

Unveiling Insights from Hematobiometry Data: A Data Science Approach Using Data from a Quito Clinical Laboratory

Miguel Ortiz¹, Paúl Campaña¹, Jhonny Pincay^{1,2} ^a and Dora Rosero³

¹*Pontificia Universidad Católica del Ecuador, Avenida 12 de Octubre 1076 y Roca, Quito, Ecuador*

²*Lucerne University of Applied Sciences and Arts, Werfstrasse 4, Lucerne, Switzerland*

³*Investigadora Independiente, Laboratorio Clínico Pura Vida, Quito, Ecuador*
{mortiz871, pacampanag, jvpincay}@puce.edu.ec, drosero@puravidalab.com

Keywords: Hematobiometry, Anemia Diagnosis, Decision Tree Learning, Data Science, Data-Driven Healthcare, Clinical Laboratory Data Analysis.

Abstract: In this applied research study, a data science approach is employed to analyze anonymized hematological data obtained from a clinical laboratory located in Quito, Ecuador. The analysis aims to examine machine learning models that could potentially be used to aid in early anemia and polycythemia detection, ultimately contributing to improved healthcare decision-making. A rigorous MLOps-driven methodology is employed, and well-established techniques such as clustering, decision trees, and neural networks are applied. These methods are evaluated to identify the most suitable approach for the specific characteristics of the data. The findings showed that clustering methods were not advisable for the type of data used for the exploration and no significative results could be obtained. However, decision trees and neural networks demonstrated superior performance in predicting the presence of these blood disorders. Additionally, the outcomes of this research have the potential to be particularly significant for Ecuador, a nation facing challenges in healthcare access and malnutrition, where early anemia detection could be highly impactful.

1 INTRODUCTION


Modern medicine stands at the forefront of a rapidly accelerating transformation, driven by the development and deployment of technology-based tools, most notably artificial intelligence (AI). One area where this impact is most pronounced is in outpatient care, specifically in the early diagnosis of diseases and in the interpretation and analysis of laboratory tests (Deo, 2015; Ahsan et al., 2022). Besides, information technology has enabled healthcare professionals to gain a more detailed and accurate understanding of patients' health, facilitating the diagnosis and monitoring of various medical conditions.

In particular, laboratory tests, such as complete blood counts, have proven essential in understanding the internal functioning of the human body and detecting disorders (Alsheref and Gomaa, 2019; Gunčar et al., 2018). However, this information can quickly grow in volume and complexity. On the other hand, technological advancements have led to increasingly sophisticated analytical solutions capable of process-

ing and analyzing large volumes of data efficiently and systematically. Furthermore, current tools and techniques enable the storage and processing of these data, providing invaluable support to medical professionals in decision-making. This progress not only optimizes diagnosis but also enhances the monitoring and treatment of patients.

This initiative aims to investigate the impact of data analysis and machine learning tools applied to the results of complete blood count tests. A primary goal is to support medical staff in diagnostic decision-making by identifying trends and applying appropriate analytical methodologies. However, the analysis of data from laboratory tests presents several challenges. The complexity of blood data, which includes a wide variety of parameters, can complicate classification and interpretation. Additionally, individual patient variability, influenced by factors such as age and gender, adds another layer of complexity to the analysis. These factors highlight the importance of applying data analytics techniques capable of handling this diversity and improving diagnostic accuracy.

In this research, the operational databases of a clinical laboratory located in Quito, Ecuador, were

^a  <https://orcid.org/0000-0003-2045-8820>

analyzed. Data science techniques and MLOps methodology were applied with two objectives: to implement data analysis models that are better suited to the dataset and to provide intelligent information in diagnosing diseases that can be identified through blood tests, such as anemia and polycythemia.

The article is structured as follows: Section 2 introduces the concepts and related works that support this initiative. Next, Section 3 describes the methods used to design a solution. Section 4 presents the results of the project implementation. Finally, Section 5 offers a comprehensive analysis of the results, a summary, and final observations.

2 THEORETICAL BACKGROUND AND RELATED WORK

This section presents the theories studied and applied in the development of this project, along with similar work in the field of hematological data analysis.

2.1 Clustering Algorithms

Clustering models aim to group datasets based on similar characteristics, dividing the data into distinct clusters, each defined by the behavior of its features' values (Kesavaraj and Sukumaran, 2013). Below, we outline the general principles of several key concepts and methods applied in this study.

