IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 90-98
www.iosrjournals.org

Robustness of Predictive Data Mining Methods under the Presence of Measurement Errors in the Context of Production Processes

Daniela F. Dianda
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística, Fac. de Ciencias Económicas y Estadística. Universidad Nacional de Rosario-CONICET[1]. Argentina.

Abstract: *One of the main objectives of data analysis in industrial contexts is prediction, that is, to identify a function that allows predicting the value of a response from the values of other variables considered as potential predictors of this outcome. The large volumes of data that current technology allows to generate and store have made it necessary to develop methods of analysis alternative to the traditional ones to achieve this objective, which allow mainly to process these large amounts of information and to predict the response in real time. Enclosed under the name of Data Mining, many of these new methods are based on automatic algorithms mostly originated in the computer field. However, the quality of the information that feeds these procedures remains a key factor in ensuring the reliability of the results. With this premise, in this work we study the effect that the presence of faults in the measurement devices that originate the information to be analyzed, can cause on the predictive ability of one of the predictive methods of data mining, the decision trees. The results are compared with those obtained using one of the traditional statistical techniques: multiple linear regression. The results obtained indicate that the effect of measurement related errors on the predictive ability of decision trees, compared to traditional regression models, depends on the nature of the measurement error.*

Keywords: *CART decision trees; Linear regression; Measurement error; Prediction Error*

I. Introduction

Quality improvement (QI) of industrial products and processes requires collection and analyses of data to solve quality related manufacturing problems. With advances in automation and computer systems, data from manufacturing processes is becoming more and more available. Although traditional data analysis tools have been successfully used in improving quality of products and processes, now better tools exist to mine massive data sets collected through computerized systems in the industry.

Knowledge discovery in databases (KDD) is being successfully used for solving QI problems involving multivariate data in various stages of product/process life cycles. One of the earliest definitions of KDD is: "the non-trivial extraction of implicit, previously unknown, and potentially useful information from data" [1]. For many years, this attempt to find useful patterns in data has been given different names, including data mining (DM), knowledge extraction, information discovery, etc. With the passage of time, the concepts have been refined. Although today some authors still use KDD and DM as synonymous, we agree with the view of many other authors who state that KDD refers to the overall process of discovering useful knowledge from data, while DM refers to a step in this process that consists of applying data analysis and discovery algorithms to produce a particular enumeration of patterns or models from the data.

These methods arise as an attempt to deal with the need to process massive volumes of data, making indispensable to turn to automated methods of analysis. While in many fields automation has been considered as a substitute for human supervision, we considered that problem solving is a human process and DM should be seen as a very powerful tool that help us in the process of knowledge discovery. As in any other analysis technique, data quality is one of the vital factors to ensure good results. Although KDD process involves steps entirely dedicated to data preparation and data cleansing, there is one source of noise in data that can be avoided prior to data collection and storage: measurement errors. Most of the available data, mainly in manufacturing processes, come from measurements. Measurement systems have also evolved as a result of the advances in technologies. Nowadays, industries have complex measurement systems integrated to the whole production process that automatically scan and register many characteristics of the product at every phase of the process. In this context, any distortion in a measurement sensor will lead to a lot of "contaminated data" in a very short period of time. It has been proved that measurement errors can cause important distortions on the performance

[1] This paper was supported by the National Scientific and Technical Research Council of Argentina (CONICET), by means of a post-doctoral fellowship.

of many traditional statistical methods used in QI activities ([2] to [7]). Therefore, it is natural to wonder if measurement errors can also affect the performance of DM algorithms.

In this work we focus on one of the data mining methods with applications in manufacturing industries to handle QI problems involving prediction goals; the decision trees.This technique is used to model a real database coming from a concrete manufacturing process. The predictive performance of the model under the presence of both random and systematic measurement errors is analyzed and compared with the performance reached by using traditional multiple linear regression (MLR) analysis.

The rest of the paper is organized as follows. Section II gives a brief note on prediction methods for quality improvement. In section III the design of the study is detailed. This includes a brief summary of the techniques of analysis used in this work, a description of the real dataset used and its context, a review on the concepts of measurement systems and measurement errors and the specification of the steps of the analysis and the criteria assumed to conducted it. Section IV presents the results of the analysis and the conclusions and discussion are provided in section V.

II. Prediction in quality improvement

In a very general way, data mining methods can be divided in terms of the goal that they try to achieve. The two primary goals of data mining in practice used to be prediction and description. Although the boundaries between them are not so sharp, we considered as prediction goal related methods those that allow using some variables or fields in the database to predict the unknown future values of other variable.

