Machine learning as a way to have more accuracy when defining milk quality classification

10.56238/homeinternationalanais-069

Matheus Henrique Lopes Cunha
Universidade de São Paulo

Hygor Santiago Lara
Universidade Estadual de Campinas

Keywords: Machine learning, Random Forest, Extra Trees Classifier, KNeighbors Classifier

1 INTRODUCTION

The classification of the quality of milk is necessary to foresee several evaluation parameters, such as its value, market analysis and how to improve it when it has not been presenting a good level. Among several extremely important factors, the analysis and search for more performance is necessary. The milk classification is a strategic tool before, during, and after any process that involves it.

In general, we can say that this analysis is an important factor for companies in this segment to define which are the best suppliers and, furthermore, what is the destination of each type of milk. With this, companies can identify the best ways for the marketing, loyalty and distribution, promoting a greater engagement of their products.

In order to classify milk, machine learning algorithms can be used as a tool. Such algorithms are based on several milk characteristics, such as pH, turbidity, temperature, fat rate, taste and the relationship that all the previous characteristics have with the evaluation of some milks that have already been classified. For Piovezan (2022, p.3) "Machine learning is an application of artificial intelligence that simplistically boils down to geometry problems. Applied models are programmed to interpret data from a data set, learn and improve themselves according to their experience within what they have been given. "

In view of the above, it was realized that this analysis is a technique that involves powerful technology, has as its great intention to speed up the classification process and bring more accurate results about the real classification of milk. With this quoted can be concluded that the study of the classification of milk is extremely important and also generator of much knowledge. For Junior (2022, p. 12) "Technology is able to open doors like never seen before, we can follow that people with visual impairment today are able to follow social networks have their engagement comment and have the same access as everyone, this is wonderful I've been saying for years that the future is to include every day more people, children have access to videos that are entertaining, and the idea is every day more people and the web have this inclusion we will have a different world, although some make some toxic environments still exist people who want these good environments and lean on them."

This work aims to create a machine learning algorithm that has a high accuracy rate when trying to predict the final milk classification. The data presented, were taken from the milk classification database, which can be found on the Kaggle website. The data was analyzed using the artificial intelligence programming language Python, through which statistical analysis and predictions with machine learning were provided.

2 METHODOLOGY

The research was designed with the objective of understanding the statistical information regarding the analysis of the milk classification. Through this analysis, create machine learning models to determine its final quality safely, in order to understand how this space was formed and developed.

The methodology used in this research is of an explanatory nature, and its nature is descriptive. The study was of the quantitative and qualitative type, using data analysis and the use of machine learning algorithms by the Python programming language, for data interpretation. Given this, the quantitative study, is usually conducted by: "Data collection is usually performed in these studies by questionnaires and interviews that present distinct and relevant variables for research, which in analysis is usually presented by tables and graphs." (DALFOVO, LANA E SILVEIRA; 2008, p.10). In the qualitative study for Carspecken (2011, p.27) "The qualitative social researcher will usually want to understand how forms of power work, specifically in real interactions that he observes and possibly participates".

The explanatory study is a method of scientific analysis that seeks to explain how it works and the performance achieved by the models of the segment studied. The study comes in the quantitative mold, since the intention is to bring an approach with numerical data, which will be presented in graphic, discursive and statistical formats, with the intent of understanding which machine learning model has the greatest applicability for the study in question. According to Dalfovo, Lana and Silveira (2008, p. 8) "In general, like experimental research, quantitative field studies are guided by a research model where the researcher starts from conceptual frames of reference as well structured as possible, from which he/she formulates hypotheses about the phenomena and situations he/she wants to study.". And the work also has qualitative features because there is meaningful and transformative data analysis, as Carspecken (2011, p.29) said: "Critical qualitative research is truly stimulating, political, meaningful, mind-expanding when truly practiced. Both fieldwork and data analysis experiences are richly meaningful and transformative".

The term machine learning can be defined according to Tome (2017, p. 24) "The elements of machine learning consist of a set of variables, called features, which can be measured or predefined, and a set of outputs, which can be known or not. The dataset consists of examples that will be used to build the model. "

Machine learning is a complex activity with features that can help both professional and personal development, the concepts and tools presented in it offer ways to create strategies that facilitate reaching the desired goal, so according to Stange (2011, p. 7) "Incremental learning requires the learning mechanism to be based on the dynamic accumulation of information extracted from the experiences performed. Machine learning using adaptivity considers the integration of symbolic machine learning techniques with adaptive techniques to solve learning problems."

