Initial Plan and Approach: AI/ML Research, Analysis, Algorithm Model and Learning Method Selection

# Team 1

Riley Dorough

Kayla Echols

Julia Wilkins

Brett Wolff

January 30th, 2022

**Table of Contents**

|  |  |  |  |
| --- | --- | --- | --- |
| **Section/Sub-section Title** | **Author** | **Page #** | |
| **1. Introduction**  **1.1 Document Purpose** | Julia Wilkins | 3 | |
| **2 Model and Methodology**   * 1. **Selected Model and Methodology**   2. **Requirements**   3. **Specifications**   4. **Analytical Conclusions** | Julia Wilkins  Riley Dorough | 3 | |
| 1. **Product and Milestone Schedule**    1. **Implementation Steps**    2. **Deliverable Schedule**    3. **Graphs and Charts** | Kayla Echols  Brett Wolff | 4 |
| **4. References** |  |  |

# Introduction

## Document Purpose

The purpose of this document is to outline the plan for this semester’s capstone project by describing the current model and methodology chosen and the milestones and deliverables as outlined in the state of work. This document is subject to change as Team 1 moves forward and adapts the plan and approach to better accomplish the goal.

# Model and Methodology

The following sections outlines the approach and goals for this project. However, the approach is subject to change as advancements are made and the team gains a better knowledge of how to best approach the project.

* 1. **Selected Model and Methodology**

The initial plan is to adopt decision tree prediction methods until causational relationships between data inputs is better understood. As more data is included into the pipeline and a better understanding of how data intersects is understood, we intend to switch to a more precise mathematical approach. Specifically, regression learning appears to be effective as it is fully understood and minimally randomly varied data sets.

* 1. **Requirements**

There are a few requirements for this project. We can assume that we will be working with the network architecture as previously outlined. We will also be using Knime to establish the machine learning pipeline. Knime will be using data from a CSV file pulled from our MS SQL Server that is collecting data from the Opto Snap Pack PLC. We will also be using network data collected from a Graylog server hosted on the Ubuntu server.

* 1. **Specifications**

As of now Team One’s current approach is to use KMeans clustering algorithm and Decision Tree learning to establish normal fluctuation in data and then predict and detect anomalies within that data. As more types of data are introduced to the pipeline different data sets will need to be manipulated in different ways before they are integrated into our anomaly detection machine learning pipeline. This approach is subject to change as the other teams share knowledge and learn from each other. Team One expects for the endo goal of this phase of the project to change as advances are made and obstacles are overcome.

* 1. **Analytical Conclusion**

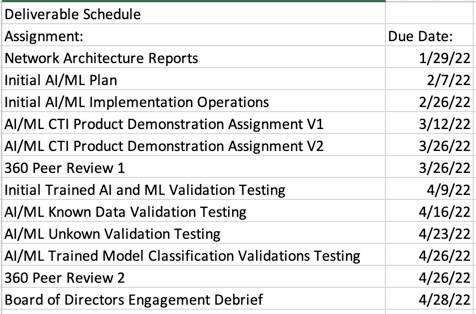
The natural variation of data available to our team needs to be understood to begin attempting to predict or recognize abnormalities. Along with this, classified safe boundaries can be established to negate the need to label any data with a high enough contrast to the common set as an “abnormality”. As an example, a data set of values that ranges from 0-10 would classify a single data point that reaches a 12 as an abnormality. While at the same time, a data set that ranges from 0-10 with a single data point of 12 which falls under a safe operating range of 0-20 would be an outlier, however, would not impact the operations of the physical systems. The scaling and span of data sets can also show the beginning/ end of trends which is different than a previous trend. The future goals of our team are to understand what data should be considered or chosen to have abnormal results. The chosen data should be normalized with the other teams eventually to the best of operational reason. Lastly, attempts will be made to derive predictable correlation between data inputs.

1. **Product Milestone Schedule**
   1. **Implementation Steps**

Initially, the implementation will be done in a controlled manner with carefully collected data loaded into CSV files. This is due to the nature of how to train the algorithm to understand what is normal and what should be brought to the attention of the researchers. This CSV will be a cumulation of both network traffic, and various numbers from the OPTO system. When it is deemed trained enough, live data will be fed into the pipeline under supervision and scored for accuracy. Once comfortable with the accuracy of the pipeline and the algorithm, it will be tested over a longer timeline with less supervision, as if it was fully released, however someone will be on stand-by in case something was to break. Should the algorithm and pipeline pass all these, it will be released fully to the client along side documentation to further evolve the ML algorithm,

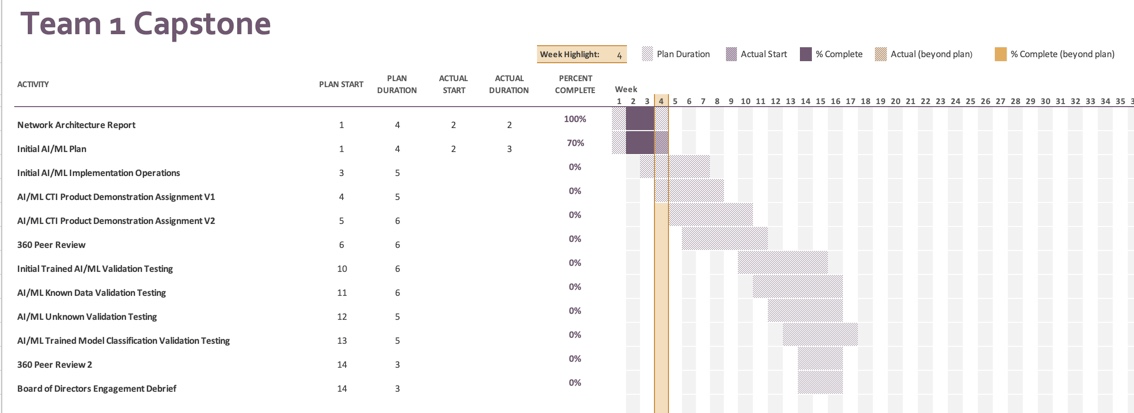
* 1. **Deliverable Schedule**

The deliverable schedule is based on the deliverables given to the team by the project manager. The expectation of the project manager for the team is to have these assignments completed by the dates as seen on the chart. The chart will be used by the team as a guide for when to have the deliverables completed unless otherwise notified by the project manager.



* 1. **Graphs and Charts**

**Gantt Chart**



**References**

Sipantzi, T., Tucker, R. (2021). System Requirements Specification Template. CSIS 471: Software Development.