Generation of predictions of weight and illness in cattle

Manuela Zapata[1]

Advisors:
Mauricio Toro[2]
Rodrigo García[3]
José Aguilar[4]

Research practice (I)
Research proposal
Mathematical Engineering
Department of Mathematical Sciences
School of Sciences
Universidad EAFIT

June 2021

[1]Department of Mathematical Sciences. Universidad EAFIT- mzapatam1@eafit.edu.co (CvLAC: *https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh* = 0001845319)
[2]Department of Informatic and Systems. Universidad EAFIT- mtorobe@eafit.edu.co
[3]Universidad del Sinú- rjgarciah@eafit.edu.co
[4]GIDITIC. Universidad EAFIT- aguilarjos@gmail.com

Abstract

Currently in the world there is a high consumption of animal products such as meat, dairy and eggs that come from different animals. Colombia is a very agricultural country and according to Banco de la República (2018), agriculture contributed 6,3% of the GDP in 2017, that number only tend to increase due to the increment on the demand. This is why the results obtained in this investigation by using Machine Learning algorithms such as decision tree, random forest, support vector machine and XGBoost are important because they can help ranchers and companies to develop strategies that can improve their profits as well as animal health. The best prediction result was obtained with the random forest algorithm, although the other algorithms provided very similar predictions and error.

Keywords: Illness in cattle, weight prediction, regression, machine learning.

1 Introduction

Currently in the world there is a high consumption of animal products such as meat, dairy and eggs that come from different animals. As the world's population increases the production methods need to be improved constantly into more efficient versions in order to fulfill the demand. In the aim of develop methods to better usage of resources and also seeking for ways to lessen the environmental impact of this activities. This project looks for ways to integrate technology and farming which is an industry that so far hasn't reached a very high technological development.

Colombia is a highly agricultural country, it breeds a wide variation of cattle and many people and their families financially depend on this industry for a living. However, despite agriculture being one of the main sources of jobs and income in Colombia, there hasn't been much research and investment in technology to improve processes and getting greater margins of profits for the ranchers and companies.

Analyzing the current situation, the main problem to solve was to find ways to implement non expensive technological solutions that companies and ranchers can use to lessen the costs of breeding and diminish the losses of animals by identifying illness on time so that the animal can be treated and sold in the market. Also, identifying growth peaks in the cattle so that the ranchers and companies can use this period to supplement the cattle's food or its food intake so that they can reach the ideal weight to be sold on the market in shorter periods of time, increasing their income.

The main contribution made with this article is that using machine learning algorithms and an artificially generated weight dataset we were able to predict cattle weight and analyze the results obtained so that the ranchers and the companies can implement strategies if they see that the growth wave of the animal its not matching with the weight it should have for its breed and age.

The importance of this contribution is that using this models made with machine learning algorithms could help ranchers and companies to increase their profits by shortening the time that cattle needs to grow. This is possible by analysing each animal's growth wave and determining if it has a normal behaviour compared to a normal growth curve for its breed and age, and if it doesn't, check for underlaying causes such as parasites affecting the animal's weight gain or a nutrient deficiency in the grass the animal is consuming.

It also can help to improve the animals health because understanding the results obtained with the algorithms can help to identify illness in cattle and apply solutions quickly rather than leaving the animal in pain until the illness is more visible or loosing the animal due to the large progress of the disease.

In the following sections the reader can find the state of the art where a broad number of projects and methods have been used to predict cattle weight around the world. The methodology used in the development of this project is mentioned with detail, so are the results obtained when this machine learning algorithms were applied. Finally, the conclusions and input for future research are explained as well.

2 State of the art

Cattle weight prediction is a very studied problem around the world because the livestock industry is an important part of the economy of several countries and also provides food and other supplies for many people. For example the article written by Song *et al.* (2018) explains the methods used by them to make this predictions using 3D tools, the variables they found were the best fit in the predictions (hip height, hip width, and rump length) and also the mathematical methods they used to make the predictions, in their case was multiple linear regression and their validation models were leave-one-out cross-validation, giving the root mean square error and mean absolute percentage error.

