|  |  |  |
| --- | --- | --- |
| CSIS 486 | Riley Dorough  Kayla Echols  Julia Wilkins  Brett Wolff | AI/ML Real Time Anomaly Detection and Intra-Team CTI Demonstration- Version 2 |
|  | **April 23, 2022** | **Module 6, Phase 2** |

Title [Kayla Echols]

* The goal of this lab is to report on the accuracy of the points for the team’s pipeline.
* Another goal of this lab report is to document what has been done for this project to be used in the next semester.

Abstract [Kayla Echols]

Team 1 was able to finalize this project by documenting the anomaly detection and explaining the level of accuracy of the team’s pipeline. The team is making the proper documentation in order to pass this project along to the next team of students. This lab explains how we classified the OT and IT data in rows in the chart and what we classified as anomalies. The team also evaluated the level of accuracy based off the nodes that were used and how the data was classified within Knime.

Introduction [Kayla Echols]

The team was able to come to finish this phase of the project by documenting everything that has been done this past semester and explaining the accuracy of our pipeline. Our team was able to put together an anomaly report framework as a result based off the anomaly detection that we have created using the pipelines. This framework will be helpful for the future capstone teams as they begin to attempt to predict the anomalies that we detected. The team was also able to prove accuracy of the pipelines and because all of the nodes that we used were manually classified, the pipeline is considered 100% accurate. The team relies on the binner nodes within Knime to classify the data with specified ranges. This report will be able to help the future capstone teams gain a better understanding of how our anomaly detection works and our accuracy within the pipelines. In order to complete this lab, the team needed to have access to the Liberty VPN, the vSphere virtual machines, and the Knime software.

Method [Riley Dorough]

**Updates Since Last Document**

* Finalized anomaly reporting utilities and documentation for the semester
* Additional focus on understanding data and gathering accuracy of various points of our pipelines

**IT & OT Data Anomaly Reports**

**Graphical user interface, application, table

Description automatically generated**

* This report features a framework designed to include four separate streams of anomaly input
* Classification Anomalies for OT Data includes any direct anomalies recorded by Team 1’s classification of OT data
* The OT Rate of Change section includes information pertaining to Team designed classification for the rate of change of OT data
* Classification Anomalies for IT Data includes any direct anomalies recorded by Team 1’s classification of IT data
* The OT Rate of Change section includes information pertaining to Team designed classification for the rate of change of IT data

**Graphical user interface, table

Description automatically generated**

**Table

Description automatically generated**

* The data collected includes only rows of data from each of the four sections which includes at least one anomaly from each category
* If our team had more time, we would attempt to flag which column contained the anomaly for additional classification and prediction
* Data prediction should be focused on for research as later teams focus on continuing phases of the project

**Review of Accuracy of Pipeline**

* Due to the manual classification of data for nodes, it is not believed that any are less than 100% accurate for each of their intended purposes
* Binner nodes often rely on range assignments for data and follow a simple ruleset
* K-means nodes and auto binning are based on even splits between the lowest data node and the highest or evenly clustering data into groups of equal size or equal range of possible values
  + Since the data is being initially classified, it is not possible for it to be in the wrong bin due to the classifier telling the data which bin it belongs to
* As data prediction begins, you will be able to reach values of accuracy less than 100% as data will need to be compared to other data or test the likelihood of data correlating based on matching anomalies over time

Results [Julia Wilkins]

The main focus of this lab was to address accuracy and to finalize our reporting. Due to the nature of how we are classifying our data, we can say that the pipelines are 100% accurate. For OT data we chose to use binners that would classify data based on a range of values that we chose. This range was chosen from what we knew about the data. For IT data we set the binners to classify data based on width of the range of the values being binned. This bins according to how similar in value the data values are. So, anything that is different to the normal values will be binned accordingly. Using this system gives us the advantage of having more control by setting the ranges for what is considered normal for our data and gives us perfect accuracy.

Our reports are currently set to pull all anomalies within a specified 24-hour period. This can be adjusted based on the desired report to be produced. This feature was added to aid in correlating data based on time. It allows users to see what anomalies occurred when and then those time stamps can be compared to see if multiple anomalies occurred at the same time.

Discussion [Brett Wolff]

For this lab, the team worked on concluding the work that was done toward anomaly detection for the semester. Using data from the pipeline and previous experiences, paired with the experiences of other teams, ranges were set up for both the OT and IT systems. This meant that anything outside that range would be an anomaly and would require either further action to prevent an attack or further investigation to see why there is an outlier. This method gave the team more control over what would need to be deemed anomalous and provided the perfect accuracy score.

Should a report also need to be pulled from the data, it was also made possible to pull just the anomalies. Added to this was data based on time, allowing the person reading the anomaly report an idea of when something happened, and allowing the ability to narrow down where to look to see if more data needs examination.

Conclusion [Brett Wolff]

This lab is a concluding lab as the semester ends. The team is wrapping up what data can be input into the pipeline and finalizing the pipeline’s ability to read the data. This was done using both forms of available data, feeding it into the pipeline and teaching the pipeline how to read and predict data.