One of the bases of the today's management strategies in industry is the ability to predict malfunctions within processes even before such situations occur. Advanced predictive models can be used to this purpose. Depending on whether the type of data to be predicted is continuous or categorical, predictive modeling can be performed building prediction or classification functions, respectively. According to several authors ([8] to [11]), three main approaches for building prediction/classification functions can be distinguished: statistical based methods, machine learning and neural networks. Within the statistical approach, relationships among variables in the data set are modeled under the assumption of an explicit underlying probability model. Machine learning and neural networks approaches belong to the group of non-parametric techniques, where no assumptions are made about any underlying distribution in the data. Under the machine learning umbrella, algorithms based on decision trees (DT) are of the most used for prediction tasks. They generates rules having "if-then" type structures, so prediction/classification results from following a sequence of logical steps. Finally, neural network (NN) algorithms were inspired by the way that biological systems like the human brain process the information. These algorithms build connections among sets of nodes, each node producing a nonlinear function of its input, in such a way that the complete network forms a structure in which each input variable or node is connected to one or more output nodes, leading to a very complex system of inter-dependencies.

Kösksalet al.[11] conduct a review of literature involving DM applications in manufacturing industries to handle QI problems. The review covers publications from 1997 through 2007, and one of the issues considered in this review was the identification of the DM methods that are commonly used for different "quality tasks[2]" in manufacturing industry. According to their results, predicting quality is the most frequently performed task. They also report that the most commonly used methods to accomplish this task are: multiple linear regression (MLR) among the statistical based method, CART among the DT-based methods and multi-layer perceptron (MLP) with backpropagation learning used with gradient descent optimization (BP(GD)) under the NN-based methods.

In this work we compare the two techniques reported as most used in the first two groups, MLR and CART DT, regarding their robustness against the presence of measurement errors.

III. Design of the study

3.1 Analysis techniques used

Linear regression analysis has been widely used in the most diverse domains to model the structural relationship between a dependent or response variable and at least one predictor or independent variable. In general, the main purpose is to use this model to predict, as accurately as possible, the values that the response variable would take for future observations of the predictors. These models are easy to construct and interpret;however, it should be taken into account that they are built under certain assumptions, which must be satisfied in order to assess the validity of the results. This could restrict the situations in which MLR can be applied ([12], [13]).Nowadays, it is not difficult to find situations in which the variables of interest are related through complex nonlinear functions, which are rarely noticed or suspected a priori, so as to be taken into

[2] Tasks involved in QI and control activities considered in Köksal et al.'s work are refereed by them as *"quality tasks"*, namely description of quality, classification of quality, predicting quality and parameter optimization

account when proposing the regression model. This motivates the need to have methods that allow modeling thiskind of relationships in a more flexible way, with respect to the requirements that must be satisfied.

Decision trees are a non-parametric method of supervised learning, whose objective is to create a model for predicting the values of a response variable of interest, based on simple decision rules inferred from the data. In general terms, a decision tree looks like a tree upside down with the root at the top and the leaves at the bottom. The algorithm works by repeatedly partitioning the data, in order that each partition generates more homogeneous groups with respect to the response variable of interest. Beginning with the root (called "root node") the tree splits into two or more branches, where each branch in turn can again be split into two or more branches. The process continues until a "terminal node" is reached, that is, a node or branch which is not further split ([14], [15], [16]).

The CART classification and regression trees [17] have a series of properties that make them extremely attractive, being therefore one of the algorithms most used in practical applications. The CART algorithm performs only binary partitions at each step, so that the final structure of the tree is determined by a network of simple questions; starting with the root node, the answer to each question determines which branch of the tree to follow and what is the next question, until a terminal node is reached. This terminal node represents the final "decision" for that observation.

3.2 Concrete data

Concrete is an essential material in civil engineering, since it is used for the construction of the most diverse structures, from houses to bridges or large buildings. One of the most important mechanical properties of this material is its compressive strength, in fact, a specific variety called high strength concrete (HSC) exists. Its higher strength compared with normal strength concrete makes it useful in constructions that need to be reduced in weight; HSC carries load more efficiently so a lower amount of material is needed.

HSC is the product of a sophisticated mixture of several materials: cement, water, fine and coarse aggregates and some other ingredients. The strength of the concrete depends on the proportions of each ingredient added to the mixture, the ratio of water to cement being the most important factor. When the cement mixes with the water creates a cementing medium; if the mixture is adequate this paste covers each particle of the rest of the ingredients and fills all the free spaces between them. When the paste sets and hardens, all the ingredients are adhered and transformed into a solid mass. The chemical reactions between the water and the cement that cause the cemented paste to harden are produced quickly at first and slowly afterwards during a long period of time, so that under normal conditions the concrete reaches its maximum resistance over time ([18], [19]). Therefore, the compressive strength is usually measured after curing of 28 days.

