**Министерство Образования и Исследований Молдовы**

**Технический Университет Молдовы**

**Факультет Вычислительной техники, Информатики и Микроэлектроники**

**Департамент Информатики и Системной Инженерии**

**Предмет:** Analiza Datelor

**Научная статья**

**На тему: «Анализ данных о диабете у молодых людей»**

**Выполнила:** Oprea A. гр. MI-213

**Кишинёв 2023**

*Статистика и анализ данных в образовании*

**Анализ данных о диабете у молодых людей**

**Опря Анастасия**

Студентка третьего курса, специальность Информационный Менеджмент, Технический Университет Молдова, Кишинев, Республика Молдова.

[anastasia.oprea@iis.utm.md](mailto:anastasia.oprea@iis.utm.md)

**Аннотация**

**Предмет.**

В данной статье проведено исследование, направленное на анализ факторов, влияющих на распространение диабета среди молодых людей. Проблема воздействия различных переменных на заболеваемость диабетом является актуальной и требует системного подхода для выявления ключевых аспектов, формирующих риск заболевания, что позволит создать оптимальные условия для профилактики и успешного управления здоровьем молодого поколения.

**Цели.**

Целью данного исследования является анализ факторов, оказывающих влияние на распределение диабета среди молодых людей, с использованием методов Exploratory Data Analysis, а также разработка анализа влияния (influence analysis) или расчет важности признаков на основе их коэффициентов и стандартных ошибок. Основной фокус направлен на изучение воздействия различных независимых переменных, таких как количество беременностей, уровень глюкозы, артериальное давление, толщина кожи, уровень инсулина, индекс массы тела (BMI), а также генетический фактор предрасположенности к диабету.

**Методология.**

Для проведения анализа данных в работе были применены методы Exploratory Data Analysis с целью оценки распределения и взаимосвязей между переменными. Методология анализа данных по диабету у молодых людей включает использование различных методов, визуализаций и статистических подходов для более глубокого понимания данных. Все методы в совокупности обеспечивают комплексный подход к анализу данных о диабете у молодых людей, позволяя выявить основные закономерности и влияние различных факторов на заболевание.

**Результаты.**

В процессе исследования данных о диабете у молодых людей были обнаружены интересные взаимосвязи. Проведенный анализ выявил, что риск развития сахарного диабета существенно связан с несколькими ключевыми переменными.

В частности, выявлено, что уровень глюкозы (Glucose) является одним из существенных факторов, оказывающих влияние на вероятность заболевания. Также отмечено, что индекс массы тела (BMI) и количество беременностей (Pregnancies) также имеют высокую степень связи с риском развития диабета у молодых людей.

Дополнительно, факторы, такие как генетическая предрасположенность к диабету (DiabetesPedigreeFunction), артериальное давление (BloodPressure), возраст (Age), уровень инсулина (Insulin) и толщина кожи (SkinThickness), также были выделены как значимые влияющие переменные.

Анализ также раскрывает, что уровень глюкозы, BMI, количество беременностей, DiabetesPedigreeFunction, артериальное давление, и возраст являются основными предикторами для определения вероятности развития диабета у молодых людей**.**

**Заключение.**

В ходе проведенного анализа данных о заболеваемости диабетом у молодых людей были выявлены существенные факторы, влияющие на вероятность развития данного заболевания. Результаты исследования показывают, что переменные, такие как уровень глюкозы, индекс массы тела (BMI), количество беременностей, генетическая предрасположенность к диабету (DiabetesPedigreeFunction), артериальное давление, возраст, уровень инсулина, и толщина кожи, существенно влияют на риск заболевания.

Анализ данных также подчеркивает, что особенно высок риск диабета у молодых людей с повышенным уровнем глюкозы, высоким BMI, а также при наличии предрасположенности к диабету по генетическим данным. Эти факторы оказывают значительное воздействие и могут быть ключевыми при формировании стратегий профилактики и управления здоровьем в данной возрастной группе.

В целом, результаты анализа данных о диабете у молодых людей выделяют важность учета множества переменных при оценке риска заболевания. Полученные выводы могут послужить основой для разработки эффективных программ по предотвращению и управлению диабетом, а также для формирования индивидуализированных подходов к заботе о здоровье молодого поколения.

