Inter-County Comparative Analysis of ID3 Decision Tree Algorithms for Disease Symptom Burden Classification and Diagnosis

Nicodemus Nzoka Maingi[1], Ismail Ateya Lukandu[2], Matilu Mwau[3]

[1]Faculty of Information Technology, Strathmore University, Ole Sangale Road, Madaraka Estate, Nairobi, Kenya

[2]Faculty of Information Technology, Strathmore University, Ole Sangale Road, Madaraka Estate, Nairobi, Kenya

[3]Kenya Medical Research Institute (KEMRI), Ministry of Health, Off Mbagathi Way, Nairobi, Kenya

Abstract: ***The ID3 decision tree algorithm provides a key method of defining decision trees that can be used to prioritize and eventually classify disease outbreak symptom burdens in the fight against disease outbreaks. The decision trees are mainly derived from the calculation of the entropy of using some predefined variables of interest, herein referred to as disease symptom burden variables (which generally point to any disease's symptoms coded into variables accordingly) and then ranking of information gain ratios of the various disease symptom burdens. The decision trees can then be compared to draw important ideas and knowledge. The comparison of the decision trees for various geographical regions (counties) provides key ideas to better understanding of the various similarities and differences, be they just pure random, geographical, or even deliberate. This comparative understanding can help the relevant authorities in better joint policy development and business continuity planning in the event of any disruptive disease outbreaks. The comparison could trigger some critical vantage points; providing better economies of scale in running joint surveillance activities as compared to individualized planning and executions, pooling efforts together to create useful and unassailable synergistic styles of execution, and finally it also allows the various teams bring in unique skills and experiences that wouldn't have been possible in separately executed endeavors. Ultimately, such efforts could also help the health and government personnel get to easily identify common attributes and results that could prove key in fighting disease outbreaks. Since the algorithm used here breaks down each disease into its constituent symptomatic burdens, it helps to cluster together those attributes or symptom burden variables that are most critical in the fight against disease outbreaks instead of the traditional focus on the general diseases alone.***

Keywords: Decision Trees, Entropy, Information gain, ID3, Disease Symptom Burdens

1. Introduction

In Kenya, the mandate to manage and mitigate disease outbreaks and their effects falls under the Ministry of Health's Disease Surveillance and Response Unit (DSRU). Modeling chronic and infectious diseases entails tracking and describing individuals and their attributes (such as disease status, date of diagnosis, risk factors and so on) as they move and change through space and time (Jacquez, et al, 2014). According to Chen et al (2005), information technology has now become an indispensable part of making nations safer and more responsiveness to disease outbreak responses and mitigation measures. Machine learning offers a principled approach for developing sophisticated, automated, and objective algorithms for the analysis of multidimensional and multimodal biomedical data (Sajda, 2006). Decision tree theory is considered to be one of the most popular approaches for representing classifiers in many studies (Rokach, 2005). The concept of decision tree construction has been applied in multiple disciplines, ranging from statistics, machine learning, pattern recognition, as well as data mining.

2. Literature Review

Machine learning deals with the challenge of designing computers that automatically learn and adapt over time through experience. Seating between computer science and statistics, it is arguably one of the most rapidly expanding technical fields; remaining at the very core of artificial intelligence and data science (Jordan et al, 2015). According to Rokach (2005), a decision tree is a classifier expressed as a recursive partition of an instance's space. It consists of nodes, starting with the first node, commonly referred to as the root node, and followed by other lower level tree nodes, commonly referred to as the leaf nodes. Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree (Mitchell, 1995) The iterative dichotomizer (ID3) is a simple decision tree learning algorithm whose basic idea is to construct the decision tree by employing a top-down, greedy search through given data sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, the terms entropy and information gain are used (Peng. 2009). Quinlan (2014) defines entropy as the amount of uncertainty in a system, and information gain as a calculated value that is used to determine the most useful attribute in the construction of a classification decision tree.

3. Methodology

The methodology employed here is mostly experimental research, coupled with evolutionary prototyping and modeling. Experimental research mainly points to the systematic, theoretical analysis of the methods applied to a field of study (Howell, 2012). The data for the two counties of interest is gathered and broken down into its constituent disease burden variables. The variable data is then used to compute the entropies, information gains and their respective rankings, which are eventually used to construct the decision trees which are eventually compared for any deductive and useful observations.

