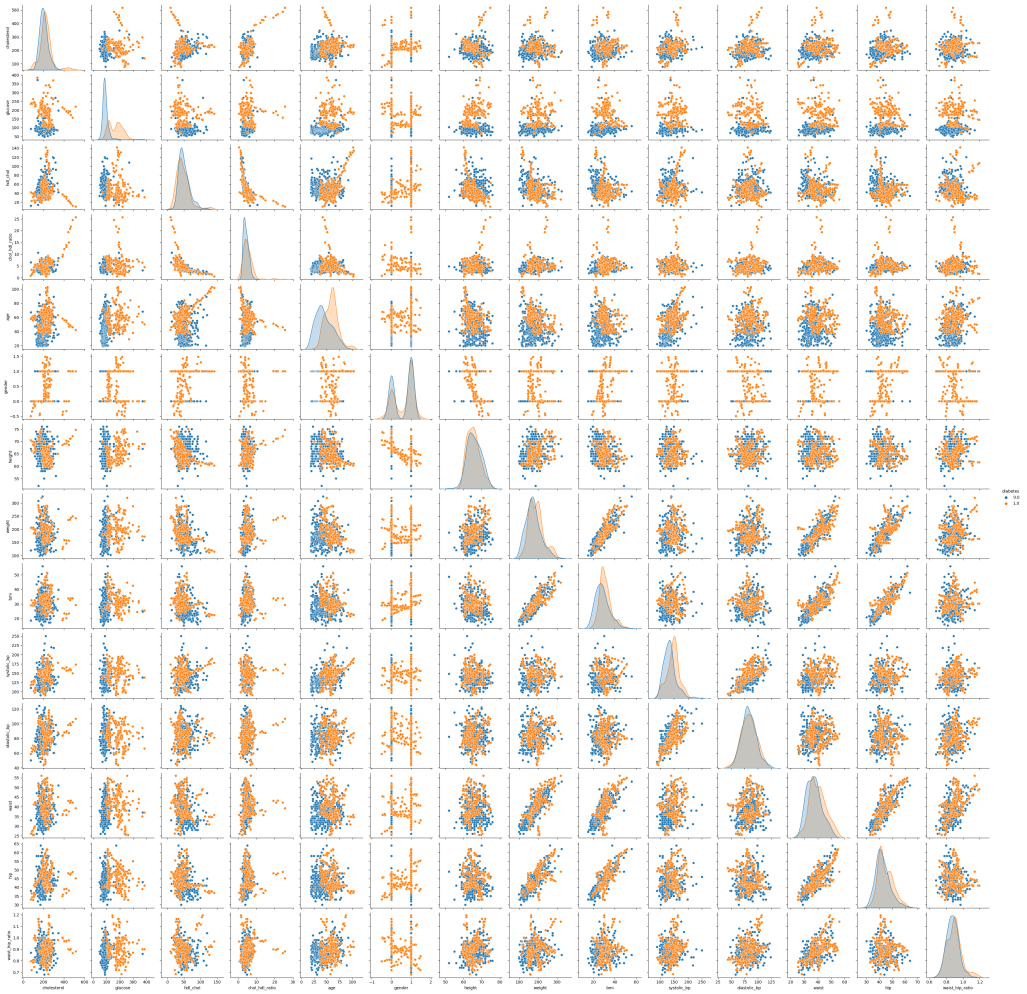
```
Data columns (total 15 columns):
#   Column           Non-Null Count  Dtype
--- ------           -------------- -----
0   cholesterol      390 non-null    int64
1   glucose          390 non-null    int64
2   hdl_chol         390 non-null    int64
3   chol_hdl_ratio   390 non-null    object
4   age              390 non-null    int64
5   gender           390 non-null    int64
6   height           390 non-null    int64
```

```
7   weight           390 non-null    int64
8   bmi              390 non-null    object
9   systolic_bp      390 non-null    int64
10  diastolic_bp     390 non-null    int64
11  waist            390 non-null    int64
12  hip              390 non-null    int64
13  waist_hip_ratio  390 non-null    object
14  diabetes         390 non-null    int64
dtypes: int64(12),object(3)
memory usage: 45.8+ KB
```

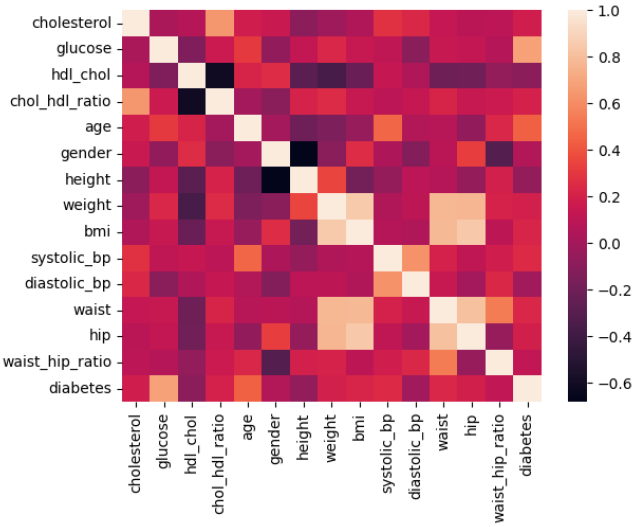
В наборе нет неполных данных, но при этом не все признаки числовые. В файле ЛР описано, как неподходящие типы преобразуются к типу float, вкратце - замена уникальных значений на числа и парсинг самих вещественных чисел из строк. Также в датасете наблюдался дизбаланс целевых классов, поэтому был проведён oversampling при помощи библиотеки imbalanced-learn [3]



