

Predictive models in Alzheimer's disease: an evaluation based on data mining techniques

Laberiano Andrade-Arenas¹, Inoc Rubio-Paucar², Cesar Yactayo-Arias³

¹Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú

²Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú

³Departamento de Estudios Generales, Universidad Continental, Lima, Perú

Article Info

Article history:

Received Jan 22, 2024

Revised Feb 22, 2024

Accepted Feb 25, 2024

Keywords:

Alzheimer's

Data mining

Decision tree

Predictive models

SEMMA methodology

ABSTRACT

The increasing prevalence of Alzheimer's disease in older adults has raised significant concern in recent years. Aware of this challenge, this research set out to develop predictive models that allow early identification of people at risk for Alzheimer's disease, considering several variables associated with the disease. To achieve this objective, data mining techniques were employed, specifically the decision tree algorithm, using the RapidMiner Studio tool. The sample explore modify model and assess (SEMMA) methodology was implemented systematically at each stage of model development, ensuring an orderly and structured approach. The results obtained revealed that 45.00% of people with dementia present characteristics that identify them as candidates for confirmation of a diagnosis of Alzheimer's disease. In contrast, 52.78% of those who do not have dementia show no danger of contracting the disease. In the conclusion of the research, it was noted that most patients diagnosed with Alzheimer's are older than 65 years, indicating that this stage of life tends to trigger brain changes associated with the disease. This finding underscores the importance of considering age as a key factor in the early identification of the disease.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Cesar Yactayo-Arias

Departamento de Estudios Generales, Universidad Continental

Avenida Alfredo Mendiola 5210, Lima, Peru

Email: cyactayo@continental.edu.pe

1. INTRODUCTION

In the current international scenario, Alzheimer's disease has emerged as a monumental health challenge. Demographic aging and the consequent increase in the prevalence of neurodegenerative diseases have transformed Alzheimer's disease into a global public health problem. This phenomenon raises crucial questions about how to effectively address and manage a condition that not only affects the individual but also places significant pressure on healthcare systems and society at large. The urgency of addressing this complex problem has led to the exploration of innovative approaches, with data mining and predictive modeling at the forefront of scientific research [1], [2].

Alzheimer's disease, despite its prevalence and devastating impact, faces significant obstacles in its early diagnosis. The problem is compounded by the lack of clear biological markers and definitive diagnostic tests according to the initial characteristics of the anomaly. Early symptoms, such as mild memory loss and subtle changes in behavior, are often mistaken for the normal aging process, resulting in a delayed diagnosis [3], [4]. This delay impedes the timely implementation of treatment and management strategies, limiting the effectiveness of interventions and exacerbating the consequences for both patients and their families. In addition, the emotional and financial burden associated with advanced care for patients with Alzheimer's

disease poses additional challenges for healthcare systems and society as a whole [5]. The complexity of Alzheimer's disease is accentuated by the multiplicity of factors that contribute to its development. Research indicates that genetics, advanced age, the presence of certain genes, and exposure to vascular risk factors, such as hypertension and diabetes, may increase the likelihood of developing the disease. In addition, the role of environmental factors, such as sleep quality and cardiovascular health, is becoming increasingly evident. Understanding these causes is crucial for the development of effective predictive models that address the variability in the manifestation of Alzheimer's disease and enable personalized interventions [6].

The relevance of this research is magnified when considering the widespread consequences of Alzheimer's disease on individuals, families, and health care systems. A robust predictive approach has the potential to radically change the course of health care in the context of Alzheimer's disease. Early anticipation of disease will not only improve the efficacy of medical interventions but also enable proactive planning for care and support for patients and their families. The importance of this research is reflected not only in its ability to improve the quality of life of those affected but also in its contribution to the sustainability and efficiency of healthcare systems in the treatment of Alzheimer's disease. Furthermore, the research is justified in its potential to advance the understanding of the underlying factors and biological pathways involved in Alzheimer's disease [7], [8]. The exploration of predictive patterns will not only address the problem of late diagnosis but will also provide valuable insights into the complexity of this disease. The relevance of this contribution is not limited to clinical care, but extends to the research arena, laying the groundwork for future developments in targeted therapies and preventive strategies. To address this challenge and mitigate the number of premature deaths associated with these risk factors, it is proposed to develop an investigation using data mining techniques [9]. Applying the sample explore modify model and assess (SEMMA) methodology will be selected as the conceptual framework to anticipate the probability of developing Alzheimer's disease and to understand which clinical variables are the most influential. The decision tree algorithm will be employed through the RapidMiner Studio tool to predict the probability of a patient contracting Alzheimer's disease, considering the most significant variables extracted from the selected database. This proposal seeks not only to provide accurate predictions but also to shed light on the key factors that contribute to the development of the disease, thus providing valuable information for prevention and early care.

This study has an ambitious but essential goal: to develop and evaluate advanced predictive models based on data mining techniques specifically tailored for Alzheimer's disease. These models aim to go beyond diagnosis, seeking to identify subtle patterns and hidden risk factors that may provide crucial clues about disease progression. Achieving this goal will not only advance the accuracy of early detection but also lay the groundwork for personalized treatment and prevention strategies, setting a new standard in the comprehensive approach to Alzheimer's disease internationally.

2. LITERATURE REVIEW

This literature review dives into the exciting field of data mining applied to brain health, with a specific focus on predicting the likelihood of Alzheimer's disease. Its purpose is to analyze research conducted by experts and scientists in this field, highlighting their valuable contributions and, at the same time, identifying limitations and opportunities for advancing this crucial facet of health care. The combination of data mining technology and brain health has proven to be a promising approach for early and accurate detection of risk factors, opening the door to more personalized and effective care for those affected by this disease.

The study proposes to develop a new feature selection model based on mutual relationships for the detection and prediction of Alzheimer's stages using magnetic resonance imaging. This approach seeks to avoid redundancies by considering the mutual relationship between features. Four machine learning classifiers were employed: decision trees, support vector machines (SVM), k-nearest neighbors, and naive Bayes. Ten-fold cross-validation was applied to prevent overfitting. The results revealed that SVM stands out as the most effective classifier for prediction, with an area under the curve of 0.936, an accuracy of 96.9%, a recall of 96.6%, and an F1 score of 96.8% [10]. On the other hand, another study aims to create a hybrid data mining model that integrates text mining with structured data to improve the diagnosis of dementia. A total of 605 medical records with 19 attributes for patients with cognitive impairment were used. A new structured attribute was generated by text mining, grouping pathological history information stored in an unstructured attribute. Classification algorithms provided predictive models for Alzheimer's disease and mild cognitive impairment. Ensemble methods were applied to improve accuracy. Results indicate that the hybrid model significantly outperforms the one based on structured data alone, achieving an accuracy of 92.6% for Alzheimer's disease and 90.2% for mild cognitive impairment [11].

In another area, the study investigates the factors associated with dementia in the elderly, using nationally representative data on older people. Seven machine learning algorithms were applied to build and

compare predictive models. A simple model-ensemble approach was used to amalgamate the results of the base models. Key factors were identified in areas such as biology, cognition, and social aspects. The results provide new evidence on factors associated with dementia in the elderly. This information will facilitate the early detection and development of preventive measures for possible signs of dementia [12], [13]. According to this research, an exhaustive analysis of the diagnosis of Alzheimer's disease is carried out in order to determine the most appropriate treatment for each patient. However, previously employed models have shown limitations when facing the predictive demands of this disease. To this end, the study focuses on a comparison between different machine learning algorithms, considering key aspects of accuracy, prediction, recall, and model training time. The results of this analysis reveal that classifier algorithms based on Bayesian approaches outperform, highlighting their ability to effectively address the diagnostic pressure associated with Alzheimer's disease. These algorithms show substantial interpretation by the model, contributing significantly to the early diagnosis of the disease [14], [15].

This is why, in another study, the implementation of a system aimed at determining the diagnosis of mild cognitive impairment through the application of deep learning techniques is proposed. The methodological approach adopted consisted of the use of a convolutional network model with dual fusion cluster graphs. In this context, the developed algorithm was applied to a public dataset related to Alzheimer's disease, yielding accuracy, sensitivity, and specificity results of 90%, 91.1%, and 94.0%, respectively. As a consequence of these findings, it is concluded that the proposed model represents a significant improvement in the diagnosis of mild cognitive impairment, consolidating itself as a valuable tool for the detection and treatment of this condition [16]. In other circumstances, the paper proposes to design a model by applying the internet of things (IoT), implemented with an eye tracking node, considering cloud-based diagnosis by using deep learning. The application of the algorithm in the model facilitates the extraction of information related to various types of eye movements to improve classification accuracy and reduce the dimension of the data. On the other hand, the technique employed can distinguish older adult patients with Alzheimer's disease, achieving an accuracy of 86%, a true positive rate of 78%, and a predictive value of 90%. The results confirmed that the algorithm implemented in the diagnosis of Alzheimer's disease has demonstrated high feasibility in solving the problem of eye tracking designed by IoT for the detection of Alzheimer's disease [17], [18].

In this research, the application of the combination of acupoints in the treatment of Alzheimer's disease is pursued, complemented by a systematic review and meta-analysis of randomized trials. To carry out this study, a method based on association rule analysis was employed, using data mining concepts and the R programming language to implement the Apriori algorithm. The evaluation of 503 association rules revealed that the combinations $\{SP6, BI10\} \geq \{HT7\}$ and $\{HT7, BI10\} \geq \{SP6\}$ were the most frequent and significant in 15 randomized controlled trials (RCTs). From these results, we conclude that the combination of acupoints, as mentioned above ($\{SP6, BI10\} \geq \{HT7\}$ and $\{HT7, BI10\} \geq \{SP6\}$), can be applied in future acupuncture protocols for the treatment of Alzheimer's disease [19], [20]. Neurodegenerative diseases, such as Parkinson's and Alzheimer's, are more prevalent in the elderly. In this research, the sunflower optimization (SFO) algorithm was used to select a specific set of features, which were subsequently processed by the Kernel extreme learning machine (KELM) to perform the classification. The results obtained indicate high diagnostic accuracy, with classification rates of 99.32% and 98.65% for the ADNI and local datasets, respectively, when considering Alzheimer's disease versus cognitively normal (CN). In the specific case of diagnosis, the accuracy reaches 99.52% and 99.45%. It is concluded that the model used has been well received, empowering medical professionals to make informed decisions in the diagnosis of these neurodegenerative diseases [21].

In this evaluation, an association rule-based learning model used as a tool for analyzing raw neuropsychological data was examined using the frequent pattern algorithm (FP-Growth). The study involved the evaluation of complex data from 84 confirmed cases of Alzheimer's disease patients and 294 participants, taking into account variables such as age and race. The results obtained using this algorithm indicated that the set of frequent items, categorized as mild, moderate, and severe Alzheimer's, was significantly relevant ($p < 0.001$ and $\eta^2 = 0.488$). The conclusion reached is that the application of the FP-Growth model is adequate as a tool for neuropsychological assessment, contributing to decision-making in various fields of health sciences [22]. In this research, a study was carried out to measure the cognitive and functional domains in patients with Alzheimer's disease. For this reason, optimization techniques were evaluated by applying classifier models such as random forest, boxed, naive Bayesian, and logistic regression to improve performance. The goal is to implement a convolutional neural network, considering the performance of the model using the Shannon entropy-based cutoff technique. The results indicate that the collection of images from an open access series database allows the categorization of images into three groups to extract features from each according to their texture. The combination of these features demonstrates an accuracy of 80% and 60% in the image analysis applied to groups 1 and 3, respectively [23], [24].

In the review of the fifteen articles, several limitations have been identified, among which are the lack of precision in the prediction of risks associated with Alzheimer's disease as well as the underuse of data mining techniques in the analysis of clinical data. A significant proposal for improvement would lie in the more effective incorporation of data mining and machine learning in future studies. This will allow a greater likelihood of making an accurate prediction about people who get Alzheimer's and understanding the important variables that shape this prediction. This strategy could contribute to the identification of more subtle patterns and risk factors, enabling more personalized and effective interventions to reduce the risks associated with Alzheimer's disease in patients with neurodegenerative conditions.

3. METHOD

3.1. Definition of the SEMMA methodology

The SEMMA process was developed by the SAS Institute in 2017 to be applied to the statistical prediction of large amounts of meaningful data, enabling the discovery of important patterns within the model. Therefore, this process determines an orderly and systematic structure that addresses complex projects based on data mining. This enables data professionals to implement robust and more accurate models for decision-making, according to the results obtained in the final stage [25], [26]. It is important to note that SEMMA should not be considered a methodology but rather a process. Although it is often confused with a methodology, the SAS Institute specifically defines it as a process. This process supports the implementation of software based on statistics and data mining, thus contributing to the improvement of efficiency and effectiveness in obtaining knowledge from data. Figure 1 shows a visual representation of the SEMMA process structure. The SEMMA process consists of key stages: sampling, exploration, modification, modeling, and finally model evaluation. Each of these stages has been designed to consider a representative sample of the population, using data collected from patients as a reference.

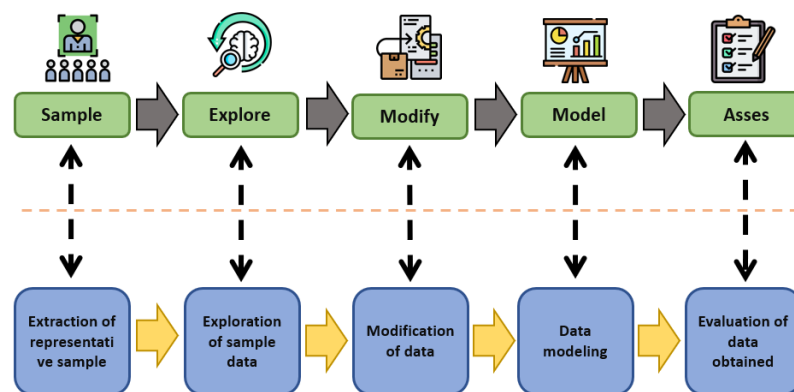


Figure 1. Stages of the SEMMA process

3.2. Stages of the SEMMA process

In this section, the stages of the SEMMA process, as defined above, have been applied to create a comprehensive predictive model. It is crucial to highlight that this phase represents the essential foundation of the model development since here we proceed to the construction based on the concepts established by the SEMMA process. The central objective of this model is to provide a solution to the problem posed, which focuses on the detection of people most vulnerable to Alzheimer's disease and the identification of the most influential variables for the model [27]. It is important to emphasize that the construction of this base is fundamental, and although it may be modified according to the needs of the project, in our case, the objectives of the project are clearly defined, which ensures a coherent development and avoids possible complications in the future.

