RECIBIDO EL 23 DE JULIO DE 2021 - ACEPTADO EL 20 DE OCTUBRE DE 2021

EARLY DETECTION ALGORITHM FOR ALZHEIMER'S DISEASE USING AUTONOMOUS LEARNING

ALGORITMO DE DETECCIÓN TEMPRANA PARA LA ENFERMEDAD DE ALZHEIMER MEDIANTE APRENDIZAJE AUTÓNOMO

Aguilar Obregón Jorge Eduardo[1]

Salcedo Parra Octavio José[2]

Rodríguez Miranda Juan Pablo[3]

Universidad Distrital Francisco José de Caldas

ABSTRACT

The current document describes the approach to a research problem that aims to generate an algorithm that allows detecting the probable appearance of Alzheimer's disease in its first phase, using autonomous learning techniques or Machine Learning, more specifically KNN (K-nearest Neighbor) with which the best result was obtained. This development will be based on a complete information bank taken from ADNI (Alzheimer's Disease NeuroImaging Initiative), with all the necessary parameters to direct the investigation to an algorithm that is as efficient as possible, since it has biological, sociodemographic and medical history data, biological specimens, neural images, etc., and in this way the early detection of the aforementioned disease was configured. A complete guide to the process will be carried out to finally obtain the KNN algorithm whose efficiency is 99%, and then discuss the obtained results.

RESUMEN

El presente documento describe el abordaje de un problema de investigación que tiene como objetivo generar un algoritmo que permita detectar la probable aparición de la enfermedad de

1 Facultad de Ingeniería. Universidad Distrital Francisco José de Caldas. Bogotá, Colombia. Correo electrónico: reyaguilar07@gmail.com
ORCID: https://orcid.org/0000-0002-2502-7846
2 Profesor Titular. Facultad de Ingeniería. Universidad Distrital Francisco José de Caldas. Bogotá, Colombia. Profesor de Planta, Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Sede Bogotá. Correo electrónico: osalcedo@udistrital.edu.co; ojsalcedop@unal.edu.co
ORCID: https://orcid.org/0000-0002-0767-8522
3 Profesor Titular. Facultad del Medio Ambiente y Recursos Naturales. Universidad Distrital Francisco José de Caldas. Bogotá, Colombia. Correo electrónico: jprodriguezm@udistrital.edu.co
ORCID: https://orcid.org/0000-0002-3761-8221

Alzheimer en su primera fase, utilizando técnicas de aprendizaje autónomo o Machine Learning, más concretamente KNN (K-Neighbor Neighbor) con el que el se obtuvo el mejor resultado. Este desarrollo se basará en un banco de información completo extraído de ADNI (Alzheimer's Disease NeuroImaging Initiative), con todos los parámetros necesarios para dirigir la investigación a un algoritmo lo más eficiente posible, ya que cuenta con datos biológicos, sociodemográficos y de historia clínica. especímenes biológicos, imágenes neuronales, etc., y de esta forma se configuró la detección precoz de la enfermedad mencionada. Se realizará una guía completa del proceso para finalmente obtener el algoritmo KNN cuya eficiencia es del 99%, para luego discutir los resultados obtenidos.

PALABRAS CLAVE

Deteccion temprana, Machine Learning, Alzheimer

KEYWORDS

Early detection, Machine Learning, Alzheimer's

INTRODUCTION

Today's completely globalized world leads us every day to a path more and more dependent on technology, and it is common to see how most fields of knowledge adapt it more and more precisely to their approach. For this reason, technology advances at extraordinary speed, and attempts to develop tools to facilitate the work of each of the fields of action.

One of the most influential fields that makes the most use of technology for its development is medicine, which has established itself as a necessary science for human preservation. The creation of Software and hardware that supplies the great technological demand that medicine has, is one of the largest of all fields of action in which technology is necessary (Giacometti, 2013). Based on this fact, this research work will be aimed at one of the most important and fastest growing diseases in the world population.

The disease is: Alzheimer's. In general terms is a disease related to cognitive degeneration, has no cure and directly affects the brain causing memory, thinking and behavior problems (Alzheimer's association, 2016). The disease has several stages of development, some longer than others, and although they last for many years. There are certain common factors among patients, which make it possible to determine whether or not they will suffer from Alzheimer's (Kivipelto et al., 2001. An early diagnosis will help establish treatment and preventive measures.

The idea of the present work with all this in mind, is to consolidate an algorithm for the early detection of Alzheimer's, and achieved through the implementation of current technological tools, of greater acceptance and expansion such as autonomous learning or Machine Learning. The technology will contribute to safeguarding the quality of life of potential Alzheimer's patients.

II. Background and related works

The sources consulted have been developed under the same objective, the AD prediction. It should be noted that there are very good works, and very complete with respect to a problem statement, but this document will focus on extracting what is relevant for the project that is currently being developed.

A. USING NEURAL NETWORKS IN THE IDENTIFICATION OF SIGNATURES FOR PREDICTION OF ALZHEIMER'S DISEASE.

In (Dantas & Valenca, 2014) they postulate a prediction model of Alzheimer's disease using a technological tool called neural networks, and for this work they take as analysis and processing parameters 120 proteins contained in the blood of the test subjects, and the probability of the patient to suffer from Alzheimer's in the long

term is predicted, in addition to this, it shows the protein group that presents the most abnormal levels and can contribute to the appearance of Alzheimer's, so that it can follow a normalization treatment of said proteins. For this last activity, use a forest algorithm for selection. The disease prediction model achieved quite promising data, reaching an efficiency of 90%.

B. A BAYESIAN NETWORK DECISION MODEL FOR SUPPORTING THE DIAGNOSIS OF DEMENTIA, ALZHEIMER'S DISEASE AND MILD COGNITIVE IMPAIRMENT.

For the early diagnosis of Alzheimer's, we refer to the study carried out in (Seixas, Zadrozny, Laks, Conci, & Muchaluat Saade, 2014). This work takes as inputs variables such as the level of education, gender, age and psychiatric factors such as the patient's depression, then applies the most common tests for the diagnosis of Alzheimer's, these data will be analyzed by a decision model of a Bayesian network to give a positive or negative diagnosis of the condition of the disease. This decision network is based on a real medical diagnosis model, that is, it had the participation of expert doctors in AD diagnosis and treatment.

C. A MACHINE INTELLIGENCE DESIGNED BAYESIAN NETWORK APPLIED TO ALZHEIMER'S DETECTION USING DEMOGRAPHICS AND SPEECH DATA.

Another work related to the previous one is the one proposed in (Land & Schaffer, 2016), in this article, the results of a Bayesian network are presented, designed for the Alzheimer's diagnosis, taking into account demographic data such as age, sex, race and degree of education, also, they make an evaluation of the way of speaking of the people who are studied, where evaluated conditions such as wealth vocabulary, syntactic complexity, density of ideas, among others. For the realization of the network, genetic algorithms are used that contribute to the selection of a subset of characteristics, and an SVM classifier to differentiate the way in which a sick person expresses himself from a healthy person.

D. LONGITUDINAL MEASUREMENT AND HIERARCHICAL CLASSIFICATION FRAMEWORK FOR THE PREDICTION OF ALZHEIMER'S DISEASE.

The next study implemented in the prediction of Alzheimer's is the one carried out in (Huang et al., 2017), and is based on analysis of MRI images, which based on the longitudinal measurement of the brain of people with MCI (mild cognitive defect), and using a hierarchical classification method, you can predict the risk of suffering AD. The longitudinal measurement of the brain is carried out with image processing methods and the classification method is carried out based on the LRC statistical model (Longitudinal Redundancy Verification).

E. EXTENDED COX PROPORTIONAL HAZARD MODEL TO ANALYZE AND PREDICT CONVERSION FROM MILD COGNITIVE IMPAIRMENT TO ALZHEIMER 'S DISEASE.

Finally, an article is presented in (Alsaedi & Qader, 2018), with a different approach, oriented to the optimization of the MMSE (Mini-Mental State Examination) method, which, as explained in the previous paragraph, is a questionnaire-type test that it is used to measure the cognitive ability of patients. This method has been judged as ineffective and for this reason, the development of this research attempts to improve its effectiveness by combining it with the Cox proportional hazard regression method.

All the works are based on a common premise, the methodology used, in which they refer to scientific and medical documents on the diagnosis of Alzheimer's disease, an attempt is made to find a similarity between the appearance and the symptoms and characteristics present. Next, the parameters to be taken for the predictive mod-

el are finally chosen and we proceed to make a categorical evaluation in which, according to the type of data, the prediction model that best adapts to them is chosen, which is applied in a specialized database in these parameters, whose final answer indicates the efficiency of the model used, which depends both on the cleaning of the chosen registers, and on the chosen method.

III. METHODOLOGY

For the development of this research work, a research method known as Analysis-Synthesis will be used, since these are processes that allow the researcher to know reality. And part of the premise is that through the analysis of each of the components or parts that characterize a reality, the relationship between each of the research objects is established. Regarding the synthesis, consider the aforementioned objects as a whole, the interrelation of the elements that identify the object. The method consists of separating the object of study into two parts and, once its essence is understood, building a whole. Analysis and synthesis are two processes that complement each other in one. For this particular case, the analysis of the parts will be each of the parameters involved in the diagnosis of Alzheimer's, and the whole will be the disease! Then, through the study of the parameters, it is intended to establish the appearance or not of Alzheimer's. To carry out this model, the following phases will be executed:

PHASE 1: Information, knowledge and theoretical analysis; This phase focuses particularly on the identification and selection of the information necessary to fit the work. The most important parameters that may be the main causes of Alzheimer's disease are identified.

An analysis and synthesis of the characteristics are made that according to medical consultation and academic literature are the most decisive in the diagnosis of AD. This phase includes the following tasks:

Collection of information sources related to Alzheimer's disease, recent and completely reliable sources were consulted, for this, medical information, neurological documents and expert opinion were used, in addition to having access to a wide medical database.

The information was categorized, dividing it by parameters, each independent of the other, but which are present in most diagnoses of Alzheimer's disease.

PHASE 2: Once the variables that the early detection algorithm will have selected, the field research was carried out, that is, the information collection process was been made in terms of tangible data according to the information that resulted from the previous step, that is, we proceeded to work with a true report that contained the information chosen in the previous phase. Reports of the diagnoses of patients who presented Alzheimer's, and whose symptoms are common.

The ADNI (Alzheimer's Disease Neuroimaging Initiative) database was selected, the information of which is complete and varied, and a study of the variables involved is carried out, which was previously verified.

All the information organized according to what interests us is stored, after organizing the information, a validation process of its format and veracity was carried out. Those that presented conflicts were discarded.

PHASE 3: Phase 3 will work completely with the technology to be used, in this case it will be Machine Learning and it was reviewed according to the information we have available, the mechanism to use, the method that was used to process the data.

A deep comparison was made between all the Machine Learning methods that we can use, selecting the one that best suits the type of information we handle.

The algorithm for the early detection of Alzheimer's disease was designed and implemented, using the Machine Learning method selected in the previous step. That is, this phase resulted in said algorithm.

PHASE 4: In this last phase, a mixture was made between the results obtained from phase 2 and phase 3, that is, the algorithm, the real information of the DataSet and the databases obtained were validated and the algorithm of early detection of Alzheimer's disease (result of phase 3), in this way, we were able to obtain the final result and determine the effectiveness of the algorithm.

IV. DESIGN

For the design, each of the parameters involved in the process of extracting common information from Alzheimer's patients is explained in detail.

After analyzing medical information related to AD, 10 parameters involved in the Alzheimer's process are selected, and from which an abundant dataSet can be obtained with which to work through Machine Learning. The process to get to those parameters will be skipped, since it is not part of the context of the document.

The first 4 parameters are relevant sociodemographic data. These are Age, Sex, Education Level, and Ethnicity. All four are involved in the Alzheimer's process, as the categories formed in each parameter may be more or less likely to develop Alzheimer's. For example, in most epidemiological studies there is a tendency for men to suffer more AD than women before the age of 50-65 years. This profile reverses above the age of 60-65, with AD predominant in women. (Cacabelos, 2001).

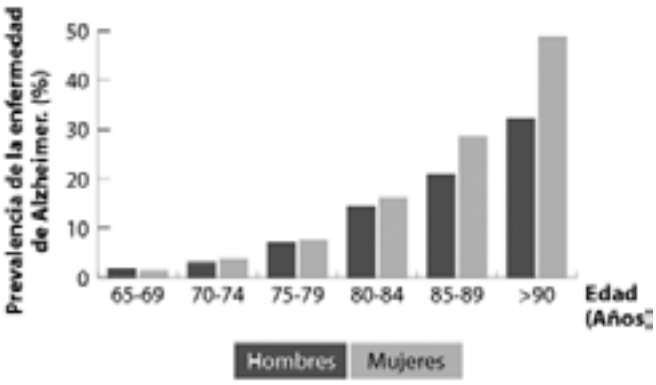

Figure 1. Prevalence of AD by sex and age.

Source: Cacabelos 2001.

The following 2 parameters are biological data, and have to do with the APOE lipoprotein genotype of the individuals. In medicine, this is the most important predictor of Alzheimer's. Taking into account both genotypes (There are 2 that each individual has), a very accurate prediction of it can be established.

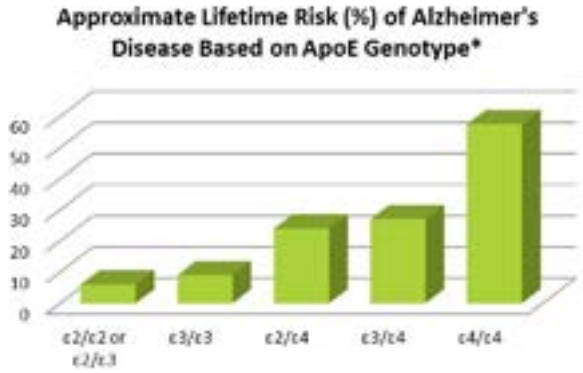

Figure 2. Risk of AD based on APOE genotype

Source: DiBattista, 2018.

Finally, the last 4 parameters are data from the clinical past of the subjects. The first is a history of Diabetes, then a history of hypertension and finally the history of Alcoholism and smoking.

The first two is part of the diseases that directly affect cardiovascular performance, and affect the correct circulation of blood to the brain.

They were chosen for this process due to their close relationship with EA. It is very important to take it into account for the diagnosis of Alzheimer's, since 80% of people with AD suffer from insulin resistance or Type 2 Diabetes (Arnold & Arvanitakis, 2018).

And the other two, which are smoking and alcoholism, were chosen due to their also close relationship with AD, at a biological level, tobacco would contribute to this deterioration of the vessels by increasing homocysteine, a blood molecule related to an increased risk of stroke, cognitive decline, and dementia, such as Alzheimer's. (Aguilar-Navarro, 2007); While on the side of Alcohol, it is indicated that light to moderate alcohol consumption is associated with a decrease in the risk of dementia in people 55 years of age or older and that the effect does not seem to depend on the type of alcoholic beverages consumed (Ruitenberg A, van Swieten, 2002).

Once these 10 parameters had been chosen, we proceeded to search for clinical DataSets that had this information. The data with which this document was developed, were taken from a dataset belonging to the ADNI (Alzheimer's Disease Neuroimaging Initiative), it is a community made up of exclusively research experts in Alzheimer's disease, it has carried out 4 separate study phases. Each one involves new characteristics and new subjects. For this research, we will focus on all phases, including the last one, ADNI-3 which as presented on its website, ADNI3 began in 2016 and includes scientists at 59 research centers in the United States and Canada. Between 1070-2000 participants are part of the study: approximately 700-800 participants transferred from ADNI2 and 370-1200 newly enrolled subjects.

The dataset was initially composed of 4565 records with that number of test subjects, hosted in the four ADNI investigations mentioned above. However, it is necessary to do a data cleaning before, as all the records may not be useful in our investigation.

The ADNI database contains many missing values. For different reasons, it may be that all the subjects in the sample have not provided the necessary information. That is why a decision is made, whether to omit the entire record or simply put the null value, categorizing said value.

The variability in data collection can generate noise in the final processing of the data model, therefore, the choice is to remove those records where no data is present. In this way, we can avoid much of the noise mentioned and the most successful result and with a higher percentage of success. The research focuses on 8 variables that will be extracted from the dataset, these are:

1. Age. In the dataset, the age is not said properly, but the patient's year of birth appears, so a numerical transformation will be made, which consists of taking the current year and subtracting the patient's birth year. In other words, it will not be taken into account if the patient was already in compliance this year. It will be a numerical data.

2. Sex. In the dataset, the gender information is consigned as a numeric field, 1 for masculine and 2 for feminine, for practical methods and for organizing the information, we will use them as well.

3. Education. In this case, the dataset offers us information already specified by educational level that is ranked from 0 to 20, and for practical terms it will remain at the same numerical level.

4. Ethnicity. This information is entered in the dataset as a numeric value with a range between 1 and 7 as follows: 1 = American Indian or Alaska Native; 2 = Asian; 3 = Native Hawaiian or other Pacific island; 4 = Black, African or African American; 5 = White; 6 = More than one race; 7 = Unknown.

5. The APOE e4 genotype. For this value, the dataset offers the information of only 2,389 subjects, of the 4,565 that were had in the beginning, and given its level of importance as a parameter, the records that do not have it should be omitted. Therefore, only 2,389 records will be used. Given that we have 2 alleles of the alipoprotein, and the dataset offers both for each patient, then we will work with both records. These two fields are numeric values that can be 2,3 or 4; representing the APOE allele e2, e3 and e4… And considering that the one that interests us is the APOE e4, since the epsilon 4 allele of APOE is the strongest known genetic risk factor for AD with a risk increased two to three times in people with one e4 allele and that increases to approximately 12 times in people with both alleles (ADNI, 2017).

6. Diabetes History. For the following parameters, it is necessary to refer to the patient's medical history, the dataset is divided into two, one for the entire patient history of the ADNI1, ADNI2 and ADNIGO study, and the other for only ADNI 3. The way in the one that stores the clinical data of the patients, is stored in 19 numerical categories that each represent a general condition in the following way: 1 = Psychiatric; 2 = Neurological; 3 = Head, eyes, ears, nose, throat; 4 = Cardiovascular; 5 = Respiratory; 6 = Hepatic; 7 = Dermatological or Integumentary; 8 = Musculoskeletal; 9 = Metabolic-Endocrine; 10 = Gastrointestinal; 11 = Hematopoietic-lymphatic; 12 =. Renal-genitourinary; 13 = Allergies or sensitivities to drugs; 14 = Alcohol Abuse; 15 = Drug abuse; 16 = History of cigarette consumption; 17 = Malignant Neoplasia; 18 = Major surgical procedures; 19 = Other.

In this case, diabetes is located in category 9, Metabolic-Endocrine, but it does not mean that if the patient has a record in his clinical history with this number, it will automatically be taken as diabetes, as other diseases such as hyperthyroid may occur or similar, so a data cleaning will be done, and another description column will be taken in which the condition presented is said through words. The conversion process will be from those two columns (If it is category 9 and also the description presents the word diabetes), to a column with a numerical value between 1 and 2, which represent with a history of diabetes and without a history of diabetes respectively.

7. Hypertension. For hypertension, the same process will be followed as for diabetes, only it is located in another category in the dataset. This is number 4, Cardiovascular Problems, but again this column will be taken with the value 4 and the text to search will be hypertension; to later transform it to the numerical value of 1 or 2 (If you have a history of hypertension or not, respectively).

8. Consumption of alcohol and cigarettes. The dataset allows knowing the clinical history of each subject, therefore determining if they have had a history with alcohol, with cigarettes or with both, for this category 14 and 16 will be taken for alcohol and cigarettes respectively. For the investigation, only the subjects who present abuses with both substances will be taken.

At the beginning there were 4565 records, but duplicates and those that had incomplete information or that were corrupted and generated noise in the predictive model were cleaned. This leaves us with a total of 2388 records to load into the model.

To carry out this process of selecting, cleaning and finalizing the data, an environment was established in the team which consists of specialized software oriented to the analysis of the data, such as Wrangler, which was selected for the rapid manipulation and extraction of the data, which allowed for a successful cleaning; The Python environment was also installed, with its respective libraries (Pandas, sklearn, matplotlib, numpy) as it is the one used for the entire Machine Learning process.

V. IMPLEMENTATION

Starting with the implementation of the Machine Learning algorithm, and having the data completely clean, we begin the process choosing the prediction model.

Taking into account that the final data required is categorical data; that indicates a yes or no in the prediction of whether or not he has Alzheimer's disease, the following three methods are chosen: Logistic Regression, since a binary result is required, that of K-nearest Neighbor (kNN), which also returns a binary result grouping its characteristics and Naive Bayes that allows treating each variable as independent of the other, which gives it a practical and simple sense when predicting the result.

The linear regression model was applied through pandas and sklearn, Python libraries.The final variable, will be Alzheimer and it will determine whether or not a patient has the likelihood of developing Alzheimer's.

The dependent variable in this case 'Y', will be the last column of the dataset (ad), which for our case will depend on all the other parameters. The model generates a calculation in which it gives the weight to each of the parameters. It does the regression calculation, and we have the result.

However, one of the most important parts here in the application of the model is the validation of the model. The algorithm learns from the data, and through such learning it must predict whether or not the test subject tends to have Alzheimer's. But like the dataset used was taken directly from an experimental association and their data was collected over several years; It is very difficult to extract a considerable amount of data again, so to test the efficiency of the method a trick of the Machine Learning application was used which is to apply the model so that it learns from all the data. Later load it with only 80% of the data and the remaining 20% use them for the efficiency check.

At the end, what you get is not only the machine learning algorithm, but its characteristics, such as effectiveness and the range of hits and misses, among others.

When applying the model and the algorithm used for the Logistic Regression, in effect, in the first instance there is a low accuracy of success. The algorithm is approximately 62% effective, using 80% of the records as training.

	precision	recall	f1-score	support
0	0.64	0.85	0.73	1161
1	0.52	0.25	0.34	746
accuracy			0.62	1907
macro avg	0.58	0.55	0.53	1907
weighted avg	0.59	0.62	0.58	1907

PREDICCION	0	1
REAL		
0	991	170
1	560	186

Table 1. Result using Linear Regression algorithm

Fountain. Author.

By applying the model and algorithm used for the KkN, it already gives us a fairly accurate prediction. In this case 99% effective. The numbers of this algorithm are as follows:

```
[[350   2]
 [  3 123]]
```

	precision	recall	f1-score	support
0	0.99	0.99	0.99	352
1	0.98	0.98	0.98	126
accuracy			0.99	478
macro avg	0.99	0.99	0.99	478
weighted avg	0.99	0.99	0.99	478

Figure 3. Result using KkN algorithm.

Source: Author.

As can be seen in figure 3, of the 2,388 records, 20% were taken to carry out the respective tests, this represents 478 records for the validation of the algorithm. The result was that, out of 352 records that do not have Alzheimer's, the algorithm successfully predicted 350 and out of 126 records in which patients have Alzheimer's, it successfully predicted 123. Although the accuracy is not 100%, it is too close, for this reason, this prediction model is chosen.

However, it is tested with the other chosen model, the Naive Bayes model, from which its own algorithm was also created, and it was found that in effect, the average of success is very low as well. In this case it is approximately 63%.

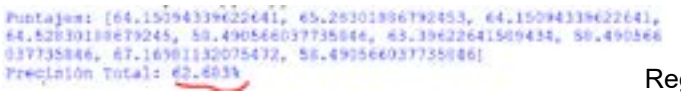

Table 2. Result using Naive Bayes algorithm

Fountain. Author.

As mentioned before, the chosen model is the K-Nearest Neighbor.

VI. ANALYSIS OF RESULTS

The algorithm for the early detection of Alzheimer's disease is non-invasive, does not require any examination or procedure of the patient; it is based solely on his medical history. Therefore, its use and implementation is quite easy.

Thanks to the studies carried out by ADNI (Alzheimer's Disease Neuroimaging Initiative), over the years and the entire number of patients enrolled in them, it was possible to obtain a fairly thick data report with a lot of relevant and important information, making the prediction algorithm possible. Although a cleaning and information extraction process was carried out, since the variables used in the research were not so specified, they were still totally necessary for the final result.

Another important point was the Python libraries involved in the creation of the Algorithm, since they allow a fairly deep data treatment by applying prediction methods and in this way, selecting the one that best suits the present investigation, which in this case was K-Nearest Neighbors, offering accuracy levels of approximately 99%.

It can be observed that there are certain parameters that have a greater probability in the development of Alzheimer's, for example the APOE Genotype, which had a higher frequency of affirmative results for the disease when it was 4, therefore, an elevation in the affirmative final result when there are certain combinations of the variables can be contrasted.

Regarding the precision of the predictive model used, a quite satisfactory result was reached, 99% is quite high and while it always has the exceptions and some errors can occur, it was greatly improved with the use of the K-Nearest Neighbor predictive model. To further reduce the chances of error, a larger dataset would be required, and one that is obviously consistent with the one you already have, because the precision is given by the amount of learning of the algorithm. But for now you it is limited to the amount of used records..

A random record will be taken and tested on the K-Nearest algorithm to see the result and check against the correct one.

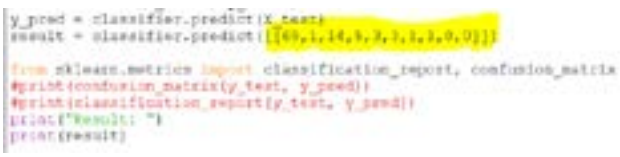

Figure 4. Predictor of specific cases.

Source: Author

According to that registry, the patient is a 69-year-old white man who did 14 years of study, who suffers from diabetes and hypertension. Besides, his APOE Genotypes are both 3. According to ADNI records, this patient suffers Alzheimer's, but it will be checked if when passing this data again, the algorithm returns this result.

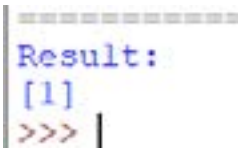

Figure 5. Test result.

Source: Autho

In effect, the algorithm has returned a positive result for Alzheimer's, that is, it was correct in the evaluation of the patient.

For classification, the K-Nearest neighbor or k-nearest neighbors (kNN) algorithm is used, and it was the one with the highest precision in practice. Perhaps to determine why, it should be remembered that it is an instance-based machine learning method. This method is characterized by memorizing only all the training examples available during the training phase. During the test phase, the data to be classified is compared with these examples based on a previously defined distance measure. The most similar data is called "nearest neighbor", however, it is also possible to include the k closest neighbors in the calculation. The influence of potential error values may be limited, although considering more than one neighbor does not inevitably lead to more precise classifications (Lüschow - Wartena, 2018). In the experiments for the present work, we established k = 6.

KNN was chosen due to its simplicity and understandability that allows a first understanding of our classification approach without having to deal with complex or advanced algorithms. There are several measures of distance to define the similarity of two data. We use the Euclidean distance that calculates the square root of the sum of all the squares of the attribute value differences:

$$\sqrt{(a_1^{(1)} - a_1^{(2)})^2 + (a_2^{(1)} - a_2^{(2)})^2 + \cdots + (a_y^{(1)} - a_y^{(2)})^2}$$

Figure 6. Euclidean Distance

Source: Author

VII. DISCUSSION OF RESULTS

With the help of previous Alzheimer's studies, it was possible to establish a relationship between certain important parameters of each patient and the disease as such, which was a quite viable way to create a prediction model of the disease. Finally, a discreet result is obtained, limiting the investigation to an affirmative or negative diagnosis for Alzheimer's disease.

The values of the parameters chosen in each of the test subjects were contrasted, establishing relationships in the probability in the development of the disease. Positive cases for Alzheimer's were studied, determining that certain factors affect more than others. For example, it

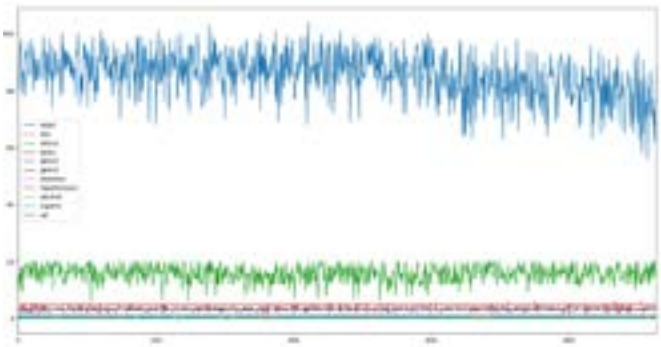

Figure 7. Variables Graph (Blue is the age), it is observed that the majority of cases occur after the age of 80. Source: author.

It was found, for example, that although all the studies suggest that the more years of education a person undertakes, the less likely they are to develop Alzheimer's, however, in practice, taking the results, it is verified that this theory would not appear to be functional, since the test subjects suffering from Alzheimer's are mostly lawyers and people with years of study greater than 10. In fact, most of those who are illiterate or have less than 10 years of study do not suffer from Alzheimer's. This obviously not in order to deny, much less go against such studies, but to look from an engineering perspective on certain curiosities present in the data management process.

Given this result, although it could be said that subjects with Alzheimer's, and with a high educational level, are more likely to develop AD, it would go against the medical studies that support the opposite. Therefore, analyzing the information, it was found that subjects with less than 10 years of education are very unlikely in the DataSet. That is, there are very few records compared to those over 10 years old. This means that it is not that they are more likely to suffer from AD, but that there are fewer records of people with little education. So it is more complicated for the algorithm to determine the prediction for people with less than 10 years of study.

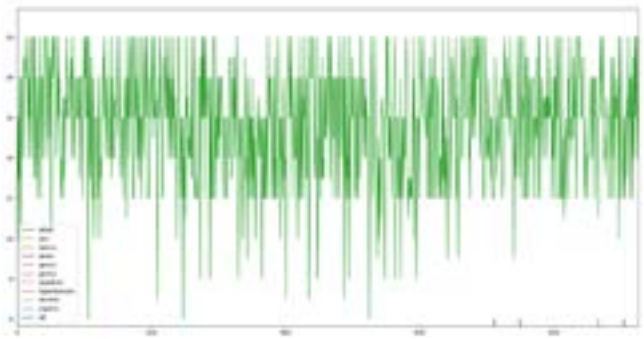

Figure 8. Years of study in test subjects suffering Alzheimer's. Source: Author

On the contrary, it is known that of the 927 subjects who have Alzheimer's, the APOE 2 Genotype, are mostly 4 as it is said in medical studies, however, it can also occur in 3 and in very distant cases in 2, and indeed, according to the Genotype 2 graph, there is only one case out of 3 with Alzheimer's, 430 with 3 and 496 with 4, which is consistent with the research.

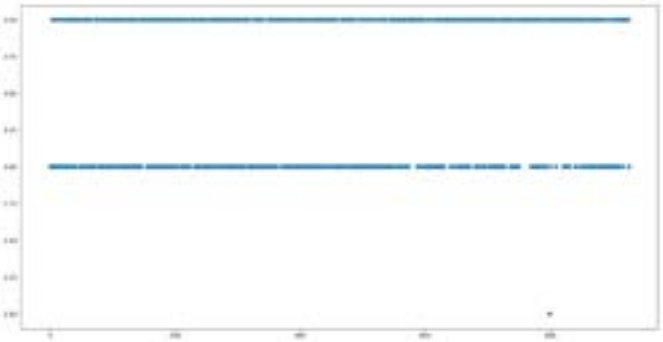

Figure 9. APOE 2 genotype for Alzheimer's patients. Source: Author.

Once all the development of the algorithm has been completed, several similarities are established with the related works that were consigned in that section of this Paper. The first of the works, although it makes a fairly accurate prediction (90%), the parameters it uses are 120 proteins present in the patients. However, the methodology is the same. Regarding the second and third related work, I think they are the most similar to the present one, since they also use both sociodemographic and other variables, in their case they use psychiatric, syntactic, linguistic parameters and Alzheimer's test results, but the methodology of using various aspects of the patient, it seems to me to be correct and similar to the one used. The last two related works present a similar methodology, however, one uses image analysis, since it processes the neural images of the patients, and the other processes the results in the tests themselves.

In conclusion, this document provides greater precision than the related works, but using the variables that have the most correlation among those chosen from the database.

VIII. CONCLUSIONS

The data is extracted and transformed to a CSV file. Once with the data obtained, it is processed using binary classification methods of machine learning, to verify the similarity, relationship and subsequent prediction of the data taken. Once the correct prediction model had been chosen, tests were carried out in the form of a prototype of the algorithm, in which certain parameters were loaded, and depending on them, a result of the disease was offered or not. The accuracy of the model is 99%, in case the parameters coincide with those of the CSV file; that is, it is completely based on those records.

While there are currently many Alzheimer's prediction algorithms, the current one takes a combination of parameters that none have, on certain occasions one or more of them is taken, but not all and in the combination that is taken. These were chosen from the beginning, taking into account that Alzheimer's can be more easily predicted if a combination of sociodemographic, physiological, and antecedents of a certain pair of diseases are taken. So it is considered that this was the success of the investigation.

The algorithm for early prediction of Alzheimer's disease using machine learning can be supplemented in the future if a user interface is applied and the resource is uploaded to the cloud, or stored on a server, offering free use to the entire community and in a simple and easily manageable way. Apart from being a prediction algorithm based on Machine Learning, it clearly depends on the data it uses to learn. Therefore, it should constantly be updated with truthful information that contributes to a more accurate precision of the environments that are proposed.

The approach of the proposed model is carried out based on premises and medical studies of the relationship of certain health antecedents in the onset of Alzheimer's, which were tested by implementing the algorithm for the early detection of Alzheimer's, making use of the information registered in the ADNI database.

The amount of information was reduced from 4565 to 2388 final records, which are complete and clean and helped the final 99% accuracy to be achieved. In addition, when selecting the KNN algorithm, it was ensured that each record had a categorical prevalence, so the prediction increased its precision by 37%.

IX. BIBLIOGRAPHIC REFERENCES

[1] Aguilar N., Reyes G. (2007) Alcohol, tabaco y deterioro cognoscitivo en adultos mexicanos mayores de 65 años, (salud pública de méxico / vol.49, suplemento 4 de 2007).

[2] Alsaedi, A., & Qader, I. A. (2018). Extended Cox Proportional Hazard Model to Analyze and Predict Conversion From Mild Cognitive Impairment To Alzheimer's Disease, 131– 136.

[3] Alzheimer's association. (2016). Información básica sobre la enfermedad de Alzheimer.

Alzheimer's Association, 16, 1–30.

[4] Alzheimer's Association. (2012). Diagnóstico de la enfermedad de Alzheimer y de demencia. Retrieved from http://www.alz.org/documents/greaterillinois/Diagnosis_.pdf

[5] Arnold S., Arvanitakis Z. (2018), Brain insulin resistance in type 2 diabetes and Alzheimer disease: concepts and conundrums. Nat Rev Neurol. 2018 Mar; 14(3): 168–181

[6] Cacabelos, R (2001). Enfermedad de Alzheimer Presente terapéutico y retos futuros. Retrivied from http://www.scielo.org.co/pdf/rcp/v30n3/v30n3a02.pdf

[7] Dantas, L., & Valenca, M. (2014). Using Neural Networks in the Identification of Signatures for Prediction of Alzheimer's Disease. Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 2014–Decem, 238–242. https://doi.org/10.1109/ICTAI.2014.43

[8] DiBattista A. M, Heinsinger N, (2018), Alzheimer's Disease Genetic Risk Factor APOE-ε4 Also Affects Normal Brain Function. Curr Alzheimer Res. 2016; 13(11): 1200–1207.

[9] Farrer, L. A. (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. JAMA: The Journal of the American Medical Association, 278(16), 1349–1356. https://doi.org/10.1001/jama.278.16.1349

[10] Giacometti-Rojas. (2013) Technological innovation and development of competitive advantage in health care. A conceptual and methodological approach, Julio, 1657-7027

[11] Huang, M., Yang, W., Feng, Q., Chen, W., Weiner, M. W., Aisen, P., … Fargher, K. (2017). Longitudinal measurement and hierarchical classification framework for the prediction of Alzheimer's disease. Scientific Reports, 7(January), 1–13. https://doi.org/10.1038/srep39880

[12] Kivipelto, M., Helkala, E., Laakso, M. P., Hänninen, T., Hallikainen, M., Alhainen, K., … Nissien, A. (2001). Midlife vascular risk factors and Alzheimer's Disease in later life: Longitudinal, population based study. Bmj, 322(June), 1447–1451. https://doi.org/10.1136/bmj.322.7300.1447

[13] Lahad MD, Thomas D., Bird MD (1996). Genetic factors in Alzheimer's disease: A review of recent advances, 1996. https://doi.org/10.1002/ana.410400604

[14] Land, W. H., & Schaffer, J. D. (2016). A Machine Intelligence Designed Bayesian Network Applied to Alzheimer's Detection Using Demographics and Speech Data. Procedia Computer Science, 95, 168–174. https://doi.org/10.1016/j.procs.2016.09.308

[15] Lüschow A. , Wartena C. (2018) Classifying Medical Literature Using k-Nearest-Neighbours Algorithm, (ORCiD: 0000-0001-5483-1529) University of Applied Sciences and Arts Hanover 2018.

[16] Peña-Casanova (1999), Enfermedad de Alzheimer, del diagnóstico a la Terapia: conceptos y hechos. Enfermedad de Alzheimer Neuropatología, 4, 14-15 .

[17] Petersen R. (2010) Alzheimer's disease: progress in prediction. The Lancet Neurology. 9 (1): 4-5.

[18] Reitz, C., Tang, M.-X., Schupf, N., Manly, J. J., Mayeux, R., & Luchsinger, J. A. (2010). A Summary Risk Score for the Prediction of Alzheimer Disease in Elderly Persons. Archives of Neurology, 67(7), 835–841. https://doi.org/10.1001/archneurol.2010.136

[19] Ruitenberg A, van Swieten JC, Witteman JCM, Mehta KM, van Duijn CM, Hofman A, et al. (2002) Alcohol consumption and risk of dementia: the Rotterdam Study. Lancet 2002;359:281-286.

[20] Sánchez, C. R. De, Nariño, D., Fernando, J., & Cerón, M. (2010). Epidemiología y carga de la Enfermedad de Alzheimer. Acta Neurológica Colombiana, 26(3), 87–94. https://doi.org/Acta Neurol Colomb 2010;26:Sup (3:1):87-94

[21] Seixas, F. L., Zadrozny, B., Laks, J., Conci, A., & Muchaluat Saade, D. C. (2014b). A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. Computers in Biology and Medicine, 51, 140– 158. https://doi.org/10.1016/J.COMPBIOMED.2014.04.010

[22] Selkoe. (2001). DJ. Alzheimer's disease: genes, proteins, and therapy. Physiol Rev. 2001; 81: 741-766.