МИНОБРНАУКИ РОССИИ

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

Кафедра физики открытых систем

Анализ многопараметрических данных
методами 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

Жусупова Асхата Сагнышевича

фамилия, имя, отчество

Научный руководитель

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

Заведующий кафедрой
физики открытых систем

д. ф - м. н., профессор		А. А. Короновский
должность, ученая степень, уч. звание	подпись, дата	Инициалы Фамилия

Саратов 2026 г.

Введение

В современную эпоху цифровых технологий объемы данных стремительно растут, и обработка многомерной информации становится ключевой задачей в различных областях — от медицины и биоинформатики до финансов и промышленности. Особую актуальность приобретают методы машинного обучения и нейронных сетей, позволяющие эффективно анализировать, классифицировать и визуализировать сложные данные [1].

Актуальность исследования обусловлена необходимостью разработки эффективных подходов к обработке многомерных данных, которые могут содержать скрытые закономерности, шумы или избыточные признаки. Методы кластеризации и классификации помогают упростить анализ, улучшить интерпретаемость данных и повысить точность прогнозирования.

Значимость работы заключается в сравнительном анализе различных методов машинного обучения и нейронных сетей для обработки многомерных данных

Практическая ценность исследования состоит в том, что применение этих методов может быть полезно в различных сферах, включая медицинскую диагностику (анализ данных о заболеваниях), биометрию (распознавание паттернов) и другие области, где требуется работа с многомерными наборами данных.

Целью выпускной квалификационной работы является реализация, освоение и сравнение методов машинного обучения и нейронных сетей для обработки многомерных данных на двух датасетах: Iris (ирисы Фишера) и Kidney Stone Dataset (данные о мочекаменной болезни).

Глава1. Теоретические сведения

Машинное обучение — это направление искусственного интеллекта, в котором компьютеры учатся принимать решения на основе данных [2]. Его цель — обучить алгоритмы следовать определенным принципам, позволяющим им анализировать информацию схожим образом.

Одной из ключевых областей машинного обучения является глубокое обучение, основанное на нейронных сетях — математических моделях, вдохновленных работой человеческого мозга [3]. Эти модели способны выявлять сложные закономерности в данных, что делает их мощным инструментом.

Машинное обучение опирается на два ключевых подхода: контролируемое ("с учителем") и неконтролируемое ("без учителя") обучение. Контролируемое обучение использует размеченные данные, где каждой входной переменной соответствует известное выходное значение. Этот подход требует участия человека для подготовки данных и применяется в задачах классификации (например, распознавание объектов) и регрессии (прогнозирование значений).
Неконтролируемое обучение работает с немаркированными данными, где выходные переменные неизвестны. Алгоритмы самостоятельно выявляют скрытые закономерности, группируя данные по сходству (кластеризация) или находя зависимости между признаками.

1.1. Методы Кластеризации

Метод k-средних

Метод k-средних (k-means) представляет собой один из наиболее распространенных алгоритмов кластеризации в машинном обучении[6]. Его основная задача заключается в разделении набора данных на заданное количество кластеров (k), где объекты внутри каждого кластера обладают высокой степенью сходства, в то время как объекты из разных кластеров

существенно различаются. Математически это выражается через минимизацию суммы квадратов расстояний от точек до центров их кластеров:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2,$$

где C_i обозначает -й кластер, а μ_i - его центр тяжести. Эта целевая функция, называемая критерием внутрикластерной дисперсии, служит мерой качества кластеризации

Метод Mean-Shift

Алгоритм Mean-Shift представляет собой мощный инструмент кластеризации, основанный на концепции ядерного сглаживания и анализе плотности распределения данных [8]. В отличие от k-means, который жестко разделяет пространство на кластеры, Mean-Shift рассматривает пространство данных как вероятностную плотность, где кластеры соответствуют локальным максимумам этой плотности.

Основная идея алгоритма заключается в последовательном смещении точек в направлении увеличения плотности данных. Этот процесс можно представить как "притяжение" точек к модам распределения, которые и становятся центрами кластеров. Математически смещение точки x вычисляется по формуле:

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)} - x$$

где $N(x)$ - окрестность точки x, а k - ядерная функция, определяющая вес соседних точек.

Ключевым параметром алгоритма является радиус окна (bandwidth), который определяет размер области поиска локальных максимумов плотности. Этот параметр существенно влияет на результаты кластеризации: слишком маленькое значение приведет к выделению множества мелких кластеров, а слишком большое - к объединению всех данных в один кластер.

В современных реализациях часто используется адаптивный bandwidth, который автоматически подстраивается под локальную плотность данных.

