Prediction of Poultry Yield Using Data Mining Techniques

[1]Akanmode, E. R

School of IT & Computing, American University of Nigeria, Yola, Adamawa state, Nigeria

[2]Dr. N.D. Oye

Department of Computer Science, ModibboAdama University, Yola, Adamawa state, Nigeria.

[3]Celestine, H. C.

Department of Computer Science, ModibboAdama University, Yola, Adamawa state, Nigeria.

Corresponding Author Email: oyenath@yahoo.co.uk

ABSTRACT

A poultry yield prediction model have been designed using a data mining and machine learning technique called Classification and Regression Tree (CART) algorithm. The developed model has been optimized and pruned using the Reduced Error Pruning (REP) algorithm to improve prediction accuracy. An algorithm to make the prediction model flexible and capable of making predictions irrespective of poultry size or population has been proposed. The model can be used by poultry farmers to predict yield even before a breeding season. The model can also be used to help farmers take decisions to ensure desirable yield at the end of the breeding season.

Keywords: Datamining; Prediction; Poultry yield, Cart Algorithm

I. INTRODUCTION

Over the years, pattern extraction from data has evolved from manual to automated processing. Early pattern extraction methods includes Bayes‘ theorem from the 1700s to regression analysis in the 1800s. The revolution of technology especially computer technology has brought about increase in large data storage, collection and manipulation hence the need for methods and techniques to efficiently discover patterns in these large data (Mucherino*et al*., 2009). The need for data exploration and extraction later brought about discoveries in Computer Science such as cluster analysis, neural networks, genetic algorithms, decision rules, decision trees and support vector machines; all of which constitute methods of data mining (Han *et al.,* 2011). Data mining is therefore the process of exploring large data sets so as to find purposeful patterns, relationships, correlations or associations within the data sets (Klosgen and Zytkow 2002). It forms the intersection linking various disciplines such as computer science, statistics, machine learning and database systems (Bozdogan, 2003). The main objective of data mining is to convert meaningless data to meaning information which results to knowledge discovery (Sumathi and Sivanandam, 2006). Data mining goes beyond just analyzing raw data. It involves establishment of practices and policies that manage full data life cycle of an organization or enterprise. Data mining also involves building of models and deduction of inference (Han *et al.,* 2011). This means that data mining goes beyond the mere extraction (mining) of data but the extraction of patterns from data to produce knowledge. One attribute data mining and database share is the storing, manipulation and extraction of data.

Data is collected and stored (database). The data is then worked upon (data mining) which results in knowledge discovery. The discovered knowledge can then be stored for further use (database). Different terms have been used to reference data mining. Terms such as: data archaeology, information harvesting, information discovery, knowledge extraction and so on. Gregory Piatetsky-Shapiro invented the term "knowledge discovery in databases" (KDD) in 1989. However, because of the popularity of the term ―data mining‖ in machine learning and artificial intelligence (AI) community, the terms KDD and data mining have been used interchangeably (Piatetsky-Shapiro *et al.,* 2011). In general, data mining encompasses six common types of tasks. They are anomaly detection, association rule learning, clustering, classification, regression and summarization (Thuraisingham, 1998). Data are basically mined to achieve one or more of these tasks. Data mining in agriculture is a recent research field (Ramesh and Vardhan, 2013). It is also considered as the future of agriculture (ElFangary, 2009). This forms the basic motivation behind this research. Thus far, some data mining applications in agriculture include: detection of diseases from animal sounds, predicting crop yield, weather and soil types forecasting, etc.

Poultry Farming in Nigeria

Poultry can simply be defined as domesticated birds reared for meat, egg and feather purposes. In Nigeria, poultry is mainly reared for meat and egg purposes. For this reason, the two main poultry breeds reared in Nigeria are broilers (for meat) and layers (for eggs). Other popular poultry breeds in Nigeria include guinea fowls, cockerels, ducks and turkeys. Poultry farming in Nigeria has been on a tremendous rise. This may be attributed to the high rate of unemployment in the country. For some individuals and states, poultry farming has become a means of revenue generation. Nigerians depend heavily on poultry husbandry to create self-employment in a bid to reduce poverty (Heise et al., 2015). Agriculture is a dominant practice in Sub-Saharan Africa countries like Nigeria and is seen as a major instrument for poverty alleviation in the Sub-Saharan region (Larsen et al., 2009). It is therefore important to introduce ideas that will improve poultry husbandry in Nigeria. This research intends to improve poultry farming by developing models that poultry farmers can use to forecast or predict yield using data mining techniques.

Statement of the problem

Prediction of yield or harvest is most farmers' problem. Farmers have often depended on previous experiences to forecast yield but this method most times turns out non-reliable and incorrect (Ramesh and Vardhan, 2013). If farmers can have an idea of what yield will be during the harvest period the farmers take adequate steps or decisions to ensure maximum yield. With data mining, patterns from poultry data that can lead to predictions can be discovered to provide poultry prediction models. The aim of this research work is to develop a prediction model using data mining techniques that can help poultry farmers to predict yield. The objective of the research is to provide local farmers with a tool in form of a model that they can apply to predict yield for upcoming breeding seasons. In the same vein, the model can help poultry farmers navigate through various decision processes as they try to cut costs (cost effective poultry farming).

Justification of Study

Poultry farming in Nigeria has been on a tremendous rise over the past decades. This may be attributed to the high rate of unemployment in the country. For some individuals and states in Nigeria, poultry farming has become a means of revenue generation. Nigerian as a sub-Sahara African country rely on agricultural activities including poultry farming to create self-employment in a bid to reduce poverty (Larsen et al., 2009; Heise et al., 2015). It is therefore important to introduce ideas that will improve poultry farming in Nigeria. The researchers intend to improve poultry farming by developing a model that poultry farmers can use to forecast or predict yield using data mining techniques. Poultry farmers like every other business man (or woman), juggle between opportunity costs, foregoing some needs in favour of others and at the same time, targeting maximum yield as possible. This study is particularly useful as it can help poultry farmers through a number of permutations of certain factors that affect poultry production and the possible yields that can result from such permutations.

Scope and Limitation

This research is restricted to Adamawa State in particular or the north-eastern region at large. This is because weather factors of Adamawa state have been considered. It has been assumed that weather conditions in other regions of the country differ from the weather conditions in Adamawa state, a north-eastern regional state in Nigeria.

Majority of the data used for the research constitute breeds of broilers and layers with quite a few on turkeys and guinea fowls. Therefore the yield prediction model is not expected to be applicable for birds such as ostriches, pigeons, parrots and so on.

II. Literature Review

Data Mining Prediction Techniques in Agricultural Research

Clustering algorithm is the technique used to identify appropriate groups of instances in a given set of data (Aggarwal and Reddy, 2014). This algorithm is used when no prior knowledge of the data is available therefore the concept of training or learning data set is practically impossible (Mucherino et al., 2009). A k-means variant (k-means clustering) of the clustering algorithm is among the most popular of the clustering algorithm, ranked among the top 10 algorithm of all the data mining algorithms (Wu et al., 2008). It is therefore no surprise that it has been applied in agricultural research. For example, Urtubia*et al.,* (2007) predicted the problems associated with wine fermentation using the k-means algorithm. The fermentation problem of wine is that the process can be too slow or stagnant (Urtubia*et al.,* 2007; Muchirino*et al.,* 2009). It is therefore important to ensure that the fermentation process concludes smoothly to produce the desired wine quality. To be able to achieve this, metabolites such as organic acid, fructose, glucose, glycerol and ethanol were collected and analysed to obtain data of the fermentation process. The data obtained from the first three days were compared with the data for the whole fermentation process. The k-means algorithm proved that the data for the first 3 days of fermentation was sufficient enough to determine the final outcome of fermentation process. This means that theentire fermentation process can be determined after 3 days and adequate measures can be taken early to improve the wine quality. The K-Nearest Neighbour (K-NN) is another classifier algorithm that works by using the popular principle ―birds of a feather move together‖ (Mucherino*et al.,* 2009). This algorithm tends to classify instances based on the class of its nearest neighbour (Kotsiantis*et al.,* 2007). Like the K-means clustering algorithm, the K-NN algorithm is also ranked among the top 10 data mining algorithms (Wu *et al.,* 2008).