In this work, a database of 1030 records is used[3] [20], without missing data, corresponding to concrete mix compositions with the following ingredients: cement, slag, fly ash, water, plasticizer, coarse aggregate and fine aggregate. Each of these components of the mixture constitutes an input variable for the process ($X_1$to X_7, respectively), all measured in kg per cubic meter of mixture. There is also an additional variable that measures the "age" of the concrete, recorded in days (X_8), i.e., the number of days of curing until the strength test. Finally, the database contains the compressive strength of the resulting mixture in each case (Y), measured in megapascals (Mpa).

3.3 Measurement errors

A measurement system can be defined as the set of devices, tools, procedures, people and environments used to assign a number to a characteristic being measured [21]. Therefore, the measurement system can be seen as a process itself, as such can be affected by several factors leading to incorrect measurements. An ideal measurement system should always provide accurate and precise measurements.

The precision of a measurement system refers to the variability observed within repeated measurements of the same unit under the same conditions. A measurement system is precise if it is capable of producing consistent results when the same unit is repeatedly measured under uniform conditions. The accuracy of the system refers to the difference between the true value of the characteristic being measured and the mean of the measured values. A measurement system is accurate if it has the ability to provide measurements that, on average, coincide with the true value of the characteristic being measured. According to these properties, the measurement error can then be decomposed into two elements: a ***systematic component*** associated to the accuracy of the measurement system, and a ***stochastic or random component*** related to itsprecision [22].

[3]The database belongs to Prof. I-Cheng Yeh(*), who donated it to UC Irvine Machine Learning Repository, University of California (https://archive.ics.uci.edu/ml/datasets.html), where it is of free access with retention of copyright notice for his owner and his published paper [18].

(*) Prof. I-Cheng Yeh (Original Owner and Donor). Department of Information Management, Chung-Hua University.Hsin Chu, Taiwan 30067, R.O.C. E-mail:icyeh@chu.edu.tw.

When data are obtained from measurements, it is necessary to take into account that the observed values could be affected by measurement errors, so that a distinction between the real or latent variable X and the empirical variableX^e must be considered [3]. The presence of systematic errors generates a constant bias in the measurements, so that the empirical variable results: $X^e = X + c$, where c is a real constant that represents the magnitude of the bias of the measurement system. The presence of a stochastic error component implies the existence of a random variable Vrepresenting it, so that the empirical variable is:$X^e = X + V$.

Any analysis conducted using data coming from measurements will work on the empirical variable X^e instead of the true variable X, so the impact of this should be carefully studied.

3.4 Analysis design

First, the original data set which is free of measurement errors is analyzed using the two techniques aforementioned, MLR and CART. The adjusted models in each case are used to establish a measure of their predictive power, which is used as a reference measure for subsequent comparisons.In the second instance, different structures of measurement errors are added to the original data, leading to new modified datasets. The models built in the first stage are used to predict the response on these modified data and a measure of the prediction error is computed from the resulting predicted values. The predictive measures obtained when models are applied to data free of error are then compared with the same measures obtained by the same model applied to data affected by measurement errors, in order to assess the impact of the presence of measurement errors.

Both random and systematic measurement errors are considered, and it is assumed that they affect the seven predictors related to weights of ingredients; because of its nature, it would be reasonable to assume that they are susceptible to error, either because of the instrument or the operator. Since each ingredient has a different range of variation, it is also assumed that each one is measured with a different device, and therefore errors are introduced independently to each of the variables.

For the case of random measurement errors, it is assumed that they are generated by a normal model with zero mean and constant variance for each variable $\sigma_{e_i}^2$.Let then be $V_i \sim N(0, \sigma_{e_i}^2)$ the random variable that represents the measurement error of the variable X_i. Under this assumption, the original values$x_i$of the variables X_i for $i = 1, \dots, 7$ are replaced by random values of new variables X_i^e, obtained as: $x_i^e = x_i + v_i$, where v_i is a quantity randomly generated by the model assumed for V_i. On the contrary, systematic errors are introduced by adding a constant quantity $c_i$to the original values of the variables $X_i$for $i = 1, \dots, 7$; that is: $x_i^e = x_i + c_i$.

