2025年臺灣國際科學展覽會
優勝作品專輯

作品編號　190032

參展科別　電腦科學與資訊工程

作品名稱　MEDTEC - Artificial Intelligence Software for medical diagnosis optimization and analysis

得獎獎項　成就證書

就讀學校　Colégio Dante Alighieri
指導教師　Fabiano Zuin Antonio
作者姓名　Ana Elisa Guirao Gomes

關鍵詞　Deep Learning; Neural Networks; Medical Diagnoses; Blood Counts.

作者照片

MEDTEC

Artificial Intelligence Software for medical diagnosis optimization and analysis

Ana Elisa Guirao Gomes
Orientor: Prof. Dr. Tiago Bodê
Co-orientor: Msc.Rodrigo Assirati
Colégio Dante Alighieri

2025

Contents

Abstract

In Brazil, approximately sixty million people suffer from or acquire some type of disease daily. However, the average time for blood count diagnoses, used to identify many of these diseases, remains very lengthy. This can lead to the worsening of conditions and delays in care, as well as a decrease in the patients' quality of life. Moreover, in some cases, the waiting period can result in irreversible situations and even the death of the affected individuals. In this landscape, technological tools such as artificial intelligence software can help reduce the time taken for diagnostic reporting. In light of this, the project involves developing software to assist in the analysis of blood counts and optimize medical diagnoses. For this purpose, the methodology was divided into three stages. In the first, titled "Medical Standardization", a survey of the standard variables related to diseases that can be identified with the help of blood counts was conducted. Among the findings, diabetes, anemia, leukemia, dengue, polycythemia, tuberculosis, leprosy, meningitis, chlamydia, schistosomiasis, spotted fever, and malaria were the main diseases detected. Furthermore, hemoglobin, leukocytes, platelets, glucose, cholesterol, ions, and hormones were the key findings concerning the primary blood indicative factors for the mentioned diseases. In the second phase, the theoretical and practical foundations of the software were developed, based on artificial neural networks. In Python, regression models were also crafted to check the feasibility of the analyses. Finally, the last stage consisted of testing with real datasets, based on 1,227 anonymized blood counts. Among the artificial intelligence algorithm models tested, Support Vector (0.02) and Multiple Linear (0.61) had the lowest performances, while Polynomial (0.97), Random Forest (1.0), and Decision Tree (1.0) showed the best results. Given that the Random Forest and Decision Tree regression models achieved an accuracy of 1.0, while the Polynomial model scored 0.97, Support Vector 0.02, and Multiple Linear Regression 0.61, it is concluded that the blood count analysis system, with Python tools like regression, proved to be highly efficient. The closer the R^2 value is to 1.0, the better the programming fits the model, ensuring accurate analyses. Aside from that, in order to expand the number of analysis possible to do be done we decided to use a second tool called "classification", with which we made a bigger dataset to be used as a model to identify blood related diseases and the behavior of complex and diverse diseases. With that in mind, we performed a second evaluation of the models by doing an accuracy test, scored 87 percentage points and with a confusion matrix. With those results, we verified that the high performance of the tests indicates that Artificial Intelligence can be avaunt-guard to the elaboration of more efficient medical diagnosis, improving people's lives quality and, overall, lowering the number of deaths in our country.

Key-words: Deep Learning; Neural Networks; Medical Diagnoses; Blood Counts.

Introduction

In Brazil, the queues for public medical care have a one year and four months estimated waiting line (ALMT, 2020). This is a worrying fact, especially when considering that between the eleven main death causes in the country, eight are diseases that might many times need medical assistance (IHME, 2018). This scenario often results in the worsen of the medical cases and the long waiting time for assistance, leading to a lot of downturns in the patient's quality of life. Besides that, time can be a determining factor when it comes to discovering diseases and finding the right diagnosis, may resulting in cases of death, as showed previously. With that in mind, the importance of early diagnosis is clear to the patients life, being the waiting lines an impaction problem that needs to be solved quickly.

Thinking about that, the technological resources, such as Artificial Intelligence (AI) softwares, can help to optimize the emission timing of the diagnosis, specially related to the complete blood count exam (Figure 1). This two resources, when combined, can analyse a huge variety and quantity of data, that, by themselves, can identify and understand the disease's behavior more rapidly. With that, the softwares represent a plausible solution to the delay mentioned, helping in the diagnosis proccess.

1 Context

1.1 Complete Blood Test

Medicine is an area with an exponential growth. By being a very broad study, new techniques, methods and equipments are developed everyday. With that, very recent fields such as genetics, nuclear medicine and bioengineering are departments that need and work a lot with innovations (ALCANTARA, 2019). Deeper into that, medical diagnosis field is an extremely relevant area today, that claims and invests a lot in new was of discovering and visualizing diseases.

According to a publication done by Hospital Israelita Albert Einstein (HIAE, 2023) in their official website, complete blood exams are extremely important during the patient's daily life and checkups. They work through a sample of blood from a determined patient using a needle and a collector. Stating there, the components present inside the blood collected will be analyzed, usually, with a numerical counting of each variable in the blood sample and with a correlation with an expressed number of reference. Between the elements that are studied it is possible to see: red blood cells, white blood cells and platelets.

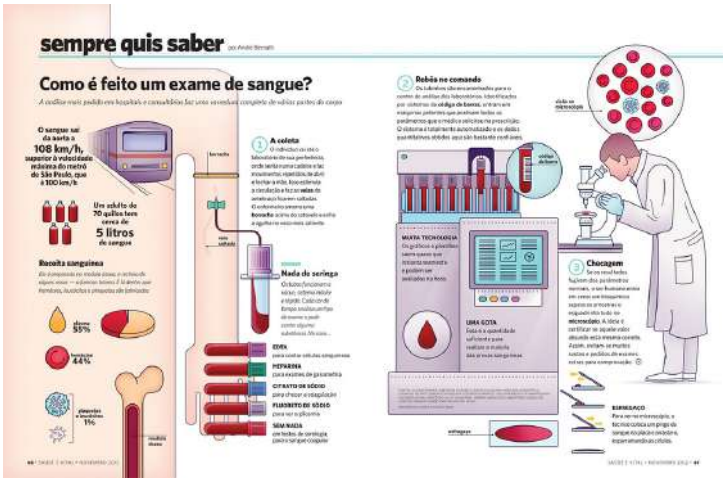

Figure 1: Complete Blood Exam - Basic procedures needed to be done during the exam.
Available at: https://encr.pw/TALQZ Accessed on 16th nov 2024

The blood exams also have the counting of five types of white blood cells (Neutrophils, Eosinophils, Basophils, Lymphocytes and Monocytes). Besides that, they can include an evaluation based on erythrograms and leukograms.

The Neutrophils are blood cells, also called polymorphonuclear leukocytes, responsible for phagocytose strange substances and adverse to the intracellular surface. This cells are produced in the bone marrow during a process called granulocytopoiesis.

Hematological values of reference in adults

		Men	Women
RBC	x 10^{12}/L	5,00 - 0,5	4,3 - 0,5
Hemoglobin	g/dL	15,0 - 2,0	13,5 - 1,5
Haematocrit	(%)	45 - 5	41 - 5
Leukocytes	x 10^{9}/L	7,0 - 3,0	
VGM	fL	92 - 9	
HGM	pg	29,5 - 2,5	
CHGM	g/dL	33 - 1,5	
RDW	CV (%)	12,8 - 1,2	
	SD (fL)	42,5 - 3,5	
	x 10^{9}/L	150 - 400	

Figure 2: Values of reference used in the blood analysis related to the biological sex
Available at: https://dante.pro/deeplAccess20nov2024.

The erythrograms evaluate the red species (red blood cells), through hemoglobin dosages and cell counting. With that,it allows the diagnose of polyglobules (FAILACE, 2015), diseases very frequent due to the small quantity of red blood cells in the sample being researched.

While that happens, leukograms are responsible for the check up of white blood cells. This part of the exam counts with an analysis of the cell's format, beyond doing a second counting. Taking that into consideration, it is indicated to diagnose infections, leukemia and other diseases related to the deformation or anomaly in the quantity of white blood cells.

In order to have a correct diagnose, it is necessary to perform an analysis with values of reference, especially sex, age, physical activity, etc (Figure 2). Being those parameters, due to that, essential to determine the normality of the cases.

In addition, Dutra (2020) study reveals the importance of blood tests in the early diagnose of leukemia. According to the author, a study realized in the service from the department of oncology from São Paulo has shown that in a huge

quantity of cases of leukemia the blood exams presented a huge amount of information that allowed the identification of the disease or the suspicion of its presence. From that, it was concluded that the blood tests are exams with a great potential to diagnose leukemia, serving as a base exam to this disease.

In that sense, it is evident that the blood tests can be important auxiliary tools to identify many different diseases (ROSENFELD, 2012). That's due to the fact that they present valuable information about blood components that may vary a lot depending on the existing condition.

The importance if the blood exams amplifies in medicine working as an essential and practical information device (ARAÚJO, 2022). It is necessary to highlight the quick identification of oncological diseases in UPA's activities (public health units), showing the aid provided to the doctors, especially with the description of anomalous leukocytes , blastocoels and abnormalities in the granulocytes series. With that, there's one more example about the exponential using of this type of exam to identify health conditions.

1.2 Artificial Intelligence

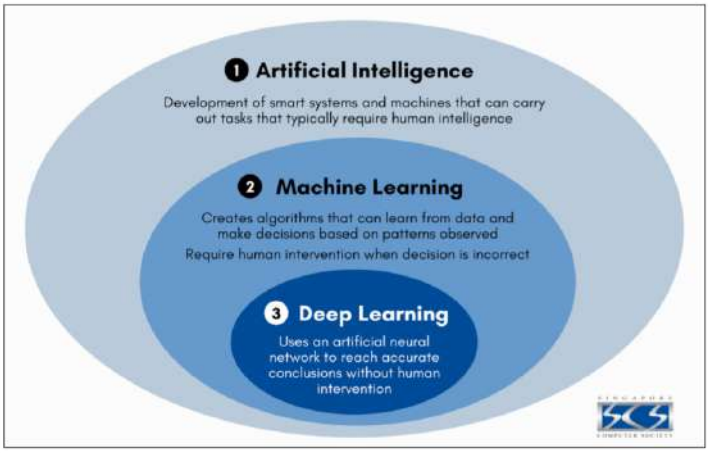

Figure 3: AI (1); Machine Learning (2) e Deep Learning Available at: https://l1nq.com/VZ8AzAccess20nov2024.

Artificial Intelligence (AI) is one scientific area that has the objective to simulate human activities in digital platforms. According to the Singapore Computer Society, this intelligence uses smart systems and machines to make activities that were once dependent on human intelligence and action. Between the many tools of AI, there is Machine Learning, Deep Learning and Softwares (LYU, 1996).

Thinking about that, one of the applications of AI is related to the medical diagnosis and the exam analysis. Between them, one exam that allows many diagnoses is blood exams. In it, information such as quantity of glucose, hemoglobin, leukocytes, plaques, quantities of hormones, minerals, water and many other substances can be detected. With that, while using artificial neural networks - method of processing data - the information collected serve as "input" of the system (first entry), the comparisons between the data patterns with the intern layer and, finally, the result of the analysis as an "output" (or last exit).

1.2.1 Adequate usage of AI

The adequate usage of Artificial Intelligence allows to complete many different tasks, but, at the same time, it carries with it great responsibilities while being used (THIEBES; LINS; SUNYAEV, 2021). Due to that, AI was divided in five types of uses that must be taken into considedration during its execution and activity time

With that in mind, the principles mentioned are:**benefits with no prejudice, autonomy, justice and relevance**.

The first term "benefits with no prejudice" relates to advantages brought by AI to the daily life in a way that it does not prejudices other people, institutions, profiles, among others. The use of this intelligence should be done with caution, always achieving the objectives of conduct on the internet, without violating digital rights, authorship, with responsibility, and in a way that benefits all those involved. Meanwhile, autonomy refers to the process of independence generated by its use, as this new technology becomes an efficient way to accomplish tasks that were previously complicated and difficult. In light of this, its users are able to be autonomous in their projects while also being responsible for their outcomes. Justice, an important term that forms part of the principles, is directly related to the constitution of the current countries and digital regulation. The internet often becomes a "no man's land," being used as a symbol of liberation through masks (pseudonyms) and without accountability for the facts. (DE SOUSA, 2013). Therefore, from this, justice should be used as a means of overseeing digital activity.Finally, the concept of relevance is closely related to the objectives of using AI, presenting reasonable arguments for its implementation in appropriate platforms and locations.

Based on them, the correct use of Artificial Intelligence can be carried out in a way that does not harm any user and does not compromise the ethics of the project. From this, one can think of a usage method that follows the principles established in the article, such as the softwares — a sequence of digital instructions to perform a specific task, constituting a program — for example.

1.2.2 Functioning

Moreover, AI models are determined by machine learning. An essential step for these programs occurs through the training of equipments with programmed scenarios, often involving various possibilities of actions and data.(HASHIMOTO et al., 2018). With each scenario introduced, the programming directs the computer to evaluate the necessary actions, and through a system of checks and error, it automatically learns about the required steps.

There is also the growing use of this programming in medicine, particularly in the fields of endoscopy and surgery (HASHIMOTO et al., 2018) that require training for both doctors and equipment, especially for the proper use of the technology. This fact can also be confirmed by the works of Johnson et al. (2018) e Chang et al. (2019), both addressing the application of new technologies in medical fields, specifically in cardiology and pathology.

1.2.3 AI on daily life

Regarding the use of AI systems in cardiology, we can mention the standardization of cardiological exam algorithms, the organization of waiting lists and patients, devices such as pacemakers, and more. Thus, this is an emerging area for new technologies to operate, primarily because it has a large structure of exams and operational sources that can be enhanced with technological use (Figure 4).

While that happens, in pathology, AI works through the visualization of cells and the recognition of morphological patterns by a computational program, for example. As mentioned, clinical, radiological, and genomic information are often used by those programs, which is fed into deep learning systems that will be explained further.