Other studies have taken a different approach in the prediction goal, in the case of the Franco *et al.* (2017) study they didn't use any specific technological tool, instead they chose some cattle body measurements as heart girth, wither height, hip width, body length and hip height and used them as their variables when they were developing new equations to predict cattle weight. Just as the study presented before this, they used regressions to make the predictions.

In the research made by Ashwini *et al.* (2019) they perform a similar strategy as the investigation mentioned before this one, but in this case they added other variables such as age and cattle breed. They also used multiple linear regressions to obtain the predictions and did it using the SPSS 21 software.

As mentioned before, this problem is an important field of study in different countries such as in Senegal where Tebug *et al.* (2018) also made predictions of cattle weight but in this case they focused in low-input systems because in their demography and environment there are not many available resources to feed the cattle. That is why they developed a prediction equation and translated it into a weigh band, to provide a simple and reliable method for cattle keepers to estimate the weight of studied cattle breed types. Just as all of the articles mentioned before they also used linear regression in their prediction equation.

A study made in Austria by Gruber *et al.* (2018) made predictions on cattle weight also using regressions and some of the same variables that other articles mentioned but they included as a significant variable animal data in the lactation and dry periods. Body measurements were tested

as single predictors and in multiple regressions according to their prediction accuracy and their correlations with body weight. For validation, data sets were split randomly into independent subsets for estimation and validation. Accuracy of the predictions was evaluated by decomposing the mean square prediction error (MSPE) into error due to central tendency, error due to regression, and error due to disturbance.

Miller *et al.* (2019) took a different approach of the problem, this study is about selection of finishing beef cattle for slaughter while the others focused on production only. The evaluation of performance is currently achieved through visual assessment and/or by weighing through a crush. Consequently, large numbers of cattle are not meeting target specification at the abattoir. Video imaging analysis (VIA) is increasingly used in abattoirs to grade carcasses with high accuracy. There is potential for three-dimensional (3D) imaging to be used on farm to predict carcass characteristics of live animals and to optimise slaughter selections. The objectives of this study were to predict liveweight (LW) and carcass characteristics of live animals using 3D imaging technology and machine learning algorithms (artificial neural networks). Sixty potential predictor variables were automatically extracted from the live animal 3D images using bespoke algorithms. Performance of prediction models was assessed using R2 and RMSE parameters following regression of predicted and actual variables for LW. This study demonstrated that 3D imaging coupled with machine learning analytics can be used to predict LW, Saleable meat yield (SMY) and traditional carcass characteristics of live animals. This system presents an opportunity to reduce a considerable inefficiency in beef production enterprises through autonomous monitoring of finishing cattle on the farm and marketing of animals at the optimal time.

3 Solution method / Methodology

For the development of this project was used the Cross Industry Standard Process for Data Mining (CRISP-MD) methodology that is described by Wirth & Hipp (2000) in terms of a hierarchical process model, comprising four levels of abstraction (from general to specific): phases, generic tasks, specialized tasks, and process instances.
The 6 phases used to implement this process were:

3.1 Business and data understanding

This initial part of the project was understanding the objectives that we wanted to reach and look for ways to accomplish them. That included formulating the problem clearly and designing a work plan and also have a broad and deep understanding of the previous work done in the same field, that is to say, have solid knowledge about the state of the art. Business understanding and data understanding are very intertwined because in order to formulate the problem correctly there needs to be a certain degree of understanding the available data or the data to be obtained and submitted to analysis.That is to say, understand the general cattle breeding methodology, obtaining data to develop the project, study previous research and understand cattle growth curves for different breeds.

3.2 Data Preparation

As mentioned before, Colombia is a highly agricultural country but this industry is mostly empirical and not very technological. This is why despite exhausting investigation to get data about cattle and its weight it was impossible to obtain. Therefore, the data had to be simulated to be used with the algorithms. The process of simulating the data consisted in analyze the growth curve of four breeds of cattle: Cebú, Blanco Orejinegro (BON), Romosinuano (ROM) and Nellore. After this curves were understood and analyzed, the statistical parameters mean and standard deviation were used to simulate the data. This were obtained from Posada *et al.* (2011), Quiceno *et al.* (2012) and Domínguez-Viveros *et al.* (2017) and the dataset was simulated by using a normal function that generates random numbers within a range using the parameters mentioned before. This process was made for all the breeds studied in this process and for different ages of the cattle.

