Dementia Prediction using Data Mining

Mariana Lindo Carvalho [a]
[a] Center, University of Minho, Campus Gualtar, Braga 4710, P

17/01/2021

Abstract

Worldwide, around 55 million people have dementia. Aiming to prevent this disease, the goal behind this paper is to predict whether a patient could develop dementia, using Data Mining (DM) models induced with classification techniques. During the DM process, the CRISP-DM Methodology was followed and the WEKA and RapiMiner software tools were used to induce the DM models. Most of the models achieved sensitivity, specificity, precision and accuracy higher than 90% and one model even had 100 % for all these four metrics.

Keywords: Data Mining, Decision Support Systems; CRISP-DM; Classification; Dementia.

1 Introduction

Nowadays, someone in the world develops dementia every 3 seconds Alireza Atri (2020). Dementia is the loss of cognitive functioning, such as, thinking, remembering and reasoning and it interferes with a person's daily life and activities. Signs and symptoms of dementia result when once-healthy neurons, or nerve cells, in the brain stop working, lose connections with other brain cells and die. This process is common in everyone, although, people with dementia experience far greater loss NIA (2021). Therefore, predicting dementia's symptoms, based on the information collected about an elderly person, would allow doctors to interview in early states, and therefore, take measures in order to prolong, with quality, a patient's life.

Dementia is more common as people grow older, but it is not a normal part of aging. There are several different forms of dementia, where Alzheimer's disease (AD) is the most common. AD accounts for about half of the affected population, followed by vascular dementia (VaD) (20–25%), mixed dementia (5–10%), Parkinson's disease, dementia with Lewy bodies, physical brain injury, Huntington's disease, Creutzfeldt–Jacob disease, frontotemporal dementia/Pick's disease and normal pressurehydrocephalus Liu CK (2005) Andersen K (1999). Although there is no proven prevention, in general, leading a healthy lifestyle may help reduce risk factors that have been associated with these diseases Alireza Atri (2020).

DM is a the set of methods and techniques for exploring and analysing large datasets in order to find unknown or hidden rules, associations or patterns. The DM techniques can be classified as descriptive, which includes clustering techniques, or predictive, in which there are classification and regression techniques Tuffery (2011). Since transitions in healthcare organizations generate large amount of data with complex nature, the interest of this organizations on using DM has been increasing Koh C. (2011). Decision Support Systems (DSS) can be defined as an interactive computer-based systems which supports several phases of the decision-making process Marek J Druzdzel (1999) Pedro Gago (2005). To provide fast and reliable decision support, DSS need to have an automated analysis of data, capable of find tendencies and extract knowledge, which can be performed by DM techniques Bâra A. (2012). Whenever DSS are applied to healthcare, they are called Clinical Decision Support Systems (CDSS) and they intend to improve healthcare delivery by enhancing medical decisions, with the use of targeted clinical knowledge, patient information, and other health information Osheroff J. (2012). CDSSs today are primarily used at the point-of-care, for the doctors to combine their knowledge with information or suggestions provided by the CDSS Filipe Portela (2015). Other advantages of the usage of Clinical DSS in medical practice include ensuring accurate and timely diagnoses for preventing diseases, lowering operating costs, improving efficiency and reducing patient inconvenience E. (2009).

This article includes six sections. After the Introduction, the second section, Background and Related Work, presents some studies similarly to this one. On Methodology, Materials and Methods the third section, the tools and methods used in this study are mentioned and described. On section four, Data Mining Process, where the Cross Industry Standard Process for Data Mining was adopted, all the processes of DM are presented, as well as the results. These results are discussed on the fifth section, being the conclusion and future work on the last section.

2 Background and Related Work

Gopi Battinenia (2019) proved the viability of using DM models to predict dementia in patients, by using data from the same source as in this article. The classification technique used was Support Vector Machines and the patients were classified as demented, non-demented or converted. In that study, an attribute selection was applied, where attributes like Sex, EDUC, SES, eTIV and ASF were excluded, since these parameters were considered to not be good enough in dementia prediction and also considering many attributes may decrease performance. The developed models acquired satisfactory results by achieving sensitivity values between 65–82% and accuracy nearly 70%.

Xinyu (2021) compared Artificial Neural Networks and XGBoost in the prediction of AD, using the same dataset that was used in this article. In his study, he did not use the Handedness variable because all observations were right-handed. He concluded that the performance of both ANN and XGBoost were successful, because model's accuracy was greater than 85%. He explained the results obtained were due to the high correlation between dementia and factors, such as nWBV and CDR.

Kruthika Kr (2018) used a multistage classifier that included Naive Bayes classifier, Support Vector Machines and K-Nearest Neighbours for AD prediction and retrieval. The results showed that the multistage classifier got a good performance, with accuracy, sensitivity and specificity between 86-97%, 82-92% and 78-90%, respectively.

João Maroco (2011) evaluated the sensitivity, specificity, overall classification accuracy, area under the ROC and Press' Q of DM classifiers like Neural Networks, Support Vector Machines, Classification Trees and Random Forests, Quadratic Discriminant Analysis and Logistic Regression in the prediction of the evolution into dementia of 400 elderly people with mild cognitive impairment. The results showed that, when taking into account sensitivity, specificity and overall classification accuracy, Random Forests and Linear Discriminant analysis rank first among all the classifiers tested in prediction of dementia using several neuropsychological tests.

3 Methodologies, Material and Methods

3.1 Methodology

The methodology followed was the model of Cross Industry Standard Process for Data Mining (CRISP-DM), which breaks the process of DM in six steps: business understanding; data understanding; data preparation; modelling; evaluation and deployment Pete Chapman (2000).

3.2 Materials

The dataset used was obtained from the Open Access Series of Imaging Studies (OASIS-2) longitudinal collection of 150 subjects of 373 MRI data. The subjects are all right-handed and include both men and women, aged 60 to 96. Each one of the 150 instances has 15 attributes, such as Sex and Age Daniel S. Marcus (2010).

3.3 Methods

The Machine Learning (ML) software Weka was used in order to analyze the data along with six approaches to induce the DM models. The DM techniques used were Naive Bayes (NB), Multilayer Perceptron (MP), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and k-Nearest Neighbours (IBK). NB is a simple learning algorithm that utilizes Bayes'rule with a strong assumption that the attributes are conditionally independent given the class Lewis (1998). The MP algorithm is an artificial neural network that has one or more layers and it can solve problems that are not linearly separable. In this algorithm, the data flows in one direction from input until output Gardner M.W (1998). LR is a regression analysis method that estimates probabilities recurring to a certain logistic function, and this way, it can measure the relations between the dependent variable with at least one independent variable Chongsheng Zhang (2017). According to Stephan Dreiseitl (2002), SVM combines linear modelling with learning based on instances, because this algorithm chooses a limited number of samples from each group and constructs a linear function building separated boundaries between datasets. If there is no linear separation feasible, than the kernel approach will be used to automatically add the training samples into a higher dimensional space and to learn a separator in that zone. According to Freund Y. (1996), RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The IBK is one of the simplest and fastest ML algorithms. This algorithm does not build a model, instead it generates a prediction for a test instance just-in-time, by using a distance measure to locate k nearest instances in the training data for each test instance and uses those selected instances to make a prediction Vijayarani S. (2013). It is also important to refer that the software RapidMiner was used for generating the Correlation Matrix (CM) of the dataset attributes.

4 Data Mining Process

4.1 Business Understanding

Nowadays there are more and more people with dementia, who, due to their condition, are subject to a series of symptoms, such as experiencing memory loss, poor judgment and confusion, difficulty speaking, understanding, expressing thoughts, reading and writing, using unusual words to refer to familiar objects and losing interest in normal daily activities or events. Thus, it would be very beneficial to predict whether patients could develop dementia based on their characteristics, and consequently, act more quickly to improve their quality of life. Therefore, the main objective of this article is to predict whether or not a patient is demented based on some of the patient's characteristics.

4.2 Data Understanding

Of all 15 instances, the target variable chosen was *Group*, that represents whether the patient is demented, non-demented or converted, which means the patient was classified as non-demented at the time of his initial visit and was subsequently characterized as demented at a later visit. Despite the variable *Group* initially had 3 classes, in this article only were considered 2 classes, demented and non-demented, because the main focus of this study is to determinate if a patient is or not demented. Figure 1 shows the data distribution of this variable and as it can be observed, 43.5% of the patients have dementia.

