



Комплексный подход к преподаванию основ машинного и глубокого обучения: методы и практические инструменты

Дмитрий Викторович Власов

Российский государственный педагогический университет им. А. И. Герцена,
Санкт-Петербург, Россия
dmitry-v-vlasov@mail.ru, <https://orcid.org/0000-0002-7534-8576>

Аннотация: Обучение современным информационным технологиям невозможно представить без освоения подходов машинного и глубокого обучения, широко применяемых для анализа и обработки больших данных. Эти подходы активно используются для решения таких задач, как регрессия, прогнозирование, классификация и кластеризация, которые имеют высокую значимость для образовательных программ инженерных направлений. Рассматривается структура курсов «Основы машинного обучения» и «Математические основы глубокого обучения», включающая ключевые алгоритмы, методики и практические инструменты, способствующие развитию компетенций студентов в области анализа данных и построения моделей искусственного интеллекта.

Ключевые слова: машинное обучение, глубокое обучение, образовательная аналитика, прогнозирование в образовании, цифровые технологии, образовательные данные, нейронные сети, предиктивная аналитика, оптимизация учебного процесса, компетенции студентов.

Для цитирования: Власов Д. В. Комплексный подход к преподаванию основ машинного и глубокого обучения: методы и практические инструменты // Известия Балтийской государственной академии рыбопромышленного флота. 2025. № 1(71). С. 188–197.

Современные образовательные программы в сфере информационных технологий требуют сочетания глубоких теоретических знаний и практических навыков, особенно в областях, связанных с машинным обучением и анализом данных [1, 2]. Эти дисциплины являются базисом для разработки систем искусственного интеллекта и автоматизации, применяемых в разнообразных отраслях – от медицины и финансов до рекомендательных платформ и интеллектуальных решений.

Образовательные технологии с использованием искусственного интеллекта приобретают все большее значение. Одним из ключевых аспектов в этой области является образовательная аналитика [3]. Она охватывает процессы сбора, анализа и интерпретации данных цифрового взаимодействия студентов и преподавателей, направленные на оптимизацию образовательного процесса.

Цифровой след становится инструментом для построения индивидуальных траекторий обучения, совершенствования методик преподавания и повышения академической успеваемости. Образовательная аналитика подразделяется на четыре категории: описательная, диагностическая, предсказательная и предписывающая. Эти направления неразрывно связаны с алгоритмами машинного обучения и методами статистической обработки данных, что обеспечивает эффективную поддержку принятия решений и автоматизацию образовательных процессов.

При обучении основам машинного обучения в педагогическом вузе ключевое внимание уделяется освоению методов анализа данных, с акцентом на описательную аналитику. Этот аспект подготовки помогает студентам не только изучать динамику своей учебной деятельности, но и развивать способность к самостоятельному анализу образовательных процессов. Практическая работа с данными, такими как частота выполнения заданий, участие в онлайн-обсуждениях и успехи в изучении алгоритмов, формирует у будущих специалистов навыки выявления трендов и обнаружения отклонений в их учебной активности.

Для преподавателей это становится способом визуально продемонстрировать возможности анализа данных, тогда как студенты получают инструменты, которые можно применять как в образовательной сфере, так и в других профессиональных контекстах. Методы описательной аналитики, таким образом, превращаются из простого способа оценки вовлеченности в мощный

механизм, позволяющий учащимся оценивать свои результаты, корректировать учебные стратегии и укреплять уверенность в работе с информацией и ее интерпретацией.

Диагностическая аналитика служит инструментом, который позволяет студентам глубже осмыслить природу своих результатов, выявляя скрытые причины и взаимосвязи. Она действует как своеобразный «диагностический сканер» для данных, помогая определить, какие факторы оказывают ключевое влияние на результаты моделей. Осваивая эти подходы, студенты учатся использовать статистические методы для анализа структуры данных, идентифицировать значимые корреляции, кластеризовать и классифицировать данные, а также оценивать их близость.

Например, при изучении неудач модели в прогнозировании, они могут обнаружить, что ошибки связаны с отсутствием важных переменных или низким качеством выборки. Такая аналитика позволяет не просто зафиксировать проблему, но и понять, что ее корень может лежать, например, в игнорировании сезонных колебаний или в наличии значительных выбросов в данных.

Ещё один пример – если группа студентов систематически демонстрирует слабые результаты при изучении определенной темы, например, методов деревьев решений, диагностика может указать на недостаток качественных практических задач или неэффективность выбранного подхода к объяснению материала. Таким образом, использование методов диагностической аналитики не ограничивается выявлением слабых мест – она становится ключом к глубокому пониманию причинно-следственных связей и возможных улучшений.