3.2.1. Sample

The process starts with the extraction of a carefully selected sample population for the application of the proposed model to the established objectives. Therefore, the most appropriate way to obtain a sample that meets the standards of the model is to select the data at random [28]. This implies that each individual in the sample population can be selected randomly, following the essential criteria for the model. This approach is known as simple random sampling.

According to SEMMA, it is established that each sample considered for the model analysis should be associated with a sample confidence level, thus ensuring that the sample is representative and adequate for the following steps [29]. In this sense, our sample meets all the standards established in the concepts offered by SEMMA, since the sample selection process, which serves as an essential starting point for the rest of the process, has been properly consolidated.

a. Sample size

The sample size is the number of elements selected from a population to form part of a representative data set. The importance of sample size lies in its direct influence on the precision and reliability of the statistical conclusions that can be drawn about the population in question. This aspect is crucial for the understanding and fulfillment of the objectives established within a project. In the context of the project, the sample size becomes an essential characteristic that encapsulates the specific objectives. Consideration of this factor is critical, as the data collected may contain inherent errors. Consequently, strategies must be implemented to manage and reduce the margin of error associated with the sample, thus ensuring greater reliability in the inferences and results obtained during the statistical analysis [30], [31].

– Formulas for finite populations

Equation (1) deals with the selection of a representative sample in an investigation, considering a finite population. Robust statistical parameters are established, such as a confidence level of 95% and a margin of error of 3%, to ensure validity and precision in the inferences. These fundamental concepts lay the foundation for the application of the equation by a statistician, allowing the necessary sample size to be calculated. In summary, attention to these parameters contributes to the robustness of the results and ensures the representativeness required for informed decision-making in research.

$$n = \frac{N \cdot n_0}{N + n_0 - 1} \quad (1)$$

where n is the sample size, N is population size, and n_0 is initial sample size.

– Formulas for infinite populations

In statistics, an infinite population refers to a theoretical set of elements that is infinitely large. This means that the population has no defined upper limit and, in theory, contains an infinite number of individuals. This concept is used as a simplifying tool in certain statistical situations as mentioned in (2).

$$n = \frac{Z^2 \cdot p \cdot (1-p)}{E^2} \quad (2)$$

where Z is critical value of the confidence level, p is estimated proportion, E is margin of error, and standard error of the mean (SEM).

Measures the expected variability between the means of different samples taken from the same population. A lower standard error of the mean (SEM) indicates a more accurate estimate of the population mean. In (3), a representation that facilitates the calculation of these errors for control and evaluation is presented. The lower the SEM, the greater the reliability of the estimate of the population mean. This indicator is crucial to evaluating the consistency and precision of drawing conclusions from samples of a given population.

$$SEM = \frac{\sigma}{\sqrt{n}} \quad (3)$$

where σ is standard deviation of the population, n is sample size, and confidence interval for the mean.

It provides an estimated range where the true population mean is likely to reside with a specified level of confidence. The precision of the estimate increases as the confidence interval narrows. Equation (4) shows how this confidence interval is represented when applied to the mean. This mathematical representation is fundamental to understanding the variability associated with the estimation of the mean and offers a valuable tool to evaluate the reliability of the results obtained in the statistical analysis.

$$\text{Confidence interval} = \text{Half} \pm (\text{Critical value} \times \text{SEM}) \quad (4)$$

b. Standard error of proportion

It is similar to SEM but applied to proportions instead of means. It measures the expected variability between the proportions of different samples from the same population. A representation of the characteristics applied to the standard error of proportion is shown in (5).

$$SE_p = \sqrt{\frac{p \cdot (1-p)}{n}} \tag{5}$$

where p is proportion of the population, n is sample size, and ratio confidence interval.

Similar to the confidence interval for the mean but applied to proportions. It provides an estimated range within which the true population proportion is likely to be found with a certain level of confidence. For this purpose, (6) shows the formula by which the proportional confidence values are represented.

$$\text{Confidence interval} = \text{Proportion} \pm (\text{Critical value} \times SE_p) \tag{6}$$

In Figure 2, a graph illustrating the application of various operators for sample selection in RapidMiner Studio is presented. This process involves the application of specific statistical criteria based on the sample, intending to effectively consolidate the data representation. The information collected in this context represents a selected sample of a population, in this case, the representation of Alzheimer's patients and their main characteristics. This statistical approach provides a solid basis for further exploration and modeling of the data within the SEMMA process.

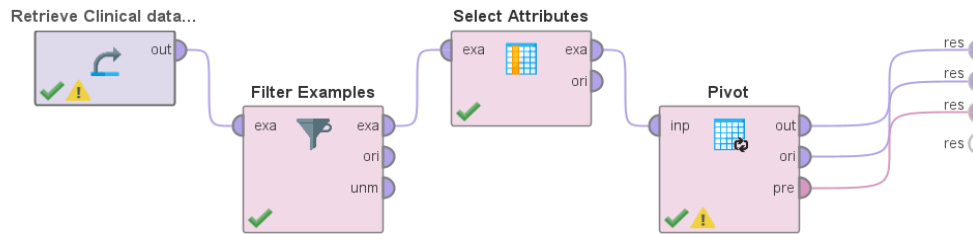


Figure 2. Sampling analysis of the data

3.2.2. Explore

In the exploration phase, the active search for patterns, trends, and outliers within the selected data is carried out. This process involves the application of advanced visual and statistical techniques, such as factor analysis, correspondence analysis, and segmentation. These tools allow us to gain a deep understanding of the project morphology and reveal emerging patterns that may not be evident at first glance [32]. Identifying trends and outliers during this stage is important, as it provides valuable initial insights into the nature of the data and the potential impact on model building. These preliminary findings not only guide the direction of the project but also facilitate informed decisions about which approaches and strategies hold the most promise for achieving the data analysis objectives [33]. Ultimately, this solid exploration phase lays the foundation for a more effective and accurate data mining process.

a. Media

The mean is a statistical metric that represents the average value of a set of data. It is obtained by adding all the values in the set and dividing the result by the total number of observations. This measure provides a fair and central representation of the data set, providing a general indicator of its trend as shown in (7).

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \tag{7}$$

where \bar{x} is the average, x_i are the individual values in the sample, and n is the sample size.

b. Median

The median is the value at the center of a set of ordered data. Unlike the mean, the median is not affected by extreme values and provides a robust measure of central tendency. Equations (8) and (9) show a representation of odd observations and the average of median values.

- The median represents the central value in an array of ordered data, dividing the set into two equal parts.
- For the number of observations (n) for odd, the median is the value at position $(n + 1)/2$ (8).
- For n that is even, the average of the median of the values at position $n/2$ and $n/2 + 1$ (9).

c. Standard deviation

The standard deviation quantifies the amplitude or variability present in a data set. It indicates how far the individual values deviate from the mean. A larger standard deviation implies greater dispersion of the data. The standard deviation measures the dispersion or variability of a data set, as mentioned in (10).

$$\sigma = \sqrt{\frac{\sum_{i=0}^n (x_i - \bar{x})^2}{n}} \tag{10}$$

where σ is the standard deviation, x_i are the individual values in the sample, \bar{x} is the average, n is the sample size.

To determine the results related to the attributes "sex," "group," and "age," which are the variables that will be used to build the proposed model, a detailed analysis has been performed using Table 1. This table provides a comprehensive description of each case, highlighting the number of valid, missing, and total cases for each particular attribute. Through this description, an accurate assessment of the completeness of the selected attributes is obtained. Accordingly, the proportion of cases with valid data and the minimal incidence of cases with missing data have been thoroughly examined, suggesting high data integrity for the attributes "sex," "group," and "age".

The mean is a reference measure that represents an average value over a large set of data. Table 2 provides a detailed description of the mean for each specified variable. In this table, key criteria such as median, variance, and standard deviation, among others, are considered. In addition, statistical percentages and the standard error applied to these concepts are included, thus providing a comprehensive understanding of the central and dispersion characteristics of the data.

Table 1. Case processing

Attributes	Cases					
	Valid		Lost		Total	
	N	Percentage	N	Percentage	N	Percentage
Sex	371	99.5%	2	0.5%	373	100,0%
Group	371	99.5%	2	0.5%	373	100,0%
Age	371	99.5%	2	0.5%	373	100,0%

Table 2. Description of the average

	Description	Statistical	Standard error
	Media	77.02	.398
	95% confidence interval for the mean	Lower limit 76.24 Upper limit 77.80	
	Stocking trimmed to 5%	76.95	
	Median	77.00	
	Variance	58.689	
Sex	Standard deviation	7.661	
	Minimum	60	
	Maximum	98	
	Range	38	
	Interquartile range	11	
	Media	.139	.127
	Kurtosis	-.424	.253
	Media	593.88	33.053
	95% confidence interval for the mean	Lower limit 528.88 Upper limit 658.87	
	Stocking trimmed to 5%	538.90	
	Median	552.00	
	Variance	405306.033	
Group	Standard deviation	636.637	
	Minimum	0	
	Maximum	2639	
	Range	2639	
	Interquartile range	873	
	Asymmetry	.949	.127
	Kurtosis	.209	.253
	Media	27.34	.191
	95% confidence interval for the mean	Lower limit 26.97 Upper limit 27.72	
	Stocking trimmed to 5%	27.82	
	Median	29.00	
	Variance	13.566	
Age	Standard deviation	3.683	
	Minimum	4	
	Maximum	30	
	Range	26	
	Interquartile range	3	
	Asymmetry	-2.366	.127
	Kurtosis	7.516	.253

3.2.3. Modify

The data collected during the modification phase of the SEMMA process is intended to address specific problems in the selected database. To accomplish this task, corrections are required, such as data cleaning, identification and management of outliers, handling of missing data, and correction of errors in the information. These actions ensure the consistency and reliability of the data, thus establishing a solid basis for moving forward in the analysis process [34], [35]. The information collected in our research will be subjected to this phase to ensure that the data obtained is reliable and of high quality, thus allowing the transition to the next stage, which is modeling.

a. Attribute selection

The use of "select attributes" allows us to identify the most important variables to incorporate into the model. In this context, it is essential to consider the relevance of these variables for the next step, which is to impute the data. This process is based on an analysis of the importance of the most representative characteristics of the selected variables, thus ensuring informed and effective data imputation.

b. Missing data handling

Missing values were identified for key variables, such as age, groups, and the delay and results of certain medical tests. Instead of eliminating the corresponding rows, we chose to impute the missing values using the median of each respective variable calculated in the previous screening process. This choice was based on the robustness of the median to outliers in medical data. For this stage, the impute operator was used to corroborate outliers in the aforementioned variables.

c. Standardization

The normalize operator is used to normalize numeric attributes in a data set by scaling attributes in numeric columns by which they are located in a particular path. This operator offers several normalization methods, including range normalization (Min-Max Scaling) and Z-score normalization (Z-Score Scaling). In the range method, the values are fitted to the specified range, while in the Z-score method, the values are scaled so that they have a mean of zero and a standard deviation of one. For the determination of this step in Figure 3, an illustration is presented highlighting the operators usable for the modification process. This is done considering certain inconsistencies within the database selected for the construction of the model.

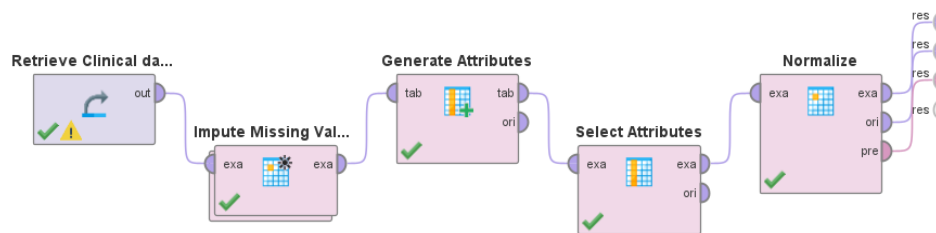


Figure 3. Modification stage analysis

3.2.4. Model

In the SEMMA framework, the "modeling" phase focuses on the creation and training of predictive models using the data previously prepared in the previous stages (sample, explore, modify). The main purpose of this stage is to develop models capable of generalizing patterns from the training data and making predictions or classifications on additional data sets [36], [37]. To build the model, we decided to use the classification-based decision tree algorithm within the RapidMiner Studio tool since we sought to determine whether the individuals in the sample contracted Alzheimer's disease according to variables such as age and sex.

a. Decision tree algorithm

Decision tree algorithms fall into the group of non-parametric machine learning algorithms used for classification and regression work. In simple terms, the decision tree makes decisions based on a series of logical questions and conditions. It starts with a root node representing the entire data set and branches into secondary nodes as successive questions are applied. Each internal node of the tree represents a question about a specific characteristic of the data, and the branches represent the possible answers to that question. The leaf nodes of the tree contain the resulting classification labels, or regression values.

– Information gain (entropy)

Information gain measures the reduction of uncertainty or entropy in a data set by performing a split based on a specific feature. The aim is to maximize this gain during the construction of the tree to obtain divisions that classify the data more effectively, as specified in (11).

$$Entropy(S) = - \sum_{i=1}^C p_i \cdot \log_2(p_i) \quad (11)$$

where S is the set of data in a node, c is the number of classes in the dataset, and p_i is the proportion of examples in the class i .

– Gini impurity

Gini impurity is a measure of how mixed the classes are in a data set. The lower the Gini impurity, the purer the nodes. During tree construction, we seek to minimize the Gini impurity to obtain splits that result in more homogeneous nodes, as described in (12).

$$Gini(S) = 1 - \sum_{i=1}^C p_i^2 \tag{12}$$

where S is the set of data in a node, C is the number of classes in the dataset, and p_i is the proportion of examples in the class i .

– Gain of information for the division (or Gini reduction)

This formula is used to evaluate the quality of a proposed split. The information gained from splitting compares the impurity of the node before the split with the weighted impurity of the child nodes after the split. A higher gain value indicates a better split. It is used to select the optimal feature to divide the data set at a given node established as stated in (13).

$$Gain(S, A) = Gini(S) - \sum_{v \in values(A)} \frac{|S_v|}{S} \cdot Gini(S_v) \tag{13}$$

where S is a set of data in a node, A is a candidate feature for splitting the dataset, $values(A)$ are the possible values of characteristic A , and $|S_v|$ is the size of the subset of data in which the characteristics A takes the value v .

Once the data preprocessing phase is completed, the next stage consists of modeling, where additional operators are applied along with those used in the previous stage. These new operators include "set role" and "select attributes," which play key roles in model configuration. In particular, "set role" helps to define the roles of the variables, while "select attributes" allows us to choose the most relevant variables for the construction of the model using the decision tree algorithm. This process allows us to classify the results obtained according to the configured model. Figure 4 visually illustrates this workflow.

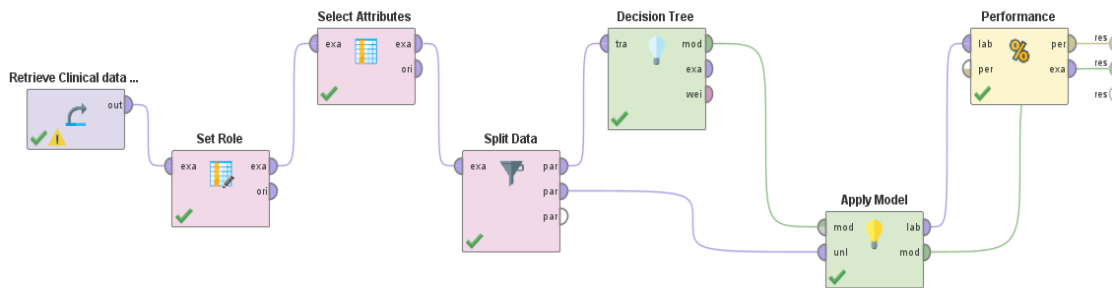


Figure 4. Modeling using decision tree algorithm

4. RESULTS

4.1. Evaluation of result

The results obtained in this study mark a significant step towards the understanding and resolution of the model. Throughout this article, we have meticulously analyzed the collected data, conducted rigorous experiments, and applied analysis related to research data extraction. In this context, we present below the main conclusions drawn from our analysis, which shed light on the trends, patterns, and relationships identified in our research. Figure 5 depicts the image of a decision tree based on the classification of Alzheimer's types described in the obtained database.

The Alzheimer's disease classification table is based on the concept of groups categorized as "converted", "demented", and "nondemented". It provides a summary of Alzheimer's cases, broken down by individuals' gender, presenting the count of individuals in each category and their respective percentages. These percentages indicate the proportion of individuals affected by Alzheimer's in relation to the total cases, differentiated by gender. Detailed information is available in Table 3.

In Figure 6, the decision tree generated by incorporating selected variables into the model, such as age, gender, and group, is depicted. The decision tree algorithm was utilized for classification, yielding the

represented outcomes. This decision tree serves as a visual representation of the process by which classification decisions are made based on the specific characteristics mentioned. The prediction results derived from the model construction are detailed in Table 4. Additionally, the evaluation metrics, including accuracy, precision, recall, and F1 score, further illuminate the model's performance and its ability to predict the likelihood of Alzheimer's development in individuals. These metrics provide a comprehensive assessment of the model's effectiveness in capturing true positive, true negative, false positive, and false negative predictions. The application of these metrics enhances the interpretability and reliability of our model, contributing valuable insights for potential clinical applications and further research in the field of Alzheimer's prediction and diagnosis.

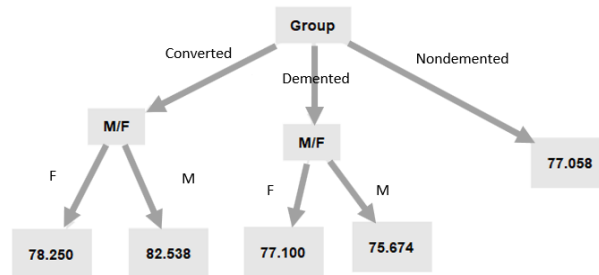


Figure 5. Decision tree on the type of Alzheimer's

Table 3. Classification of results according to the variable "Group"

Groups	Male	Count M	Female	Count F
Converted	82.538	24	78.250	13
Demented	75.674	86	77.100	60
Nondemented	77.058		190	

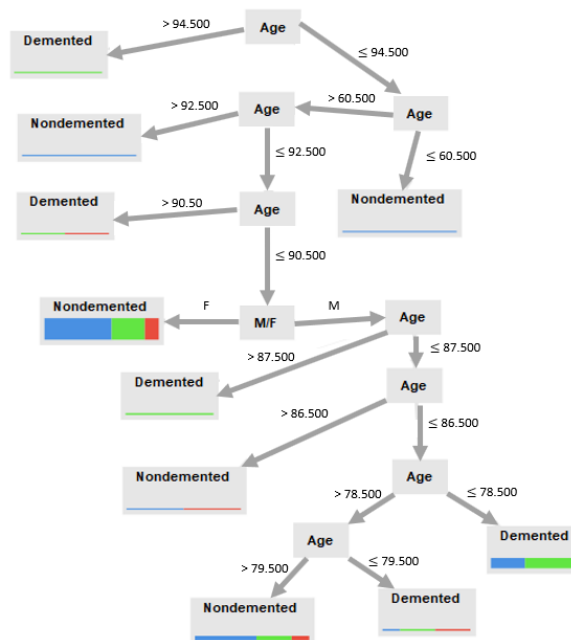


Figure 6. Tree using selected variables

Table 4. General prediction results

	True nondemented	True demented	True converted	Class precisión
Pred. Nondemented	38	26	8	52.78%
Pred. Demented	19	18	3	45.00%
Pred. Converted	0	0	0	0.00%
Class recall	66.67%	40.91%	0.00%	

4.1.1. Results analysis

In evaluating the results, it was found that 38 instances were correctly classified as "truly non-demented," suggesting a robust ability of the model to recognize normal cognitive health. However, only 18 instances were accurately identified as "truly demented." This finding highlights the importance of improving the sensitivity of the model to detect instances of dementia, as early detection can be crucial for treatment and disease management. Surprisingly, no instances were correctly detected in the "true" category, underscoring the need for further analysis to identify possible limitations in the classification approach used.

4.1.2. Evaluation metrics

a. Precision

Measuring the proportion of correctly predicted positive instances among all instances predicted as positive is critical to assessing the accuracy of a classification model. This metric, known as the true positive rate or sensitivity, provides crucial information about the model's ability to correctly identify positive instances within the data set. It is especially relevant in medical applications, such as dementia diagnosis, where early detection can significantly influence the patient's treatment and quality of life.