In both cases, random and systematic errors, the study considers different values of $\sigma_{e_i}^2$and c_i, in order to evaluate the effect of varying the magnitude of the measurement errors. For $\sigma_{e_i}^2$ a grid of values proportional to the magnitude of the variability of the corresponding variable X_i is considered; that is: $\sigma_{e_i}^2 \in \{0.01j\sigma_{e_i}^2,\ j = 1,\dots,100\ \forall\ i=1,\dots,7$.Similarly, a grid of values for the constant $c_i$is assumed, which in this case are proportional to the average of the variable that will be affected; that is: $c_i \in \{0.01j\ \bar{X}_i, j = 1, \dots, 100\}\ \forall\ i = 1, \dots, 7$.The analysis is implemented in R-project software. The original database is randomly partitioned into train and test sets (70:30). Measurement errors are added to observations in the test set.The predictive power of models is assessed through the usual measure to evaluate predictive models, the root of the mean squared error (RMSE), computed from predictions on the test set. Regarding the fit of the regression model, stepwise selection of variables is applied to select the predictors to consider in the model. For the decision tree, on the other hand, the control parameters are set so as to let the tree to growuntil each terminal node is pure, and then proceed to prune it back. The prune is made according to the contribution that each subsequent split of nodes provides to the reduction of the prediction error. The criterion established for pruning is to prune the tree by selecting the first value of complexity for which the next partition generates an error reduction of less than 0.1%.

IV. Results

4.1 Data free of measurement errors.

As it has been aforementioned, a test set was split off from the original database. Table 1 shows descriptive measures of all the variables, by group (training - test) and for the overall dataset as well, where it can be seen that groups are balanced.

The stepwise regression procedure identifies the quantities of cement, superplasticizer, slag, water and fly ash, as well as the age of the cement, as relevant variables (Table 2). Analysis of ~~the~~ variance for the model fitted with those selected predictors yields a p-value lower than 0.0001 and a coefficient of determination equal to 0.62 ($R^2 = 0.62$).

The estimated coefficients of the model indicate that the strength of the concrete increases with age and increases as well by increasing the amounts of all ingredients except for water, whose increase in the mixture produces the opposite effect.

Table 1: Descriptive measures of variables under study.

	Complete dataset			Training dataset			Test dataset		
Variable	**Min; Max**	**Mean**	**SD**	**Min; Max**	**Mean**	**SD**	**Min; Max**	**Media**	**DE**
Cement	102; 540	281.17	104.51	102; 540	286.20	105.21	102; 540	269.40	102.06
Slag	0; 359.4	73.90	86.28	0; 359.4	74.86	87.59	0; 359.4	71.65	83.22
Fly ash	0; 200.1	54.19	64.00	0; 200	52.02	63.16	0; 200.1	59.25	65.74
Water	121.8; 247	181.57	21.35	121.8; 247	181.90	20.83	121.8; 246.9	180.90	22.56
Superplasticizer	0; 32.2	6.20	5.97	0; 32.2	6.16	5.89	0; 32.2	6.30	6.17
Coarse aggregate	801; 1145	972.92	77.75	801; 1145	972.90	78.22	801; 1134.3	972.90	76.79
Fine aggregate	594; 992.6	773.58	80.18	594; 992.6	771.80	81.63	594; 992.6	777.70	76.63
Age	1; 365	45.66	63.17	1; 365	44.47	59.66	3; 365	48.43	70.71
Strength	2.33; 82.60	35.82	16.71	2.33; 82.60	36.32	16.90	4.83; 81.75	34.64	16.21

Table 2: Stepwise regression results.

Coefficient	Estimation	Std. error	t-value	p-value
Intercept	32.156	5.175	6.214	<0.001
Cement($\boldsymbol{X_1}$)	0.104	0.005	20.723	<0.001
Superplasticizer($\boldsymbol{X_5}$)	0.209	0.105	2.003	0.045
Age($\boldsymbol{X_8}$)	0.125	0.007	18.244	<0.001
Slag($\boldsymbol{X_2}$)	0.086	0.006	14.451	<0.001
Water($\boldsymbol{X_4}$)	-0.235	0.026	-9.013	<0.001
Fly ash($\boldsymbol{X_3}$)	0.070	0.010	7.232	<0.001

The resulting model is used to predict the strength of the concrete obtained under each of the conditions given in the records of the test dataset. Those values are then compared with the true known values of strength under each condition and the RMSE is computed, which results 10.483 ($RMSE_{MLR} = 10.483$).

On the other hand, the complete decision treehas 460 nodes. Table 3 shows the values of the complexity parameter and the relative errors in training and cross validation for the first 17 splits of the tree. The criterion assumed for pruning the tree indicates that the tree should stop partitioning after 12 splits, since the improvement in the error rate for the next split is lower than 0.1%.

Table 3: Complexity table of the decision tree.