Эти навыки критически важны для будущих специалистов в области машинного обучения, поскольку позволяют не только корректировать проблемы, но и предотвращать их на стадии проектирования, формируя аналитический подход к решению задач.

Прогнозная (предиктивная) аналитика играет центральную роль в обучении основам машинного обучения и глубокого обучения, позволяя предугадывать события на основании анализа данных. Этот подход объединяет методы статистического анализа, моделирования и алгоритмов машинного обучения, предоставляя студентам мощные инструменты для предсказания ключевых образовательных аспектов. Например, использование этих методик позволяет определить темы, которые требуют углубленного изучения, или выявить области, где студенты могут столкнуться с трудностями, чтобы вовремя скорректировать учебный план или перераспределить временные ресурсы.

Одним из практических примеров является анализ текущей и прошлой активности студентов, на основе которого можно предсказать, кто из обучающихся находится в группе риска снижения успеваемости. Это дает возможность заранее предложить индивидуальную помощь, организовать дополнительные занятия или консультации.

Еще один пример – оценка эффективности внедрения новых методик обучения: анализируя данные о результатах предыдущих изменений, можно прогнозировать, какие из них окажут наибольшее положительное влияние на образовательные результаты. Таким образом, прогнозная аналитика не только способствует углублению понимания математических моделей, но и формирует у студентов навыки стратегического планирования, помогая им адаптироваться к требованиям учебного процесса и повышать свою успеваемость.

Наконец, прескриптивная аналитика открывает широкие возможности для создания адаптивных учебных стратегий, позволяя не только анализировать образовательные данные, но и генерировать конкретные рекомендации для их улучшения [4]. Этот подход, основанный на алгоритмах машинного обучения и искусственного интеллекта, способствует индивидуализации учебного процесса и его оптимизации.

Например, анализируя академическую активность и успеваемость студентов, системы аналитики могут выявлять проблемные темы и предлагать углубленные учебные материалы или дополнительные модули для самостоятельной работы.

Системы прескриптивной аналитики также могут использоваться для формирования учебных планов с учетом индивидуальных потребностей студентов и их предрасположенности к определенным формам обучения. Например, если группа учащихся сталкивается с затруднениями в освоении темы глубоких нейронных сетей, аналитические системы могут рекомендовать разработку интерактивных симуляторов, проведение воркшопов или изменение последовательности изучения тем для улучшения усвоения материала.

Еще одна важная область применения – динамическая адаптация учебной нагрузки. Например, системы могут оценивать уровень сложности текущих заданий и автоматически подбирать материалы, соответствующие текущим знаниям и навыкам студентов. Если модель показывает, что группа студентов готова к более сложным задачам, преподаватели могут использовать рекомендации для увеличения



сложности практических кейсов, одновременно поддерживая баланс с учетом индивидуальных возможностей.

Прескриптивная аналитика также может быть использована для улучшения организации командной работы. Анализируя взаимодействия студентов в групповых проектах, такие системы способны оптимизировать распределение ролей и задач в командах, повышая продуктивность и снижая дисбаланс в нагрузке. В результате студенты получают возможность максимально раскрыть свой потенциал, а преподаватели – инструменты для быстрой и эффективной адаптации образовательного процесса, что особенно актуально в динамичных условиях современного обучения.

Освоение методов описательной, предиктивной, диагностической и прескриптивной аналитики в рамках образовательных курсов по машинному и глубокому обучению не только способствует совершенствованию образовательного процесса, но и закладывает прочную базу для формирования у студентов всестороннего понимания принципов и применения технологий искусственного интеллекта.

Ключевым аспектом успешного обучения методам искусственного интеллекта является не только практическое освоение инструментов, но и глубокое понимание теоретических основ обработки больших данных. Студенты должны изучать математические основы, включая линейную алгебру, теорию вероятностей и статистику, которые лежат в основе тестирования гипотез, выбора метрик качества моделей и задач оптимизации, таких как градиентный спуск и его модификации [5, 6]. Эти знания являются основополагающими для построения и настройки сложных моделей машинного обучения и искусственного интеллекта [6].

В процессе обучения важно сосредоточиться как на углубленных методах, так и на базовых техниках исследовательского анализа данных (EDA). Это включает визуализацию данных с использованием современных библиотек, таких как Matplotlib, Seaborn или Plotly [7], устранение пропущенных значений (например, с помощью методов импутации) и изучение взаимосвязей между переменными, включая корреляционный и причинно-следственный анализ. Особое внимание следует уделять таким аспектам, как выявление выбросов, балансировка несбалансированных классов и стандартизация данных, так как эти процессы критически важны для успешной работы моделей [5].