- For "nondemented" the precision is $38/(38+19+0)=52.78\%$.
- For "demented" the precision is $18/(18+26+0)=45.00\%$.
- For "converted" the precision is $0/(0+8+3)=0.00\%$.

b. Recall (recovery or sensitivity)

The proportion of correctly identified positive instances among all true positive instances is known as the true positive rate. This metric is critical in the evaluation of model performance in classification problems, as it provides information about the model's ability to correctly detect positive samples from the population. A high true positive rate indicates a good ability of the model to identify positive samples, while a low rate may indicate deficiencies in the model's detection ability. It is an important metric in evaluating the effectiveness of a model in detecting positive events or cases in a data set.

- For "nondemented" the recall is $38/(38+26+8)=66.67\%$.
- For "demented" the recall is $18/(18+26+3)=40.91\%$.
- For "converted" the recall is $0/(0+8+3)=0.00\%$.

4.2. Model comparison

In this section, a comparison is made between the algorithms used in data mining, evaluating their effectiveness in solving specific problems. We have chosen to represent this evaluation using a Cartesian plane, considered a distinctive means of differentiating between the various algorithms. Figure 7 offers a comparative view between models such as naive Bayes, decision tree, and rule induction. Rule induction stands out as the most outstanding algorithm with a score of 1.0, while the others present lower results. This visual approach provides us with a clear understanding of the relative capabilities of these algorithms in the tool, making it easier to identify which ones might be most effective in addressing the particular problem at hand.

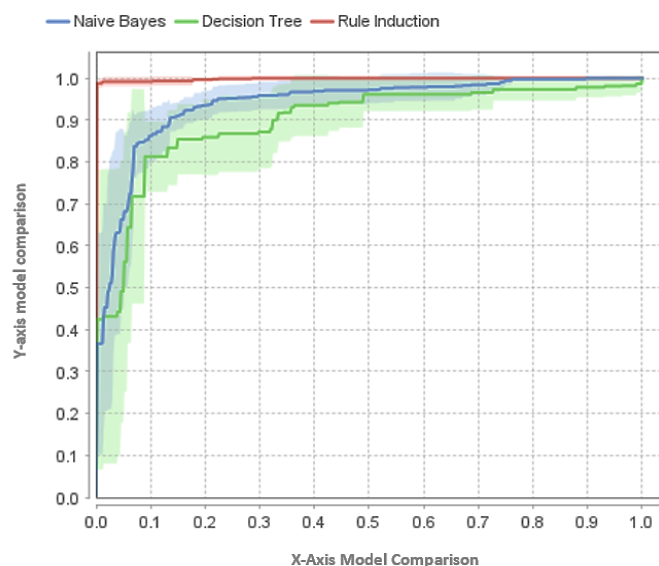


Figure 7. Comparison of algorithmic models

4.3. Comparison of methodologies

SEMMA and CRISP-DM chose the KDD method because it best suited the proposed data mining project. The KDD methodology stands out for its holistic approach, which covers all stages of data mining, from data selection and preparation to model evaluation. Additionally, KDD has proven to be quite effective in identifying valuable patterns and insights in large data sets, which was critical for our project. KDD outperformed SEMMA and CRISP-DM in terms of versatility and ability to address the specific challenges of our data mining project in a more efficient and comprehensive manner, as shown in Table 5.

Table 5. Comparison of methodologies

Comparison of attributes	SEMMA Methodology	CRIPS-DM Methodology	KDD Methodology
Main focus	Developed by the SAS Institute, SEMMA is a methodology focused on the application of statistical and modeling techniques [38].	Developed by a consortium including SPSS, CRISP-DM is a widely used methodology that defines a standard process for data mining projects [39].	KDD is a broader approach than SEMMA and refers to the entire process of knowledge discovery in databases, including the identification of useful patterns [40].
Characteristics of its phases	The "explore" phase is highlighted and places a strong emphasis on exploring data to understand the structure and relationships before modeling.	CRISP-DM has well-defined phases, including business understanding, data understanding, data preparation, modeling, evaluation, and deployment.	KDD encompasses various techniques and methodologies, making it broader and less specific than SEMMA or CRISP-DM.
Lifecycle	SEMMA follows a linear sequence in which each phase is performed after the previous one.	Similar to KDD, CRISP-DM follows an iterative approach, allowing adjustments and improvements based on feedback.	KDD is presented as an iterative process in which the results obtained can feed back into previous phases.

5. DISCUSSION

In the results of our research, it is evident that Alzheimer's-related dementia issues have been effectively addressed through data mining and the use of decision tree algorithms. A certain alignment is identified with a previous study focused on predicting Alzheimer's from magnetic resonance images using decision trees, support vector machines, nearest neighbors, and naive Bayes [10]. However, our model differs in the obtained results, emphasizing that the support vector machine algorithm stands out as the most significant classifier in such predictions. In another study with higher concordance, data mining was employed by integrating structured texts to enhance dementia diagnosis, successfully identifying 92.6% of confirmed Alzheimer's cases. Nevertheless, certain differences are observed compared to our model, which utilizes a hybrid approach without a clear specification of the concept [11].

In the following research, there is some concordance with our proposed model as machine learning algorithms are applied. However, this study focuses on predicting factors associated with the onset of Alzheimer's in the elderly, according to the authors [12], [13]. Another study aligns with our results by conducting a thorough diagnosis of the disease to determine personalized treatments, focusing on key concepts such as accuracy, prediction, recall, and training time [14], [15]. Similarly, there is alignment with a study that implemented a computer system to determine the diagnosis of mild cognitive impairment using deep learning techniques, successfully having the model correctly anticipate mild impairment and distinguishing it from our approach based on other data mining techniques [16]. On the other hand, in research that implemented an IoT-based model utilizing the cloud and deep learning techniques with eye tracking, there is not a meticulous match with our model, as it does not emphasize the Alzheimer's prediction goal [17], [18].

In the authors' research [19], [20], the combination of data mining techniques, with an emphasis on association rule analysis, is highlighted. This relates to our research, as these techniques could be applied in future investigations to tailor treatments for each patient. Conversely, in another study on neurodegenerative diseases such as Parkinson's and Alzheimer's, machine learning algorithms were used, but the difference lies in that the diagnosis obtained was more focused on extracting specific features for model classification [21]. Another study related to the results obtained within our model presents some differences in the tools and algorithms used but shares the same objective of identifying patients with Alzheimer's disease [22]. Finally, in research that aimed to implement a convolutional network, no substantial match is found with the goal of our research, as it seeks the categorization of images into groups using certain algorithms to measure the cognitive and functional domains of Alzheimer's patients [23], [24].

