Title:
Machine Learning Based Anomaly Detection and Risk Mitigation in Large-Scale Cloud-Based Software Systems

Objective:
The project aims to develop and evaluate a machine learning-based system for the detection of operational anomalies in Big Data cloud-based software systems. The project will aim at the mitigation of performance failure and unexpected behavior risks in software engineering in Big Data processing platforms through the application of methods in applied machine learning, cloud computing, and software trace analysis.

Research Questions:
What is the current baseline performance of machine learning-based anomaly detection models on structured operational logs from distributed systems (LogHub datasets)?

Which log attributes (such as event templates, component patterns, and time-based session features) are most predictive of anomalies in operational logs?

How can feature engineering methods (such as session windowing, event clustering, and sequence modeling) improve the anomaly detection performance on structured logs?

Tools and Techniques:

Dataset: https://github.com/logpai/loghub/tree/master/HDFS
https://github.com/logpai/loghub/tree/master/BGL
https://github.com/logpai/loghub/tree/master/Zookeeper
https://github.com/logpai/loghub/tree/master/OpenStack
https://github.com/logpai/loghub/tree/master/Thunderbird

Programming Languages: Python

Machine Learning Libraries: Scikit-learn, TensorFlow(keras)

Data Handling: Pandas

Exploratory Data Analysis: Elasticsearch, Logstash

Anomaly Detection Methods: Isolation Forest, Autoencoders, Hidden Markov Models

Version Control and Collaboration: GitHub

Report Writing: LaTeX

Communication Tool: Stack