Отчет по лабораторной работе №0
по курсу «Искусственный интеллект»

Тема: Data Mining и исследование данных

Студент группы 8О-308 Красоткин Семён, № по списку 10.

Работа выполнена:6.06.2022

Преподаватель: Самир Ахмед

Отчет сдан:

Итоговая оценка:

1. Цель работы

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.

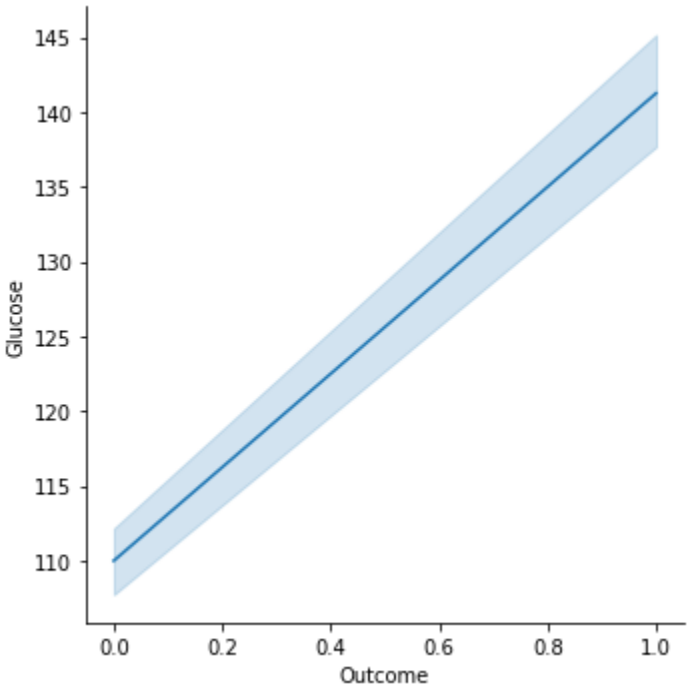
2. Ход работы

Взял датасет для обнаружения диабета.

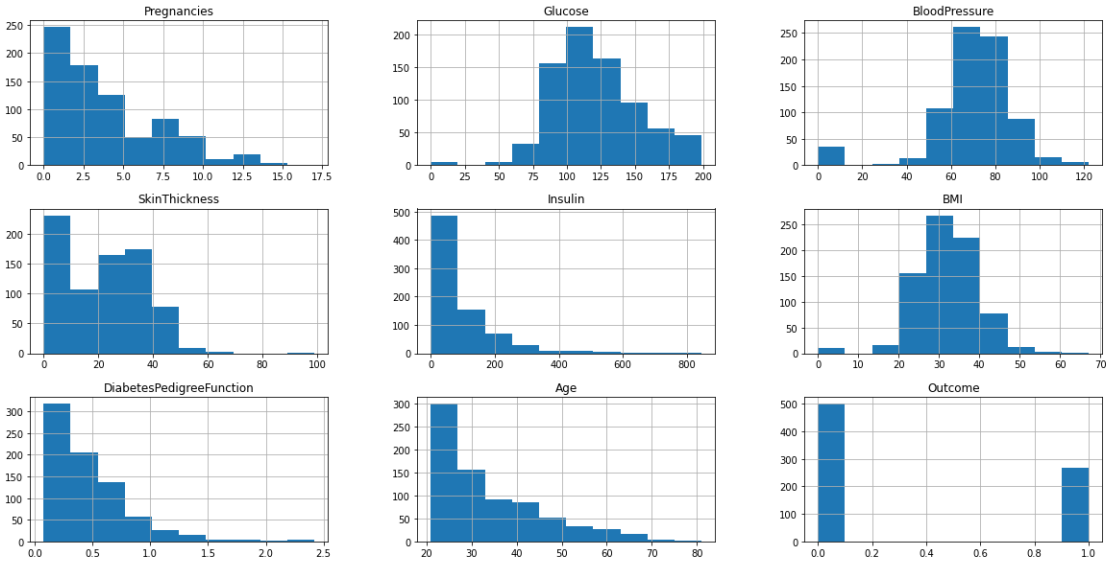
В критерии входит:

- Число беременностей
- Уровень глюкозы
- Кровяное давление
- Толщина кожи на трицепсе.
- Уровень инсулина.
- ИМТ
- Функция предсказания диабета
- Возраст
- Болен или нет.

