



МИНОБРНАУКИ РОССИИ  
Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«МИРЭА – Российский технологический университет»  
**РТУ МИРЭА**

---

**Институт информационных технологий (ИИТ)**  
**Кафедра прикладной математики (ПМ)**

### **КУРСОВАЯ РАБОТА**

по дисциплине «Технологии организации обработки и хранения статистических данных»

**Тема курсовой работы:** «Анализ признаков и рисков сахарного диабета для раннего прогнозирования»

Студент группы ИМБО-02-22      Ким Кирилл Сергеевич

  
(подпись)

Руководитель  
курсовой работы

старший преподаватель  
Моралева А.А.

  
(подпись)

Работа представлена к защите      «15» декабря 2023 г.

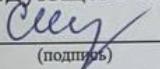
Допущен к защите      «15» декабря 2023 г.

Москва 2023 г.



МИНОБРНАУКИ РОССИИ  
Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«МИРЭА – Российский технологический университет»  
**РТУ МИРЭА**

**Институт информационных технологий (ИИТ)**  
**Кафедра прикладной математики (ПМ)**

Утверждаю  
И.о. заведующего кафедрой ПМ  
  
(подпись) Смоленцева Т.Е.

«21» сентября 2023 г.

**ЗАДАНИЕ**

**на выполнение курсовой работы**

по дисциплине «Технологии организации обработки и хранения статистических данных»

Студент Ким Кирилл Сергеевич

Группа ИМБО-02-22

**Тема** «Анализ признаков и рисков сахарного диабета для раннего прогнозирования»

**Исходные данные:** выбранные студентом данные.

**Перечень вопросов, подлежащих разработке, и обязательного графического материала:**

Характеристика деятельности организации и общее описание признаков, влияющих на развитие сахарного диабета (описание признаков, типов данных, общее текстовое описание сахарного диабета и графическое влияние признаков на риск)

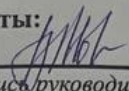
Графические модели зависимости влияния признаков на риск сахарного диабета (описание путем представления графиков по признакам)

Анализ риска возникновения сахарного диабета (выявление зависимостей между признаками)

Раннее прогнозирование сахарного диабета.

**Срок представления к защите курсовой работы:**

**Задание на курсовую работу выдал**

  
Подпись руководителя

до «15» декабря 2023 г.  
Моралева А.А.

(ФИО руководителя)

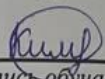
«21» сентября 2023 г.

Ким К.С.

(ФИО обучающегося)

«21» сентября 2023 г.

**Задание на курсовую работу получил**

  
Подпись обучающегося

**ОТЗЫВ**  
на курсовую работу  
по дисциплине «Технологии организации обработки и хранения  
статистических данных»

Студент


Ким Кирилл Сергеевич

ИМБО-02-22

Характеристика курсовой работы

Критерий	Да	Нет	Не полностью
Соответствие содержания курсовой работы указанной теме	✓		
Соответствие курсовой работы заданию	✓		
3. Соответствие рекомендациям по оформлению текста, таблиц, рисунков и пр.			✓
Полнота выполнения всех пунктов задания			✓
Логичность и системность содержания курсовой работы			✓
Отсутствие фактических грубых ошибок	✓		

Рекомендуемая оценка: удовлетворительно, хорошо, отлично  
(нужное подчеркнуть)

  
Подпись руководителя

Моралева А.А.  
(ФИО руководителя)

«25» декабря 2023 г.

## СОДЕРЖАНИЕ

ВВЕДЕНИЕ .....	5
1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ .....	6
1.1 Сахарный диабет .....	6
1.2 Корреляционно-регрессионный анализ .....	13
2 ПРАКТИЧЕСКАЯ ЧАСТЬ .....	18
2.1 Корреляционно-регрессионный анализ признаков и рисков сахарного диабета для раннего прогнозирования.....	18
ЗАКЛЮЧЕНИЕ .....	32
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ .....	33
ТЕОРЕТИЧЕСКАЯ ЧАСТЬ .....	33
ПРАКТИЧЕСКАЯ ЧАСТЬ .....	34
ПРИЛОЖЕНИЯ.....	35
Приложение А .....	36

# ВВЕДЕНИЕ

Сахарный диабет является одним из наиболее распространенных хронических заболеваний во всем мире. Количество людей, страдающих сахарным диабетом, увеличивается с каждым годом, что делает проблему еще более актуальной.

Он имеет серьезные последствия для здоровья пациентов и требует постоянного контроля уровня сахара в крови и правильного лечения.

Раннее прогнозирование появления сахарного диабета может помочь предпринять меры по предотвращению развития заболевания или управлению им.

Таким образом, тема данной курсовой работы является актуальной для медицины и человека в целом.

Целью данной курсовой работы является — провести анализ и построить модель прогнозирования риска сахарного диабета.

Задачи, решаемые в данной курсовой работе:

1. Изучение научной и методической литературы о сахарном диабете.
2. Поиск/сбор данных о сахарном диабете.
3. При необходимости подготовить данные, провести очистку и предобработку.
4. Проанализировать исходные данные, построив графики зависимостей.
5. Использование знания математической статистики и современных средств обработки данных.
6. Обучение оформлению официальных документов.

# 1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

## 1.1 Сахарный диабет

Сахарный диабет — это хроническое заболевание эндокринной системы человека, характеризующиеся длительным повышением концентрации глюкозы в крови и сопутствующими изменениями процессов обмена веществ. Представляет серьёзную угрозу для здоровья и жизни больного, так как провоцирует развитие тяжелых сопутствующих заболеваний. В основе сахарного диабета лежит инсулиновая недостаточность, что приводит к увеличению сахара в крови и появлению его в моче.

Это происходит из-за того, что поджелудочная железа вырабатывает недостаточное количество гормона — инсулина, который регулирует углеводный обмен в организме. При наследственной предрасположенности к заболеванию его возникновение могут спровоцировать переедание, злоупотребление сладким, нервное перенапряжение, вирусная инфекция. Диабет может развиваться после краснухи, кори, гриппа и других вирусных заболеваниях. Самостоятельно выявить заболевание невозможно, но некоторые отклонения в организме могут быть первыми тревожными сигналами.

К таковым признакам можно отнести: частое и обильное мочеиспускание, ощущение постоянной жажды, кожный зуд, постоянные скачки весовых соотношений (масса тела то резко увеличивается, то моментально снижается), проявление кожных высыпаний.

На начальной стадии болезни сахарный диабет практически незаметен, все симптомы усиливаются в тот момент, когда недуг активизировался. Диагностировать болезнь при первичном осмотре и беседе с пациентом невозможно, для того чтобы поставить точный диагноз нужно обязательно

провести лабораторные исследования. К сожалению, несвоевременное лечение сахарного диабета приводит к серьезным осложнениям для больного. Различают 2 типа сахарного диабета:

Сахарный диабет 1 типа (инсулин зависимый диабет) — это деструкция клеток поджелудочной железы, которая приводит к полной инсулиновой недостаточности. При 1 типе диабета практически все клетки поджелудочной железы, которые выделяют инсулин, разрушаются вследствие чего железа не в состоянии продуцировать инсулин. Инфекция сама по себе не разрушает клетки поджелудочной железы, она включает иммунную систему, клетки которой и уничтожают клетки поджелудочной. Данный тип составляет 10 % от общего числа больных сахарным диабетом. Болеют дети и взрослые до 30 лет. При диабете 1 типа пациент вынужден постоянно вводить инсулин, который нужен для нормального передвижения глюкозы в организме.

Основной метод лечения — инъекции инсулина. Поэтому и называется 1 тип диабета — инсулин зависимый диабет. Этот тип диабета активно прогрессирует, быстро развиваются осложнения и стадия декомпенсации.

Сахарный диабет 2 типа (инсулин независимый) — это нарушение углеводного обмена с развитием гипергликемии. При данном типе поджелудочная железа не разрушается и продолжает вырабатывать инсулин, но в организме развивается резистентность (снижение чувствительности) клеток к инсулину. В результате этого в клетки не поступает нужного количества глюкозы, даже в присутствии инсулина. При наличии данного типа диабета у кого-то в роду, вероятность развития у потомка на протяжении жизни составляет 40 %. Также, часто его развитию способствуют ожирение, перенапряжение и стрессовые ситуации. Его коррекция может быть достигнута диетой, снижением массы тела и сахароснижающими таблетками.

При диабете 2 типа не нужно вводить инсулин, так как проблема не в выработке инсулина, а именно в усвоении глюкозы тканями. Но, по мере

прогрессирования диабета, выделение инсулина клетками поджелудочной железы снижается и тогда приходится назначать инсулин. [1.1]

Сахарный диабет является сложным заболеванием, которое трудно поддается лечению. При его развитии в организме происходит нарушение углеводного обмена и снижение синтеза инсулина поджелудочной железой, в результате чего глюкоза перестает усваиваться клетками и оседает в крови в виде микрокристаллических элементов. Точные причины, по которым начинает развиваться данный недуг, ученым установить до сих пор не удалось. Но благодаря им были выявлены факторы риска, которые могут спровоцировать возникновение этого заболевания у людей разностной возрастной категории.

Прежде чем рассматривать факторы риска развития сахарного диабета, необходимо сказать, что это заболевание, как уже было сказано выше, имеет два типа, и каждый из них имеет свои особенности. Диабет 1 типа характеризуется системными изменениями в организме, при которых нарушается не только углеводный обмен, но и функциональность поджелудочной железы. По каким-то причинам ее клетки перестают вырабатывать инсулин в нужном количестве, в результате чего сахар, проникающий в организм вместе с пищей, не подвергается процессам расщепления и, соответственно, не может усваиваться клетками.

Самыми распространенными осложнениями этой болезни являются следующие состояния:

- гипергликемия — повышение уровня сахара в крови за пределы нормы (свыше 7 ммоль/л);
- гипогликемия — снижение уровня глюкозы в крови за пределы нормы (ниже 3,3 ммоль/л);
- гипергликемическая кома — повышение уровня сахара в крови свыше 30 ммоль/л;
- гипогликемическая кома — снижение уровня глюкозы в крови ниже 2,1 ммоль/л;



- диабетическая стопа — снижение чувствительности нижних конечностей и их деформация;
- диабетическая ретинопатия — снижение остроты зрения;
- тромбофлебит — образование в стенках сосудов бляшек;
- гипертония — повышение артериального давления;
- гангрена — некроз тканей нижних конечностей с последующим развитием абсцесса;
- инсульт и инфаркт миокарда.

Основными факторами его развития являются:

- наследственная предрасположенность;
- вирусные заболевания;
- интоксикация организма;
- неправильное питание;
- частые стрессы.

В возникновении СД1 главную роль играет наследственная предрасположенность. Если кто-то из членов семьи страдает от этого недуга, то риски его развития у последующего поколения составляют примерно 10-20 %.

При этом следует отметить, что в данном случае речь идет не об установленном факте, а о предрасположенности. То есть если мать или отец болеют СД1, это вовсе не значит, что у их детей также будет диагностировано это заболевание. Предрасположенность говорит о том, что если человек не будет проводить профилактические мероприятия и будет вести неправильный образ жизни, то у него есть большая вероятность стать диабетиком в течение нескольких лет.

Однако и в этом случае необходимо учитывать, что если от диабета страдают сразу оба родителя, то вероятность возникновения его у их ребенка значительно повышается. И нередко именно в таких ситуациях это заболевание диагностируется у детей еще в школьном возрасте, хотя они еще не имеют вредных привычек и ведут активный образ жизни.

Вирусные заболевания — еще одна причина, по которой может развиваться СД1. Особенно опасными в этом случае являются такие болезни, как паротит и краснуха. Учеными уже давно было доказано, что эти заболевания негативно сказываются на работе поджелудочной железы и приводят к повреждению ее клеток, снижая, таким образом, уровень инсулина в крови.

Следует отметить, что в зоне риска не только уже рожденные дети, но и те, которые еще находятся в утробе матери. Любые вирусные заболевания, которые переносит беременная женщина, могут спровоцировать развитие у ее ребенка сахарного диабета 1 типа.

Многие люди, работающие на заводах и предприятиях, где используются химические вещества, также подвержены большому риску. Действие химикатов негативно сказывается на работе всего организма, в том числе и на функциональности поджелудочной железы.

Химиотерапии, которые проводятся для лечения различных онкологических заболеваний, также оказывают токсичное действие на клетки организма, поэтому их проведение тоже в несколько раз увеличивает вероятность развития СД1 у человека.

Неправильное питание является одним из самых распространенных причин развития сахарного диабета 1 типа. Ежедневный рацион современного человека содержит в себе огромное количество жиров и углеводов, что оказывает сильную нагрузку на пищеварительную систему, в том числе и на поджелудочную. Со временем ее клетки повреждаются, и синтез инсулина нарушается. Необходимо также отметить, что из-за неправильного питания СД1 может развиваться и у детей в возрасте 1-2 лет.

Симптомы обоих типов сахарного диабета проявляются тогда, когда уровень глюкозы в крови превышает 6,0 ммоль/л и тогда, когда глюкоза попадает в мочу, где ее в норме быть не должно.

Для сахарного диабета 1 типа характерно быстрое развитие симптомов (в течение нескольких недель, месяцев).

Сахарный диабет 2 типа может длительное время себя не проявлять. Его можно выявить случайно, при очередном медицинском осмотре или сдаче анализов по поводу другого заболевания или при медицинском осмотре. Симптомы могут развиваться годами, быть мало выраженными.

Основные симптомы являются:

1. Полиурия — повышенное мочеотделение. По мере увеличения уровня глюкозы в крови, повышается ее содержание и в моче. Почки реагируют первыми и начинают выделять больше жидкости из организма, для того чтобы разбавить концентрацию глюкозы в моче. Характерно усиление мочеотделения в ночное время. В результате этого происходит усиленное мочевыделение (до 2 литров мочи в сутки).
2. Полидипсия — это неутолимая жажда, сухость во рту, что является следствием выделения большого количества жидкости из организма в виде мочи. Больные начинают много пить, чтобы утолить жажду и восполнить потерю жидкости с мочой.
3. Полифагия — это постоянное чувство голода. Это связано с нарушением обмена веществ, точнее сказать из-за того, что клетки не способны поглощать и перерабатывать глюкозу без помощи инсулина. Резкое снижение веса, похудение особенно характерны для больных с диабетом 1 типа. Это связано с повышенным разрушением белков и жиров из-за отсутствия глюкозы в энергетическом обмене клеток. Парадоксально то, что похудение развивается, несмотря на повышенный аппетит больного.

Симптомы, которые могут сопровождать диабет, но не являются обязательными:

- сухость кожных покровов;
- сухость во рту;
- мышечная слабость;
- гнойничковые образования и заболевания кожи;

- длительное и плохое заживление ран и рубцов;
- быстрая утомляемость;
- частые головные боли;
- зуд половых органов (чаще после мочеиспускания).

Выделяются три степени тяжести заболевания:

- I. степень (легкая) — при этом повышение глюкозы в крови не превышает 8 ммоль/л натощак. При этом нет значительных колебаний глюкозы в течение суток, допустимы следы глюкозы в моче до 20 г/л. Возможны начальные проявления осложнений (ангионейропатии).
- II. степень (средняя) — уровень глюкозы в крови натощак достигает 14 ммоль/л. Глюкоза в моче увеличивается до 40 г/л. Компенсация состояния достигается диетой и приемом сахароснижающих лекарственных препаратов. Возможны проявления осложнений.
- III. степень (тяжелое течение) — уровень глюкозы натощак более 14 ммоль/л, в моче уровень глюкозы превышает 50 г/г. В этой стадии больные нуждаются в постоянной инсулинотерапии и ярко выражены сопутствующие осложнения.

Рассмотрим разновидности фаз компенсации сахарного диабета:

1. Фаза компенсации, при которой больной чувствует себя хорошо, а с помощью терапии легко можно добиться нормальных цифр глюкозы в крови. В моче глюкозы не содержится.
2. Фаза субкомпенсации. При данной фазе не удастся снизить уровень глюкозы в крови ниже 13,9 ммоль/л. Появляется сахар в моче. Ацетона в моче не содержится.
3. Фаза декомпенсации (самая тяжелая) — проводимая терапия не дает эффекта, и уровень сахара поднимается выше 14,0 ммоль/л. Количество глюкозы в моче увеличивается и появляется ацетон. Возможно развитие гипергликемической комы.

В данной работе рассмотрены такие понятия, как типы сахарного диабета, основные симптомы, тяжести заболевания, основные факторы, разновидности фаз компенсации сахарного диабета. Таким образом, раннее выявление признаков и рисков развития сахарного диабета имеет важное значение для предотвращения его возникновения для медицины. Он разнообразен и может быть подвергнут корреляционно-регрессионному анализу.

## **1.2 Корреляционно-регрессионный анализ**

Корреляционно-регрессионный анализ является наиболее широко распространенным и гибким приемом обработки статистической информации.

Корреляционно-регрессионный анализ — это один из самых распространенных методов изучения отношений между численными величинами. Его основная цель состоит в нахождении зависимости между двумя параметрами и ее степени с последующим выводением уравнения. То есть, корреляционно-регрессионный анализ представляет из себя объединение методов корреляционного и регрессионного анализов. [1.2]

Задачами корреляционно-регрессионного анализа являются:

- установление типа уравнения регрессии;
- определение параметров уравнения регрессии и оценка значимости параметров;
- оценка тесноты и направления связи между переменными; — оценка значимости уравнения регрессии;
- определение прогнозных значений зависимой переменной и оценка полученного прогноза.

Так как в корреляционно-регрессионном анализе используются методы корреляционного и регрессионного анализа, рассмотрим эти методы подробнее.

Корреляционный анализ — раздел математической статистики, в котором изучаются задачи выявления статистических зависимостей между случайными величинами путем оценок различных коэффициентов корреляции. Методы корреляционного анализа дают хорошие результаты тогда, когда данные эксперимента можно считать выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

Невозможно управлять явлениями, предсказывать их развитие без изучения характера, силы и других особенностей связей. Поэтому методы исследования, изменения связей составляют чрезвычайно важную часть методологии научного исследования, в том числе и статистическую.

Связи между изучаемыми переменными подразделяются на функциональные и статистические. При функциональной связи определенному значению одной переменной величины соответствует строго определенное значение другой переменной.

При изменении одной из них на определенную величину, другая переменная изменяется на величину, в соответствии с видом функции, связывающей переменные.

Статистической называется связь между переменными или признаками, когда определенному значению факторного признака соответствует несколько различных значений результативного признака. Частным случаем статистической связи является корреляционная, которая проявляется в среднем, в массе наблюдений, как статистическая закономерность.

При корреляционной связи с изменением факторного признака на определенную величину изменяется среднее значение результативного признака. Обычно корреляционная зависимость представляется как

функциональная зависимость между переменными в виде уравнения регрессии.

Корреляционной связью называют важнейший частный случай статистической связи, состоящий в том, что разным значениям одной переменной соответствуют различные средние значения другой. С изменением значения признака  $x$  закономерным образом изменяется среднее значение признака  $y$ ; в то время как в каждом отдельном случае значение признака  $y$  (с различными вероятностями) может принимать множество различных значений. [1.3]

Тесноту связи изучаемых явлений оценивает Коэффициент Пирсона ( $K_n$ ).

Коэффициент Пирсона используется для изучения связи между двумя качественными признаками, каждый из которых состоит более чем из двух групп. Вычисляют по формуле:

$$K_n = \sqrt{\frac{\varphi^2}{1+\varphi^2}},$$

где  $\varphi^2$  — показатель взаимной сопряженности:

$$\varphi^2 = \sum \frac{n_{xy}^2}{n_x * n_y},$$

где  $n_x$  — объемы признака  $X$  по группам;

$n_y$  — объемы признака  $Y$  по группам;

$n_{xy}$  — объемы выборок, относящихся к  $X$  и  $Y$  одновременно.

