УДК 004.8

*Черных В.В., кандидат технических наук, доцент, ФГБОУ ВО «Луганский государственный университет имени Владимира Даля»*

*Балалаечников А.В., старший преподаватель, ФГБОУ ВО «Луганский государственный университет имени Владимира Даля»*

# ОЦЕНКА ЭФФЕКТИВНОСТИ РАЗЛИЧНЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ПРОГНОЗИРОВАНИИ УСПЕВАЕМОСТИ СТУДЕНТОВ

*Аннотация: В статье рассматриваются возможности применения машинного обучения для прогнозирования успеваемости студентов в условиях современного образовательного пространства с постоянно растущими объемами данных. Анализируются ключевые аспекты оценки эффективности различных алгоритмов машинного обучения. Акцентируется внимание на важности адекватных метрик, строгой валидации моделей и учета этических соображений при внедрении технологий прогнозирования.*

*Ключевые слова: машинное обучение, алгоритм, прогнозирование успеваемости, регрессия, модель, категориальные признаки, обучающая выборка, тестовая выборка*

В условиях современного образовательного пространства, характеризующегося экспоненциальным ростом объема данных, машинное обучение (МО) приобретает статус мощного инструмента для получения ценных аналитических выводов [1]. Одной из наиболее перспективных областей применения МО является прогнозирование успеваемости обучающихся [2]. Точное предсказание успешности студентов в обучении может помочь преподавателям и административному аппарату учебного заведения заранее выявлять потенциально отстающих, предлагать им необходимую поддержку и оптимизировать образовательные программы с целью повышения общего уровня успеваемости.

Тем не менее, эффективность различных методов машинного обучения в решении данной задачи может значительно варьироваться. В данной статье будут проанализированы ключевые аспекты оценки эффективности основных алгоритмов МО, применяемых для прогнозирования успеваемости студентов, а также рассмотрены распространенные метрики для их объективной оценки.

Рассмотрим основные алгоритмы для прогнозирования успеваемости студентов.

**Линейная регрессия.** Линейные регрессионные модели могут служить отправной точкой для прогнозирования числовых оценок (линейная регрессия) или вероятности успешного завершения курса (логистическая регрессия). Они хорошо работают при наличии линейных зависимостей между признаками и целевой переменной (рис. 1).

Такие модели могут учитывать успеваемость по предыдущим курсам, посещенные занятия и выполненные работы, что является достаточно простым, но эффективным способом прогнозирования, который помогает выявить студентов, у которых есть проблемы с обучением, до наступления экзаменационной сессии.

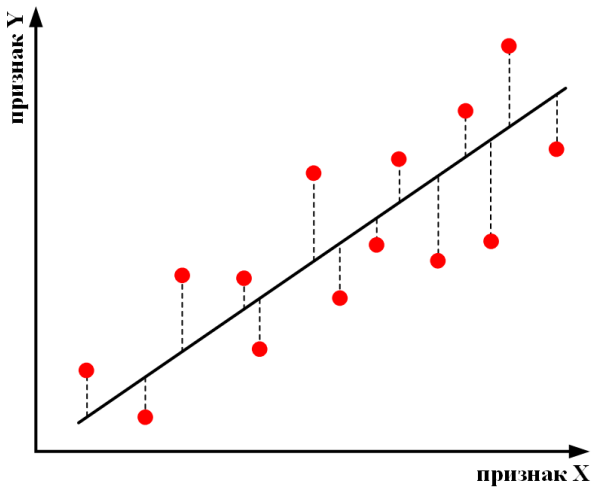


Рис.1. Графическое представление модели линейной регрессии

Для прогнозирования среднего балла студентов в третьем семестре ($Б3_{ср}$) необходимо иметь данные об их успеваемости за два предыдущих семестра ($Б1_{ср}$ за первый семестр и $Б2_{ср}$ за второй семестр соответственно). Далее следуем по алгоритму.

1) Собрать данные об успеваемости студентов, включающие $Б1_{ср}$, $Б2_{ср}$ и фактический $Б3_{ср}$ для каждого студента.

2) Подготовить данные для дальнейшего использования (например, очистить от ненужных пропусков) и при необходимости выполнить

масштабирование признаков. В нашем случае $Б1_{ср}$ и $Б2_{ср}$ – это независимые переменные (признаки), а $Б3_{ср}$ – зависимая переменная (целевая переменная).