Complexity	Number of splits	Relative training error	Relativevalidation error	Std. error
2.41E-01	0	1.0000000	1.00935	0.048339
1.87E-01	1	0.7586929	0.76530	0.037851
6.59E-02	2	0.5720161	0.57773	0.029817
6.42E-02	3	0.5060859	0.54496	0.028787
4.07E-02	4	0.4418724	0.46804	0.025410
3.82E-02	5	0.4011390	0.45844	0.023822
3.22E-02	6	0.3629869	0.41659	0.021231
1.95E-02	7	0.3308067	0.37844	0.019360
1.89E-02	8	0.3113339	0.36729	0.017996
1.83E-02	9	0.2924311	0.36368	0.017950
1.73E-02	10	0.2740841	0.34855	0.017643
1.01E-02	11	0.2568203	0.29966	0.015162
1.00E-02	12	0.2467083	**0.28522**	0.014215
9.37E-03	13	0.2367019	**0.28619**	0.014347
8.80E-03	14	0.2273370	0.30085	0.019211
8.36E-03	15	0.2185320	0.29626	0.019093
7.86E-03	16	0.2101720	0.29640	0.019186
7.84E-03	17	0.2023080	0.29166	0.019148
…	…	…	…	…

Note: Only the first 17 rows of the complete table are shown.

The pruned tree has then 13 terminal nodes and the variables used in tree construction coincide with those selected by the stepwise regression and adds the quantity of coarse aggregate (X_6). This means that the tree would lead to a decision going over a network of binary rules based on the age of the concrete and on the quantities of water, cement, slag, superplasticizer and coarse aggregate used in the mixture. Fig. 1 shows the plotting of the pruned tree. Fig. 2 shows the training and cross validation relative error against the number of splits, for the first one hundred splits; the vertical dashed line is located at the number of splits at which the tree was pruned.

The selected tree is run on the test set records and the resulting predicted values are compared to the true known values of the response, yielding to $RMSE_{CART} = 9.359$.

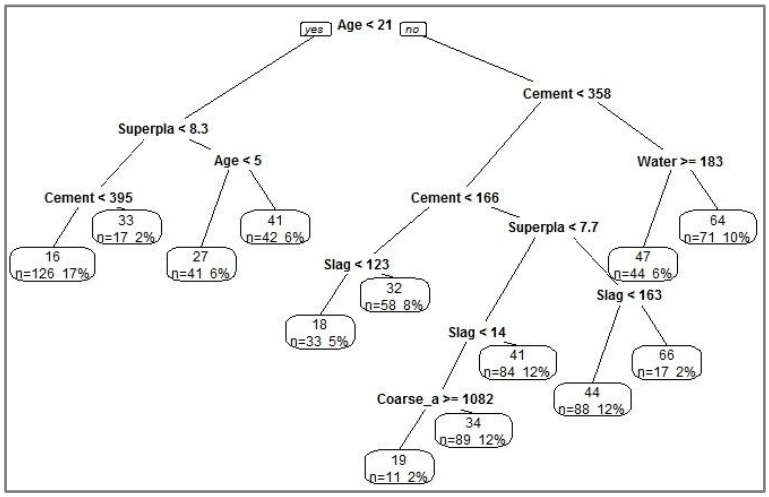

Figure 1: Pruned decision tree.

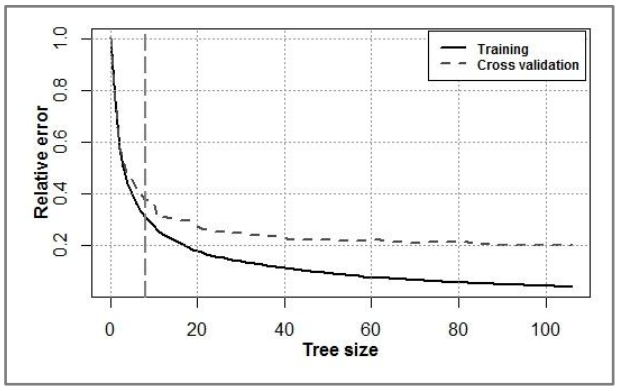

Figure 2: Relative error in training set and cross validation based relative error.
Note: Vertical dashed line at number of splits of the pruned tree.

4.2 Prediction under the presence of random measurement errors

Since random measurement errors needs to be added to the original data by randomly generating a value of a normal distribution, the process is repeated several times in order to avoid the possible noise accounted for a single random selection. Therefore, for each predictor X_i, $i = 1, \dots, 7$ and for each magnitude of measurement error $(j = 1, \dots, 100)$ 500 samples of $n = 309$ (number of records in the test set) random values of a normal distribution with mean zero and variance $\sigma_{e_i}^2$ are simulated. Each of these samples is added in turn to the original values of the predictors, leading to a "contaminated" dataset with which the response is predicted by the regression model and the pruned tree both selected in the previous section. In both cases RMSE is computed. Repeating this process for each of the samples leads to 500 values of RMSE, which are at the end summarized in mean.