Метод главных компонент (PCA)

Метод главных компонент представляет собой мощный инструмент анализа данных, основанный на линейном ортогональном преобразовании. Основная цель PCA — выявление наиболее информативных направлений в данных и последующее проецирование на эти направления с минимальной потерей информации [6]. Алгоритм особенно полезен при необходимости устранения мультиколлинеарности, сжатия данных или их визуализации в пространстве меньшей размерности.

Математическая основа метода опирается на сингулярное разложение матрицы данных или вычисление собственных векторов ковариационной матрицы. Главные компоненты представляют собой направления максимальной дисперсии данных и вычисляются как решение задачи оптимизации:

$$w_{(1)} = arg\max_{\|w\|=1} \left(\sum_i (x_i * w)^2 \right)$$

где $w_{(1)}$ соответствует первой главной компоненте, дающей максимальную дисперсию проекции данных.

1.2. Методы классификации

Метод k-ближайших соседей (k-NN) является классическим алгоритмом контролируемого обучения, используемый для задач классификации и регрессии [13]. Основным преимуществом k-NN является его простота и высокая точность при наличии хорошо размеченных данных. Однако алгоритм чувствителен к масштабу признаков и к шумам в данных. Для эффективной работы требуется предварительная нормализация данных, а вычислительные затраты на этапе классификации растут с увеличением размера выборки [14].

Его ключевая идея заключается в отнесении новой точки данных к тому классу, который преобладает среди её k ближайших соседей в пространстве признаков.

Метод Random Forest (случайный лес) — это ансамблевый алгоритм машинного обучения, который строит множество решающих деревьев на случайных подвыборках данных и случайных подмножествах признаков, а затем объединяет их предсказания для повышения точности и устойчивости модели.

Среди множества архитектур нейронных сетей, применяемых для распознавания речи, можно выделить многослойный перцептрон (MLP) Типичная нейронная сеть состоит из входного, скрытого и выходного слоев [16]. Сначала данные поступают на входной слой, где происходит обнаружение признаков [17]. Затем скрытый слой (слои) анализирует и обрабатывает эти входные признаки, а выходной слой отображает конечный результат (рис. 2).

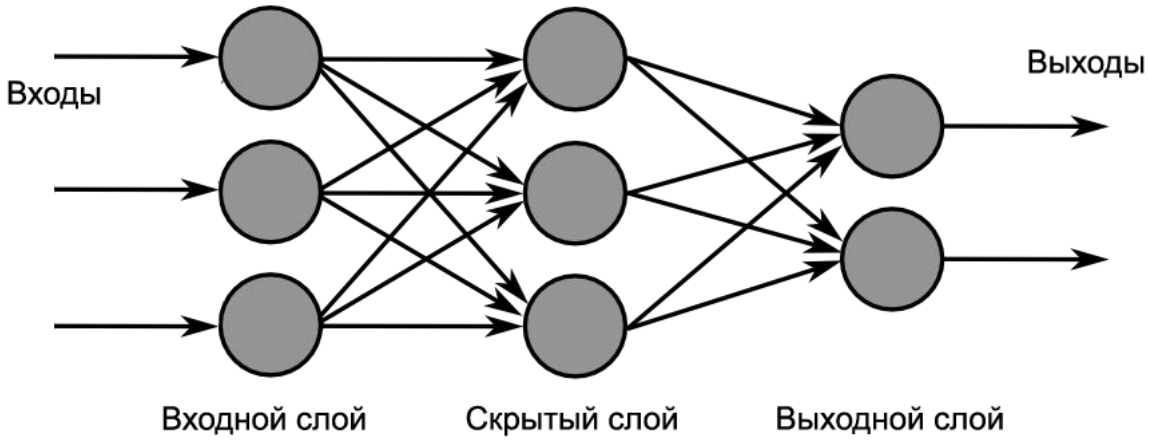

Рисунок 1. Структурная схема нейронной сети с входным, скрытым и выходным слоями.

Глава 2. Основные практические результаты

В рамках практической части исследования рассматривается применение классических методов машинного обучения и нейросетевых подходов на двух стандартных наборах данных. Основной задачей нашего исследования является решение проблемы классификации на этих датасетах.

Первый используемый набор данных - "Iris". Это классический датасет, содержащий информацию о 150 образцах цветков ириса, относящихся к трем видам: Iris setosa, Iris versicolor и Iris virginica. Набор включает четыре числовых признака: длину и ширину чашелистика (sepal length и sepal width), длину и ширину лепестка (petal length и petal width). Целевая переменная представляет собой категориальную величину с тремя классами, соответствующими видам ириса. В нашем исследовании этот датасет будет использоваться для решения задачи многоклассовой классификации, что позволит сравнить эффективность различных алгоритмов машинного обучения.

