Aplicación Software para la clasificación de pacientes con Alzheimer basado en técnicas de data mining

Valeria Andrea Álvarez Zárate
Departamento de Ingeniería de Sistemas y Computación
Universidad del Norte
Barranquilla, Colombia
valeriaz@uninorte.edu.co

Luis Eduardo Sepulveda Cobo
Departamento de Ingeniería de Sistemas y Computación
Universidad del Norte
Barranquilla, Colombia
lesepulveda@uninorte.edu.co

Jesus David Padilla Woltmann
Departamento de Ingeniería de Sistemas y Computación
Universidad del Norte
Barranquilla, Colombia
woltmannj@uninorte.edu.co

Profesor asesor de proyecto
Ph.D. Wilson Nieto Bernal
Profesor Asistente
Dpto. Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
wnieto@uninorte.edu.co

***Abstract*— In Colombia, 342,956 people over 60 will suffer from some type of dementia, of which about 75% will suffer from Alzheimer's disease (AD). For this reason, the team developed a software tool that allows the classification of patients with respect to the level of dementia they have from the data entered by doctors when conducting a new consultation. using the OASIS open database. A Python script was created which is integrated in Django and is supported by the analyzed open data. Random Forest technique was implemented using sets of CART classifier trees, which use the Gini index to separate the criteria. Using this model and the results obtained in the area under the ROC curve (AUC), the program successfully classifies patients with a 92% probability among subjects with sanity and dementia, in addition, a classification with a probability of 81% with respect to the Clinical Dementia Rating (CDR) of the patients is done. The risk factors associated with each patient are obtained from the information provided in the forms and evaluations.**

Keywords: Alzheimer's disease (AD), Random forest, CART, Gini Index, Clinical Dementia Rating (CDR), Risk Factors.

A. INTRODUCCIÓN

a. Definición del Problema

El alzheimer es una enfermedad neurodegenerativa caracterizada por el atrofiamiento progresivo de la corteza cerebral lo cual conduce a la pérdida de la memoria, generando déficits cognitivo y potencial pérdida de las funciones motoras del paciente. La enfermedad del alzheimer es tipo de demencia más común que pueden llegar padecer las personas, siendo esta una enfermedad incurable es de vital importancia hacer una detección temprana de la misma. (Fulton et al.,2019)

La demencia afecta alrededor de 50 millones de personas a nivel mundial, dentro de esta cifra cerca del 60% de estos casos son presentados en países de bajos y medios recursos. El Alzheimer ocupa entre el 60% y 70% de los casos actuales de demencia (Organización mundial de la salud, 2019), de este modo la mayoría de los casos de demencia presentados son de Alzheimer, es por ello que es necesario tener un mayor control y seguimiento de estas enfermedades, *Alzheimer's Association* (2019) afirma que sólo el 16% de personas de la tercera edad reciben evaluaciones cognitivas regulares durante chequeos de rutina, estos datos son alarmantes ya que el Alzheimer es una enfermedad mortal, de hecho es la quinta causa principal de muerte entre las personas de 65 años en adelante, disminuyendo de esta manera la esperanza de vida de aquellos que la padecen.

Cada 3 segundos una persona en el mundo desarrolla demencia, para el año 2015 46.8 millones de personas padecían alzheimer y según predicciones de *Alzheimer's Disease International* (2015) esas cifras se duplicarán cada 20 años . En la actualidad la mayoría de personas que padecen Alzheimer no han recibido un diagnóstico oficial, impidiendo de este modo la detección temprana de esta enfermedad, sólo el 16% de personas de la tercera edad reciben evaluaciones cognitivas regulares durante chequeos de rutina (Alzheimer's Association, 2019).

La Organización iberoamericana de seguridad social se estima que para el año 2020, en Colombia 342.956 personas mayores de 60 años padecerán algún tipo de demencia, de las cuales cerca del 75% sufrirán de alzheimer, siendo en su mayoría mujeres las afectadas por esta enfermedad (OISS, 2019), El ministerio de Salud y protección social colombiano afirma que las personas que padecen alzheimer viven en promedio 7,1 años (MinSalud, 2017), es por ello que la detección temprana del Alzheimer es vital para aumentar la calidad de vida de la población colombiana y asegurar de mismo modo el control y manejo

de esta enfermedad dentro de la población colombiana, para dicho cometido existen diferentes pruebas que ayudan a medir el nivel de progresión de la enfermedad, entre dichos tests se encuentran el tests Minimental (MMSE), la clasificación clínica de la demencia (CDR) y Evaluación cognitiva de Montreal (MoCA), los cuales son tests de vital importancia en el contexto colombiano, ya que son los más utilizados en el país.

Los tests de detección de alzheimer son de vital importancia no solo para medir el nivel de progresión de esta patología, sino también para permitir a los médicos proporcionar intervenciones en el período pre-clínico asintomático. Si bien esta enfermedad no posee una cura, es posible ralentizar la progresión de la misma y su sintomatología en su etapa temprana (Fulton et al., 2019). Es por esto que a lo largo de este proyecto el grupo de trabajo se centrará en la implementación de dos técnicas de clasificación basada en la aplicación de estos tests para determinar si un paciente padece alzheimer y poder tener un monitoreo adecuado de la condición del paciente. Las técnicas utilizadas son: Random forest y Decision trees, esto debido a que la base de datos con la cual estaremos tratando trabaja de manera positiva con estos algoritmos. Por otro lado en el estado del arte, estos algoritmos son de los más implementados para la clasificación de datos sobre todo aquellos en los cuales se manejan un gran número de campos. Más adelante en el proyecto se hablará a fondo de estos métodos y su implementación.

b. Mapa del problema

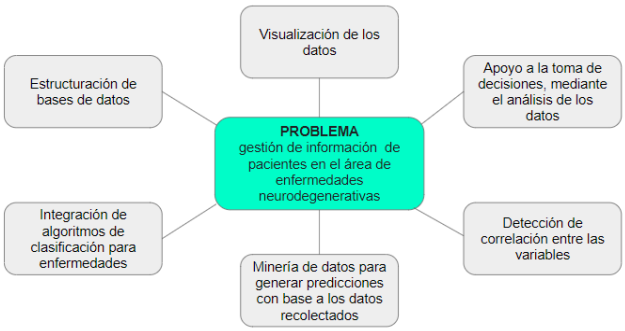


Figura 1. Mapa del problema

c. Justificación

El Alzheimer es una enfermedad que, lastimosamente, no tiene cura, lo único que se puede hacer frente a esta e es ralentizar el progreso de la misma, un gran porcentaje de la población sufre de esta enfermedad, es por ello que tener una buena percepción de la data disponible es importante para aliviar los efectos del Alzheimer y es vital para abrir camino a una futura cura de la enfermedad (Ertek, G., Tokdil, B., & Günayddn, b., 2018)

Entender los posibles factores de riesgo ante la enfermedad, ayuda no solo a tener una mejor comprensión de las causas de esta, sino también detectar aquellas variables que potencian y estimulan el progreso del Alzheimer. Es por ello que tener los datos correctamente clasificados y registrados es un requisito fundamental para el seguimiento y estudio de la enfermedad.

Dentro del tratamiento del Alzheimer una herramienta de registro de los datos, y clusterización ayudaría a en gran medida a analizar el posible progreso de la enfermedad en los pacientes, para esto dentro de este proyecto se estará trabajando con técnicas de minería de datos acorde al comportamiento de la "open data" de la cual se estará haciendo uso a lo largo del desarrollo de la aplicación web para la clínica de la memoria, con el fin de crear correlaciones entre los datos, encontrar posibles factores de riesgo y facilitar la visualización de la data.

d. Objetivo General

Diseñar e implementar una solución de software para la clasificación de personas con Alzheimer, implementando técnicas de análisis de minería de datos para ayudar a la toma de decisiones de los doctores y facilitar el tratamiento de los pacientes.

e. Objetivos Específicos

- Elaborar la revisión sistemática de la literatura relacionada con el desarrollo de software para control de registros de pacientes, mediante la implementación de algoritmos de clustering y minería de datos.
- Modelar y diseñar arquitectura de la solución para el seguimiento de pacientes con síntomas de alzheimer.
- Desarrollar el Prototipo de la herramienta software para la clasificación de pacientes con alzheimer utilizando minería de datos
- Validar el prototipo de la herramienta software para la clasificación de pacientes con alzheimer implementando minería de datos

B. Revisión de Literatura

a. Mapa Conceptual

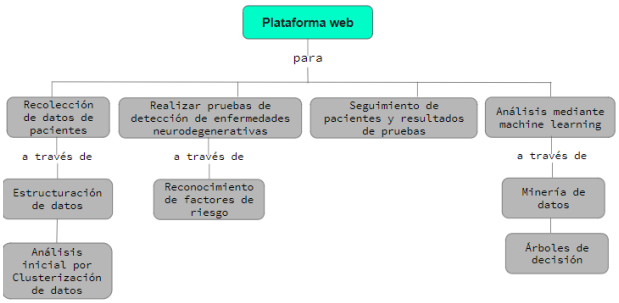


Figura 2. Mapa conceptual de la plataforma a desarrollar

b. Marco conceptual

En la actualidad muchos hospitales cuentan con sistemas automatizados para el registro de sus pacientes, sin embargo pocos de estos sistemas apoyan al tratamiento de estos analizando los factores de riesgos que pueden existir y facilitando el diagnóstico de los enfermos. La demanda de métodos ágiles, confiables y sensibles para la detección de la demencia crece cada vez más, dando énfasis a la rentabilidad en el sector de la atención médica. A medida que el volumen de pacientes de demencia vaya en aumento, los profesionales de la salud necesitarán herramientas más rápidas y eficientes para la clasificación adecuada de los pacientes (Mani, S., Dick, M., Pazzani, M., Teng, E., Kempler, D., & Taussig, I., 1999), para esto la implementación de técnicas de minería de datos resulta una solución factible.

La minería de datos puede ser definida como el proceso de hallar patrones y tendencias anteriormente desconocidos

dentro de la base de datos y usar la información encontrada para construir modelos de predicción (Koh, H. C., & Tan, G., 2011). En el área de la salud se han utilizado técnicas de minería de datos para el análisis y clusterización de enfermedades neurodegenerativas.

Fulton et al.(2019) realizaron un estudio para predecir la presencia del Alzheimer en los pacientes haciendo uso datos sociodemográficos, clínicos y de resonancia magnética. En dicho estudio una máquina impulsada por gradiente (GBM) predijo la presencia del Alzheimer en función del género, la edad, la educación, el estado socioeconómico y el examen de estado minimental (MMSE). Por otro lado una red residual con 50 capas (ResNet-50) predijo la presencia y la gravedad de la clasificación de demencia clínica (CDR) a partir de MRI (clasificación de múlti-clases). El estudio concluyó que los modelos GBM pueden ayudar a proporcionar una detección inicial basada en análisis sin imágenes, mientras que los modelos de red ResNet-50 pueden ayudar a identificar a los pacientes con EA automáticamente antes de la revisión del proveedor, ayudando de este modo a los familiares de las personas con alzheimer a reducir y planear costos a lo largo del avance de la enfermedad.

Ertek, G., Tokdil, B., & Günayddn, b. (2018) analizaron los factores de riesgo para la identificación del alzheimer, usando sólo datos no relacionados con imagen, usando técnicas de aprendizaje automático y minería de datos. Los métodos aplicados incluyen análisis de árbol de clasificación, análisis de conglomerados, visualización de datos y análisis de clasificación. Abdullah, S., et al (2019) realizaron el mismo proceso teniendo como enfoque la enfermedad del parkinson para la identificación de factores de riesgo y predicción de caídas, implementando tres métodos estadísticos para la clasificación: árboles de decisión, bosques aleatorios y regresión logística.