Корреляционный коэффициент Пирсона может принимать значения в диапазоне —  $1 < K_n < 1$ .

По значению эмпирического корреляционного отношения судят о тесноте связи между признаками. Обычно придерживаются следующей шкалы:

$0,1 < K_n \leq 0,3$  — связь слабая;

$0,3 < K_n \leq 0,5$  — связь заметная;

$0,5 < K_n \leq 0,7$  — связь умеренно тесная;

$0,7 < K_n \leq 0,9$  — связь тесная;

$K_n > 0,9$  — связь очень тесная.

После того как с помощью корреляционного анализа выявлено наличие статистических связей между переменными и оценена степень их тесноты, обычно переходят к математическому описанию зависимостей, то есть к регрессионному анализу.

Регрессионный анализ применяется в тех случаях, когда необходимо отыскать непосредственно вид зависимости  $x$  и  $y$ . При этом предполагается, что независимые факторы являются не случайными величинами, а результативный показатель  $y$  имеет постоянную, независимую от факторов дисперсию и стандартное отклонение.

Рассмотрим метод логистической регрессии.

Под линейностью имеется в виду, что переменная  $y$  предположительно находится под влиянием переменной  $x$  в зависимости

$$y = \frac{\exp(z)}{1 + \exp(z)}; z = b_0 + b_1 x.$$

где  $b_0$  — постоянная величина (или свободный член уравнения); [1.8]

$b_1$  — коэффициент регрессии, определяющий наклон линии, вдоль которой рассеяны данные наблюдения. Это показатель, характеризующий изменение переменной  $y_i$  при изменении значения  $x_i$  на единицу. Если  $b_1 > 0$ , переменные  $x_i$  и  $y_i$  положительно коррелированные, если  $b_1 < 0$  — отрицательно коррелированные;



Построение логистической регрессии сводится к оценке параметров  $b_1$  и  $b_0$ . Классический подход к оцениванию параметров логистической регрессии основан на методе максимального правдоподобия (ММП). ММП позволяет получить такие оценки параметров  $b_1$  и  $b_0$ , при которых оценка  $\hat{\theta}$  достигает максимума.

$$\hat{\theta} = \max_{i=1} L(x_1, x_2, \dots, x_n; \theta),$$

где  $x_1, x_2, \dots, x_n$  — фиксированные числа.

Используйте различные метрики, такие как средняя квадратичная ошибка (RMSE) или псевдо- $R^2$  Макфаддена.

Проанализируем коэффициенты регрессии, чтобы определить, какие признаки оказывают наибольшее влияние на развитие сахарного диабета. Оценим и объясним риски и связи между признаками и диабетом.

Предскажем риски: используя обученную модель для предсказания риска развития сахарного диабета у новых пациентов на основе их признаков. Это поможет в раннем прогнозировании и определении групп риска.

Рассмотрим понятие корреляционно-регрессионного анализа, а также методы корреляционно-регрессионного анализа: нахождение корреляционной связи с помощью коэффициента Пирсона; построение парной логистической регрессии с помощью метода максимального правдоподобия (ММП).

Таким образом, корреляционно-регрессионный анализ позволяет нам лучше понять факторы, которые влияют на заболевание сахарным диабетом. Полученные результаты могут быть использованы для анализа признаков прогнозирования.

Рассмотрим применение корреляционно-регрессионного анализа на примере анализа риска возникновения сахарного диабета.

## 2 ПРАКТИЧЕСКАЯ ЧАСТЬ

### 2.1 Корреляционно-регрессионный анализ признаков и рисков сахарного диабета для раннего прогнозирования.

В практической работе проведём корреляционно-регрессионный анализ Набор данных «Для прогнозирования риска сахарного диабета включает в себя следующие столбцы данных:

1. `patient_number` — код (номер) пациента;
2. `cholesterol` — уровень холестерина;
3. `glucose` — уровень глюкозы и др.

Ниже приведен нормальный диапазон содержания различных растворенных веществ в крови:

1. Уровень холестерина — менее 200 ммоль/л;
2. Глюкоза 80-120 ммоль/л;
3. Уровень холестерина ЛПВП (Липопротеиды высокой плотности) — более 60 ммоль/л;
4. Систолическое АД — 120 мм рт. ст.;
5. Диастолическое АД — 80 мм рт. ст.

Максимальный уровень холестерина, обнаруженный в данных, составляет 443 ммоль/л, глюкозы 385 ммоль/л, холестерина ЛПВП 120 ммоль/л, систолического АД 250 мм рт. ст., диастолического 124 мм рт. ст.

Минимальное значение холестерина ЛПВП — 12 ммоль/л.

Вес 75 % пациентов составляет от 99 до 200 килограммов.

На рисунке 2.1 представлена статистика.

№	Метка	Вид	Гистограмма	Диаграмма размаха	Минимум	Максимум	Среднее	Стандарт...	Пропуски	Уникаль...
1	12 patient_number	○			1	390	196	112,72754...	0	
2	12 cholesterol	○			78	443	207	44,666005...	0	
3	12 glucose	○			48	385	107	53,798188...	0	
4	12 hdl_chol	○			12	120	50	17,279069...	0	
5	90 chol_hdl_ratio	○			1,5	19,3	4,5246153...	1,7366336...	0	
6	12 age	○			19	92	47	16,435911...	0	
7	ab gender	☀		Недоступно	4	6	5,1692307...	0,9868424...	0	2
8	12 height	○			52	76	66	3,9188668...	0	
9	12 weight	○			99	325	177	40,407823...	0	
10	90 bmi	○			15,2	55,8	28,775641...	6,6009149...	0	
11	12 systolic_bp	○			90	250	137	22,859527...	0	

Рисунок 2.1 — Статистика

Для обработки данных и проведения корреляционно-регрессионного анализа используем аналитическую low-code платформу Loginom.

Импортируем наши наборы данных в сценарий пакета. Для этого в рабочую область сценария добавляем два узла «Excel файл» и для каждого в настройках указываем путь к нужному набору данных.

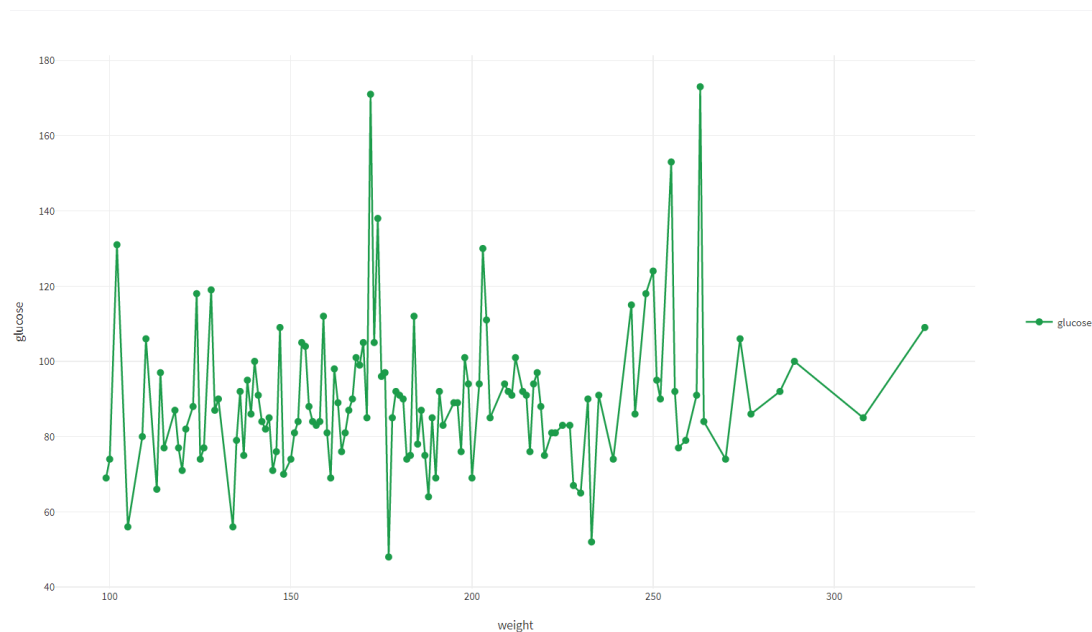
На Рисунке 2.2 представлено исходные данные

#	12 patie...	12 cholesterol	12 glucose	12 hdl_chol	90 ch...	12 age	ab gender	12 hei...	12 weight	90 bmi	12 systolic_bp	12 diastolic_...	12 waist	12 hip	90 waist_hip_ra...	ab diabetes
1	1	193	77	49	3,90	19	female	61	119	22,50	118	70	32	38	0,84	No diabetes
2	2	146	79	41	3,60	19	female	60	135	26,40	108	58	33	40	0,83	No diabetes
3	3	217	75	54	4,00	20	female	67	187	29,30	110	72	40	45	0,89	No diabetes
4	4	226	97	70	3,20	20	female	64	114	19,60	122	64	31	39	0,79	No diabetes
5	5	164	91	67	2,40	20	female	70	141	20,20	122	86	32	39	0,82	No diabetes
6	6	170	69	64	2,70	20	female	64	161	27,60	108	70	37	40	0,93	No diabetes
7	7	149	77	49	3,00	20	female	62	115	21,00	105	82	31	37	0,84	No diabetes
8	8	164	71	63	2,60	20	male	72	145	19,70	108	78	29	36	0,81	No diabetes
9	9	230	112	64	3,60	20	male	67	159	24,90	100	90	31	39	0,79	No diabetes

Рисунок 2.2 — Фрагмент табличного представления исходных данных

Для проведения корреляционно-регрессионного анализа какой-либо отрасли требуется отфильтровать набор данных, выделив людей, у кого нет сахарного диабета и есть сахарный диабет. Нам требуется выделить столбец «diabetes». Воспользуемся узлом «Фильтр строк».

Визуализация результата в виде таблицы представлена на Рисунке 2.3.



**Рисунок 2.3 — Визуализация в виде таблицы набора данных с отфильтрованным набором данных, кого нет сахарного диабета**

На Рисунке 2.4 показан отфильтрованный набор данных, выделив людей, у кого есть сахарный диабет.

