**Тема: Предсказание риска развития диабета**

Чубара Дарья

*Технический университет Молдовы*

*Факультет Вычислительной Техники,*

*Информатики и Микроэлектроники*

e-mail: daria.ciubara@iis.utm.md

**Абстракт**

В данной работе представлен анализ набора данных по предсказанию диабета, включающего медицинскую и демографическую информацию о пациентах, а также их статус по диабету (положительный или отрицательный). Данные включают такие характеристики, как возраст, пол, индекс массы тела (ИМТ), гипертония, заболевания сердца, курение, уровень HbA1c и уровень глюкозы в крови.

Набор данных может быть использован для построения моделей машинного обучения с целью предсказания развития диабета у пациентов на основе их медицинской и демографической истории. Это может быть полезно для медицинских специалистов при выявлении пациентов, подверженных риску развития диабета, и для разработки персонализированных планов лечения.

Кроме того, набор данных может быть использован исследователями для изучения взаимосвязей между различными медицинскими и демографическими факторами и вероятностью развития диабета.

**Введение**

Данное исследование нацелено на разработку прогностических моделей для предсказания развития диабета, используя обширный датасет, состоящий из более чем 100 тысяч записей о пациентах. Информация включает в себя медицинские и демографические характеристики, такие как возраст, пол, индекс массы тела (ИМТ), наличие гипертонии, заболевания сердца, курение, уровень HbA1c и уровень глюкозы в крови.

Основной целью проекта является создание эффективных моделей машинного обучения для предсказания развития диабета на основе предоставленных данных. В рамках исследования применены различные методы статистического анализа и машинного обучения, включая логистическую регрессию. Анализ выявил взаимосвязи между различными медицинскими факторами и вероятностью развития диабета.

Результаты данного исследования представляют важный вклад в область предсказания диабета, предоставляя ценную информацию для медицинских специалистов. Эти результаты могут быть использованы для выявления рисковых групп пациентов и разработки персонализированных стратегий лечения, способствуя более эффективному управлению этим серьезным заболеванием.

**Материалы и Методы**

В данном исследовании проводится анализ набора данных по предсказанию диабета, состоящего из более чем 100 тысяч записей о пациентах. Датасет был получен из различных источников, включая научные исследования и медицинские учреждения. Важными характеристиками датасета являются медицинские и демографические параметры, такие как возраст, пол и т.д.

В отличие от некоторых исследований, данная работа не требовала предварительной очистки данных от пустых значений, поскольку они в датасете отсутствовали. Процесс анализа начинается с оценки взаимосвязей между различными параметрами, применяя метод логистической регрессии. Предоставленные коэффициенты модели используются для интерпретации влияния каждого параметра на вероятность развития диабета.

Тестирование модели включает в себя анализ коэффициентов с использованием различных методов, таких как изучение коэффициентов и построение графика предсказаний и метрик.

**Результаты**

В ходе проекта были получены результаты, описанные в этой части статьи. В выбранном мною датасете всего 8 признаков, так или иначе влияющих на развитие диабета. Некоторые из них прямо указывают на то, что человек находится в группе риска, другие же только затрудняют точный показатель диагноза. Вот все переменные моего датасета:

1. Возраст (Age): Возраст пациентов.

2. Пол (Gender): Пол пациентов.

3. Индекс массы тела (BMI): Показатель массы тела.

4. Гипертония (Hypertension): Наличие или отсутствие гипертонии.

5. Болезнь сердца (Heart Disease): Наличие или отсутствие болезни сердца.

6. История курения (Smoking History): Информация о курении в медицинской истории.