En trabajo realizado por Alashwal, H., El Halaby, M., Crouse, J., Abdalla, A., & Moustafa, A. (2019) estudiaron y revisaron los métodos de agrupamiento que se han aplicado a los conjuntos de datos de enfermedades neurológicas, especialmente la enfermedad de Alzheimer(EA). El objetivo de dicho estudio es proporcionar información sobre qué técnica de agrupamiento es más adecuada para dividir a los pacientes con EA basándose en su similitud. Implementando diferentes técnicas de Clustering utilizando métodos no supervisados concluyeron que el análisis de agrupamiento puede señalar varias características que subyacen a la conversión de AD en etapa temprana a AD avanzada.

Muchos hospitales y empresas dependen del manejo y control de datos de sus clientes, es por ello que tener una herramienta para el manejo de los registros a través de servidores de bases de datos se vuelve vital para ejercer su trabajo, Farooqui, N., & Mehra, R. (2018) dentro de su estudio buscan facilitar el manejo y recuperación de los datos mediante un sistema de información médica que simplifique el manejo de la data de pacientes, dentro de su artículo proponen una metodología para la construcción de una data warehouse para un sistema de información médica usando técnicas de minería de datos. El almacén de datos consiste de una base de datos informativa que se crea mediante la transformación de la base de datos operativa. Un analista de datos puede analizar los datos y tomar decisiones, de esta forma el sistema propone las sugerencias y predice las enfermedades con la ayuda de los datos.

Dentro de la medicina la minería de datos ha sido ya ampliamente utilizada, Delshi Howsalya Devi, R., & Deepika, P. (2015) realizaron un estudio donde comparaban diferentes técnicas de clustering para el diagnóstico del cáncer de seno, se acuerdo con los resultados del trabajo experimental, compararon cinco técnicas de agrupación como DBSCAN, Farthest first, canopy, LVQ y agrupación jerárquica en el software Weka.

C. METODOLOGÍA

Dentro de este proyecto de investigación se realizarán dos procesos en paralelo, en primer lugar el desarrollo de la plataforma web para el control de registros de datos de pacientes con alzheimer y como segundo proceso de minería de datos de los pacientes para permitir realizar analitica de los datos y detectar correlaciones en los datos.

A. Desarrollo web

Como primera metodología se implementará el método para desarrollo ágil Scrum, dicho método nos permite reaccionar rápidamente en caso de necesitar realizar cambios inesperados dentro el proyecto. El método Scrum es actualmente uno de los más utilizados al realizar desarrollo de software ágil ya que este trabaja mediante sprints, el propósito de estos se enfoca en establecer entregables en cortos periodos de tiempo comenzando con planeación y terminando en revisión del entregable. Scrum es un método ideal para pequeños grupos de desarrolladores, ya que permite a equipos auto-organizados producir pequeñas piezas de software entregables en cortos lapsos de tiempo.

Mahalakshmi, M., & Sundararajan, M. (2015) afirman que Scrum es una de las metodologías más populares y poderosas ya que consta de muchas ventajas y del mismo es implementada en el campo del software de manera exitosa. Las características de Scrum son Reuniones, Rol de Scrum y los artefactos. Dichos roles son:

- Dueño de producto: en este proyecto el dueño de producto corresponde a los médicos expertos de la clínica de la memoria de la universidad del norte, quienes se encargan de establecer los requerimientos del proyecto.
- Scrum Master: este es el encargado de lograr las metas trazadas, realizando revisiones y retroalimentaciones mediante la supervisión del equipo, el encargado este rol es el profesor supervisor del proyecto.
- Miembros de equipo: un grupo auto.organizado que trabaja en la implementación de las tareas asignadas, en este caso el equipo de trabajo está constituido por los estudiantes encargados de desarrollar el proyecto.

Xuesong Zhang, & Dorn, B. (2011) identificaron la importancia de implementar buenas prácticas dentro del desarrollo con metodología Scrum tales como:

- Scrum diarios: el objetivo de este es coordinar y mejorar la comunicación del equipo, de este modo permite facilitar la colaboración dentro de los miembros de equipo, como medio de trabajo el equipo de desarrollo utiliza whatsapp y skype ya que permite comunicación de manera inmediata y remota.
- Backlogs: dentro de la metodología scrum existen dos tipos de backlogs para mantener registro de la lista de trabajo a lo largo del ciclo de vida del proyecto. El contenido e importancia de cada ítem de la lista del Product Backlog se deriva de la lista otorgada por el dueño de producto y mantenido y actualizado por el líder del equipo, el esfuerzo de desarrollo asociado lo establece el equipo en su conjunto. En el Sprint Backlog prioriza y expande cada item del Product Backlog en tareas más detalladas de este modo el equipo puede trabajar cada tarea de manera más efectiva y realizar commits menos complejos dentro de la iteración del sprint. Dentro del desarrollo de esta investigación se estará utilizando esta metodología para poder agilizar, mejorar y acomodar los items necesarios de acuerdo con la prioridad que tengan, permitiendo desglosar los items del Product Backlog en tareas más específicas y de este mismo modo remover tareas cuando el equipo lo considere necesario.
- Sprints: esta es la unidad básica del la metodología de desarrollo Scrum. los Sprints usualmente tienen corta duración, todos los Sprints del Scrum mantienen igual duración , en este proyecto se trabajará con Sprint de 5 días, ya que es un proyecto de corta duración. Existen dos tipo de reuniones de Sprints los de planeación (los cuales se realizan al comienzo de cada Sprint) y los de revisión (que es al finalizar todas las tareas del Sprint, en esta reunión se realizan anotaciones sobre las modificaciones necesarias que serán asignadas al siguiente Sprint). Para el desarrollo de la plataforma de registros el equipo de trabajo realiza estas reuniones de manera virtual, utilizando correos, Skype y Whatsapp, y el tracking de las tareas se realizó utilizando excel.

B. Data Mining

Para Minar los datos de los pacientes de ALzheimer se utilizará la metodología de minería de datos CRISP-DM (Proceso estándar de la industria cruzada para la minería de datos) , Bin-Hezam, R., & E., T. (2019) afirma que: "CRISP-DM es uno de los modelos de proceso más utilizados para el análisis predictivo de datos, Las fases del ciclo de vida del proyecto son la comprensión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación, la implementación y el monitoreo". La manera en que se comportan las seis fases del modelo CRISP-DM se relacionan entre sí se puede observar en la Figura 3.

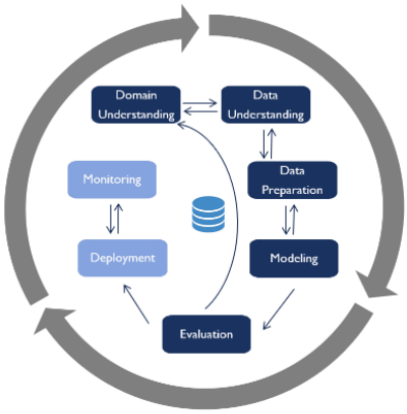


Figura 3. Fases del modelo CRISP-DM adaptado por Bin-Hezam, R., & E., T. (2019).

Olegas NIAKŠU (2015) Define cada fase de la siguiente forma:

1. Comprensión del negocio (Business understanding): en esta fase el equipo de trabajo se encarga de establecer los objetivos del proyecto de analitica de datos implementando mineria de datos, estableciendo los criterios exito que definen el proyecto. Este proyecto se centra en la comprensión del problema "Alzheimer" dado que no se está realizando un negocio sino una investigación, para esto se utilizan como criterios de éxito la capacidad de identificar factores de riesgo mediante la identificación de correlaciones entre las variables.
2. Comprensión de los datos (Data understanding): en esta fase comienza con la recopilación de datos iniciales y el acceso al conjunto de datos. Los problemas de calidad de los datos deben identificarse y se crean los supuestos iniciales que los conjuntos de datos pueden ser de interés para los pasos posteriores. El dataset con el cual se estará trabajando a lo largo del proyecto será una base de datos de libre acceso llamada OASIS (Open Access Series of Imaging Studies).
3. Preparación de datos (Data preparation): La fase de preparación de datos cubre todas las actividades

necesarias para preparar previamente el conjunto de datos final. Las actividades de la fase de preparación de datos dependen en gran medida de las características y la calidad de los datos sin procesar originales. Algunas de las tareas características de la preparación de datos incluyen la elección de una tabla, proyecciones de atributos y registros, transformación de atributos, clasificación, normalización, eliminación de ruido y muestreo.

4. Modelado (Modeling): En esta fase, se realiza una selección adecuada de técnicas de modelado, algoritmos o combinaciones de los mismos. Luego, se eligen los valores óptimos de los parámetros del algoritmo. La calidad del modelo se evalúa formalmente. Para evaluar la calidad del modelo, se utilizan métricas que son populares en la minería de datos y estadística: sensibilidad, precisión, especificidad y curva ROC (característica de funcionamiento del receptor).
5. Evaluación (Evaluation): Antes del despliegue final del modelo, es esencial evaluar cuidadosamente, revisar los pasos de construcción del modelo y asegurarse de que los objetivos comerciales se alcancen adecuadamente. El resultado final de esta fase: la elección de si los resultados de DM se pueden usar en entornos prácticos.
6. Implementación y el monitoreo (Deployment): en esta fase los conocimientos adquiridos deberían estructurarse y presentarse al usuario final de forma comprensible.

C. Revisión de la Literatura

Fase 0. Para apoyar el proceso de investigación dentro del proyecto se usaron las siguientes fuentes:

- Sibila +
- IEEE Xplore
- ResearchGate
- Web of Science
- Latvia University of Life Sciences and Technologies
- Cornell university
- IOP Science
- USA national library of medicine
- citeseerx

siendo las principales fuentes Web of Science y IEEE Xplore, ya que las otras fuentes fueron utilizadas para buscar documentos específicos

Fase 1. Búsqueda enfocada en la temática del alzheimer y la minería de datos

La búsqueda por tema se realizó en las 2 fuentes principales:

Concepto	IEEE Xplore	Web of Science
Health database data mining	874	1083
Scrum for web development	25	45
Alzheimer data mining	109	243
total	1008	1371

Implementando filtros con palabras clave

Palabra	IEEE Xplore	Web of Science
CRISP	5	8
Agile	23	32
Dementia	31	87
total	59	127

Fase 2. Lectura del Abstract y conclusión

Fuente	No. de artículos
IEEE Xplore	25
ResearchGate	10
Web of Science	7
Latvia University of Life Sciences and Technologies	1
Sibila +	3
Cornell university	4
IOP Science	6
USA national library of medicine	2
citeseerx	1
Total	59

Fase 3. Lectura en profundidad.

