Сначала расскажу чуть теории о данной работе, чтобы было больше понимания, что и зачем мы делаем.

Данные, используемые для обнаружения вторжений, представляют собой журналы сетевой активности, включающие информацию о соединениях, используемых протоколах, переданных байтах, флагах пакетов и многом другом. Эти данные собираются из сетевых трафиков и системных журналов, а затем используются для анализа потенциальных угроз.

При изучении структуры данных важно понять, какие признаки представлены в датасете. Обычно это числовые и категориальные признаки, отражающие параметры сетевого соединения. Например, продолжительность сессии, количество переданных и полученных байтов, тип протокола, используемый сервис и статус соединения. Чаще всего в таких наборах данных есть метка, указывающая, является ли трафик нормальным или вредоносным.

Визуализация данных помогает лучше понять их распределение и взаимосвязи между признаками. Для этого строят гистограммы, диаграммы рассеяния и тепловые карты корреляции. Гистограммы позволяют увидеть распределение значений, диаграммы рассеяния помогают выявить аномалии, а тепловая карта показывает, какие признаки коррелируют друг с другом.

Перед применением алгоритмов машинного обучения необходимо провести предобработку данных. Этот этап включает в себя обработку пропущенных значений, кодирование категориальных признаков и нормализацию числовых данных. Если в данных присутствуют пропуски, их можно удалить или заменить на средние значения. Категориальные признаки, такие как тип протокола или флаг пакета, кодируются с помощью методов **Label Encoding** или **One-Hot Encoding**. Числовые признаки нормализуются, чтобы привести их к одному масштабу и улучшить работу моделей.

Label Encoding и **One-Hot Encoding** — это два основных метода кодирования категориальных признаков, используемых в машинном обучении. Они необходимы, поскольку многие модели не умеют работать с текстовыми данными и требуют числового представления.

Label Encoding присваивает каждой категории уникальное числовое значение. Например, если у нас есть признак "протокол" с категориями TCP, UDP и ICMP, то они могут быть закодированы как 0, 1 и 2 соответственно. Этот метод прост в реализации, но может создать проблему, если числовые значения будут восприниматься моделью как упорядоченные, что не всегда соответствует смыслу данных.

One-Hot Encoding преобразует каждую категорию в отдельный бинарный признак. Для примера с протоколами создаётся три новых столбца: TCP (1, если это TCP, иначе 0), UDP (1, если это UDP, иначе 0) и ICMP (1, если это ICMP, иначе 0). Такой подход исключает проблему ложного порядка, но увеличивает размерность данных, особенно если категорий много.

Выбор метода зависит от контекста: Label Encoding используется, если модель может обрабатывать порядковые зависимости, а One-Hot Encoding предпочтителен, когда категории равнозначны и порядок их появления не имеет значения.

Теперь о наборах данных, которые мы будем использовать в работе. **NSL-KDD**, **CICIDS2017** и **UNSW-NB15** — это популярные наборы данных, используемые для исследования и разработки систем обнаружения вторжений. Каждый из них содержит информацию о сетевых соединениях и помогает анализировать поведение нормального и вредоносного трафика.

NSL-KDD является улучшенной версией KDD'99, созданной для устранения его недостатков, таких как дублирование записей и дисбаланс классов. Данные включают сетевые соединения с различными атрибутами, такими как продолжительность сессии, количество переданных байтов и использованный протокол. Метки классов указывают, является ли трафик нормальным или относится к одному из видов атак, включая DoS, R2L, U2R и Probe.

CICIDS2017 он включает реальные сетевые атаки, такие как ботнеты, атаки с использованием эксплойтов и вредоносного ПО, а также трафик обычных пользователей. В этом датасете представлены детализированные сетевые характеристики и метки классов, что делает его полезным для обучения моделей обнаружения вторжений.

UNSW-NB15 в отличие от NSL-KDD, он содержит больше признаков, таких как уровень энтропии пакетов, флаги TCP и характеристики потока данных. Этот набор данных отличается сложностью и реалистичностью, что делает его актуальным для современных исследований в области сетевой безопасности.

Данные для обнаружения вторжений представляют собой записи сетевого трафика, которые содержат информацию о соединениях, протоколах, объемах переданных данных и других характеристиках. Они используются для выявления аномальной активности в сети, которая может свидетельствовать о кибератаках. Популярные наборы данных, такие как NSL-KDD, CICIDS2017 и UNSW-NB15, содержат примеры как нормального, так и вредоносного трафика, что позволяет обучать и тестировать системы обнаружения вторжений.

Первым шагом в анализе данных является их загрузка и изучение структуры. Важно определить, какие признаки содержатся в наборе, какие из них числовые, а какие категориальные, а также проверить наличие пропущенных значений. Анализ данных помогает выявить закономерности и возможные проблемы, такие как несбалансированность классов или избыточность информации.

Визуализация данных играет важную роль в анализе сетевого трафика и обнаружении вторжений, так как помогает выявить скрытые закономерности, аномалии и взаимосвязи между признаками. Для этого часто используются гистограммы, диаграммы рассеяния и тепловые карты корреляций, каждая из которых выполняет свою функцию в исследовании данных.

