# BIRLA INSTITUTE OF INFORMATION TECHNOLOGY & SCIENCE, PILANI FIRST SEMESTER 2024-2025

DSECLZG628T **DISSERTATION**

Dissertation Title : To establish baseline for threat detection Name of Supervisor : Prathibha Panduranga Rao

Name of Student : Goyal Taruchit Tarun Chitra ID No. of Student : 2022DC04496

Courses Relevant for the Project & Corresponding Semester:

1. Introduction to Data Science (Semester 1)
2. Machine Learning (Semester 2)
3. Artificial and Computational Intelligence (Semester 2)
4. Data Visualization and Interpretation (Semester 2)

# Abstract

Cybersecurity is key for any organization; attackers keep evolving and learning new ways to evade cyber-attack detection deployed by organizations. By analyzing the events, security operations center (SOC) can detect threats and make existing detections more effective. While analyzing network dataset for detecting cyber-attacks, the volume of records and dimensionality of records generated are very high. As the result, building automated analysis and detection of potential threats can lead to noisy outcomes. In most of the historical research, the power of advanced graph or deep learning models are leveraged for handling high-dimensional dataset. But it comes at the cost of extensive tuning, computation power and time. Thus, the dissertation aims to leverage optimization algorithms for feature selection which enables to handle high dimensional dataset efficiently and effectively, allowing to identify the most optimal set of features for training the models, improving model’s overall performance. Most often in network dataset, the features are not linearly correlated, thus, for handling non-linearity of features, optimization algorithms are useful. The features are then used to train two models: the first model performs binary classification to differentiate an attack from a normal event. The second model performs a multi-class classification to identify the type of attack. This enables to handle both the models independently and make choices which allow to get optimal results for the specific objectives of each model. Finally, the project evaluates each model based on the corresponding subset of optimal features obtained from each optimization algorithm and rank the outcomes. Thus, the projects demonstrate mitigating dependence on advance and complex models for higher accuracy, and rather use existing optimization algorithms with Machine Learning algorithms to achieve the same. This also allows to define baseline of results using Machine Learning algorithms, which can be later used as a benchmark for more advanced models.