2.1.1 K-Means

The K-Means algorithm is widely used for data clustering. According to (Pascual et al., 2007), its main idea is to define k centroids representing different clusters. The process begins by assigning each data point to the nearest centroid, followed by recalculating the centroids and redistributing the points. This cycle repeats until there are no significant changes in the clusters. The centroids, which can be adjusted by the user, are crucial for determining the optimal number of clusters needed for accurate and meaningful classification.

2.1.2 DBSCAN

DBSCAN is a density-based clustering algorithm. According to (Pascual et al., 2007), it relies on the concepts of core points, border points, and noise. A core point is defined as one that has a specified minimum number of neighboring points within a defined radius. The algorithm begins by arbitrarily selecting a point p . If p is a core point, a cluster is formed by

including all objects that are density-reachable from p . If p is not a core point, the algorithm moves to the next object. This process repeats until all objects have been processed. Points that are not assigned to any cluster are considered noise, while points that are neither noise nor core points are designated as border points.

DBSCAN generates clusters based on point density, offering significant flexibility in data grouping and ensuring thorough, accurate classification by evaluating all points.

2.2 Classification Algorithms

Classification models are essential for organizing and analyzing large volumes of information. These models assign an instance with an unknown class to a specific class within a predefined set, dividing data records into distinct classes based on the values and relationships among variables (Kesavaraj and Sukumaran, 2013).

Among classification algorithms, decision trees are widely used. These predictive models are based on inductive learning from observations and logical structures, using training data to gradually learn patterns that connect variables (Martínez et al., 2009).

Building a decision tree involves iteratively dividing the dataset into smaller subsets. At each step, the variable that best separates the data into distinct classes is chosen, using impurity measures such as *entropy* or the *Gini* index. This approach allows the tree to capture the underlying structure of the data, facilitating the identification of complex patterns and relationships.

2.3 Neural Networks

Artificial Neural Networks (ANNs), or simply neural networks, are inspired by the functioning of the human brain. They consist of a large number of interconnected neurons that process information (Wang, 2003). Their value lies in their ability to make inferences based on learning from previous data. They are part of several methods depicted under the concept of computational intelligence (Pincay, 2022).

Today, neural networks are widely applied to solve problems such as pattern recognition, image segmentation, and facial recognition, among others.

2.4 Related Work

This section reviews existing research that contributes to the objectives of this initiative.

In the work of (Meena et al., 2019a), the authors investigated childhood anemia and the relationship between maternal health and diet during pregnancy with the child's anemic status. They compared two techniques—decision trees and association rules—to determine the most suitable approach and proposed a model to reduce anemia risk in healthcare. The results are presented as rules, along with recommendations for doctors, parents, and governments to reduce the risk of childhood anemia.

In a similar study (Noor et al., 2019), a non-invasive method was proposed for detecting anemia using images of the palpebral conjunctiva. By analyzing photographs of conjunctival color and applying decision trees and Support Vector Machine (SVM) algorithms, a rapid diagnosis was achieved with approximately 82% accuracy. However, the authors noted limitations due to the dataset size and emphasized the need for more comprehensive testing.

Another initiative with significant findings was presented by Mena et al. (Meena et al., 2019b). By comparing decision trees and association rules, the researchers proposed a model to reduce anemia risk by identifying influential factors in child nutrition in India. Factors such as location, state, and gender were identified as influential in determining anemia presence in children. Additionally, recommendations were provided for doctors, parents, and the government based on identified rules to help reduce the risk of childhood anemia.

Other related works include (Schober and Vetter, 2021), (Mansoori et al., 2024), and (Parthvi et al., 2020); all highlight the importance and feasibility of applying data mining and data science in supporting the diagnosis of hematological diseases.

In contrast to these related studies, to our knowledge, no studies in this field have been conducted in Ecuador that focus on blood test analysis. This gap underscores the value and contribution of this research initiative.

3 METHODOLOGY AND USE CASE

For this study, the MLOps methodology was employed (Cloud, 2023). Based on this methodological framework, the development of this research proposal considered seven phases: i) business analysis and assessment of its information system (IS) structure; ii) data extraction; iii) data validation; iv) data preparation; v) modeling; vi) model evaluation; and, for ML-related models, the additional step vii) model validation. Figure 1 illustrates these phases and their

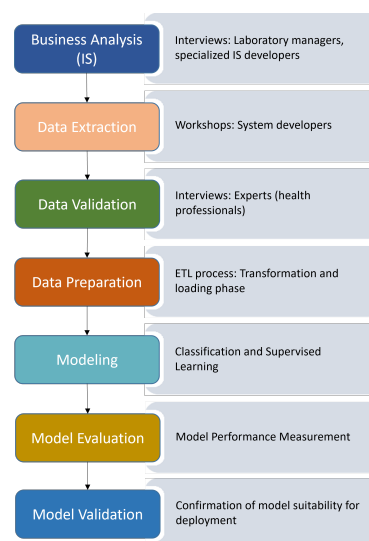


Figure 1: Depiction of the method followed in this study.

intermediate actions.