Второй анализируемый набор данных - "Kidney-stone", посвященный урологическим исследованиям. Он содержит информацию о 90 пациентах с камнями в почках, включая семь признаков: удельный вес мочи (gravity), уровень pH (ph), осмоляльность (osmo), электропроводность (cond), концентрацию мочевины (urea), концентрацию кальция (calc) и бинарный показатель наличия камней (target). В нашем исследовании основной задачей для этого датасета будет бинарная классификация - предсказание наличия или отсутствия камней в почках на основе предоставленных медицинских показателей.

Результаты кластеризации

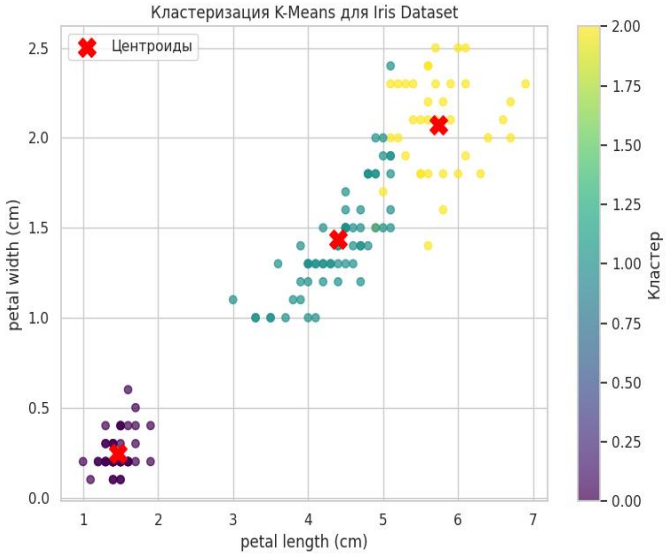

Рисунок 4. Кластеризация на Iris Dataset методом k-средних.

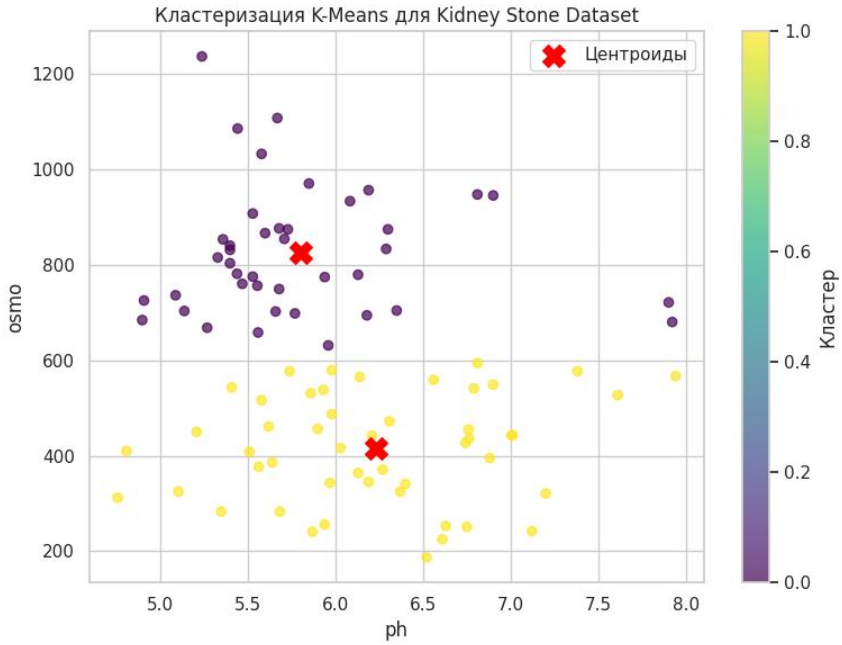

Рисунок 5 . Кластеризация на Kidney Stone Dataset методом k-средних.

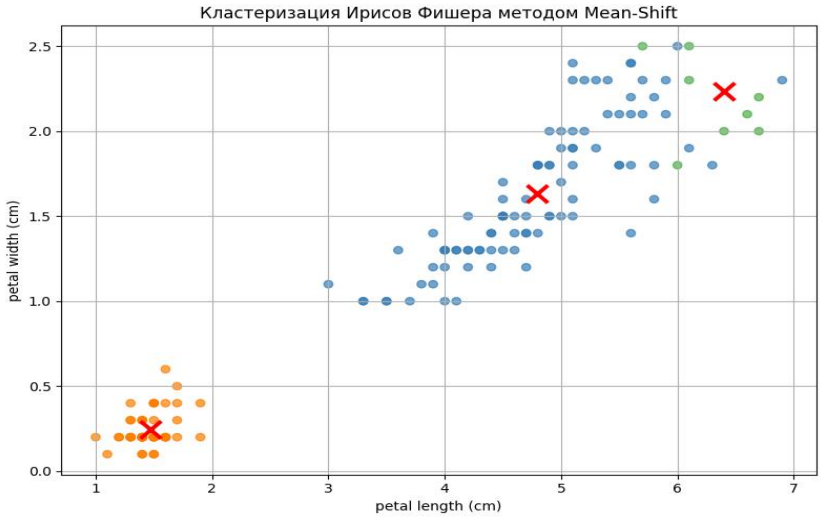

Рисунок 6. Кластеризация методом Mean-Shift на Iris Dataset.

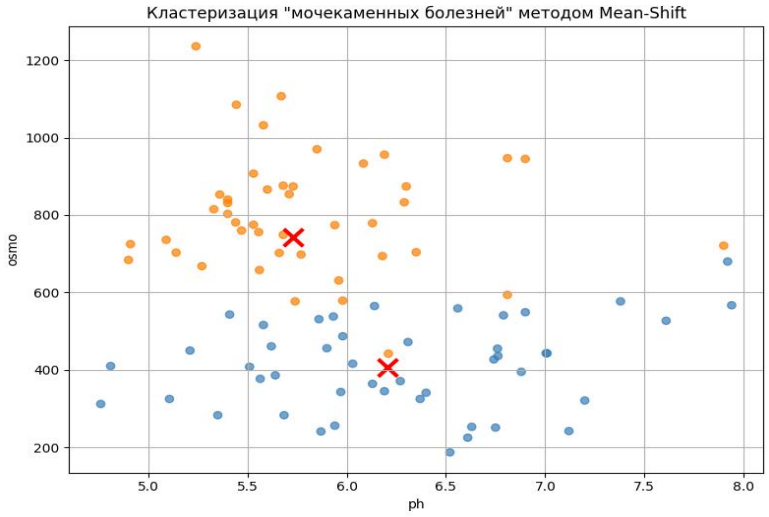

Рисунок 7. Кластеризация методом Mean-Shift на Kidney Stone Dataset.

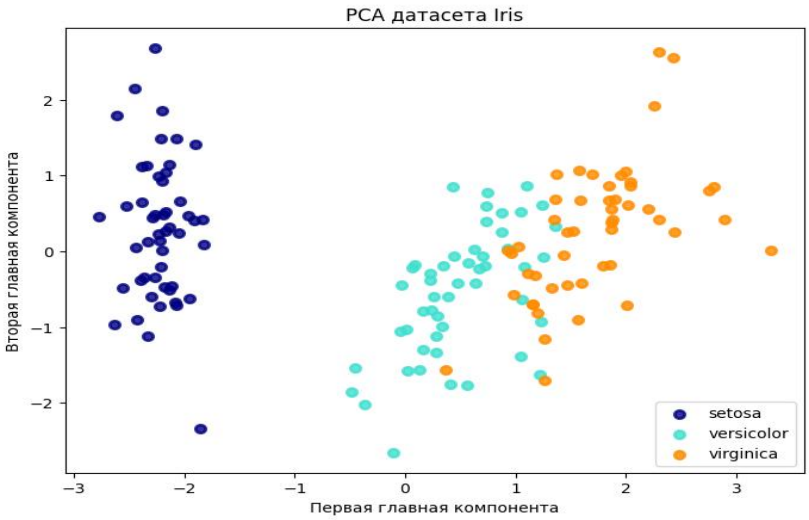

Рисунок 8. Кластеризация Iris методом PCA: проекция на первые две главные компоненты (PC1 и PC2).

