Machine Learning Applied to Milk Sample Classification

Mia León and Diego Ossa,
Universidad de Lima, Carrera de Ingeniería Industrial
Lima, Perú
20172282@aloe.ulima.edu.pe; 20172430@aloe.ulima.edu.pe

José Antonio Taquía Gutiérrez
Universidad de Lima, Instituto de Investigación Científica,
Carrera de Ingeniería Industrial.
Lima, Perú
jtaquia@ulima.edu.pe

Abstract

The document presents the results of the evaluation of the milk sample classification process through the modeling of machine learning techniques, with random forest being the most accurate according to its accuracy percentage of 96%. The paper presents the results of the evaluation of the milk sample classification process through the modeling of machine learning techniques. This research aimed to discriminate the presence or absence of adulterants, which allows the obtaining of dairy products suitable for human consumption. Also, accelerate and specify the inspection process of these samples. The relevance of the present study can be understood from the product under analysis: milk. This is mass consumption, especially in children. Therefore, it is considered relevant to demonstrate efficiently that quality products are provided to the population and this document is a contribution to the credibility of the integrity of dairy products.

Keywords

Milk, Machine Learning, Random Forest, K-Nearest Neighbors, Neural Networks

1. Introduction

The safety of milk is relevant for the mass production of dairy suitable for human consumption, due to the close relationship between raw milk and its derivatives such as cheese, yogurt, evaporated milk, among others. Milk has been recommended by organizations, such as UNESCO and FAO, as an indispensable food sustenance, especially for infants. Therefore, its production and distribution represents a significant part of food security strategies in various countries (Fuentes et al., 2013). To achieve the right standards, good hygiene practices must be implemented during production and storage. According to Pathot (2019), in his study on hygienic practices and bacteriological quality of milk, the establishment of standards, the use of effective application, the education of dairy staff and producers on various aspects of milk hygiene and handling technique are relevant. Milk is part of a remarkable base for the reproduction of microorganisms because of its high nutritional percentage; Because of this, the health of this must be constantly evaluated. Raw milk in its optimum condition must not have residues or sediments, be tasteless or have unusual color and odor; in addition, a minimum content of bacteria, the absence of chemical substances and normal acidity indices in its composition. Distinguishing the integrity of dairy will promote the implementation of standards, which ensure healthy products (Gonzales et al., 2019). The present research seeks to classify milk samples, according to the presence or absence of adulterants, which allow the obtaining of dairy products suitable for human consumption. The traditional method of identifying properties of this product is given through experimentation; however, classification methods and techniques linked to the use of machine learning show better results than traditional methods; In addition, they are very useful, at present, for society, since they are transforming industrial processes, which bring with them improvements in terms of efficient productive operations and agile decision making. In addition, it allows the collection of data, prediction of events based on these, creation of value-added models through the use of automation techniques, accelerate and specify the process of sample inspection.

1.1. Objectives

The purpose of the research is to classify milk samples with the support of machine learning techniques, which will discriminate the presence or absence of adulterants, which will allow the obtaining of dairy products suitable for human consumption. Also, accelerate and specify the inspection process of these samples. Research methods and data collection were established to evaluate the technical data collected and design, based on machine learning, an automated model, which speeds up the examination of the information collected and provides a more effective result on its optimal quality condition.

2. Literature Review

The application of automations in the food industry allows an effective management of the quality of products for human consumption. The implementation of systems for the detection of food fraud allows their in-depth examination in terms of integrity and content (Mengucci et. al., 2021). These use artificial intelligence, machine learning or deep learning algorithms (Kumar et. al., 2021). Systems are presented to estimate the characteristics subject to evaluation and define the precision when adapted to a given situation. In relation to the above, articles related to analysis of accuracy parameters as a result of automations were studied.

