МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра физики открытых систем

Анализ и обработка многомерных данных с использованием методов машинного обучения и нейронных сетей

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 4 курса 4041 группы
направления 09.03.02 «Информационные системы и технологии»
код и наименование направления
института физики
наименование факультета, института, колледжа
Лаврухина Виталия Витальевича
фамилия, имя, отчество

Научный руководитель
профессор, д.ф.-м.н., профессор _____ А.Н. Павлов
должность, ученая степень, уч. звание подпись, дата Инициалы Фамилия

Заведующий кафедрой физики открытых систем
полное наименование кафедры
д.ф.-м.н., профессор _____ А.А. Короновский
должность, ученая степень, уч. звание подпись, дата Инициалы Фамилия

Саратов 2025 г.

ВВЕДЕНИЕ

Современные технологии обработки данных сталкиваются с необходимостью анализа сложных многомерных наборов, содержащих скрытые закономерности, шумы и избыточные признаки [1]. Такие данные широко распространены в медицине, биологии, физике и других научных областях. Рост объёма и сложности информации требует разработки эффективных методов, способных не только обрабатывать большие массивы данных, но и выявлять в них значимые зависимости [2]. Методы машинного обучения и нейронных сетей предоставляют мощные инструменты для решения этих задач, включая кластеризацию, классификацию, регрессию и снижение размерности [3].

Актуальность данного исследования обусловлена необходимостью разработки эффективных подходов к обработке многомерных данных, которые позволяют не только улучшить понимание их структуры, но и повысить точность прогнозирования. В работе рассматриваются два набора данных: классический набор "Ирисы Фишера" и медицинские данные о пациентах с мочекаменной болезнью (Kidney Stone Dataset). Эти наборы представляют собой примеры данных с разной структурой и сложностью, что позволяет провести сравнительный анализ методов машинного обучения и нейронных сетей.

Целью работы является сравнительный анализ методов кластеризации и классификации для обработки многомерных данных, оценка их эффективности и выявление оптимальных подходов для каждого типа данных.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

1. Теоретическая часть

Метод k-средних имеет итеративный алгоритм кластеризации, минимизирующий сумму квадратов расстояний от точек до центроидов кластеров. Число кластеров k задаётся заранее.

Формула (обновление центроидов):

$$\mu_j = \frac{1}{|C_j|}\sum_{x_i \in C} x_i, \quad (1)$$

где μ_j — новый центроид j -го кластера, C_j — множество точек, принадлежащих кластеру j, $|C_j|$ — количество точек в кластере j, x_i — i-я точка данных.

Метод сдвига среднего (Mean Shift) — непараметрический метод кластеризации, основанный на ядерном сглаживании. Ищет локальные максимумы плотности данных, адаптируясь к их форме.

Формула (сдвиг центра):

$$m(x) = \frac{\Sigma_{x_i \in N(x)} K(x_i - x) \cdot x_i}{\Sigma_{x_i \in N(x)} K(x_i - x)}, \quad (2)$$

где $m(x)$ — сдвиг точки x направлении максимума плотности, N(x) — окрестность точки x (задаётся радиусом или ядром), $K(x)$ — ядерная функция (например, гауссова), x_i — точки в окрестности N(x)N(x).

Метод главных компонент (PCA) - метод снижения размерности, находящий ортогональные направления максимальной дисперсии данных. W — матрица собственных векторов ковариационной матрицы.

Формула (проекция):

$$z = W^T(x - \mu), \quad (3)$$

где z — проекция данных в новое пространство, W — матрица собственных векторов ковариационной матрицы, x — исходный вектор данных, μ — среднее значение данных.

Случайный лес (Random Forest) — ансамбль решающих деревьев, обученных на бутстрап-выборках. Уменьшает переобучение за счёт усреднения предсказаний.

Формула (прогноз для классификации):

$$\hat{y} = mode(\{f_t(x)\}_{t=1}^{T}), \quad (4)$$

где $\hat{y}$— итоговый прогноз модели, $\{f_t(x)$ — предсказание t -го дерева в ансамбле, T — общее количество деревьев.

Многослойный перцептрон (MLP)- искусственная нейронная сеть с прямым распространением сигнала, использующая нелинейные функции активации σ для обучения сложным паттернам.

Формула (прямое распространение):

$$a^{(l)} = \sigma\left(W^{(l)}a^{(l-1)} + b^{(l)}\right), \quad (5)$$

где $a^{(l)}$ — активации нейронов на слое l, $W(l)$ — матрица весов слоя l, $b(l)$ — вектор смещений слоя l, σ — функция активации (ReLU, sigmoid и др.).

Рекуррентная нейронная сеть (RNN) - сеть с цикличными связями, сохраняющая состояние между шагами. Подходит для обработки последовательностей (текст, временные ряды).

Формула (скрытое состояние):

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b), \quad (6)$$

где h_t — скрытое состояние на шаге t, W_h — матрица весов для предыдущего состояния,W_x— матрица весов для входных данных, x_t— вход на шаге t, b — вектор смещения, σ — функция активации (ReLU, sigmoid и др.).Каждый метод имеет свои преимущества и ограничения в зависимости от типа данных и задачи.

2. Практическая часть

В рамках исследования были проведены эксперименты на двух наборах данных:

- Iris Dataset: классический набор, содержащий 150 образцов ирисов с 4 признаками. Алгоритмы кластеризации и классификации показали высокую точность, что подтверждает их применимость для данных с четкой структурой.
- Kidney Stone Dataset: медицинские данные о 90 пациентах с 7 признаками.

Результаты показали, что сложность и неоднородность данных требуют более тщательного выбора методов и их настройки.

В ходе исследования были получены графики, наглядно демонстрирующие результаты применения различных методов кластеризации и классификации к наборам данных Iris и Kidney Stone. Ниже представлено описание каждого графика и его интерпретация.

1. Результаты кластеризации методом Mean-Shift

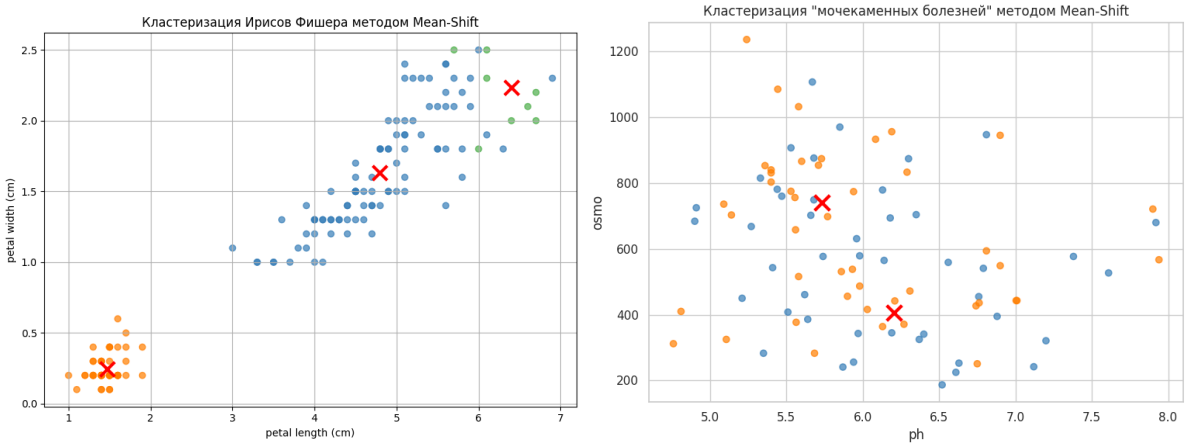

Рисунок 1. График для Iris Dataset - слева, для Kidney Stone Dataset - справа.

На рис. 1 показано разделение данных на кластеры с учётом плотности распределения. Видно, что алгоритм успешно выделил три группы, соответствующие видам ирисов, однако в некоторых областях наблюдается перекрытие кластеров из-за чувствительности метода к параметру bandwidth. Кластеризация данных о пациентах с мочекаменной болезнью оказалась менее точной. Алгоритм выделил несколько кластеров, но их границы размыты, что связано с высокой вариативностью и зашумленностью медицинских данных. Точность составила 54.44%.

