

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

**МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ
КЛАССИФИКАЦИИ ЗАБОЛЕВАНИЯ ДИАБЕТОМ**
АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Магистрантки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Авдеевой Дарьи Александровны

Научный руководитель
доцент, к. э. н.

А. Р. Файзлиев

Заведующий кафедрой
д. ф.-м. н., доцент

С. П. Сидоров

Саратов 2025

СОДЕРЖАНИЕ

1	Введение	3
2	Метод опорных векторов	4
3	Метрики оценки эффективности	5
4	Разработка и исследование модели классификации заболевания диа- бетом	7
5	Заключение	9
6	Публикации	11
7	Ключевые слова	12

1 Введение

В современных условиях проблема сахарного диабета приобретает характер глобальной медико-социальной проблемы. По данным Международной диабетической федерации наблюдается устойчивая тенденция к росту распространённости заболевания, что позволяет говорить о неинфекционной пандемии. За последние десятилетия число людей с диабетом увеличилось более чем в три раза, достигнув к 2025 году 537 миллионов случаев.

Существующие методы диагностики диабета, несмотря на свою надёжность, обладают рядом существенных ограничений:

- необходимость инвазивных процедур для определения уровня глюкозы;
- временной лаг в выявлении заболевания;
- зависимость результатов от субъективного опыта специалиста;
- значительные временные и финансовые затраты на комплексное обследование.

Целью данного исследования является изучение методов машинного обучения и разработка программного кода для их сравнительного анализа на данных о заболеваниях диабетом.

2 Метод опорных векторов

Метод опорных векторов (Support Vector Machine, SVM) представляет собой мощный инструмент машинного обучения, предназначенный для решения задач классификации и регрессии. Основная идея метода заключается в поиске оптимальной разделяющей гиперплоскости, которая максимально эффективно разделяет данные на классы.

SVM применялся для бинарной классификации пациентов (диабет/нет) с использованием линейного и радиального ядер (RBF) для работы с нелинейными данными. Метод основан на построении оптимальной гиперплоскости, максимизирующей зазор между классами. Для ускорения обучения использовался алгоритм SMO (Sequential Minimal Optimization). Результаты показали, что точность SVM составила 78.6%, при этом радиальное ядро продемонстрировало лучшую эффективность при работе с гетерогенными медицинскими данными.

3 Метрики оценки эффективности

Для комплексной оценки качества прогностических моделей диабета использовался набор метрик, каждая из которых отражает различные аспекты качества классификации:

- Ассигасу (общая точность) составила 83.7% для нейронной сети. Эта метрика показывает долю правильно классифицированных случаев среди всех наблюдений. Однако в случае дисбаланса классов ассигасу может быть вводящей в заблуждение, поэтому использовались дополнительные метрики.
- F1-score достиг значения 0.81, представляя собой гармоническое среднее между точностью (precision) и полнотой (recall). Данная метрика особенно важна в медицинской диагностике, где необходимо учитывать как точность положительных диагнозов, так и полноту выявления больных пациентов.
- ROC-AUC (площадь под рабочей характеристикой приёмника) составила 0.85, что свидетельствует о хорошей способности модели различать классы. Значение близкое к 1 указывает на высокую диагностическую эффективность модели.

Матрица ошибок предоставила детальную информацию о типах ошибок модели:

- ложноположительные результаты (здоровые пациенты, ошибочно диагностированные как больные);
- ложноотрицательные результаты (больные пациенты, ошибочно классифицированные как здоровые);
- верно классифицированные положительные случаи;
- Верно классифицированные отрицательные случаи.

Для более полной оценки качества моделей также учитывались следующие важные показатели. Чувствительность (Sensitivity) отражает способность модели правильно идентифицировать больных пациентов. Этот показатель особенно важен в медицинской диагностике, так как позволяет оценить, какой процент действительно больных пациентов будет правильно выявлен моделью.

Специфичность (Specificity) показывает способность модели правиль-

но идентифицировать здоровых пациентов. Высокая специфичность важна для минимизации количества ложноположительных диагнозов, что помогает избежать ненужных дополнительных обследований.

Positive Predictive Value (PPV) представляет собой вероятность того, что пациент действительно болен при положительном результате теста. Этот показатель помогает оценить надежность положительного диагноза, выданного моделью.

Negative Predictive Value (NPV) отражает вероятность того, что пациент здоров при отрицательном результате теста. Высокая NPV важна для уверенности в отрицательных диагнозах и минимизации пропусков заболевания. Такой комплексный подход к оценке моделей позволил получить полное представление об их диагностической эффективности и выбрать наиболее подходящий алгоритм для практического применения.

4 Разработка и исследование модели классификации заболевания диабетом

В рамках исследования была проведена комплексная работа по созданию и тестированию прогностических моделей диабета. Исследование базировалось на наборе данных Pima Indians Diabetes Database, включающем информацию о 768 пациентках с 8 признаками: уровень глюкозы, ИМТ, возраст и другие показатели. Особенностью набора данных являлся дисбаланс классов - 65% здоровых и 35% больных пациентов.