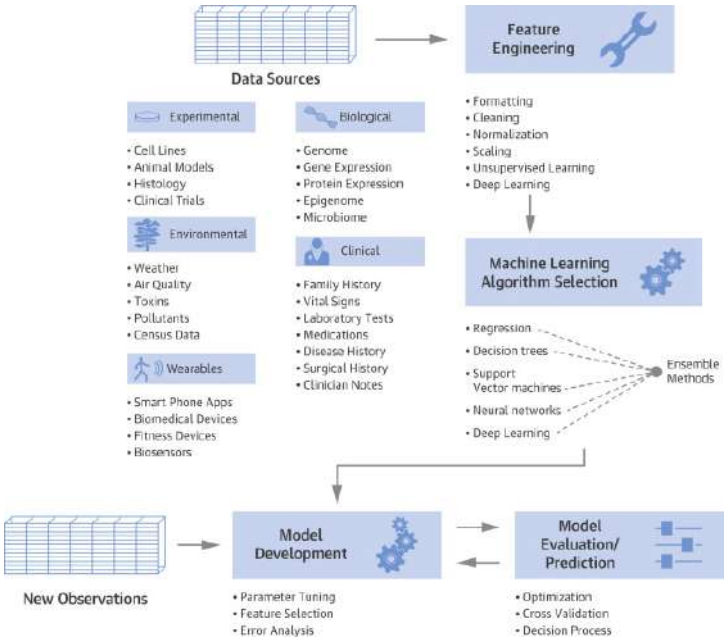

Figure 4: Using of AI in medical fields
Available at:https://doi.org/10.1016/j.jacc.2018.03.521 Accessed on 26 nov 2024.

1.3 Machine Learning

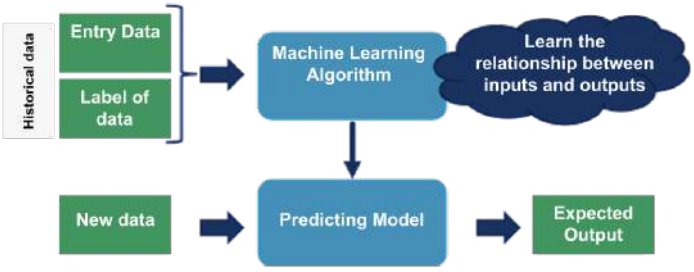

Figure 5: Simulation of a machine learning program
Available at:https://dante.pro/ml Access on 26 nov 2024.

Within the field of Artificial Intelligence, one can find Machine Learning — serving as a kind of subdivision - which includes the ability of programs to learn to recognize patterns and information. They primarily operate through training with data and problem-solving, without human interference. (JANIESCH; ZSCHECH; HEINRICH, 2021).

Such a system, with tests and examples, can, thus, contribute to the quick solution of problems that require the identification of data, information, images, and patterns in our daily lives.(RUDIN et al., 2022). In this way, Machine Learning programs, usually, make daily tasks faster and more efficient.

In this scenario, these systems function through pattern recognition, similarly to how humans do during their development or growth(JANIESCH; ZSCHECH; HEINRICH, 2021). With that, the machine interacts with the data and programming, drawing conclusions based on its training, and even being able to predict outcomes. This represents a form that goes beyond the general limits of Artificial Intelligence.

1.3.1 Tools

The field of Machine Learning can be divided into several tools. The first of these includes Artificial Neural Networks, which will be discussed in more detail later in this report.

Artificial neural networks, as the name suggests, simulate the human nervous system in programs with interconnected data, information, scenarios, or references (WU; FENG, 2018). These elements are processed in the model's internal layers, which act as a kind of filter. It is at this stage that some patterns between the data are established, contributing to the machine's learning process. With the patterns in place, the program provides, as a response and final result, some predictions made based on the correlations found, thus offering increasingly accurate results as the number of correlations increases.

There are also methods such as linear models. In these systems, the program is induced to establish linear relationships between the data (SU; YAN; TSAI, 2012).Normally, the information to be analyzed has a direct dependency on one another. This means that when one is altered, its dependent counterpart also changes, revealing a pattern of direct influence that, in Machine Learning, can be used for the machine's learning process.

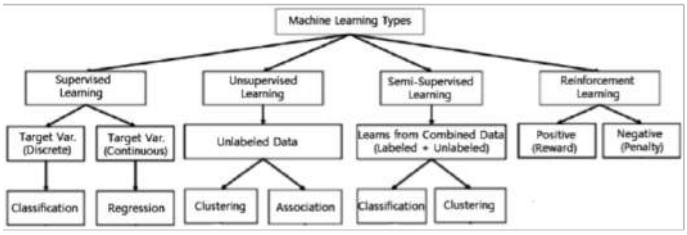

Figure 6: Different types of human intervention in AI
Available at:https://doi.org/10.1007/s42979-021-00592-x Accessed on 26 november 2024

There are several forms of supervision within this field, which can include varying levels of human influence on the results and tests, from greater to lesser involvement (SARKER, 2021). Currently, Machine Learning and Deep Learning systems are used in various areas of daily life. An example of this is their use in security systems, smart cities, online commerce, healthcare centers (the focus of this project), and agriculture.

These examples highlight the relevance of Machine Learning systems, which are increasingly influential in the contemporary world. Besides that, a research made by "Estadão" (Brazilian journal) in April 2024 demonstrated that AI could generate savings of 25 to 50 percent in terms of time and costs during the pre-clinical phase of development.

1.4 Deep Learning

Deep Learning is another division inside AI and Machine Learning. This area's main objective is to reach conclusions and results without the need for human interference. To achieve this, artificial neural networks are used, an important tool that operates through "input" (data entry) and "output" (data exit) (Dong et al.2020).

With this in mind, Deep Learning, by operating specifically in a particular area of machine learning, allows programs to learn on their own. This can greatly contribute to the machine's understanding of medical data, making it even more efficient in its function.

1.4.1 Performance

Deep Learning often outperforms other AI models and even human activity in terms of performance (Liu et al.2020).This is due to the machine's deep learning process, which incorporates not just a single layer of relationships but multiple layers, forming a more complete analysis. Additionally, in deep learning programs, it is often not necessary to adjust the dataset,

allowing work with "raw" data as well (MAYERICH et al., 2023). With this in mind, these models are typically used on a large scale for high-importance issues that require the use of various types of data, such as images, PDFs, graphs, etc.

1.4.2 Usage

Currently, deep learning is increasingly present in the daily life of large cities, being widely used in medicine, particularly with image-based exams (SUGANYADEV et al., 2021). Additionally, the authors emphasize the importance of digital resolution systems in medical fields. It has been observed that these models are strongly utilized in areas such as neurology, pathology, pulmonology, orthopedics, and more. Typically, deep learning is used in disease diagnosis and exams, as seen in the identification of certain factors in images of organs, for example.

A great advantage of this type of programming is its diagnostic processing time. In a matter of seconds, multiple images are often processed and identified, which greatly facilitates efficient medical care.

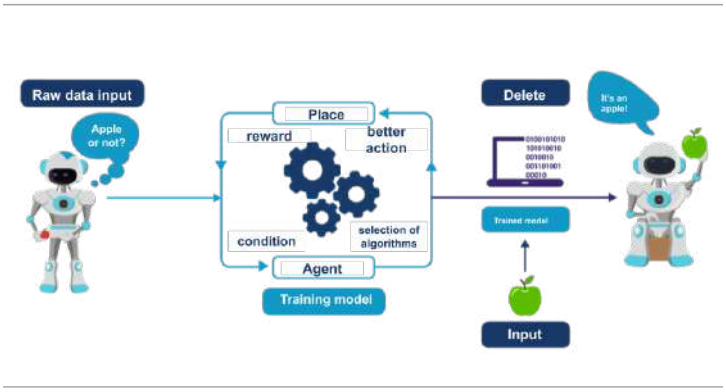

Figure 7: How models of deep learning work
Available at: https://solvimm.com/blog/o-que-e-machine-learning/ Accessed on 26 nov 2024.

In the previous image, we have a simple example of an analysis using Deep Learning (Figure 7). First, there is the identification of a problem or scenario that will be developed in the program. In the case of the image, the problem would be identifying the "apple" algorithm in a data sample. This analogy is simple, but it is important to emphasize the significance of this step in building the model, as without it, there is no clear direction regarding the action that should be taken with the input data. After that, in the internal layers of the programming, the best tool is selected to perform the desired task. In this context, the task would be to identify the presence of the algorithm. The model then begins to test various analysis patterns on the data set until it successfully identifies the desired item. Finally, with the tests that were able to perform the task, a sort of pattern is created, making the machine increasingly capable of executing the same command. The final step involves presenting the data containing the algorithm or installing a new set of data to be analyzed.

1.5 Neural Networks

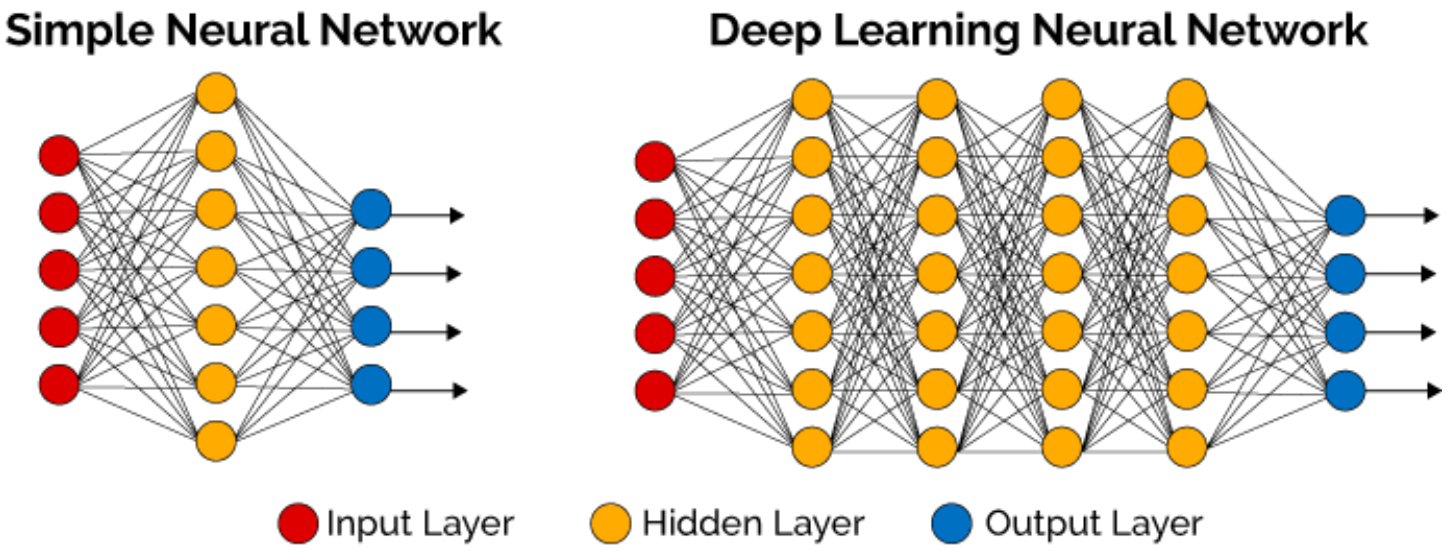

Figure 8: Simulation of a simple neural network (with input, processing and output)
Available at:https://dante.pro/redesneurais Accessed on 26 nov 2024.

Artificial neural networks are a flexible programming method that can be modified for various purposes and scenarios, along with graphic processing units (GPUs) (JANIESCH; ZSCHECH; HEINRICH, 2021).They are based on the human neural system and function similarly to the process of neuron transmission in our body, and can be divided into three main parts.

In Deep Neural Networks (More complex neural networks), there is typically an expansion of the range and scope of the networks, which, through various internal layers, can establish a series of connections between the input data. This type of tool is capable of developing more accurate analyses of a given data set, requiring less human intervention.

As an example of this, one of the steps in building Artificial Neural Networks is called "Feature Scaling." In this step, the model itself identifies irregularities in the input data and, once again without intervention, is able to select the best data from raw inputs to guide its analysis (JUSZCZAK; TAX; DUIN, 2002).

There are various types of programming and tools that use these networks. However, because there is a well-defined structure for neural learning, with systems for identifying the best patterns to establish relationships between the input data, a vast variety of models are formed, suited to different tasks. For this reason, the field of neural networks is often able to more complexly handle a variety of problems and data sets.

1.5.1 Inside division

The first part functions as a data receptor that directs the information to the intermediate layers. In this receptor, tools such as "Feature Scaling," "Pre-processing," "Test," among others, are applied. These tools serve as a pre-processing stage for the data, selecting the best ones to be used in training the model.

After the data is processed, through connection layers, the models operate in the internal layers, which typically contain a large structure that analyzes each input and establishes relationships between them. These relationships can occur in two ways: with supervised learning or unsupervised learning (HAYKIN, 2001).

Supervised learning works with a pre-selection of data in the input layer. From this, it can be stated that in such a model, the first layer is not as complete, meaning, for example, it does not require tools like "Feature Scaling."

On the other hand, unsupervised learning tends to assign more tasks to the model itself, teaching it in a more comprehensive way. In this program, the data is introduced in a raw form, without any type of filtering. This approach is precisely designed to guide the machine in selecting the best data sets for future training and predictions.

After selecting the type of learning, various relationships between the data will be formed, establishing a pattern. This pattern will serve as an analysis algorithm for future processed data sets.

The final layer, ultimately, constitutes the output of the data, now already processed. In this output, the data that matches the pattern of the algorithm used will be presented, thereby contributing to the evaluation of the model and its future predictions.

1.6 Programming Language - Python

To shape software programs or any type of digital equipment, it is necessary to use a programming language. However, there are various types of languages that can be used, such as Python, Java, Ruby, R, PHP, LaTeX, among others. These languages are typically used for data processing, transforming the information provided. (GOTARDO, 2015).

For the present project, the chosen programming language is Python. This type of programming language is very versatile, working particularly well with numerical analyses, for example. Additionally, it allows the creation of graphs, spreadsheets, tables, networks, and more.

With this in mind, the programming begins primarily with the introduction of libraries. Among them, we can mention Numpy, Pandas, Scikit-learn, and others. These libraries contain a range of commands and possible actions that the model can perform, offering a wide variety of functions.

1.6.1 Scikit-learn

For computing Machine Learning models in Python, a very interesting library is Scikit-learn, as mentioned earlier (PEDREGOSA et al., 2011). This module contains commands for a vast range of scenarios, including both supervised and unsupervised learning, and is widely used in academic and commercial fields.

