0.1. *Чеглов Е.Р.* Распознавание сгенерированного программного кода в студенческих работах на основе метода случайного леса и анализа стилистических признаков

В последние годы использование генеративных моделей искусственного интеллекта студентами становится всё более распространённым не только в учебном процессе, но и в соревновательных форматах — на хакатонах, олимпиадах и других конкурсах программирования [1]. Это создаёт серьёзные вызовы для академической среды, где оценка оригинальности решений является ключевым элементом контроля знаний. Существуют различные подходы к автоматическому распознаванию сгенерированного кода [2]. Однако их применение требует значительных вычислительных ресурсов и крупных размеченных корпусов данных. В данном исследовании рассматривается альтернативный подход — бинарная классификация программного кода с использованием метода случайного леса и анализа стилистических и статистических признаков.

Для исследования был собран датасет, включающий 500 решений, написанных студентами, и 450 решений, сгенерированных различными моделями ИИ для тех же задач. Перечень моделей формировался на основе анонимного опроса, в котором приняли участие более 60 студентов. Среди наиболее часто используемых инструментов были отмечены ChatGPT, Gemini, DeepSeek, Qwen, Microsoft Copilot, Grok, Claude, а также отечественные модели YandexGPT и GigaChat. Каждая модель генерировала пять решений: одно — имитируя типичное поведение студента, три — как различные варианты ответа по просьбе, и одно — по специальному промпту, побуждающему имитировать студенческий стиль с намеренными стилистическими неточностями. Все решения были проверены в тестирующей системе, после чего проведена подготовка данных для обучения модели. Датасет был сохранён и обрабатывался в формате JSON.

Для извлечения признаков был разработан скрипт на Python с использованием инструментов scikit-learn для обучения и валидации модели. Анализировались такие характеристики, как длина кода, частота ключевых слов и другие метрики. На первой итерации модель случайного леса обучалась на 30 признаках с параметрами по умолчанию (150 деревьев, фиксированное случайное состояние). Полученные результаты показали высокие показатели на обучающей выборке (F1 ≈ 0.997) и F1 ≈ 0.90 на тестовой, при средней F1 кросс-валидации 0.8187 ± 0.0394, что указывало на склонность к переобучению. На следующем этапе часть признаков была исключена, их количество сократилось до 18, а гиперпараметры модели были изменены: глубина деревьев ограничена до 10, минимальное количество образцов для разбиения увеличено до 5, добавлен баланс классов. В результате F1 на обучающей выборке снизился до ≈ 0.98, тестовая F1 составила 0.895, а средняя F1 кросс-валидации — 0.8178 ± 0.0531. Эти изменения свидетельствуют о снижении переобучаемости и достижении моделью естественного потолка качества, обусловленного информативностью признаков.

Таким образом, проведённое исследование показало, что анализ стилистических и статистических характеристик кода позволяет достичь высоких результатов в задаче распознавания сгенерированных решений в студенческой среде при относительно низких вычислительных затратах. В перспективе планируется расширение датасета за счёт включения новых тематических блоков и построение инфографики, отражающей вариативность качества распознавания в зависимости от тематики задач. Также планируется исследовать применение глубоких нейросетевых моделей и комбинированных методов для повышения точности классификации.

Научный руководитель — к.ф.-м.н. Пестунов А. И.

Список литературы

[1] Кириенко Д. П. Нейросети: плюсы и минусы // Тр. Конф. «Всероссийский съезд учителей информатики». Красноярск: Сибирский федеральный университет, 2025. С. 300–303.

[2] Oedingen M., Engelhardt R., Denz R., Hammer M., Konen W. ChatGPT Code Detection: Techniques for Uncovering the Source of Code // AI. 2024. N. 5(3). P. 1066–1094.