Fig. 3 shows the means of RMSE for both methods against the magnitude of the measurement error, the latter expressed as the percentage of the variance of each variable. It can be seen that the predictive power of both the regression model and the decision tree is affected by the presence of measurement errors. The effect is shown as an increase in the RMSEs compared to those obtained with the free of errors data. In both techniques the effect becomes more important as the magnitude of error increases.

In relative terms, the effect seems to be more harmful for CART based predictions, unless the magnitude of errors is high (more than 75% of the variance of the corresponding variable), in which case regression based predictions are more affected. This can be seen in Fig. 4, where the plotted curves represent, for each technique, the ratio of the RMSE obtained from data with measurement errors relative to the RMSE reached with the original free of errors data, against the magnitude of errors.

Finally, in order to compare the techniques with each other, the ratio between the CART based RMSE and the regression based RMSE is computed for both datasets with and without measurement errors. The result shows that the linear regression model has a better predictive performance than CART decision tree, for lower measurement error magnitudes (less than 30% of the original variance), whereas CART is definitely better when errors are of higher magnitude.

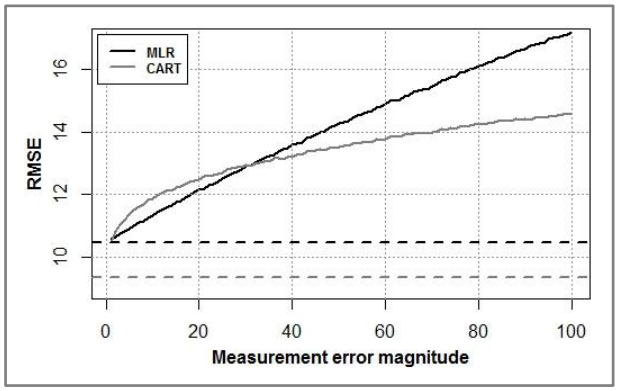

Figure 3: Mean RMSE of MLR and CART obtained from data affected by random measurement errors (solid lines) and RMSE values of MLR and CART from data free of error (dashed lines), against error magnitude.

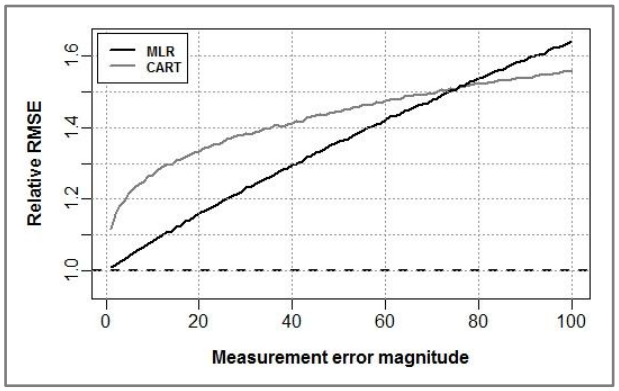

Figure 4: RMSE of MLR and CART from data with random measurement errors relative to the RMSE obtained from data free of errors, against the error magnitude.

Note: Dashed horizontal line at ordinate value 1 for reference.

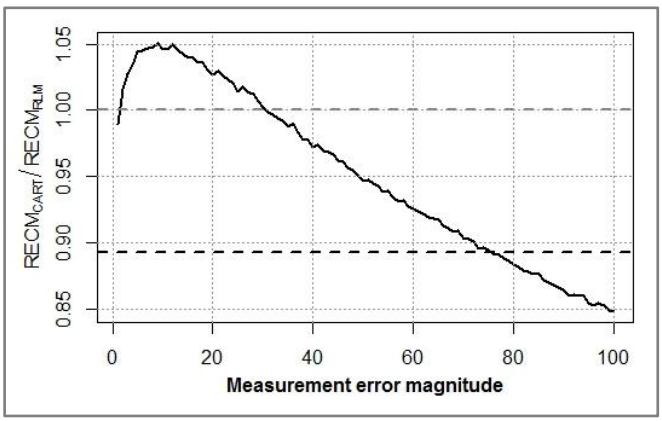

Figure 5: Ratio of CART RMSE to MLR RMSE from data affected by random errors (solid lines) and from data free of error (dark dashed line), against the measurement error magnitude.

Note: lighter dashed line at ordinate value 1 for reference.

4.3 Prediction under the presence of systematic measurement errors

In this case each record in the original test set is modified by adding a constant error, proportional to the mean of the corresponding variable. Thus, a new "contaminated" data set is generated for each of the magnitudes of measurement error considered. As before, these data are used to predict the response using the two models selected in section 4.1, and RMSE for both methods is computed from the resulting predicted values.The RMSE curves for CART and MLR are shown in Fig. 6, along with the RMSE values of both techniques obtained from the data free of errors. It is possible to observe that under the presence of this type of errors, the predictive power of the decision tree is seriously affected, whereas the effect on the MLR model performance is less important. In fact, the increase in the MLR prediction error induced by the presence of measurement errors is almost negligible when the magnitude of errors is less than 40% of the original means.