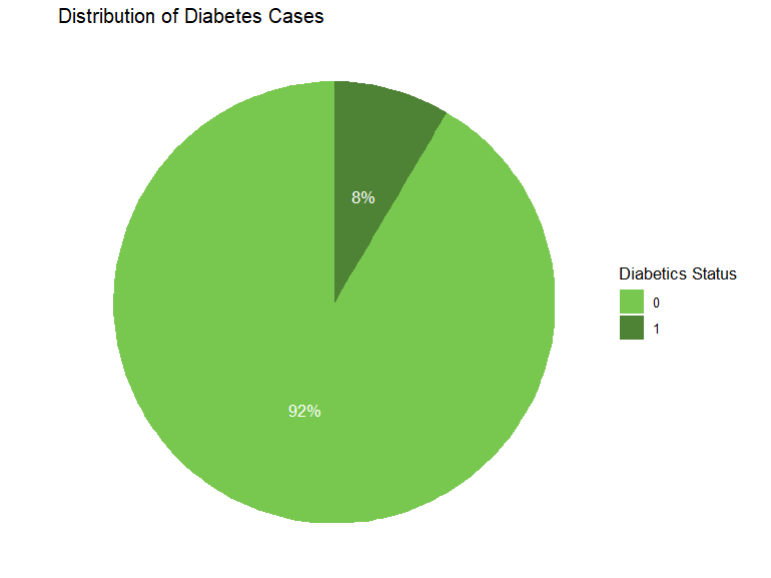
7. Уровень HbA1c (HbA1c Level): Уровень HbA1c в крови.

8. Уровень глюкозы в крови (Blood Glucose Level): Уровень глюкозы в крови.

9. Статус Диабета (Diabetes Status): Диагноз диабета (положительный или отрицательный). Зависимая переменная датасета.

Эти переменные используются для построения модели машинного обучения для предсказания вероятности развития диабета на основе медицинских и демографических данных пациентов.

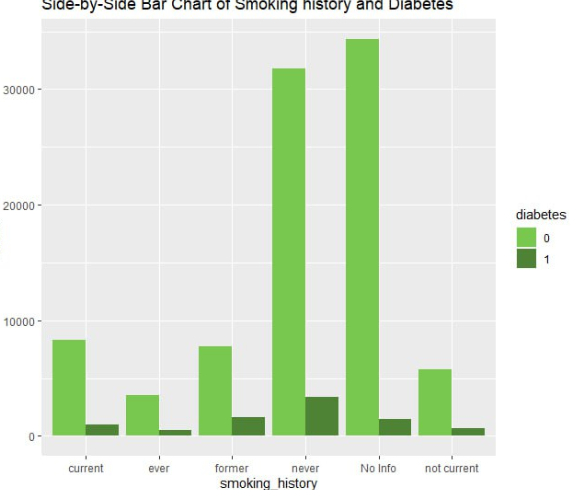
Сначала было важно преобразить данные, отмеченные как "character", в формат "factor", поскольку многие статистические модели в R демонстрируют более эффективную работу с факторными переменными. Моя зависимая переменная распределена неравномерно, всего 8% датасета составляют люди, болеющие диабетом.



**Рис.1 Распределение зависимой переменной**

Помимо этого были и переменные. Которые не несли важной информации.

Примером может послужить история курения, которая указывает, что люди, не курившие никогда в жизни, склонны к заболеванию диабетом куда больше тех, кто курил или курит всю жизнь. Поэтому эта переменная не была включена при создании модели логистической регрессии.

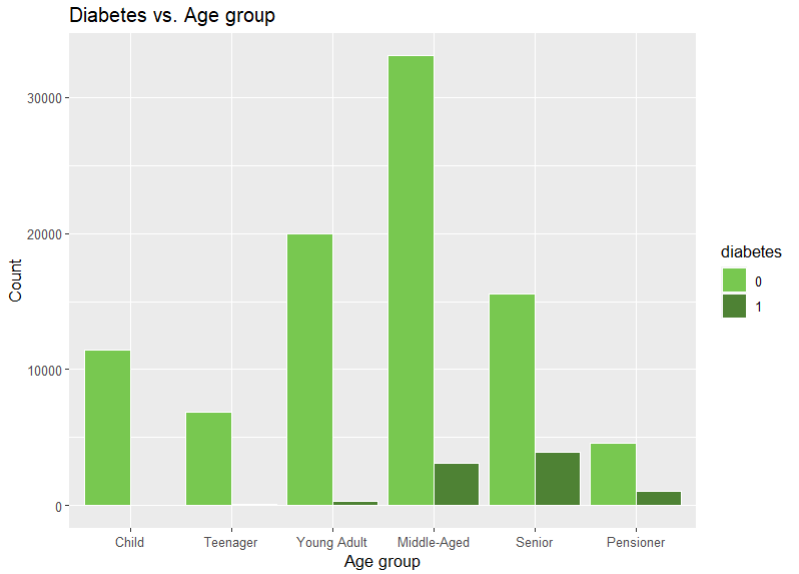


**Рис.2 История курения**

Одной из важных переменных датасета была та, что отвечает за возраст опрошенных людей. Разделив все возраста на несколько групп, я могу наглядно увидеть, в какой из этих групп больше всего больных людей. Группы были разделены так:

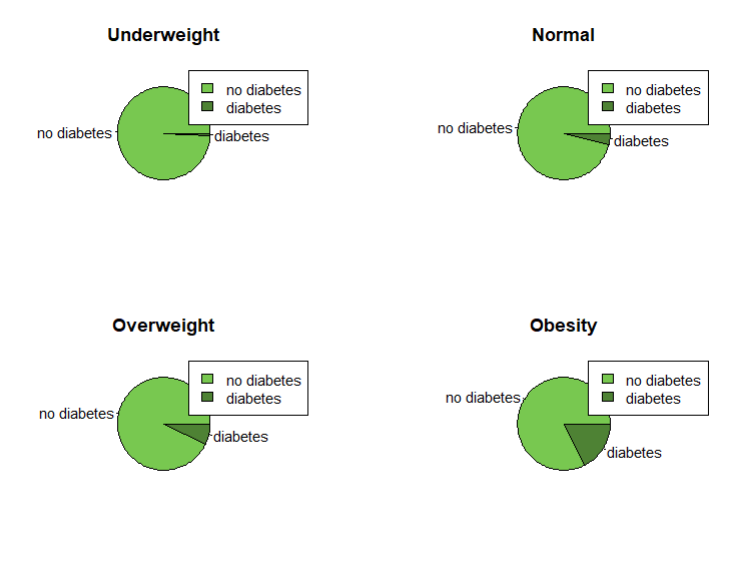
1. Child (Ребенок): 0-12 лет
2. Teenager (Подросток): 13-19 лет
3. Young Adult (Молодой взрослый): 20-35 лет
4. Middle-Aged (Средний возраст): 36-60 лет
5. Senior (Старший возраст): 61-80 лет
6. Pensioner (Пенсионер): 81 лет и старше

Наиболее распространена болезнь именно в группе Senior, следом за которой лидирует средний возраст.



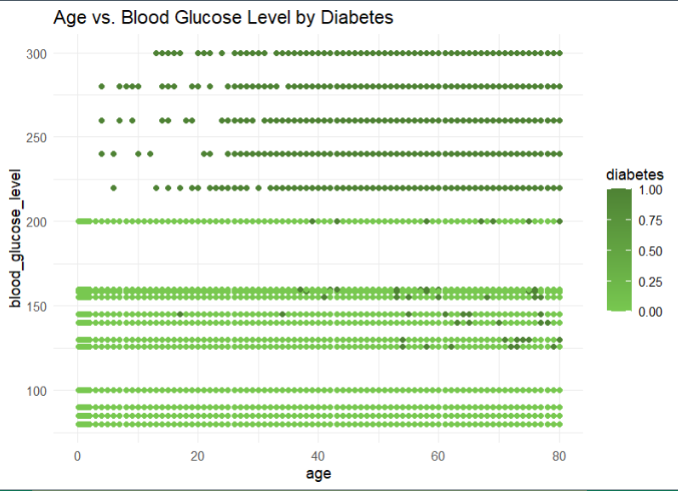
**Рис.3 Взаимосвязь возраста и наличия диабета**

Далее по такому же принципу группы людей были разделены на категории по индексу массы тела. Нормальный индекс массы тела для взрослого человека составляет 19-25. Ниже него идет недовес. Далее небольшой лишний вес (от 26 до 30) и ожирение (выше 30). Заметно, что пусть больные люди есть во всех категориях, чем выше вес, тем выше процент больных.