The various ways of using machine learning are also essential factors, because it is from them that the most efficient model to be used in each situation is defined, these models can be called attributes, for Almeida (2014, p.19) "The greater the presence of irrelevant and redundant attributes, the greater the difficulty of learning the classifier during the training stage. One way to remove the irrelevant attributes is to select the most important attributes for the classification, that is, those that have greater power to differentiate between positive and negative news.".

The classifiers chosen for the machine learning models that were used to achieve the results of this study use the Random Forest, Extra Trees Classifier, and KNeighbors Classifier methods. The models were used in order to see which classifier would be the most efficient.

The Random Forest method creates several decision trees and takes the results that have been found in the majority of the models as the most important. Tome's definition (2017, p. 30)" Random Forest uses the bagging method, whose central idea is the creation of several samples from the database for learning classifiers, where the final result will be given by the majority vote of the classifiers.".

The other model analyzed, the Extra Trees Classifier, is intended to generate several small trees and make decisions based on these small decision trees. The Extra Trees Classifier model adds another layer of randomness to decision forests. The additional randomization step is introduced at the node during tree training. Instead of searching for the optimal cut-off point, a random threshold value is selected for each feature. Soon after, the search space is reduced, leading to faster training. The downside is that the size and depth of the forest is increased due to the cuts being too low (MAIER et al., 2015).

With the KNeighbors Classifier model, equity is made between the variables that are closest, thus performing a relationship calculation between the variables. Instead of considering only a single nearest neighbor in the data set, the k-th nearest neighbor model uses an arbitrary number k, of neighbors, deciding the output value by voting. This means that for each test point, how many neighbors are ranked 0 and how many are ranked 1 are counted, and the highest frequency binary value is decided. (MÜLLER; GUIDO, 2017).

As for not generating overfitting and underffiting, thus having a more reliable model, because when we throw the data into the models immediately, it tends to have its predictions biased, the division of the data into test data and training data was used. For Lopes (2018, p.8) "In problems where the main objective is to choose which model has the greatest predictive power one should be careful with overfitting and underfitting, the former being when the model fits the training data so well that it makes terrible out-of-sample predictions, and the latter concerns when the model does not fit well even in the training set. ”

Before applying the machine learning models, statistical tests were performed to see the importance of each attribute and consequently whether the attribute could be irrelevant to the model. The boxplot was analyzed to detect outliers and correlation to analyze possible relationships between variables. In addition, the data was also checked for normality to validate that the number of samples is sufficient for the model.

After the static analyses, the GridSearchCV function was used, with cross-validation on 10 folds, to define the best parameters for each model. With this already defined, the data was divided into 80% for training and 20% for testing, where the 1059 samples were divided into 847 for training and 212 for testing, randomly selected by the function train_test_split. For Campos and Miguel (2013, p. 204) "The objective of the implementation of the technical process standard is to reduce the amount of changes made in the parameters of process control in the introduction of a new product, contributing to a more effective preparation (setup) machine, reducing losses in productivity and quality, making it possible to eliminate the variability of the specifications that occur during production.”

The models were evaluated for their accuracy, sensitivity and precision, calculated using their confusion matrix. Another important analysis for this study is the noise test. It was verified up to what percentage of noise inclusion in the models still maintain good performance. For Sano and Filho (2013, p. 37) "The crucial need for more efficiency, efficacy and effectiveness (3Es) of government actions is intrinsically related to the issue of social development, because its possibilities are often curtailed, due to the limits that arise when the actors involved in public management are not committed to these concepts, resulting in negative impacts on the lives of all citizens.”

The methods and tools that present inefficiency need to be improved. Thus, it was of great importance for the study the use of several techniques, because in some very positive results were found and consequently were maintained and others were inefficient and discarded. For De Figueiredo and Cabral (2020, p.86) "According to the concepts pointed out, it is important to note that the technique known as Machine Learning (ML) has gained great prominence in several areas. The machine learning technique is used so that computers are programmed to learn from past experience, that is, this programming does not just reproduce what was fed into the system with the insertion of data, but the system has a cognitive capacity of its own, which enables the condition to continuously learn from experience, whether with successes or failures."

In this context, the goal of the algorithm is to produce a classifier that can predict what the final milk quality classification will be, even when the information is not very clear to statistical analysis.