The data for this research consists of anonymized clinical laboratory results from complete blood count tests. Working with this data requires an understanding of the business rules and the data's significance, as well as identifying models that support pattern recognition and knowledge generation. In this regard, MLOps proved to be the most suitable methodology for this research.

3.1 Method

In the business analysis phase, interviews were conducted with personnel responsible for sample processing and result recording. During these interviews, we identified the parameters evaluated in the complete blood count (CBC) and the tolerance ranges for each parameter across women, men, and children. Depending on the equipment, technique, and reagents used by the clinical laboratory, the CBC test evaluates different blood components crucial for diagnosing a variety of medical conditions.

The objective of this initial phase was to understand the business, its processes, and the context of CBC testing in a clinical laboratory in Quito. It also enabled the identification of the data's complexity, comprehensiveness, and quality.

In the data extraction phase, anonymized data was obtained directly from the laboratory's specialized system, with assistance from laboratory staff and system developers.

During the data validation phase, and based on interviews and information gathered from healthcare experts, the resulting components or parameters from the CBC test were identified for this

Table 1: Sample of the original data structure.

orderid	age (years)	sex	date	result
77871	43	M	02/01/2020	HEMATOCRIT: 53.5
77871	43	M	02/01/2020	HEMOGLOBIN: 17.3
77871	43	M	02/01/2020	PLATELETS: 259
77871	43	M	02/01/2020	WHITE BLOOD CELLS: 8.57
77871	43	M	02/01/2020	NEUTROPHILS: 5.38
77871	43	M	02/01/2020	LYMPHOCYTES: 2.49
77871	43	M	02/01/2020	NEUTROPHILS
77871	43	M	02/01/2020	LYMPHOCYTES
77871	43	M	02/01/2020	RED BLOOD CELL COUNT: 5.84
77871	43	M	02/01/2020	MEAN CORPUSCULAR VOLUME: 91.6
77871	43	M	02/01/2020	MEAN CORPUSCULAR HGB: 29.6
77871	43	M	02/01/2020	MEAN CORPUSCULAR HGB CONC.: 32.3
77871	43	M	02/01/2020	RDW CV: 12
77871	43	M	02/01/2020	MID
77871	43	M	02/01/2020	MID: 0.70
77871	43	M	02/01/2020	MPV: 10.2
77871	43	M	02/01/2020	PDW: 16.6
77871	43	M	02/01/2020	PCT: 0.265
77871	43	M	02/01/2020	RDW SD: 46.4

study. All variables (19 components) were analyzed and cross-referenced to identify those relevant for diagnosing anemia and polycythemia (MedlinePlus, 2022). Key components for these conditions included Hematocrit, Hemoglobin, Red Blood Cell Count, Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), and Mean Corpuscular Hemoglobin Concentration (MCHC) (Institute, 2024). The HDW CV (Red Cell Distribution Width - Coefficient of Variation), representing the distribution width of erythrocytes, was excluded. While HDW CV supports anemia diagnosis and other medical conditions, it does not contribute to polycythemia diagnosis (Mansoori et al., 2024).

In the data preparation phase, the data was structured to represent one test per record, with each component or parameter in a separate column. Table 1 presents the original data structure of a test, while columns such as age, sex (1 = M, 0 = F), and result were reorganized to facilitate analysis.

In the modeling phase, we aimed to identify clustering conditions within the data. Models for K-means and DBSCAN were developed, and with the variables related to anemia and polycythemia estimation, machine learning models such as neural networks and decision trees were created.

For the evaluation phase, we applied the scoring function and the confusion matrix. In the validation

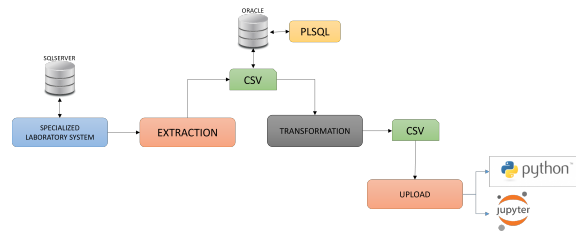


Figure 2: Applied ETL Process.

phase, we assessed whether the models met the deployment requirements; this phase focused solely on the machine learning models.