With this in mind, the models work by structuring the libraries followed by identifying the dependent and independent variables within the provided dataset. This step provides the model with the resources to formulate coherent predictions in the future. Additionally, there is the naming of omitted data, which compromises the model's analysis, leading to either the exclusion or substitution of such data, depending on the type of program being used and the accuracy required for the analysis (more or less precise).

After this step, it is necessary to split the dataset into training and testing sets. Therefore, a percentage of the data should be specifically allocated for training, while another portion is reserved for testing, or as a control group. This is similar to the division made in scientific experiments and, for this reason, it should be standardized. Thus, there are two main types of models in Python: Regression models and Classification models.

1.7 Regression

Regression models are those that perform an analysis correlating two variables. Typically, one of the variables will be the dependent variable, and the second will be the independent variable. In other words, a change in the independent variable leads to a change in the dependent variable.

In order for this type of analysis to be performed, it is necessary to work with a Cartesian layout. This means that the input data into the model will be arranged along the horizontal (X) and vertical (Y) axes, in order to produce an XY relationship graph as a byproduct.

The analysis that includes two directly influencable variables typically constitutes a model called Simple Linear Regression. In this model, the relationships are necessarily linear, meaning the predictions are either descending or ascending.

It is possible, however, in this type of analysis, to establish relationships between multiple variables. Models of this type are called Multiple Linear Regression, and they can handle more complex analyses, not only in a two-dimensional plane like the Cartesian plane, but also in three-dimensional planes. An example of this is the Decision Tree model, which, in addition to the X and Y variables, includes a third variable.

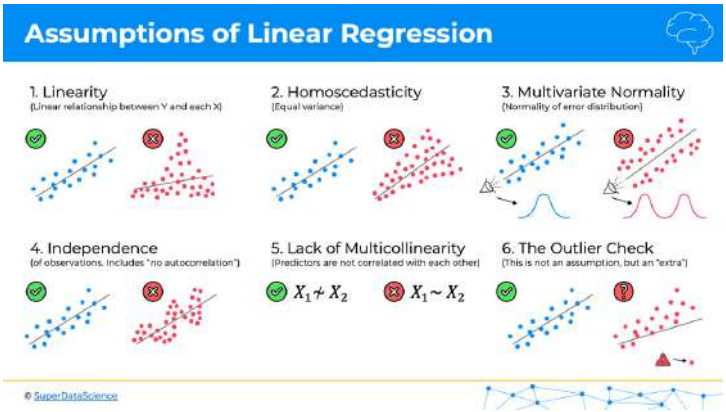

Figure 9: Issues of simple linear regression
Available at: https://dante.pro/regressao Accessed on 26 nov 2024

1.7.1 Linear model suppositions

However, when it comes to linear programming, it is important to pay attention to some assumptions made by the model itself. First, one must be aware of the linearity of the resulting graph. The line of relationships should remain in an average position between the data points, and should not be positioned above or below this range. If it does, the established relationship might be skewed by a specific data point, rather than reflecting the overall dataset.

Additionally, there should be a variety of homogeneous data, without fixed positioning on the graph or abnormalities in the relationship line. It is also necessary to identify any independent and isolated data points that could potentially skew the model, leading to inaccurate predictions.

1.8 Classification

Classification analysis models, also part of Machine Learning, as the name suggests, do not focus so much on correlating the input data, but rather on associating each data point with a classification within a predefined category.

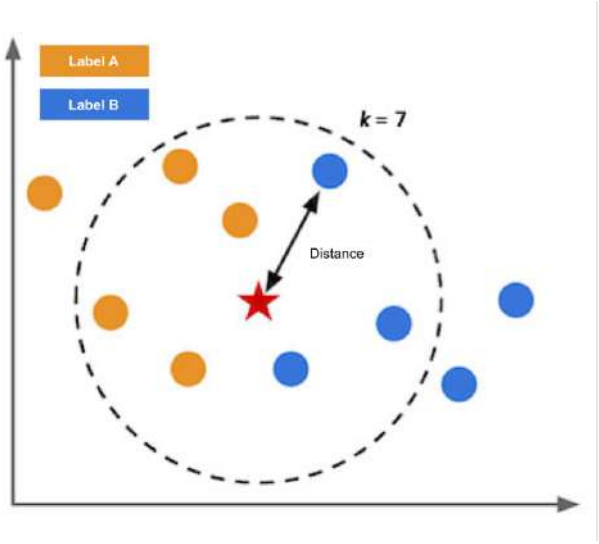

Figure 10: Example of the KNN technique
Available at: https://dante.pro/knnn Accessed on 26 nov 2024

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Figure 11: Formula used to calculate the Euclidean distance
Available at: https://dante.pro/euclidiana Acessed on 26 nov 2024

To achieve this, classification models share strong similarities with regression models, as they also rely on data input and the division of data into training and testing sets, as previously mentioned in regression models.

Based on this, what truly differentiates the two analyses is their objective. While regression models focus on linking two variables in order to obtain a pattern of relationships, classification models work by using data from a provided set, often with larger and more diverse raw data, to categorize the data into different classes.

One of the tools used in this model is the "K-Nearest Neighbors" (K-NN) algorithm (HO, LI, and SAYAMA, 2023). This tool operates based on the distance between each data point, or training sample, in order to classify the values (HARRISON, 2019).

This technique can be used in both two-dimensional or three-dimensional interfaces, taking into account the spatial distance between points in a more complex and selective manner (GÉRON, 2022).

To calculate this distance, the Euclidean distance system is often used. This type of measurement works by spacing two vectors at least in a two-dimensional plane, or it can also be applied to planes with more dimensions. To calculate it, it is necessary to sum the square root of the difference of 'x' (horizontal axis) and 'y' (vertical axis), always analyzing the correlation with its dimension.

With this, classification and regression models can be used for the composition of Machine Learning in Python, differing in terms of the analysis purpose.

There are other types of models that work with data classification. Among them are: Linear SVM, RBF SVM, Gaussian Process Decision Tree, Random Forest, Neural Net, AdaBoost, Naive Bayes, and QDA.

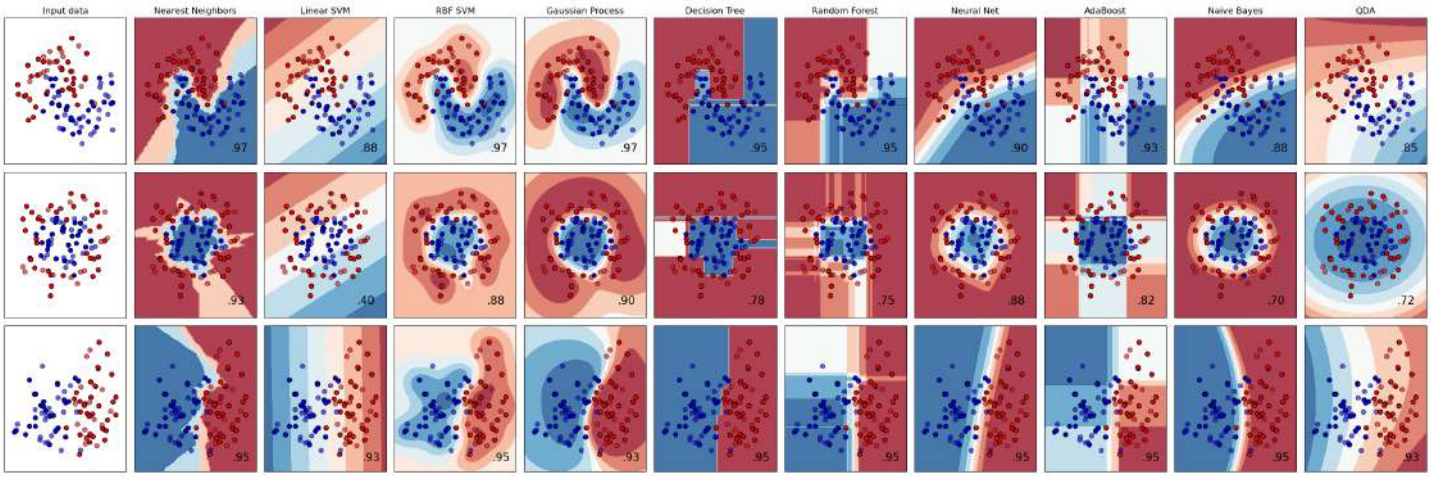

Figure 12: Classification models
Available at:https://dante.pro/scikit Acessed on 26 nov 2024

1.9 R² Equation

The R² Equation can be referred to as the 'Coefficient of Determination.' This formula calculates how well the model fits the relationships between the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2}$$

Figure 13: Statistical formula used for model accuracy
Available at: https://dante.pro/r2 Accessed on 26 nov 2024

Its understanding is simple, functioning through a mathematical value. The closer the result obtained by the model is to 1.0, the better the model fits the dataset. However, the opposite could indicate some flaw in the programming.

With this in mind, because it can determine the model's accuracy in the desired scenario, the use of the R² Equation or coefficient of determination is very important for the program to function properly.

This formula, in the project, will be used to evaluate the models worked on and present the best model for the diagnoses with blood tests, always checking the proximity to the maximum value.

1.10 Confusion Matrix

The confusion matrix is a tool widely used during the machine learning training phase to evaluate correct classifications and predicted classifications for each hypothesis raised, that is, for each output provided (MONARD, BARANAUSKAS, 2003). This method consists of creating a matrix, usually 2x2, involving the true positives, those that were originally positive and correctly classified, false positives, those negative instances incorrectly classified as positive, true negatives, those originally negative and correctly classified as negative, and false negatives, those originally positive but classified as negative by the model.

Thus, it is possible to quantify the cases predicted by the machine and perform the necessary analyses and adjustments for better performance. The ideal confusion matrix would have the second column (elements a12 and a22, with the classification aij) with zero values, indicating zero errors in the model and, therefore, extremely high accuracy and precision in real classification and predicted classification.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Figure 14: Model of confusion matrix 2x2 used for model evaluation
Available at: https://acesse.dev/wWJV5Access26nov2024

2 Problem

How to increase the efficiency of diagnoses for human diseases that can be detected through blood test exams?

3 Solution

To increase efficiency, it would be possible to develop a machine learning system based on artificial neural networks and deep learning, which would allow the analysis of a percentage of diseases, thus contributing to a faster and more effective

diagnosis. These systems (JANIESCH, 2021, LIU, 2020), through tests and examples, are capable of performing various tasks, analyses, and interpretations without human interference.

4 Methodology

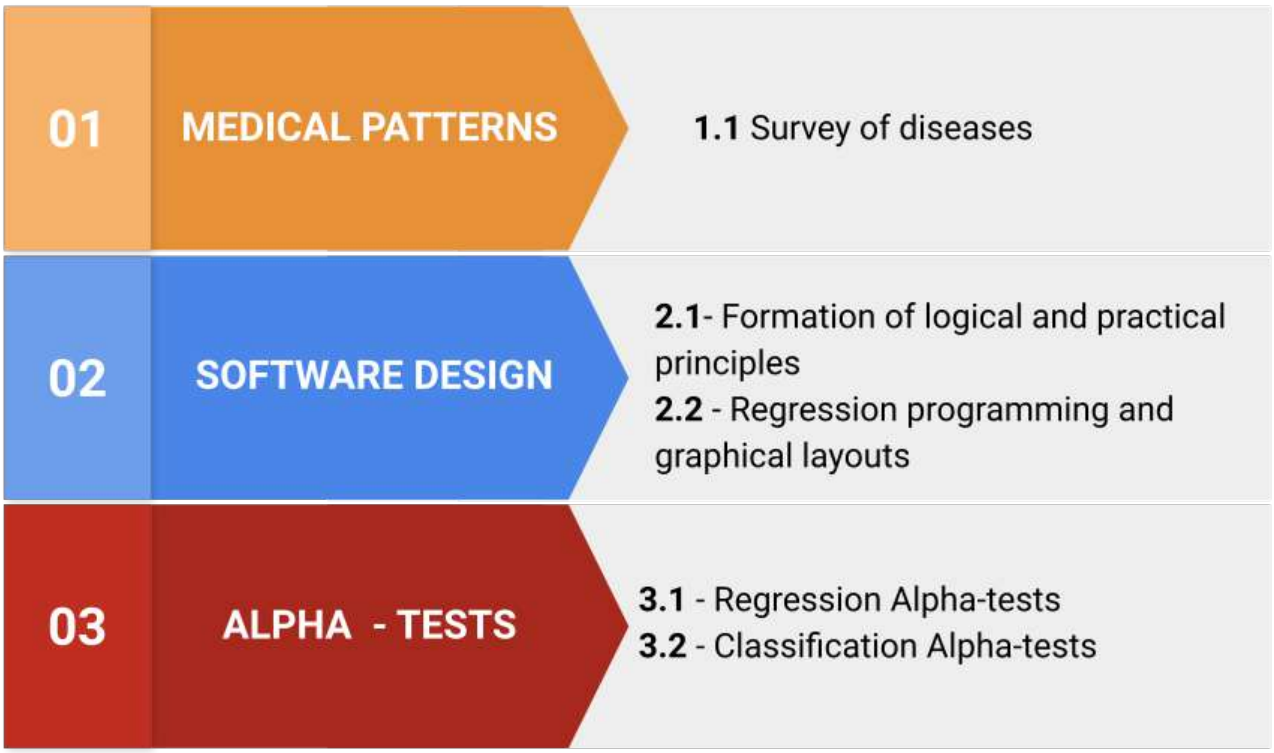

Figure 15: Methodology 1. Analysis of the medical patterns ; 2. Software design; 3. Alpha-tests. Source: Own autorship

The methodology for better efficiency was divided into 3 main stages: Medical Patterns, Software Design, and Alpha-Tests

In the first phase, Medical Patterns, a survey of variables – symptoms, characteristics, rarity, etc. – related to genetic diseases was conducted along with a survey of diseases that can be identified or minimally evidenced in blood tests.

In the second stage, Software Design, the development of theoretical foundations was enhanced through diagrams, networks, and representations, as well as the development of practical foundations, done in a similar way to the construction of various regression systems, which are very common in Python programming.

Finally, in the last stage, Alpha-Tests, two test datasets were developed: an initial one with a reduced dataset and a final one with a more extensive dataset. This phase of the project also involved testing with classification models, where we used a dataset of 2241 exams and classified the data with clustering and classification methods, performing a final evaluation of both stages.