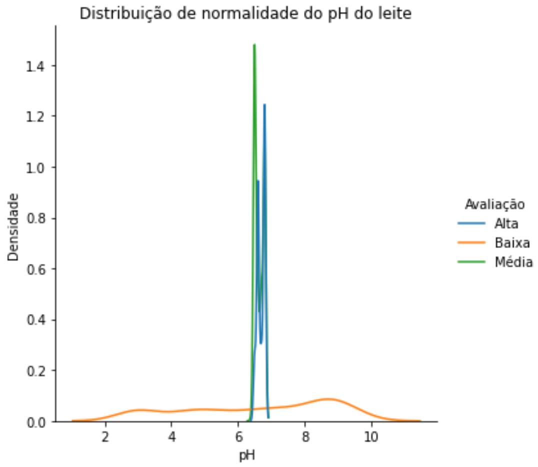
3 CONCLUSION

The central objective of this article is to show how machine learning can be used for the measurement of milk quality evaluation. Presenting the positive and negative points seeking to evolve the aspects related to machine learning tools within this segment. For Domingos (2017, p. 40) "Machine learning is the scientific method using steroids. It follows the same process of generating, testing, and discarding or refining hypotheses. However, while a scientist might spend his or her entire life creating and testing a few hundred hypotheses, a machine learning system can do the same in a fraction of a second. Machine learning automates discovery. So it's not surprising that it is revolutionizing science as well as business."

The elaboration of the model was based on the statistical analysis of some milk characteristics, having in mind the relationship that these characteristics have with the final evaluation of the milk. During the process, the characteristics related to the pH of the milk, the temperature it presented at the moment of the analysis, its taste, odor, fat level, turbidity, and the color it presented at the moment were analyzed. This measurement brought satisfactory results to be used in the models.

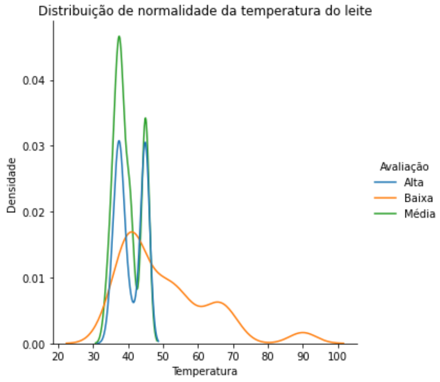
It is also important to point out that normality tests were performed on the variables that were introduced in the machine learning model. Analyses were made with the normality distribution graphs. And the tests found that all variables tested positive for normality and consequently were used in this work.

Regarding the pH, the analyses concluded that the pH of low grade milk varies between 3 and 9.5. The medium grade milk, on the other hand, has a pH ranging between 6.5 and 7. The high grade milk has a pH between 6.05 and 7. In the graph below we can see how the low grade milk is more dispersed and those with medium and high evaluation, have a higher concentration at a certain point of the normality distribution graph.



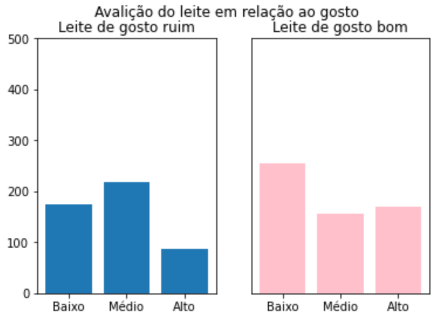
Translation:
Distribuição de normalidade do ph do leite: normality distribution of milk ph
Avaliação: evaluation
Alta: high:
Baixa: low
Média: average

The temperature is another point that had an interesting analysis, the low grade milk is between 34º and 90º. The medium classification has between 34º and 45º. The high grade milk is between 35º and 44º. In the normality distribution chart below, it can be seen that the low quality milk has a great dispersion, but maintains its peak at 40°, while the medium and high quality have their normality values between 30° and 40°.



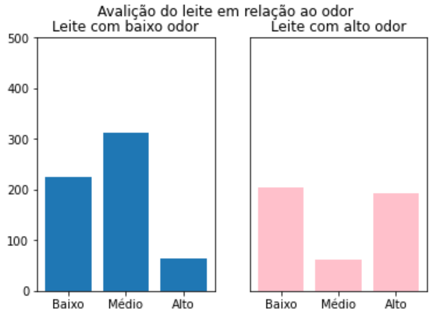
Translation:
Distribuição de normalidade da temperatura do leite: normality distribution of milk temperature
Avaliação: evaluation
Alta: high:
Baixa: low
Média: average

For taste it was analyzed in two parts, milk that tasted good and milk that did not taste good. Thus seeing which probability has more chance of being of certain classification. The milk with a bad taste presents in majority a medium classification, having more than two hundred samples, in second place comes the low classification with approximately 20 elements less than the previous variable and in last place comes the high classification which has less samples with bad taste. Already the milk of good taste has mostly the low classification, in sequence comes the milk of high classification that presents a large drop in quantity in relation to the previous variable, soon after comes the milk of average classification, which has almost the same amount of variables that the previous sample.



