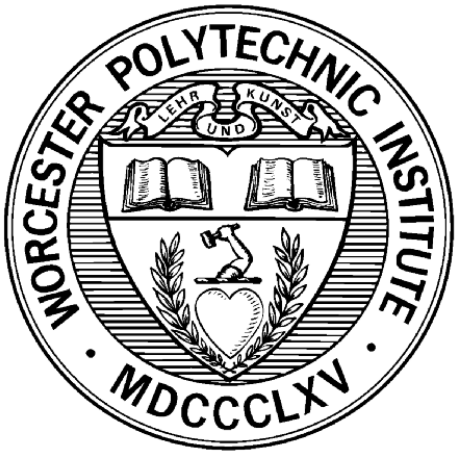
DESIGNING A DATA-DRIVEN FRAMEWORK FOR SMART AND AUTONOMOUS FREIGHT FARMING



A Major Qualifying Project Report

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science.

Submitted By:

Shivangi Pandey

Steven Yevchak

Sponsor: Freight Farms

Advisors: Yanhua Li, Krishna Kumar Venkatasubramanian

ABSTRACT

Indoor hydroponic farming has become an industry changing technology that has allowed for crop growth in areas of the world where it would never have been expected before. Freight Farms' Leafy Green Machine allows for farmers to grow crops throughout the year, by controlling the climate inside the farm; it also allows the farmer to not concern themselves with the external environment. However, the farm is not able to predict how the climate of the farm will change based on its current sensor readings and equipment states. To allow the farmer to see how the farm will behave in the future, machine learning algorithms can utilize these readings and states to predict the future climate readings and notify the farmer of any harmful changes. This project seeks to build a predictive machine learning model to add further measures to help maintain the Leafy Green Machine's self-sustaining climate.

ACKNOWLEDGEMENTS

We would like to thank our sponsor, Freight Farms, for their assistance during the project. We would also like to thank our advisors, Professor Krishna Venkatasubramanian and Professor Yanhua Li for their feedback and support throughout our project experience. We would also like to give a special thank you to Hang Cai for his time, assistance, feedback, support and guidance throughout the duration of our project experience.

TABLE OF CONTENTS

TABLE OF FIGURES

Chapter 1: Introduction

Traditionally, farming was seen to be a laborious and time-consuming profession that yielded unpredictable results. Harvesting would only occur once a year due to seasonal restrictions, which could be further effected by natural disasters, animal infestations, or crop disease. To remedy the problem of a single harvest, hydroponics was developed to grow crops without soil. This method allows for the use a nutrient rich water solution to feed plants. Hydroponics allowed for farmers to grow crops in non-traditional locations. Recently, the consumer market for hydroponic crops and farming has grown significantly. Crops yielded from hydroponics are both cost-effective and easy to grow. One of the leaders in vertical hydroponic farming is Freight Farms.

Freight Farms' Leafy Green Machine (LGM) uses vertical towers, and yield more crops than the conventional one acre of farmland. As opposed to the traditional 1 harvest per year, a freight can have anywhere from 8 to 12 harvests per year. The marketable yield of crops of a LGM is up to 93% as opposed to 75% from one acre of farmland.

The LGM can be set up anywhere around the world, the only necessary requirements are access to level ground, electricity, and water. This allows parts of the world that are not suited for traditional agriculture to produce crops year-round. The controlled environment permits optimal yields for any given crop, provided the correct parameters are maintained in the LGM. To maintain these parameters, it is important to know when the farm waivers outside the given ranges.

Leafy green vegetables, such as different types of lettuce, spinach and herbs are grown in the LGM. Freight Farms provides its farmers with the optimal growing conditions for each of these crops. While some conditions, like amount of sunlight, are simple to control, others, like

pH and electrical conductivity of water, cannot be controlled by the flip of a switch. They are influenced by many equipment inside the farm as well as the environment outside the farm.

The LGM provides sensor data detailing the current conditions in the farm (i.e. air temperature, humidity, pH, and other readings useful readings). The state of the equipment (on or off), used to manipulate the environment inside the farm, is also given.

This project's goal was to devise a data-driven solution, using the provided sensor and equipment data as features, to build a predictive model. The label was determined by the sensor reading the farmer is attempting to predict, allowing for a model to be trained on the farm's existing dataset to provide predictions about the future climate.

The datasets provided by a given LGM required extensive preprocessing, before they were used by the classifiers. The primary classifier used was random forest. For the predictive model to be considered successful it should predict out a given number of minutes with 80% accuracy. To demonstrate the dataset could be predicted with that high level of accuracy two parameters were chosen, air temperature and humidity. Three equipment, lights, coolbot, and main pump, were chosen as attributes that affected the given sensor readings.

These five inputs were used as features for classifiers. To further authenticate the accuracy of the resulting model, the results were cross validated and tested against data that was withheld from the model during training.

Initially, we saw promising results, indicating it would be possible to predict sensor ranges at least 10 minutes into the future with 80% accuracy. At that time, it was our goal to predict if the reading for the sensor would fall within an acceptable range. A major problem we encountered was that the farm was almost always with an acceptable range for the sensor readings; there were very few instances of the farm being outside the optimal ranges, giving us

insufficient amounts of data for the classifiers to train on. We then shifted our focus to predicting ranges that contained similar amounts of data. The results were less successful with the best classifiers only getting 75% accuracy for predicting the shortest duration in the future of 5 minutes.

The discussion will start with background information about Freight Farms, hydroponics, and related work in Chapter 2. Chapter 3 will introduce the problem of this project. Chapter 4 articulates the approach taken for building the model's framework. Chapter 5 further delves into the different experiments that were conducted, with their respective results. Possible future work and recommendations are described for this project in Chapter 6. Finally, Chapter 9 presents the conclusions of this project.

Chapter 2: Background

To understand this project, it is pertinent to have an understanding of hydroponics as a whole and Freight Farms' Leafy Green Machine. Hydroponics have evolved over the past few decades and now allow farming in an indoor environment. Freight Farms has built upon hydroponics to build a self-sufficient freight that produces a constant optimal climate crops.

2.1 History of Hydroponics

In 1634, Jan Baptist van Helmont conducted the "Willow Tree Experiment." While he was put on house arrest by the Spanish Inquisition officers for studying plants scientifically. The theory at the time was that plants got their nutrients from the soil. Helmont studied this theory by weighing a willow try and dry soil. Once he planted the tree, he watered it and let it grow for five years. The tree had grown significantly and increased in mass. When he weighed the soil after drying it, he noticed it had the same mass. Therefore, he concluded the tree grew by drawing its nutrients from water. His experiment revolutionized botany and changed the way plant growth was studied [1].

Building off of the discovery of Helmont, in the 19th century Julius von Sachs and W. Knop discovered the necessary nutrients plants needed to for plant growth, nitrogen, phosphorus, and potassium, earning them the names "The Fathers of Water Culture." In the 1920s Hoagland's Solution, developed by Dennis Hoagland, took account of the micronutrients necessary for plant growth, magnesium, sulfur and iron. His solution provided the essential nutrients for crop growth wherever weather and sun permitted [2].

During World War II, the American military used hydroponics to grow vegetables in the Middle East and Pacific Islands to avoid transporting foods. This allowed for soldiers to grow fresh food in the harshest environments. This practice continued through the Korean War. In the

1960s, the Nutrient Film Technique (NFT) was developed as a drip system to provide nutrients directly to the roots of the plant. In the 1970s, after Nixon's crackdown on the Mexican border for stopping marijuana from being brought into the United States, companies started producing equipment use hydroponics in an indoor environment. After further research over the past few decades, companies have begun building indoor systems for year-round crops [3].

2.2 Freight Farms

Freight Farms was founded in 2010 by Brad McNamara and Jon Friedman. It began with a Kickstarter campaign with a prototype for the current Leafy Green Machine. The goal was to produce crops year-round in climates and locations not hospitable to agriculture. Crops grown include many types of lettuce, hearty greens, and herbs.

Leafy Green Machine

Growth Cycle

The growth cycle in the Leafy Green Machine is concerned specifically with germination, seedling growth, and mature plant growth [4].

The germination process takes 5 to 14 days. Seedling trays are filled with growth plugs which each contain 1 to 2 seeds based on how likely the specific type of seed will sprout. To determine the number of seeds the farmer plants he desired number of plants and divides it by the germination rate. For example, if 275 Bambi Bib Lettuce seedlings are desired, they are divided by the Bambi Bib Lettuce germination rate which is 0.99. Leading to 728 seeds being planted. Seeded trays are then placed on the germination shelf where a humidity dome is placed on top of the tray to preserve moisture during germination. Once the seeds germinate and 2

leaves form, the tray is moved to the seedling trough. If more than one plant sprouted in a single cell they are separated [4].

Seedling growth stage is approximately 2 weeks long. During this time seedlings receive water and light periodically. True leaves form during this stage. A healthy seedling will have roots wrap around the bottom of the grow plug and have the stem strengthen, allowing for it to be pulled from the tray. Healthy roots are white and not slimy. When seedling leaves turn yellow, it can indicate many potential problems. The plant may not be getting the right nutrients or receiving unequal doses of nutrient A and B. Other causes are EC not being set to 700, poor water quality, and the pH not being in the range of 5.8 to 6.2. Slow growth comes from similar problems to leaves turning yellow like EC and pH unbalance and root rot. It can also come from exposure to differing air temperature and humidity or exposure to tower conditioning in early stages of growth. Some slow growth may just be normal growth for some crops the look smaller compared to other. Herbs, kale, and swiss chard seedlings look smaller than lettuce seedlings.

At the end of seedling stage, the plants are transplanted from trays into vertical towers for the mature plant growth stage; plants grow from seedlings to mature plants in the towers. The plant in the growth plug is planted perpendicular to the tower with the wicking strips lying flat against the growth medium; it is set at least half an inch from the front edge of the tower. The bottom third of the plant's grow should be in contact with the wicking strip. Both over contact and lack of contact are undesirable. Over contact can oversaturate the plant, potentially causing stem rot. The grow plug should not stick out past the front face of the grow medium, this differs based on if the plant needs its stem supported or not. Plants should be monitored throughout growth checking for adequate access to water, proper airflow and signs of disease and pests.

Wilted plants may be the result of a lack of water from the emitter becoming clogged or wicking strop placement being incorrect [4].

Before plants are harvested the main pump should be turned off. New seedlings can immediately be placed in the tower. Mature plants are harvested with their grow plug intact. Plants that are harvested by trimming should be removed from the tower and place on the harvest rack where they can be trimmed. Basil is trimmed right above its second set of leaves and can be harvested every two weeks. Kale, swiss chard, and collards are harvested by breaking off outer larger leaves from the base of the plant. Once trimming is completed, they should be placed back on the tower. [4]

Climate Ranges

Crop	Arugula Mustard Greens Lettuce	Basil	Kale Swiss Chard Spinach	Cilantro Mint Dill
Lights On / Off	14:00 8:00	14:00 8:00	14:00 8:00	14:00 8:00
Temperature Day / Night	63° 60°	78° 68°	63° 60°	68° 63°
Humidity Day / Night	65% 65%	65% 65%	65% 65%	65% 65%
Main Nutrient Day / Night	1200 1200	1600 1600	1800 1800	1300 1300
Seedling Nutrient Day / Night	700 700	700 700	700 700	700 700
CO_2 PPM	1100	1250	1250	1100

Figure 1:Ideal Climate Ranges for Crops

Crops need to grow within specific ranges. For lettuce, kale, and basil, the ranges should be as follows: the seedling EC should be 600 to 900 µs/cm2; the seedling and main pH should be 5.8 to 6.2; the seedling and main water temp should be 35ºF to 140ºF; the CO_2 level should range from 500 ppm to 2000 ppm. For lettuce, the air temperature should be 53ºF to 73ºF, have a

humidity from 50% to 80% rH and have a main EC from 1000 to 1500 µs/cm2. For kale, the air temperature should be 53ºF to 73ºF, have a humidity from 60% to 80% rH and have a main EC from 1600 to 1800 µs/cm2. For basil, the air temperature should be 63ºF to 88ºF, have a humidity from 60% to 80% rH and have a main EC from 1400 to 1800 µs/cm2. [4]

Sensors and Equipment

During the Nursery Stage, the seedlings germinate in an aluminum workstation for about three weeks before being put in vertical towers. The space can grow up to 3,600 seedlings with the help of an irrigation system and a LED lighting array. After the germination period, the seedlings are transferred to one of 256 towers in the four rows of 7' vertical towers. The container makes use of a drip irrigation system and a strip LED lighting system for the growth stage of the crops.

Climate Control

The climate of the LGM can be set to the ideal environment for each crop being grown inside the container. Environment sensors measure different factor, such as temperature, humidity, CO2, nutrient levels, and etc., to maintain the ideal climate for the crops by interacting with the farm controller. The container makes use of a multi-planes airflow and intercrop aeration system for air circulation. The ventilation system maintains the temperature and humidity of the environment with a 24,000 BTU air conditioner.

Lighting System

Freight Farms uses a high efficiency LED lighting system to imitate sunlight for the crops. This is done by emitting red and blue colored light for photosynthesis. Approximately,

128 strips run for 18 hours a day. This gives the plants enough light to maximize the growth cycle. The six-hour break allows for the plants to get time to rest and avoid using electricity during the day.

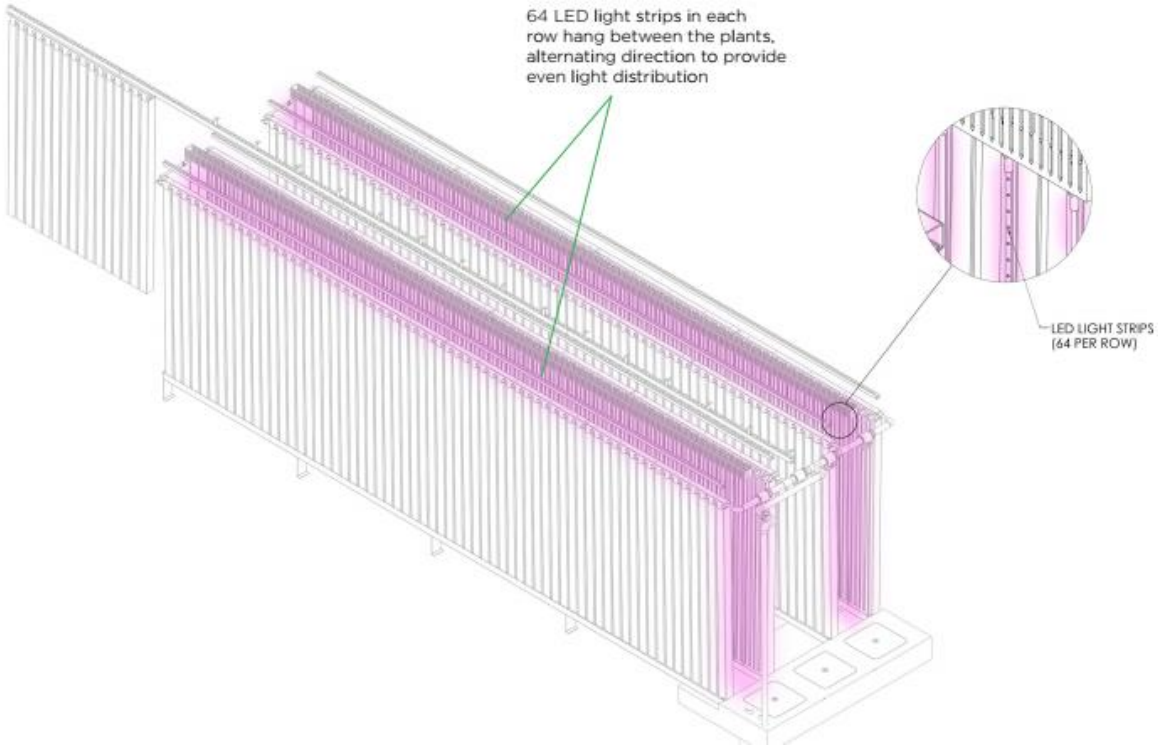


Figure 2: LED Lighting System

Irrigation System

The LGM utilizes a close-looped hydroponic irrigation system to deliver a water solution, rich with nutrients, to the plant roots. The ensures all plants grow uniformly. The nutrient dosing panel interact with the temperature, pH, and EC sensors to control the water conditions and check for the optimal levels of nutrients for growth.

