Материалы IV Всероссийской научно-практической конференции с междун. участием «ИИ в образовании. Современные достижения и перспективы применения: в генерации знаний, управлении, обучении, оценке результатов обучения и формировании компетенций обучающихся», 2025, №5 (98).

Информационно-коммуникационные технологии в педагогическом образовании, 2025. № 5 (98). *infed.ru*

УДК 004.852

Е. А. Чудакова, Е. В. Решетникова

E. A. Chudakova, E. V. Reshetnikova

Чудакова Екатерина Алексеевна, студент, КГПИ ФГБОУ ВО «КемГУ», г. Новокузнецк, Россия.

Решетникова Елена Васильевна, к. т. н., зав. кафедрой математики, физики и математического моделирования, КГПИ ФГБОУ ВО «КемГУ», г. Новокузнецк, Россия.

Chudakova Ekaterina Alekseevna, student, Kuzbass Humanitarian Pedagogical Institute of Kemerovo State University, Novokuznetsk, Russia.

Reshetnikova Elena Vasilievna, Candidate of Technical Sciences, Associate Professor, Kuzbass Humanitarian Pedagogical Institute of Kemerovo State University, Novokuznetsk, Russia.

ПОСТРОЕНИЕ АНСАМБЛЕЙ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ВОЗНИКНОВЕНИЯ ДЕПРЕССИИ У СТУДЕНТОВ

LEARNING ENSEMBLES OF MACHINE LEARNING MODELS FOR PREDICTING THE OCCURRENCE OF DEPRESSION IN STUDENTS

Е. В. Решетникова, Е. А. Чудакова 2025-06-23

Аннотация. *Статья посвящена исследованию возможности применения машинного обучения для прогнозирования депрессии у студентов. Предложен подход, использующий ансамблевые модели. Исследование способствует развитию инструментов ранней диагностики и может быть использовано для улучшения психического благополучия студентов.*

Annotation. *This article is devoted to the study of the possibility of using machine learning to predict depression in students. An improved approach is proposed, including ensemble models and hyperparameter optimization. The research contributes to the development of early diagnostic tools and can be used to improve students' mental well-being.*

Ключевые слова: *модель машинного обучения, депрессия студентов, ансамбль моделей, логистическая регрессия, метод опорных векторов, деревья решений.*

Keywords: *machine learning model, student depression, model ensemble, logistic regression, support vector machine method, decision trees.*

Данные для построения моделей представляют собой информацию о студентах и их психологическом состоянии. Число объектов в наборе данных 27901, что позволяет достаточно качественно строить модели машинного обучения. Каждый объект набора данных характеризуется четырнадцатью признаками, среди которых можно выделить академические показатели (учебная нагрузка, средний балл за весь период обучения), психологическое состояние (академический стресс, суицидальные мысли, наследственные заболевания), социально-бытовые условия (финансовый стресс, питание, сон) и демографические характеристики (возраст, пол, место жительства). Такая структура позволяет анализировать взаимосвязи между различными аспектами жизни студентов и их психологическим состоянием.

Для преобразования категориальных признаков выбран метод One-Hot Encoding, поскольку большинство признаков имеют небольшое число уникальных значений, числовые данные стандартизированы.

Проведены исследования по применению алгоритмов логистической регрессии, метода k-ближайших соседей (KNN) и метода опорных векторов (SVC) для прогнозирования депрессии студентов. Тренировочная выборка, сформирована из 60 % от общего набора данных [1]. Все модели реализованы с применением библиотеки scikit-learn [2]. Для оценки качества моделей использовались следующие метрики: accuracy – общая доля правильных предсказаний модели, precision (точность) – доля истинных положительных случаев среди всех предсказанных положительных, recall (полнота) – доля найденных положительных случаев среди всех предсказанных положительных, F1-мера – гармоническое среднее между precision и recall, AUC – площадь под ROC-кривой, которая показывает способность модели отличать положительные классы от отрицательных.