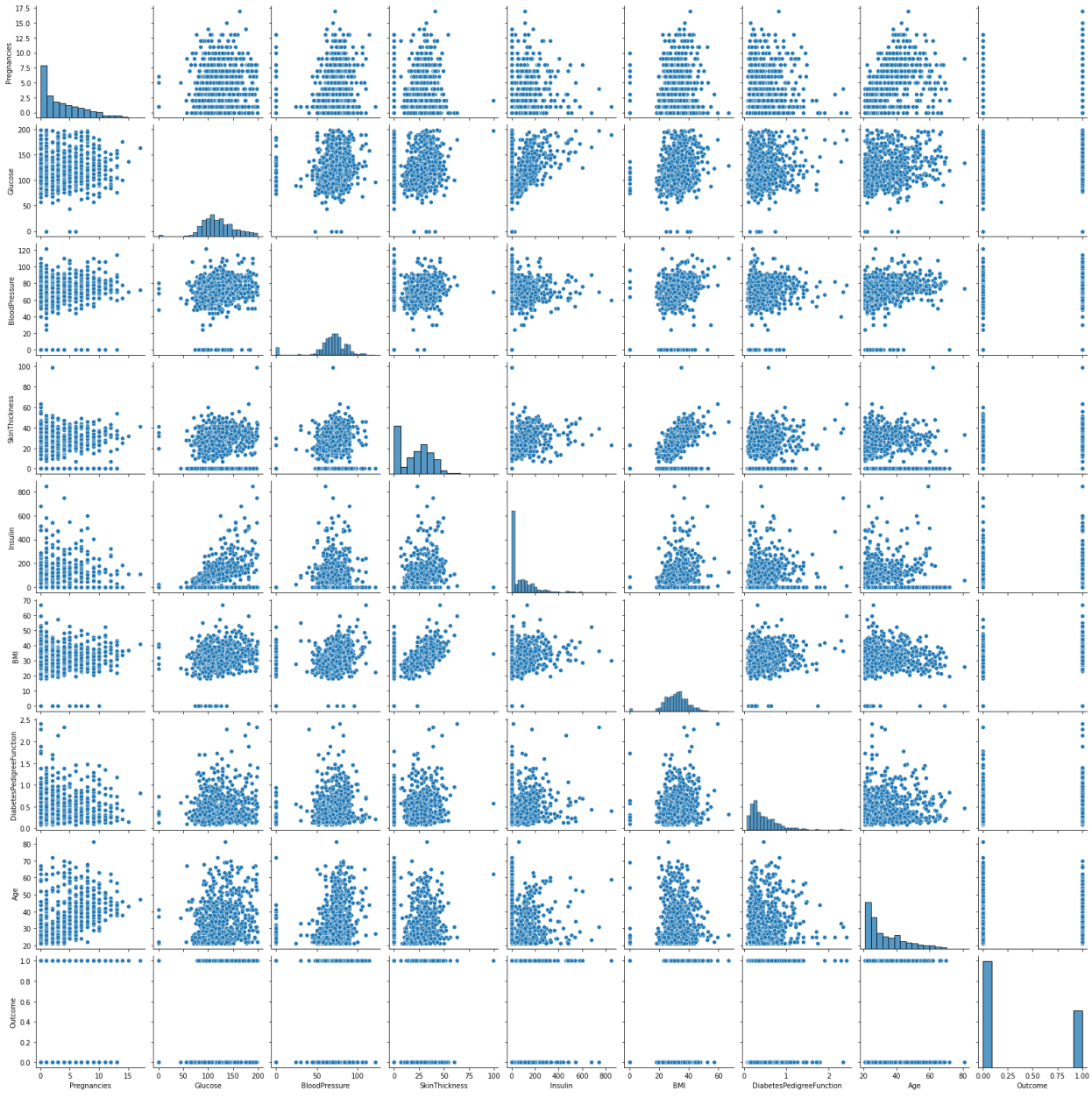
Без умаления общности виден линейный характер:



Гистограммы признаков не выявляют аномальных отклонений, за исключением быть может кровяного давления, но непонятно насколько это скажется потом.

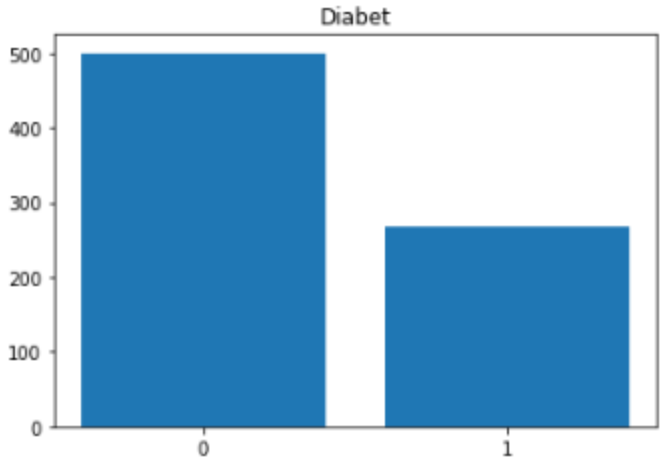


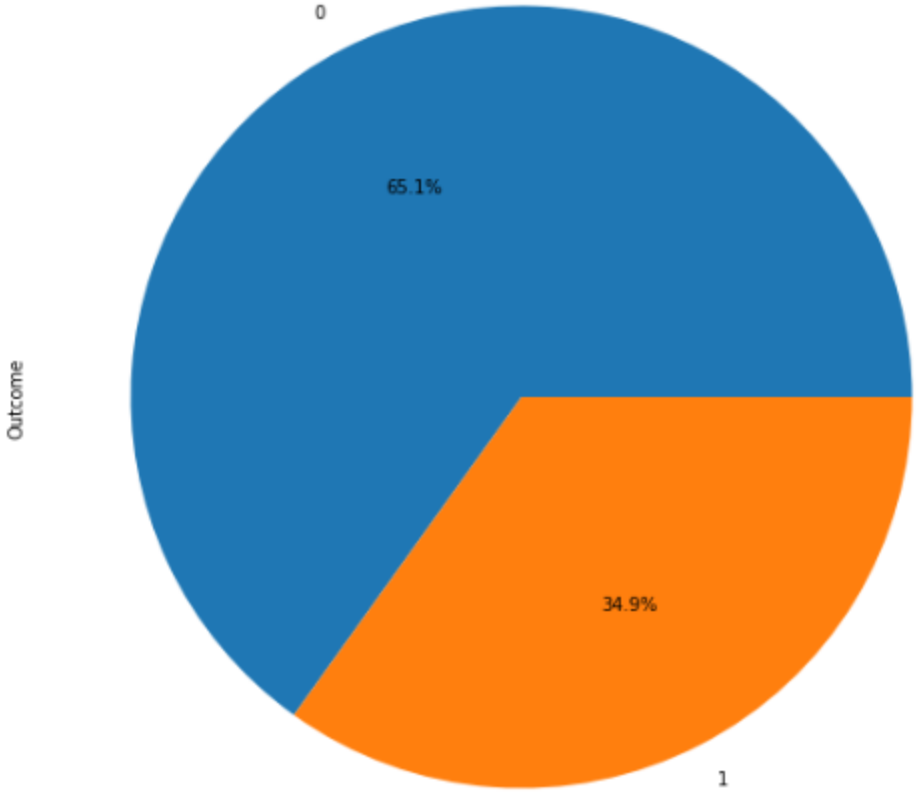
Посмотрю на графики зависимостей и ещё учитывая класс “есть ли у человека диабет”:


Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome

Из этого возможно, следует, что данные плохо линейно разделимы.

