МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

Аналитическая записка

исследовательского проекта на тему

**“Предсказание наличия метаболического синдрома”**

Выполнил: Зайнулин Владислав Сергеевич

Томск – 2024

**Аннотация**

Данная аналитическая записка представляет результаты исследования по предсказанию наличия метаболического синдрома. В работе проведен статистический анализ данных. На основе этих данных была построена базовая модель машинного обучения с использованием алгоритма CatBoost для предсказания наличия метаболического синдрома.

**Проблема**

Метаболический синдром - сложное состояние, характеризующееся набором метаболических нарушений, включая ожирение, гипертонию, нарушенный обмен глюкозы и дислипидемию. Эти факторы совместно увеличивают риск развития сердечно-сосудистых заболеваний и сахарного диабета второго типа.

Однако, несмотря на широкое распространение метаболического синдрома, его раннее обнаружение и прогнозирование остаются сложной задачей. Существующие методы диагностики, такие как измерение индекса массы тела (ИМТ), уровня глюкозы в крови натощак и артериального давления, могут быть недостаточно точными или не позволяют предсказать развитие метаболического синдрома до появления клинических симптомов.

Таким образом, ключевой проблемой, которую адресует данный исследовательский проект, является необходимость разработки более точных методов прогнозирования метаболического синдрома.

**Цель**

Целью данного исследования является анализ демографических, клинических и лабораторных данных пациентов с целью выявления основных признаков и зависимостей, связанных с наличием метаболического синдрома. Используя статистические методы анализа данных, выявить важные показатели, которые могут быть связаны с риском развития метаболического синдрома. Это позволит создать основу для дальнейших исследований и разработки эффективных методов диагностики.

**Источник данных**

Для выполнения исследования и построения модели предсказания наличия метаболического синдрома были использованы данные, доступные на платформе Kaggle.

Kaggle — это крупнейшая в мире платформа для соревнований по анализу данных (Data Science) и машинному обучению, где специалисты по анализу данных и машинному обучению со всего мира могут делиться, обсуждать и соревноваться в решении различных задач.

Для данного исследования был выбран набор данных, содержащий информацию о пациентах, включая различные демографические, клинические и лабораторные данные, а также наличие или отсутствие метаболического синдрома.

Важно отметить, что данные были предоставлены исключительно для научных целей и анонимизированы в соответствии с законодательством о защите персональных данных. Каждый пациент в наборе данных имеет уникальный идентификатор, и конфиденциальная информация о них не была доступна исследователям.

Исходные данные содержат следующие признаки:

* seqn - Порядковый идентификационный номер
* Age - Возраст человека
* Sex - Пол человека
* Marital - Семейное положение человека
* Income - Уровень дохода или информация о доходах
* Race - Этническое или расовое происхождение человека
* WaistCirc - Измерение обхвата талии
* BMI - Индекс массы тела - показатель состава тела
* Albuminuria - Измерение, связанное с содержанием альбумина в моче
* UrAlbCr - Соотношение альбумина и креатинина в моче
* UricAcid - Уровень мочевой кислоты в крови
* BloodGlucose - Уровень глюкозы в крови - показатель риска развития диабета
* HDL - Уровень холестерина липопротеинов высокой плотности ("хорошего" холестерина)
* Triglycerides - Уровень триглицеридов в крови
* MetabolicSyndrome - наличие или отсутствие метаболического синдрома у человека

Источник данных: https://www.kaggle.com/datasets/antimoni/metabolic-syndrome/data

**Анализ данных**

***Первичный анализ***

Прежде чем приступить к анализу конкретных зависимостей и построению моделей, важно изучить свойства и особенности данных. Для этого проведен первичный анализ структуры датасета.

Количественные данные включают в себя такие признаки, как "seqn", "Age", "Income", "WaistCirc", "BMI", "Albuminuria", "UrAlbCr", "UricAcid", "BloodGlucose", "HDL" и "Triglycerides".