5 Results

5.1 Results Phase 1 - Medical Patterns

In this stage of the project, we created a table where 13 diseases, namely: diabetes, anemia, leukemia, dengue, polycythemia, meningitis, chlamydia, tick-borne disease, inflammatory diseases (in general), tuberculosis, leprosy, schistosomiasis, and malaria, which can be identified in blood tests or whose test aids in identification, were detected. During the analysis of the table, the patterns of each column were categorized with the same color in order to make the visualization more dynamic.

Among them, factors such as hemoglobin, leukocytes, platelets, glucose, cholesterol, ions, and hormones were analyzed in order to find variables that would allow their identification among the others.

Doença	Hemoglobina	Leucócitos	Plaquetas	Glicose	Colesterol	Ions	Hormônios
Diabetes	Alterável	10000 + mm ³	140.000 a 440.000	126g +	- 130 mg/dl	136-145 mmol/l	Insulina: 126g +
Anemia	- 12 g/dL ou - 14g/dL	- 4.500/mm³	<150 000 / mm3	- 99 mg/dL	- 130 mg/dl	136-145 mmol/l	Eritropoético
Leucemia	- 12 g/dL	90 mil, 100 mil	- 100.000/mm³	70 a 99 mg/dl	- 130 mg/dL	136-145 mmol/l	-
Dengue	17+ g/dL	- 2000/ mm³	- 100.000/mm3	Relativamente alta (+ 100)	- 130 mg/dL	136-145 mmol/l	-
Policitemia	5.400000 + uL	10000 + mm³	440000 + variável	70 a 99 mg/dl	- 130 mg/dL	136-145 mmol/l	Eritropoietina alta
Meningite	13/17 g/dL	1.000/mm3.	140.000 a 440.000	≤ 18 mg/dL	- 130 mg/dl	136-145 mmol/l	-
Clamídia	17+ g/dL	10000 + mm³	440000 + variável	70 a 99 mg/dl	- 130 mg/dl	136-145 mmol/l	-
Doença do Carrapato	- 12 g/dL ou - 14g/dL	- 4500/ mm³	- 140 000 mm³	70 a 99 mg/dl	- 130 mg/dL	136-145 mmol/l	Hormônio do crescimento baixo
Doença inflamatória	- 15 g/dL variável	10000 + mm³	440000 + variável	Relativamente alta (+ 100)	- 130 mg/dL	136-145 mmol/l	Variação na insulina e hormonio tiroidiano
Tuberculose	- 12 g/dL ou - 14g/dL variável	10000 + mm³	440000 + variável	70 a 99 mg/dl	- 190 mg/dL + 40 mg/dL	136-145 mmol/l	-
Hanseníase	- 12 g/dL ou - 14g/dL variável	10000 + mm³ variável	440000 + variável	Relativamente alta (+ 100) variável	- 190 mg/dL + 40 mg/dL	136-145 mmol/l	Variação em hormônios endócrinos
Esquistossomose	- 12g/dL	10000 + mm³ variável	- 140 000 mm³	-70 mg/dL variável	- 40 mg/dL reduzido	- 136 mmol/l	-
Malária	- 12 g/dL ou - 14g/dL variável	10000 + mm³ variável	- 20 000 mm³	70 a 99 mg/dl	190 mg/dL + aumentado	- 136 mmol/l	-

Figure 16: Medical Patterns. Major diseases with blood count related factors

Thus, we found that regarding the elements that make up the blood, platelets showed the greatest variation among the types of diseases, with 20,000 mm^3/dL of blood in the case of Malaria and 440,000 mm^3/dL in cases of Polycythemia, Chlamydia, inflammatory diseases, Tuberculosis, and Leprosy. Additionally, Meningitis was the only disease that showed 18 mg/dL of glucose, which are possibly the factors that most differentiate one disease from the others.

5.2 Results Phase 2 - Software Design

5.2.1 Logical and Practical Principals (neural networks)

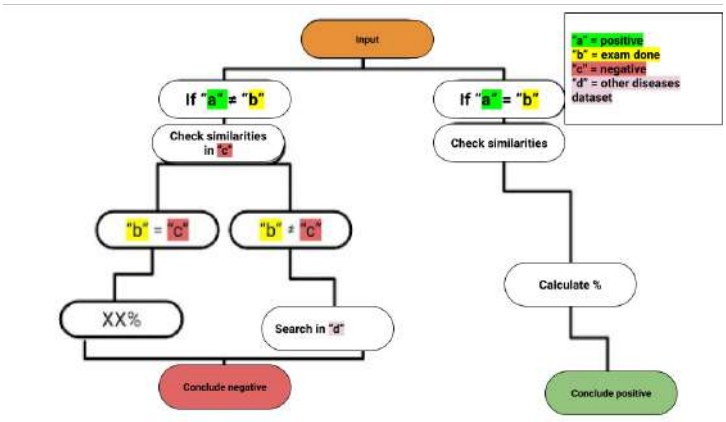

Figure 17: Artificial neural network developed to assist in the optimization of diagnosis through blood test exams. Logical principle

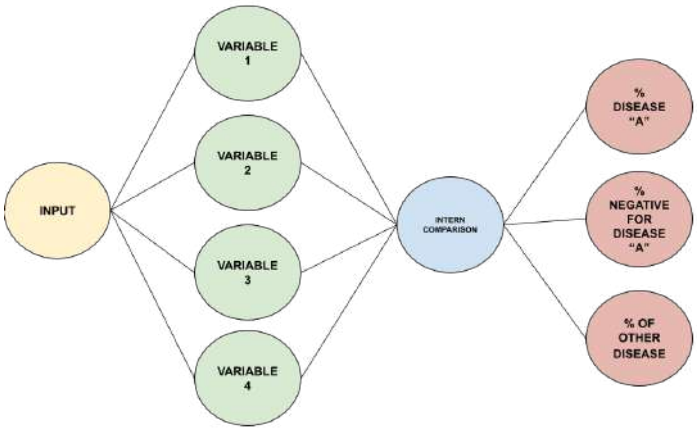

Figure 18: Artificial neural network developed to assist in the optimization of diagnosis through blood test exams. Practical principle

The first image (Figure 17) represents the logical principle of the software in the form of a flowchart. As the starting point ('input'), we have 'data input,' which generates two possible outputs: 'test result compatible with positive diagnosis' and 'test result incompatible with positive diagnosis.

In the case of the first output ('test result compatible with positive diagnosis'), the software will look for matches with negative test results. If the test result is compatible with the negative test, the software will calculate the percentage of likelihood for the disease and conclude it is negative. However, if the test result is incompatible with the negative test, the software will search for similarities in blood tests of other diseases and conclude it is negative for the analyzed disease.

Meanwhile, if the test result is compatible with the positive test for the disease (second output), the software will check for matches between the tests, calculate the percentage likelihood of the disease, and finally conclude it is positive for that condition.

The second image (Figure 9), in turn, represents the application of neural networks in the software, being developed based on the previous flowchart. It includes the 'input,' which is the data entry of the blood test, leading to comparisons between various internal layers that include: the number of red blood cells, white blood cells, platelets, glucose, and hormones. These values will then be compared with documented quantities and generate three final 'outputs': the percentage likelihood of presenting the analyzed disease, of not presenting the disease, and of presenting another disease.

5.2.2 Programming in regression

The second phase of the project consisted of the software design with the introduction of mechanisms linked to Python (programming language). Thus, several regression models were researched and developed, tools used for adapting various datasets to the programming, with different datasets and fictional scenarios to test and understand the different types of analyses and graphs generated by the program. Among the programmed and found models were: Data Pre-processing (Figure 11), Simple Linear Regression (Figure 12), Multiple Linear Regression (Figure 13), Polynomial Regression (Figure 14), SVR Regression (Figure 15), Decision Tree Regression (Figure 16), and Random Forest Regression (Figure 17).

Importing the libraries

```
[ ] import numpy as np
    import matplotlib.pyplot as plt
    import pandas as pd
```

Importing the dataset

```
[ ] dataset = pd.read_csv('Data.csv')
    X = dataset.iloc[:, :-1].values
    Y = dataset.iloc[:,-1].values
```

```
[ ] from google.colab import drive
    drive.mount('/content/drive')
```

```
[ ] print(X)
```

```
[['France' 44.0 72000.0]
 ['Spain' 27.0 48000.0]
 ['Germany' 30.0 54000.0]
 ['Spain' 38.0 61000.0]
 ['Germany' 40.0 nan]
 ['France' 35.0 58000.0]
 ['Spain' nan 52000.0]
 ['France' 48.0 79000.0]
 ['Germany' 50.0 83000.0]
 ['France' 37.0 67000.0]]
```

```
[ ] print(Y)
```

```
['No' 'Yes' 'No' 'No' 'Yes' 'Yes' 'No' 'Yes' 'No' 'Yes']
```

Figure 19: Data Pre-Processing regression model initially set up to verify the most important characteristics for programming a software.

The Data Pre-Processing regression model was set up to verify important factors for programming any software. As the name suggests, this model involved data processing in an initial stage, using key variables such as country names, numerical quantities, and 'yes' or 'no' responses, under the fictitious relationship of the number of people per country and the presence (yes/no) of a country in a company (for example).

With this in mind, the Pre-Processing model shows important characteristics for the Software Design, such as the division of the dataset and the processing of numerical and textual data.

Visualising the Training set results

```
plt.scatter(X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'orange')
plt.title('Salary and Experience (Training set)')
plt.xlabel('Years of Experience')
plt.ylabel('Salary Given')
plt.show()
```

Figure 20: Simple Linear Regression Model. (Dataset obtained from: 'Machine Learning A-Z')

After the Data Pre-Processing regression model, the second program created was the Linear Regression Model or Simple

Linear Regression Model. This tool was tested to verify possible analyses in a dataset with linear data, meaning data with a practically fixed pattern. Thus, two main variables were used to understand this model: Salary and Years of Experience of a fictional employee. From this, it can be observed, through the obtained graph, that within the dataset, some relationships between the two variables were identified that appeared to be linear, as indicated by the line and the red dots (Figure 12), something that can be very common in datasets with blood test results, as is the focus of the project.

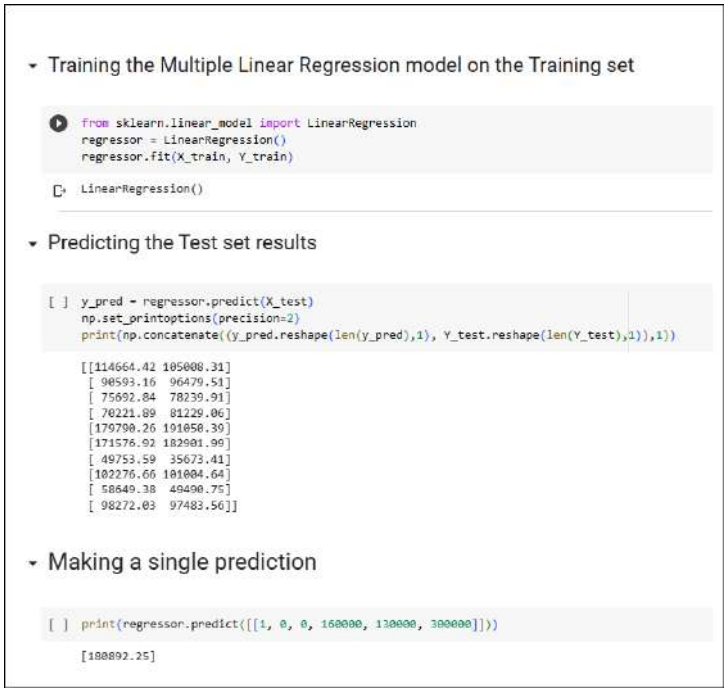

Figure 21: Modelo de regressão linear múltipla (Multiple Linear Regression). (Conjunto de dados obtido com: “Machine Learning A-Z”)

The Multiple Linear Regression Model, in turn, was developed with the aim of verifying a programming model that works with multiple linear variables, meaning variables that depend on each other. To test this model, a dataset with two main variables was used: Cost of Living and Places. Therefore, the program needed to relate the cost of living in each of the places in the dataset and predict which locations would be more or less expensive. Thus, it can be seen that through this model, it was possible to make predictions about different places, as indicated in the section 'Making a single prediction' (Figure 13).

Having this completed, the next four models worked with larger datasets that contained as main variables the positions of fictional employees in a company and their respective salaries. Thus, because they contained very similar datasets, the models presented very similar physical analyses, with well-designed graphs, but increasingly precise and adapted to the data:

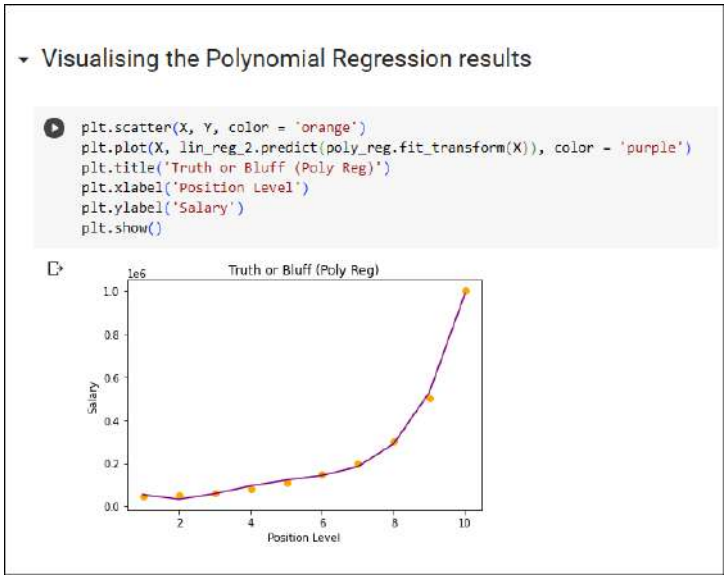

Figure 22: Polynomial Regression Model (Dataset obtained from: 'Machine Learning A-Z')

The first model tested with the new dataset was the Polynomial Regression model. As we can see in the graph, this type of program was able to relate the variables of salary and job position in a more homogeneous way, meaning with a more well-adapted analysis to the input data.