The K-NN classifier was prescribed as an efficient method for estimating soil water parameter (Mucherino*et al.,* 2009) using crop simulation systems such as CROPSYST (Stockle*et al.,* 1994), DSSAT (Jones *et al.,* 1998) or any crop simulation system. Soil parameters such as the lower limit of plant water availability (LL), the drained upper limit and plant extractable soil water (PESW) are most likely to be unavailable. K-NN algorithm can be used on available information such as soil texture and organic carbon to obtain the unavailable parameters (Mucherino*et al.,* 2009). This shows that K-NN classifier can be used to predict unknown variables from known ones. ElFangary (2009) developed a model for improving cow and buffalo production in Egypt. The research used Pearson's Coefficient to analyse and find correlations between variables such as pregnancy, death, diseases, vaccines and the various interval of the animals' production to develop the model. The Artificial Neural Network (ANN) algorithm is another powerful classifier used for prediction. A typical example of its application was demonstrated by (Kondo *et al.,* 2000) to predict that certain categories of oranges are relatively sweeter by measuring the sugar and acid content of oranges. A three-layer artificial neural network was used to predict that oranges with attributes: reddish color, medium size, low height and glossy appearance are relatively sweeter. Another application of ANN in agriculture was conducted on pigs to detect the presence of diseases via their sounds (Moshou*et al.,* 2001). Initially sound samples of

354 sounds were trained. The sounds consist of coughs from different pigs, metal clanging, grunts, and background noise. Sounds such as cough and metal clanging were difficult to distinguish because they have similar frequency range (Mucherino*et al.*, 2009). The neural network was further trained to distinguish the similar sound. Once that was done, result showed sound recognition correctness greater than 90%.

Similarly, ANN was used to detect watercore in apples (Shahin *et al.,* 2001). Watercore is an interior apple disorder (Mucherino*et al.,* 2009; Herremans, 2014). An ANN was able to identify good apples from bad ones based on their watercore severity. This study was necessary because watercore is an internal disorder and consumers could only discover it after purchase of the apple (Mucherino*et al.,* 2009). The Support Vector Machine (SVM) technique is normally restricted to discriminate between two classes (Mucherino*et al.,*2009; Campilho and Kamel, 2014). Gill *et al.,* (2006) used meteorological and soil moisture to develop SVM predictions for four and seven days forecast of soil moisture. Just like Moshou*et al.,* (2001) research on pigs, Fagerlund (2007) used SVM to distinguish and recognize different bird species based on birds' sounds. Bird sound data were used to train a SVM classifier in conjunction with a binary decision tree. N-fold cross validation was then used to obtain the optimal classifier model that identifies birds.
Crop Yield has been predicted using Multiple Linear Regression (MLR) and Density-Based Clustering Data Mining technique (Ramesh and Vardhan, 2015). Rajeshwari and Arunesh (2016) used three Classification techniques: Naïve Bayes, JRip and J48 (also called C4.5 algorithm) to analyse and predict soil types: red and black. JRip and J48 algorithms are decision tree algorithm proposed by William Cohen and Ross Quinlan respectively. This researcher shows that both decision tree algorithms produced higher prediction accuracy rate compared to the Naïve Bayes technique. JRip and J48 produced 98.18% and 97.27% prediction accuracy while Naïve Bayes technique produced 86.36% prediction accuracy. Chowdhury and Ojha (2017) performed disease diagnosis on mushrooms using Naïve Bayes, Sequential Minimal Optimization (SMO) and Ripple-Down Rule Learner (RIDOR) Classification techniques. They concluded that the Naïve Bayes technique provides better results for mushroom disease diagnosis.

Data Mining Techniques in Poultry Farming

Study shows that very few research have been carried out in poultry farming and production. Thus far, no research has been done to predict poultry production or yield using CART. This constitutes a setback because very little literature is available upon which this research can complement and vice versa. Vale *et al.,* (2008) used decision tree, a prediction tool to estimate mortality rate in broilers when they are exposed to heat wave. The research further strengthens the claim that high temperatures have a negative effect on broilers. Sadeghi *et al.,* (2015) proposed a procedure to distinguish healthy broilers from unhealthy ones based on the sounds they make. The researcher used Fisher Discriminant Analysis (FDA) to classify the healthy broilers from the unhealthy ones. This research is particularly efficient for the early detection of diseases among broilers to enable farmers take appropriate measures.

Comparisons between Various Prediction Techniques

Clustering algorithms are generally easy to implement however, the algorithm require that output classes be identified upfront (Tiwari *et al.,* 2013, Jones, 2015). This is particularly a setback for this research since no prior knowledge of the outcome (yield) of the proposed prediction model is known since yield is as determined by factors such as vaccine, disease, feed and season. Like k-means algorithm, the KNN is relatively easy to implement. It can also be used to classify qualitative and quantitative data attributes (Banks *et al.,* 2011). However, result of the algorithm does not always yield a compact representation of the sample distribution; given room to errors as irrelevant samples will also be equally classified (Elder, 2009). In addition to this setback, the choice of the number of neighbours (K) can produce different results (Banks et al., 2011). Large computational time can also be an issue because the algorithm requires that the distance to every training pattern to be calculated (de Albornoz and Terashima, 2005). The ANN classifier is a fast learning algorithm which can automatically learn from training dataset. However, the algorithm is hard to interpret and apply to solve real life problems (Braspenning and Thuijsman, 1995; Patan, 2008). We are compelled to feel that this technique might be too complicated for an average farmer to understand and utilise.For SVM, Abe (2005) suggested the following advantages and disadvantages of SVM. The advantages are: strong generalization ability of the dataset provides global optimum solution and robust to outliers. Disadvantages include restriction to two classes thereby making multi-classification problem difficult and extended training time. Poultry yield is a continuous variable not a categorical variable. It therefore doesn't make sense to apply the SVM since the research goal is not to classify yield into two classes but to predict yield.

Decision tree is machine learning and data mining technique that produce models which are easy to interpret and understand (Rokach and Maimon, 2014). This technique is also capable to model variables that have a non-linear relationship with each other (Raut and Nichat, 2017). Decision trees work well with all variable types irrespective of whether it is categorical or continuous or both (Siau, 2008). Decision trees make use of a greedy algorithm which makes it very sensitive to outliers in the training set. In addition to this drawback, the greedy algorithm may result in error predictions at the leaves if an error occurs at corresponding higher level nodes (Rokach and Maimon, 2008). However, to handle the problem of error prediction, large amount of training data sets can be used to train the model (Mitchell, 1977; Aggarwal, 2015). Multiple Linear Regression (MLR) technique is only suitable when the dependent and independent variables share linear relationships (Wendler and Gröttrup, 2016). This implies that situations where no linear relationship exists between some or all of the variables; linear regression techniques (SLR and MLR) are not suitable. The Fisher Discriminant Analysis (FDA) is similar to MLR. It produces fast, direct and concise analytical model solutions which can easily be programmed by IT personnel. It also requires few instances of a dataset to build models. The FDA is however sensitive to outliers, can't handle discrete independent variables or missing values as well as suitable only for linear phenomena (Tuffery, 2011).

After critically assessing these prediction data mining techniques that have been applied in agricultural research, we discover that poultry data works well with decision tree algorithm. This is because decision tree works well with all kinds of data (categorical and continuous data). Decision tree models are also easy to understand and interpret (this is particularly necessary if the model is to be used by local poultry farmers).Vale *et al.* (2008) has also used decision tree to predict broiler mortality rate. This research was however restricted to the impact environmental attributes (environmental temperature) have on broilers. This research did not use key attributes such as: diseases, vaccination, feed type, etc. to predict overall poultry yield.Another similar research for identifying poultry disease based on their sound has been done by Sadeghi *et al*. (2015). While this research is useful for the early detection of diseases among the poultry birds, the research did not provide procedures for predicting overall poultry yield.

III. Methodology

Research Framework

The first step of building any model is the collection of dataset. Most times, the data are inconsistent and contain errors making the data unfit for implementing the model. To resolve this, the data mining task of anomaly detection called data pre-processing is required (Tan, 2006). The data is then divided into two sets: the training data set and the validation data set. The training data set is used to build the model using the CART algorithm (regression tree) and the validation set is used to optimize the model by pruning it. The post pruning technique known as Reduced Error Pruning (REP) will be applied on the fully grown tree to reduce model overfitting and increase prediction accuracy (Mitchell, 1977). The model is then tested with the validation data set, a process referred to as cross validation. REP and cross validation form part of the pruning process. The pruned tree produces a smaller, précised prediction tree model which we propose to be the poultry prediction model. These steps have been illustrated diagrammatically in the Figure 1.

CART Algorithm

CART is an umbrella term popularized by Breiman et al., (1984) to describe the similar procedures of both classification trees and regression trees as a decision tree algorithm (Brieman et al, 1984). The CART algorithm follows a procedure called recursive partitioning algorithm that seeks to repeatedly partition a large dataset space into smaller rectangles or subsets aiming to contain as pure as possible, elements of the same class or category (Han et al., 2011; Niu, 2017).Though, classification tree and regression tree algorithms share a common decision tree name known as CART, there is a major difference between both (Aggarwal, 2015). The Classification tree is mainly used to classify categorical attributes / variables while the regression tree on the other hand is used to classify and predict continuous or numeric values (Champandard, 2003). A categorical variable can be viewed as a label or quantity used to represent a class for example: colour (red, green, blue) or age group (young, adult, elderly) and so on. Numeric/continuous variables on the other hand are numbers that can take any value (Hoffmann, 2016). Yield, the target variable to be predicted is a continuous variable. This is the reason why the regression tree algorithm of CART has been chosen to build the prediction model.

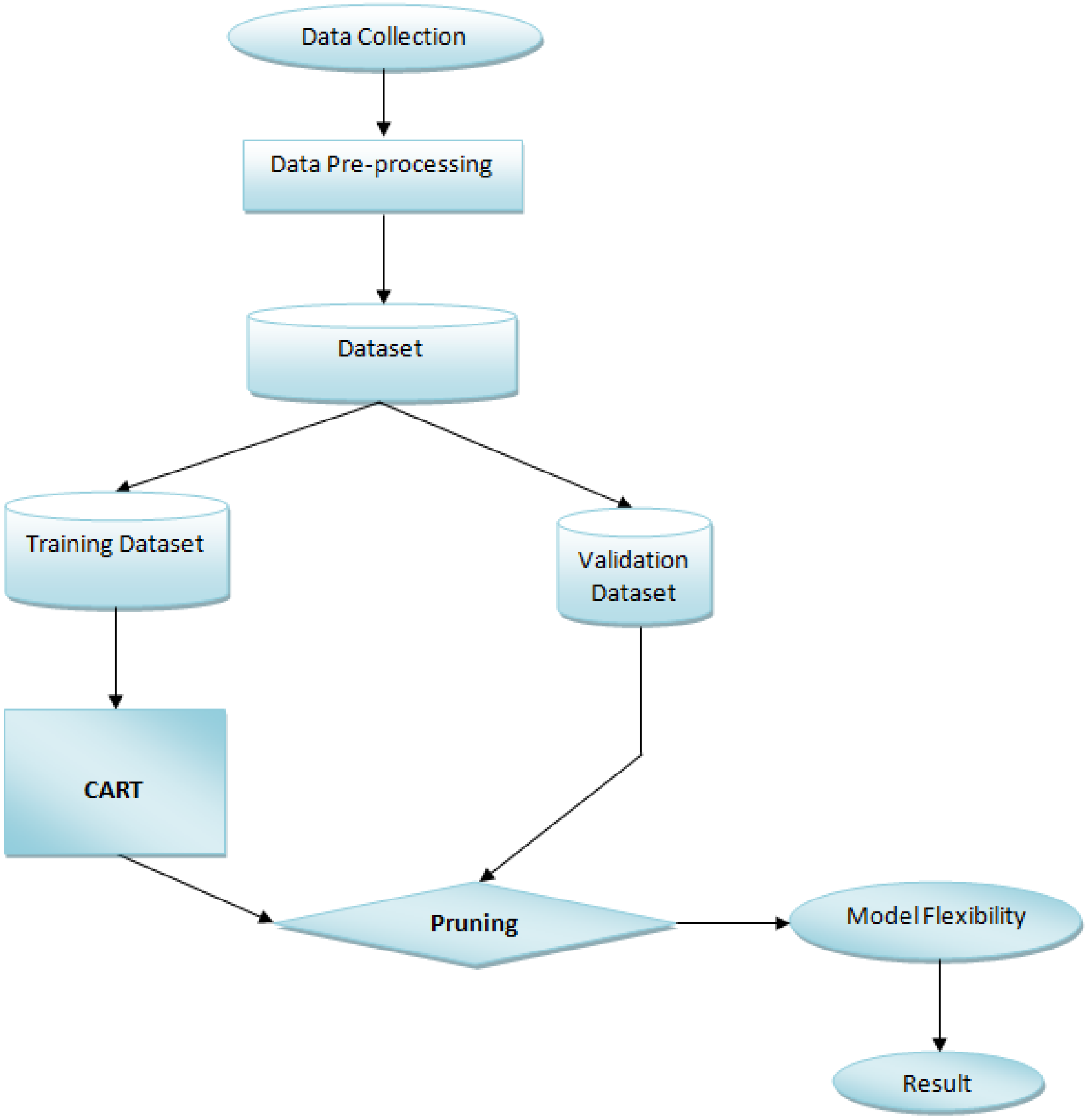


Figure 1: Conceptual framework of the research

Structure of the CART Model

The CART decision tree model consists of nodes, branches and leaves (Wolff et al., 2011; Sucar, 2011; Bhattacharyya and Kalita, 2013). The nodes represent decisions to be taken as one navigates through the tree (Rokach and Maimon, 2005). The model is built in a top-to-bottom manner (Rokach and Maimon, 2005; Wang, 2008; Han et al., 2011). The topmost decision node is called the root decision node while the terminal nodes are called leaves (Beretti et al., 2016).