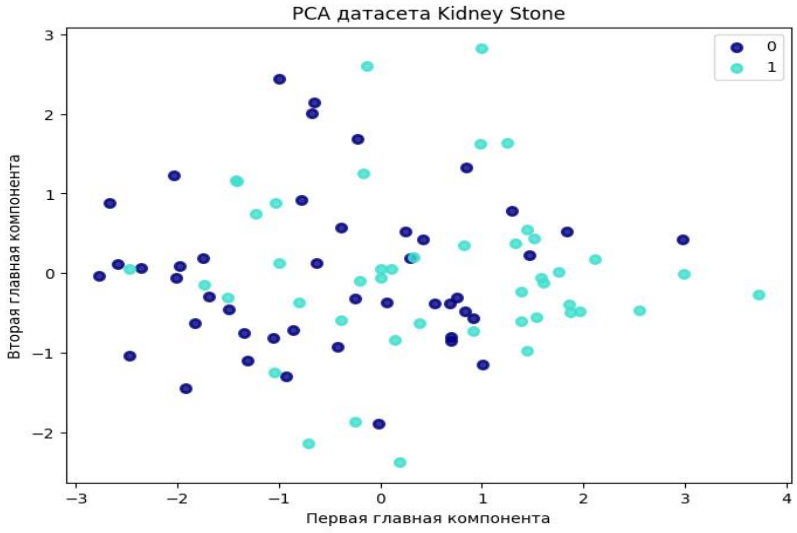

Рисунок 9. Кластеризация методом PCA на многомерном массиве данных Kidney Stone Dataset на первые две главные компоненты (PC1 и PC2).

Результаты классификации

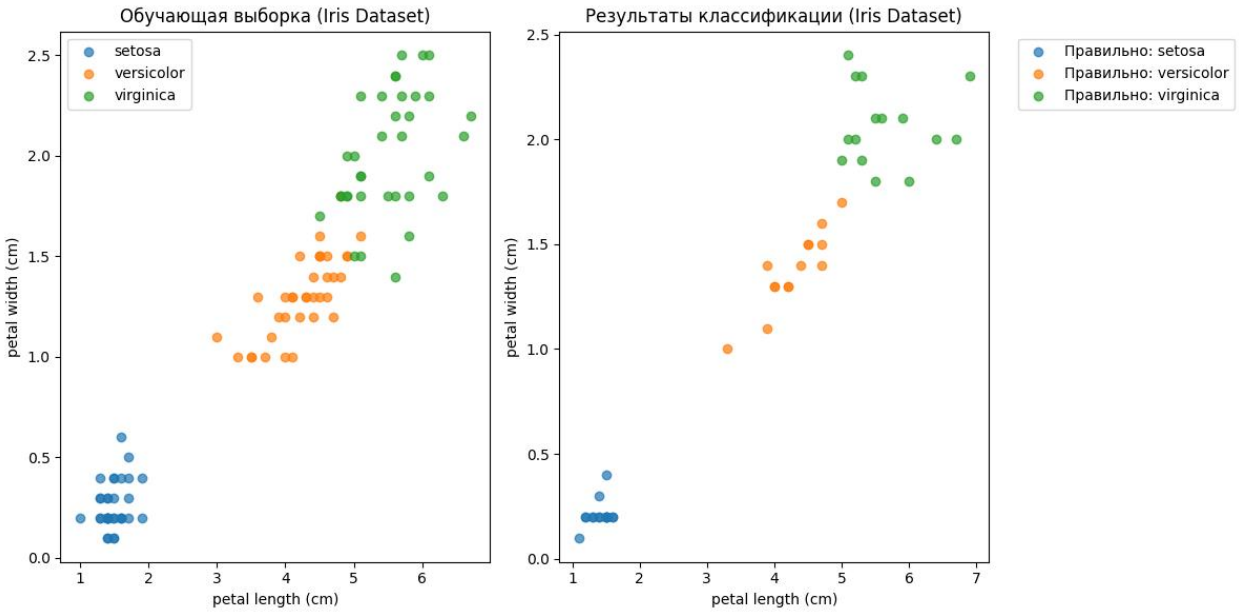

Рисунок 10. Визуализация обучающей выборки (слева) и результатов классификации методом k-ближайших соседей на Iris Dataset (справа).