3) Применить алгоритм линейной регрессии для обучения модели на собранных данных. Модель будет стремиться найти линейную зависимость между $Б1_{ср}$ и $Б2_{ср}$ с одной стороны, и $Б3_{ср}$ – с другой. Это можно выразить математически следующим образом:

$$Б3_{ср} = \delta_0 + \delta_1 \cdot Б1_{ср} + \delta_2 \cdot Б2_{ср},$$

где $Б3_{ср}$ – прогнозируемый средний балл за третий семестр;

$Б1_{ср}$ – средний балл за первый семестр;

$Б2_{ср}$ – средний балл за второй семестр;

$\delta_0$ – свободный член;

$\delta_1$ и $\delta_2$ – коэффициенты регрессии, которые показывают, как изменение $Б1_{ср}$ и $Б2_{ср}$ влияют на $Б3_{ср}$.

Алгоритм линейной регрессии подберёт оптимальные значения $\delta_0$, $\delta_1$ и $\delta_2$, которые сведут к минимуму ошибку между предсказанными и фактическими значениями $Б3_{ср}$ на обучающих данных.

4) После обучения модель необходимо оценить на независимом наборе данных (тестовой выборке) с использованием метрик регрессии, таких как средняя абсолютная ошибка (MAE), среднеквадратическая ошибка (MSE) или корень из среднеквадратической ошибки (RMSE). Это дает возможность оценить, насколько эффективно модель способна обобщать информацию на новые данные.

5) После того как модель продемонстрировала приемлемые результаты на тестовой выборке, её можно применять для прогнозирования $Б3_{ср}$ для новых студентов, имея значения их $Б1_{ср}$ и $Б2_{ср}$.

Линейная регрессия может быть полезным инструментом для первоначального анализа и построения базовых моделей прогнозирования успеваемости благодаря своей простоте и интерпретируемости. Однако для достижения более высокой точности часто требуется применение более сложных

алгоритмов, способных улавливать нелинейные зависимости и взаимодействия между признаками.

**Деревья решений и ансамблевые методы.** Данные методы МО способны улавливать сложные нелинейные зависимости и взаимодействия между признаками. Ансамблевые методы (например, случайный лес, градиентный бустинг) часто демонстрируют высокую точность прогнозирования за счет объединения результатов работы множества деревьев (рис. 2).

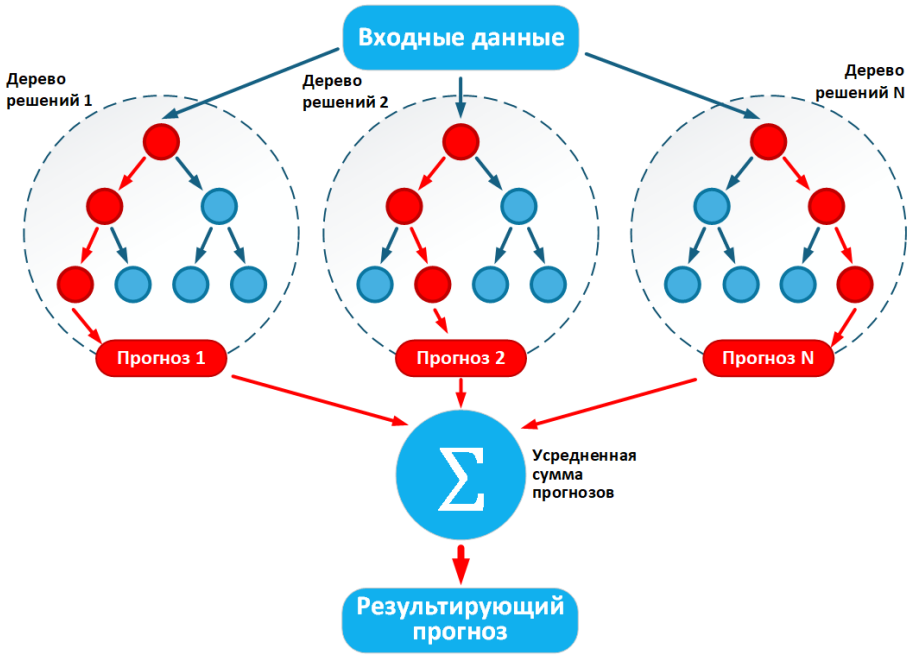


Рис. 2. Пример использования ансамблевого метода «Случайный лес»

Допустим, необходимо спрогнозировать, успешно ли студент сдаст экзамен по определенной дисциплине (прогнозирование «успешно» или «неуспешно»). Для этого имеем следующие данные о студентах:

– средний балл за предыдущие курсы ($БП_{ср}$);
– количество пропущенных занятий по дисциплине;
– результаты промежуточных контрольных работ (К1, К2);
– наличие/отсутствие дополнительной факультативной подготовки по дисциплине (бинарный признак).

Алгоритм предполагает выполнение следующих шагов.

