Understanding Stroke Risk Profiles in Middle-Aged Adults: A Genetic Algorithm-Based Feature Selection Aproach

Ligia Ferreira de Carvalho Gonçalves[1][a], Caio Davi Rabelo Fiorini[1][b], Daniel Rocha Franca[2][c], Marta Dias Moreira Noronha[3][d], Mark Alan Junho Song[3][e] and Luis Enrique Zárate Galvez[3][f]

[1]*Data Science and Artificial Intelligence, Pontifícia Universidade Católica de Minas Gerais, Rua Claudio Manuel, Belo Horizonte, Brazil*

[2]*Computer Science, Pontifícia Universidade Católica de Minas Gerais, Rua Claudio Manuel, Belo Horizonte, Brazil*

[3]*Institute of Exact Sciences and Computer Science, Pontifícia Universidade Católica de Minas Gerais, Rua Claudio Manuel, Belo Horizonte, Brazil*

{ligiacarv.goncalves, caiodavi442, danielfrancamdt}@gmail.com, {martanoronha, song, zarate}@pucminas.br

Keywords: Stroke, Machine Learning, Genetic Algorithm, Decision Tree, Middle-Aged, Rules.

Abstract: Data mining and machine learning techniques have been widely used in the knowledge extraction process of medical databases, one highlight being their use to improve diagnostic systems. Decision trees are supervised black box machine learning models that, although simple, are easy to interpret. In this work, we propose the use of these techniques to describe the profile of middle-aged adults (40-59) diagnosed with stroke, a disease that in Brazil was one of the main causes of death in previous years. The genetic algorithm was applied to extract the best characteristics so that the Decision Tree algorithm could then be used in the database provided by the 2019 National Health Survey to obtain the most comprehensive rules and identify the most relevant attributes for describing the profile of these individuals. The conclusions indicate that the rules generated for middle-aged adults are mainly about routine habits, such as work or salt consumption.

1 INTRODUCTION

Cerebrovascular Accident (CVA), also known as a stroke, according to the Brazilian Stroke Society, can be characterized by the appearance of a sudden neurological deficit caused by a problem in the vessels/arteries or veins of the brain. In Brazil, according to the Civil Registry Transparency Portal, among cardiovascular diseases, stroke was the main cause of death in 2023. In 2024 alone, from January to March, about 20 thousand deaths were recorded (BRASIL, 2024). Although the occurrence of stroke is higher among individuals in older age groups (Rajati et al., 2023), there has recently been a worrying increase in the number of cases among middle-aged adults, a group that until now, was believed to have a low predisposition to this condition. The survey conducted by the American Heart Association discusses the growth of cases among adults under 49 years of age, a fact that is relevant to understanding the situation of the disease among the middle-aged population.

Machine learning (ML) is an area that explores the study and development of computational models that learn through datasets, and which has received great attention in the field of medicine (Paixão et al., 2022) being used to improve clinical diagnostic systems, and specifically in cardiology, it has been responsible for a large part of the scientific productions about aiding in the diagnosis of cardiovascular diseases.

Given the data and discussions presented, it is essential to identify and understand the factors that characterize the profile of middle-aged adults, aged 40 to 59, diagnosed with stroke. One way to understand how the population behaves is by applying computer learning models, which allow us to discover whether the same conditions (extracted rules) characterize the occurrence of stroke in the population.

This work is divided into 5 sections: Introduction, Related Work, Materials and Methods, Results and Discussions, and Conclusions. As a data source, the

[a] https://orcid.org/0009-0000-7601-2938
[b] https://orcid.org/0009-0005-3606-7623
[c] https://orcid.org/0009-0008-9457-1221
[d] https://orcid.org/0000-0002-2992-8422
[e] https://orcid.org/0000-0001-7315-3874
[f] https://orcid.org/0000-0001-7063-1658

Gonçalves, L. F. C., Fiorini, C. D. R., Franca, D. R., Noronha, M. D. M., Song, M. A. J. and Galvez, L. E. Z.
Understanding Stroke Risk Profiles in Middle-Aged Adults: A Genetic Algorithm-Based Feature Selection Aproach.
DOI: 10.5220/0013184600003911
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 2: HEALTHINF*, pages 623-630
ISBN: 978-989-758-731-3; ISSN: 2184-4305
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

recent study by the Brazilian Institute of Geography and Statistics (IBGE), National Health Survey (PNS) 2019, a survey carried out through questionnaires in 2019 throughout the national territory [1][2][3][4].

