

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Using Multiple-Instance Learning Techniques to Rank Maize Ears According to Their Traits

Karamot Kehinde Biliaminu



Integrated Master in Informatics and Computing Engineering

Supervisor: Prof. João Pedro Mendes Moreira

Co-Supervisor: Prof. Pedro Mendes Moreira

June, 2017

Using Multiple-Instance Learning Techniques to Rank Maize Ears According to Their Traits

Karamot Kehinde Biliaminu

Integrated Master in Informatics and Computing Engineering

June, 2017

Abstract

Multiple-Instance Learning (MIL) is a sub-field of machine learning. Its main goal is to do accurate predictions on new data based on a predictive model generated from previously group of labeled bags of data, known as training data, containing many instances. MIL has many real world important applications such as image retrieval or text categorization and medical diagnosis problems.

It is often difficult for crop breeders to predict yield by combining different yield components to produce better plants with superior performance. Data analysis is one area that is striving to let farmers have an idea of their expected yield preharvest. Accurate early yield prediction will improve agricultural strategies plan, proper resources allocation and improve management of maize ear cultivation with consequent increase in productivity. Most experiments on maize ears traits only considered ear evaluation and maize improvement without yield prediction. One of the experiments that included yield prediction was PR. NDCG measure which was developed to rank maize evaluation for Sousa Valley Best Ear Competition.

The focus of this work was to make an intelligent regression models recognition and analysis by running some MIL algorithms to predict and assign real value to maize yield from randomly group N vary parameter sizes of maize ear traits and soil parameters of maize population dataset. Furthermore, this dissertation also ranked the models per result and establish a relationship between variables.

Acknowledgements

Special thanks to Almighty God for the gift of life and for seeing me through my wonderful two years stay here.

I would like to appreciate my supervisor, Prof. Joao Mendes Moreira, without whom this work would not have been successful for his patience, understanding and always been there to offer clear and insightful advice. His professionalism and clarity of thought were important throughout this project.

I'm grateful and will always be to European Commissions for the opportunity given to study in this great university, especially at Faculty of Engineering. I appreciate all the lecturers in the department for the knowledge they imparted in me.

To my friends, Paula Fortuna, Miguel Sandim, Valter Silva, Sara Silva, Sara Paiva and Boris for their support and encouragement.

I would also like to thank my family, for their inspiration, care and support. I love you guys.

“The best way to predict the future is to create it.”

Peter Drucker

Contents

1.	Introduction	1
1.1.	Context.....	1
1.2.	Motivations and Goals	2
1.3.	Multiple Instance Regression	2
1.4.	Report Structure	3
2.	MIL Concepts.....	4
2.1.	Application Cases.....	4
2.1.1.	Drug Activity Prediction.	4
2.1.2.	Image Retrieval.	4
2.1.3.	Text Categorization (Andrews et al. (2002))	5
2.2.	Predictive Analytics.....	6
2.2.1.	Regression Analysis.....	7
2.3.	Learning Algorithms	7
2.3.1.	Linear regression.....	7
2.3.2.	Least Absolute Shrinkage and Selection Operator (Lasso)	8
2.3.3.	Multivariate Adaptive Regression Splines (MARS)	8
2.3.4.	K Nearest Neighbor(K-NN)	8
2.3.5.	Artificial neural networks.....	9
2.3.6.	Support Vector Machines (SVMs).....	10
2.3.7.	Random forest	10
2.4.	Application of Predictive Analytics on Yield Prediction.....	10
2.5.	Summary	11
3.	Dataset and Methodology	12
3.1.	Dataset	12
3.2.	Data Description	13
3.3.	Task	18
3.4.	Methodology.....	18
3.5.	Summary	19
4.	Experimental Setup	19
4.1	Assign predicted value to a bag after averaging each instance in the bag	19
4.1.1	Recursive Partitioning (Rpart)	19

4.1.2.	Multivariate Adaptive Regression Splines (MARS)	21
4.1.3.	Least Absolute Shrinkage and Selection Operator (LASSO)	23
4.1.4.	Random Forest	24
4.1.5.	K-Nearest Neighbor	25
4.2.	Represent a bag with the average predicted value of observations	27
4.2.1.	Rpart	28
4.2.2.	MARS	28
4.2.3.	LASSO	29
4.2.4.	Random Forest	30
4.2.5.	KNN	31
4.3.	Comparison of Results	32
4.3.1.	Rpart	32
4.3.2.	MARS	34
4.3.3.	LASSO	35
4.3.4.	Random Forest	37
4.3.5.	KNN	38
5	Conclusion.....	45
5.1.	Future work	45

References

List of Figures

1.1	Multiple Instance Learning Process (Dietterich, Lathrop, and LozanoPerez 1997).....	1
2.1	Predictive Analytics Evolution	6
2.2	Neural Network Process (Hastie, Tibshirani, and Friedman 2009)	10
3.1	World corn production 2016/2017	12
3.2	CRISP-DM Data mining life cycle (Wikipedia 2016)	18
4.1	Rpart residual plot	21
4.2	Rpart predicted values plot	22
4.3	MARS scatter plot and histogram of residual error	23
4.4	Plot and histogram of MARS predicted values	24
4.5	LASSO plot and histogram of residual error	25
4.6	Plot and histogram of LASSO predicted values	26
4.7	Scatter plot and histogram of RF model predicted values	27
4.8	Plot and histogram of KNN predicted values	28
4.9	Boxplot of all the models	29
4.10	Rpart plot for the second MIL approach	30
4.11	MARS plot for the second MIL approach	31
4.12	LASSO plot for the second MIL approach	32
4.13	RF plot for the second MIL approach	33
4.14	KNN plot for the second MIL approach	34
4.15	Boxplot of rpart prediction of both MIL approaches and actual value	35
4.16	Rpart comparison histogram	35
4.17	MARS comparison box plot	37
4.18	MARS comparison histogram	37
4.19	LASSO comparison box plot	39
4.20	LASSO comparison histogram	39
4.21	RF comparison box plot	40
4.22	RF comparison histogram	41
4.23	KNN comparison box plot	42
4.24	KNN comparison box histogram	43

List of Tables

2.1	MIL Application problem characteristics (Marc-Andr�e Carbonneau 2016).....	5
2.2	Learning Algorithms Characteristics (Hastie, Tibshirani et al.)	11
3.1	First sample field values overview of the first 29 records with some traits	14
3.2	Second sample field values overview of the first 29 records with some traits	16
3.3	Dataset traits description	17
4.1	Rpart summary for the first MIL approach	21
4.2	Results for the Rpart model the first MIL approach	21
4.3	MARS summary for the first MIL approach	22
4.4	Results for the MARS model for the first MIL approach	23
4.5	LASSO summary for the first MIL approach	24
4.6	Results for the LASSO model the first MIL approach	24
4.7	RF model results for the first MIL approach	26
4.8	Random Forest summary for the first MIL approach	26
4.9	Results for the KNN model the first MIL approach	27
4.10	KNN summary for the first MIL approach	27
4.11	Models Results summary	28
4.12	Rpart result for the second MIL approach	30
4.13	MARS result for the second MIL approach	31
4.14	LASSO results for the second MIL approach	32
4.15	RF result for the second MIL approach	32
4.16	KNN result for the second MIL approach	33
4.17	Rpart model results of both MIL approaches	34
4.18	MARS model results of both MIL approaches	36
4.19	LASSO model results of both MIL approaches	38
4.20	RF model results of both MIL approaches	40
4.21	KNN model results of both MIL approaches	42
4.22	Summary of all models' predictive results on both approaches and actual value	44

Abbreviations

APRS	Axis Parallel Rectangles
Caret	Classification and Regression Training
CERES	Crop-Environment Resource Synthesis
CRISP-DM	Cross Industry Standard Process for Data Mining
DD	Diverse Density
KNN	K-Nearest Neighbor
LASSO	Least Absolute Shrinkage and Select Operators
MARS	Multivariate Adaptive Regression Splines
MIL	Multiple-Instance Learning
MIR	Multiple-Instance Regression
RF	Random Forest
Rpart	Recursive Partitioning
SVMS	Support Vector Machines

1. Introduction

Supervised learning is a machine learning/data mining methodology which requires training data set that consists of two variables set, inputs and outputs (Babenko 2008). Inputs are measured, used in predicting outputs and always have influence on outputs.

Multiple-Instance Learning (MIL) is a variation of supervised machine learning for solving problems with incomplete knowledge. It has received much attention recently due to its representation of various real-world problems and has various applications, ranging from drug activity prediction to algae, medical diagnosis and stock market predictions. Any Multiple-Instance problem must establish a relationship between instances and bag-level class that contain them before applying MIL algorithms due to the influence this has on algorithms performance (James and Eibe 2004).

MIL assumptions can be grouped into two major categories: standard assumption and collective assumption. Standard assumption follows hereditary polymorphic which assumes a data set bags of instances is negative when it contains only negative instances and positive when it contains at least one positive instance while collective assumption states considering more than one instance in the dataset bags to define bag labels (Marc-Andre et al. 2016).

The goal of MIL is either to learn a classifier which assigns correct labels to individual instances or to accurately predict the labels of the bags without inducing labels of each individual instance (James and Eibe 2004; Pappas and Popescu-belis 2017).

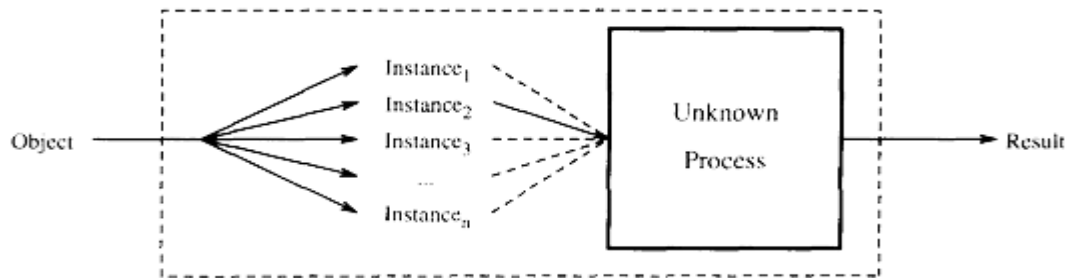


Figure 1.1: Multiple Instance Learning Process (Dietterich, Lathrop, and LozanoPerez 1997)

1.1. Context

Having the ability to predict yield would be of great benefit in achieving maximum maize yield cultivation for crop breeders, which is one of the goals of agricultural production. Prediction allows proper business planning and logistics for agricultural markets decisions and crop management strategies (Wagstaff and Roper 2008). The accuracy in predicting the responses of agronomic traits is crucial for the success of the plants (Fritsche-Neto et al. 2012).

There are three approaches to solve the problem of assigning a real value to a bag or an instance in MIL regression task. First approach which is perhaps the simplest one is to assign the value of the instance closest to target or best fit in regression to bag. Another approach is to represent a labeled bag with the average value of some instances, the last approach is to use a regressor identifier at bag level (Marc-Andre et al. 2016).

Some of the MIL algorithms for prediction will be treated later in chapter two.

For this research, an instance is a maize ear and a bag is a set of maize ears sampled from the same field for which the yield is known.

1.2. Motivations and Goals

Predicting crop is a Multiple Instance Regression problem which is a challenge that needs proper attention (Wagstaff and Roper 2008). Accuracy of yield prediction will improve the management of maize ear cultivation with a consequent increase of productivity, which results in more income for crop breeders. Maize ear kernel row arrangement, ear diameter, kernel number per ear, kernel weight per ear, kernel number per row, cob color, cob diameter and plant height are some of the factors that contribute to yield quality. Only few researches and experiments have been done regarding yield predictions, mostly were on crop evaluation (Mendes-Moreira et al. 2014). The motivation of this study is to investigate this area of task prediction problem, with the intention of better understanding the scenario to effectively create and use tools to solve learning problems in this domain where this assumption is.

One of the goals of this dissertation is to predict yield using some learning algorithms. This task is done applying two MIL approaches; In the first approach, find the average of each instances in N vary observations of a bag, assign the value to that instance then predict a value for the target. For the second approach, predict a value for each observation in N vary bag, then average the predicted values to represent the bag. In the second approach, each bag composes varied patterns of n size observations containing different instances of maize ear traits and soil characteristics. Furthermore, another objective is to do a comprehensive analysis of regression methods and of the two approaches by estimating prediction results performance using Root Mean Squared Error (RMSE). To provide a description of predictors influence on yield and to establish relationships between them is also an important step for this work. Lastly, to build a new model that allows prediction of the value of yield variable given the value of predictors is also a focus of this dissertation.

1.3. Multiple Instance Regression

Multiple Instance Regression (MIR) is an extension of single instance regression to multiple instance setting. It's a popular MIL task that aims to predict accurately a numerical outcome for every future bag based on the training bags (Herrera et al. 2016). The difficulty of MIR problems depends on its ambiguous of multiple descriptions for every bags and lack of information relating

the descriptions to the bag label (Herrera et al. 2016; Ray 1999), also on the type and variability of instances within each bag

Ray and Page pioneered the area of MIR and proposed an EM primary instance regression (PIR) method which assumes that the label of a bag is determined by exactly one primary instance, often refer to as prime instance and that the rest of the items in the bag are noisy observations (Ray 1999). The solution proposed by them was to train a linear predictor for prime instances, but it was not specified how to detect the prime instance (Wang, Lan, and Vucetic 2011) . The assumption of prime instance was replaced by (Wagstaff and Lane 2007), with assumptions that bag instances have different relevance and that bag label is a relevance-weighted average of instance-level predictions. They proposed an approximation that simultaneously determines relevant instances and trains a line predictor (Wang, Lan, and Vucetic 2011). Kiri and Wagstaff approach provides additional degrees of freedom in locating a high-quality regression fit than Ray and Page approach.

In 2008, Wagstaff et al. proposed another method which learns the internal structure of bags using clustering, in 2009 the method was adapted by Zhang and Zhou to map bags into single instance feature space (Pappas, Pappas, and Popescu-belis 2015).

In Pappas et al. 2015 work, they proposed new MIR model which assigns individual relevance values to each instance of a bag

1.4. Report Structure

The remainder of this report is organized as follows. Chapter 2 focuses on Multiple-Instance Learning application areas and their main concepts. Additionally, identification of some learning algorithms and their comparison. Chapter 3 describes the dataset analysis and the methodology used in the study. Chapter 4 presents the demonstration and discussion of the results obtained from experimental setups of assigning value of a single instance to a bag and representing a bag with the average values of instances in each bag. Also, compares performance of each model on the two approaches. Finally, chapter 5 concludes the work realized during the thesis and discuss future perspective.

2. MIL Concepts

2.1. Application Cases.

This chapter presents a background review of Multiple Instance Learning concepts, it is divided into three sections. The first section discusses different application areas, the contributors and work already done in this field. Also, summary of data impacts on application field performance. Section two gives an overview of predictive analytics and its application to the area of yield prediction. Third section describes MIL algorithms and follows with a comparison of handling data while accomplishing tasks.

2.1.1. Drug Activity Prediction.

Dietterich et al. introduced MIL in 1997 to solve the problem of drug activity prediction. The goal was to determine the ability of a molecule binding strength whether it will make a new drug or not by observing its activity when binding to binding site (Dietterich, Lathrop, and LozanoPerez 1997). Drugs are small molecules with alternative shapes possible by rotating bonds. For a molecule to produce a new drug, it must be of required shape and bind at low energy.

The learner has knowledge about individual molecules, but not about the shapes they can take on, this represents the MIL framework. Each molecule is regarded as a labeled bag of instances and the shapes of the molecules as instances without individual labels. A bag is labeled positive if at least one of the shapes of the molecule conforms with the binding site and negative if none of the shapes binds. All shapes of qualified molecule are regarded as positive instances and all shapes of unqualified molecules as negative instances. The solution to this problem proposed three axis-parallel rectangles (APRs) algorithms of a noise-tolerant standard, outside-in and inside-in (Dietterich, Lathrop, and LozanoPerez 1997). APRs algorithm works under the assumption that positive instances are located in a single cluster or region in feature space (Marc-Andre et al. 2016).

2.1.2. Image Retrieval.

The work on image retrieval proposed Diverse Density (DD) algorithm in 1998 by Maron and Ratan. The problem is to find a concept point in feature space that is close to sub-images from every positive bag and far from all negative bags (Ratan et al. 1999). An image is represented by a bag of instances of sub-segments/sub-images. This problem follows the MIL framework standard: an image is considered a labeled bag while instances are various sub-images in a bag (Ratan et al. 1999). An image is regarded positive if one of its instances is not far from maximum Diverse Density point located by feature space search carried out by the algorithm.

The ability of DD framework to perform well on other problems of MIL, such as drug activity, stock selection and image classification, makes it a bench mark for other algorithms. DD algorithm also works under the assumption that positive instances are located in a single cluster or region in feature space (Marc-Andr'e Carbonneau 2016).

2.1.3. Text Categorization (Andrews et al. (2002))

Formulating this as MIL depends on problem contents and can be done at different levels. Documents can be bags with passages as instances, or the passages can be bags with paragraphs as instances. Words can be regarded as instances, but passages and paragraphs can be as well (Yang 2005).

MIL categories are applicable to different fields, reacting to problems differently. Table 2.1 shows a clarification of data impacts on application field performance, according to Marc-Andre et al. paper published in December 2016, “*Multiple Instance Learning: A Survey of Problem Characteristics and Applications*”.

Problem Characteristics

Application Fields	Instance classification	Real-Valued outputs	Low witness rate	Intra-bag similarities	Instance co-occurrence	Structure in bags	Multimodal positive distribution	Non-modellable negative distribution	Label noise	Different label spaces
Drug activity prediction		x		xx			x	x		
DNA Protein identification	xx	x	x	xx		xx	x	x		
Binding sites identification	xx	x		xx			x	x		
Image Retrieval			x	x	xx	xx	xx	xx	x	xx
Object localization in image	xx		x	x	x	x	xx	xx	xx	x
Object localization in video	xx		x	x	x	xx	xx	xx	xx	x
Computer aided diagnosis	x	x	x	x	x		x		xx	x
Text classification	x		x		xx		xx	x	x	x
Web mining	x		x	x	x	x	x	x		x
Sound classification	x			x	x	xx	x	x	x	
Activity recognition	x				x	xx	x	x	x	x

Table 2.1: MIL Application problem characteristics (Marc-Andr e Carbonneau 2016)

x= Moderate performance impact

xx= Large performance impact

MIL is basically divided into three main categories: prediction, retrieval and categorization serving as a framework guide for solving many real world problems. Any problem emulating MIL must fall under one of the categories.

2.2. Predictive Analytics

Predictive analytics is the process of using data mining, statistics, machine learning techniques to predict the future using present or historical data. The goal is to provide assessment of what will happen in future regarding what has happened. It gives better intelligence and insight about how to make the best decisions, drawing on the right information at the right time.

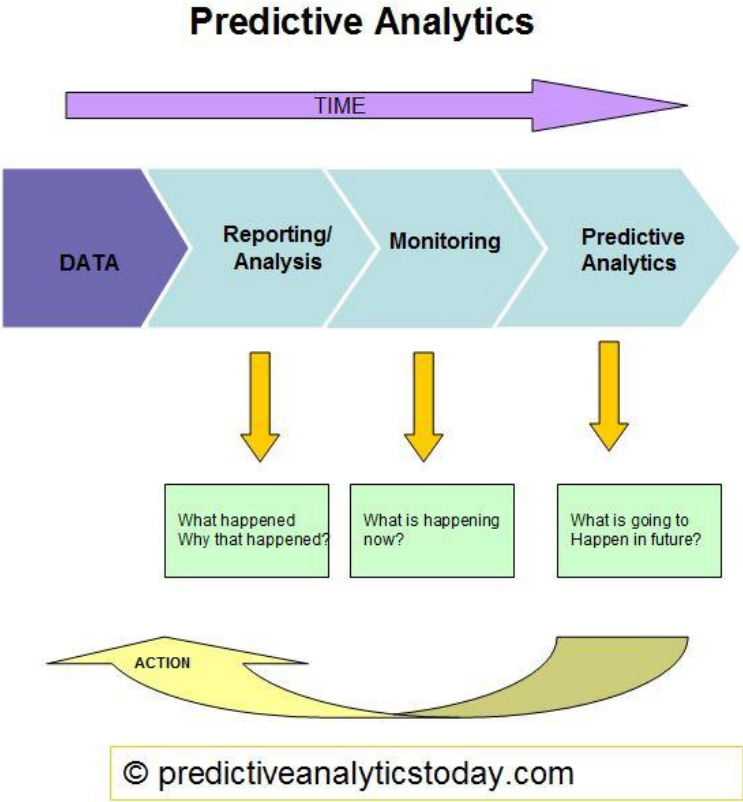


Figure 2.1: Predictive Analytics Evolution

Predictive analytics relies on exploiting the relationships between explanatory variables and predicted variables from past occurrences to predict the outcome of an unknown event. Explanatory variables are also known as independent variables, they can be measured and determine variation in outcome of dependent variables. Predictive analytics can be categorized into three different types, they are commonly used together since each tackle different decision.

Predictive Models are used for analyzing likely of an event to occur, having insight into the future by predicting what will happen. Predictive model outcomes in future, can be a real value outcomes or binary outcomes representing the probability of an event to occur or not.

Descriptive Models centered on providing an insight into past occurrences description. They enable learning from past behaviors and how they might influence future outcomes. Decision Models encompass set of rules for the outcome of any action.

2.2.1. Regression Analysis

Regression Analysis is one of the most widely used type of predictive analytic. It is used for prediction of quantitative target variables. It is also used for evaluating the impact of the variables on one another. Target variable is often known as dependent or output variable and independent variables as predictors. They determine at least partially the quantity of the quantitative target variable.

2.3. Learning Algorithms

We review some algorithms already proposed in solving MIL problems in this section. One of the issues in Multiple-Instance representation is the selection of an appropriate algorithm. The selection should be based on problem contents, since not all algorithms work efficiently in all problems, another is determining what is the bag and instances in the bag (Zhou 2004)

2.3.1. Linear regression

Linear regression is a means to study and model the relationships between explanatory variable(s) and dependent variable. The dependent variable is continuous, explanatory variables can be continuous or discrete and nature of the regression line is function of each of the explanatory variables, holding the others fixed, and the contributions of different explanatory variables to the predictions are additive. Linear regression is sensitive to outliers which can affect predicted values and subject to over-fitting.

Linear regression equation to find prediction y:

$$y = a + bx + e$$

Where: y = dependent variable

a = intercept

x = explanatory variables

b = slope of the line

e = regression error

Linear regression analysis can be divided into simple and multiple regression analysis by their number of independent variables.

Simple linear regression analyzes relationship between dependent variable and one independent variable. Relationships in simple linear regression can be an exact relationship between the two variables where they fall exactly on linear regression line or relationship which are not exactly but there are trend and scatter relationship between the two variables.

Multiple linear regression analyzes relationship between dependent variable and more than one independent variable. It usually suffers from multicollinearity, highly correlated independent variables and error components are uncorrelated with one another. In multiple linear regression, only relevant variables must be included in the model and linearity must be assumed.

The general form of multiple regression:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_ix_i + e$$

2.3.2. Least Absolute Shrinkage and Selection Operator (Lasso)

LASSO is a regression method applicable when too large coefficients are not needed and there is a need for prediction and interpretation accuracy. It produces some coefficients that are exactly zero by minimizing residual sum of squares subject (Tibshirani 1996). It can reduce the variability and improve the accuracy of linear regression models. LASSO is convenient when there is automatic variable selection, it handles multicollinearity by picking only one of the variables and shrinks others to zero. The shrinkage process allows for better interpretation of the model and identifies the most important variables strongly associated with the response variable. LASSO uses tuning parameter lambda to control the strength of the penalty. Lambda and model coefficients are inversely related, as lambda increases, more coefficients are reduced to zero that is fewer predictors are selected and there is more shrinkage of the non-zero coefficient. Least Angle Regression and Shrinkage (LARS) can be modified to compute regularization path of LASSO.

2.3.3. Multivariate Adaptive Regression Splines (MARS)

MARS is a fully interpretable regression method which requires no previous assumptions to best fit model (Mendes-Moreira et al. 2014). It uses piecewise linear segments to describe complex relationships between variables (Leathwick et al. 2005) and fits linear segments with different slopes to describe non-linear responses between variables (Hastie, Tibshirani, and Friedman 2009). MARS does not need long training process and hence can save lots of model building time (Lee and Chen 2005). It combines the strength of both regression tree and multivariate linear regression. MARS model operates in two-stage process. Firstly, it allows continuous selection of explanatory variables, they can interact with each other or be restricted to enter in only as additive components. In the second stage, variables are eliminated in order of least useful explanatory variables among the previously selected set determined by the generalized cross-validation (GCV) criterion (BALSHI et al. 2009; Chou et al. 2004). MARS performs well for predictive modeling of continuous outcomes, it's more interpretable and distinguishes well between signal and noise variables (Crino and Brown 2007).

2.3.4. K Nearest Neighbor(K-NN)

This method depends on the distance between samples. Bags are labeled according to their closeness to the target using nearest analysis of all neighbors and those that regard themselves as a neighbor in the training set (Zhou) (Marc-Andr e Carbonneau 2016). After specification of the

target, the average of neighbors is used to define the prediction value. Different metrics are used for calculating the distance depending on predictors types (Kuhn and Johnson 2013). For accuracy result and equal contribution of predictors in calculating distance, predictors should be scaled and centered prior to performing K-NN. K-NN uses the values of other training set point to calculate the value of an unknown given a point function (Navot et al. 2006). Bags with similar measures are assumed to belong to the same class label (James and Eibe 2004).

Formulas for Euclidean and Minkowski Distance Metrics;

Euclidean Distance: Is a straight-line distance between two samples.

$$\sqrt{\left(\sum_{j=1}^p (y_{aj} - y_{bj})^2\right)}$$

(Kuhn and Johnson)

Where \mathbf{y}_a and \mathbf{y}_b are two samples

Minkowski Distance: Is a generalization of the Euclidean distance:

$$\sqrt[q]{\left(\sum_{j=1}^p |y_{aj} - y_{bj}|^q\right)}$$

(Kuhn and Johnson)

2.3.5. Artificial neural networks

Neural Networks is an iterative backward approach made up of three-layers perceptron: of input layer, hidden layer and output layer containing interconnecting elements called nodes. The input layer receives values of independent variables as inputs, hidden layer adds together values from different nodes of input layer together with a small value for each node called bias/beta and the result is called $y_{\text{predicted}}$ in the output layer. At output layer, a comparison of predicted value and actual prediction value is done. If the two are not equivalent or close, the iteration is repeated by adjusting the bias value until neural produces accurate prediction for most observations. Once this is achieved the model is used to apply predictions.

The algorithms can be expressed into two steps of feed forward computation from input layer down to output layer and backpropagation from output layer to the hidden layer for weight updates. The process will keep running until the value of error function becomes small (Hnin, Pa, and Thu 2017)

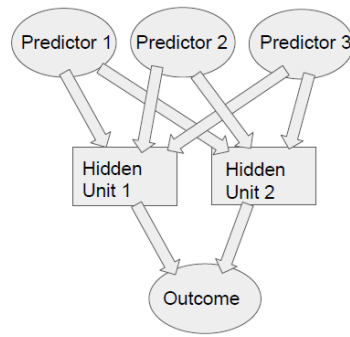


Figure 2.2: Neural Network Process (Hastie, Tibshirani, and Friedman 2009)

2.3.6. Support Vector Machines (SVMs)

SVMs is a method which uses hyperplanes to separate two classes or positive instances from negative instances of the training data (Tong and Koller 2001), and then minimizes the margin between them while ensuring that all points are classified correctly. There are two ways of defining margins (Babenko 2008). The first method is by ignoring negative instances in positive bag and only one positive instance can be a support vector but in the second method, negative instances and multiple positive instances can be support vectors.

2.3.7. Random forest

This method splits data into subsets of trees by randomly selecting data and variables to develop decision trees from a training dataset. An entity of the tree is created by using bootstrap samples of training data and random feature in tree induction (Svetnik et al. 2003). Random forest is made up of decision trees with nodes which can have two or more branches and leaf nodes which are used to represent decision on numerical targets. A decision tree is built top-down from the root node by grouping instances with similar value on the same branch. All decision trees are used to create a model and conduct voting for each of the observations. Trees provide accurate prediction due to making mistake at different nodes.

2.4. Application of Predictive Analytics on Yield Prediction.

Yield prediction is a significant challenge in agricultural sector (González-sanchez and Frausto-solis 2014). At the beginning of new planting season, plant breeders desire having rough yield estimate for all the crops involved so as to maximize production (Frausto-Solis, Gonzalez-Sanchez, and Larre 2009). In the past, yield prediction was done traditionally where farmers relied on previous experience on particular field and crop, which is usually inaccurate (Ruß 2009). Due to inaccuracy of the traditional method, more efficient methods were developed with majority being crop specific mechanistic models while those not are available for crop types variety through parameter fitting (González-sanchez and Frausto-solis 2014). The models simulate different processes and predict their interacting effects on crop growth and yield. Though some of the models are moderately accurate but are expensive in terms of time and money.

Crop-Environment Resource Synthesis (CERES) Maize model was developed specifically to simulate maize development, growth and yield. It was first released in 1986 and since then, different versions of it has been generated by slight changes in the original model (L et al. 2005)

2.5. Summary

In this chapter, we covered a brief overview of MIL, starting with its application areas and then followed by an overview of predictive analytics and its application. Learning algorithms and its application field problem characteristics was also presented, finishing with an explanation of how learning methods handle data problems.

MIL algorithms have different ways of handling data issues while accomplishing tasks. Table 2.2 summarizes MIL algorithms manner in handling data characteristics as stated in Hastie, Trevor, Robert Tibshirani, and Jerome Friedman 2009 textbook “*The Elements of Statistical Learning Data Mining, Inference and Prediction.*”. Good is the best which means the characteristic has no effect in the algorithm performance, Fair means the performance can be disturbed but not severe while bad means severe influence in performance of the algorithm.

Characteristics	Neural nets	SVM	Trees	MARS	k-NN kernels
Handling of multiple data types	++	++	***	***	++
Handling of missing values	++	++	***	***	***
Robustness to outliers	++	++	***	++	***
Input transformations insensitive	++	++	***	++	++
Computational scalability	++	++	***	***	++
Ability to deal with irrelevant inputs	++	++	***	***	++
Linear features extraction	***	***	++	++	+*+
Interpretability	++	++	+*+	***	++
Predictive power	***	***	++	+*+	***

Table 2.2 : Learning Algorithms Characteristics(Hastie, Tibshirani et al.)

Key: Good= ***

Fair= +*+

Bad= ++

3. Dataset and Methodology

This chapter gives a brief description of dataset and the methodology employed in this work. It is divided into two sections with section one describing the dataset and section two the methodology used during this research.

3.1. Dataset

Maize crop also known as corn, *Zea mays* is the most popular crop of the grains class widely cultivated throughout the world, with the United States, China, and Brazil being the top three maize-producing countries. It plays an important role in the world economy and is a valuable ingredient in manufactured items that affect a large proportion of the world population (Nemati et al. 2009). It was first cultivated in Mexico about 10,000 years and has been evolving since then. There are many factors which contribute to quality.

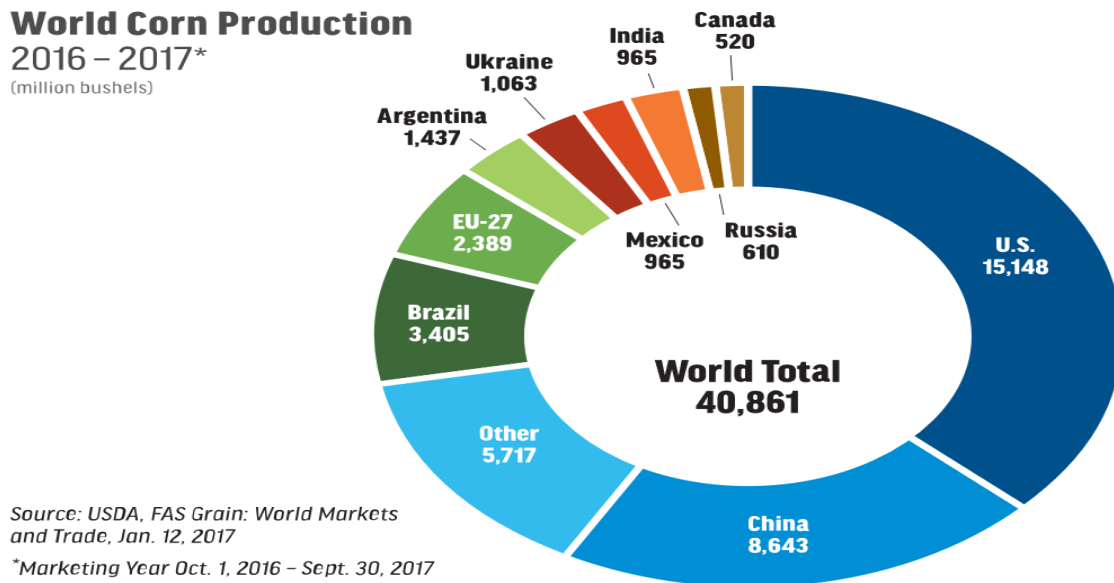


Figure 3.1: World Corn Production 2016/2017

Some common types of corn according to (Wikipedia 2014) are

- Dent corn *Zea mays indentata*: This type of corn is made up of dented kernel, consist of a soft large number of glucose and easy to mill. Kernel indentation usually shows its maturity,
- Flint corn *Zea mays indurata*: Is the common subspecies of corn named for its hard kernels, which come in a multitude of colors. It is usually milled in a certain way and retains a distinct texture and flavor when cooked.
- Popcorn, *Zea mays everta*: Is the type of corn used in making popcorn. Pressure builds within the kernel when heated, afterward result in pop explosion
Flour corn: The purpose of this variety, is corn flour for human consumption

- Sweet corn *Zea mays saccharata*: This variety has the health benefits of increasing acidic levels of ferulic, which fights against cancer. It contains high sugar content.

The stages involved in producing desired maize with characteristics that satisfy needs are four; the first stage is domestication for human use, second stage is germplasm collection from different sources, the third step is introducing maize germplasms from one region or country to another, then acclimatize it to the changed climate. Stage four is hybridization to create genetic variation by crossing two plants or lines of dissimilar genotype.

3.2. Data Description

The dataset used for this study was collected from Maize Breeding Station known as NUMI (NUcleo de melhoramento de Milho, Braga, Portugal). NUMI was established after successful hybridization of North American and Portuguese germplasms, it is responsible for overall national programs and the production of important hybrids. The dataset comprises of maize ears characteristics of 20% Portuguese and 80% North American dent and flint maize types. The yield components considered are but not limited to; kernel length, row per ear, kernel per ear, size and weight, i.e. traits of maize ear. Other factors that affect yields are eco-geographical, physiological, seed development processes and evolutionary but they are not to be considered in this work. The dependent variable is yield and independent variables are the yield components and soil characteristics.

Two datasets in different worksheets of a single Microsoft Excel Workbook was used for this problem. The first consists of data for 4801 observations, each observation contains information on 46 variables. The observations are grouped using serial number, maximum observation in each serial number is 10.

The first dataset sample supposed to have 5300 as total observations, because it is a multiple 10 of the second sample. It was impossible to achieve this due to removing observations with missing values. So, 499 observations were missing. Table 3.1 shows the first 29 observations of the first dataset sample.

10 Serial	Genotype	EW	Alt	Alt 1E	PI	EW	L	DE1	DE2	Yield	DE3	DE4
1	Pigarro C0(C3) 08	12995	250	143	1	12995	141	56	53	6473.513628	44	42
1	Pigarro C0(C3) 08	11340	227	112	2	11340	176	49	46	6473.513628	37	36
1	Pigarro C0(C3) 08	17285	242	140	3	17285	140	66	55	6473.513628	55	44
1	Pigarro C0(C3) 08	22088	228	137	4	22088	154	73	65	6473.513628	62	45
1	Pigarro C0(C3) 08	19268	220	130	5	19268	169	63	59	6473.513628	53	46
1	Pigarro C0(C3) 08	2315	277	130	6	2315	220	53	51	6473.513628	40	39

1	Pigarro C0(C3) 08	2759	280	167	7	2759	210	58	55	6473.513628	55	45
1	Pigarro C0(C3) 08	2333	206	120	8	2333	200	60	55	6473.513628	47	45
1	Pigarro C0(C3) 08	2445	225	100	9	2445	170	64	60	6473.513628	49	50
1	Pigarro C0(C3) 08	1353	260	162	10	1353	142	53	50	6473.513628	45	43
2	Amiudo (C3) 97	11943	182	90	1	11943	161	39	38	5938.269478	33	33
2	Amiudo (C3) 97	11648	174	55	2	11648	150	42	42	5938.269478	32	32
2	Amiudo (C3) 97	11355	186	130	3	11355	126	45	45	5938.269478	37	37
2	Amiudo (C3) 97	12887	205	100	4	12887	142	44	44	5938.269478	34	35
2	Amiudo (C3) 97	12595	212	110	5	12595	140	45	44	5938.269478	36	33
2	Amiudo (C3) 97	11302	216	120	6	11302	182	45	42	5938.269478	34	33
2	Amiudo (C3) 97	13608	217	110	7	13608	164	42	42	5938.269478	35	35
2	Amiudo (C3) 97	13574	220	138	8	13574	152	47	44	5938.269478	40	38
2	Amiudo (C3) 97	11047	190	120	9	11047	178	40	38	5938.269478	33	32
2	Amiudo (C3) 97	8536	192	112	10	8536	124	42	40	5938.269478	36	35
3	Broa - 102	8771	234	102	1	8771	126	46	40	5215.891396	37	31
3	Broa - 102	10608	210	100	2	10608	112	45	43	5215.891396	39	37
3	Broa - 102	14399	214	100	3	14399	123	55	50	5215.891396	58	42
3	Broa - 102	13910	222	117	4	13910	144	55	50	5215.891396	48	39
3	Broa - 102	14955	220	143	5	14955	161	56	52	5215.891396	41	41
3	Broa - 102	1843	179	126	6	1843	211	50	50	5215.891396	40	40
3	Broa - 102	1360	128	110	7	1360	164	50	43	5215.891396	40	45
3	Broa - 102	1420	199	110	8	1420	151	49	47	5215.891396	40	40
3	Broa - 102	1276	230	127	9	1276	165	45	50	5215.891396	40	40

Table 3.1: First sample Field values overview of the first 29 records with some traits

The second dataset sample consists of 530 observations, built by averaging each trait in the serial numbers (that is, for trait EW in serial 1 of the first dataset, we average the 10 observations of EW in the serial 1, then represent the value with EW in serial 1 of second dataset. So, in the second dataset instead of having 10 EW values in serial 1, we have only a single value). Table 3.2 gives an overview of the first 30 observations of our second dataset sample.

10 Serial	Genotype	EW	Alt	Alt 1E	PI	L	DE1	DE2	Yield	DE3	DE4
1	Pigarro C0(C3) 08	94.1 81	26 0	162	10	172. 2	59.5	54.9	6473.513 628	48.7	43.5
2	Amiudo (C3) 97	118. 495	19 2	112	10	151. 9	43.1	41.9	5938.269 478	35	34.3
3	Broa - 102	69.7 24	21 5	136	10	146. 7	50.3	47.3	5215.891 396	42.8	39.5
4	Pop45xPig	207. 307	25 0	128	10	173. 9	51.7	50.6	10257.75 69	43.4	43.3
5	Amiudo C0S0 03	122. 447	23 2	120	10	161. 2	45.6	43.8	4898.046 743	35.5	35.2
6	Amiudo C0S0 84	105. 481	19 6	90	10	148. 8	42.2	40.1	6219.654 703	34	32.8
7	Broa - 164	175. 984	17 8	124	10	173. 6	55	52.1	7926.873 494	44.3	41.7
8	Broa - 70	142. 852	22 5	106	10	150. 4	52.1	49.4	5328.747 394	37.8	36.5
9	Aljezudo2 006	191. 023	27 0	120	10	199. 8	49.6	48.5	6204.734 901	39	38.1
10	Bastos C0S0 96	164. 515	24 6	149	10	161. 3	53.2	51.6	5617.553 888	42	40.5
11	Algarro 08	201. 886	24 2	114	10	190. 6	50.9	49	7343.812 055	39.2	39.2
12	Pigarro 2008	49.1 65			10	107	37.5	36.5	6734.652 189	33.5	25
13	CMSPH3	109. 006	13 0	78	10	146. 3	48.3	45.7	5771.578 282	40.8	36.2
14	Broa- 172	84.1 14	15 0	86	10	137. 2	37.7	36.7	4883.895 858	31.1	30.6
15	Broa - 136	87.6 6	21 7	110	10	143. 888 888 9	41	39.77 77777 8	4619.342 149	32.55 55555 6	32.44 44444 4
16	Broa - 142	107. 136	18 7	100	10	137	45	42.9	3415.331 119	36.6	35.6
17	Broa - 186	116. 362 222 2	17 5	80	10	154. 888 888 9	42	41.33 33333 3	6680.236 362	35.55 55555 6	34.22 22222 2

18	Broa - 148	99.6 3	16 5	74	10	140. 75	42.125	41.75	4794.169 606	34.87 5	33.5
19	ACC Nº03972 94	126. 255	18 0	87	10	143. 7	52.666 66667	49.55 55555 6	4442.038 985	11111 1	40.33 33333 3
20	Broa - 214	104. 252	25 0	133	10	153. 1	43.5	42	5075.758 923	35.3	33.9
21	Broa - 34	133. 001 111 1	23 5	107	10	144. 444 444 4	53.555 55556	50.22 22222 2	4944.229 759	33333 3	41.11 11111 1
22	ACC Nº03974 94	116. 51	20 3	67	10	135. 6	51.1	48.6	3462.817 366	43.9	42.2
23	Bulk 1990/91	124. 219	18 3	118	10	145. 1	48.1	46.4	4444.575 778	36.5	34.5
24	Broa - 92	260. 488 888 9	21 6	100	10	164. 444 444 4	44.444 44444	42.66 66666 7	3391.363 656	55555 6	35.22 22222 2
25	Broa - 83	125. 483	21 4	98	10	146. 7	45.6	44.3	4338.644 936	36.8	36.5
26	Broa - 57	163. 088	19 6	112	10	168. 5	54.9	51.8	6530.942 203	42.6	35.4
27	Broa - 48	125. 998	18 0	105	10	143. 6	53.8	47.4	4746.654 741	44.5	37
28	BS22	154. 39	17 7	80	10	176. 7	46.3	45.2	4932.961 954	39	39.1
29	Broa - 213	136. 747	18 8	98	10	166. 7	47.1	45.9	7123.826 499	37.2	37.2

Table 3.2: Second Sample Field values Overview of the first 29 records with some traits.

Each data sample contain information about a distinct analysis bag labelling topic. The model variables, codes and units are identified in table 3.3.

Traits	Codes	Scale/Unit	Description
Root Lodging	R	%	Percentage of plants leaning more than 30 ⁰ from vertical
Stack Lodging	S	%	Percentage of stalk quality and damage caused by insect attack
Stand	Stand_pl_ha	Plant ha ⁻¹	Thousands of plants per hectare
Ear Weight	EW	g	Ear Weight adjusted to 15% of grain Moisture
Cob Weight	CW		Weight of cob
Grain Yield	Yield15	Mg ha ⁻¹	Yield Quantity
Kernel Weight	KW	g	Kernel weight per ear adjusted to 15% of grain moisture
Cob Weight/Ear Weight	CW_EW	%	Percentage of cob weight in the ear weight
Ear Length	L	cm	Ear Length
Ear Diameter	DE(1,2,3,4)	cm	DE1,DE3: Large Diameter in the bottom and top of ear. DE2,DE4: Small Diameter 90 ⁰ rotation from large diameter
Kernel row Number	R(1,2)	n ⁰ (number)	Row number in the 1/3 bottom and top of the ear respectively
Fasciation	Fa	1 to 9	1- Without fasciation 9- Maximum fasciation
Determinated/Indeterminated	D_I		Average value of determinated and indeterminated 2-Determinated ears 1-Indeterminated ears
Convulsion	CV	0-5	Kernel row arrangement in the ear 0- without convulsion 5-Maximum convulsion
	KC		
Kernel Dept	KD	cm	Measure of one kernel in the middle ear
Flint/Dent	F_D	1 to 9	1-Popcorn 2-Flint 3-Medium flint 4-low flint, 5-50% flint & 50% dent, 6-low dent, 7-medium dent, 8-high dent, 9-sweet maize
Plant Height	H	cm	Height from the stalk basis to the last leaf insertion
Thousand Kernel Weight	SW	g	Thousand kernel weight at 15% moisture content
Kernel Number	KN		Kernel number per ear
Kernel per row	NC	n ⁰	Kernel number per row
Cob Diameter	DC(1,2,3,4)	cm	DC(1,3)-Large diameter DC(2,4)-Small diameter

Table 3.3: Dataset Traits Description.

3.3. Task

This work presents a predictive regression analytic on how maize ears traits affect yield. The tasks analyzed separately two MIL predictive approaches to assign value to a bag: 1) Find the average of each instance in N vary observations of a bag, assign the value to that instance then predict a value for the target;2) Predict a value for each observation in N vary bag, then average the predicted values to represent the bag. Different MIL models were applied, and then performance of each model on the two topics were compared. The first dataset sample of 4801 observations was used for experimental setup of the second topic while the second data sample with 530 observations was used for the analyzes of the first topic.

Data were analyzed using R programming language in conjunction with Rstudio and several packages. R language was chosen due to its powerful integration with multiple programming languages and graphical analytical skills and because it is a standard tool in the area.

3.4. Methodology

This section describes the Cross Industry Standard Process for Data Mining (CRISP-DM), the methodology chosen for this work.

CRISP-DM

Is the most commonly used data mining approach by data miners (Wikipedia 2016). CRISP-DM is an industry proven methodology and process model to guide data mining, which also allows the creations of model that fit particular needs (International Business Machines Corporation USA 2013). The process is divided into six phases but only five phases were applicable in this work.

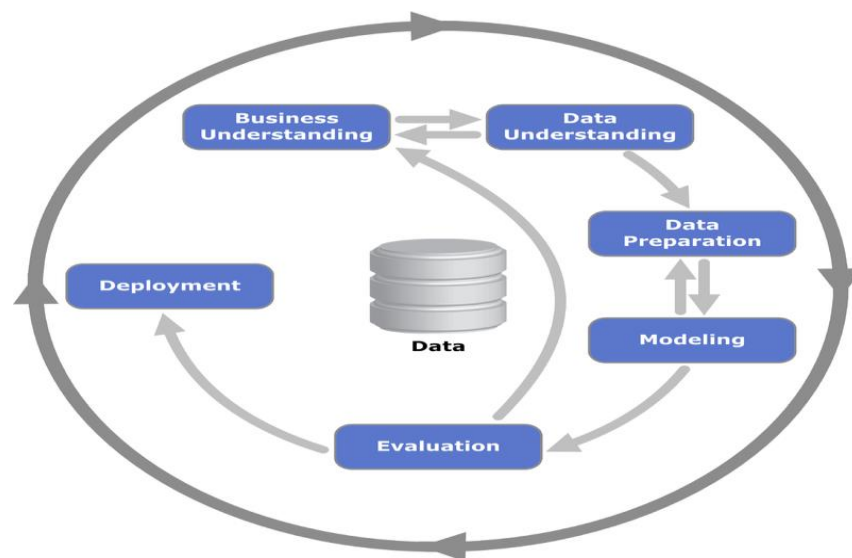


Figure 3.2: CRISP-DM Data mining life cycle (Wikipedia 2016)

1. Business Objective: This is the fundamental of any data mining problem. Thorough understanding, clear definition of problem goals and objectives of the task, then converting the knowledge into data mining problem (Chapman et al. 2000). In this study, the objective was to predict maize yield using based on maize ear traits.
2. Data Understanding: Data was studied to determine any abnormality in its quality.
3. Data Preparation: Convert initial dataset into final dataset that was used for modeling. The two worksheets and variables were renamed to suit R programing language name conversion. All observations with missing values were removed.
4. Modeling: This is the phase where the main task was performed. Modelling techniques to be used were selected, activities such as parameters settings, running the models, model description and assessing model result were done in this phase
5. Evaluation: Evaluation, comparison and review of model(s) built to ascertain achieving meeting business objective. Two of the most common regression models' accurate metrics were used; Mean Squared Error(MSE) for quantifying error and Root Mean Squared Error(RMSE) which measures the difference between actual and predicted values

$$\text{RMSE} \quad \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$\text{MSE} \quad \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Where

\hat{y} = predicted value

y = real value

i = observation

3.5. Summary

This chapter explains the methodology together with accurate metrics used in the dissertation period to estimate models' performance.

4. Experimental Setup.

This chapter consists of three sections which present the models predictive results. Section 1 shows the experimental setup of the first MIL approach; find the average of each instances in N vary observations of a bag, assign the value to that instance then predict a value for the target. Section 2 describes the result of the second approach; Predict a value for each observation in N vary bag, then average the predicted values to represent the bag. Section 3 compares the performance of individual model on the two approaches.

Recursive partitioning, Multivariate Adaptive Regression Splines, Least Absolute Shrinkage and Selection Operator, Random Forest and K-Nearest Neighbor were the models used to accomplished the task. Caret package (Classification And Regression Training) that contains tools for data splitting, data preprocessing, features selection and model tuning using resampling was used to split our data into 10-fold.

10-fold cross validation splits data into 10 different groups of about same size for training and testing. The process of training and testing is carried out 10 times. At each time, using 9 groups of the data in training and testing on the remaining 1 group of data to validate.

4.1.Assign Predicted value to a bag after Averaging Each Instance in the bag

This section presents the predictive results obtained from the exploration of different predictive models on the dataset sample containing a total of 530 observations using the first MIL approach to assign value to a bag; which is to find the average of each attribute in the N observations of a bag obtaining one instance per target value and then predict a value for the target. The dataset was built by averaging individual trait of observations in each N vary sizes bag which composes of maximum of 10 observations of the other dataset sample, that is, each observation/row in this dataset represents average value of traits in bag N of the other dataset sample (each bag N is a unique serial number).

All models were constructed under the same conditions using all other variables present in the data to predict yield value except for different packages which suit distinct models. Each model prediction values are presented with the RMSE and MSE values.

4.1.1. Recursive Partitioning (Rpart)

Recursive partitioning is a decision tree induction algorithm that allows for modeling of relationships and detection of interactions among variables (Zeileis, Hothorn, and Hornik 2017). It is a process of uncovering hidden patterns in data subdividing it into significant subgroups (Crino and Brown 2007). Rpart package was used for the implementation and caret createFolds function for splitting the data into 10 groups. The model did not require predictors transformation, the summary and results are presented in tables 4.1 and 4.2 respectively. Figure 4.1 and figure 4.2 illustrate the model predicted value.

Min.	2090
1 st Quantile	4175
Median	5269
Mean	5083
3 rd Quantile	5531
Max	9761

Table 4.1: rpart summary for first MIL approach.

RMSE	MSE
2074.237	4302458

Table 4.2: Results for the rpart model for first MIL approach.

Residual, difference between observed and predicted values, shows that some predicted values were less than and others greater than the actual values. The residual is distributed with a residual range from negative 3000 to positive 6000.

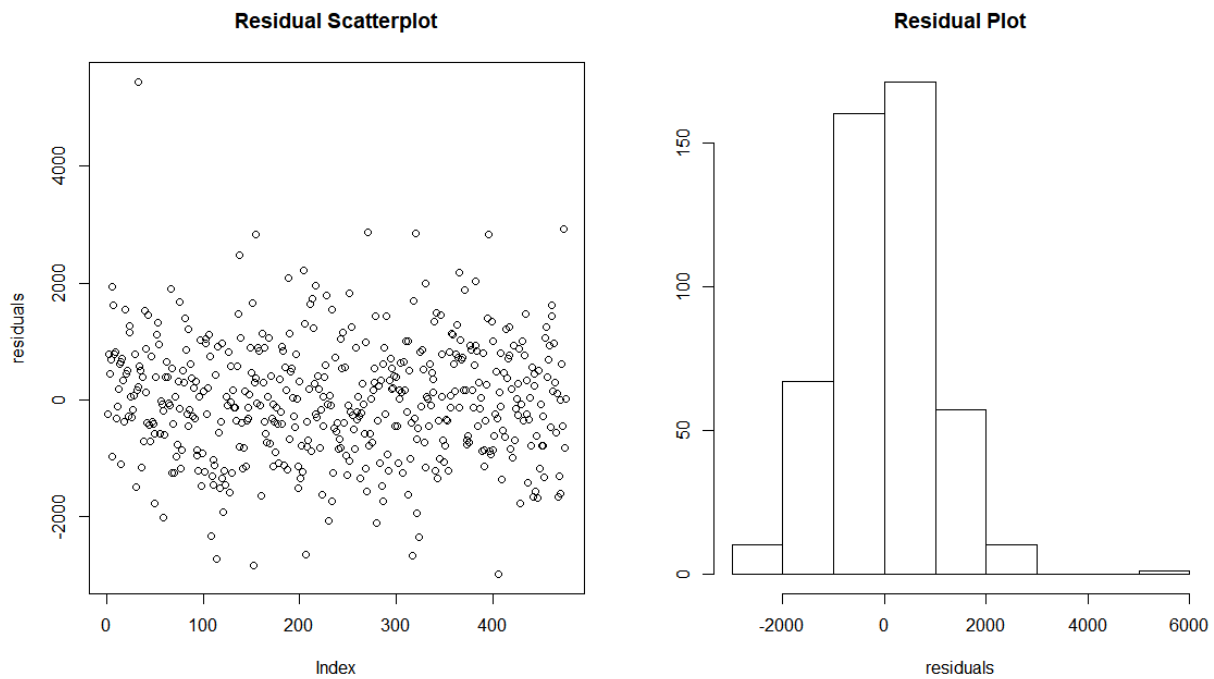


Figure 4.1: Rpart Residual Plot

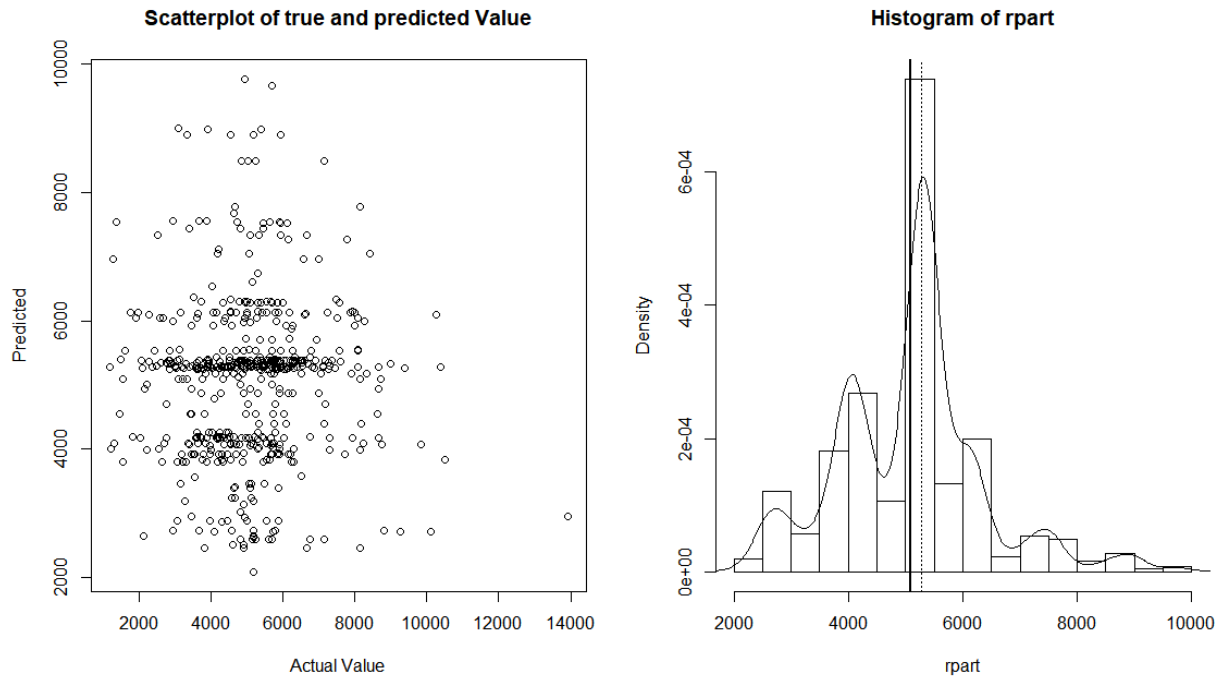


Figure 4.2: rpart Predicted Values

Median= Dash line
Mean = Thick line

4.1.2. Multivariate Adaptive Regression Splines(MARS)

MARS is an extension of linear models that automatically models nonlinearities and interactions between variables. Earth package was used for implementation of MARS. The model, like rpart, has many residuals values between -1000 and 1000, with no much difference between median and mean. createFolds function in caret was used for splitting the data into 10 groups. The model did not require predictors transformation. Tables 4.3 and 4.4 show MARS model summary and results for the first MIL approach. Figure 4.3 and 4.4 illustrate MARS performance.

Min.	-100.8
1 st Quantile	4147
Median	5112
Mean	5127
3 rd Quantile	6030
Max	26400

Table 4.3: MARS summary for the first MIL approach.

RMSE	MSE
2436.181	5934979

Table 4.4: Results for the MARS model for the first MIL approach.

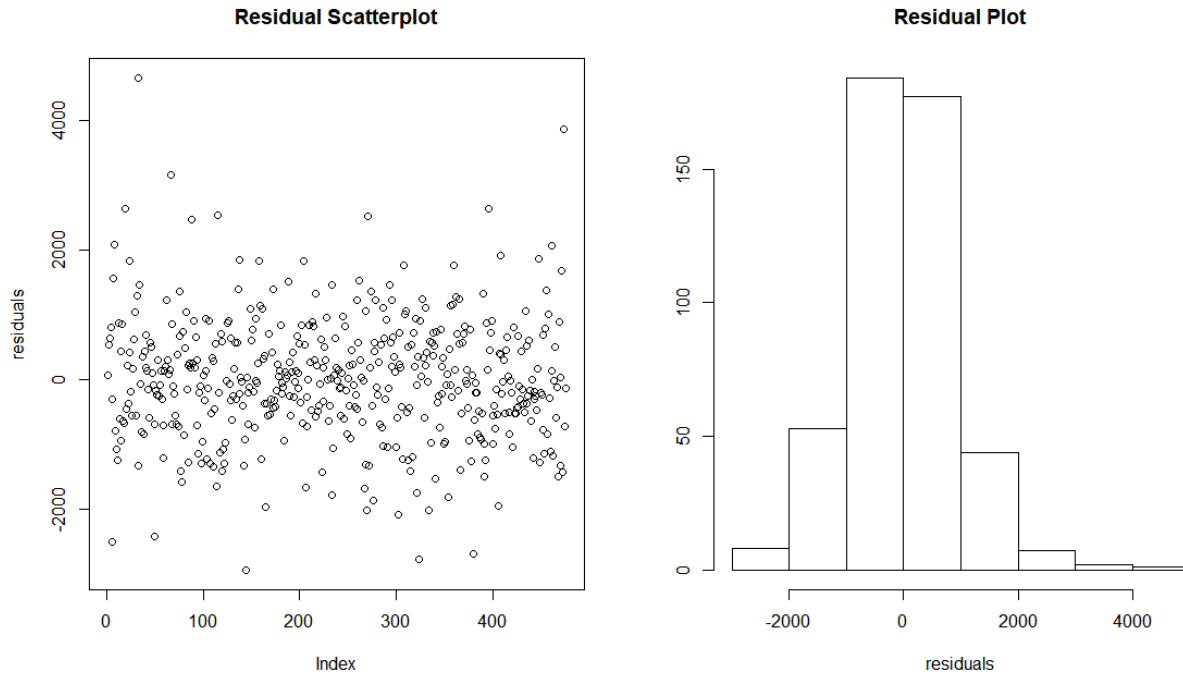


Figure 4.3: MARS scatter plot and Histogram of residual error

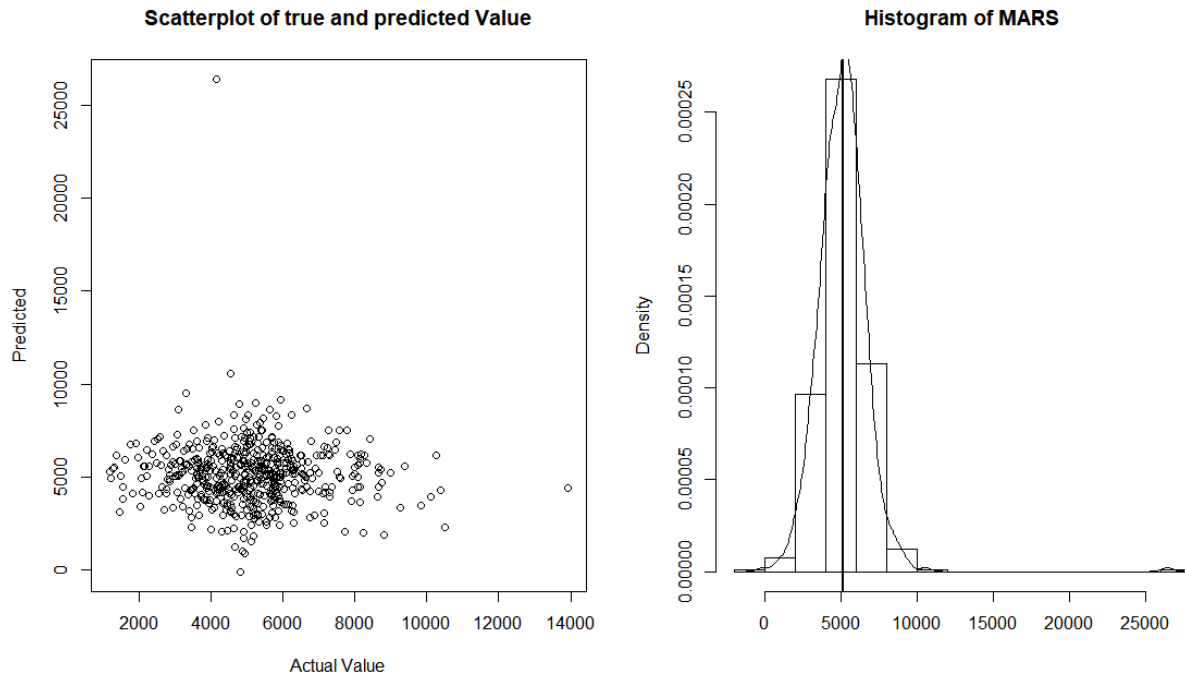


Figure 4.4: Plot and histogram of MARS predicted values

4.1.3. Least Absolute Shrinkage and Selection Operator(LASSO)

LASSO is an extension of linear regression where data values are shrunk towards a central point. Its loss function yields a piecewise linear solution path (Wieringen, n.d.). glmnet was used as the package for implementation. The model did not require predictors transformation. Tables 4.5 and 4.6 present LASSO model summary and results for the first approach. Figures 4.5 and 4.6 display LASSO performance graphically.

Min.	-751.9
1 st Quantile	4221.0
Median	5020.0
Mean	5076.0
3 rd Quantile	5884.0
Max	12580.0

Table 4.5: LASSO summary for first MIL approach

RMSE	MSE
2222.91	4941329

Table 4.6: Results for the LASSO model for the first MIL approach

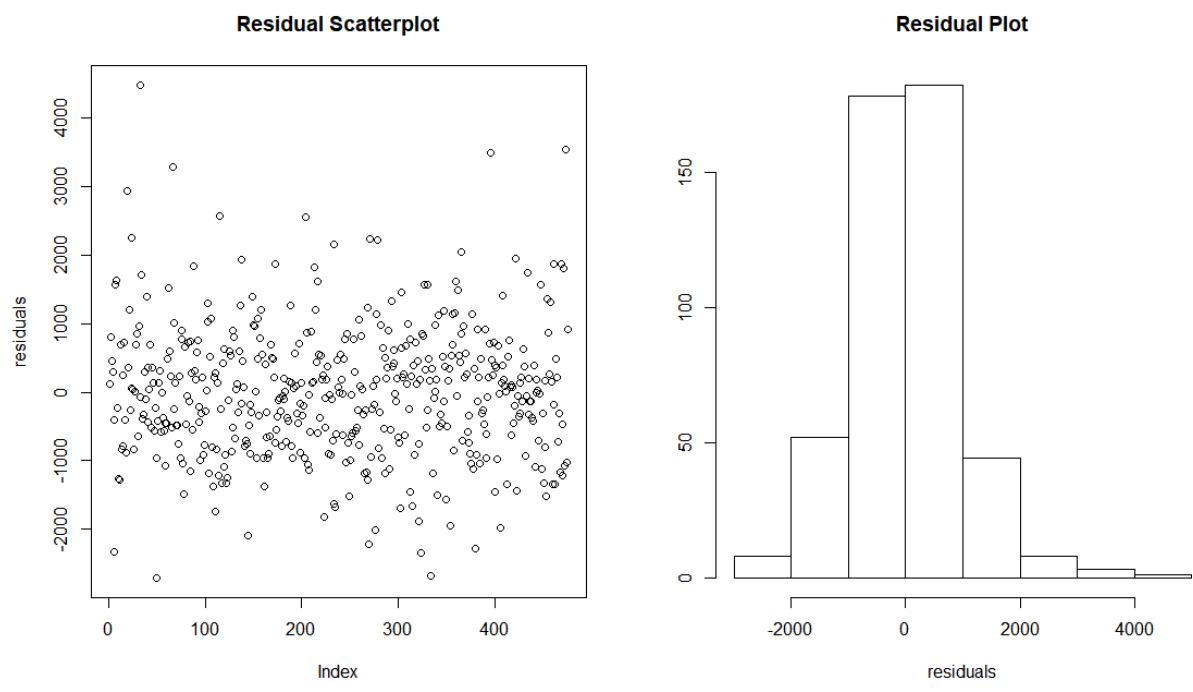


Figure 4.5: LASSO plot and Histogram of residual error

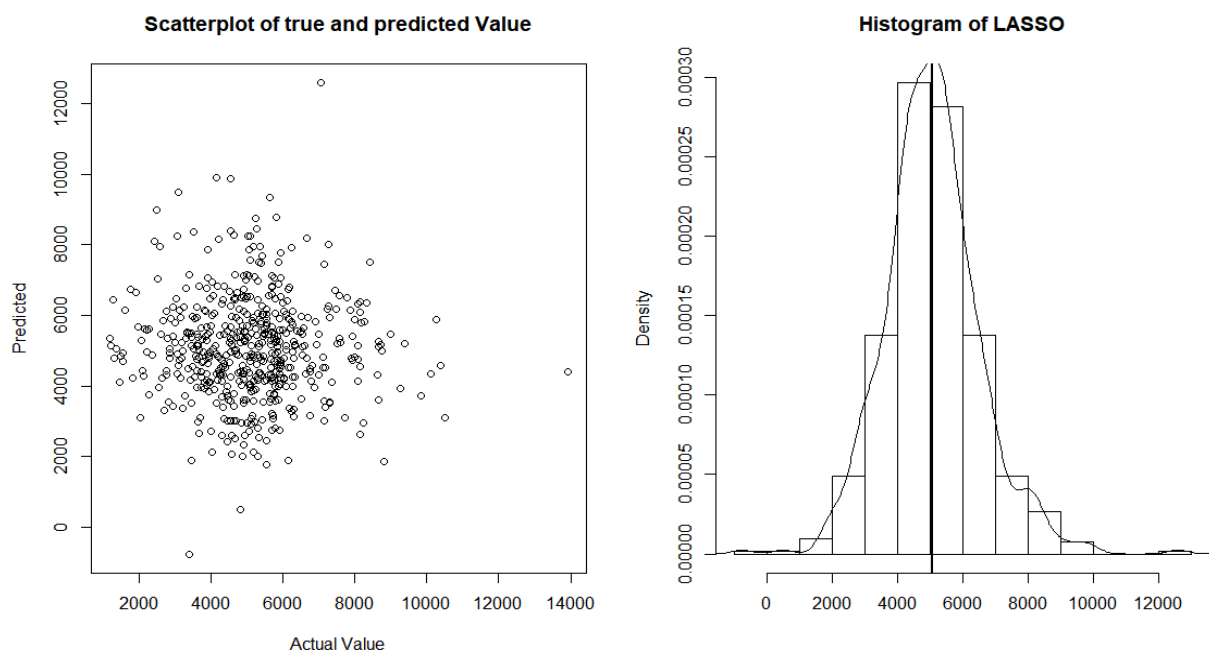


Figure 4.6: Plot and Histogram of LASSO predicted values.

4.1.4. Random Forest

Is a combination of different regression trees and each node contains a distribution for the continuous variables output, at each node, best predictors are randomly chosen to split that node (Svetnik et al. 2003). It uses random package for implementation. Tables 4.7 and 4.8 represent random forest model results and summary for the first approach. The model predicted values range between 2294 to 8325. Random forest performance is represented graphically in figure 4.7

RMSE	MSE
1949	3800702

Table 4.7: Random Forest model results for first MIL approach.

Min.	2294
1 st Quantile	4358
Median	5057
Mean	5088
3 rd Quantile	5682
Max	8325

Table 4.8: Random Forest summary for the first MIL approach.

Below is a two graphs side by side on same figure of a scatter plot of predicted value against the actual value and density against histogram of predicted value.

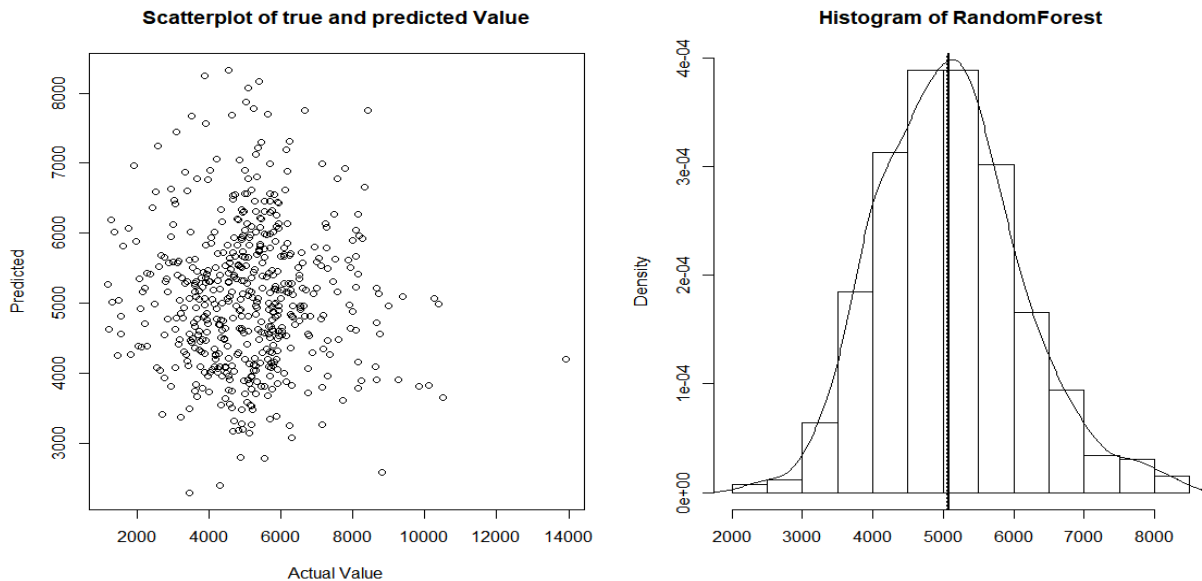


Figure 4.7: Scatter plot and Histogram of RandomForest model predicted values

Median=Dash line
Mean=Thick line

4.1.5. K-Nearest Neighbor(knn)

Knn predicts the target based on distance functions, takes advantage of closest points. Kknn package was used for implementation. No predictors transformation was done in this model. Tables 4.9 and 4.10 display KNN model results and summary for the first approach. It has a range of predicted values from 2103 to 9394. The two tables summarize its results. Figure 4.8 illustrates its performance.

RMSE	MSE
1950.561	3804686

Table 4.9: Results for the K-NN model for the first MIL approach

Min.	2103
1 st Quantile	4442
Median	5048
Mean	5138
3 rd Quantile	5750
Max	9394

Table 4.10: K-NN summary for first MIL approach.

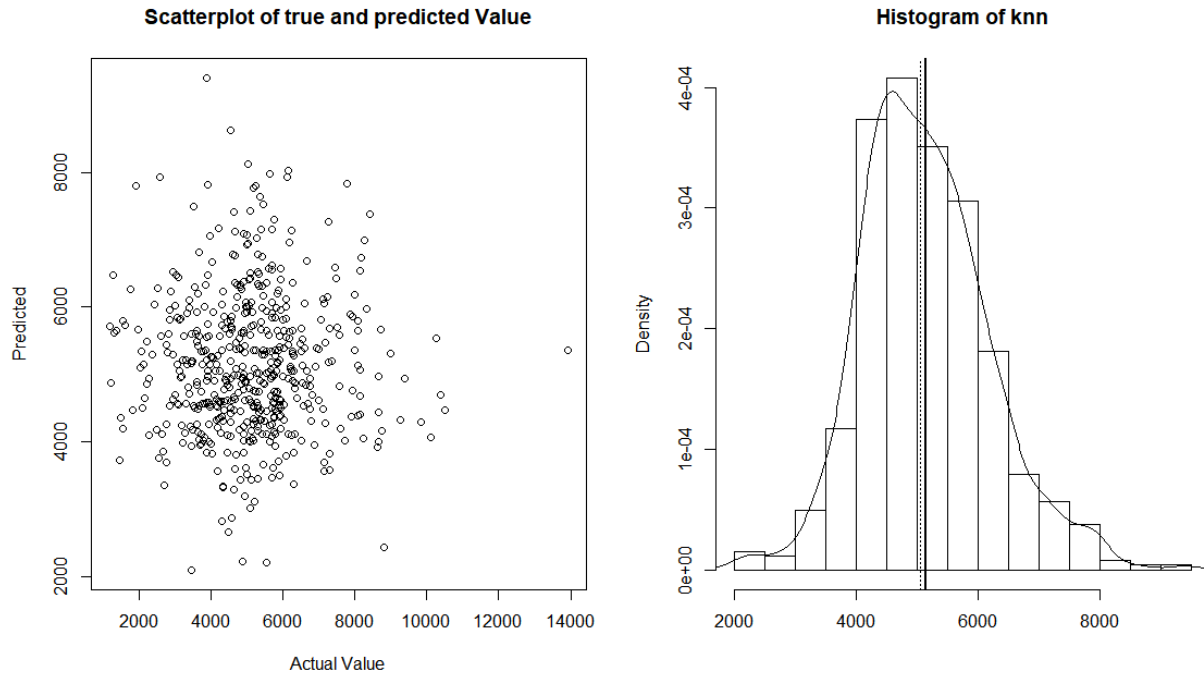


Figure 4.8: Plot and Histogram of KNN predicted values

4.1.6 Aggregated Results

The models were successful at predicting the target, yield value. RandomForest and KNN have close range RMSE values. In any prediction task, the probability of having differences between actual and predicted values is high. The difference in the two depends on models and how they handle data. Table 4.11 represents the models' results summary.

Model	RMSE	MSE
Rpart	2074.237	4302458
MARS	2436.181	5934979
LASSO	2222.91	4941329
RandomForest	1949.539	3800702
Knn	1950.561	3804686

Table 4.11: Models Results Summary

The boxplot visualized each model performance. The top and bottom lines of the rectangle are the 3rd and 1st quartiles respectively. 2nd quartile is the line that divides the box into two parts. The length of the rectangle from top to bottom is the interquartile range. The upper dash denotes maximum value or 3rd quartile plus 1.5 times the interquartile range, whichever is smaller and lower dash denote minimum value or the 1st quartile minus 1.5 times the interquartile range, whichever is larger.

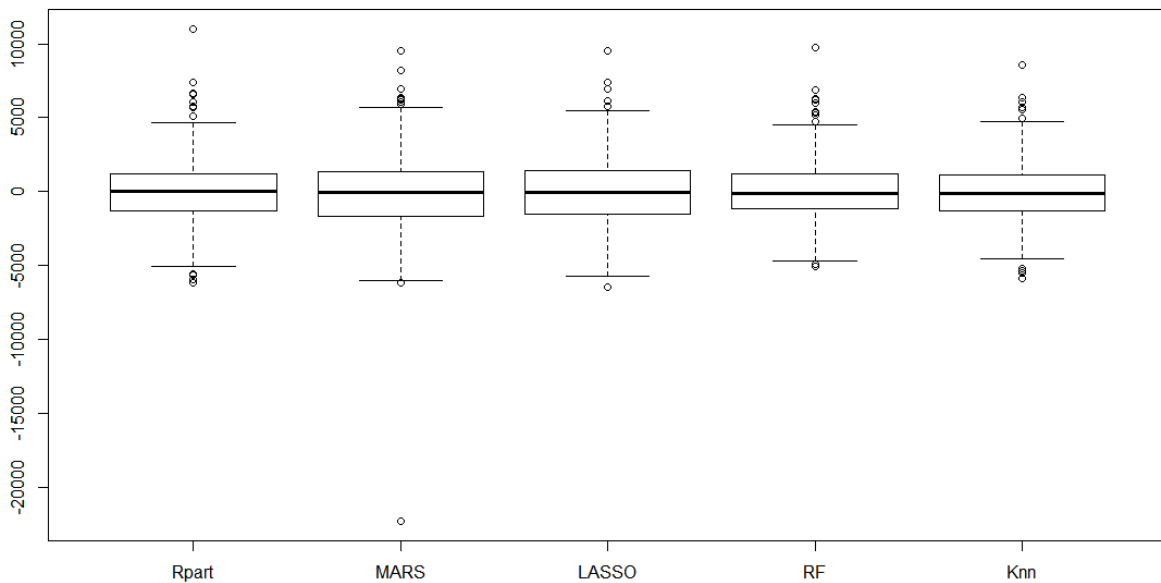


Figure 4.9: Box plot for the models.

Observations about the box plot

- All the models have a high level of agreement with each other.

- All have median at same level
- All have low and high observations often known as outliers compared to all others. MARS has extremely low outliers compared to other models
- Rpart, RF and Knn have almost the same maximum and minimum values, and MARS and LASSO have the same maximum and minimum values.

4.2. Represent a bag with the average predicted value of observations.

This section describes the results obtained from the exploration of the five models mentioned above on the dataset sample containing a total of 4801 numerical values using the second MIL approach of predict a value for each observation in N vary bag, then average the predicted values to represent the bag. The dataset sample is a multiple 10 of the other sample, we supposed to have a total of 5300 observations but it was impossible to achieve this due to eliminating missing values, which also caused us to have vary size of observations in each bag. Each predictive result is a prediction of vary n sizes of observations with 10 as maximum observation. This approach considers relevance with respect to instances in a bag, it uses the weighted average of a bag content to represent it (Amar et al., n.d.; Marc-Andre et al. 2016). This method also works in two stages. First do a prediction for each observation in a bag, the second stage is to average the predicted values in the bag, and represent the bag with the value. Average predicted value of each bag i from the dataset represents the bag.

In this yield prediction task, observations with same serial number were grouped under same bag (that is, serial numbers were used to represent bags). Prediction was done for each row in a serial number and average of the result was calculated to represent the serial number.

Same as section 1, all models were constructed under the same conditions using all other variables present in the data to predict yield value except for different packages which suit distinct models. Each model prediction values are presented with the RMSE and MSE values.

4.2.1. Rpart

Has light tailed distribution skew at both ends. The deviations from the straight line are minimal, it indicates normal distribution. No predictor transformations were done while generating the model, createFolds function in caret was used to split the data into 10 groups. The model predicted value ranges from 2338 to 8911. Table 4.12 presents rpart summary result and figure 4.10 is three graphs two on the upper and one lower figures to graphically display rpart model performance results. The first side on upper part is scatter plot of predicted against actual values, the second side is density plot of predicted value and the lower is the normal q-q plot of the model predicted value distribution.

MSE	RMSE
3338750	1827.225

Table 4.12: rpart result for the second MIL approach.

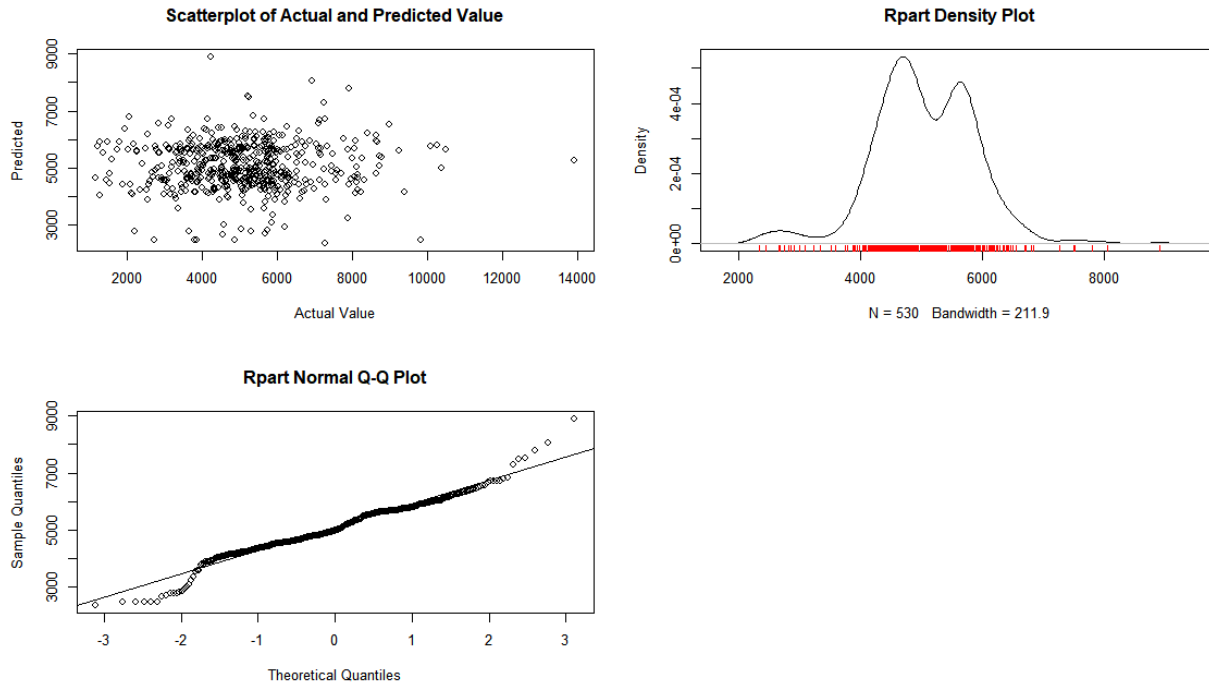


Figure 4.10: rpart plot for the second MIL approach.

4.2.2. MARS

The model is skew at both ends with light tailed. The deviations from the straight line are minimal. Minimum value predicted is 597.6 and maximum value predicted is 8896.1. Table 4.13 shows the result summary and figure 4.11 illustrates the model analysis.

MSE	RMSE
3772796	1942.369

Table 4.13: MARS result for the second MIL approach.

The first side on upper part of the plot below is scatter plot of predicted against actual values, the second side is density plot of predicted value and the lower part is the normal q-q plot of the model predicted value distribution.

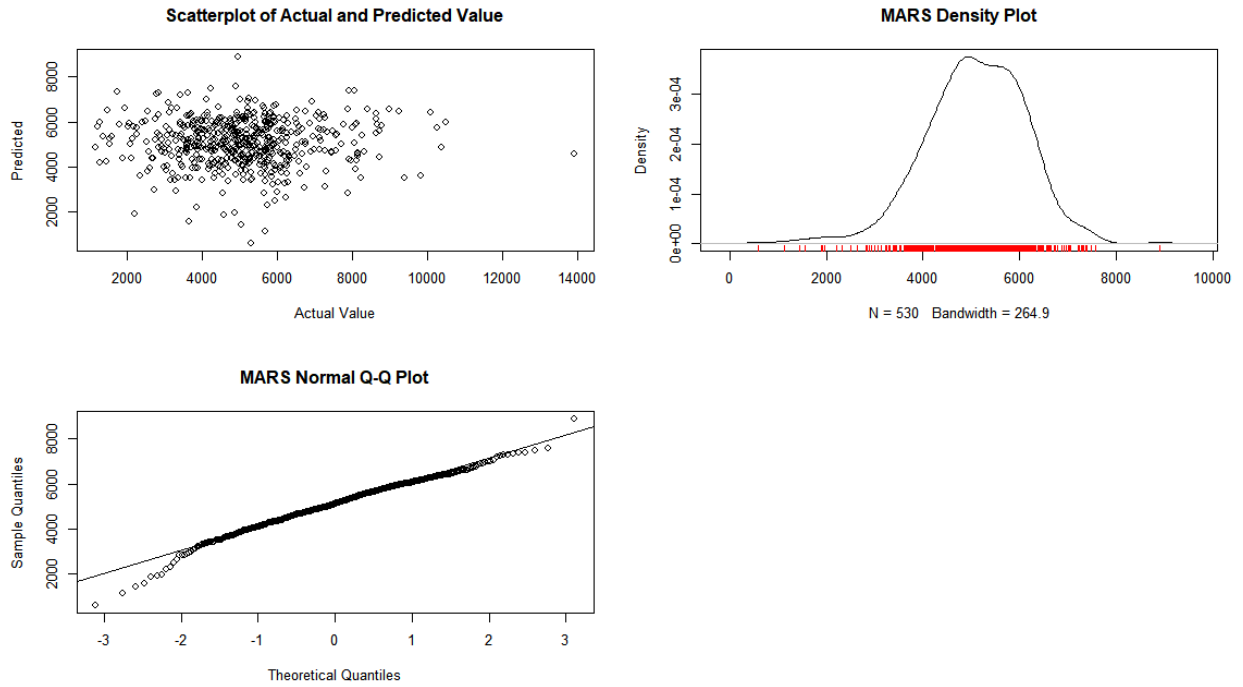


Figure 4.11: MARS plot for the second MIL approach

4.2.3. LASSO

The result is heavy tailed and right skewed. The predicted value has a range between negative 3043 to positive 7940 with 5000 as highest predicted value. Table 4.15 presents model result and figure 4.12 shows the scatter plot of predicted against actual values, density plot and normal q-q plot of predicted value.

MSE	RMSE
3013958	1736.075

Table 4.14: LASSO result for the second MIL approach.

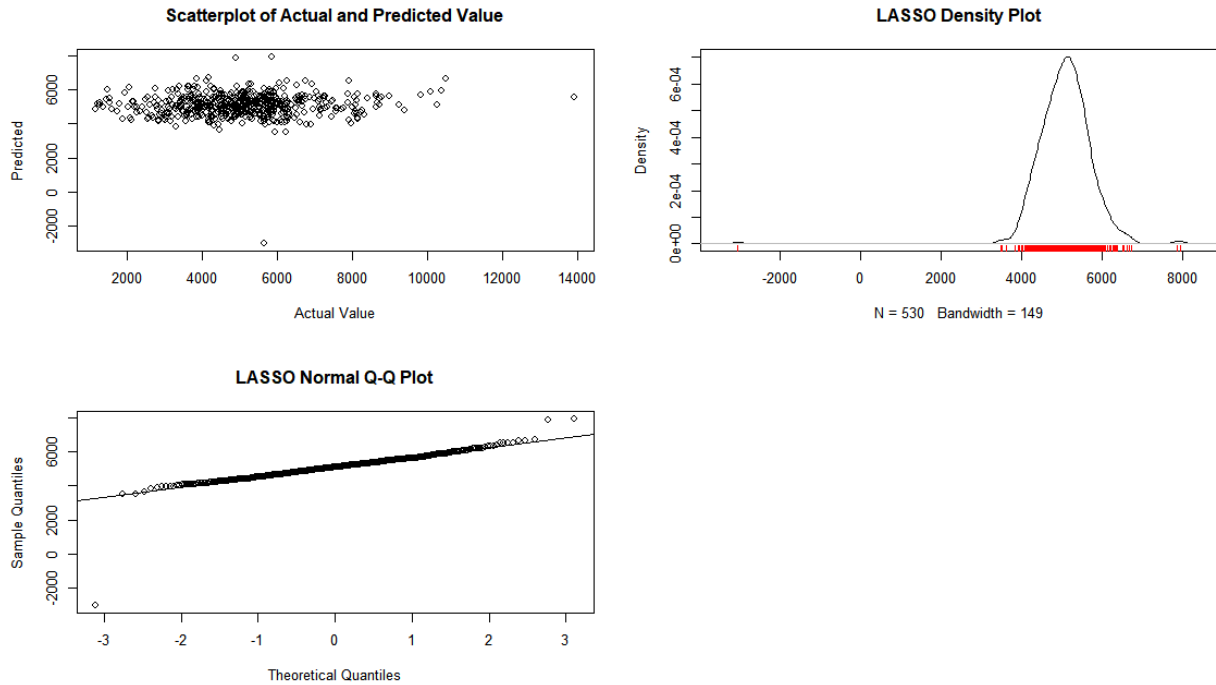


Figure 4.12: LASSO plot for the second MIL approach.

4.2.4. Random Forest

The model is skewed at the right side. The deviations from the straight line are minimal. This indicates normal distribution. Has range of 2892 to 8300 predicted value with peak at 5000. Table 4.15 presents result and figure 4.13 shows the scatter plot of predicted against actual values, density plot and normal q-q plot of predicted value.

MSE	RMSE
3192029	1786.625

Table 4.15: RF result for the second MIL approach

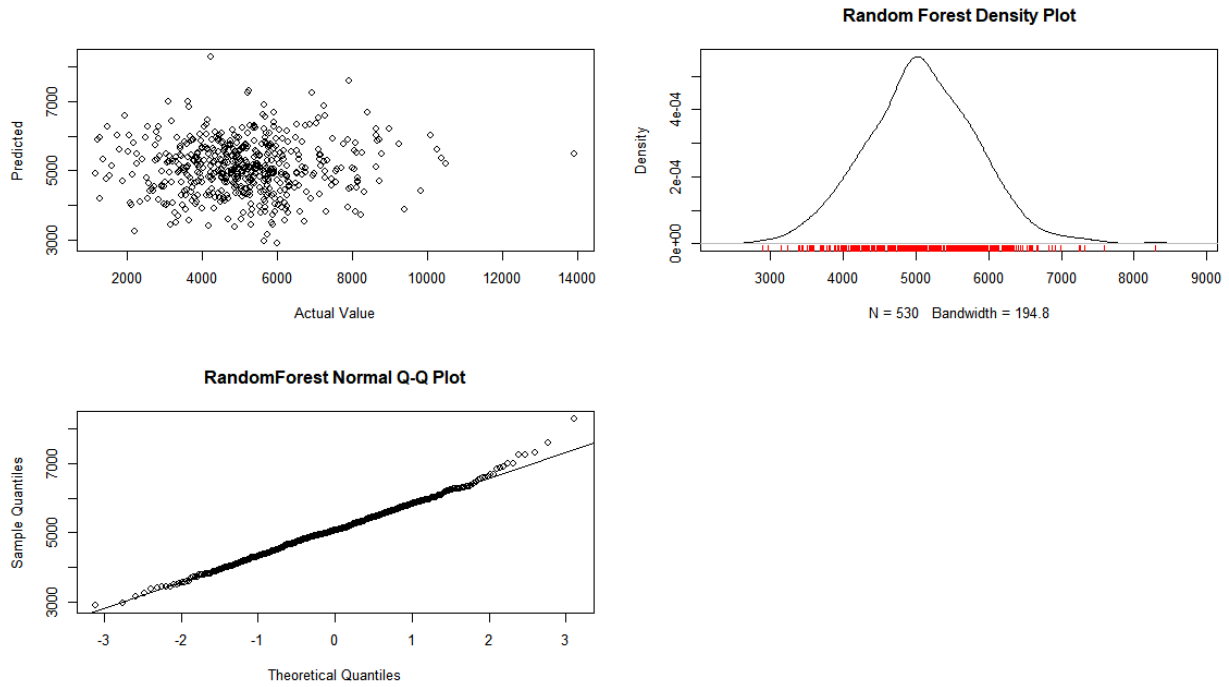


Figure 4.13: RF plot for second MIL approach

4.2.5. KNN

Result has right skew and normally distributed. Highest predicted value at 5000 with range between 3544 to 9091. Table 4.16 presents result and the result is illustrated in figure 4.14. The first side on upper part is scatter plot of predicted against actual values, the second side is density plot of predicted value and the lower is the normal q-q plot of the model predicted value distribution.

MSE	RMSE
2904267	1704.191

Table 4.16: KNN result for the second MIL approach

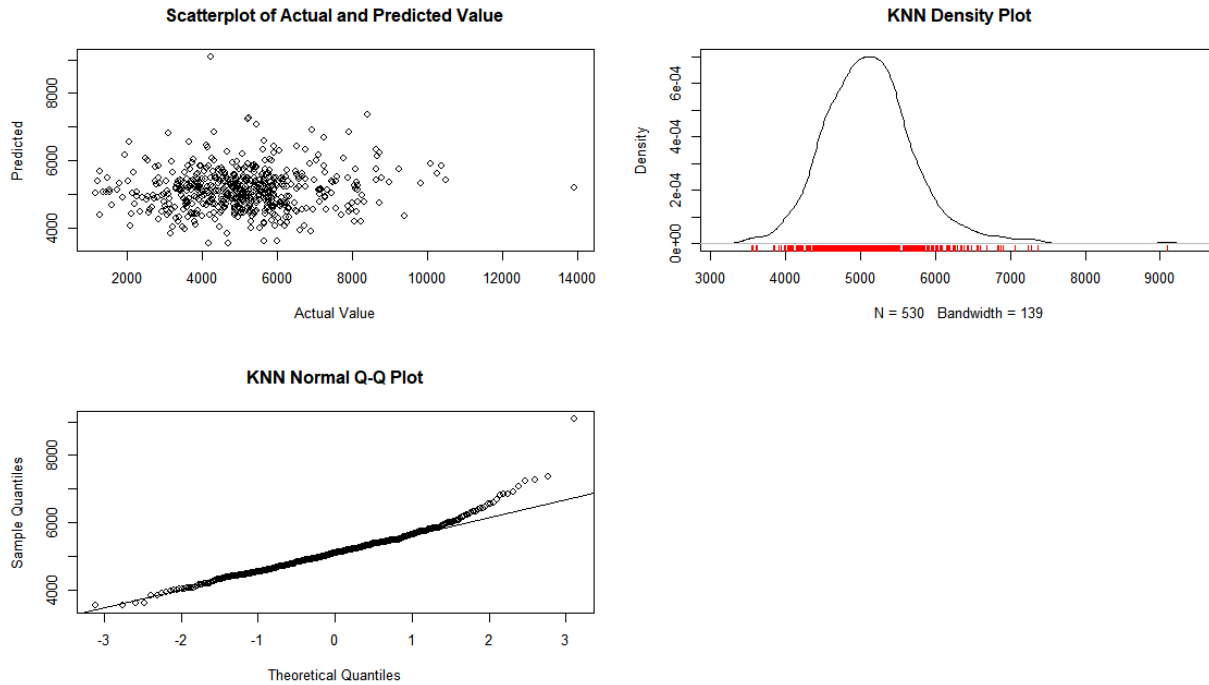


Figure 4.14: KNN plot for second MIL approach

4.3. Comparison of Results

This section compares the performance of individual model on the two approaches of whether to first calculate average of each trait in a bag then predict a single value for the bag or predict values for each row in the bag then calculate the average to represent the bag, and the actual values.

4.3.1 Rpart

Table 4.17 is rpart model MSE and RMSE values on the two MIL approaches of whether to calculate average of each trait in a bag then predict a single value for the bag or predict values for each row in the bag then calculate the average to represent the bag. Figures 4.15 and 4.16 present box plot and histogram respectively comparing the predicted values of rpart model on the two MIL approaches together with the actual yield value.

Approach	MSE	RMSE
Single Instance	4302458	2074.237
Multiple Instances	3338750	1827.225

Table 4.17: Rpart model results of both MIL approaches

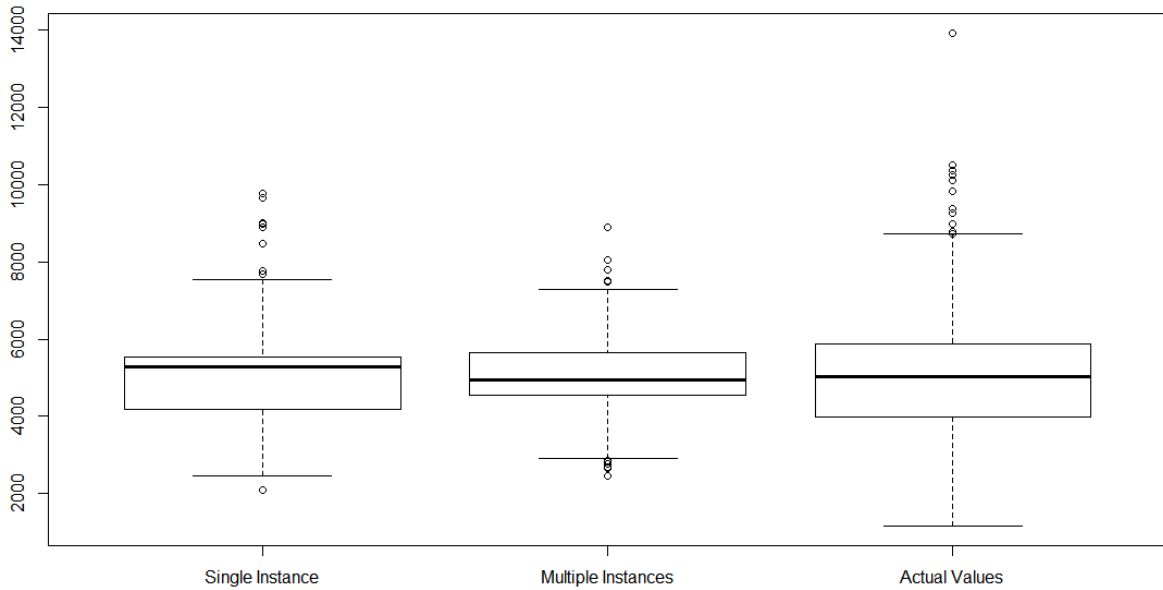


Figure 4.15: Box plot of Rpart prediction of both MIL approaches and actual value

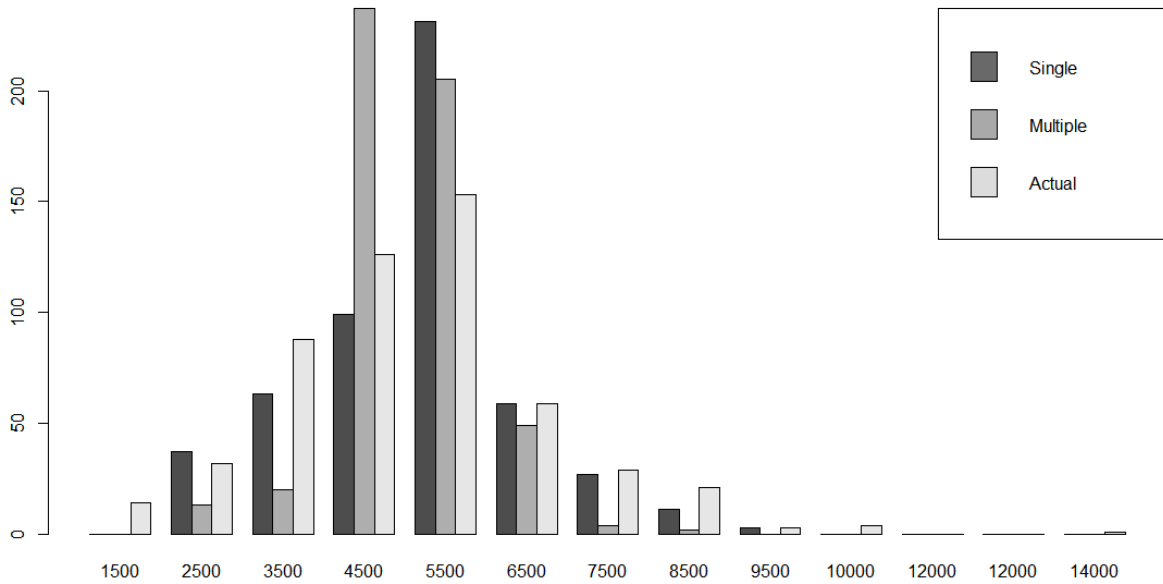



Figure 4.16: Rpart comparison histogram

Single = First Approach; average each trait, then predict

Multiple = Second Approach; predict for each observation then average the value to represent a bag

Actual  = Actual value of the target which is yield

Comparison

- Rpart median on represent a labeled bag with average predicted value of some instances and the median of the actual values are at the same level.
- The box plots are not of the same size, this suggests a difference between rpart performance on the two approaches and the actual values.
- Box plots of both single and multiple are short with multiple having shorter box plot, this shows that there are more similar predicted values in the second approach.
- Single instance RMSE error is 2074.237 while that of multiple is 1827.225

It can be concluded that rpart performs better on represent a labeled bag with the average predicted value of some instances.

