Journal of Management & Technology
e-ISSN: 2177-6652

MACHINE LEARNING AS A WAY TO BE MORE PRECISE WHEN DEFINING MILK QUALITY CLASSIFICATION

APRENDIZADO DE MÁQUINA COMO FORMA DE SER MAIS PRECISO NA DEFINIÇÃO DA CLASSIFICAÇÃO DA QUALIDADE DO LEITE

EL APRENDIZAJE AUTOMÁTICO COMO FORMA DE SER MÁS PRECISOS A LA HORA DE DEFINIR LA CLASIFICACIÓN DE LA CALIDAD DE LA LECHE

How to cite:
Cunha, Matheus H. L. Santiago, Hygor. (2023). Machine learning as a way to be more precise when defining milk quality classification. Revista Gestão & Tecnologia. v. 23, nº 4, 2023, v. 23, nº 4, p: 222-237

Hygor Santiago
Engenheiro mecânico formado pela UFSJ
http://orcid.org/0000-0002-4835-5498

Matheus Cunha
Graduado em Markentig pela USP
https://orcid.org/0009-0005-1370-3498

Scientific Editor: José Edson Lara
Organization Scientific Committee
Double Blind Review by SEER/OJS
Received on 09/11/2022
Approved on05/12/2023

Abstract

Objective: Machine learning algorithm that manages to have a high rate of use in relation to the prediction of the evaluation of milk quality.
Methodology: Is explanatory in nature, quantitative and qualitative methods were used, and data was used to understand the final analyses.
Originality: The article shows how the models were developed, the tests applied before the implementation of the models, the utilization rate of each model and also an analysis of which is the most efficient model for a specific situation.
Mains results: The specifications of each machine learning model and its impact on the development of the models that were used in the work were determined by tests and applications made in the Python programming language; the positive and negative results were considered to arrive at a final position on what was the best way to use the algorithms in this case.
Theorical contributions: This work contributes to the literature on computer science world and in agricultural world too.
Social contributions: The article concludes that the application of Machine learning models in milk quality classification can help many companies or organizations that need to streamline processes and increase the accuracy rate when measuring milk classification, to have a high improvement in this process and consequently increase productivity and profitability.

Keywords: *Machine learning, Random Forest, Extra Trees Classifier, KNeighbors Classifier,* Milk, Agribusiness

Resumo

Objetivo: Algoritmo de aprendizado de máquina que consegue ter alto índice de utilização em relação à predição da avaliação da qualidade do leite.
Metodologia: É de natureza explicativa, foram utilizados métodos quantitativos e qualitativos e os dados foram utilizados para compreensão das análises finais.
Originalidade: O artigo mostra como os modelos foram desenvolvidos, os testes aplicados antes da implementação dos modelos, a taxa de utilização de cada modelo e também uma análise de qual é o modelo mais eficiente para uma situação específica.
Principais resultados: As especificações de cada modelo de aprendizado de máquina e seu impacto no desenvolvimento dos modelos utilizados no trabalho foram determinados por testes e aplicações feitas na linguagem de programação Python; foram considerados os resultados positivos e negativos para se chegar a uma posição final sobre qual a melhor forma de utilização dos algoritmos neste caso.
Contribuições teóricas: Este trabalho contribui para a literatura no mundo da ciência da computação e também no mundo agrícola.
Contribuições sociais: O artigo conclui que a aplicação de modelos de Machine Learning na classificação da qualidade do leite pode ajudar muitas empresas ou organizações que precisam agilizar processos e aumentar o índice de precisão na medição da classificação do leite, a ter uma alta melhoria neste processo e consequentemente aumentar a produtividade e rentabilidade.

Palavras-chave: Aprendizado de máquina, Random Forest, Classificador Extra Trees, Classificador KNeighbors, Leite, Agronegócio