Категориальные данные включают в себя такие признаки, как "Sex", "Marital" и "Race".

Категориальные признаки включают в себя следующие значения:

Sex (пол человека):

* Male (мужской)
* Female (женский)

Marital (семейное положение):

* Divorced (Разведен(а))
* Married (Женат(а))
* Separated (Раздельно проживающий(ая))
* Single (Не женат(а))
* Widowed (Овдовевший(ая))
* nan (Неизвестно)

Race (этническое или расовое происхождение):

* Asian (Азиат(ка))
* Black (Афроамериканец(ка))
* Hispanic (Испанец(ка))
* MexAmerican (Мексиканец(ка))
* Other (Другой)
* White (Белый)

Целевая переменная – MetabolicSyndrome - бинарная (0: отсутствие метаболического синдрома; 1: наличие метаболического синдрома).

Были выявлены пропуски в следующих столбцах:

* "Marital": 208 пропусков
* "Income": 117 пропусков
* "WaistCirc": 85 пропусков
* "BMI": 26 пропусков

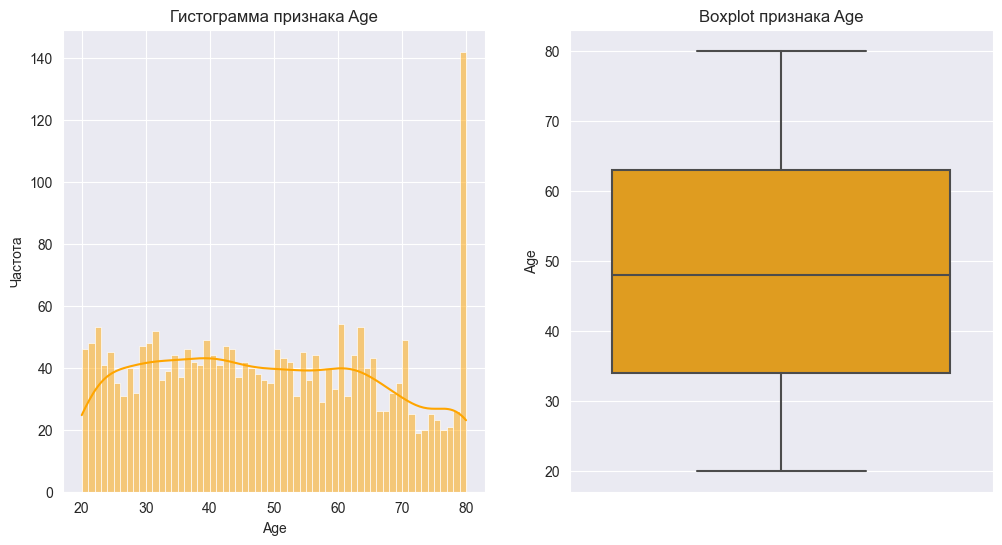
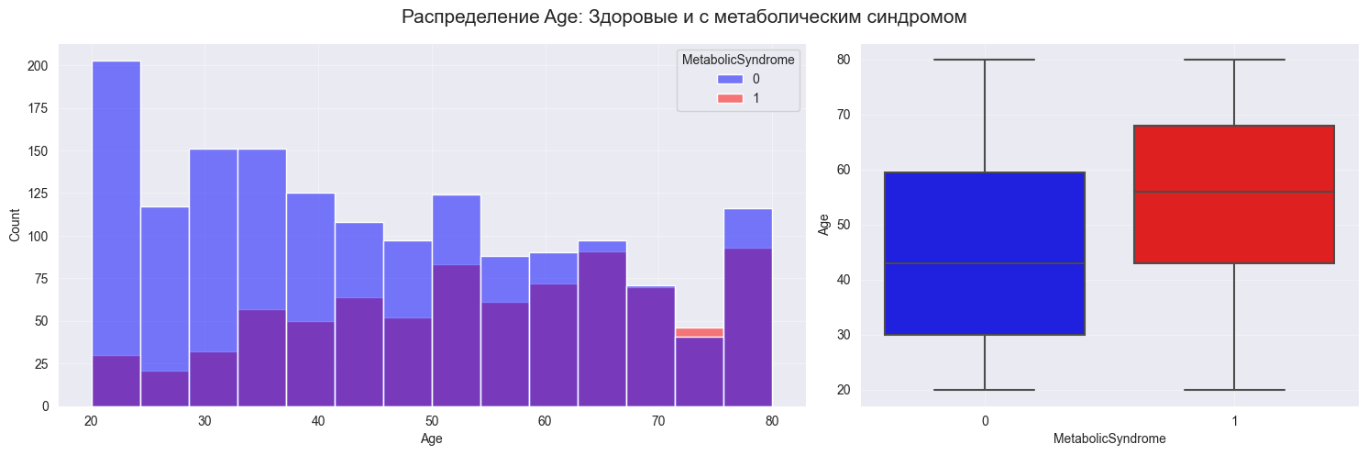
Не было обнаружено дубликатов в датасете.