The branches of the CART model represent paths leading from one decision node to another. The leaves represent the final decisions reached based on prior decision steps taken along corresponding decision paths (Tjoa and Trujillo, 2010).A locally optimized linear model (regression) is formed at the leaves which are the predicted target values (Aggarwal, 2015). The predicted value at the leaf node is usually the average of the values in a particular class after a split (Witten and Eibe, 2005).At each node, starting from the root node, the CART algorithm attempts to asks a ―yes‖ or ―no‖ (binary) question and an appropriate path is followed either left or right (splitting the node) to subsequent decision nodes down the tree (Mitchell, 1997). The same process is repeated on each node, splitting the nodes continuously (recursive partitioning) until a decision is reached at the leaves (Beretti et al., 2016). Due to the binary splitting of decision nodes in the CART decision tree, CART is essentially a binary tree (Hill et al., 2006; Aggarwal, 2014; Niu, 2017). Figure 2 illustrates the structure of CART.

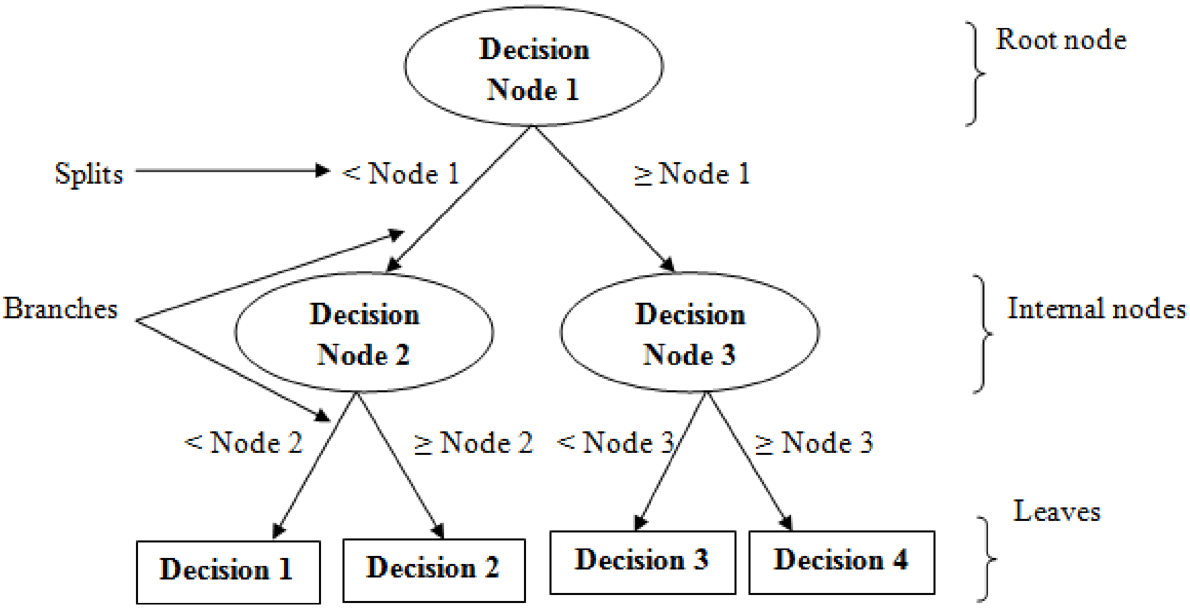


Figure 2: structure of CART

Building the CART Model

Just like every other tree induction algorithm, building a CART model requires some criteria –the splitting criteria, the stopping criteria and pruning criteria (Aggarwal, 2014; Aggarwal, 2015).

Splitting Criteria

To be able to classify similar data at various points in a dataset space, a criterion to determine what attributes of the data to split and the particular point at which the splitting should occur is necessary. That is where the splitting criteria come to play. The splitting criteria of a CART is the measure use to determine the best variable to split as well as the most appropriate points to split the variable so as to achieve classification purity (Diday et al, 2013). Purity in this case is the measure of the homogeneity of elements or attribute in a particular class (Witten and Eibe, 2005; Aggarwal, 2015). If a particular class/node is said to be 100% pure, it means that the class/node consist of 100% similar elements with no error or outlier (dissimilar element/attribute). To achieve pure classification splits, for classification trees, some splitting criteria have been proposed such as Entropy, Gini index and Twoing (Mitchell, 1997; Wu and Kumar, 2009; Issac and Israr, 2014). For a regression tree model however, a splitting criteria that involves an error based measure or measure of variance is considered more appropriate because of the continuous numerical implication of the attributes of the target variable (Witten and Frank, 2005; Aggarwal, 2015). To classify variables or attributes with respect to a target numeric continuous variable, a locally optimised linear model is obtained from each hierarchical partitioning of the decision node at the leaves of the tree (Aggarwal, 2015). To obtain a true representation of the value of every split, the average of all the values of the split is computed and used (Witten and Frank, 2005; Moolayil, 2016). This is indicated at the leaves of the tree and along decision paths along the tree. One common variant measurement splitting criteria for a regression tree is the Standard Deviation Reduction (SDR) measure (Witten and Frank, 2005; Moolayil, 2016).

Pruning Criteria

Pruning a tree requires cutting off branches from the tree so as to improve accuracy and reduce overfitting (Mitchell, 1997; Witten and Eibe, 2005). Pruning is a way of making complex and large trees simpler and precise. This is in accordance to Occam‟s razor theory which states that a simpler and less complex a model is, the more accurate it is (Hall et al., 2011).Pruning techniques/criteria that involves the use of a validation dataset are called post pruning techniques. Post pruning requires that a tree model be fully grown from top to bottom and then pruned bottom to top (Aggarwal, 2015). This pruning technique is quite different from the pre pruning technique which requires that the tree be stopped early enough before it begins to over fit (Mitchell, 1997). The problem with pre pruning however is that there is the uncertainty of the ‗early point‟ to stop the tree growth (Aggarwal, 2015). Mitchell, (1997) also suggested that growing the tree fully is the most practical approach for tree induction models. For this reason, we decided to use a post pruning technique. Some post-pruning criteria include cost complexity pruning, reduced-error pruning and rule-based pruning (Mitchell, 1997).

Regression Tree Model for Poultry Prediction

For the poultry prediction model, the SDR measure as prescribed by Witten and Frank (2005) will be used. The splitting process continues until no further splitting is feasible (when partitions are as pure as possible). Though this stopping criterion will result to a large tree, it is however the most pragmatic criterion for any tree induction model (Mitchell, 1997). The resulting tree will be pruned using the REP criterion/technique to optimise a regression tree model that will predict the target variable *yield* using the predictor variables *vaccine*, *season, feed*, and *disease* all of which will form decision nodes of the regression tree model.

Training and Validation Dataset

For the purpose of developing algorithms and models for machine learning, a training dataset and validation datasets are required (Hall et al., 2011).CART algorithms generally require a significantly large amount of training dataset (Aggarwal, 2015). Though there isn‟t any specified percentage of dataset to be set aside as training dataset, certain literatures suggest over 50% of the total dataset. For the purpose of this research, we decide to utilize the first 11-breeding period (55% of the total dataset) for our training dataset while the remaining 9 breeding periods will be used as the validation dataset to prune the regression tree and validate the model as shown in Table 1 and 2.