3.2 Case Study

Table 1 outlines the original structure of the data, which served as the foundation for the transformation process. This transformation was conducted in PLSQL on a virtual server running ORACLE LINUX as the operating system and ORACLE-XE as the database. In this environment, the table structures were created, and the data was cleaned and transformed to produce the final dataset. Anonymized data from hematological biometric tests conducted on various regular patients of a laboratory in Quito from 2020 to 2023 was utilized. It is important to note that these parameters may vary for patients residing in different regions and geographic conditions than those of Quito.

Based on the aforementioned structure and the final dataset, K-Means (Ahmed et al., 2020) and DBSCAN (Deng, 2020) clustering models were applied to identify potential relationships among hematological biometric parameters. Machine learning models, including neural networks (Alsheref and Gomaa, 2019) and decision trees (Noor et al., 2019), were also employed to identify parameters that could support the development of a diagnostic support system for anemia and polycythemia in the future.

4 IMPLEMENTATION

All models were implemented using the Python programming language (Bonaccorso, 2019). For clustering models, the Pandas library was used, and for machine learning, the Sklearn library (Raschka and Mirjalili, 2019) was applied. Figure 2 shows the ETL (Extract, Transform, Load) process followed in this research proposal, as well as the various technological resources applied in the process.

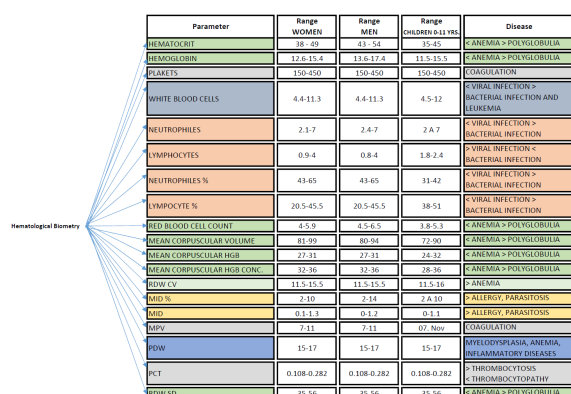


Figure 3: Components: Tolerance Values and Medical Conditions for Hematological Biometrics.

4.1 Data Extraction and Analysis

The Data Extraction and Analysis phase in the MLOps methodology defines two activities: 1) understanding the business context, and 2) extracting and understanding the data.

4.1.1 Business Understanding

The laboratory under study has been legally established since 2006 by Ecuadorian governmental regulatory organizations (ACESS, the Agency for Quality Assurance of Health Services and Prepaid Medicine). It is a medium-sized laboratory providing clinical laboratory services, occupational health, and industrial safety and hygiene services.

The laboratory has thousands of hematological biometrics records. Its managers understand that this information has not been fully utilized and consider it important to identify trends or knowledge that could support medical diagnoses.

4.1.2 Data Extraction and Understanding

The data extracted from the laboratory’s specialized systems went through the ETL process defined in Figure 2. The *result* column in the original dataset (see Table 1) required transformation, generating 19 columns in the final dataset, each defining one of the 19 components or parameters reported by the hematological biometric test, totaling 6551 records (tests conducted between 2020 and 2023). For each of these components, individual tolerance ranges were identified for women, men, and children, as well as associated medical conditions if values fell outside these ranges. Figure 3 illustrates the tolerance values and medical conditions for each of these components.

Table 2: Sample of the final data structure.

ORDER	77871	77873
AGE	43	72
SEX	1	1
ADMISSION_DATE	2/1/2020	2/1/2020
HEMATOCRIT	53.5	52.7
HEMOGLOBIN	17.3	15.8
PLATELETS	259	280
WHITE BLOOD CELLS	8.57	12.9
NEUTROPHILS	5.38	9.43
LYMPHOCYTES	2.49	2.28
NEUTROPHILS	62.8	73.1
LYMPHOCYTES	29	17.7
RED BLOOD CELLS	5.84	5.75
MEAN CORPUSCULAR VOLUME	91.6	91.6
MEAN CORPUSCULAR HGB	29.6	27.4
MEAN CORPUSCULAR HGB CONC.	32.3	30
RDW CV	12	18.2
MID	8.2	9.2
MID	0.7	1.19
MPV	10.2	7.9
PDW	16.6	15.9
PCT	0.265	0.22
RDW SD	46.4	70
SEDIMENTATION		
ANEMIA	1	1
POLYCYTHEMIA	0	0
ANEMIA CONF LVL	1	2
POLYCYTHEMIA CONF LVL	0	0

4.2 Data Preparation

Based on the established ranges for each component, each test was analyzed to define a variable indicating whether the result suggests conditions of anemia or polycythemia. A confidence level variable was used (NIV CONF ANE, NIV CONF POL) to measure the reliability of the pathology assigned to each test. This variable increases based on the number of parameters indicating either anemia or polycythemia. An algorithm in PLSQL was developed to analyze the tolerance ranges for each component across the 6551 tests, identifying the results and recording the values in the variables ANEMIA, POLYCYTHEMIA, ANEMIA CONFIDENCE LEVEL, and POLYCYTHEMIA CONFIDENCE LEVEL. Table 2 displays the final structure of the dataset (the fieldnames have been translated to the English language from Spanish).

4.3 Modeling

4.3.1 Decision Tree

A decision tree classification model was used. The model structure has a depth of 5 levels for both pathologies, as shown in Figures 4 and 6. This struc-

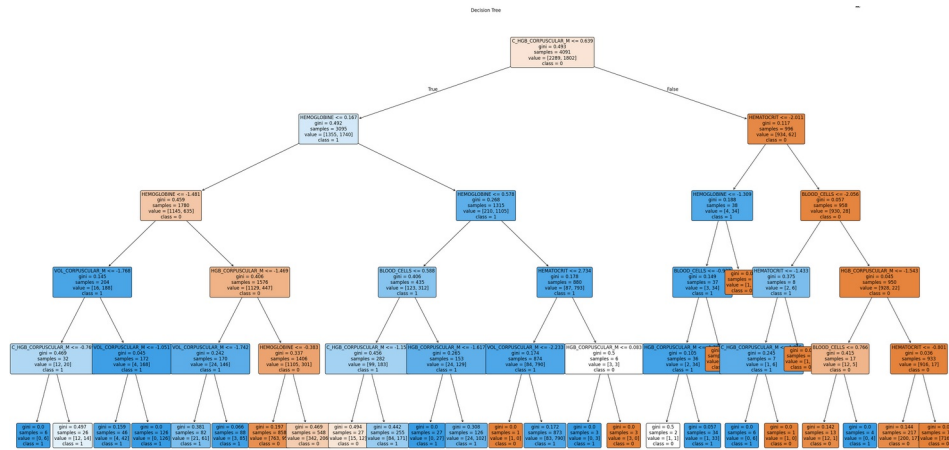


Figure 4: Anemia Decision Tree.

ture allows the model to successfully predict (Class 0 = NO ANEMIA, Class 1 = ANEMIA) without overfitting the data.

For the decision tree in anemia, it is observed that the root node's dominant variable is M CORPUSCULAR HGB, with a *Gini* index of 0.493, indicating an impurity or mixture of classes. In the following tree levels, decisions are split between other characteristics or components, suggesting that the decision for anemia is not solely based on one parameter. However, based on the *Gini* index, it can be deduced that hematocrits are more likely to lead to an anemia diagnosis. It is notable that most intermediate nodes have low *Gini* values, indicating good class separation for prediction (ANEMIA, NO ANEMIA). The tree leaves indicate that variations in HEMOGLOBIN and HEMATOCRIT values affect M CORPUSCULAR HGB and CORPUSCULAR HGB; as seen in Figure 4.

As shown in Figure 5, in the univariate diagonal comparison, the non-anemia values are higher than the anemia values. On the other hand, in anemia variables like HEMATOCRIT, HEMOGLOBIN, and MEAN CORPUSCULAR HEMOGLOBIN CONCENTRATION, clear peaks denote bimodality and suggest significant differences in the data.

In the bivariate comparison of HEMOGLOBIN, HEMATOCRIT, RED BLOOD CELLS, MEAN CORPUSCULAR HEMOGLOBIN CONCENTRATION, and MEAN CORPUSCULAR HEMOGLOBIN, a linear correlation is observed, confirming a relationship among these variables. Several points show clear class separation, suggesting these variables may be good discriminators for the target.