Resumen

Objetivo: Algoritmo de aprendizaje automático que logra tener un alto índice de utilización en relación a la predicción de la evaluación de la calidad de la leche.
Metodología: Es de naturaleza explicativa, se utilizaron métodos cuantitativos y cualitativos, y se utilizaron datos para comprender los análisis finales.
Originalidad: El artículo muestra cómo se desarrollaron los modelos, las pruebas aplicadas antes de la implementación de los modelos, la tasa de utilización de cada modelo y también un análisis de cuál es el modelo más eficiente para una situación específica.
Principales resultados: Las especificaciones de cada modelo de aprendizaje automático y su impacto en el desarrollo de los modelos que se utilizaron en el trabajo fueron determinadas mediante pruebas y aplicaciones realizadas en el lenguaje de programación Python; Se consideraron los resultados positivos y negativos para llegar a una posición final sobre cuál era la mejor forma de utilizar los algoritmos en este caso.
Contribuciones teóricas: Este trabajo contribuye a la literatura sobre el mundo de la informática y también en el mundo agrícola.
Contribuciones sociales: El artículo concluye que la aplicación de modelos de Machine Learning en la clasificación de la calidad de la leche puede ayudar a muchas empresas u organizaciones que necesitan agilizar procesos y aumentar la tasa de precisión al medir la clasificación de la leche, a tener una alta mejora en este proceso y en consecuencia aumentar la productividad. y rentabilidad.

Palabras clave: Aprendizaje automático, Bosque aleatorio, Clasificador de árboles adicionales, Clasificador KNeighbors, Leche, Agronegocios

1. INTRODUCTION

The classification of the quality of milk is necessary to predict 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 factors of extreme importance, it is necessary to analyze and search for more performance. Milk classification is a strategic tool before, during and after any process involving it.

In general, it can be said that this analysis is an important factor for companies in this segment to be able to define which are the best suppliers and, in addition, what is the destination given to each type of milk. With this, companies can identify the best ways for commercialization, loyalty and distribution, promoting greater engagement of their products.

Machine learning algorithms can be used to classify the milk as a tool. Such algorithms are based on several milk characteristics, such as pH, turbidity, temperature, fat content, 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 is simply summarized in geometry problems. Applied models are programmed to interpret data from a data set, learn and improve according to their experience within what was provided. ”

In view of the above, it should be noted that this analysis is a technique that involves powerful technology, with the main purpose of streamlining the classification process and bringing more accurate results on the actual classification of milk. With that said, it can be concluded that the study of milk classifications is extremely important and also generates a lot of knowledge. For Junior (2022 , p. 12) “Technology is capable of opening doors like never before, we can monitor which visually impaired person today is able to follow social networks, have their engagement comment and have the same access as everyone else, this is wonderful I have been saying for years that the future is inclusive to include more people every day, children have access to videos that they have fun with, and the idea is that more people and the web have this inclusion like we will have a different world, although some make it some toxic environments yet there are 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 in relation to trying to predict the final classification of the milk. The data presented were taken from the milk classification database, which can be found on the Kaggle website. The data were analyzed through the artificial intelligence of the Python programming language, through which statistical analyzes and predictions were provided with *machine learning.*

2. Methodology

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

The methodology used in this research is of explanatory character, and has as its descriptive nature. The study was quantitative and qualitative, using data analysis and the use

of *machine learning algorithms* using the Python programming language for data interpretation. In view of this, the quantitative study is usually carried out by: “Data collection is usually carried out in these studies through questionnaires and interviews that present distinct and relevant variables for research, which in analysis is usually presented by tables and graphs”. (Dalfovo, Lana & 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 in which he 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 studied segment. The work comes in the quantitative mold, as the intention is to bring an approach with numerical data, which will be presented in graphic, discursive and statistical format, in order to be able to understand what the machine learning model is. That has greater 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 as well structured as possible, from which he formulates hypotheses about the phenomenal and situations he wants to study.”. And the work also has qualitative traits, as there is significant and transformative data analysis, as Carspecken (2011, p.29) said: “Critical qualitative research is really stimulating, political, meaningful, it expands the mind when truly practiced. Both the 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 do not. The data set is formed by examples that will be used to build the model.”

The *machine learning* is a complex activity with characteristics that can help both in professional and personal development, the concepts and tools presented in it offer ways to create strategies that facilitate reaching the desired goal. According to Stange (2011, p. 7) “Learning incremental requires that the learning mechanism be based on the dynamic accumulation of information extracted from the experiences carried out. Machine learning using

adaptability considers the integration of symbolic machine learning techniques with adaptive techniques for solving learning problems."

The various ways to use the *machine learning*, are also essential factors, as 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 attributes and redundant, the greater the difficulty of learning the classifier during the training stage. One way to remove irrelevant attributes is by selecting the most important attributes for classification, that is, those that have the greatest 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 methods of *Random Forest, Extra Trees Classifier* and *KNeighbors Classifier*. The models were used in order to see which would be the most efficient classifier.

