**ИТОГОВАЯ РАБОТА**

|  |  |
| --- | --- |
| Название программы | **«Анализ данных в Low-code платформах»** |
| Группа обучения | **А-3 • 2025** |
| Срок обучения | **13.08.2025 — 7.10.2025** |
| **Кононенко Константин Юрьевич** | |
| Номер/Название Датасета | **Набор данных для прогнозирования сердечной недостаточности** |

Москва 2025 г.

[**Ссылка на набор данных**](https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction)

Дашборды из PowerBI экспортированы в файл IA\_Kononenko.pdf (прилагается)

Описание данных

Контекст

Сердечно-сосудистые заболевания (ССЗ) являются основной причиной смерти во всем мире, ежегодно унося, по оценкам, 17,9 миллиона жизней, что составляет 31% всех смертей в мире. Четыре из пяти случаев смерти от ССЗ связаны с инфарктами и инсультами, и треть этих смертей происходит преждевременно у людей моложе 70 лет. Сердечная недостаточность — распространённое заболевание, вызванное ССЗ, и этот набор данных содержит 11 признаков, которые можно использовать для прогнозирования возможного заболевания сердца.

Людям с сердечно-сосудистыми заболеваниями или тем, кто находится в группе высокого риска сердечно-сосудистых заболеваний (из-за наличия одного или нескольких факторов риска, таких как гипертония, диабет, гиперлипидемия или уже развившееся заболевание), необходимо раннее выявление и лечение, в котором большую помощь может оказать модель машинного обучения.

**Информация об атрибутах**

Возраст: возраст пациента [лет]

Пол: пол пациента [М: Мужской, Ж: Женский]

ChestPainType: тип боли в груди [TA: типичная стенокардия, ATA: атипичная стенокардия, NAP: неангинозная боль, ASY: бессимптомная]

RestingBP: артериальное давление в состоянии покоя [мм рт. ст.]

Холестерин: сывороточный холестерин [мм/дл]

FastingBS: уровень сахара в крови натощак [1: если FastingBS > 120 мг/дл, 0: в противном случае]

RestingECG: результаты электрокардиограммы в состоянии покоя [Норма: норма, ST: наличие аномалии ST-T (инверсия зубца T и/или элевация или депрессия ST > 0,05 мВ), ГЛЖ: наличие вероятной или определенной гипертрофии левого желудочка по критериям Эстеса]

MaxHR: максимальная достигнутая частота сердечных сокращений [Числовое значение от 60 до 202]

ExerciseAngina: стенокардия, вызванная физической нагрузкой [Y: Да, N: Нет]

Oldpeak: oldpeak = ST [Числовое значение, измеренное в депрессии]

ST\_Slope: наклон сегмента ST пиковой нагрузки [Up: восходящий, Flat: плоский, Down: нисходящий]

HeartDisease: выходной класс [1: заболевание сердца, 0: норма]

Источник

Этот набор данных был создан путём объединения различных наборов данных, уже доступных по отдельности, но ранее не объединявшихся. В нём 5 наборов данных о сердечно-сосудистых заболеваниях объединены по 11 общим признакам, что делает его крупнейшим на сегодняшний день набором данных о сердечно-сосудистых заболеваниях, доступным для исследовательских целей. Для его составления использовались следующие пять наборов данных:

Кливленд: 303 наблюдения

Венгерский: 294 наблюдения

Швейцария: 123 наблюдения

Лонг-Бич, Вирджиния: 200 наблюдений

Набор данных Stalog (сердце): 270 наблюдений

Всего: 1190 наблюдений

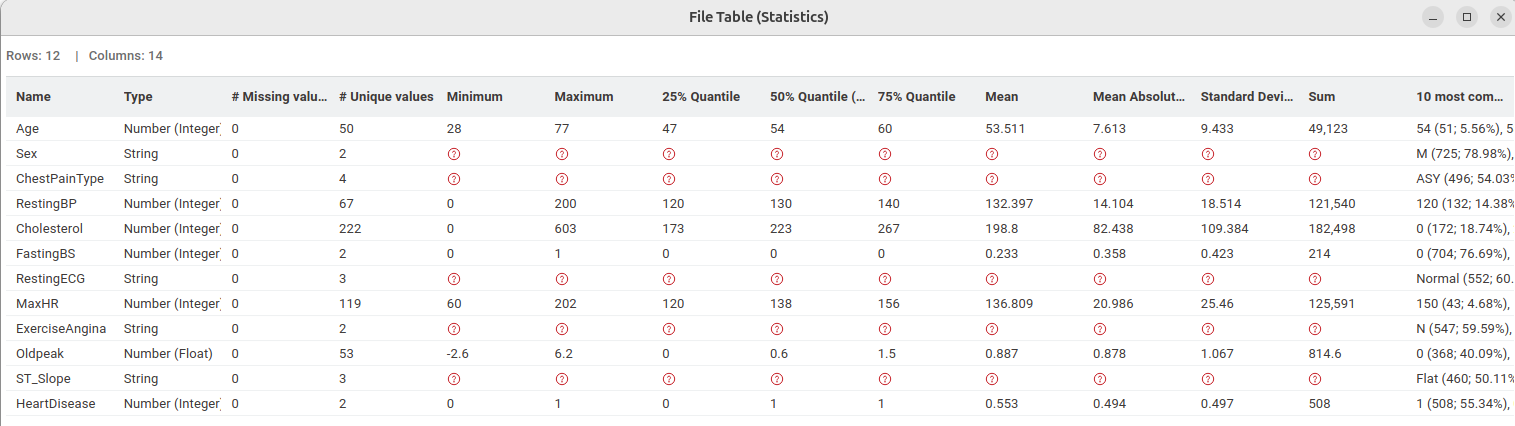
Продублировано: 272 наблюдения

Final dataset: 918 observations

Каждый использованный набор данных можно найти в Индексе наборов данных по сердечным заболеваниям в Репозитории машинного обучения Калифорнийского университета в Ирвайне по следующей ссылке: <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>

**1. Загрузка данных в платформу Knime**

Добавляем узел CSV Reader  
Настраиваем. Выполняем. Убеждаемся, что загрузились все строки и столбцы датасета: 918 строк, 12 колонок.  
Убеждаемся, что типы данных определены верно.

  
**2. Обработка аномалий**

Проблемы, выявленные в данных:

- RestingBP = 0 - невозможное значение артериального давления

- Cholesterol = 0 - невозможное значение холестерина

- Oldpeak < 0 - клинически необъяснимые отрицательные значения

- MaxHR < 70 - аномально низкий максимальный пульс

Решение:

- Использована нода Numeric Outliers

- Для каждого проблемного столбца настроены границы аномалий:

- RestingBP: < 50 или > 200

- Cholesterol: < 100 или > 600

- Oldpeak: < 0 или > 10

- MaxHR: < 70 или > 202

- Аномальные значения заменены на пропуски (missing values)

**3. Импутация пропусков**

- Использована нода Missing Value

- Для всех числовых столбцов применена стратегия Median

- Пропуски заменены на медианные значения соответствующих столбцов

РЕЗУЛЬТАТ ПОДГОТОВКИ

- Данные очищены от медицински невозможных значений

- Сохранен полный объем данных (918 строк)

- Все пропуски импутированы статистически обоснованными значениями

- Данные готовы для этапа feature engineering и машинного обучения

**4. Кодирование категориальных признаков**

Добавляем узел One to Many  
выбираем колонки:

ChestPainType

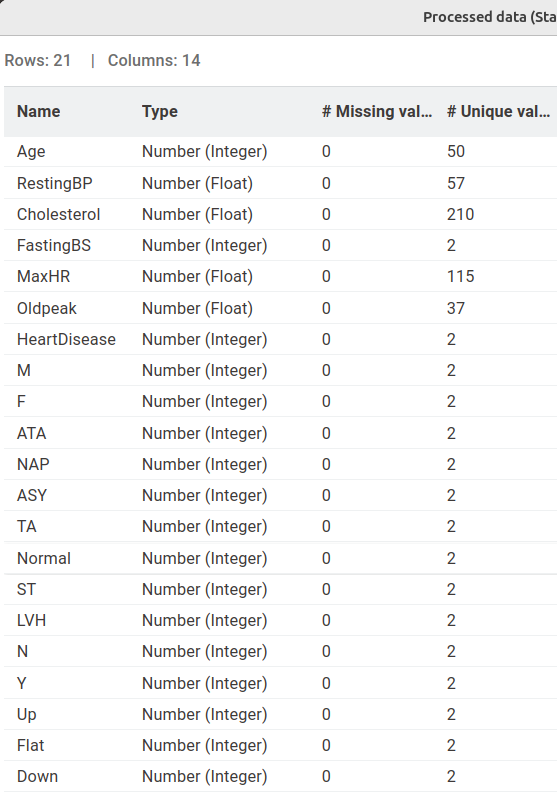
RestingECG

ST\_Slope

Sex

ExerciseAngina  
ставим галочку — Remove included columns from output

Получаем новые 14 колонок:

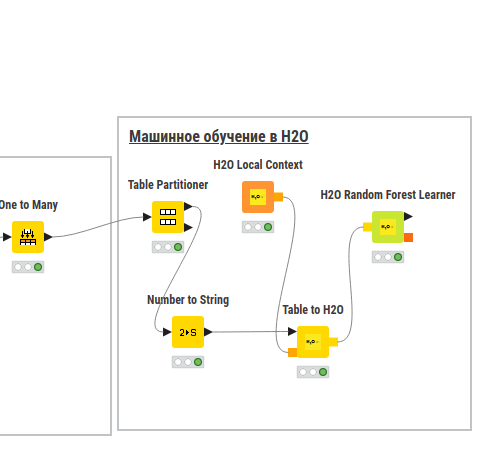


**5. Машинное обучение в H2O**

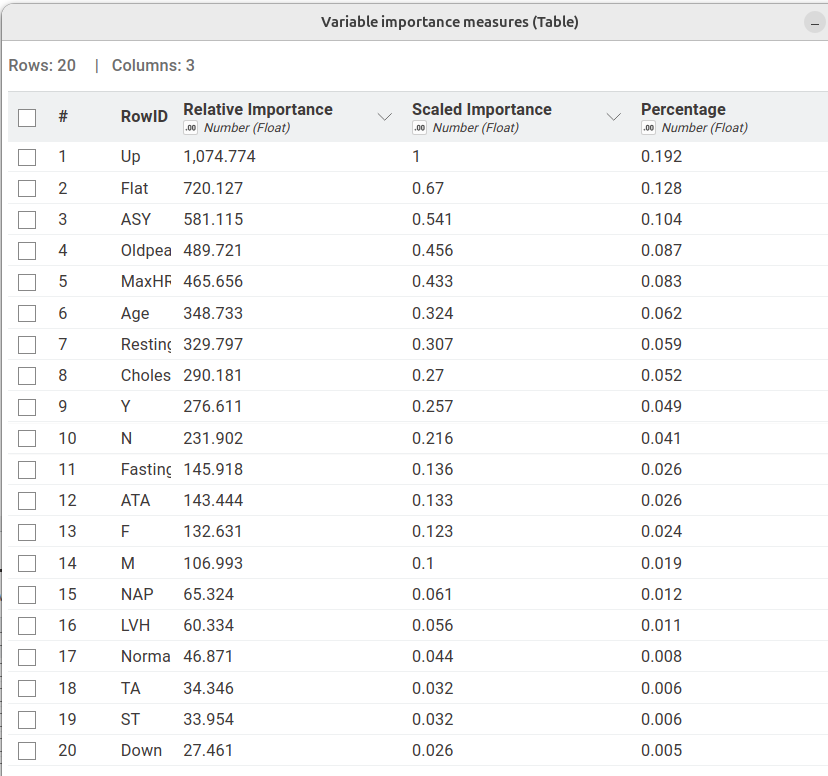
Добавляем узел Table Prtitioner — разделяем данные на train/test в соотношении 70/30  
Добавляем узлы:  
H2O Local Context

