ПРАКТИЧЕСКАЯ РАБОТА №15
Исследование вопросов этики и безопасности искусственного интеллекта

Цель работы

Освоить методы и инструменты по теме «исследование вопросов этики и безопасности искусственного интеллекта» на уровне самостоятельного магистерского исследования; получить измеримый результат и использовать его для последовательного развития индивидуального семестрового проекта.

Краткие теоретические сведения

Responsible AI включает безопасность, справедливость, прозрачность, конфиденциальность и ответственность. Финальная оценка интеллектуальной системы должна учитывать не только метрики модели, но и качество данных, ограничения, возможный ущерб, способы неправильного использования и контроль после внедрения.

Связь с предыдущими практическими работами

В ПЗ №15 необходимо использовать результаты предыдущих этапов индивидуального проекта. Финальный аудит, сравнительный анализ и защита индивидуального проекта. Не заменяйте ранее подготовленные данные или постановку задачи без обоснования; изменения фиксируйте в отчёте.

Методические указания

Работа выполняется магистрантом самостоятельно. Все решения должны быть обоснованы: недостаточно получить программный результат, необходимо объяснить выбор данных, метода, метрик и интерпретировать полученные результаты.

Код должен быть воспроизводимым: фиксируйте random_state/seed там, где это применимо, подписывайте этапы Notebook и не удаляйте результаты ключевых экспериментов. Используйте только данные, которые разрешено обрабатывать в выбранной среде.

Практическая работа является частью сквозного индивидуального семестрового проекта. Результаты предыдущих работ повторно не создаются, а используются как вход для текущего этапа. Если метод текущей недели объективно не соответствует предметной задаче, магистрант выполняет небольшой связанный эксперимент и отдельно обосновывает, почему этот компонент не следует включать в итоговую архитектуру.

Рекомендуемый порядок выполнения: 1) сформулировать исследовательский вопрос; 2) подготовить входные данные; 3) реализовать базовый вариант; 4) провести контролируемый эксперимент; 5) сравнить результаты; 6) интерпретировать ошибки; 7) зафиксировать вклад текущей работы в семестровый проект.

Задания

Задание 1. Проведите аудит данных, модели, интерфейса и сценария использования проекта.

Задание 2. Определите не менее 5 рисков: ошибки, смещения, конфиденциальность, безопасность, неправильное использование и др.

Задание 3. Для каждого риска оцените вероятность/последствия и предложите меры контроля.

Задание 4. Сравните минимум две исследованные модели/варианта и обоснуйте итоговую архитектуру.

Задание 5. Сформулируйте ограничения, область допустимого применения и план дальнейшего развития прототипа.

Минимальный программный пример

```
risks = [
   {"risk":"ошибочная классификация", "probability":3, "impact":4},
   {"risk":"утечка данных", "probability":2, "impact":5},
   {"risk":"смещение данных", "probability":3, "impact":3}
]
for r in risks:
   r["score"] = r["probability"] * r["impact"]
print(sorted(risks, key=lambda x: x["score"], reverse=True))
```

Пример предназначен только для демонстрации принципа. Магистрант должен адаптировать код к собственной задаче, данным и структуре проекта; простое копирование примера не считается выполнением индивидуального задания.

Пример выполнения

Финальный аудит системы обнаружения фишинговых URL

В итоговой таблице фиксируются качество классификации, ложные срабатывания, устойчивость к новым типам URL, риски утечки данных и сценарии неправильного доверия к прогнозу. Система должна позиционироваться как инструмент поддержки аналитика, а не как безошибочный источник решения.

Элемент анализа	**Пример фиксации результата**
Исходные условия	Указать набор данных, разбиение, seed и используемые признаки.
Метод	Указать модель/режим и ключевые параметры.
Результат	Привести измеримые метрики или формализованные критерии.
Интерпретация	Объяснить, почему результат получен и какие ошибки обнаружены.
Решение	Обосновать, включается ли компонент в итоговый прототип.

Индивидуальное задание по семестровому проекту

Финальный аудит, сравнительный анализ и защита индивидуального проекта

1. Выполните текущий этап исключительно на теме собственного проекта.
2. Добавьте минимум одно самостоятельное решение, которого нет в приведённом примере.
3. Зафиксируйте минимум один альтернативный вариант и объясните, почему он принят или отклонён.
4. Сохраните артефакт текущего этапа: Notebook, таблицу, модель, промпт, схему workflow или матрицу рисков — в зависимости от темы работы.
5. В конце раздела напишите 5–8 предложений о том, как результат текущей ПЗ изменил или уточнил архитектуру индивидуальной системы.

Требования к отчёту

Отчёт должен содержать: тему и цель; краткое теоретическое обоснование; описание исходных данных; ход выполнения заданий; ключевые фрагменты кода и результаты их выполнения; таблицы/графики при необходимости; анализ ошибок; индивидуальный раздел

семестрового проекта; вывод. Рекомендуемый объём — 5–8 страниц без искусственного увеличения объёма скриншотами.

К защите магистрант должен уметь объяснить каждую существенную строку собственного кода, происхождение данных, причины выбора метода и смысл полученных метрик. Невоспроизводимый результат без объяснения не считается завершённым экспериментом.

Результат практической работы

После выполнения ПЗ №15 магистрант представляет отчёт и обновлённый артефакт индивидуального проекта. Результат текущей работы должен быть пригоден для использования на следующем этапе курса.

Контрольные вопросы

1. Какова основная исследовательская задача практической работы №15?
2. Какие исходные данные и предположения используются в работе?
3. Почему выбранный метод соответствует поставленной задаче?
4. Какие альтернативные методы можно было бы использовать?
5. Какие метрики или критерии позволяют оценить результат?
6. Какие источники ошибок или смещения возможны?
7. Как обеспечить воспроизводимость эксперимента?
8. Какие риски информационной безопасности необходимо учитывать?
9. Как результаты данной работы используются в индивидуальном семестровом проекте?
10. Какой следующий эксперимент логично выполнить для улучшения результата?

Список источников

1. Бессмертный И. А., Авдюшина А. Е., Кугаевских А. В., Королева Ю. А., Рущенко Н. Г. Системы искусственного интеллекта: методическое пособие. – Санкт-Петербург: Университет ИТМО, 2024. – 281 с.

2. Соснило А. И. Атлас искусственного интеллекта для бизнеса и власти: учебное пособие. – Санкт-Петербург: Университет ИТМО, 2022. – 98 с.

3. Сидоркина И. Г. Системы искусственного интеллекта: учебное пособие. – Москва: КНОРУС, 2018. – 248 с.

4. Баланов А. Н. Искусственный интеллект. Понимание, применение и перспективы: учебник для вузов. – Санкт-Петербург: Лань, 2024. – 312 с.

5. Калинин М. О., Крундышев В. М. Основы искусственного интеллекта. Безопасность искусственного интеллекта: учебное пособие. – Санкт-Петербург, 2024. – 220 с.

6. Андреянов Н. В., Евдокимова Т. С., Павлов А. Д. [и др.]. Методы искусственного интеллекта: учебное пособие. – Казань: КНИТУ-КАИ, 2024. – 392 с.