Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №0 по курсу «Искусственный интеллект»
Тема: Анализ и подготовка данных

Студент: К. А. Спиридонов
Преподаватель: Самир Ахмед
Группа: М8О-407Б-19
Дата:
Оценка:
Подпись:

Москва, 2022

Задача

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. Если вы заинтересовались этим направлением, то можно будет в дальнейшем что-то придумать)

1 Описание

Для задачи был выбран датасет «Stroke Prediction Dataset».

Проблема, которую решает датасет:

По данным Всемирной организации здравоохранения (ВОЗ), инсульт занимает второе место среди причин смерти в мире, на него приходится около 11% всех смертей. Этот набор данных используется для прогнозирования вероятности инсульта у пациента на основе таких входных параметров, как пол, возраст, различные заболевания и статус курения. Каждая строка в данных содержит соответствующую информацию о пациенте.

Признаки датасета:

Attribute Information

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

2 Анализ данных

Привожу результаты из ноутбука, т.к. там довольно понятно всё описано

Признаки и их типы

```
1 df.dtypes
```

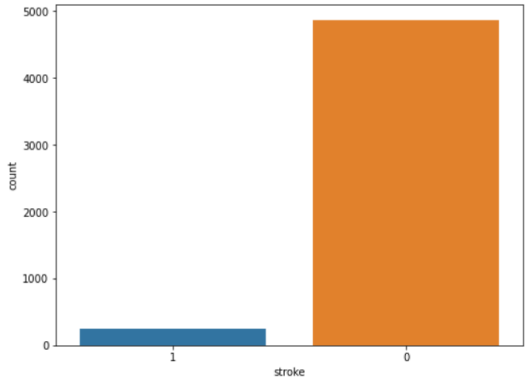
```
id                   int64
gender              object
age                float64
hypertension         int64
heart_disease        int64
ever_married        object
work_type           object
Residence_type      object
avg_glucose_level  float64
bmi                float64
smoking_status      object
stroke               int64
dtype: object
```

Можно заметить, что hypertension, heart disease и stroke представлены как int, но мы знаем, что это категориальные переменные. Поэтому давай конвертируем их в объектный тип

Затем посмотрел как соотносятся данные целевого признака

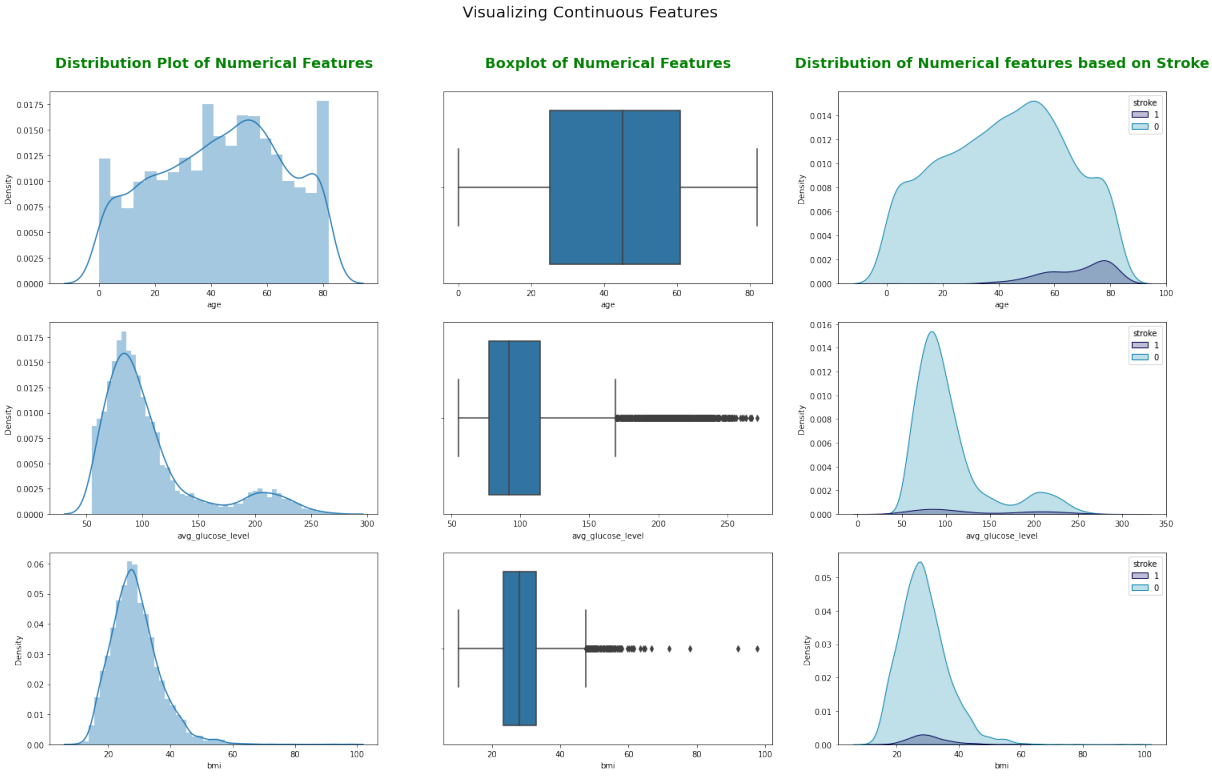
Анализ целевого знаяения - stroke

```
[ ] 1 plt.figure(figsize=(8,6))
    2 sns.countplot(x = 'stroke', data = df)
    3 plt.show()
```



