**Lily Nurliana Abdullah** – PhD, Associate Professor, Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia; e-mail: liyana@upm.edu.my. ORCID: https://orcid.org/0000-0001-8704-2390.

*Поступила в редакцию 01.06.2026*
*Поступила после доработки 31.08.2026*
*Принята к публикации 02.09.2026*

https://doi.org/10.53360/2788-7995-2026-3(23)-26

МРНТИ: 76.01.85

**А.К. Үкібай[1], Е.А. Дайнеко[2], А.Б. Ақбаян[3], Д.С. Баймагамбетова[2], Г.Б. Бекешова[4*]**
[1]Филиал Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ» в городе Алматы, проспект аль-Фараби, 71, корпус 12, Алматы, 050040,
[2]Satbayev University, ул. Сатпаева, 22а, Алматы, 050013, Казахстан,
[3]Narxoz University, ул. Жандосова, 55, Алматы, 050035, Казахстан,
[4]Евразийский национальный университет им. Л.Н. Гумилева,
010000 Республика Казахстан, г. Астана, ул. Сатпаева, 2
*e-mail: gulvirabauyrzhanovna@gmail.com.

# РАЗРАБОТКА ВЕБ-ПРИЛОЖЕНИЯ ДЛЯ ПРЕДВАРИТЕЛЬНОЙ ОЦЕНКИ РИСКА САХАРНОГО ДИАБЕТА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

***Аннотация:*** *Сахарный диабет является одним из наиболее распространенных хронических неинфекционных заболеваний и требует раннего выявления лиц с повышенным риском. В данной статье представлена разработка веб-приложения для предварительной оценки риска сахарного диабета на основе симптоматических и демографических данных пользователя. Система реализована на языке Python с использованием фреймворка FastAPI и модуля машинного обучения, обученного на открытом наборе данных UCI Early Stage Diabetes Risk Prediction Dataset (520 наблюдений, 16 признаков). Сравнение классификаторов Logistic Regression, Support Vector Machine и Random Forest выполнено с использованием повторной стратифицированной 10-кратной кросс-валидации (5 повторов, 50 оценок для каждой модели). Наилучшие и наиболее устойчивые показатели получены для модели Random Forest: accuracy 98.27 ± 1.60%, F1-мера 98.59 ± 1.31%, ROC-AUC 99.93 ± 0.15%. Преимущество Random Forest над Support Vector Machine и Logistic Regression подтверждено статистически (скорректированный парный t-критерий, 5×2cv t-критерий и критерий знаковых рангов Уилкоксона, p < 0.05). Выполнен анализ вклада признаков тремя независимыми методами (значимость по критерию Джини, перестановочная значимость и значения SHAP), согласованно выделившими полиурию, полидипсию, пол и возраст как наиболее информативные признаки. На основе полученного ранжирования построен сокращенный опросник из 9 признаков, для которого статистически значимого снижения качества классификации по сравнению с полным набором из 16 признаков не выявлено (accuracy 97.62 ± 1.87%, p = 0.246), что позволяет сократить объем пользовательского ввода на 43.8%. Итоговая модель интегрирована в веб-приложение, отображающее уровень риска, визуальный индикатор и блок профилактических рекомендаций. Разработанное приложение следует рассматривать как программный прототип для предварительной самооценки и повышения осведомленности пользователя; оно не предназначено для постановки диагноза и требует дальнейшей внешней клинической валидации на независимых данных.*

***Ключевые слова:*** *сахарный диабет; машинное обучение; веб-приложение; FastAPI; оценка риска; кросс-валидация; статистическая проверка гипотез; значимость признаков; SHAP; отбор признаков; scikit-learn; ранняя диагностика.*

## Введение

Сахарный диабет (СД) относится к числу наиболее распространенных хронических неинфекционных заболеваний и представляет собой одну из ключевых глобальных медико-

©А.К. Үкібай, Е.А. Дайнеко, А.Б. Ақбаян, Д.С. Баймагамбетова, Г.Б. Бекешова, 2026

социальных проблем. По данным Международной диабетической федерации (IDF), в 2025 году количество взрослых пациентов с диабетом во всем мире составило около 589 миллионов человек, а к 2050 году ожидается рост данного показателя до более чем 853 миллионов [1]. Наиболее распространенной формой заболевания является сахарный диабет 2 типа (СД2), на долю которого приходится около 90-95% случаев. СД2 характеризуется постепенным развитием и длительным бессимптомным течением, что нередко приводит к позднему выявлению заболевания и развитию осложнений.

Раннее выявление лиц с повышенным риском развития диабета является важной задачей современной системы здравоохранения. Существенная часть факторов риска СД2 связана с образом жизни и включает низкую физическую активность, нарушение питания, ожирение, хронический стресс и нарушения сна [2]. Использование доступных неинвазивных инструментов предварительного скрининга может способствовать более раннему изменению поведенческих факторов и снижению вероятности развития осложнений [3].

В последние годы методы машинного обучения (ML) и цифровые медицинские платформы активно применяются для решения задач прогнозирования и поддержки принятия решений в здравоохранении [4, 5]. В области диабета значительное число исследований посвящено построению классификационных моделей на основе медицинских наборов данных [6]. Однако многие существующие работы ограничиваются сравнением алгоритмов машинного обучения без последующей интеграции моделей в полнофункциональные пользовательские системы. Кроме того, значительная часть существующих цифровых решений ориентирована преимущественно на пациентов с уже установленным диагнозом, а не на предварительную оценку риска [7, 8].

Применение методов машинного обучения для прогнозирования и предварительной оценки риска сахарного диабета активно исследуется в последние годы. Наиболее распространенные подходы основаны на использовании классификационных алгоритмов, обучаемых на медицинских наборах данных, содержащих симптоматические, демографические и клинические признаки пациентов. Среди наиболее часто используемых алгоритмов выделяются Logistic Regression, Support Vector Machine, Decision Tree и Random Forest.

A. Mujumdar и V. Vaidehi выполнили сравнительный анализ нескольких алгоритмов машинного обучения на наборе данных PIMA Indians Diabetes Database и показали, что Random Forest обеспечивает наиболее высокую точность классификации благодаря ансамблевому механизму принятия решений и снижению чувствительности к шуму данных [9]. R. Krishnamoorthi et al. предложили гибридный подход к отбору признаков, сочетающий методы фильтрации и обертки, что позволило повысить качество классификации и сократить размерность входных данных [10]. В ряде последующих исследований также отмечается высокая эффективность методов машинного обучения при анализе симптоматических и клинических признаков диабета [11, 12].

Одновременно с развитием алгоритмов машинного обучения активно развиваются цифровые медицинские платформы, ориентированные на автоматизацию мониторинга и предварительной оценки состояния пациентов. A. Rghioui et al. предложили систему мониторинга диабета на основе интернета вещей (IoT), объединяющую датчики и облачные ML-сервисы для анализа состояния пользователя в реальном времени [13]. Однако подобные решения требуют специализированного оборудования и устойчивой аппаратной инфраструктуры. Это ограничивает их применение в условиях ограниченных ресурсов и массового пользовательского доступа.

В исследованиях I. Dankwa-Mullan et al. и G. Fagherazzi et al. подчеркивается, что эффективность систем искусственного интеллекта в здравоохранении определяется не только качеством алгоритмов, но и доступностью интерфейсов, удобством взаимодействия и возможностью интеграции в повседневные пользовательские сценарии [14, 15]. Современные подходы к цифровому здравоохранению рассматривают ML-модели как часть более широких интеллектуальных систем поддержки принятия решений, включающих обработку пользовательских данных, визуализацию результатов и механизмы взаимодействия с пользователем [5, 16, 17]. Дополнительное внимание в современных исследованиях

уделяется вопросам интерпретируемости, надежности и практической интеграции AI-систем в клинические и пользовательские сценарии [18, 19].

Несмотря на значительное количество исследований в области прогнозирования диабета, многие существующие работы сосредоточены преимущественно на сравнении алгоритмов и анализе точности классификации на открытых наборах данных. При этом существенно меньше внимания уделяется вопросам практической интеграции моделей машинного обучения в легковесные пользовательские веб-системы, пригодные для быстрого предварительного скрининга без использования специализированного оборудования. Кроме того, современные исследования подчеркивают важность прозрачности и интерпретируемости медицинских AI-систем, особенно в задачах поддержки принятия решений и взаимодействия с пользователем [20].

В данной работе представлено веб-приложение для предварительной оценки риска сахарного диабета на основе симптоматических и демографических данных пользователя. Разработанная система объединяет серверную архитектуру на базе Python/FastAPI, модуль машинного обучения и браузерный пользовательский интерфейс для выполнения быстрой предварительной оценки риска без необходимости лабораторных исследований или специализированного оборудования. Такой подход позволяет рассматривать систему как легковесный программный прототип, ориентированный на применение в условиях ограниченной инфраструктуры и сценариях удаленного предварительного скрининга.