2 RELATED WORKS

Given the impact that the chronic disease CVA has on the age group described in the previous section, a search was carried out in repositories based on terms in English and Portuguese, of the keywords: stroke, risk factors, and machine learning. Several studies investigate the risk factors that indicate vulnerability to stroke, while other authors apply machine learning techniques to identify patterns and predict their occurrence.

In the work by Yousufuddin and Young (2019), the authors discuss the relationship between the aging process and the occurrence of stroke. They concluded that the presence of factors such as hypertension and diabetes increases the risk of stroke as the individual ages, while the relative risk associated with factors such as cigarette consumption and high blood pressure decreases over time. In addition, in Noche et al. (2020), the authors point out the knowledge gap concerning understanding the risk factors for middle-aged adults (aged 40-60). Their research discusses precisely the importance of understanding the phenomenon of stroke recurrence in this age group.

In Dritsas and Trigka (2022), the authors proposed the use of ML models for stroke prediction using the following algorithms: Naive Bayes, Random Forest, Logistic Regression, K-Nearest Neighbors, Stochastic Gradient Descent, Decision Tree, Multilayer Percepton, Majority Voting, and Stacking. The results obtained showed that the classification by Stacking had a better performance when compared to the other methods used, with an AUC of 98.9%, F-Measure, Precision, and Recall equal to 97.4%, and accuracy of 98%.

3 MATERIALS AND METHODS

3.1 Description of the Database

In this work, we use data from the National Health Survey (PNS) 2019, carried out by the Brazilian Institute of Geography and Statistics (IBGE) in partnership with the Ministry of Health. The survey collected data on the general health and lifestyle situation of the population from Brazil. It has 293,726 records and 1,088 attributes. For this study, we made a cut-off for the chronic disease Cerebrovascular Accident (CVA) based on the age group, containing 88,861 records referring to individuals not diagnosed with stroke and only 1,975 for those with a positive diagnosis. Finally, we performed a data balance. That is, the final database obtained has 1108 records, equally divided into positive and negative diagnoses.

3.2 Understanding the Problem and Conceptual Selection of Attributes

Understanding the problem domain is an important step in the process of knowledge discovery, which aims to build more representative learning models. The previous understanding allows us to observe the complexity of each problem and the discovery of useful and non-obvious knowledge about the domain considered. Also, due to the high dimensionality of the PNS 2019 database, the conceptual selection of attributes is a relevant strategy. In Figure 1, it is possible to see the conceptual model that was developed for this study. The model was built using the CAPTO method, recently proposed in Zarate et al. (2023).

The method is an approach that proposes the capture of knowledge, both explicit and tacit, for the understanding of a problem domain prior to the application of machine learning algorithms to reduce dimensionality, selecting the attributes that best represent the problem domain. Table 1 shows the selected attributes considering the conceptual model constructed by the CAPTO method.

3.3 Pre-Processing and Data Preparation

Due to the high rate of missing data in the selected attributes, it was necessary to perform combinations between attributes to minimize the impact caused by them. For this, 6 new attributes were created they are:

BMI: Body Mass Index, based on the attributes Weight (P00103) and Height (P00403). The categorization was carried out based on the ranges defined

[1] https://avc.org.br/

[2] https://transparencia.registrocivil.org.br/painel-registral/especial-covid/

[3] https://www.heart.org/en/

[4] https://www.pns.icict.fiocruz.br/

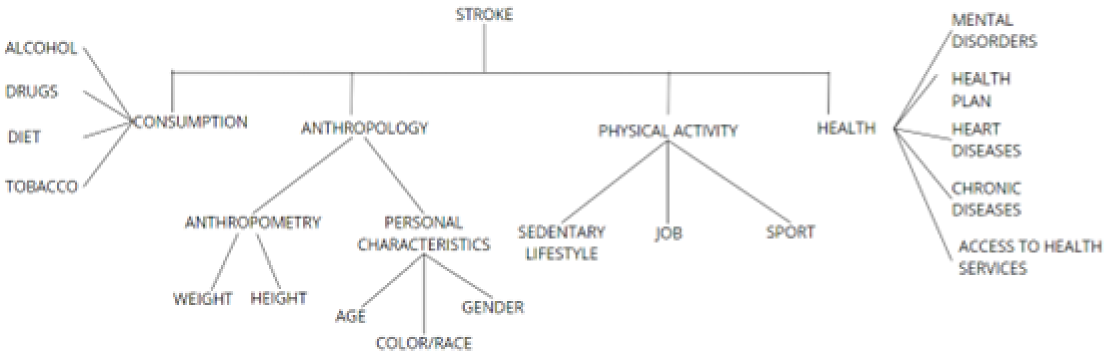

Figure 1: Vertical Conceptual Model.