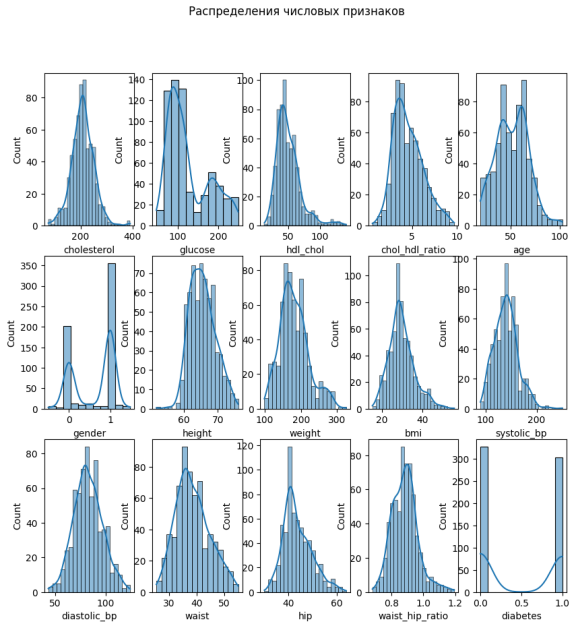
Построю графики для каждой пары признаков. Синим отмечены пациенты, у которых диабет, оранжевым - у которых диабета нет:



Построю корреляционную матрицу для признаков:

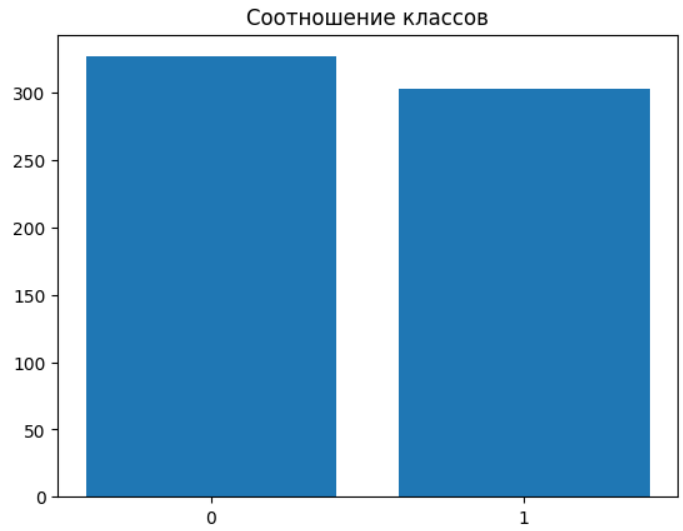


Так же построю гистограммы для числовых признаков:



Выбросов не было обнаружено, так как датасет довольно маленький.

Соотношение классов объектов:



Объектов разных классов примерно одинаковое количество после выполнения oversampling'а, но всё-таки не 50/50 т.к. я удалял выбросы из данных.

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.

В ходе поиска подходящего датасета я перепробовал несколько, и во всех из них требовался oversampling, был выбран именно этот набор данных, так как относительно большое количество примеров класса, который представлен в меньшем количестве, благодаря чему oversampling может быть более точным. Я мог бы просто скопировать данные класса, который представлен в меньшем количестве, и в контексте обучения моделей машинного обучения это даже было бы валидно, насколько я понял, но я решил проверить позволяют ли синтетически сгенерированные данные построить оптимальную модель для классификации по признакам.

Был проанализирован набор данных Diabetes classification for beginner [1], результаты получились интересные, некоторые признаки сильно коррелируют с конечным результатом, но не настолько, чтобы можно было определить, болен человек или нет, имея только их.

Список литературы

[1] *Beginner's Classification Dataset*
URL: `https://www.kaggle.com/datasets/houcembenmansour/predict-diabetes-based-on-di`
(дата обращения: 30.09.2022).

[2] *Exploratory data analysis with Pandas — mlcourse.ai*
URL: `https://mlcourse.ai/book/topic01/topic01_pandas_data_analysis.html`
(дата обращения: 30.09.2022).

[3] *Oversampling documentation*
URL: `https://imbalanced-learn.org/stable/over_sampling.html`
(дата обращения: 30.09.2022).