Entropy Determination:

$$\text{Entropy (Decision)} = \sum_{n=1}^{\infty} -\text{p(Decision)}.\log_2 \text{p(Decision)}$$

Equation (1)

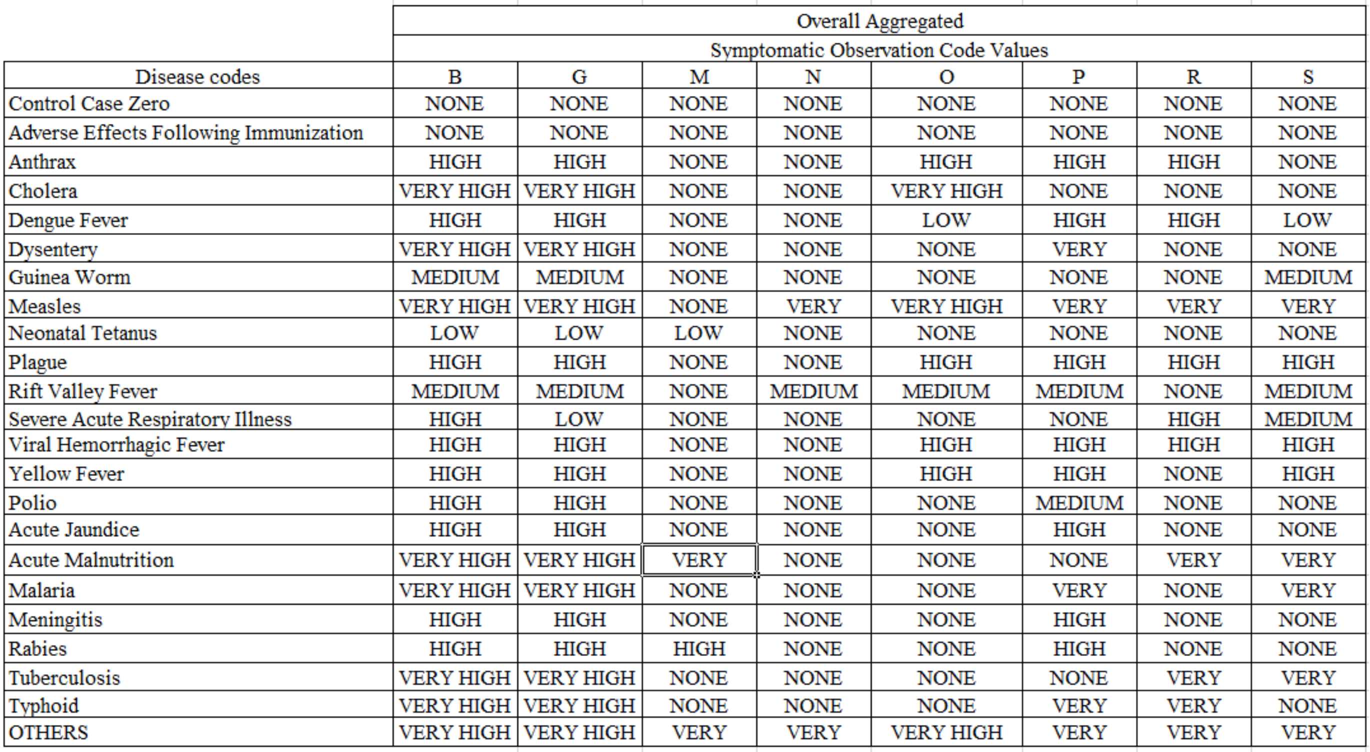
Information Gains Determination:

$$\text{I G (Decision, Variable)} = \text{Entropy(Decision)} - \sum_{n=1}^{\infty} [\text{p(Decision|Variable)}.\text{Entropy(Decision|Variable)}]$$

Equation (2)

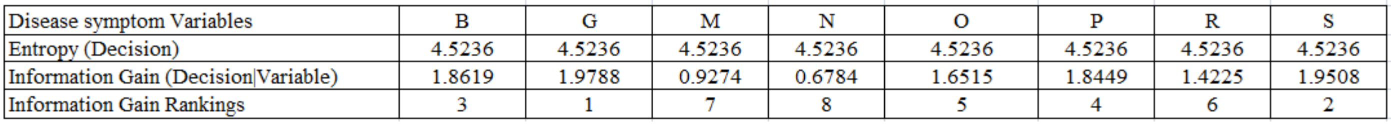
4. Data Results and Analysis

Table 1: Nairobi Country Disease Symptom Burdens Variable Data (2015 – 2018)