Table 1: Variables Extracted from the PNS 2019 database from the Concept Map.

Variables Extracted from the PNS 2019 database from the Concept Map		
Dimension	**Aspect**	**PNS Variables**
Anthropology	Personal Characteristics	C8, C6, C9
	Anthropometry	P1a, P4a
Consumption	Diet	P9a, P15, P18, P20b, P25a, P26a, P26b
	Alcohol	P27, P28a, P29
	Drugs	No information available in the database
	Smoking	P50, P51, P52, P53, P54, P54a, P54b, P54c, P54d, P54e, P54f, P54g
Health	Access to health services and Health insurance	100102
	Mental Illnesses	Q92, Q110a
	Heart Disease	Q2a, Q63a
	Chronic Diseases	Q30a, Q60, Q63a, Q68
	Sedentary lifestyle	P45a, P45b
Physical Activity	Work	P38, P39, P40, P41, P42, P43, P44, E17, E19
	Sport	P34, P35, P37, P36

by the World Health Organization (WHO) where BMI = Weight/Height^2.

- If BMI <18.5: "Underweight"
- If 18.5 <= BMI <24.99: "Normal weight"
- If 35 <= BMI <39.99: "Obesity grade II"
- If 25 <= BMI <29.99: "Overweight"
- If BMI >= 40: "Obesity grade III"

Daily_Smoking: To create the attribute containing information on the daily consumption of products containing tobacco, the following PNS attributes were used: P05402 (Industrialized Cigarettes), P05405 (Straw or Hand-Rolled Cigarettes), P05408 (Clove or Bali Cigarettes), P05411 (Pipes), P05414 (Cigars or Cigarillos), P05417 (Hookah (sessions)), P05421 (Other). The procedure applied consisted of the sum of response values between the attributes (Daily_Smoking= P05402 + P05405 + P05408 + P05411 + P05414 + P05417 + P05421). When an empty record (NaN) is found, it is considered zero (Non-smoker), so the individual whose result of the sum of all the columns associated with it was equal to zero was considered a non-smoker. The standardization proposed by the Government of Canada[5], entitled "Tobacco Use Statistics – Terminology", was used as a basis for categorization. The procedure adopted is given below:

- If cigarettes == 0: "Non-smoker"
- If 11 <= cigarettes <= 20: "Moderate Smoker"
- If 0 <cigarettes <= 10: "Light Smoker"
- If cigarettes >20: "Heavy Smoker"

Weekly_Alcohol_Intake: The strategy to create this attribute combines two answers: the weekly frequency of consumption (P02801) and the number

[5] https://www.canada.ca/en.html/

of doses consumed per occasion (P029). When the weekly frequency is not available, but there is information that the individual in question consumes alcohol less than once a month, the developed function adjusts the amount consumed, dividing it by four, to estimate a weekly average of consumption. It was used to categorize the standards defined by the National Institute on Alcohol Abuse and Alcoholism (NIAAA)[6].

For the Female Gender

- If Doses == 0: "Don't drink"
- If 1 <= Doses <= 7 servings: "Low consumption"
- If 8 <= Doses <= 14 servings: "Moderate Consumption"
- If Doses >14: "High Consumption"

For the Male Gender

- If Doses == 0: "Don't drink"
- If 1 <= Doses <= 14 servings: "Low Consumption"
- If 15 <= Doses <= 28 servings: "Moderate Consumption"
- If Doses >28: "High Consumption"

Weekly_Moderate_Physical_Activity and **Weekly_Intense_Physical_Activity:** Using the WHO definition of moderate (M.Activity_Weekly = ([P04001*(P04101*60)] + P04102) + ([P042*(P04301*60)] + P04302)) and intense physical activity (I.Activity_Weekly = ([P035*(P03701*60)] + P03702) + ([P03904*(P03905*60)] + P03906)), attributes regarding the weekly minutes dedicated to these activities were created. The following attributes were used: P035 (Days per week spent exercising, practicing physical exercise or sport), P037 (Duration in hours and minutes – P035), P039 (Days, hours and minutes per week spent by the interviewee doing heavy activities), P04001 (Days per week spent walking or cycling), P041 (Duration in hours and minutes – P04001), P042 (Days per week that involve commuting to carry out usual activities) and P043 (Duration in hours and minutes – P042). For categorization, the activity guide published by the Federal Government in 2021[7] is used as a reference.

