Foodborne Disease DetectionUsing Machine Learning

Salma Yousaf and Shakeel Khan

September 19, 2020

Foodborne Disease DetectionUsing Machine Learning

Salma Yousaf , Shakeel Khan, June 2020, Department of Computer Scinece, Riphah International University

Abstarct

Foodborne disease outbreak, arises when number of people affected with the same type of illness from the use of the same type of infected food or drinks. Almost 48 million people in the US get ill with Foodborne diseases by using the infected food and drinks per year. Foodborne is not part of well known outbreak disease as outbreaks diseases provides the detailed information of the disease.

In this paper I a used the dataset of Foodborne disease outbreaks containing the data from year 1998 to 2015. It contains the different attributes of the data like year, month, state, location, food, illness etc. Rapid Miner tool of data science is used for the analysis of this dataset, which is one of the best visualizing tool. Three types of algorithms are applied on the dataset. Two types of clustering algorithms are also applied on this dataset.

Keywords--- Clustering, Generalized Linear Model, Random Forest, Gradient boosted tree.

Introduction

Rapid miner is one of the best tool which is used for data mining. Data mining techniques are used now days in many fields of technology.

Foodborne diseases are the cause of illness of many people. It arises when some or more people affected with the same type of illness with the use of infected a\or polluted drinks or foods.

This dataset provides data on foodborne disease outbreaks reported to CDC from 1998 through 2015. Data fields include year, state (outbreaks occurring in more than one state are listed as "multistate"), location where the food was prepared, reported food vehicle and contaminated ingredient, etiology (the pathogen, toxin, or chemical that caused the illnesses), status (whether the etiology was confirmed or suspected), total illnesses, hospitalizations, and fatalities.

In this research paper I used the tool of Rapid Miner for analyzing the dataset of Foodborne disease outbreak. Data mining techniques Generalized linear model, Random forest and gradient boosted trees algorithms are applied on this dataset. We analyze the prediction chart and simulation prediction and its impact factor prediction graph for all these three algorithms. Two types of clustering algorithms are also used. These clustering algorithms are k-Means clustering and x-Means clustering.

Problem Statement

To detect the foodborne disease outbreak from year 1998 to 2015 using machine learning techniques.

Literature Review

1. In this paper they show that the highly accurate MIC prediction models can be produced with less than 500 genomes. This is one of the largest MIC modeling studies to be published. Their strategy for developing whole-genome sequence-based models for surveillance and clinical diagnostics can be readily applied to other important human pathogens. [1]

2. More complex and globalised patterns of food production and distribution have resulted in outbreaks that are sometimes global in scale, such as the 2001 outbreak of Salmonella infection caused by peanuts imported into Australia and several other countries. On the other hand, large-scale commercial food processing may also decrease food contamination, as safety procedures are often stricter. [2]

3. “PulseNet USA” detects nearly all foodborne outbreaks of pathogenic bacteria. This is a bit odd because PulseNet has not only been very efficient in detecting foodborne disease but has thereby positively impacted public health and saved millions of dollars since it was founded 20 years ago PulseNet is now undergoing profound changes as it both expands internationally to protect consumers in other countries and invests heavily—financially and scientifically.

4. In this paper they take the reporting delays into consideration and apply a Bayesian hierarchical model for this forecast problem. The Bayesian hierarchical model was established to predict the daily true number of patients using the number of visiting patients. We propose several scoring rules to assess the performance of different now casting procedures.

5. This paper introducing a gravity-based approach to model food-flows from supermarkets to consumers and demonstrating how models of consumer shopping behavior can be used to improve computational methodologies to infer the source of an outbreak of foodborne disease. The value of considering shopping behavior in computational approaches for inferring the source of an outbreak is illustrated through an application example to identify a retail brand source of an outbreak.

Research Methodology

Predict

In Predict process I selected the column of illness in a task and then moved forward to prepare the target. Then it shows the visualization of the targeted column shown. Then in next step Input will be selected. Three types of models are applied on the dataset of Foodborne Disease outbreaks. Prediction charts, simulation prediction and important factor for prediction chart of these algorithms are applied. In clustering two types of clustering algorithms k-Means clustering and x-means clustering are applied. The other three algorithms are:

1. Generalized Linear Model
2. Random Forest
3. Gradient Boosted Trees

Generalized Linear Model Prediction /chart

Generalized linear model prediction chart is shown in the fig 1.1. It shows the true values for the illness below the red dotted line of chart and predicting values are above this line. It contain the more true values as compared to the predicting values.