Кроме того, студенты должны быть знакомы с современными практиками, такими как использование библиотек автоматизации обработки данных, включая Scikit-learn, Pandas. Это позволяет минимизировать рутинные задачи и сосредоточиться на интерпретации результатов. Расширяя свои навыки, обучающиеся смогут не только корректно подготавливать данные, но и эффективно интегрировать их в более сложные процессы, такие как построение и оптимизация моделей искусственного интеллекта [6, 8].

Ключевым аспектом успешного обучения методам искусственного интеллекта является не только практическое освоение инструментов, но и глубокое понимание теоретических основ обработки больших данных. Студенты должны изучать математические основы, включая линейную алгебру, теорию вероятностей и статистику, которые лежат в основе тестирования гипотез, выбора метрик качества моделей и задач оптимизации, таких как градиентный спуск и его модификации [5, 6]. Эти знания являются основополагающими для построения и настройки сложных моделей машинного обучения и искусственного интеллекта.

В процессе обучения важно сосредоточиться как на углубленных методах, так и на базовых техниках исследовательского анализа данных (EDA). Это включает визуализацию данных с использованием современных библиотек, таких как Matplotlib, Seaborn или Plotly [7], устранение пропущенных значений (например, с помощью методов импутации) и изучение взаимосвязей между переменными, включая корреляционный и причинно-следственный анализ. Особое внимание следует уделять таким аспектам, как выявление выбросов, балансировка несбалансированных классов и стандартизация данных, так как эти процессы критически важны для успешной работы моделей [5].

Кроме того, студенты должны быть знакомы с современными практиками, такими как использование библиотек автоматизации обработки данных, включая Scikit-learn, Keras, Tensorflow [8], PyTorch [9], PyCaret [10] и другими. Это позволяет минимизировать рутинные задачи и сосредоточиться на интерпретации результатов. Расширяя свои навыки, обучающиеся смогут не только корректно подготавливать данные, но и эффективно интегрировать их в более сложные процессы, такие как построение и оптимизация моделей искусственного интеллекта [6, 8].

Учебная программа для двух рассматриваемых дисциплин включает детальное изучение основополагающих моделей машинного обучения, таких как линейная регрессия, логистическая

регрессия, деревья решений и их ансамбли, включая случайный лес и градиентный бустинг. Кроме того, студенты осваивают начальные модели глубокого обучения, такие как полносвязные нейронные сети с одним или несколькими слоями. Эти алгоритмы закладывают базу для изучения более сложных подходов, таких как рекуррентные и сверточные нейронные сети, что позволяет студентам подготовиться к решению задач в областях компьютерного зрения, обработки естественного языка и анализа временных рядов [5, 6].

Для достижения максимальной эффективности практических занятий учебный процесс базируется на использовании современных технологий и инструментов. Платформа Jupyter notebooks обеспечивает интерактивную среду для выполнения кода, анализа данных и их визуализации, что делает ее незаменимым инструментом при изучении алгоритмов машинного обучения. SciKit-Learn, популярная библиотека Python, предлагает широкий набор инструментов для реализации классических моделей, предварительной обработки данных и оценки качества моделей.

Параллельно используются TensorFlow и PyTorch – мощные фреймворки, которые предоставляют гибкие средства для разработки и обучения моделей глубокого обучения, включая поддержку ускоренного вычисления на GPU, что особенно важно при обработке больших объемов данных или работе с многослойными нейронными сетями [8, 9].

Для расширения практического опыта студенты знакомятся с автоматизацией рабочих процессов машинного обучения с помощью таких инструментов, как MLflow и Weights & Biases. Эти платформы позволяют отслеживать эксперименты, визуализировать метрики моделей и управлять жизненным циклом проектов, что приближает учебный процесс к реальной профессиональной практике. Важно также акцентировать внимание на обработке данных: обучение включает такие аспекты, как использование pandas и NumPy для трансформации данных, построение визуализаций с Matplotlib и Seaborn, а также анализ данных временных рядов с помощью библиотеки statsmodels [7].

В процессе подготовки студентов по направлению ИВТ особое внимание уделяется разработке гибкого подхода к преподаванию основ машинного обучения и математических основ глубокого обучения. С учетом их загруженности учебой и работой курсы ориентированы на предоставление практических, применимых знаний, которые можно интегрировать в профессиональную деятельность. Для этого структура курсов фокусируется на задачах, максимально соответствующих реальным условиям и времени, которым располагают студенты.

На начальных занятиях по основам машинного обучения вводится ключевое понятие о том, что машинное обучение – это одна из областей искусственного интеллекта, направленная на создание алгоритмов, которые способны обучаться и совершенствоваться на основе данных. Важным акцентом становится идея автоматизации аналитических процессов: студенты изучают, как использование машинного обучения помогает существенно снизить необходимость ручного программирования, особенно при работе с большими массивами данных.

На простых примерах объясняется, как алгоритмы анализируют данные, выявляют закономерности и применяют их для прогнозирования или принятия решений. Такой подход помогает сформировать у студентов представление о том, как новые технологии могут быть использованы для решения практических задач.

Особое внимание уделяется знакомству с процессом CRISP-DM [11] (Cross Industry Standard Process for Data Mining), который является универсальной структурой для выполнения проектов анализа данных и машинного обучения. Студенты изучают шесть ключевых этапов этого процесса: определение целей бизнеса, анализ данных, их предварительная обработка, разработка моделей, их оценка и последующее развертывание.

Преподавание организовано так, чтобы студенты видели практическую применимость этого подхода, включая его итеративный характер, что позволяет учитывать изменения в данных или требованиях к проекту. Такой системный взгляд помогает будущим специалистам эффективно организовывать работу с данными, решать задачи анализа и разрабатывать модели, готовые к применению в реальных условиях.

CRISP-DM также визуализируется с помощью диаграмм и схем, которые помогают студентам понять цикличность и гибкость процесса, что особенно важно при выполнении проектов в условиях изменяющихся данных или бизнес-требований. На практических занятиях студенты выполняют задания, включающие построение полных рабочих циклов, что позволяет им закрепить полученные теоретические знания и освоить навыки их применения.



При этом делается целенаправленное сравнение CRISP-DM и Agile. Они представляют собой два разных подхода, используемых в разработке, однако каждый из них имеет свои уникальные аспекты. Agile, будучи общей методологией управления проектами в сфере программной разработки, фокусируется на гибкости, коротких циклах поставки и активном взаимодействии с заказчиком.

Этот подход идеально подходит для разработки программных продуктов, требующих частых доработок и обратной связи. В то же время CRISP-DM, специально разработанный для анализа данных и задач машинного обучения, ориентирован на детализированный и последовательный процесс обработки данных, начиная с их подготовки и заканчивая развертыванием моделей.

При объяснении различий студентам акцентируется внимание на том, что CRISP-DM предлагает четкие фазы, такие как анализ бизнес-требований, исследование данных и оценка моделей, что делает его более подходящим для Data Science. Эта структура обеспечивает целенаправленный подход к решению задач, отличающихся высокой степенью неопределенности и экспериментального характера, что часто встречается в проектах, связанных с анализом данных. Agile, напротив, подходит для более универсальных проектов, где упор делается на быстрое реагирование на изменения и адаптацию к новым требованиям.

Для закрепления этих концепций студенты выполняют практические работы, придерживаясь структуры CRISP-DM. Одним из ключевых этапов процесса является создание базовой модели (baseline model), которая служит отправной точкой для оценки качества более сложных алгоритмов. Эти модели, будучи простыми и зачастую не оптимизированными, позволяют установить минимальную планку качества, которую должны превзойти более сложные методы. Примеры включают случайное угадывание в задачах классификации или линейную регрессию без сложных преобразований данных. Использование базовых моделей помогает студентам понять, как дополнительные этапы, такие как инженерия признаков или настройка гиперпараметров, влияют на конечные результаты.

Важным элементом обучения является и практическое развертывание моделей на облачных платформах или локальных серверах [12].

Это позволяет студентам познакомиться с такими инструментами, как Docker, Kubernetes и платформами вроде Yandex.Cloud или Google Cloud, что готовит их к реалиям промышленной разработки. Таким образом, интеграция CRISP-DM с современной инфраструктурой машинного обучения становится важным звеном в подготовке студентов к профессиональной деятельности.

Дисциплина «Основы машинного обучения» делится на несколько ключевых тем и соответствующих практических заданий.

Тема 1. «Разведочный анализ данных (EDA)». Введение в курс начинается с изучения методов разведочного анализа данных, который помогает студентам понять структуру и особенности исходных данных. В рамках темы рассматриваются:

1. Визуализация данных с помощью библиотек Python (Matplotlib, Seaborn, Bokeh, HoloViews).

2. Обработка пропущенных значений.

3. Анализ выбросов и аномалий.

4. Работа с числовыми и категориальными признаками.

5. Тестирование гипотез с использованием различных статистических тестов.