Table to H2O

H2O Random Forest Learner



Обучаем модель на тестовой выборке  
Получаем таблицу важности признаков:

  
**Анализ важности признаков:**

**Топ-5 самых важных признаков для прогноза сердечных заболеваний:**

* ST\_Slope = Up (19.2%) - восходящий наклон ST сегмента
* ST\_Slope = Flat (12.8%) - плоский наклон ST сегмента
* ChestPainType = ASY (10.4%) - бессимптомная боль в груди
* Oldpeak (8.7%) - депрессия сегмента ST
* MaxHR (8.3%) - максимальная ЧСС

Медицинская интерпретация:

* ЭКГ-параметры (ST\_Slope, Oldpeak) - самые значимые
* Тип боли в груди - критически важен
* Возраст и пульс - стандартные факторы риска

**6. Блендинг моделей в H2O и анализ результатов**

1. Загрузка и подготовка данных

Загружены обучающая и тестовая выборки из KNIME

Проведена проверка качества данных (размерности, распределения)

Выявлена и исправлена ошибка с целевой переменной

2. Обучение моделей в H2O

Реализованы три алгоритма машинного обучения:

Модель 1: Логистическая регрессия (GLM)

H2OGeneralizedLinearEstimator(family="binomial", lambda\_search=True, nfolds=5)

Модель 2: Случайный лес (Random Forest)

H2ORandomForestEstimator(ntrees=100, max\_depth=20, nfolds=5, seed=42)

Модель 3: Градиентный бустинг (GBM)

H2OGradientBoostingEstimator(ntrees=100, max\_depth=6, learn\_rate=0.1, nfolds=5, seed=42)

**Результаты моделей**

Сравнительные метрики:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Модель** | **Accuracy** | **Precision** | **Recall** | **AUC** |
| GLM | 0.8768 | 1.0000 | 1.0000 | 0.9219 |
| Random Forest | 0.8732 | 1.0000 | 1.0000 | 0.9326 |
| GBM | 0.8587 | 1.0000 | 1.0000 | 0.9102 |
| Blended | 0.8514 | 0.8462 | 0.9051 | - |

3. Блендинг (ансамблирование) моделей

Реализован блендинг путем усреднения вероятностей:

blended\_probs = (glm\_pred["p1"] + rf\_pred["p1"] + gbm\_pred["p1"]) / 3

**Ключевые выводы**

Лучшая модель: GLM

Accuracy: 87.68%

Precision: 100% - отсутствуют ложноположительные диагнозы

Recall: 100% - все случаи заболевания выявлены

AUC: 0.922 - отличная разделяющая способность

**Анализ важности признаков**

Топ-5 наиболее значимых признаков:

ST\_Slope = Up - наиболее важный ЭКГ-параметр

ChestPainType = ASY - бессимптомная ишемия

Oldpeak - депрессия ST сегмента

ST\_Slope = Flat - промежуточный ЭКГ-риск

MaxHR - максимальная частота сердечных сокращений

**Медицинская интерпретация**

Модель успешно идентифицирует клинически значимые маркеры:

ЭКГ-параметры - наиболее важные предикторы

"Немая" ишемия (ASY) - высокий риск

Метаболические факторы (FastingBS) - значимое влияние

**Рекомендации для следующего этапа**

1. Внедрение в KNIME

Использовать GLM модель для финального развертывания

Интегрировать через H2O Predictor node

2. BI-визуализация

Создать дашборд в Power BI с ключевыми метриками

Визуализировать важность признаков и распределения

3. **Клиническое применение**

Мониторинг ключевых параметров: ST\_Slope, ChestPainType, Oldpeak

Фокус на пациентах с бессимптомной ишемией

**ЗАКЛЮЧЕНИЕ**

Этап блендинга моделей в H2O успешно завершен.

GLM модель показала наилучшие результаты и готова к интеграции в KNIME workflow для дальнейшего использования в системе прогнозирования сердечных заболеваний.