Из графика видно, что мы имеем дисбаланс классов.

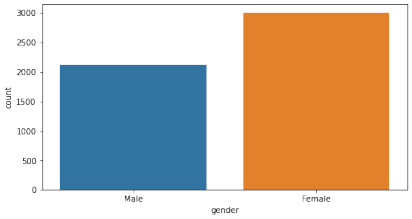
Чтобы решить это проблему, будем исползовать RandomOverSampler()
Посмотрел как соотносятся количественные признки:

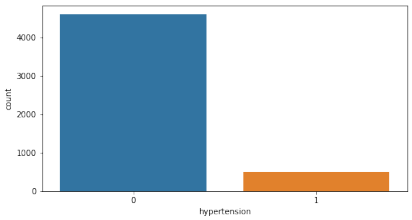


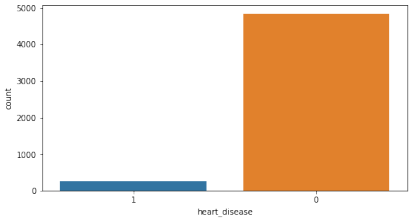
У нас явно много выбросов в столбцах avg_glucose_level и BMI.
Распределение по возрасту, основанное на инсульте, показывает, что у пожилых людей гораздо больше шансов заболеть инсультом по сравнению с более молодыми людьми.

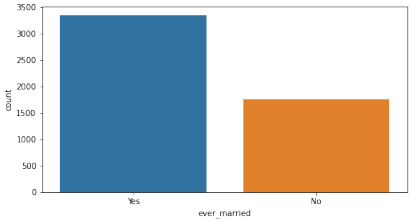
Анализ категориальных признаков был таким:

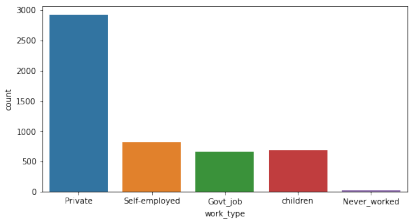
Count plot for Categorical Features

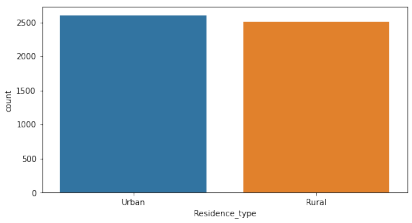


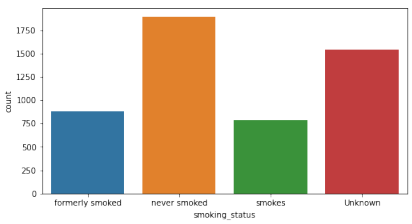




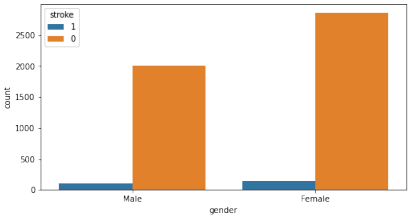


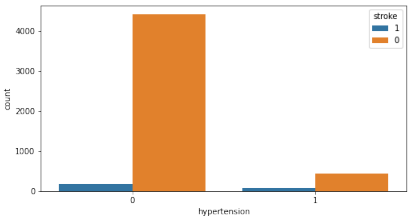


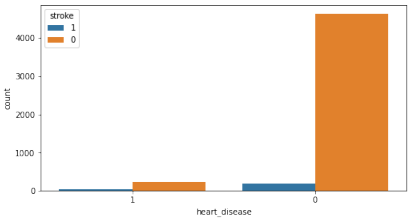


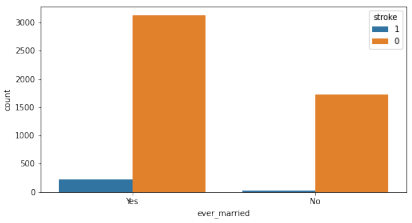


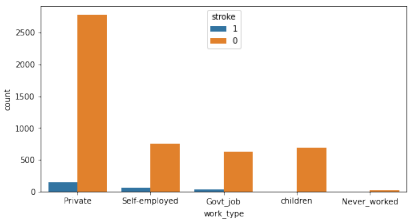
Count plot for Categorical Features Based on Stroke

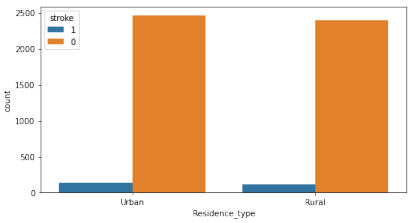


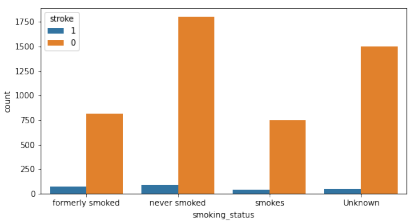






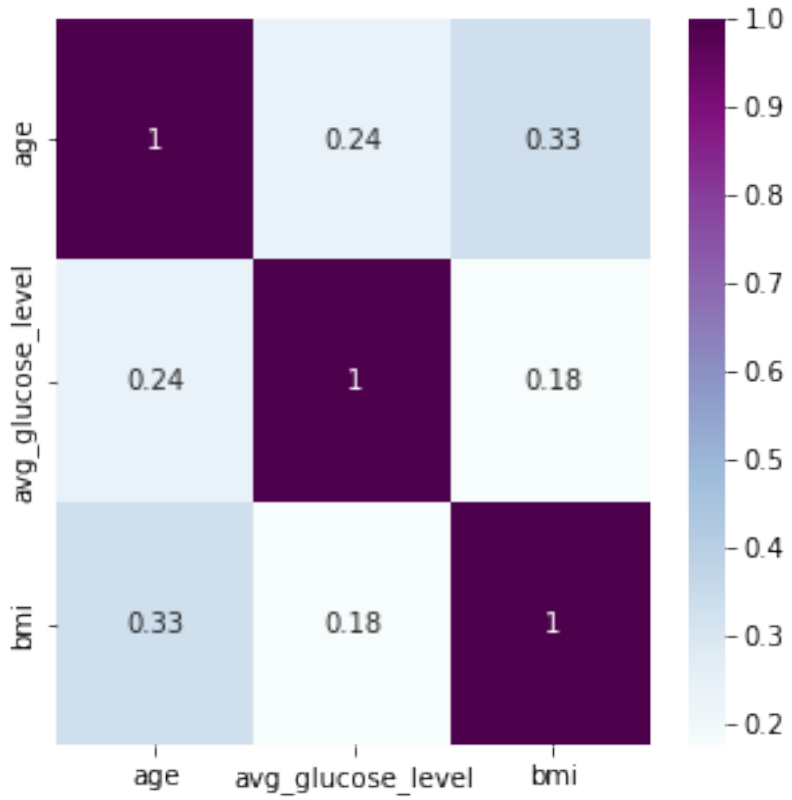






Женатые люди чаще страдают от инсульта по сравнению с неженатыми.
У городских жителей инсульты случаются чаще, чем у жителей сельской местности.

Так выглядит корреляционная матрица:



3 Выводы

Выполнив лабораторную работу, я научился базов навыкам анализа данных. Познакомился с сайтам kaggle, на котором есть много различных датасетов. Так же я познакомился с библиотекой pandas, в которой есть множество полезных инструментов для работы с данными.