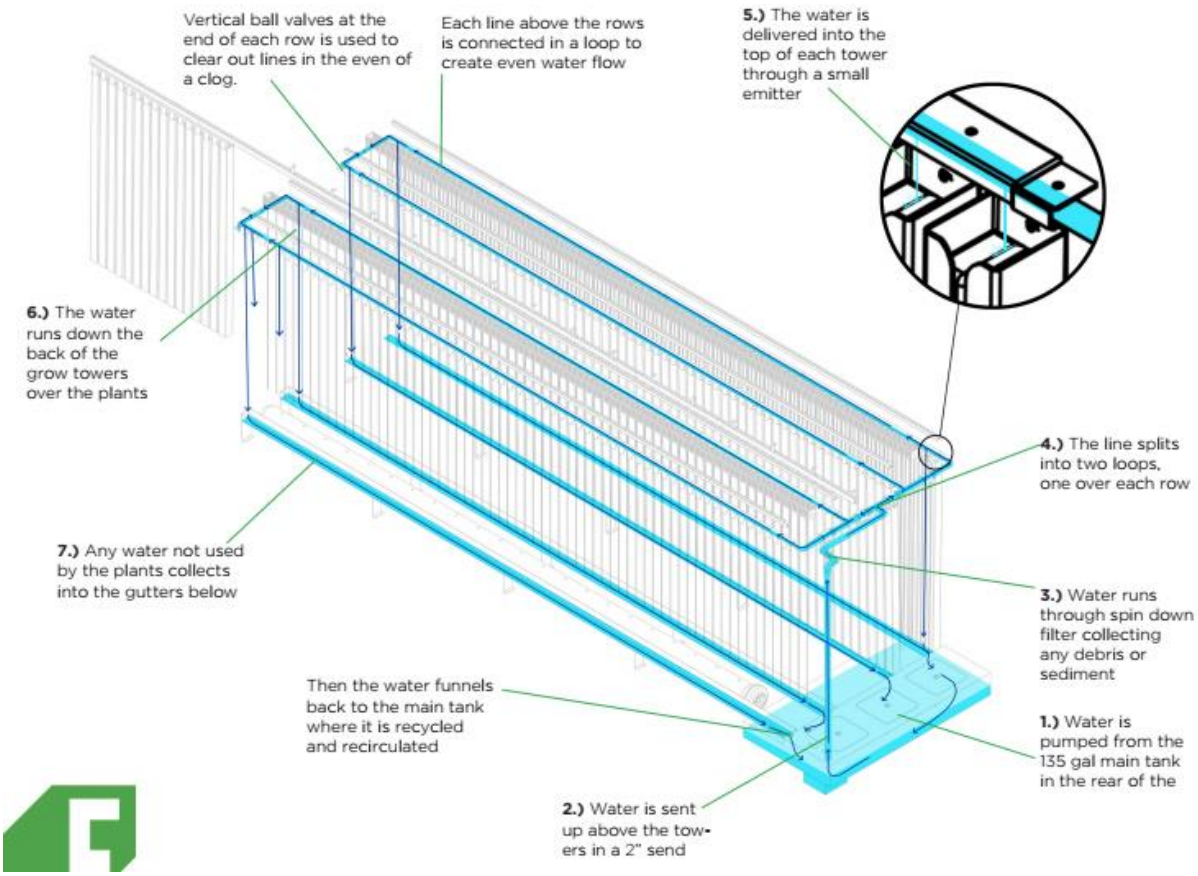


Figure 3: Freight Irrigation System

In-farm Controller

The container comes with an in-farm controller that communicates with the sensors to maintain an optimal environment for the crops. The farm's data is displayed on a weatherproof screen so the information can be easily accessed by the farmer.

Farm Camp

To ensure the LGM is operates correctly, Freight Farms offers Farm Camp as a means to provide the necessary tools to each new farmer. The camp is composed of in-farm lessons and classroom sessions.

2.3 Related Work

In one study, neural networks are incorporated in a predictive greenhouse control strategy for inside air temperature. The authors of the paper modeled air temperature based on the relative humidity and the outside temperature, along with incorporating solar radiation. They experimented with multiple different training and learning models and compared them. They utilized Multi-Objective Genetic Algorithms in this study [5]. Another study was done using simulations with the goal of predicting air temperature an hour in the future. That study used a particle swarm optimization algorithm to predict temperature with success in their simulations [6].

A third study used neural networks to predict the temperature and humidity in a greenhouse. That study also manipulated the greenhouse environment in addition to its prediction. The experiments were also largely concerned with energy efficiency of the greenhouse and how predictive models could lead to a more efficient use of machines controlling the environment. The work in this paper differs by using historical data with not ability to influence the environment of the greenhouse. The results also were not simulations, but tested on different part of the data set the classifiers had not been trained on. The main classifier used for prediction was a random forest unlike the commonly used neural network which was tried but found inaccurate [7].

Another group took the approach of developing an algorithm to predict greenhouse conditions which would optimize profits for the tomato crop's production. The algorithm used two different programs, one that calculated crop yield, and a second that calculated the energy costs of the greenhouse in reference to the external climate. In conjunction, the two algorithms

predicted the set of climate parameters for each harvesting period. The overall goal was to minimize energy costs and maximize crop yield [8].

Chapter 3: Problem Statement

Our goal was to build a predictive model that would understand the different equipment states to predict the future sensor data and the environment of the farm. Our hope is that the farmer can use this predictive model to see the projected future environment of the farm and decide if it is optimal or not. To make the preliminary model, we used the following equipment data:

- Coolbot (AC)
- Main Pump
- Lights,

and the following sensor data:

- Temperature
- Humidity.

Chapter 4: Approach

To build a predictive model, first a large amount of data needed to be obtained and processed into a usable dataset. Pre-existing models were studied and tested on the respective dataset. The main motivation was to find the optimal model and parameters for the data.

4.1 Data

Freight Farms provided us with 9 months of data from 1/1/17 to 8/31/17 from the farm located at their headquarters in Boston. This dataset included sensor and equipment readings for the freight. The original raw dataset contained instances where there was an overlap in data between months and windows of time where days of data was missing. In order to build our model, we needed to preprocess the data so that it could account for these anachronisms.

Data Preprocessing

To preprocess the data, we first cleaned the data to remove the overlapped data and ignored the missing data readings. We wanted to ensure the data was continuous. Once the data was continuous, we preprocessed the data so that it could be used by the predictive models.

First, the data was divided by its start time and end time. This was done in the following manner, the data was either divided into 4 months of continuous data (January to April, March to June, or May to August) or kept in its original 9 month dataset. The goal was to use the first 70%

of the dataset for training the model that would be used for prediction and then the next 30% for testing the model, as described in the figure below.

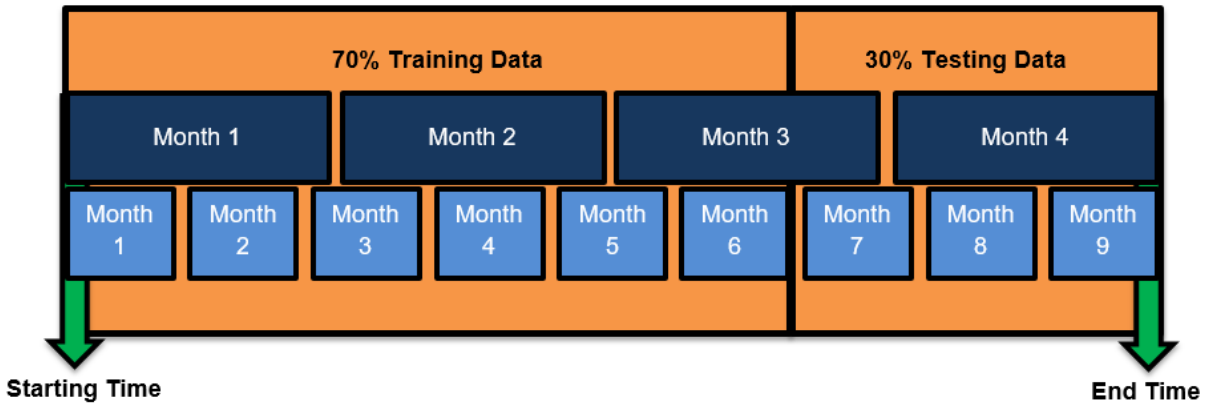


Figure 4:Training and Testing Data

Before the data could be divided into testing and training, it had to be preprocessed and but in a data frame so the model could extract features for training and predicting. The model used three equipment ($E_{Equipment}$) states and two sensor (S_{Sensor}) readings and calculated the average value for the allotted chunk (C_x) size (i.e. 5 mins, 10 mins, 15 mins, etc.) to build a feature within a block (B_X) of data. A block of data is a specific size (30mins, 1 hr, 2hrs, 3 hrs, etc.) that was used for training and testing the model.

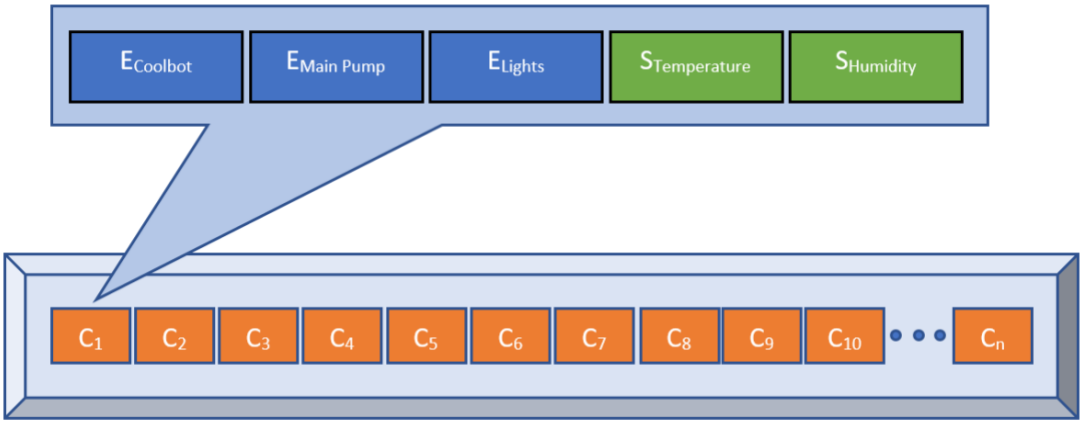


Figure 5:Preprocessed Data Block

The model would use these blocks to train the data and predict the next readings for the next n minutes using C_n. The goal was to use the previous block as the current state of the environment so that it could predict the next n chunks of data readings for the farm.

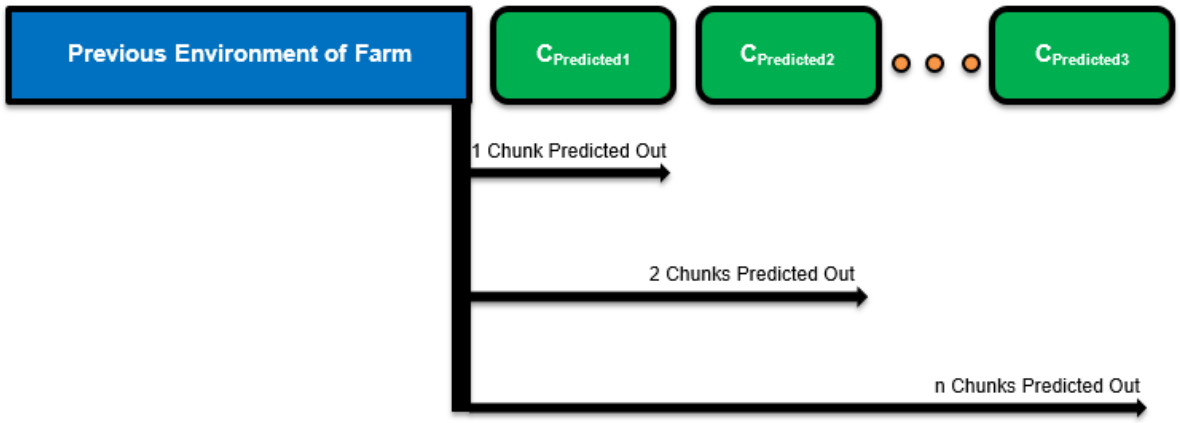


Figure 6: Predicting Data

4.2 Model

Random Forest

Random Forests are a type of ensemble learning method. They can be used for classification, regression and many other statistical methods; they were designed for decision tree classifiers. The forest contains multiple decision trees, where each tree is created using a set of random vectors. Each vector is generated from an established probability distribution. The random forest is preferred over decision trees because they help to avoid overfitting of the model to the training set. [5]

Cross Validation

Cross-validation is training method that allows for a model to train on as much data as possible to improve itself. This method divides the data equally into partitions, where each partition is used k – 1 times and once for testing. The process of utilizing all of the partitions for

training and testing is done k times and allows for the entire data set to be covered by the model for training and testing. [6]

K-Nearest Neighbor

K-Nearest Neighbor is a proximity based algorithm that finds the distance between the i^{th} object to all its other neighbors of object. The algorithm then sorts the distances of the neighbors from the object in decreasing order, while keeping track of which object the distance belongs to. The algorithm then returns the neighbors associated with first K distances from the sorted list.[7]

Gaussian Naïve Bayes

The Naïve Bayes algorithm converts the data into a frequency table and finds the likelihood for a possible outcome. It then uses the Naïve Bayesian equation to calculate the probability of the outcome occurring. Once it has calculated the outcome, it chooses the option with the highest probability of occurring. This is done assuming a gaussian or normal data distribution. [8]

Chapter 5: Tests and Results

Six experiments were performed on the preprocessed dataset. As parameters were tuned the results generally improved for the first five experiments, indicating that the predictive model could predict up to 10 minutes in the future. The changing variables in the first five experiments were block and chunk size; this effected the amount of data the model had to train and test on.

The first five experiments used the same values for temperature and humidity ranges. Ranges were changed in the final experiment, proving it to be more difficult for the model to predict the future temperature and humidity accurately.

5.1 Experiment 1

Methodology

The dataset used was from the beginning of May to the end of August. Data labels for temperature ranges were: “Too Hot” > 85°F; 85°F ≥ “Hotter than Expected” > 70°F; 70°F ≥ “Expected Range” ≥ 60°F; 60°F > “Colder than Expected” ≥ 55°F; “Too Cold” < 55°F; chunk size was C = 5, block size was B = 11, and predicted chunk was P = 1. This meant that 55 minutes of data was used, and the label for that block of data the next 5 minute chunk slotted into the correct air temp range. These experiments were run as above to find ideal random forest parameters in a realistic environment that will predict the state of farm in the next five minutes, forest size is the variable changing.

Results

Cross Validation was also done on the same dataset using a random forest classifier with the same parameters as above. The first 70% of the chronological May to August dataset, or the

training dataset, was used in the cross validation set. The number of trees was used to as the factor to contribute the most to the variance of the model, but it had little effect.

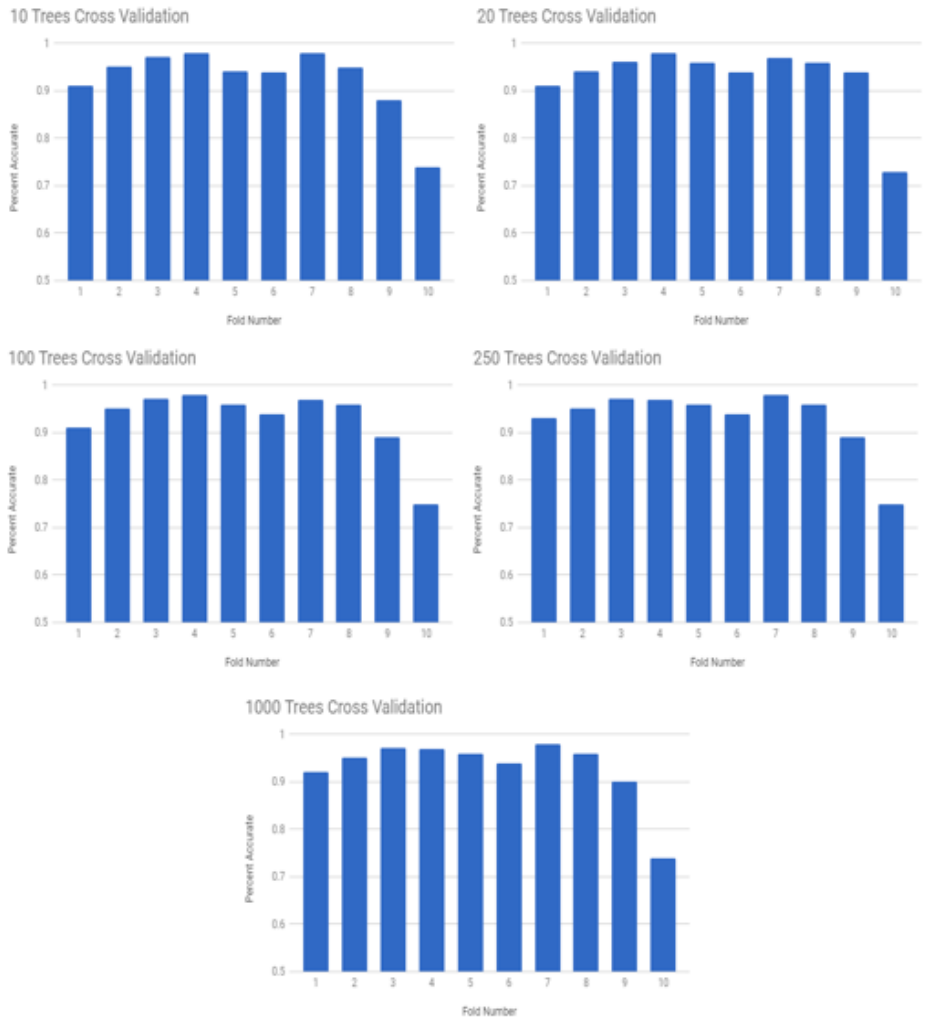


Figure 7: Cross Validation Results Using Different Trees

Cross validation put the random forest at around 92% to 94% accurate as shown in the Figure 7. The results are similar between the number of trees used in the classifier with only minor differences between the best and worst classifiers. The best classifier, by a small margin, used 100 trees.