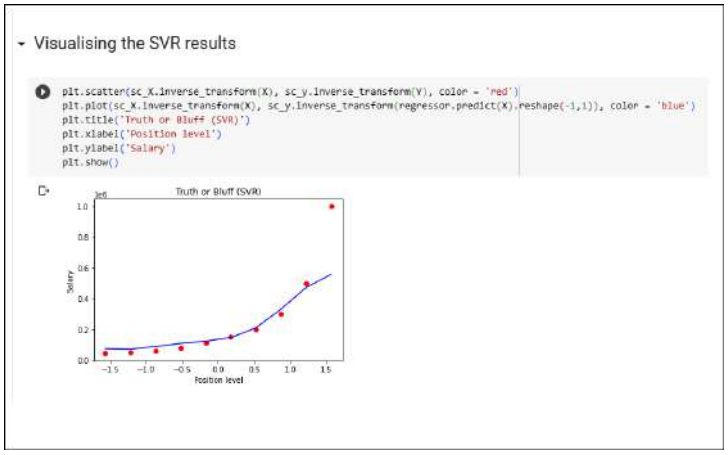

Figure 23: Support Vector Machine Regression (SVR) Model (Dataset obtained from: 'Machine Learning A-Z')

With this in mind, the Support Vector Regression (SVR) model worked with the same dataset as the previous model (Polynomial Regression), but with a different analysis tool. This program works by categorizing each of the data entries into two different areas and forming a line between them with similar factors. This line serves as a kind of pattern between the data, allowing them to be analyzed more accurately and used to make predictions. Therefore, it can be seen in Figure 13 that the SVR model was able to find the similarities in the model, creating a graph with the main predictions.

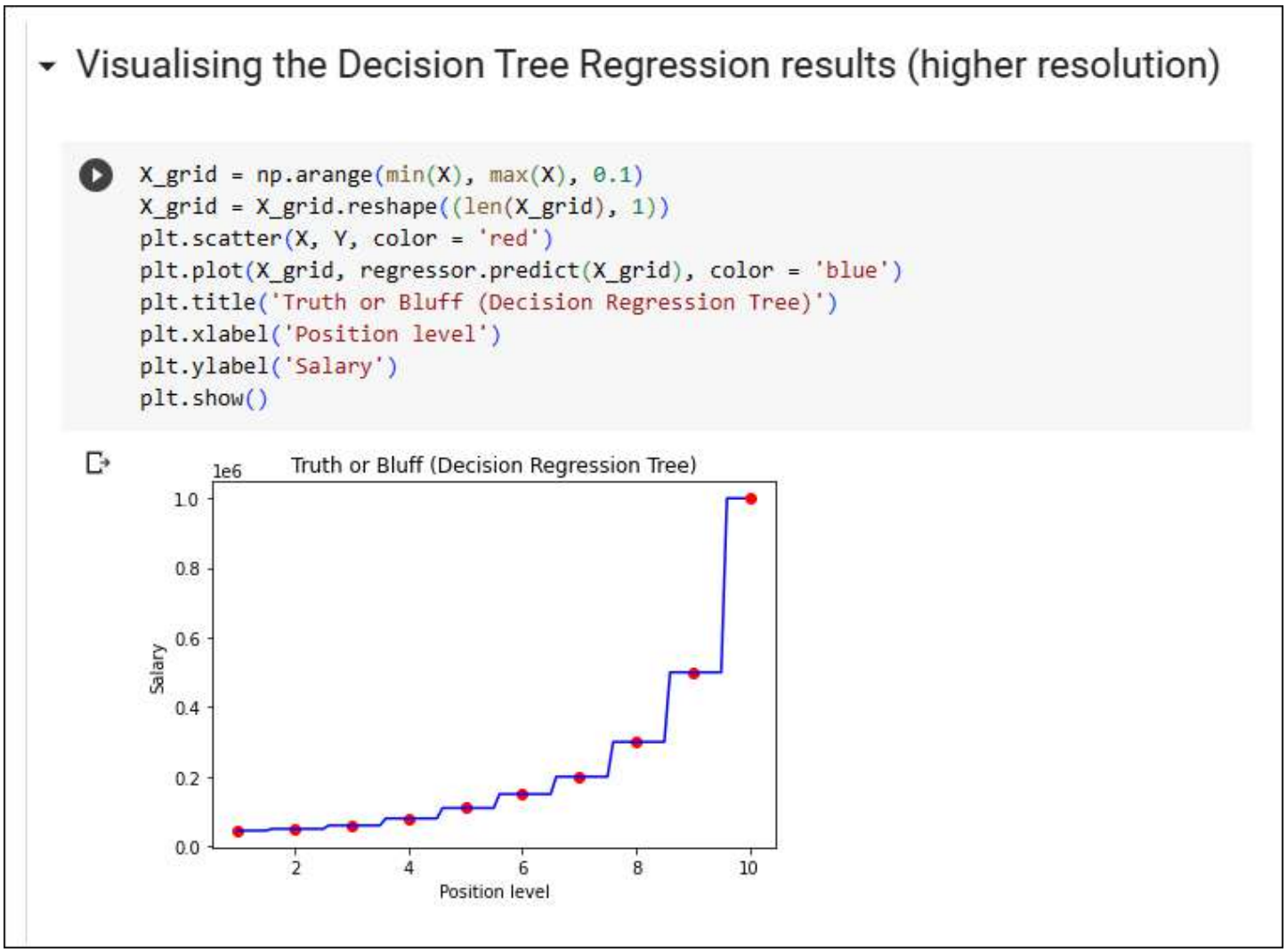

Figure 24: Decision Tree Regression (Dataset obtained from: 'Machine Learning A-Z')

Like the previous models, the Decision Tree Regression also used the variables of job position and salary, but, as expected, with a different type of construction. This program, as presented, works in parallel with the Linear Regression Model (Figure 12), meaning each data point has a linear analysis, functioning similarly to a margin of error. From this, the result can be visualized in the graph with straight lines at each red point (Figure 16).

Visualising the Random Forest Regression results (higher resolution)

```
[ ]  X_grid = np.arange(min(X), max(X), 0.01)
     X_grid = X_grid.reshape((len(X_grid), 1))
     plt.scatter(X, y, color = 'red')
     plt.plot(X_grid, regressor.predict(X_grid), color = 'blue')
     plt.title('Truth or Bluff (Random Forest Regression)')
     plt.xlabel('Position level')
     plt.ylabel('Salary')
     plt.show()
```

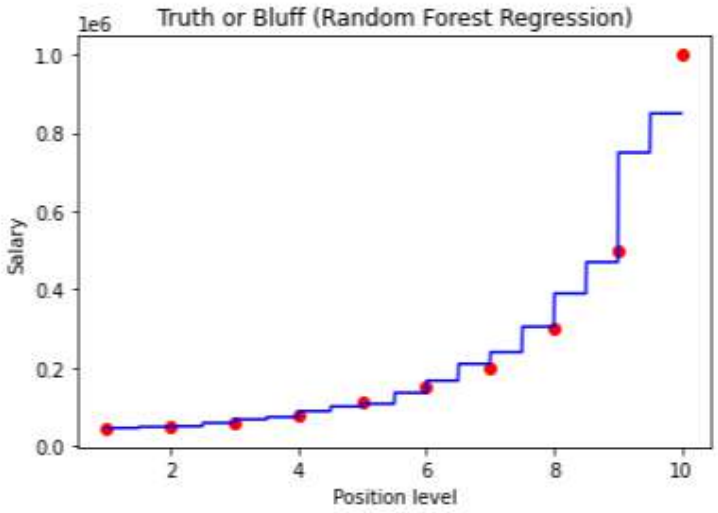

Figure 25: Random Forest Regression (Dataset obtained from: 'Machine Learning A-Z')

Finally, the last model found and tested was the Random Forest Regression, which, as the name suggests, functions as a maximizer of the Decision Tree Regression model (Figure 16), incorporating many of its functions. With this in mind, it can be observed in the last graph a pattern very similar to the previous one, with the presence of straight lines under each red point, indicating a similar analysis.

Based on this, it can be concluded that there is an analysis pattern among the models, making it possible to group the first three—Data Pre-Processing, Simple Linear Regression, and Multiple Linear Regression—into one group, and the last four—Polynomial Regression, SVR, Decision Tree Regression, and Random Forest Regression—into another. This division can assist in the future programming of the blood test dataset by making the analysis easier to understand.

5.3 Results Phase 3 - Alpha Tests

5.3.1 Alpha test with regression

During Phase 3, two tests were developed. The first included a reduced dataset that served as a model for a second system, this time with the final and larger dataset, containing 1,227 blood test records, about 40 of which were related to diseases or blood anomalies. After the tests, the models were evaluated using a specific formula (R^2) that shows the adaptability of the programming to the dataset.

ion	Glicose	Hemoglobina	WBC	Leucócitos	Doença X
899	89	5.5	3	6.3	1
325	57	2.2	4	4	0
656	98	3	10	23	1
777	12	2.2	5	12	0

Figure 26: Example of the reduced dataset used in Alpha Test 1. Including: ions, glucose, hemoglobin, wbc, leukocytes and Disease X, respectively.

Modelo	R^2
Random Forest	1.0
Polynomial	0.97
Support Vector	0.02
Multiple Linear	0.61
Decision Tree	1.0

Figure 27: Comparison table of the models' adaptability. Including: the models and their R^2 score

After the tests, we observed that two models achieved maximum adaptation to the dataset, 1.0, Random Forest Regression and Decision Tree Regression, among other models that had median adaptations, such as Multiple Linear Regression with 0.61, or poor adaptations, such as Support Vector Regression, which obtained only 0.02. The result from the first two models indicates that the regression models were able to adapt to the dataset and, possibly, make good predictions and analyses with it.

5.3.2 Classification Alpha - tests

In the second part of Phase 3, the alpha tests were actually adapted to the healthcare scenario. Thus, a simulation of a diagnosis for two diseases was executed
For this, Anemia and Leukemia were selected, diseases with specific modifications in blood components. The first disease, anemia, as mentioned earlier, corresponds to the most common hematological alterations in the global population (DOS SANTOS, 2024). It occurs with a decrease in hemoglobin and erythrocytes, which assist in the transport of oxygen to all body cells. Leukemia, on the other hand, is a disease that affects the production of erythroblasts and megakaryocytes, white blood cells involved in immune defense (PATERSON, 1952).

From this, using changes in these two components, with the binary system (1 for positive cases and 0 for negative cases), we verified the correlations between the blood test data and both diseases. Thus, we developed a classification program in Python, using the K-NN method.

First, we used the Cluster method to perform graphical analyses between two blood test variables. The result of these models typically classifies the data into the presence of typical space bubbles. In other words, the formation of bubbles indicates data differentiation, allowing them to be classified as positive or negative.

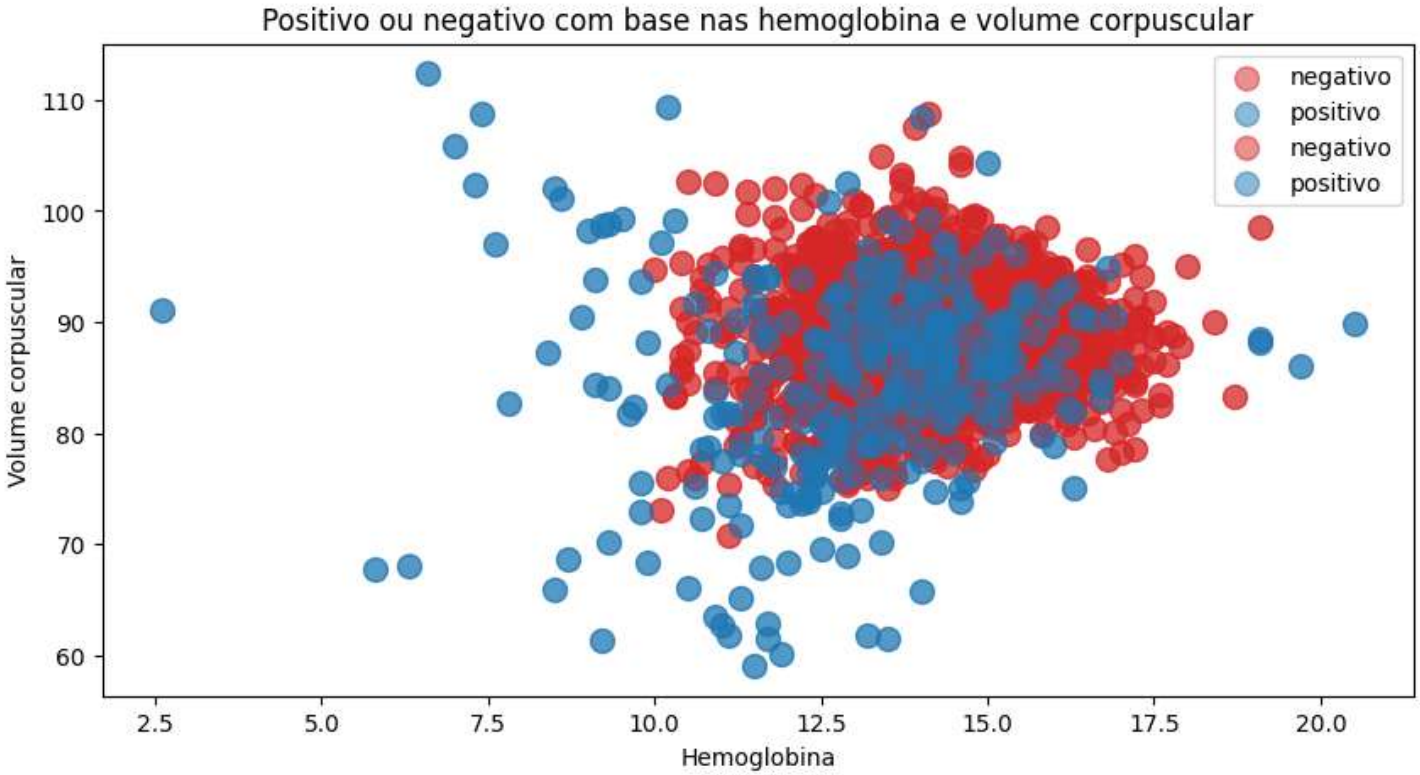

Figure 28: First Cluster analysis with the variables "White Blood Cells" and "Red Blood Cells". Red data means negative and blue data positive.

It can be observed that in the first case, the correlation between white blood cells and red blood cells did not show the typical bubble formation. There was a high overlap between red and blue bubbles. It should be noted that the blue points correspond to altered tests containing some diseases, and in this specific context, the divergent bubbles could indicate anemia, for example. Thus, the relationship between these two variables is not suitable for quick diagnosis.

