Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Изображение выглядит как текст, керамические изделия, фарфор

Автоматически созданное описание

**Курсовая работа по дисциплине**

**«Методы машинного обучения»**

**на тему:**

**«Определение диабета на основе диагностических измерений»**

**ИСПОЛНИТЕЛЬ:**

Ханмурзин Т.И. ИУ5-31М

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

"\_\_"\_\_\_\_\_\_\_\_\_\_\_2021 г.

Москва 2021

# Оглавление

[Оглавление 2](#_Toc90950316)

[Задание 3](#_Toc90950317)

[Подготовка данных 4](#_Toc90950318)

[Загрузка датасета книг 4](#_Toc90950319)

[Устранение пропусков данных 4](#_Toc90950320)

[Обработка категориальных признаков 5](#_Toc90950322)

[Нормализация числовых признаков 6](#_Toc90950323)

[Масштабирование признаков 6](#_Toc90950324)

[Отбор признаков 7](#_Toc90950325)

[Результат работы моделей 8](#_Toc90950326)

[AutoML 9](#_Toc90950327)

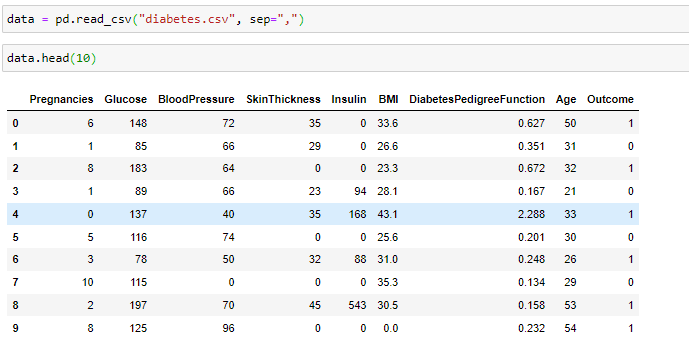
# Задание

1. Поиск и выбор набора данных для построения модели машинного обучения. На основе выбранного набора данных строится модель для задачи классификации.
2. Для выбранного датасета решить следующие задачи:
   1. устранение пропусков в данных;
   2. кодирование категориальных признаков;
   3. нормализацию числовых признаков;
   4. масштабирование признаков;
   5. обработку выбросов для числовых признаков;
   6. обработку нестандартных признаков (которые не является числовым или категориальным);
   7. отбор признаков, наиболее подходящих для построения модели;
3. Обучить модель и оценить метрики качества для двух выборок:
   1. исходная выборка, которая содержит только минимальную предобработку данных, необходимую для построения модели (например, кодирование категориальных признаков).
   2. улучшенная выборка, полученная в результате полной предобработки данных в пункте 2.
4. Построить модель с использованием произвольной библиотеки AutoML.
5. Сравнить метрики для трех полученных моделей.

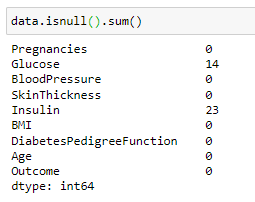
# Подготовка данных

## Загрузка датасета

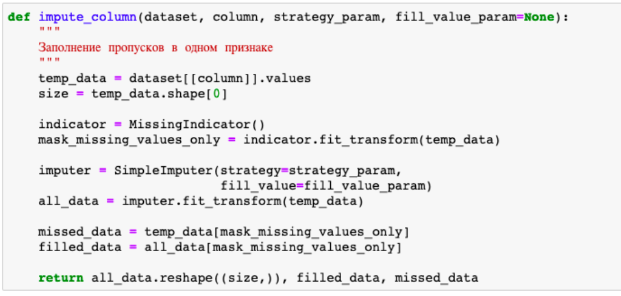
Используется датасет пациенток от 21 лет из Национального института диабета, болезней органов пищеварения и почек. Цель набора данных – определить, есть ли у пациента диабет, на основе определенных диагностических измерений, включенных в набор данных.



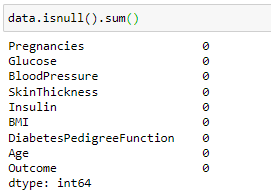
## Устранение пропусков данных



С помощью функции impute\_column убираем пропуск по каждому признаку.

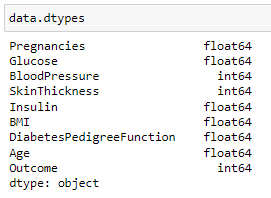


Колонки заполнены.

