# Проблемы в данных и решения

* + Проблема: Наличие выбросов в медицинских показателях
  + Решение: Винсоризация выбросов
  + Проблема: Сильный дисбаланс в количестве пропусков между признаками
  + Решение: Замена пропусков медианой для SkinThickness и Insulin и средним значением для Glucose, BloodPressure, BMI

# Гипотезы для моделирования

1. Возраст является важным фактором риска развития диабета
2. Высокие значения глюкозы у пациента значительно повышают риск диабета
3. Высокий индекс массы тела у пациента значительно повышает риск диабета

# Ключевые выводы

1. Выполнено исследование качества данных:

* Выявлены и обработаны скрытые пропуски
* Проведен анализ распределений всех переменных
* Обнаружен дисбаланс классов (65% без диабета и 35% с диабетом)
* Выполнен статистический анализ выбросов методом IQR

1. Проведено исследование взаимосвязей:

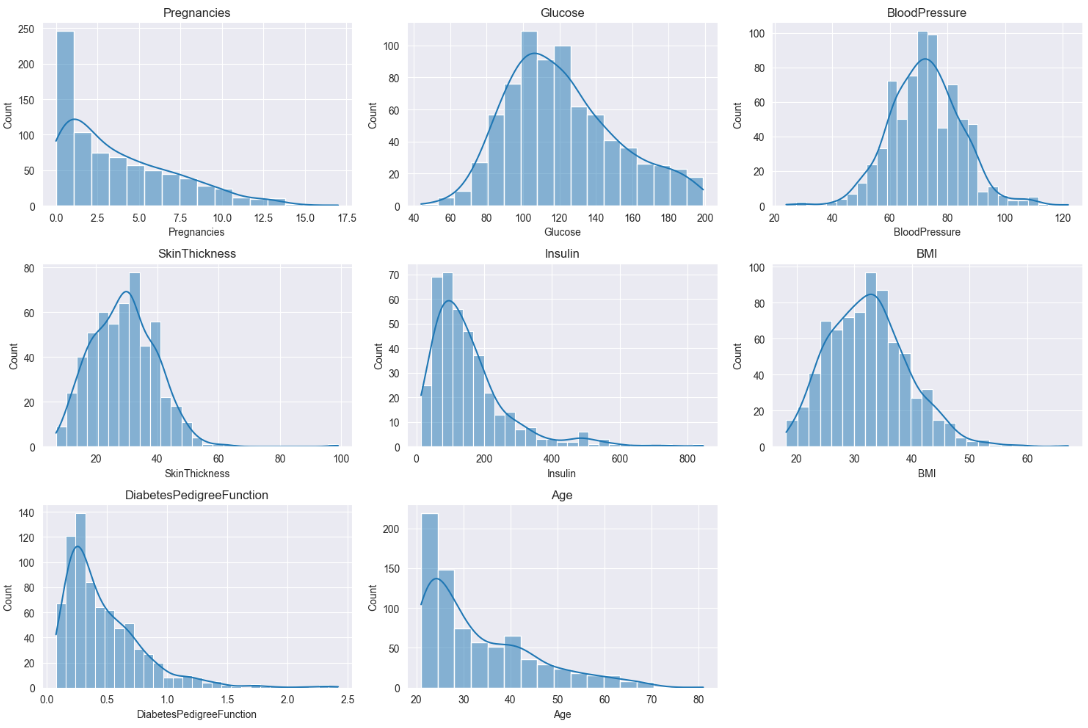
* Применены статистические тесты для анализа значимости признаков
* Построена корреляционная матрица для выявления взаимосвязей (уровень глюкозы наиболее важный предиктор диабета)

1. Осуществлена предобработка данных:

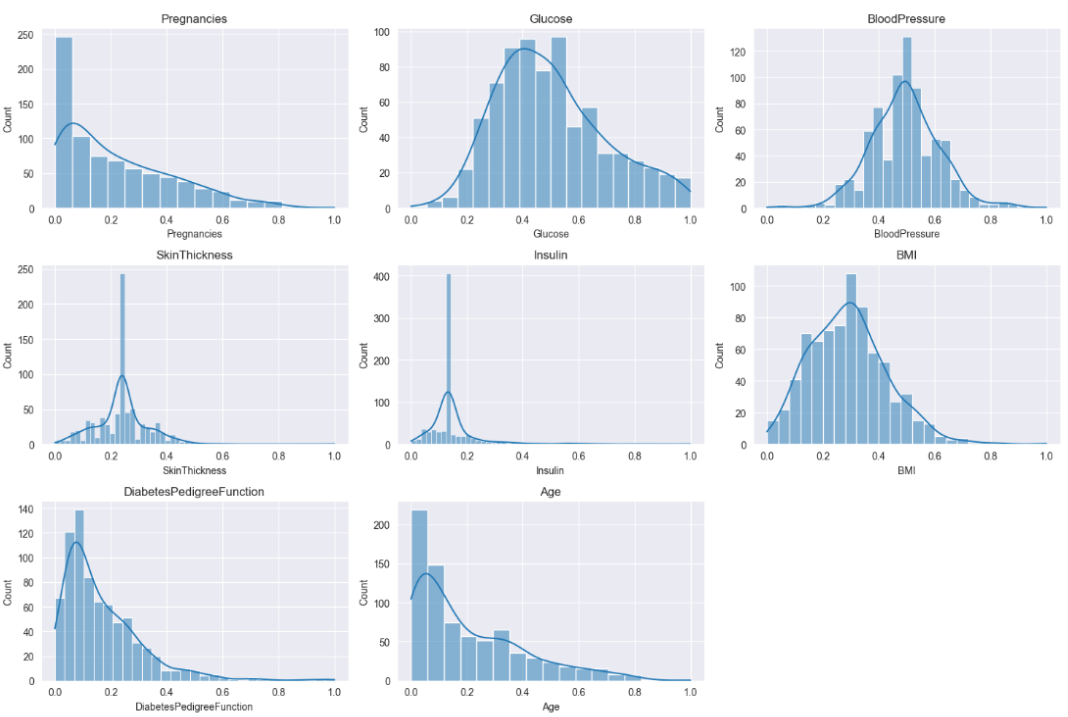
* Импутация пропусков с использованием медианы и среднего
* Масштабирование признаков
* Обработка выбросов методом винсоризации

# Ключевые графики

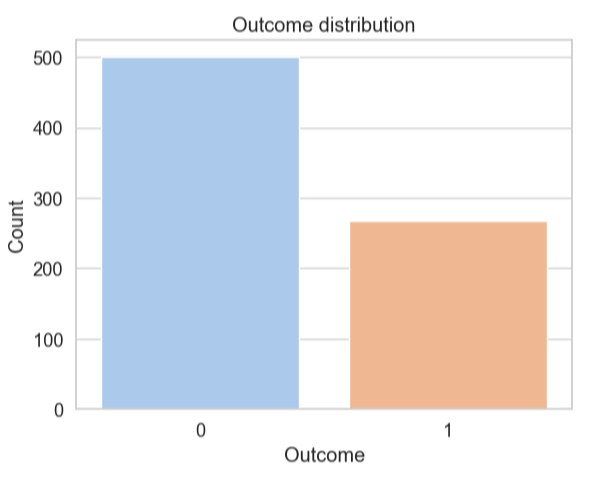
## До нормализации:



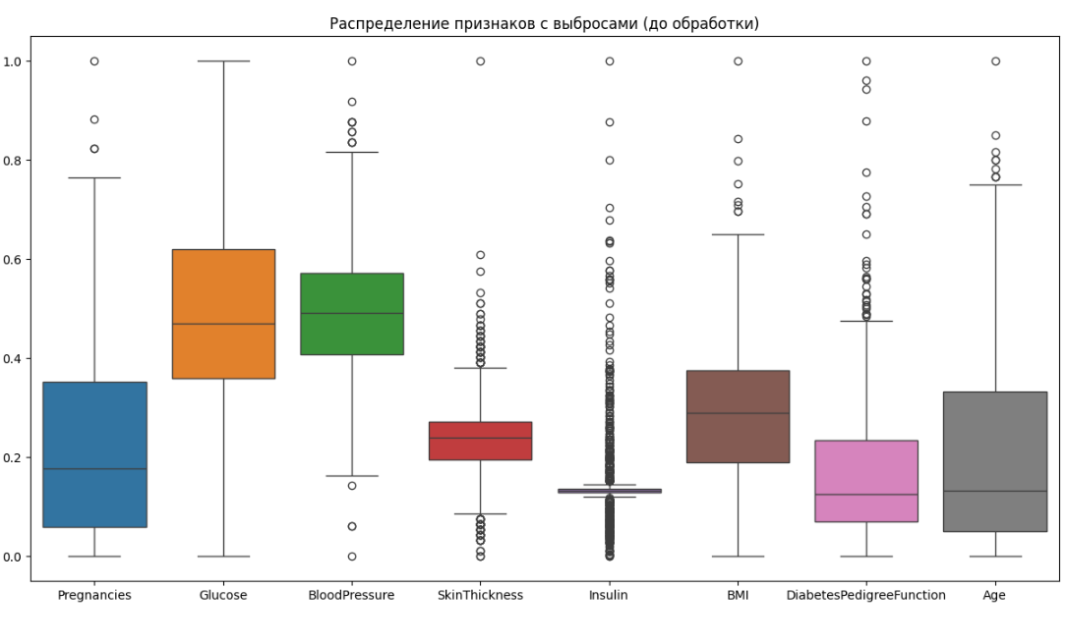
## После нормализации:



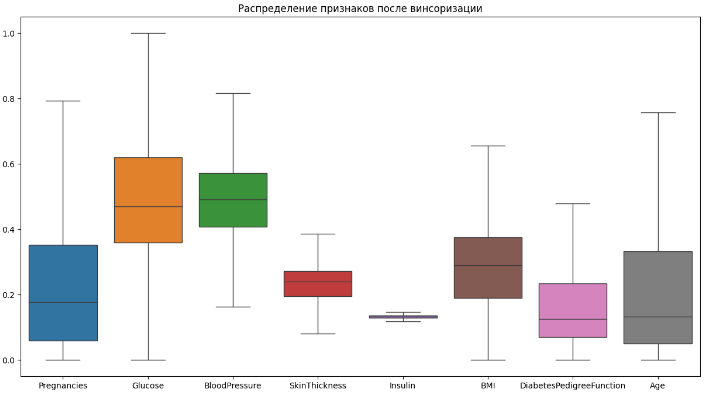
## Распределение классов:



## Распределение выбросов:



## Распределение выбросов после винсоризации:



## Корреляционная матрица:

