Bachelor Thesis

Classifying Vascular Age based on Brazilian Reference Values: A Bayes' Theorem and Least Squares approach

Rodrigo Santos da Silva

Advisors:

Prof. PhD. Thiago Damasceno Cordeiro
Prof. PhD. Glauco Estácio Gonçalves

Maceió, May 2023

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário: Jone Sidney A. de Oliveira – CRB-4 –1485

S586c Silva, Rodrigo Santos da.

Classifying vascular age based on brazilian reference values a bayes' theorem and least squares approach / Rodrigo Santos da Silva. - Maceió: AL, 2023.

39 f.: il.

Orientador: Thiago Damasceno Cordeiro

Coorientador: Glauco Estácio Gonçalves

Monografia (Trabalho de conclusão de curso em Sistemas de Informação) – Universidade Aberta do Brasil. Universidade Federal de Alagoas, Instituto de Computação. Maceió: AL. 2022.

Inclui bibliografia: f.28.

1. Mobil-O-Graph. 2. Teorema de Bayes. 3. Pressão Sistólica Central. I. Título.

CDU:004

Rodrigo Santos da Silva

Classifying Vascular Age based on Brazilian Reference Values: A Bayes' Theorem and Least Squares approach

Thesis presented as a partial requirement for obtaining a Bachelor's degree in Computer Science from the Computing Institute at the Federal University of Alagoas.

Advisors:

Prof. PhD. Thiago Damasceno Cordeiro

Prof. PhD. Glauco Estácio Gonçalves

Maceió, May 2023

Acknowledgments

I would like to express my gratitude to Prof Thiago Cordeiro for all the support, not only in this thesis but in all my graduation time, for allowing me to become a better professional, a better researcher and showing me the kind of professor that I want to be. Special thanks to Prof. Glauco Estácio, who helped me with guidance and knowledge. Additionally, I would like to express my gratitude to Prof. Aydano Machado and Prof. Álvaro Sobrinho for generously dedicating their time and participating in the chair.

I would like to thank my family, in special my parents Ivone and Roberval, my stepmother Marlene, and my siblings, Vitória and Lorenzo, for always being there for me and supporting my studies. Thank you for showing me the importance of education.

Furthermore, I want to recognize my classmates and friends, whom I had the honor to study with and who were with me during all the tough times. A special thanks to Márcio Henrique, Michael, João Falcão, and Pedro Henrique for being by my side since we were in middle school, at IFAL, and now in college.

There are no words to represent how Fernanda was important in this whole process, as a wonderful partner, friend, and coworker. Thank you for always being there for me. Meeting you in the college was a great gift.

Also, a special thanks to all my friends from IFAL that have been by my side, even though we are not studying together anymore. A special thanks to Aguida, Clara, Gabriel, Haniel, Juliane, Karen, Luísa and Henrique.

Lastly, a special thanks to my friends Bruno, Gabriela, Roger, Vívian, Sophia, Arthur, José Neto, Lucas, Lilian, João Pedro, Gabriel, João Ayalla, Rafael, Laryssa, and Lucas Lisboa for the pleasure of working, have fun, and learning with you all.

And thank you, dear reader, for having an interest in this work.

“If you only do what you can do you will never be more than you are now”

– Master Shifu

Resumo

Doenças cardiovasculares, de acordo com a (Organização Mundial da Saúde) OMS, são as principais responsáveis pelos casos de morte na última década. Como uma prova disso, em 2019 um total de 17.9 milhões foram em decorrência de problemas cardiovasculares, o que representa 32% das mortes no mundo. Além disso, tais doenças tem um papel crucial em casos de morte nas populações de países de baixa e média renda. Assim, um dos métodos aplicados para avaliar as condições de cardiovasculares de um indivíduo é o cálculo da idade cardiovascular. Hoje, existem técnicas que realizam esse cálculo, como o dispositivo Mobil-O-Graph que consegue calcular a idade cardiovascular com base em 5 variáveis hemodinâmicas (as mesmas usadas nesse trabalho). Porém, tal dispositivo foi desenvolvido e calibrado utilizando populações europeias, que possuem estilos de vida diferentes de outras populações, o que torna difícil a identificação da idade vascular de indivíduos brasileiros, por exemplo. Assim, mediante ao estudo que levanta valores de referência para a população brasileira, torna-se possível desenvolver métodos de classificação de idade cardiovascular para outras populações, como o caso da população brasileira nesse trabalho. Assim, esse estudo tem como objetivo desenvolver um método de classificação de idade cardiovascular, utilizando o método do Teorema de Bayes, baseado na distribuição da população brasileira de 2010, via a não publicação de dados novos em decorrência da pandemia da Covid-19, de acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE), para realizar os cálculos de probabilidade da idade cardiovascular de um indivíduo. Além disso, utiliza-se o método de mínimos quadrados para minizar o erro entre os dados previstos e as funções acumulativas calculadas com base nos valores de referência. Os resultados mostram que foi possível calcular classes de idades cardiovasculares com base nos valores das variáveis hemodinâmicas informadas. Além disso, foi desenvolvida uma aplicação web com o intuito de testar a técnica, utilizando dados de pacientes reais.

Keywords: Mobil-O-Graph, Teorema de Bayes, Naive Bayes, Mínimos Quadrados, Velocidade da Onda de Pulso, Índice de Aumentação, Pressão Diastólica Central, Pressão Sistólica Central, Pressão de Pulso.

Abstract

Cardiovascular diseases, according to the World Health Organization (WHO), have been the leading cause of death in the last decade. As evidence of this, in 2019 a total of 17.9 million deaths were due to cardiovascular problems, representing 32% of global deaths. Furthermore, such diseases play a crucial role in mortality cases among populations in low- and middle-income countries. Therefore, one of the methods applied to assess an individual's cardiovascular conditions is the calculation of cardiovascular age. Nowadays, there are techniques that perform this calculation, such as the Mobil-O-Graph device, which can calculate cardiovascular age based on 5 hemodynamic variables (the same variables used in this study). However, this device was developed and calibrated using European populations, which have different lifestyles compared to other populations, making it difficult to identify the vascular age of individuals from, for example, Brazil. Thus, through a study that establishes reference values for the Brazilian population, it becomes possible to develop methods for classifying cardiovascular age for other populations, such as the Brazilian population in this study. Therefore, the objective of this study is to develop a method for classifying cardiovascular age using Bayes' Theorem, based on the distribution of the Brazilian population in 2010, due to the non-publication of new data as a result of the Covid-19 pandemic, according to the Brazilian Institute of Geography and Statistics (IBGE), to calculate the probability of an individual's cardiovascular age. In addition, the least squares method is used to minimize the error between the predicted data and the cumulative functions calculated based on reference values. The results show that it was possible to calculate cardiovascular age classes based on the provided hemodynamic variable values. Furthermore, a web application was developed to test the technique using data from real patients.

Keywords: Bayes' Theorem, Naive Bayes, Least Squares, Pulse Wave Velocity, Augmentation Index, Central Diastolic Pressure, Central Systolic Pressure, Pulse Pressure.

Contents

List of Abbreviations and Acronyms

PWV	*Pulse Wave Velocity*
AI	*Augmentation Index*
CSP	*Central Systolic Pressure*
CDP	*Central Diastolic Pressure*
PP	*Pulse Pressure*
API	*Pulse Pressure*
CDFs	*Cumulative distribution functions*
CORS	*Cross-Origin Resource Sharing*

List of Figures

1 Introduction

1.1 Motivation

Cardiovascular diseases have played a crucial role in human mortality. According to the WHO (World Health Organization), it has been the leading cause of death worldwide in the last two decades. Just in 2019, a total of 17.9 million deaths in the global population was estimated, which represents 32% of all deaths worldwide, and 85% of these were due to heart attacks and strokes [1]. Furthermore, it is a fact that 75% of cardiovascular diseases occur in populations of low and middle-income countries (BOWRY et al., 2015).