Цель работы – разработка и экспериментальное обоснование воспроизводимой программной системы предварительной оценки риска сахарного диабета по симптоматическим и демографическим признакам, включая определение минимального набора признаков, достаточного для сохранения качества классификации. Для достижения указанной цели решались следующие задачи: построение воспроизводимого конвейера предварительной обработки данных и обучения моделей машинного обучения на открытом наборе данных UCI Early Stage Diabetes Risk Prediction Dataset; сравнительная оценка классификаторов Logistic Regression, Support Vector Machine и Random Forest с использованием повторной стратифицированной 10-кратной кросс-валидации; статистическая проверка значимости различий между классификаторами; анализ вклада отдельных признаков в прогноз методами Gini importance, перестановочной значимости и SHAP; построение сокращенного симптоматического опросника и проверка сохранения качества классификации при его использовании; интеграция итоговой модели в веб-приложение и определение границ применимости полученного решения.

Основные особенности разработанной системы включают:

– воспроизводимый конвейер предварительной обработки данных, обучения и выбора модели машинного обучения;

– протокол оценки качества, основанный на повторной стратифицированной 10-кратной кросс-валидации и статистической проверке значимости различий между моделями;

– анализ вклада признаков тремя независимыми методами (значимость по критерию Джини, перестановочная значимость, значения SHAP) и построенный на его основе сокращенный симптоматический опросник из 9 признаков вместо 16;

– сравнительную оценку классификаторов Logistic Regression, Support Vector Machine и Random Forest на наборе данных UCI Early Stage Diabetes Risk Prediction Dataset (520 наблюдений);

– интеграцию итоговой модели в развертываемое веб-приложение на основе Python/FastAPI с визуальным представлением результата оценки риска и генерацией профилактических рекомендаций для пользователя.

Научная новизна работы состоит не в самом факте объединения модели машинного обучения и веб-сервиса: интеграция ML-моделей в веб-приложения на основе REST API является известным инженерным подходом и неоднократно описывалась в литературе, поэтому в настоящей работе она рассматривается как инженерная основа реализации, а не как научный результат. Новизна предлагаемого решения заключается в следующем.

Во-первых, предложена схема отбора симптоматических признаков, в которой ранжирование выполняется одновременно по трем независимым критериям – значимости по критерию Джини, перестановочной значимости и значениям SHAP, – а решение о сокращении набора принимается не по эвристически заданному порогу, а по результату статистической

проверки различия качества сокращенной и полной моделей. Применение предложенной схемы позволило уменьшить число признаков с 16 до 9 (на 43.8%) без статистически значимого снижения качества классификации ($p = 0.246$).

Во-вторых, полученный сокращенный набор признаков непосредственно определяет структуру пользовательской формы веб-приложения: сокращение опросника уменьшает время его заполнения и вероятность ошибок самооценки, что является практически значимым результатом именно для систем предварительного самоскрининга, работающих без участия медицинского персонала.

В-третьих, сравнение классификаторов выполнено не по единственному разбиению выборки, а с использованием повторной стратифицированной кросс-валидации и трех статистических критериев проверки значимости различий, что повышает достоверность вывода о преимуществе ансамблевой модели по сравнению с работами, ограничивающимися однократным разбиением данных.

Научно-практическая значимость работы заключается в создании воспроизводимого программного прототипа системы предварительного скрининга риска диабета, объединяющего сокращенный симптоматический опросник, статистически обоснованный выбор модели, механизм анализа вклада признаков и веб-интерфейс в рамках единой программной архитектуры. Разработанный подход может рассматриваться как основа для дальнейшего расширения системы, адаптации на локальных клинических данных и последующей внешней валидации.

**Условия и методы исследования**

Условия проведения исследования определяются характером используемых данных и постановкой задачи. Во-первых, используются исключительно симптоматические и демографические признаки, регистрируемые со слов пациента; лабораторные показатели (уровень глюкозы крови, HbA1c, результаты перорального глюкозотолерантного теста) в исследовании не применяются. Во-вторых, все данные получены в одном лечебном учреждении, что ограничивает демографическую и географическую репрезентативность выборки. В-третьих, независимая внешняя выборка и проспективное клиническое наблюдение отсутствуют, поэтому все приводимые оценки качества являются внутренними оценками модели на открытом наборе данных. В-четвертых, вычислительные эксперименты выполнялись в среде Python 3.11 с использованием библиотек scikit-learn, pandas, numpy, SciPy, SHAP и joblib при фиксированном значении генератора псевдослучайных чисел (random_state = 42), что обеспечивает воспроизводимость приведенных результатов.

Для обучения и оценки модели использовался открытый набор данных UCI Early Stage Diabetes Risk Prediction Dataset, содержащий 520 записей, сформированных на основе анкетирования пациентов Sylhet Diabetes Hospital (Bangladesh) [21]. Набор данных включает возраст, пол и 14 симптоматических признаков, связанных с ранними проявлениями сахарного диабета, включая полиурию, полидипсию, резкую потерю веса, слабость, полифагию, нарушение зрения, зуд кожи, раздражительность, замедленное заживление ран, частичный парез, мышечную скованность, алопецию и ожирение. Целевая переменная представлена бинарным классом, отражающим наличие или отсутствие риска диабета.

Распределение классов является умеренно несбалансированным: 320 наблюдений относятся к положительному классу и 200 – к отрицательному, что соответствует соотношению приблизительно 1.6:1. Предварительная обработка данных включала преобразование категориальных признаков в числовой формат с использованием методов кодирования категориальных переменных библиотеки scikit-learn. После обработки формировались матрица признаков X и целевой вектор y.

Кодирование признаков выполнялось по фиксированной схеме: бинарные симптоматические признаки, принимающие значения «да»/«нет», преобразовывались в значения 1 и 0; признак «пол» кодировался бинарно; числовой признак «возраст» подвергался стандартизации. Порядок признаков фиксировался и сохранялся вместе с моделью, что обеспечивает идентичность преобразования входных данных на этапе обучения и на этапе эксплуатации веб-приложения и исключает рассогласование структуры вектора признаков при обработке пользовательского запроса.

Для обеспечения корректности оценки качества модели набор данных был разделен на обучающую и тестовую выборки в соотношении 80/20 с использованием стратифицированного разбиения train_test_split, что позволило сохранить исходное распределение классов в обеих выборках.

Оценка качества моделей выполнялась в два этапа. На первом, основном этапе использовалась повторная стратифицированная k-кратная кросс-валидация при k = 10 и пяти независимых повторах, что дает 50 оценок качества для каждой модели. Стратификация обеспечивает сохранение исходного соотношения классов 1.6:1 в каждом фолде, а повторение процедуры с различными разбиениями снижает зависимость результата от конкретного разбиения выборки. Данный протокол выбран как существенно более надежный по сравнению с однократным разделением данных, которое при объеме выборки 520 наблюдений дает оценки качества с высокой дисперсией. На втором этапе выполнялась контрольная оценка на отложенной тестовой выборке, не участвовавшей в обучении.

Для проверки статистической значимости различий между классификаторами использовались три критерия. Скорректированный парный t-критерий применялся к 50 значениям метрик, полученным при повторной кросс-валидации; данный критерий учитывает зависимость между перекрывающимися обучающими подвыборками и корректирует оценку дисперсии, что снижает вероятность ошибки первого рода по сравнению с обычным парным t-критерием [26]. Дополнительно применялся парный 5×2cv t-критерий [25], рекомендованный для сравнения алгоритмов классификации на выборках ограниченного объема, и непараметрический критерий знаковых рангов Уилкоксона, не требующий предположения о нормальности распределения разностей [27]. Уровень значимости во всех проверках принимался равным 0.05.

Анализ вклада признаков выполнялся тремя независимыми методами. Первый метод – значимость признаков по критерию Джини, вычисляемая непосредственно в модели Random Forest как суммарное уменьшение неоднородности узлов при разбиении по данному признаку [29]. Второй метод – перестановочная значимость, определяемая как снижение F1-меры модели при случайной перестановке значений одного признака (50 перестановок для каждого признака); данный метод не зависит от внутренней структуры модели и позволяет проверить устойчивость полученного ранжирования. Третий метод – значения SHAP, вычисляемые алгоритмом TreeExplainer и позволяющие оценить как величину, так и направление вклада каждого признака в итоговый прогноз [28].

На основе согласованного ранжирования признаков строился сокращенный набор. Признаки последовательно включались в модель в порядке убывания значимости; для каждого набора из k признаков (k = 1, …, 16) выполнялась повторная стратифицированная 10-кратная кросс-валидация, после чего полученный результат сравнивался с результатом полной модели по скорректированному парному t-критерию. Минимально достаточным считался наименьший набор признаков, для которого различие с полной моделью не является статистически значимым (p > 0.05) [30].