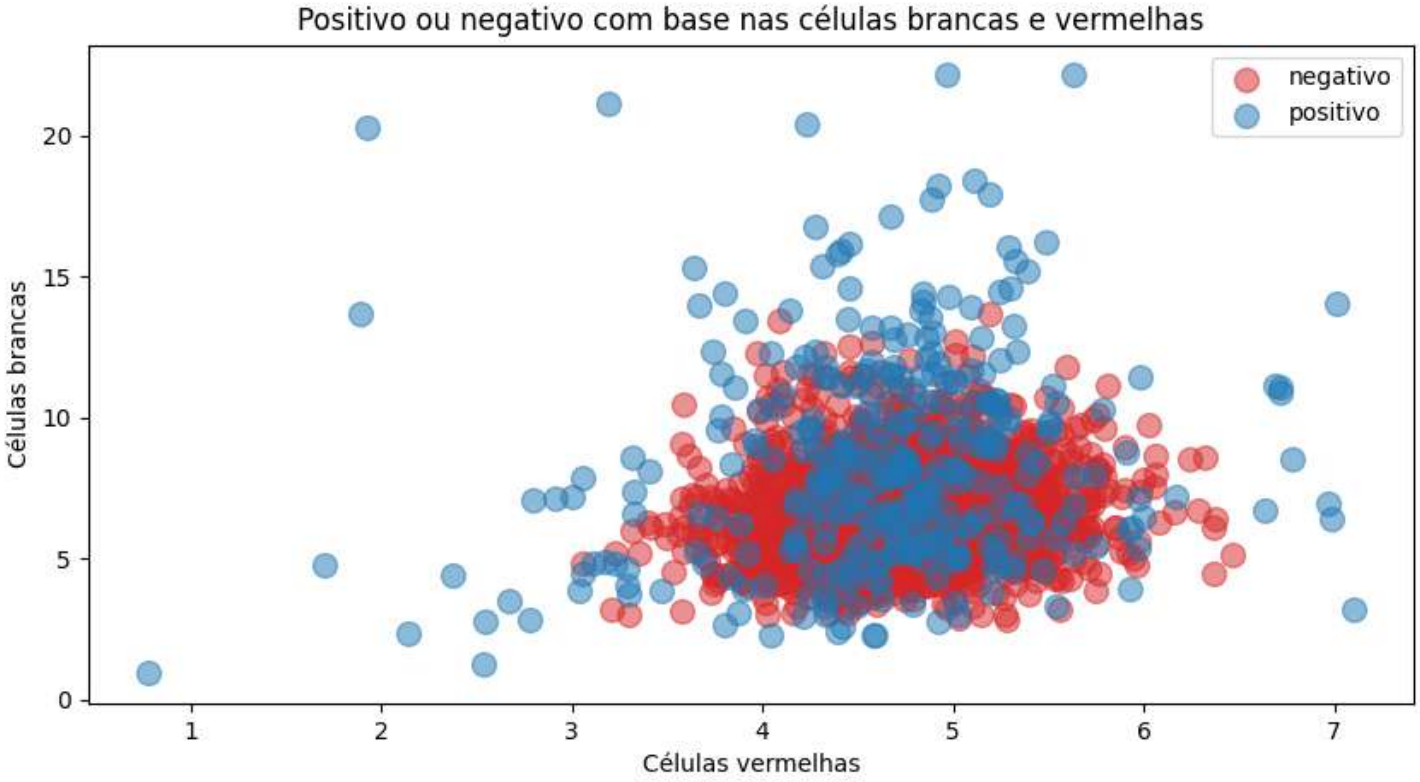

Figure 29: Second Cluster analysis with the variables "Hemoglobin" and "Corpuscular Volume". Red data means negative and blue data positive.

The second Cluster analysis also did not result in typical bubbles. In this experiment, the blue bubbles again indicate altered tests, which may show patterns with fewer than 3 red blood cells in the graph and more than 15 white blood cells.

Therefore, this correlation was also not useful for disease identification.

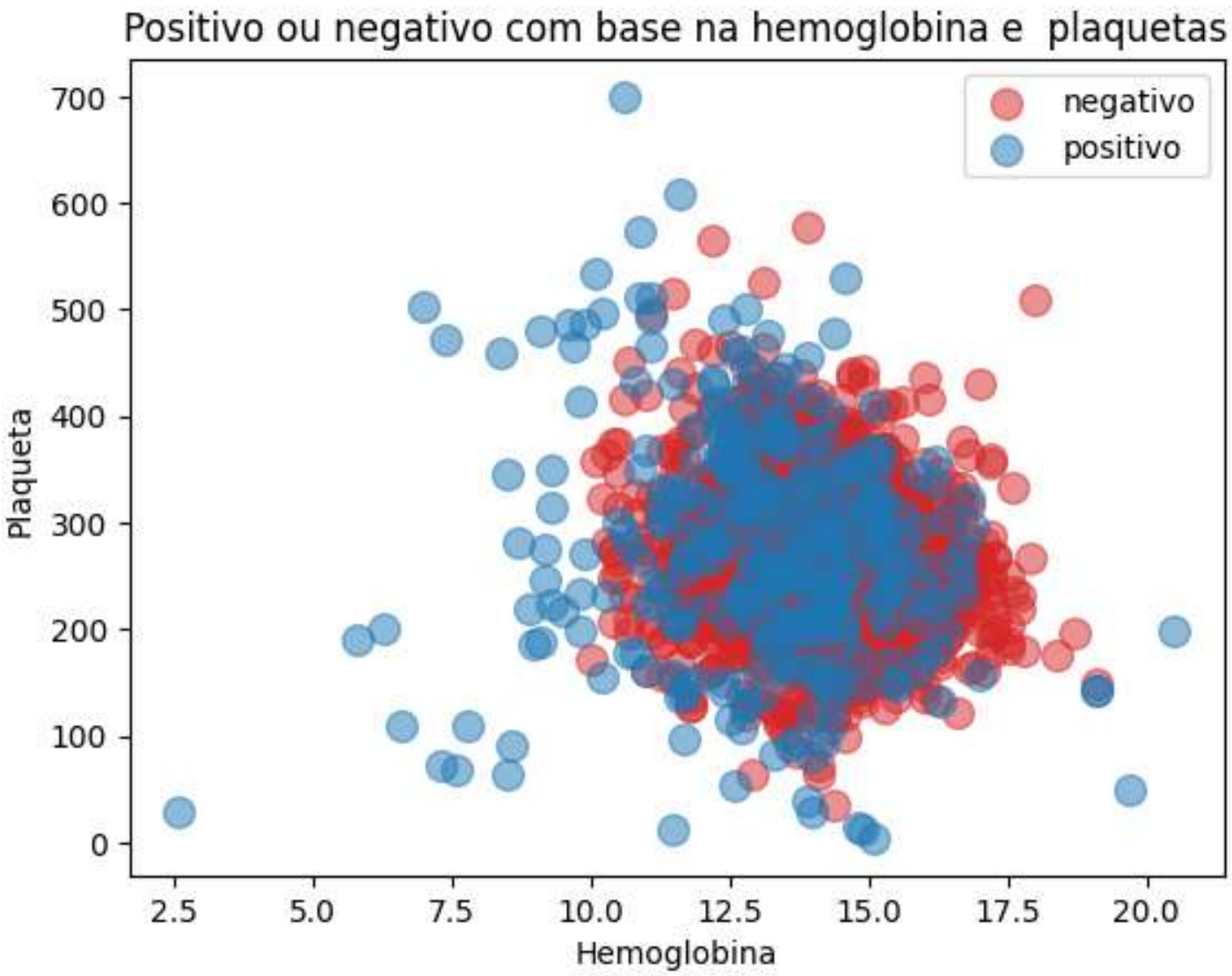

Figure 30: Third Cluster analysis with the variables "Hemoglobin" and "Platelets". Red data means negative and blue data positive.

The bubble formation in the third analysis was also not clear, marked by specific points in peripheral areas and a central mass of red and blue bubbles.

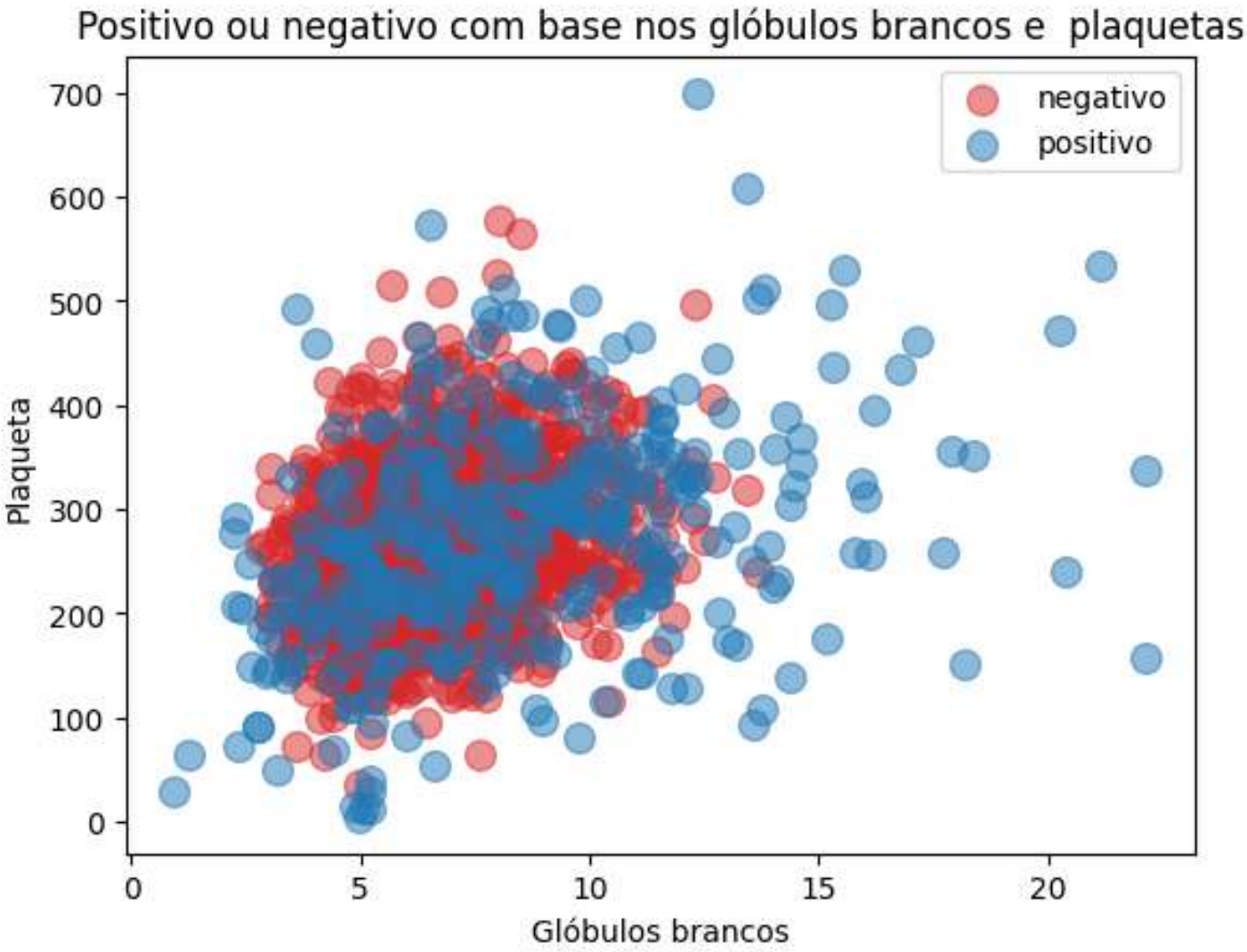

Figure 31: Fourth Cluster analysis with the variables "White Blood Cells" and "Platelets". Red data means negative and blue data positive.

The fourth relationship again did not show typical bubbles. In this case, there was a variation of blue altered bubbles on the right side of the graph, causing the overlap of bubbles to be limited to the left side.

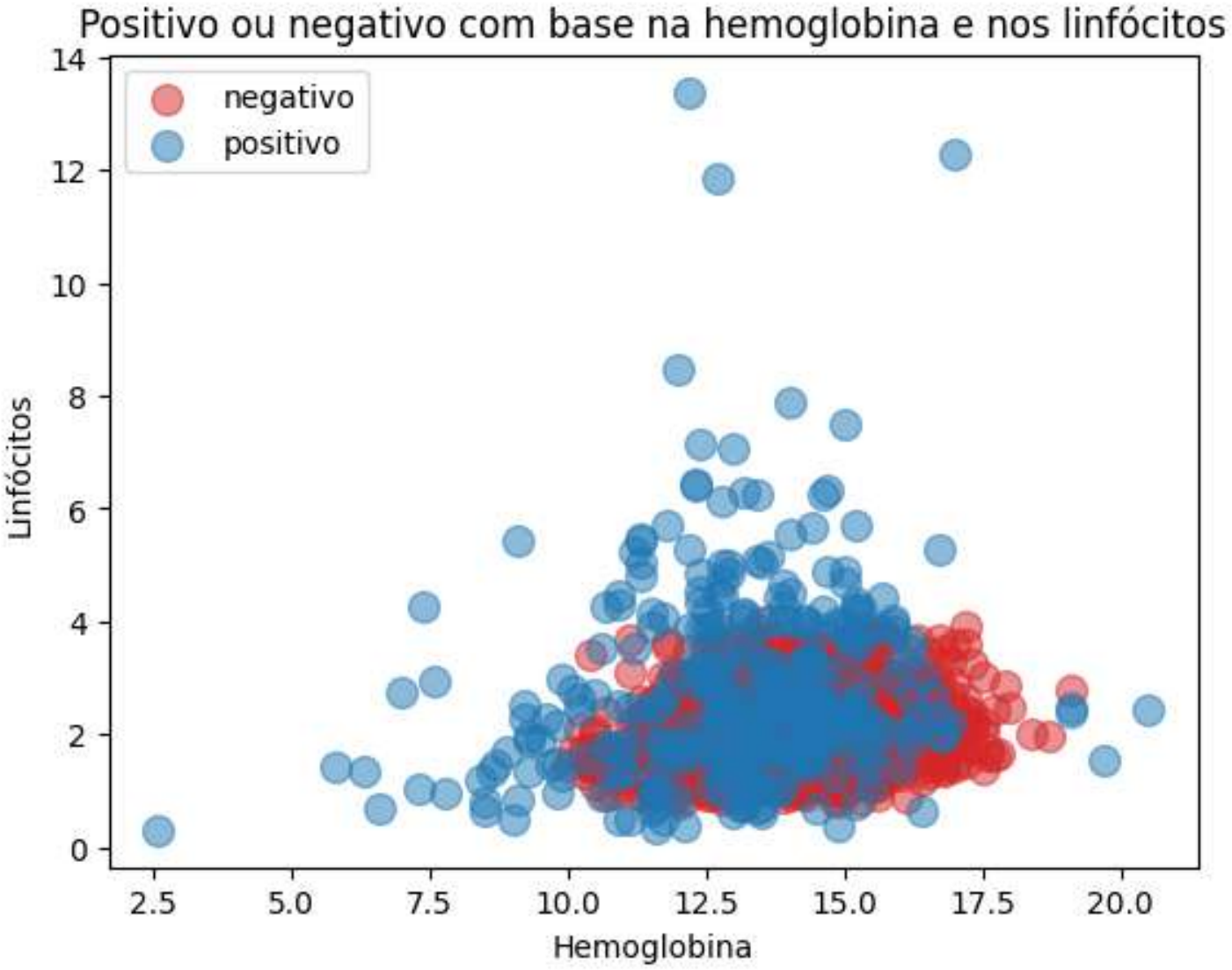

Figure 32: Fifth Cluster analysis with the variables "Hemoglobin" and "Lymphocytes". Red data means negative and blue data positive.