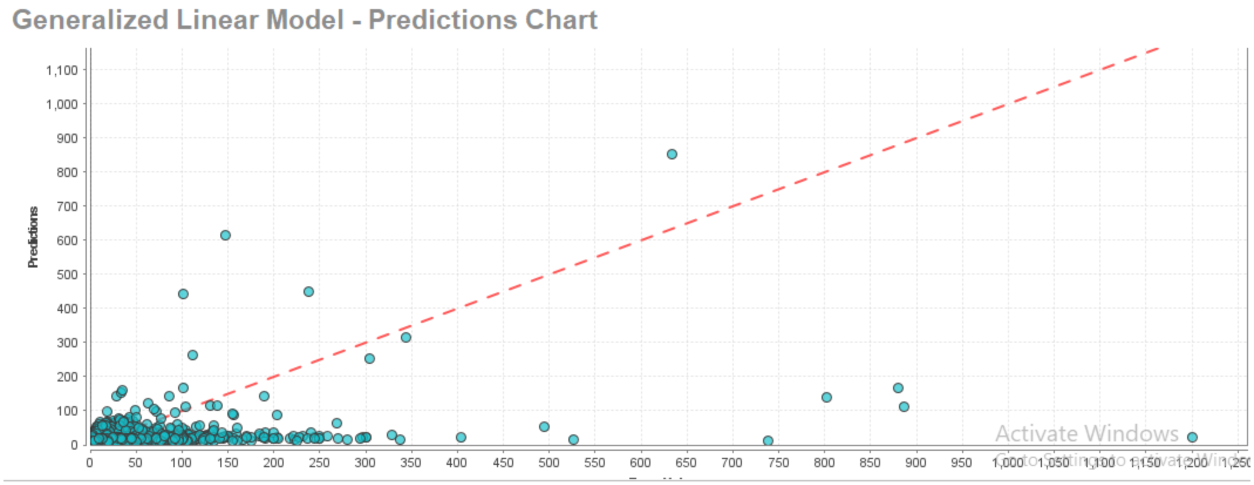
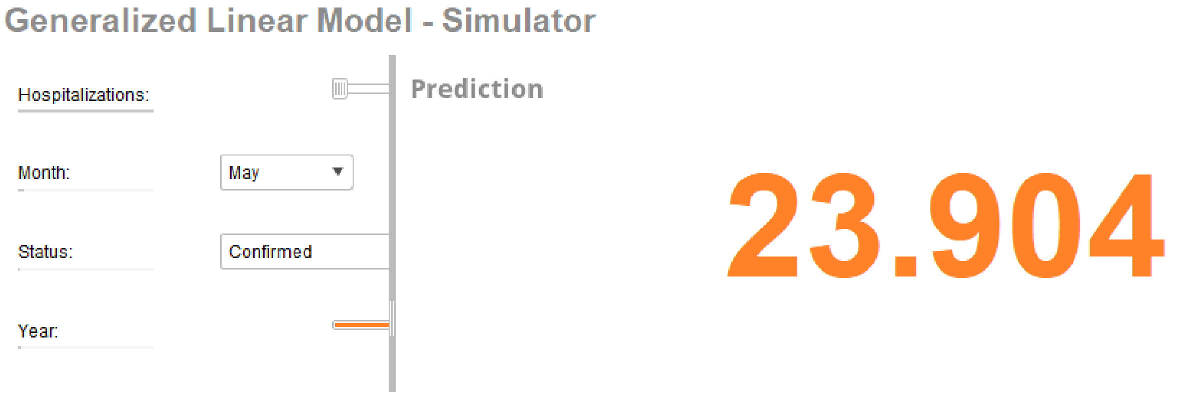


Figure 1.1 Generalized linear model predictions chart

Generalized Linear Model Simulation Prediction

Generalized linear model simulation prediction value is shown for the required dataset. Fig 1.2 shows the second largest predictive value.



Generalized Linear Model - Simulator

Hospitalizations:

Month: May

Status: Confirmed

Year:

Prediction

23.904

Figure 1.2 Generalized Linear Model Simulation Prediction value

Figure 1.3 shows the important factors for prediction for generalized linear model. Green color shows the support prediction and red color shows the contradict prediction. It means it has no contradict prediction.