1) Сбор и подготовка данных, которые включают все перечисленные признаки и результат экзамена («успешно» или «неуспешно»). Затем проводим

предварительную обработку данных (обработка пропусков, кодирование категориальных признаков, если они есть в большем количестве).

2) Обучение модели дерева решений на обучающей выборке. Алгоритм будет рекурсивно разбивать данные на подмножества на основе наиболее информативных признаков, стремясь создать однородные группы студентов с точки зрения результата экзамена.

3) Обучение ансамблевых моделей.

Случайный лес (Random Forest). Необходимо обучить множество независимых деревьев решений на случайных подмножествах данных и случайных подмножествах признаков. Для прогнозирования результата экзамена для нового студента каждое дерево выдает свой прогноз, и итоговый прогноз определяется путем голосования (для классификации) или усреднения (для регрессии, если нам необходимо спрогнозировать оценку).

Градиентный бустинг (Gradient Boosting). Необходимо последовательно обучить множество слабых моделей (обычно – деревьев решений), причем каждая последующая модель будет пытаться исправить ошибки, допущенные предыдущими моделями. Итоговый прогноз представляет собой усредненную сумму прогнозов всех моделей.

4) После обучения моделей необходимо оценить их производительность на независимой тестовой выборке с использованием метрик классификации, таких как точность (доля правильно классифицированных студентов), полнота (доля фактически успешных студентов, правильно предсказанных как успешные), точность (доля студентов, предсказанных как успешные, которые действительно являются успешными), F1-мера (гармоническое среднее полноты и точности) и AUC-ROC (оценивает способность модели различать классы на разных порогах классификации), а после – сравнить результаты различных моделей.

5) Выбрать лучшую модель на основе результатов оценки в целях ее дальнейшего использования для прогнозирования результата экзамена для новых студентов.

Деревья решений представляют собой интерпретируемый, но часто менее точный метод. Ансамблевые методы обычно обеспечивают более высокую точность прогнозирования успеваемости, но при этом теряют в интерпретируемости и могут быть более требовательны к вычислительным ресурсам и настройке.

**Метод опорных векторов.** Данный алгоритм машинного обучения используется как для задач классификации, так и для регрессии. Метод опорных векторов ищет гиперплоскость, которая максимально отделяет разные классы данных. Эта гиперплоскость определяется таким образом, чтобы максимизировать расстояние до ближайших точек данных, которые называются опорными векторами (рис. 3).

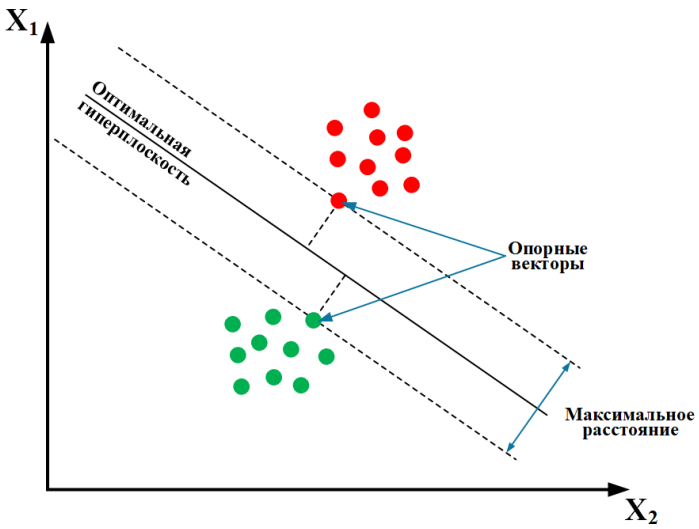


Рис. 3. Классификация при помощи метода опорных векторов

Предположим, что нам необходимо спрогнозировать итоговый балл студента по курсу «Объектно-ориентированное программирование» (числовое значение от 0 до 100). У нас имеются следующие данные о студентах:

– средний балл за курс «Технологии программирования» ($БТП_{ср}$);

– время, затраченное на выполнение домашних заданий по курсу «Объектно-ориентированное программирование» в неделю ($t_{ДЗ}$);

– результаты промежуточного контроля по курсу «Объектно-ориентированное программирование» в виде тестовых заданий (Т1, Т2).

Последовательность действий в данном случае следующая.

1) Сбор исторических данных, включающих перечисленные признаки и фактический итоговый балл по «Объектно-ориентированному программированию».

2) Провести предварительную обработку данных. Для этого обработать пропущенные значения, если такие есть. Масштабировать числовые признаки, так как метод опорных векторов чувствителен к масштабу признаков. Кодировать категориальные признаки, если они присутствуют (в нашем примере все признаки числовые).