В ходе анализа было выявлено, что у целевого признака наблюдается заметный дисбаланс классов. Дисбаланс классов может оказать существенное влияние на процесс анализа данных и построение моделей прогнозирования.

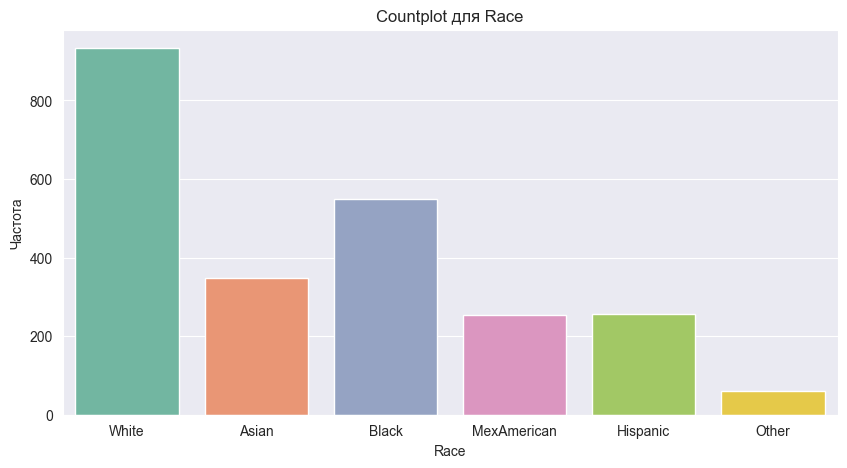
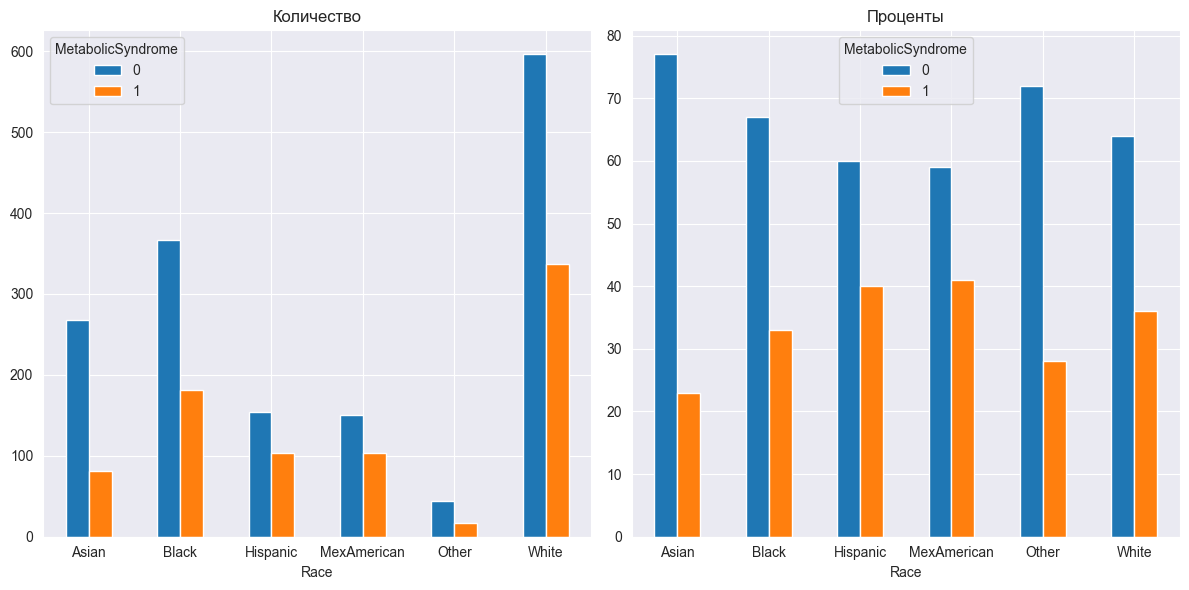
***Анализ взаимосвязи между признаками***