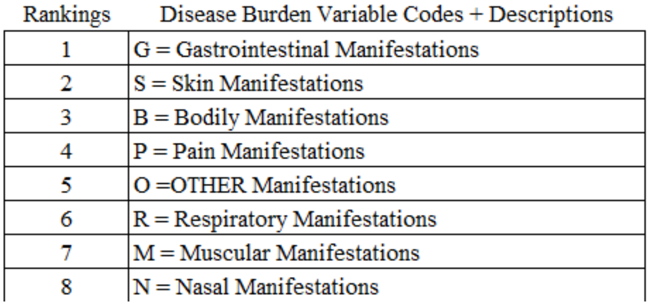
	Overall Aggregated							
	Symptomatic Observation Code Values							
Disease codes	B	G	M	N	O	P	R	S
Control Case Zero	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
Adverse Effects Following Immunization	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
Anthrax	HIGH	HIGH	NONE	NONE	HIGH	HIGH	HIGH	NONE
Cholera	VERY HIGH	VERY HIGH	NONE	NONE	VERY HIGH	NONE	NONE	NONE
Dengue Fever	HIGH	HIGH	NONE	NONE	LOW	HIGH	HIGH	LOW
Dysentery	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY	NONE	NONE
Guinea Worm	MEDIUM	MEDIUM	NONE	NONE	NONE	NONE	NONE	MEDIUM
Measles	VERY HIGH	VERY HIGH	NONE	VERY	VERY HIGH	VERY	VERY	VERY
Neonatal Tetanus	LOW	LOW	LOW	NONE	NONE	NONE	NONE	NONE
Plague	HIGH	HIGH	NONE	NONE	HIGH	HIGH	HIGH	HIGH
Rift Valley Fever	MEDIUM	MEDIUM	NONE	MEDIUM	MEDIUM	MEDIUM	NONE	MEDIUM
Severe Acute Respiratory Illness	HIGH	LOW	NONE	NONE	NONE	NONE	HIGH	MEDIUM
Viral Hemorrhagic Fever	HIGH	HIGH	NONE	NONE	HIGH	HIGH	HIGH	HIGH
Yellow Fever	HIGH	HIGH	NONE	NONE	HIGH	HIGH	NONE	HIGH
Polio	HIGH	HIGH	NONE	NONE	NONE	MEDIUM	NONE	NONE
Acute Jaundice	HIGH	HIGH	NONE	NONE	NONE	HIGH	NONE	NONE
Acute Malnutrition	VERY HIGH	VERY HIGH	VERY	NONE	NONE	NONE	VERY	VERY
Malaria	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY	NONE	VERY
Meningitis	HIGH	HIGH	NONE	NONE	NONE	HIGH	NONE	NONE
Rabies	HIGH	HIGH	HIGH	NONE	NONE	HIGH	NONE	NONE
Tuberculosis	VERY HIGH	VERY HIGH	NONE	NONE	NONE	NONE	VERY	VERY
Typhoid	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY	VERY	NONE
OTHERS	VERY HIGH	VERY HIGH	VERY	VERY	VERY HIGH	VERY	VERY	VERY

Table 2: Nairobi County ID3 Entropy, Information Gains and Rankings



Disease symptom Variables	B	G	M	N	O	P	R	S
Entropy (Decision)	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236
Information Gain (Decision|Variable)	1.8619	1.9788	0.9274	0.6784	1.6515	1.8449	1.4225	1.9508
Information Gain Rankings	3	1	7	8	5	4	6	2

Table 3: Nairobi County Disease Symptom Burden Variables Rankings Mapped to their Codes and Descriptions



Rankings	Disease Burden Variable Codes + Descriptions
1	G = Gastrointestinal Manifestations
2	S = Skin Manifestations
3	B = Bodily Manifestations
4	P = Pain Manifestations
5	O =OTHER Manifestations
6	R = Respiratory Manifestations
7	M = Muscular Manifestations
8	N = Nasal Manifestations

Nairobi County ID3 Decision Tree Construction (based on Information Gains Rankings of Disease Burden Variables)