In Figure 6, the root node is HEMOGLOBIN, considered determinant due to its low *Gini* index; this tree

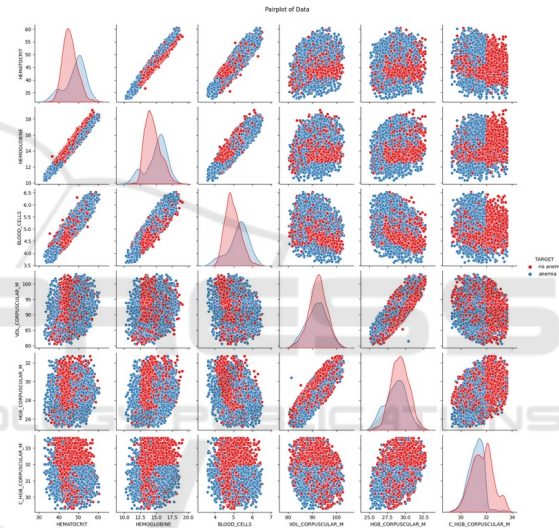


Figure 5: Anemia Multivariable Scatter Matrix.

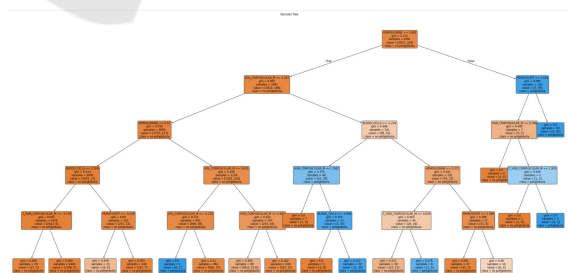


Figure 6: Polycythemia Decision Tree.

shows a high probability of diagnosing polycythemia based on M CORPUSCULAR VOL and HEMATOCRIT.

In Figure 7, in the univariate diagonal comparison, non-polycythemia diagnoses are more prevalent than polycythemia. However, bimodality is less evident in

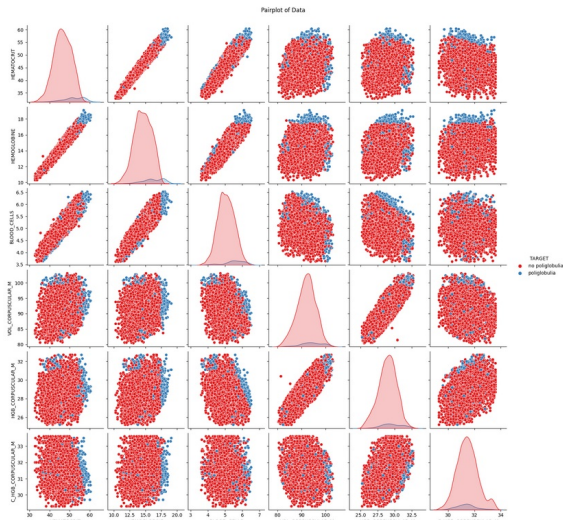


Figure 7: Polycythemia Multivariable Scatter Matrix.

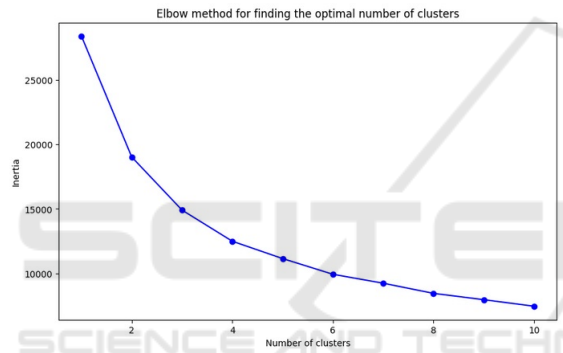


Figure 8: Elbow Plot for Cluster Identification.

this diagnosis.

In the bivariate comparison of highly correlated variables, the data maintain a linear dispersion pattern.

4.3.2 K-Means

The optimal number of clusters was selected using the elbow method, determining three clusters as optimal; as shown in Figure 8.

In Figure 9, the centroids of the three clusters are relatively close to each other, as variable values fall within similar ranges due to normal patient conditions. Those further from these centroids tend to have anemia, polycythemia, or another medical condition.

4.3.3 DBSCAN

Figure 10 shows the k-nearest distance, with a value of 0.6 applied in the model.

The data are grouped into three clusters, as shown in Figure 11. The most prominent cluster (blue) repre-



Figure 9: K-Means Cluster Scatter Plot.