Гистограммы позволяют оценить распределение значений числовых признаков. Они показывают, насколько данные сосредоточены в определенных диапазонах, и помогают выявить выбросы или дисбаланс в выборке. Например, анализ распределения количества переданных байтов может показать, что большинство соединений используют небольшие объемы данных, но есть редкие случаи аномально больших значений, что может указывать на атаки типа DoS.

Диаграммы рассеяния помогают визуализировать зависимость между двумя числовыми признаками. Каждая точка на таком графике представляет одно сетевое соединение, а

расположение точек отражает характер взаимосвязи. Если точки образуют четкий тренд, это может свидетельствовать о линейной зависимости между признаками. Если же они распределены хаотично, связи может не быть. Диаграммы рассеяния полезны для выявления кластеров в данных или аномалий, таких как отдельные точки, сильно отличающиеся от основной массы.

Тепловая карта корреляций используется для анализа взаимосвязей между всеми числовыми признаками в датасете. Она представляет собой матрицу, в которой значения корреляции между признаками закодированы цветом. Если два признака имеют высокую положительную корреляцию, значит, они изменяются вместе, а если отрицательную – увеличение одного признака сопровождается уменьшением другого. Это помогает отобрать наиболее информативные признаки и избежать дублирования данных при построении модели.

Предобработка данных включает в себя несколько важных шагов. Если в данных есть пропущенные значения, их необходимо либо удалить, либо заменить на средние или наиболее частые значения. Категориальные признаки, такие как названия протоколов или флагов пакетов, необходимо преобразовать в числовой формат с помощью методов Label Encoding или One-Hot Encoding. Числовые признаки часто нормализуются, чтобы привести их к единому масштабу, что помогает моделям машинного обучения работать более эффективно.

Практическое задание: Знакомство с данными для обнаружения вторжений

Цель:
Изучить структуру и особенности данных, используемых для обнаружения вторжений, выполнить их визуализацию и предобработку.

Системы обнаружения вторжений (IDS) анализируют сетевой трафик, выявляя аномальные и подозрительные активности. Для их обучения используются наборы данных, содержащие примеры нормального и вредоносного трафика. Одним из таких датасетов является NSL-KDD, представляющий собой улучшенную версию KDD'99, с устраненными дублирующимися записями и улучшенным балансом классов.

Анализ данных включает изучение их структуры, распределения признаков и выявление закономерностей. Визуализация помогает лучше понять данные и обнаружить возможные аномалии. Предобработка включает обработку пропущенных значений, кодирование категориальных признаков и нормализацию, что повышает точность работы моделей машинного обучения.

1. **Загрузка и анализ данных**

1.1 Загрузите набор данных NSL-KDD из репозитория GitHub
https://github.com/HoaNP/NSL-KDD-DataSet
1. 2 Импортируйте его в среду Python с помощью библиотеки pandas.
1.3 Исследуйте структуру данных:
1.4 Выведите первые 5 строк датасета.
1.5 Определите количество строк и столбцов.
1.6 Определите типы признаков (числовые, категориальные).
1.7 Проверьте наличие пропущенных значений.

```
import pandas as pd

# Загрузка данных
url = "https://raw.githubusercontent.com/HoaNP/NSL-KDD-DataSet/main/KDDTrain+.txt"
df = pd.read_csv(url, header=None)

# Просмотр структуры данных
print(df.head())
print(df.info())
print(df.isnull().sum())
```

2. Визуализация данных

1. **Постройте гистограммы для анализа распределения числовых признаков.**
2. **Создайте диаграммы рассеяния для выявления взаимосвязей между признаками.**
3. **Постройте тепловую карту корреляций для выявления зависимостей между числовыми переменными.**

Код:

```
import matplotlib.pyplot as plt
import seaborn as sns

# Гистограммы
df.hist(figsize=(12, 8), bins=30)
plt.show()

# Диаграмма рассеяния
sns.scatterplot(x=df[0], y=df[4], hue=df[41])
plt.show()

# Тепловая карта корреляции
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(), annot=True, cmap="coolwarm")
plt.show()
```

3. Предобработка данных

1. **Обработка пропущенных значений** (если имеются, замените средним или наиболее частым значением).
2. **Кодирование категориальных признаков:**

   Label Encoding – для упорядоченных категорий.

   One-Hot Encoding – для неупорядоченных категорий (например, протокол).

3. **Нормализация числовых данных:**

Min-Max Scaling (приведение значений к диапазону [0,1]).

Standardization (преобразование к среднему 0 и стандартному отклонению 1).

```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder, MinMaxScaler

# Кодирование категориальных признаков
label_encoder = LabelEncoder()
df[1] = label_encoder.fit_transform(df[1])  # Кодирование протокола

# One-Hot Encoding для сервиса и флагов
df = pd.get_dummies(df, columns=[2, 3])

# Нормализация числовых признаков
scaler = MinMaxScaler()
df[df.select_dtypes(include=['number']).columns] =
scaler.fit_transform(df.select_dtypes(include=['number']))
```

4. Попробуйте провести анализ на тестовом наборе данных NSL-KDD.

1. Выполните балансировку классов с помощью техники **oversampling** или **undersampling**.

Формат сдачи:
Jupyter Notebook (.ipynb) с кодом и комментариями.