Among the vast array of cardiovascular diseases, this work focuses on arterial stiffness, a disease in which stiffening and reduction in the elasticity of arterial walls occur. This disease can impact how arteries regulate themselves according to different levels of pressure in blood flow, which can lead to problems such as arteriosclerosis, increased chances of arterial obstructions, stroke, heart attack, and others (O'ROURKE; MANCIA, 1999).

This disease can be derived from various factors, such as chronological age, an unhealthy lifestyle (bad diet, smoking, alcoholism), or other diseases like diabetes and hypertension. However, the problem arises in cases where arterial stiffness is at a headmost level than expected for their chronological age, as the process of arterial stiffening, due to advanced age is a natural occurrence.

In this sense, the computer science area, which consists in the field that has the computing area as focus, can develop the most various tools, such as a tool for assisting humanity and improving quality of life, and can provide various solutions aimed at heart diseases, and arterial stiffness conditions in this case, through collaboration with cardiology and other areas related to the cardiovascular system. Thus, through this aspiration, one of the many contributions of computer science, in partnership with medicine, was the development of the Mobil-O-Graph, a

[1] <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>

device capable of detecting, based on representative indices of the health of the cardiovascular system, the patient's vascular age. The device can perform a pulse wave analysis with 8 hemodynamic parameters, i.e., it is a device that provides an analysis of the state of blood vessels and their functioning without being invasive. In addition, the device has a step-by-step expansion of MAPA (Ambulatory Blood Pressure Monitoring) measurement, as well as ease of use and device integration[2].

However, even though this device brings great benefits to civil society, it was developed and configured according to the indices of the European population, with arterial stiffening being a problem affecting all populations. Thus, through the work (PAIVA et al., 2020), it was possible to map reference values for the Brazilian population.

1.2 Problem

Although the mentioned work brings results that may be useful for humanity and solve the problem of classifying vascular age, there was no presentation of how the inference process is done, making it difficult for various health professionals to realize their medical diagnoses.

1.3 Objective

Develop a classification method that aims to find the vascular age of Brazilian patients, with or without risk factors, using the variables augmentation index, pulse wave velocity, central systolic pressure, central diastolic pressure, and pulse pressure, according to the projection of its population presented in the reference study, using Bayes' Theorem and Least Squares due to the structure of the reference values in the reference work (PAIVA et al., 2020).

To achieve the overall goal of this work, the following specific objectives were considered:

1. Analyze the data distributions of the reference values in work (PAIVA et al., 2020).

2. Apply the concepts of Bayes' Theorem to solve the classification problem for vascular ages.

3. Apply least squares to minimize the error between the model and the actual values.

4. Develop an interface for visualization of the estimates.

1.4 Thesis Structure

The thesis consists of the following structure:

[2] <https://www.iem.de/en_US/mobil-o-graph>

- Introduction: it presents the motivation to develop this work, such as the problem and objectives.

- Background: it intends to present different contents for the understanding of the proposed solution.

- Methodology: it presents how all the contents shown in the background can be put together to develop the solution.

- Results: results that the developed methodology obtained.

- Final considerations: Revision of the developed work and future objectives.

- References: references utilized in this work.

2

Background

This chapter will address the main points and elements of the literature used to develop the work. Among them, we have Bayes' theorem, Naive Bayes, and Least Squares. Furthermore, we present the hemodynamic variables used to calculate vascular age, such as Pulse Wave Velocity (PWV), Augmentation Index (AIx), Central Systolic Pressure (CSP), Central Diastolic Pressure (CDP), and Pulse Pressure (PP). Each one of these variables has reference values in this order: fifth (also known as the median), tenth, twenty-fifth, seventy-fifth, and ninetieth percentiles.

2.1 Bayes' Theorem

The Bayes' Theorem, named in honor of its creator, statistician, and philosopher Thomas Bayes, is a famous theorem in the field of statistics, which involves the calculation of a particular probabilistic event, given that a series of occurrences, or conditions, related to the event have occurred previously. This theorem has played a role in various areas and it has even developed other ways of analyzing problems, such as Bayesian Logic, where the individual thinking about the issue also influences the probability of an event occurring, and the Bayesian Inference Process, one of the ways of performing statistical inference.

Such theorem can be expressed in the following form:

$$P(y|x_1,\ldots,x_n) = P(y)\frac{P(x_1,\ldots,x_n|y)}{P(x_1,\ldots,x_n)} \tag{1}$$

where y corresponds to the event that we are analyzing, $(x_1,\ldots,x_n)$ to a series of conditional events that occurred before y, $P(y)$ corresponds to the probability of the event y occurs, $P(x_1,\ldots,x_n)$ to the probability of the previous events have occurred, and $P(x_1, \ldots, x_n|y)$ corresponds to the probability of y will occur given $(x_1,\ldots,x_n)$ have already occurred.

To illustrate this theorem, we will use an example that was presented in a column of the BBC, where they mention the relevance of Bayes' Theorem in various areas in the current world, such as computing, health, economics, and other factors, dealing precisely with scenarios where there is a particular subjectivity of a probability [3]. In this example, a sporadic disease is considered, where one in 10 thousand suffers (this is also known as previous probability). The examination for detection of this disease is highly accurate, capable of correctly identifying the problem in 99% of cases, as well as in cases where the individual does not have the condition.

Thus, the test was taken in a population of 1 million individuals. In this scenario, in the group of people that have the disease, which is one hundred, 99 will be diagnosed correctly (True positive) and one will be diagnosed without the disease (false negative), the worst scenario. Now, in the group that doesn't have the disease, which is an amount of 999,900, the test will determine 989,901 as without the disease (true negative) and 9.999 as with the disease (false positive). This means that, despite having coverage of 99% of the cases in groups with and without the disease, the test predicted 9,999 people with the disease, although they are not sick. Such example can be seen in the table 2.1.

Table 2.1: Diagnostics for example disease

Total	With disease	Without disease	Percentage
100	99	1	1%
999.000	9.999	989.901	99%

Thus, in case a patient is diagnosed with the disease, the real probability of him having the condition is the true positive cases (99) divided by the sum of true positive cases with false positive cases (99 + 9.999), which is less than 1%:

$$\text{Real Chances} = \frac{99}{9.999+99} = \frac{99}{10.098} \approx 0.09\% \tag{2}$$

Therefore, without knowing the previous probability, it is hard to say if an outcome is true or false. Also cited in the article, a revision of the cases of breast cancer in 2016 showed that 60% of the women who took tests yearly for ten years had at least one wrong positive diagnosis for the disease.

This directly reflects in the problem of specificity, where specificity measures a test's ability to correctly detect the absence of a condition or disease in people who do not have it. The problem occurs when a test is less specific than expected, which means a test can give many false positive results, as in our example.

The Bayes theorem is of utmost importance for statistics, and since statistics heavily influences the Machine Learning area, it also has Bayes' influence through the algorithm known as Naive Bayes. Naive Bayes is a supervised learning algorithm where conditional independence between pairs of attributes of a given data set is assigned to the established target class. In other

[3] <https://www.bbc.com/portuguese/internacional-59701523>

words, the algorithm assumes that the variables analyzed in the problem have equal relevance for the result.

The algorithm works by first building a probability table describing each predictor variable's frequency concerning the output variable. This table is then used to calculate the probability of a particular output variable given a set of predictor variables. The predictor variables that result in the highest probability are then presented as the answer for that specific context.

To avoid zero values entering the calculation of probabilities, all frequency variables are given an additional count of 1 unit. This is known as Laplace smoothing.