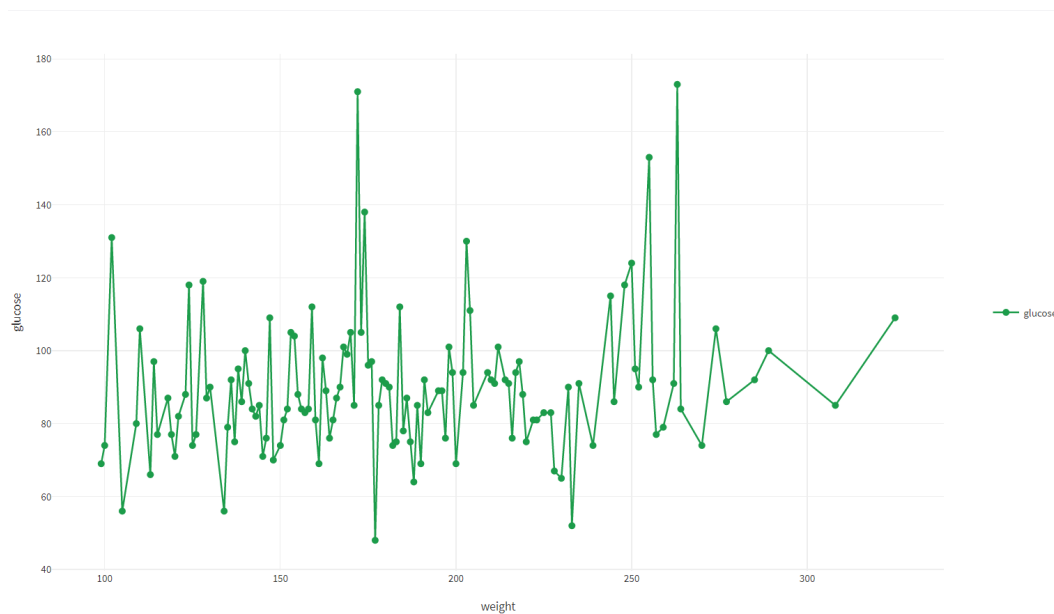
#	ab diabetes	12 pat...	12 cholesterol	12 glucose	12 hdl_chol	90 chol_hdl_ra...	12 age	ab gender	12 hei...	12 weight	90 bmi	12 systolic_bp	12 diastolic...	12 waist	12 hip	90 wais...
1	Diabetes	41	220	60	66	3,30	26	male	70	150	21,50	136	88	33	39	0,85
2	Diabetes	65	194	269	38	5,10	29	female	69	167	24,70	120	70	33	40	0,83
3	Diabetes	83	191	155	58	3,30	31	female	62	237	43,30	140	87	53	56	0,95
4	Diabetes	118	245	119	26	9,40	36	male	66	179	28,90	150	92	37	42	0,88
5	Diabetes	140	203	299	43	4,70	38	female	69	288	42,50	136	83	48	55	0,87

**Рисунок 2.4 — Визуализация в виде таблицы набора данных с отфильтрованным набором данных, кого есть сахарного диабета**

Общепризнано, что повышенный вес и пожилой возраст являются двумя основными факторами, вызывающими диабет.

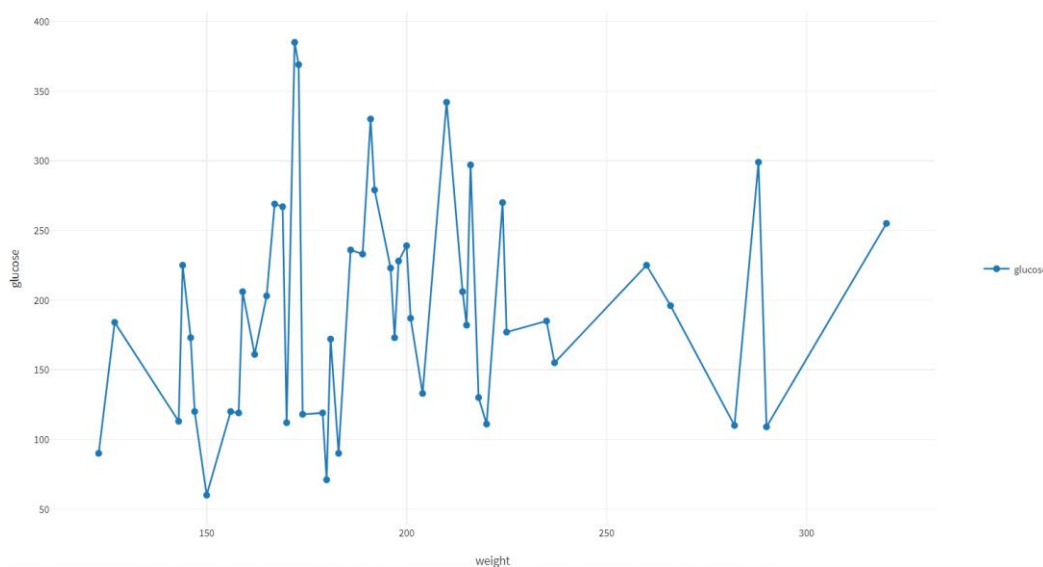
Пациенты с сахарным диабетом имеют более высокий уровень глюкозы и больший вес по сравнению с недиабетическими пациентами.

На Рисунке 2.5 показан график зависимости "glucose" от "weight", у которого нет сахарного диабета.



**Рисунок 2.5 — График зависимости "glucose" от "weight"**

На Рисунке 2.6 показан график зависимости "glucose" от "weight", у которого есть сахарный диабет.

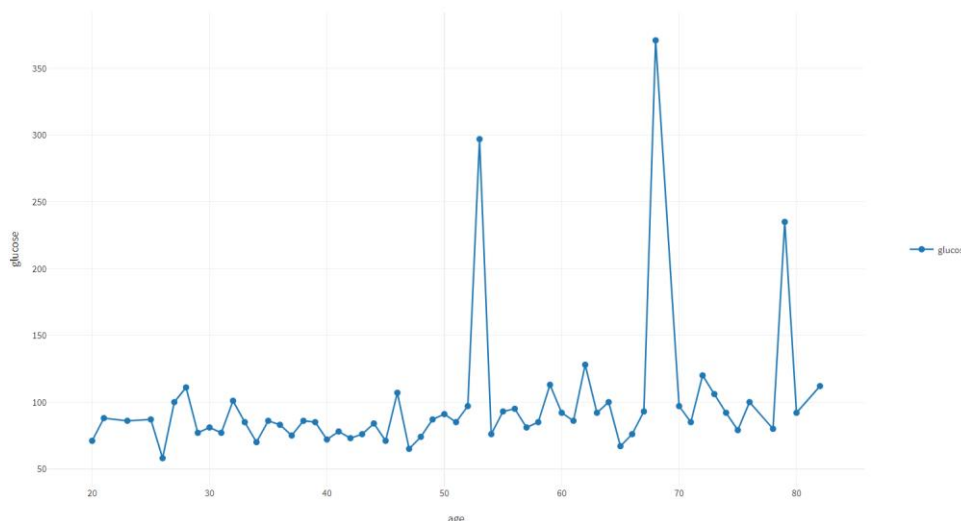


**Рисунок 2.6 — График зависимости "glucose" от "weight"**

Требуется отфильтровать набор данных, выделив отдельно мужчин и отдельно женщин.

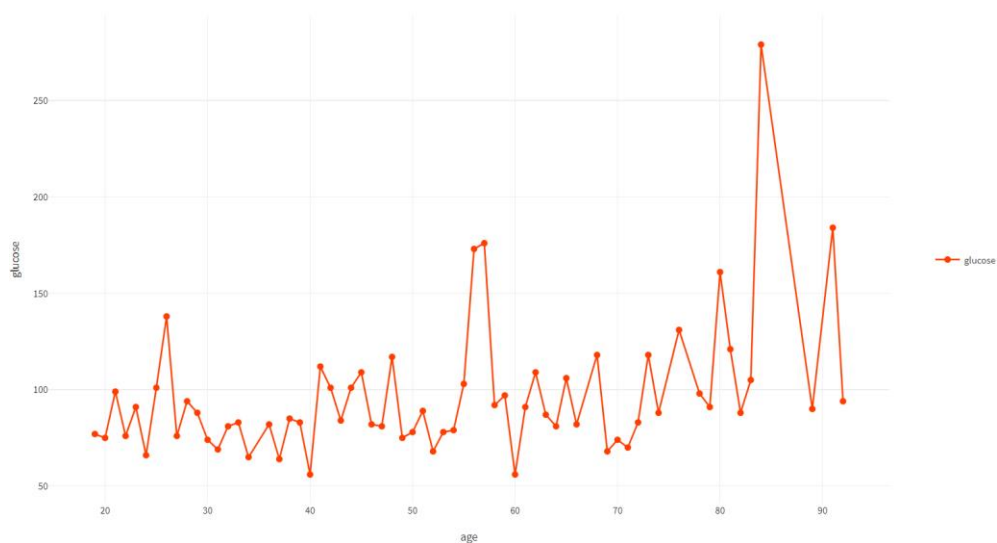
Возраст напрямую не определяется, но более высокий уровень глюкозы у пожилых людей может быть причиной развития у них диабета, также у мужчин в возрасте от 40 до 80 лет уровень глюкозы в крови выше, чем у женщин.

На Рисунке 2.7 показан график зависимости "glucose" от "age", глюкозы от возраста мужчин.



**Рисунок 2.7 — График зависимости "glucose" от "age"**

На Рисунке 2.8 показан график зависимости "glucose" от "age", глюкозы от возраста женщин.



**Рисунок 2.8 — График зависимости "glucose" от "age"**

Артериальное давление напрямую не связано с сахарным диабетом, поскольку пациенты с самым высоким артериальным давлением считаются недиабетическими.

На Рисунке 2.9 показан график зависимости "glucose" от "systolic\_hp", у которых нет сахарного диабета.