3) Метод опорных векторов использует ядро для преобразования исходных данных в более высокое измерение, чтобы сделать их линейно разделимыми, что значительно расширяет возможности алгоритма, позволяя ему эффективно работать со сложными данными. Поэтому на данном этапе нужно выбрать тип ядра: линейное, полиномиальное, Гауссово радикальное базисное, сигмоидное и т.д. Выбор ядра зависит от предполагаемого характера зависимости между признаками и итоговым баллом.

4) Обучение модели провести на обучающей выборке, используя выбранное ядро и настраивая гиперпараметры (например, параметр регуляризации C, параметры ядра).

5) Оценить производительность обученной модели на независимой тестовой выборке с применением метрик регрессии, таких как средняя абсолютная ошибка (MAE), среднеквадратическая ошибка (MSE) или корень из среднеквадратической ошибки (RMSE). Затем выполнить кросс-валидацию для оценки обобщающей способности.

6) После оценки и настройки модели, ее можно использовать для прогнозирования итогового балла по курсу «Объектно-ориентированное программирование» для новых студентов, основываясь на их значениях признаков ($БТП_{ср}$, $t_{ДЗ}$, Т1, Т2).

Метод опорных векторов является мощным алгоритмом машинного обучения, который может быть эффективен для прогнозирования успеваемости студентов, особенно в задачах классификации с нелинейными зависимостями.

Однако его производительность сильно зависит от правильного выбора ядра и настройки гиперпараметров, а интерпретация модели может быть затруднительной. При работе с большими объемами данных могут возникнуть вычислительные сложности. Поэтому при выборе метода необходимо учитывать характеристики данных и требования к интерпретируемости модели.

**Нейронные сети.** Глубокие нейронные сети обладают высокой способностью к обучению сложным закономерностям в данных, особенно при обработке больших объемов информации. Структура нейронной сети состоит из нескольких ключевых компонентов, которые взаимодействуют между собой для выполнения задач, таких как классификация, регрессия, распознавание образов и другие. Основные элементы, из которых состоит нейронная сеть, представлены на рис. 4:

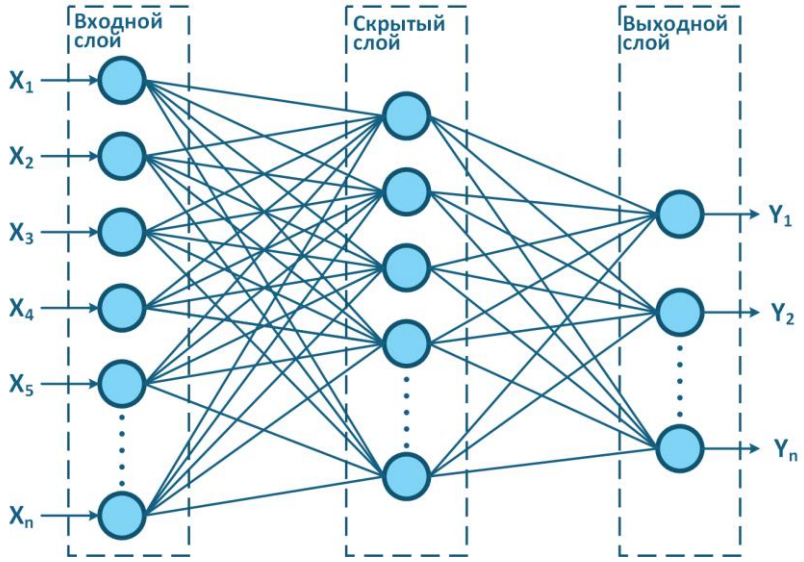


Рис. 4. Структура нейронной сети

Допустим, нам необходимо спрогнозировать итоговую оценку студента по курсу «Кроссплатформенное программирование» (числовое значение от 0 до 100). Мы располагаем следующими данными о студентах:

– результаты промежуточного контроля в виде тестирования ($Тест_{ПК}$);

– количество часов, проведенных за самостоятельным изучением материала ($t_c$);

– средняя оценка за лабораторные работы ($ЛР_{ср}$);

– активность на онлайн-платформе курса (АПл);

– средний балл за предыдущие курсы, связанные с программированием ($БПр_{ср}$).

Для того, чтобы спрогнозировать итоговую оценку, используя имеющиеся признаки, необходимо выполнить следующие шаги.

1) Собрать исторические данные, включающие все перечисленные признаки и фактический итоговый балл по «Кроссплатформенное программирование».