In this work, we were inspired by the concept of "Naive" that Naive Bayes has, so we can consider all the variables independently, although they correlate. Although Naive Bayes was first used and it is presented as a way to detect spam on e-mails, as an example, it can be applied in the most different areas, such as in diagnostic of heart diseases (VEMBANDASAMY; SASIPRIYA; DEEPA, 2015)

2.2 Least Squares

Least Squares is a mathematical optimization method that seeks to find a particular group of data that minimizes the sum of the squares of the differences, also known as residuals, between the initial estimated value and the observed values. In other words, the Least Squares method can find a representation that reduces the error between estimated values and the entire data group, as we can see following:

$$S = \sum_{k=1}^{N} (y_k^o - y_k)^2 \tag{3}$$

where S represents the sum of the squared differences between the observed data points and the predicted values, y_k^o corresponds to the observed values of y, also known as real values, and y_k corresponds to the calculated values of y, i.e., the predicted values. The variable N is the total number of samples, and $k = 1, \ldots, N$ represents the N-th sample.

As said in (BLAIS, 2010), Least Squares is a famous method in engineering (LIU et al., 2006) and experimental science, besides being almost ubiquitous in analysis and digital data processing applications in cases necessary to optimize the results.

As an example to show how the Least Square works, we will consider the example presented in (AGUIRRE, 2014), in his subsection about the Least Squares method. So, consider the following equation as a normal equation for least squares:

$$X^T y = [X^T X]\theta \tag{4}$$

where θ is a vector of n parameters, and $\theta \in \mathbb{R}^n$. Also, X is a vector of independent variables in a system, where $X \in \mathbb{R}^n$, also known as regressors vector, and y is our dependent variable and is a scalar, where $y \in \mathbb{R}$.

In order to isolate θ, we can multiply it by the inverse matrix of $[X^T X]$

$$\theta = [X^T X]^{-1} X^T y \tag{5}$$

Now, consider our system, where we know the parameter vectors, we will call it $\hat{\theta}$, and that an error occurs ξ when we try to explain the observed value y from the vectors X and $\hat{\theta}$, which means:

$$y = X\hat{\theta} + \xi \tag{6}$$

where $\xi \in \mathbb{R}^n$ is the vector of errors that happened when we try to explain y as $X\hat{\theta}$. Thus, it would be interesting that $\hat{\theta}$ was a vector that minimizes ξ em in some way. So, we can define a value (J_{LS}) that represents the quality of the adjustment of $X\hat{\theta}$ that can be calculated like this:

$$J_{LS} = \sum_{i=1}^{N} \xi(i)^2 = \xi^T \xi = ||\xi||^2 \tag{7}$$

replacing ξ from equation 6 we have:

$$J_{LS} = (y - X\hat{\theta})^T (y - X\hat{\theta}) = y^T y - y^T X\hat{\theta} - \hat{\theta}^T X^T y + \hat{\theta}^T X^T X\hat{\theta} \tag{8}$$

In order to minimize the J_{LS} in respect of $\hat{\theta}$ it is necessary that $(\frac{\partial J_{LS}}{\partial \hat{\theta}}) = 0$. Doing so, we have:

$$\frac{\partial J_{LS}}{\partial \hat{\theta}} = -(y^T X)^T - X^T y + (X^T X + X^T X)\hat{\theta} = -2X^T y + 2X^T X\hat{\theta} \tag{9}$$

making 9 equals to zero, we have:

$$\theta = [X^T X]^{-1} X^T y \tag{10}$$

which is equal to the equation 5. Thus, in order to minimize $\hat{\theta}$, it is necessary that:

$$\frac{\partial J_{LS}}{\partial \hat{\theta}}^2 = 2X^T X > 0 \tag{11}$$

which is true, because $2X^T$ is positive by his construction. Therefore, the equation 5 is the estimator that gives the value of $\hat{\theta}$ that minimizes the sum of the squared errors. In resume, we have:

$$\hat{\theta}_{LS} = \arg_\theta \min J_{LS} = [X^T X]^{-1} X^T y \tag{12}$$

where $\arg_\theta \min J_{LS}$ indicates the argument, that belongs to the domain of θ, that minimizes the cost function J_{LS}

2.3 Hemodynamic Variables

As said before, we use the same five variables that the Mobil-O-Graph uses as reference values to diagnose vascular age. In this section, those five variables will be enlightened.

2.3.1 Pulse Wave Velocity

Pulse Wave Velocity (PWV) is a well-known cardiology metric, representing the speed at which a pressure wave travels through the circulatory system. This indicator, which is calculated in a non-invasive manner, can be calculated as follows:

$$\mathrm{PWV} = \frac{d}{tt} \tag{13}$$

where d represents the distance between two measured points, e.g., from the heart (initial point) to a human limb (endpoint), such as the arm. The variable tt is when the wave travels from the initial point to the endpoint. Such calculation and index are expressed in image 1, where such image was developed by us.

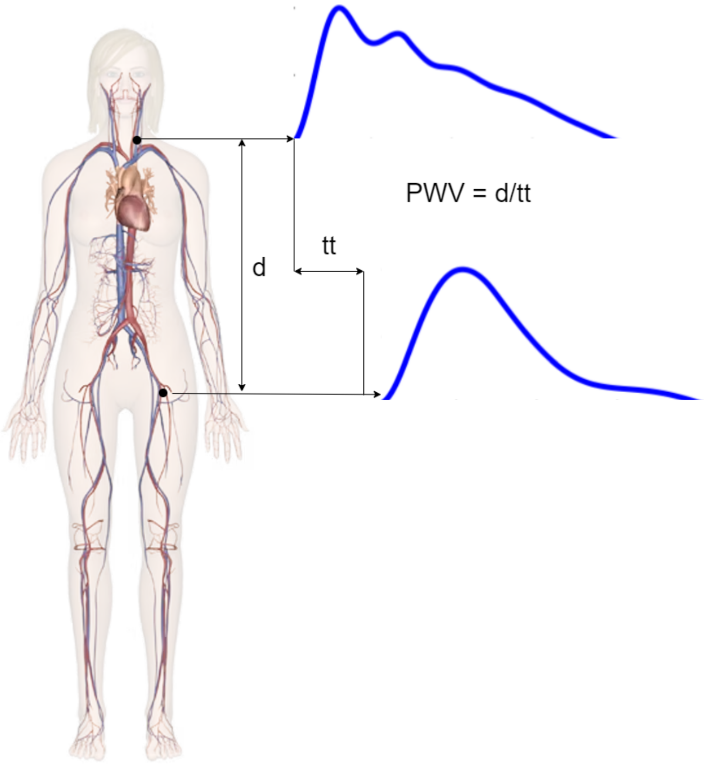


Figure 1: Representation of PWV

Thus, PWV is commonly used in problems such as arterial stiffness. Thanks to its calculation, the higher the PWV, the higher the arterial stiffness, where arterial stiffness represents the reduced ability of the artery to expand and contract with each heartbeat. Thus, PWV can work with arterial stiffness and other related problems, such as atherosclerosis (KIM; KIM, 2019) and the diagnosis of vascular age.

It’s important to inform you that, although (PAIVA et al., 2020) presents us reference values for individuals with and without cardiovascular risk (where each one of them has reference values for women and men), cardiologists always consider the individual that is taking the test as a person with cardiovascular risk. Therefore, these are the reference values for the PWV in table 2.2 for women and 2.3, where these and next reference values presented in this work where all collected from Paiva’s work:

Table 2.2: PWV reference values for women

Age range	Median	10th	25th	75th	90th
< 30	5.3	4.7	5.0	5.6	6.0
30 - 39	5.8	5.3	5.5	6.2	6.7
40 - 49	6.8	6.0	6.4	7.2	7.7
50 - 59	7.9	7.1	7.5	8.3	8.8
60 - 69	9.3	8.4	8.8	9.8	10.4
70+	11.8	10.2	10.8	12.9	14.0

Table 2.3: PWV reference values for men

Age range	Median	10th	25th	75th	90th
< 30	5.5	5.0	5.3	5.8	6.3
30 - 39	6.1	5.5	5.8	6.4	6.7
40 - 49	6.8	6.2	6.4	7.1	7.5
50 - 59	7.9	7.1	7.5	8.3	8.8
60 - 69	9.2	8.4	8.7	9.7	10.2
70+	11.2	9.9	10.4	12.1	13.2

2.3.2 Augmentation Index

The Augmentation Index (AIx) is another metric related to the cardiovascular system, correlated to central diastolic pressure, heart rate, and even an individual height and gender. It represents the shape of the pressure wave in the circulatory system. This pressure wave is calculated by the difference between systolic blood pressure and the pressure wave created by the reflection of pressure waves from the peripheral vessels to the heart.