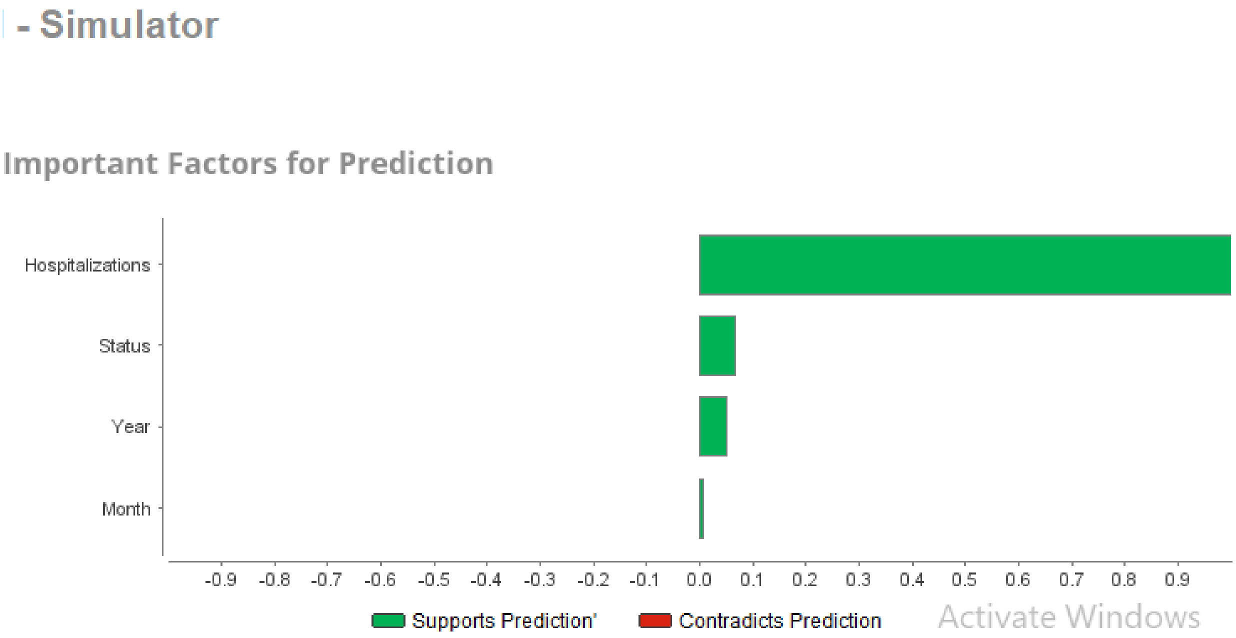


Figure 1.3 Generalized Linear Model Important Factor for prediction chart

Random Forest- Prediction /chart

Random Forest prediction chart is shown in the fig 1.4. It shows the true values for the illness below the red dotted line of chart and predicting values are above this line. It also contains the more true values as compared to the predicting values.