3.3 Modeling

In this section four models were applied: Decision trees, random forest, xgboost and support vector machine. This were chosen because they are all based in regression and would be easy to compare results by using the metrics mean square error (MSE), root mean square error (RMSE), regression score function (R^2) and mean absolute percentage error (MAPE).

3.4 Evaluation

During this phase, the artificially created datasets were used to train and test the algorithms mentioned before. The metrics MSE, RMSE, R^2 and MAPE where applied to all of the algorithms in order to compare results with this standard measures, make conclusions and find improvement points to take into account in future research in this topic. The algorithms were trained and tested with the dataset of each breed and also with the dataset containing the data from all breeds. This was made to compare the results obtained in more detail.

3.5 Deployment

This last section of the project consists in organising the results and algorithms developed so that researchers, ranchers and companies can use them and interpret them easily to improve their businesses and increase their profits by implementing strategies based on the results obtained from the algorithms and models.

4 Results

4.1 Cebu

Figure 1: Cebu breed animal. Wikipedia (2021)

Model	MSE	RMSE	R2	MAPE
Random Forest Regressor	2443.0437	49.2286	0.533	0.1766
Decision Tree Regressor	2448.7522	49.2979	0.532	0.1768
XGBoost	2448.7491	49.2979	0.532	0.1768
Support Vector Machine (SVM)	2448.7522	49.2979	0.532	0.1768

Figure 2: Models results for Cebu breed.

The MSE is a statistical estimator that measures the difference between the estimator and what is estimated, in this case, the difference between the original weights and the predicted ones. Its equation is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{1}$$

Comparing all the models, it is clear that all of them have very similar values. However, the model with the better MSE is the Random Forest Regressor because it has the lower value, which means that the original values and the predicted ones are very similar.
The RMSE is the sqared root of the MSE. It is an accuracy measure and is useful because it adds the magnitudes of the errors between the predicted and original values. The goal is to obtain zero,

that would mean that the prediction is equal to the original values. This outcome is almost never achieved so the next best thing is to find the lowest RMSE possible.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{2}$$

Analyzing the results for this measures the same conclusion as the MSE is reached: The model with lowest RMSE is the Random Forest Regressor, despite all models having very similar values, this one is better by a few decimals.

R^2 is the coefficient of determination. It measures the proportion between the variances of the original values and the predicted ones. If R^2 is equal to 1 it means that there is a perfect linear fit in the model, and if it is 0 it means that there is non-representativeness of the linear model. The objective is that the value obtained is as closer to 1 as possible. Analyzing the results obtained with all the models for this breed, the conclusion is that once again the best model is the Random Forest Regressor by one decimal despite the results for the other models are very similar.

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} \tag{3}$$

MAPE is a method of predicting the accuracy of a model taking as variables the original values and the predicted ones. Usually is expressed as a percentage.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \tag{4}$$

The ideal MAPE is equal or below 10%, however this is hard to accomplish. However the MAPE for this models weren't the ideal value, the ones obtained were between 20 and 30%, and that is a good value. The model with the best MAPE is the Random Forest Regressor again by decimals despite all models having very similar values.

4.2 Blanco Orejinegro (BON)

Figure 3: Blanco Orejinegro breed animal. Generalidades de la ganadería bovina (2014)

Model	MSE	RMSE	R2	MAPE
Random Forest Regressor	2216.5493	46.6016	0.3877	0.1749
Decision Tree Regressor	2214.1214	46.5913	0.3877	0.1745
XGBoost	2214.1189	46.5913	0.3877	0.1745
Support Vector Machine (SVM)	2214.1214	46.5913	0.3877	0.1745

Figure 4: Models results for BON breed.

In this breed the best MSE was obtained by the XGBoost model, and contrary to the Cebu breed the worst result was obtained by the Random Forest Regressor. This outcome can be explained by the data because the simulation of the data for Cebu breed started when the cattle was 7 months old and ended when the cattle was 18 months, but for the BON breed the simulation started when the cattle was 10 months and ended at 17 months. Although each simulation had the same amount of data for each period of time, the fact that BON breed has less months than the Cebu breed can explain the change in the accuracy of the predictions for Random Forest Regressor.