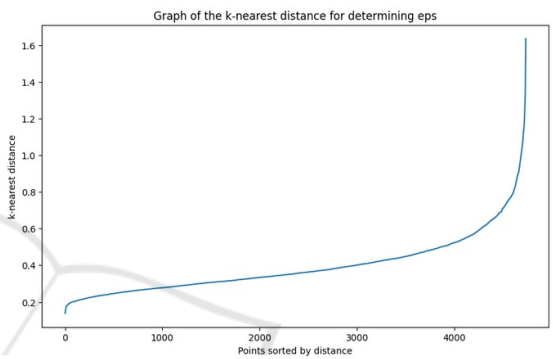


Figure 10: K-Nearest Distance Plot to Identify Eps Density.

sents patients with normal medical conditions, while the additional two clusters (red and green) represent patients with anemia and polycythemia, respectively.

4.3.4 Neural Networks

The network structure consists of an input layer, two hidden layers with 50 neurons each, and an output layer. This structure allows the model to capture important patterns in the data without overfitting.

Figure 12 shows two density curves. For non-anemia patients (first curve), the model predicts values very close to actual values. In anemia patients

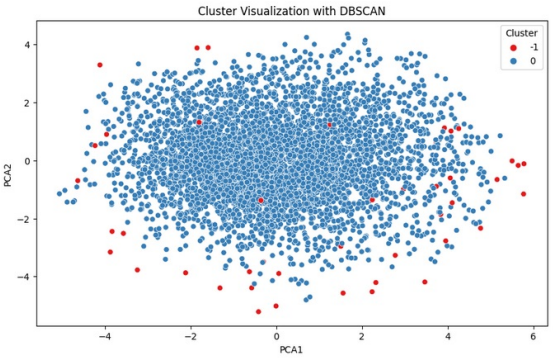


Figure 11: DBSCAN Cluster Scatter Plot.

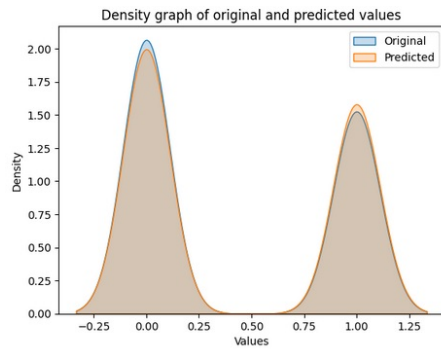


Figure 12: Density Plot for Real vs. Predicted Anemia Values.

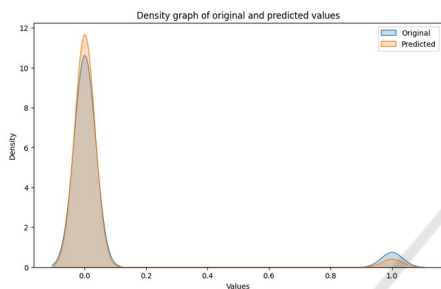


Figure 13: Density Plot for Real vs. Predicted Polycythemia Values.

(second curve), the model slightly overpredicts anemia with a 0.7

In Figure 13, the density plot for polycythemia cases shows real and predicted values. As the positive values are few (6.6% of the population), the model is less efficient, with a 3.4% false negative rate.

4.4 Model Evaluation and Validation

To evaluate and validate these models, a confusion matrix was used, an example of which is shown in Figure 14, corresponding to the anemia decision tree.

Accuracy, Precision, Recall, and F1-Score metrics were also calculated. Table 3 presents the results for

Confusion matrix	
TN = 905 905	FP = 104 104
FN = 137 137	TP = 608 608

Figure 14: Anemia Decision Tree Confusion Matrix.

Table 3: Evaluation and Validation Metrics.

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
Decision Tree				
Anemia	0.86	0.85	0.82	0.84
Polycythemia	0.95	0.89	0.42	0.57
Neural Networks				
Anemia	0.85	0.81	0.84	0.83
Polycythemia	0.96	0.85	0.42	0.57

each model, showing that the sensitivity in predicting positive cases is good and that predictions are correct a high percentage of the time, except for models applied to polycythemia. Although polycythemia models have high accuracy and a high positive prediction ratio, class imbalance is evident due to the limited polycythemia cases identified.

5 CONCLUSIONS

This study presents a detailed analysis of using anonymized hematological biometrics data to develop models that facilitate the diagnosis of blood disorders, such as anemia and polycythemia. The project followed an MLOps methodology across seven phases: business analysis, data extraction, data validation, data preparation, modeling, evaluation, and model validation. In the initial phase, an in-depth analysis of the clinical laboratory context was conducted, identifying relevant processes. This was followed by the extraction and comprehension of available data, which were prepared through an ETL process that ensured compliance with healthcare data protection regulations. Once the dataset was prepared, clustering and classification techniques were applied, leading to the evaluation and modeling of the resulting models.