6. CONCLUSION

In this research, we delve into the intricate issue of Alzheimer's disease, whose symptoms, linked to decreased brain activity, include a marked slowdown in cognitive function and memory loss. The main objective of our study was to predict the probability of a patient developing Alzheimer's, and this challenge was met with the greatest scientific Rigor, exploring in depth the development of the subject. The application of the SEMMA methodology emerged as an essential component of our scientific approach. Each stage of SEMMA became a crucial tool for extracting and analyzing the most significant variables from the database, guiding us through rigorous statistical processes. Our model revealed a decrease in the number of Alzheimer's cases in the database, although a significant percentage of cases associated with the disease persist. Our study suggests that a higher proportion of individuals associated with the disease should use the developed model, as it has yielded positive results for controlling the illness. This finding underscores the need for continuous adjustments to progressively reduce Alzheimer's cases over time. We also found a correlation between the interpretation of our model and the models proposed by previously researched investigators. The application of advanced machine learning algorithms meant a substantial advance in the predictive capacity of our model, with the identification of key variables such as gender, age and group that contributed fundamentally to the accuracy and reliability of our predictions. As we celebrate our achievements, we recognize the inherent limitations that require attention and further research to refine our model. The dynamics of the medical context lead us to consider the possible inclusion of new variables or the expansion of the sample to improve the robustness of our approach. On the horizon of future research, we propose exploring new variables and data sources that enrich the complexity of the model, providing a more holistic understanding of the factors influencing the probability of contracting Alzheimer's, which future studies could explore, taking our research criteria as a reference. External validation of the model in different patient cohorts is essential to assess its generalization and applicability in different clinical settings. Finally, this study has achieved significant advances in predicting the probability of developing Alzheimer's, highlighting the importance of the variables identified and the applicability of the SEMMA methodology. At the conclusion of this research, we maintain the vision that our work will lay the foundation for future research that will contribute to the understanding and effective management of Alzheimer's disease. Our findings provide conclusive evidence that this disease is strongly associated with the aging process of the individual or with familial inheritance, which triggers the illness.




REFERENCES

- [1] Q. Pan, S. Wang, and J. Zhang, "Prediction of Alzheimer's disease based on bidirectional LSTM," *Journal of Physics: Conference Series*, vol. 1187, no. 5, Apr. 2019, doi: 10.1088/1742-6596/1187/5/052030.
- [2] R. Bin-Hezam and T. E., "A machine learning approach towards detecting dementia based on its modifiable risk factors," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, 2019, doi: 10.14569/IJACSA.2019.0100820.
- [3] R. S. Nelson *et al.*, "Neurodegenerative pathologies associated with behavioral and psychological symptoms of dementia in a community-based autopsy cohort," *Acta Neuropathologica Communications*, vol. 11, no. 1, Jun. 2023, doi: 10.1186/s40478-023-01576-z.
- [4] H. Cai, Y. Pang, X. Fu, Z. Ren, and L. Jia, "Plasma biomarkers predict Alzheimer's disease before clinical onset in Chinese cohorts," *Nature Communications*, vol. 14, no. 1, Oct. 2023, doi: 10.1038/s41467-023-42596-6.
- [5] A. Alberdi *et al.*, "Smart home-based prediction of multidomain symptoms related to Alzheimer's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 6, pp. 1720–1731, Nov. 2018, doi: 10.1109/JBHI.2018.2798062.
- [6] H. Yang and P. A. Bath, "The use of data mining methods for the prediction of dementia: evidence from the English longitudinal study of aging," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 345–353, Feb. 2020, doi: 10.1109/JBHI.2019.2921418.
- [7] C. C. Brück, F. J. Wolters, M. A. Ikram, and I. M. C. M. de Kok, "Projected prevalence and incidence of dementia accounting for secular trends and birth cohort effects: a population-based microsimulation study," *European Journal of Epidemiology*, vol. 37, no. 8, pp. 807–814, Aug. 2022, doi: 10.1007/s10654-022-00878-1.
- [8] H. Estiri, A. Azhir, D. L. Blacker, C. S. Ritchie, C. J. Patel, and S. N. Murphy, "Temporal characterization of Alzheimer's disease with sequences of clinical records," *eBioMedicine*, vol. 92, Jun. 2023, doi: 10.1016/j.ebiom.2023.104629.
- [9] S. Singh and L. K. Bhatt, "Targeting cellular senescence: a potential therapeutic approach for Alzheimer's disease," *Current Molecular Pharmacology*, vol. 17, Jul. 2023, doi: 10.2174/1874467217666230601113430.
- [10] V. G. Shankar, D. S. Sisodia, and P. Chandrakar, "A novel discriminant feature selection-based mutual information extraction from MR brain images for Alzheimer's stages detection and prediction," *International Journal of Imaging Systems and Technology*, vol. 32, no. 4, pp. 1172–1191, Jul. 2022, doi: 10.1002/ima.22685.
- [11] L. B. Moreira and A. A. Namen, "A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia," *Computer Methods and Programs in Biomedicine*, vol. 165, pp. 139–149, Oct. 2018, doi: 10.1016/j.cmpb.2018.08.016.
- [12] H. Arifin *et al.*, "Meta-analysis and moderator analysis of the prevalence of malnutrition and malnutrition risk among older adults with dementia," *International Journal of Nursing Studies*, vol. 150, Feb. 2024, doi: 10.1016/j.ijnurstu.2023.104648.
- [13] Y. Pu, D. Beck, and K. Verspoor, "Graph embedding-based link prediction for literature-based discovery in Alzheimer's disease," *Journal of Biomedical Informatics*, vol. 145, Sep. 2023, doi: 10.1016/j.jbi.2023.104464.
- [14] S. Chai, X. Li, Y. Ye, J. Sun, H. Cai, and Z. Wang, "Silent information regulator 1: a potential target of semaglutide in the treatment of Alzheimer's disease," *Chinese Journal of Tissue Engineering Research*, vol. 28, no. 20, pp. 3235–3239, 2024.
- [15] A. Khan and M. Usman, "Early diagnosis of Alzheimer's disease using informative features of clinical data," in *Proceedings of*