Thus, AIx can indicate how much a reflected pressure wave from the periphery will amplify systolic pressure. Therefore, a high AIx suggests increased wave pressure and may be associated with cardiovascular diseases (NüRNBERGER et al., 2002) such as arterial stiffness, as the arteries will have a reduced capacity to absorb the reflected pressure wave and can provide information about the circulatory system as a whole. Furthermore, these are the reference values for AIx for women in 2.4 and for men in 2.5:

Table 2.4: AIx reference values for women

Age range	Median	10th	25th	75th	90th
< 30	28	11	20	34	38
30 - 39	26	11	18	32	37
40 - 49	25	10	17	34	38
50 - 59	24	8	14	33	39
60 - 69	28	11	18	37	44
70+	33	17	25	42	48

Table 2.5: AIx reference values for men

Age range	Median	10th	25th	75th	90th
< 30	16	2	8	23	30
30 - 39	15	3	9	21	27
40 - 49	15	2	8	23	30
50 - 59	15	3	7	24	32
60 - 69	17	3	9	26	34
70+	22	4	12	31	41

2.3.3 Central Systolic Pressure

Central Systolic Pressure (CSP) represents the blood pressure in the major blood vessels that carry blood from the heart to the rest of the body, the central arteries. It measures the pressure in the aorta, the main blood vessel that carries blood from the heart to the rest of the body.

This way, the CSP can tell how much pressure the heart generates in each beat and how much pressure the aorta and other vessels receive. In the same way as PWV, AIx, and the other indexes from this work, CSP can be calculated non-invasively. It can be estimated from the PWV and shows how the CSP is related to the arterial stiffness problem (O'ROURKE, 1990). Furthermore, these are the reference values for CSP for women in 2.6 and for men in 2.7:

Table 2.6: CSP reference values for women

Age range	Median	10th	25th	75th	90th
< 30	118	102	109	127	131
30 - 39	120	102	110	130	143
40 - 49	121	104	112	134	146
50 - 59	124	106	114	135	146
60 - 69	127	105	115	141	154
70+	131	108	118	146	165

Table 2.7: CSP reference values for men

Age range	Median	10th	25th	75th	90th
< 30	123	107	114	132	144
30 - 39	125	108	116	133	141
40 - 49	123	108	115	131	141
50 - 59	124	105	114	134	144
60 - 69	123	103	112	136	149
70+	125	102	111	140	156

2.3.4 Central Diastolic Pressure

Central Diastolic Pressure (CDP) represents the minimum pressure in the aorta while the heart relaxes and fills with blood (this phase is called diastole). CDP can be measured invasively through catheterization or non-invasively using techniques such as tonometry, a medical procedure used to measure the pressure inside the eye. It can measure arterial stiffness and estimates CDP.

Monitoring CDP can provide important information about cardiovascular health and help guide treatment decisions (MCEVOY et al., 2016), just because Abnormal CDP readings may indicate underlying cardiovascular disease, such as hypertension, atherosclerosis, or heart failure, and may require further investigation or treatment. Furthermore, these are the reference values for CDP for women in 2.8 and for men in 2.9:

Table 2.8: CDP reference values for women

Age range	Median	10th	25th	75th	90th
< 30	82	68	73	90	97
30 - 39	86	71	77	95	105
40 - 49	86	71	78	94	103
50 - 59	84	71	77	92	100
60 - 69	81	67	74	90	98
70+	81	66	72	89	97

Table 2.9: CDP reference values for men

Age range	Median	10th	25th	75th	90th
< 30	83	72	77	93	100
30 - 39	88	75	80	96	103
40 - 49	90	75	82	97	104
50 - 59	88	75	80	97	103
60 - 69	85	71	77	93	101
70+	82	68	74	91	98

2.3.5 Pulse Pressure

Pulse Pressure (PP) can be defined as the difference between the CSP and CDP values, subtracting the diastolic blood pressure from the systolic. It is another important indicator of cardiovascular health and broader pulse pressure may be associated with an increased risk of cardiovascular diseases, such as coronary artery disease or heart failure. This occurs in order that a broader pulse pressure indicates that the arteries are stiffer and less able to absorb the force of the blood flow, which can put additional strain on the heart.

However, a narrow pulse pressure can also indicate bad health conditions, such as congestive heart failure(HAIDER et al., 2003), where the heart is unable to pump enough blood to meet the body's needs, and Shock, which is a medical emergency indicating that the body's organs and tissues are not receiving enough blood and oxygen. Furthermore, these are the reference values for PP for women in 2.10 and for men in 2.11:

Table 2.10: PP reference values for women

Age range	Median	10th	25th	75th	90th
<30	34	24	28	41	48
30-39	34	24	28	38	46
40-49	35	25	29	43	53
50-59	39	28	32	47	58
60-69	44	30	36	55	66
70+	50	33	41	63	77

Table 2.11: PP reference values for men

Age range	Median	10th	25th	75th	90th
< 30	38	26	31	46	52
30 - 39	36	25	31	41	48
40 - 49	33	23	28	37	46
50 - 59	34	25	28	41	49
60 - 69	37	25	31	46	58
70+	42	28	34	52	66

2.4 Final Considerations

In this chapter, the main theoretical contents that serve as a basis for understanding the results of this work were presented. In the next chapter, we will use the Bayes' Theorem, along with the notion of Naive Bayes, as the calculation method for the vascular ages of the individuals analyzed and Least Squares to optimize the values found by the algorithm.

We also consider it essential to say that all the hemodynamic variables mentioned in this work, and the others that exist, are better used for diagnosing possible heart problems when used together.

3

Methodology

3.1 Context

In the previous chapters, we talked about how Cardiovascular diseases still are a global problem and how arterial stiffness can be well diagnosed using Mobil-O-Graph. However, it still needed reference values so the diagnostic could be made, and even so, it required reference values for other populations besides the one the device was calibrated in. Now we will use all the Hemodynamic Variables that we presented, along with the Bayes Theorem, Naive Bayes concept, and Least Squares, to represent these reference values in an actual diagnostic that reflects the reality of the Brazilian population.

First, we need to say that each possible class representing one possible diagnostic is an age range of ten years. So, our possible diagnostics are from classes representing an age rate from under 30 to over 70 years old, these are the six age ranges:

1. < 30 years old
2. 30 to 39 years old
3. 40 to 49 years old
4. 50 to 59 years old
5. 60 to 69 years old
6. 70 years old or more

It's also important to say that to be considered an individual with a healthy cardiovascular age, the individual needs to be in an age range that includes his chronological age or even an age range even lower than his actual age.

On top of that, we consider it fundamental to explain that the distribution of the Brazilian population used in this thesis was according to the data presented by the Brazilian Institute of Geography and Statistics (IBGE)[4], in the year 2010, due the fact that the institute was unable to carry out a data survey during 2021, the year the study was realized, nor has it yet published data from the 2022 survey, the date of writing of this work, where all this delay occurred due to the Covid-19 pandemic. The image 2 shows the distribution of the Brazilian population in 2010.