2. Результаты кластеризации методом K-means

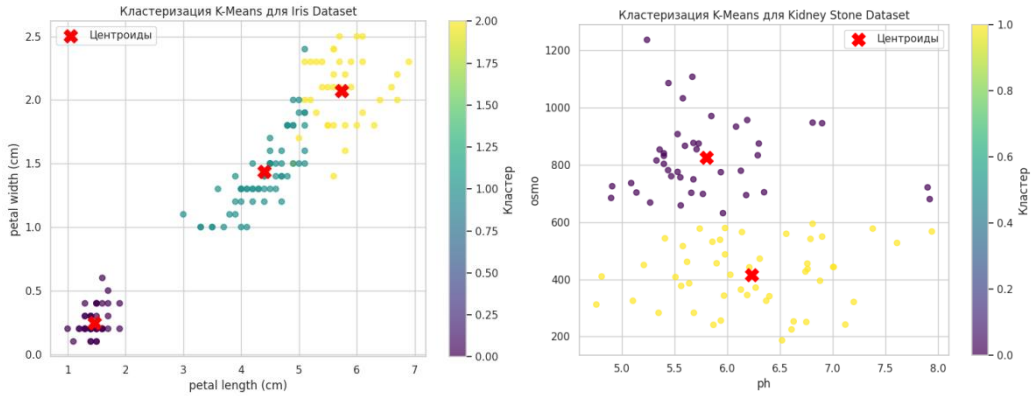

Рисунок 2. График для Iris Dataset - слева, для Kidney Stone Dataset - справа.

K-средних четко разделил данные на три кластера, соответствующие видам ирисов (рис. 3). Границы кластеров хорошо различимы, что подтверждает высокую точность метода (89.33%) для данных с явной структурой. Для медицинских данных K-means показал менее точные результаты (53.33%). Кластеры перекрываются, что указывает на необходимость предварительной обработки данных или выбора другого числа кластеров.

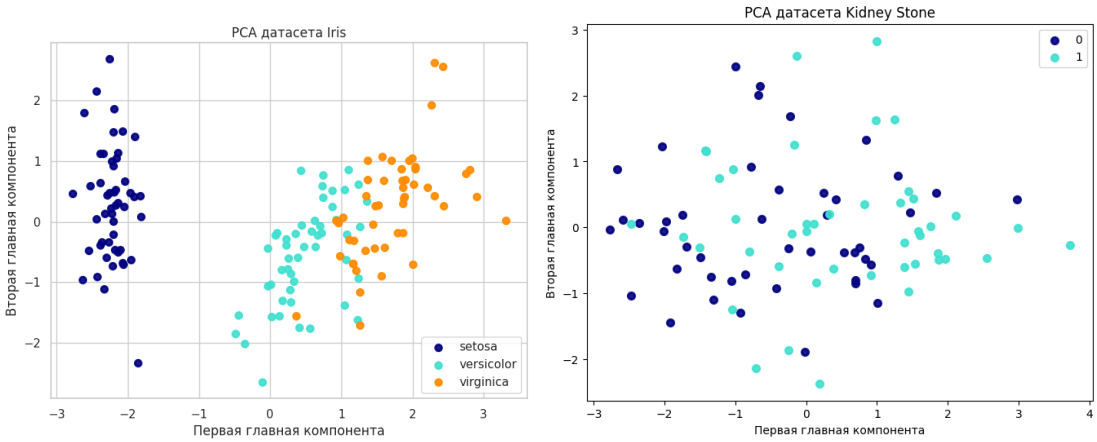

Рисунок 3. График для Iris Dataset - слева, для Kidney Stone Dataset - справа

После применения PCA данные визуализированы в двумерном пространстве. Видно, что три вида ирисов хорошо разделяются, что подтверждает сохранение 95% дисперсии. PCA сохранил 76% дисперсии, но разделение данных менее четкое. Это связано с более сложной структурой медицинских данных, где признаки сильно коррелируют.