Table 1: Training dataset

Breeding Period (2006)	Vaccine Administered	Disease Breakout	Season	Feed Type	Yield
1	not enough	high	dry	Low	356
2	Enough	low	dry	Low	352
3	Enough	low	rainy	High	390
4	Enough	low	rainy	Low	384
5	not enough	low	rainy	High	380
6	Enough	low	dry	Low	375
7	not enough	high	rainy	Low	347
8	not enough	high	dry	High	365
9	enough	low	dry	High	375
10	not enough	high	rainy	Low	345
11	enough	low	rainy	High	400

Breeding Period (2006)	Vaccine Administered	Disease Breakout	Season	Feed Type	Yield
12	Enough	High	Rainy	Low fat	387
13	Enough	Low	Dry	High fat	383
14	Not enough	High	Dry	Low fat	350
15	Enough	Low	Dry	Low fat	365
16	Not enough	High	Rainy	High fat	346
17	Not enough	Low	Rainy	High fat	372
18	Not enough	High	Dry	Low fat	347
19	Enough	Low	Dry	High fat	387
20	Enough	Low	Dry	High fat	384

Building the Regression Tree Model

Building a regression tree model using SDR splitting criterion is summarised into the following algorithm/ steps.

Step 1: Calculate the standard deviation of target variable

Step 2: Separate attributes of each predictor variable of the dataset

Step 3: Calculate the standard deviation of variables based on their attributes

Step 4: The standard deviation of target variable before separating predictor variable is separated from resulting standard deviation from step 3 after separating predictor variables

- The result from step 4 is the Standard Deviation Reduction

Step 5: Select variable with the largest/highest SDR as decision node

Step 6: The attributes of selected variable from step 5 is separated

Step 7: Based on the separated attributes of selected variable from 5, calculate SD of attribute sets.

- Attribute set with SD > 0 is split further (go to step 3)

Step 8: Repeat process recursively until all non-leaf variables (decision nodes) are processed

Step 9: For final processed variables with more than one attribute leading to leaf nodes, calculate the average as the final value for the leaf node (target value).

Reduced-Error Pruning (REP)

The reduced-error pruning (by Quinlan 1987) is a post pruning technique done bottom to top (Mitchell, 1997). This technique views every decision node in the tree as a pruning candidate. It involves replacing a set of decision nodes with the most common classification and assigning it to affiliate leafs

(Mitchell, 1997). The replacement is done only if the resulting pruned tree supports the validation dataset. This is because classifications irregularities that may occur with the training dataset are unlikely to occur with the validation dataset (Mitchell, 1997). The reduced-error pruning technique is been used in this research because of its simplicity and speed (Mitchell, 1997).

REP algorithm/steps are given as:

Step 1: Break full tree into sub trees

Step 2: Prune each sub tree by replacing the decision node with the most common decision node to form a pruned tree

Step 3: Test pruned tree against validation dataset

Step 4: Select pruned sub tree with the least classification error.

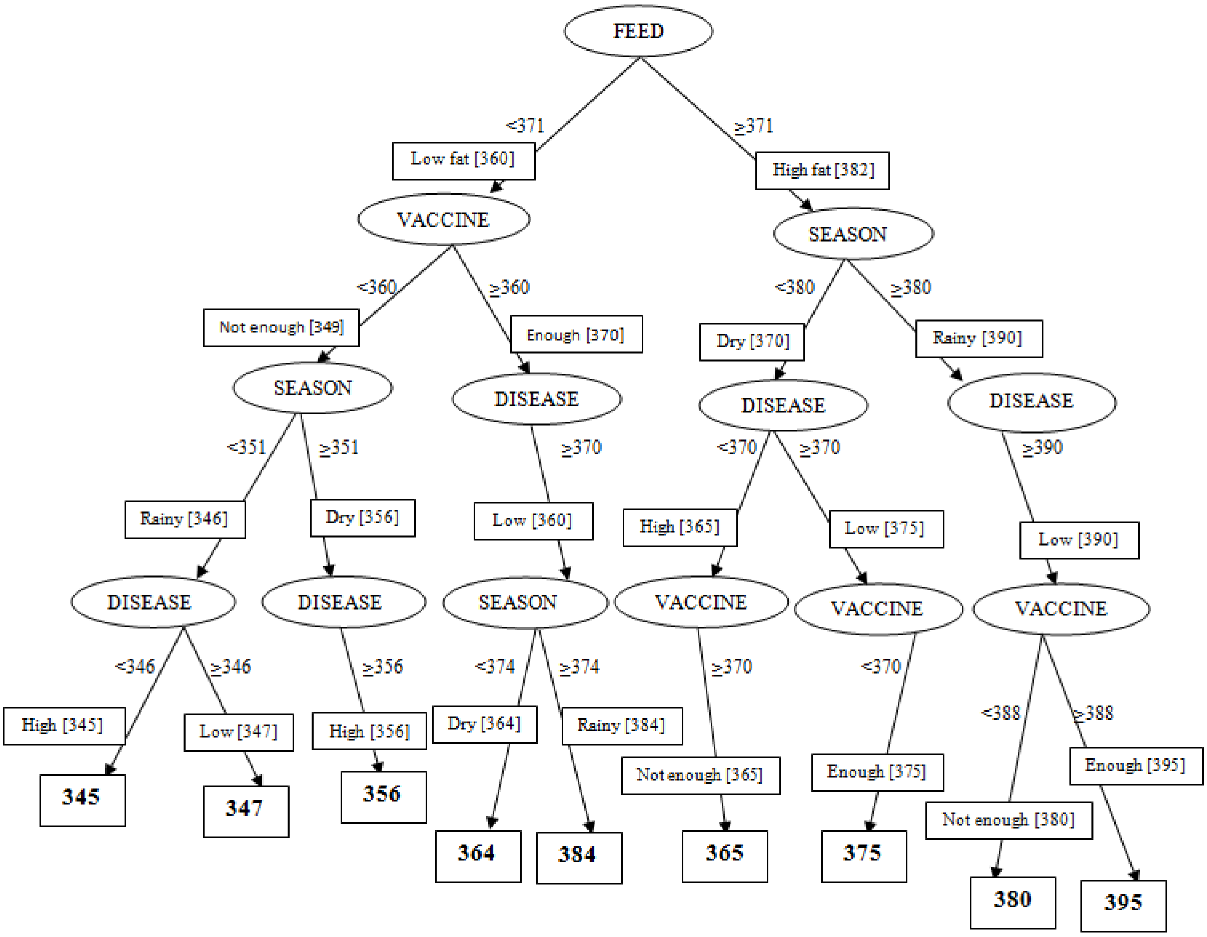
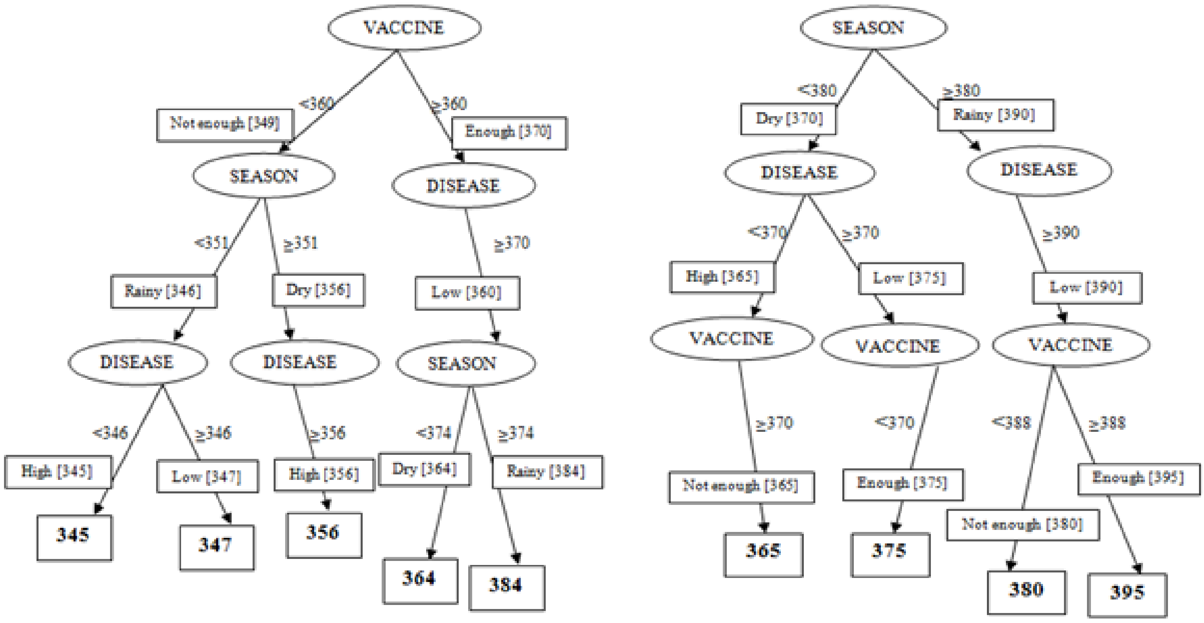