In instance, it is possible to consider, the case of the estimation of the age of certain foods, defined after the use of regression of support vectors and the regression of Gaussian processes by means of DoFP. In this situation, an accuracy of approximately 93% was obtained (Takruri et. al., 2020). On the other hand, the evaluation of the quality of milk powder is mentioned, where partial least squares regression, support vector machine and extreme learning machine were used, obtaining 90% specificity (Feng et.al., 2019). In addition, the study of food and beverages for contamination detection is presented. In it, data was collected using RFID and XGBoost was executed for the training of a model that provided an accuracy of 90%, which benefits consumer safety (Sharif et. al, 2021). In the dairy industry, the application of FTIR to obtain data on the composition of milk was studied to subsequently be subjected to techniques such as random forest and determine whether or not the product has been adulterated (Neto et. al., 2019). This case was 98% accurate. Likewise, in the case of other foods, where laser-induced decomposition spectroscopy identified adulterations in their composition. For its examination, it was evaluated by linear discriminant analysis and random forest, where in both cases precisions greater than 90% were obtained (Stefas et.al., 2021). On the other hand, the principal component analysis regression algorithm was also used, obtaining an approximate accuracy of 94% (Alaiz-Rodriguez & Parnell, 2020).

Likewise, these techniques have been used in the field of livestock. For example, for disease detection. Through the inspection of the milk applying infrared analysis, the data was collected to be examined through neural networks, achieving 94% accuracy (Denholm et. al., 2020). Also, the detection of subclinical ketosis in cows was studied through model design and data examination, where the best was the one related to a support vector classification, obtaining a specificity of 73%, which refers to a useful tool. for handling dairy products (Satoła & Bauer, 2021). Likewise, through predictions of neural networks and partial least squares, adequate feeding in dairy systems was evaluated in terms of dry matter (Dórea et. al., 2018). In this case, a determination of 70% was obtained, being the use of neural networks the best alternative. Additionally, the training of data obtained from milk production according to environmental conditions was investigated. In this situation, the random forest was changed and a prediction of more than 80% accuracy was obtained, which serves for the future trend of input production (Bovo et. al., 2021). As well as the implementation of models for the evaluation of milk content and feed consumption by cattle, in which highly accurate results were obtained for the maintenance of an ideal level of quality of the product subjected to high temperatures (Sources et al., 2020). As well as the monitoring of milk quality through infrared spectroscopy, according to the analysis of food sources, and the application of partial least squares as a method with greater precision for the creation of strategies that cooperate with the consumer (Frizzarin et. al., 2021). Also, to evaluate the health of the cattle, information was collected through traxial acceleration to identify factors on their feeding. To achieve this, the KNN, the support vector machine and neural networks were used, where precisions greater than 90% were obtained (Shen et. al., 2020). This benefits the health of the cattle, promoting a future product that is innocuous and suitable in quality.It should be noted the use of machine learning in precision agriculture, which refers to the revolution of technology in this field. Its knowledge-based application improves the productivity and quality of products intended for human consumption (Sharma et. al., 2021).

It is worth considering, on the other hand, the use of the electronic nose, a tool that imitates the sense of smell based on the use of an array of electrochemical sensors and a data identification system. For example, the detection of volatile organic compounds with the use of this tool and its classification by means of a support vector machine is mentioned, where 98% accuracy was obtained (Huang et. al., 2021). Likewise, its application is useful to predict food safety and detect contamination. Therefore, they were defined for the classification of the data, support vector machine and linear discriminant analysis, which presents a higher accuracy than other algorithms (Keerthana & Santhi, 2020). It is relevant to mention that the e-nose is a low-cost and non-invasive tool (Mu et. al., 2020). Likewise, the addition of a brix meter is proposed as a standardization instrument to classify the sweetness of pineapples through the use of KNN with an accuracy of up to 82% (Yan et. al., 2021).

3. Methods

In the present study, the classification of milk was carried out - using machine learning - to determine the presence of adulterants within the selected samples. To achieve this goal, classification methods such as random forest, neural networks and k-nearest neighbor, generally used by the food industry, will be used.

The technologies and statistical methods used to determine the quality parameters in milk samples will be detailed below.

3.1 Random Forest

The algorithms provided by random forest modeling are better suited for medium and large data sets, especially by having a better understanding of the field of estimators in greater quantity. Random forest is one of the systems most predisposed to collect better performance when learning about data (Schonlau & Yuyan Zou, 2020).

Random forest in turn allows working with estimators with a high degree of heterogeneity in a consistent way which aims to obtain a high degree of learning or prediction with data sources complicated to handle in order to obtain flexible and forceful results (Athey et al., 2019).

3.2 K-Nearest Neighbor

KNN is presented as a simple alternative method to use related to machine learning. To make use of this tool the database must be prepared in advance to be executed by KNN in R. After the KNN algorithm has predicted any outcome, the performance of the model should be reviewed. It is worth mentioning that factors such as the value of k, distance and choice of predictors have a crucial impact on model performance (Zhang, 2016). The entire mentioned process collaborates with tasks of creating predictive graphs, classification and regression diagrams (Kang, 2021).

The way in which samples are classified by k-nearest neighbor classification is quite simple, since different examples are segmented according to their function with respect to the nearest "neighbors". The ideal is to take more than one "neighbor" into account so its name (KNN) is derived. These classifications are within a given runtime all recorded in a runtime memory where all cases or situations with similar "neighbors" characteristics converge. As the classification is based on examples, it is also known as example-based classification or case-based classification (Cunningham & Delany, 2021).

3.3 Neural Networks

Neural networks are a system made up of a variety of processing units, where each of them performs a simple calculation by receiving an input vector and providing only one output. It can be classified according to its topology (hidden layers) or the commonly supervised learning algorithm. They represent a useful tool used for the development of learning for a machine in order to develop systems that generate information, recognize patterns and predict behaviors of a sample based on a database automatically (Sarmiento-Ramos, 2019). This tool has great potential in the field of research related to engineering applications. In addition, it should be noted that it has the ability to operate with large populations (Mera & Ochoa, 2021).

4. Data Collection

The data analyzed for the present research was obtained from the dataset of Neto et al. (2019) where commercial quality milk samples were collected. Some of them subjected to an adulterating agent for the purpose of being examined later. These have the following substances in their composition (analytical grade): sucrose, starch, sodium bicarbonate, hydrogen peroxide and formaldehyde. Even milk can present additional adulterants to those incorporated. Based on these data, it is sought to classify milk in terms of safety for human consumption; that is, determine whether this is crude or not. Also, demonstrate that Fourier spectrometry is an appropriate method for its evaluation. This was applied to the total of samples collected in order to obtain data on infrared spectra simple to model.

Modeling techniques were defined, based on the file obtained, to determine the accuracy of its classification. The techniques used were random forest, k-nearest neigbors and neural networks. In instance, columns containing information such as dates, days, and times were discarded. The ingredients were rated with numbers from 0 to 5. Also, the binary variable linked to the rawness of milk with values of 0 and 1. Input data were 2/3 of the sample records, randomly selected. After training the data by means of random forest, the amount was reduced to 531 estimators in order to remain with those whose presence would influence the final result.

5. Results and Discussion

5.1 Numerical Results

As a result of the execution of the aforementioned techniques, parameters linked to the accuracy of the proposed models were obtained considering the 969 records of milk with different adulterant ingredients. First, the precision parameter determines the accuracy of a model based on its relationship between predicted positives and actual positives. This indicates increased costs, due to false positives (Figure 1 and Figure 3).

Precision = True Positive / True Positive + False Positive

It is also worth noting the existence of the confusion matrix, which graphically represents the index of false positives and true negatives, which will later determine the accuracy of the model. This 2x2 matrix has two axes: x and y. The x-axis, linked to the antecedents and the y-axis to consequentials. Based on its location in the array, the following is interpreted in relation to the data:

- True – True: The milk is raw in the dataset and after the execution of the model.
- False – False: The milk is not raw in the dataset and after the execution of the model.
- True – False: In the dataset the milk is raw; however, the model defines it as non-raw (false positive).
- False – True: In the dataset milk is not raw; however, the model defines it as raw (true negative).

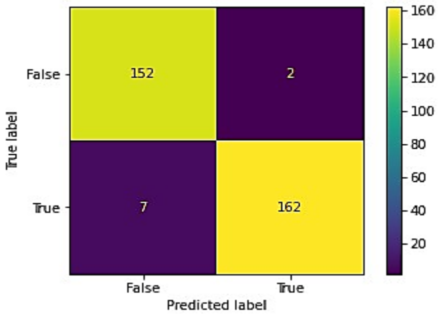

Figure 1. Confusion Matrix – Random Forest

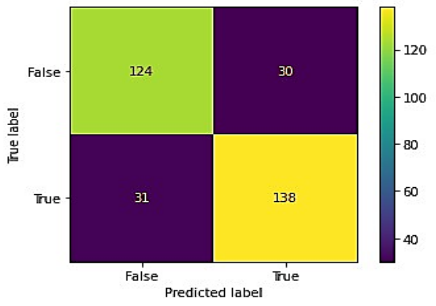

Figure 2. Confusion Matrix – KNN

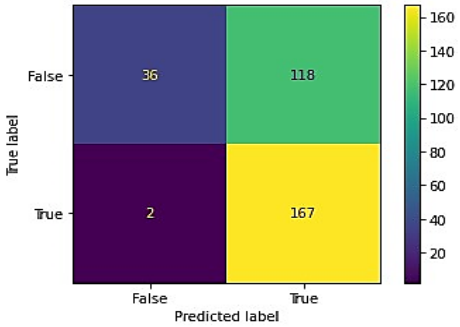

Figure 3. Confusion Matrix – Neural Networks

On the other hand, *recall* is responsible for calculating how many real positives the model captures and labels them as true positives. This parameter is used for the selection of the most suitable model to counteract the high costs linked to false positives. *Recall = True Positive / True Positive + False Negative*

Likewise, there is the *accuracy* parameter, which provides information on the number of times the model had to be executed to achieve the greatest number of true positives.

Finally, *F1*, which allows to calculate a balance between the precision and recall parameters, in order to link false positives and true negatives. This validates that the most efficient model is chosen.

*F1 = 2 * Precision * Recall / Precision + Recall*

Next, the numerical results obtained after applying the three techniques will be presented and the accuracy parameters will be determined.

Table 1. Accuracy Parameters - Random Forest

Technique/Parameter	Random Forest
Precision	0.98
Recall	0.94
Accuracy	0.96
F1	0.96

Table 2. Accuracy Parameters - KNN

Technique/Parameter	KNN
Precision	0.81
Recall	0.87
Accuracy	0.82
F1	0.84

Table 3. Accuracy Parameters - Neural Networks

Technique/Parameter	Neural Networks
Precision	0.82
Recall	0.2
Accuracy	0.54
F1	0.33

5.2 Graphical results

Next, comparative graphs will be shown, which allow you to appreciate the values of the previous point. This collaborates with the choice of the most appropriate technique for the case study (Figure 4-6).

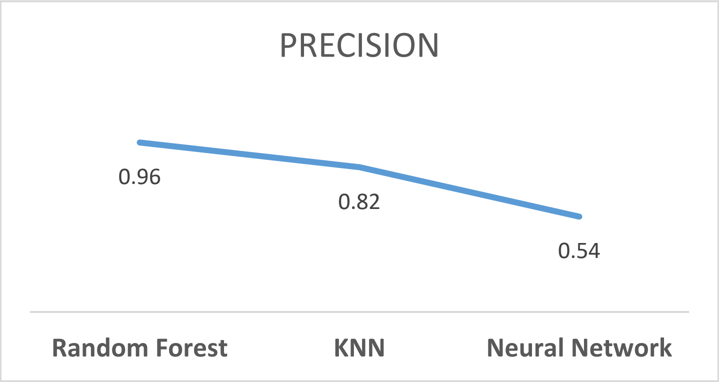

Figure 4. Comparison of accuracy percentages

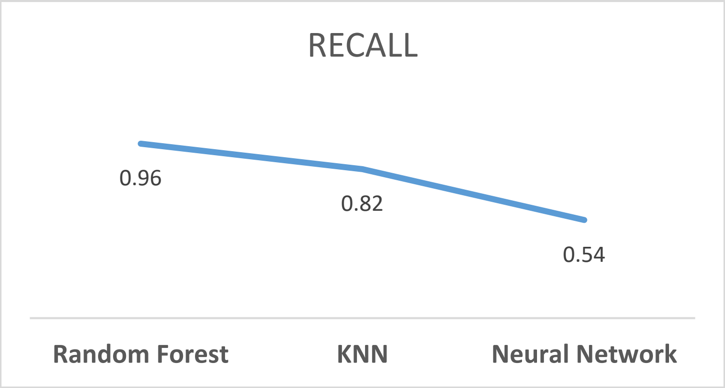

Figure 5. Recall Percentage Comparison

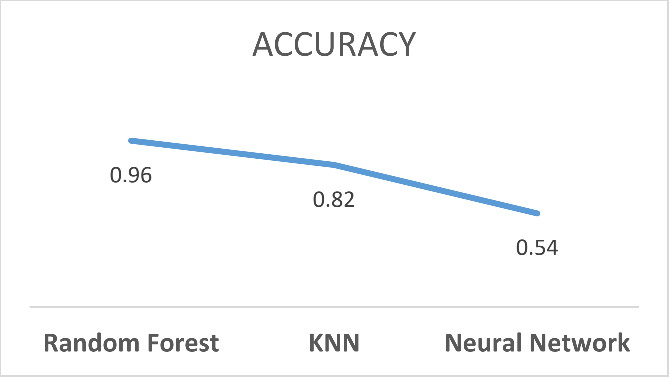

Figure 6. Comparison of accuracy percentages

In instance, it is claimed that the random forest technique provides a more accurate model, because its value is the highest in terms of accuracy and prediction of data. Likewise, it is visualized that with respect to KNN there is no wide variation in terms of what is mentioned; However, neural networks show reduced values and present different values in each execution of the model due to the random choice of data from the beginning.

5.3 Proposal Improvements

With the results collected, it was decided to implement the improvement proposal within the random forest technique. The method of reducing variables was used with the random tree sampling technique. For this, the boruta library from the dataset of Neto et al. (2019) was used. It went from a total of 531 estimators to 112, in order to have significant information for the final result of the model. This applied to a real scenario is responsible for minimizing costs and times during the analysis of samples.

5.4 Validation

After data training, random forest was applied to the 112 estimators, which had been selected and an accuracy of 0.99 was achieved. Although the improvement was 1%, it is shown that borura contributes to the accuracy of input data minimizing those anomalous data that enter the model.

In order to specify the selected techniques and the appropriate machine learning models for the analysis of milk samples, tests were carried out that compared the performance of the methods described in the methodology. Starting with the random trees they managed to obtain between 94% to 99% accuracy working with 2/3 of the randomly established data. For KNN working in the same way with the records, a level of accuracy between 77% and 81% was achieved. Finally, for neural networks it was not possible to ensure a certain level of accuracy, since the input data being random so that the values that the neural networks threw are very volatile. As much as all the methods proved to have acceptable performances, what inclines us to choose one of them is the nature of their way of operating, we are talking about the randomness trees that their branched structure effectively serves to classify the different classes of adulterants (ingredients) within the samples and after this selection we begin to train the dataset within this method to impact aspects such as time or correlativity within trees. It is worth mentioning that in the multiclass section we are talking about: peroxide, sucrose, formaldehyde, raw, bicarbonate, starch and within the binary class if it is raw milk or not raw milk.