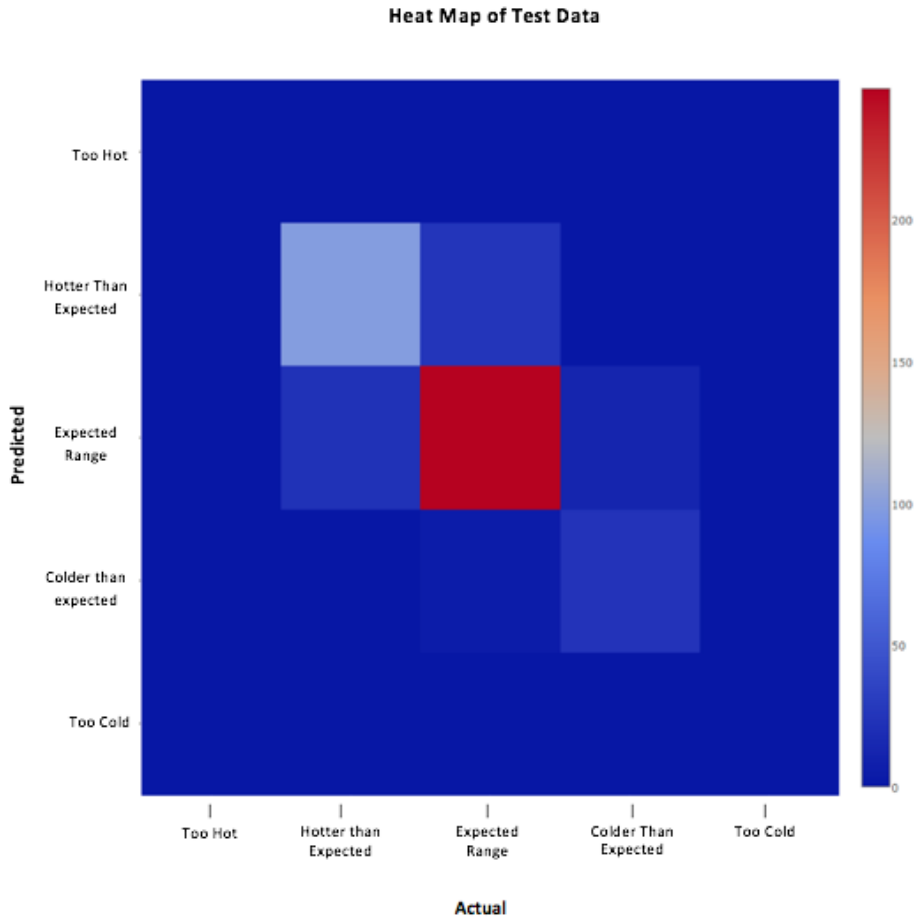


Figure 8: Heat Map of Confusion Matrix Using 100 Trees

Using the results from cross validation the random forest was trained on all the training data and then tested on the remaining test data. Figure 8 shows the confusion matrix with 100 trees, the best result from cross validation, in the random forest. The vertical labels are the predicted values while the horizontal labels are the actual value. For instance, in the matrix the slot containing 24 indicates that the model predicted the block to be in the "Expected Range" how "Expected Range" however they were actually "Hotter than Expected". The diagonal from top left to bottom right indicates correct predictions. This model was 84.72% accurate in predicting the test data.

This experiment served as a basis for future experiments. It allowed for us to test the random forest model built.

5.2 Experiment 2

Methodology

This dataset used was again from the beginning of May to the end of August. Chunk size was C = 5 and block size varied so that B = [1 hour = (11,10,9), 2 hours = (23,22,21), 3 hours = (35,34,33), 30 mins = (5,4,3)]. The predicted chunk varied so that P = (1, 2, 3). This meant that 9 chunks or 45 minutes of data was used to predict the 3rd 5-minute average. In other words, 45 minutes was used in the classifier and the label range was determined based on the average from 55 to 60 minutes. Similarly, 50 minutes was used with the label being the 2nd five minutes. This same treatment was applied to other sets of block sizes so that the classifier would be predicting 10 and 15 minutes out. The same process was done for humidity, where the ranges for humidity were “Too Humid” > 80%, 70%, ≤ “More Humid than Expected” ≤ 80%, 60% ≤ “Expected Range” ≤ 70%, 50%< “Less Humid than Expected” < 60%, “Not Humid Enough” < 50. Air temp was also used with the same range parameters as the previous experiment. The number of trees used remained the same as the previous experiment, for the graphs the number of trees giving best result was chosen. Cross validation and 70/30 was run on the formatted dataset. The goal was to see how the model would predict for humidity data compared to temperature.

Results

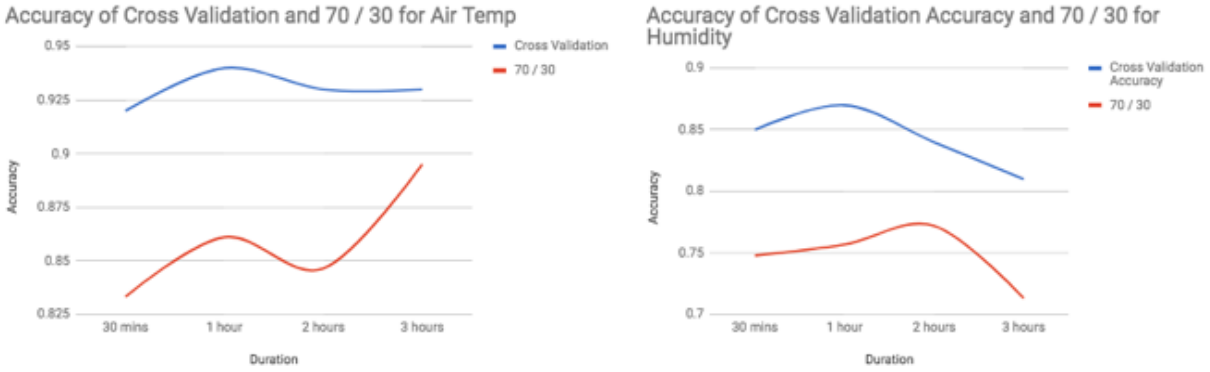


Figure 9:Cross Validation Results from Experiment 2

The cross validation results and 70/30 prediction results show a trend that cross validation is more accurate than 70/30. It also showed the prevailing trend that air temp typically improves with more data and the humidity generally does worse. Figure 9 only shows the accuracy at predicting 5 minutes out and that relatively high accuracy prediction for a short interval is possible under the given parameters.

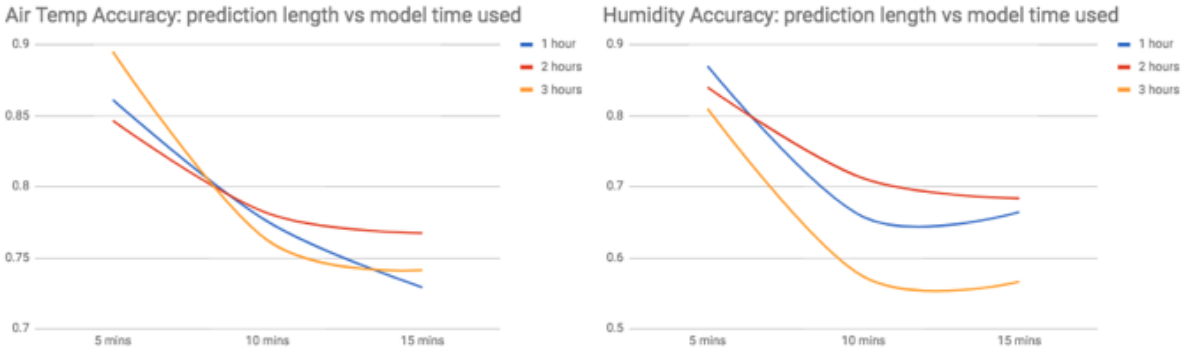


Figure 10: Accuracy Results Based on Time Predicted

Figure 10 shows only three of the main block sizes and the accuracy of predictions for 5, 10 and 15 minutes out. Accuracy falls the further out predictions are made and similar to the

cross validation results more data improves accuracy of the classifier for air temp and gives worse accuracy for humidity. The classifier is also better at predicting air temp than humidity.

This experiment allowed for us to test our model on humidity from the sensor readings and understand how well the model predicted on humidity data.

5.3 Experiment 3

Methodology

This dataset used all data provided by Freight Farms breaking it up into three sections from January to April, March to June, and May to August. Chunk size was C = 5 and block size varied so that B = [1 hour = 12, 2 hours = 24, 3 hours = 36, 30 mins = 6]. The predicted chunk varied so that P = (1, 2, 3). Using these parameters for the experiments removes the problem of unequal data for the number of chunks predicted out. The same number of chunks was used regardless of the chunk that was being predicted on. As in previous iterations, the first 70% of the data was used for training and the next 30% was used for testing. Our goal was to see how these changes would affect the accuracy of the model compared to how the model was setup in previous experiments.

Results

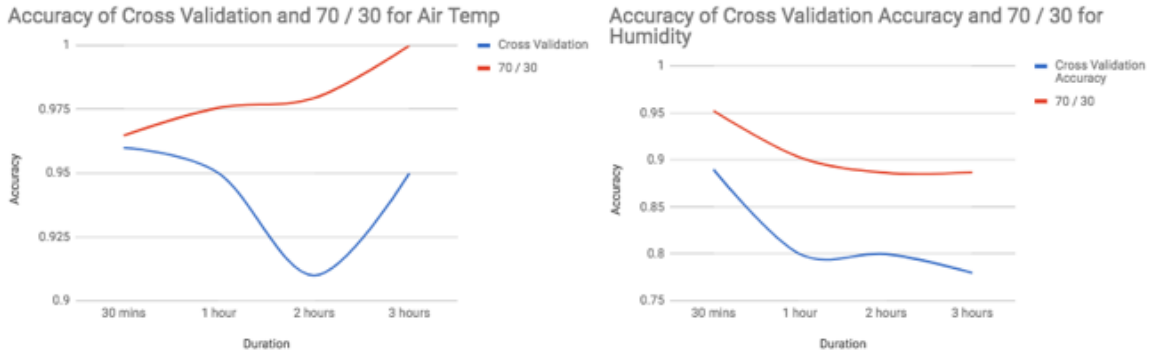


Figure 11: Cross Validation Results from January to April (Experiment 3)

Data from January to April in Figure 11 shows an unusual case were the 70/30 prediction accuracy outperformed the cross validation for both air temp and humidity. It then showed some inconsistent results for whether more data was beneficial for the prediction. One reason for this may be that there are less total blocks being used by the random forest for block of a greater size. 3 hour blocks have significantly less blocks of data than the number of 30 minute blocks.

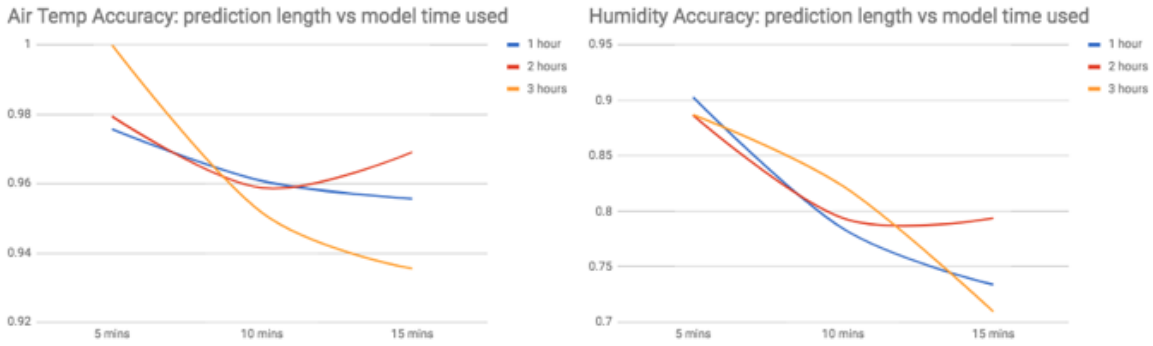


Figure 12: Accuracy Results for January to April (Experiment 3)

Similar results here show worse prediction the farther out time becomes and inconsistent results for the different block sizes. Overall, air temp was quite a bit more accurate at over 90% accuracy in its predictions than humidity which ranged from 70% to 90%.

This experiment introduced changing block and chunk sizes. This became the basis for future experiments where different block sizes and chunk sizes were tested along with sliding window and different classifiers.

5.4 Experiment 4

Methodology

This dataset used all data provided by Freight Farms breaking it up into three sections from January to April, March to June, and May to August. Chunk size was C = 5 and block size varied so that B = [1 hour = 12, 2 hours = 24, 3 hours = 36, 30 mins = 6]. The predicted chunk varied so that P = (1, 2, 3). These are the same parameters as the previous experiment; the only change was on the approach, instead of starting the next block of formatted data at the end of the last block the blocks overlap. Two blocks that are next to each other in the formatted data will differ by only two chunks, the first and last chunk. This sliding window method makes the number of blocks for each of the given block size B equal. As in previous iterations, the first 70% of the data was used for training and the next 30% was used for testing. More consistent results were expected with the same parameters given the methodology change.

Results

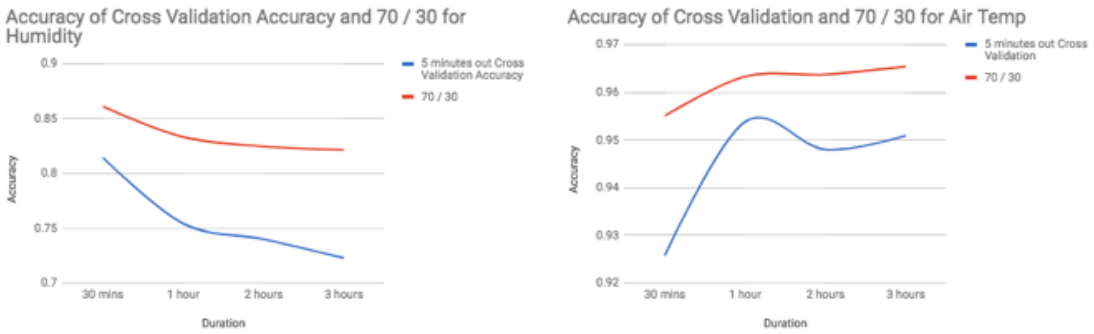


Figure 13: Cross Validation Results January to April (Experiment 4)

The data in Figure 13 remained abnormal for 70/30 data being more accurate than cross validation. However, the more consistent trend of air temp getting better prediction accuracy with more data and humidity getting slightly worse is more easily visible. The best tree is still being chosen for each point on the figure.

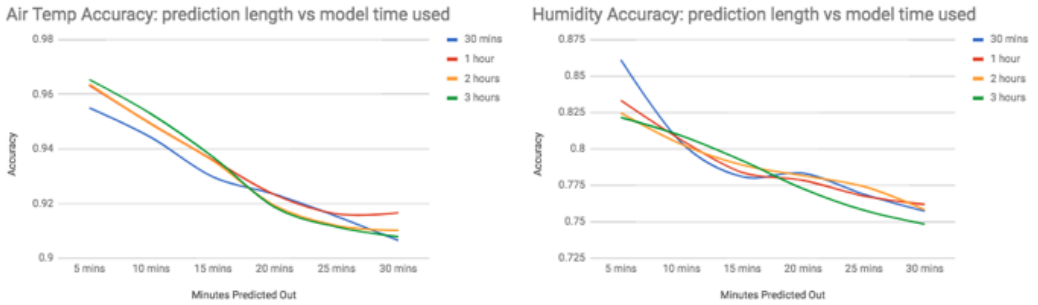


Figure 14: Accuracy Results from January to April (Experiment 4)

Figure 14 shows the block size not be the major factor for either air temp or humidity. The downward trend in prediction accuracy the further out predicted is expected. There is however little difference between block size of 3 hours and a block size of 30 minutes at only a couple percent.

From this experiment on, we switched our focus to fine tuning our parameters and testing different classifiers.

5.5 Experiment 5

Methodology

This dataset used all data provided by Freight Farms, breaking it up into three sections from January to April, March to June, and May to August. Chunk size was C = (5, 10, 15) and block size varied so that B = [1 hour = 12, 2 hours = 24, 3 hours = 36, 30 mins = 6] when C = 5, B = [1 hour = 6, 2 hours = 12, 3 hours = 18, 30 mins = 3] when C = 10, and B = [1 hour = 4, 2 hours = 8, 3 hours = 12, 30 mins = 2] when C = 15. The predicted chunk varied so that P = (1, 2, 3). The number of chunks per block was also reduced respective to chunk size to keep the overall time for the classifier to learn on equal for all chunk sizes at: 30 minutes; 1 hour; 2 hours; 3 hours. We hoped that the additional fine tuning would prove the model to become more accurate.