For Moderate Physical Activity

- If 0 <= Minutes <= 149: "Level 1"
- If 150 <= Minutes <= 299: "Level 2"
- If 300 <= Minutes <= 449: "Level 3"
- If 450 <= Minutes <= 599: "Level 4"
- If Minutes >= 600: "Level 5"

For Intense Physical Activity

- If 0 <= Minutes <= 74: "Level 1"
- If 75 <= Minutes <= 149: "Level 2"
- If 150 <= Minutes <= 224: "Level 3"
- If 225 <= Minutes <= 299: "Level 4"
- If Minutes >= 300: "Level 5"

Working_journey: It refers to the number of hours worked weekly. It resulted from the sum of the values between the attributes E017 (Total hours worked per week in the main job) and E019 (Total hours worked per week in other jobs). If both instances were empty (NaN), this value was considered zero, so the final attribute was not affected and there was no change in the data used. In other words, if all the attributes used were null, the final attribute would have a balance equal to zero, meaning that this person is not employed. For its categorization, World Health Organization (WHO) recommendations were used as basis.

- If Hours == 0: "Not employed"
- If 41 <= Hours <= 54: "Excessive Working Hours"
- If Hours <= 40: "Normal Working Hours"
- If Hours >= 55: "High-Risk Excessive Working Hours"

Diet_Classification (Score): Weights were assigned to each food item of the respondent and were based on its importance for constructing a diet that prevents the occurrence of cardiovascular diseases. Thus, regular consumption of fruits (P018) and vegetables (P00901) was associated with reduced risk of stroke due to their high content of antioxidants, fiber, and micronutrients in Dauchet et al. (2006). Regular consumption of fish (P015) is also responsible for cardiovascular health due to the anti-inflammatory effects as it is a food rich in fatty acids according to Mozaffarian and Rimm (2006). On the other hand, excessive consumption of sweets (P02501), fast food (P02602), and soft drinks (P02002) is related to an increased risk of obesity, hypertension, and cardiovascular diseases Malik et al. (2006). To conclude the treatment of these attributes, the score attribute was created, which corresponds to the sum of the products between the frequency of consumption of each food and its respective weight. To categorize this attribute, cut-off values were defined in the score attribute to separate the categories related to the type of food.

[6] https://www.niaaa.nih.gov/

[7] https://www.gov.br/saude/pt-br/composicao/saps/ecv/publicacoes/guia-de- atividade-fisica-paraBrazilian population/view/

- If Score <= 0: “Unhealthy Diet”
- If Score >5: “Healthy Eating”
- If 0 <Score <= 5: “Unhealthy Diet”

In all the attributes created, the OrdinalEncoder encoding method was applied, to encode categorical attributes in which their elements have a hierarchy among themselves. The rest of the attributes were already coded by the way the answers were arranged in the PNS 2019 questionnaire, maintaining the original coding. The only change made was in the coding of the attribute ”Salt Consumption’”. An inversion in the order was made, where the value ”5”, which originally corresponded to the lowest level, now represents the highest consumption, while the value ”1”, which previously corresponded to the highest level, now represents the lowest consumption.

After the steps described above, the resulting database was divided into a new database referring to the age group of 40 to 59 years considered in this study. A correlation and entropy analysis are applied to analyze respectively the existence of attributes with little variance and information. In other words, identify attributes that present little information and attributes that are highly related to the target attribute, to avoid direct classification of a dominant attribute. After this procedure, it was not possible to eliminate attributes based on these analyses, because the entropy values were very close, making it difficult to establish a cut-off value and the correlation analysis. It means that no attribute had a high correlation with the class attribute.

Finally, it was checked if there were inconsistencies after coding, in order to identify instances that have the values of all the identical attributes, but belonging to different classes. Records that met this condition were eliminated from both databases. Table 2 describes the attributes present in the databases after pre-processing and data preparation.

