Домашнее задание: Практическая работа 3

Дисциплина	Применение ML в кибербезопасности
Тема	Машинное обучение в контексте кибербезопасности
Форма проверки	**Домашнее задание с проверкой преподавателем** *Совет: выполняйте домашнее задание сразу после изучения темы*
Имя преподавателя	Юрий Иванов
Время выполнения	3 часа
Цель задания	Научиться методологии анализа и решения задачи ML
Инструменты для выполнения ДЗ	Для выполнения задания используйте google, medium.com, препринты (arxiv.org), github Результаты выполнения задания должны быть внесены в документ Google Doc
Правила приема работы	1. Для выполнения задания создайте документ Google Doc и внесите все полученные данные и ваш план решения задачи ML. 2. Прикрепите ссылку на Google Doc. Важно: убедитесь в том, что по ссылке есть доступ. Название файла должно содержать фамилию и имя студента и номер ДЗ (ДЗ 3)
Критерии оценивания	**Задание считается выполненным, если:** - выполнены все этапы задания - прикреплена ссылка на файл Google Doc с выполненным заданием - доступы к материалам открыты - получено не менее 4 баллов за задание **Задание не выполнено, если:** - не выполнены все этапы задания - файл с заданием не прикреплен - отсутствует доступ по ссылке - получено менее 4 баллов за задание
Дедлайн	

Важно
Практическая работа в формате обзорного аналитического исследования.

Обзорное аналитическое исследование проводится перед практическим решением любой задачи ML.

Описание задания
Вам предстоит представить себя в роли инженера по машинному обучению, который получил задачу в формате проблемного кейса.

При такой постановке задачи практически отсутствуют исходные данные.

Вы должны понять предметную область задачи путем анализа существующих решений:

- сформулировать основную гипотезу
- определиться с набором данных для обучения
- выбрать используемые подходы и модели.

Методика погружения в предметную область

1. При анализе существующих подходов к решению аналогичных задач ML вы должны сравнить методы обнаружения вирусов:
    - статический анализ — в вашу систему будет поступать исполняемый файл
    - динамический анализ — вы получаете анализ поведения файла в песочнице (виртуальной среде для выполнения вредоносных программ)
2. Определить преимущества и недостатки этих методов.
3. Выбрать, какой из методов вы предполагаете использовать (если бы решали задачу на самом деле).
4. Понять, что такое вредоносный файл и чем он (по каким признакам) отличается от безопасного.

Полученная задача (проблемный кейс)
Разработать сервис (программу) для обнаружения вредоносных программ windows (PE, Portable Executable)

По итогам анализа вы должны сформулировать план решения этой задачи с точки зрения ML. Разрабатывать само решение задачи не нужно.

Задание
Провести обзорное аналитическое исследование методов решения задачи по обнаружению вредоносных программ windows (PE, Portable Executable) с использованием методов машинного обучения

Этапы выполнения задания:

Этап 1
Повтор материалов предыдущих лекций и вебинаров

Этап 2

Используя поиск google, medium.com, препринты (arxiv.org), найдите примеры 3-5 интересных подходов решения задачи обнаружения вредоносных программ и дайте их краткое описание (3-5 предложений):

- используемая модель ML
- используемый набор признаков
- отличительные особенности подхода

Этап 3

Проведите анализ существующих решений для обнаружения вредоносных приложений на github.com
Приведите примеры 3-5 лучших, по вашему мнению, решений для обнаружения вредоносных приложений с кратким описанием (3-5 предложений):

- используемая модель ML
- используемый набор признаков
- отличительные особенности подхода

Этап 4

Проведите поиск существующих наборов данных для обучения, используя следующие ресурсы: kaggle.com, google.com, medium.com

Сохраните ссылки на 3-5 наборов данных с кратким описанием (объем, авторы, признаки) для последующего внесения в таблицу в Google Doc

Этап 5

Проанализируйте и обобщите на основе исследований, проведённых на этапах 1-4, различные подходы решения задачи обнаружения вредоносных файлов

Сформулируйте 3-5 основных подхода по плану (1-2 абзаца на каждый):

- какие используются признаки и как они извлекаются
- какая модель ML используется
- особенности подхода

Этап 6
Создайте таблицу в Google Doc в соответствии шаблоном, приведенным ниже

Внесите результаты предыдущих этапов в таблицу:

	Действие	Результат исследования
1	Анализ по сайтам google.com medium.com препринтам (arxiv.org)	*Примеры 3-5 интересных подходов задачи обнаружения вредоносных программ с кратким описанием:* - *используемая модель ML*

	научным статьям	- *используемый набор признаков* - *отличительные особенности подхода* *(3-5 предложений)*
2	Анализ существующих решений на github	*Примеры 3-5 лучших, по вашему мнению, решений для обнаружения вредоносных приложений с кратким описанием* - *используемая модель ML* - *используемый набор признаков* - *отличительные особенности подхода* *(3-5 предложения)*
3	Поиск существующих наборов данных для обучения на kaggle.com, google.com, medium.com	*Ссылки на 3-5 наборов данных с кратким описанием (объем, авторы, признаки)*
4	Анализ различных подходов к решению задачи обнаружения вредоносных файлов	*Описание 3-5 основных подхода по плану:* - *какие используются признаки и как они извлекаются* - *какая модель ML используется* - *особенности подхода* *(1-2 абзаца на каждый)*

Этап 7

Опишите в документе Google Doc (после таблицы) предлагаемое вами решение задачи.

На основании проведенного анализа опишите предлагаемый вами способ решения поставленной задачи ML по следующему плану:

1. Какой подход (метод обнаружения вирусов) будете использовать и почему
2. Каким образом будете собирать данные или какие датасеты будете использовать
3. Будет ли выполняться валидация/очистка датасета и каким образом
4. Опишите набор признаков, используемых для обучения
5. Какие модели попробуете обучить в качестве базового решения. Обоснуйте выбор модели
6. Предполагается ли использование ансамблирования
7. Какие целевые метрики будут использоваться для оценки качества модели

Этап 8

Проверьте доступ к документу Google Doc
Прикрепите ссылку на Google Doc в ответ на задание.
Название файла должно содержать фамилию и имя студента и номер ДЗ (ДЗ 3)

Критерии оценки задания экспертом

Критерии оценивания	Баллы за критерий (MAX)
Найдены примеры 3-5 интересных подходов решения задачи обнаружения вредоносных программ и дано их краткое описание (3-5 предложений): - используемая модель ML - используемый набор признаков - отличительные особенности подхода	2
Приведены примеры 3-5 лучших решений для обнаружения вредоносных приложений с кратким описанием (3-5 предложений): - используемая модель ML - используемый набор признаков - отличительные особенности подхода	2
Приведены ссылки на 3-5 наборов данных с кратким описанием (объем, авторы, признаки)	2
Сформулированы 3-5 основных подхода по плану (1-2 абзаца на каждый): - какие используются признаки и как они извлекаются - какая модель ML используется - особенности подхода	2
На основании проведенного анализа описан предлагаемый способ решения поставленной задачи ML по заданному плану	2
Всего баллов за задание	**10**