Практическая работа сосредоточена на анализе данных об успеваемости студентов по математике, исследовании факторов, влияющих на оценки, устранении пропусков и выбросов, а также визуализации полученных результатов.

Тема 2. «Базовые модели машинного обучения». После освоения основ анализа данных студенты переходят к изучению базовых моделей машинного обучения. Эта тема включает:

1. Линейная и полиномиальная регрессия.

2. Метрики качества регрессионных моделей (MAE, MSE, R^2).

3. Понятия переобучения и недообучения, мультиколлинеарности и регуляризации.

4. Деревья решений и ансамблевые алгоритмы (случайные леса, бустинг, бэггинг).

5. Автоматический подбор гиперпараметров (GridSearchCV, RandomizedSearchCV, Optuna, Hyperopt).

6. Кросс-валидация и методы оценки качества моделей.

Практическая работа включает применение линейной регрессии для анализа успеваемости и переход к классификации, используя пороговые значения для преобразования задач.

Тема 3. «Временные ряды». Следующим шагом является изучение временных рядов, их анализа и моделирования. Рассматриваются:

1. Основные компоненты временных рядов (тренды, сезонность, циклы, случайные колебания).
2. Методы тестирования стационарности (тест Дики-Фуллера) и приведения рядов к стационарности.
3. Модели временных рядов (AR, MA, ARIMA, SARIMA, ARCH, GARCH).
4. Прогнозирование и оценка точности моделей (MAE, MSE, RMSE).
5. Использование нейронных сетей и деревьев решений для анализа временных рядов.

Практическое задание посвящено прогнозированию посещаемости занятий, позволяя студентам работать с реальными данными и выявлять тенденции.

Тема 4. «Кластеризация». Заключительная крупная тема посвящена кластеризации. Студенты изучают:

1. Основные понятия и цели кластеризации.
2. Алгоритмы кластеризации: k-means, DBSCAN, OPTICS.
3. Оценка качества кластеризации (индекс силуэта, внутрикластерные расстояния).
4. Метрики расстояния (Евклидово, Манхэттенское, косинусная близость).
5. Методы снижения размерности (PCA, UMAP).
6. Применение кластеризации в реальных задачах (сегментация рынка, биоинформатика).

Практическая работа сосредоточена на сегментации студентов по анкетным данным, анализе бинарных признаков и применении подходящих алгоритмов кластеризации.

Финальная часть курса данной дисциплины посвящена выполнению самостоятельного проекта. Студенты выбирают собственный набор данных, выполняют полный цикл анализа, включая предварительную обработку, моделирование и оценку результатов. Завершающий этап проекта включает развертывание обученной модели в облаке, что позволяет студентам продемонстрировать не только технические, но и прикладные навыки работы с искусственным интеллектом. Работа может выполняться индивидуально или в командах, способствуя развитию профессиональных и коммуникационных навыков.

Практические занятия по курсу «Математические основы глубокого обучения» организованы таким образом, чтобы теория и практика гармонично сочетались. Каждое занятие начинается с краткого объяснения ключевых концепций, сопровождающегося решением задач в Jupyter notebooks. В начале курса студенты знакомятся с понятием глубокого обучения и его отличиями от классического машинного обучения, включая архитектурные и вычислительные аспекты.

Особое внимание уделяется алгоритму обратного распространения ошибки (backpropagation). Теоретический материал подробно раскрывается, включая последовательное математическое обоснование рекуррентных формул, что позволяет студентам глубже понять процесс корректировки весов в многослойных нейронных сетях.

Это также помогает осознать критическую важность аппаратного ускорения с помощью дорогостоящих современных GPU-чипов для обучения нейронных сетей за приемлемое количество времени. После изложения базового алгоритма рассматриваются примеры дополнительных слоев, таких как слой нормализации, и их влияние на общий процесс обучения, вводится обобщающее понятие обучаемых параметров (веса полносвязных слоёв – это один из примеров). Эти примеры демонстрируют, как различные параметры интегрируются в общую математическую модель.

Практические задания фокусируются на реализации базовых архитектур нейронных сетей и их обучении на небольших наборах данных. Студенты получают возможность протестировать различные гиперпараметры, экспериментируя с настройками, такими как скорость обучения и размер батча. Этот подход позволяет формировать у обучающихся понимание как математической, так и инженерной стороны глубокого обучения. Ниже изложены ключевые темы и практические задания.