3.4 Multi-Objective Genetic Algorithm

Through all the steps applied as conceptual selection, pre-processing and data preparation, the original dataset that contained 1088 attributes gave rise to a dataset with 17 attributes. As a final step to obtain the learning model, the new dataset was submitted to an attribute selection process based on a multi-objective genetic algorithm (GA), NSGA-II (Non-dominated Sorting Genetic Algorithm II). The objectives were to select the least number of attributes, with the lowest error rate in the decision tree-based classification model. The algorithm implementation was carried out in Python, with the DEAP library (https://github.com/DEAP).

3.4.1 Coding of Individuals

Candidate individuals (containing selected attributes) were treated in a binary manner, assembling the chromosomes with values 0 and 1. The value One represents the presence of the attribute and the value Zero represents the absence of it. The chromosome is represented by 17 positions.

3.4.2 Objective Function

For this study, two aspects of the prediction problem for the data set were considered by the objective function Lowest Percent Error Rate of Decision Tree Classifier Prediction: the accuracy of the model, and the smallest possible number of characteristics.

Lowest Percent Error Rate = 1 – Accuracy = 1 – TP) / TP) + FP), in which TP is True Positive and FP is False Positive.

As a tiebreaker between two individuals, the one with the highest fitness was chosen, which consequently means the one with the lowest error rate of Precision, followed by the one with the smallest number of characteristics. To increase the robustness and reliability of the result, a cross-validation was performed for each evaluated individual, with the number of folds equal to 10.

3.4.3 Defining the Size of the Population

Preliminary tests were carried out to define which parameter value ranges would be the most appropriate to carry out the experiments and to establish a stopping criterion based on the number of generations. This was done to have a comprehensive understanding of the experiments in relation to the computational effort required.

During the preliminary experiments, significant changes in the individuals of the population between generations were analyzed, seeking to observe the number of generations from which the convergence of the results is observed. If there were no significant changes among individuals in the population, this could indicate that the results were in the direction of convergence. The values considered adequate for the population size were 200, 400, and 600. For the number of generations, the values of 200 and 400 were chosen.

Other parameters of the genetic algorithm, such as the crossover probability (Pc), ranged between 0.7 and 0.9. The mutation probability (Pm) was set at 0.1, since the goal of the preliminary tests was to find a range of values for the population size and the number of generations. It is worth noting that these values used as experimental parameters were based on

Table 2: Attributes in the Database After the Pre-Processing steps.

Attribute	**Attribute description**
Gender	Refers to the gender of the interviewee.
Color/Race	Refers to the interviewee's color/race.
Salt_Consumption	Interviewee's perception of their salt consumption.
Heavy Domestic Work	If the interviewee does heavy cleaning in their domestic activities.
Hypertension	Whether the respondent has been diagnosed with hypertension.
Diabetes	Whether the respondent has been diagnosed with diabetes.
High Cholesterol	Whether the respondent has been diagnosed with high cholesterol.
Heart Disease	Whether the respondent has been diagnosed with any heart disease.
Depression	Whether the respondent has been diagnosed with depression.
Mental Illness	Whether the interviewee has been diagnosed with a mental disorder.
BMI_category	Respondent's Body Mass Index category.
Working Hours	Level of hours in the interviewee's working day.
Smoker_category	Interviewee's level of tobacco consumption.
Weekly alcohol category	Interviewee's weekly alcohol consumption.
Diet_Classification	Type of diet of the interviewee.
Category_moderate_cent	Moderate physical activity level practiced by the interviewee.
Category_vigorous_cent	Interviewee's level of intense physical activity.
STROKE	Target column: Whether the respondent has been diagnosed with stroke.

the articles by Bento and Kagan (2008), Santos et al. (2018), Oh et al. (2004), and Fernández et al. (2019)

After carrying out these experiments, the population size was defined as ranging between 200 and 400. The Table 3 shows the parameters and their adjusted values for the global experiments.

4 RESULTS AND DISCUSSION

A total of 240 experiments were conducted, representing the different possible combinations between the 20 seeds and the elements listed in Table 3. For each of the seeds, 12 experiments were carried out, ensuring that all the parameterized possibilities for that seed were tested. Different seeds were used to expand the search space where the algorithm began to track its analysis through the genetic space of individuals. For each experiment, a set of 10 non-dominant candidates (hall of fame) was found. The most frequent attribute is gender, and the least frequent but not less important, is ethnicity. It is worth noting that frequency is merely a heuristic for the creation of a synthetic individual and does not influence whether that attribute is considered important for the analyzed cause or not.