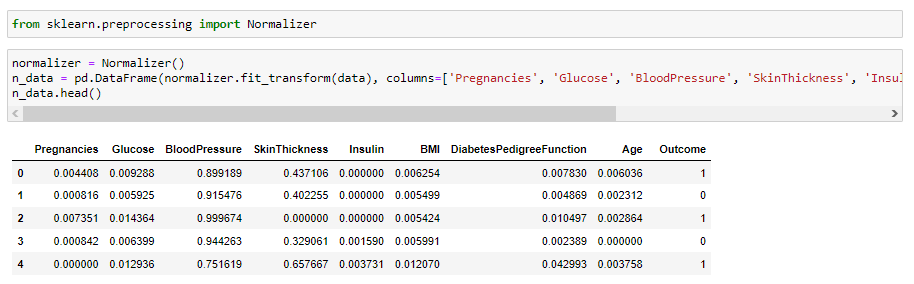


## Обработка категориальных признаков

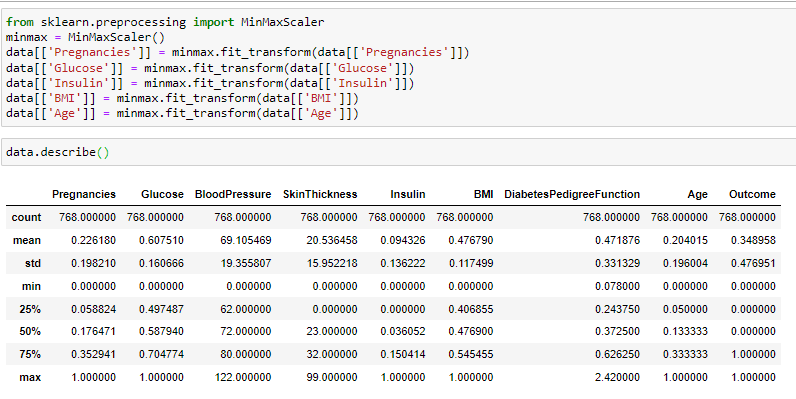
Датасет не содержит категориальных признаков.



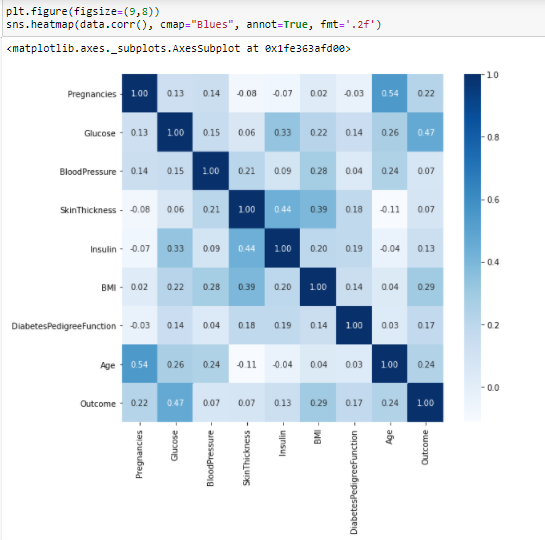
## Нормализация числовых признаков



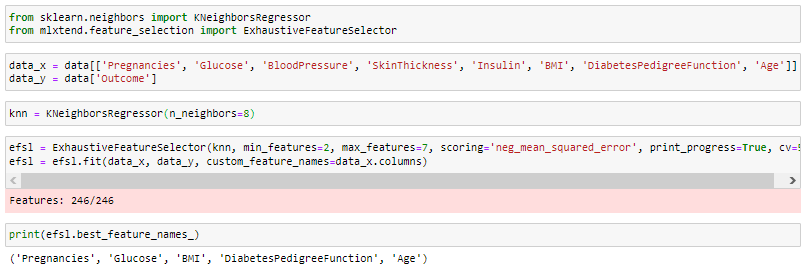
## Масштабирование признаков



## Отбор признаков



Выберем признаки не отталкиваясь от субъективного мнения с помощью ExhaustiveFeatureSelector.



# Результат работы моделей

При подготовленных данных:

LinR – 87,5232%

KNeighborsClassifier – 79,4363%

Tree – 88,2384%

GB – 87,6372%

RandomForestClassifier – 88,2785%

Если данные не подготовлены:

LinR – 72,4783%

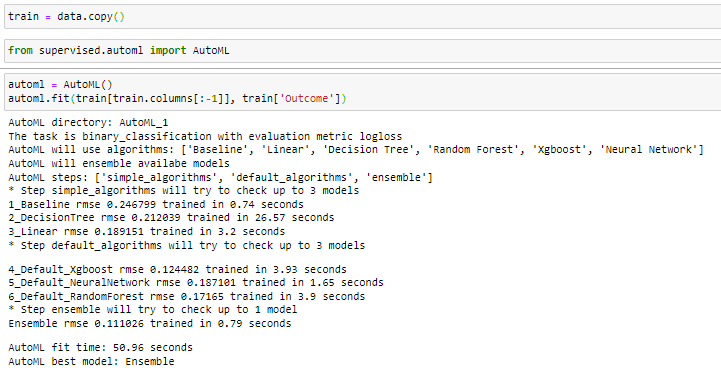
KNeighborsClassifier – 63,8842%

Tree – 73,5374%

GB – 72,9456%

RandomForestClassifier – 76,2345%

# AutoML



Для полностью подготовленного датасета точность Ансамблевой модели составляет ~88,8% при обучении самой AutoML в 50.96 секунд, хороший результат для таких данных.

**Выводы**

Можно сделать вывод, что предобработка данных очень важна для решения задач машинного обучения, это доказывают результаты оценки точности обучения.

**Список источников**

1. Методы машинного обучения. Лекции. Московский государственный технический университет имени Н.Э.Баумана. Автор: Гапанюк Ю.Е. - 2021