3. Результаты классификации методом k-NN

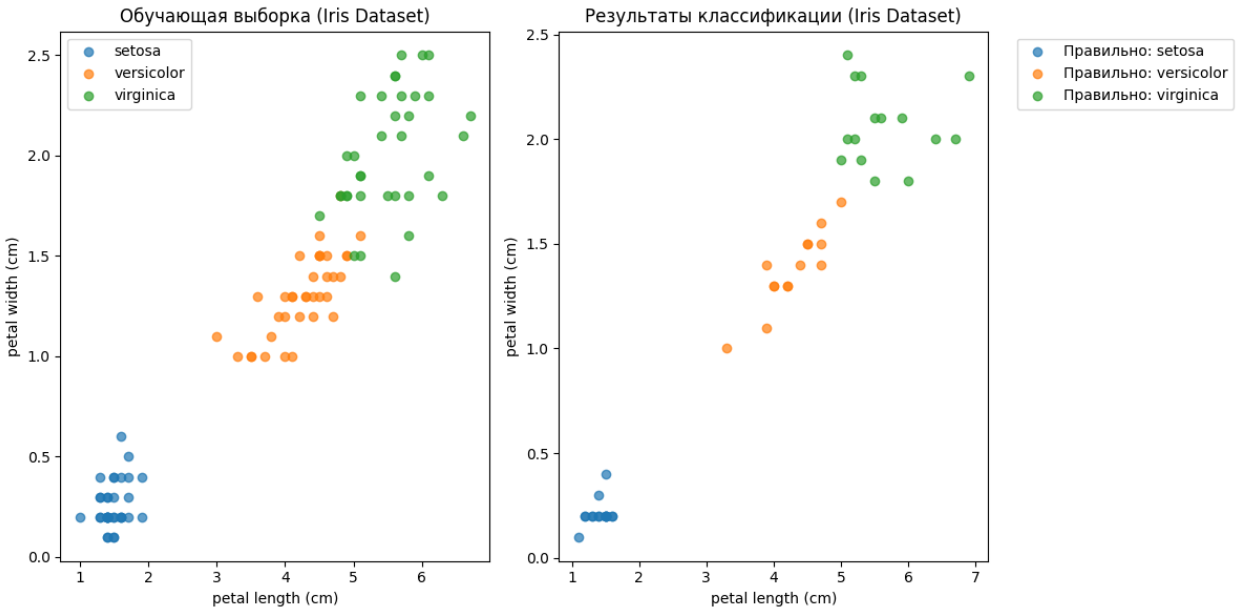

Рисунок 7. График для Iris Dataset (слева, обучающая выборка; справа, результаты классификации):

На рисунке 7 показано распределение данных и границы принятия решений. k-NN обеспечил 100% точность, так как данные хорошо разделяются в пространстве признаков.

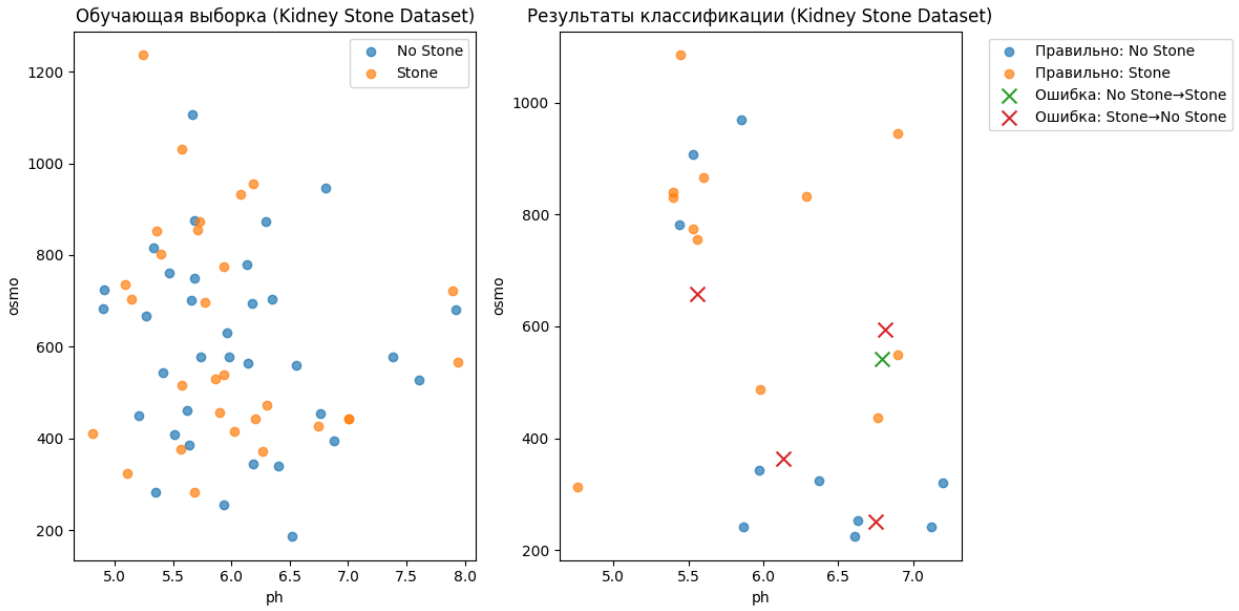

Рисунок 8. График для Kidney Stone Dataset (слева, обучающая выборка;; справа, результаты классификации):

На рис. 8 границы между классами менее четкие, что объясняется наличием выбросов и перекрывающихся областей. Точность снизилась до 81.48%.