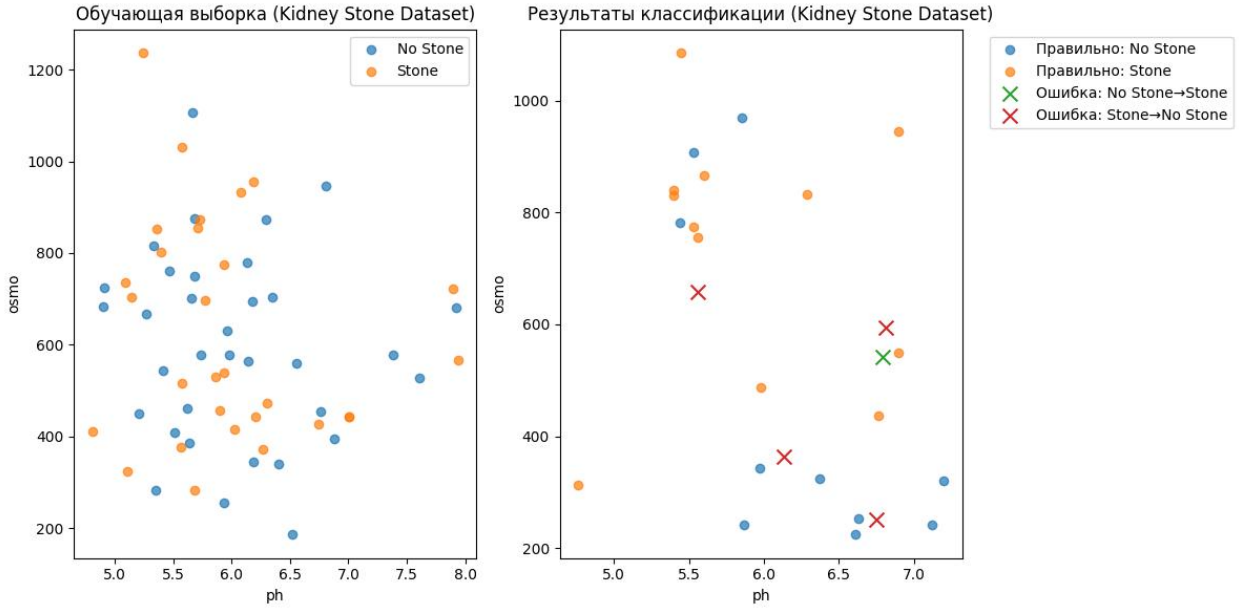

Рисунок 11. Визуализация обучающей выборки (слева) и результаты классификации методом k-ближайших соседей на Kidney Stone Dataset

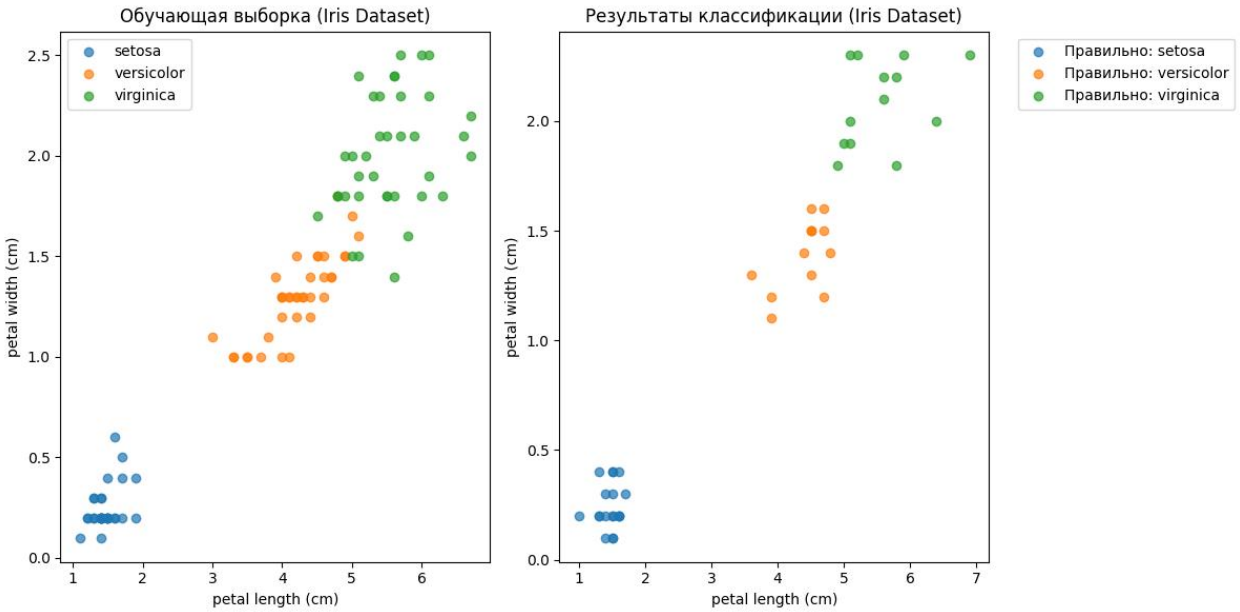

Рисунок 12. Распределение обучающей выборки по признакам (Petal_width и Petal_length) на данных Iris – слева, справа результаты классификации Random forest.