The clustering analysis enabled the identification of common characteristics among patients, though diagnostic accuracy posed challenges. To address this, decision tree and neural network models were developed for the classification of anemia and polycythemia. Both models exhibited similar performance according to evaluation metrics, demonstrating strong diagnostic capabilities for anemia. In terms of F1-score, the models showed better performance in identifying anemia than polycythemia. Specifically, the recall for polycythemia was 0.42, indicating limited capability in correctly identifying positive cases. Conversely, both models showed an adequate capacity to diagnose anemia cases accurately.

The results indicate that while the developed models are suitable for diagnosing anemia, they have limitations in identifying polycythemia, suggesting a need for dataset improvements to address these limitations. Additionally, these models could be valuable in settings with limited access to healthcare services, such as certain regions in Ecuador.

In conclusion, the models developed are promising for anemia diagnosis but require further improvement for precise detection of polycythemia.

Future work will focus on enriching the datasets and extending the models to identify other diseases, such as leukemia. Early diagnosis is critical to improving recovery rates and preventing health deterioration. This study lays the groundwork for implementing automated diagnostic systems that could significantly benefit populations with limited access to healthcare services, contributing to enhanced management and treatment of hematological disorders.

REFERENCES

- Ahmed, M., Seraj, R., and Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9:1295.
- Ahsan, M. M., Luna, S. A., and Siddique, Z. (2022). Machine-learning-based disease diagnosis: A comprehensive review. In *Healthcare*, volume 10, page 541. MDPI.
- Alsheref, F. K. and Gomaa, W. H. (2019). Blood diseases detection using classical machine learning algorithms. *International Journal of Advanced Computer Science and Applications*, 10(7).
- Bonaccorso, G. (2019). *Hands-on unsupervised learning with Python: implement machine learning and deep learning models using Scikit-Learn, TensorFlow, and more*. books.google.com.
- Cloud, G. (2023). Mlops: canalizaciones de automatización y entrega continua en el aprendizaje automático.
- Deng, D. (2020). Dbscan clustering algorithm based on density. *2020 7th international forum on electrical ...*
- Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20):1920–1930.
- Gunčar, G., Kukar, M., Notar, M., Brvar, M., Černelč, P., Notar, M., and Notar, M. (2018). An application of machine learning to haematological diagnosis. *Scientific reports*, 8(1):411.
- Institute, N. H. G. R. (2024). Linfocito - glosario parlante de términos genómicos y genéticos. Accessed: 2024-05-30.
- Kesavaraj, G. and Sukumaran, S. (2013). A study on classification techniques in data mining. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, pages 1–7. IEEE.
- Mansoori, A., Gohari, N. F., Etemad, L., and ... (2024). White blood cell and platelet distribution widths are associated with hypertension: data mining approaches. *Hypertension ...*
- Martínez, R. E. B., Ramírez, N. C., Mesa, H. G. A., Suárez, I. R., Trejo, M., León, P. P., and Morales, S. L. B. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista médica de la Universidad Veracruzana*, 9(2):19–24.
- MedlinePlus (2022). Biblioteca nacional de medicina usa. Consultado: 2024-05-30.
- Meena, K., Tayal, D. K., Gupta, V., and Fatima, A. (2019a). Using classification techniques for statistical analysis of anemia. *Artificial Intelligence in Medicine*, 94:138–152.
- Meena, K., Tayal, D. K., Gupta, V., and Fatima, A. (2019b). Using classification techniques for statistical analysis of anemia. *Artificial Intelligence in Medicine*, 94:138–152.
- Noor, N. B., Anwar, M. S., and Dey, M. (2019). Comparative study between decision tree, svm and knn to predict anaemic condition. In *2019 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITH-CON)*, pages 24–28.
- Parthvi, A., Rawal, K., and Choubey, D. K. (2020). A comparative study using machine learning and data mining approach for leukemia. In *2020 International Conference on Communication and Signal Processing (ICCSPP)*, pages 0672–0677.
- Pascual, D., Pla, F., and Sánchez, S. (2007). Algoritmos de agrupamiento. *Método Informáticos Avanzados*, pages 164–174.
- Pincay, J. (2022). *Computational Intelligence*, pages 33–56. Springer International Publishing, Cham.
- Raschka, S. and Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. books.google.com.
- Schober, P. and Vetter, T. (2021). Logistic regression in medical research. *Anesthesia & Analgesia*.
- Wang, S.-C. (2003). *Artificial Neural Network*, pages 81–100. Springer US, Boston, MA.