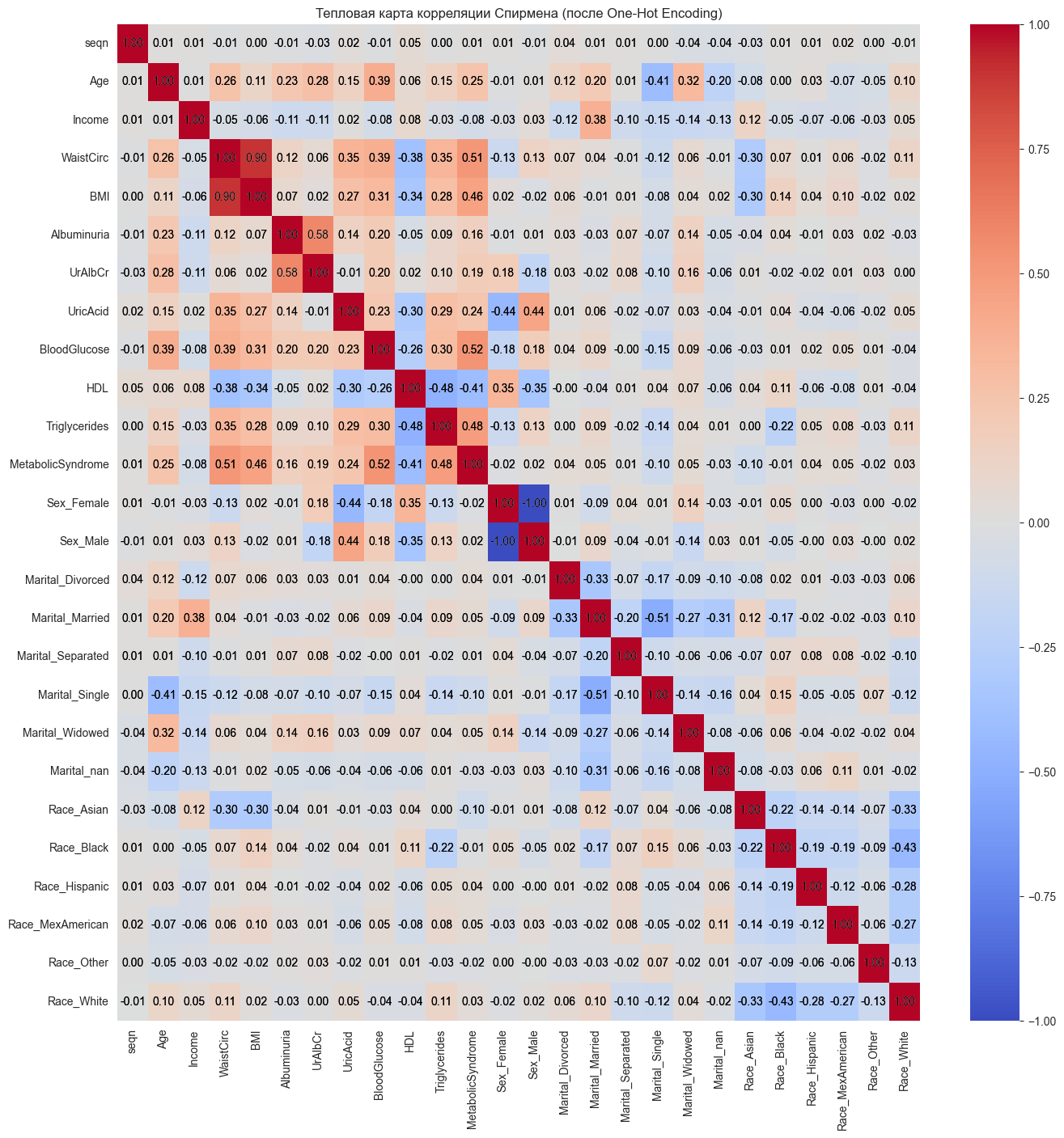
Для более глубокого понимания метаболического синдрома необходимо проанализировать взаимосвязи между различными признаками и целевой переменной - наличием метаболического синдрома у пациентов.