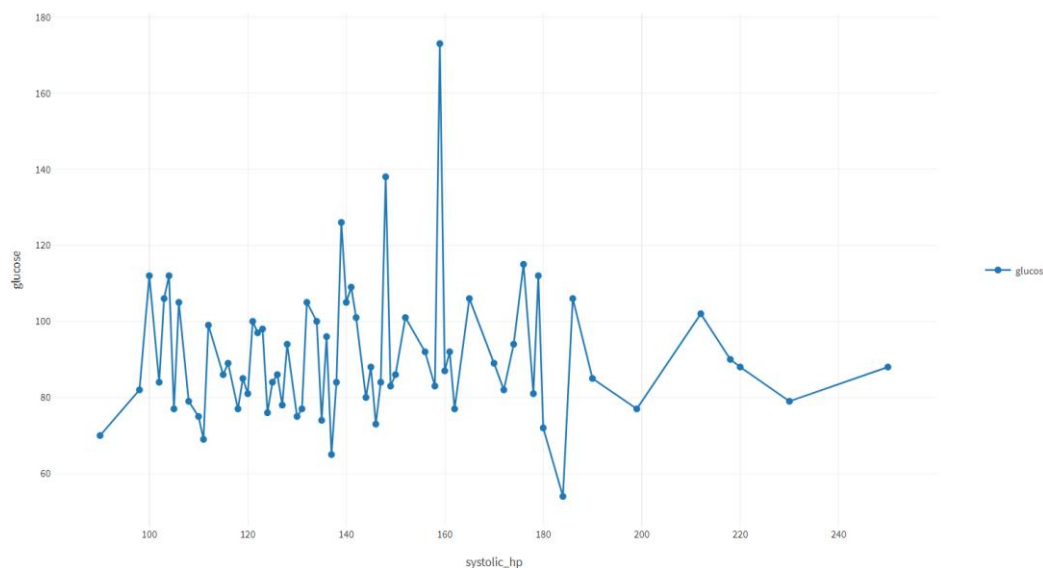


Рисунок 2.9 — График зависимости "glucose" от "systolic\_hp"

На Рисунке 2.10 показан график зависимости "glucose" от "systolic\_hp", у которых есть сахарный диабет.

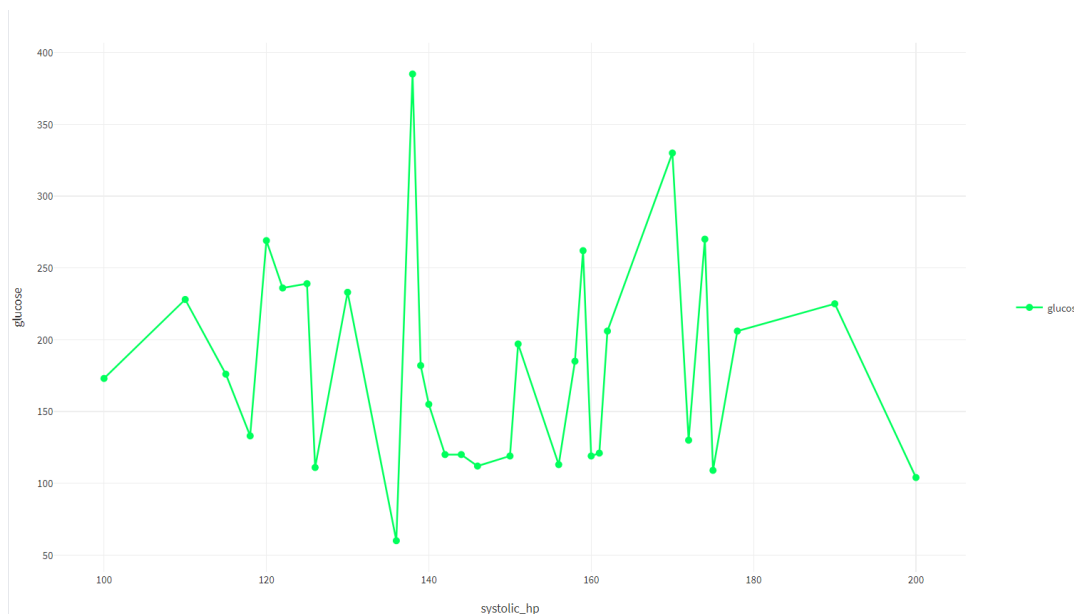
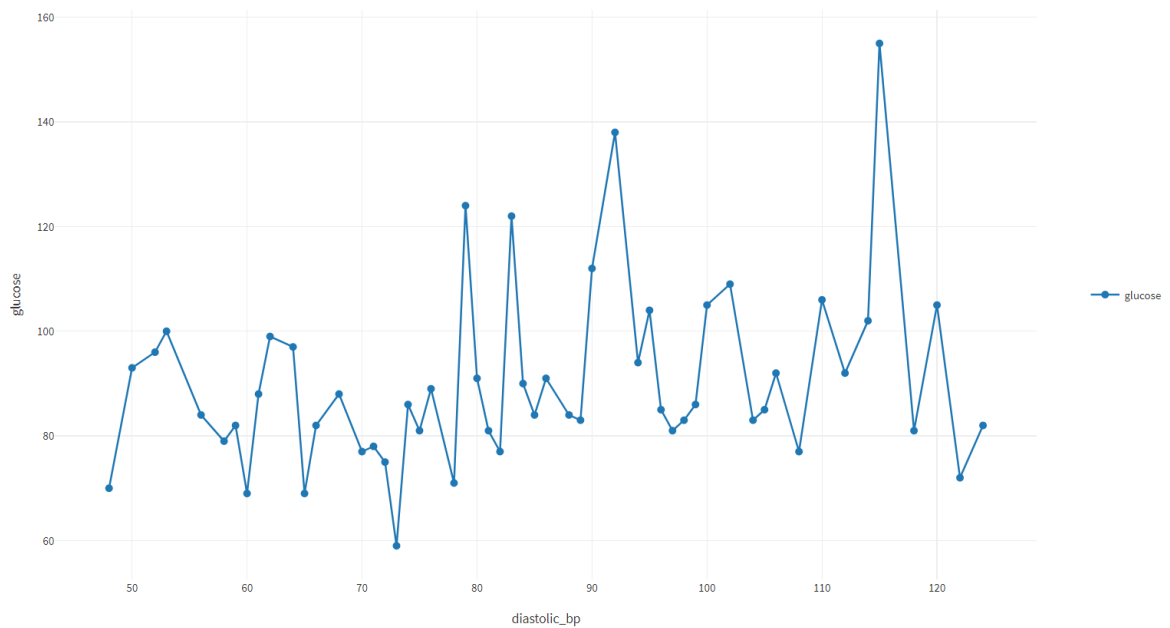


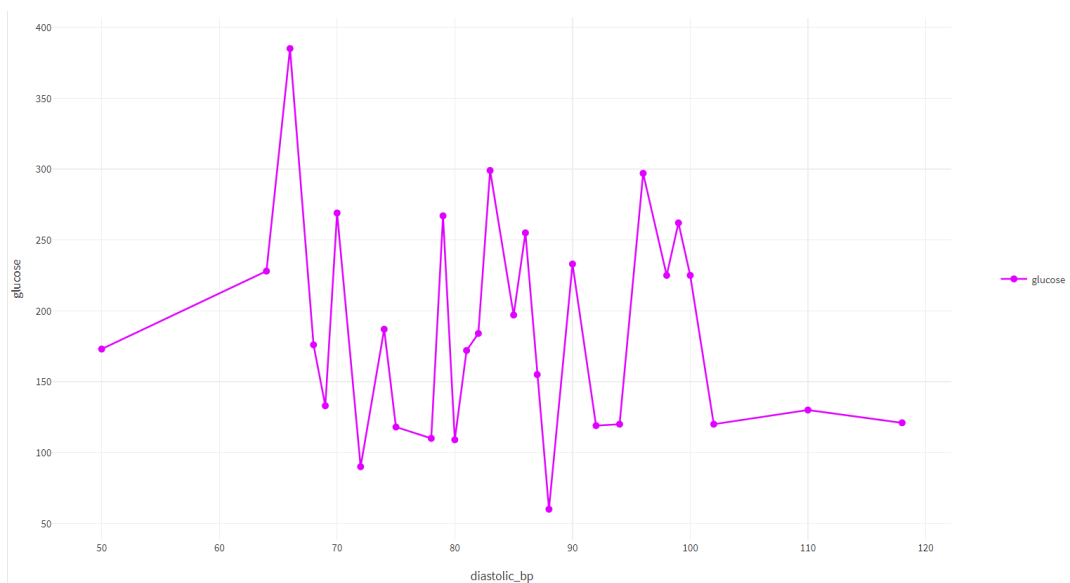
Рисунок 2.10 — График зависимости "glucose" от "systolic\_hp"

На Рисунке 2.11 показан график зависимости "glucose" от "diastolic\_bp", у которых нет сахарного диабета.



**Рисунок 2.11 — График зависимости "glucose" от "diastolic\_bp"**

На Рисунке 2.12 показан график зависимости "glucose" от "diastolic\_bp", у которых есть сахарный диабет.