4. Результаты классификации методом Random Forest

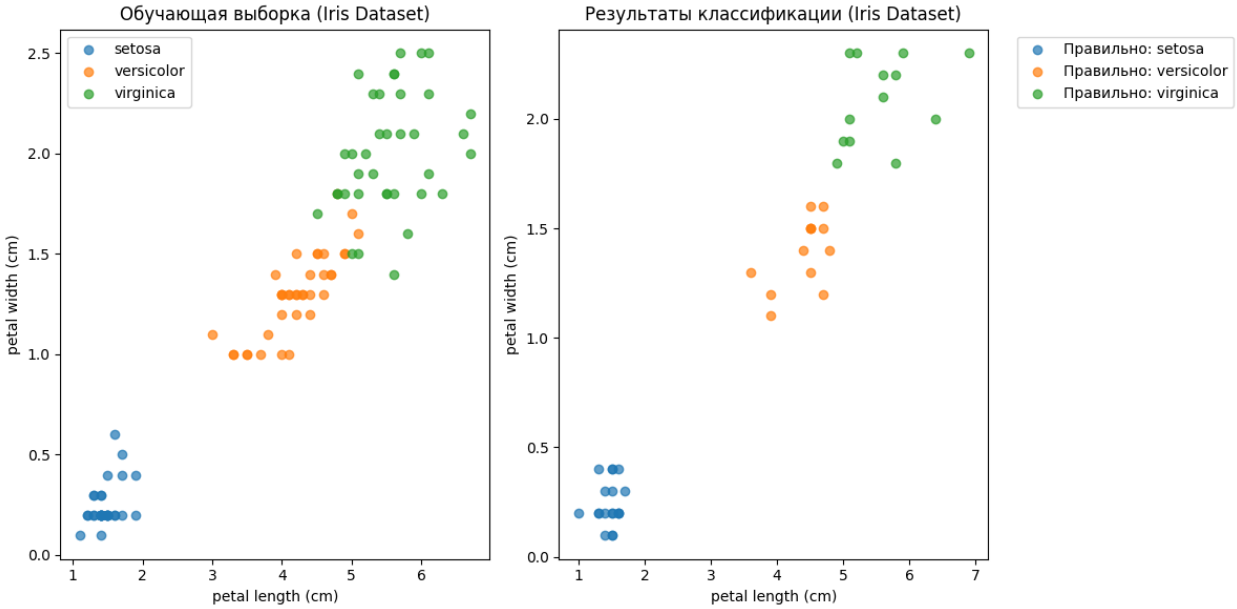

Рисунок 9. График для Iris Dataset (слева, обучающая выборка; справа, результаты классификации)

Random Forest достиг 100% точности. На рис 9. результатов классификации все точки соответствуют истинным классам.