Гиперпараметры моделей задавались следующим образом: для логистической регрессии – решатель lbfgs, максимальное число итераций 1000, параметр регуляризации C = 1.0; для метода опорных векторов – радиально-базисное ядро (RBF) с параметрами C = 1.0 и gamma = «scale»; для случайного леса – 300 деревьев решений без ограничения глубины с критерием Джини. Специальные методы искусственной балансировки классов не применялись, поскольку исходная степень несбалансированности выборки (примерно 1.6:1) является умеренной и учитывается стратификацией разбиений.

Следует учитывать, что используемый набор данных имеет ряд ограничений. Во-первых, данные были собраны в рамках одной клинической базы, что ограничивает демографическую и географическую репрезентативность выборки. Во-вторых, большинство признаков представлены в бинарной форме и основаны на субъективной оценке симптомов пользователями, что может приводить к различиям в интерпретации симптомов и ошибкам самооценки. Кроме того, объем выборки в 520 наблюдений достаточен для разработки исследовательского программного прототипа, однако недостаточен для подтверждения клинической универсальности модели. В связи с этим результаты настоящей работы следует рассматривать как внутреннюю техническую оценку модели на открытом наборе данных, а не как клинически подтвержденную систему диагностики диабета.

Обучение моделей выполнялось в рамках воспроизводимого конвейера машинного обучения. После разделения данных категориальные признаки преобразовывались в числовой формат, а этапы предварительной обработки и обучения модели выполнялись последовательно внутри единого ML-конвейера. Для предотвращения утечки данных масштабирование признаков выполнялось исключительно на обучающей выборке с последующим применением параметров преобразования к тестовым данным.

Для сравнительного анализа были выбраны три классификатора: Logistic Regression, Support Vector Machine (SVM) и Random Forest Classifier. Логистическая регрессия использовалась как базовая интерпретируемая модель, SVM – как нелинейный классификатор для разделения сложных границ между классами, а Random Forest – как ансамблевый алгоритм, устойчивый к шуму данных и нелинейным зависимостям между признаками.

Все модели были реализованы средствами библиотеки scikit-learn с использованием конвейера Pipeline, объединяющего этап предварительной обработки признаков и последующее обучение классификатора; такая организация исключает утечку информации из проверочных данных при выполнении кросс-валидации.

Качество моделей оценивалось с использованием метрик accuracy, precision, recall и F1-score. Для задач предварительного медицинского скрининга особое значение имеет показатель recall (sensitivity), поскольку модель должна минимизировать вероятность пропуска пользователей с потенциально высоким риском диабета.

После завершения сравнительной оценки модель с наилучшими показателями качества сохранялась для последующей интеграции в веб-приложение. Такой подход обеспечивает связь между этапом машинного обучения и практическим использованием модели внутри пользовательской системы предварительной оценки риска.

В таблице 1 представлены входные признаки, используемые моделью машинного обучения для прогнозирования риска сахарного диабета. Все признаки соответствуют параметрам, вводимым пользователем через веб-интерфейс системы.

Таблица 1 – Исходные признаки, используемые в модели прогнозирования риска диабета

| № | Наименование признака | Тип | Описание |
|---|---|---|---|
| 1 | Возраст | Числовой | Возраст пациента в годах (1-120) |
| 2 | Пол | категориальный признак | Биологический пол (мужской / женский) |
| 3 | Полиурия | Бинарный | Частое мочеиспускание |
| 4 | Полидипсия | Бинарный | Повышенная жажда |
| 5 | Резкая потеря веса | Бинарный | Необъяснимая быстрая потеря массы тела |
| 6 | Слабость | Бинарный | Общая слабость |
| 7 | Полифагия | Бинарный | Повышенный аппетит |
| 8 | Генитальный кандидоз | Бинарный | Грибковая инфекция половых органов |
| 9 | Нарушение зрения | Бинарный | Размытое или ухудшенное зрение |
| 10 | Зуд кожи | Бинарный | Кожный зуд |
| 11 | Раздражительность | Бинарный | Повышенная раздражительность |
| 12 | Замедленное заживление | Бинарный | Медленное заживление ран |
| 13 | Частичный парез | Бинарный | Частичная мышечная слабость |
| 14 | Мышечная скованность | Бинарный | Скованность мышц |
| 15 | Алопеция | Бинарный | Выпадение волос |
| 16 | Ожирение | Бинарный | Индекс массы тела > 30 кг/м² |

Разработанная система представляет собой легковесное веб-приложение для предварительной оценки риска сахарного диабета, реализованное на языке Python с использованием фреймворка FastAPI. Архитектура приложения построена по модульному принципу и объединяет веб-интерфейс, серверную логику обработки запросов, модуль предварительной обработки данных и модель машинного обучения в рамках единой программной системы. Использование FastAPI обеспечивает обработку HTTP-запросов и упрощает развертывание приложения на минимальной серверной инфраструктуре.

Общая архитектура системы и взаимодействие основных компонентов представлены на рисунке 1.

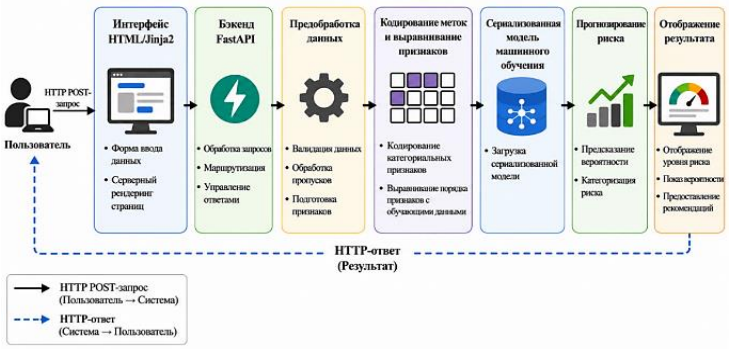



Рисунок 1 – Архитектура веб-приложения для предварительной оценки риска диабета

Представленная архитектура отражает последовательность обработки пользовательских данных внутри системы, начиная от ввода симптоматической информации через веб-интерфейс и заканчивая формированием итоговой оценки риска на основе ML-модели. Подобное разделение компонентов упрощает поддержку системы, обеспечивает воспроизводимость обработки данных и позволяет независимо модифицировать отдельные модули приложения без изменения общей логики работы системы.

Архитектура приложения включает несколько основных компонентов: модуль обработки пользовательских запросов, модуль предварительной обработки данных, ML-модель прогнозирования риска диабета и систему визуализации результатов. После отправки формы данные пользователя передаются на сервер через HTTP POST-запрос, где выполняется преобразование категориальных признаков, выравнивание структуры входных данных и подготовка вектора признаков для последующего прогнозирования. Для обеспечения воспроизводимости обработки используются сохраненные параметры кодирования признаков и фиксированный порядок входных переменных.

Модель машинного обучения загружается в память приложения в виде предварительно сериализованного объекта, что позволяет сократить время отклика системы и снизить вычислительные затраты при выполнении прогнозирования. После обработки входных данных модель вычисляет вероятность риска диабета, которая далее преобразуется в категорию риска и передается в пользовательский интерфейс вместе с текстовой интерпретацией результата и профилактическими рекомендациями.

Следует отметить, что используемые категории низкого, умеренного и высокого риска являются эвристическими границами визуальной интерпретации результата внутри программного прототипа. Они не представляют собой клинически валидированные диагностические пороги и не должны использоваться для постановки медицинского диагноза. Их основная задача заключается в упрощении восприятия результата пользователем и формировании рекомендации о необходимости дальнейшего медицинского обследования.

Общая схема процесса машинного обучения, включающая подготовку данных, обучение и оценку моделей, а также интеграцию итоговой модели в веб-приложение, представлена на рисунке 2.

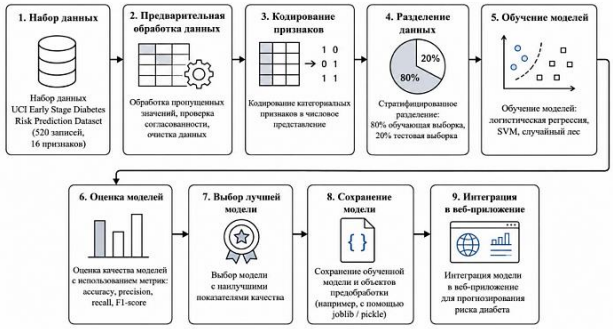



Рисунок 2 – Конвейер машинного обучения для обучения, оценки и интеграции модели в веб-приложение

**Результаты исследований**

Все эксперименты проводились с использованием Python 3.11 и библиотек scikit-learn, pandas, numpy, SciPy, SHAP и joblib. Оценка качества выполнялась по описанному выше двухэтапному протоколу: повторная стратифицированная 10-кратная кросс-валидация на всем наборе данных (50 оценок для каждой модели) и контрольная оценка на отложенной тестовой выборке, полученной стратифицированным разбиением в соотношении 80/20, что соответствует 416 обучающим и 104 тестовым наблюдениям. Дополнительное увеличение выборки и методы искусственной балансировки классов не применялись.

Результаты контрольной оценки обученных классификаторов на отложенной тестовой выборке представлены в таблице 2. Наилучшие показатели качества были достигнуты моделью Random Forest, обеспечившей accuracy 97.15% и F1-меру 97.42%; окончательный выбор модели для интеграции в веб-приложение выполнен по результатам кросс-валидации и статистической проверки значимости различий, приведенным ниже.

Таблица 2 – Сравнительная эффективность классификаторов на тестовом наборе

| Classifier | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| Logistic Regression | 88.55 | 87.92 | 91.20 | 89.52 |
| SVM | 92.37 | 91.81 | 93.73 | 92.72 |
| Random Forest | 97.15 | 96.83 | 97.94 | 97.42 |

Графическая сравнительная оценка качества рассмотренных классификаторов, включающая ROC-кривые и матрицы ошибок, представлена на рисунке 3.

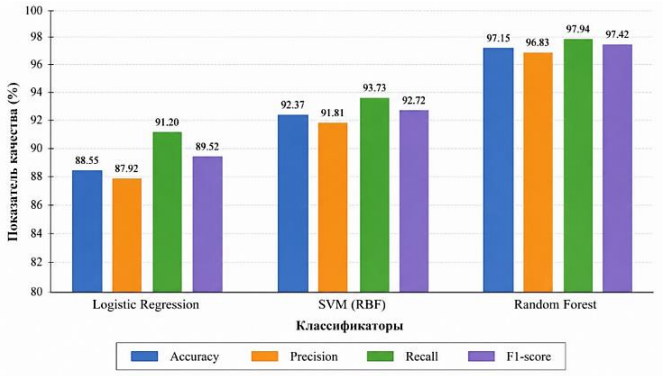



Рисунок 3 – Сравнительная оценка качества классификаторов машинного обучения

Оценка качества по единственному разбиению выборки обладает высокой дисперсией и не позволяет судить об устойчивости различий между моделями. Поэтому основная оценка выполнялась методом повторной стратифицированной 10-кратной кросс-валидации при пяти независимых повторах. Усредненные значения метрик и их стандартные отклонения приведены в таблице 3.

Таблица 3 – Результаты повторной стратифицированной 10-кратной кросс-валидации (среднее ± стандартное отклонение, %)

| **Классификатор** | **Accuracy** | **Precision** | **Recall** | **F1-мера** | **ROC-AUC** |
|---|---|---|---|---|---|
| Logistic Regression | 92.50 ± 3.67 | 94.49 ± 3.99 | 93.44 ± 4.39 | 93.87 ± 3.03 | 97.29 ± 1.82 |
| SVM (RBF) | 96.31 ± 2.46 | 97.60 ± 2.35 | 96.44 ± 3.42 | 96.97 ± 2.04 | 99.67 ± 0.37 |
| Random Forest | 98.27 ± 1.60 | 98.68 ± 1.90 | 98.56 ± 2.36 | 98.59 ± 1.31 | 99.93 ± 0.15 |

Модель Random Forest показала наиболее высокие значения всех рассмотренных метрик при наименьшем стандартном отклонении, что свидетельствует не только о более высоком среднем качестве, но и о большей устойчивости модели к изменению разбиения выборки. Особое значение для задач предварительного скрининга имеет показатель recall: его значение для Random Forest (98.56 ± 2.36%) превышает значения для SVM (96.44 ± 3.42%) и логистической регрессии (93.44 ± 4.39%), что соответствует меньшей доле пропущенных пользователей с повышенным риском.

Для проверки того, являются ли наблюдаемые различия статистически значимыми, а не следствием случайного разбиения данных, выполнялось попарное сравнение моделей по трем критериям. Результаты статистической проверки приведены в таблице 4.

Таблица 4 – Статистическая проверка значимости различий между классификаторами

| Сравниваемые модели | Разность accuracy, п.п. | Скорректированный парный t-критерий | 5×2cv t-критерий | Критерий Уилкоксона (F1-мера) |
|---|---|---|---|---|
| Random Forest – SVM | +1.96 | t = 2.48; p = 0.016 | t = 3.61; p = 0.015 | W = 56.5; p < 0.001 |
| Random Forest – Logistic Regression | +5.77 | t = 4.21; p < 0.001 | t = 5.01; p = 0.004 | W = 1.0; p < 0.001 |
| SVM – Logistic Regression | +3.81 | t = 2.83; p = 0.007 | – | W = 61.5; p < 0.001 |

Преимущество модели Random Forest над методом опорных векторов является статистически значимым на уровне 0.05 по всем трем примененным критериям (p = 0.016 по скорректированному парному t-критерию, p = 0.015 по 5×2cv t-критерию и p < 0.001 по критерию знаковых рангов Уилкоксона), а преимущество над логистической регрессией – значимым на уровне 0.01. Таким образом, выбор Random Forest в качестве основной модели системы имеет статистическое обоснование и не является следствием результата единственного разбиения выборки. Распределения значений F1-меры по 50 разбиениям для трех рассмотренных классификаторов представлены на рисунке 4.

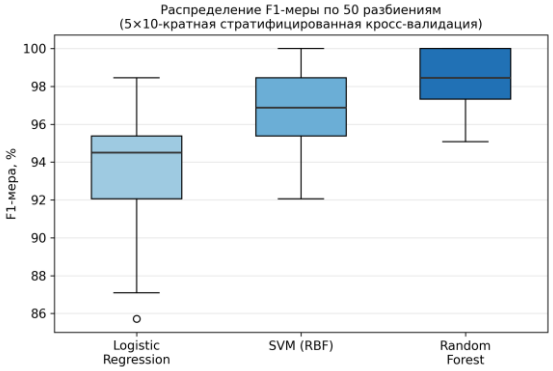



Рисунок 4 – Распределение F1-меры трех классификаторов по 50 разбиениям повторной стратифицированной 10-кратной кросс-валидации

Высокая эффективность алгоритма Random Forest связана с использованием ансамблевого подхода, основанного на объединении нескольких деревьев решений. Подобная архитектура позволяет снизить чувствительность модели к шуму данных и обеспечивает устойчивую обработку бинарных симптоматических признаков. В условиях рассматриваемого набора данных ансамблевый подход оказался более эффективным по сравнению с линейной моделью Logistic Regression и нелинейным классификатором SVM.

Логистическая регрессия продемонстрировала наиболее интерпретируемое, но менее гибкое поведение вследствие ограничения линейной зависимостью между признаками и целевой переменной. Модель SVM с RBF-ядром обеспечила более высокое качество классификации за счет способности учитывать нелинейные границы между классами, однако итоговые показатели оказались ниже результатов Random Forest.

Для определения того, какие именно признаки определяют решение модели, выполнялся анализ их вклада тремя независимыми методами. Результаты ранжирования признаков по критерию Джини и по перестановочной значимости приведены на рисунке 5, распределение значений SHAP – на рисунке 6.

Все три метода дают согласованное ранжирование признаков. Наибольший вклад в прогноз вносят полиурия и полидипсия: случайная перестановка значений этих признаков приводит к снижению F1-меры на 12.3 и 10.2 процентных пункта соответственно, тогда как для остальных признаков это снижение не превышает 4.3 процентных пункта. Далее по значимости следуют пол, возраст, резкая потеря веса и частичный парез.

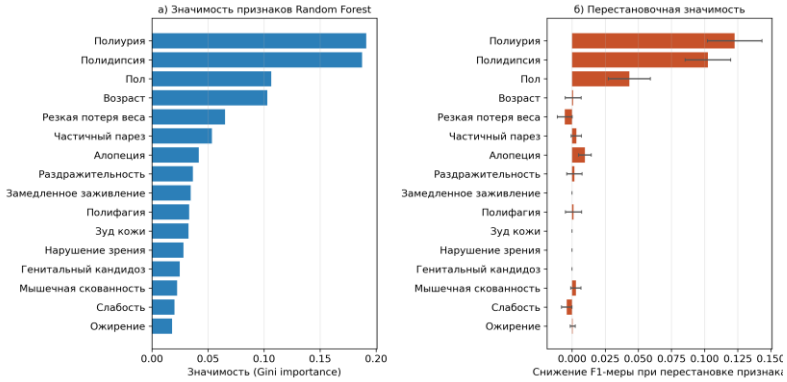



Рисунок 5 – Ранжирование признаков по значимости:
*а) значимость по критерию Джини; б) перестановочная значимость (снижение F1-меры при перестановке признака)*

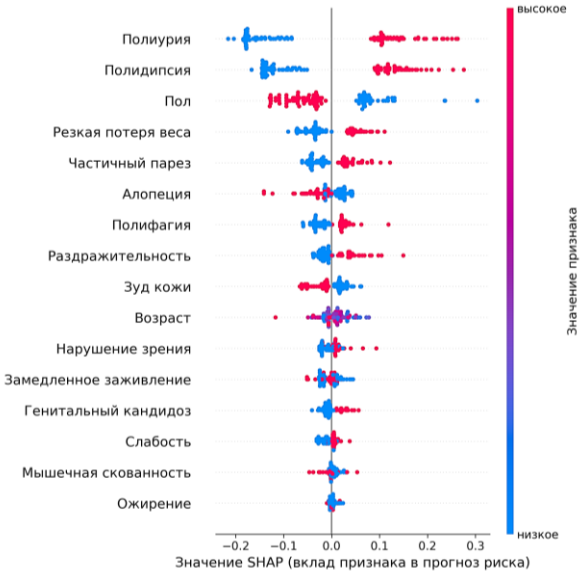



Рисунок 6 – Распределение значений SHAP для модели Random Forest: величина и направление вклада признаков в прогноз риска диабета

Полученный результат согласуется с клиническими представлениями о ранних проявлениях сахарного диабета, для которых полиурия и полидипсия являются классическими симптомами, что дополнительно подтверждает содержательную корректность построенной модели. Диаграмма значений SHAP показывает также направление вклада признаков: наличие полиурии и полидипсии увеличивает прогнозируемую вероятность риска, тогда как мужской пол и наличие алопеции в рассматриваемой выборке связаны с ее снижением. Последнее наблюдение следует интерпретировать с осторожностью, поскольку оно отражает статистические особенности состава конкретной выборки, а не универсальную биологическую закономерность.

Полученное ранжирование использовалось для построения сокращенного симптоматического опросника. Признаки последовательно включались в модель в порядке убывания значимости, и для каждого набора выполнялась повторная стратифицированная 10-кратная кросс-валидация с последующим сравнением с полной моделью из 16 признаков по скорректированному парному t-критерию. Результаты приведены в таблице 5 и на рисунке 7.

Как следует из таблицы 5, при числе признаков от одного до восьми качество сокращенной модели статистически значимо ниже качества полной модели ($p < 0.05$). Начиная с девяти признаков различие перестает быть статистически значимым ($p = 0.246$): сокращенная модель обеспечивает accuracy 97.62 ± 1.87% и F1-меру 98.07 ± 1.51% против 98.27 ± 1.60% и 98.59 ± 1.31% для полного набора.

Таблица 5 – Качество модели Random Forest при последовательном сокращении набора признаков

| k | Последний добавленный признак | Accuracy, % | F1-мера, % | ROC-AUC, % | p (сравнение с полным набором) |
|---|---|---|---|---|---|
| 1 | Полиурия | 82.31 ± 4.50 | 83.98 ± 4.45 | 84.22 | < 0.001 |
| 2 | Полидипсия | 86.92 ± 3.96 | 88.95 ± 3.50 | 90.56 | < 0.001 |
| 3 | Пол | 89.42 ± 2.75 | 91.79 ± 2.10 | 94.62 | < 0.001 |
| 4 | Возраст | 91.35 ± 2.86 | 92.77 ± 2.51 | 96.95 | < 0.001 |
| 5 | Резкая потеря веса | 93.81 ± 3.08 | 94.84 ± 2.63 | 97.97 | < 0.001 |
| 6 | Частичный парез | 93.92 ± 2.67 | 94.97 ± 2.27 | 98.35 | < 0.001 |
| 7 | Алопеция | 95.08 ± 2.58 | 95.98 ± 2.11 | 98.98 | 0.001 |
| 8 | Раздражительность | 96.85 ± 2.06 | 97.45 ± 1.65 | 99.33 | 0.014 |
| 9 | Замедленное заживление | 97.62 ± 1.87 | 98.07 ± 1.51 | 99.48 | 0.246 |
| 10 | Полифагия | 97.85 ± 1.66 | 98.24 ± 1.37 | 99.60 | 0.371 |
| 11 | Зуд кожи | 98.00 ± 1.72 | 98.37 ± 1.42 | 99.70 | 0.389 |
| 12 | Нарушение зрения | 97.92 ± 1.67 | 98.31 ± 1.37 | 99.90 | 0.390 |
| 13 | Генитальный кандидоз | 98.08 ± 1.80 | 98.43 ± 1.48 | 99.91 | 0.565 |
| 14 | Мышечная скованность | 98.19 ± 1.69 | 98.53 ± 1.39 | 99.92 | 0.820 |
| 15 | Слабость | 98.31 ± 1.62 | 98.62 ± 1.32 | 99.94 | 0.878 |
| 16 | Ожирение (полный набор) | 98.27 ± 1.60 | 98.59 ± 1.31 | 99.93 | – |

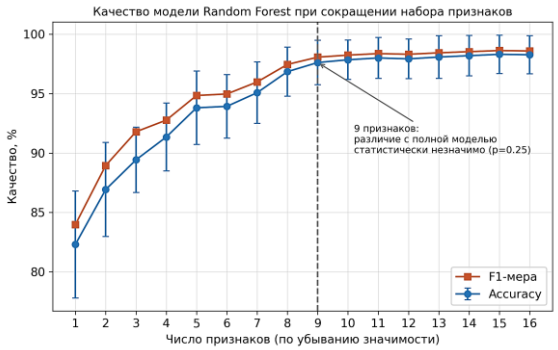



Рисунок 7 – Качество модели Random Forest при последовательном сокращении набора признаков (повторная стратифицированная 10-кратная кросс-валидация)

Таким образом, минимально достаточным является набор из девяти признаков – возраст, пол, полиурия, полидипсия, резкая потеря веса, частичный парез, алопеция, раздражительность и замедленное заживление ран, – что соответствует сокращению объема пользовательского ввода на 43.8% без статистически значимой потери качества классификации. Полученный результат имеет непосредственное практическое значение для систем самоскрининга, поскольку сокращение опросника уменьшает время его заполнения и снижает вероятность ошибок самооценки пользователя.

Следует отметить, что высокие значения accuracy и F1-score могут быть связаны не только с качеством алгоритма, но и с особенностями используемого набора данных. Большинство признаков представляют собой симптоматические показатели, непосредственно связанные с ранними проявлениями диабета, что формирует относительно высокую разделимость классов внутри рассматриваемой выборки. Аналогичные тенденции высокой эффективности моделей машинного обучения на symptom-based diabetes datasets ранее отмечались в работах A. Mujumdar и V. Vaidehi, а также R. Krishnamoorthi et al. [9,10].

Полученные результаты показывают, что использование методов машинного обучения позволяет эффективно выполнять предварительную оценку риска диабета на основе ограниченного набора симптоматических признаков. При этом разработанная система ориентирована не на клиническую постановку диагноза, а на реализацию легковесного программного инструмента предварительного скрининга и пользовательской оценки риска.

Главный интерфейс веб-приложения реализован в виде формы пользовательского ввода, предназначенной для предварительной оценки риска сахарного диабета. Взаимодействие пользователя с моделью машинного обучения выполняется через веб-браузер без установки дополнительного программного обеспечения. Форма содержит демографические и симптоматические признаки, описанные в таблице 1.

Возраст вводится пользователем через числовое поле, пол – посредством выпадающего списка, а остальные признаки представлены бинарными элементами выбора. Верхняя часть формы пользовательского ввода представлена на рисунке 8.

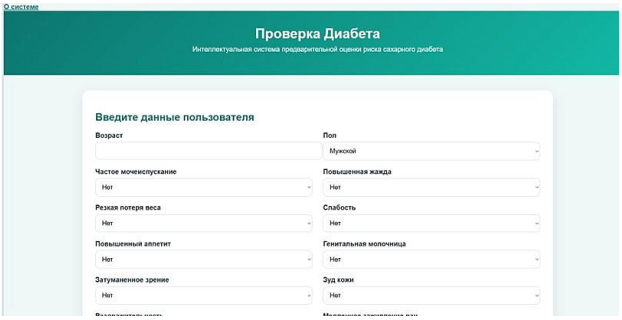



Рисунок 8 – Верхняя часть формы ввода пользовательских данных

Нижняя часть интерфейса содержит оставшиеся симптоматические признаки и кнопку запуска процедуры оценки риска (рис. 9).

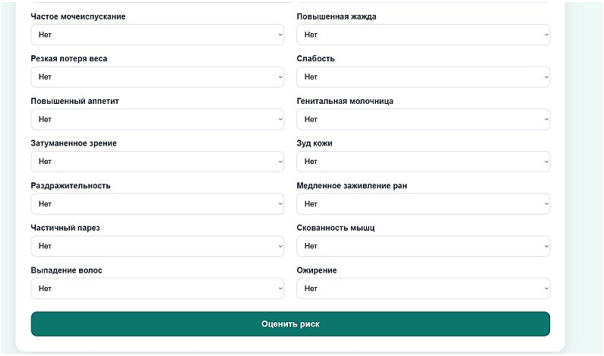



Рисунок 9 – Нижняя часть формы ввода данных и кнопка запуска оценки риска

Передача пользовательских данных выполняется посредством HTTP POST-запроса; на стороне клиента контролируются допустимый диапазон значений возраста и заполнение обязательных полей.

После отправки формы приложение формирует страницу результатов оценки риска, содержащую вероятность риска, категорию риска и визуальный индикатор уровня риска. Для повышения удобства восприятия используется цветовая индикация: зеленый цвет соответствует низкому риску, желтый – умеренному, красный – высокому. Пример страницы результатов представлен на рисунке 10.

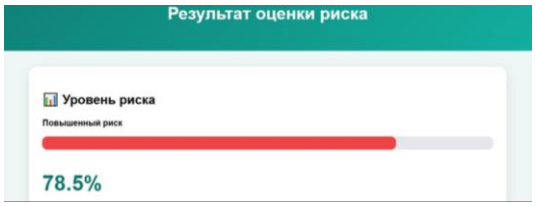



Рисунок 10 – Страница отображения результатов предварительной оценки риска диабета

В зависимости от рассчитанного уровня риска приложение формирует текстовые рекомендации, ориентированные на необходимость контроля состояния здоровья, коррекции образа жизни и обращения за медицинской консультацией. Рекомендации носят исключительно информационный характер и не предназначены для самостоятельной постановки диагноза или назначения лечения (рисунок 11).

Дополнительно приложение поддерживает механизм сохранения результатов внутри пользовательской сессии браузера, что позволяет просматривать ранее выполненные оценки риска без повторного ввода данных. Доступ к истории результатов осуществляется через маршрут /history.

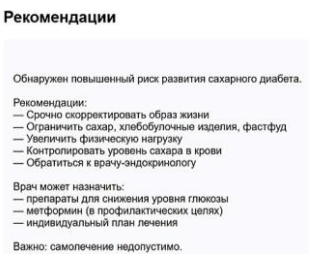



Рисунок 11 – Блок пользовательских рекомендаций для результата с повышенным риском диабета

При описании разработанной системы учитывались общие принципы прозрачной отчетности для медицинских моделей прогнозирования. В работе раскрыты источник данных, набор входных признаков, схема разбиения выборки, используемые классификаторы, метрики оценки и ограничения применимости модели. Вместе с тем представленная система не претендует на полное соответствие требованиям клинической валидации, поскольку не включает независимую внешнюю выборку, проспективное исследование и анализ влияния приложения на реальные решения пациента или врача. Эти элементы должны рассматриваться как обязательные этапы последующего развития системы [22].

**Обсуждение научных результатов**

Полученные результаты показывают, что методы машинного обучения способны эффективно выполнять предварительную оценку риска сахарного диабета на основе ограниченного набора симптоматических признаков. Наилучшие показатели качества были достигнуты классификатором Random Forest, обеспечившим accuracy 97.1% на тестовой выборке. Полученное значение находится в верхней части диапазона результатов, ранее опубликованных для рассматриваемого набора данных и сопоставимых symptom-based diabetes datasets [9, 10].

Оценка, полученная методом повторной стратифицированной 10-кратной кросс-валидации, подтверждает результат контрольной проверки на отложенной выборке: accuracy модели Random Forest составила 98.27 ± 1.60%, F1-мера – 98.59 ± 1.31%, ROC-AUC – 99.93 ± 0.15%. Различия с методом опорных векторов и логистической регрессией являются статистически значимыми, что позволяет обоснованно говорить о преимуществе ансамблевой модели на рассматриваемом наборе данных, а не о случайном превосходстве на отдельном разбиении.

Высокие показатели классификации во многом связаны с особенностями используемого набора данных. Большинство признаков представляют собой симптомы, непосредственно ассоциированные с ранними проявлениями диабета, что формирует относительно высокую разделимость классов. Дополнительным фактором является использование ансамблевого подхода Random Forest, обеспечивающего устойчивость модели к шуму данных и нелинейным взаимодействиям между признаками. По сравнению с Logistic Regression и SVM, ансамблевый классификатор продемонстрировал более стабильные результаты при работе с бинарными симптоматическими признаками.

Анализ вклада признаков показал, что решение модели определяется преимущественно небольшим числом симптоматических признаков. Доминирующая роль полиурии и полидипсии согласуется с клиническими представлениями о ранних проявлениях сахарного диабета и подтверждает содержательную интерпретируемость построенной модели. Практическим следствием этого наблюдения является возможность сокращения симптоматического опросника с 16 до 9 признаков без статистически значимой потери качества классификации, что снижает нагрузку на пользователя при самостоятельном заполнении формы и сокращает время получения предварительной оценки риска.

Вместе с тем полученный сокращенный набор признаков определен на одном открытом наборе данных и отражает статистические особенности конкретной выборки. Значимость отдельных признаков, в частности пола и алопеции, может быть обусловлена составом выборки Sylhet Diabetes Hospital и не должна интерпретироваться как установленная биологическая закономерность. Проверка устойчивости полученного ранжирования признаков и состава сокращенного опросника на независимых данных является необходимым этапом дальнейшего исследования.

Несмотря на высокие значения accuracy и F1-score, представленная система не предназначена для клинической диагностики диабета. Приложение ориентировано исключительно на предварительный пользовательский скрининг и должно рассматриваться как вспомогательный инструмент оценки риска. Результаты модели не заменяют лабораторные исследования, включая измерение уровня глюкозы крови, HbA1c и проведение перорального глюкозотолерантного теста. При получении высокого уровня риска пользователь должен рассматривать результат как основание для обращения к врачу и прохождения дополнительного обследования, а не как самостоятельное медицинское заключение [23].

Использование симптоматических данных, вводимых пользователем самостоятельно, приводит к ряду ограничений, связанных с субъективностью самооценки симптомов. Пользователи могут недооценивать отдельные признаки, ошибочно интерпретировать симптомы или вводить социально предпочтительные ответы. Подобные ограничения характерны для большинства неинвазивных опросниковых систем предварительной оценки риска и должны учитываться при интерпретации результатов.

С точки зрения пользовательского сценария разработанная система близка к неинвазивным инструментам оценки риска, основанным на анкетировании, включая шкалу FINDRISC. Однако между ними существует принципиальное различие. FINDRISC представляет собой клинически валидированную шкалу с фиксированной системой балльной оценки, тогда как представленное приложение использует модель машинного обучения, обученную на открытом наборе симптоматических данных. Поэтому разработанную систему следует рассматривать не как замену клинически валидированных шкал, а как программный прототип, демонстрирующий возможность интеграции симптоматического опросника, ML-модели и веб-интерфейса в единую систему предварительной оценки риска [24].

При разработке приложения учитывались рекомендации по ответственному использованию систем искусственного интеллекта в здравоохранении, согласно которым ML-инструменты должны дополнять клинические решения, а не заменять их [25,26]. В работе раскрыты источник данных, ограничения набора данных, схема оценки моделей и границы применимости разработанной системы. Вместе с тем приложение не прошло независимую внешнюю клиническую валидацию и не оценивалось в условиях реального медицинского использования. Эти направления следует рассматривать как ключевые этапы дальнейшего развития системы, включая расширение набора данных, тестирование на независимых выборках и последующую оценку клинической применимости модели [22].

**Заключение**

В данной работе разработано веб-приложение для предварительной оценки риска сахарного диабета на основе симптоматических и демографических данных пользователя. Система объединяет модель машинного обучения, серверную архитектуру на базе Python/FastAPI и пользовательский веб-интерфейс в рамках единого программного решения. Сравнение классификаторов Logistic Regression, Support Vector Machine и Random Forest выполнено на открытом наборе данных UCI Early Stage Diabetes Risk Prediction Dataset с использованием повторной стратифицированной 10-кратной кросс-валидации и статистической проверки значимости различий. Наилучшие и наиболее устойчивые результаты показала модель Random Forest (accuracy 98.27 ± 1.60%, F1-мера 98.59 ± 1.31%, ROC-AUC 99.93 ± 0.15%), преимущество которой над остальными рассмотренными классификаторами подтверждено статистически ($p < 0.05$).

Полученные результаты подтверждают техническую возможность интеграции методов машинного обучения в легковесные веб-системы предварительной оценки риска. При этом разработанное приложение следует рассматривать как исследовательский программный прототип, а не как клиническую диагностическую систему. Используемый набор данных обладает рядом ограничений, включая ограниченную выборку, отсутствие внешней клинической валидации и использование симптоматических признаков, основанных на самооценке пользователей. В связи с этим результаты модели не могут заменять лабораторную диагностику и профессиональное медицинское заключение.

Практическая значимость работы заключается в создании воспроизводимой программной системы, объединяющей сокращенный симптоматический опросник,

статистически обоснованный выбор модели, анализ вклада признаков и веб-интерфейс в рамках единого инструмента предварительной оценки риска диабета. Показано, что число вводимых пользователем признаков может быть сокращено с 16 до 9 без статистически значимой потери качества классификации, что имеет непосредственное значение для практического применения систем предварительного самоскрининга. Разработанный подход может служить основой для дальнейшего развития пользовательских медицинских AI-систем и адаптации модели к локальным клиническим данным.