- the International Conference on Machine Vision and Applications*, Apr. 2018, pp. 56–60, doi: 10.1145/3220511.3220515.
- [16] Y. Zhao *et al.*, “Global joint information extraction convolution neural network for Parkinson’s disease diagnosis,” *Expert Systems with Applications*, vol. 243, Jun. 2024, doi: 10.1016/j.eswa.2023.122837.
- [17] Y. Yin, H. Wang, S. Liu, J. Sun, P. Jing, and Y. Liu, “Internet of things for diagnosis of Alzheimer’s disease: a multimodal machine learning approach based on eye movement features,” *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11476–11485, Jul. 2023, doi: 10.1109/JIOT.2023.3245067.
- [18] A. Abd-alrazaq *et al.*, “The performance of artificial intelligence-driven technologies in diagnosing mental disorders: an umbrella review,” *npj Digital Medicine*, vol. 5, no. 1, Jul. 2022, doi: 10.1038/s41746-022-00631-8.
- [19] S. Feng *et al.*, “Discovery of acupoints and combinations with potential to treat vascular dementia: a data mining analysis,” *Evidence-Based Complementary and Alternative Medicine*, vol. 2015, pp. 1–12, 2015, doi: 10.1155/2015/310591.
- [20] Y.-C. Wang, C.-C. Wu, A. P.-H. Huang, P.-C. Hsieh, and W.-M. Kung, “Combination of acupoints for Alzheimer’s disease: an association rule analysis,” *Frontiers in Neuroscience*, vol. 16, Jun. 2022, doi: 10.3389/fnins.2022.872392.
- [21] K. Balasubramanian, N. P. Ananthamoorthy, and K. Ramya, “Prediction of neuro-degenerative disorders using sunflower optimisation algorithm and Kernel extreme learning machine: a case-study with Parkinson’s and Alzheimer’s disease,” *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 236, no. 3, pp. 438–453, Mar. 2022, doi: 10.1177/09544119211060989.
- [22] K. A. Happawana and B. J. Diamond, “Association rule learning in neuropsychological data analysis for Alzheimer’s disease,” *Journal of Neuropsychology*, vol. 16, no. 1, pp. 116–130, Mar. 2022, doi: 10.1111/jnp.12252.
- [23] L. Meng and Q. Zhang, “Research on early diagnosis of Alzheimer’s disease based on dual fusion cluster graph convolutional network,” *Biomedical Signal Processing and Control*, vol. 86, Sep. 2023, doi: 10.1016/j.bspc.2023.105212.
- [24] I. Abunadi, “Deep and hybrid learning of MRI diagnosis for early detection of the progression stages in Alzheimer’s disease,” *Connection Science*, vol. 34, no. 1, pp. 2395–2430, Dec. 2022, doi: 10.1080/09540091.2022.2123450.
- [25] J. A. Rocha *et al.*, “Human activity recognition through wireless body sensor networks (WBSN) applying data mining techniques,” *Smart Innovation, Systems and Technologies*, pp. 327–339, 2022, doi: 10.1007/978-981-16-5036-9_31.
- [26] D. Jacob and R. Henriques, “Educational data mining to predict bachelors students’ success,” *Emerging Science Journal*, vol. 7, pp. 159–171, Jul. 2023, doi: 10.28991/ESJ-2023-SIED2-013.
- [27] R. Perez-Siguas *et al.*, “Crime prediction and citizen security plans using big data in metropolitan lima,” *International Journal of Engineering Trends and Technology*, vol. 70, no. 10, pp. 144–154, Oct. 2022, doi: 10.14445/22315381/IJETT-V70I10P215.
- [28] L. M. O’Connor, B. A. O’Connor, J. Zeng, and C. H. Lo, “Data mining of microarray datasets in translational neuroscience,” *Brain Sciences*, vol. 13, no. 9, Sep. 2023, doi: 10.3390/brainsci13091318.
- [29] B. Yang, W. Bao, and S. Hong, “Alzheimer-compound identification based on data fusion and forgeNet_SVM,” *Frontiers in Aging Neuroscience*, vol. 14, Jul. 2022, doi: 10.3389/fnagi.2022.931729.
- [30] S. Rajayyan and S. M. Mohamed Mustafa, “Prediction of dementia using machine learning model and performance improvement with cuckoo algorithm,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 4, pp. 4623–4632, Aug. 2023, doi: 10.11591/ijece.v13i4.pp4623-4632.
- [31] L. M. O’Connor, B. A. O’Connor, S. Bin Lim, J. Zeng, and C. H. Lo, “Integrative multi-omics and systems bioinformatics in translational neuroscience: a data mining perspective,” *Journal of Pharmaceutical Analysis*, vol. 13, no. 8, pp. 836–850, Aug. 2023, doi: 10.1016/j.jpha.2023.06.011.
- [32] Z. Zhao *et al.*, “Conventional machine learning and deep learning in Alzheimer’s disease diagnosis using neuroimaging: a review,” *Frontiers in Computational Neuroscience*, vol. 17, Feb. 2023, doi: 10.3389/fncom.2023.1038636.
- [33] H. Albalawi, “An experimental study on evaluating Alzheimer’s disease features using data mining techniques,” *Journal of Information & Knowledge Management*, vol. 22, no. 01, Feb. 2023, doi: 10.1142/S0219649222500782.
- [34] S. L. Morgan *et al.*, “Most pathways can be related to the pathogenesis of Alzheimer’s disease,” *Frontiers in Aging Neuroscience*, vol. 14, Jun. 2022, doi: 10.3389/fnagi.2022.846902.
- [35] Y. Nian *et al.*, “Mining on Alzheimer’s diseases related knowledge graph to identify potential Alzheimer’s disease-related semantic triples for drug repurposing,” *BMC Bioinformatics*, vol. 23, no. S6, Sep. 2022, doi: 10.1186/s12859-022-04934-1.
- [36] J. C. Mundt, D. M. Freed, and J. H. Greist, “Lay person-based screening for early detection of Alzheimer’s disease: Development and validation of an instrument,” *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 55, no. 3, pp. P163–P170, May 2000, doi: 10.1093/geronb/55.3.P163.
- [37] A. Z. Klein, A. Magge, K. O’Connor, and G. Gonzalez-Hernandez, “Automatically identifying Twitter users for interventions to support dementia family caregivers: annotated data set and benchmark classification models,” *JMIR Aging*, vol. 5, no. 3, Sep. 2022, doi: 10.2196/39547.
- [38] R. Turrisi, M. Squillario, G. Abate, D. Uberti, and A. Barla, “An overview of data integration in neuroscience with focus on Alzheimer’s disease,” *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2024, doi: 10.1109/JBHI.2023.3268729.
- [39] J. Bokrantz, M. Subramanian, and A. Skoogh, “Realising the promises of artificial intelligence in manufacturing by enhancing CRISP-DM,” *Production Planning & Control*, pp. 1–21, Jul. 2023, doi: 10.1080/09537287.2023.2234882.
- [40] S. López-Torres *et al.*, “IoT monitoring of water consumption for irrigation systems using SEMMA methodology,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 222–234, 2020, doi: 10.1007/978-3-030-44689-5_20.

BIOGRAPHIES OF AUTHORS







Laberiano Andrade-Arenas    doctor in systems and computer engineering. Master’s in systems engineering. Graduated with a master’s degree in University Teaching. Graduated with a master’s degree in accreditation and evaluation of educational quality. Systems Engineer. scrum fundamentals certified, a research professor with publications in Scopus-indexed journals. He can be contacted at email: landrade@uch.edu.pe.



Inoc Rubio-Paucar     bachelor in systems and computer engineering. He has a background in database management and computer system design, with a focus on artificial intelligence applications, machine learning, and data science. His research interests are in the area of computer science. He can be contacted at email: Enoc.Rubio06@hotmail.com.



César Yactayo-Arias     obtained a bachelor's degree in administration from Universidad Inca Garcilazo de la Vega and a master's degree in education from Universidad Nacional de Educación Enrique Guzmán y Valle, he is a doctoral candidate in administration at Universidad Nacional Federico Villarreal. Since 2016 he has been teaching administration and mathematics subjects at the Universidad de Ciencias y Humanidades and since 2021 at the Universidad Continental. Currently, he also works as an administrator of educational services at the higher level, he is the author and co-author of several refereed articles in journals, and his research focuses on TIC applications to education, as well as management using computer science and the internet. He can be contacted at e-mail: yactayocesar@gmail.com.