**Ключевые слова:** диабет, молодые люди, анализ данных, исследование

**Введение**

* 1. **Актуальность темы**

Современное общество сталкивается с ростом случаев диабета среди молодых людей. Анализ этих данных не только информативен о текущей ситуации, но также позволяет прогнозировать будущие тенденции заболевания, что важно для создания эффективных стратегий здравоохранения. Исследование данных о диабете у молодежи выявляет потенциальные факторы риска, такие как генетика, образ жизни и питание, что влияет на разработку целенаправленных стратегий профилактики и управления на ранних стадиях.

Заболевание диабетом в молодом возрасте имеет серьезные последствия для долгосрочного здоровья. Исследование данных о диабете у молодежи глубже раскрывает особенности заболевания, становясь фундаментом для эффективных стратегий предотвращения и лечения. Исследование данных о диабете у молодежи поддерживает переход к персонализированной медицине, улучшая диагностику, терапию и предотвращение диабета.

Учитывая вышеописанные аспекты, тема анализа данных о диабете среди молодежи остается востребованной и обладает потенциалом для активного вклада в область здравоохранения и медицинских исследований.

**1.2 Цель и задачи**

В данной научной статье по анализу данных о диабете среди молодых людей, является выявление ключевых переменных, которые могут оказывать влияние на развитие заболевания. Это исследование направлено на более глубокое понимание факторов риска и формирование основы для разработки эффективных стратегий профилактики и управления диабетом среди молодых людей.

Проведение анализа направлено на выявление потенциальных факторов риска в развитии диабета среди молодых людей. Это включает анализ воздействия различных переменных, таких как беременность, уровень глюкозы, артериальное давление, толщина кожи, уровень инсулина, индекс массы тела (BMI), генетические предпосылки и возраст, на вероятность заболевания диабетом. Основная задача заключается в выделении ключевых факторов, которые могут быть связаны с увеличенным риском развития диабета у молодежи.

**Методология**

**2.1 Описание датасета**

В процессе исследования применялся обширный датасет, включающий информацию о 769 молодых людях. Основная цель заключалась в тщательном анализе воздействия различных факторов на развитие диабета. Этот объем данных предоставляет уникальную возможность более глубокого понимания динамики развития диабета и взаимосвязей между различными параметрами, оказывающими влияние на возникновение этого заболевания.

Данные в датасете, используемом для анализа диабета у молодых людей, могут содержать информацию о следующих переменных:

* Беременность (Pregnancies): Эта переменная обычно представляет собой число беременностей, которые перенесла женщина. В контексте анализа диабета, это может быть полезной информацией для определения риска развития диабета у женщин.
* Глюкоза (Glucose): Уровень глюкозы в крови является одним из ключевых показателей для диагностики диабета. Эта переменная представляет уровень глюкозы в крови.
* Артериальное давление (Blood Pressure): Артериальное давление измеряется с помощью двух значений: систолического (верхнего) и диастолического (нижнего) давления. Эти значения могут быть важными для оценки состояния сердечно-сосудистой системы у молодых людей с диабетом.
* Толщина кожи (Skin Thickness): это измерение толщины складки кожи на верхней части предплечья. Может быть связано с инсулинорезистентностью.
* Инсулин (Insulin): Уровень инсулина в крови может быть важным показателем для диагностики и управления диабетом.
* Индекс массы тела (ИМТ, BMI - Body Mass Index): ИМТ вычисляется на основе веса и роста человека и используется для оценки избыточного или недостаточного веса. Высокий ИМТ может быть фактором риска для развития диабета.
* Родословная диабета (Diabetes Pedigree Function): Эта переменная может содержать информацию о наличии диабета у близких родственников, что может свидетельствовать о генетической предрасположенности к болезни.
* Возраст (Age): Возраст пациента может играть важную роль при анализе диабета, так как риск развития болезни может увеличиваться с возрастом.
* Исход (Outcome): это бинарная переменная, которая обычно указывает на наличие или отсутствие диабета у пациента (например, 1 - диабет присутствует, 0 - диабет отсутствует).