4.3.2 MARS

This section describes MARS model on the two MIL approaches; of whether to calculate average of each trait in a bag then predict a single value for the bag or predict values for each row in the bag then calculate the average to represent the bag, and the actual values. Table 4.18 shows the model performance MSE and RMSE values. Figure 4.17 is a box plot and figure 4.18 is histogram comparing the predicted values of MARS model on the two MIL approaches together with the actual yield value

Approach	MSE	RMSE
Single Instance	5934979	2436.181
Multiple Instances	3772796	1942.369

Table 4.18: MARS model results of both MIL approaches

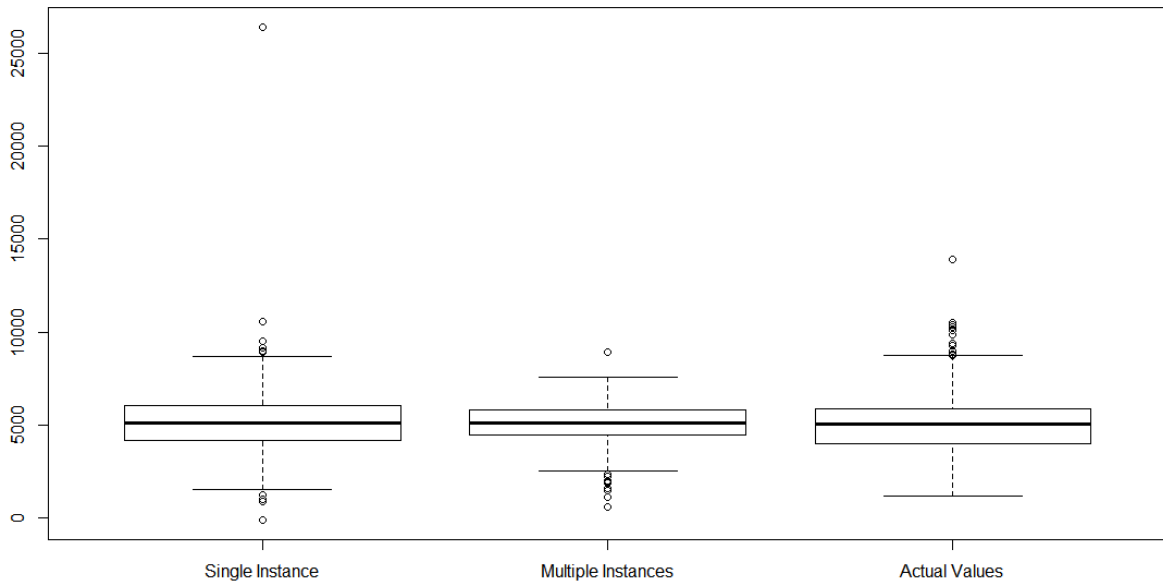


Figure 4.17: MARS comparison box plot

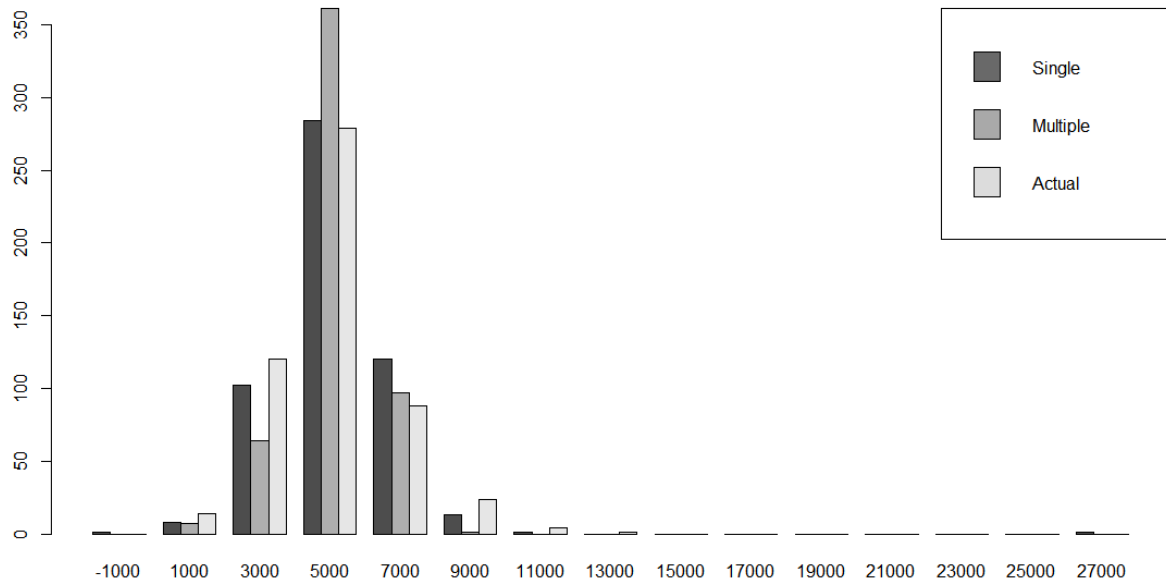




Figure 4.18: MARS comparison histogram

Single = First Approach; average each trait, then predict

Multiple  = Second Approach; predict for each observation then average the value to represent a bag

Actual  = Actual value of the target which is yield

Comparison

- All the three have same median
- The four sections of the boxplots are uneven, the long upper whisker means values are varied amongst the most positive quartile group, and very similar for the least positive quartile group.
- No much differences between box plots.
- Single instance RMSE error is 2436.181 while that of multiple is 1942.369

4.3.3 LASSO

This subsection describes performance of the model on whether to calculate average of each trait in a bag then predict a single value for the bag or predict values for each row in the bag then calculate the average to represent the bag, and the actual values. Table 4.19 gives a summary of LASSO performance on both MIL approaches. Single instance represents first approach while multiple instances represents second approach. Figures 4.19 and 4.20 present box plot and histogram comparing the predicted values of LASSO model on the two MIL approaches together with the actual yield value

Approach	MSE	RMSE
Single Instance	4941329	2222.91
Multiple Instances	3013958	1736.075

Table 4.19: LASSO model results of both MIL approaches.

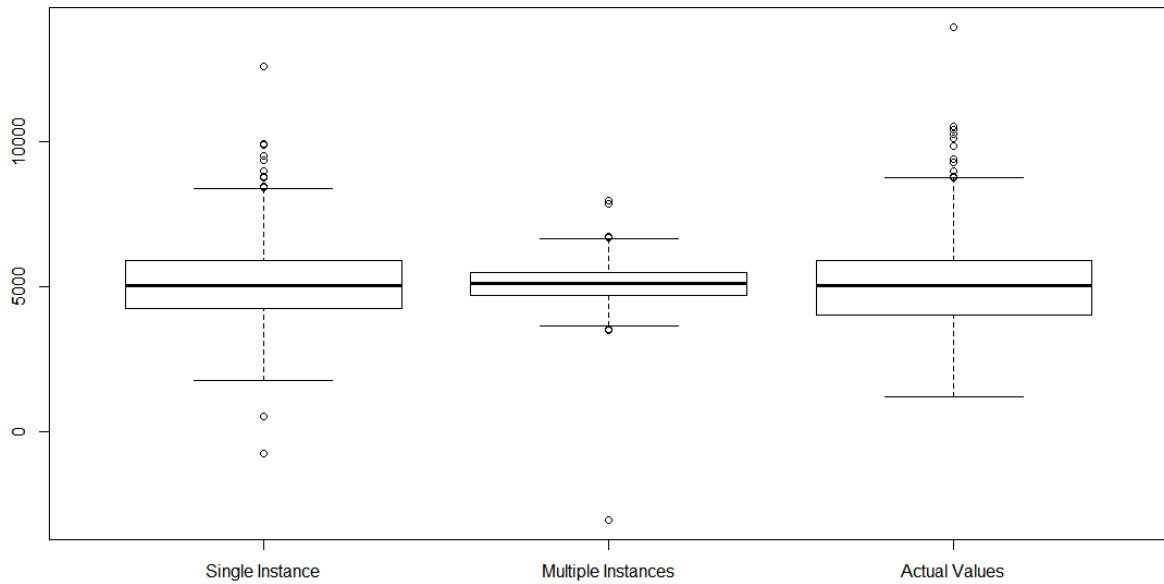


Figure 4.19: LASSO comparison box plot

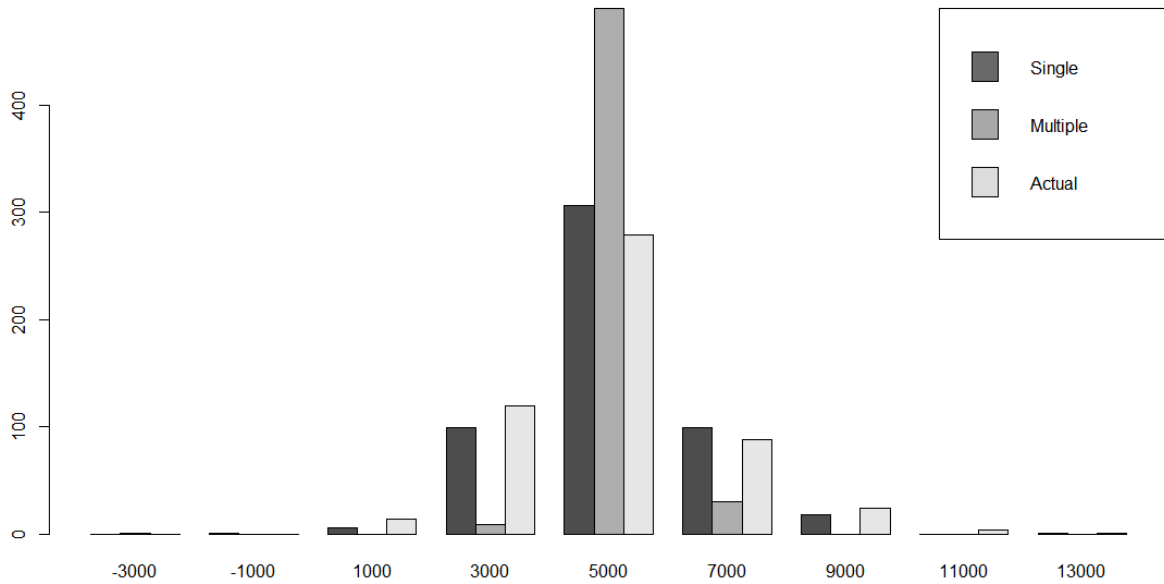


Figure 4.20: LASSO comparison histogram

Single = First Approach; average each trait, then predict

Multiple = Second Approach; predict for each observation then average the value to represent a bag

Actual = Actual value of the target which is yield

Comparison

- Single instance RMSE error is 2222.91 while that of multiple is 1736.075
- The medians divide the 3rd and 1st quartiles in to two.
- Values have a high level of agreement with each.
- There are more similar predicted values.
- The medians are at the same level.

4.3.4 Random Forest

This subsection compares random forest performance on the two approaches; whether to calculate average of each trait in a bag then predict a single value for the bag or predict values for each row in the bag then calculate the average to represent the bag, and the actual values. Table 4.20 presents the summary results of the two approaches. Figures 4.21 and 4.22 present plots comparing the predicted values of random forest model on the two MIL approaches together with the actual yield value

Approach	MSE	RMSE
Single Instance	3800702	1949.539
Multiple Instances	3192029	1786.625

Table 4.20: RF model results of both MIL approaches.

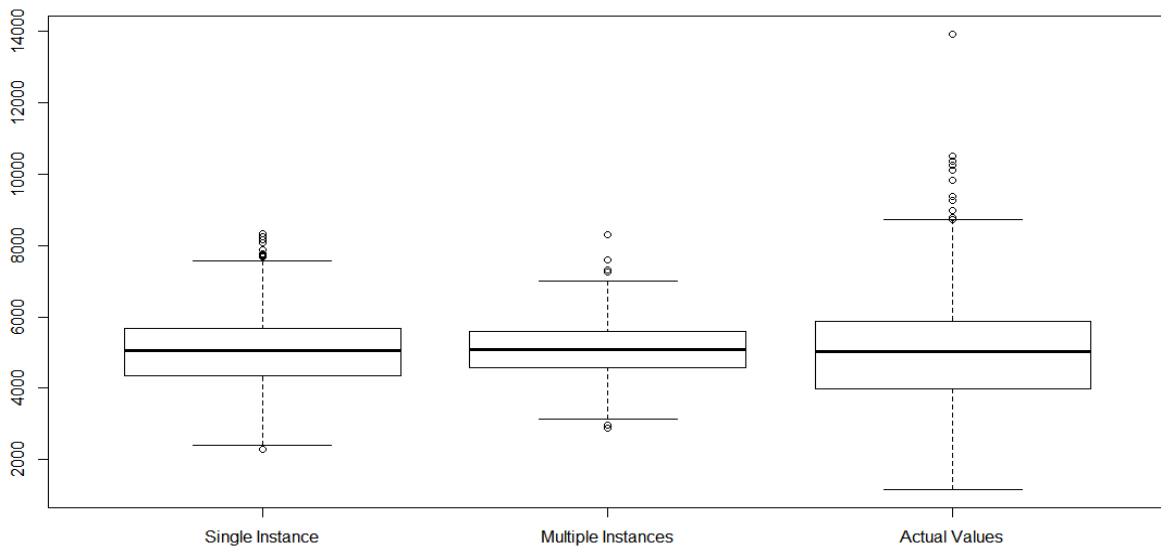


Figure 4.21: RF comparison box plot

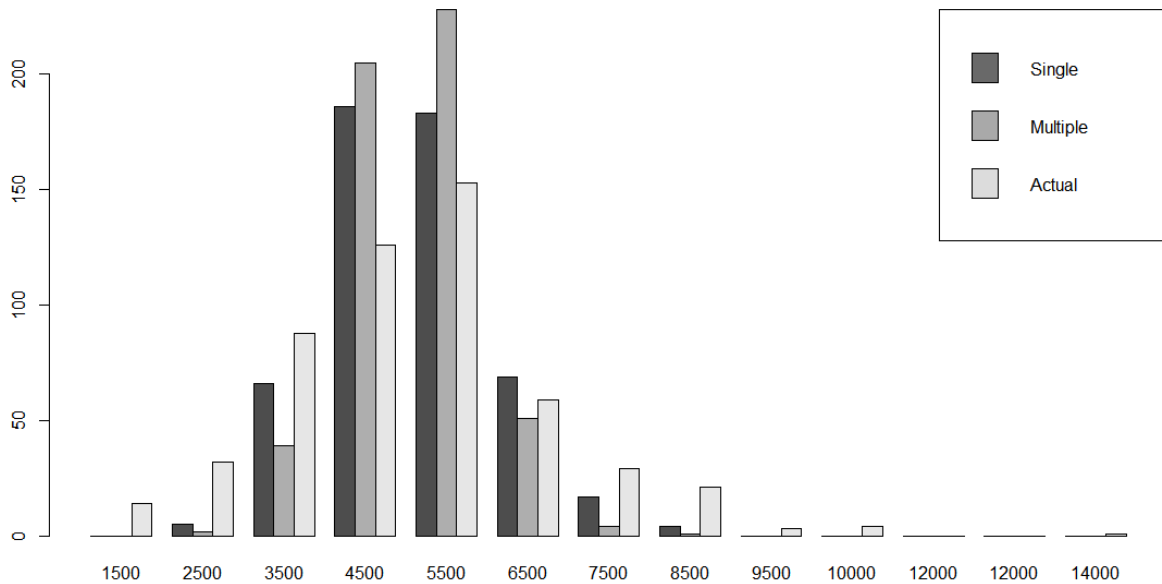


Figure 4.22: RF comparison histogram

Single = First Approach; average each trait, then predict

Multiple = Second Approach; predict for each observation then average the value to represent a bag

Actual = Actual value of the target which is yield

Comparison

- Single instance RMSE error is 1949.539 while that of multiple is 1786.625
- Values have a high level of agreement with each other in second approach.
- The medians are at the same level.

Using RMSE value, random forest performance is better on second approach than first approach.

4.3.5 K-NN

The subsection compares knn predicted value on the two approaches of whether to first calculate average of each trait in a bag then predict a single value for the bag or predict values for each row in the bag then calculate the average to represent the bag, and the actual values. Table 4.21 is knn result on the two approaches. Figures 4.23 and 4.24 present plots comparing the predicted values of the model on the two MIL approaches together with the actual yield value.