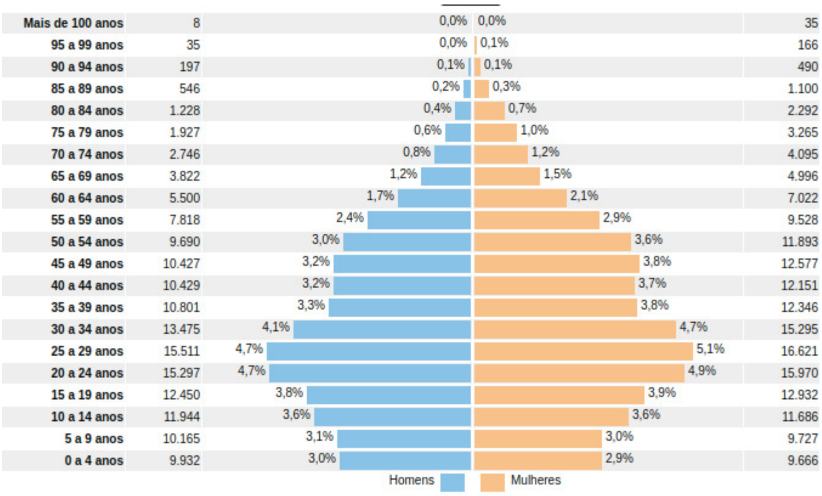


Figure 2: Age pyramid

At least, the reference values in (PAIVA et al., 2020) refer to the quantiles, median, and percentiles each hemodynamic variable has for that specific age range. As said before, they are presented in the 10 percentile, first quantile, median, third quantile, and 90 percentile. Also, we have reference values for four groups: men with no cardiovascular risk, women with no cardiovascular risk, men with cardiovascular risk, and women with cardiovascular risk.

3.2 Main idea

Let C be the discrete variable that represents an age range of a patient and S be the continuous variable that represents the CSP of the patient. We want to compute P$(C = c_1|S = s)$, where c_1 is one of the ages range of C. Applying the Bayes' Theorem, we have:

$$P(C = c_1|S = s) = \frac{P(C = c_1) * f_{S|C}(s|c1)}{f_S(s)} \tag{14}$$

Now, remember that (PAIVA et al., 2020) gives us the reference value for our five hemodynamic variables for each age range we are considering. In this way, we represented the reference value as $F_{S|C}(s|c)$. Thus, we can find that:

[4] <https://www.ibge.gov.br/censo2010/apps/sinopse/webservice/frm_piramide.php?codigo=320530>

$$f_{S|C}(s|c_1) = \frac{dF_{S|C}(s|c_1)}{ds} \tag{15}$$

we know that:

$$f_S(s) = \sum_{k=c_1}^{c_5} f_{S|C}(s|k) * P(C = k) \tag{16}$$

In this way, if we have a way to compute P(C = c), we can compute P(C = c1|S = s). In addition, we are going to also consider D as the continuous variable that represents the CDP of a patient. Therefore, we have:

$$P(c_1|s,d) = \frac{P(C = c_1) * f_{S,D|C}(s,d|c_1)}{f_{S,D}(s,d)} \tag{17}$$

if we assume that S and D are conditionally independents, we have:

$$f_{S,D}(s,d) = f_{S|C}(s,c) * f_{D|C}(d,c) \tag{18}$$

This structure allows us to work with the five hemodynamic variables presented earlier. Let's say X is a set of our possible variables (S, D, PWV, AI, PP), and x represents the values of the variables (s, d, pwv, ai, pp) in our set. In this way, we have:

$$P(c|x) = \frac{P(C = c) * f_{X|C}(x|c)}{f_X(x)} \tag{19}$$

Again, if we assume that (S, D, PWV, AI, PP) are K, and are conditionally independent, we have:

$$f_K(k) = \prod_{k=s}^{pp} f_{K|C}(k,c) \tag{20}$$

We compute this process using the Python language, along with the Scipy and Numpy libraries. This is the main code, which is also present on Github:

```
import scipy.stats as st
from scipy.stats import norm
from scipy.optimize import minimize
import numpy as np

def __computeF(x):
  mu = x[2]

  def obj_func(sigma, x):
    n = norm.ppf([0.1,0.25,0.5,0.75,0.9])
```

```
        x = np.array(x)
        return sum(((x - x[2])/sigma[0] - n)**2)

    arg = minimize(obj_func,x0=np.array([0.1]),args=(x), method="BFGS")
    sigma = arg.x[0]

    Fn = st.norm(mu,sigma)
    return Fn
```

As an example of the behavior of our reference values, we plotted the Cumulative distribution functions (CDFs) interpolation and normalization of the CSP for women with cardiovascular risk. These functions can be seen in the figures 3, 4, 5, 6, 7, and 8.