6. Conclusion

In the present work, we submitted milk samples with adulterants and their respective spectral data in search of ways to predict and classify according to the interest of the research. A total of 969 milk registers were used with different adulterant ingredients which were used within classifying models such as KNN, random trees and neural networks all with the ability to recognize the elements of the composition within the milk.

After testing the models on different occasions in order to achieve a consistent and effective one, we achieved satisfactory results in both random trees and KNN. Due to its nature of procedure and structure, random forest allows a better classification of milk components accurately between 96% and 99% after training the data to discard non-significant data and make the development of the model more accurate, so randomness trees end up being the ideal way to operate to work with data related to laboratory samples and compositional analysis. Finally, we consider expanding this work with more development within the field of neural networks, as they are a preferable tool when predicting the behavior or nature that milk samples may present and thus make more profitable what is covered in this project.

References

Alaiz-Rodriguez, R., & Parnell, A. A Machine Learning Approach for Lamb Meat Quality Assessment Using FTIR Spectra. *IEEE Access*, 2020.

Athey, S., Tibshirani, J., & Wager, S. Generalized random forests. *Annals of Statistics,* 2019.

Bovo, M., Agrusti, M., Benni, S., Torreggiani, D., & Tassinari, P. Random forest modelling of milk yield of dairy cows under heat stress conditions. *Animals,* 2021.

Cunningham, P., & Delany, S. k-Nearest Neighbour Classifiers - A Tutorial. *ACM Computing Surveys,* 25, 2021.

Denholm, S. J., Brand, W., Mitchell, A. P., Wells, A. T., Krzyzelewski, T., Smith, S. L., . . . Coffey, M. P. Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning. *Journal of Dairy Science,* 2020.

Dorea, J., Rosa, G., Weld, K., & Armentano, L. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. *Journal of Dairy Science,* 2018.

Feng, L., Zhu, S., Chen, S., Bao, Y., & He, Y. Combining fourier transform mid-infrared spectroscopy with chemometric methods to detect adulterations in milk powder. *Sensors,* 2019.

Frizzarin, M., O'Callaghan, T., Murphy, T., Hennessy, D., & Casa, A. Application of machine-learning methods to milk mid-infrared spectra for discrimination of cow milk from pasture or total mixed ration diets. *Journal of Dairy Science,* 2021.

Fuentes, G., Ruiz, R., Sanchez, J., Avila, D., & Escutia, J. Microbiological analysis of milk of organic origin: desirable attributes for its transformation Agriculture. *Society and Development,* 14, 2013.

Fuentes, S., Viejo, C., Cullen, B., Tongson, E., Chauhan, S., & Dunshea, F. Artificial intelligence applied to a robotic dairy farm to model milk productivity and quality based on cow data and daily environmental parameters. *Sensors,* 2020.

Gonzales, J., Portocarrero, S., & Abanto, M. Quality of dairy products produced in the Amazonas Region, Peru. *UNTRM Scientific Journal: Natural Sciences and Engineering,* 6, 2019.

Huang, Y., Doh, I., & Bae, E. Design and validation of a portable machine learning-based electronic nose. *Sensors,* 2021.

Kang, S. k-Nearest Neighbor Learning with Graph Neural Networks. *Mathematics,* 12, 2021.

Keerthana, S., & Santhi, B. Survey on applications of electronic nose. *Journal of Computer Science,* 2020.

Kumar, I., Rawat, J., & Mohd, N. Opportunities of Artificial Intelligence and Machine Learning in the Food Industry. *Journal of Food Quality,* 2021.

Mengucci, C., Rabiti, D., Urbinati, E., Picone, G., Romano, R., Aiello, A., . . . Capozzi, F. Spotting frozen curd in PDO buffalo mozzarella cheese through insights on its supramolecular structure acquired by 1H TD-NMR relaxation experiments. *Applied Sciences,* 2021.