Approach	MSE	RMSE
Single Instance	3804686	1950.561
Multiple Instances	2904267	1704.191

Table 4.21: KNN model results of both MIL approaches

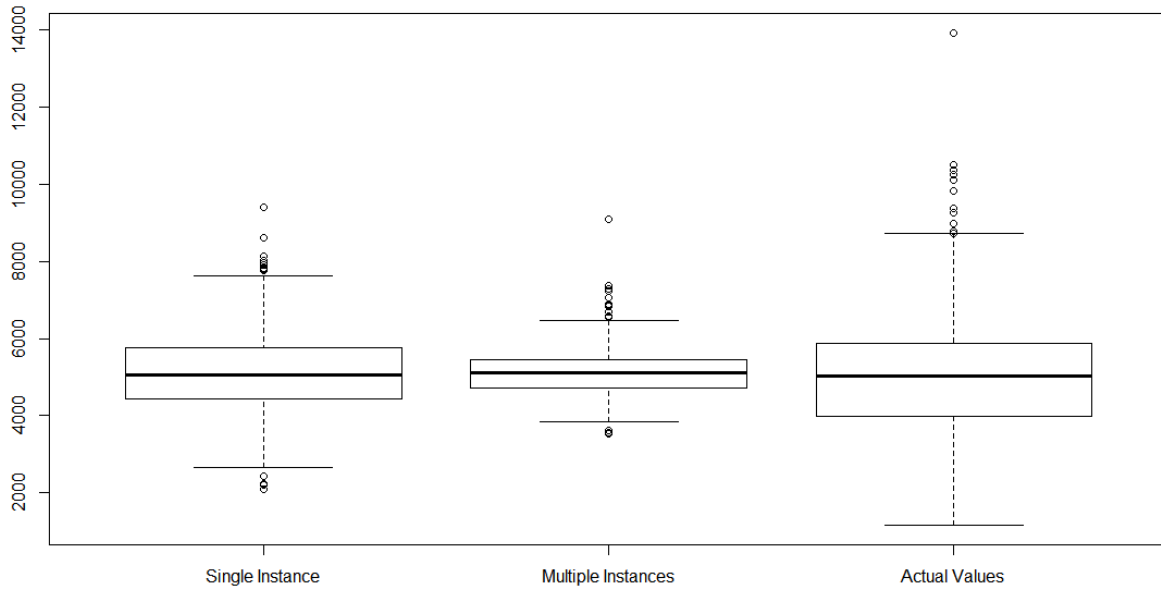


Figure 4.23: K-NN comparison box plot

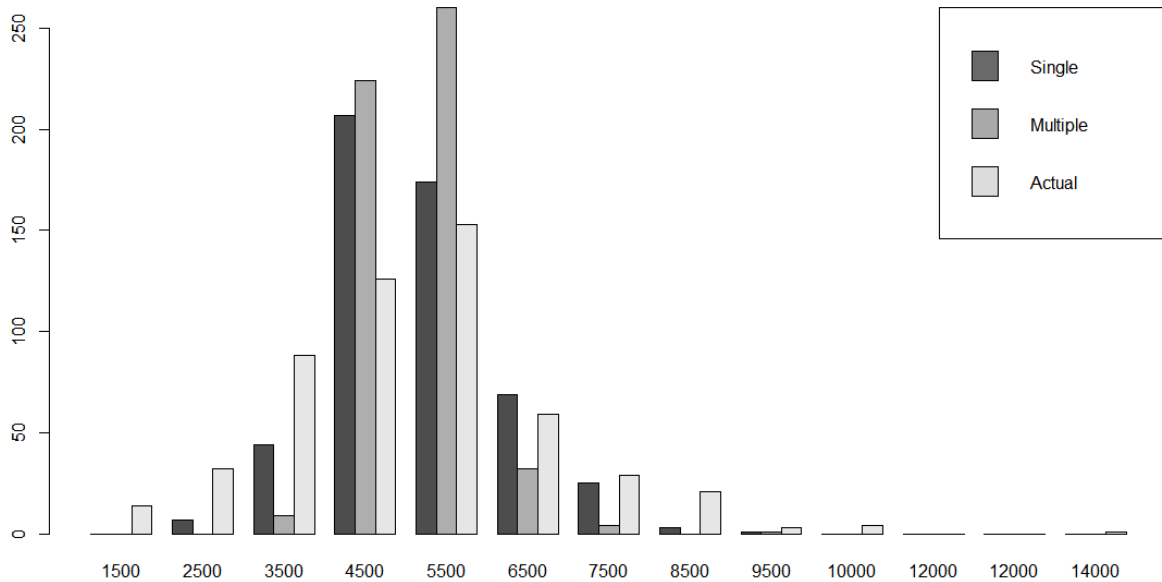


Figure 4.24: K-NN comparison histogram

Single = First Approach; average each trait, then predict

Multiple = Second Approach; predict for each observation then average the value to represent a bag

Actual = Actual value of the target which is yield

Comparison

- Single instance RMSE error is 1950.561 while that of multiple is 1704.191
- Values have a high level of agreement with each other in second approach.
- The medians are at the same level.

Using RMSE as the evaluation measure, KNN performance is better on second approach than first approach.

		MSE	RMSE	Min	1 st Quartile	Median	Mean	3 rd Quartile	Max
Rpart	Single	4302458	2074.2	2090	4175	5269	5083	5531	9761
	Multiple	3338750	1827.2	2338	4554	4971	5062	5660	8911
MARS	Single	5934979	2436.2	-100.8	4147	5112	5127	6030	26400
	Multiple	3772796	1942.4	597.8	4433	5113.6	5076.8	5815.9	8896.1
LASSO	Single	4941329	2222.9	-751.9	4221	5020	5076	5884	12580
	Multiple	3013958	1736.1	-3043	4689	5089	5079	5466	7940
RF	Single	3800702	1949.5	2294	4358	5057	5088	5682	8325
	Multiple	3192029	1786.6	2892	4570	5072	5085	5587	8300
KNN	Single	3804686	1950.6	2103	4442	5048	5138	5750	9394
	Multiple	2904267	1704.2	3544	4726	5106	5127	5452	9091
Actual				1172	3990	5035	5070	5887	13919

Table 4.22: Summary of all models predictive results on both MIL approaches and Actual value

5 Conclusion

Two prediction topics of MIL approaches for assigning real values were analyzed: assign the value of the instance closest to the target, and represent a labeled with average value of some instances. The task was done using a combined predictive modelling approach to predict yield value.

Two experimental setups were carried out for the two approaches. In each setup, different datasets were used and five experiments were conducted for each. Each of the five experiments represent the five predictive models used. No modification of dataset for any model, all models were constructed under the same conditions using all other variables present. RMSE (Root Mean Squared Error) was used to measure the performance of all models. There are differences in the values obtained due to differences in the models.

For the first approach, assign the value of the instance closest to the target,

- RF and KNN models proved to achieve best results with RMSE value of 1949.539 and 1950.561 respectively.

In the second approach, represent a labeled with average value of some instances.,

- KNN model achieved the best result, presenting RMSE value of 1704.191.
- Though all the models achieved better results in the second approach.

We can conclude that K-NN model performs better than all other models.

Learning algorithms can offer cost saving and efficiency to yield prediction due to them being more reliable and error free. Analysts should direct their focus to agricultural sector to grow the sector and this will be of great benefit to all.

5.1. Future Work

This dissertation focused on first two MIL Predictive analytics approaches of assigning real value to regression task problem. For future work,

- Focus should be on another approach to assign value to a bag.
- Predictive analysis application to other crop types in NUMI.
- Improvement of the dataset for future work, there were many missing values.

References

- Amar, Robert A, Daniel R Dooly, Sally A Goldman, and Qi Zhang. n.d. “Multiple- Instance Learning of Real- Valued Data.”
- Babenko, B. 2008. “Multiple Instance Learning: Algorithms and Applications.” *View Article PubMed/NCBI Google Scholar*, 1–19.
http://vision.ucsd.edu/~bbabenko/data/bbabenko_re.pdf%5Cnpapers3://publication/uuid/2CDB4FD4-9E25-4F12-826C-E67049137B7C.
- BALSHI, MICHAEL S, A DAVID McGUIRE, PAUL DUFFY, MIKE FLANNIGAN, JOHN WALSH, and JERRY MELILLO. 2009. “Assessing the Response of Area Burned to Changing Climate in Western Boreal North America Using a Multivariate Adaptive Regression Splines (MARS) Approach.” *Global Change Biology* 15 (3). Blackwell Publishing Ltd: 578–600. doi:10.1111/j.1365-2486.2008.01679.x.
- Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. 2000. “Crisp-Dm 1.0.” *CRISP-DM Consortium*, 76. doi:10.1109/ICETET.2008.239.
- Chou, Shieu-Ming, Tian-Shyug Lee, Yuehjen E Shao, and I-Fei Chen. 2004. “Mining the Breast Cancer Pattern Using Artificial Neural Networks and Multivariate Adaptive Regression Splines.” *Expert Systems with Applications* 27 (1): 133–42. doi:<https://doi.org/10.1016/j.eswa.2003.12.013>.
- Crino, S, and D E Brown. 2007. “Global Optimization With Multivariate Adaptive Regression Splines.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. doi:10.1109/TSMCB.2006.883430.
- Dietterich, T G, R H Lathrop, and T LozanoPerez. 1997. “Solving the Multiple Instance Problem with Axis-Parallel Rectangles.” *Artificial Intelligence* 89 (1–2): 31–71. doi:Doi 10.1016/S0004-3702(96)00034-3.
- Frausto-Solis, Juan, Alberto Gonzalez-Sanchez, and Monica Larre. 2009. “A New Method for Optimal Cropping Pattern.” In *MICAI 2009: Advances in Artificial Intelligence: 8th Mexican International Conference on Artificial Intelligence, Guanajuato, M{é}xico, November 9-13, 2009. Proceedings*, edited by Arturo Hernández Aguirre, Raúl Monroy Borja, and Carlos Alberto Reyes García, 566–77. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-05258-3_50.
- Fritsche-Neto, Roberto, Rafael Augusto Vieira, Carlos Alberto Scapim, Glauco Vieira Miranda, and Luciano Moreira Rezende. 2012. “Updating the Ranking of the Coefficients of Variation from Maize Experiments.” *Acta Scientiarum - Agronomy* 34 (1): 99–101. doi:10.4025/actasciagron.v34i1.13115.
- González-sanchez, Alberto, and Juan Frausto-solis. 2014. “Predictive Ability of Machine Learning Methods for Massive Crop Yield Prediction,” no. June. doi:10.5424/sjar/2014122-4439.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. “The Elements of Statistical Learning Data Mining, Inference and Prediction.pdf.” *Springer Series in Statistics*.
- Herrera, Francisco, Sebastián Ventura, Rafael Bello, Chris Cornelis, Amelia Zafra, Dánel Sánchez-Tarragó, and Sarah Vluymans. 2016. “Multi-Instance Regression.” In *Multiple Instance Learning: Foundations and Algorithms*, 127–40. Cham: Springer International Publishing. doi:10.1007/978-3-319-47759-6_6.
- Hnin, H M, W P Pa, and Y K Thu. 2017. “Back-Propagation Neural Network Approach to

- Myanmar Part-of-Speech Tagging.” *Advances in Intelligent Systems and Computing*. doi:10.1007/978-3-319-48490-7_25.
- International Business Machines Corporation USA. 2013. “IBM SPSS Modeler CRISP-DM Handbuch.”
- James, Foulds, and Frank Eibe. 2004. “A Review of Multiple-Instance Assumptions.” *The Knowledge Engineering Review* 0 (January): 1–24. doi:10.1017/S0000000000000000.
- Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. Springer New York Heidelberg Dordrecht London. doi:10.1007/978-1-4614-6849-3.
- L, Francisco Xavier, Kenneth J Boote, Benigno Ru, and Federico Sau. 2005. “Testing CERES-Maize Versions to Estimate Maize Production in a Cool Environment” 23: 89–102. doi:10.1016/j.eja.2005.01.001.
- Leathwick, J. R., D. Rowe, J. Richardson, J. Elith, and T. Hastie. 2005. “Using Multivariate Adaptive Regression Splines to Predict the Distributions of New Zealand’s Freshwater Diadromous Fish.” *Freshwater Biology* 50 (12): 2034–52. doi:10.1111/j.1365-2427.2005.01448.x.
- Lee, Tian-Shyug, and I-Fei Chen. 2005. “A Two-Stage Hybrid Credit Scoring Model Using Artificial Neural Networks and Multivariate Adaptive Regression Splines.” *Expert Systems with Applications* 28 (4): 743–52. doi:https://doi.org/10.1016/j.eswa.2004.12.031.
- Marc-Andre, Carbonneau, Granger Eric, Gagnon Ghyslain, and Cheplygina Veronika. 2016. “Multiple Instance Learning A Survey of Problem Characteristics and Applications.pdf.”
- Mendes-Moreira, Pedro M R, João Mendes-Moreira, António Fernandes, Eugénio Andrade, Arnel R Hallauer, Silas E Pêgo, and M C Vaz Patto. 2014. “Is Ear Value an Effective Indicator for Maize Yield Evaluation?” *Field Crops Research* 161: 75–86. doi:10.1016/j.fcr.2014.02.015.
- Navot, Amir, Lavi Shpigelman, Naftali Tishby, and Eilon Vaadia. 2006. “Nearest Neighbor Based Feature Selection for Regression and Its Application to Neural Activity.” *Advances in Neural Information Processing Systems* 18: 995.
- Nemati, Ali, Mohammad Sedghi, Rauf Seyed Sharifi, and Mir Naser Seiedi. 2009. “Investigation of Correlation between Traits and Path Analysis of Corn (*Zea Mays* L.) Grain Yield at the Climate of Ardabil Region (Northwest Iran).” *Notulae Botanicae Horti Agrobotanici Cluj-Napoca* 37 (1): 194–98.
- Pappas, Nikolaos, Nikolaos Pappas, and Andrei Popescu-belis. 2015. “Explaining the Stars : Weighted Multiple- Instance Learning for Aspect-Based Sentiment Analysis Explaining the Stars : Weighted Multiple-Instance Learning for,” no. January 2014. doi:10.3115/v1/D14-1052.
- Pappas, Nikolaos, and Andrei Popescu-belis. 2017. “Explicit Document Modeling through Weighted Multiple-Instance Learning Explicit Document Modeling through Weighted,” no. February.
- Ratan, A L, O Maron, W E L Grimson, and T Lozano-Perez. 1999. “A Framework for Learning Query Concepts in Image Classification.” doi:10.1109/CVPR.1999.786973.
- Ray, Soumya. 1999. “Multiple Instance Regression.”
- Ruß, Georg. 2009. “Data Mining of Agricultural Yield Data: A Comparison of Regression Models.” In *Advances in Data Mining. Applications and Theoretical Aspects: 9th Industrial Conference, ICDM 2009, Leipzig, Germany, July 20 - 22, 2009. Proceedings*, edited by Petra Perner, 24–37. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-03067-3_3.

- Svetnik, Vladimir, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. 2003. "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling." *Journal of Chemical Information and Computer Sciences* 43 (6): 1947–58. doi:10.1021/ci034160g.
- Tibshirani, Robert. 1996. "Regression Selection and Shrinkage via the Lasso." *Journal of the Royal Statistical Society B*. doi:10.2307/2346178.
- Tong, Simon, and Daphne Koller. 2001. "Support Vector Machine Active Learning with Application to Text Categorization.pdf." *Journal of Machine Learning Research (2001)* 45-66.
- Wagstaff, Kiri L, and T Lane. 2007. "Salience Assignment for Multiple-Instance Regression."
- Wagstaff, Kiri L, and Alex Roper. 2008. "Multiple-Instance Regression with Structured Data." *Proceedings of the 4th International Workshop on Mining Complex Data*.
- Wang, Zhuang, and Bo Han. n.d. "Aerosol Optical Depth Prediction from Satellite Observations," 165–76.
- Wang, Zhuang, Liang Lan, and Slobodan Vucetic. 2011. "Mixture Model for Multiple Instance Regression and Applications in Remote Sensing," 1–12.
- Wieringen, Wessel Van. n.d. "Lasso Regression." *Department of Epidemiology and Biostatistics, VUmc & Department of Mathematics, VU University Amsterdam, The Netherlands*.
- Wikipedia. 2014. "Maize History," 1–16. <https://en.wikipedia.org/wiki/Maize>.
- . 2016. "Cross Industry Standard Process for Data Mining," 1–3. doi:10.1017/S0269888906000737.
- Yang, Jun. 2005. "Review of Multi-Instance Learning and Its Applications." *Technical Report, School of Computer Science Carnegie Mellon University*.
- Zeileis, Achim, Torsten Hothorn, and Kurt Hornik. 2017. "Model-Based Recursive Partitioning Model-Based Recursive Partitioning" 8600 (June). doi:10.1198/106186008X319331.
- Zhou, Zhi-Hua. 2004. "Multi-Instance Learning: A Survey." *Technical Report, Nanjing University*, 32.