Дальнейшее развитие системы должно включать внешнюю валидацию на независимых выборках, адаптацию модели к особенностям локальной популяции, расширение набора входных признаков, повышение интерпретируемости модели, сравнение с клинически валидированными шкалами риска, включая FINDRISC, а также разработку механизмов безопасной обработки пользовательских медицинских данных.

**Список литературы**


1. Duncan B.B. IDF Diabetes Atlas 11th Edition 2025: Global Prevalence and Projections for 2050 / B.B. Duncan, D.J. Magliano, E.J. Boyko // Nephrology Dialysis Transplantation. – 2025. – Vol. 41. – P. 7-9. https://doi.org/10.1093/ndt/gfaf177.
2. Ismail L. Association of Risk Factors with Type 2 Diabetes: A Systematic Review / L. Ismail, H. Materwala, J. Al Kaabi // Computational and Structural Biotechnology Journal. – 2021. – Vol. 19. – P. 1759-1785. https://doi.org/10.1016/j.csbj.2021.03.003.
3. Transforming Diabetes Care Through Artificial Intelligence: The Future Is Here / I. Dankwa-Mullan et al // Population Health Management. – 2019. – Vol. 22, № 3. – P. 229-242. https://doi.org/10.1089/pop.2018.0129.
4. Machine Learning and Artificial Intelligence Based Diabetes Mellitus Detection and Self-Management: A Systematic Review / J. Chaki et al // Journal of King Saud University – Computer and Information Sciences. – 2022. – Vol. 34. – P. 3204-3225. https://doi.org/10.1016/j.jksuci.2020.06.013.
5. Rajkomar A. Machine Learning in Medicine / A. Rajkomar, J. Dean, I. Kohane // New England Journal of Medicine. – 2019. – Vol. 380. – P. 1347-1358. https://doi.org/10.1056/NEJMra1814259.
6. Comparative Approaches for Classification of Diabetes Mellitus Data: Machine Learning Paradigm / Maniruzzaman Md. Et al // Computer Methods and Programs in Biomedicine. – 2017. – Vol. 152. – P. 23-34. https://doi.org/10.1016/j.cmpb.2017.09.004.
7. Shan R. Digital Health Technology and Mobile Devices for the Management of Diabetes Mellitus: State of the Art / R. Shan, S. Sarkar, S.S. Martin // Diabetologia. – 2019. – Vol. 62. – P. 877-887. https://doi.org/10.1007/s00125-019-4864-7.
8. Diabetes Digital App Technology: Benefits, Challenges, and Recommendations. A Consensus Report by the European Association for the Study of Diabetes (EASD) and the American Diabetes Association (ADA) Diabetes Technology Working Group / G.A. Fleming et al // Diabetes Care. – 2020. – Vol. 43. – P. 250-260. https://doi.org/10.2337/dci19-0062.
9. Mujumdar A. Diabetes Prediction Using Machine Learning Algorithms / A. Mujumdar, V. Vaidehi // Procedia Computer Science. – 2019. – Vol. 165. – P. 292-299. https://doi.org/10.1016/j.procs.2020.01.047.
10. A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques / R. Krishnamoorthi et al // Journal of Healthcare Engineering. – 2022. – Vol. 2022. – Art. 1684017. – P. 1-10. https://doi.org/10.1155/2022/1684017.
11. Abushawish A.Y.I. Prediction of Early-Stage Diabetes Using Machine Learning / A.Y.I. Abushawish, A.B. Nassif // Proceedings of the 2023 Advances in Science and Engineering Technology International Conferences (ASET). – IEEE, 2023. – P. 1-4.
12. An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study / U.E. Laila et al // Sensors. – 2022. – Vol. 22. – Art. 5247. https://doi.org/10.3390/s22145247.
13. An IoT Based Diabetic Patient Monitoring System Using Machine Learning and Node MCU / A. Rghioui et al // Journal of Physics: Conference Series. – 2021. – Vol. 1743. – Art. 012035. https://doi.org/10.1088/1742-6596/1743/1/012035.

14. Fagherazzi G. Digital Diabetes: Perspectives for Diabetes Prevention, Management and Research / G. Fagherazzi, P. Ravaud // Diabetes & Metabolism. – 2019. – Vol. 45. – P. 322-329. https://doi.org/10.1016/j.diabet.2018.08.012.
15. Beam A.L. Big Data and Machine Learning in Health Care / A.L. Beam, I.S. Kohane // JAMA. – 2018. – Vol. 319. – P. 1317. https://doi.org/10.1001/jama.2017.18391.
16. Obermeyer Z. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine / Z. Obermeyer, E.J. Emanuel // New England Journal of Medicine. – 2016. – Vol. 375. – P. 1216-1219. https://doi.org/10.1056/NEJMp1606181.
17. Topol E.J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence / E.J. Topol // Nature Medicine. – 2019. – Vol. 25. – P. 44-56. https://doi.org/10.1038/s41591-018-0300-7.
18. A Guide to Deep Learning in Healthcare / A. Esteva et al // Nature Medicine. – 2019. – Vol. 25. – P. 24-29. https://doi.org/10.1038/s41591-018-0316-z.
19. Ivanovs M. Perturbation-Based Methods for Explaining Deep Neural Networks: A Survey / M. Ivanovs, R. Kadikis, K. Ozols // Pattern Recognition Letters. – 2021. – Vol. 150. – P. 228-234. https://doi.org/10.1016/j.patrec.2021.06.030.
20. Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques / M.M.F. Islam et al // Data Analytics and Management. – 2020. – P. 113-125.
21. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods / G.S. Collins et al // BMJ. – 2024. – Vol. 385. – Art. e078378. https://doi.org/10.1136/bmj-2023-078378.
22. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2025 / N.A. ElSayed et al // Diabetes Care. – 2025. – Vol. 48 (Suppl. 1). – P. S27-S49. https://doi.org/10.2337/dc25-S002.
23. Lindström J. The Diabetes Risk Score / J. Lindström, J. Tuomilehto // Diabetes Care. – 2003. – Vol. 26, № 3. – P. 725-731. https://doi.org/10.2337/diacare.26.3.725.
24. Contreras I. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review / I. Contreras, J. Vehí // Journal of Medical Internet Research. – 2018. – Vol. 20. – Art. e10775. https://doi.org/10.2196/10775.
25. Dietterich T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms / T.G. Dietterich // Neural Computation. – 1998. – Vol. 10, № 7. – P. 1895-1923. https://doi.org/10.1162/089976698300017197.
26. Nadeau C. Inference for the Generalization Error / C. Nadeau, Y. Bengio // Machine Learning. – 2003. – Vol. 52, № 3. – P. 239-281. https://doi.org/10.1023/A:1024068626366.
27. Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets / J. Demšar // Journal of Machine Learning Research. – 2006. – Vol. 7. – P. 1-30.
28. Lundberg S.M. A Unified Approach to Interpreting Model Predictions / S.M. Lundberg, S.-I. Lee // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 4765-4774.
29. Breiman L. Random Forests / L. Breiman // Machine Learning. – 2001. – Vol. 45, № 1. – P. 5-32. https://doi.org/10.1023/A:1010933404324.
30. Saeys Y. A Review of Feature Selection Techniques in Bioinformatics / Y. Saeys, I. Inza, P. Larrañaga // Bioinformatics. – 2007. – Vol. 23, № 19. – P. 2507-2517. https://doi.org/10.1093/bioinformatics/btm344.

**А.К. Үкібай[1], Е.А. Дайнеко[2], А.Б. Ақбаян[3], Д.С. Баймагамбетова[2], Г.Б. Бекешова[4]***
[1]«МИФИ» Ұлттық ядролық зерттеу университеті» федералды мемлекеттік автономды жоғары білім беру мекемесінің Алматы қаласындағы филиалы,
әл-Фараби даңғылы, 71, 12-корпус, Алматы, 050040, Қазақстан,
[2]Satbayev University, Сәтбаев көшесі, 22а, Алматы, 050013, Қазақстан,
[3]Narxoz University, Жандосов көшесі, 55, Алматы, 050035, Қазақстан,
[4]Л.Н. Гумилев атындаағы Еуразия ұлттық университеті,
Астана, Қазақстан, 0100000, Қазақстан Республикасы, Астана қ., Сатпаев к-сі, 2

**МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНА ОТЫРЫП ҚАНТ ДИАБЕТІ ҚАУПІН АЛДЫН АЛА БАҒАЛАУҒА АРНАЛҒАН ВЕБ-ҚОСЫМШАНЫ ӘЗІРЛЕУ**