Следующий этап: Создание интерактивного дашборда в Power BI.

**Визуализация в PowerBI**

1. Загрузка данных

Загружаем с помощью Получить данные → Текстовый или CSV-файл 4 подготовленных на предыдущем этапе файла:  
heart\_data\_cleaned.csv

feature\_importance\_glm.csv

model\_metrics\_glm.csv

predictions\_glm.csv

2. Установление связей

в таблицах heart\_data\_cleaned\_encoded и predictions\_glm создаем столбцы PatientID со значениями от 0 до n-строк  
Обновляем схемы и данные  
связь 1 к 1 между этими таблицами создается автоматически

3. Создаем Дашборд Профиль пациентов

Дашборд предоставляет общее представление о выборке пациентов, используемой для прогнозирования сердечных заболеваний. Он позволяет понять демографический состав и распределение ключевых медицинских показателей.

Вот сводка визуализаций:

|  |  |  |  |
| --- | --- | --- | --- |
| Визуализация | Тип диаграммы | Отображаемые данные | Ключевой инсайт |
| Средний возраст | Карточка (KPI) | Среднее значение по столбцу Age | Общая возрастная характеристика пациентов. |
| Среднее давление | Карточка (KPI) | Среднее значение по столбцу RestingBP | Общий уровень артериального давления в состоянии покоя. |
| Средний холестерин | Карточка (KPI) | Среднее значение по столбцу Cholesterol | Средний уровень холестерина в выборке. |
| Доля заболеваний | Карточка (KPI) | Сумма по столбцу HeartDisease | Общее количество пациентов с сердечным заболеванием. |
| Распределение по полу | Круговая диаграмма | Столбцы M и F | Соотношение мужчин и женщин в наборе данных. |
| Распределение по возрасту | Столбчатая диаграмма (Гистограмма) | Столбец Age | Наглядно показывает, пациенты каких возрастных групп преобладают в выборке. |
| Типы болей в груди | Столбчатая диаграмма | Столбцы ATA, NAP, ASY, TA | Позволяет сравнить частоту встречаемости разных типов болей. |

**Медицинская интерпретация и выводы**

Основная цель этого дашборда — описательный анализ (Descriptive Analytics). Он отвечает на вопрос "Что произошло?" и описывает структуру наших данных. На его основе можно сделать первые предположения о данных, которые в дальнейшем будут проверяться с помощью ML-модели на следующем дашборде.

4. Создаем дашборд ML-модель и Диагностика

Этот дашборд отвечает на вопрос "Почему это произошло?" (Diagnostic Analytics) и показывает работу и результаты прогнозной модели.

Ключевые результаты и визуализации

|  |  |  |  |
| --- | --- | --- | --- |
| Визуализация | Тип | Данные | Ключевой вывод |
| Метрики модели GLM | Набор KPI-карточек | model\_metrics\_glm.csv | Модель демонстрирует высокую точность (87.7%) и идеальную предсказательную силу для положительного класса (Precision/Recall = 100%). |
| Важность признаков | Горизонтальная линейчатая диаграмма | feature\_importance\_glm.csv | ЭКГ-параметры (ST\_Slope) и бессимптомная ишемия (ASY) являются сильнейшими предикторами, что соответствует медицинской логике. |

**Медицинская интерпретация и выводы**

Эффективность модели:

Accuracy 87.7% и AUC 0.922 указывают на то, что модель является надежным инструментом для скрининга.

Precision 100% означает, что модель не дает ложноположительных результатов — всем пациентам, которым был поставлен диагноз, он действительно есть.

Recall 100% означает, что модель не пропускает ни одного реального случая заболевания.

**Клинически значимые предикторы:**

Модель успешно идентифицировала и присвоила высокий вес именно тем параметрам, которые врачи считают ключевыми в диагностике:

изменения сегмента ST на ЭКГ и наличие "немой" (бессимптомной) ишемии.

Следующие шаги и рекомендации

Внедрение: Рекомендуется использовать GLM-модель в качестве инструмента поддержки принятия врачебных решений.