Для проведения анализа взаимосвязей между количественными признаками и целевым, были выполнены следующие шаги:

1. Построение гистограммы и boxplot признака: для каждого количественного признака были построены гистограмма и boxplot, чтобы визуально оценить их распределение и наличие выбросов.
2. Построение гистограммы и boxplot для каждого класса целевого признака: для более детального анализа были построены гистограммы и boxplot для каждого класса целевого признака, сравнивая распределение количественных признаков в каждой из категорий.
3. Основные характеристики признака: для каждого количественного признака была проведена описательная статистика с помощью метода describe, что позволило получить основные числовые характеристики данных.
4. Проверка на нормальность данных: для каждого количественного признака была выполнена проверка на нормальность распределения с помощью теста Шапиро-Уилка как для всего признака, так и для значений каждого класса целевого признака.
5. Проверка равенства дисперсий: была выполнена проверка на равенство дисперсий между группами с помощью теста Левена.
6. Проверка предположения о зависимости с целевым признаком: для определения наличия статистически значимой разницы между группами с разными значениями целевого признака, была использована непараметрическая проверка гипотезы с помощью теста Манна-Уитни.
7. Проверка корреляции с целевым признаком: для оценки силы и направления связи между количественными признаками и целевым признаком был проведен анализ корреляции с использованием коэффициента корреляции Пирсона, Спирмена и Кендалла.

Для анализа взаимосвязей между категориальными признаками и целевым, были выполнены следующие этапы:

1. Построение countplot: для каждой категориальной переменной был построен countplot, который отображает количество наблюдений в каждой категории.
2. Построение countplot для каждого класса целевого признака: для дальнейшего анализа были построены countplot для каждой категориальной переменной, разделенные по классам целевого признака, как в количественном, так и в процентном выражении, чтобы оценить распределение категорий внутри каждого класса целевого признака.
3. Проверка на связь с помощью Хи-квадрат и коэффициента Крамера: для оценки статистической значимости взаимосвязи между категориальной переменной и целевым признаком был применен тест Хи-квадрат. После этого был рассчитан коэффициент Крамера, который позволяет оценить силу связи между переменными.

Для дальнейшего анализа взаимосвязей между признаками была построена тепловая карта корреляции с использованием коэффициента корреляции Спирмена. После применения метода кодирования One-Hot Encoding для обработки категориальных переменных, была получена матрица, отражающая степень корреляции между всеми парами признаков.

Использование коэффициента корреляции Спирмена обусловлено тем, что ни один из признаков не имел нормального распределения, что делает использование коэффициента корреляции Пирсона некорректным. Коэффициент корреляции Спирмена позволяет оценить степень монотонной связи между признаками, не требуя нормальности распределения данных.

Тепловая карта корреляции помогла выявить как сильные, так и слабые связи между признаками. Обнаруженные высокие значения корреляции между определенными признаками могут указывать на наличие мультиколлинеарности, что может привести к нестабильности модели.

Интересным является наблюдение, что признаки Sex\_Male и Sex\_Female коррелируют с такими признаками, как WaistCirc, UrAlbCr, BloodGlucose, HDL, UricAcid и Triglycerides. Анализ графиков boxplot показал, что для мужчин и женщин значения этих признаков в среднем различаются. Такие различия могут иметь важное значение для прогнозирования наличия метаболического синдрома.

Хотя признак Sex сам по себе не имеет статистически значимой ассоциации с целевой переменной, его комбинация с другими признаками может оказать существенное влияние на прогнозирование. Это указывает на необходимость учета взаимодействия между признаками при построении модели, а также на потенциальную важность признака Sex в общем контексте моделирования метаболического синдрома.

Таким образом, данные методы анализа помогли выявить значимые взаимосвязи между признаками.

**Модель для предсказаний**

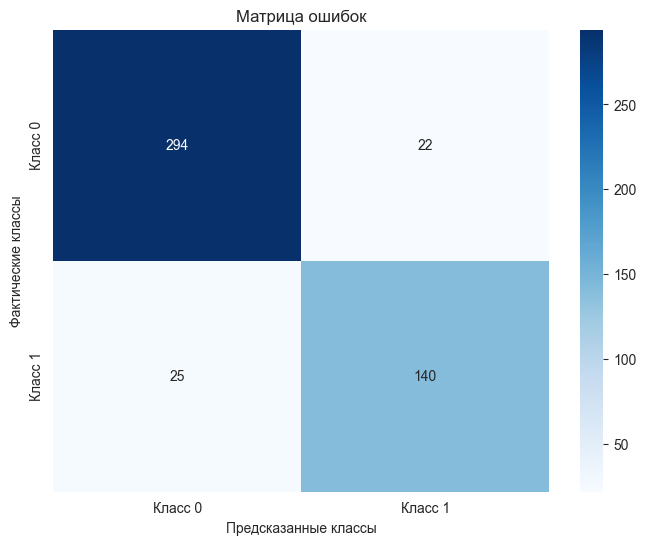
После проведения анализа данных, было построение модели для прогнозирования наличия метаболического синдрома. В качестве признаков для модели были использованы все доступные признаки, за исключением признака seqn, так как порядковый номер не влияет на целевую переменную.

Для создания модели был выбран алгоритм CatBoostClassifier, который является мощным инструментом для решения задач классификации с категориальными признаками. Были использованы стандартные параметры данного классификатора.

После завершения процесса обучения и тестирования модели были получены следующие результаты:

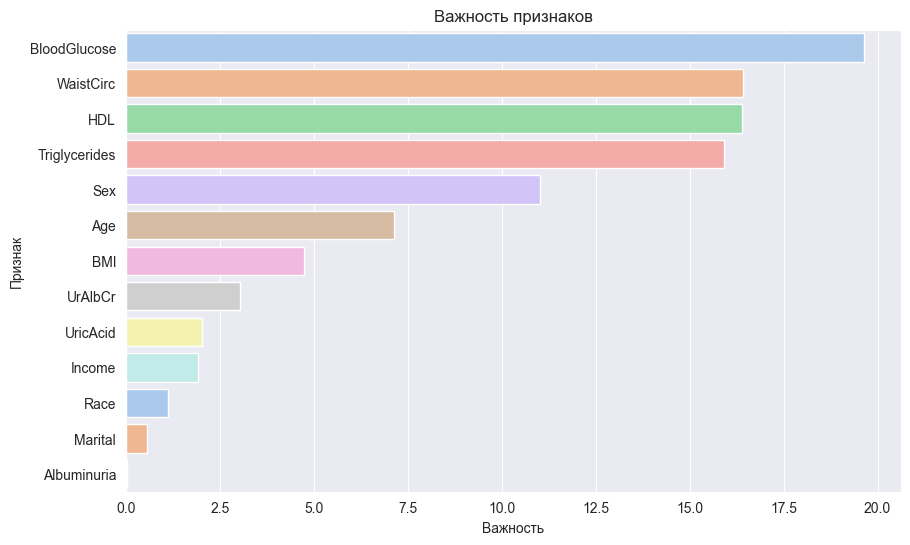
* Accuracy: 0.90 - указывает на общую точность модели в правильном определении классов.
* Precision: 0.86 - показывает долю истинно положительных результатов среди всех предсказанных положительных результатов.
* Recall: 0.85 - отражает долю истинно положительных результатов среди всех действительно положительных случаев.
* F1 Score: 0.90 - является гармоническим средним между Precision и Recall и представляет собой общую оценку точности модели.
* AUC-ROC Score: 0.98 - оценивает способность модели различать между классами.