The fifth experiment consisted of the variables 'Hemoglobin' and 'Leukocytes'. Again, the formation of identifying bubbles was not consistent, with a concentration of blue bubbles in the upper part of the graph. These bubbles, being farther from the others, symbolize cases of leukemia and infection, for example.

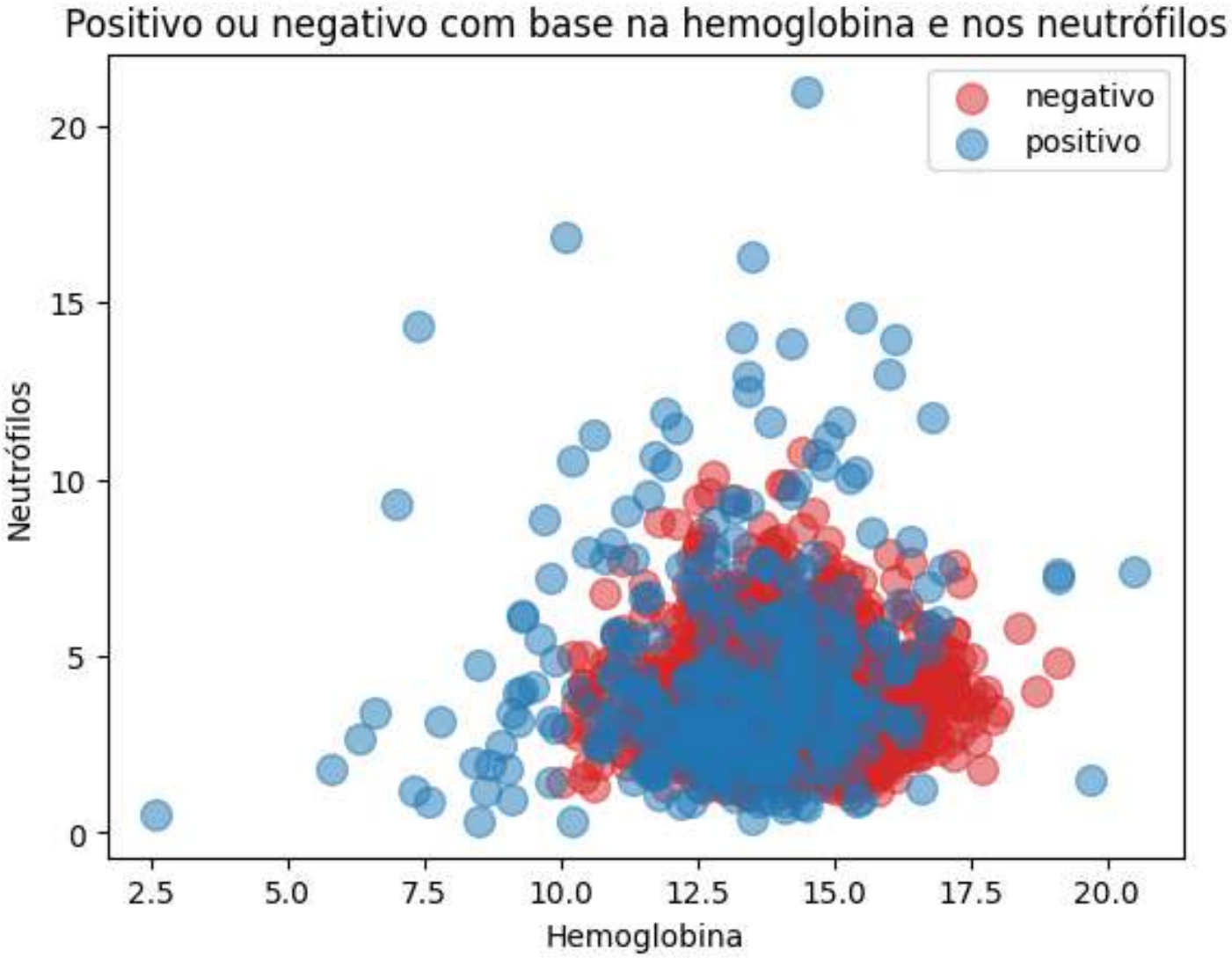

Figure 33: Sixth Cluster analysis with the variables "Hemoglobin" and "Neutrophils". Red data means negative and blue data positive.

The sixth correlation between 'Hemoglobin' and 'Neutrophils' did not form circles, showing diagonal concentrations of blue bubbles. This lateral concentration may indicate cases of bacterial infections within the dataset, helping to identify them among the others.

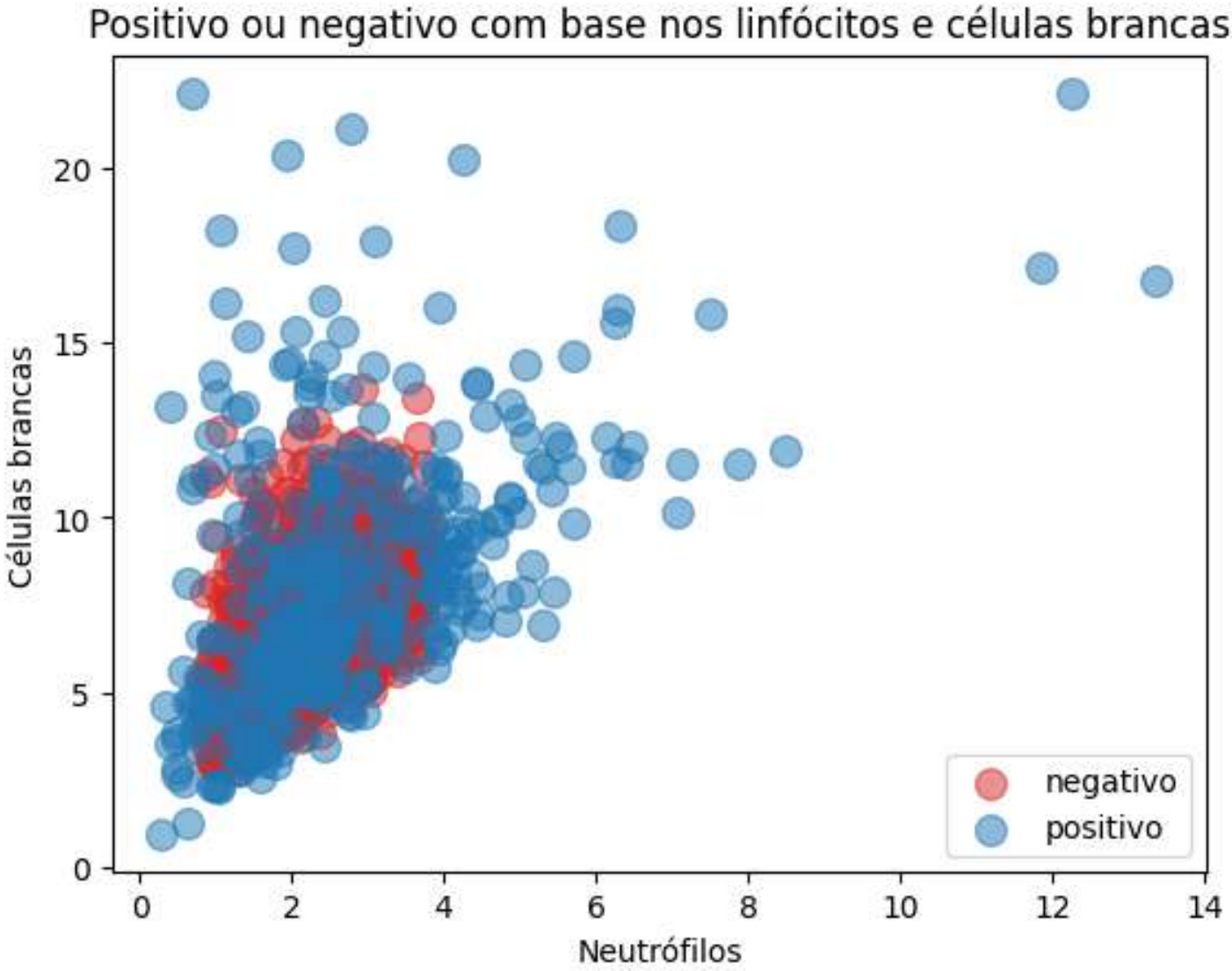

Figure 34: Seventh Cluster analysis with the variables "Lymphocytes" and "White Blood Cells". Red data means negative and blue data positive.

The seventh analysis used data on lymphocytes and white blood cells. This experiment showed a different morphology from the others, with a diagonal concentration of overlap and the remaining blue points scattered, indicating viral infections.

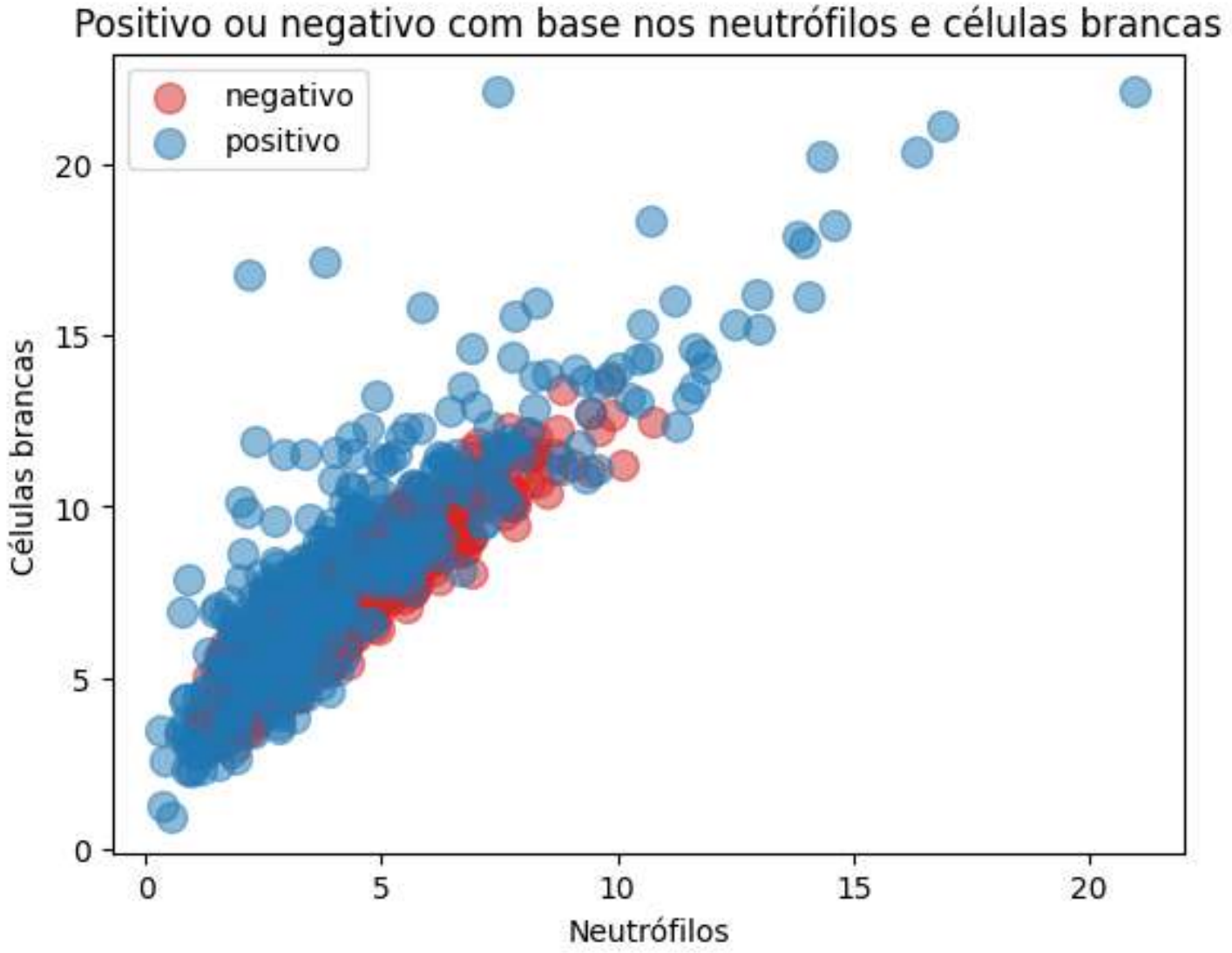

Figure 35: Eighth Cluster analysis with the variables "Neutrophils" and "White Blood Cells". Red data means negative and blue data positive.

In the eighth analysis, the red and blue points formed around a diagonal line, with the blue points above the line indicating a Leukopenia alteration.

After these analyses, we concluded that the formation of clusters and the identification of diseases for diagnoses using two variables is not easy to apply. Therefore, we decided to proceed with new analyses using the 'KNN' (K-Nearest Neighbors) tool.