**2.2 Источник данных**

Источник данных для проведения исследования представлен на платформе Kaggle, широко известной как ресурс для обмена исследовательскими наборами данных. Поскольку датасет обладает публичным доступом и используется в образовательных целях, он предоставляет возможность для обширного анализа и применения результатов исследования в практических аспектах образовательной сферы. Основная цель анализа этого датасета заключается в выявлении ключевых факторов, оказывающих влияние на развитие диабета у молодых людей. Это исследование способствует формированию более глубокого понимания динамики и взаимосвязей в сфере заболеваемости диабетом. Полученные результаты могут послужить основой для разработки рекомендаций, направленных на предотвращение или сдерживание распространения диабета.

**2.3 Инструменты и этапы исследования**

Первым шагом в работе над проектом было проведение разведочного анализа данных (EDA). Разведочный анализ данных (Exploratory Data Analysis) представляет собой предварительное исследование датасета с целью выявления его основных характеристик и взаимосвязей между признаками. Этот этап также направлен на уточнение выбора методов, используемых при построении модели машинного обучения [1]. Для реализации этого процесса мы воспользовались языком R - открытым исходным кодом, широко используемым в качестве статистического программного обеспечения и инструмента для анализа данных [2]. Этот язык обладает обширным набором удобных библиотек, специально предназначенных для анализа и визуализации данных.

Для более глубокого постижения воздействия факторов на развитие диабета среди молодежи были применены разнообразные методы визуализации. Были построены графики рассеяния, многомерные гистограммы и изучено воздействие факторов. Эти визуализации помогли выявить тенденции и закономерности в представленных данных.

Для построения и оценки моделей машинного обучения применялись методы логической регрессии. Качество моделей было оценено с использованием различных метрик, таких как коэффициент детерминации, среднеквадратичная ошибка и средняя абсолютная ошибка. В итоге эта методология анализа факторов, влияющих на развитие диабета среди молодежи, позволяет выявить ключевые факторы, которые следует учитывать при разработке стратегий лечения и предотвращения развития заболевания. Результаты исследования могут быть использованы для более эффективного диагностирования и лечения.

**Результаты**

**3.1 Разведочный анализ данных**

В ходе реализации данного этапа работы, были созданы различного рода графики, визуализирующие различные моменты дата сета для выявления интересных закономерностей и прочей всевозможной информации.

В начале анализа данных была построена диаграмма рассеяния (Scatter plot) для визуализирования зависимости между двумя переменными.

Когда на диаграмме рассеяния строятся точки, каждая точка представляет конкретное сочетание значений двух переменных. Таким образом, она позволяет визуально оценить, существует ли какая-либо структура или закономерность в распределении данных.

Каждый график показывает связь между двумя переменными с разделением цветом в зависимости от значения переменной Outcome.

Графики поочередно строятся для следующих переменных:

* Pregnancies и Age с группировкой по Outcome. (см. Приложение, Рисунок 3)
* Glucose и Age с группировкой по Outcome. (см. Приложение, Рисунок 3)
* BloodPressure и Age с группировкой по Outcome. (см. Приложение, Рисунок 4)
* SkinThickness и Age с группировкой по Outcome. (см. Приложение, Рисунок 4)
* Insulin и Age с группировкой по Outcome. (см. Приложение, Рисунок 3)
* BMI и Age с группировкой по Outcome. (см. Приложение, Рисунок 2)
* DiabetesPedigreeFunction и Age с группировкой по Outcome. (см. Приложение, Рисунок 1)

Далее была построена Многомерная гистограмма (Multivariate Histogram) для визуализации распределения данных в двух или более измерениях. Эта гистограмма состоит из блоков (бинов), каждый из которых представляет комбинацию значений переменных.

Основная идея многомерной гистограммы заключается в том, чтобы разделить пространство на ячейки и подсчитать количество наблюдений, которые попадают в каждую ячейку. Каждая измеренная переменная представляет собой ось, а блоки образуются в тех местах, где значения переменных соответствуют друг другу.

