CAPSTONE PROJECT – ANOMALY DETECTION

**AnomaData** **(Automated Anomaly Detection for Predictive Maintenance)**

**Problem Statement:**

Many different industries need predictive maintenance solutions to reduce risks and gain actionable insights through processing data from their equipment.

Although system failure is a very general issue that can occur in any machine, predicting the failure and taking steps to prevent such failure is most important for any machine or software application.

Predictive maintenance evaluates the condition of equipment by performing online monitoring. The goal is to perform maintenance before the equipment degrades or breaks down.

This Capstone project is aimed at predicting the machine breakdown by identifying the anomalies in the data.

The data we have contains about 18000+ rows collected over few days. The column ‘y’ contains the binary labels, with 1 denoting there is an anomaly. The rest of the columns are predictors.

**Your focus in this exercise should be on the following:**

The following is recommendation of the steps that should be employed towards attempting to solve this problem statement:

* **Exploratory Data Analysis:** Analyze and understand the data to identify patterns, relationships, and trends in the data by using Descriptive Statistics and Visualizations.
* **Data Cleaning:** This might include standardization, handling the missing values and outliers in the data.
* **Feature Engineering:** Create new features or transform the existing features for better performance of the ML Models.
* **Model Selection:** Choose the most appropriate model that can be used for this project.
* **Model Training:** Split the data into train & test sets and use the train set to estimate the best model parameters.
* **Model Validation:** Evaluate the performance of the model on data that was not used during the training process. The goal is to estimate the model's ability to generalize to new, unseen data and to identify any issues with the model, such as overfitting.
* **Model Deployment:** Model deployment is the process of making a trained machine learning model available for use in a production environment.

**Timeline**

We expect you to do your best and submit a solution within 2 weeks.

**Deliverables**

Please share the following deliverables in a zip file.

* A report (PDF) detailing:
* Description of design choices and Performance evaluation of the model
* Discussion of future work
* The source code used to create the pipeline

**Tasks/Activities List**

Your code should contain the following activities/Analysis:

* Collect the time series data from the CSV file linked here.
* Exploratory Data Analysis (EDA) - Show the Data quality check, treat the missing values, outliers etc if any.
* Get the correct datatype for date.
* Feature Engineering and feature selection.
* Train/Test Split - Apply a sampling distribution to find the best split
* Choose the metrics for the model evaluation
* Model Selection, Training, Predicting and Assessment
* Hyperparameter Tuning/Model Improvement
* Model deployment plan.

**Success Metrics**

Below are the metrics for the successful submission of this case study.

* The accuracy of the model on the test data set should be > 75%(Subjective in nature)
* Add methods for Hyperparameter tuning.
* Perform model validation.

**Bonus Points**

* You can package your solution in a zip file included with a README that explains the installation and execution of the end-to-end pipeline.
* You can demonstrate your documentation skills by describing how it benefits our company.

**Data**

The dataset for this project can be accessed by clicking the link provided below.

[Anoma\_data.csv](https://kh3-ls-storage.s3.us-east-1.amazonaws.com/DS%20Project%20Guide%20Data%20Set/AnomaData.xlsx)