Fuente	No. de artículos
IEEE Xplore	11
ResearchGate	4
Web of Science	1
Latvia University of Life Sciences and Technologies	1

Sibila +	1
Cornell university	1
IOP Science	1

USA national library of medicine	1
citeseerx	1
Total	22

Fase 4. Tabla de Referencias finales

No.	Titulo del articulo	Palabras claves	fuente
1	Fulton et al.(2019). Classification of Alzheimer's Disease with and without Imagery using Gradient Boosted Machines and ResNet-50. Brain Sciences, 9(9), 212. doi: 10.3390/brainsci9090212	Alzheimer’s disease; extreme gradient boosting; deep residual learning; convolutional neural networks; machine learning; dementia	ResearchGate
2	Ertek, G., Tokdil, B., & Günayddn, b. (2018). Supplement for 'Risk Factors and Identifiers for Alzheimer's Disease: A Data Mining Analysis'. SSRN Electronic Journal. doi: 10.2139/ssrn.3151516	Alzheimer’s Disease (AD), analysis of risk factors, identifying tests that can help diagnose AD. non-image data, techniques from machine learning and data mining. classification tree analysis, cluster analysis, data visualization, and classification analysis. assess possible risks, preventive measures.	ResearchGate

3	Mani, S., Dick, M., Pazzani, M., Teng, E., Kempler, D., & Taussig, I. (1999). Refinement of Neuro-psychological Tests for Dementia Screening in a Cross Cultural Population Using Machine Learning. Artificial Intelligence In Medicine, 326-335. doi: 10.1007/3-540-48720-4_35	Knowledge Discovery from Databases (KDD) approach. Decision tree learners (C4.5, CART). rule inducers (C4.5 Rules, FOCL) Reference classifier (Naive Bayes)	ResearchGate
4	Latifah, E., Abdullah, S., & Soemartojo, S. (2018). Identifying of factor associated with parkinson's disease subtypes using random forest. Journal Of Physics: Conference Series, 1108, 012064. doi: 10.1088/1742-6596/1108/1/012064	Classification by = { Tremor dominant (TD), Postural Instability Gait Difficulty (PIGD) } Random forests CART C4.5	IOP Science
5	Abdullah, S., et al (2019). Assessing the predictive ability of the UPDRS for falls classification in early stage Parkinson's disease. Cornell university.	Risk factor decision trees, random forests, and logistic regression, Gini index criterion used for decision trees and random forests, Bayesian model averaging receiver operating characteristics (ROC) sensitivity.	Cornell university
6	Alashwal, H., El Halaby, M., Crouse, J., Abdalla, A., & Moustafa, A. (2019). The Application of Unsupervised Clustering Methods to Alzheimer's Disease. Frontiers In Computational Neuroscience, 13. doi: 10.3389/fncom.2019.00031	k-means, k-means mode, Multi-layer clustering Hierarchical agglomerative clustering	USA national library of medicine

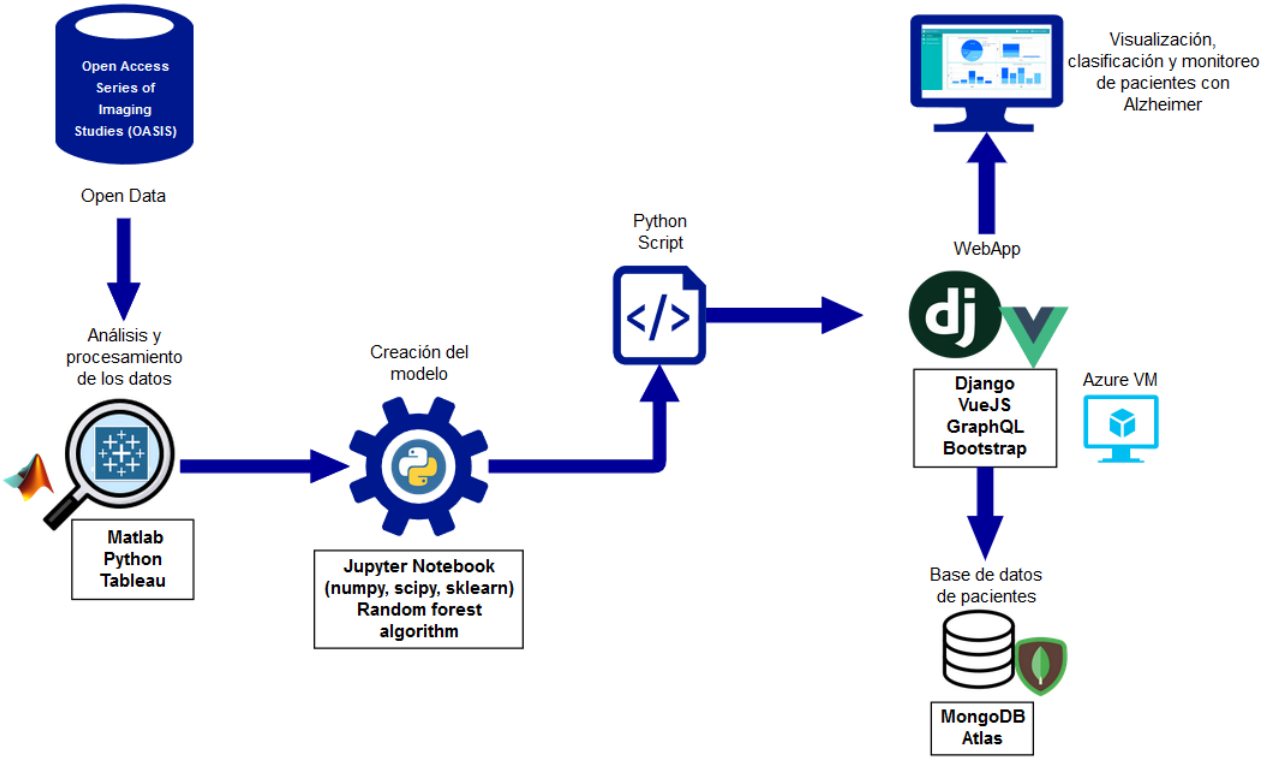
7	Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of healthcare information management, 19(2), 65.	CRISP-DM Methodology	citeseerx
8	Olegas NIAKŠU (2015). CRISP Data Mining Methodology Extension for Medical Domain. Baltic J. Modern Computing, Vol. 3	CRISP-DM Phases: Business understanding, Data understanding, Data preparation, Modeling Evaluation, Deployment	Latvia University of Life Sciences and Technologies
9	Farooqui, N., & Mehra, R. (2018). Design of A Data Warehouse for Medical Information System Using Data Mining Techniques. 2018 Fifth International Conference On Parallel, Distributed And Grid Computing (PDGC). doi: 10.1109/pdgc.2018.8745864	Data Mining, Clinical Data, Database Server, Data Warehouse, Medical Information System	IEEE Xplore
10	Delshi Howsalya Devi, R., & Deepika, P. (2015). Performance comparison of various clustering techniques for diagnosis of breast cancer. 2015 IEEE International Conference On Computational Intelligence And Computing Research (ICCIC). doi: 10.1109/iccic.2015.7435711	Cluster, machine learning, Data Mining. Farthest first algorithm, Hierarchical Clustering	IEEE Xplore

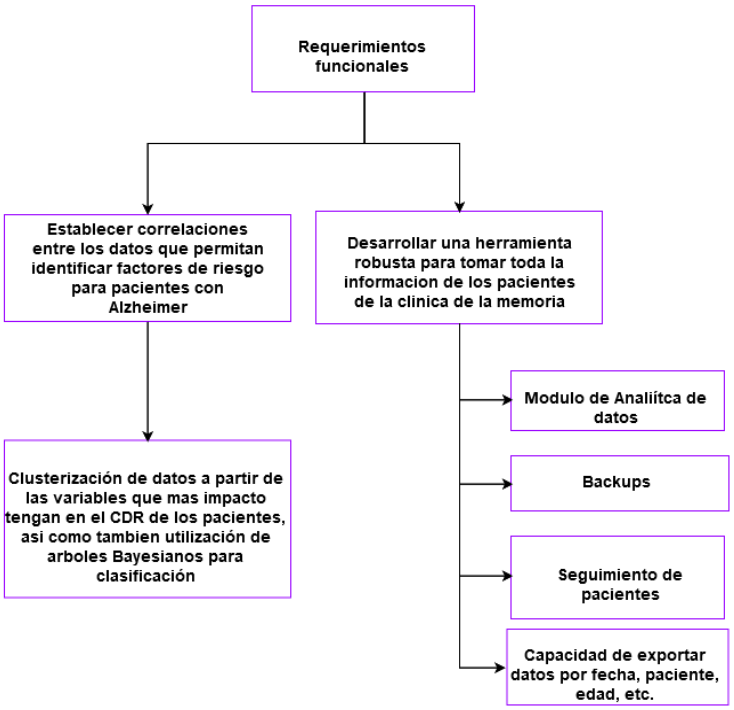
11	Patel, K., & Thakral, P. (2016). The best clustering algorithms in data mining. 2016 International Conference On Communication And Signal Processing (ICCSP). doi: 10.1109/iccsp.2016.7754534	Cluster size, Clustering Algorithms, Data normalization, Time complexity of algorithm	IEEE Xplore
12	Hunt, M., von Konsky, B., Venkatesh, S., & Petros, P. Bayesian networks and decision trees in the diagnosis of female urinary incontinence. Proceedings Of The 22Nd Annual International Conference Of The IEEE Engineering In Medicine And Biology Society (Cat. No.00CH37143). doi: 10.1109/iembs.2000.900799	Data mining, big data, data classification Bayesian Network, Decision Tree, Expert System	IEEE Xplore
13	Ranganatha, S., Anusha, C., Vinay, S., & Pooja Raj, H. (2013). Medical data mining and analysis for heart disease dataset using classification techniques. National Conference On Challenges In Research & Technology In The Coming Decades (CRT 2013). doi: 10.1049/cp.2013.2485	Data mining, classification, entropy, gain, health informatics, ID3, Naïve Bayesian.	IEEE Xplore
14	Xiaofeng Zhao, Liyan Jiao, Jinping An, Li Wang, & Lipin Jia. (2010). A data mining method on the study of medical information. 2010 International Conference On Computer Application And System Modeling (ICCASM 2010). doi: 10.1109/iccasm.2010.5623058	Data mining, Apriori algorithm, medical data	IEEE Xplore

15	Ilayaraja, M., & Meyyappan, T. (2013). Mining medical data to identify frequent diseases using Apriori algorithm. 2013 International Conference On Pattern Recognition, Informatics And Mobile Engineering. doi: 10.1109/icprime.2013.6496471	Frequent Diseases; Data Mining; Medical Data; Association Rule; Apriori Algorithm	IEEE Xplore
16	Monali Dey et al.(2014) Study and Analysis of Data mining Algorithms for Healthcare Decision Support System.International Journal of Computer Science and Information Technologies, Vol. 5	Naive Bayes, C4.5, healthcare decision support,Neural network	Sibila +
17	Lyman, J., Scully, K., & Harrison, J. (2008). The Development of Health Care Data Warehouses to Support Data Mining. Clinics In Laboratory Medicine, 28(1), 55-71. doi: 10.1016/j.cll.2007.10.003	Data mining, data warehouses, online analytical processing, Data acquisition and processing	ResearchGate
18	Yu, H., & Wang, D. (2012). Research and Implementation of Massive Health Care Data Management and Analysis Based on Hadoop. 2012 Fourth International Conference On Computational And Information Sciences. doi: 10.1109/iccis.2012.225	EMR (Electronic Medical Record) , PHR(Personal Healthcare Record), Hadoop, HDFS	IEEE Xplore

19	Chen, B., Liao, Q., & Tang, Z. (2007). A Clustering Based Bayesian Network Classifier. Fourth International Conference On Fuzzy Systems And Knowledge Discovery (FSKD 2007). doi: 10.1109/fskd.2007.8	Bayesian network classifiers, Bayesian network structure learning algorithm(CBNA), Hierarchical clustering.	IEEE Xplore
20	Bin-Hezam, R., & E., T. (2019). A Machine Learning Approach towards Detecting Dementia based on its Modifiable Risk Factors. International Journal Of Advanced Computer Science And Applications, 10(8). doi: 10.14569/ijacsa.2019.0100820	Machine learning; classification; data mining; data preparation; dementia; modifiable risk factors	IEEE Xplore
21	Mahalakshmi, M., & Sundararajan, M. (2015). Tracking the student's performance in Web-based education using Scrum methodology. 2015 International Conference On Computing And Communications Technologies (ICCCT). doi: 10.1109/iccct2.2015.7292779	Scrum framework, Proposed Framework, Artifacts	IEEE Xplore
22	Xuesong Zhang, & Dorn, B. (2011). Agile Practices in a Small-Scale, Time-Intensive Web Development Project. 2011 Eighth International Conference On Information Technology: New Generations. doi: 10.1109/itng.2011.187	Agile; Scrum; Web development; Project management	IEEE Xplore

D. ARQUITECTURA DE LA SOLUCIÓN





A. Explicación de la arquitectura

La arquitectura lógica de la solución se divide en 2, una para el desarrollo de la herramienta software y otra para la minería de datos, que finalmente se juntan.

La herramienta es desplegada en una aplicación web, en la cual todos los datos de los pacientes y de los tests que son recolectados en la herramienta se almacenan en la nube a partir del servidor MongoDB Atlas. Todos los datos son tratados por el backend desarrollado en Django por medio de una API que posteriormente el Frontend consume utilizando GraphQL al momento de que se realiza alguna consulta en la aplicación.