Посмотрю на соотношение классов.



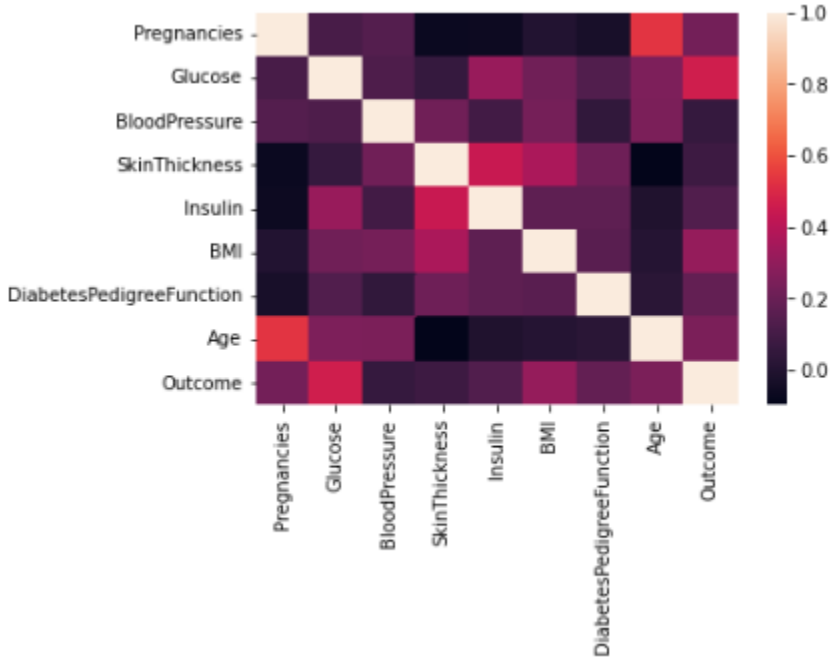


Они не сбалансированы. Я применил оверсемплинг.

Посмотрю на корреляцию:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.108788	0.144639	-0.067582	-0.060599	0.004458	-0.028003	0.533481	0.225113
Glucose	0.108788	1.000000	0.126307	0.059712	0.320267	0.215145	0.134161	0.249338	0.467631
BloodPressure	0.144639	0.126307	1.000000	0.217749	0.093494	0.230434	0.043347	0.245527	0.058003
SkinThickness	-0.067582	0.059712	0.217749	1.000000	0.442795	0.366982	0.211698	-0.098301	0.080102
Insulin	-0.060599	0.320267	0.093494	0.442795	1.000000	0.168833	0.168290	-0.004939	0.135294
BMI	0.004458	0.215145	0.230434	0.366982	0.168833	1.000000	0.156719	0.009610	0.311697
DiabetesPedigreeFunction	-0.028003	0.134161	0.043347	0.211698	0.168290	0.156719	1.000000	0.022266	0.184406
Age	0.533481	0.249338	0.245527	-0.098301	-0.004939	0.009610	0.022266	1.000000	0.246268
Outcome	0.225113	0.467631	0.058003	0.080102	0.135294	0.311697	0.184406	0.246268	1.000000

И тепловую карту:



Похоже, что многие признаки друг с другом коррелируют.

3. Выводы

Выглядит как будто модель готова к дальнейшим алгоритмам обучения, но кажется, что возникнут проблемы с проведением прямой.

В ходе работы применил оверсемплинг дабы сбалансировать классы