Figure 3: Complete regression tree

Pruning the Regression Tree

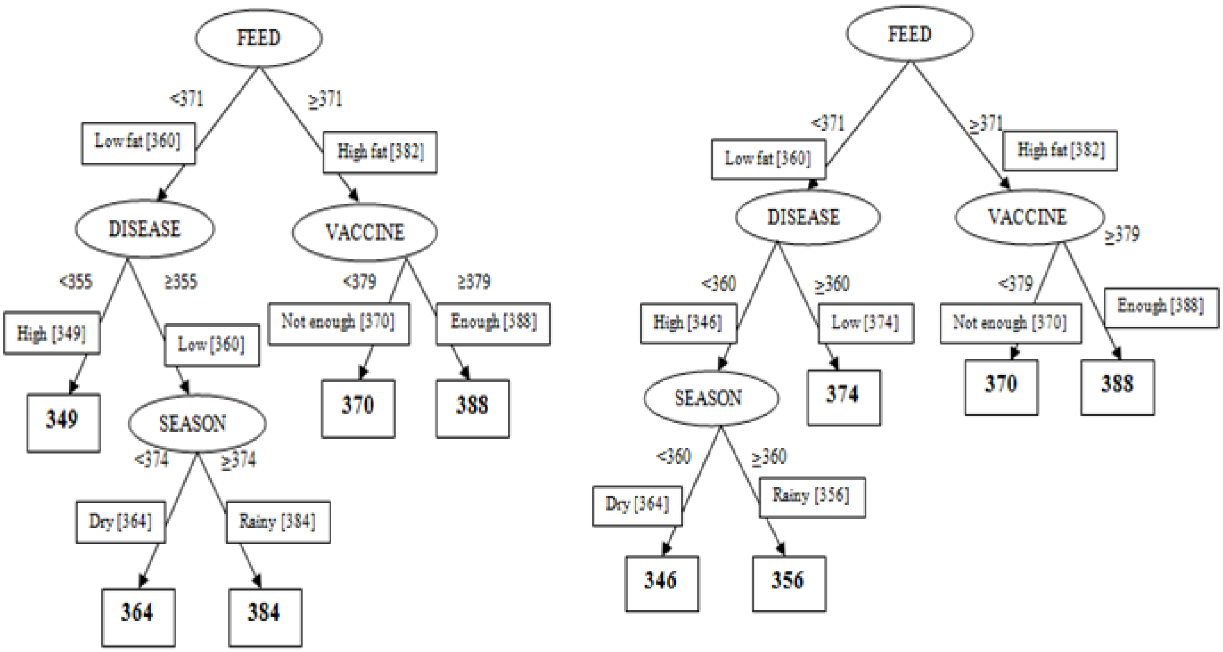
Pruning is required to reduce overfitting. Pruning is also a concept supported by Occam‘s theory which states that the smaller a model, the more accurate it is (Hall et al., 2011). We begin the REP technique as prescribe by Mitchell (1997) by dividing the tree into two sub trees: sub tree A and sub tree B as shown in Figure 4.1. Two pruned regression tree have been obtained (see Figures 4.2).



Sub tree A

Sub tree B

Figure 4.1: Sub trees A and B



Pruned tree A

Pruned tree B

Figure 4.2: Pruned tree A and B

Cross Validation

The variables of the validation data set have been rearranged in the same pattern as the pruned tree. Misclassified classes of pruned trees A and B have been indicated with bold italics as shown in Table 2 and Table 3 respectively.

Table 2: Misclassification table of pruned tree A

Breeding Period (2006)	Feed Type	Vaccine Administered	Disease Breakout	Season	Yield
13	High fat	Enough	Low	Dry	383
16	***High fat***	***Not enough***	***High***	***Rainy***	***346***
17	High fat	Not enough	Low	Rainy	372
19	High fat	Enough	Low	Dry	387
20	High fat	Enough	Low	Dry	384
12	***Low fat***	***Enough***	***High***	***Rainy***	***387***
14	Low fat	Not enough	High	Dry	350
15	***Low fat***	***Enough***	***Low***	***Dry***	***365***
18	Low fat	Not enough	High	Dry	347

Table 3: Misclassification table of pruned tree B

Breeding Period (2006)	Feed Type	Vaccine Administered	Disease Breakout	Season	Yield
13	High fat	Enough	Low	Dry	383
16	***High fat***	***Not enough***	***High***	***Rainy***	***346***
17	High fat	Not enough	Low	Rainy	372
19	High fat	Enough	Low	Dry	387
20	High fat	Enough	Low	Dry	384
12	***Low fat***	***Enough***	***High***	***Rainy***	***387***
14	Low fat	Not enough	High	Dry	350
15	Low fat	Enough	Low	Dry	365
18	Low fat	Not enough	High	Dry	347

We propose that pruned tree B be our selected model for predicting poultry yield. Pruned tree B has been selected because of it contains less classification errors (22%, indicated in bold italics) compared to pruned tree A (33%, also indicated in bold italics).

Applying the Prediction Model to Predict Poultry Yield

The main objective of this research is to develop a model that poultry farmers can use to predict poultry yield. A regression tree model has been developed in that respect. However, it is necessary to present this model in such a way that the local poultry farmers irrespective of the population size of their respective poultry farms can apply and utilize.

The CART algorithm has been applied on a sample size of 400 poultry birds to demonstrate how we can develop a prediction model. As a result, the predicted yields have been with respect to a population size of 400 poultry birds. The issue of a non-flexible model therefore arises. We present a simple algorithm to modify the developed model such that the model makes percentile prediction. This way, predictions can be achieved by simply multiplying the percentile prediction with whatever poultry sample size. The algorithm has been presented below.

Algorithm

Poultry percentile prediction model, regression tree, N

```
N = sample population

If Feed = Feed $_{\text{high fat}}$ then

    Vaccine $_{\text{not enough}}$ = $\frac{[\frac{370*100}{400}]}{100}$ * N
    Vaccine $_{\text{enough}}$ = $\frac{[\frac{388*100}{400}]}{100}$ * N
Else if
Feed = Feed $_{\text{low fat}}$ then

        Disease $_{\text{low}}$ = $\frac{[\frac{374*100}{400}]}{100}$ * N

Else if

        Disease = Disease $_{\text{high}}$ then

        Season $_{\text{dry}}$ = $\frac{[\frac{346*100}{400}]}{100}$ * N

        Season $_{\text{rainy}}$ = $\frac{[\frac{356*100}{400}]}{100}$ * N

    End if

End.
```

The output is shown in Figure 5.