The same information but in relative terms is shown in Fig. 7. Each RMSE obtained from contaminated data was divided by the RMSE obtained from the original data. The serious distortion of predictive power of CART model is more than notable; under the presence of systematic measurement errors the CART based RMSE was observed up to almost twice the value reached from the original data.

Regarding the comparison of both methods to each other, recall that CART algorithm showed a better predictive performance than MLR when applied to the original data ($RMSE_{CART} = 9.359$vs $RMSE_{MLR} = 10.483$). However, the presence of systematic measurement errors generates the opposite behavior (Fig. 8). The RMSE obtained by CART is greater than that achieved by the linear regression model. Depending on the magnitude of the measurement error, the prediction error with CART can be up to 60% greater than the generated by the regression model.

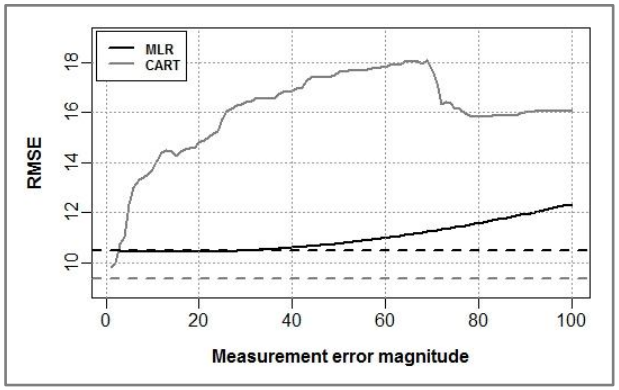

Figure 6: Mean RMSE of MLR and CART obtained from data affected by systematic measurement errors (solid lines) and RMSE values of MLR and CART from data free of error (dashed lines), against error magnitude.

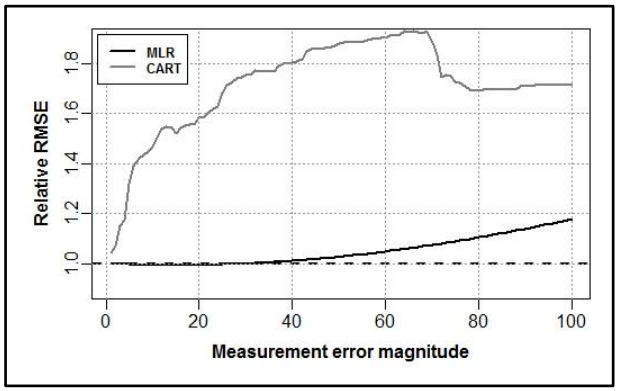

Figure 7: RMSE of MLR and CART from data with systematic measurement errors relative to the RMSE obtained from data free of errors, against the error magnitude.

Note: Dashed horizontal line at ordinate value 1 for reference.

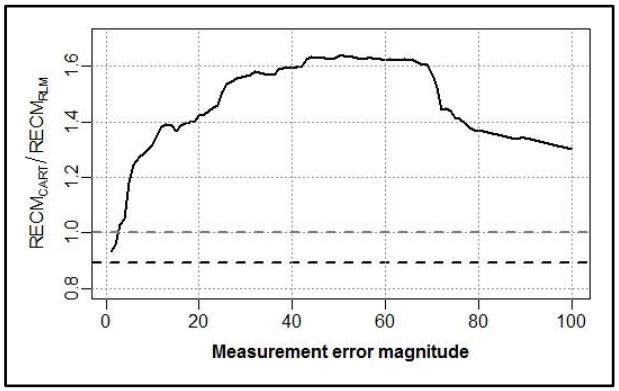

Figure 8: Ratio of CART RMSE to MLR RMSE from data affected by systematic errors (solid lines) and from data free of error (dark dashed line), against the measurement error magnitude.

Note: lighter dashed line at ordinate value 1 for reference.

V. Conclusion and Discussion

Decision tree algorithms have emerged as a non-parametric alternative to identify or discover the relationships existing in a set of variables, with the aim of creating a model that allows predicting the value of a response as a function of input variables and their relationships. Nowadays it is perhaps one of the most used data mining tools in various disciplines, because of its many advantages: it is easy to understand and interpret, it requires little data preparation, it can incorporate both numerical and categorical variables, among others.

From the practical point of view, however, the question that has motivated this work is how robust these algorithms are to make predictions when the data begin to incorporate external "noise", in particular, the additional variability coming from measurement errors, which is likely to occur every time a measurement system is used to get the information about the process.