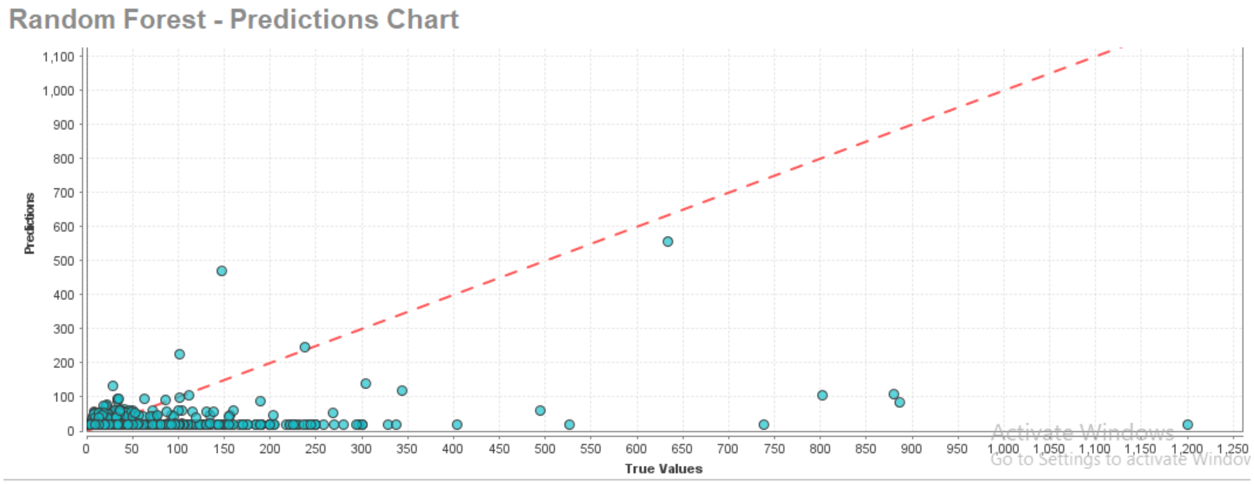
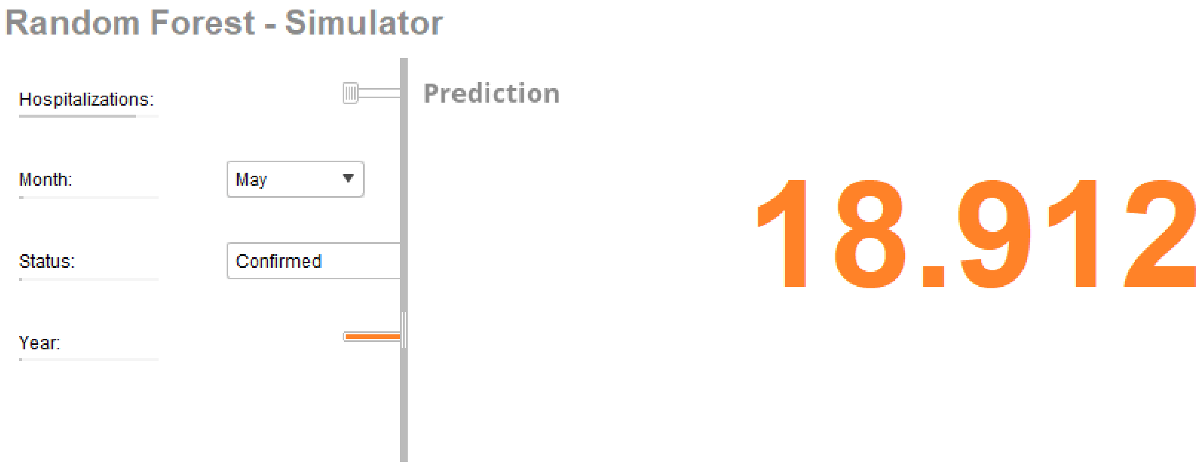


Figure 1.4 Random Forest- Prediction /chart

Random Forest Simulation Prediction

Random Forest simulation prediction value is shown for the required dataset. Fig 1.5 shows the shows the smallest predictive value.



Random Forest - Simulator

Hospitalizations:

Month: May

Status: Confirmed

Year:

Prediction

18.912

Figure 1.5 Random Forest Simulation Predictions

Figure 1.6 shows the important factors for prediction for Random Forest. Green color shows the support prediction and red color shows the contradict prediction. It means it has no contradict prediction.

