

Project Proposal

Server Failure Prediction using Machine Learning

Abdullah Alshahrani

Master's Student, Department of Artificial Intelligence Midocean University

1. Project Overview

In modern IT infrastructures, server reliability is essential for maintaining system availability and business continuity. However, predicting server failures before they occur remains a major challenge for organizations that rely on traditional monitoring tools.

This project aims to design and develop a **machine learning-based predictive model** capable of analyzing server performance data (such as temperature, disk activity, and system errors) to **predict potential failures in advance**.

By enabling early detection of server issues, this project will help IT teams take **proactive maintenance actions**, reduce downtime, and improve overall system reliability.

2. Project Goal

The main goal of this project is to develop an **AI-powered predictive maintenance system** that can:

- Analyze real-time and historical server data.
- Predict whether a server is likely to fail within a specific time period.
- Provide alerts to IT administrators before a critical failure occurs.

Specifically, the model will:

- **Develop a predictive model** to identify servers at risk of hardware failure.
- **Improve operational efficiency** by enabling preventive maintenance.
- **Provide insights** into the key performance indicators that most influence server health.

3. Approach

Data Collection:

The project will use the **Backblaze Hard Drive Failure Dataset** available on *Kaggle*, which includes millions of records from real-world server drives.

The dataset contains attributes such as:

- SMART error metrics (disk health indicators)
- Temperature and power-on hours
- Operational status (0 = healthy, 1 = failed)

Data Processing:

- **Data cleaning:** remove missing or inconsistent values.
- **Feature selection:** choose the most relevant indicators affecting failures.
- **Data normalization:** scale numeric values for model efficiency.

Model/Method Selection:

The project will test several machine learning algorithms for classification and prediction, including:

- Random Forest Classifier
- XGBoost
- Logistic Regression

These models will be trained and evaluated to determine the one with the highest predictive accuracy.

Testing and Evaluation:

Model performance will be evaluated using metrics such as:

- Accuracy
 - Precision and Recall
 - F1-score
 - Confusion Matrix and ROC Curve
-

4. Project Steps and Timeline

Phase	Tasks	Duration
Phase 1	Project approval and title confirmation	Week 1–2
Phase 2	Data preprocessing and exploratory analysis	Week 3–4
Phase 3	Model building and training	Week 5–6
Phase 4	Model testing and optimization	Week 7
Phase 5	Dashboard design and final reporting	Week 8–9
Phase 6	Presentation and submission	Week 10

5. Expected Results

By the end of this project, the following outcomes are expected:

- A **trained predictive model** capable of accurately forecasting server failures.
- **Identification of critical parameters** that most influence hardware health.
- A **simple interactive dashboard** displaying each server's health status and failure probability.
- **Reduced downtime** and improved maintenance efficiency in server operations.

6. Resources Needed

- **Software:** Python, Jupyter Notebook, Scikit-learn, XGBoost, Pandas, Matplotlib, Streamlit.
- **Dataset:** Backblaze Hard Drive Failure Dataset (Kaggle).
- **Hardware:** Laptop or cloud environment with sufficient processing power (CPU/GPU).

7. References

- Backblaze Hard Drive Failure Dataset – Kaggle (2024).
- Breiman, L. (2001). *Random Forests*. Machine Learning Journal.
- Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD.