Предсказание риска диабета

Максим Назаров 1

*1- Технический университет Молдовы (UTM)*

g-mail: mxvnazarov@gmail.com

**Абстракт**

В данном исследовании представлен анализ медицинских и демографических данных, относящихся к пациентам, а также оценка их диабетического статуса (положительного или отрицательного). С целью более глубокого понимания данных и выделения ключевых тенденций, мы провели анализ данных с использованием исследовательского метода. Этот метод включал в себя визуализацию данных, что позволило выявить потенциальные закономерности и взаимосвязи между различными переменными.

**Введение**

Данное исследование в направлении повышения точности прогнозирования вероятности развития диабета. Этот обширный набор данных становится ценным ресурсом для создания моделей машинного обучения, способных предсказывать вероятность диабета у пациентов на основе их медицинской и демографической истории.

Полученные результаты могут быть весьма полезными для медицинских специалистов в выявлении потенциальных кандидатов, подверженных риску развития диабета, и в формировании индивидуализированных планов лечения. Кроме того, этот набор данных предоставляет возможность исследователям изучить взаимосвязи между различными медицинскими и демографическими факторами и вероятностью развития диабета.

Целью проекта является глубокий анализ риска развития диабета, выявление факторов, воздействующих на его появление, а также разработка моделей для точных прогнозов вероятностей. Исследование охватывает 100 000 строк данных, что позволяет выявить не только общие тенденции, но и уникальные особенности, связанные с данной болезнью.

Для достижения поставленных целей использовались разнообразные методы статистического анализа и машинного обучения, включая логистическую регрессию, метод случайного леса и метод опорных векторов. Эти методы способствовали изучению взаимосвязей между различными аспектами развития диабета, такими как возраст, пол, индекс массы тела, история курения, гипертония, и их значимость в контексте исследования.

Таким образом, данное исследование представляет собой значимый шаг в области предсказания вероятности развития диабета, обогащая наше понимание связей между различными переменными и предоставляя важные инсайты для дальнейших медицинских исследований.

**Материалы и Методы**

Для данного исследования мы воспользовались обширной коллекцией медицинских и демографических данных пациентов, включающей характеристики, такие как возраст, пол, индекс массы тела (ИМТ), наличие гипертонии, сердечно-сосудистых заболеваний, история курения, уровень HbA1c и уровень глюкозы в крови.

Начальным этапом исследования была предварительная обработка данных, а затем, с целью анализа взаимосвязей между различными признаками, были использованы методы корреляционного анализа и визуализации данных. Все вычисления проводились с использованием языка программирования R и соответствующих библиотек, предварительно установленных в рамках пакетов.

Следующим этапом было создание и обучение моделей для прогнозирования риска развития диабета. Мы применили разнообразные методы машинного обучения, включая логистическую регрессию, метод случайного леса и метод опорных векторов. Датасет был разделен на обучающую (70%) и тестовую (30%) выборки, что позволило эффективно оценить производительность моделей.

Неотъемлемой частью проекта стал этап тестирования выбранных моделей. Для этого применялись различные методы анализа ошибок, оценки точности и достоверности предсказаний, такие как изучение признаков, построение матрицы ошибок и использование метрик оценки точности (Accuracy). Эти шаги позволили не только создать модели с высокой прогностической способностью, но и более глубоко понять их эффективность и применимость в контексте исследования.

**Результаты**

В ходе проекта были получены результаты, которые могут быть использованы для дальнейших исследований данной тематики. В выбранном датасете 9 признаков, из которых 7 числовых и 2 категориальных. Так признаками датасета стали: возраст, пол, индекс массы тела (ИМТ), наличие гипертонии, сердечно-сосудистых заболеваний, история курения, уровень HbA1c и уровень глюкозы в крови.

Первым делом необходимо было представить данные типа «character» в «factor», так как многие статистические модели в R работают более эффективно с факторными переменными. Также проверено количество пропущенных значений, удивительно, но их не было, все данные чистые. Также зависимая переменная «diabetes» распределена неравномерно, см.рис.1, где 91500 записей людей, не болеющих диабетом и лишь 8500 имеют положительные результаты. Следующими шагами было анализ влияния других переменных на зависимую.

Изображение выглядит как снимок экрана, диаграмма, круг, Графика

Автоматически созданное описание

**Рис. 1 Распределение зависимой переменной**

После дальнейшего анализа было замечено, что при построении гистограммы возраста, заметно, что датасет имеет обширную возрастную категорию от детей до взрослого поколения, это можно увидеть на рисунке 2. Также видно, что преобладает больше людей, которым 50 лет, на это указывает мода.

**Изображение выглядит как График, диаграмма, текст, линия

Автоматически созданное описание**

**Рис. 2 Распределение возраста пациентов**

Также можно отметить, что половина признаков имеют лишь 2 уникальных значения, такие как гендер (мужской/женский), наличие гипертонии (да/нет) и сердечно-сосудистых заболеваний(да/нет). Остальные имеют больше уникальных значений, возраст (102), индекс массы тела (247), история курения (6 уровней), уровень HbA1c (18) и уровень глюкозы в крови(18).

Рассмотрев подробнее уровни признака курения и построив гистограмму, видно, что больше всего данных, не имеющих информацию о курении. Также самый высокий показатель диабета у людей никогда не куривших (рис. 3).