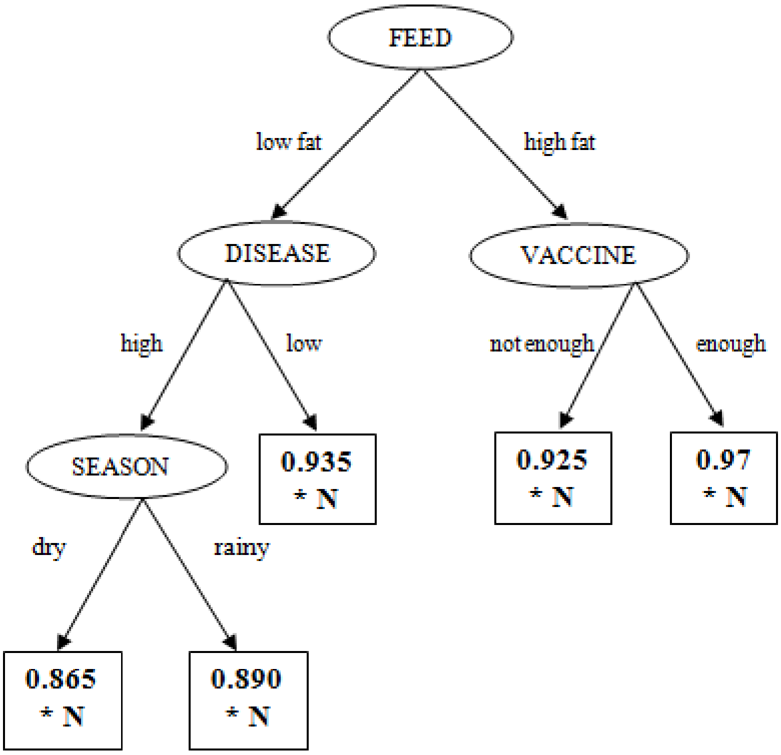


Figure 5: Percentile prediction model

IV. Conclusion

Data mining techniques generally unravel hidden patterns in data. Knowledge can be discovered from this hidden pattern. Poultry data have been collected and mined in this research and patterns which can result in yield prediction have been discovered using regression tree of the CART algorithm. To achieve this, we employed the SDR technique to hierarchically split the data rather than other splitting techniques like Gini index and entropy because of the numerical and continuous implication of the target variable ‗Yield'.To avoid model over fitting and improve accuracy of the model, a post pruning technique called REP have been used. In line with post pruning techniques, a validation data set was set aside to test the performance of two pruned model trees. The model tree that performed better with the validation data set was chosen as our proposed prediction model.To make the proposed model flexible, we presented another algorithm that converts predictions into percentiles based on the predictions of the proposed model. This algorithm makes prediction for whatever poultry population by multiplying the resulting predictions at the leaf nodes with the poultry population (N). CART algorithms have been applied for prediction purposes with high prediction accuracy. This can largely be attributed to the fact that CART is a machine learning algorithm that is well grounded in rigorous statistics and probability theory (Wu and Kumar, 2009). A CART model for predicting poultry yield has been developed in this study and it has been pruned to provide optimal results.

REFERENCES

[1.] Abe, S. (2005). *Support vector machines for pattern classification* (Vol. 53). London: Springer.

[2.] Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

[3.] Aggarwal, C. C. (Ed.). (2014). *Data classification: algorithms and applications*. CRC Press.

[4.] Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering. *Algorithms and Applications, Chapman & Halls*.

[5.] Azzalini, A., &Scarpa, B. (2012). *Dataanalysis and data mining: An introduction.* OUP USA.

[6.] Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education*, *7*(3), 112-118.

[7.] Banks, D., House, L., McMorris, F. R., Arabie, P., & Gaul, W. A. (Eds.). (2011). *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*. Springer Science & Business Media.

[8.] Berretti, S., Thampi, S. M., &Dasgupta, S. (Eds.). (2016). *Intelligent systems technologies and applications*. Springer International Publishing.

[9.] Bhattacharyya, D. K., &Kalita, J. K. (2013). *Network anomaly detection: A machine learning perspective*. CRC Press.

[10.] Bozdogan, H. (Ed.). (2003). *Statistical data mining and knowledge discovery*. CRC Press.

[11.] Braspenning, P. J., &Thuijsman, F. (1995). *Artificial neural networks: an introduction to ANN theory and practice* (Vol. 931). Springer Science & Business Media.

[12.] Breiman, L., Friedman, J., Stone, C. J., &Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

[13.] Campilho, A., &Kamel, M. (Eds.). (2014). *Image Analysis and Recognition: 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22-24, 2014, Proceedings* (Vol. 8814). Springer.

[14.] Champandard, A. J. (2003). *AI game development: Synthetic creatures with learning and reactive behaviors*. New Riders.

[15.] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 15.

[16.] Cherkassky, V., &Mulier, F. M. (2007). *Learning from data: concepts, theory, and methods*. John Wiley & Sons.

[17.] Chowdhury, D. R., &Ojha, S. (2017). ―An Empirical Study on Mushroom Disease Diagnosis: A Data Mining Approach‖, *International Research Journal of Engineering and Technology,* 4(1), 529-534.

[18.] deAlbornoz, A. G. Á., &Terashima-Marín, H. MICAI 2005: Advances in Artificial Intelligence.

[19.] deSá, J. P. M., Silva, L. M., Santos, J. M., & Alexandre, L. A. (2013). *Minimum error entropy classification*. Springer.

[20.] Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., &Burtschy, B. (Eds.). (2013). *New approaches in classification and data analysis*. Springer Science & Business Media.

[21.] Digby, B. (2001). *It's a World Thing*. Oxford University Press, USA.

[22.] El Fangary, L. M. (2009, December). Mining Data of Buffalo and Cow Production in Egypt. In *Frontier of Computer Science and Technology, 2009. FCST'09. Fourth International Conference on* (pp. 382-387). IEEE.

[23.] Elder, J. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

[24.] Ellen, S. (2012). Slovin's Formula Sampling Techniques.

[25.] Fagerlund, S. (2007). *Bird Species Recognition Using Support Vector Machines*, EURASIP Journal on Advances in Signal Processing 2007, Article ID 38637, 1–8.

[26.] Fasina, F. O., Wai, M. D., Mohammed, S. N., &Onyekonwu, O. N. (2007). Contribution of poultry production to household income: a case of Jos South Local Government in Nigeria. *Family Poultry*, *17*(1&2), 30-34.

[27.] Gill, M. K., Asefa, T., Kemblowski, M. W., & McKee, M. (2006). Soil moisture prediction using support vector machines. *JAWRA Journal of the American Water Resources Association*, *42*(4), 1033-1046.

[28.] Hall, M., Witten, I., & Frank, E. (2011). Data mining: Practical machine learning tools and techniques. *Kaufmann, Burlington*.

[29.] Han, J., Jian, P., & Michelin, K. (2006). Data Mining, Southeast Asia Edition.

[30.] Han, J., Pei, J., &Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[31.] Heise, H., Crisan, A., &Theuvsen, L. (2015). The poultry market in Nigeria: market structures and potential for investment in the market. *International Food and Agribusiness Management Review*, *18*, 197-222.

[32.] Herremans, E., Melado-Herreros, A., Defraeye, T., Verlinden, B., Hertog, M., Verboven, P., ...&Wevers, M. (2014). Comparison of X-ray CT and MRI of watercore disorder of different apple cultivars. *Postharvest biology and technology*, *87*, 42-50.