En relación a la minería de datos, se utiliza la metodología CRISP-DM para la manipulación de datos públicos. Utilizando MATLAB se analiza la correlación entre los datos y crea una proyección de estos donde se analizan los porcentajes de varianza para establecer el nivel de confiabilidad de los mismos dentro del dataset. A partir de las herramientas Tableau prep y excel se realiza un proceso ETL para la limpieza del dataset. Posteriormente utilizando Python y las librería de numpy, scipy y scikit-learn se analizan estos datos para establecer el mejor modelo que se ajuste a ellos. Finalmente a partir del modelo creado en python se desarrolla un script que se ejecuta en el backend de la Aplicación y se visualiza en el frontend.

La presentación de la información se da por medio de Vue.js utilizando como librería Vue Bootstrap para asegurar la responsividad de la aplicación. A partir de esto se da la visualización de factores de riesgo y clasificación a partir de los datos recogidos en el formulario. Adicionalmente se cuenta con el módulo de Dashboard el cual es creado usando ZingChart.js para establecer los gráficos y finalmente los módulos de gestión de la información y monitoreo de los pacientes.

E. DATA Y MODELO

A. Explicación del Dataset

El caso de estudio hará uso de data extraída de Open Access Series of Imaging Studies (OASIS), se usará dicha base de datos ya que esta fue creada con el fin de brindar un libre acceso a la información para la comunidad científica y para la investigación al compilar y distribuir libremente conjuntos de datos de neuroimagen, buscando facilitar futuros descubrimientos en neurociencia básica y clínica. Esta base de datos contiene información de pacientes con y sin alzheimer entre 60 y 96 años, de los cuales 64 individuos fueron diagnosticados con la enfermedad del Alzheimer de carácter leve o moderado durante su primera visita usando el sistema de clasificación CDR. Otros 14 de los individuos se caracterizaron como no dementes en el momento de una o más exploraciones y luego se determinó clínicamente que tenían Alzheimer en el momento de una exploración posterior. (Marcus et al., 2010).

El dataset consiste en una coleccion de 354 observaciones a un total de 142 pacientes, dado que dentro de los datos recolectados también se recopila el seguimiento de los pacientes, es probable que un sujeto aparezca más de una vez en los registros, todos los pacientes son diestros, los datos incluyen hombres y mujeres, de igual manera se tiene registro de su estado socioeconómico de los sujetos. Demented y non-demented son las clases en las que se clasifica si un paciente tiene o no Alzheimer. Converted hace referencia a cuando el paciente desarrolla la enfermedad a lo largo de las pruebas (Ertek, G., Tokdil, B., & Günayddn, b., 2018), dentro de nuestro caso de estudio no incluimos a los convertidos para generar mayor diferenciación entre las dos clases, y de esta forma el clasificador sea más definitivo.

En la tabla 1. se puede observar la explicación de cada uno de los campos de la base de datos y el valor que toman para su interpretación, y si fueron o no tomados en cuenta para nuestro caso de estudio.

Atributo	**Explicación y Rango de valores**	**Incluido**
Group	se etiqueta como Demented con valor de -1, non-demented con valor de 1, or converted con valor de 0.	Sí, sin embargo todos los convertidos son tomados en cuenta como dementes para facilitar la clasificación de los individuos.
MRIID	ID de la prueba, que contiene un valor único para cada valoración registrada enumeradas del 1 a 354.	No
SubjectID	ID del paciente de 1 a 142. Un paciente puede estar realizando seguimiento por lo tanto debe tener un ID que lo identifique y diferencie de los demás pacientes.	No, ya que no es un dato que sea necesario ya que nuestro caso de estudio no tendrá en cuenta si un paciente se convierte o no, por lo tanto el seguimiento es irrelevante para el proceso de clasificación.
Visit	Visita del paciente de 1 a 5.	Si
MRDelay	El retraso de un sujeto desde la última visita.	Si

CDR	Clasificación de demencia clínica. 0 = no demencia, 0.5 = Alzheimer muy leve, 1 = Alzheimer leve, 2 = Alzheimer moderado. (Morris, 1993)	Si, se utilizaron solo 3 niveles de clasificación, sano (0), demencia cuestionable (1) y demencia leve (1 y 2)
Gender	Masculino (M) o Femenino (F)	Si, dentro de nuestra implementación M toma el valor numérico de 1 y F de -1
Age	edad del sujeto en observación	Si
EDUC	Nivel de educación	Si
SES	Estado socioeconómico, que se evalúa mediante el Índice de Posición Social de Hollingshead. 1 (estado más alto) a 5 (estado más bajo). (Hollingshead, 1957)	Si
MMSE	Valor del examen del estado mini mental. 0 (peor valor) a 30 (mejor valor). (Folstein, Folstein y McHugh, 1975)	Si
eTIV	Volumen intracraneal total estimado (cm3) (Buckner et al., 2004)	Si
nWBV	Volumen normalizado de todo el cerebro, expresado como un porcentaje de todos los vóxeles (Fotenos et al., 2005)	Si
ASF	Factor de escala del atlas; factor de escala de volumen para el tamaño del cerebro.	Si

Tabla 1. La explicación y los rangos de valores de los atributos del conjunto de datos OASIS.

Los atributos se pueden clasificar entonces en:
Información demográfica

- Edad
- EDUC
- SSE

Información clínica

- MMSE
- CDR
- eTIV
- nWBV
- ASF

La base de datos OASIS hace especial énfasis en la recolección de información acerca de las imágenes longitudinales del cerebro ya que estas han demostrado ser útiles para estudiar el envejecimiento normal y enfermo. Se ha demostrado que la disminución del volumen total del cerebro, medida a partir de una MRI adquirida longitudinalmente, evoluciona a un ritmo cercano al 0.5% por año en adultos mayores no dementes, algo mayor que la observada en adultos más jóvenes. Se ha demostrado que los volúmenes de todo el cerebro y las estructuras asociadas con la memoria se atrofian a un ritmo significativamente mayor en el deterioro cognitivo leve y la EA temprana. De este modo las medidas longitudinales de la estructura cerebral están surgiendo como herramientas para rastrear la progresión de la enfermedad y como medidas de resultado adjuntas en ensayos clínicos (Marcus et al., 2010).

B. Preparación de la data

Antes de iniciar a trabajar con los datos de OASIS, fue necesario estandarizar los valores de forma que todos los valores sean numéricos para así agilizar el tratamiento de los datos y proceder a analizar el comportamiento de los mismos.

En primera instancia para establecer la matriz de correlación se decidieron quitar los campos de ID del paciente e ID de la consulta, ya que se consideró que no tendrian ningun aporte al análisis.
Posteriormente se realizó una transformación de los 3 grupos en los que se categoriza el paciente de acuerdo a la siguiente tabla:

Grupo al que pertenece el paciente	**Valor al que se transformó**
Demente	-1

No-Demente	1
Convertido	0

Para iniciar el proceso de comprensión de los datos se hizo uso de una matriz de correlación entre las diferentes variables del dataset, esto con el objetivo de evaluar la fuerza y dirección de la relación entre los diferentes campos de OASIS modificado por el grupo del proyecto. De este modo, si la correlación entre los elementos es es alta y positiva significa que están altamente correlacionadas. También puede verse el caso en que la correlación sea alta y negativa, en ambos casos al estar las variables fuertemente correlacionadas por lo tanto estas miden la misma destreza o característica. Si los elementos no están altamente correlacionados, entonces los elementos pudieran medir diferentes características o no estar claramente definidos.

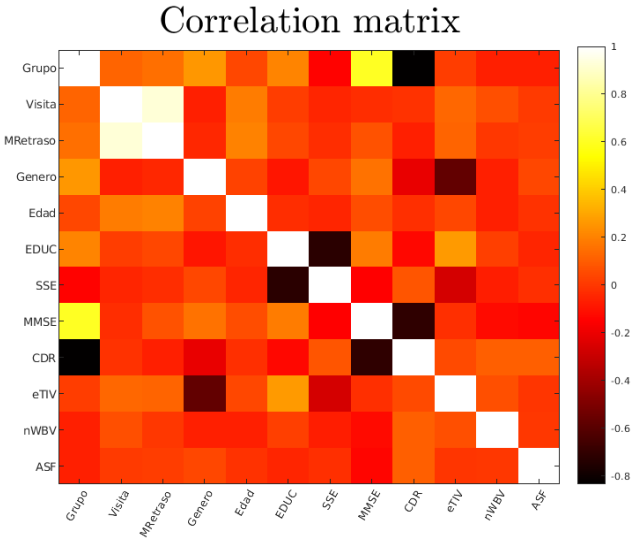


Figura 4. Matriz de correlación

En esta matriz de correlación podemos ver que las correlaciones más fuertes son:

- CDR y Grupo: correlación negativa.
- EDUC y SSE: correlación negativa.
- CDR y MMSE:correlación negativa.
- eTIV y Género: correlación negativa.
- MMSE y Grupo: correlación positiva.

En la matriz de correlación podemos observar que la correlación más fuerte se genera entre CDR y Grupo, es por ello que se consideran estas dos variables cómo las determinantes para identificar si un paciente padece Alzheimer o no.

Igualmente se realizaron las proyecciones en 2 y 3 dimensiones y se estableció el porcentaje de varianza en ambos casos.

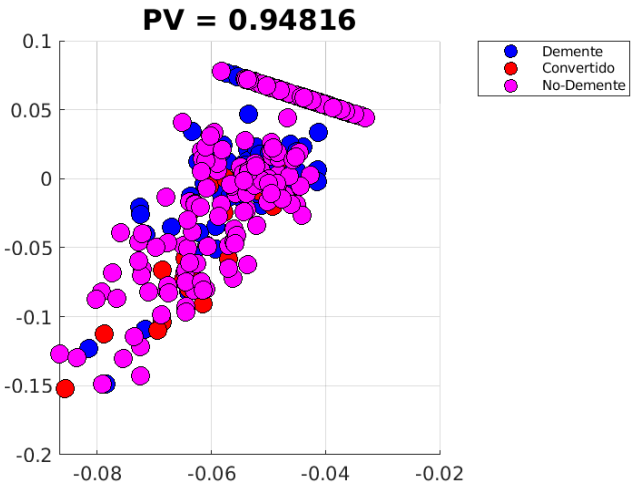


Figura 5. Proyección en 2 dimensiones

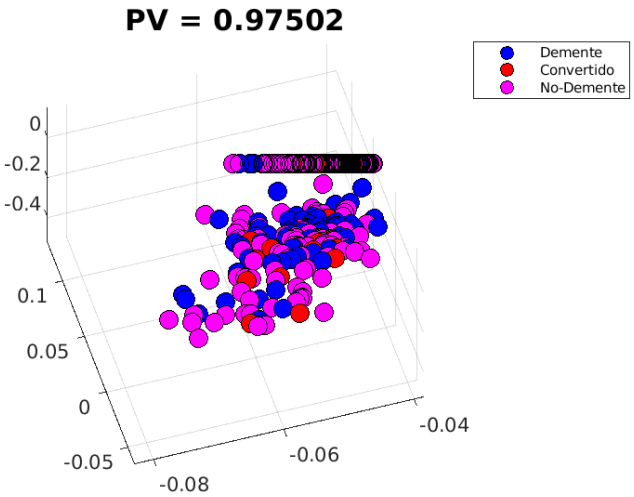


Figura 6. Proyección en 3 dimensiones

En la figura 5 se puede observar la proyección en 2 dimensiones cuya varianza es de 0.94816, mientras que en la figura 6 se muestra la proyección en 3 dimensiones cuya varianza es 0.97502. como en ambos casos la varianza es mayor a 0.9 podemos asegurar que los datos del dataset OASIS son altamente confiables.

Una vez comprobado el comportamiento de los datos se procede entonces a realizar diagramas de caja y bigotes, diagramas de dispersión y diagramas de barra para representar de forma visual el comportamiento de los datos, usando la base de datos OASIS modificada con base a los datos cuya correlación fue alta.

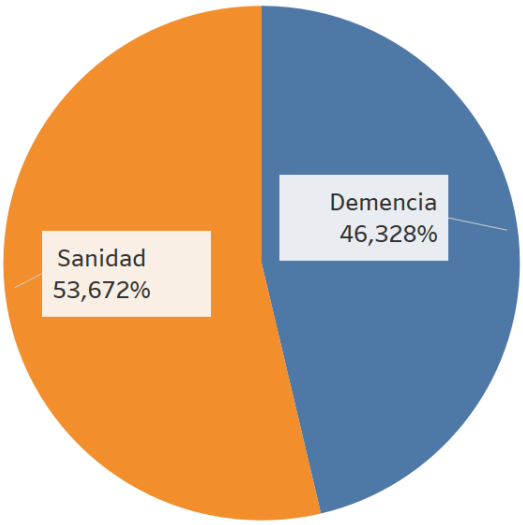


Figura 7. Gráfica circular Demencia vs. Sanidad

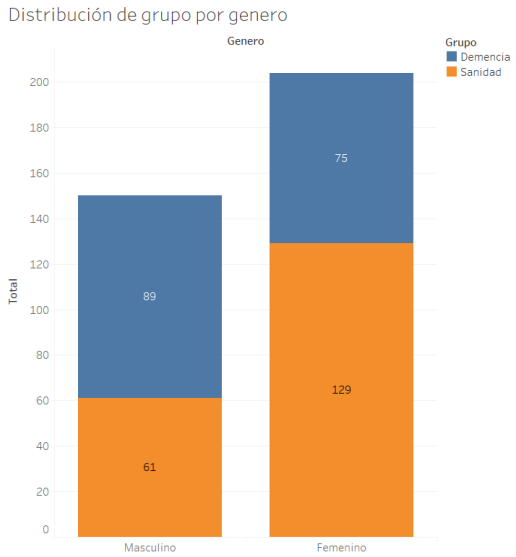


Figura 8. Distribución de Grupo por Género.

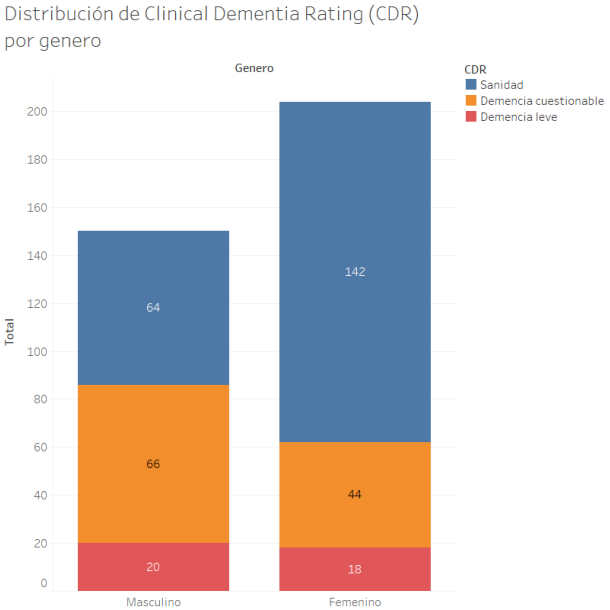


Figura 9. Distribución de CDR por Género

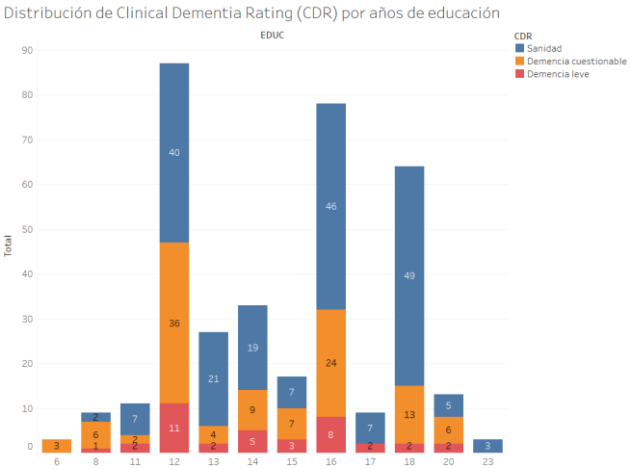


Figura 10. Distribución de CDR por años de educación

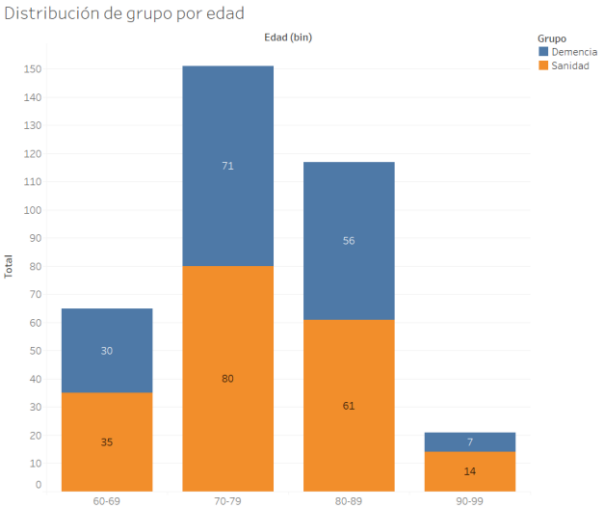


Figura 11. Distribución de grupo por Edad

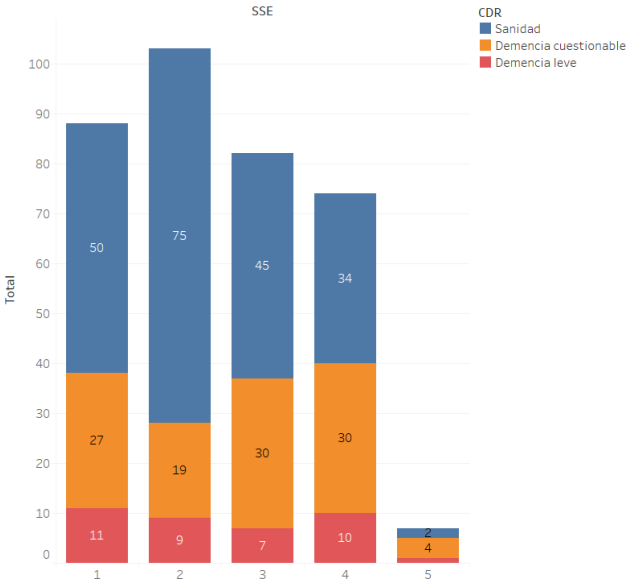


Figura 12. Distribución de CDR por SSE

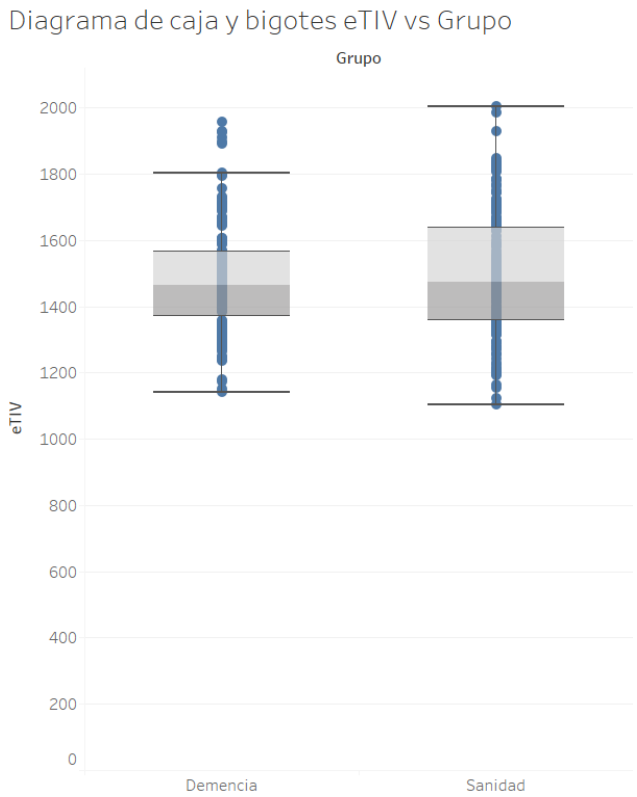


Figura 13. Diagrama de caja y bigotes etiV vs.Grupo

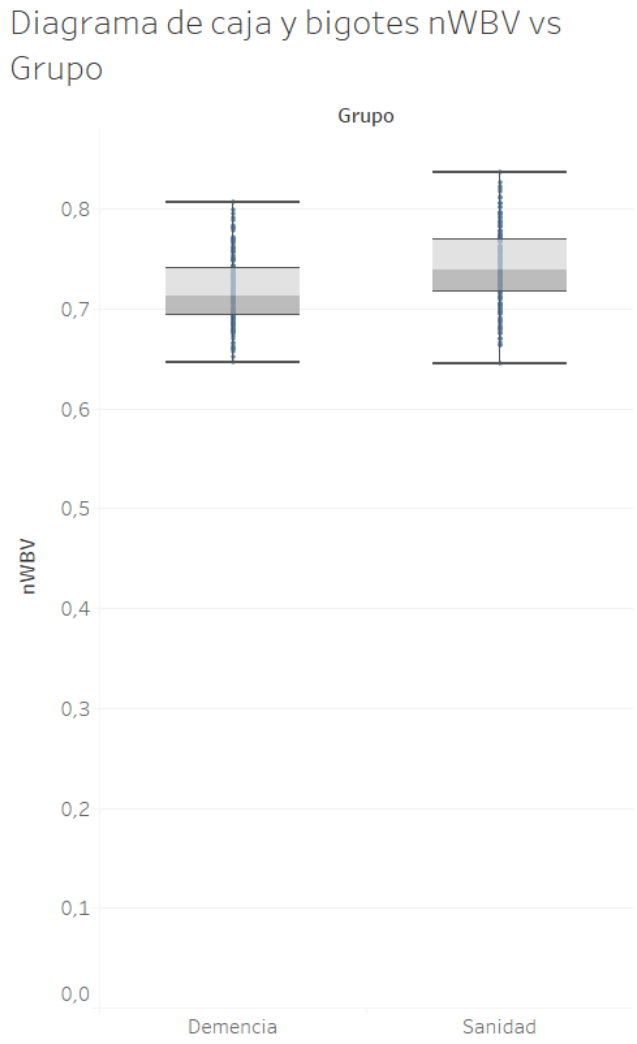


Figura 14. Diagrama de caja y bigotes nWBV vs Grupo

A partir de los datos observados se puede decir lo siguiente:

- Los pacientes con demencia tienden a tener menor cantidad de años de educación
- El volumen cerebral de los pacientes sin demencia es mayor a los de los pacientes con demencia
- En este dataset los hombres fueron más propensos a poseer demencia que las mujeres

Una vez terminada la exploración de los datos el equipo de trabajo escogió los algoritmos con los cuales trabajaría la información para generar un seguimiento adecuado y apoyo a la toma de decisiones.

C. Modelos

Como mencionado anteriormente, los dos métodos a utilizar serán árboles de decisión y random forest.

Un árbol de decisión es es un modelo basado en árbol que divide los datos repetidamente de acuerdo con valores de corte específicos en las características. Se crean diferentes subconjuntos mediante la división, separando las instancias para que pertenezcan a un subconjunto. Los subconjuntos intermedios son los nodos internos, mientras que los subconjuntos finales son los nodos hoja. Los árboles de decisión son más útiles si la relación entre las características y el objetivo son no lineales o si hay interacciones entre las características. (Bin-Hezam, R., & E., T. 2019).

Los árboles de decisión tienen como mayor desventaja el hecho que son propensos al sobreajuste, esto se refiere a cuando el algoritmo continúa haciendo crecer el árbol para reducir el error para los datos de entrenamiento, esto resulta en un mayor error para los datos de prueba (Latifah, E. et al., 2018). es por ello que se decidió usar también el método de random forest para comparar la calidad de los resultados entre ambas implementaciones.