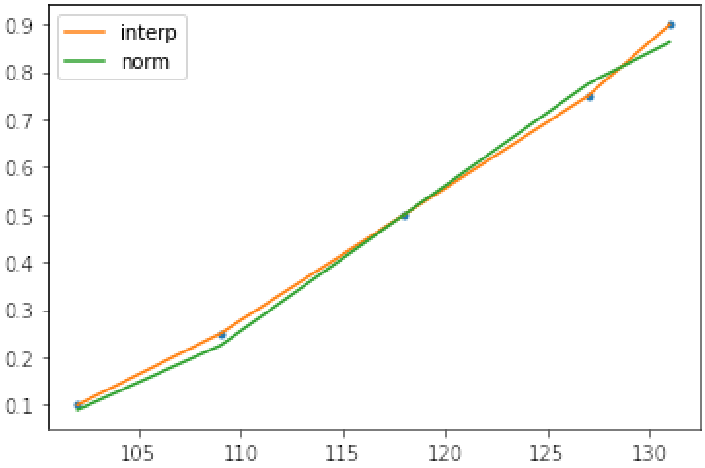


Figure 3: Values for ages under 30 years old

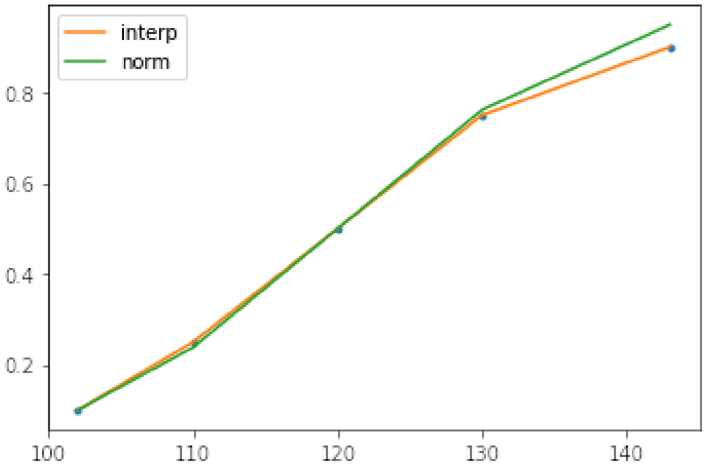


Figure 4: Values for ages between 30 - 39

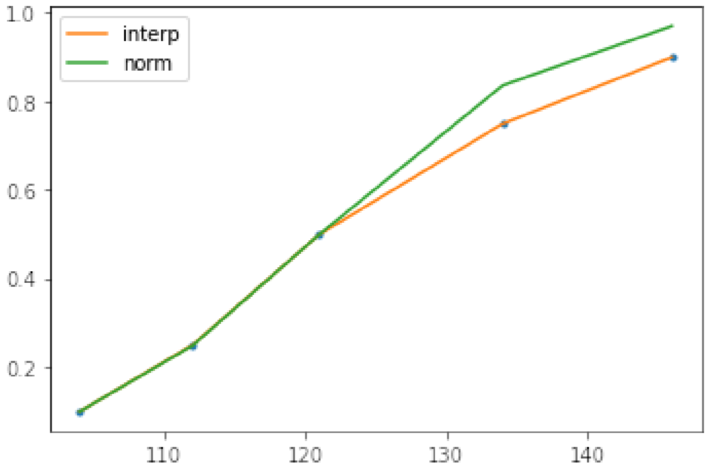


Figure 5: Values for ages between 40 - 49

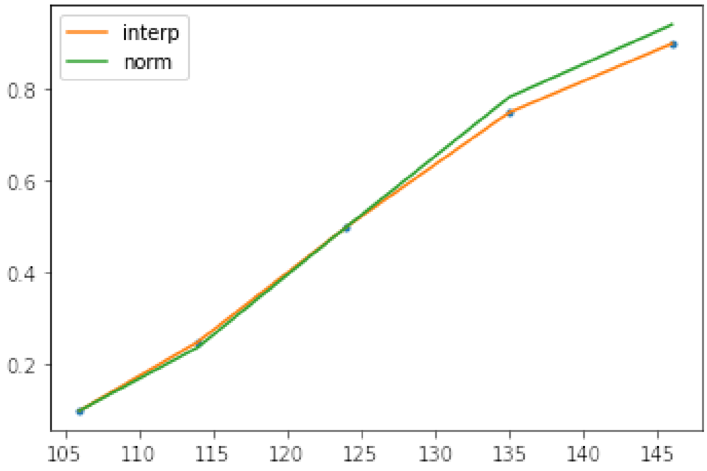


Figure 6: Values for ages between 50 - 59

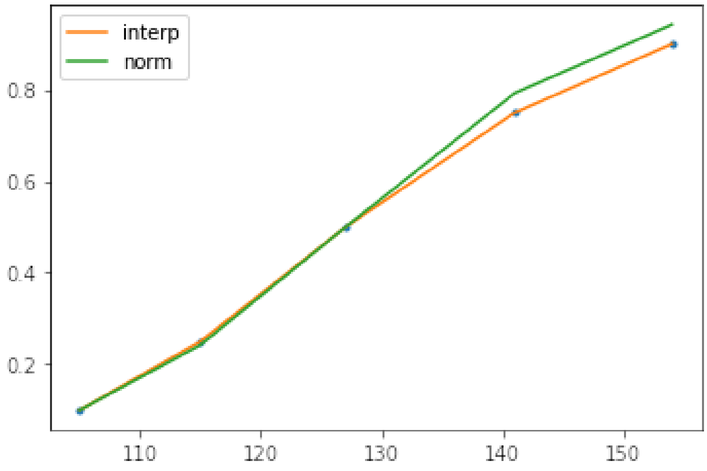


Figure 7: Values for ages between 60 - 69

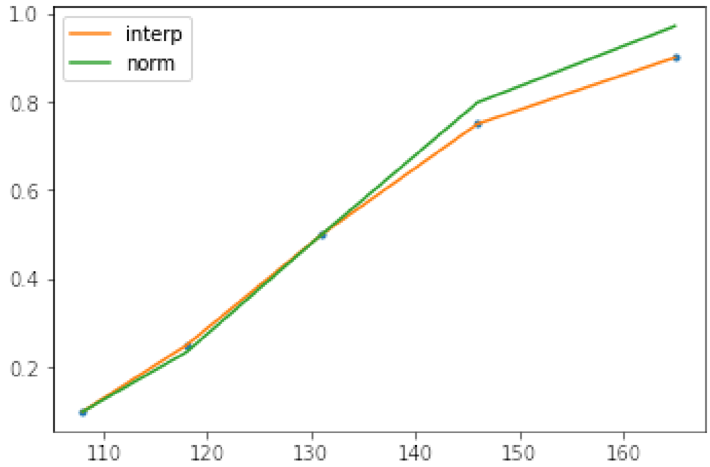


Figure 8: Values for ages over 70 years old

We used CDFs as they are often used in statistical analysis to describe the probability distribution of a set of data. In many cases, it is necessary to estimate the CDF at points that are not directly observed in the data or that are very close to the observed data but may contain some degree of noise or uncertainty, which is really our case in this problem due to the condition of not having access to the data that were used to create the reference values. To address this issue, interpolation can be used to estimate the CDF at these intermediate points.

Interpolation is a mathematical technique for estimating the value of a function at points between observed data points. In the context of CDFs, interpolation can be used to estimate the probability of a certain value falling between two observed data points. In Python, the SciPy library provides the interp1d function, which can be used for linear interpolation of a CDF.

Normalization is another important technique in statistical analysis, particularly when comparing CDFs of different data sets or testing whether a given data set conforms to a known distribution. Normalization involves scaling the values of a CDF so that they fall within a specified range or conform to a standard distribution. For example, the cumulative distribution of a normal distribution is well-known and can be used as a reference for comparison.

```
import matplotlib.pyplot as plt
from scipy.interpolate import interp1d
from scipy.stats import norm

x = [102,109,118,127,131]
F = [0.1,0.25,0.5,0.75,0.9]

import scipy.stats as st
mu = x[2]
sigma = (x[0]-x[2])/-1.282
```

```
Fn = st.norm(mu,sigma)

plt.plot(x,F,’.’)
Fa = interp1d(x, F)
plt.plot(x,Fa(x),label="interp")
plt.plot(x,Fn.cdf(x),label="norm")
plt.legend()
plt.show()
```

In the Python code provided, which refers to the CSP reference values for women, interpolation is used to estimate the CDF at intermediate points using the interp1d function. Normalization is used to compare the estimated CDF to a normal distribution, which is done by fitting the normal distribution to the data using the mean and standard deviation estimated from the data.

3.3 Conclusion

In this section, we present how the techniques presented in section 2 were able to be put all together in order to solve our vascular age classification. We presented how Bayes Theorem could work with all the variables together, considering the concept of naive, from Naive Bayes, where we consider all variables independent of each other, and how we can use Least Squares in order to reduce the possible noise that we might have in real cases.

4

Results

4.1 API

In order to compute, test and evaluate all the hypotheses presented in the Methodology, such as our objectives presented in the Introduction of this thesis, we developed a web API (Application Programming Interface) using Python (as the programming language), Flask[5] (a web microframework for develop web API), Scipy and Numpy (both libraries of the Python language, commonly used in scientific purposes to have more mathematical and statistical functions to work), and, at least, it was used the library Pandas to manipulated datasets or even data arranged in tables.

The web API was hosted on Heroku, a cloud service platform that supports multiple programming languages, and the source code was hosted on Github [6]. In figure 9 we have the components diagram of the API.

In this diagram, we have the current structure of our backend project, where we host our web API. App is the component where the flask module is created and connected with the CORS (Cross-Origin Resource Sharing) module, which allows us to connect resources from different origins to consume our API, such as our frontend. Also, we have the blueprint module so that the endpoints can be handled from different files.