[33.] Hill, T., Lewicki, P., &Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. StatSoft, Inc..

[34.] Hoffmann, J. P. (2016). *Regression Models for Categorical, Count, and Related Variables: An Applied Approach*. Univ of California Press.

[35.] Issac, B., &Israr, N. (Eds.). (2014). *Case Studies in Secure Computing: Achievements and Trends*. CRC Press.

[36.] Jones, J. W., Tsuji, G. Y., Hoogenboom, G., Hunt, L. A., Thornton, P. K., Wilkens, P. W., ... & Singh, U. (1998). Decision support system for agrotechnology transfer: DSSAT v3. In *Understanding options for agricultural production* (pp. 157-177). Springer Netherlands.

[37.] Jones, M. T. (2015). *Artificial Intelligence: A Systems Approach: A Systems Approach*. Jones & Bartlett Learning.

[38.] Klösgen, W., &Zytkow, J. M. (2002). *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc.

[39.] Kondo, N., Ahmad, U., Monta, M., &Murase, H. (2000). Machine vision based quality evaluation of Iyokan orange fruit using neural networks. *Computers and electronics in agriculture*, *29*(1), 135-147.

[40.] Kotsiantis, S. B., Zaharakis, I., &Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

[41.] Larson, N. I., Story, M. T., & Nelson, M. C. (2009). Neighborhood environments: disparities in access to healthy foods in the US. *American journal of preventive medicine*, *36*(1), 74-81.

[42.] Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, *45*(37), 870-877.

[43.] Moolayil, J. (2016). *Smarter Decisions–The Intersection of Internet of Things and Decision Science*. Packt Publishing Ltd.

[44.] Moshou, D., Chedad, A., Van Hirtum, A., De Baerdemaeker, J., Berckmans, D., & Ramon, H. (2001). An intelligent alarm for early detection of swine epidemics based on neural networks. *Transactions of the ASAE*, *44*(1), 167.

[45.] Moshou, D., Chedad, A., Van Hirtum, A., De Baerdemaeker, J., Berckmans, D., & Ramon, H. (2001). Neural recognition system for swine cough. *Mathematics and Computers in Simulation*, *56*(4), 475-487.

[46.] Mucherino, A., Papajorgji, P. J., &Pardalos, P. M. (2009). *Data mining in agriculture* (Vol. 34). Springer Science & Business Media.

[47.] Niu, G. (2017). *Data-Driven Technology for Engineering Systems Health Management*. Springer.

[48.] Patan, K. (2008). *Artificial neural networks for the modelling and fault diagnosis of technical processes*. Springer.

[49.] Piatetsky-Shapiro, G., & Parker, G. (2011). Lesson: Data mining, and knowledge discovery: An introduction. *Introduction to Data Mining, KD Nuggets*.

[50.] Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, *27*(3), 221-234.

[51.] Rajeswari, V., &Arunesh, K. (2016). Analysing soil data using data mining classification techniques. *Indian Journal of Science and Technology*, *9*(19).

[52.] Ramesh, D., &Vardhan, B. V. (2013). Data mining techniques and applications to agricultural yield data. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(9), 3477-80.

[53.] Ramesh, D., &Vardhan, B. V. (2015). Analysis of crop yield prediction using data mining techniques. *International Journal of Research in Engineering and Technology*, *4*(1), 47-473.

[54.] Raut, A. B., &Nichat, M. A. A. (2017). Students Performance Prediction Using Decision Tree. *International Journal of Computational Intelligence Research*, *13*(7), 1735-1741.

[55.] Rokach, L., &Maimon, O. (2005). The Data Mining and Knowledge Discovery Handbook: A Complete Guide for Researchers and Practitioners.

[56.] Rokach, L., &Maimon, O. (2008). *Data mining with decision trees: theory and applications*.

[57.] Rokach, L., &Maimon, O. (2014). *Data mining with decision trees: theory and applications*. World scientific.

[58.] Ryan, T. P. (2013). *Sample size determination and power*. John Wiley & Sons.

[59.] Sadeghi, M., Banakar, A., Khazaee, M., &Soleimani, M. R. (2015). An Intelligent Procedure for the Detection and Classification of Chickens Infected by Clostridium Perfringens Based on their Vocalization. *RevistaBrasileira de CiênciaAvícola*, *17*(4), 537-544.

[60.] Shahin, M. A., Tollner, E. W., & McClendon, R. W. (2001). AE—Automation and Emerging Technologies: Artificial Intelligence Classifiers for sorting Apples based on Watercore. *Journal of agricultural engineering research*, *79*(3), 265-274.

[61.] Siau, K. (Ed.). (2008). *Advanced Principles for Improving Database Design, Systems Modeling, and Software Development*. IGI Global.

[62.] Stockle, C. O., Martin, S. A., & Campbell, G. S. (1994). CropSyst, a cropping systems simulation model: water/nitrogen budgets and crop yield. *Agricultural Systems*, *46*(3), 335-359.

[63.] Sucar, L. E. (Ed.). (2011). *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions: Concepts and Solutions*. IGI Global.

[64.] Sumathi, S., &Sivanandam, S. N. (2006). *Introduction to data mining and its applications* (Vol. 29). Springer.

[65.] Tan, P. N. (2006). *Introduction to data mining*. Pearson Education India.

[66.] Thuraisingham, B. (1998). *Data mining: technologies, techniques, tools, and trends*. CRC press.

[67.] Tiwari, V., Tiwari, B., Thakur, R. S., & Gupta, S. (2013). Pattern and data analysis in healthcare settings.

[68.] Tjoa, A. M., & Trujillo, J. (2010). *Data Warehousing and Knowledge Discovery*. Springer Berlin/Heidelberg.

[69.] Tuffery, S. (2011). *Data mining and statistics for decision making* (Vol. 2). Chichester: Wiley.

[70.] Urtubia, A., Pérez-Correa, J. R., Soto, A., &Pszczolkowski, P. (2007). Using data mining techniques to predict industrial wine problem fermentations. *Food Control*, *18*(12), 1512-1517.

[71.] Vale, M. M., Moura, D. J. D., Nääs, I. D. A., Oliveira, S. R. D. M., & Rodrigues, L. H. A. (2008). Data mining to estimate broiler mortality when exposed to heat wave. *Scientia Agricola*, *65*(3), 223-229.

[72.] Wang, J. (Ed.). (2008). *Data warehousing and mining: Concepts, methodologies, tools, and applications: Concepts, methodologies, tools, and applications* (Vol. 3). IGI Global.

[73.] Wang, J. (Ed). (2014). *Encyclopedia of Business Analytics and optimization.* IGI Global.

[74.] Wendler, T., &Gröttrup, S. (2016). *Data Mining with SPSS Modeler: Theory, Exercises and Solutions*. Springer.

[75.] Witten, I. H., Frank E. (2005). *Data Mining: Practical machine learning tools and techniques*, *2*, 127-143.

[76.] Wolff, K. E., Palchunov, D. E., Zagoruiko, N. G., &Andelfinger, U. (Eds.). (2011). *Knowledge Processing and Data Analysis: First International Conference, KONT 2007, Novosibirsk, Russia, September 14-16, 2007, and First International Conference, KPP 2007, Darmstadt, Germany, September 28-30, 2007. Revised Selected Papers* (Vol. 6581). Springer Science & Business Media.

[77.] Wu, X., & Kumar, V. (2009). The top ten algorithm in data mining. *International Standard Book*, *13*, 978-1.

[78.] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ...& Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37