Графики поочередно строятся для следующих переменных:

* Pregnancies и Age с группировкой по Outcome. (см. Приложение, Рисунок 6)
* Glucose и Age с группировкой по Outcome. (см. Приложение, Рисунок 5)
* BloodPressure и Age с группировкой по Outcome. (см. Приложение, Рисунок 7)
* SkinThickness и Age с группировкой по Outcome. (см. Приложение, Рисунок 7)
* Insulin и Age с группировкой по Outcome. (см. Приложение, Рисунок 8)
* BMI и Age с группировкой по Outcome. (см. Приложение, Рисунок 7)
* DiabetesPedigreeFunction и Age с группировкой по Outcome. (см. Приложение, Рисунок 8)

Следующим шагом была проведена статистика сводки (Summary statistics), которая представляет собой сжатую, обобщенную информацию о наборе данных. Эта информация включает основные числовые характеристики и структуру данных, что позволяет получить представление о его основных свойствах без необходимости анализа всего объема данных. Сводная статистика часто используется для быстрого обзора данных и выявления ключевых моментов. (см. Приложение, Рисунок 9)

Так же была проведена сводка результатов, которая была использована для представления основных выводов и ключевых данных, полученных в анализе данных. Это краткое, обобщенное изложение помогает быстро понять основные точки и сделать выводы без детального рассмотрения всего объема данных или доклада. (см. Приложение, Рисунок 10)

**3.2 Построение модели**

Для построения модели была выбрана логическая регрессия исходя из моих данных. Логистическая регрессия, является статистическим методом, который используется для моделирования вероятности бинарного (двоичного) исхода зависимой переменной от одной или нескольких независимых переменных. Этот метод широко используется в области статистики и машинного обучения для классификации, предсказания вероятности событий и оценки влияния факторов.

Цель построения моделей логистической регрессии для различных пар переменных с группировкой по Outcome заключается в исследовании связи между этими переменными и вероятностью наличия диабета. Рассмотрим несколько пар переменных, таких как Pregnancies и Age, Glucose и Age, BloodPressure и Age, SkinThickness и Age, Insulin и Age, BMI и Age, DiabetesPedigreeFunction и Age, с группировкой по Outcome.

Для каждой пары переменных определены следующие цели:

1. **Pregnancies и Age с группировкой по Outcome :**

Изучить, как количество беременностей (Pregnancies) и возраст (Age) влияют на вероятность развития диабета (Outcome). Модель может предсказывать, как эти переменные связаны с риском диабета у разных групп. (см. Приложение, Рисунок 11)

1. **Glucose и Age с группировкой по Outcome:**

Анализ влияния уровня глюкозы в крови (Glucose) и возраста (Age) на вероятность наличия диабета. Модель может помочь определить, какие значения глюкозы и возраста связаны с повышенным риском диабета. (см. Приложение, Рисунок 12)

1. **BloodPressure и Age с группировкой по Outcome:**

Понять, как уровень артериального давления (BloodPressure) и возраст (Age) влияют на вероятность диабета. Модель может предоставить информацию о том, какие значения давления и возраста связаны с более высоким риском. (см. Приложение, Рисунок 13)

1. **SkinThickness и Age с группировкой по Outcome:**

Исследовать влияние толщины кожи (SkinThickness) и возраста (Age) на вероятность диабета. Модель может выделить значимые факторы, связанные с риском развития заболевания. (см. Приложение, Рисунок 13)

1. **Insulin и Age с группировкой по Outcome:**

Оценить, как уровень инсулина (Insulin) и возраст (Age) соотносятся с вероятностью наличия диабета. Модель может выделить важные аспекты, влияющие на риск. (см. Приложение, Рисунок 13)

1. **BMI и Age с группировкой по Outcome:**

Исследовать взаимосвязь между индексом массы тела (BMI) и возрастом (Age) с учетом наличия диабета. Модель может помочь выявить, какие значения BMI и возраста ассоциированы с более высоким или низким риском. (см. Приложение, Рисунок 14)

1. **DiabetesPedigreeFunction и Age с группировкой по Outcome:**

Анализировать, как генетическая предрасположенность к диабету (DiabetesPedigreeFunction) и возраст (Age) связаны с вероятностью развития диабета. Модель может выявить, какие значения этих переменных важны для предсказания риска. (см. Приложение, Рисунок 14)

Так же был отображен в графики метод анализ влияния (influence analysis) (см. Приложение, Рисунок 15), который представляет собой метод в статистике и анализе данных, который используется для выявления наблюдений, оказывающих значительное влияние на результаты статистического анализа. Этот метод основан на оценке того, насколько изменение или удаление конкретных наблюдений может повлиять на параметры модели или статистические выводы.

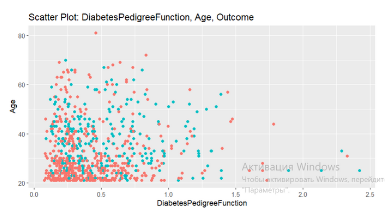
**Вывод**

В ходе исследования диабета у молодежи выявлены важные взаимосвязи и факторы, оказывающие существенное влияние на риск развития этого заболевания. Уровень глюкозы, индекс массы тела (BMI), количество беременностей, генетическая предрасположенность, артериальное давление, возраст, уровень инсулина и толщина кожи существенно влияют на вероятность развития диабета молодых людей.

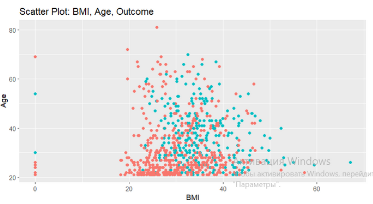
Анализ подчеркивает, что особенно высок риск диабета у молодых людей с повышенным уровнем глюкозы, высоким BMI и генетической предрасположенностью. Эти факторы играют ключевую роль в формировании стратегий профилактики и управления здоровьем в данной возрастной группе.

Обобщенные результаты подчеркивают важность комплексного подхода к оценке риска диабета с учетом разнообразных переменных. Полученные выводы имеют потенциал для создания эффективных программ по предотвращению и управлению диабетом, а также для формирования индивидуализированных стратегий заботы о здоровье молодого поколения.

**Приложение:**

****

**Рисунок 1** Диаграмма рассеяния для переменной DiabetesPedigreeFunction и Age с группировкой по Outcome

****

**Рисунок 2** Диаграмма рассеяния для переменной BMI и Age с группировкой по Outcome

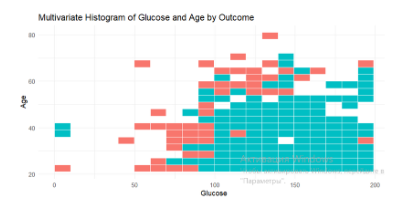
****

**Рисунок 3** Диаграмма рассеяния для переменной Glucose, Pregnancies, Insulin и Age с группировкой по Outcome

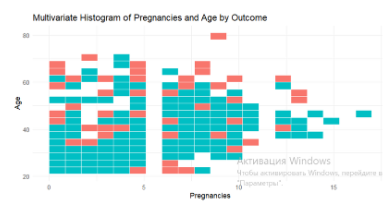
****

**Рисунок 4** Диаграмма рассеяния для переменной BloodPressure, SkinThickness

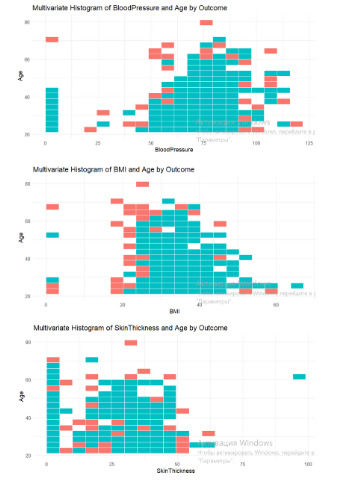
и Age с группировкой по Outcome



**Рисунок 5** Многомерная гистограмма для переменной Glucose и Age с группировкой по Outcome



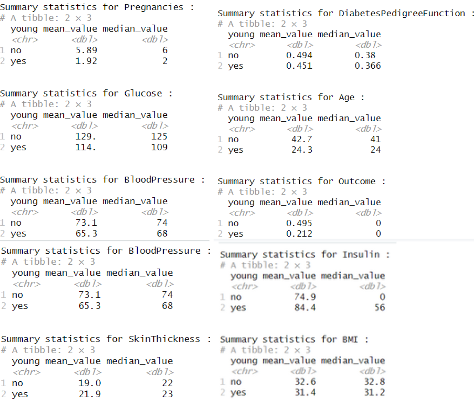
**Рисунок 6** Многомерная гистограмма для переменной Pregnancies и Age с группировкой по Outcome