Results

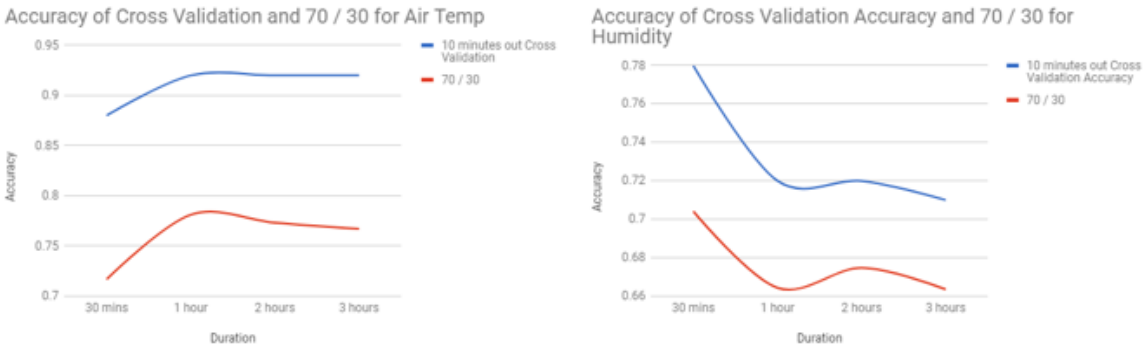


Figure 15: Cross Validation Results for January to April, Chunk Size 10 (Experiment 5)

In Figure 15, results for chunk size 10 are shown Accuracy of Cross Validation and 70 / 30 graphs show how accurate each time interval was at predicting the next chunks label, predicting the range of the next 10 minutes for chunk size 10. The graphs include both the ten-

fold cross validation results as well as test data results where the classifier was trained on the first 70% of the data and tested on the last 30%. The number of trees resulting in the highest accuracy for each block was used. Cross validation results are better than the test data results, but remain mostly consistent with air temp accuracy improving with a larger training set and humidity accuracy declining with a larger training set.

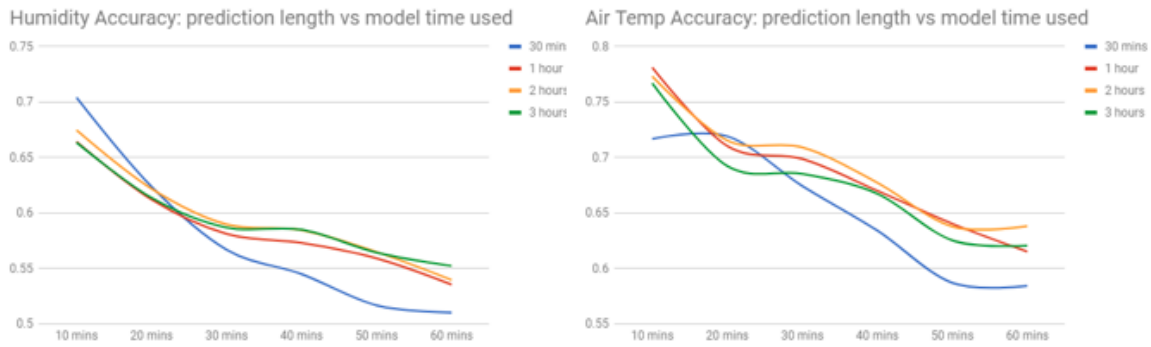


Figure 16:Accuracy Results for January to April, Chunk Size 10 (Experiment 5)

In Figure 16, chunk size 10 was also used. Accuracy of prediction length vs the model time graphs show the best result of a classifier predicting a certain time number of chunks out after having been trained for the given amount of time. As expected, the data shows prediction accuracy drops the further out the prediction. For air temp, generally having a larger training set, 3 hours vs 30 mins, resulted in higher accuracy. The opposite was true for humidity.

5.6 Experiment 6

Methodology

This dataset used all data provided by Freight Farms breaking it up into three sections from January to April, March to June, and May to August. Chunk size was C = (5, 10, 15) and block size varied so that B = [1 hour = 12, 2 hours = 24, 3 hours = 36, 30 mins = 6] when C =

5, B = [1 hour = 6, 2 hours = 12, 3 hours = 18, 30 mins = 3] when C = 10, and B = [1 hour = 4, 2 hours = 8, 3 hours = 12, 30 mins = 2] when C = 15. Same parameters as the previous experiment except for label ranges. The 5 predefined ranges for air temp and humidity were changed to 5 generated ranges based on the sensor readings for a given time duration. The ranges were calculated for each block size as well as for each time duration i.e. May to August. This decision was made because the outer ranges of the predefined ranges were virtually empty this led the classifier to mostly ignore them. The middle range also had the largest amount of data labels so it was previous classifiers may have been fairly accurate based on how the data was distributed. To solve this problem every data point for each sensor was sorted before the 5 ranges were made so they would hold about an equal number of sensor readings. This was done for each sensor and for each time duration. Some sensor ranges ended with only 1% to 2% between its starting and ending values, which gave presented problems for the random forest classifier. In addition, 8 other classifiers in addition to random forest to try to find a model that would be more accurate for the dataset. Classifiers used: C-Support Vector Classifier; quadratic classifier; passive aggressive classifier (type of linear classifier); neural network; Linear classifiers (SVM, logistic regression, i.e.) with SGD training; k-nearest neighbor; Gaussian Naive Bayes; decision tree classifier.

Results

Overall the accuracy results plummeted from the range changes. When the random forest classifier had a difficult problem to solve with even amounts of instances of each classifier accuracy could not be maintained near the other experiment levels. Random forest remains the best classifier. The two closest are k-nearest neighbor and gaussian naive bayes, with other

classifiers having inconsistent to erratic results. In their best-case scenarios those three were 60% to 70% accurate.

K-Nearest Neighbor

Figure 17, 18, and 19 from Iteration 6, K-Nearest Neighbor Classifier for chunk sizes 5, 10, and 15 showing prediction accuracy for air temp and humidity. Dataset used is from January to August and predictions are 1 to 6 chunks out and block size from 30 minutes to 3 hours were used.

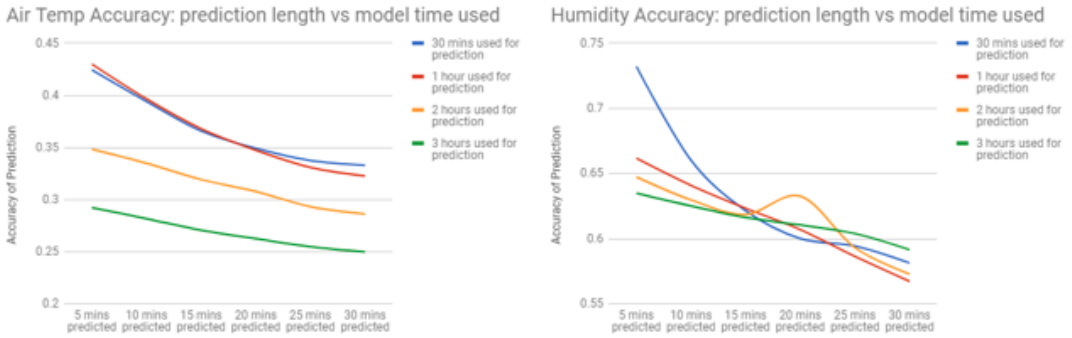


Figure 17: K-Nearest Model; Block Size 5 (Experiment 6)

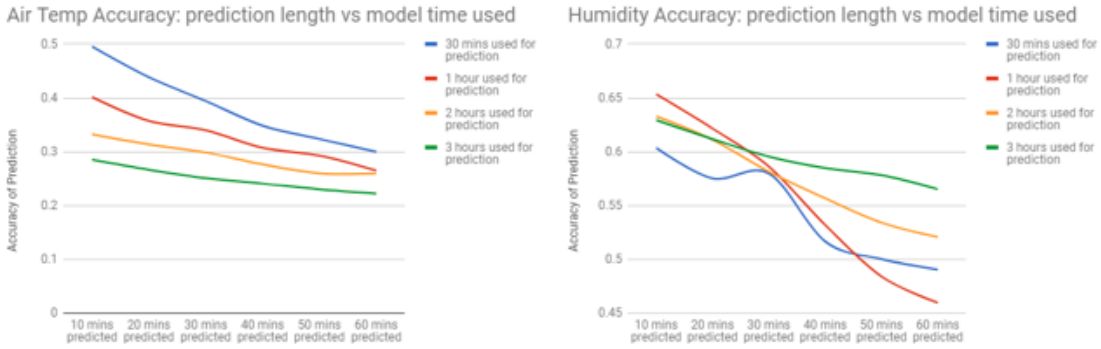


Figure 18 : K-Nearest Model; Block Size 10 (Experiment 6)

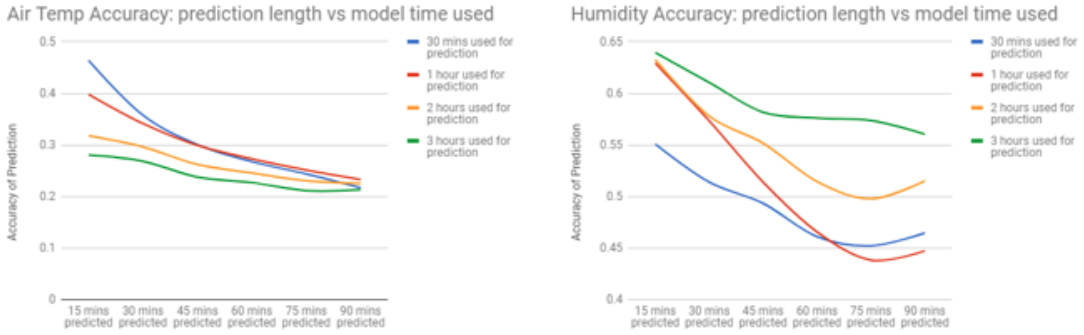


Figure 19: K-Nearest Model; Block Size 15 (Experiment 6)

Gaussian Naive Bayes

Figures 20, 21 and 22 show the accuracy results for the Gaussian Naive Bayes classifier for block size 5, 10 and 15 respectively. This was one of the few consistent classifiers as well produced decent accuracy. It was the best classifier for predicting humidity. Being one the better classifiers still did not get good enough prediction accuracy to realistically useful for predicting outside of 15 minutes for air temp where just above 60% accuracy was attained. Humidity faired a bit better with the best results getting above 70% accuracy predicting 15 minutes out. The figures show the expected downward trend of prediction accuracy the further out predicted. Interestingly having larger block sizes, more data, hurt the classifiers performance in most cases.

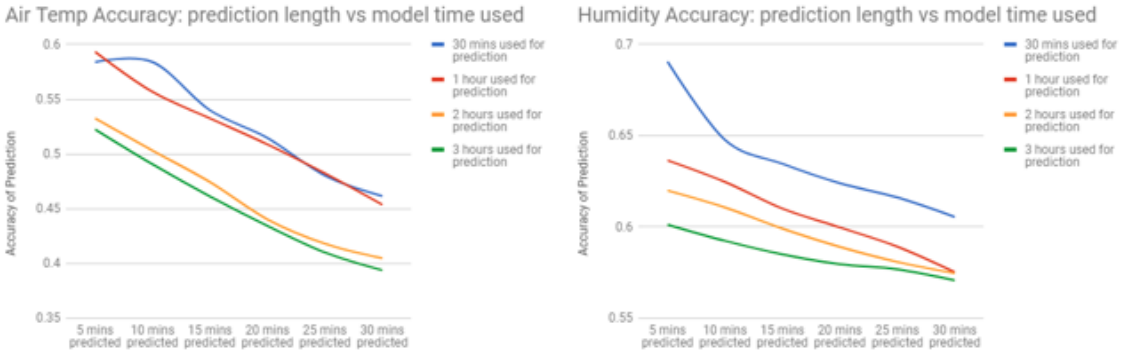


Figure 20: Gaussian Naive Bayes; Block Size 5 (Experiment 6)

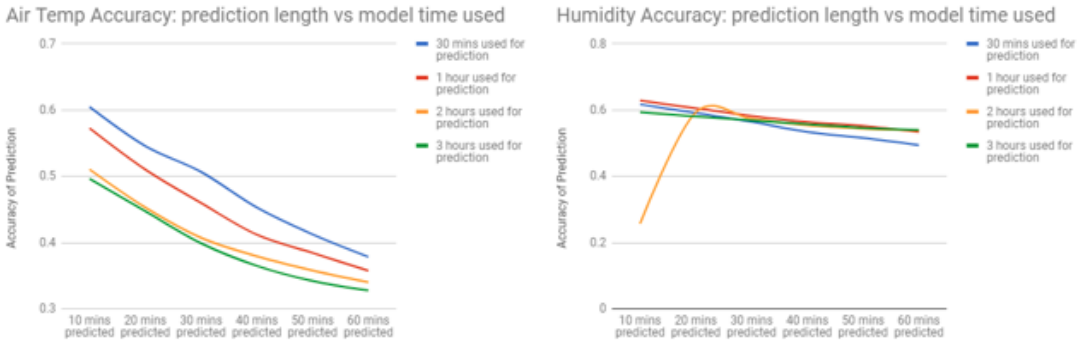


Figure 21: Gaussian Naive Bayes; Block Size 10 (Experiment 6)

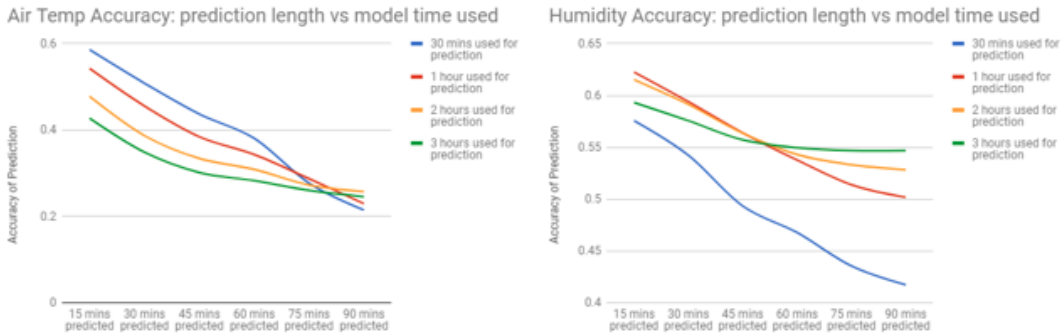


Figure 22: Gaussian Naive Bayes; Block Size 10 (Experiment 6)

Random Forest Classifier

Figures 23, 24 and 25 show the accuracy results for the Random Forest classifier for block size 5, 10 and 15 respectively. This was another one of the few consistent classifiers as well produced the best overall accuracy, and the best air temp prediction accuracy. Being one the better classifiers still did not get good enough prediction accuracy to realistically useful for predicting outside of 15 minutes for air temp where just above 70% accuracy was attained. Humidity faired a bit better with the best results getting above 60% accuracy predicting 15 minutes out. The figures show the expected downward trend of prediction accuracy the further

out predicted. Block size had little effect on performance showing inconsistent results as to whether more data help the classifiers accuracy.