Наибольшая эффективность прогнозирования достигнута при применении метода опорных векторов с линейным ядром и коэффициентом регуляризации C = 4,1. Анализ метрик качества показал стабильность работы этой модели: на тестовой выборке достигнута точность 0.8496, полнота 0.8777, F1-мера 0.8634 при значении AUC-ROC 0.9131. Минимальное расхождение между показателями на тренировочной (precision = 0.8589, recall = 0.8944, F1 = 0.8763, AUC = 0.9242) и тестовой выборках свидетельствует о высокой обобщающей способности алгоритма и отсутствии значимого переобучения.

Несмотря на высокие показатели метрик линейных моделей, их остаточная ошибка классификации свидетельствует о возможном наличии нелинейных взаимосвязей между предикторами (признаками) и целевой переменной. В связи с этим следующий этап исследования посвящен построению решающих деревьев [3] и ансамблей моделей [4], которые потенциально могут обеспечить выявление скрытых закономерностей в данных.

Для настройки гиперпараметров решающего дерева использована кросс-валидация. Выборка разбивалась на 5 частей, что способствует более качественной оценке точности. Варьировались максимальная глубина дерева, минимальное количество объектов для разделения узла и минимальное количество объектов в листе дерева. Также проводились эксперименты по использованию различных критериев информативности для разделения узла – индекса Джини (gini) и энтропийного критерия (entropy).

Обученная модель решающего дерева (рис. 1), несмотря на менее сбалансированные результаты по сравнению с линейной SVC-моделью, продемонстрировала высокую эффективность в выявлении случаев депрессии, правильно классифицировав 87 % студентов. Этот показатель сопоставим с результатами SVC-модели и особенно важен для целей исследования. Анализ precision показал, что модель решающего дерева допускает меньше ошибок при прогнозировании случаев депрессии (0.84) по сравнению с идентификацией здоровых студентов (0.81).

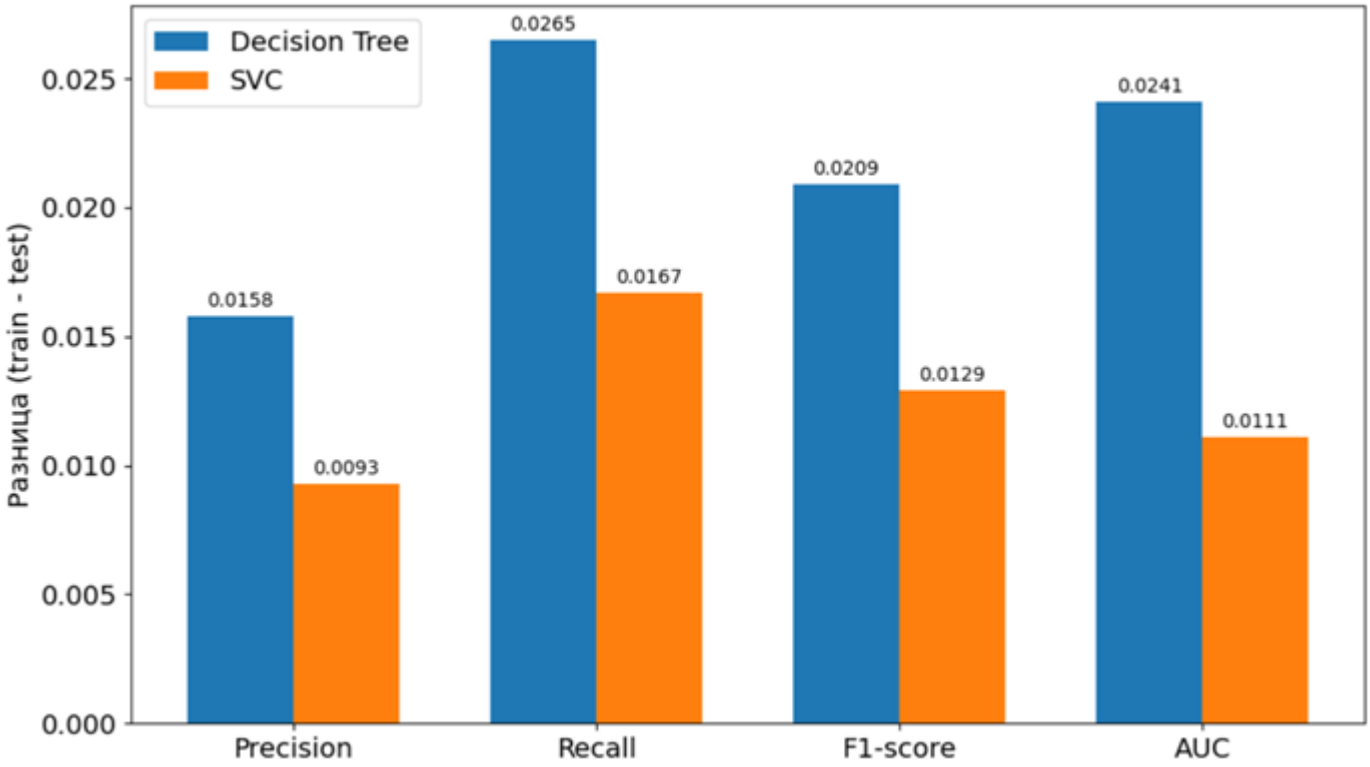

Рисунок 1. Сравнение дисбаланса точности прогнозов на тренировочной и тестовой выборках решающего дерева и SVC

Такой результат согласуется с общей стратегией медицинской диагностики, где ложноположительные результаты считаются менее критичными, чем ложноотрицательные. Показатель F1-score составил 0.85 для класса студентов, страдающих депрессией против 0.79 для здоровых студентов, что подтверждает общую эффективность нелинейной модели для поставленной задачи.

Выбранная конфигурация гиперпараметров (максимальная глубина дерева = 7, минимальное количество объектов для разделения узла = 2, минимальное количество объектов в листе дерева = 4, критерий информативности = gini) позволила достичь оптимального баланса между сложностью модели и ее обобщающей способностью. Важно отметить, что разница в accuracy между тренировочной и тестовой выборками составляет всего 0.03, что свидетельствует об отсутствии переобучения модели. Однако, эта модель все же не превзошла по эффективности метод опорных векторов с линейным ядром (рис. 2).

Как известно деревья показывают себя лучше при построении ансамблей. Для ансамблирования использовался метод VotingClassifier из библиотеки scikit-learn, который агрегирует предсказания моделей путем голосования. Ансамбль построен на основе трех моделей: SVC, логистическая регрессия и решающее дерево. Вычислительные эксперименты по подбору весов моделей, включенных в ансамбль, показали, что увеличение веса решающего дерева чаще приводит к улучшению результатов, тогда как влияние весов других моделей носит менее выраженный характер. Однако ни один из ансамблей, построенных методом голосования, не показал эффективности, большей, чем ранее построенная модель SVC.