For the RMSE happens the same thing as in the MSE, that is that the model with the worst performance is Random Forest Regressor, again only by few decimals. The other models obtained the same value.

In the R^2 all models obtained the same value. In comparison to the Cebu breed this one shows a worse result because the values are closer to 0. This can be explained by the high standard deviation used to simulate the data used in this model.

In regards to the MAPE all models obtained basically the same results because since in MAPE an outcome between 20 to 30% is a good outcome we can conclude that all models had good results.

4.3 Romosinuano (ROMO)

Figure 5: Romosinuano breed animal. CONtextoganadero (2016)

Model	MSE	RMSE	R2	MAPE
Random Forest Regressor	1794.8983	41.9597	0.238	0.1724
Decision Tree Regressor	1793.5583	41.9398	0.2388	0.1722
XGBoost	1793.5564	41.9397	0.2388	0.1722
Support Vector Machine (SVM)	1793.5583	41.9398	0.2388	0.1722

Figure 6: Models results for ROMO breed.

For the MSE in this breed the best outcome was obtained by the XGBoost model because it has the lower value. However, and as was shown in the other breeds all of the models have almost the same values.
In the RMSE the best value was also obtained by the XGBoost model by a few decimals because it has the closest value to zero amongst the models. However, all the models have almost the same result.
Evaluating the R^2 it evident that there is a tie between all models, but since the values obtained are so close to 0 it is relevant to say that there is no significant correlation between the variables in

this breed. This can be explained by the high standard deviation used to simulate the data used in this model.
For the MAPE, just as with the previous breeds the values obtained are very similar with each other and are also good because they are between 20 and 30%.

4.4 Nellore

Figure 7: Nellore breed animal. Consultagro (2018)

Model	MSE	RMSE	R2	MAPE
Random Forest Regressor	425.7545	20.3496	0.9507	0.0505
Decision Tree Regressor	425.5359	20.3517	0.9507	0.0506
XGBoost	425.5360	20.3517	0.9507	0.0506
Support Vector Machine (SVM)	425.5359	20.3517	0.9507	0.0506

Figure 8: Models results for Nellore breed.

For this breed and in the MSE statistic the best value was obtained by Decision Tree Regressor and Support Vector Machine and the worst value was obtained by the Random Forest Regressor due to having the highest value.

The best RMSE was obtained by the Random Forest Regressor because it has the lowest value, although as it was presented in the other breeds, all models have almost the same values. The difference is just of a few decimals.

As far as the R^2 this breed has the better values amongst all breeds evaluated in this project because it is the closest values to 1. This indicates that the data for this breed has a very good linear fit in the model.

This breed also obtained the best MAPE amongst all the breeds evaluated here, and also a very good value in general because it is below 10%. This improvement can be explained by the data, because this breed has more information than the other breeds, that is to say that the weights for this breed were reported during 16 months while the weights for the other breeds were reported only during 7 to 10 months.

4.5 All breeds

Model	MSE	RMSE	R2	MAPE
Random Forest Regressor	1696.7108	40.9370	0.7723	0.1428
Decision Tree Regressor	1697.8975	40.9518	0.7722	0.1428
XGBoost	1697.8949	40.9517	0.7722	0.1428
Support Vector Machine (SVM)	1697.8975	40.9518	0.7722	0.1428

Figure 9: Models results for all breeds.

In the global perspective, the model with better MSE is the Random Forest Regressor. This is consistent with the results obtained with each breed, because in general this was the model with lower numbers, despite all models having very similar values.

For the RMSE, as well as in the previously mentioned measure the model with the best result is the Random Forest Regressor because it has the lowest values even though it is just for a few decimals.

In the R^2 statistic the best outcome was reached by the Random Forest Regressor because just for one decimal it is closer to 1. In general is correct to affirm that all the models represent a good linear fit.

Finally, the MAPE obtained by all models was the same. This is a very good outcome that was surpassed only by the outcome obtained in the Nellore breed. However, it is safe to say that the predictions made with the models are accurate values.