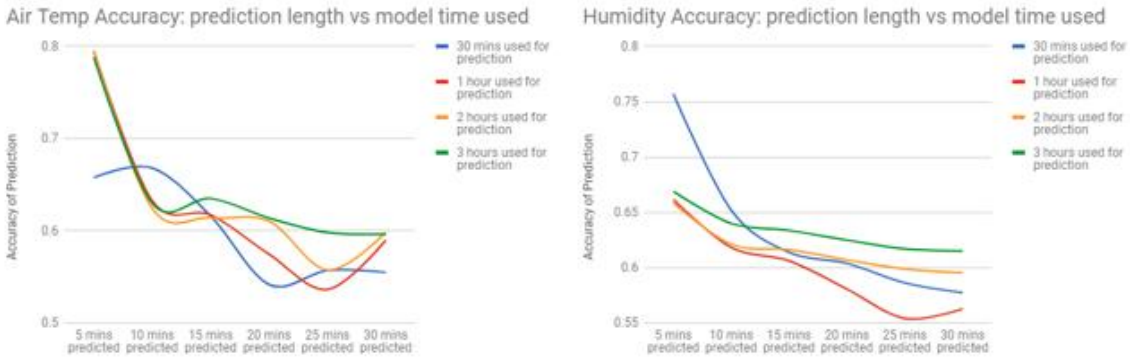


Figure 23: Random Forest; Block Size 5 (Experiment 6)

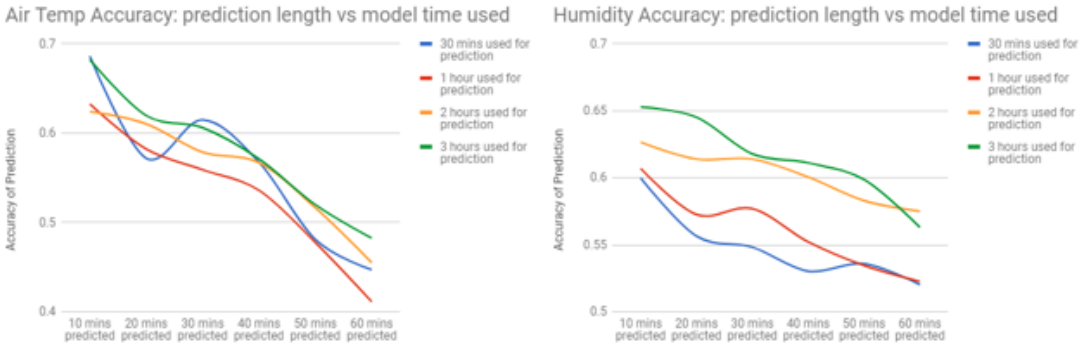


Figure 24:Random Forest; Block Size 10 (Experiment 6)

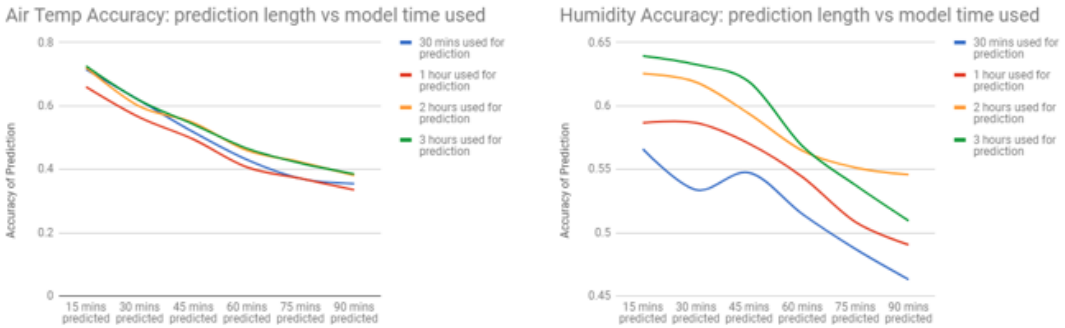


Figure 25:Random Forest; Block Size 15 (Experiment 6)

After changing the ranges for temperature and humidity to ensure we had equal data in each range, we realized that building a predictive model for the Leafy Green Machine's environment is a much more difficult task than initially expected.

Chapter 6: Future Work and Recommendations

The project focused on building a predictive model for a freight's environment. Further work can be done to improve the model and build upon it. To ensure that the model has enough data to train on, more data without any discrepancies, such as missing days and overlapping data, should be provided. This will allow for the model to improve its accuracy and be tested on data from multiple different environments. Different equipment and sensor combinations can be used to predict trends of other sensor reading such as pH, EC and nutrients and understand how each piece of equipment affects each other.

Another possible course that can be taken is watching how the sensor data behaves and instead of predicting the environment, the model can detect anomalies. A model can also be built to see how the outside environment affects the freight's environment. Though Freight Farms says its environment stays normalized throughout the year, the climate can change based on the farmer walking in and out of the farm and letting heat escape or get into the farm. Many different approaches can be taken to build upon the data-driven framework we built in this project.

Chapter 7: Conclusions

The goal of this project was to design a data-driven framework that could accurately predict relevant conditions inside of a freight farm. Once accurate predictions were made, they could be used to alert the farmer when the conditions are predicting to be outside of an acceptable range. Two crucial conditions to plant growth were experimented with, air temperature and humidity. Three key equipment to controlling those condition were chosen, lights, coolbot, and main pump. To be considered effective, prediction results would need to be at least 80% accurate at the given prediction range.

Two main methods were attempted. The first used ranges set by air temperature and humidity parameters from real world crops to create five ranges that indicated if the condition of the freight farm was at an acceptable level for the crops. The best results for that method came from Experiment 4, which indicated that prediction up to 10 minutes out could be done with 80% accuracy for both air temperature and humidity with the random forest classifier. While that seemed like a success, the lack of data in all ranges meant the model was learning to solve a trivial problem. The second method used ranges that were generated so that each range would have an equal number instances. The results were poor for random forest and the 8 other classifiers used with none being able to predict the next 5 minute average with 80% accuracy. Random forest classifier was the closest with over 75% accuracy at predicting the next 5 minute average for both air temperature and humidity. Prediction accuracy needs to be improved for predictions further into the future for this to be useful for freight farmers.

New learning models could be used as well as incorporating more equipment states and adding derived features to these models could improve accuracy.

References

[1] D. R. Hershey, "Digging Deeper into Helmont's Famous Willow Tree Experiment," *The American Biology Teacher,* vol. 53, no. 8, pp. 458-460, 1991.

[2] J. S. Douglas, Hydroponics, Bombay: Oxford UP, 1975.

[3] "History of Hydroponics," [Online]. Available: http://hydroponicgardening.com/history-of-hydroponics/water-culture-hydroponics-history/. [Accessed 10 November 2017].

[4] "Grow Cycle," 2016. [Online]. Available: https://www.freightfarms.com/. [Accessed 15 July 2017].

[5] W. P. Warren S. McCulloch, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics,* vol. 5, no. 5, pp. 115-133, 1943.

[6] O. D. J. K. S. S.-S. Y. S. Koby Crammer, "Online Passive-Aggressive Algorithms," *Journal of Machine Learning Research,* vol. 7, pp. 551-585, 2006.

[7] R. S. Z. R. Ridella S, "Circular backpropagation networks for classification," *IEEE Transactions on Neural Networks,* vol. 8, no. 1, pp. 84-97, 1997.

[8] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics,* vol. 7, pp. 111-132, 1936.

[9] T. S. Ferguson, ""An inconsistent maximum likelihood estimate," *Journal of the American Statistical Association,* vol. 77, no. 380, pp. 831-834, 1982.

[10] M. Pelillo, "Alhazen and the nearest neighbor rule," *Pattern Recognition Letters,* vol. 38, no. 1, pp. 34-37, 2014.

[11] P. L. George H. John, "Estimating Continous Distributions in Bayesian Classifiers," in *Proc. Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.

[12] J. R. Quinlan, "Induction of Decision Trees," *Machin Learning,* vol. 1, pp. 81-106, 1986.

[13] T. K. Ho, "Random Decision Forests," in *Proc. of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 1995.

[14] P. d. M. O. J. B. C. J.P. Coelho, "Greenhouse air temperature predictive control using the particle swarm optimisation algorithm," *Computers and Electronics in Agriculture,* vol. 49, no. 3, pp. 330-344, 2005.

[15] "SE-Structures and Environment," *Journal of Agricultural Engineering Research,* vol. 78, no. 4, pp. 407-413, 2002.

[16] D. J. J. S. Z. F. a. N. F. Q. Liu, "A WSN-based prediction model of microclimate in a greenhouse using extreme learning approaches," in *18th International Conference on Advanced Communication Technology* , Pyeongchang, 2016.

[17] E. F. A. R. P.M. Ferreira, "Neural network models in greenhouse air temperature prediction, In Neurocomputing," *Neurocomputing,* vol. 43, no. 1-4, pp. 51-55, 2002.

[18] A. E. R. a. C. M. F. P. M. Ferreira, "Genetic assisted selection of RBF model structures for greenhouse inside air temperature prediction," in *Proceedings of 2003 IEEE Conference on Control Applications*, Istanbul, 2003.

[19] E. C. E. C. M. L. A.E. Ruano, "Prediction of building's temperature using neural networks models," *Energy and Buildings,* vol. 38, no. 6, pp. 682-694, 2006.

[20] "Seed Station Guide," 2016. [Online]. Available: https://www.freightfarms.com/. [Accessed 15 July 2016].

[21] M. S. V. K. P.-N. Tan, Introduction to Data Mining, Pearson, 2005.

Appendix

A.1: Data Exploration

Introduction

Before we could begin building our model, we explored a preliminary dataset provided to us by Freight Farms. The dataset was used to understand how each variable, CO_2, air temperature, pH, electrical conductivity, and humidity, behaves in the farm environment. We conducted data exploration in hope of answering the following question:

How does the presence of humans affect the environment of the Freight Farm container?

Methodology

Initial Data

A set of preliminary data was provided in five JSON files. *Wpi-sensors-meta.json* and *wpi-sensors.json* gave information about the temperature, humidity, CO2 levels, pH, and electrical conductivity for different parts of the freight containers. *Wpi-images.json* contained urls of the images taken in the farm approximately every three minutes. *Wpi-equipment-meta.json* and *wpi-equipment.json* contained data from the equipment, which is used to control the environment of the container.

Data Format Conversion

To study the sensor data provided, the JSON files were converted into three CSV files. Python was chosen for tool development due to its portability and external library support. *Wpi-sensors-meta.json* and *wpi-sensors.json* were converted into *ff_sensor.csv,* using *sensor_data_to_csv.py,* which allowed for the data to be viewed in data analytics programs, such

as Excel. *Wpi-equipment-meta.json* and *wpi-equipment.json* were similarly converted to *equipment_data.csv,* using *equipment_data_to_csv.py. Wpi-images.json* was converted to *image_data.csv,* using *image_data_to_csv.py*. Once the data was in CSV format it was ready visualized for trends and behavior.

Data Visualization

Plotly is a python-graphing library, which is used to visualize different kinds of data. The tool creates graphs over a specified data set and allows for the customization of graph type, range, and multiple axes. In the scope of the project, the library is currently being used to visualize the sensor and labeled image data. Our hope was to expand this to equipment data once accurate equipment data was provided.

Two scripts were used for the visualization of sensor and labeled image data: *genPlots.py* and *genPlotsByDay.py*. These scripts automate the graph creation process by requiring two to three inputs: the number of data sets (either 1 or 2), the set(s) of data to graphed for example humidity1, airtemp1, etc. The first script, *genPlots.py,* plots the inputted sensors so they can be compared against one another. The second script, *genPlotsByDay.py*, has the same parameters as the first. In addition, it segments the data into days and graphs the labeled image data. In the future, it is likely these will be combined into one tool with an additional argument to specify if images and segments by day are desired.

Analyzing and Labeling Image Data

Two scripts, *downImages.py* and *lookup_image.py*, were used as tools to analyze the image data. *downImages.py* downloaded all the images so they could be viewed as a set, as opposed to individual images looked up by their urls. This allowed finding patterns and early

observations on the types of images of the farm; this will be used for classification in the future. The second script, *lookup_image.py*, allowed for image lookup. A date was provided as an argument and three images would be downloaded and saved. The images represent a 9 to 12 minutes time slice before the date provided in the command line argument. Utilizing these tools with graphing tools described above showed some correlation between sensor data the farm being in use.

Pedestrian Detection

To correlate the farm being used with sensor data, the image data needed to be labeled. Two tools were explored for the task. The first tool was Pedestrian Detection, a python library, which uses OpenCV Image Recognition to determine the presence of people in images. This tool did not work well with our images. The tool was recognizing random objects in the farm as people. The problems likely were lack of face in many images and significant portions of the body not being in frame. An example of the tool attempting to recognize people in the environment can be seen below.

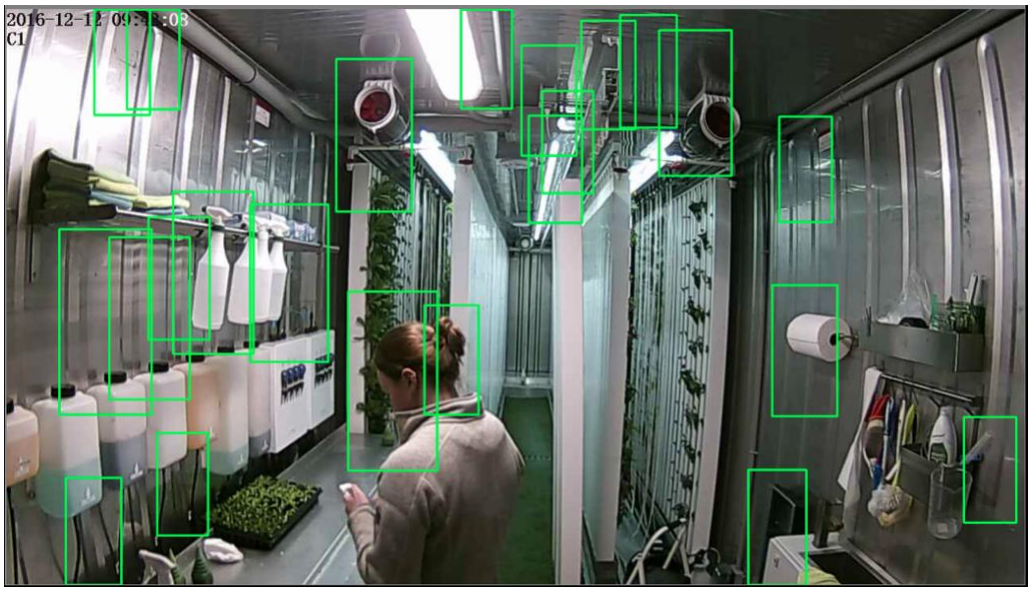


Figure 26: ClarifAI Pedestrian Detection

ClarifAI

The next tool used was ClarifAI, which assigns images tags base on how likely a certain feature is in the image. While this tool was equally as bad at determining if a person was in an image, it did tag similar looking images consistently. The shift went from finding people in the images to determining whether or not the farm was in use. This was done by checking if the main lights were on in the farm container. The initial *image_data_to_csv.py* was edited in order to implement labels to the images. The first tag was used to determine if the farm was in use. During testing, the first tag only produced 1% false positives. This allowed for using the method for labeling the image data. Eventually, ClarifAI will be further trained to improve image detection accuracy.

Three main image types were present in the data set: lights on, purple images, and monochrome images.

When the lights were on, it was found that people were present in the images of the environment; the tag "indoors" was used to detect this type of image.

Figure 27: Indoor Tag

When the LED Lighting System was on, it produced a purple hue in the images; it was found that no people were present in these images. The tag “no person” was used to detect this type of image.

Figure 28: No Person Tag

No people were present in the monochrome pictures; ironically, the tag “people” was used to detect this type of image.

Figure 29: People Tag

Findings

We visually studied the trends in the sensor data for the following variables: CO_2, air temperature, pH, electrical conductivity, and humidity. We focused on the trends of the readings in the main tank and disregarded the seedling data. The trends for each variable were studied first studied individually, meaning as a whole throughout the full week. Then they were studied how they were dependent on the presence of humans in the container. Finally, we studied the data by sectioning of the data into days. Our goal was to see how each variable contributed to answering our overarching question, “How does the presence of humans affect the environment of the freight farm?”

CO_2

CO_2 was the first variable studied by the group. We found that after humans had been found to be in the farm, the CO_2 levels in the environment would increase depending on the duration of the human's’ presence (this is excluding the data from days one and two) and then

begin to decrease after the humans would leave the farm. This led to us believing that CO_2 will be an important variable to study in future data sets in relation to human presence.

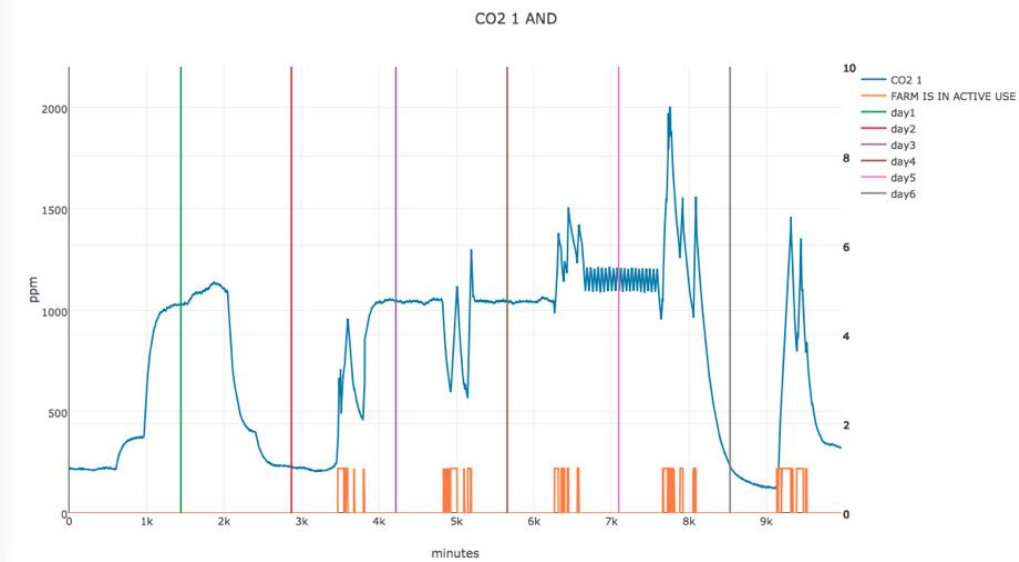


Figure 30: CO2 Levels

Air Temperature and Humidity

We then studied air temperature and humidity inside the freight farm. We noticed that the temperature would drastically drop throughout the week. Our first thought was that the door was being opened and the heat from the environment would escape. We then noticed that the temperature would drop around the same time every day when studied in conjunction to humidity. We think that the drops in temperature and humidity occur around the same time due to a combination of the fans turning on and the botanists coming into the environment to run tests on the vegetables being grown.