Since converted patients were not considered, all dataset columns that were related to this patients were not used. The Handedness column was also discarded because all observations were right-handed. Therefore, the 10 columns of the final dataset after data preparation are *Group*, Sex, Age, Socioeconomic Status (SES), Education level (EDUC), Mini Mental State Evaluation (MMSE), Clinical Dementia Ratio (CDR), estimated Total Intracranial Volume (eTIV), normalized Whole Brain Volume (nWBV) and Atlas Scaling Factor (ASF).

In order to compare some of the relations between the different attributes, it was built a CM. This matrix is represented in Figure 2, where is possible to verify that regardless of whether a person is in dementia or not, the related factors are not absolutely independent. Dewey M (1999) claims that sex is an important risk factor for dementia, in particular for AD, because the incidence of AD in women is higher than in men after the age of 85 years. The same team has also reported that the risk of AD declines in men but not in women after the age of 90 years. In addition, Ruitenberg A (2001) concluded that the overall incidence of VaD is lower in women than in men. Although, several factors might complicate this associations, e.g. sex steroid hormones, lifestyle, ethnicity, and genetic polymorphisms of sex-related genes. Therefore, it is important to consider all these factors while exploring the association between sex and risk of dementia. In Figure 2, it is possible to verify that the correlation coefficient between the attributes Sex and *Group* is not too strong, which could probably be due to the lack of consideration of other factors, as mentioned by Dewey M (1999).

Launer LJ (1999) showed that the risk of all subtypes of dementia increases with age, specially after 65 years. However, in Figure 2, it is not possible to identify a strong correlation, which is probably related to the fact that all the patients in the dataset had ages between 60 to 96 years, and so, almost all patients are over 65 years. Therefore the age effect is not visible.

Regarding to sociodemographic characteristics as SES and EDUC, this last contains the years of education of each patient and SES is the Socioeconomic status as assessed by the Hollingshead Index of Social Position and classified into categories from 1 (highest status) to 5 (lowest status) Daniel S. Marcus (2010). Figure 2 shows that EDUC and SES are highly negatively correlated. This is predictable and reasonable, as a highly educated person tends to have a better socioeconomic status. Regarding to the correlation of EDUC with *Group*, poor education has been cited by Ott A. (1997) as a risk factor for AD, especially in males. However, in Figure 2, it is not possible to identify a strong correlation between this variable and *Group*, which probably is due to the fact that the EDUC is not an independent variable, it is also conditioned by other factors like sex, and so, this dependency may complicate EDUC individual association to dementia.

MMSE is a 30-point questionnaire that reflects the cognitive ability of a person, where 0 and 30 are the worst and the best score, respectively. According to Xinyu (2021), low scores of MMSE are more common in demented people. This influence of MMSE in dementia is also observable in Figure 2, because, MMSE and *Group* are highly negatively correlated.

Regarding to CDR, the CDR Scoring Table is useful for characterizing and tracking a patient's level of dementia. If CDR is equal to zero, the patients are non-demented and if CDR value is 0.5, the patients have very mild dementia. When CDR is equal or higher than 1, the patients will face the tendency to have dementia, with moderate to severe cognitive impairment Hughes CP (1982). This strong relation between CDR and *Group* is perfectly visible in Figure 2, due to the high correlation coefficient between these two variables.

The nWBV is expressed as a percent of all voxels in the atlas-masked image that are labeled as gray or white matter by the automated tissue segmentation process. Although no study was found that proved the influence of nWBV on the existence of dementia, from Figure 2, it is possible to verify that these nWBV and *Group* have a considerably high negative correlation coefficient. In order to analyze this relationship more deeply, the graph in Figure 3 was drawn up. This graph which shows that people with dementia tend to have lower values of nWBV in relation to those who are not demented.

The ASF is an one-parameter scaling factor that allows the comparison of eTIV based on differences in human anatomy Daniel S. Marcus (2010). In Figure 2, the variables ASF and eTIV have the highest correlation with a coefficient very close to 1. Therefore it is predictable that this variables have similar contribution on the prediction of dementia.

Besides nWBV, for SES, eTIV and ASF there was not found any study that proved the direct association of these factors to dementia. It is also important to refer that all the studies mentioned in this section affirm that the factors are not independent and their dependency may complicate their individual association to dementia.

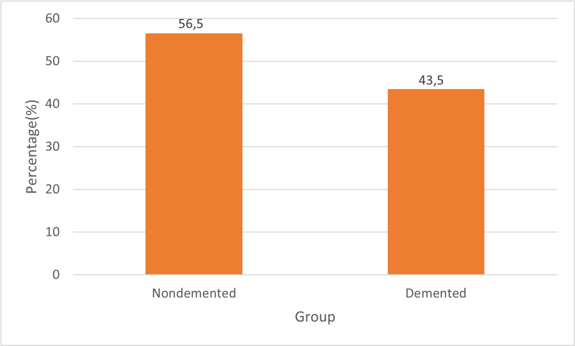

Fig. 1: Data distribution of the target variable *Group*.

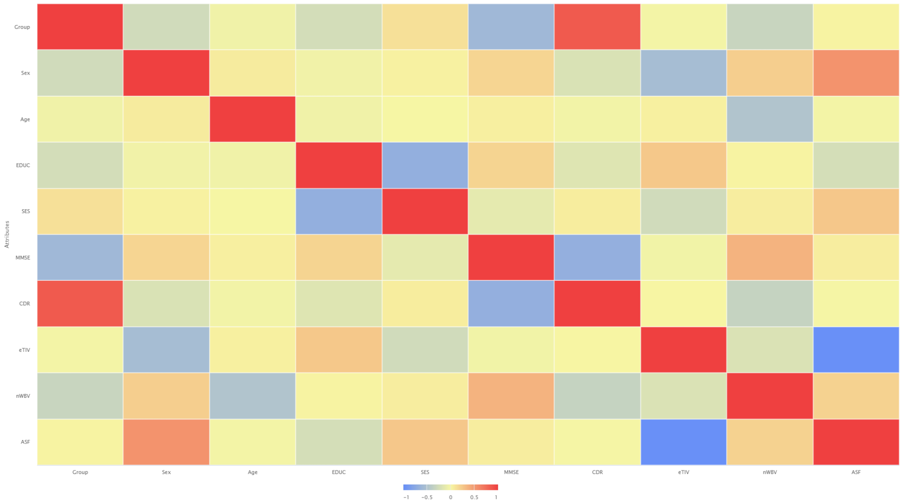

Fig. 2: CM of the dataset attributes.

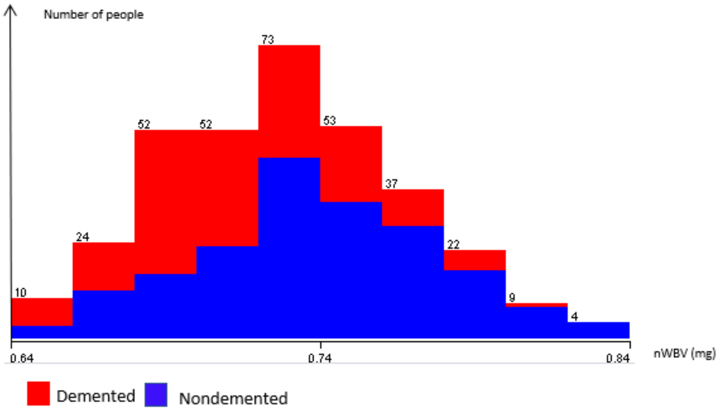

Fig. 3: Relation between the attribute nWBV and the *Group* class.

4.3 Data Preparation

After selecting the data, a pre-processing phase started with the removal of the variables that were related to converted patients, as mentioned before. After that, all of the missing values in the polynominal attributes SES and MMSE were filled-up with the mode of the respective attribute. Furthermore, the numeric values in the dataset were ranged from 0 to 2004.480, which could affect the DM performance. To avoid this, all this values were normalized and converted to numbers between 0 and 1, with the point used as decimal separator. In the case of the variables with continuous values, 0 and 1 correspond respectively to the minimum and maximum values.

4.4 Data Modelling

This section, will explore the different Data Mining Models (DMM) used to get all the results. The DMM have the following formula:

DMM = {Approach, Scenarios, DM Techniques (DMT), Data Approaches (DA), Target} (1)

For this article, it was decided to give to each DMM the following methods: Approach: {Classification}; Scenarios: {S1; S2; S3; S4}; DMT: {NB, MP, LR, SVM, RF, IBk}; DA: {Cross Validation 10 folds, Cross Validation 50 folds, Percentage Split - 66% of the data for training}; Target: {Group}.

The final objective is to predict if a patient has dementia, which means that the appropriated approach is classification and the target is to understand if the patient has dementia or not. The Scenarios (S) can be described as S1: {All variables}, S2: {Group, Age, MMSE, CDR, eTIV, nWBV}, S3: {Group, Sex, Age, EDUC, MMSE, CDR} and S4: {MMSE, CDR, nWBV, Group}. The scenarios were chosen in order to evaluate which attributes are the most relevant to predict the dementia in a patient. Therefore, the scenario S2 has the variables used in Gopi Battinenia (2019)'s article, S3 has the variables that are proved to be related to dementia in studies and S4 has the variables chosen by the AtribbuteSelection Filter in Weka and in which it is possible to verify that the chosen variables are the ones that have the three highest correlation coefficients with the *Group* class. The DMT used were the methods listed in section 3.3. At the end, there were 72 different models to compare and analyze: DMM = {1 Approach, 4 S, 6 DMT, 3 DA, 1 Target}.

4.5 Evaluation

The induced models evaluation was performed by calculating the statistic metrics as it follows, based on the results given by the confusion matrix. This matrix provides the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). With these results, it is possible to determinate sensitivity, specificity, accuracy and precision. In the Tables 1 to 4 are mentioned the models that, for each DMT, achieved the best sensitivity, specificity, accuracy and precision results, respectively. It is also important to refer that in this study it was considered true be demented and false be non-demented.

Table 1. DMM with the best sensitivity results for each DMT.

DMT	Scenario	DA	Sensitivity(%)
NB	S1,S2,S3,S4	Cross Validation 10 folds	98.65
MP	S1	Cross Validation 10 (or 50) folds	99.32
	S2	Cross Validation 10 folds	
LR	S1	Cross Validation 50 folds	99.32
SVM	S1	Percentage Split 66%	100
RF	S1,S2,S4	Cross Validation 10 (or 50) folds	98.65
IBk	S1	Cross Validation 10 (or 50) folds	100

Table 2. DMM with the best specificity results for each DMT.

DMT	Scenario	DA	Specificity(%)
NB	S1,S2,S3,S4	Any of them	100
MP	S1,S3,S4	Any of them	100
	S2	Cross Validation 10 folds and Percentage Split 66%	
LR	S1	Cross Validation 50 folds	100
	S4	Any of them	
SVM	S1	Cross Validation 50 folds	100
	S2,S3	Cross Validation 10 (or 50) folds	
	S4	Any of them	
RF	S1,S2,S4	Any of them	100
	S3	Percentage Split 66%	
IBk	S1	Cross Validation 10 (or 50) folds	100
	S2,S3,S4	Percentage Split 66%	

Table 3. DMM with the best accuracy results for each DMT.

DMT	Scenario	DA	Accuracy(%)
NB	S1,S2,S3,S4	Cross Validation 10 folds	99.40
MP	S1	Cross Validation 10 (or 50) folds	99.70
	S2	Cross Validation 10 folds	
LR	S1	Cross Validation 50 folds	99.70
SVM	S1	Cross Validation 10 folds	99.40
	S2	Cross Validation 10 (or 50) folds	
RF	S1,S2,S4	Cross Validation 10 (or 50) folds	99.40
IBk	S1	Cross Validation 10 (or 50) folds	100

Table 4. DMM with the best precision results for each DMT.

DMT	Scenario	DA	Precision(%)
NB	S1,S2,S3,S4	Any of them	100
MP	S1,S3,S4	Any of them	100
	S2	Cross Validation 10 folds and Percentage Split 66%	
LR	S1	Cross Validation 50 folds	100
	S4	Any of them	
SVM	S1	Cross Validation 50 folds	100
	S2,S3	Cross-validation 10 (or 50) folds	
	S4	Any of them	
RF	S1,S2,S4	Any of them	100
	S3	Percentage Split 66%	
IBk	S1	Cross Validation 10 (or 50) folds	100
	S2,S3,S4	Percentage Split 66%	

5 Discussion

From the analysis of the best results, it is possible to verify that the best achieved sensitivity, specificity, accuracy and precision values were all 100%. This values belong all to the model of the scenario S1 that used IBk as the DMT and Cross Validation (10 or 50 folds) as the DA. Thus, it is possible to claim that the most suitable model, from all the 72 induced models, is DMM = {Classification, S1, IBk, Cross Validation 10 (or 50) folds, Group}.

The analysis of the obtained results also allowed to conclude that models which used Cross Validation (with 10 or 50 folds) as the DA achieved better results, namely of sensitivity and accuracy, when compared to those which used Percentage Split. The reason behind this is that the Cross Validation method uses all the data for training, while Percentage Split only uses a certain percentage and algorithms effectively learn more when they use more data for training.

Another important fact to mention is that although all the models presented very good results, the models of scenario S1 always obtained the best results, being present in all rows of tables 1 to 4. This may mean that, in fact, all of the dataset attributes have an influence in whether or not a patient has dementia, although there may not yet be strong scientific evidence to prove it. It also shows that the combination of all attributes has much more effect in whether a person is demented or non-demented, than just considering some of the attributes. The scenarios S2 and S4 had a very similar performance, presenting very often the same results. Since the scenario S4 only has 4 attributes, this allows to conclude that the number of attributes used does not affect model's performances. The scenario with the worst performance was S3, probably because it does not include the attribute nWBV, which has the third highest coefficient correlation with the *Group* class, after CDR and MMSE.

It was also verified that the algorithms that had the best results were MP, LR and specially IBk and that all the other 3 techniques had less good results. Regarding to the NB technique, this may not had the best performance because of the assumption this technique makes about the independence of the variables. This assumption might have not worked well on the used dataset due to the high correlation that exists between the attributes. The models that used RF as the DMT probably did not had the best performances, because RF is more suitable for larger datasets. Regarding to the models that use SVM as the

DMT, they had the lowest sensitivity, accuracy and precision, which means SVM is the less suitable DMT for this dataset.

In this particular dataset, the selection of the variables performed an important role during results analysis. As was seen in Figure 2, CDR, MMSE and nWBV are obviously essential indexes of detecting dementia and therefore they have a high influence on the *Group* class. This made all the models produced have very good performances, no matter what kind of method is used, or data approach is considered, because, the results still showed a high match with the data. The problem is the variables in this dataset are too directly related to dementia and even are the standard of determining if someone has dementia, which made the comparison of results less important than what they should be. The variables not directly reflecting dementia such as smoke habits may give a clearer effect. However, it does not mean the analysis is meaningless. Even with the direct indexes, IBk method still was the best one, with unbeatable results.

6 Conclusion and Future Work

This study aimed to predict, demonstrating the utility of clinical decision support systems, whether a patient has or not dementia, by allying real data with DMM. The best results were obtained when inducing IBk algorithm with Cross Validation, achieving 100% of sensitivity, accuracy, precision and specificity. This result is probably due to the high correlation between dementia and the dataset attributes, namely CDR, MMSE and nWBV. This study also has some limitations. The habit of handedness may have an impact on brain and therefore affect the possibility of getting dementia. However, the dataset had only collected right-handed individuals, so the study did not consider the situation of the impact of handedness on dementia. Furthermore, the age group of dataset focuses on 60-96, so the analysis before 60 is not considered. Future studies should focus on collecting data on left-handed individuals and widening the range of age to explore the risk factors of dementia for young-age group.

Acknowledgements

I am thankful to the Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, and J. Morris to provide access on OASIS longitudinal studies. The Italian Ministry of Economic Development (MISE) no. F/080034/01/X35 funded this study. Gopi Battineni Ph.D.bursaries were supported in part by the University of Camerino.

References

Alireza Atri, Glenn Rees, P. B. (2020). Dementia statistics. , Alzheimer's Disease International, London, UK. Available at `https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/`.

Andersen K, Launer LJ, D. M. e. a. (1999). *Basics of Alzheimer's Disease: What It Is and What You Can Do* (1nd ed.). Beauvais, France: EURODEM Incidence Research Group.

Bâra A., L. I. (2012). *Improving Decision Support Systems with Data Mining Techniques* (1nd ed.). London, UK: INTECH Open Access Publisher.