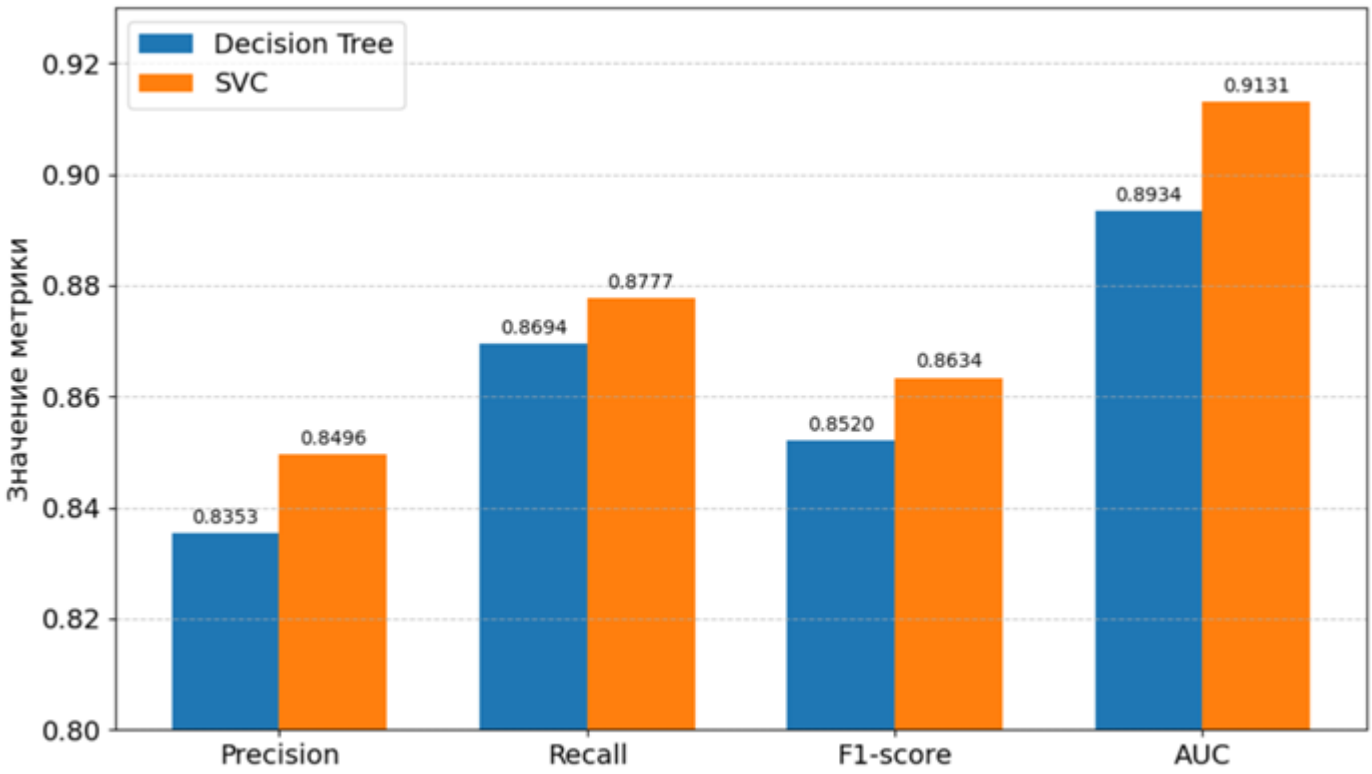

Рисунок 2. Сравнение значений различных метрик на тестовых выборках
для решающего дерева и SVC

В отличие от голосования, где используются фиксированные модели, другая ансамблевая модель – бэггинг предполагает построение ансамбля независимых моделей одного типа обученных на различных бутстреп-выборках исходных данных. Такой подход позволяет уменьшить дисперсию предсказаний, повысить устойчивость ансамбля к шуму в данных, улучшить обобщающую способность.

Проведено исследование по влиянию количества используемых в ансамбле моделей на качество прогноза. При построении ансамбля логистических регрессий максимальный F1-score (0.8634) достигнут при использовании 36 моделей. Дальнейшее увеличение числа моделей не приводит к значительному улучшению качества, а в некоторых случаях может даже снижать эффективность из-за избыточности (рис. 3).

Ансамбль на основе решающих деревьев показывает максимальное значение F1-score = 0.8612 при 26 моделях в ансамбле (рис. 4). Однако дальнейшее добавление деревьев не дает значимого улучшения. Применение меньшего количества ухудшает результаты из-за недостаточной стабильности ансамбля.

Вычислительные эксперименты по подбору оптимального количества моделей SVC в ансамбле показали, что максимальное значение F1-меры (0.8644) достигается при использовании 8 базовых моделей (рис. 5).

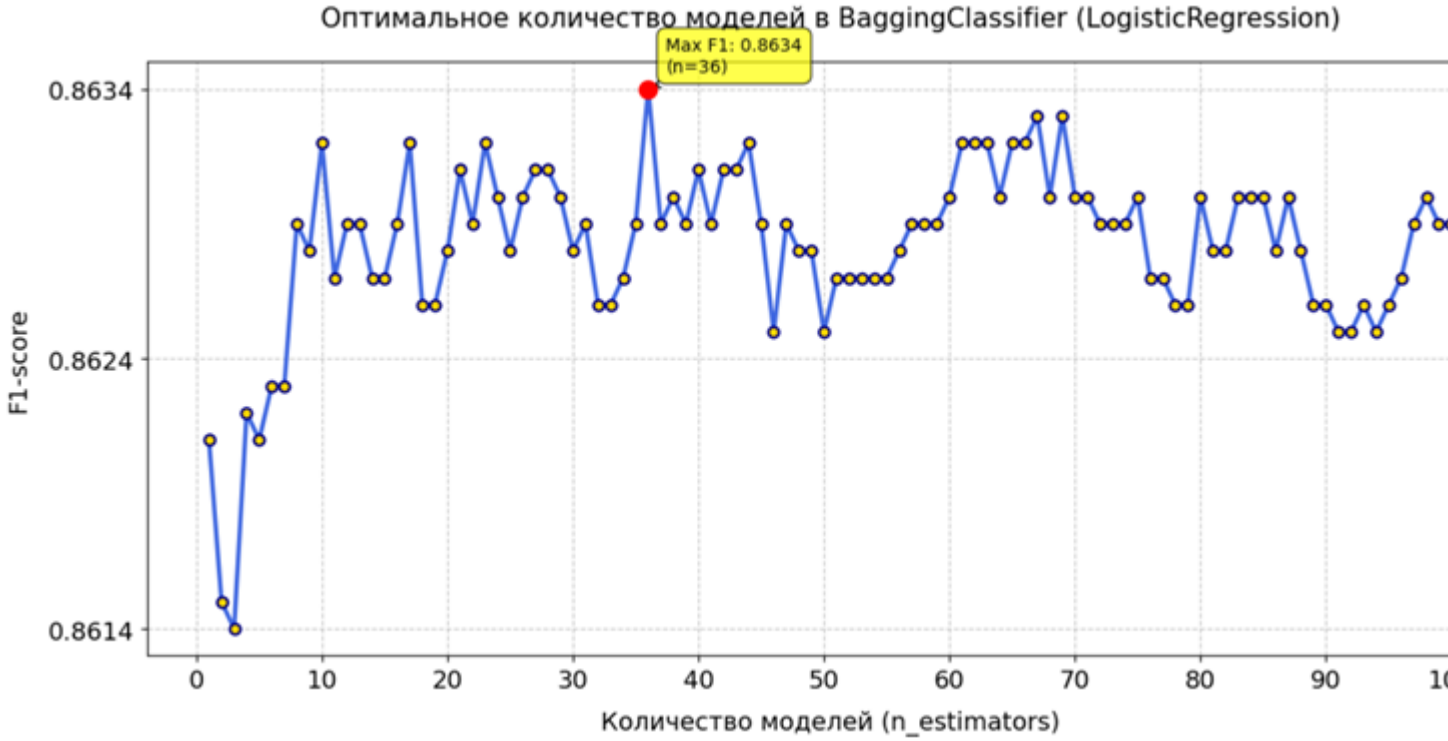

Рисунок 3. Зависимость F1-меры от количества моделей в ансамбле BaggingClassifier (Logistic Regression)