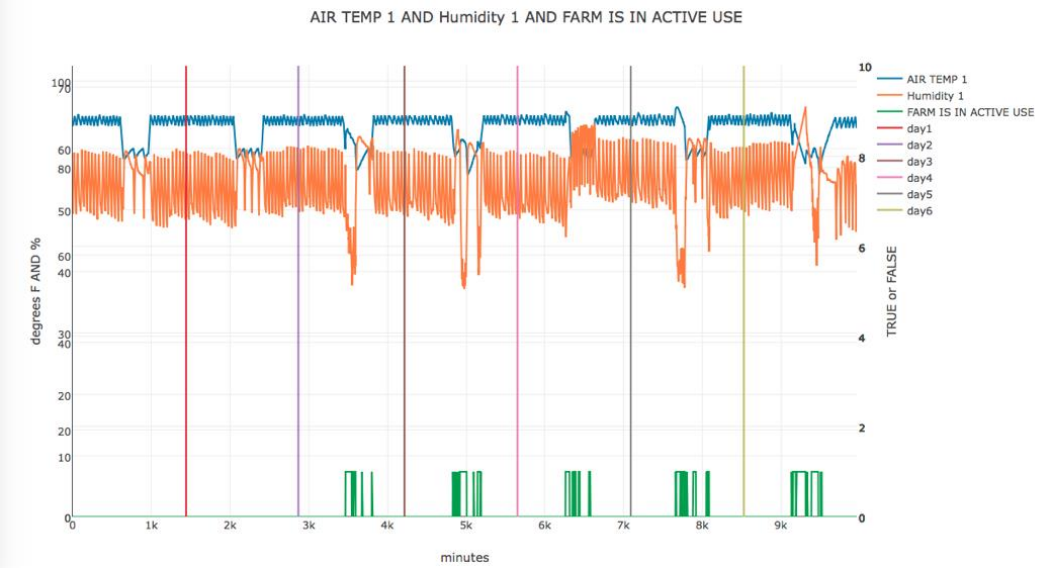


Figure 31: Air Temperature and Humidity

pH and Electrical Conductivity

We then studied pH and electrical conductivity (EC) in the freight farm environment. It was found that EC is dependent on the pH levels in the environment. The graph below shows that EC and pH somewhat reflect each other, meaning that as pH increases, EC decreases and vice versa. When the data was looked at in conjunction with human presence in the environment, we found that pH and EC had no correlation with people being in the farm.

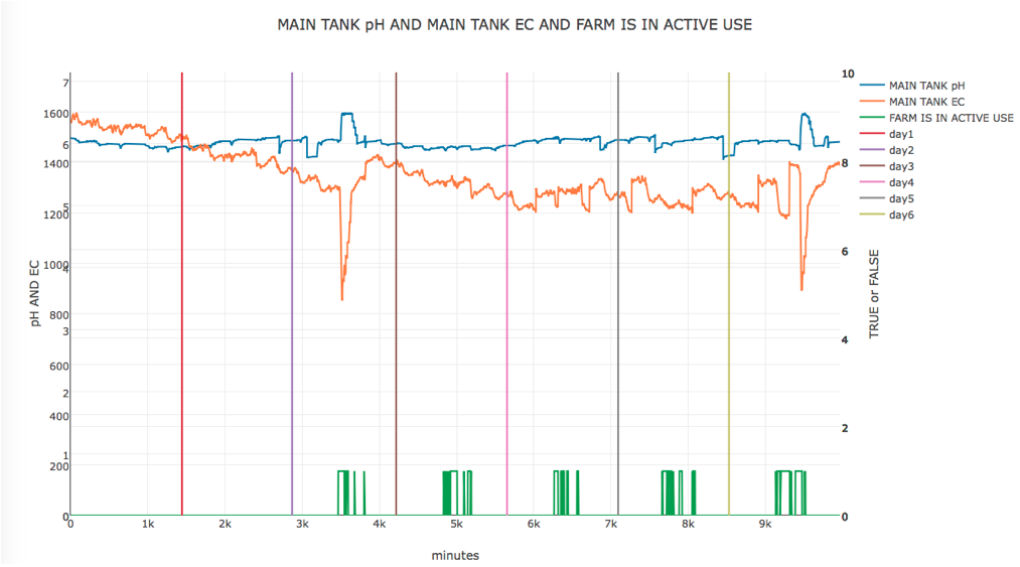


Figure 32: Main Tank pH and EC When Farm in in Active Use

CO_2 Levels Analysis

CO_2 level changes, in human presence, were further analyzed to understand the changes occurring in the environment. The mean, variance, and percent difference between the maximum and minimum level were calculated of the CO_2 levels for the segments of time the farm was active. The values of time segments of the same length were averaged. This allowed for there to be one entry for each length of time. The graphs were generated using google sheets. Thus far, it seems that the percent difference between the minimum and maximum CO_2 levels can prove to be useful in the future.

One problem that could have been run into in the future, was not accounting for the number of people in the container during certain time segments. A higher number of people in the environment can lead to a greater CO_2 level.

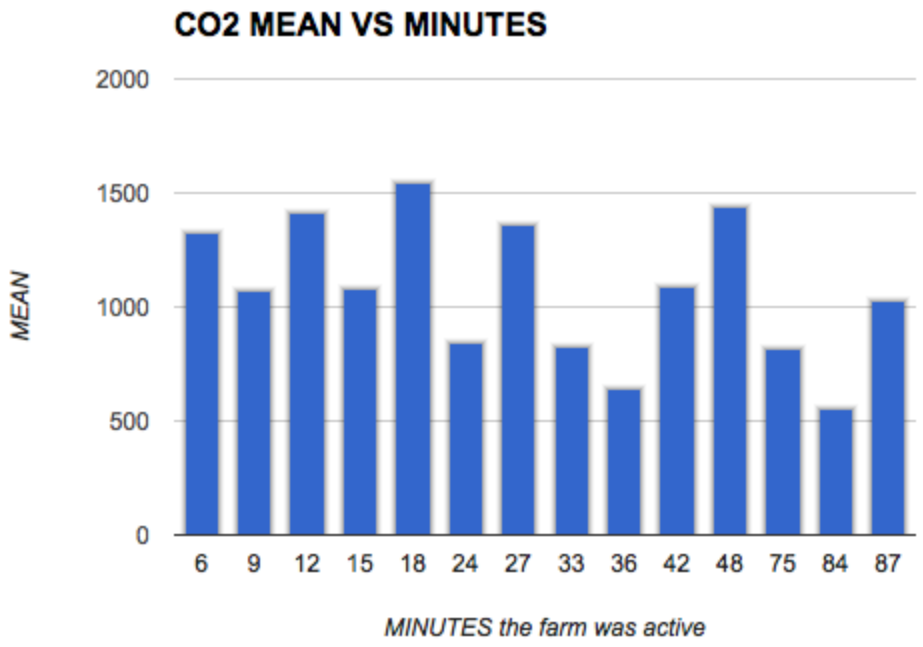


Figure 33: CO2 Levels vs Minutes

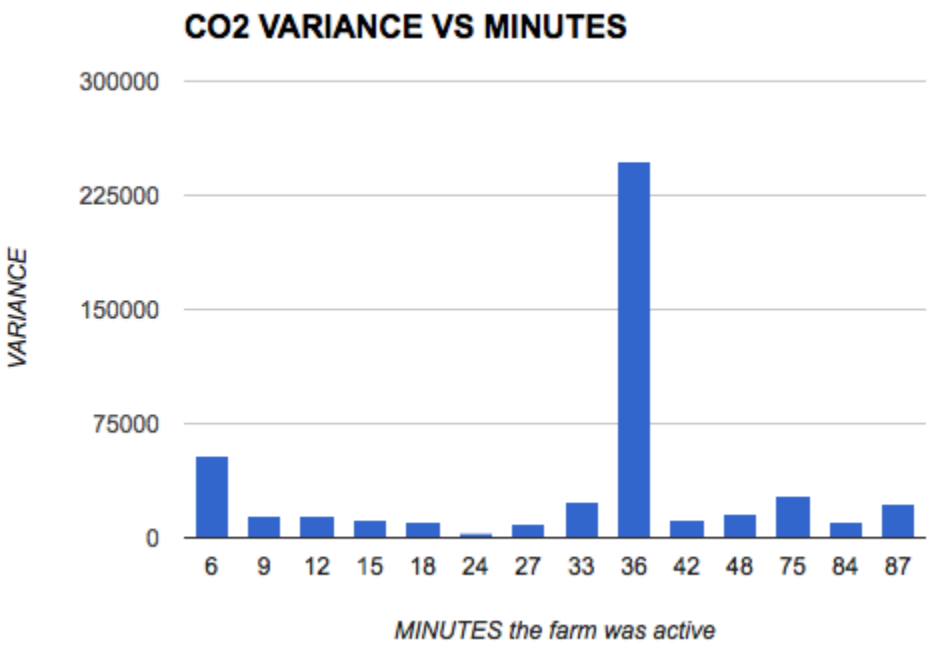


Figure 34: CO2 Variance vs Minutes

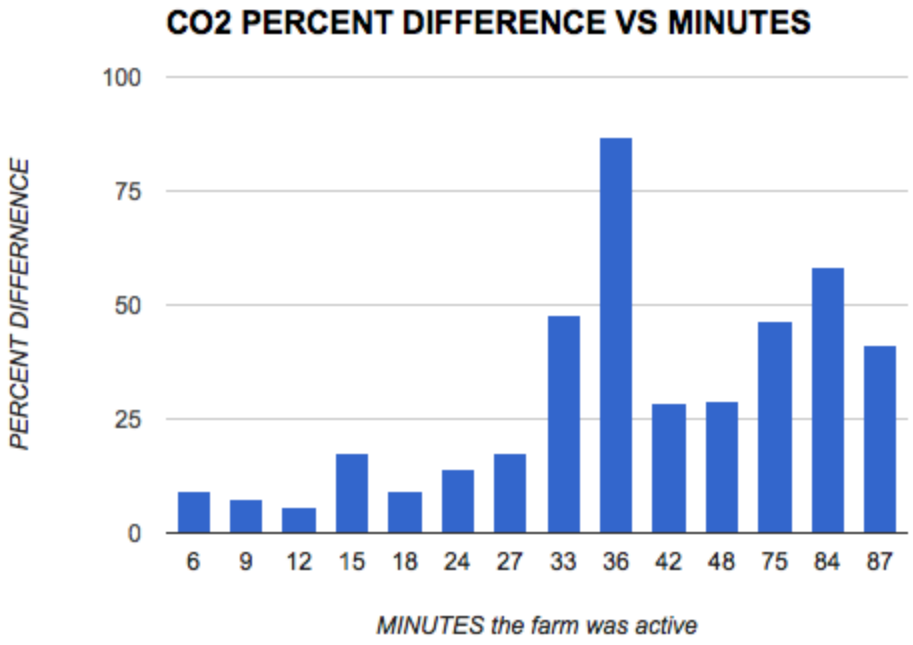


Figure 35: CO2 Percent Difference vs Minutes

Conclusions

After studying each of the variables, we concluded that CO_2 and humidity were the variables that could be used to indicate that a person is in the farm environment. When the CO_2 levels increased, the humidity in the farm started jumping drastically between and out of its range of 60 to 80%. These jumps in CO_2 and humidity occurred whenever a person was present in the environment.

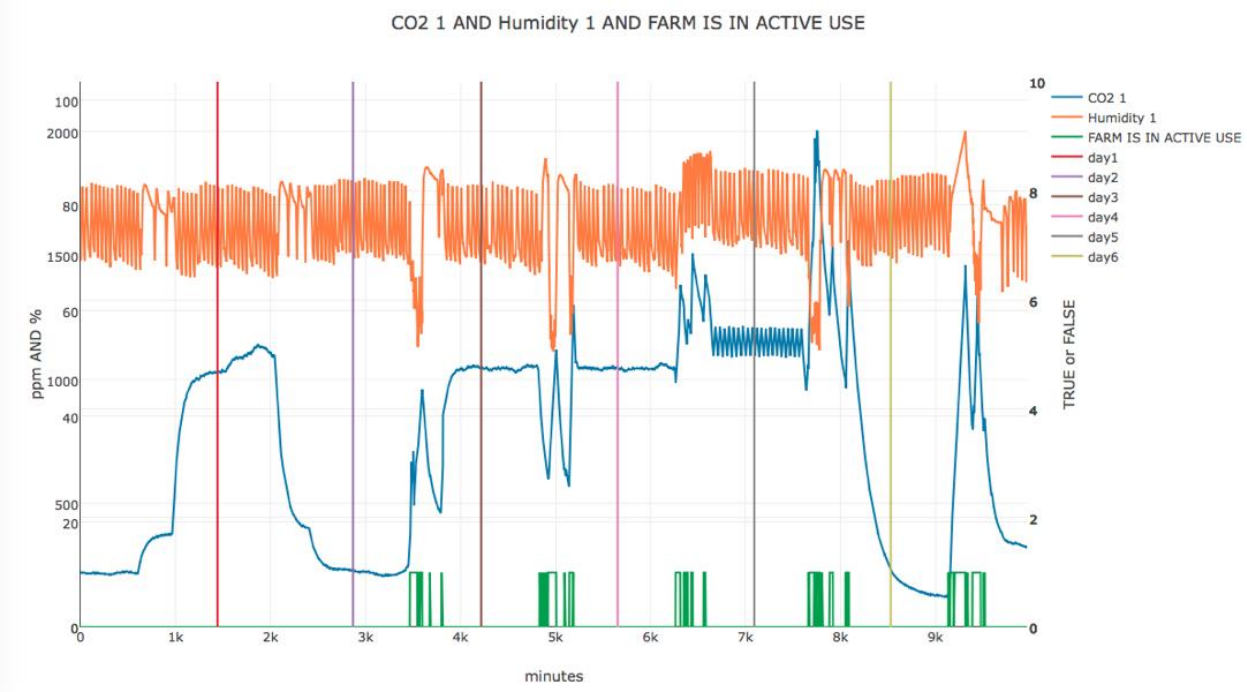


Figure 36: CO2 and Humidity Levels

Studying the preliminary data allowed us to explore the different trends each variable exhibited in the freight farm environment. This allowed for us to gain on understanding of how the farm works and what we can look for in future datasets to understand how human presence affects the environment.

A.2 Further Detail on Experiments

Experiment 3

March to June

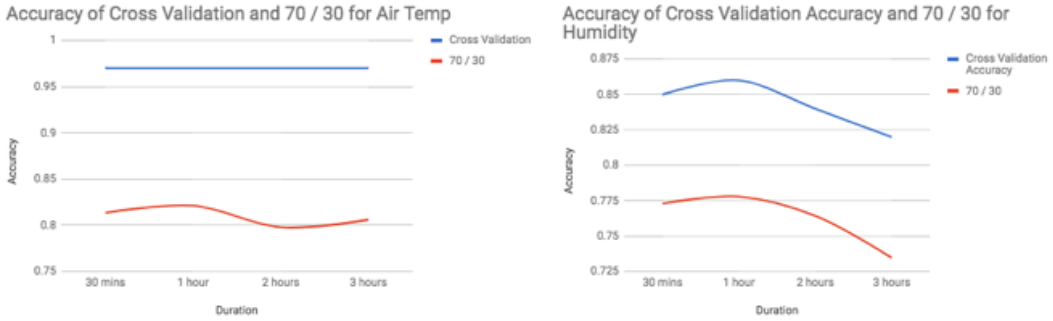


Figure 37: Cross Validation Results March to June

From Iteration 3, the accuracy of cross validation and 70/30 of air temp and humidity for prediction of the next 5 minutes using different block sizes on the March to June dataset.

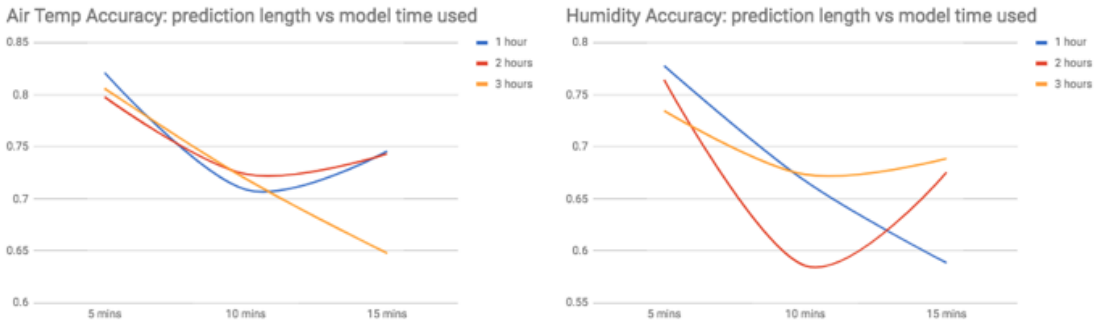


Figure 38: Accuracy Results for March to June

From Iteration 3, accuracy of 70/30 air temp and humidity prediction accuracy 5 to 15 minutes out using block sizes of 3 hours, 2 hours, and 1 hour. March to June dataset.