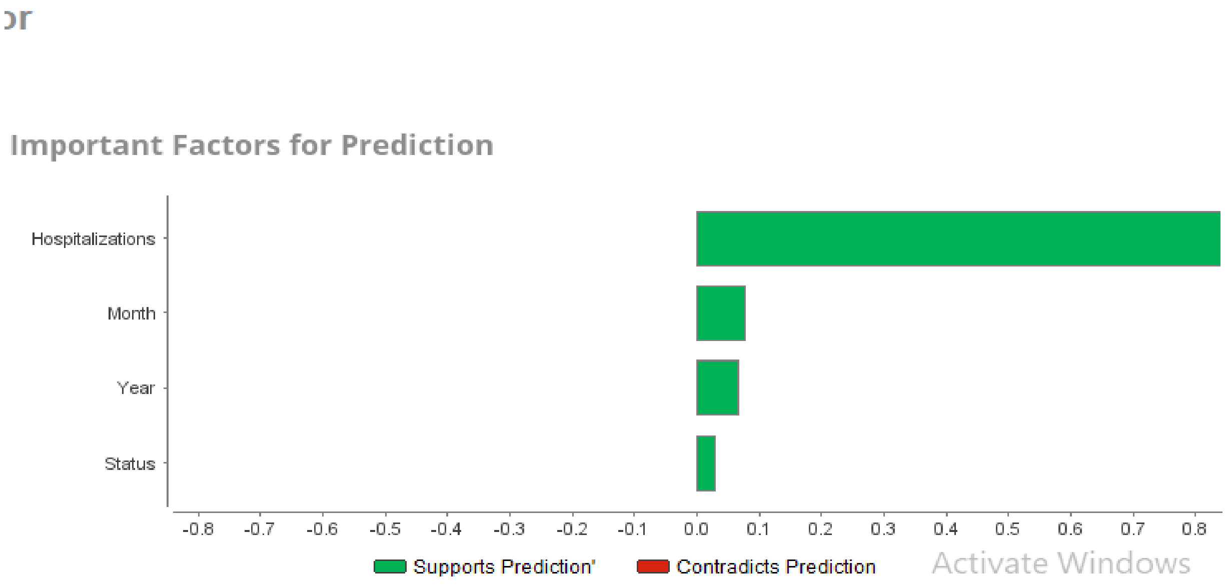


Figure 1.6 Random Forest Important Factor for prediction chart

Gradient Boosted Trees- Prediction /chart

Gradient Boosted Tree prediction chart is shown in the fig 1.7. It shows the true values for the illness below the red dotted line of chart and predicting values are above this line. It also has the more true values as compared to the predicting values.

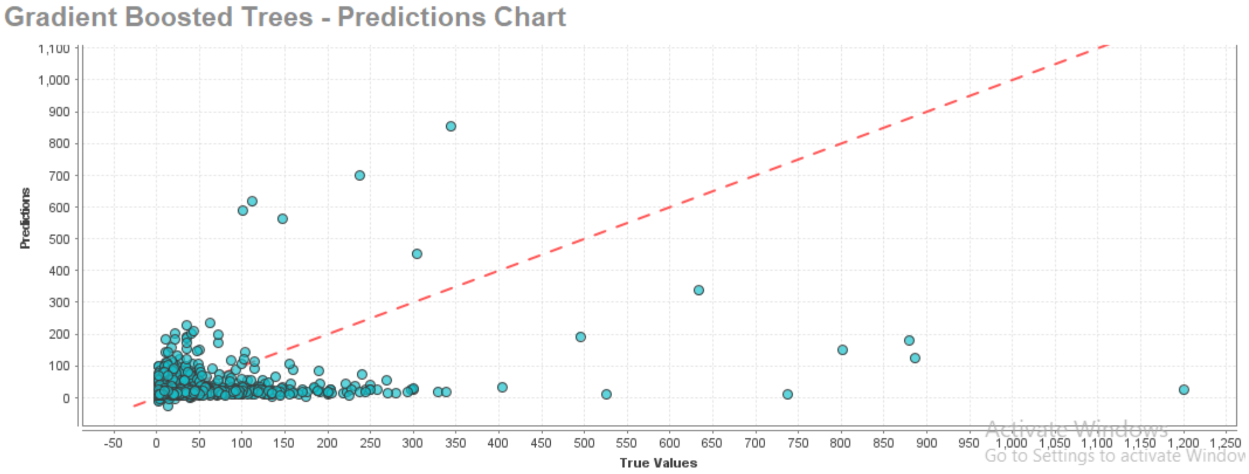
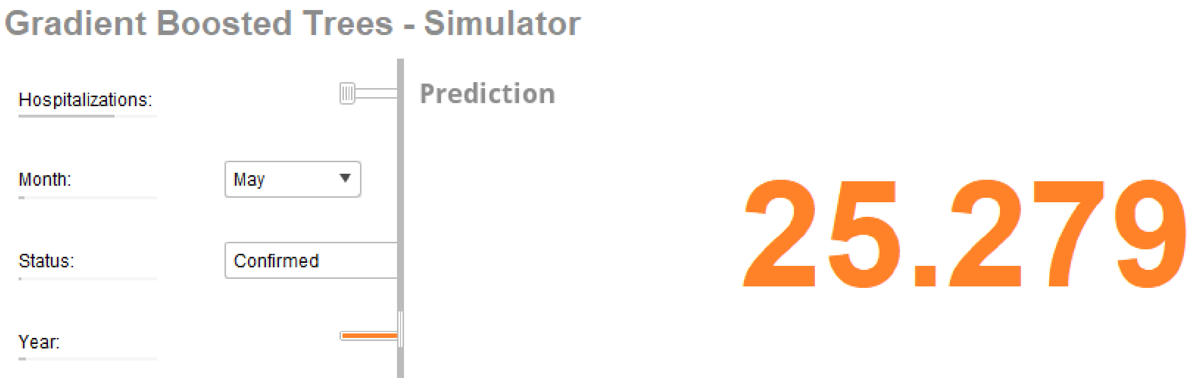