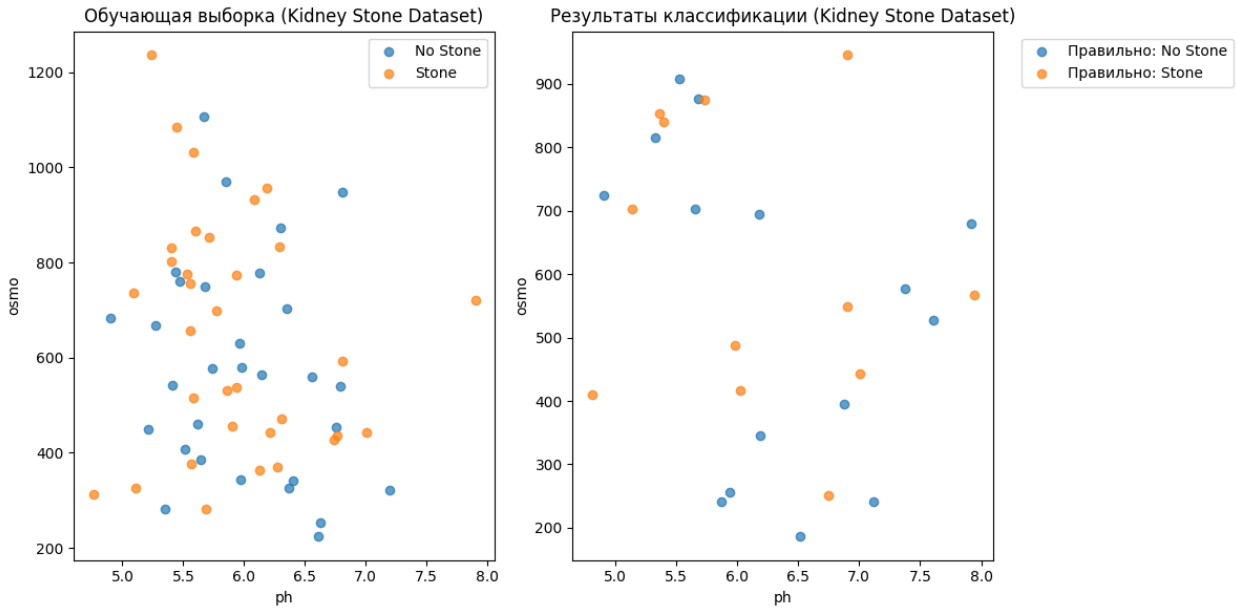

Рисунок 10. График для Kidney Stone Dataset (слева, обучающая выборка; справа, результаты классификации):

Метод также показал 100% точность, что делает его наиболее надежным для медицинских данных. Рис. 10 демонстрирует четкое разделение классов.

6. Результаты классификации методом MLP (многослойный перцептрон)

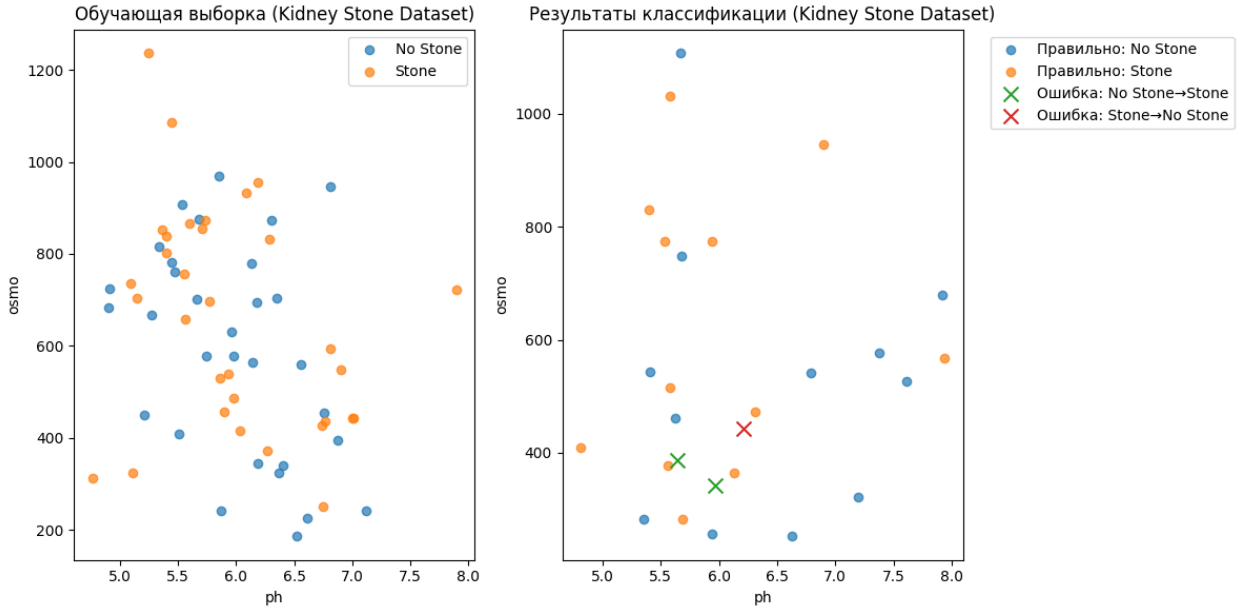

Рисунок 11. График для Iris Dataset (слева, обучающая выборка; справа, результаты классификации)

MLP достиг 98% точности. На рис.11 видны небольшие ошибки, связанные с перекрытием классов вблизи границ.