Mera, L., & Ochoa, J. Redes neuronales convolucionales para la clasificación de componentes independientes de rs-fMRI. *Atención Primaria ,* 19, 2021.

Mu, F., Gu, Y., Zhang, J., & Zhang, L. Milk source identification and milk quality estimation using an electronic nose and machine learning techniques. *Sensors,* 2020.

Neto, H., Tavares, W., Ribeiro, D., Alves, R., Fonseca, L., & Campos, S. On the utilization of deep and ensemble learning to detect milk adulteration. *BioData Mining,* 2019.

Pathot, Y. D. Prácticas Higiénicas y Calidad Bacteriológica de Leche: Una revisión . *International Journal of Research-Granthaalayah ,* 16, 2019.

Satoła, A., & Bauer, E. Predicting subclinical ketosis in dairy cows using machine learning techniques. *Animals,* 2021.

Sarmiento-Ramos, J. L. Applications of neural networks and deep. *UIS Ingenierias,* 18, 2019.

Schonlau, M., & Yuyan Zou, R. The random forest algorithm for statistical. *Stata Journal,* 2020.

Sharif, A., Abbasi, Q., Arshad, K., Ansari, S., Ali, M., Kaur, J., . . . Imran, M. Machine learning enabled food contamination detection using rfid and internet of things system. *Journal of Sensor and Actuator Networks,* 2021.

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access,* 2021.

Shen, W., Cheng, F., Zhang, Y., Wei, X., Fu, Q., & Zhang, Y. Automatic recognition of ingestive-related behaviors of dairy cows based on triaxial acceleration. *Information Processing in Agriculture,* 2020.

Stefas, D., Gyftokostas, N., Kourelias, P., Nanou, E., Kokkinos, V., Bouras, C., & Couris, S. A laser-based method for the detection of honey adulteration. *Applied Sciences,* 2021.

Takruri, M., Abubakar, A., Alnaqbi, N., Shehhi, H., Jallad, A., & Bermak, A. DoFP-ML: A Machine Learning Approach to Food Quality Monitoring Using a DoFP Polarization Image Sensor. *IEEE Access,* 2020.

Yan, J., Jin, M., Xu, Z., Chen, L., Zhu, Z., & Zhang, H. Recognition of Suspension Liquid Based on Speckle Patterns Using Deep Learning. *IEEE Photonics Journal,* 2021.

Zhang, Z. Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine,* 7,2016.

Acknowledgements

This paper has received the support from the Universidad de Lima - Scientific Research Institute (https://www.ulima.edu.pe/en/scientific-research).

Biographies

Mia León is a final year industrial engineering student at the Universidad de Lima. She is an analyst in the area of information technology in a recognized company in the banking sector at the national level. Interested in predictive analysis and operational research applied to the management of both industrial and business processes.

Diego Ossa is an industrial engineering student at the Universidad de Lima who is currently performing logistics functions in the private sector. With a marked interest in supply chain and predictive models for industrial processes.

José Antonio Taquía is a Doctoral Researcher from Universidad Nacional Mayor de San Marcos and holds a Master of Science degree in Industrial Engineering from University of Lima. He is a member of the School of Engineering and Architecture teaching courses on quantitative methods, predictive analytics, and research methodology. He has a vast experience on applied technology related to machine learning and industry 4.0 disrupting applications. In the private sector he was part of several implementations of technical projects including roles as an expert user and in the leading deployment side. He worked as a senior corporate demand planner with emphasis on the statistical field for a multinational Peruvian company in the beauty and personal care industry with operations in Europe and Latin America. Mr. Taquía has a strong background in supply chain analytics and operations modeling applied at different sectors of the industry. He is also a member of the Scientific Research Institute at the Universidad de Lima being part of the industrial and circular economy groups. His main research interests are on statistical learning, predictive analytics, and industry 4.0.