2) Провести предварительную обработку данных: обработать пропущенные значения, масштабировать числовые признаки (например, стандартизация или нормализация) для улучшения сходимости обучения нейронной сети, кодировать категориальные признаки, если они присутствуют (в нашем примере все признаки числовые).

3) Провести разделение данных на обучающую, валидационную (используется для настройки гиперпараметров сети и предотвращения переобучения) и тестовую выборки.

4) Определить архитектуру нейронной сети:

– количество входных нейронов должно соответствовать количеству признаков (в нашем случае 5);

– количество скрытых слоев и нейронов (гиперпараметр, который часто определяется экспериментально с использованием валидационной выборки (слишком мало слоев/нейронов может привести к недообучению, а слишком много – к переобучению);

– функции активации для нейронов скрытых слоев;

– количество выходных нейронов (в задаче регрессии с одной числовой целевой переменной (итоговый балл) обычно используется один выходной нейрон с линейной функцией активации).

5) Обучить нейронную сеть на обучающей выборке с использованием алгоритма обратного распространения ошибки и оптимизатора. В процессе обучения сеть корректирует веса связей между нейронами, чтобы минимизировать функцию потерь (например, среднеквадратическую ошибку - MSE) между

предсказанными и фактическими итоговыми баллами. Мониторинг производительности на валидационной выборке помогает вовремя остановить обучение для предотвращения переобучения.

6) Провести настройку гиперпараметров, используя валидационную выборку.

7) Оценить производительность лучшей модели на независимой тестовой выборке с использованием метрик регрессии (MAE, MSE, RMSE, R-squared).

8) Использовать обученную нейронную сеть для прогнозирования итогового балла по курсу «Кроссплатформенное программирование» для новых студентов на основе их значений входных признаков.

Нейронные сети являются мощным инструментом для прогнозирования успеваемости студентов, способным улавливать сложные закономерности в данных и достигать высокой точности. Однако их применение требует большого объема данных и значительных вычислительных ресурсов. Проблема интерпретируемости также является важным фактором, который следует учитывать при выборе этого алгоритма для образовательных задач.

В заключение хочется отметить, что оценка эффективности различных методов МО в прогнозировании успеваемости студентов показывает, что не существует универсально лучшего алгоритма. Производительность модели сильно зависит от конкретного набора данных, целей прогнозирования и выбранных метрик оценки.

Использование адекватных метрик, строгая валидация моделей и учет контекстуальных и этических соображений являются ключевыми элементами успешного внедрения систем прогнозирования успеваемости студентов. Дальнейшие исследования в этой области будут способствовать разработке более точных и надежных инструментов для поддержки учащихся и оптимизации образовательных траекторий, а также будет сосредоточены на интеграции образовательной аналитики с существующими информационными системами в учебных заведениях [3].

**Список использованных источников**

1. Вилкова К.А. Учебная аналитика в традиционном образовании: ее роль и результаты / К.А. Вилкова, У.С. Захарова // Университетское управление: практика и анализ. – 2020. – Т. 24. – № 3. – С. 59–76

2. Fernandes E. et al. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil // Journal of Business Research. 2019. No. 94. P. 335-343.

3. Царькова Е.Г. Учебная аналитика в дистанционном обучении: особенности применения и перспективы развития / Е.Г. Царькова // Прикладная психология и педагогика. – 2022. – Т. 7. – № 3. – С. 54–66.

4. Алпатов А.В. Применение машинного обучения для анализа образовательных результатов студентов вузов // Информационные и математические технологии в науке и управлении. 2023. № 4(32). С. 67-78.

5. Егорова Е.С., Попова Н.А. Data Mining в образовании: прогнозирование успеваемости учащихся // Моделирование, оптимизация и информационные технологии. 2023. № 11(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1325 (дата обращения: 30.04.2025).

---

*Chernykh V.V., Candidate of Technical Sciences, Federal State Budgetary Educational Institution of Higher Education "Lugansk State University named after Vladimir Dahl"*

*Balalaechnikov A.V., Senior Lecturer, Federal State Budgetary Educational Institution of Higher Education "Lugansk State University named after Vladimir Dahl"*

EVALUATION THE EFFECTIVENESS OF VARIOUS MACHINE LEARNING METHODS IN PREDICTING STUDENT ACADEMIC PERFORMANCE

*Abstract: The article examines the possibilities of using machine learning to predict student performance in the context of a modern educational environment with ever-growing volumes of data. Key aspects of evaluating the effectiveness of various machine learning algorithms are analyzed. Attention is focused on the importance of adequate metrics, strict validation of models, and consideration of ethical considerations when implementing forecasting technologies.*

*Keywords: machine learning, algorithm, performance forecasting, regression, model, categorical features, training set, test set*