May to August

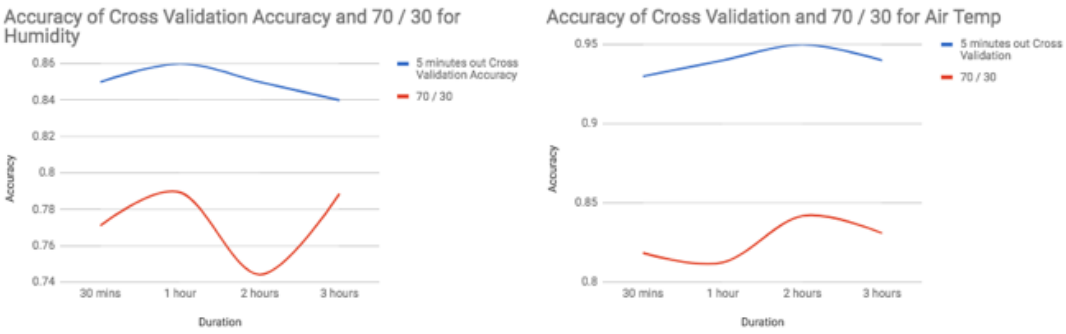


Figure 39: Cross Validation Results for May to August

From Iteration 3, accuracy of 70/30 air temp and humidity prediction accuracy 5 to 15 minutes out using block sizes of 3 hours, 2 hours, and 1 hour. May to August dataset.

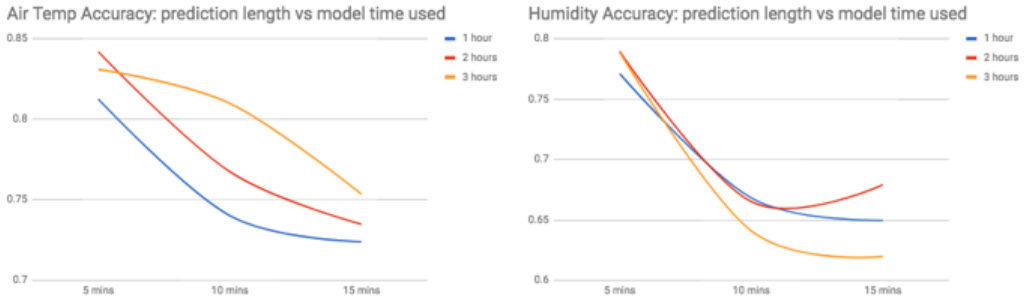


Figure 40: Accuracy Results for May to August

From Iteration 3, the accuracy of cross validation and 70/30 of air temp and humidity for prediction of the next 5 minutes using different block sizes on the May to August dataset.

Experiment 4

March to June

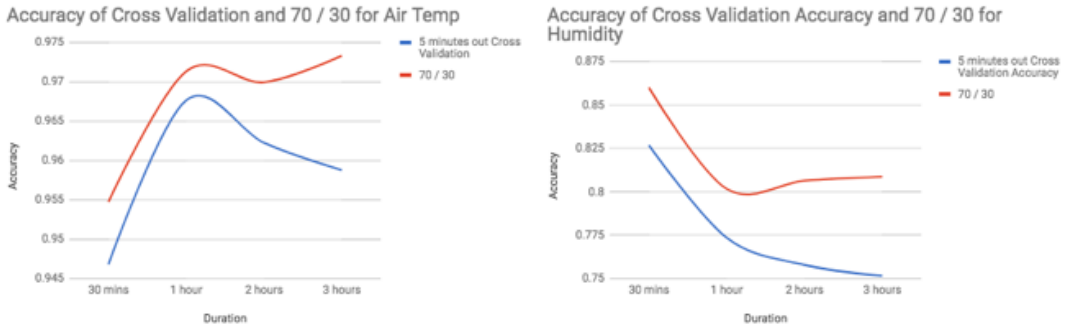


Figure 41: Cross Validation Results for March to June

From Iteration 4, the accuracy of cross validation and 70/30 of air temp and humidity for prediction of the next 5 minutes using different block sizes on the March to June dataset.

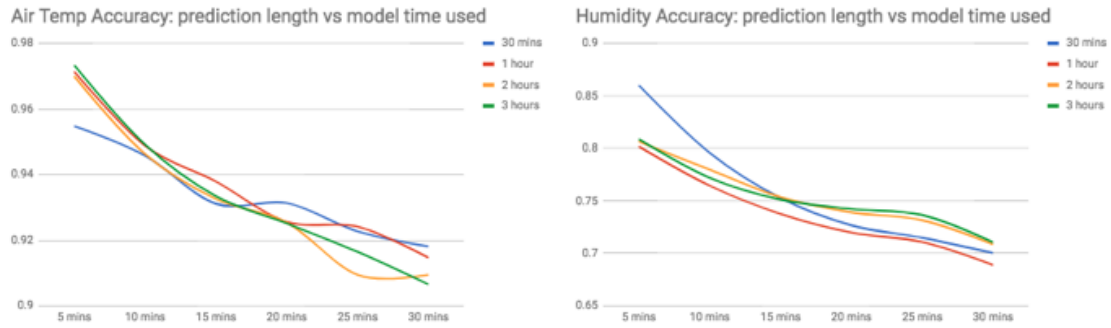


Figure 42: Accuracy Results for March to June

From Iteration 4, accuracy of 70/30 air temp and humidity prediction accuracy 5 to 15 minutes out using block sizes of 3 hours, 2 hours, 1 hour, and 30 minutes. March to June dataset.

May to August

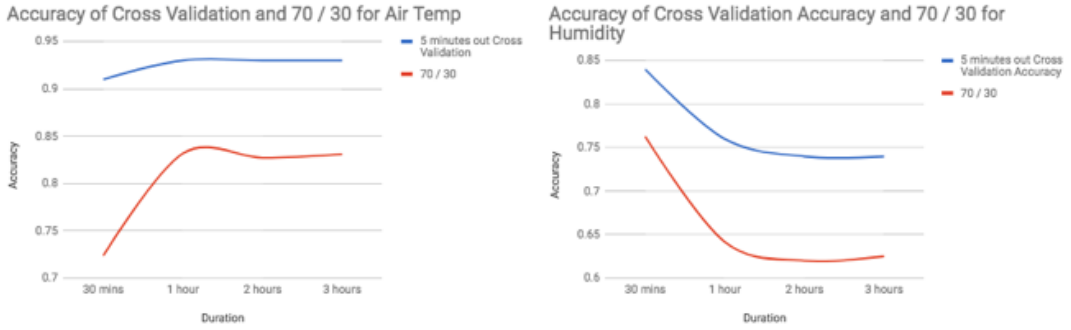


Figure 43: Cross Validation Results for May to August

From Iteration 4, the accuracy of cross validation and 70/30 of air temp and humidity for prediction of the next 5 minutes using different block sizes on the May to August dataset.

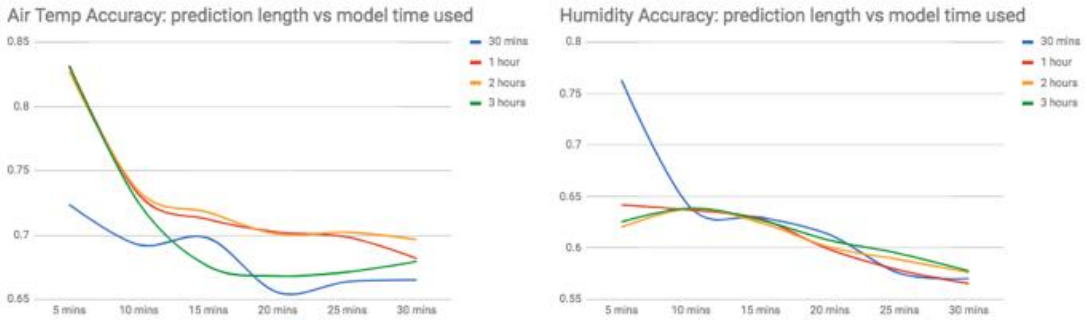


Figure 44: Accuracy Results for May to August

Experiment 6

C-Support Vector Classifier

Figure 45, 46 and 17 from Iteration 6, C-Support Vector Classifier for chunk sizes 5, 10, and 15 showing prediction accuracy for air temp and humidity. Dataset used is from January to August and predictions are 1 to 6 chunks out and block size from 30 minutes to 3 hours were used.

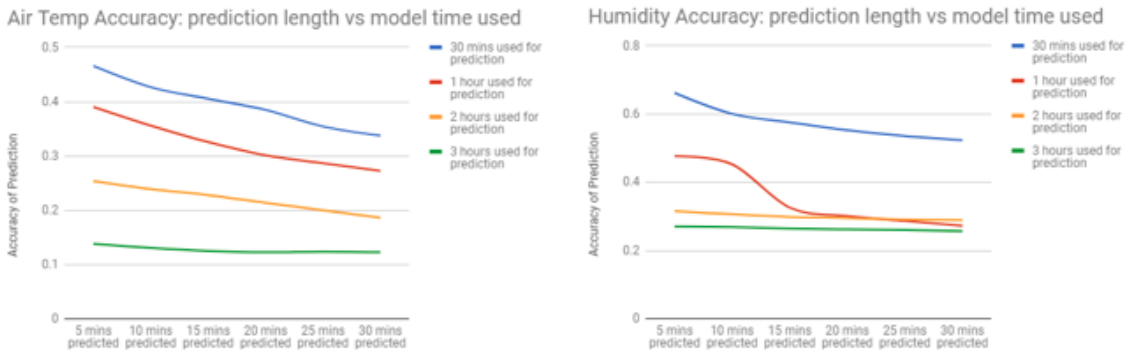


Figure 45: C-Support Vector Block Size 5 (Experiment 6)

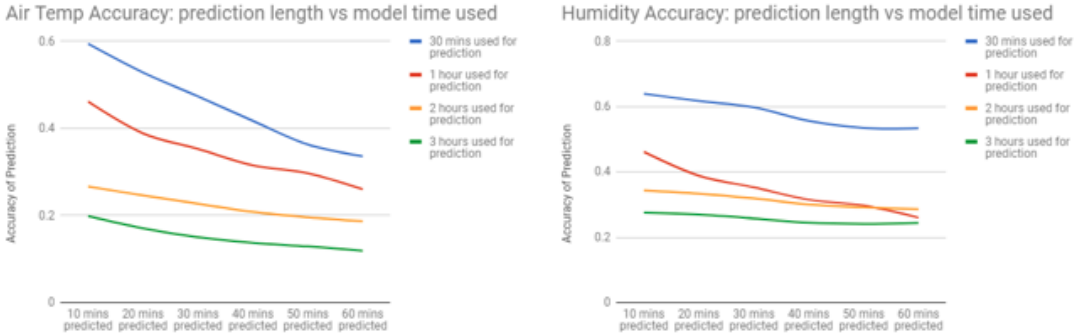


Figure 46: C-Support Vector, Block Size 10 (Experiment 6)

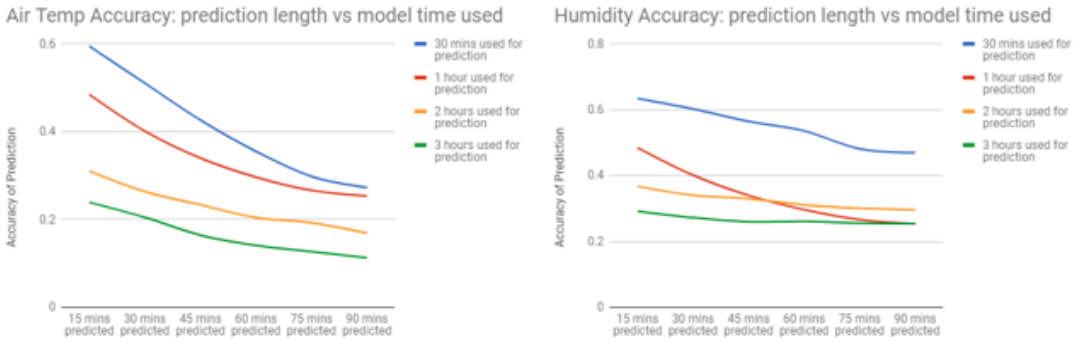


Figure 47: C-Support Vector, Block Size 15 (Experiment 6)

Quadratic Classifier

Figure 48, 49, and 50 from Iteration 6, Quadratic Classifier for chunk sizes 5, 10, and 15 showing prediction accuracy for air temp and humidity. Dataset used is from January to August and predictions are 1 to 6 chunks out and block size from 30 minutes to 3 hours were used.

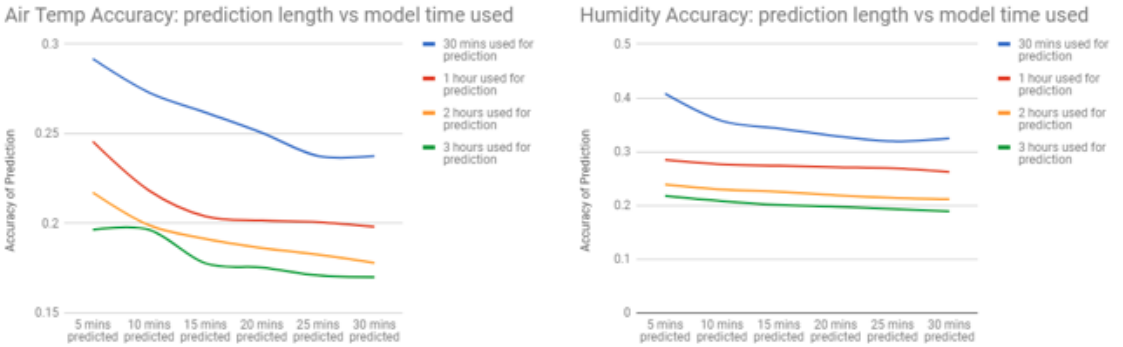


Figure 48:Quadratic Classifier, Block Size 5 (Experiment 6)

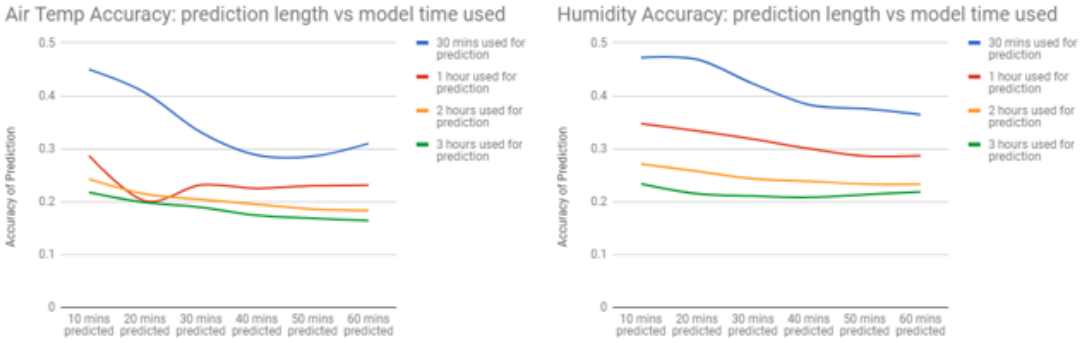


Figure 49:Quadratic Classifier, Block Size 10 (Experiment 6)

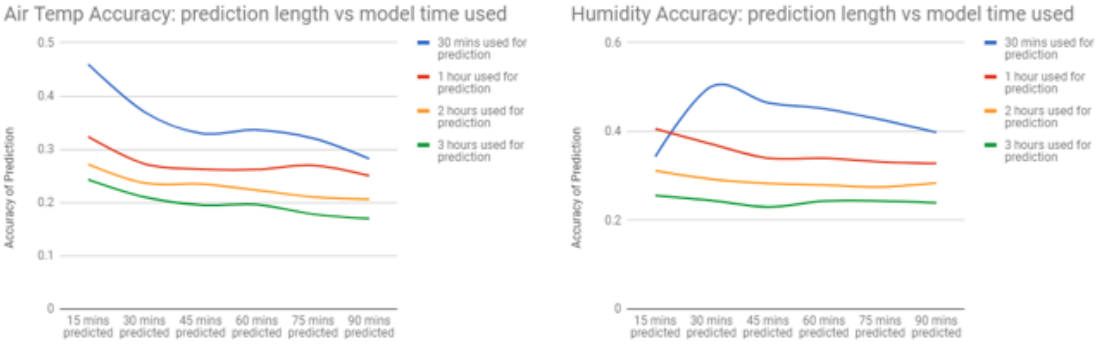


Figure 50: Quadratic Classifier, Block Size 15 (Experiment 6)

Passive Aggressive Classifier (type of linear classifier)

Figure 51, 52, and 53 from Iteration 6, Passive Aggressive Classifier for chunk sizes 5, 10, and 15 showing prediction accuracy for air temp and humidity. Dataset used is from January

to August and predictions are 1 to 6 chunks out and block size from 30 minutes to 3 hours were used.