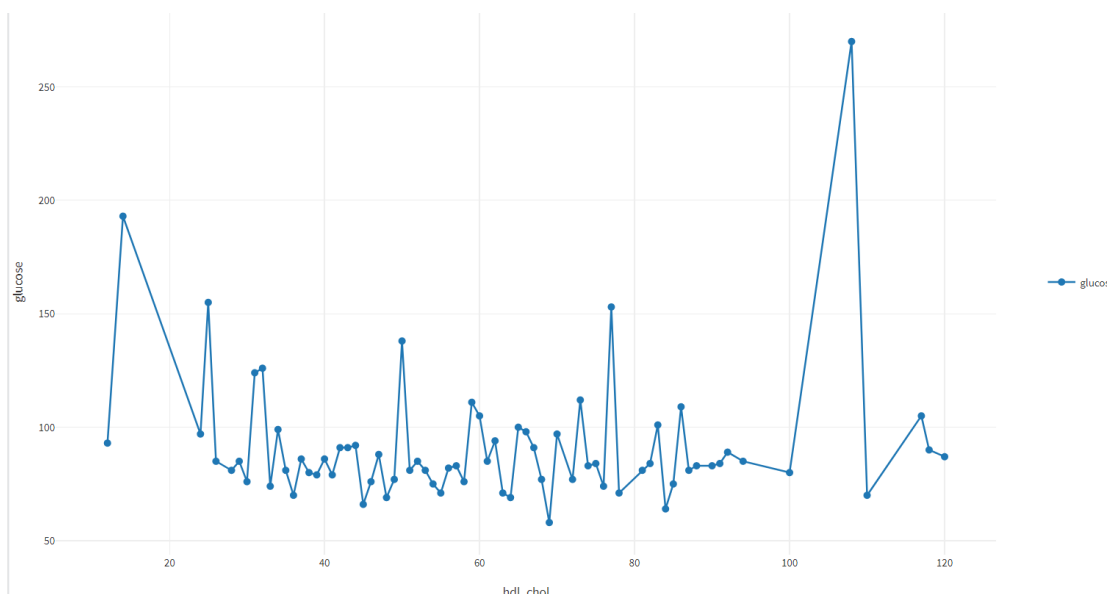


**Рисунок 2.12 — График зависимости "glucose" от "diastolic\_bp"**

Пациенты с сахарным диабетом имеют более низкий уровень холестерина ЛПВП.

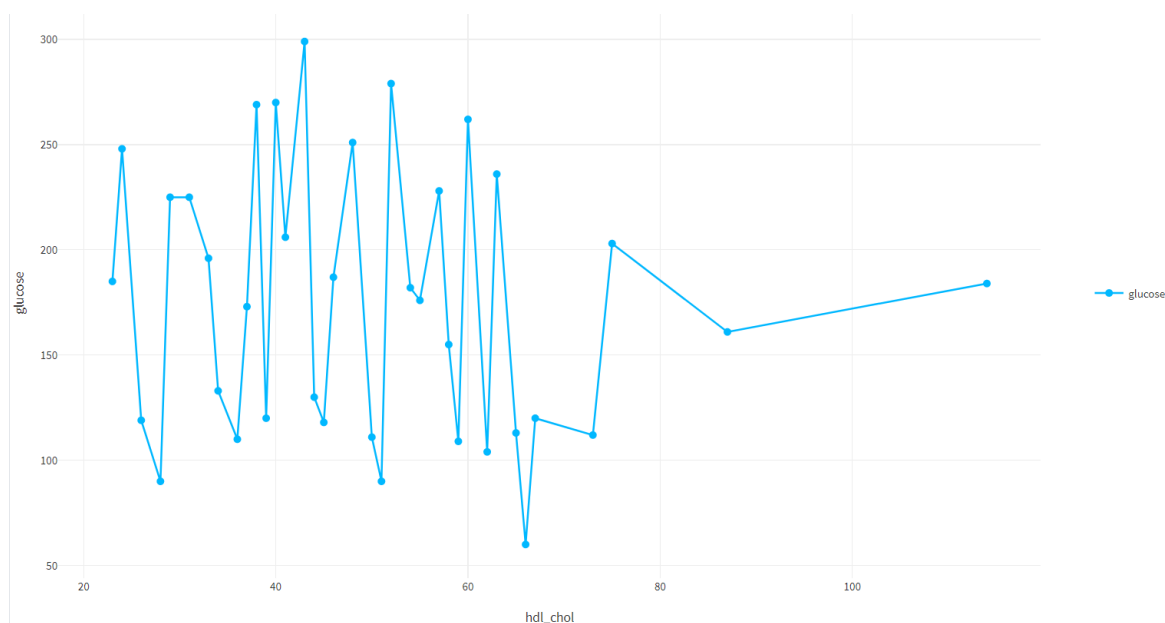
На Рисунке 2.13 показан график зависимости "glucose" от "hdl\_chol", у которых нет сахарного диабета.





**Рисунок 2.13 — График зависимости "glucose" от "hdl\_chol"**

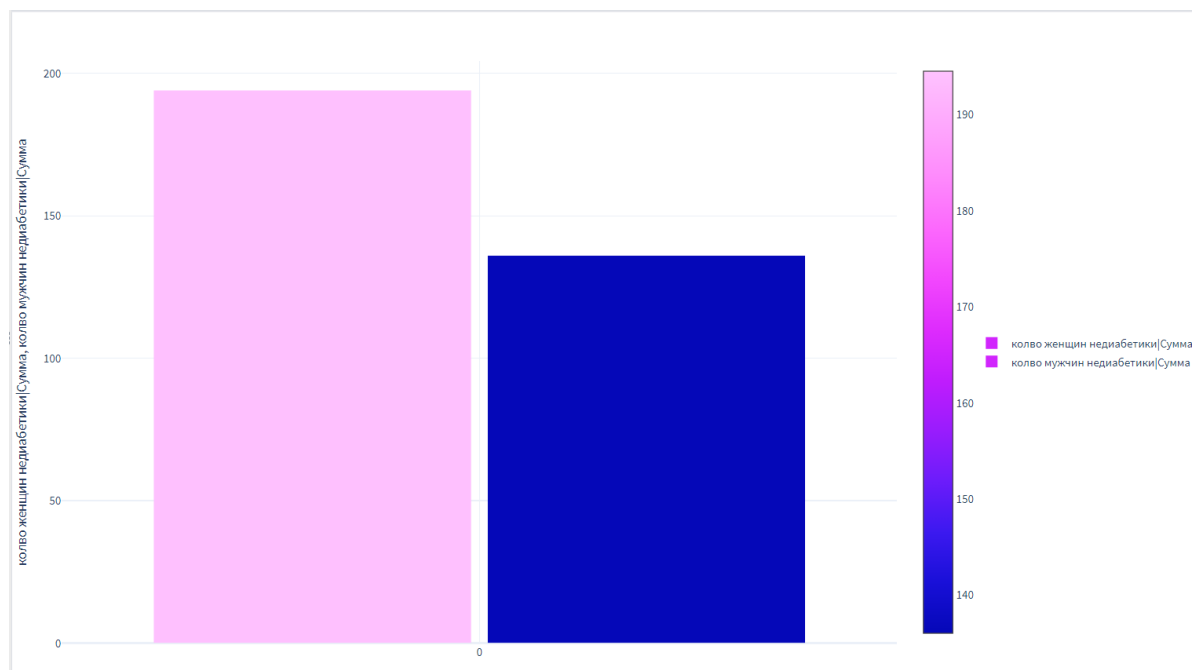
На Рисунке 2.14 показан график зависимости "glucose" от "hdl\_chol", у которых есть сахарный диабет.



**Рисунок 2.14 — График зависимости "glucose" от "diastolic\_bp"**

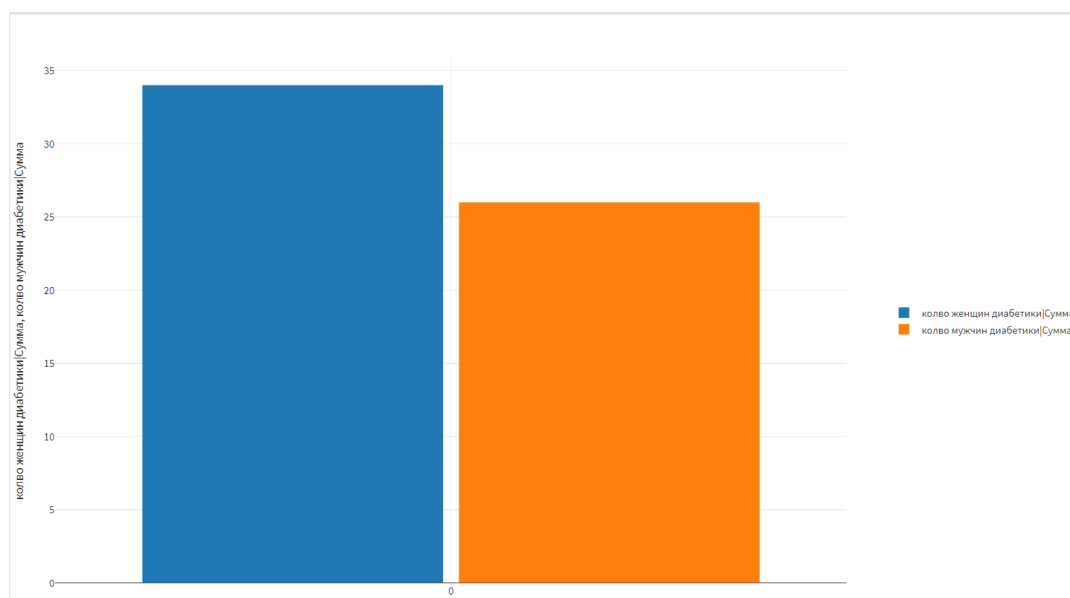
Женщин, страдающих диабетом, больше, чем мужчин, страдающих диабетом.

На Рисунке 2.15 показана столбчатая диаграмма.