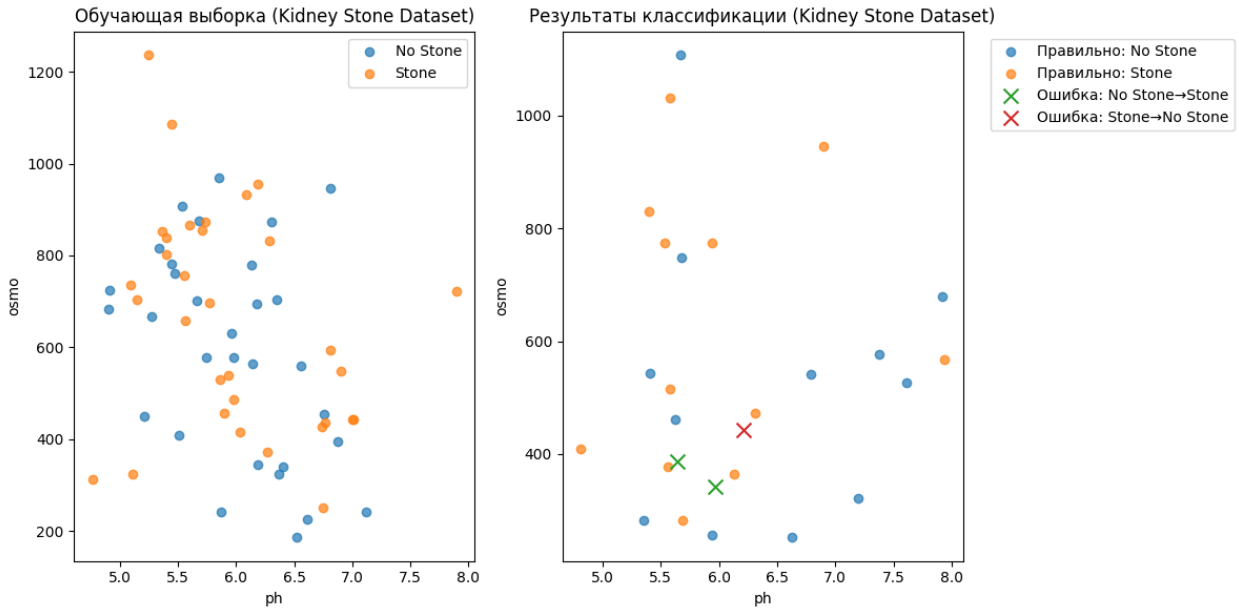

Рисунок 12. График для Kidney Stone Dataset (слева, обучающая выборка; справа, результаты классификации):

Точность составила 89%. Рис. 12 показывает, что нейросеть справилась с большинством случаев, но ошибки возникают в областях с высокой неопределенностью.

7. Результаты классификации методом RNN (рекуррентная сеть)

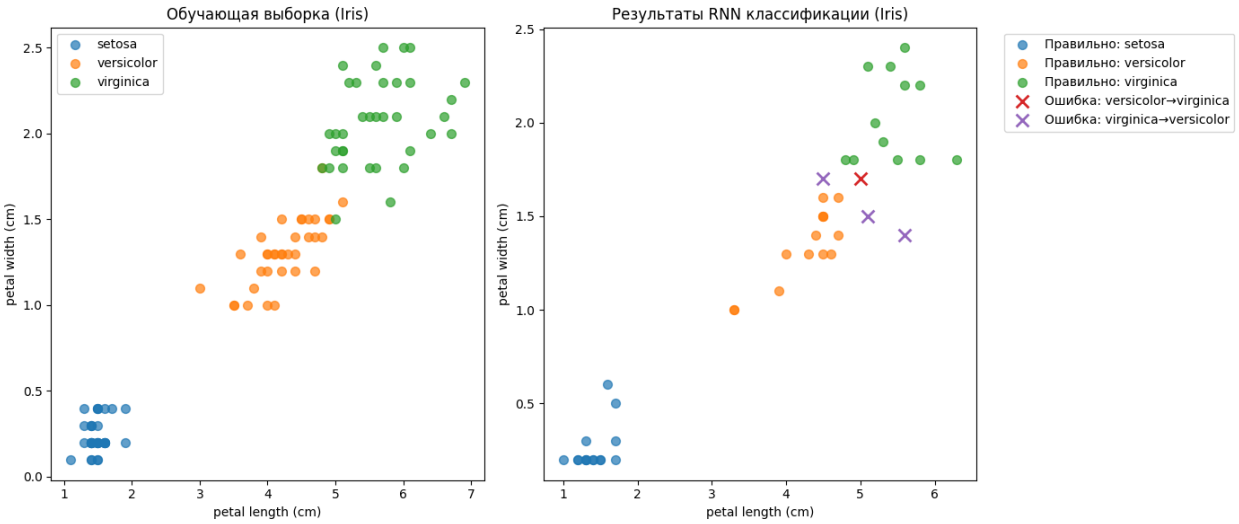

Рисунок 13. График для Iris Dataset (слева, обучающая выборка; справа, результаты классификации):

RNN показала высокую точность (91.1%%), но рис. 13 демонстрирует небольшие отклонения в предсказаниях.