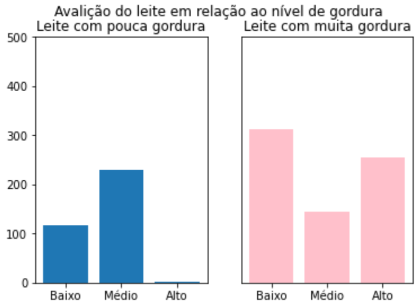
Translation:
Avaliação do leite em relação ao gosto: evaluation of milk in relation to taste
Leite de gosto ruim: bad-tasting milk
Leite de gosto bom: good-tasting milk
Alto: high:
Baixo: low
Médio: average

Then the interference that odor has on the classification of milk was analyzed, dividing it into two categories: low odor and high odor milk, and consequently aligning them with the classification of milk. The milk with low odor had mostly a medium classification, followed by low evaluation, and lastly the high classification, which had few samples, when compared to the two previous variables. The milk with high odor has low and high quality samples, with almost the same size, but the low quality sample is slightly larger and the medium classification had few samples.



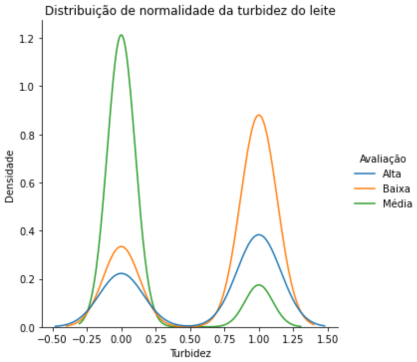
Translation:
Avaliação do leite em relação ao odor: evaluation of milk in relation to odor
Leite com baixo odor: low odor milk
Leite com alto odor: high odor milk
Alto: high:
Baixo: low
Médio: average

And the fat level of the milk was also a factor that had positive results, with the division of milk into high fat and low fat milk. It was realized which classification fits each variable. The low fat milk had practically no samples with high quality level, had more than three hundred samples with medium evaluation level and with approximately one hundred samples less appears the low evaluation milk. And the high fat milk had the evaluation with the most emphasis on the low classification, with more than three hundred samples in it, then comes the high evaluation milk with samples with values very close to the sample that is in first place, last with a much smaller sample, comes the medium evaluation milk.



Translation:
Avaliação do leite em relação ao nível de gordura: milk evaluation in relation to fat level
Leite com pouca gordura: low-fat milk
Leite com muita gordura: high-fat milk
Alto: high:
Baixo: low
Médio: average

Turbidity was divided into high and low turbidity milk, having analyzed how much each classification fits a certain level of turbidity. Regarding its normality, we have in the graph below the division into two peaks, in which the peak referring to 0 on the axis, represents milk with low turbidity and peak 1 on the x-axis, represents milk with high turbidity. The peak 0 has as its highest point the medium quality milk, followed by low and high classification milk, but both having a number well below the first. The peak 1, on the other hand, has as its highest point the low classification milk, followed by high and medium quality milk with the peak well below the highest.



Translation:
Distribuição de normalidade da turbidez do leite: normality distribution of milk turbidity
Avaliação: evaluation
Alta: high:
Baixa: low
Média: average

With the analysis performed, the machine learning models were implemented, the first was the Random Forest, being made first the test to know the best parameters to be used that were made with GridSearchCV. Soon after the model was trained and tested, with the accuracy test, precision, sensitivity and confusion matrix, being done soon after, the results obtained were:

Accuracy	99.05%
Confusion matrix	2 erros
Medium sensitivity	99.01%
Medium accuracy	99.15%

Then KNeighbors Classifier model, popularly known as Knn, was used, the parameters were set in the same way as the first models, through GridSearchCV. And model was trained and tested, then the fine results were:

Accuracy	99.05%
Confusion matrix	2 erros
Medium sensitivity	99.03%
Medium accuracy	99.01%

And lastly the Extra Tree Classifier model was tested, in which the standard procedure of first defining the parameters and then training and testing the model was maintained. And the final results are:

Accuracy	99.05%
Confusion matrix	2 erros
Medium sensitivity	99.01%
Medium accuracy	99.15%

After the analysis is done and with its efficiency tested, one realizes that it has that the results of the models are very similar, and this is due to the fact that the database has few samples, moreover as it is relatively small the number of samples, the data used for testing and training, are exactly the same. The samples are not bad, but if they were larger, they would be more diverse and more efficient in comparing the models.