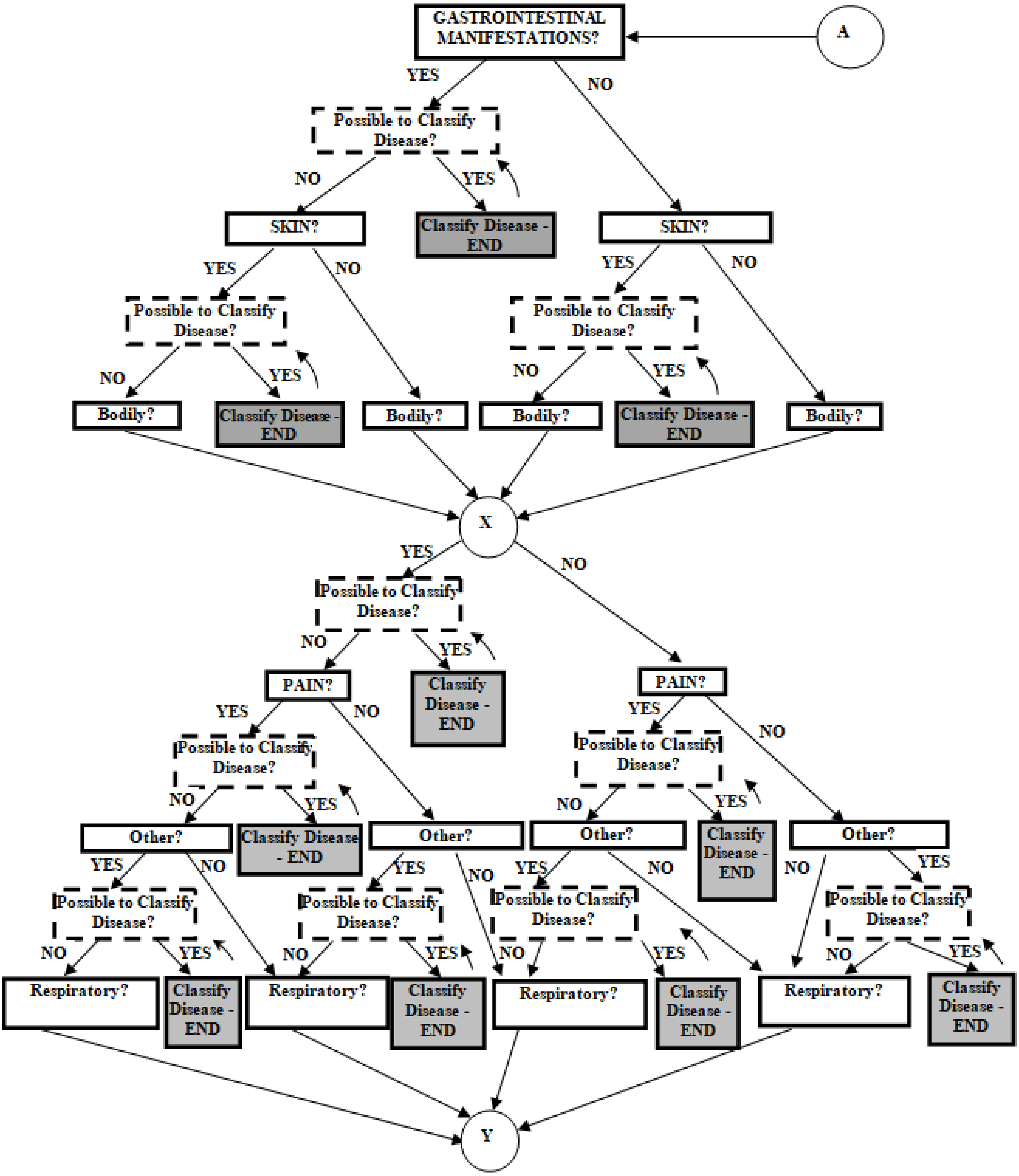


Figure 1: Nairobi County ID3 Decision Tree

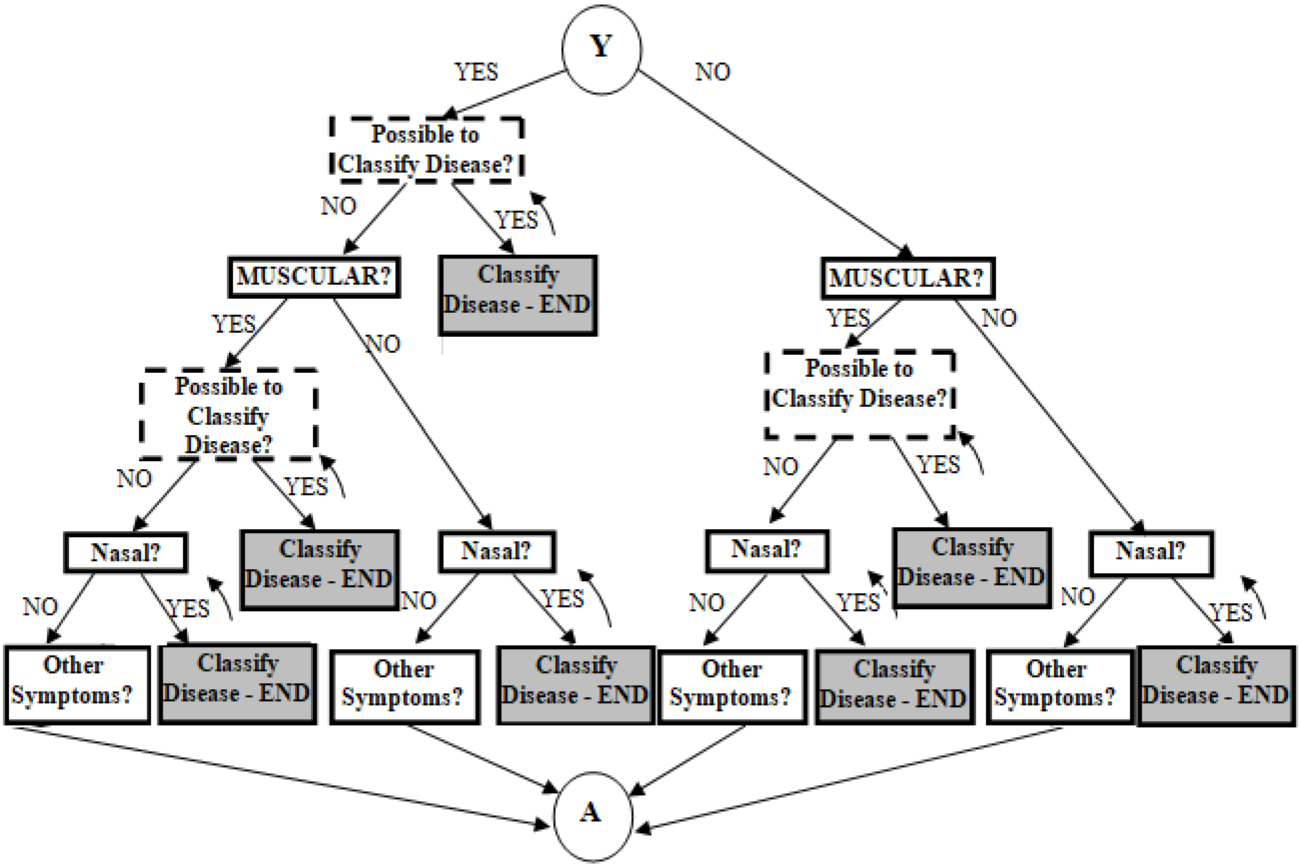


Figure 2: Nairobi County ID3 Decision Tree (Continuation)

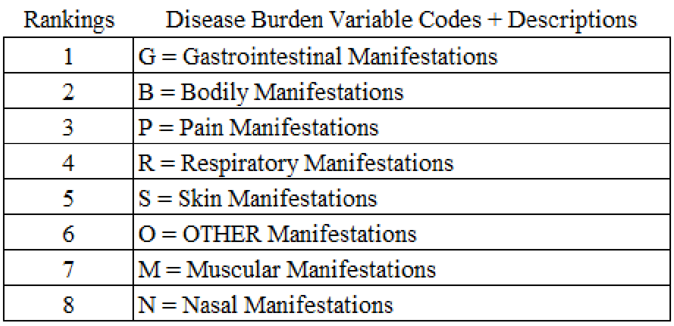
Table 4: Kajiado Country Disease Symptom Burdens Variable Data (2015 – 2018)

	Overall Aggregated							
	Symptomatic Observation Code Values							
Disease codes	B	G	M	N	O	P	R	S
Control Case Zero	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
Adverse Effects Following Immunization	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
Anthrax	LOW	LOW	NONE	NONE	LOW	LOW	LOW	NONE
Cholera	HIGH	HIGH	NONE	NONE	HIGH	NONE	NONE	NONE
Dengue Fever	LOW	LOW	NONE	NONE	LOW	LOW	LOW	LOW
Dysentery	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY HIGH	NONE	NONE
Guinea Worm	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
Measles	HIGH	HIGH	NONE	HIGH	HIGH	HIGH	HIGH	HIGH
Neonatal Tetanus	LOW	LOW	LOW	NONE	NONE	NONE	NONE	NONE