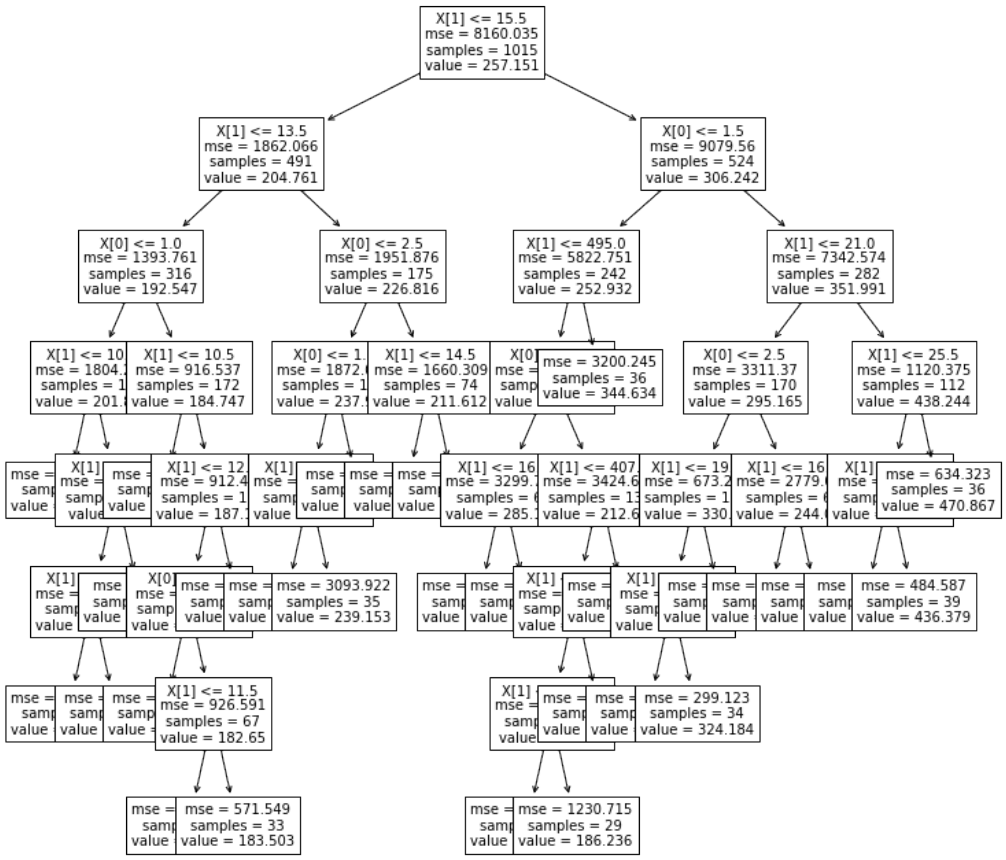

Figure 10: Decision tree for all breeds where X[0] and X[1] are the features of the model, that is to say original values and predicted values.

5 Conclusions and future research

An important source of error could have been the usage of artificial data to train and test the algorithms instead of real data. That is why for future research will be very useful to obtain real data to train and test the models because this will improve the model's accuracy. Work with real data can also help to limit the outlier values, because as in this project the data was generated using a normal distribution and a random function to generate values between a range, there were a lot of outliers and its presence could have been an important source of error in the models predictions. Since the data used in this project was synthetic, generated by the same function and also with the same volume of information could have led to all models having almost the same outcome, because as it was seen during the research, each model had almost the same values. If real data had been used it is safe to say that even though the results in the predictions could have been very similar because all the models are based on regression but there would be a much clear difference between model results and their predictions.

The normal distribution of the simulated data is what could have lead to all models having almost the same values in the statistics measures taken to evaluate performance of the models (MSE,

RMSE, R^2 and MAPE).
The reason why Random Forest Regressor model has slightly better results than the other methods is that it uses a bootstrap aggregation method, that is to say that it uses a random sampling with replacement and this helps the model to have better understanding of the variance and the bias of the data.

For future research the dataset to use with the models to for making predictions has to have the highest volume of data as possible because it would let the model to learn from much more information, identify outliers and therefore avoid making predictions based on them what would lead to having a much more accurate predictions, less error and better fitted model.
Another important thing to consider for future research is to use the models make the predictions using more variables because this would help to get higher accuracy, lessen the error and have a better fit in the model. If more variables are added then the models will have much more information to learn from and train with to obtain better outcomes and conclusions. This variables could be the kind of breeding, the purpose of the animal (meat, milk), the kind of grass they are fed with, the weather, amongst others.