Random forest es un enfoque en conjunto que ofrece una alternativa sólida a los árboles de decisión. Este método consta de dos pasos importantes. Primero, la muestra se obtiene tomando una muestra aleatoria con el reemplazo de datos (bootstrapping). Segundo, se realiza una selección aleatoria de características, en este proceso cierto tamaño de la variable candidata se toma aleatoriamente de todos los predictores que se usaron para dividir un nodo. Como resultado, los dos pasos anteriores producirán muchos árboles para formar un bosque con las diferencias de tamaño y forma de los árboles (Latifah, E. et al., 2018).

Una vez desarrollados los modelos se procede a validar la calidad de los mismos, para validar el modelo se usa el accuracy, el F1, la matriz de confusión el ROC y el AUC. La matriz de conflicto es un método utilizado para describir

de manera rigurosa el desempeño de las clasificaciones realizadas, por otro lado también tenemos el porcentaje de accuracy del algoritmo al momento de arrojar los resultados de clasificación, es decir, la proporción de todos los sujetos que están correctamente clasificados, mientras que F1 es el promedio ponderado de precisión y recuperación. La característica de funcionamiento del receptor (ROC) evalúa el rendimiento real de un modelo al tiempo que considera todos los posibles umbrales, mientras que su área bajo la curva (AUC) resume los cambios de los umbrales de sensibilidad y 1- especificidad.

Existen diferentes formas de construir un árbol de decisión, en esta ocasión se decidió trabajar con el algoritmo CART (Classification and Regression Trees) que utiliza el método de Gini, este método usa el índice Gini el cual es utilizado para medir cuánto afecta cada especificación mencionada directamente en el caso resultante. El método Gini es utilizado para crear puntos dividido calculando el índice gini mediante la siguiente fórmula:

$$\square\square\square\square(\square) = 1 \sum_{\square} \quad - \square_{\square}^{2}$$

El índice de Gini describe los nodos hijos no homogéneos del nodo t debido a la división basada en la variable $\square_{\square}$ El índice de Gini tiene un rango de valores entre 0 y 1. Cuanto menor sea el índice de Gini, mayor será la homogeneidad de un nodo, de modo que será mejor el proceso de separación de objetos en clases existentes.Se necesita una comparación de los niveles no homogéneos entre los nodos principales (antes de dividir) con los niveles no homogéneos del nodo secundario (después de dividir) para determinar qué tan bien una variable predictora está en dividir un nodo. Un predictor que produce la máxima diferencia del índice de Gini entre el nodo primario y el nodo secundario para ser seleccionado como el mejor divisor, el proceso de división continuará hasta que no se genere un aumento en la homogeneidad en el nodo secundario. (Latifah, E. et al., 2018)

F. PROTOTIPO

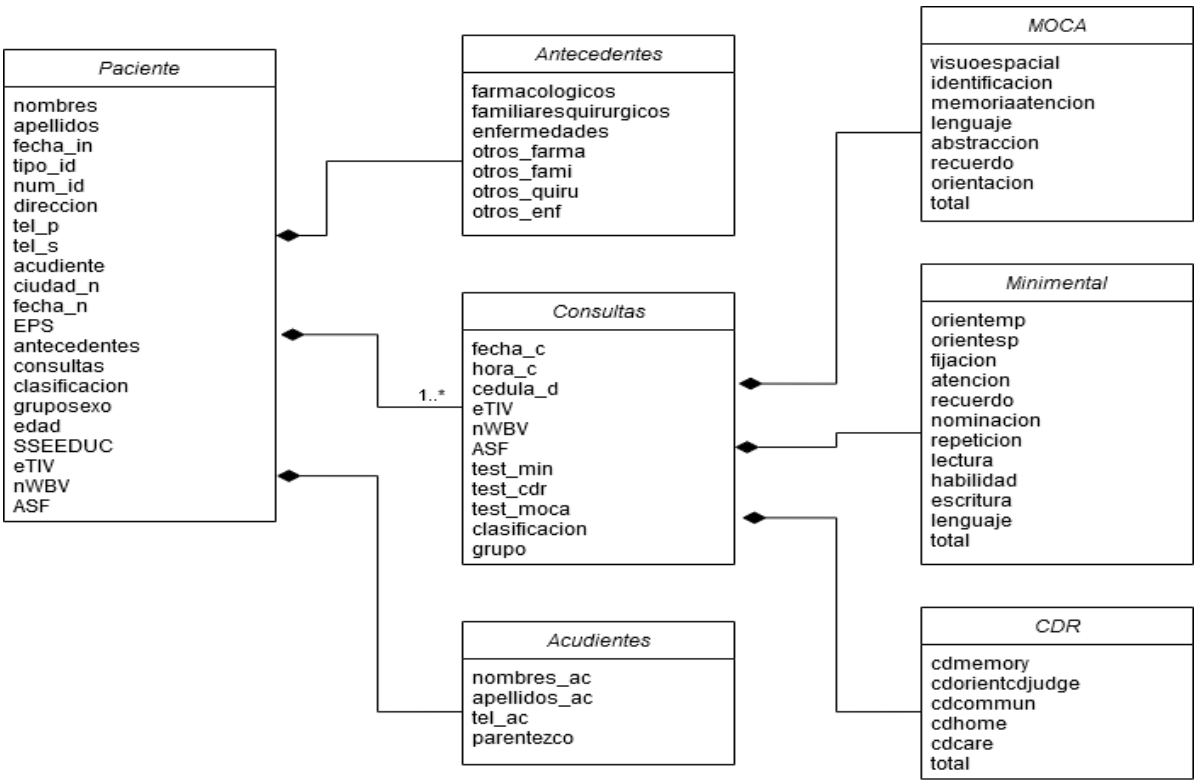


Figura 15. Diagrama de clases del prototipo.

Para la implementación del prototipo para la clínica de la memoria se establecieron primero las clases con las cuales contaría este. Como se puede observar en la Figura 15, las principales clases son:

- **Pacientes:** los cuales cuentan con antecedentes, acudientes y varias consultas, ya que a un mismo paciente se le puede estar realizando seguimiento.
- **Antecedentes:** ya que es importante para saber si hay posibles factores de riesgo dentro de estos, además de que todo centro medico debe tener claridad de los antecedentes médicos de sus pacientes.
- **Acudientes:** ya que se está tratando con pacientes a los cuales se le monitorea la posibilidad de padecer Alzheimer, una enfermedad degenerativa de la memoria, es necesario tener una persona responsable del sujeto de prueba.
- **Consultas:** Esta es la clase más importante para el seguimiento de los pacientes, ya que en estas consultas se les realiza a los pacientes las pruebas que ayudan a determinar si un paciente padece o no de Alzheimer. Todos los sujetos de prueba deben tener al menos una consulta.

Una vez se tiene claridad en la estructuración y comportamientos entre las clases, se procede al desarrollo de la herramienta web para la clínica de la memoria.

Figura 16. Landing page de la clínica de la memoria

En el landing page de la clínica de la memoria se puede ingresar una nueva consulta de un paciente o puede dirigirse directamente a ver la analitica de los datos con los que ya cuenta la clínica.

Cuando se dirige el doctor a generar una nueva consulta primeramente se piden los datos personales y los antecedentes del paciente, como se puede observar en la Figura 17, luego se procede a recolectar los datos de la consulta. El doctor realiza las pruebas correspondientes e ingresa los datos relacionados al volumen intracraneal (Figura 18) del paciente obtenidos de estudios que se haya realizado el sujeto de prueba. Al escoger una de las pruebas a realizar se despliega en pantalla las preguntas correspondientes a esta, como se puede ver en la Figura 19, donde el doctor se encarga de seleccionar aquellas preguntas que el pacientes logra responder adecuadamente.

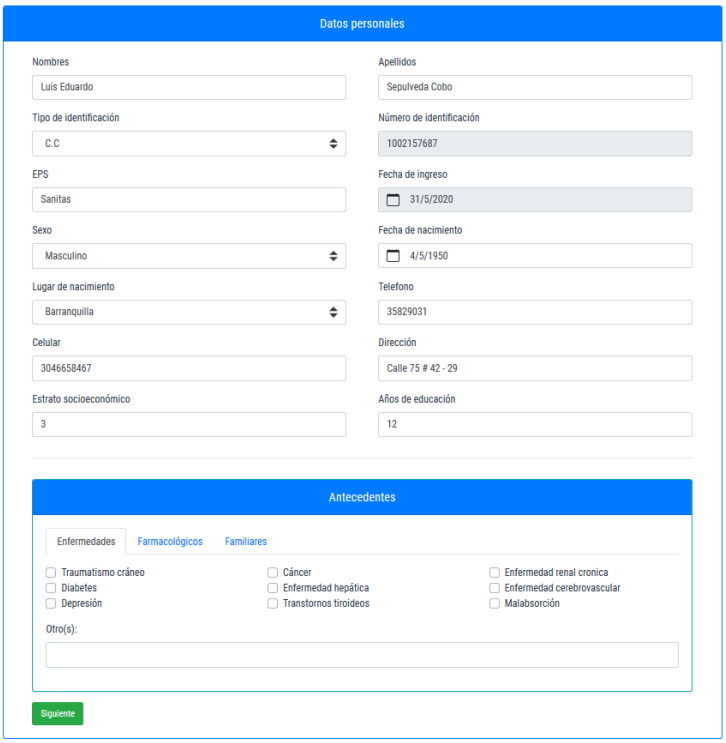


Figura 17. Pagina para ingresar datos personales y antecedentes del paciente

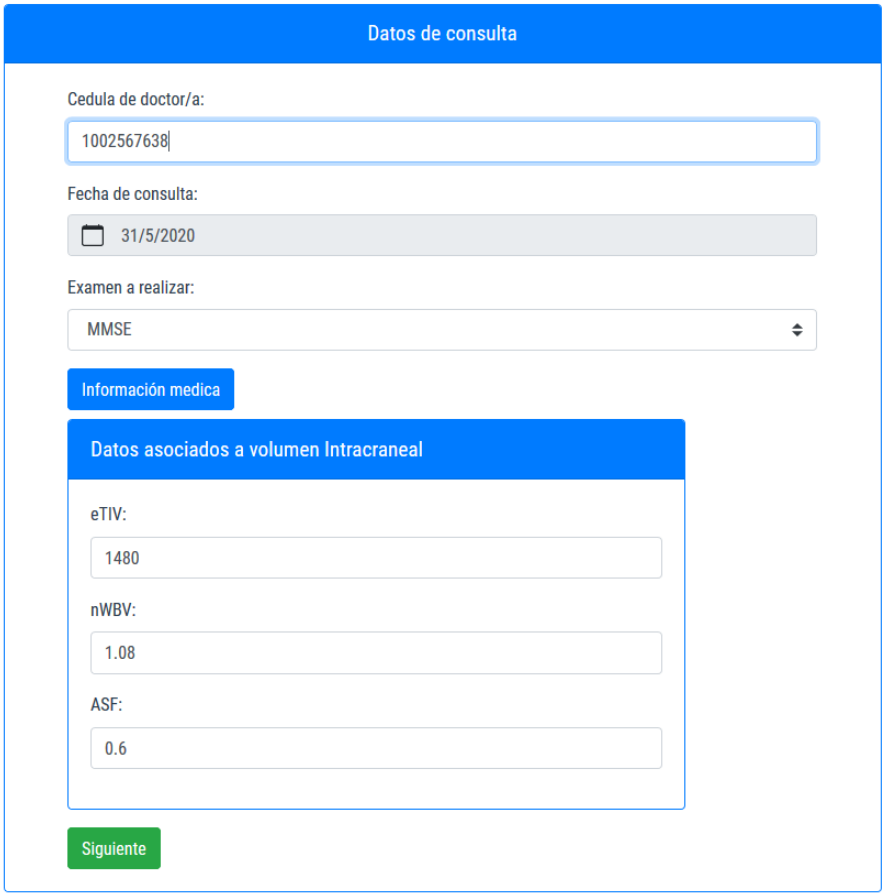


Figura 18. Pagina para ingresar datos del volumen intracraneal

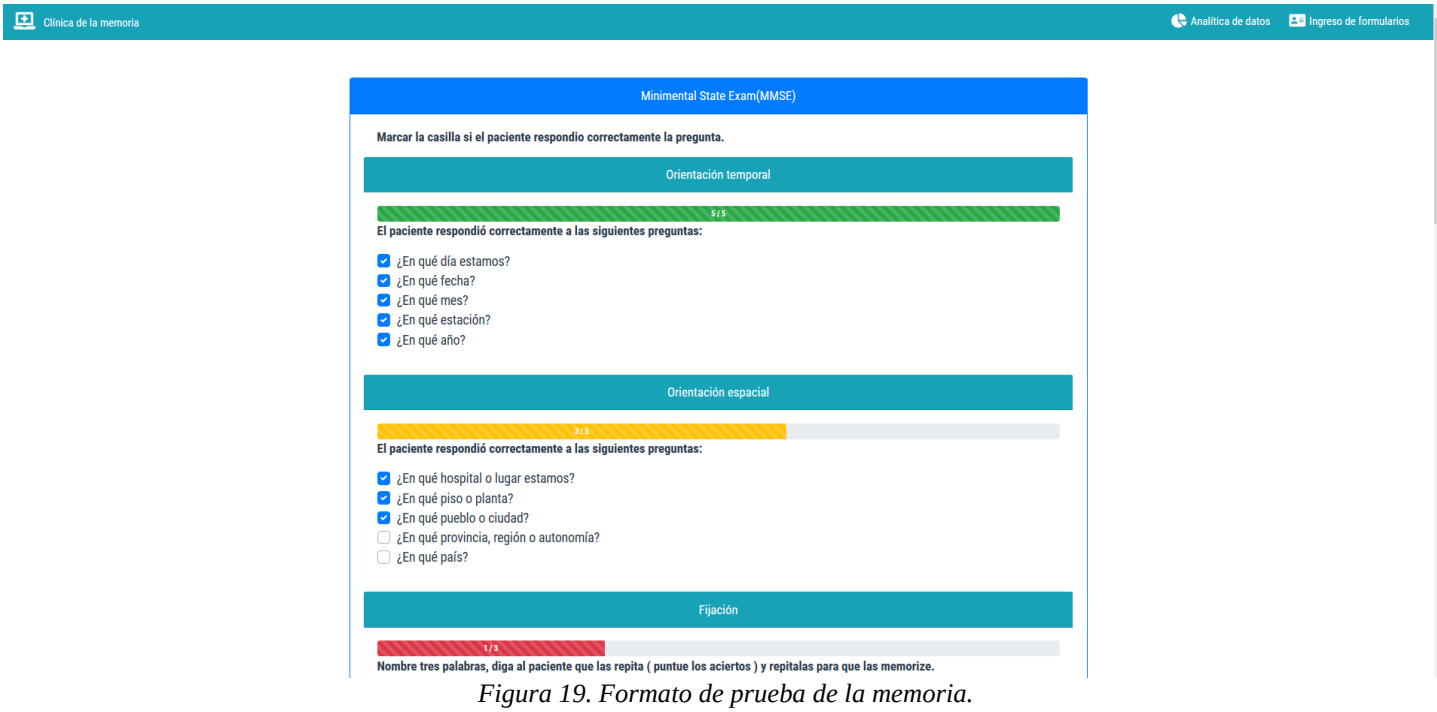


Figura 19. Formato de prueba de la memoria.

En la figura 20 se pueden observar los resultados del paciente, en esta página aparecen los resultados de la prueba de la memoria realizada, de igual forma se muestra si el paciente es diagnosticado o no con demencia de acuerdo al grupo al que pertenece y al CDR junto con su respectiva probabilidad, adicionalmente se presentan los factores de riesgos que posee el sujeto junto con el porcentaje de importancia de cada factor.

En la figura 21 se muestra el monitoreo de los pacientes, como ha sido la evolución del mismo a lo largo de las consultas realizadas, los datos de la consulta tales como fecha, hora, clasificación por grupo y CDR, entre otros. De este modo se le facilita al doctor el tratamiento y seguimiento del paciente, mostrando gráficas que reflejan el desempeño del paciente en las pruebas realizadas.

En la figura 22 se muestra la gestión de pacientes, como se pueden realizar búsqueda dentro de la base de datos haciendo uso de diferentes filtros tales como la edad, el municipio y el sexo de los pacientes. en la figura 23 se muestra una dashboard con la distribución de los pacientes de este modo la persona encargada del tratamiento de los pacientes puede tener claridad del comportamiento de la población de prueba generando claridad en qué grupos debe estar más atento de acuerdo a lo que muestran los datos. Por último en la figura 23 se muestra que los resultados obtenidos en las pruebas pueden ser descargados por el doctor encargado y permitiéndole a este escoger qué datos desea descargar.

Figura 20. Resultados del paciente

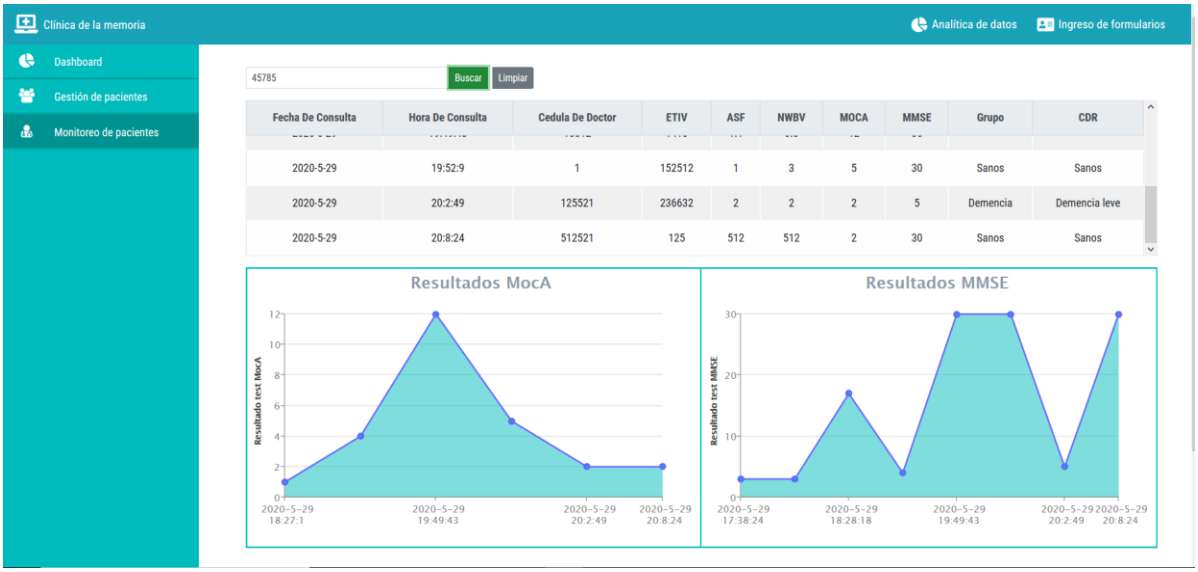


Figura 21. Monitoreo de paciente

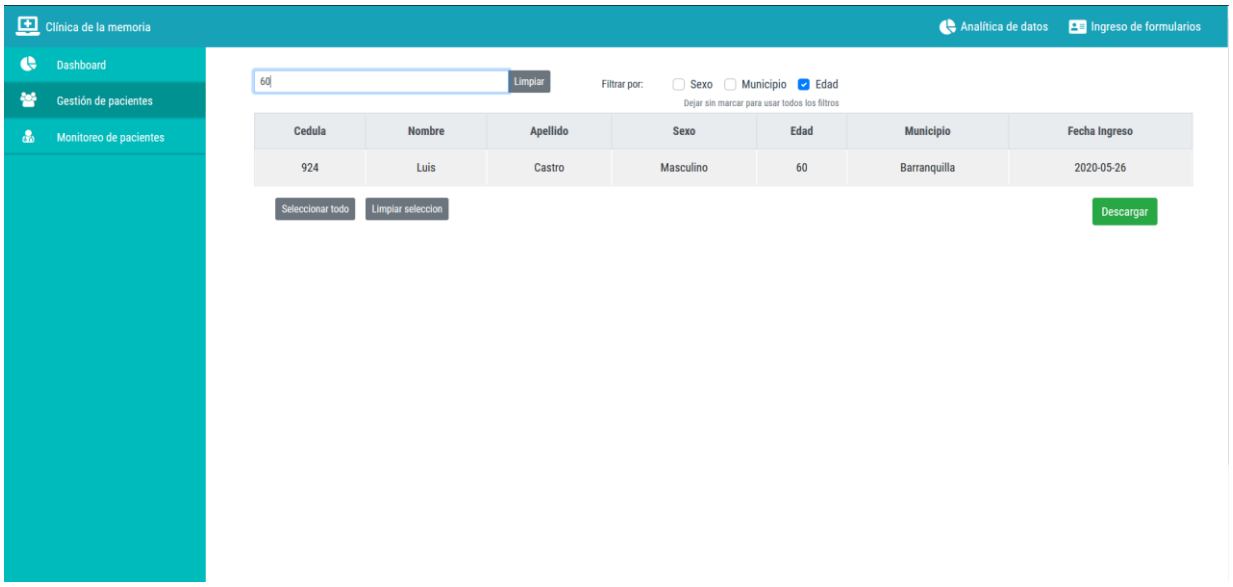


Figura 22. Gestión de pacientes

Figura 23. Dashboard con Distribución de los pacientes.