Полученные метрики свидетельствуют о высоком качестве модели. Высокие значения Accuracy, Precision, Recall, F1 Score и AUC-ROC Score указывают на то, что модель хорошо справляется с задачей прогнозирования наличия метаболического синдрома на основе доступных признаков.

Также была построена матрица ошибок, которая позволяет оценить производительность модели более детально, разделяя предсказанные классы на истинно положительные (true positive), ложно положительные (false positive), истинно отрицательные (true negative) и ложно отрицательные (false negative) случаи.

Из матрицы ошибок видно, что модель сделала 294 правильных предсказаний относительно отсутствия метаболического синдрома (true negatives), 140 правильных предсказаний относительно наличия метаболического синдрома (true positives), 22 случая ложного предсказания наличия метаболического синдрома (false positives) и 25 случаев ложного предсказания отсутствия метаболического синдрома (false negatives).

Анализ матрицы ошибок позволяет лучше понять, в каких случаях модель ошибается и в каких случаях ее предсказания наиболее точны. Эта информация может быть использована для дальнейшего улучшения модели и оптимизации ее производительности.

И, конечно, были определены важности признаков. Результаты показали, что наибольшее влияние на прогнозирование наличия метаболического синдрома оказывают уровень глюкозы в крови (BloodGlucose), окружность талии (WaistCirc), уровень холестерина высокой плотности (HDL) и уровень триглицеридов (Triglycerides). Пол (Sex) и возраст (Age) также оказывают существенное влияние на прогнозирование. Другие признаки, такие как индекс массы тела (BMI) и уровень дохода (Income), также имеют некоторое влияние, но менее значимо. Раса (Race), семейное положение (Marital) и уровень альбуминурии (Albuminuria) оказывают наименьшее влияние на прогнозирование.

Эти результаты помогают лучше понять, какие признаки следует учитывать при дальнейшем улучшении модели и оптимизации ее производительности.

**Итоги**

В данной работе был проведен статистический анализ признаков, а также построена базовая модель для прогнозирования наличия метаболического синдрома. Результаты анализа позволили выявить важные признаки, которые оказывают влияние на предсказания модели.

Для улучшения результатов в будущем можно рассмотреть несколько направлений:

1. Подбор лучших гиперпараметров модели: тюнинг гиперпараметров может значительно улучшить производительность модели и ее способность к обобщению.
2. Обработка выбросов: устранение выбросов в данных может помочь улучшить качество модели и сделать ее более устойчивой к шумам в данных.
3. Использование более сложных моделей: возможно, стоит рассмотреть использование более сложных алгоритмов машинного обучения или их ансамблей для достижения лучших результатов.
4. Дополнительный анализ данных: в дальнейшем исследовании можно также углубиться в анализ данных, рассмотреть новые признаки или выполнить дополнительные эксперименты для более полного понимания проблемы и улучшения модели.

Эти шаги помогут не только улучшить текущую модель, но и создать более точные и надежные инструменты для прогнозирования наличия метаболического синдрома, что может иметь важное значение для клинической практики и здравоохранения.