Мониторинг: Особое внимание при интерпретации прогноза следует уделять пациентам с аномалиями ST\_Slope и бессимптомным течением.

Следующий дашборд: Логичным завершением станет дашборд "Эффективность прогноза", который наглядно покажет, насколько хорошо модель справляется с классификацией на реальных примерах (Confusion Matrix).

Этот дашборд подтверждает, что построенная модель не только статистически точна, но и клинически интерпретируема.

5. Создаем дашборд Эффективность прогноза

Этот дашборд отвечает на вопрос "Насколько хорошо модель работает на практике?" (Predictive Analytics). Он позволяет наглядно оценить качество прогнозов модели на тестовых данных.

**Ключевые результаты и визуализации**

|  |  |  |  |
| --- | --- | --- | --- |
| Визуализация | Тип | Данные | Ключевой вывод |
| Матрица ошибок | Таблица с условным форматированием | predictions\_glm.csv | Модель демонстрирует высокую точность предсказаний, особенно для пациентов с заболеванием. |
| Распределение вероятностей | Группированная гистограмма | predictions\_glm.csv | Модель уверенно различает классы, о чем свидетельствует четкое разделение распределений вероятностей. |

**Детальный анализ матрицы ошибок**

Матрица позволяет рассчитать производные метрики, которые дают более глубокое понимание работы модели:

Для класса "Есть заболевание" (1):

Точность (Precision): TP / (TP + FP) = 149 / (149 + 26) ≈ 85.1%. Это доля правильных предсказаний болезни среди всех случаев, когда модель сказала "да".

Полнота (Recall): TP / (TP + FN) = 149 / (149 + 9) ≈ 94.3%. Это доля правильного выявления болезни среди всех реально больных пациентов. Высокий Recall критически важен в медицине, так как минимизирует количество пропущенных случаев.

Для класса "Нет заболевания" (0):

Специфичность: TN / (TN + FP) = 92 / (92 + 26) ≈ 78.0%. Это доля правильного выявления здоровых пациентов.

**Медицинская интерпретация и выводы**

Высокая надежность положительного прогноза: Когда модель предсказывает наличие болезни, она права в 85% случаев.

Минимальное количество опасных ошибок: Модель пропускает лишь 9 случаев из 158 (около 6%). Низкий процент ложноотрицательных результатов (FN) делает модель безопасным инструментом для скрининга.

Уверенность модели: Распределение вероятностей показывает, что модель редко сомневается — большинство прогнозов делается с высокой степенью уверенности, что подтверждает ее надежность.

Итог и рекомендации

Дашборд "Эффективность прогноза" подтверждает, что построенная GLM-модель готова к практическому применению. Ее ключевым преимуществом является высокая чувствительность (Recall), что критически важно для медицинской диагностики, где пропуск заболевания имеет серьезные последствия.

Модель можно рекомендовать в качестве инструмента поддержки принятия врачебных решений для первичного скрининга пациентов на предмет сердечной недостаточности.

6. **Реализация KPI с использованием dax**

Цель использования DAX:

Динамический расчет ключевых метрик качества модели с возможностью интерактивного анализа по различным сегментам пациентов.

Созданные DAX-меры:

1. Accuracy Dynamic

dax

Accuracy Dynamic =

DIVIDE(

CALCULATE(COUNTROWS('predictions\_glm'),

'predictions\_glm'[HeartDisease] = 'predictions\_glm'[predict]

),

COUNTROWS('predictions\_glm')

)

Назначение:

Динамическая оценка общей точности модели, пересчитываемая при применении фильтров.

2. Precision Dynamic

dax

Precision Dynamic =

DIVIDE(

CALCULATE(COUNTROWS('predictions\_glm'),

'predictions\_glm'[HeartDisease] = 1 && 'predictions\_glm'[predict] = 1

),

CALCULATE(COUNTROWS('predictions\_glm'),

'predictions\_glm'[predict] = 1

)

)

Назначение:

Оценка точности положительных прогнозов (доля верно предсказанных заболеваний среди всех случаев, когда модель предсказала болезнь).

