

**GENERACIÓN DE DATOS SINTETICOS PARA EVALUAR LA ENFERMEDAD
INFECCIOSA LEUCOSIS BOVINA****GENERATION OF SYNTHETIC DATA TO EVALUATE THE BOVINE
LEUKOSIS INFECTIOUS DISEASE**

 MSc. Javier Antonio Ballesteros-Ricaurte*,  MSc. Juan Sebastián González-Sanabria*,  PhD. Hugo Ordóñez**

* **Universidad Pedagógica y Tecnológica de Colombia**, Facultad de Ingeniería, Grupo de Investigación en Manejo de Información - GIMI.
Av Central del Norte 39-115, Tunja, Boyacá, Colombia.
Tel.: 57 (608) 7405626.

E-mail: {javier.ballesteros, juansebastian.gonzalez}@uptc.edu.co

** **Universidad del Cauca**, Facultad de Ingeniería, Departamento de Sistemas.
Calle 5 N. 4-70, Popayán, Cauca, Colombia.
Tel.: +57 (602) 8209900.
E-mail: hugoordonez@unicauca.edu.co

Cómo citar: Ballesteros-Ricaurte, J. A., González-Sanabria, J. S., & Ordóñez, H. (2023). GENERACIÓN DE DATOS SINTETICOS PARA EVALUAR LA ENFERMEDAD INFECCIOSA LEUCOSIS BOVINA. REVISTA COLOMBIANA DE TECNOLOGIAS DE AVANZADA (RCTA), 1(41), 115–122. <https://doi.org/10.24054/rcta.v1i41.2556>

Derechos de autor 2023 Revista Colombiana de Tecnologías de Avanzada (RCTA).
Esta obra está bajo una licencia internacional [Creative Commons Atribución-NoComercial 4.0](https://creativecommons.org/licenses/by-nc/4.0/).



Resumen: Los proyectos que se desarrollan en el sector de salud animal se enfrentan a limitaciones tecnológicas y científicas debido a la falta de información consistente y confiable, y a los altos costos que supone la recolección de información para los ganaderos. Así mismo, las limitaciones jurídicas en la divulgación de información por razones como las leyes de protección de datos conllevan atrasos en el desarrollo de políticas y estrategias, lo mismo que en la toma de decisiones. Ante esta falta de disponibilidad de información surge como solución la generación de datos sintéticos a partir de un conjunto de datos originales. Así, en este artículo se presenta un estudio a través del cual se valoraron tres métodos para generar datos sintéticos que reflejan el comportamiento de una enfermedad bovina en un conjunto de datos reales. El trabajo se basó en comparar algoritmos de aprendizaje automático, herramientas y métodos basados en modelos para mejorar el realismo de los datos sintéticos referidos al comportamiento de la enfermedad. El objetivo trazado fue encontrar el mejor modelo para la generación de datos sintéticos utilizando el caso de la enfermedad infecciosa mastitis bovina, ya que no se cuenta con suficientes datos sobre ella. Para validar los datos sintéticos fue necesario contrastar el conjunto de datos original y la información sintética, en busca de que el método seleccionado generara datos sintéticos con cualidades similares a las del conjunto de datos original.

Palabras clave: Aprendizaje automático, Enfermedades infecciosas bovinas, Datos sintéticos, Leucosis.

Abstract: The projects that are conducted in the animal health sector face technological and scientific limitations due both to the lack of consistent and reliable information, and to the high costs of collecting information for farmers. Likewise, legal limitations on the disclosure of information for reasons such as data protection laws lead to delays in the development of policies and strategies, as well as in decision-making. Given this lack of information availability, the generation of synthetic data from a set of original data emerges as a solution. Thus, this paper presents a study through which three methods to generate synthetic data that reflect the behavior of a bovine disease in a set of real data were evaluated. The work was based on comparing machine learning algorithms, tools, and model-based methods to improve the realism of synthetic data of disease behavior. The goal was to find the best model for the generation of synthetic data using the case of the bovine mastitis infectious disease, since there is not enough data for it. In order to validate the synthetic data, it was necessary to contrast the original data set and the synthetic information, looking for the selected method to generate synthetic data with qualities similar to those of the original data set.

Keywords: Machine learning, Bovine infectious diseases, Synthetic data, Leucosis.

1. INTRODUCTION

Bovine infectious diseases (that occur when a bovine is infected by a pathogen from another bovine or another animal) cause economic losses to livestock farmers and companies in the dairy industry; they are also considered a public health issue (Ballesteros-Ricaurte et al., 2021a).

One of the most important aspects of producing meat or dairy cattle is the herd's reproductive performance. Thus, any pathology interfering with this needs to be addressed (Andrade Becerra et al., 2014a; Pulido-Medellín et al., 2017b). While the factors affecting reproduction have diverse origins, infectious etiologies are the most common. Infectious reproductive diseases (bacterial, viral and parasitic) in farms affect the females' reproductive performance and productivity (Pulido-Medellín et al., 2017a).

The epidemiology of non-infectious diseases is a research field focused mainly on the risk factors associated with the possibility of developing the disease. On the other hand, in the case of infectious diseases, the primary risk factor for contracting them is the presence of infectious cases in the local population (Andrade Becerra et al., 2014b).

Infectious diseases are subject to official control by the governmental entities in each country, and those which are not are a way of classifying. These diseases must be controlled in the regions where meat and milk are produced. Entities must implement prevention campaigns, and if they find a source of contagion, they have to report it and take the necessary measures to avoid outbreaks.

Moreover, it is important to know the behavior of these diseases, the control measures that should be implemented when an outbreak is identified, the social and economic problems that may arise as a consequence, and the implications of an outbreak on a farm.

The bovine diseases that are not subject to official control also negatively impact the farms' reproductive and productive indexes. They entail significant economic consequences because they affect the progress of the livestock industry (Pulido-Medellín et al., 2017a). Even though they can be detected and treated properly, the lack of timely information does not allow to make timely decisions regarding their presence. Thus, the producer has to prevent and control this type of disease through differentiated schemes.

Real data collection is a technique that can have disadvantages, such as high labor and equipment costs and time (González Martínez, 2021; Tan et al., 2019). In addition, having access to original data does not mean they can be freely used. There are situations where the administrators or owners of the farms and companies do not want (or cannot) provide data on bovine diseases. Sometimes, this happens because these people do not have tools or information systems that allow them to store, organize, and control the information generated in the administration or management of the herds. Therefore, not having real information becomes a constraint (inconvenience) in knowing about the potential diseases that bovines can have and using that information as a basis for planning prevention actions. The generation of synthetic and

anonymized data emerges at this point as an excellent opportunity to solve this situation (Shah et al., 2020).

Synthetic data derive from the reproduction of artificial information based on historical data or reliable information sources. Mathematics, statistics, computing sciences, and artificial intelligence (AI) algorithms work as a whole to create significant synthetic information supported by simulations, which can be analyzed and compared with real data. This process is known as AI axioms. They are knowledge derived from data to address a specific task through machine learning (González Martínez, 2021).

Synthetic data generation can help solve several problems due to its easy and agile creation, which avoids collecting real data (Olmedo Vélez & Narváez Tello, 2021). There are synthetic data generation tools that can help preserve the data's privacy, test systems or serve as training data for machine learning algorithms. Additionally, synthetic or artificial data generation is a robust alternative and technique to mask sensitive data and avoid risks for third parties. The data are randomly generated with restrictions to hide confidential or private information. However, it is possible to withhold certain statistical information regarding the relations between attributes and patterns in the original data (González Martínez, 2021; Surendra & Mohan, 2017a).

According to (Surendra & Mohan, 2017b), synthetic data are classified into three categories:

- Fully synthetic: The information is entirely artificial.
- Partially synthetic: Only the values of the selected attribute or characteristic are replaced with synthetic values without disclosing sensitive and risky information.
- Hybrid synthetic: They have information from real and synthetic data since they were generated from original and synthetic data. They have the advantages of fully and partially synthetic data.

In the following sections, we will compare three techniques to generate synthetic data if there is a case of bovine mastitis (a condition for which there is limited data of this type). The comparison is based on the results of the research conducted by the Grupo de Investigación en Medicina Veterinaria y Zootecnia (Gidimevetz) from Universidad Pedagógica y Tecnológica de Colombia - UPTC.

2. METHODS AND MATERIALS

Data are necessary for the development of projects in diverse areas. Thus, they are considered an asset within organizations. However, as we previously mentioned, their use has certain restrictions since they include sensitive information (Lopez-Martin et al., 2018a). With the aim of solving part of these issues, synthetic data generators are used to develop and implement probabilistic models to generate synthetic or artificial information and consolidate real (but anonymous) information that can be disseminated, in addition to being consistent, dynamic, and scalable (Lopez-Martin et al., 2018b).

It is possible to benefit from synthetic data generation when the dataset and the methods (according to the use cases) are suitable. Data generation has gained relevance in recent years and will contribute largely to technology development. For example, Shah et al. (Shah et al., 2020) state that with the correct data, machine learning models can be created for domains such as autonomous vehicles, health, social with a focus on crime processing (Ordóñez et al., 2020), finance, and demographics (Ballesteros-Ricaurte et al., 2021b), among others. The usage of synthetic data can increase the accuracy of the machine learning models, which would contribute to the advance mentioned above. Table 1 describes several models for the generation of this type of data.

Table 1: Models for synthetic data generation

| Models | Description | Algorithm | Application |
|--|---|-----------------------------|---|
| Generative models | They aim to find the correlation between the variables of a dataset whose behavior cannot be defined by a standard mathematical model (Yale et al., 2020a). | Adversarial neural networks | Several applications on science and technology. Prediction of genetic models, generation of unique images or amplification of patterns in images (Beattie, 2020). |
| Generative adversarial networks (GANs) | Deep neural network architectures composed of two networks: generator and discriminator (adversarial) (Goncalves et al., 2020)- (Goodfellow et al., 2020). | Deep neural networks | Diverse areas |

| | | | |
|---------------------------------|--|----------------|---|
| Variational autoencoders (VAEs) | They can produce synthetic data that follow the same patterns as the large datasets they are fed (Lopez-Martin et al., 2018a). | Neural network | Image generation, characteristic extraction and reinforcement learning. |
| SMOTE | It is an oversampling approach to deal with imbalanced datasets. It is an algorithm to address the class imbalance problem (Sposito et al., 2020). | Own algorithm | Diverse areas where there are imbalanced data. |

The literature review and the data available on bovine mastitis led to the use of the GAN in this comparative study. The documentation indicates that this model can be used with any type of data, does not require a minimum amount of data, and, according to several authors, it performs better than other methods (Andrade Becerra et al., 2014a; Lopez-Martin et al., 2018b; Yale et al., 2020b).

On the other hand, there are several tools that use techniques to generate synthetic data. They have advantages and disadvantages.

- Advantages:
 - Implementation of an algorithm.
 - It is not necessary to program.
 - The user only has to upload the original data file.
 - The size of the dataset is not restricted.
- Disadvantages:
 - The license for use makes the data available for other users in open access, compromising the privacy of the information.
 - There is not enough information on these tools.

Some platforms allow using the algorithms through programming languages (particularly Python). We highlight the following:

- Gretel AI: Its objective is to help developers share and collaborate safely with confidential data in real-time (Gretel Labs, 2020). It uses a sequence-to-sequence architecture to train a text dataset and learn to predict the following characters of the sequence. This platform allows the creation of synthetic data from any

type of text data, structured or not. It is recommended to use a simple data format. Gretel AI works with formats such as CSV (comma-separated values), Pandas DataFrames or line-delimited unstructured text.

The size of the dataset recommended for this tool must be over 5,000 records. If the dataset is highly dimensional (for example, with more than 15 columns), Gretel Labs (developer of the platform) recommends having more than 15,000 records (Gretel Labs, 2020). The tool allows to create artificial and secure datasets, preserving the same information as the original dataset but with greater warranties regarding the personal or secret data protection in the original data.

- Mostly AI: This platform (free for the community) analyzes the data to extract their structure and types. It benefits from deep generative neural networks with incorporated privacy mechanisms to learn the patterns, the structure, and the variation specific for the dataset to be analyzed automatically (MOSTLY AI Inc., n.d.). The data are coded in a specific format, necessary for the real training. The data engine determines automatically when the training is complete and stops the process before the model starts to overfit. The model represents the parameters that have obtained the best training results and is used to generate the synthetic data that can be exported later on.

Mostly works with any type of structured data that can be presented in a tabular form. It does not accept them if they are unstructured (for example, fixed images, audio or video). The data must be a CSV file. Usually, the larger the dataset, the better the quality of the synthesized data. Mostly has a minimum of 100 rows. The maximum number of columns and rows will be determined by the type of licence the user acquired.

Gretel AI and Mostly AI can use the data without any type of issue. Each tool uses its own algorithm and generates the synthetic data required. Both offer the possibility of evaluation when comparing the original data with the synthetic data, but the behavior of their algorithms, the processes, and the way they relate the data are unknown. Moreover, the documentation on the tools is not enough to know if they consider the GAN process.

The following section will present the results of the comparison previously mentioned, the GAN's results with Gretel's and Mostly AI's results to find

the best model for generating synthetic data in the case of bovine mastitis.

3. RESULTS AND DISCUSION

To validate the quality of the synthetic data, we used metrics (Goncalves et al., 2020) that measure the grade in which the properties of the real data are captured and transferred to the synthetic dataset (Yale et al., 2020a). Next, the process for generating such data is described.

3.1 Infectious bovine diseases dataset

The dataset used as a basis for the synthetic data generation corresponds to the information collected from 863 cows located on farms in the municipality of Toca (Boyacá Department, Colombia) (Andrade Becerra et al., 2014b). An exploratory analysis (Raschka & Mirjalili, 2019) was carried out to determine the presence of atypical data, the distribution of the data, and the relations between characteristics. Figure 1 presents a summary of the relations present in the dataset.

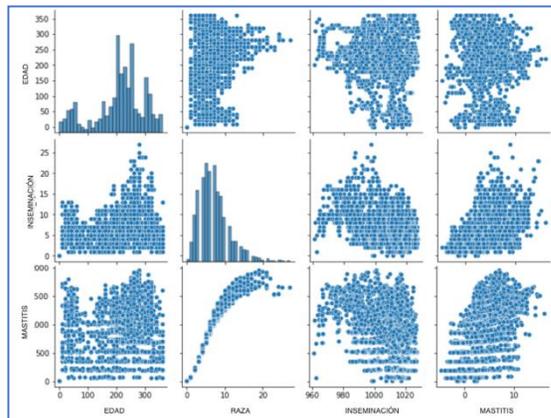


Fig. 1. Dispersion matrix

3.2 Implementation

In this process, the GAN was implemented in Python using the libraries (Keras, NumPy and Pandas) and other methods such as (generator and discriminator). Additionally, the tools mentioned above were configured to compare the results obtained in each case.

3.2.1 Application of the GAN

The GAN was configured in the Anaconda environment with its corresponding libraries to execute the algorithm. Then, the dataset was uploaded. To generate the synthetic dataset, the

training parameters were configured. The training dataset corresponded to 80% of the total data, and there were 30 epochs (the number of times the algorithm sees the complete dataset). Then, the model was trained. After the training process, the data necessary for the exercise were generated, in this case, 10,000. Once the process was completed, these data were exported to a Microsoft Excel® format file.

3.2.2 Gretel AI

In this case, the instructions and recommendations established in the tool's documentation were followed. The local environment and the Microsoft Excel® file with the data to be uploaded (in CSV format) were configured, and the Gretel API key was generated. This key allowed free access to the public version of the tool and the automated validation of synthetic data records and reports for the data quality.

The default configuration of the tool trained the model with 70% of the data. As a result, we obtained the synthetic data file (downloadable in CSV format) and a report describing the algorithm's behavior. In this case, the data generation was moderated with 56% of quality, 100% of correlation of the variables used, and 51% of stability in the usage of the algorithm, as shown in Figure 2.

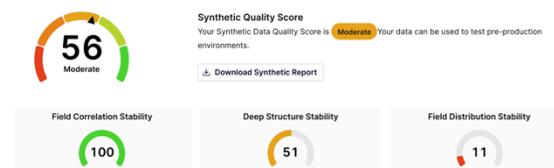


Fig. 2. Results with Gretel AI

3.2.3 Mostly AI

The first step in creating a synthetic dataset with this tool was eliminating the sensitive column, which was done by Mostly with its default configuration. A fairness constraint was added to the model's parameter optimization during training to generate synthetic data. In the dataset, the violation of statistical parity was penalized within each mini-lot by increasing the training loss by a number proportional to the difference between the fraction of each segment. Adding the fairness constraint extends the goal of the software from generating accurate and private synthetic data to generating accurate, private, and fair synthetic data.

After feeding and training the dataset with the additional fairness constraint, a synthetic version of

the dataset was generated. Once the distribution of the input was evaluated, the tool carried out an analysis of the amount of original data and decided how many times it had to execute the algorithm to determine the synthetic dataset.

A user was created to use the tool. Each test was a new project. The only step executed with the user role was uploading the file with the original data in CSV format and indicating the number of desired data. The results in Figure 3 were extracted from the downloadable report created by Mostly after the analysis.

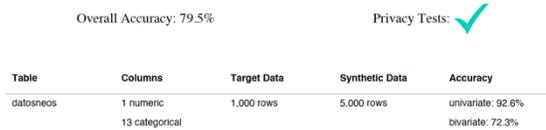


Fig. 3. Quality assurance report by Mostly AI

After executing the six steps (Figure 2), the synthetic data generation with the tool had a general accuracy of 79.5%. In this case, we started with 1000 data and 5000 were generated with an accuracy of 92.6% in the class object *Result*.

3.2.4 Comparison

To evaluate the quality of the synthetic data generators, there are metrics that measure the grade in which the properties of the real data are captured and transferred to the synthetic dataset (Goncalves et al., 2020). Additionally, other comparison criteria are considered, and they can be observed in Table 2. Those were considered in this research since the literature and the experiments completed indicate that they were relevant in determining the variables, the dataset distribution, and the amount of data that can be generated with each tool.

Table 2: Comparison criteria

| Criteria | Tool | | |
|-----------------|---------------------------------|---|--|
| | Gretel AI | Mostly AI | GAN |
| Type of license | Free version with restrictions. | Free version with restrictions, and the proprietary version that must be purchased. | Free |
| Algorithm | Not specified | Deep learning | Networks |
| Documentation | Limited | Limited | It is available in different books and articles. |

| Configurat ion | Not all the configurati on of the tool can be accessed. | The configuration of the tool cannot be accessed. | The configuratio n can be modified as often as needed when programm g the algorithm. |
|-------------------|--|--|---|
|-------------------|--|--|---|

The cross-classification metric measures the quality with which a synthetic dataset captures the statistical dependence structures that exist in the real data and determines the dependence through predictions generated for one variable based on the other variables (using a classifier). This research considered two metrics: F1 Score (Raschka & Mirjalili, 2019), which combines accuracy and recall, and CrCI-RS (Goncalves et al., 2020), which entails training with the real data and testing the data retained from the real and synthetic datasets. Table 3 presents the results of the F1 and CrCI-RS metrics applied to the three models. The GAN obtained the best results.

Table 3: Metrics results

| Metrics | Tools | | |
|---------|-----------|-----------|-------|
| | Gretel AI | Mostly AI | GAN |
| F1 | 55% | 81% | 88% |
| CrCI-RS | 62% | 79.5% | 84.5% |

CrCI-RS is especially useful for assessing if the statistical properties of the real data are similar to those of the synthetic. This metric can be used in algorithms such as decision trees, logistic regression, and neural networks (Goncalves et al., 2020). The real data available are divided into testing and training sets. A classifier is trained in the second (real), and applied to the first (maintain real) and the synthetic data. The classification accuracy metric is calculated in both sets. The performance of the classification depends on the chosen classifier. In general, a value close to 1 is ideal for the cross-classification metric.

Figure 4 presents the results in percentage after applying the metric to the three systems used in the synthetic data generation. The accuracy metric, that is, the relation between the number of correct predictions and the total number of predictions, is the basis for implementing the CrCI-RS cross-classification metric.

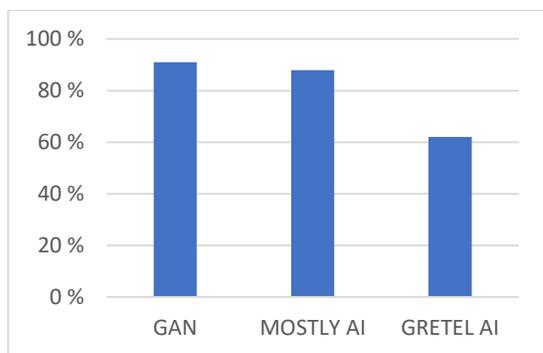


Fig. 4. Results of the CrCl-RS metric

The exercise allows establishing that the GAN is the best alternative for the synthetic data generation of bovine mastitis, given that the established criteria and the experiments' results show that it has advantages over the other two tested tools. One of them is that since the GAN requires programming, it is possible to adjust the configuration with the aim of obtaining better results in the data generation.

4. CONCLUSIONS

The same process was followed with each of the assessed methods in this study. Given a set of real samples, the model was adjusted, and then synthetic samples were generated from the learnt model. After learning with the first ones, the model was expected to be able to extract relevant statistical properties from the data. Additionally, the experimental analysis showed that there is not a unique method that outperforms the others in all the metrics considered. However, some methods might be useful in practice since they provide synthetic samples with statistical properties from the synthetic data equivalent to those of the real data. The GAN was particularly outstanding among the analyzed methods.

The synthetic data generation can find patterns among the different characteristics that usually would not be visible with common analysis methods.

The machine learning algorithms depend on large amounts of training data. Moreover, the cost of the data acquisition is high, and its availability is very limited in the case of bovine infectious diseases. Thus, synthetic data generation is relevant as an alternative for projects focused on this type of condition.

The development of this study opens the possibility of the future assessment of other machine learning

algorithms to validate their behavior in synthetic data generation. Moreover, the set of synthetic data generated for the purposes of this work should be validated in order to check whether the latter alone are enough to predict the presence of bovine mastitis.

REFERENCIAS

- Andrade Becerra, R., Caro Carvajal, Z., Pulido Medellín, M., Porras Vargas, J., & Vargas Abella, J. (2014a). Prevalencia de bacterias causantes de mastitis en fincas lecheras de Toca (Boyacá, Colombia). *Ciencia y Agricultura, 11*, 47–53.
- Andrade Becerra, R., Caro Carvajal, Z., Pulido Medellín, M., Porras Vargas, J., & Vargas Abella, J. (2014b). Prevalencia de bacterias causantes de mastitis en fincas lecheras de Toca (Boyacá, Colombia). *Ciencia y Agricultura, 11*, 47–53.
- Ballesteros-Ricaurte, J.-A., Avendaño-Fernández, E., González-Amarillo, A.-M., & Granados-Comba, A. (2021a). Mapeo científico en la búsqueda de información. Caso de estudio: enfermedades infecciosas en bovinos. *Revista Científica, 42*(3), 265–275. <https://doi.org/10.14483/23448350.17532>
- Ballesteros-Ricaurte, J.-A., Avendaño-Fernández, E., González-Amarillo, A.-M., & Granados-Comba, A. (2021b). Mapeo científico en la búsqueda de información. Caso de estudio: enfermedades infecciosas en bovinos. *Revista Científica, 42*(3), 265–275. <https://doi.org/10.14483/23448350.17532>
- Beattie, D. (2020). *The art of code*.
- Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology, 20*(1), 1–40. <https://doi.org/10.1186/s12874-020-00977-1>
- González Martínez, E. F. (2021). *Generador de datos sintéticos para el monitoreo de transacciones con factores de riesgo de lavado de activos. (Tesis de Maestría)*. (Universida).
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144. <https://doi.org/10.1145/3422622>
- Gretel Labs. (2020). *Gretel IA*. <https://gretel.ai>
- Lopez-Martin, M., Carro, B., & Sanchez-Esguevillas, A. (2018a). Variational data

- generative model for intrusion detection. *Knowledge and Information Systems*, 60(1), 569–590. <https://doi.org/10.1007/s10115-018-1306-7>
- Lopez-Martin, M., Carro, B., & Sanchez-Esguevillas, A. (2018b). Variational data generative model for intrusion detection. *Knowledge and Information Systems*, 60(1), 569–590. <https://doi.org/10.1007/s10115-018-1306-7>
- MOSTLY AI Inc. (n.d.). *Mostly*. 2020. Retrieved October 20, 2020, from <https://mostly.ai>
- Olmedo Vélez, V., & Narváez Tello, C. (2021). *Generación de un conjunto de datos sintéticos mediante técnicas de aprendizaje automático para análisis de fraude (Trabajo de grado)* (E. P. Nacional, Ed.; Escuela Po).
- Ordóñez, H., Cobos, C., & Bucheli, V. (2020). Machine learning model for predicting theft trends in Colombia. *RISTI - Revista Iberica de Sistemas e Tecnologías de Informacao*, 2020(E29), 494–506.
- Pulido-Medellín, M., González-Ariza, W., Bayona-Ríos, H., & Chavarro-Tulcán, G. (2017a). Determinación de Leucosis enzootica bovina mediante las claves Hematológicas de Göttigen y Elisa en Boyacá. *Rev. Fac.Cs. Vets.*, 58(1), 10–16.
- Pulido-Medellín, M., González-Ariza, W., Bayona-Ríos, H., & Chavarro-Tulcán, G. (2017b). Determinación De Leucosis enzoótica Bovina mediante Las cLaves HematoLógicas De göttingen y eLisa en Boyacá, coLomBia Enzootic Bovine Leukosis Assessment by Hematology Göttingen Keys and ELISA in Boyacá, Colombia. *Rev. Fac. Cs. Vets.*, 58(1), 10–16.
- Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (Segunda Ed). Marcombo.
- Shah, S., Gandhi, D., & Kothari, J. (2020). Machine learning based Synthetic Data Generation using Iterative Regression Analysis. In *Fourth International Conference on Electronics, Communication and Aerospace Technology* (pp. 1093–1100). <https://doi.org/10.1109/ICECA49313.2020.9297491>
- Spositto, O., Blanco, G., Matteo, L., & Levi, M. (2020). SMOTE , Algoritmo para balanceo de clases en un estudio aplicado a la ganadería . *XXVI Congreso Argentino de Ciencias de La Computación - CACIC*, 289–298.
- Surendra, H., & Mohan, H. S. (2017a). A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing. *International Journal of Scientific & Technology Research*, 6(3), 95–101.
- Surendra, H., & Mohan, H. S. (2017b). A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing. *International Journal of Scientific & Technology Research*, 6(3), 95–101.
- Tan, C., Behjati, R., & Arisholm, E. (2019). A model-based approach to generate dynamic synthetic test data: A conceptual model. In *IEEE 12th International Conference on Software Testing, Verification and Validation Workshops, ICSTW 2019* (pp. 11–14). IEEE. <https://doi.org/10.1109/ICSTW.2019.00026>
- Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020a). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416, 244–255. <https://doi.org/10.1016/j.neucom.2019.12.136>
- Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020b). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416, 244–255. <https://doi.org/10.1016/j.neucom.2019.12.136>