Figure 1.7 Gradient Boosted Trees- Prediction /chart

Gradient Boosted Trees – Simulator Prediction

Gradient Boosted Tree simulation prediction value is shown for the required dataset. Fig 1.8 shows the shows the largest predictive value among the all predictions.



Gradient Boosted Trees - Simulator

Hospitalizations:

Month: May

Status: Confirmed

Year:

Prediction

25.279

Figure 1.8 Gradient Boosted Trees – Simulator Prediction

Figure 1.9 shows the important factors for prediction for Random Forest. Green color shows the support prediction and red color shows the contradict prediction. It means it has no contradict prediction.

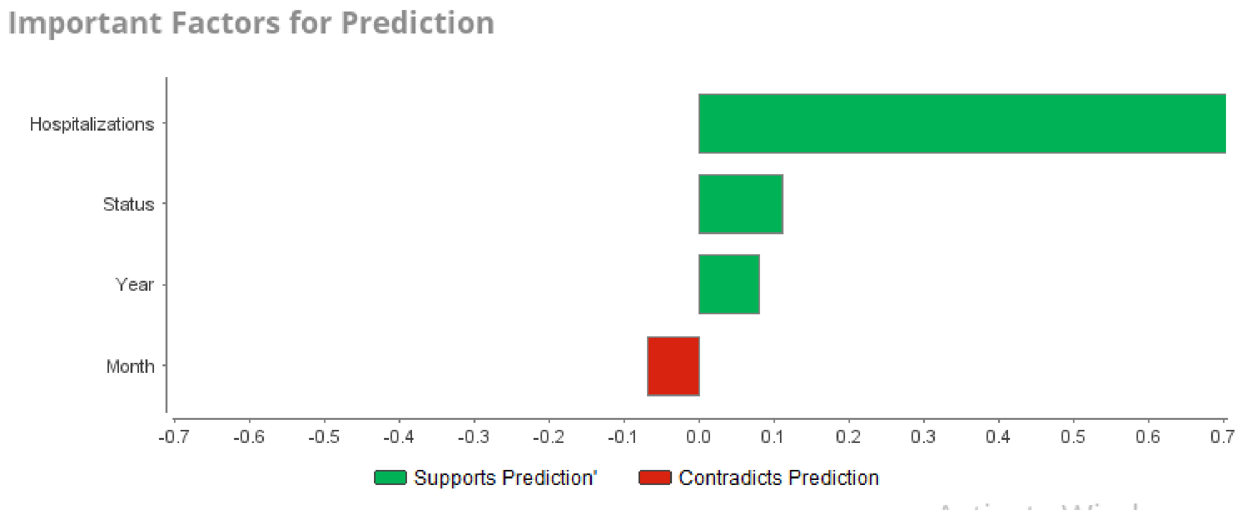


Figure 1.9 Gradient Boosted Trees Important Factors for Prediction

Clustering

Two types of clustering algorithms are applied: k means and x means. Two clusters are made namely cluster 0 and cluster 1 for the dataset of Foodborne Diseases outbreaks.

k-Means – Cluster Tree

K means Cluster tree is shown in the fig 2.0 for the dataset of Foodborne disease outbreaks. It's a not a huge cluster tree and can easily be shown in one frame clearly.