Изображение выглядит как текст, снимок экрана, диаграмма, Шрифт

Автоматически созданное описание

**Рис. 3 Гистограмма уровней курения у болеющих и неболеющих**

После анализа зависимостей категориальных признаков с целевой переменной, были проанализированы численные. Построив матрицу корреляции, было замечено, что корреляций нет, ни положительных, ни отрицательных, где показано на рис.4, это означает, что нет чёткой тенденции или закономерности. Самая высокая корреляция составила 0.4 и 0.42 между диабетом и уровня HbA1, зависимой переменной и уровня глюкозы в крови. (но это не считается хорошей корреляцией).

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

**Рис. 4 Матрица корреляции для численных переменных**

Следующим шагом были проанализированы зависимости целевой переменной уровня HbA1. Уровень HbA1c (гемоглобин А1c) является одним из важных показателей здоровья человека, связанным с уровнем глюкозы в крови. HbA1c измеряет процент гемоглобина, который связан с глюкозой в крови в течение определенного времени. Этот показатель предоставляет информацию об уровне средней концентрации глюкозы в крови за последние 2-3 месяца.

Для людей с диабетом уровень HbA1c является важным индикатором эффективности управления уровнем глюкозы. Более низкий уровень HbA1c часто связывают с лучшим контролем сахарного диабета. Нормальные значения могут различаться в зависимости от лаборатории, но обычно нормальный уровень HbA1c составляет менее 5.7%. Значения от 5.7% до 6.5% могут указывать на предиабет, а 6.5% и выше часто связывают с диабетом.

Этот показатель является важным инструментом для врачей в оценке риска развития диабета, а также для пациентов с диабетом в мониторинге их уровня глюкозы и эффективности лечения. Для этого было решено изменить данные этого признака, чтобы посмотреть какое текущее состояние уровня гемоглобина в данном датасете.

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

**Рис. 5 График уровня гемоглобина в крови**

Из представленной графика видно, что распределение уровня HbA1c в данной группе пациентов разделяется на три основные категории: "Normal" (Норма), "Pre-Diabetic" (Предиабет) и "Diabetic" (Диабет). Подсчет общего количества наблюдений показывает, что большинство пациентов находятся в категории "Normal", что соответствует нормальному уровню глюкозы в крови

Однако, при дальнейшем рассмотрении диагнозов по диабету внутри каждой категории HbA1c, становится ясно, что с увеличением уровня HbA1c увеличивается и процент случаев диабета. Это подтверждает связь между повышенным уровнем HbA1c и повышенным риском развития диабета.

Общее количество случаев диабета составляет 8,500 из 99,982 наблюдений, что подчеркивает важность мониторинга и контроля уровня HbA1c для предотвращения и управления диабетом. Следовательно, анализ данных указывает на значимость данного биомаркера для оценки риска и принятия мер по поддержанию здоровья пациентов.

Следующим этапом осуществляется разработка и оценка прогностических моделей для оценки вероятности риска диабета на основе важных характеристик. Проект продвигался к построению нескольких моделей предсказания, выявляя, что модель показывает наилучшие результаты при использовании всех признаков. Данные были разделены на обучающий и тестовый наборы в соотношении 70% к 30% (70 000 и 30 000 случаев соответственно).

Был представлен статистический анализ результатов регрессионной модели, указывающий на то, что параметры, такие как пол, возраст, артериальное давление, индекс массы тела (ИМТ), курение, уровень HbA1c и уровень глюкозы в крови, имеют значительную связь с риском в данной модели. Все параметры показали положительные оценки с низкими p-значениями, что свидетельствует о статистически значимой положительной связи с риском. Важность каждого параметра определялась его оценкой и статистической значимостью (p-значение). Оценка положительного знака указывает на положительную связь с риском, а отрицательного – на отрицательную.

Исходно была построена модель множественной логистической регрессии. Она продемонстрировала значения точности (Accuracy) 0,9625, что является высоким показателем для данной задачи. В последующем была проведена анализ матрицы ошибок для модели классификации, показав следующие результаты:

- True Positive (TP): 14 627 правильных предсказаний положительного класса.

- True Negative (TN): 14 233 правильных предсказаний отрицательного класса.

- False Positive (FP): 631 неправильное предсказание положительного класса.

- False Negative (FN): 509 неправильных предсказаний отрицательного класса.

Дополнительные метрики, такие как Precision (точность), Recall (чувствительность) и F1-мера, подтвердили высокий уровень качества модели.

Precision = 0.958

Recall (чувствительность) =0.967

F1-score = 0.963

После этого были выбраны две дополнительные модели - случайного леса и опорных векторов, показавшие точности 0,972 и 0,965 соответственно.

Итак, на основе рассмотренных метрик можно сделать вывод о высокой эффективности построенной модели в решении задачи классификации, что подтверждается точностью в 96,25% и дополнительными метриками Precision, Recall и F1-мера.

**Обсуждение**

Исследование подчеркивает важность комплексного анализа данных риска диабета у пациентов. Эти результаты свидетельствуют о перспективности применения моделей данных для точной оценки риска диабета. Однако, следует учитывать ограничения данных и потенциальные искажения. Дальнейшие исследования могут включать в себя более глубокий анализ взаимосвязи различных параметров и разработку более совершенных моделей для прогнозирования наличия диабета.

**Код**

Код можно найти по ссылке https://github.com/Suuuuiuuuus/DS.git

**Данные**

Данные найти по ссылке https://github.com/Suuuuiuuuus/DS.git