Тема 1. «Логистическая регрессия как основа нейронных сетей». Курс начинается с изучения логистической регрессии, рассматриваемой как базовая модель, на которой строятся фундаментальные принципы нейронных сетей. Изучаются вопросы:

1. Различие между линейной и логистической регрессией.
2. Логит-функция и метод максимального правдоподобия.
3. Интерпретация коэффициентов, шансов, доверительных интервалов.
4. Диагностика модели: мультиколлинеарность, анализ остатков.
5. Метрики оценки качества: ROC-кривая, accuracy, precision, recall, F-мера.
6. Многоклассовая логистическая регрессия и регуляризация.



Практическое задание посвящено анализу влияния учебных ресурсов на успеваемость студентов. Исследуется связь между доступностью ресурсов, такими как библиотеки и стипендии, и вероятностью успешной сдачи экзаменов.

Тема 2. «Последовательные нейронные сети с полносвязными слоями».

После освоения логистической регрессии студенты переходят к изучению полносвязных нейронных сетей. В теме рассматриваются вопросы:

1. Архитектура последовательных нейронных сетей.
2. Структура полносвязных слоев: нейроны, веса, функции активации (ReLU, Sigmoid, Tanh).
3. Алгоритмы обучения: градиентный спуск, обратное распространение ошибки.
4. Регуляризация: L1, L2, Dropout.
5. Метрики классификации: accuracy, precision, recall, F1-мера.
6. Инструменты реализации: TensorFlow, Keras, SciKeras.

Практическая работа направлена на разработку модели для оптимизации учебного расписания, анализируя взаимосвязь между временем проведения занятий и производительностью студентов.

Тема 3. «Сверточные нейронные сети для анализа изображений». Далее студенты изучают сверточные нейронные сети (CNN), их архитектуру и применение. Изучаемые вопросы:

1. Компоненты CNN: сверточные, пулинговые, полносвязные слои.
2. Сверточные слои: фильтры, padding, stride.
3. Пулинг: максимальный и средний пулинг.
4. Многоклассовая классификация с использованием softmax.
5. Оптимизация CNN: регуляризация, использование Adam, SGD.
6. Гиперпараметры: количество фильтров, глубина архитектуры, размер батчей.

Практическая работа включает классификацию учебных материалов (диаграммы, текст, графики) с использованием модели CNN. Студенты формируют датасет из изображений, взятых из открытых источников.

Тема 4. «Самостоятельная реализация нейронной сети». Заключительный этап курса сосредоточен на самостоятельной программной реализации последовательной нейронной сети. Основные аспекты:

1. Алгоритмы прямого и обратного распространения.
2. Выбор функции активации (например, сигмоиды).
3. Интерпретация весов и их влияние на обучение.

Практическое задание состоит в классификации статуса успеваемости студентов на основе их учебных и социальных активностей. Студенты анализируют данные о посещаемости занятий, времени на обучение, использовании библиотек и образовательных платформ, создавая модель для предсказания уровня успеваемости.

Ниже приведены некоторые примеры задач, которые предлагаются студентам.

Пример 1. Разведочный анализ данных и моделирование успеваемости учеников по экзамену по математике. Задание направлено на исследование факторов, влияющих на успеваемость учащихся по математике, и построение прогнозной модели, позволяющей оценить вероятность достижения высоких баллов на выпускных экзаменах. Анализ проводится на основе предоставленного датасета, содержащего информацию о социально-экономических и образовательных характеристиках учеников, таких как возраст, пол, уровень образования родителей, тип школы и время, затрачиваемое на обучение.

Шаги проведения разведочного анализа данных (EDA):

1. Оценка качества данных:

- Проверка наличия пропущенных значений, их доли в каждом признаке.
- Анализ выбросов в числовых переменных (например, возраст, баллы за экзамен).
- Проверка категориальных признаков на наличие некорректных значений.

Статистическое резюме:

- Вычисление основных метрик (среднее, медиана, стандартное отклонение) для числовых переменных.

- Анализ распределений и выявление асимметрий (например, с помощью гистограмм).

2. Корреляционный анализ:

- Построение корреляционной матрицы для определения взаимосвязей между признаками.
- Проверка линейной зависимости между баллами за экзамен и факторами, такими как *studytime* (время на учёбу помимо школы в неделю) или *famsup* (семейная образовательная поддержка).

Визуализация данных:

- Гистограммы, боксплоты и scatter-плоты для анализа распределений и аномалий.
- Визуализация отношений между целевой переменной *score* (баллы по госэкзамену по математике) и ключевыми предикторами.

Изучаемые студентами подходы к подбору машинных моделей:

Базовая линейная регрессия:

- Модель используется для проверки основного уровня качества предсказания, используя только ключевые количественные признаки: *studytime*, *absences* (количество пропущенных занятий), *Medu* (уровень образования матери), *Fedu* (уровень образования отца).

