**Домашнее задание (выполняется в команде)**

| Дисциплина | Применение ML в кибербезопасности |
| --- | --- |
| Тема | Итоговое домашнее задание |
| Форма проверки | **Задание, проверяемое преподавателем** |
| Имя преподавателя | Юрий Иванов |
| Время выполнения | с 24.10.2024 до 23:59 29.10.2024 |
| Цель задания | получить навык выполнения полного цикла работ при решении задачи машинного обучения |
| Инструменты для выполнения ДЗ | Для выполнения задания используйте google.com, medium.com, препринты (arxiv.org), github, python, jupyter, anaconda и ml библиотеки  Результаты выполнения задания должны быть внесены в документ Яндекс презентация |
| Правила приема работы | Прикрепите ссылку на документ [презентацию](https://disk.yandex.ru/i/uOVimqbYdaoPdg) в ответ на задание  Важно: убедитесь в том, что по ссылке есть доступ |
| Критерии оценки | **Максимальное количество баллов за итоговое задание - 10 баллов**    **Задание считается выполненным, если:**  - выполнены этап 1 и этап 2 (обязательно) либо выполнены все этапы задания  - презентация оформлена в соответствии с требованиями  - прикреплена ссылка на файл с выполненным заданием  - доступ к файлу открыт  - полученная сумма баллов за задание не менее 5  **Задание не выполнено, если:**  - не выполнены этап 1 и этап 2 задания  - не оформлена презентация  - файл с заданием не прикреплен или отсутствует доступ по ссылке  - полученная сумма баллов за задание менее 5 |
| Дедлайн | см. В ЛМС |

**Вводная информация**

Задание выполняется в предварительно сформированных командах

Перед выполнением задания необходимо определить роли каждого из участников команды

Возможные варианты распределения ролей при решении задачи машинного обучения:

* ML инженер (подготовка, упаковка модели, написание сопутствующих сервисов)
* Data Engineer (Сбор и обработка данных)
* Data Scientist (проведение экспериментов, обучение модели)
* Аналитик (анализ задачи, планирование подходов и экспериментов)
* Менеджер (обоснование проекта, подхода, защита проекта)
* и др.

**Цель задания**

Реализовать полный цикл решения задачи машинного обучения,

включающий в себя следующие этапы:

* сбор и обработку данных
* обучение модели
* упаковку модели
* публикацию модели

**Задание**

Разработать модель машинного обучения (с использованием методов NLP), которая сможет с высокой долей вероятности распознавать спам

\*Дополнительное задание (необходимо выполнить для получения максимальной оценки)

Разработать дополнительную модель, которая сможет определять категорию текста письма в формате: реклама, предложение работы, знакомства и т.д.

**Этапы выполнения задания**

**Этап 0**

Повторите материалы лекций, вебинаров и практических заданий по дисциплине

Проведите анализ существующих решений задач машинного обучения в сфере кибербезопасности

**Этап 1**

**Сбор и обработка данных**

Используя любые доступные ресурсы, найдите и выберите датасет (один или несколько) для обучения антиспам алгоритма

\*Использование нескольких датасетов является существенным фактором при оценивании задания

Результат:

Набор данных готов для обучения

*Подсказка:*

В качестве датасетов могут использоваться:

[Enron Spam Dataset](http://www.kaggle.com/datasets/wanderfj/enron-spam/code)

[SpamAssassin](https://www.kaggle.com/datasets/beatoa/spamassassin-public-corpus)

[LingSpam](https://www.kaggle.com/datasets/mandygu/lingspam-dataset)

[SpamBase](https://archive.ics.uci.edu/ml/datasets/Spambase)

**Этап 2**

**Обучение модели ML**

1. Выберите архитектуру модели ML и обучите ее

Важно: при обучении модели необходимо минимизировать ложно-положительные срабатывания

*Подсказки:*

* при использовании бустинговых моделей данные предварительно необходимо векторизовать
* возможно использование Трансформеров в качестве экстрактора эмбедингов
* используйте метрики F1, Recall и Precision для оценки модели
* не забудьте про кросс-валидацию

1. Протестируйте полученную классификацию с использованием (ссылка в чате)

*Подсказка:*

Возможно, тестирующий набор необходимо предварительно обработать

Результат:

Модель ML обучена. Подготовлен код

**\*Этап 3**

**Упаковка модели ML**

Упакуйте модель в микросервис в виде Docker + FastAPI/Flask

Результат:

Модель упакована в микросервис

**\*Этап 4**

**Публикация модели**

Разместите модель на публичном хостинге/сервисе

Результат:

Модель размещена и доступна для демо работы

**Этап 5**

**Важно!** *Оформляется одна презентация на всю команду*

Откройте и сделайте копию [презентации](https://disk.yandex.ru/i/uOVimqbYdaoPdg)

Внесите всю информацию по выполнению задания в презентацию

Презентация необходима для экзамена, где будет проведена защита вашей работы

Информация о регламенте защиты будет представлена дополнительно

**Этап 6**

**Важно!** *Каждый участник команды отправляет ссылку на презентацию своей команды в ответ на задание*

Проверьте доступ к документу Google презентация

Прикрепите ссылку на Google презентацию в ответ на задание

Название файла должно содержать фамилию и имя студента, номер группы, в которой он выполнял итоговое задание и название ДЗ (Итоговое задание)

**Критерии оценки задания экспертом**

Оценка учитывает степень проработки задания

Выставляется общая командная оценка, которая идет в зачет каждому члену команды

|  |  |
| --- | --- |
| Критерии оценивания | Баллы за критерий (MAX) |
| Готова тетрадка/ код | 3 |
| Точность выше 86% | 2 |
| Проведен анализ решений, проведено тестирование нескольких моделей | 2 |
| Модель упакована в микросервис или развернута в облаке | 2 |
| Даны ответы на дополнительные вопросы  (см. Google презентацию) | 1 |
| **Всего баллов за задание** | **10** |