**Рисунок 2.15 — Сравнение между женщинами и мужчинами, которые не диабетики**

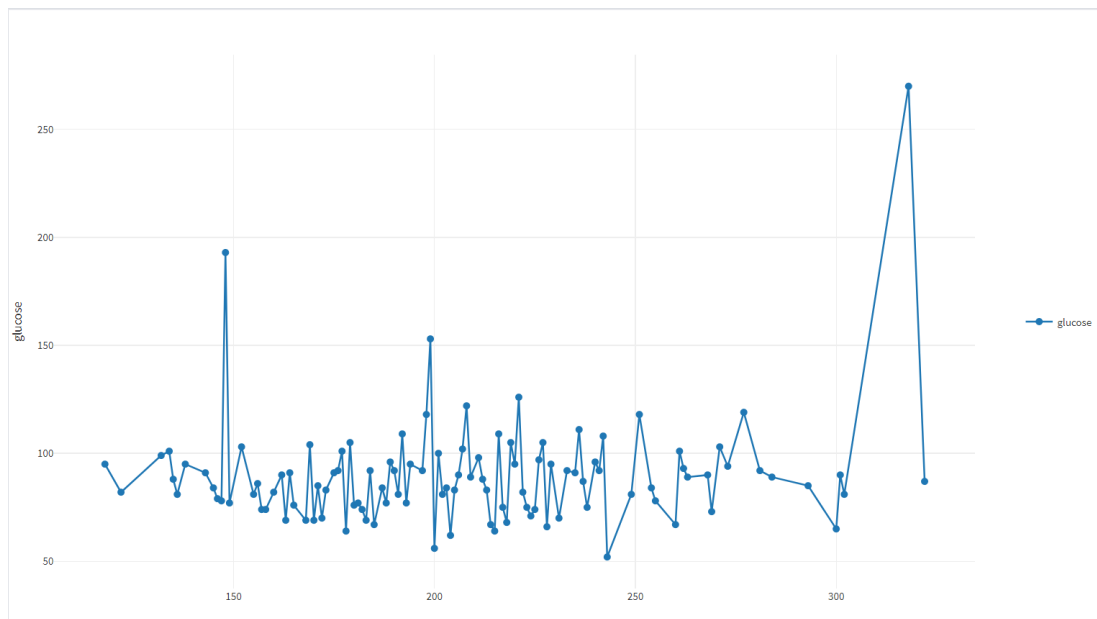
На Рисунке 2.16 показана столбчатая диаграмма.



**Рисунок 2.16 — Сравнение между женщинами и мужчинами, которые диабетики**

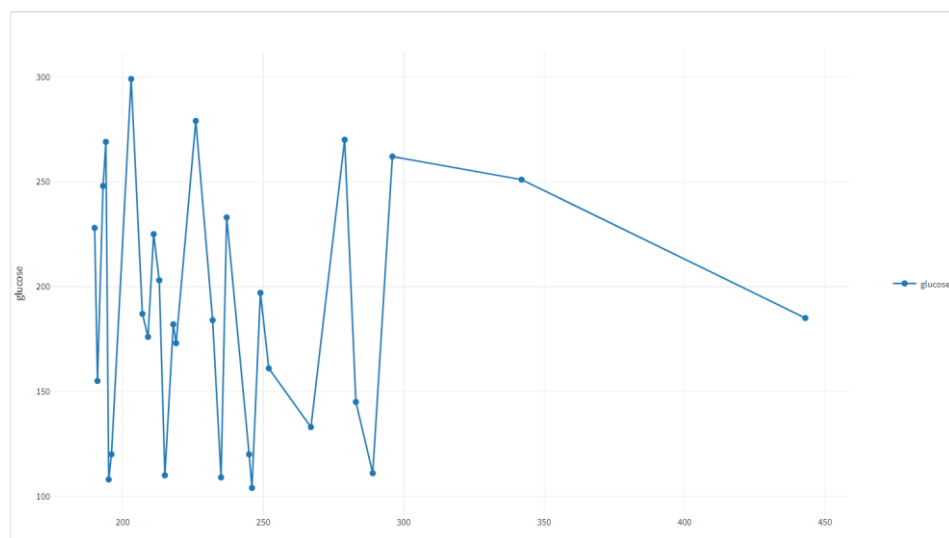
Более высокий уровень холестерина наблюдается у пациентов с сахарным диабетом

На Рисунке 2.17 показан график зависимости "glucose" от "cholesterol", у которых нет сахарного диабета.



**Рисунок 2.17 — График зависимости "glucose" от "cholesterol"**

На Рисунке 2.18 показан график зависимости "glucose" от "cholesterol", у которых есть сахарного диабета.



**Рисунок 2.18 — График зависимости "glucose" от "cholesterol"**

Для проведения корреляционно-регрессионного анализа необходимо разделить показатели на независимые (x) и зависимые (y) переменные. В данном случае зависимой переменной будет показатель «Люди, болеющие диабетом», а независимыми — показатели «Холестерин», «Глюкоза» и

«холестерин ЛПВП». Столбцы «Код пациента» в корреляционно-регрессионном анализе не участвуют.

Сначала нужно найти значение коэффициента Пирсона и по нему определить тесноту связи между зависимой и независимыми переменными. Для вычисления коэффициента Пирсона используем узел «Корреляционный анализ». Результат представлен на Рисунке 2.19.

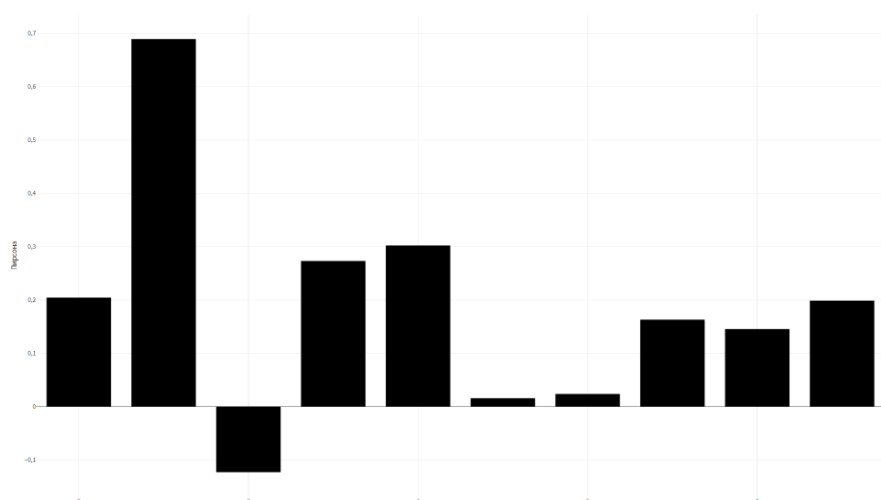
#	ab Поле1.И...	ab Поле1.Мет...	ab Поле2.Имя	ab Поле2.Мет...	9.0 Пирсо...
1	patients	patients	cholesterol	cholesterol	0,20
2	patients	patients	glucose	glucose	0,69
3	patients	patients	hdl_chol	hdl_chol	-0,12
4	patients	patients	chol_hdl_ratio	chol_hdl_ratio	0,27
5	patients	patients	age	age	0,30
6	patients	patients	gender	gender	0,02
7	patients	patients	height	height	0,02
8	patients	patients	weight	weight	0,16
9	patients	patients	bmi	bmi	0,15
10	patients	patients	systolic_bp	systolic_bp	0,20

**Рисунок 2.19 — Результат вычисления коэффициента Пирсона**

Из Рисунка 2.19 мы видим, что коэффициент корреляции Пирсона между показателем «Глюкозой» и показателем «Диабетиков» равен 0,69.

Это означает, что связь между показателем «Диабетиков» и показателем «Глюкозой» является тесной.

Посмотрим на сравнение значений коэффициента Пирсона всех независимых переменных. Визуализация представлена на Рисунке 2.20.



**Рисунок 2.20 — Сравнение коэффициентов Пирсона**

Теперь, когда известна степень тесноты связи между показателем «год» и зависимым показателем, можно проводить поиск значения коэффициента уравнения регрессии. В нашем случае уравнение имеет вид:

$$y = \frac{\exp(z)}{1+\exp(z)}; z = b_0 + b_1x.$$

То есть нам требуется найти значения неизвестных  $b_0$  и  $b_1$ .  $y$  — зависимая переменная, то есть показатели "diabetes", а  $x$  — независимая переменная, то есть показатель "glucose".

Для поиска значений уравнения воспользуемся узлом «Логистическая регрессия». Результаты выполнения узла показаны на Рисунке 2.21.

#	ab Имена входных пол...	ab Метки входных полей	ab Уникальные значения	9.8 diabetes Коэффициенты
1	<Константа>	<null>	<null>	-6,02
2	glucose	glucose	<null>	0,03

**Рисунок 2.21 — Результат вычисления коэффициентов уравнения логистической регрессии, где  $x$  — показатель "glucose"**

На Рисунке 2.21 показано, что неизвестные уравнения

$$y = \frac{\exp(z)}{1+\exp(z)}; z = b_0 + b_1x; b_1 = 0,03; b_0 = -6,02.$$

То есть, уравнение принимает вид

$$y = \frac{\exp(z)}{1+\exp(z)}; z = 0,03 * x - 6,02.$$

С помощью найденных коэффициентов можно построить линию регрессии для независимого показателя. Узел «Логистическая регрессия» автоматически вычисляет значения  $y$  в зависимости от  $x$  для найденного уравнения. Результаты вычисления значений  $y$  представлены на Рисунке 2.22.

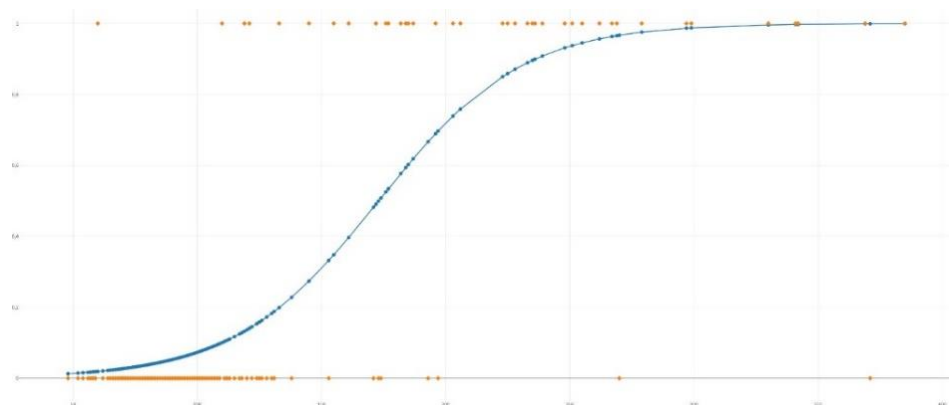
0/1 Событие Прогноз	9.0 Вероятность события Прогноз	0/1 Событие Факт	ab diabetes Прогноз	ab gender	ab diabetes
false	0,04	false	No diabetes	male	No diabetes
false	0,04	false	No diabetes	male	No diabetes
true	0,70	false	Diabetes	male	No diabetes

**Рисунок 2.22 — Результат прогноза логистической регрессии**

Из Рисунок 2.22 мы можем увидеть, какие значения принимает уравнение логистической регрессии (показатель «Вероятность события|Прогноз») в зависимости от  $x$  (показатель "glucose").