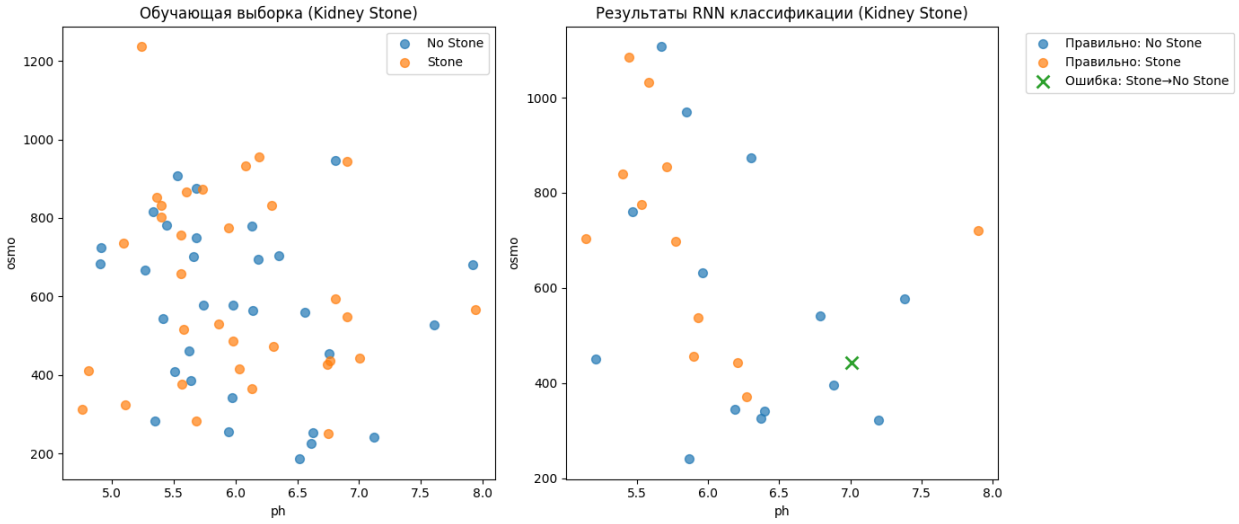

Рисунок 14. График для Kidney Stone Dataset (слева, обучающая выборка; справа, результаты классификации):

Точность улучшилась до 96.3%.

4. Сравнительный анализ методов

На основе проведенных экспериментов были сделаны следующие выводы:

- Для данных с четкой структурой (Iris) наиболее эффективными являются k-NN и Random Forest.
- Для сложных медицинских данных (Kidney Stone) наилучшие результаты показали Random Forest, RNN и MLP, в то время как Mean-Shift требуют дополнительной оптимизации.
- Методы снижения размерности, такие как PCA, полезны для визуализации и упрощения данных, но их эффективность зависит от структуры набора.

ЗАКЛЮЧЕНИЕ

В данной работе был проведен сравнительный анализ методов машинного обучения и нейронных сетей для обработки многомерных данных. Исследование показало, что выбор метода зависит от структуры и сложности данных:

- Для данных с явной группировкой (Iris) эффективны простые методы, такие как k-NN и K-means.
- Для сложных и зашумленных данных (Kidney Stone) предпочтительны Random Forest, RNN и MLP.
- Алгоритмы, чувствительные к параметрам (Mean-Shift), требуют дополнительной настройки.

Результаты работы могут быть применены в медицинской диагностике, биологии и других областях, где требуется обработка многомерных данных. Дальнейшие исследования могут быть направлены на оптимизацию нейросетевых архитектур для работы с медицинскими наборами данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Бишоп, К. М. Распознавание образов и машинное обучение / К. М. Бишоп. – Springer, 2006. – 738 с.

[2] Гудфеллоу, И. Глубокое обучение / И. Гудфеллоу, Й. Бенджио, А. Курвиль. – MIT Press, 2016. – 800 с.

[3] Педрегоса, Ф. Scikit-learn: Машинное обучение на Python / Ф. Педрегоса и др. // Journal of Machine Learning Research. – 2011. – Т. 12. – С. 2825–2830.