Furthermore, we have the Main component, which it has two main endpoints to consume our application:

1. "get_cvrf": The endpoint that receives all our variables needed on the calculation and returns the age range classes.
2. "home": An endpoint that intends to see if the connection, such as the web app, is responding;

[5] <https://flask.palletsprojects.com/en/2.2.x/>

[6] <https://github.com/Raksantos/blood-age-api>

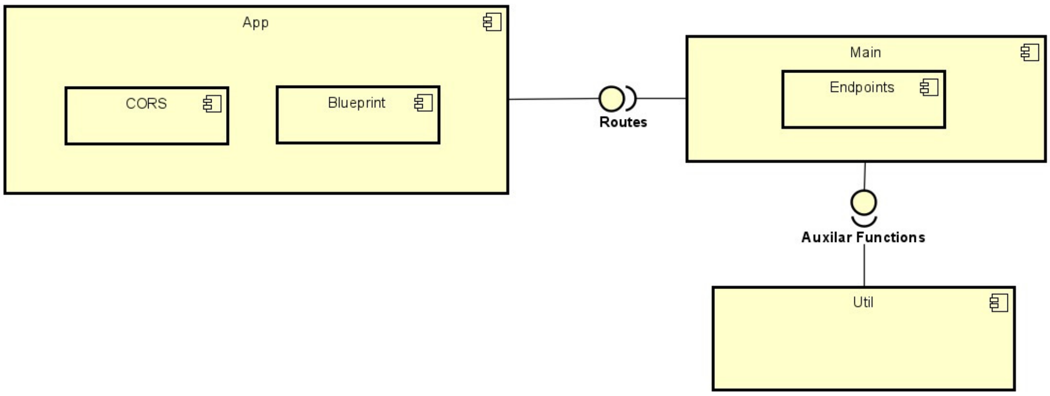


Figure 9: Deployment diagram API

Lastly, we have the Util component, where we wave the utils functions to make our computations, such as:

1. "__computeF_min_quadratic()": Method to calculate the Least Squares method, using the function minimize from the class Optmize from Scipy.
2. "__tables()": Return our reference values for each hemodynamic variable;
3. "__get_frequencies_computed()": Compute our frequencies using the Least Squares method
4. "compute_class()": Our method to compute the classes and return the class with higher probability;

4.2 Frontend

Such as the web API, the frontend was also hosted on Heroku, and the source code can also be found on Github[7]. The hosted web page can be found on [8]. We used the React.js [9] framework, a well-known microframework used to develop Single Page Applications (SPA). In 10 we have the components diagram of the frontend.

Now, in this diagram, we have the structure of the frontend project. App is the main component where react app is instanced, and we have the main structure of the home screen, such as the Form and the Footer. Also, we have our styles defined using the styles-component library.

In Components, we have two main components that may vary according to our needs: the Alerts, responsible for showing us the return from the backend, and the Form component, where we have our input fields.

[7] <https://github.com/Raksantos/blood-age>

[8] https://blood-age.herokuapp.com/

[9] <https://react.dev/>

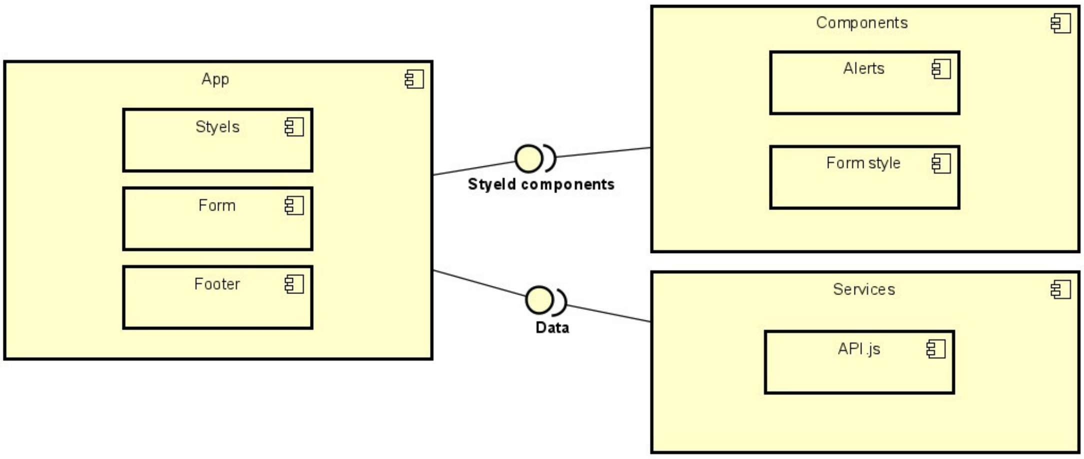


Figure 10: Deployment diagram frontend

Lastly, we have the Services, with the API component responsible for consuming our web API defined previously.

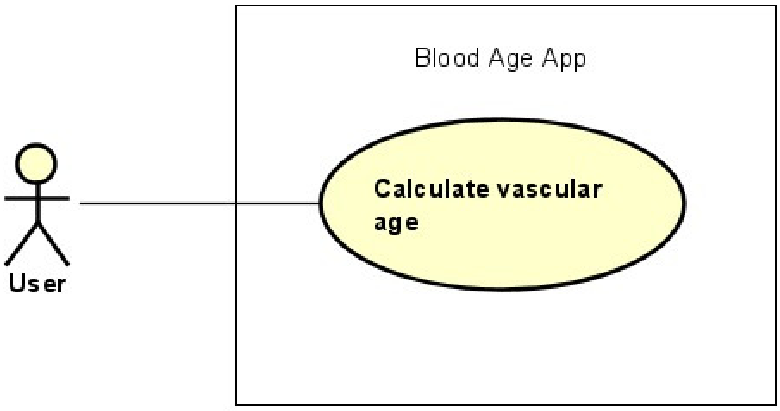


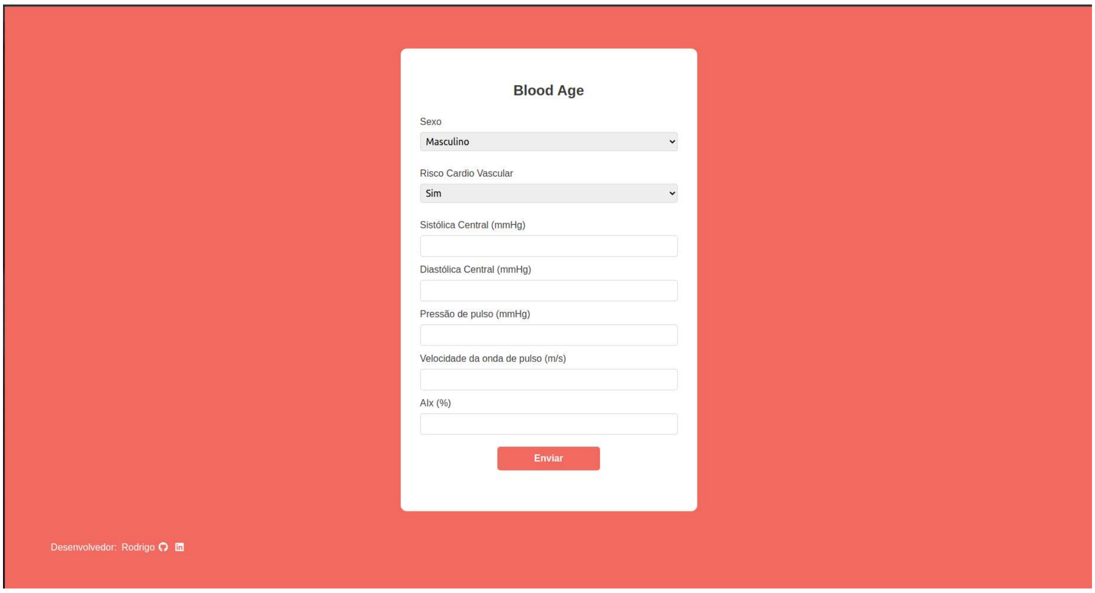
Figure 11: Use Case Diagram

Our last diagram is the Use Case diagram, where we represent how our user can interact with the system. It can be seen in figure 11.