References

Ashwini, J Patel, Sanjay, Patel, Amipara, GJ, Lunagariya, PM, Parmar, DJ, & Rank, DN. 2019. Prediction of body weight based on body measurements in crossbred cattle. *Int. J. Curr. Microbiol. App. Sci*, **8**(03), 1597–1611.

Banco de la República, de Colombia. 2018. *Informe de la junta directiva al Congreso de la República.*

Consultagro, Vzla. 2018. *Bovinos de carne: Nellore.*

CONtextoganadero, una lectura rural de la realidad colombiana. 2016. *Promoverán raza Romosinuano a través de un concurso de fotografía.*

Domínguez-Viveros, Joel, Urbina-Valenzuela, Alfredo Ramón, Palacios-Espinoza, Alejandro, Callejas-Juárez, Nicolás, Ortega-Gutiérrez, Juan Ángel, Espinoza-Villavicencio, José Luis, Padrón-Quintero, Yamariz, & Rodríguez-Castro, Manuel. 2017. Caracterización del crecimiento de bovinos cebú en pruebas de comportamiento en pastoreo. *Ecosistemas y recursos agropecuarios*, **4**(11), 341–348.

Franco, Marcia de Oliveira, Marcondes, Marcos Inácio, Campos, José Maurício de Souza, Freitas, Denise Ribeiro de, Detmann, Edenio, & Valadares Filho, Sebastião de Campos. 2017. Evaluation of body weight prediction Equations in growing heifers. *Acta Scientiarum. Animal Sciences*, **39**(2), 201–206.

Generalidades de la ganadería bovina, blog. 2014. *Raza Blanco Orejinegro.*

Gruber, Leonhard, Ledinek, Maria, Steininger, Franz, Fuerst-Waltl, Birgit, Zottl, Karl, Royer, Martin, Krimberger, Kurt, Mayerhofer, Martin, & Egger-Danner, Christa. 2018. Body weight prediction using body size measurements in Fleckvieh, Holstein, and Brown Swiss dairy cows in lactation and dry periods. *Archives animal breeding*, **61**(4), 413–424.

Miller, Gemma A, Hyslop, James J, Barclay, David, Edwards, Andrew, Thomson, William, & Duthie, Carol-Anne. 2019. Using 3D imaging and machine learning to predict liveweight and carcass characteristics of live finishing beef cattle. *Frontiers in Sustainable Food Systems*, **3**, 30.

Posada, Sandra, Rosero, Ricardo, Rodríguez, Norberto, & Costa, Ana. 2011. Estimación de parámetros de curvas de crecimiento de ganado Nellore criado en confinamiento. *Revista MVZ Córdoba*, **16**(3), 2701–2710.

Quiceno, Jaime, Martinez, Rodrigo, Mateus, Henry, Gallego, Jaime, & Medina, Pedro. 2012. Crecimiento en pastoreo rotacional de toretes de razas criollas Romosinuano y Blanco Orejinegro en Colombia. *Revista MVZ Córdoba*, **17**(1), 2891–2899.

Song, X, Bokkers, EAM, van der Tol, PPJ, Koerkamp, PWG Groot, & Van Mourik, S. 2018. Automated body weight prediction of dairy cows using 3-dimensional vision. *Journal of dairy science*, **101**(5), 4448–4459.

Tebug, Stanly Fon, Missohou, Ayao, Sourokou Sabi, Souahibou, Juga, Jarmo, Poole, Elizabeth Jane, Tapio, Miika, & Marshall, Karen. 2018. Using body measurements to estimate live weight of dairy cattle in low-input systems in Senegal. *Journal of Applied Animal Research*, **46**(1), 87–93.

Wikipedia, La enciclopedia libre. 2021. *Bos primigenius indicus*.

Wirth, Rüdiger, & Hipp, Jochen. 2000. CRISP-DM: Towards a standard process model for data mining. *In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Springer-Verlag London, UK.