The CART algorithm was applied to a real dataset, free of measurement error, to construct a regression tree to predict the strength of concrete as a function of a series of variables that represent the components of the mixture and the curing time. This model was then used to predict the response on new datasets, which are generated by modifying the original data through the addition of different structures of measurement errors. The comparison of the prediction error obtained when data are affected by measurement errors with the value reached when data are free of error allows analyzing the effect that measurement errors could cause on the predictive power of the models.

The results across all the scenarios showed that the prediction error is always affected by the presence of measurement errors, whether random or systematic, and even if they are of small magnitude. However, the presence of systematic errors has shown much more detrimental to the predictive power of trees than the presence of random errors, generating increases in the RMSE of up to about 60%.

Additionally, the comparison of the predictive performance of the CART algorithm with that of a multiple linear regression model, reveals that, under free of error data, CART has a better predictive behavior than MLR. This result holds also under the presence of random measurement errors, even though CART is more adversely affected by measurement errors. However, the contrary occurs when a constant bias affects the measurements. In this case, the distortion in the predictive ability of CART greatly exceeds that accounted for MLR model.

These results, although corresponding to the study on a particular case, show the importance of constantly ensuring the quality of the information collected. This has even more relevance in the current industrial contexts in which automation allows the generation and the storage of a large amount of data in very short periods of time, which real-time analysis identifies corrective actions almost immediately.

References

[1] W.J. Frawley, G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview, *AI Magazine, 13(3), 1992, 57-70.*

[2] H-J. Mittag, Measurement Error Effect on Control Chart Performance, *Annual Quality Congress, 49(0), 1995, 66-73.*

[3] H-J. Mittag, Measurement Error Effects on the Performance of Process Capability Indices, *Frontiers in Statistical Quality Control, 5, 1997, 195-206*

[4] S. Bordignon and M. Scagliarini, Statistical Analysis of Process Capability Indices with Measurement Errors, *Quality and Reliability Engineering International, 18, 200, 321-332*.

[5] D. Shishebori and A.Z. Hamadani, The Effect of Gauge Measurement Capability and Dependency Measure of Process Variables on the MCp, *Journal of Industrial and Systems Engineering, 4(1), 2009, 59-76*.

[6] M. Scagliarini, Multivariate process capability using principal components analysis in the presence of measurement errors, *Advances in Statistical Analysis, 95, 113-128.*

[7] D. Dianda, M. Quaglino, J. Pagura, M.L. De Castro, Efecto del error de medición en índices de capacidad de procesos, *Saberes, 8(2), 2016, 91-110.*

[8] D. Michie, D.J. Spiegelhalter and C.C. Taylor, *Machine learning, neural and statistical classification*(NY: Ellis Horwood, 1994)

[9] C. Apté, Data Mining: An Industrial Research Perspective, *IEEE Computational Science & Engineering, 4(2), 1997, 6-9*.

[10] M.H. Dunham, Data mining.Introductory and advanced topics (N. J. Prentice Hall: Pearson Education, 2003)

[11] G. Köksal, I. Batmaz and M.C. Testik, A review of data mining applications for quality improvement in manufacturing industry, *Expert Systems with applications, 38, 2011, 13448-13467.*

[12] D.C. Montgomery, E.A. Peck and G.G. Vining, *Introduction to linear regression analysis, 5thEd.* (NJ: J. Wiley & Sons, Inc., 2012)

[13] B. Jorgensen, *The theory of linear models* (NY: Chapman and Hall, Inc., 1993)

[14] R. Nisbet, J. Elder and G. Miner, *Handbook of statistical analysis and data mining applications* (Amsterdam: Elsevier Inc., 2009).

[15] G. Williams, *Data Mining with Rattle and R*, (NY: Springer, 2011)

[16] D. Larose and C. Larose,*Data mining and predictive analytics, 2nd. Ed.*(NJ: J. Wiley & Sons, Inc., 2015)

[17] L. Breiman, J. Friedman, R.Olshen and C. Stone, *Classification and Regression Trees* (Florida: Chapman & Hall/CRC Press, 1984)

[18] I.C. Yeh, Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research, 28(12), 1998, 1797-1808.*

[19] C. Deepa, K.SathiyaKumari and V.PreamSudha, Prediction of the compressive strength of high performance concrete mix using tree based modeling, *International Journal of Computer Applications, 6(5), 2010, 18-24.*

[20] M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[21] AIAG-Automotive Industry Action Group,*Measurement System Analysis, 3rd Ed.*(Detroit, MI:AIAG, 2002).

[22] D.C. Montgomery, *Introduction to Statistical Quality Control, 6th Ed.* (NY: J. Wiley & Sons, Inc., 2009).