After all the process to test the model and see its validity, proving a good performance, another test was done to test up to what percentage of noise the models can maintain a good performance. And from this it was verified that when it passes 50% noise, the models tend to have a performance with accuracy well below the desired value. The results of the decrease in accuracy are shown in the following table:

Accuracy by Noise percentage					
Model	10%	20%	30%	40%	50%
Random Forest	92.92%	89.15%	83.49%	80.18%	78.30%
Knn	89.62%	76.88%	69.81%	66.03%	64.62%
Extra Tree	91.03%	86.79%	79.71%	74.52%	67.45%

The present article aimed to analyze the performance of machine learning algorithms for the definition of milk quality classification. Positive results of this tool were obtained, and the importance of such techniques for a more effective performance of the predictions was also evidenced. In a broader context, it can be seen that even being a great innovation, the machine learning models within this concept still need to be improved. Even in the face of the difficulties presented, the study shows that the tool has an effective result.

The most efficient model was the Random Forest, even though the performance of the others were very similar. It was the one that obtained the best performance in relation to the noise test, thus being the model with the most efficient result and also considered the best model. It presented 99.05% of accuracy against the original data and a lower performance when faced with noise. In future works, more robust models should be sought.

It is concluded that this work provides a tool to identify the milk quality classification. It is believed that the use of classifiers will allow us to reach a parameter that will define whether the milk classification will be low, medium or high. This will be useful for entities looking for more efficient ways to market milk..

REFERÊNCIAS

ALMEIDA, Filipe Guedes de Oliveira. Classificadores de polaridade de notícias utilizando ferramentas de machine learning: o caso da Vale SA. 2014.

CAMPOS, Roni CP; MIGUEL, Paulo A. Cauchick. Melhoria do processo de produção por meio da aplicação do Desdobramento da Função Qualidade. Sistemas & Gestão, v. 8, n. 2, p. 200-209, 2013.

CARSPECKEN, Phil Francis. Pesquisa qualitativa crítica: conceitos básicos. Educação & Realidade, v. 36, n. 2, p. 395-424, 2011.

DALFOVO, Michael Samir; LANA, Rogério Adilson; SILVEIRA, Amélia. Métodos quantitativos e qualitativos: um resgate teórico. Revista interdisciplinar científica aplicada, 2008.

DE FIGUEIREDO, Carla Regina Bortolaz; CABRAL, Flávio Garcia. Inteligência artificial: machine learning na Administração Pública: Artificial intelligence: machine learning in public administration. International Journal of Digital Law, v. 1, n. 1, p. 79-96, 2020.

DOMINGOS, Pedro. O algoritmo mestre: como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo. Novatec Editora, 2017.

JUNIOR, Bendev. Transformando códigos em sonhos: conselhos que gostaria de receber ao entrar na área da tecnologia. SEVEN publicações acadêmicas, 2022.

LOPES, Lucas Pereira. Predição do preço do café Naturais Brasileiro por meio de modelos de statistical machine learning. Sigmae, v. 7, n. 1, p. 1-16, 2018.

MAIER, O. et al. Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. Journal of neuroscience methods, [S.l.], v.240, p.89–100, 2015.

MÜLLER, Andreas C.; GUIDO, Sarah. Introduction to Machine Learning with Python: a guide for data scientists. 2. ed. Sebastopol: O’reilly Media, Inc., 2017. 392 p.

PIOVEZAN, Raphael Paulo Beal et al. Método de aprendizagem de máquina visando prever a direção de retornos de exchange traded funds (ETFs) com utilização de modelos de classificação e regressão. 2022.

SANO, Hironobu; MONTENEGRO FILHO, Mário Jorge França. As técnicas de avaliação da eficiência, eficácia e efetividade na gestão pública e sua relevância para o desenvolvimento social e das ações públicas. Desenvolvimento em questão, v. 11, n. 22, p. 35-61, 2013.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. Understanding machine learning: from theory to algorithms. [S.l.]: Cambridge university press, 2014.

TOMÉ, Vívian Tostes. Utilização de machine learning para categorização dos gastos de bitcoin no Brasil. 2017. Tese de Doutorado.