Objective

As a Data Scientist at Sleek, you will be entrusted with exploring and unraveling insights from diverse cyber-related data.

This assignment aims to evaluate your proficiency in learning the domain, recognizing potential opportunities, and deploying advanced machine learning (ML) and deep learning (DL) techniques to extract insights that would otherwise remain obscure.

Here, your task is to analyze the CICIDS2017 dataset. Your mission involves importing the dataset, conducting comprehensive exploratory data analysis (EDA), and deriving three compelling insights. At least two of these insights must harness the power of ML or DL techniques to extract valuable information. The third insight must stem from some recommendation system of your liking.

Please ensure your insights are not only convincing but also easily digestible for others to comprehend.

Tasks

Importing and Familiarization

- ☐ Obtain the CICIDS2017 dataset from its official repository or other reputable sources. Ensure you access the latest version available.
- ☐ Load the dataset into your preferred data analysis environment. Preferably, use Python for this and all sections.
- ☐ Familiarize yourself with the structure of the dataset and make sure you understand the features. No prior knowledge is needed.

EDA

- ☐ Explore the dataset thoroughly to understand its characteristics and nuances.
- ☐ Utilize visualization techniques to gain insights into feature distributions and potential patterns.

Deriving Insights

- [ ] Derive two interesting insights from the dataset by applying classical ML or DL techniques.
- [ ] Come up with a third insight of your liking. You can apply a statistical model, a mathematical one, or another.
- [ ] Clearly articulate each insight along with the rationale behind it.

Deliverables

- [ ] A detailed documentation of your EDA process, including data preprocessing steps, visualizations, and key findings.
- [ ] Include any challenges encountered during the analysis and how you addressed them.
- [ ] Three insightful findings derived from the dataset.
- [ ] Ensure that the code is well-documented and organized for easy understanding and reproducibility.

Submission Guidelines

Please ensure your completed assignment is submitted as a GitHub repository, including clear documentation and instructions. You should dedicate up to one full day of work to the exercise, spread out over a maximum of three days.

Note

A partial solution, showing off your way of thinking and capabilities, is better than no solution at all. Feel free to reach out for any clarification or assistance during the assignment.

Good luck!