Chongsheng Zhang, Changchang Liu, X. Z. G. A. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications* (82), 128–150.

Daniel S. Marcus, Anthony F. Fotenos, J. G. C. J. C. M. R. L. B. (2010). Open access series of imaging studies: Longitudinal mri data in nondemented and demented older adults. *Journal of Cognitive Neuroscience 22*(12), 2677–2684.

Dewey M, Andersen K, L. L. e. a. (1999). Gender differences in the incidence of ad and vascular dementia: the eurodem studies. *EURODEM Incidence Research Group Neurology.*

E., B. (2009). *Clinical decision support systems: state of the art* (1nd ed.). Rockville, EUA: AHRQ publication.

Filipe Portela, Manuel Filipe Santos, J. M. A. A. F. R. S. (2015). *Real-time decision support using data mining to predict blood pressure critical events in intensive medicine patients. In Ambient Intelligence for Health* (1nd ed.). Braga, PT: Springer.

Freund Y., S. R. (1996). Experiments with a new boosting algorithm, machine learning. In *Proceedings of the Thirteenth International Conference*, pp. 148–156. ATT Research.

Gardner M.W, D. S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment 32*(14), 2627–2636.

Gopi Battinenia, Nalini Chintalapudib, F. A. (2019). Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (svm). *ScienceDirect.*

Hughes CP, Berg L, D. W. C. L. M. R. (1982). A new clinical scale for the staging of dementia. *Br J Psychiatry 140*, 566–572.

João Maroco, Dina Silva, A. R. M. G. I. S. A. d. M. (2011). Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res Notes.*

Koh C., T. G. (2011). Data mining applications in healthcare. *Journal of healthcare information management 19*(2), 65.

Kruthika Kr, Rajeswari, M. H. (2018). Multistage classifier-based approach for alzheimer's disease prediction and retrieval. *Informatics in Medicine Unlocked 14.*

Launer LJ, Andersen K, D. M. e. a. (1999). Rates and riskfactors for dementia and alzheimer's disease. *EURODEM IncidenceResearch Group and Work Groups.*

Lewis, D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: ECML-98*, pp. 4–15. Springer Berlin Heidelberg.

Liu CK, Tai CT, L. R. e. a. (2005). *Gender differences in the incidence of AD and vascular dementia: the EURODEM Studies* (1nd ed.). Chicago, EUA: Appl Psychol Res.

Marek J Druzdzel, R. R. F. (1999). *Data mining applications in healthcare. Journal of healthcare information management* (1nd ed.). New York, EUA: Marcel Dekker, Inc.

NIA (2021). What is dementia? symptoms, types, and diagnosis. Technical report, NIA Alzheimer's and related Dementias Education and Referral (ADEAR) Center, Maryland, EUA. Available at `https://www.nia.nih.gov/health/what-is-dementia`.

Osheroff J., e. a. (2012). *Improving Outcomes with Clinical Decision Support: An Implementer's Guide* (1nd ed.). Chicago, EUA: HIMSS Publishing.

Ott A., Breteler M., B. e. a. (1997). Atrial fibrillation and dementia in a population-based study: the rotterdam study. *Stroke Association 28*, 316–321.

Pedro Gago, Manuel Filipe Santos, S. P. C. J. N. L. G. (2005). A knowledge discovery based intelligent decision support system for intensive care medicine. *Journal of decision systems 14*(3), 241–259.

Pete Chapman, Julian Clinton, R. K. T. K. T. R. C. S. R. W. (2000). *Crisp-dm 1.0 step-bystep data mining guide* (1nd ed.). Chicago, EUA: SPSS Inc.

Ruitenberg A, Ott A, v. S. J. e. a. (2001). Incidence of dementia: does gender make a difference? *Neurobiology Aging*.

Stephan Dreiseitl, L. O.-M. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics 35*(5), 352–359.

Tuffery, S. (2011). *Data mining and statistics for decision making* (1nd ed.). Paris, France: WILLER.

Vijayarani S., M. M. (2013). An up-to-date comparison of state-of-the-art classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering 2*.

Xinyu, S. (2021). Application and comparison of artificial neural networks and xgboost on alzheimer's disease. In *Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing*, pp. 101–105. Association for Computing Machinery.