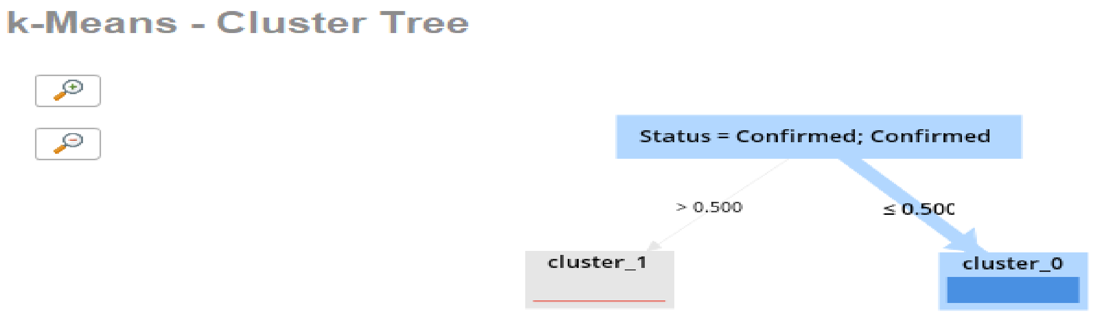


Figure 2.0 k-Means – Cluster Tree

x-Means – Cluster Tree

X means Cluster tree is shown in the fig 2.1 for the dataset Foodborne disease outbreak. It’s a huge cluster tree as compared to the k-Means cluster tree and difficult to shown in one frame clearly.

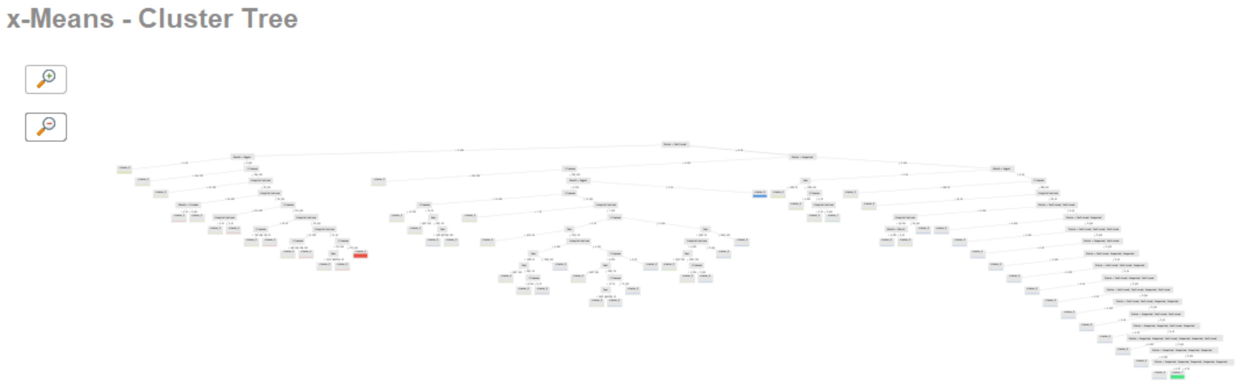


Figure 2.1 x – Means – Cluster Tree

Results/Conclusion

We see the different results in the prediction and in the clustering. In prediction Gradient boosted tree shows the largest values than generalized linear model and Random Forest for the dataset. In clustering K means and X means algorithms are used. They show the different results. The more precise one is X means in the case of this dataset as it made a huge and detailed cluster tree.

Future work

To get the more precise and accurately predicted values regarding Foodborne disease dataset other models and clustering algorithms should be applied. More accurate and better results can be produce when applied the different models which are heavier and take long time to process. We can also change some attributes in the dataset and can increase the number of entries for more accurate results. Results can also be more accurate when there is no distortion in the dataset. All missing values and dirt should be removed from the dataset for better output.

References

1. Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., ... & Davis, J. J. (2019). Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella. *Journal of clinical microbiology*, *57*(2).

2. Hall, G. V., D'Souza, R. M., & Kirk, M. D. (2002). Foodborne disease in the new millennium: out of the frying pan and into the fire?. *Medical Journal of Australia*, *177*(11), 614-618.

3. Ribot, E. M., & Hise, K. B. (2016). Future challenges for tracking foodborne diseases: PulseNet, a 20-year-old US surveillance system for foodborne diseases, is expanding both globally and technologically. *EMBO reports*, *17*(11), 1499-1505.

4. Wang, X., Zhou, M., Jia, J., Geng, Z., & Xiao, G. (2018). A Bayesian approach to real-time monitoring and forecasting of Chinese foodborne diseases. *International journal of environmental research and public health*, *15*(8), 1740.

5. Schlaich, T., Horn, A. L., Fuhrmann, M., & Friedrich, H. (2020). A Gravity-Based Food Flow Model to Identify the Source of Foodborne Disease Outbreaks. *International Journal of Environmental Research and Public Health*, *17*(2), 444.