*Қант диабеті ең кең таралған созылмалы жұқпалы емес аурулардың бірі болып табылады және жоғары қауіп тобындағы адамдарды ерте анықтауды талап етеді. Бұл мақалада пайдаланушының симптоматикалық және демографиялық деректеріне негізделген қант диабеті қаупін алдын ала бағалауға арналған веб-қосымшаны әзірлеу сипатталған. Жүйе Python/FastAPI, HTML/CSS/JavaScript технологиялары және UCI Early Stage Diabetes Risk Prediction Dataset ашық деректер жиынтығында оқытылған машиналық оқыту модулі негізінде жүзеге асырылды. Жұмыс барысында үш классификатор салыстырылды: Logistic Regression, Support Vector Machine және Random Forest Classifier. Кейінге қалдырылған тестілік іріктемеде ең жақсы нәтижені Random Forest моделі көрсетті және ол қауіп ықтималдығын есептеу үшін веб-қосымшаға енгізілді. Пайдаланушы жасын, жынысын және симптоматикалық белгілер жиынтығын енгізгеннен кейін жүйе қауіп деңгейін, визуалды индикаторды және профилактикалық ұсыныстар блогын көрсетеді. Әзірленген қосымшаны алдын ала өзін-өзі бағалауға және пайдаланушының хабардарлығын арттыруға арналған бағдарламалық прототип ретінде қарастыру қажет. Ол диагноз қоюға арналмаған және тәуелсіз деректер негізінде одан әрі сыртқы клиникалық валидацияны талап етеді. Модельдерді салыстыру қайталанатын стратификацияланған 10 еселік кросс-валидация арқылы орындалды (5 қайталау, әр модель үшін 50 бағалау). Ең жоғары әрі тұрақты нәтижені Random Forest моделі көрсетті (accuracy 98.27 ± 1.60%, F1-өлшемі 98.59 ± 1.31%, ROC-AUC 99.93 ± 0.15%), оның SVM және Logistic Regression модельдерінен артықшылығы статистикалық тұрғыдан расталды ($p < 0.05$). Белгілердің үлесі үш тәуелсіз әдіспен (Джини критерийі бойынша маңыздылық, пермутациялық маңыздылық және SHAP мәндері) талданып, полиурия, полидипсия, жыныс және жас ең ақпаратты белгілер ретінде анықталды. Алынған дәрежелеу негізінде 16 белгінің орнына 9 белгіден тұратын қысқартылған сауалнама құрылды: классификация сапасының статистикалық маңызды төмендеуі байқалмады (accuracy 97.62 ± 1.87%, $p = 0.246$), бұл пайдаланушы енгізетін деректер көлемін 43.8%-ға қысқартуға мүмкіндік береді.*

***Түйін сөздер:*** *қант диабеті; машиналық оқыту; веб-қосымша; FastAPI; симптомдарды тану; қауіп-қатерді бағалау; денсаулықты мониторингтеу; scikit-learn; ерте диагностика. кросс-валидация; статистикалық тексеру; белгілердің маңыздылығы; SHAP.*

**A.K. Ukibay[1], E.A. Dayneko[2], A.B. Akbayan[3], D.S. Baymagambetova[2], G.B. Bekeshova[4]***

[1]Almaty Branch of the National Research Nuclear University MEPhI,
71 al-Farabi Ave., building 12, Almaty, 050040, Kazakhstan,
[2]Satbayev University, 22a Satpayev str., Almaty, 050013, Kazakhstan,
[3]Narxoz University, 55 Zhandossov str., Almaty, 050035, Kazakhstan
[4]L.N. Gumilyov Eurasian National University,
Astana, Kazakhstan 0100000, Republic of Kazakhstan, Astana, Satpayev street, building 2

## DEVELOPMENT OF A WEB APPLICATION FOR PRELIMINARY DIABETES RISK ASSESSMENT USING MACHINE LEARNING METHODS

*Diabetes mellitus is one of the most common chronic non-communicable diseases and requires early identification of individuals at increased risk. This article presents the development of a web application for preliminary diabetes risk assessment based on users' symptomatic and demographic data. The system was implemented using Python/FastAPI, HTML/CSS/JavaScript, and a machine learning module trained on the open UCI Early Stage Diabetes Risk Prediction Dataset. Three classifiers were compared in this study: Logistic Regression, Support Vector Machine, and Random Forest Classifier. The best performance on the held-out test set was achieved by the Random Forest model, which was integrated into the web application to calculate risk probability. The user enters age, gender, and a set of symptomatic features, after which the system displays the risk level, a visual indicator, and a block of preventive recommendations. The developed application should be considered a software prototype for preliminary self-assessment and increasing user awareness. It is not intended for diagnosis and requires further external clinical validation on independent data. Model comparison was performed using repeated stratified 10-fold cross-validation (five repetitions, 50 estimates per model). The Random Forest model achieved the best and most stable performance (accuracy 98.27 ± 1.60%, F1-score 98.59 ± 1.31%, ROC-AUC 99.93 ± 0.15%), and its superiority over SVM and Logistic Regression was confirmed statistically (corrected paired t-test, 5×2cv t-test and Wilcoxon signed-rank test, $p < 0.05$). Feature contributions were analysed by three independent methods (Gini importance, permutation importance and SHAP values), which consistently identified polyuria, polydipsia, gender and age as the most informative features. Based on this ranking, a reduced questionnaire of nine features was constructed, for which no statistically significant loss of classification quality was observed compared with the full set of sixteen features (accuracy 97.62 ± 1.87%, $p = 0.246$), reducing the amount of user input by 43.8%.*

***Key words:*** *diabetes mellitus; machine learning; web application; FastAPI; symptom recognition; risk assessment; health monitoring; scikit-learn; early diagnosis. cross-validation; statistical significance testing; feature importance; SHAP; feature selection.*

**Сведения об авторах**

**Айдана Қуанышбекқызы Үкібай** – Филиал Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский ядерный университет «МИФИ» в городе Алматы», Казахстан; e-mail: aidanakuanishbekkyzy@gmail.com.

**Евгения Александровна Дайнеко** – PhD, ассоциированный профессор, НАО «Казахский национальный исследовательский технический университет имени К.И. Сатпаева» (Satbayev University), Алматы, Казахстан; e-mail: y.daineko@satbayev.university.

**Ақбаян Бекарыстанқызы** – PhD, Narxoz University, Алматы, 050035, Казахстан; e-mail: akbayan.b@gmail.com. ORCID: https://orcid.org/0000-0003-3984-2718.

**Динара Жалеловна Баймагамбетова** – преподаватель, Satbayev University, Алматы, Казахстан; e-mail: d.baimagambetova@satbayev.university.

**Гульвира Бауыржановна Бекешова*** – магистр технических наук, старший преподаватель, Евразийского национального университета имени Л.Н. Гумилева, Астана, Казахстан; e-mail: gulvirabauyrzhanovna@gmail.com. ORCID: https://orcid.org/0000-0002-1635-4693.

**Авторлар туралы мәліметтер**

**Айдана Қуанышбекқызы Үкібай** – Алматы қаласындағы «Ұлттық зерттеу ядролық университеті «МИФИ» федералдық мемлекеттік автономды жоғары білім беру мекемесінің филиалы, Алматы қ., Қазақстан; e-mail: aidanakuanishbekkyzy@gmail.com.

**Евгения Александровна Дайнеко** – PhD, қауымдастырылған профессор, «Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті» КеАҚ (Satbayev University), Алматы қ., Қазақстан; e-mail: y.daineko@satbayev.university.

**Ақбаян Бекарыстанқызы** – PhD, Narxoz University, Алматы қ., Қазақстан; e-mail: akbayan.b@gmail.com. ORCID: https://orcid.org/0000-0003-3984-2718.

**Динара Жалеловна Баймагамбетова** – оқытушы, Satbayev University, Алматы қ., Қазақстан; e-mail: d.baimagambetova@satbayev.university.

**Гульвира Бауыржановна Бекешова*** – техника ғылымдарының магистрі, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, аға оқытушы, Астана қ., Қазақстан; e-mail: gulvirabauyrzhanovna@gmail.com. ORCID: https://orcid.org/0000-0002-1635-4693.

**Information about the authors**

**Aidana Ukibay** – Branch of the National Research Nuclear University MEPhI in Almaty, Kazakhstan; e-mail: aidanakuanishbekkyzy@gmail.com.

**Evgeniya Aleksandrovna Dayneko** – PhD, Associate Professor, Satbayev University (Kazakh National Research Technical University named after K.I. Satpayev), Almaty, Kazakhstan; e-mail: y.daineko@satbayev.university.

**Akbayan Bekarystankyzy** – PhD, Narxoz University, Almaty, Kazakhstan; e-mail: akbayan.b@gmail.com. ORCID: https://orcid.org/0000-0003-3984-2718.

**Dinara Baymagambetova** – Lecturer, Satbayev University, Almaty, Kazakhstan; e-mail: d.baimagambetova@satbayev.university.

**Gulvira Bekeshova*** – Senior Lecturer at the Department of Information Security, IT Faculty at the L.N. Gumilyov ENU, Astana, Kazakhstan; e-mail: gulvirabauyrzhanovna@gmail.com. ORCID: https://orcid.org/0000-0002-1635-4693.

*Поступила в редакцию 27.06.2026*
*Поступила после доработки 03.09.2026*
*Принята к публикации 04.09.2026*