To evaluate the contribution of each attribute more frequently, chromosomes were created and used to build learning models based on Decision Trees. For the training process, cross-validation with k = 10 was used. The best results were achieved with 14 to 17 attributes, from the attribute Gender to Hypertension. After testing the synthetic chromosomes to determine the best results, we concluded that the chromosome containing 14 attributes achieved the highest F1 score. These attributes are: Gender, Weekly Alcohol Category, Diabetes, Physical ActivityI Category, Diet Classification, Smoking Category, Mental Illnesses, Heavy Housework, Physical ActivityM Category, Salt Intake, High Cholesterol, Depression, Heart Disease, and Hypertension.

After discovering the best attributes, a decision tree was created based on these attributes. Consequently, the following values were obtained for the evaluation metrics: average precision, average accuracy, average sensitivity, and average F1 score: 0.6955, 0.6931, 0.6955, and 0.6919, respectively. The model generated earlier with all the attributes achieved an average accuracy of 0.65. For comparative purposes, only this metric will be used. Therefore, we can conclude that the genetic algorithm produced better results with fewer attributes, demonstrating its effectiveness.

In Table 4 , it is possible to observe that the absence of heart disease as well as the presence of con-

Table 3: Parameters of the Genetic Algorithm for the Experiments.

Parameters	**Values**
Population Initialization	Random
Representation	Binary
Crossover Operator	Two Points
Crossover Probability (PC)	70%, 90%
Mutation Operator	One Point
Mutation Probability (PM)	1%
Population Size	[200, 400, 600]
Number of Generations	[200, 400]
Crossover Selection Method	Tournament = 2
Composition of the New Generation	Not dominated individuals
Stop Criterion	Number of generations

Table 4: Main Rules and their Respective Coverage.

Extracted Rules	**Coverage**
If the individual does not have hypertension, but has heart disease, high cholesterol, depression, and does not smoke, then they do not have a stroke.	[102 cases without stroke, 86 cases with stroke]. Total: 188 cases.
If the individual has hypertension, heart disease, depression, does not smoke, and their level of intense physical activity is low, then they have a stroke.	[83 cases without stroke, 146 cases with stroke]. Total: 229 cases.
If the individual has hypertension, heart disease, depression, does not smoke, and their level of intense physical activity is high, then they have a stroke.	[46 cases without stroke, 148 cases with stroke]. Total: 194 cases.
If the individual does not have hypertension, heart disease, high cholesterol, depression, and does not smoke, then they do not have a stroke.	[74 cases without stroke, 14 cases with stroke]. Total: 88 cases.
If the individual does not have hypertension, heart disease, high cholesterol, depression, does not smoke, and their level of intense physical activity is low, then they do not have a stroke.	[9 cases without stroke, 5 cases with stroke]. Total: 14 cases.

ditions such as depression, hypertension, smoking habits, and the practice of intense physical activity are related to the rules used to describe the profile of individuals diagnosed with Stroke. As for the rules associated with the profile of individuals with negative diagnoses, a variation is noticeable in the presence and absence of certain attributes, for example, tobacco consumption and high cholesterol. In addition, hypertension is indeed a condition that is not associated with these individuals; however, the presence of the attributes heart and mental diseases is guaranteed.

Thus, for positive diagnoses, the attributes present in the rules are more consistent in the sense that there is no variation in their state. In other words, there are no discrepancies regarding the presence or absence of these attributes in the classification of instances, which enhances the predictability and the description of their profile. On the other hand, the rules associated with negative diagnoses exhibit greater variation in the state of the attributes, meaning that the attributes may be present in some instances and absent in others. This variability makes it more challenging to accurately extract patterns and characteristics for this group of individuals.

5 CONCLUSIONS

The use of Genetic Algorithms (GA) in this study proved to be advantageous because it allowed the exploration of a wide variety of combinations of attributes, generating consistent rules for stroke prediction. One of the main advantages of GA is its ability to find near-optimal solutions in complex search spaces, which has proven effective in identifying relevant patterns. The rules extracted, such as those related to hypertension, heart disease, and lifestyle habits, provided a clearer picture of the factors that contribute to stroke risk, showing that GA can be a useful tool to support medical diagnoses.