Процесс разработки модели включал несколько последовательных этапов. На первом этапе проводилась предобработка данных, которая заключалась в замене нулевых значений медианой и нормализации признаков. Затем осуществлялось разделение выборки на обучающую (80%) и тестовую (20%) части. Для оценки стабильности моделей применялась 5-fold кросс-валидация.

В ходе исследования было проведено сравнение восьми различных алгоритмов машинного обучения:

- Логистическая регрессия
- Случайный лес
- Метод опорных векторов
- К-ближайших соседей
- Деревья решений
- Градиентный бустинг
- Наивный байесовский классификатор
- Искусственная нейронная сеть

Для каждой модели проводилась оценка качества с помощью метрик точности, полноты, F1-меры и ROC-AUC. Дополнительно анализировались чувствительность, специфичность, PPV и NPV.

Наилучшие результаты показала искусственная нейронная сеть с архитектурой, состоящей из трех слоев (64-32-1 нейрон), использующая Dropout регуляризацию (0.2-0.5) и функции активации ReLU/sigmoid. Точность этой модели достигла 84.9%, что является наивысшим показателем среди всех протестированных алгоритмов.

Второй по эффективности оказался метод случайного леса, определив-

ший наиболее значимые признаки: уровень глюкозы (важность 0.47), ИМТ (0.22) и возраст (0.15). Точность модели составила 82.3%. Другие методы показали результаты в диапазоне 72-81%.

Особое внимание было уделено анализу влияния различных факторов на прогноз. Установлено, что наибольшее воздействие на результат классификации оказывают уровень глюкозы (корреляция 0.47 с целевым признаком), ИМТ (0.22) и возраст (корреляция 0.54 с числом беременностей).

Для улучшения качества моделей проводился подбор гиперпараметров с помощью GridSearchCV. Наибольший прирост точности был достигнут для случайного леса при следующих параметрах:

- Количество деревьев: 100
- Максимальная глубина: 20
- Минимальное количество объектов для разделения: 5

Финальным этапом исследования стало построение ROC-кривых для всех моделей, что позволило провести сравнительный анализ их диагностической эффективности. Результаты демонстрируют преимущество нейросетевых подходов и метода опорных векторов в задачах медицинской диагностики по сравнению с традиционными методами.

5 Заключение

В ходе исследования была проведена комплексная работа по разработке и тестированию моделей машинного обучения для классификации диабета. На основе анализа данных Pima Indians Diabetes Database были созданы и протестированы различные алгоритмы прогнозирования.

В процессе исследования был проведен тщательный анализ исходных данных, включающий визуализацию и предобработку. Особое внимание было уделено обработке пропущенных значений, где нулевые показатели были заменены медианой соответствующих признаков. Данные были нормализованы для обеспечения корректной работы моделей машинного обучения.

На первом этапе были протестированы классические методы классификации: логистическая регрессия, случайный лес, метод опорных векторов, k-ближайших соседей, деревья решений, градиентный бустинг и наивный байесовский классификатор. Наилучшие результаты показал случайный лес, достигший точности около 82.3%.

Следующим этапом стало создание и обучение искусственной нейронной сети с архитектурой, состоящей из трех слоев (64-32-1 нейрон). Модель использовала функции активации ReLU и sigmoid, а также Dropout регуляризацию. Нейронная сеть продемонстрировала точность 84.9% на тестовой выборке, что стало лучшим результатом среди всех протестированных моделей.

В ходе исследования были определены наиболее значимые признаки для прогнозирования диабета: уровень глюкозы в крови, индекс массы тела и возраст пациента. Эти параметры показали наибольшую корреляцию с наличием заболевания.

Для оценки качества моделей использовались различные метрики: точность, полнота, F1-мера, а также ROC-AUC. Дополнительно были построены ROC-кривые для всех моделей, что позволило провести сравнительный анализ их диагностической эффективности.

Практическая значимость работы заключается в создании прогностической модели, способной выявлять предрасположенность к диабету на ранних стадиях. Разработанная система может быть интегрирована в существующие медицинские информационные системы для поддержки принятия врачебных

решений.

Дальнейшие исследования могут быть направлены на улучшение качества моделей путем расширения набора признаков, увеличения объема обучающих данных и оптимизации гиперпараметров. Также перспективным направлением является интеграция разработанных решений в системы телемедицины для обеспечения удаленного мониторинга пациентов.

Полученные результаты демонстрируют эффективность применения современных методов машинного обучения для решения задач медицинской диагностики и могут служить основой для создания практических инструментов поддержки принятия врачебных решений в области диабетологии.

6 Публикации

Работа прошла апробацию на различных научных площадках, в частности, на ежегодной студенческой конференции «Актуальные проблемы математики и механики», которую проводил механико-математический факультет СГУ в апреле 2025 года, в секции «Анализ данных», а также на XIII Международной молодежной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками» в ноябре 2024 года.

7 Ключевые слова

Машинное обучение, классификация диабета, искусственные нейронные сети, метод опорных векторов, медицинская диагностика, прогностическая модель, предиктивная аналитика, медицинская информатика, анализ данных, сахарный диабет.