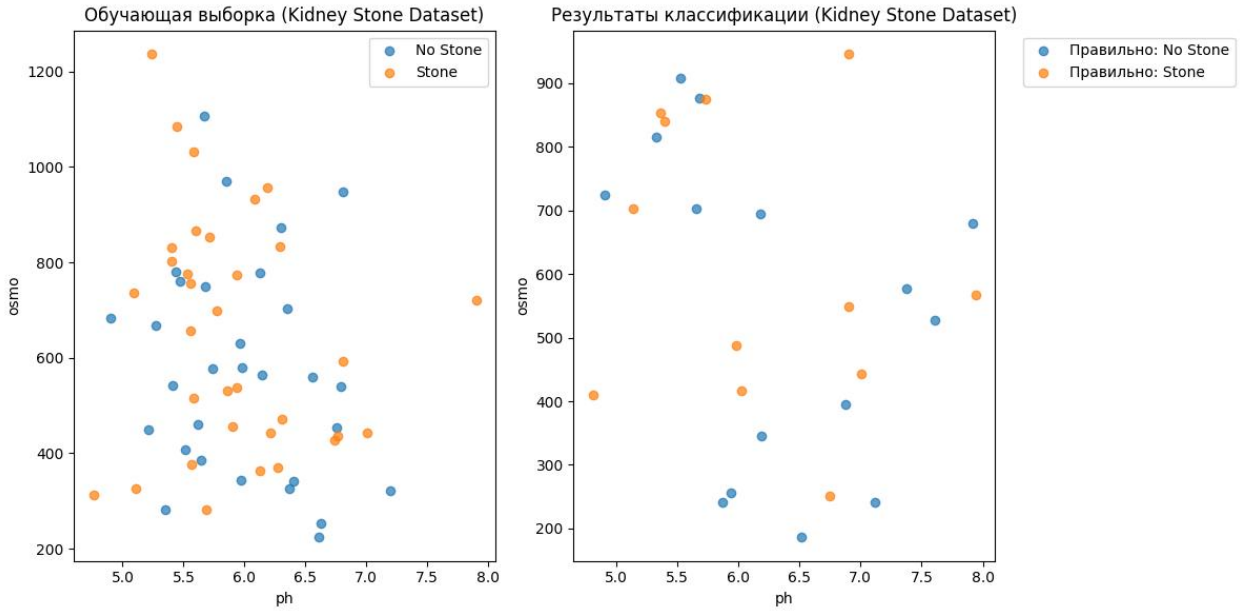

Рисунок 13. Распределение обучающей выборки по первым двум признакам (слева) и результаты классификации Random forest (справа) для Kidney Stone Dataset.

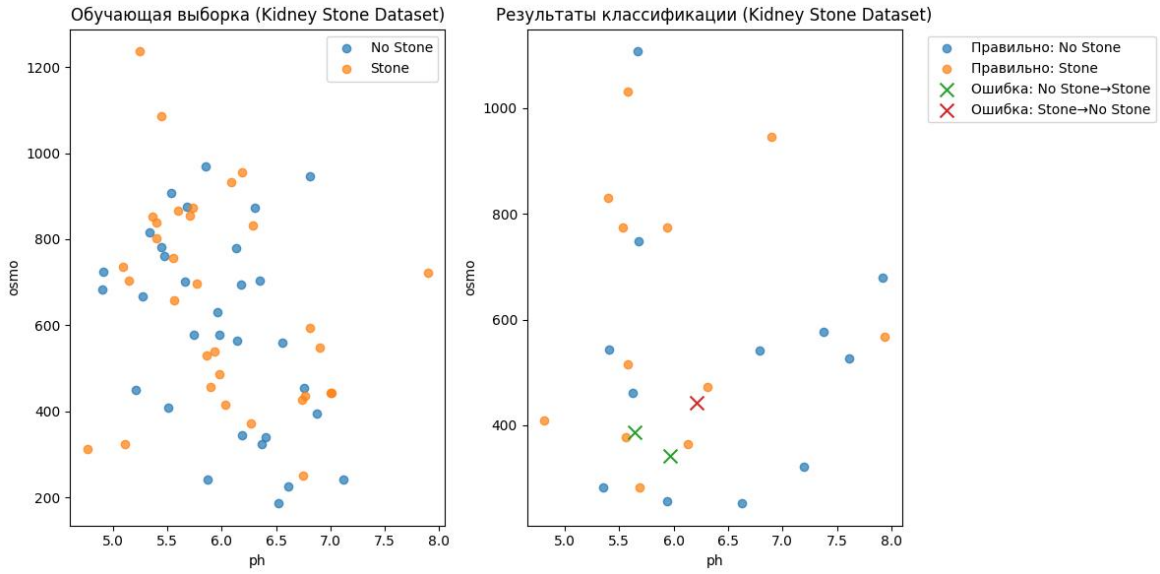

Рисунок 14. Распределение данных обучающей выборки Kidney Stone по третьему и четвертому признакам – слева, справа результаты классификации MLP.