However, parameterizing the algorithm was a significant challenge. The choice of parameters, such as the number of generations and the mutation rate, directly impacts the effectiveness of GA. Because NSGA-II scales rapidly depending on the parameters, it was not possible to exploit the entire search space efficiently. This limited the number of subjects assessed, which could be improved with methods that allow better exploration of the start space. This challenge reflects the need for further investigation into the best parameter setting to maximize coverage of individuals of interest.

Another important point was the transformation of categorical data to numerical data, a necessary action for the use of the decision tree. Some inaccuracies brought by the transformation may impact the quality of the extracted rules. Still, in this work, it is minimized by using standard measures from reliable sources to preserve data quality. Thus, the interpretation of the results is not so affected by the categorization performed on the numerical data. Even so, the use of a classification algorithm capable of dealing directly with categorical data, in Python, without the need for transformation, would be a solution to reduce these inaccuracies and improve the accuracy of the model. Finally, the rules extracted were consistent with the theory about the disease, as demonstrated by the rules that relate hypertension, heart disease, and the practice of physical activities to the occurrence of stroke. However, the limitations imposed by the dataset, which did not include a wider range of 17 attributes, restrict the predictive potential of the model.

For future work, the first focus would be the use of classification algorithms or techniques that, in conjunction with genetic algorithms, can better explore the search space in an efficient and computationally feasible manner. Ideally, this would include the possibility of covering the entire search space or most of it, ensuring the discovery of more varied solutions.

Another crucial point would be the use of algorithms that allow working directly with categorical data, without the need to transform them into numerical data. This approach avoids the inaccuracies introduced by the transformation and contributes to more reliable results. Furthermore, a data transformation methodology that preserves as much information as possible from the original dataset, especially records related to the occurrence of stroke, could be developed. Such a methodology would help to preserve the richness of the data, providing more robust and reliable analyses.

ACKNOWLEDGEMENTS

The authors thank The National Council for Scientific and Technological Development of Brazil (CNPQ); The Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) (Grant PROAP 88887.842889/2023-00 – PUC/MG, Grant PDPG 88887.708960/2022-00 – PUC/MG - INFORMATICA and Finance Code 001); Minas Gerais State Research Support Foundation (FAPEMIG) under grant number APQ-01929-22, and the Pontifical Catholic University of Minas Gerais, Brazil.

REFERENCES

Bento, E. P. and Kagan, N. (2008). Algoritmos genéticos e variantes na solução de problemas de configuração de redes de distribuição. *Revista Controle & Automação*, 19(3):302–305.

Dauchet, L., Amouyel, P., Hercberg, S., and Dallongeville, J. (2006). Fruit and vegetable consumption and risk of coronary heart disease: A meta-analysis of cohort studies1. *The Journal of Nutrition*, 136(10):2588–2593.

Dritsas, E. and Trigka, M. (2022). Stroke risk prediction with machine learning techniques. *Sensors*, 22(13).

Fernández, C., Pantano, N., Godoy, S., Serrano, E., and Scaglia, G. (2019). Parameters optimization applying monte carlo methods and evolutionary algorithms. *Revista Iberoamericana de Automática e Informática Industrial*, 16(2):89–99.

Malik, V. S., Schulze, M. B., and Hu, F. B. (2006). Intake of sugar-sweetened beverages and weight gain: a systematic review. *The American journal of clinical nutrition*, 84 2:274–88.

Mozaffarian, D. and Rimm, E. (2006). Mozaffarian d, rimm eb. fish intake, contaminants, and human health: evaluating the risks and the benefits. jama 296, 1885-1899. *JAMA: the journal of the American Medical Association*, 296:1885–99.

Noche, R., Biffi, A., Sansing, L., Shoamanesh, A., Benavente, O., Falcone, G., and Sheth, K. (2020). Abstract 156: Recurrent stroke in middle-aged lacunar stroke survivors: Understanding risk factors and vulnerability in an important target population. *Stroke*, 51.

Oh, I.-S., Lee, J.-S., and Moon, B.-R. (2004). Hybrid genetic algorithms for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1424–1429.

Santos, B. C., Nobre, C. N., and Zárate, L. E. (2018). Multiobjective genetic algorithm for feature selection in a protein function prediction context. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8.

Yousufuddin, M. and Young, N. (2019). Aging and ischemic stroke. *Aging*, 11.

Zarate, L., Petrocchi, B., Maia, C. D., Felix, C., and Gomes, M. P. (2023). Capto - a method for understanding problem domains for data science projects. *Concilium*.