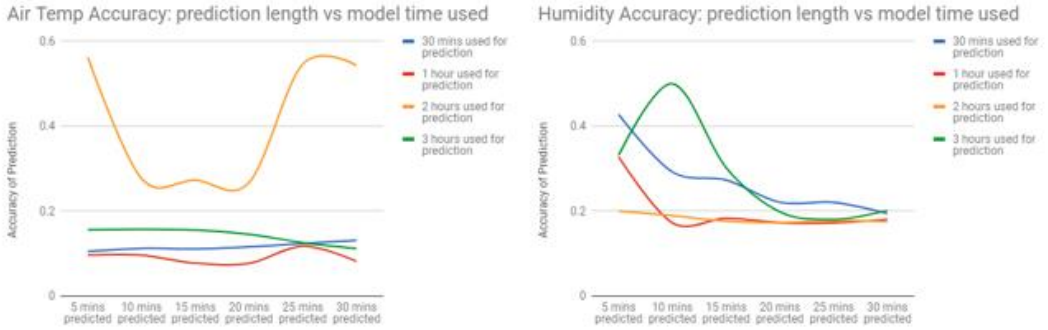


Figure 51:Passive Aggressive Classifier, Block Size 5 (Experiment 6

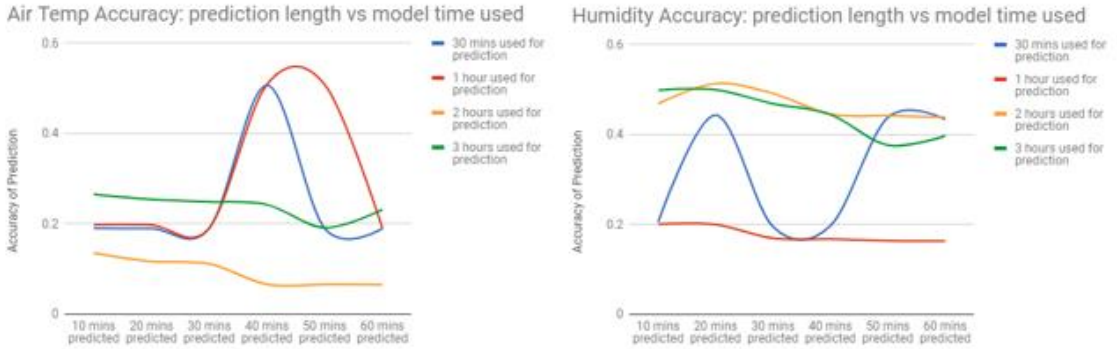


Figure 52: Passive Aggressive Classifier, Block Size 10 (Experiment

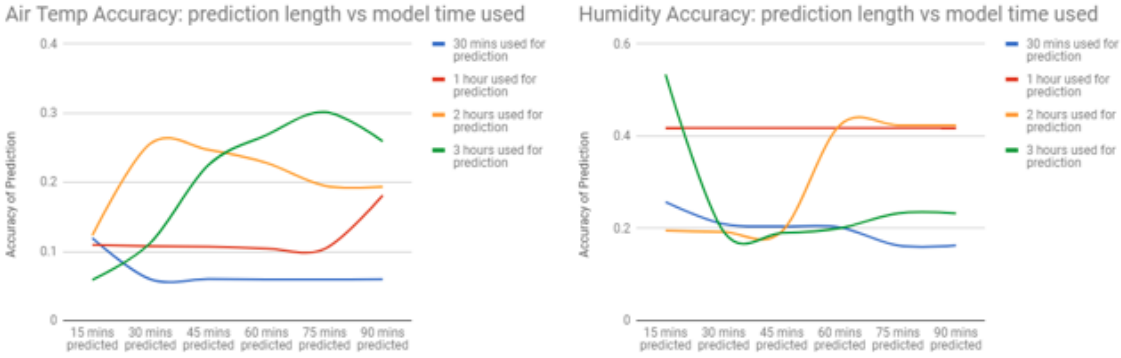


Figure 53: Passive Aggressive Classifier, Block Size 15 (Experiment 6)

Neural Network

Figure 54, 55, and 56 from Iteration 6, Neural Network Classifier for chunk sizes 5, 10, and 15 showing prediction accuracy for air temp and humidity. Dataset used is from January to August and predictions are 1 to 6 chunks out and block size from 30 minutes to 3 hours were used.

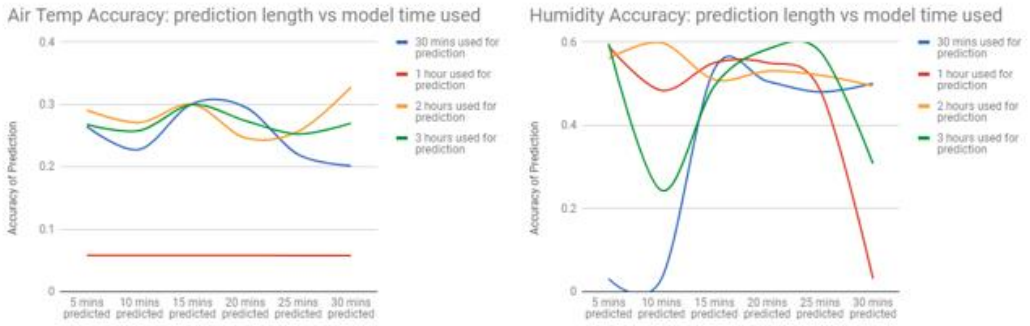


Figure 54:Neural Network, Block Size 5 (Experiment 6)

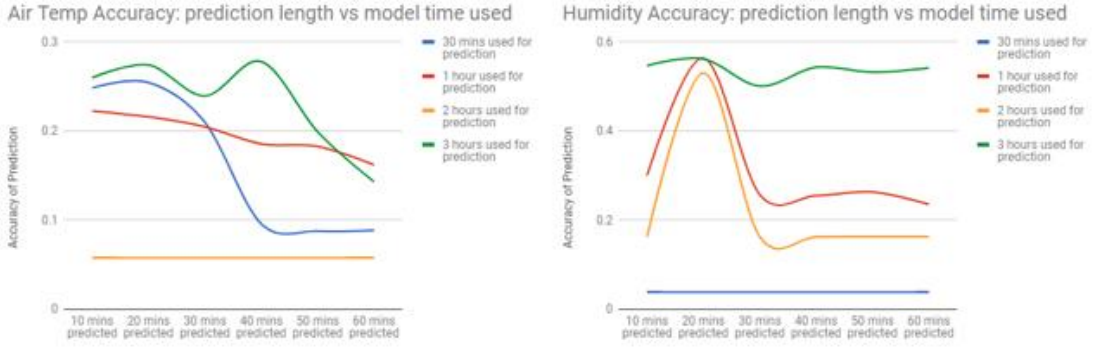


Figure 55: Neural Network, Block Size 10 (Experiment 6)

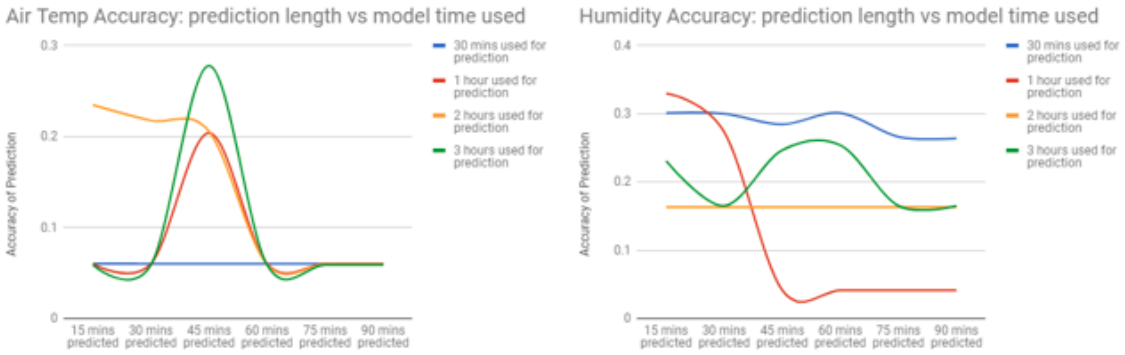


Figure 56: Neural Network, Block Size 15 (Experiment 6)

Linear Classifiers (SVM, logistic regression, i.e.) with SGD training

Figure 57, 58, and 59 from Iteration 6, Linear Classifiers (SVM, logistic regression, i.e.) with SGD training Classifier for chunk sizes 5, 10, and 15 showing prediction accuracy for air temp and humidity. Dataset used is from January to August and predictions are 1 to 6 chunks out and block size from 30 minutes to 3 hours were used.

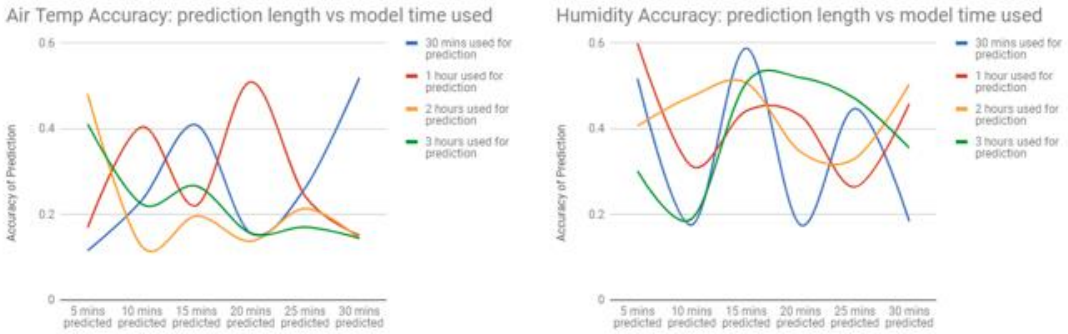


Figure 57: Linear Classifiers, Block Size 5 (Experiment 6)

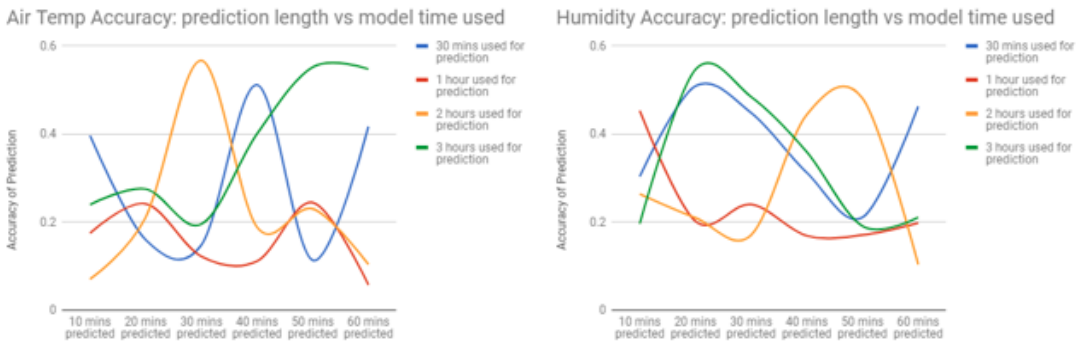


Figure 58: Linear Classifiers, Block Size 10 (Experiment 6)

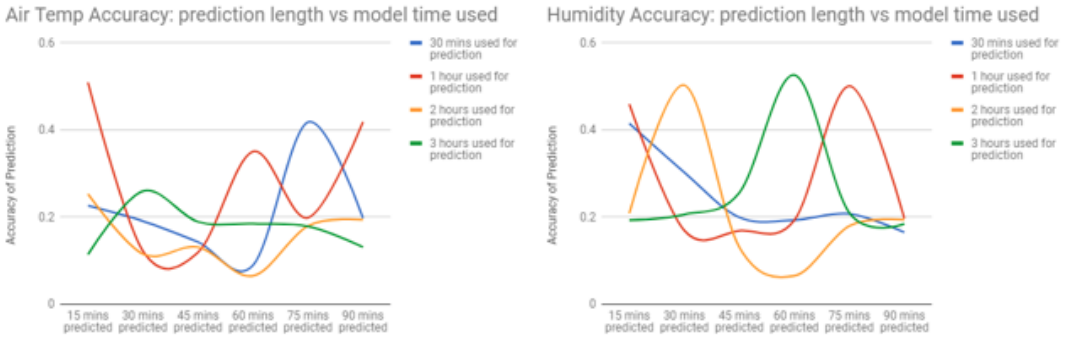


Figure 59: Linear Classifiers, Block Size 15 (Experiment 6)

Decision Tree Classifier

Figures 60, 61 and 62 show the accuracy results for the Random Forest classifier for block size 5, 10 and 15 respectively. This was another one of the few consistent classifiers as well produced the best overall accuracy, and the best air temp prediction accuracy. Being one the better classifiers still did not get good enough prediction accuracy to realistically useful for predicting outside of 15 minutes for air temp where just above 70% accuracy was attained. Humidity faired a bit better with the best results getting above 60% accuracy predicting 15 minutes out. The figures show the expected downward trend of prediction accuracy the further out predicted. Block size had little effect on performance showing inconsistent results as to whether more data help the classifiers' accuracy.

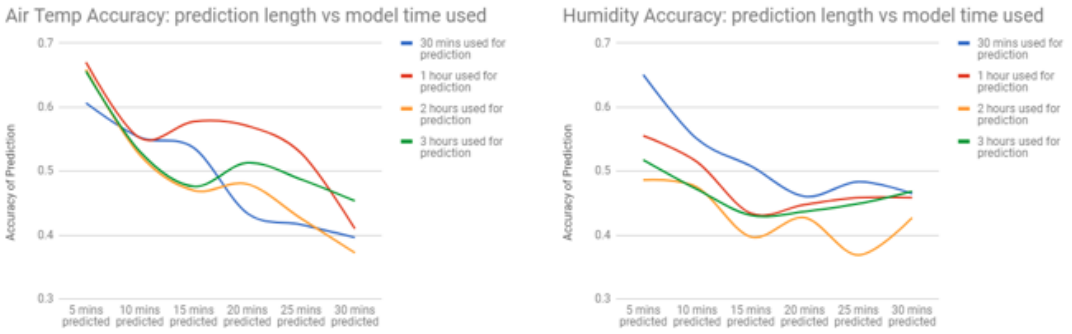


Figure 60: Decision Tree Classifier, Block Size 5 (Experiment 6)

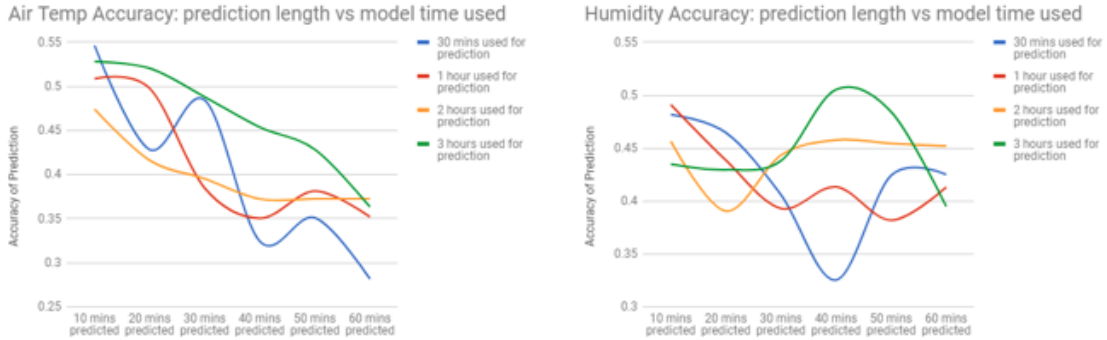


Figure 61: Decision Tree Classifier, Block Size 10 (Experiment 6)

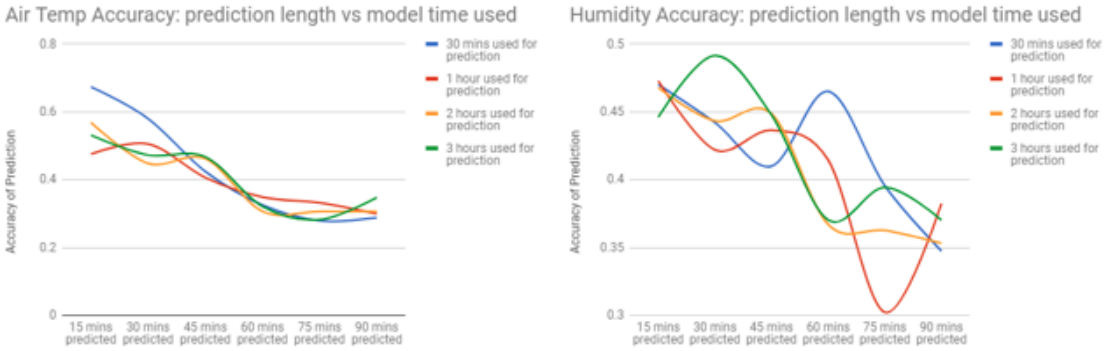


Figure 62: Decision Tree Classifier, Block Size 15 (Experiment 6)