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

The other model analyzed is *Extra Trees Classifier*, is intended to generate several small trees and make decisions based on these small decision trees. The *Extra Trees model Classifier*, adds another layer of randomness to decision forests. The additional randomization step is introduced into the node during tree training. Instead of looking for the ideal 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 model Classifier*, 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 neighbors model uses an arbitrary number *k*, of neighbors, deciding the output value by means of voting. This means that for each test point, how many neighbors are classified as 0 and how many are classified as 1 are counted, being decided by the highest frequency binary value. (Müller & Guido, 2017).

In relation to not generating *overfitting* and *underfitting*, thus having a more reliable model, because when we throw the data in the models immediately, it tends to get its predictions vitiated, we used the division of the data into test and training data. To Lopes (2018, p.8) “In problems where the main objective is to choose which model has the greatest predictive power, care must be taken with *overffiting* and *underffiting* , the first being when the model fits so well to training data that makes poor out-of-sample predictions, and the second concerns when the model does not fit well even on the training set. ”

Before applying *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 were also checked for normality to validate whether the number of samples is sufficient for the model.

GridSearchCV function was used, with cross validation in 10 *folds*, to define the best parameters for each model. With that already defined, the data were 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 *train_test_split function*. For Campos and Miguel (2013, p. 204) “The purpose of implementing the technical process standard is to reduce the amount of changes made to process control parameters when introducing a new product, contributing to machine setup. More effectively, reducing productivity and quality losses, making it possible to eliminate the variability of 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 to what percentage of inclusion of noise in the models still maintain good performance. For Sano and Filho (2013, p. 37) “The crucial need for more efficiency, effectiveness and effectiveness (3Es) of government actions is intrinsically related to the issue of social development, as 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.”

Methods and tools that are inefficient need to be improved. Thus, it was of great importance for the study to use several techniques, as 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 punctuated concepts, it is important to point out that the technique known as *machine learning* or, simply, Machine Learning (ML) has gained great prominence in several areas. the *machine learning* technique is used so the computers are programmed to learn from past experience, that is, this programming not only reproduces what was fed into the system with the insertion of data, but the system has its own cognitive capacity, which enables the condition of continuously learn from experience, whether with successes or failures.”

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

3. CONCLUSION

The main objective of this article is to show how the *machine learning* can be used to measure milk quality assessment. Presenting the positive and negative points, seeking to evolve aspects related to machine learning tools within this segment. For Domingos (2017, p. 40) “The *machine learning* is the scientific method on steroids. It follows the same process of generating, testing, and discarding or refining hypotheses. However, whereas a scientist might spend his entire life creating and testing a few hundred hypotheses, a *machine learning system* can do the same in a fraction of a second. The *machine learning* automate the discovery. So, it is not surprising that it is revolutionizing science as well as business.”

The elaboration of the model was based on the statistical analysis of some characteristics of the milk, considering 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 had at the time of analysis, the taste, odor, fat level, turbidity and the color it had at the time were analyzed. This measurement brought satisfactory results to be used in the models.

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

Regarding the pH, the analyzes came to the conclusion that low classification milk has a pH with variations between 3 and 9.5. Medium classification milk, on the other hand, has a pH varying between 6.5 and 7. High classification milk has a pH between 6.05 and 7. The graph below shows how low classification milk is more dispersed and those with medium and high ratings show a higher concentration at a certain point on the normality distribution graph.

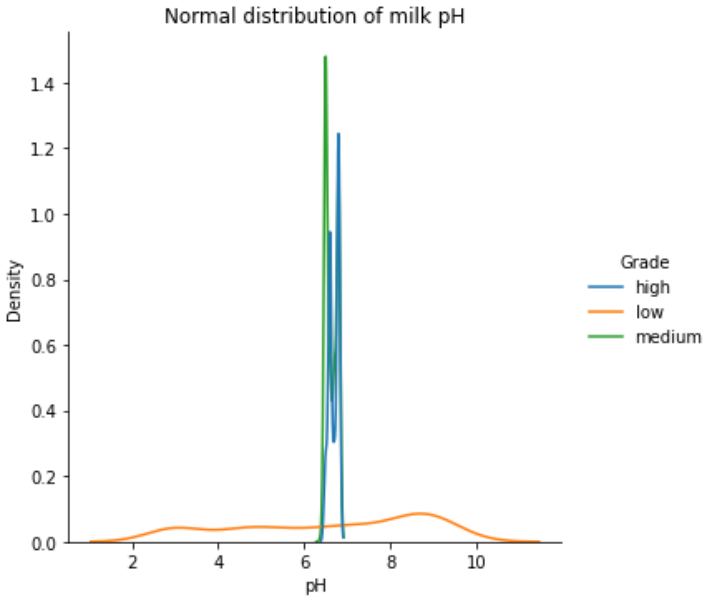

Figure 1: Normal distribution of milk pH
Source: research data

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

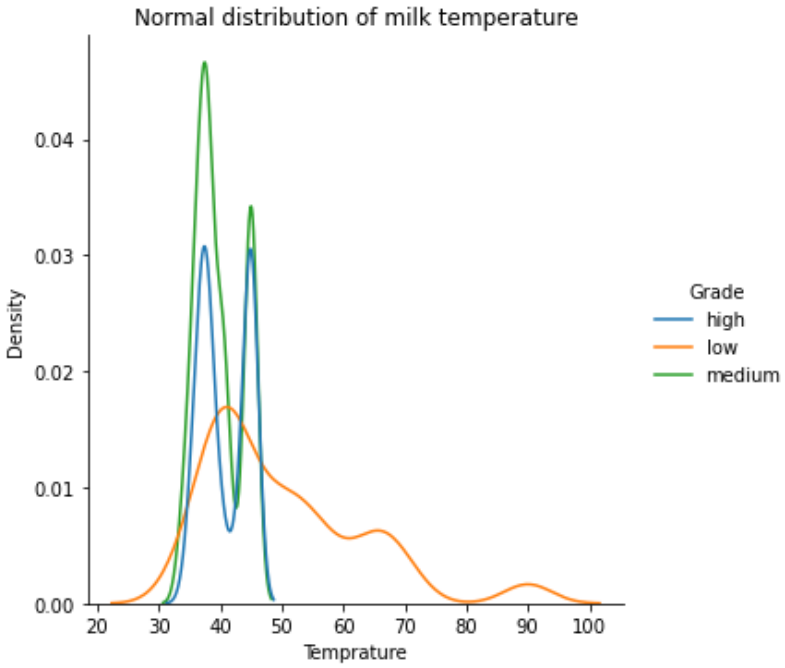

Figure 2: Normal distribution of milk temperature
Source: research data

For taste it was analyzed in two parts, the milk that tasted good and the milk that did not taste good. So seeing which probability has more chance of being of a certain classification. Bad-tasting milk mostly presents an average classification, with 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 bad tasting. On the other hand, milk with a good taste has mostly a low classification, followed by high-class milk, which has a large drop in quantity compared to the previous variable, then comes medium-class milk, which has almost the same amount of variables than the previous sample.

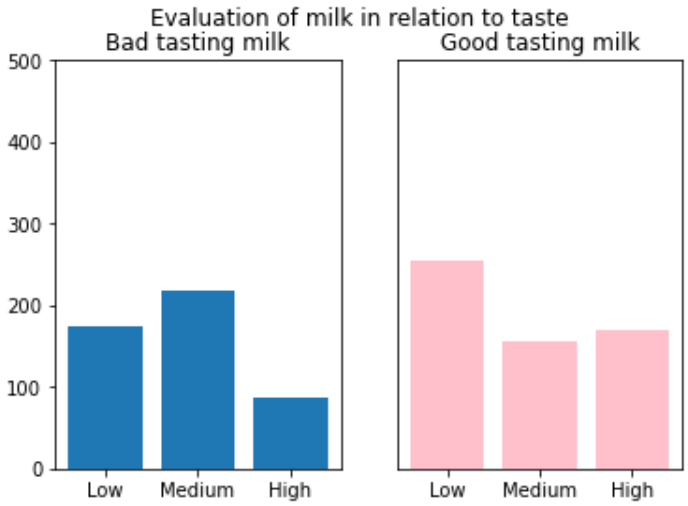

Figure 3: Evaluation of milk in relation to taste
Source: research data

Soon after, the interference that odor has on milk classification was analyzed, dividing it into two categories: low-odor milk and high-odor milk, and consequently aligning them with milk classification. Milk with low odor mostly had the average classification, followed by the low evaluation, with the high classification in last place, which presented few samples, when comparing the two previous variables. High odor milk, on the other hand, has low and high quality samples, practically the same size, with the low quality sample being slightly larger and the average classification having few samples.

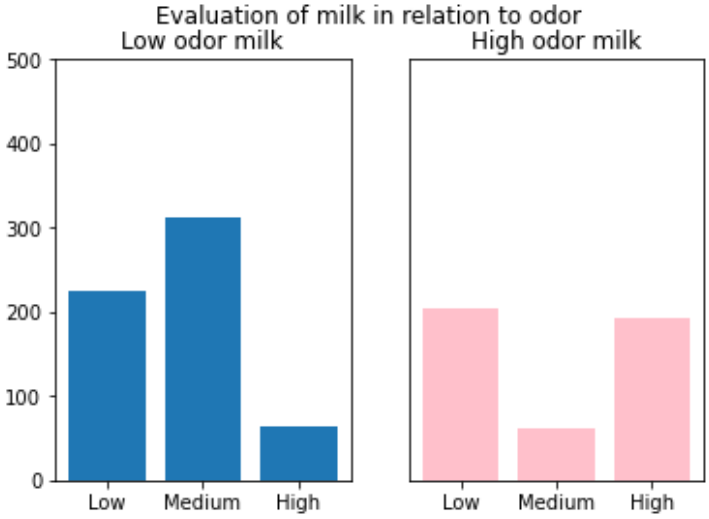

Figure 4: Evaluation of milk in relation to odor
Source: research data

And the level of fat in the milk was also a factor that had positive results, with the division of milk into: milk with a lot of fat and milk with little fat. It was noticed which classification fits each variable. Low-fat milk had practically no samples with a high level of quality, there were more than three hundred samples with an average level of evaluation and with approximately one hundred samples less, low-evaluation milk appears. And the milk with a lot of fat had the evaluation with the most emphasis on the low classification, having more than three hundred samples in it, soon after comes the high evaluation milk with samples with values very close to the sample that is in first place, in last with a much smaller sample, comes the average evaluation milk.

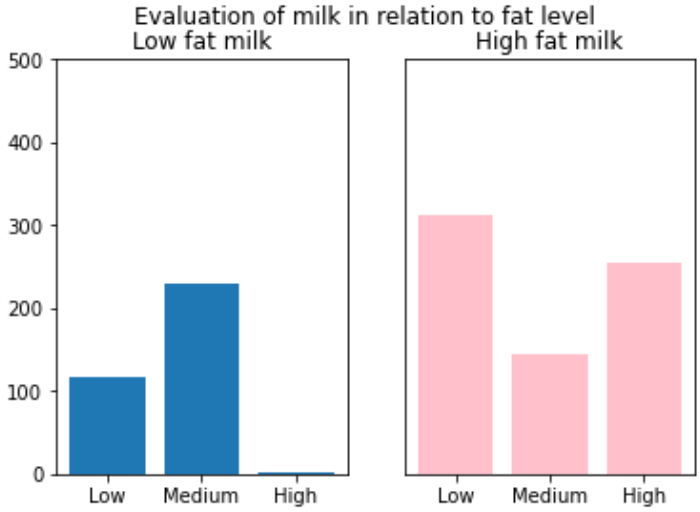

Figure 5: Evaluation of milk in relation to fat level
Source: research data

Turbidity was divided into high and low milk, analyzing how much each classification fits into a given turbidity level. Regarding its normality, the graph below shows 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. Peak 0 has medium quality milk as its highlight, followed by low and high grade milk, but both having a much lower number than the first. Peak 1, on the other hand, has low-grade milk as its highest point, accompanied by high- and medium-quality milk with a much lower peak compared to the highest.

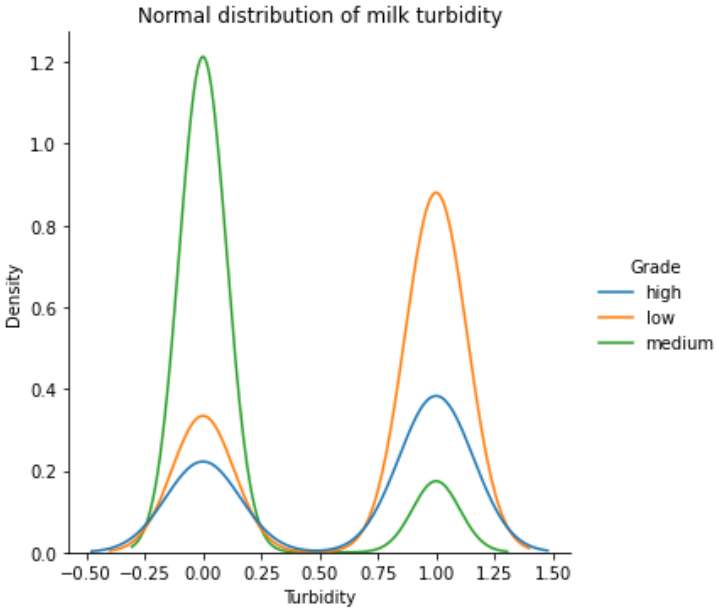

Figure 6: Normal distribution of milk turbidity
Source: research data

With the analysis carried out, the *machine learning models* were implemented. The first was the *Random Forest*, the test being done first to know the best parameters to use which were done with *GridSearchCV*. Soon after the model was trained and tested, with the test of accuracy, precision, sensitivity and confusion matrix, being done right after, the results obtained were:

Accuracy	99.05%
confusion matrix	2 mistakes
medium sensitivity	99.01%
average accuracy	99.15%

Figure 7: statistic parameters of Randon Foresty with GridSearchCV
Source: research data

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

Accuracy	99.05%
confusion matrix	2 mistakes
medium sensitivity	99.03%
average accuracy	99.01%

Figure 8: statistic parameters of KNeighbors with GridSearchCV
Source: research data

And finally, the *Extra Tree model was tested Classifier*, in which the standard procedure of first defining the parameters and then training and testing the model was maintained, ending with the performance test that it presented. And the end results are:

Accuracy	99.05%
confusion matrix	2 mistakes
medium sensitivity	99.01%
average accuracy	99.15%

Figure 8: statistic parameters of Extra Tree model with GridSearchCV
Source: research data

After the analysis carried out and with its efficiency tested, it is noticed that the results of the models are very similar, and this is due to the fact that the database has few samples, in addition to the relatively small 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 diversified and more efficient when comparing the models.

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

Accuracy by percentage of noise					
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%

Figure 9: Accuracy by percentage of noise
Source: research data

This article aimed to analyze the performance of *machine learning algorithms* to define milk quality classification. Positive results were obtained from this tool and the importance of such techniques for more effective performance of forecasts was also evident. In a broader

context, it can be seen that even though it is a great innovation, *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 *Random Forest*, even though the use of the others was 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 being considered the best model. It presented 99.05% % of accuracy in face of the original data and inferior performance when faced with noise. In future work, 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 reaching a parameter that will define whether the milk classification will be low, medium or high. This will be useful for entities that seek more efficient ways to market milk.

REFERENCES

Almeida, F. G. de O. (2014). Classificadores de polaridade de notícias utilizando ferramentas de machine learning: o caso da Vale SA.

Campos, R. C. P., & Miguel, P. A. C. (2013). Melhoria do processo de produção por meio da aplicação do Desdobramento da Função Qualidade. Sistemas & Gestão, 8(2), 200-209.

Carspecken, P. F. (2011). Pesquisa qualitativa crítica: conceitos básicos. Educação & Realidade, 36(2), 395-424.

Dalfovo, M. S., Lana, R. A., & Silveira, A. (2008). Métodos quantitativos e qualitativos: um resgate teórico. Revista interdisciplinar científica aplicada.

De Figueiredo, C. R. B., & Cabral, F. G. (2020). Inteligência artificial: machine learning na Administração Pública: Artificial intelligence: machine learning in public administration. International Journal of Digital Law, 1(1), 79-96.

Domingos, P. (2017). O algoritmo mestre: como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo. Novatec Editora.

Junior, B. (2022). Transformando códigos em sonhos: conselhos que gostaria de receber ao entrar na área da tecnologia. SEVEN publicações acadêmicas.

Lopes, L. P. (2018). Predição do preço do café Naturais Brasileiro por meio de modelos de statistical machine learning. Sigmae, 7(1), 1-16.

Maier, O., et al. (2015). Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. Journal of neuroscience methods, 240, 89–100.

Müller, A. C., & Guido, S. (2017). Introduction to Machine Learning with Python: a guide for data scientists (2ª ed.). Sebastopol: O'reilly Media, Inc.

Piovezan, R. P. B., et al. (2022). Método de aprendizagem de máquina visando prever a direção de retornos de exchange traded funds (ETFs) com utilização de modelos de classificação e regressão.

Sano, H., & Montenegro Filho, M. J. F. (2013). 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, 11(22), 35-61.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: from theory to algorithms. Cambridge: Cambridge University Press.

Tomé, V. T. (2017). Utilização de machine learning para categorização dos gastos de bitcoin no Brasil. Tese de Doutorado.