Plague	LOW	LOW	NONE	NONE	LOW	LOW	LOW	LOW
Rift Valley Fever	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
Severe Acute Respiratory Illness	LOW	NONE	NONE	NONE	NONE	NONE	LOW	LOW
Viral Hemorrhagic Fever	LOW	LOW	NONE	NONE	LOW	LOW	LOW	LOW
Yellow Fever	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
Polio	LOW	LOW	NONE	NONE	NONE	LOW	NONE	NONE
Acute Jaundice	HIGH	HIGH	NONE	NONE	NONE	HIGH	NONE	NONE
Acute Malnutrition	VERY HIGH	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY HIGH	VERY HIGH
Malaria	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY HIGH	NONE	VERY HIGH
Meningitis	LOW	LOW	NONE	NONE	NONE	LOW	NONE	NONE
Rabies	MEDIUM	MEDIUM	MEDIUM	NONE	NONE	MEDIUM	NONE	NONE
Tuberculosis	HIGH	HIGH	NONE	NONE	NONE	NONE	HIGH	HIGH
Typhoid	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY HIGH	VERY HIGH	NONE
OTHERS	VERY HIGH	VERY HIGH	VERY HIGH	VERY HIGH	VERY HIGH	VERY HIGH	VERY HIGH	VERY HIGH