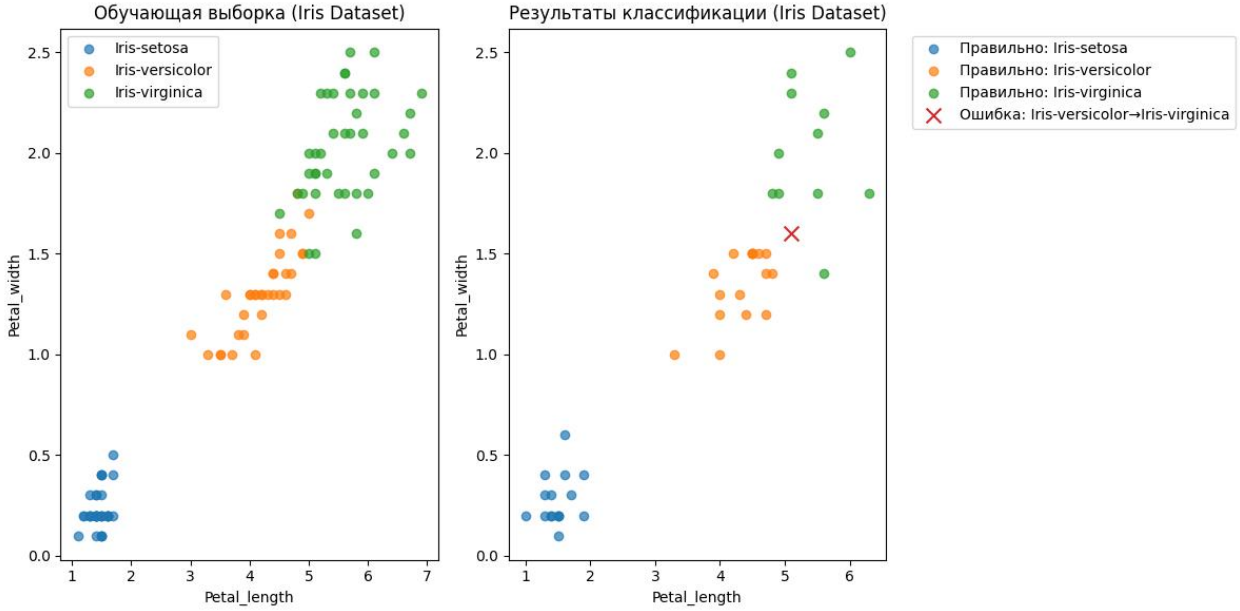

Рисунок 15. Распределение данных обучающей выборки Iris Dataset по третьему и четвертому признакам – слева, справа результаты классификации MLP.

Заключение

В ходе исследования были применены различные методы машинного обучения для анализа двух наборов данных: Iris Fisher's и Kidney Stone Dataset. Каждый из рассмотренных алгоритмов показал разную эффективность в зависимости от характера данных и поставленной задачи. Mean-Shift продемонстрировал способность автоматически определять количество кластеров, но его точность оказалась сравнительно низкой (54,0–72,0%). Это связано с чувствительностью к выбору параметра `bandwidth` и возможным перекрытием кластеров. K-средних показал высокую точность (89,3%) на данных Iris, но низкую (53,3%) на Kidney Stone Dataset, что указывает на необходимость предварительного анализа структуры данных и выбора оптимального числа кластеров. PCA позволил эффективно визуализировать данные в двумерном пространстве, сохраняя значительную долю дисперсии. Однако для медицинских данных (Kidney Stone Dataset) разделимость классов оказалась менее выраженной, что может потребовать дополнительных методов обработки. KNN показал высокую точность (100% для Iris, 81,4% для Kidney Stone Dataset), но его эффективность сильно зависит от выбора метрики расстояния и параметра k. Random Forest показал 100% точность на Iris Dataset благодаря явной разделимости setosa (Рисунок 12), а также слабому перекрытию versicolor и virginica, которое алгоритм корректно учел, на Kidney Stone Dataset Random Forest также показал 96,8% точность, что может указывать на идеальную разделимость классов для выбранных признаков. Многослойный перцептрон (MLP) продемонстрировал хорошую обобщающую способность (97,8% для Iris, 88,9% для Kidney Stone Dataset), однако на медицинских данных наблюдалось переобучение, что требует дополнительной регуляризации. Таким образом, выбор метода машинного обучения должен основываться на природе данных, требуемой точности и интерпретируемости результатов.