These are the images from the web page hosted, where the image 12 is the home page and 13 is the result of the application after the values were informed:

4.3 Results from patients

In all, 15 patients of both male and female sex were used to validate the system. As explained before, all of those patients were considered with cardiovascular risk, although the website gives the option to consider the patient as without cardiovascular risk. Listed in 4.1 we have the values for each hemodynamic variable of the tested patients:



Blood Age

Sexo

Masculino

Risco Cardio Vascular

Sim

Sistólica Central (mmHg)

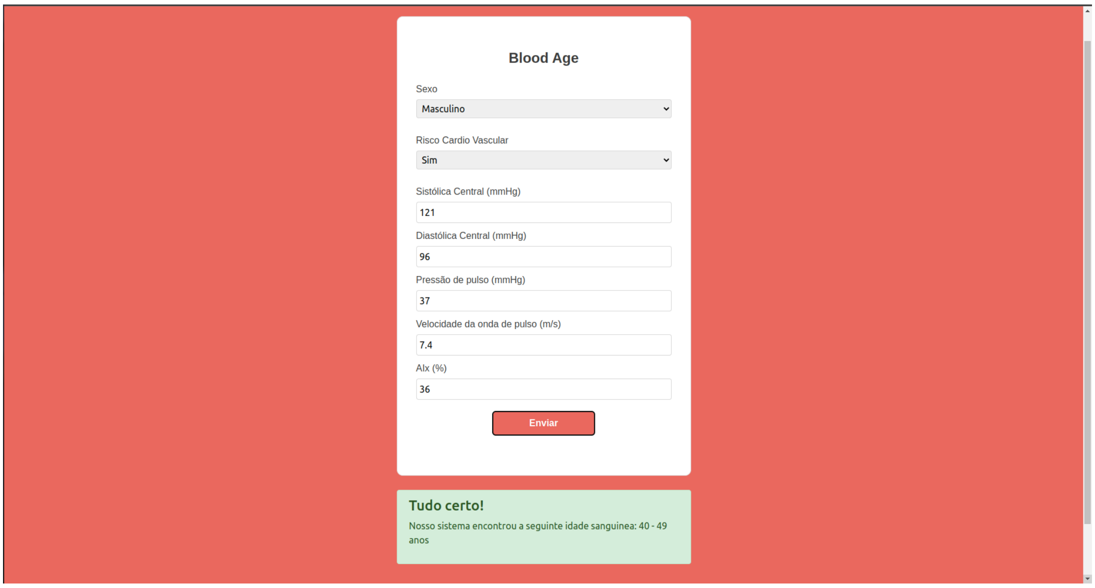
Diastólica Central (mmHg)

Pressão de pulso (mmHg)

Velocidade da onda de pulso (m/s)

AIx (%)

Figure 12: Application Homepage



Blood Age

Sexo

Masculino

Risco Cardio Vascular

Sim

Sistólica Central (mmHg)

121

Diastólica Central (mmHg)

96

Pressão de pulso (mmHg)

37

Velocidade da onda de pulso (m/s)

7.4

AIx (%)

36

Tudo certo!

Nosso sistema encontrou a seguinte idade sanguinea: 40 - 49 anos

Figure 13: Test case

Table 4.1: Results

Sex	CSP	CDP	PP	PWV	AIx
F	111	78	42	10.2	35
M	104	67	60	5.4	11
M	107	80	44	5.0	21
M	138	106	44	9.8	31
M	107	70	47	6.4	3
F	131	82	78	10.9	32
M	115	69	53	11.3	8
M	111	76	41	6.9	23
F	94	63	41	9.6	29
M	121	96	37	7.4	36
M	117	79	62	5.5	23
M	98	70	41	7.7	23
F	97	63	46	6.0	24
M	129	76	62	11.7	28

And in 4.2 we have, respectively, the chronological age, vascular age from Mobil-O-Graph, and vascular age from our system:

We also reviewed the results with a cardiology specialist, who presented great enthusiasm for the work and told us that the results aligned with what was expected for the reported values. So, it is normal that the app presents an equal or an age range under the chronological age.

Table 4.2: Results

Chronological age	Mobil age	Blood app age
72	74	60 - 69
28	31	< 30
24	27	< 30
64	71	60 - 69
45	43	30 - 39
70	78	70+
76	80	70+
50	47	40 - 49
72	70	60 - 69
51	52	40 - 49
24	33	< 30
57	55	50 - 59
43	38	30 - 39
76	83	70+

5

Final considerations

This work has successfully developed and tested, with the revision of a doctor, a method to classify vascular age for the Brazilian population. In the work testing process, fifteen exams were used, which gave us age ranges related to the reference values for the patients. We expect this to be a very encouraging result due to the possibility of calculating the vascular age of the individual, considering the Brazilian reality and without the need of having the Mobil-O-Graph in hands to measure all the hemodynamic variables.

Furthermore, with the web application hosted, we expect more doctors to use and test the application and, consequently, our work to find possible failed cases or cases that weren't correctly mapped.

For future works, we expect to have access to the original dataset used to develop the reference variables to test other methodologies that may be even more precise than the one we use now. We can try classification machine learning algorithms to solve this problem, which can present us with a more precise method where we do not need to consider all the hemodynamic variables independently as we did and be able to calculate a specific vascular age, just like the Mobil-O-Graph does, and not an age range.

In addition, we might be able to find the correlation between the hemodynamic variables, such as the PWV and the AI (precisely because the AI is calculated based on the PWV) and try to reduce the number of variables needed to realize the classification. Thus, we might be able to develop a web application that can reach not only doctors interested in calculating a vascular age range but citizens that might want to calculate the vascular age and use this to look for a cardiologist.

Finally, we hope this work to be able to help other possible works in development, as well as professionals who work with problems related to cardiovascular age, such as cardiologists or researchers, and civil society.

Bibliography

AGUIRRE, L. A. *Introdução à Identificação de Sistemas - Técnicas Lineares e Não Lineares Aplicadas a Sistemas: Teoria e Aplicação*. [S.l.: s.n.], 2014. v. 4.

BLAIS, J. A. Least squares for practitioners. *Mathematical Problems in Engineering*, v. 2010, 2010. ISSN 1024123X.

BOWRY, A. D. K. et al. The burden of cardiovascular disease in low- and middle-income countries: Epidemiology and management. *Canadian Journal of Cardiology*, v. 31, p. 1151–1159, 2015. ISSN 0828-282X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0828282X15005073>.

HAIDER, A. W. et al. Systolic blood pressure, diastolic blood pressure, and pulse pressure as predictors of risk for congestive heart failure in the framingham heart study. *Annals of Internal Medicine*, v. 138, 2003. ISSN 00034819.

KIM, H. L.; KIM, S. H. Pulse wave velocity in atherosclerosis. *Frontiers in Cardiovascular Medicine*, v. 6, 2019. ISSN 2297055X.

LIU, Y. S. et al. Automatic least-squares projection of points onto point clouds with applications in reverse engineering. *CAD Computer Aided Design*, v. 38, 2006. ISSN 00104485.

MCEVOY, J. W. et al. Diastolic blood pressure, subclinical myocardial damage, and cardiac events: Implications for blood pressure control. *Journal of the American College of Cardiology*, v. 68, 2016. ISSN 15583597.

NüRNBERGER, J. et al. Augmentation index is associated with cardiovascular risk. *Journal of Hypertension*, v. 20, 2002. ISSN 02636352.

O'ROURKE, M. Arterial stiffness, systolic blood pressure, and logical treatment of arterial hypertension. *Hypertension*, v. 15, 1990. ISSN 0194911X.

O'ROURKE, M. F.; MANCIA, G. Arterial stiffness. *Journal of Hypertension*, v. 17, n. 1, 1999. ISSN 0263-6352. Disponível em: <https://journals.lww.com/jhypertension/Fulltext/1999/17010/Arterial_stiffness.1.aspx>.

PAIVA, A. M. G. et al. Reference values of office central blood pressure, pulse wave velocity, and augmentation index recorded by means of the mobil-o-graph pwa monitor. *Hypertension Research*, v. 43, p. 1239–1248, 2020. ISSN 1348-4214. Disponível em: <https://doi.org/10.1038/s41440-020-0490-5>.

VEMBANDASAMY, K.; SASIPRIYA, R.; DEEPA, E. Heart diseases detection using naive bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, v. 2, 2015.