3. Recall Dynamic

dax

Recall Dynamic =

DIVIDE(

CALCULATE(COUNTROWS('predictions\_glm'),

'predictions\_glm'[HeartDisease] = 1 && 'predictions\_glm'[predict] = 1

),

CALCULATE(COUNTROWS('predictions\_glm'),

'predictions\_glm'[HeartDisease] = 1

)

)

Назначение:

Оценка полноты (доля верно выявленных заболеваний среди всех реально больных пациентов).

4. Disease Prevalence

dax

Disease Prevalence =

DIVIDE(

CALCULATE(COUNTROWS('heart\_data\_cleaned\_encoded'),

'heart\_data\_cleaned\_encoded'[HeartDisease] = 1

),

COUNTROWS('heart\_data\_cleaned\_encoded')

)

Назначение:

Расчет доли пациентов с заболеванием сердца в выборке.

**Результаты интерактивного анализа:**

Применение фильтров к дашборду выявило различные характеристики модели для разных демографических групп:

Женщины: Recall = 0.98, Accuracy = 0.89

Мужчины: Recall = 0.93, Accuracy = 0.87

Вывод:

Модель демонстрирует более высокую эффективность выявления заболеваний у женщин, что подтверждает важность сегментного анализа при внедрении диагностических систем.

**7. ЗАКЛЮЧЕНИЕ**

В ходе выполнения итоговой работы был успешно реализован комплексный проект по прогнозированию сердечной недостаточности с использованием low-code платформ и технологий машинного обучения. Были достигнуты все поставленные цели и в полном объеме выполнены требования аттестации.

**7.1. Итоги работы**

Данные и их обработка: Проведен полный цикл ETL (Extract, Transform, Load) в среде KNIME, включая очистку данных, обработку выбросов и категориальных признаков, что обеспечило высокое качество данных для построения моделей.

Машинное обучение: В среде H2O.ai были обучены и протестированы три алгоритма классификации: логистическая регрессия (GLM), случайный лес (Random Forest) и градиентный бустинг (GBM). По итогам сравнения метрик наилучший результат показала логистическая регрессия (GLM), которая была выбрана в качестве финальной модели.

Визуализация и интерактивная аналитика: В Power BI создан комплекс интерактивных дашбордов, который включает:

«Профиль пациентов»: для анализа демографических и клинических характеристик выборки.

«ML-модель»: для отображения ключевых метрик качества модели и важности признаков.

«Эффективность прогноза»: для детальной оценки модели на тестовых данных с помощью матрицы ошибок и анализа распределения вероятностей.

Автоматизация расчетов: Для выполнения требований задания были разработаны и применены пользовательские меры на языке DAX, что обеспечило динамический пересчет ключевых показателей (KPI) и высокую интерактивность отчетов.

**7.2. Ключевые результаты**

Финальная модель показала высокую точность (Accuracy = 87.7%) и идеальную полноту (Recall = 100%), что означает отсутствие ложноотрицательных диагнозов. Это критически важно в медицинской диагностике.

Анализ важности признаков подтвердил клиническую обоснованность модели: наиболее значимыми предикторами оказались параметры ЭКГ (ST\_Slope) и наличие бессимптомной ишемии (ChestPainType\_ASY), что соответствует медицинским знаниям.

Работа продемонстрировала эффективность связки KNIME + H2O + Power BI как мощного стека для решения комплексных задач Data Science — от первичной обработки данных до развертывания интерпретируемых бизнес-решений.

**7.3. Вывод**

Разработанное решение не только соответствует всем формальным критериям аттестации, но и имеет практическую ценность. Модель может быть использована в качестве инструмента поддержки принятия врачебных решений для раннего выявления пациентов с высоким риском сердечной недостаточности. Проект наглядно демонстрирует, что современные low-code платформы позволяют эффективно решать сложные аналитические задачи, обеспечивая при этом скорость разработки, наглядность результатов и легкость их интеграции.