Так как мы знаем значения уравнения в зависимости от  $x$ , можно построить линию регрессии с помощью визуализатора «Диаграмма» узла «Логистическая регрессия», где осью  $X$  будет показатель "glucose", а осью  $Y$  — показатель «Вероятность события». Результат построения линий представлен на Рисунок 2.23.

На Рисунок 2.23 видно, что линии регрессии проходят сквозь поля скопления точек, обозначающих зависимость Вероятности от глюкозы.



**Рисунок 2.23 — Диаграмма построения линии регрессии с помощью уравнения логистической регрессии**

Это говорит о том, что модель классифицирована не очень хорошо, результаты представлены на Рисунок 2.24.

10	9.0 R2	Псевдо- $R^2$ Макфаддена	0,42
11	9.0 AdjustedR2	Псевдо- $R^2$ Макфаддена (скорр.)	0,40

**Рисунок 2.24 — Результат вычисления коэффициента детерминации для показателя "glucose"**

На Рисунке 2.24 видно, что Псевдо- $R^2$ -Макфаддена = 0,42. Это означает, что модель является плохой. [2.1]

Так как данные не сбалансированы, для правильного прогнозирования необходимо скорректировать, чтобы получить модель с хорошей точностью.

Понимание основных факторов, связанных с развитием диабета, позволяет разрабатывать более эффективные методы скрининга и профилактики, что в конечном итоге способствует улучшению здоровья населения и снижению затрат на лечение осложнений диабета.

Таким образом, проведя корреляционно-регрессионный анализ мы выяснили, что основным фактором заболевания сахарным диабетом является повышенная глюкоза. Модель может предсказать, заболеет ли человек сахарным диабетом.

Мы вывели уравнения регрессии и построили линии регрессии, с помощью которых есть возможность заранее предсказания заболеет человек сахарным диабетом. Что даёт нам возможность использовать в медицинской сфере, которое мы предсказали с помощью уравнений регрессии, что будет сказываться в лучшую сторону для людей, которая будет уведомлять о риске заболеваний сахарном диабетом.

## ЗАКЛЮЧЕНИЕ

Сахарный диабет — это нарушение обмена веществ, характеризующееся повышением содержания сахара в крови.

Заболевание возникает в результате дефектов выработки инсулина, дефекта действия инсулина или обоих этих факторов. Помимо повышенного уровня сахара крови, заболевание проявляется выделением сахара с мочой, обильным мочеиспусканием, повышенной жаждой, нарушениями жирового, белкового и минерального обменов и развитием осложнений.

Цель данной курсовой работы — провести корреляционно-регрессионный анализ признаков и рисков сахарного диабета для раннего прогнозирования достигнута.

В ходе выполнения данной курсовой работы проведен корреляционно-регрессионный анализ влияния выпуска валовой продукции в отрасли на количество работников в отрасли с использованием low-code платформы Loginom.

Задачи, выполненные в данной курсовой работе:

- изучены научная и методическая литературы о заболеваниях сахарным диабетом и корреляционно-регрессионном анализе;
- выполнен корреляционно-регрессионный анализ собранных данных;
- использованы знания математической статистики с использованием современных средств обработки данных: аналитической платформы Loginom;
- пройдено обучение оформлению официальных документов. [2.2]



# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

## ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

- 1.1. Риск развития сахарного диабета в Москве/ Гемотест  
[Электронный ресурс]. <https://gemotest.ru/moskva/catalog/chastishchut/sakharnyy-diabet/risk-razvitiya-sakharnogo-diabeta/>
- 1.2. Корреляционно-регрессионный анализ / Научный словарь-справочник «Справочник24» [Электронный ресурс]  
[https://spravochnik.ru/ekonomicheskij\\_analiz/korrelyacionno-regressionnyy\\_analiz/](https://spravochnik.ru/ekonomicheskij_analiz/korrelyacionno-regressionnyy_analiz/)
- 1.3. Корреляционно-регрессионный анализ: пример, задачи, применение. Метод корреляционно-регрессионного анализа / Интернет-портал BusinessMan.ru [Электронный ресурс].  
<https://businessman.ru/new-korrelyacionno-regressionnyj-analiz-primer-zadachi-primeneniye.html>
- 1.4. Бондаренко П.С., Горелова Г.В., Кацко И.А., Жминько А.Е., Соловьёва Т.В., Кацко С.А., Куижева С.К., Митус А.А., Паклин Н.Б., Сенникова А.Е.. Эконометрика. Практикум: учебно-практическое пособие / коллектив авторов; под ред. И.А. Кацко. — Москва: КНОРУС, 2019. — 218 с.
- 1.5. Александрова, О.В. Статистические методы решения технологических задач: учебное пособие / Александрова О.В., Мацеевич Т.А., Кирьянова Л.В., Соловьев В.Г.. — Москва: Издательство МИСИ — МГСУ, 2017. — 154 с.
- 1.6. Круценюк, К.Ю. Корреляционно-регрессионный анализ в эконометрических моделях: учебное пособие / К. Ю. Круценюк. — Норильск: НГИИ, 2018. — 108 с.
- 1.7. Ганичева, А. В. Прикладная статистика: учебное пособие для спо / А. В. Ганичева. — Санкт-Петербург: Лань, 2021. — 164 с.

- 1.8. Метод корреляционно-регрессионного анализа / Студенческие реферативные статьи и материалы "Studref" [Электронный ресурс]. [https://studref.com/591347/ekonomika/metod\\_korrelyatsionno\\_regressiionnogo\\_analiza](https://studref.com/591347/ekonomika/metod_korrelyatsionno_regressiionnogo_analiza)
- 1.9. Кийко, П. В. Эконометрика. Регрессионные модели: учебное пособие / П. В. Кийко, Н. В. Щукина. — Омск: Омский ГАУ, 2021. — 83 с.
- 1.10. Ситилаб / Лишний вес и диабет — какая зависимость. <https://citilab.ru/articles/lishnii-ves-i-diabet/>
- 1.11. Факторный анализ / Loginom Help / Документация — <https://help.loginom.ru/userguide/processors/scrutiny/factor-analysis.html>

## ПРАКТИЧЕСКАЯ ЧАСТЬ

- 2.1. Loginom / Метрики линейных регрессионных моделей. <https://loginom.ru/blog/quality-metrics>
- 2.2. Loginom / Документация [Электронный ресурс]. <https://help.loginom.ru/userguide/>

## **ПРИЛОЖЕНИЯ**

Приложение А — Графический материал.

## Приложение А

На Рисунке А.1 представлен сценарий проекта в аналитической платформе Loginom, полученный в результате выполнения данной курсовой работы.

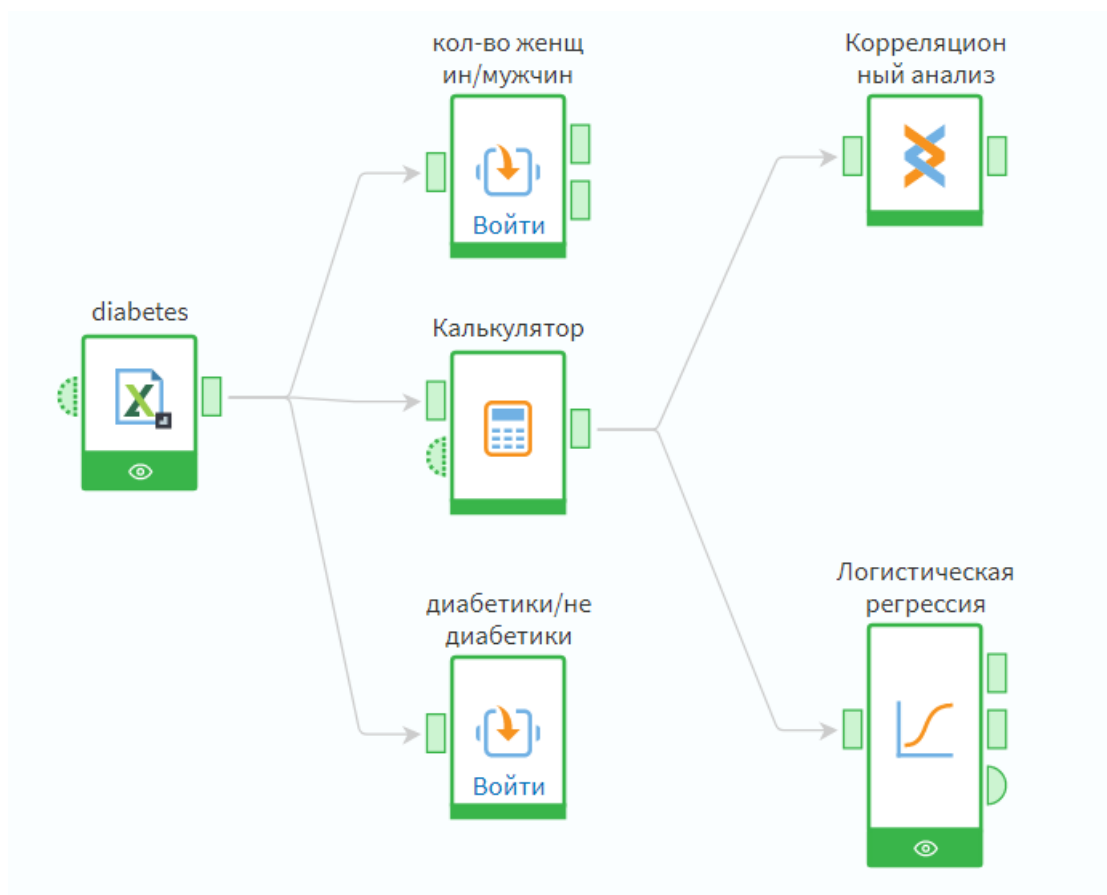


Рисунок А.1 — Сценарий проекта в аналитической платформе Loginom

Этапы выполнения сценария:

1. Импорт исходных данных (Excel файлы).
2. Предобработка исходных данных.
3. Корреляционно-регрессионный анализ.