Table 5: Kajiado County ID3 Entropy, Information Gains and Rankings

Disease symptom Variables	B	G	M	N	O	P	R	S
Entropy (Decision)	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236	4.4594	4.5236
Information Gain (Decision|Variable)	2.1227	2.1422	0.9274	0.5132	1.3062	1.9701	1.5694	1.5645
Information Gain Rankings	2	1	7	8	6	3	4	5

Table 6: Kajiado County Disease Symptom Burden Variables Rankings



Rankings	Disease Burden Variable Codes + Descriptions
1	G = Gastrointestinal Manifestations
2	B = Bodily Manifestations
3	P = Pain Manifestations
4	R = Respiratory Manifestations
5	S = Skin Manifestations
6	O = OTHER Manifestations
7	M = Muscular Manifestations
8	N = Nasal Manifestations

Kajiado County ID3 Decision Tree Construction (based on Information Gains Rankings of Disease Burden Variables)

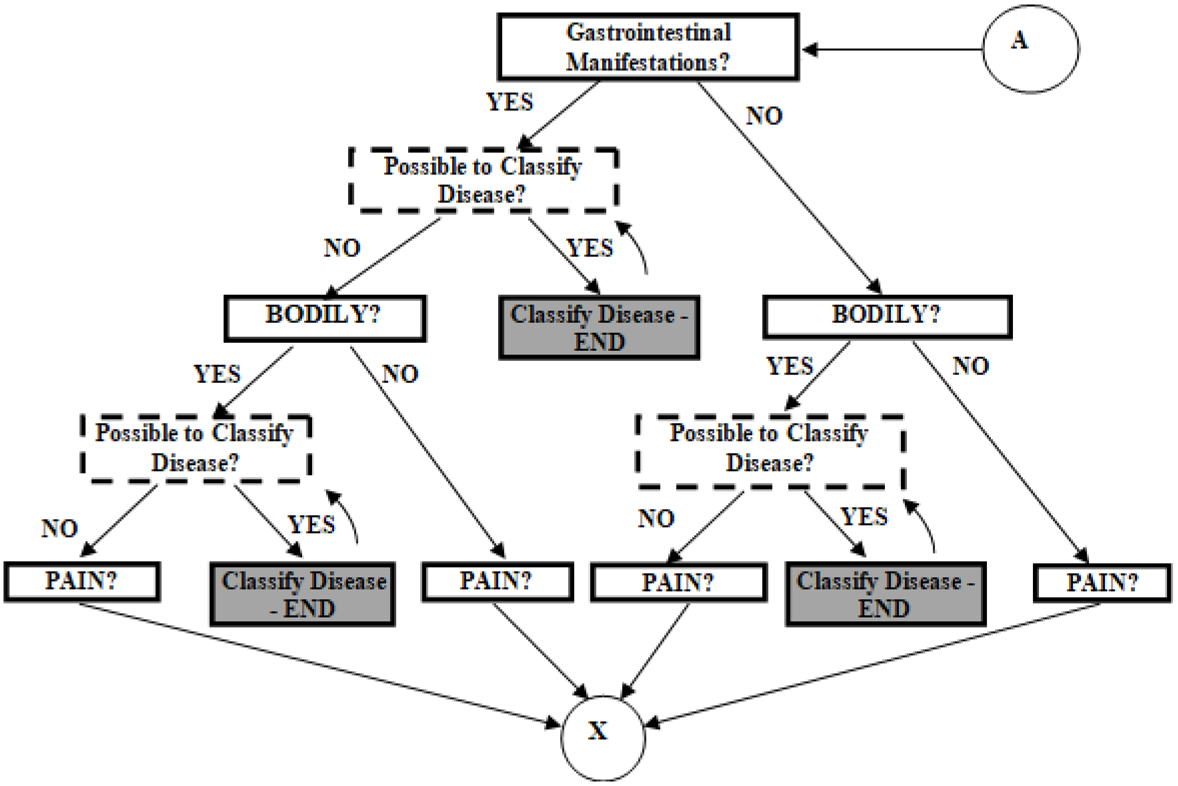


Figure 3: Kajiado County ID3 Decision Tree

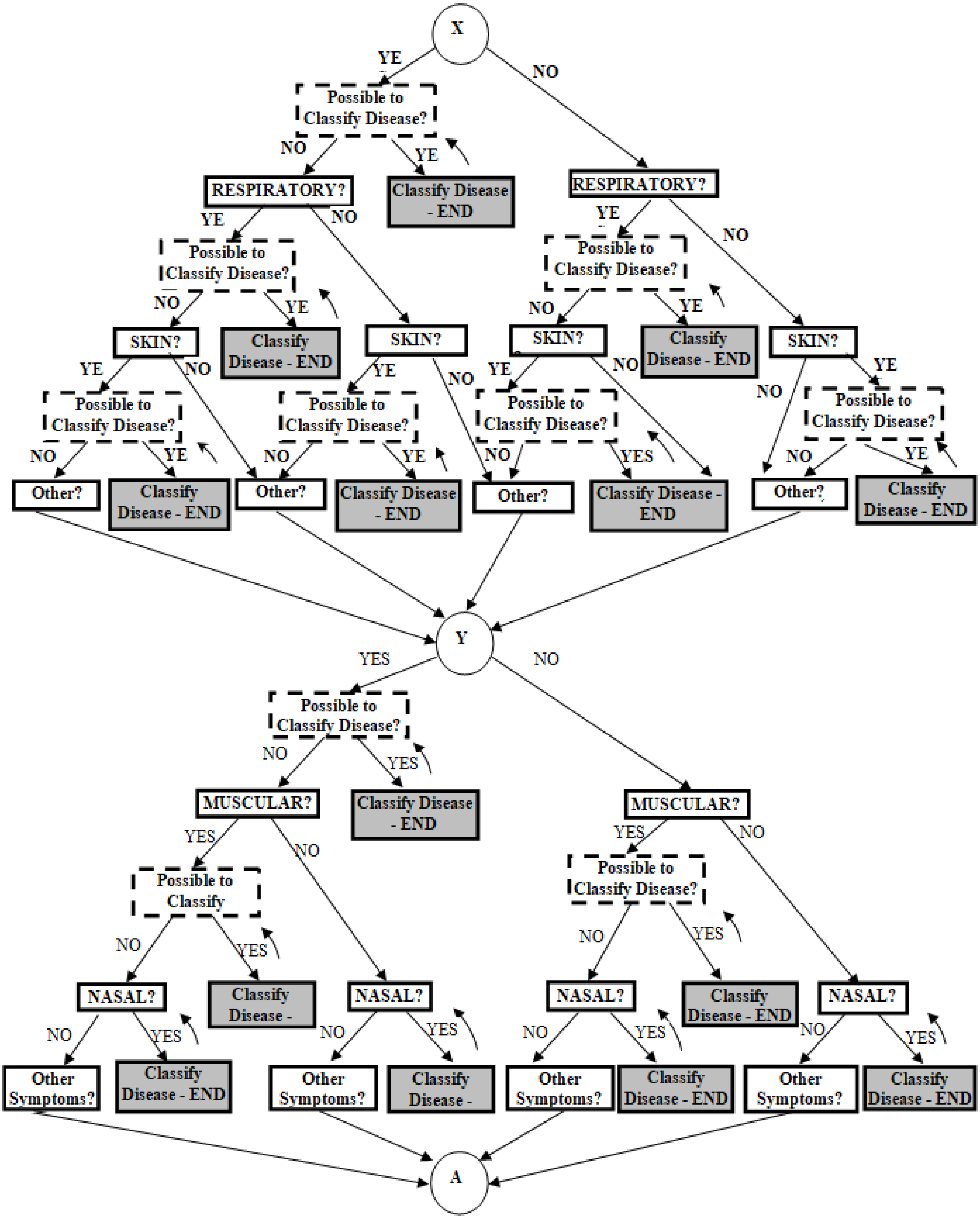
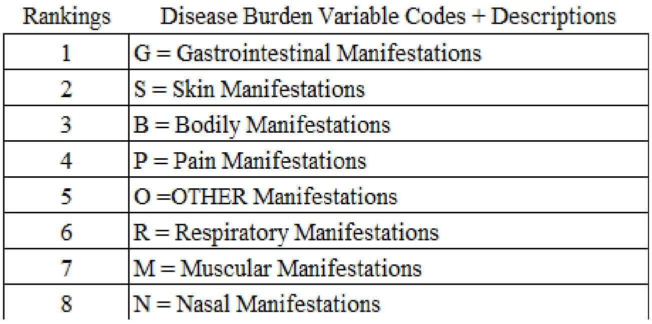


Figure 4: Kajiado County ID3 Decision Tree (Continuation)