- Интерпретация коэффициентов модели для понимания вклада каждого признака.

Модели с регуляризацией:

- Ridge и Lasso-регрессия для борьбы с мультиколлинеарностью и улучшения обобщающей способности модели.

- Полиномиальные признаки:

- Введение взаимодействий между ключевыми признаками (например, влияние *studytime* и *famsup*).

- Построение полиномиальных моделей для учета нелинейных взаимосвязей.

Автоматизация подбора гиперпараметров:

- Использование алгоритмов перебора GridSearchCV или RandomizedSearchCV для выбора оптимальных параметров моделей.

- Применение кросс-валидации для оценки стабильности модели.

Получаемый студентами результат. На выходе студенты создают отчет в формате Jupyter notebook, включающий код, визуализации, интерпретации и сравнительный анализ качества моделей. Итоговая модель может быть использована для прогнозирования результатов экзаменов и выработки рекомендаций по повышению успеваемости, таких как увеличение времени на самостоятельное обучение или предоставление дополнительной образовательной поддержки.

Пример 2. Анализ эмоционального состояния и учебной мотивации студентов. Студентами составляется анкета для анализа эмоционального состояния и учебной мотивации студентов разработана с целью выявления факторов, влияющих на успеваемость, эмоциональное состояние и мотивацию учащихся в процессе обучения. Вопросы анкеты охватывают ключевые аспекты учебной деятельности, включая уровень стресса, уверенности в себе, вовлеченность в учебный процесс, социальную поддержку и результаты обучения.

Особое внимание при этом уделено сбору данных, связанных с преодолением препятствий, использованием учебных ресурсов и влиянием окружения. Такой подход позволяет получить самим студентам всестороннюю картину условий, в которых они обучаются, и выявить возможные точки для улучшения своего собственного образовательного процесса.

Эмоциональное состояние:

1. Уровень стресса перед экзаменами (по шкале от 1 до 5).
2. Уровень уверенности в своих силах перед экзаменами (по шкале от 1 до 5).
3. Частота выгорания: ежемесячно, раз в семестр, редко, никогда.
4. Причины стресса: сложность материала, нехватка времени, личные обстоятельства, давление со стороны родителей или преподавателей.
5. Методы борьбы со стрессом: физическая активность, медитация, поддержка близких, игнорирование проблемы.
6. Оценка качества сна во время подготовки к экзаменам (по шкале от 1 до 5).
7. Частота отсутствия аппетита или переедания из-за стресса: часто, иногда, редко, никогда.

Учебная мотивация:

1. Основные мотивы учебы (выберите не более двух): карьерный рост, интерес к предмету, социальное давление, достижение личных целей, стипендия.
2. Уровень вовлеченности в учебный процесс (по шкале от 1 до 5).
3. Влияние преподавателя на мотивацию: положительное, нейтральное, отрицательное.
4. Влияние однокурсников на мотивацию: положительное, нейтральное, отрицательное.
5. Частота выполнения домашних заданий вовремя (в процентах).
6. Использование дополнительных ресурсов: пособия, онлайн-курсы, репетиторы, семинары.



Социальные факторы:

1. Наличие постоянного рабочего места для учебы дома: да, нет.
2. Частота обсуждения учебных вопросов с семьей: ежедневно, раз в неделю, редко, никогда.
3. Уровень поддержки семьи (по шкале от 1 до 5).
4. Участие в студенческих клубах: да, нет.
5. Время на общественную деятельность (часы в неделю).

Результаты оценки анкеты нейросетью (множественные выходные слои нейросети):

1. Достижение целей за семестр: да, частично, нет.
2. Средние баллы по ключевым дисциплинам.
3. Частота получения высоких оценок: часто, иногда, редко.
4. Основные успехи за семестр (короткий свободный текст).
5. Препятствия к достижениям: сложность материала, нехватка времени, отсутствие мотивации, личные проблемы.

Контекстные данные:

1. Специальность и курс обучения.
2. Уровень загруженности программой (по шкале от 1 до 5).
3. Часы самостоятельного обучения в неделю.
4. Участие в оплачиваемой работе: да (часы в неделю), нет.

Реализация задания студентами. Для реализации задания студенты разрабатывают и проводят опрос среди своей учебной группы или других студентов университета. Опрос может быть выполнен как в бумажной, так и в электронной форме, с использованием платформ для сбора данных, таких как Yandex Forms или аналогичные сервисы. Студенты самостоятельно собирают данные, проверяют их на корректность и отсутствующие значения, проводят предварительную обработку, включая кодирование категориальных переменных и нормализацию числовых данных.