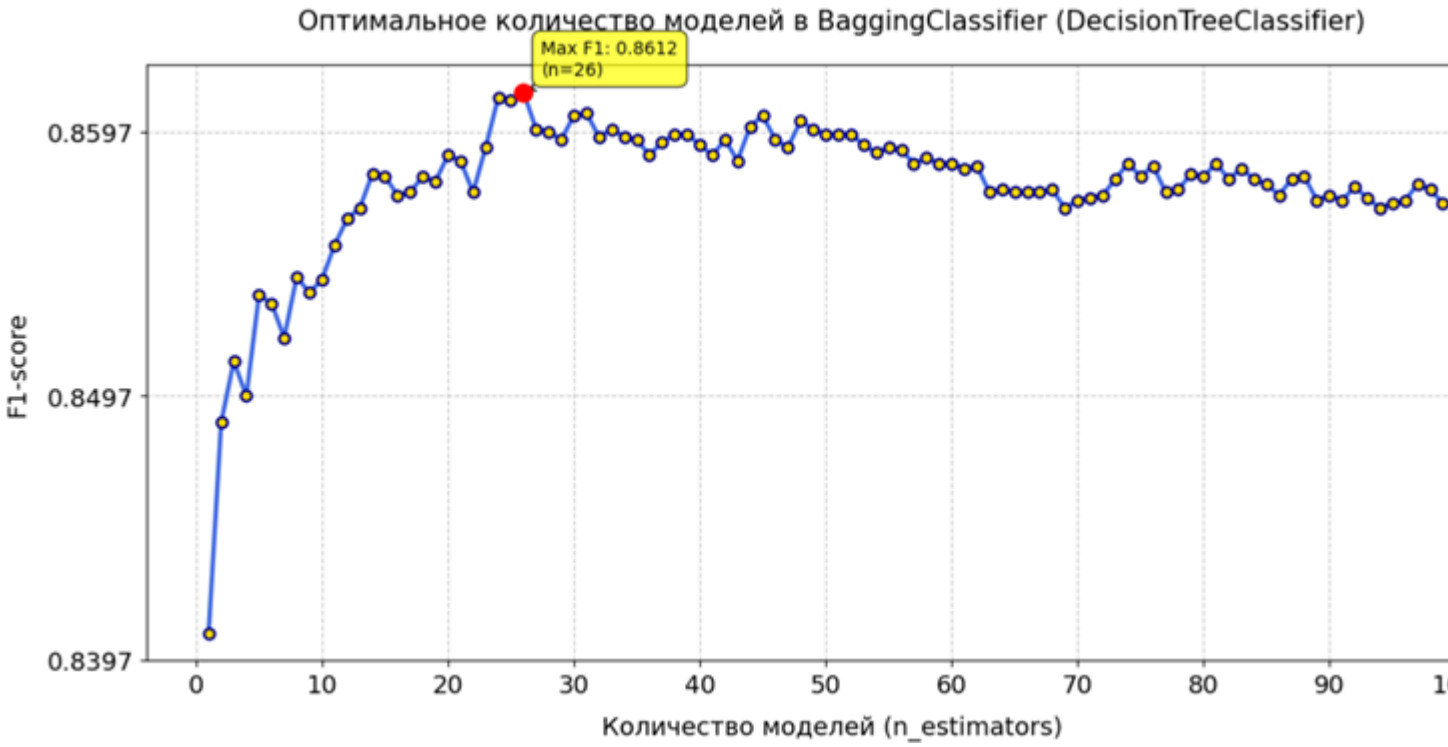

Рисунок 4. Зависимость F1-меры от количества моделей в ансамбле BaggingClassifier (Decision Tree Classifier)

Следует отметить, что полученные показатели эффективности каждой модели с использованием бэггинга улучшились по сравнению с одиночными моделями. Несмотря на длительное время обучения, связанное с вычислительной сложностью метода опорных векторов, ансамбль (бэггинг) именно этих моделей продемонстрировал наилучшие значения метрик по сравнению со всеми построенными моделями, включая одиночные SVC, логистическую регрессию, решающие деревья и различные их ансамбли.

Полученные результаты имеют практическое значение для разработки систем психологического мониторинга в учебных заведениях.

Рисунок 5. Зависимость F1-меры от количества моделей в ансамбле BaggingClassifier (SVC)

Список литературы

1. Платонов, А. В. Машинное обучение : учебное пособие для вузов / А. В. Платонов. – 2-е изд. – Москва : Издательство Юрайт, 2025. – 89 с. – (Высшее образование). – ISBN 978-5-534-20732-3. – URL: https://urait.ru/bcode/558662 (дата обращения: 03.03.2025). – Текст : электронный.
2. Scikit-learn: Machine Learning in Python : сайт. – 2007 - . – URL: https://scikit-learn.org/stable/index.html (дата обращения: 03.03.2025). – Текст : электронный.
3. Шалев-Шварц, Ш. Идеи машинного обучения : учебное пособие / Ш. Шалев-Шварц, Бен-ДавидШ. ; перевод с английского А. А. Слинкина. – Москва : ДМК Пресс, 2019. – 436 с. – ISBN 978-5-97060-673-5. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/131686 (дата обращения: 16.04.2025). – Режим доступа: для авториз. пользователей.
4. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. – Москва : ДМК Пресс, 2015. – 400 с. –

Материалы IV Всероссийской научно-практической конференции с междун. участием «ИИ в образовании. Современные достижения и перспективы применения: в генерации знаний, управлении, обучении, оценке результатов обучения и формировании компетенций обучающихся», 2025, №5 (98).

ISBN 978-5-97060-273-7. – URL: https://e.lanbook.com/book/69955 (дата обращения: 16.04.2025). – Текст : электронный.

Е. В. Решетникова, Е. А. Чудакова 2025-06-23