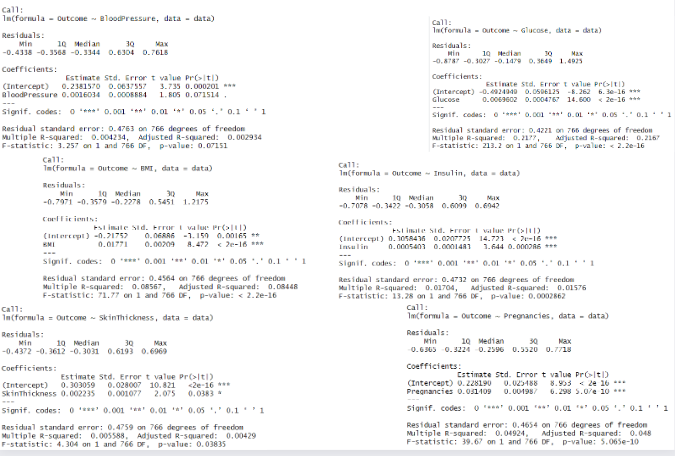
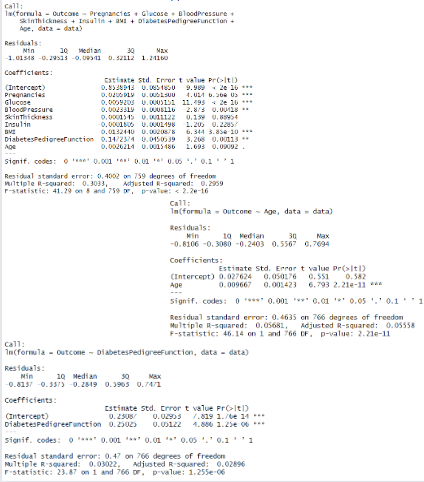
**Рисунок 7** Многомерная гистограмма для переменной BloodPressure, BMI, SkinThickness и Age с группировкой по Outcome



**Рисунок 8** Многомерная гистограмма для переменной DiabetesPedigreeFunction, Insulin и Age с группировкой по Outcome



**Рисунок 9** Cтатистика сводки



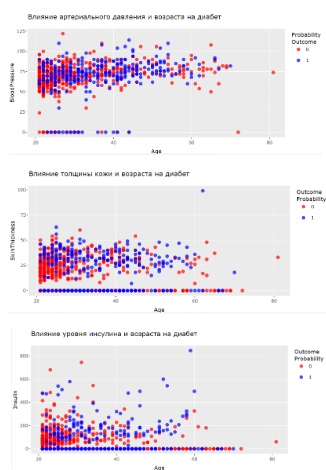
**Рисунок 10** Cводка результатов



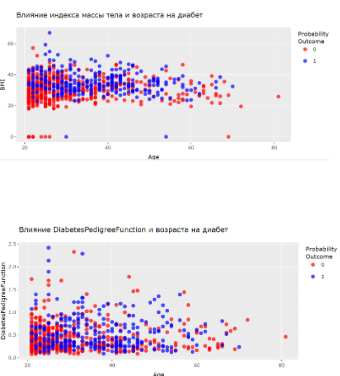
**Рисунок 11** Логистическая модель Pregnancies и Age с группировкой по Outcome



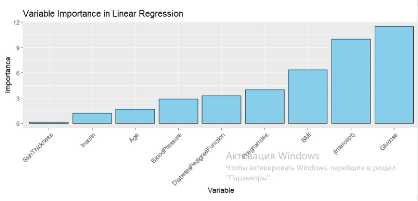
**Рисунок 12** Логистическая модель Glucose и Age с группировкой по Outcome



**Рисунок 13** Логистическая модель BloodPressure, SkinThickness, Insulin и Age с группировкой по Outcome

****

**Рисунок 14** Логистическая модель BMI, DiabetesPedigreeFunction и Age с группировкой по Outcome



**Рисунок 15 Анализ влияния**

**Библиография**

1. Kapatsa, H. (2023, 17 декабря). Разведочный анализ данных (EDA), ч. 1. Взято из <https://mlaccessible.medium.com/eda-в-машинном-обучении-простыми-словами-ч-1-663fd5231196>
2. Акод (2023, 17 декабря). Введение в язык программирования R. Взято из <https://ravesli.com/introduction-r/>