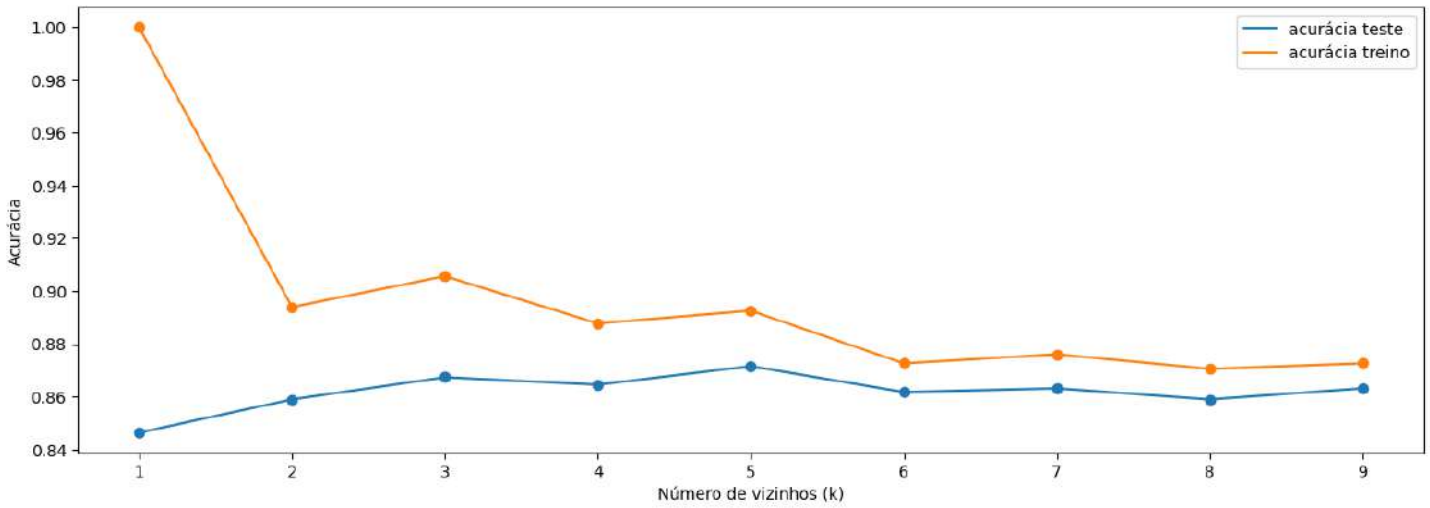

Figure 36: KNN comparing table, with 10 different euclidean distances. Orange line means training acuracy and blue line means test acuracy.

Defining ten KNN distances, we realized that the number with highest acuracy for the tests is number 5, that presented 0.87 or 87 percentage points of adaptability.

In this context, we created a confusion matrix to identify the origin of the data presented as "False Positives," meaning data that were negative but were classified as positive, "False Negatives," positive data classified as negative, "True Positives," positive data classified as positive, and "True Negatives," negative data that were correctly classified.

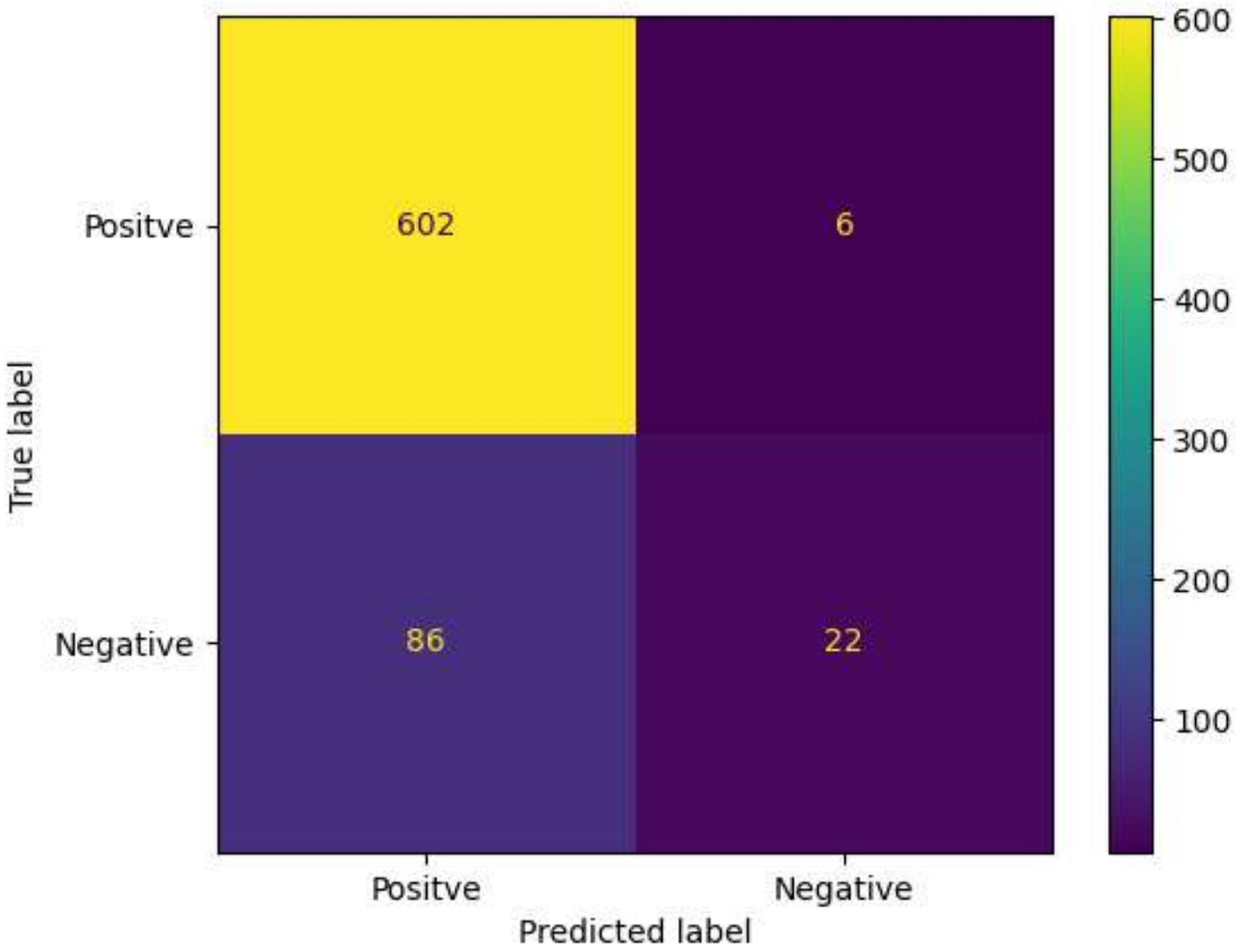

Figure 37: Confusion matrix indicating the fp, fn, tp and tn values.

We identified within the test group 602 positive tests classified as positive and 22 negative tests classified as negative. Meanwhile, 6 positive tests were classified as negative and 86 negative tests were classified as positive.

6 Conclusions

Given that the classification model showed, through the KNN comparison chart, the dataset's adaptability to the number of 5 KNN and an accuracy of 0.87, I conclude that my hypothesis—that it would be possible to create an artificial intelligence software that analyzes a percentage of diseases and promotes a faster and more extensive diagnosis—was supported. Since the KNN tool employs Euclidean distance between computed values, resulting in complex analyses and counting of correct and incorrect predictions, as seen with the model's high accuracy. In this way, it will be possible to further assist both the public and private healthcare systems in our country, contributing to disease prevention and saving many lives through its use.

Bibliography

ALCANTARA, R. L. *et al.* **A tecnologia de CRISPR-Cas9 na terapia gênica do câncer de pulmão**. Revista Brasileira Militar de Ciências, v. 5, n. 13, 2019.

ALMT. Comissão de Saúde da Assembleia Legislativa de Mato Grosso. Tempo médio de espera para atendimento pelo SUS é de um ano e 4 meses, diz levantamento de comissão da ALMT. TV Centro América, 11 mar. 2020. Disponível em: https://g1.globo.com/mt/mato-grosso/noticia/2020/03/11/tempo-medio-de-espera-para-atendimento-pelo-sus-e-de-um-ano-e-4-meses-diz-levantamento-de-comissao-da-almt.ghtml. Acesso em: 25 set. 2023.

ARAÚJO, C. N. M. *et al.* **Perfil de doenças onco-hematológicas, triadas por hemograma, em pacientes atendidos em unidades de pronto atendimento (UPA 24H)**. Hematology, Transfusion and Cell Therapy, Elsevier, v. 44, p. S546, 2022.

CHANG, H. Y. *et al.* **Artificial intelligence in pathology**. Journal of Pathology and Translational Medicine, The Korean Society of Pathologists and the Korean Society for Cytopathology, v. 53, n. 1, p. 1–12, 2019.

DE SOUSA, Márcio Morais. A CONFIANÇA NA TERRA DE NINGUÉM: Uma análise da aplicabilidade da boa-fé na internet. **PROJEÇÃO, DIREITO E SOCIEDADE**, v. 4, n. 2, p. 11-29, 2013.

DONG, H.; DONG, H.; DING, Z.; ZHANG, S.; CHANG. **Deep Reinforcement Learning**. Springer, 2020.

DUTRA, R. A. *et al.* **A importância do hemograma no diagnóstico precoce da leucemia**. Revista Eletrônica Acervo Saúde, v. 12, n. 7, p. e3529, 2020.

FAILACE, Renato. Hemograma: Manual de Interpretação. Porto Alegre: Artmed Editora, 2015.

FENG, J.-W.; WU, Y.-C. **Development and application of artificial neural network**. Wireless Personal Communications, Springer, v. 102, p. 1645–1656, 2018.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**. "O'Reilly Media, Inc.", 2022.

GOTARDO, R. **Linguagem de programação**. Rio de Janeiro: Seses, p. 34, 2015.

HARRISON, M. **Machine Learning–Guia de referência rápida: trabalhando com dados estruturados em Python**. Novatec Editora, 2019.

HASHIMOTO, D. A.; ROSMAN, G.; RUS, D.; MEIRELES, O. R. **Artificial intelligence in surgery: promises and perils**. Annals of Surgery, NIH Public Access, v. 268, n. 1, p. 70, 2018.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2001.

HO, K. K. W.; LI, N.; SAYAMA, K. C. **Equip public managers with data analytics skills: a proposal for the new generation of MPA/MPP programs with data science track**. Library Hi Tech, Emerald Publishing Limited, 2023.

Hospital Israelita Albert Einstein (HIAE). Vida Saudável. Disponível em: [URL do artigo específico]. Acesso em 11 de julho de 2023.

Institute for Health Metrics and Evaluation (IHME). Findings from the Global Burden of Disease Study 2017. Seattle, WA: IHME, 2018. Disponível em https://dante.pro/datahealth

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. **Machine learning and deep learning**. Electronic Markets, Springer, v. 31, n. 3, p. 685–695, 2021.

JOHNSON, K. W. *et al.* **Artificial intelligence in cardiology**. Journal of the American College of Cardiology, American College of Cardiology Foundation Washington DC, v. 71, n. 23, p. 2668–2679, 2018.

JUSZCZAK, P.; TAX, D.; DUIN, R. P. W. **Feature scaling in support vector data description**. In: *Proc. asci.* Citeseer, p. 95–102, 2002.

KIM, M. *et al.* **Elucidating the effects of curcumin against influenza using in silico and in vitro approaches**. Pharmaceuticals, MDPI, v. 14, n. 9, p. 880, 2021.

LIU, L. *et al.* **Deep learning for generic object detection: A survey**. International Journal of Computer Vision, Springer, v. 128, p. 261–318, 2020.

LYU, M. R. **Handbook of software reliability engineering**. IEEE computer society press Los Alamitos, v. 222, 1996.

MAYERICH, D.; SUN, R.; GUO, J. **Deep Learning**. In: **Microscope Image Processing**. Elsevier, 2023. p. 431–456.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32, 2003.

NASCIMENTO, A. G. *et al.* **Portal GenomaUSP: materiais didáticos para o ensino de Genética**. Genética na Escola, v. 17, n. 2, p. 237–238, 2022.

PEDREGOSA, F. *et al.* **Scikit-learn: Machine learning in Python**. The Journal of Machine Learning Research, JMLR. org, v. 12, p. 2825–2830, 2011.

RASCHKA, S.; PATTERSON, J.; NOLET, C. **Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence**. Information, MDPI, v. 11, n. 4, p. 193, 2020.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. Editora Manole Ltda, 2003.

RODRÍGUEZ, J. A. *et al.* **De Cancerología**, 2013.

ROSENFELD, R. **Hemograma**. Jornal Brasileiro de Patologia e Medicina Laboratorial, SciELO Brasil, v. 48, p. 244, 2012.

RUDIN, C. *et al.* **Interpretable machine learning: Fundamental principles and 10 grand challenges**. Statistic Surveys, The American Statistical Association, the Bernoulli Society, the Institute, v. 16, p. 1–85, 2022.

SARKER, I. H. **Machine learning: Algorithms, real-world applications and research directions**. SN Computer Science, Springer, v. 2, n. 3, p. 160, 2021.

SU, X.; YAN, X.; TSAI, C.-L. **Linear regression**. Wiley Interdisciplinary Reviews: Computational Statistics, Wiley Online Library, v. 4, n. 3, p. 275–294, 2012.

SUGANYADEV I. S.; SEETHALAKSHMI, V.; MANIKANDAN, S. **Brain tumor segmentation using deep learning algorithms**. Journal of Ambient Intelligence and Humanized Computing, Springer, v. 12, p. 14159–14171, 2021.

THIEBES, S.; LINS, S.; SUNYAEV, A. **Trustworthy artificial intelligence**. Electronic Markets, Springer, v. 31, p. 447–464, 2021.

WU, Y.-c.; FENG, J.-w. *Development and application of artificial neural network.* Wireless Personal Communications, **Springer**, v. 102, p. 1645-1656, 2018.

XIONG, A.-S. *et al.* **Chemical gene synthesis: strategies, softwares, error corrections, and applications**. FEMS Microbiology Reviews, Federation of European Microbiological Societies, v. 32, n. 3, p. 522–540, 2008.

【評語】190032

The predicted medical diagnosis is very useful to help doctors and patients to get an earlier information efficiently. This work can be much better improved by providing symptoms and medical history, vital signs at the same time with the lab data to get an accurate predicted medical diagnosis with acceptable results that can really help patients.

More electronic medical records should be attained to make the job well done.