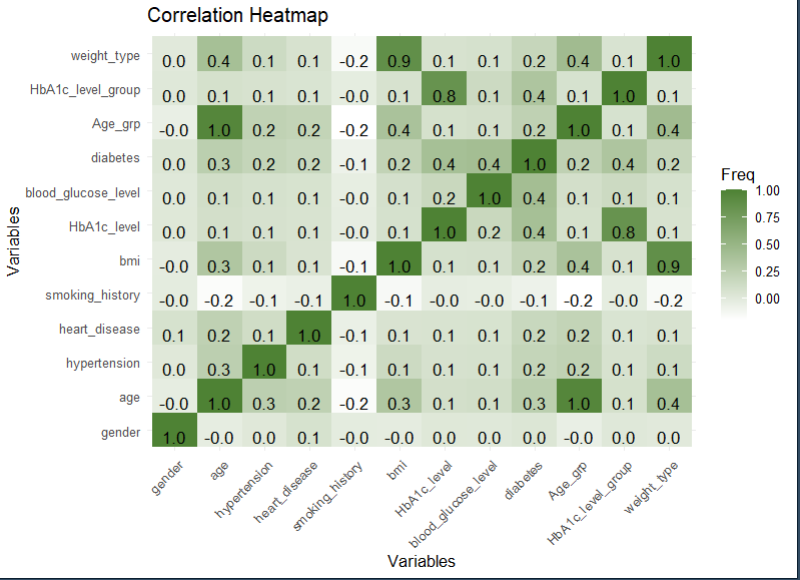
**Рис.4 Взаимосвязь веса и наличия диабета**

Далее рассматриваем важность уровня глюкозы. Нормальный показатель глюкозы находится в пределах - 70 -130. Пограничный - 131 - 180. Высокий 181- 250. А более 250 - опасный. Чем выше уровень глюкозы, тем больше процент больных. А там, где ее уровень достигает опасного, абсолютно каждый пациент является больным.



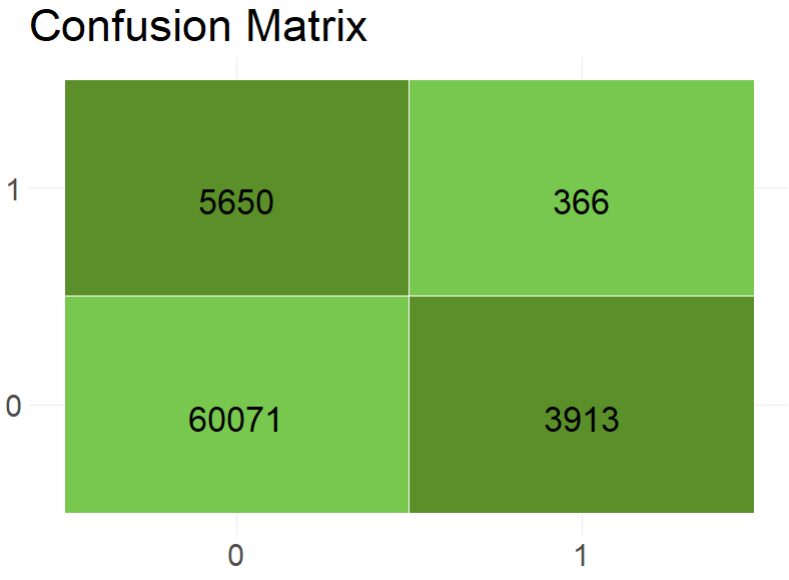
**Рис.5 Взаимосвязь уровня глюкозы и наличия диабета**

Построив матрицу корреляции, было замечено, что корреляций нет, ни положительных, ни отрицательных, это означает, что нет четкой тенденции или закономерности.



**Рис.6 Матрица корреляции**

Была построена модель множественной логистической регрессии. Она показала значения точности (Accuracy) 0.95. Позже была реализована матрица ошибок (confusion matrix) для модели классификации. В нашей матрице ошибок такие результаты:



**Рис.7 матрица ошибок**

Также были выведены метрики для анализа модели, которые предоставили следующие результаты:

* Полнота (Recall): 0.95
* Точность (Precision): 0.69
* F1-мера: 0.80

ROC Curve плавная, имеет высокий AUC, высокую чувствительность и специфичность, отсутствие перекосов.



**Рис.8 ROC Curve**

**Обсуждение**

Представленный датасет для прогнозирования диабета представляет собой ценный ресурс для исследований в области медицинского прогнозирования. Анализ медицинских и демографических данных имеет потенциал для разработки точных методов предсказания развития диабета.

В рамках дальнейших исследований рекомендуется провести более глубокий анализ взаимосвязей между различными параметрами датасета с целью улучшения предсказательных моделей. Это может включать в себя исследование влияния каждого параметра на вероятность развития диабета и разработку более сложных моделей машинного обучения для повышения точности прогнозов. Такие усилия могут значительно улучшить эффективность диагностики и разработки персонализированных подходов к лечению диабета.

**Код**

Код можно найти по ссылке

**Данные**

Данные можно найти по ссылке