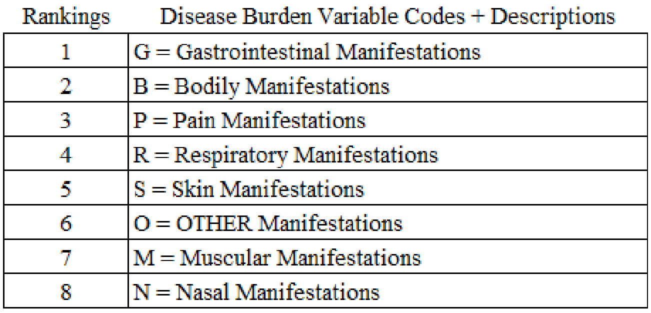
5. Results and Discussion

Table 7: Nairobi County Information Gain Rankings



Rankings	Disease Burden Variable Codes + Descriptions
1	G = Gastrointestinal Manifestations
2	S = Skin Manifestations
3	B = Bodily Manifestations
4	P = Pain Manifestations
5	O =OTHER Manifestations
6	R = Respiratory Manifestations
7	M = Muscular Manifestations
8	N = Nasal Manifestations

Table 8: Kajiado County Information Gain Rankings



Rankings	Disease Burden Variable Codes + Descriptions
1	G = Gastrointestinal Manifestations
2	B = Bodily Manifestations
3	P = Pain Manifestations
4	R = Respiratory Manifestations
5	S = Skin Manifestations
6	O = OTHER Manifestations
7	M = Muscular Manifestations
8	N = Nasal Manifestations

Comparatively, the two counties share a number of common observations:

Both the data sets generate information gain values for the gastrointestinal disease burden variable that ranks first.

However, there is a slight variation in the attributes ranking for the second, third, fourth, fifth and sixth attributes i.e. the Nairobi data set ranks Bodily manifestations second, then Pain comes third, Respiratory manifestation comes fourth, with Skin manifestations coming fifth, before crowning it with Other Manifestations coming sixth. The Kajiado county data varies in that it ranks Skin manifestations second, followed by Bodily manifestations, then Pain, then Other, and finally Respiratory comes sixth.

For both data sets, the Muscular and Nasal manifestations rank seventh and eighth respectively. In a sense, the decision trees have similar root nodes as well as the last two leaf nodes i.e. for muscular and nasal manifestations variables.

6. Conclusion

From the observations made from both decision trees, it could be argued that both counties have different disease burden challenges and that their focus or strategy should be slightly different given they only share common root nodes and the last two leaf nodes; the rest of the disease burden variables differ in the order of information gain ordering or ranking. From the results analyzed, generally, it could be construed that both Nairobi and Kajiado counties should totally mind each its own effort as the data does not closely tally across the board.

7. Recommendations for future research

Given that the decision trees generated from the two counties' data sets seem to differ, further validation of this position could be done by using an alternative algorithm in generating the information gains and consequent decision trees. The researcher could additionally employ the C4.5 or the Classification and Regression Trees (CART) algorithms to compare the outcome with the already analyzed ID3.. Once this is done, then the assertion that the two counties have different disease burden challenges could be easily validated or debunked, their proximate geographical locations notwithstanding.

References

[1] Chen, H., Wang, F. (2005). Artificial Intelligence for Homeland Security.

[2] Howell, K. E. (2012). An Introduction to the Philosophy of Methodology. Sage

[3] Jacquez, G. M., Greiling, D. A., Kaufmann, A. M. (2014). Design and Implementation of a Space-Time Intelligence System for Disease Surveillance.

[4] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[5] Peng, W., Chen, J., & Zhou, H. (2009). An implementation of ID3-decision tree learning algorithm. From web. arch. usyd. edu. au/wpeng/DecisionTree2. pdf Retrieved date: May, 13.

[6] Quinlan, J. R. (2014). C4. 5: Programs for Machine Learning. Elsevier.

[7] Rokach, L., & Maimon, O. (2005). Decision trees. In Data mining and knowledge discovery handbook (pp. 165-192). Springer, Boston, MA.

[8] Sajda, P. (2006). Machine learning for detection and diagnosis of disease. Annu. Rev. Biomed. Eng., 8, 537-565.

Author Profile

Nicodemus Maingi received the B.Sc. in Applied and Statistical Mathematics from Egerton University in Njoro, Nakuru, Kenya, before proceeding for his postgraduate studies at the University of Nairobi, where he undertook an M.Sc. in Information Systems (Artificial Intelligence and Knowledge-Based Systems). He has been lecturing and doing research at Strathmore University's Faculty of Information Technology from August 2001 to date. In 2010, he founded the HP-Strathmore Research Laboratory (commonly referred to as HP Lab), a novelty research lab through which undergraduate and postgraduate students are mentored and challenged with various industry-level research and consultancy projects in order to prepare them for the industry work environment once they graduate. The HP Lab has brought on board various research and consultancies both from private and public sector player, making it a catalyst in innovation and driving technology to tackle local problems using local talent. Mr. Maingi is now completing his doctoral studies in Health Informatics at Strathmore University's School of Graduate Studies.