Далее собранный датасет используется в рамках лабораторных работ для построения и анализа моделей машинного обучения, таких как полносвязные нейронные сети с множественными выходными слоями для предсказания различных аспектов успеваемости студентов.

Результаты работы. По итогам выполнения задания студенты создают структурированный датасет, который включает собранные анкеты в табличной форме.

Этот датасет служит основой для обучения моделей, которые предсказывают ключевые показатели, такие как вероятность достижения целей, средние баллы по дисциплинам, частота высоких оценок и основные препятствия к успеху.

В ходе работы студенты анализируют важность признаков, интерпретируют результаты моделей и формируют рекомендации по улучшению образовательного процесса. Дополнительно они приобретают навыки работы с данными, проектирования опросов, предобработки данных и построения предсказательных моделей, что способствует развитию их компетенций в области анализа данных и машинного обучения.

Обучение основам машинного обучения и математическим принципам глубокого обучения нацелено на формирование у студентов ИВТ практических компетенций в разработке и применении современных моделей искусственного интеллекта, которые находят широкое применение в образовательных задачах и за их пределами.

Список источников

1. Власова, Е. З. Большие данные в анализе ИТ-успешности преподавателей университета / Е. З. Власова, Д. В. Власов // Известия Балтийской государственной академии рыбопромыслового флота : психолого-педагогические науки. – 2024. – № 1(67). – С. 7–12.
2. Власова, Е. З. Анализ готовности преподавателей высшей педагогической школы к профессиональной деятельности в условиях электронного обучения на основе больших данных / Е. З. Власова, Д. В. Власов // Цифровая экосистема педагогического образования. Актуальные вопросы. Достижения. Инновации. – Санкт-Петербург : НИЦ АРТ, 2022. – С. 40–52.
3. Власов, Д. В. Цифровой след и перспективы его использования в высшем образовании / Д. В. Власов // Балтийский морской форум : материалы XI Международного Балтийского морского форума. В 8-ми томах, Калининград, 25–30 сентября 2023 года. – Калининград : Издательство БГАРФ, 2023. – С. 36–42.
4. Шафорост, Н. В. Анализ алгоритмов машинного обучения для разработки модели рекомендательной системы по созданию учебных планов / Н. В. Шафорост, Д. В. Власов // Современное образование: традиции и инновации. – 2023. – № 2. – С. 226–230.

5. Hastie, T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction : учебное пособие / Hastie, T., Tibshirani, R., Friedman, J. – 2-е изд. – New York : Springer, 2009. – 745 p. – ISBN 978-0-387-84858-7
6. Goodfellow, I. Deep Learning : учебное пособие / Goodfellow, I., Bengio, Y., Courville, A. – Cambridge : MIT Press, 2016. – 800 p. – ISBN 978-0-2620-3561-3.
7. McKinney, W. Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter : учебное пособие – 3-е изд. – Sebastopol : O'Reilly Media, 2022. – 579 p. – ISBN 978-1-098-10403-0.
8. Géron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems : учебное пособие / Géron, A. – 3-е изд. – Sebastopol, CA : O'Reilly Media, 2022. – 864 p. – ISBN 978-1-0981-2597-4
9. Paszke, A. PyTorch: An Imperative Style, High-Performance Deep Learning Library / A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshine, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala. – Текст : электронный // Advances in Neural Information Processing Systems. – Vol. 32: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, 8–14 декабря 2019 / Neural Information Processing Systems Foundation. – Vancouver : NeurIPS, 2019. – URL: <https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library> (дата обращения: 23.01.2025).
10. Sarangpure N. Automating the Machine Learning Process Using PyCaret and Streamlit / N. Sarangpure, S. Tamboli, J. Doye, V. Dhamde, A. Roge, S. Patle. – Текст : электронный // Proceedings of the 2nd International Conference for Innovation in Technology (INOCON), Nagpur, 24–26 ноября 2023 / Institute of Electrical and Electronics Engineers. – Nagpur : IEEE, 2023. – URL: <https://ieeexplore.ieee.org/document/10101357> (дата обращения: 23.01.2025).
11. What is CRISP DM? – URL: <https://www.datascience-pm.com/crisp-dm-2> / (дата обращения 24.01.2025).
12. Шафорост, Н. В. Развертывание модели машинного обучения с использованием современного фреймворка FastAPI / Н. В. Шафорост, Д. В. Власов // Современное образование: традиции и инновации. – 2023. – № 3. – С. 102–106.

Информация об авторе

Д. В. Власов – кандидат физико-математических наук, доцент.