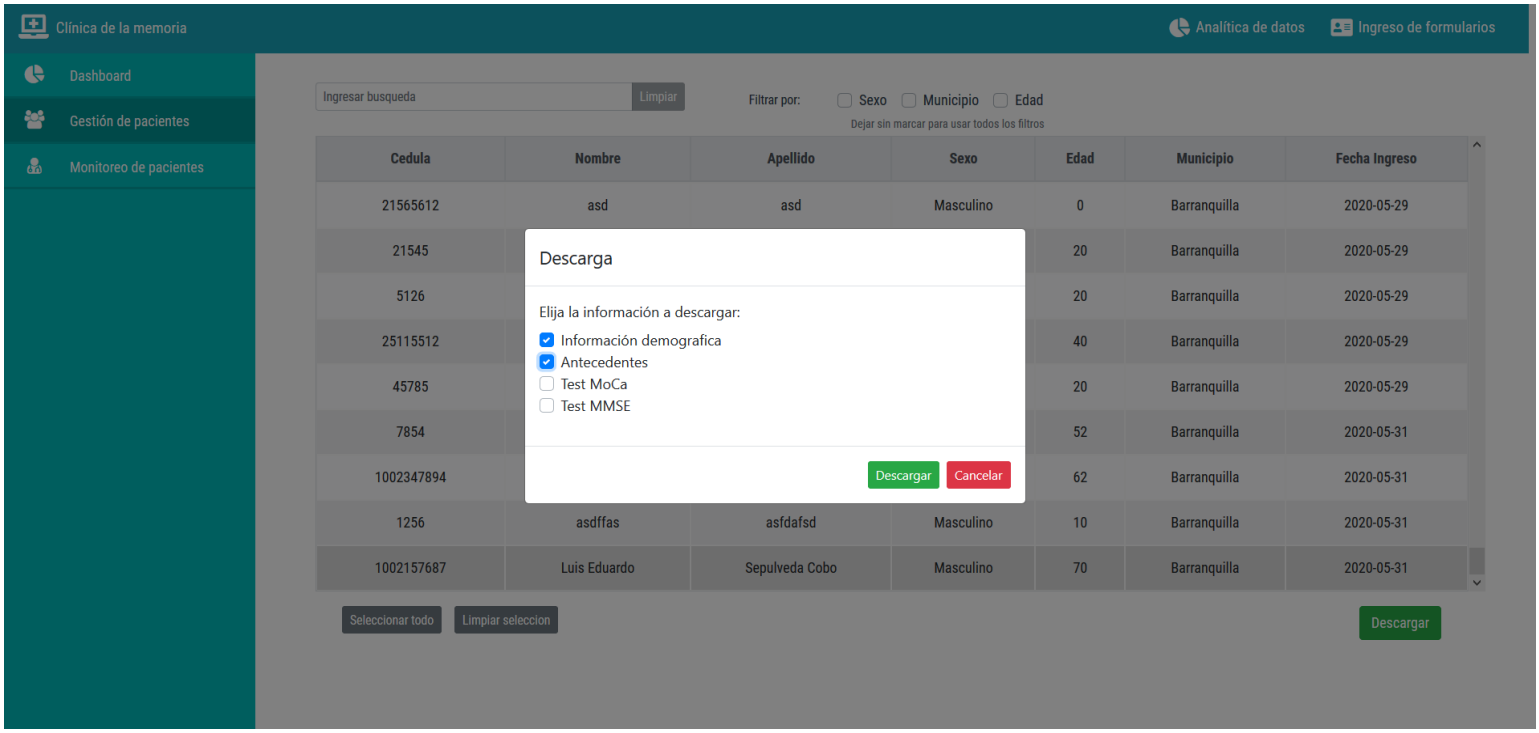


Figura 24. Descarga de los Datos

G. ANÁLISIS Y RESULTADOS

A. *Resultados del árbol de decisión basados en el grupo*

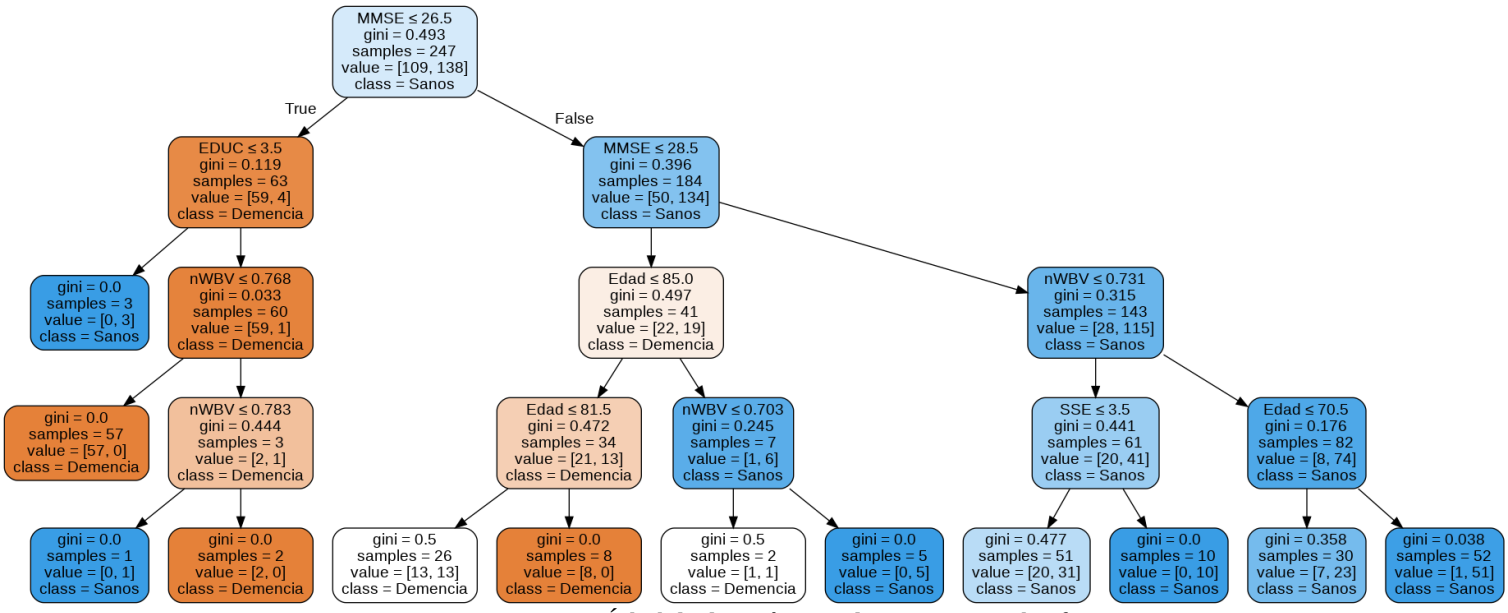


Figura 25. Árbol de decisión usando grupo para clasificar.

El árbol de decisión obtenido al clasificar a los pacientes de acuerdo al grupo que pertenecen, demente o sano, arroja un nivel de accuracy del 67%, al ser un valor menor del 70% consideramos que las clasificaciones hechas por este modelo no son de alta confiabilidad.

Luego de crear el árbol de decisión procedemos a realizar la validación del mismo, para esto se calcula el ROC del modelo como se puede observar en la Figura 26, el área bajo la curva de este modelo arrojó un nivel de accuracy del 74%.

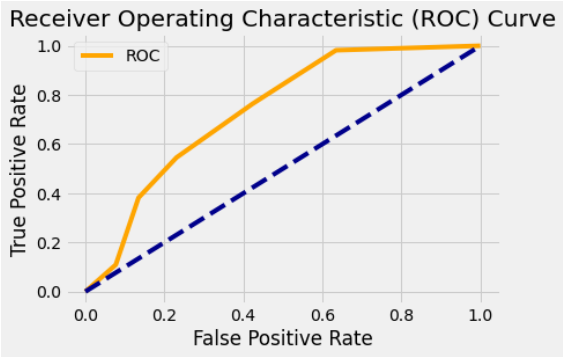


Figura 26. Gráfica de ROC Curve árbol de decisión usando grupo para clasificar.

Posteriormente procedemos a calcular la matriz de confusión de este modelo, como se puede observar en la Figura 27. Al calcular la precisión promedio del modelo está arrojó un 64%, se calculó que la recuperación de precisión fue de 78% y el F1 fue de 72%.

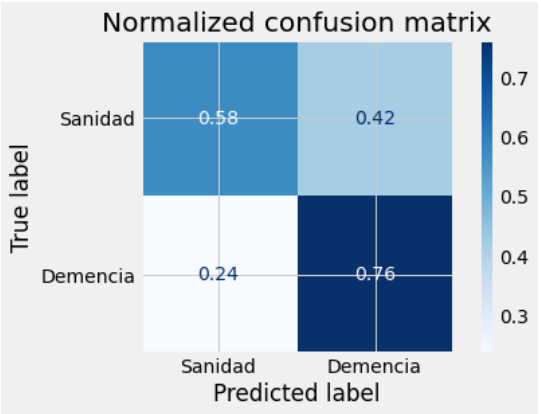


Figura 27. Matriz de confusión del árbol de decisión usando grupo para clasificar.

Este modelo arrojó 31 verdaderos positivos, el 21 falsos positivos, 12 falsos negativos y 43 verdaderos negativos, es decir que este modelo clasifica correctamente 31 pacientes con demencia y 43 sanos.

B. Resultados del Random Forest basados en el grupo

Al construir el modelo de Random forest a partir de clasificación por grupos los resultados obtenidos fueron un nivel de accuracy del 83%, con un número promedio de nodos de 84 y un promedio de profundidad máxima de 10.

Posteriormente se valida el modelo calculando el ROC, como se puede observar en la Figura 28, el AUC de este arrojó un nivel de accuracy del 93%

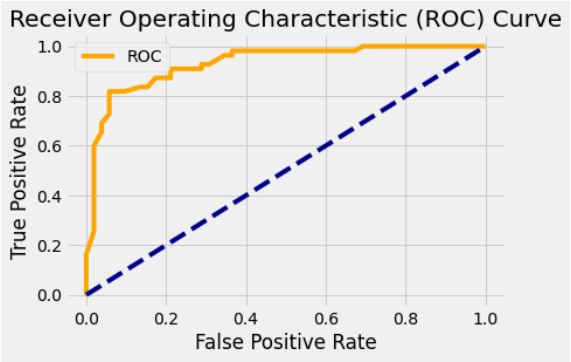


Figura 28. Gráfica de ROC Curve Random Forest usando grupo para clasificar.

Una vez calculado el AUC, se procede a realizar la matriz de confusión, la cual se puede ver en la Figura 29. Al calcular la precisión promedio del modelo está arrojó un 77%, se calculó que la recuperación de precisión fue de 91% y el F1 fue de 85%.

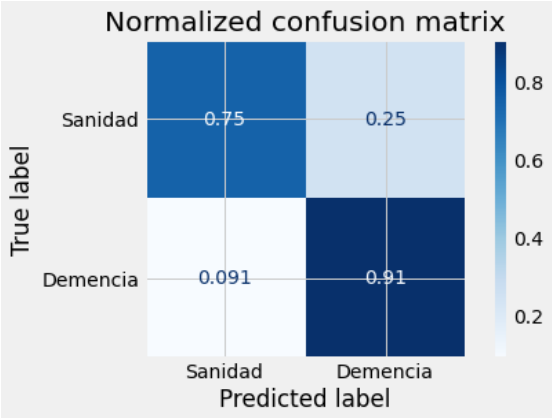
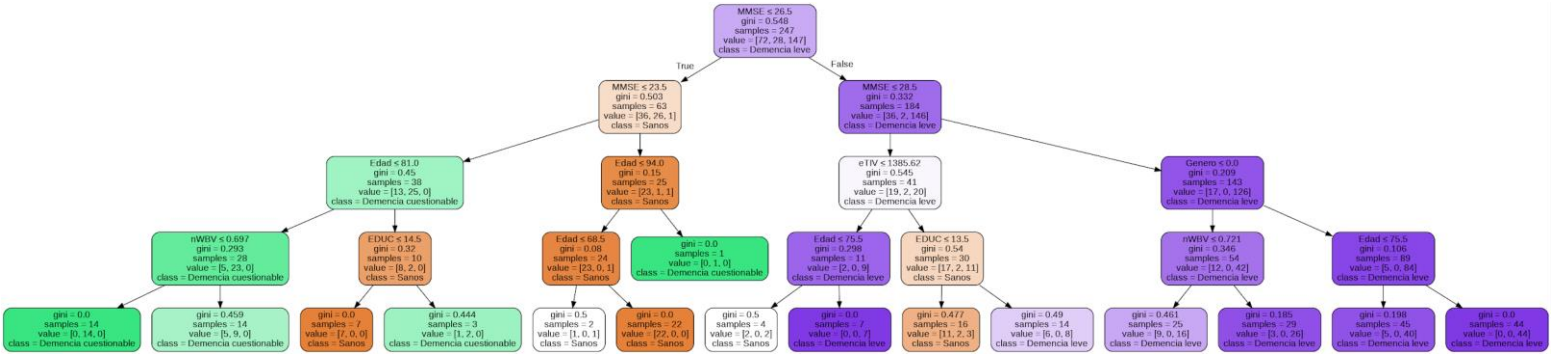


Figura 29. Matriz de confusión Random Forest usando grupo para clasificar.

Este modelo arrojó 39 verdaderos positivos, el 13 falsos positivos, 5 falsos negativos y 50 verdaderos negativos, es decir que este modelo clasifica correctamente 39 pacientes con demencia y 50 sanos.

C. Resultados del árbol de decisión basados en el CDR



El árbol de decisión obtenido al clasificar a los pacientes de acuerdo al resultado del CDR arroja un nivel de accuracy del 64%, Luego de crear el árbol de decisión procedemos a realizar la validación del mismo, para esto se calcula el AUC de este modelo el cual arrojó un nivel de accuracy del 74%.

Posteriormente procedemos a calcular la matriz de confusión de este modelo, como se puede observar en la Figura 31, Se calculó que la precisión promedio del modelo es del 65%, la recuperación de precisión fue de 65% y el F1 fue de 65%.

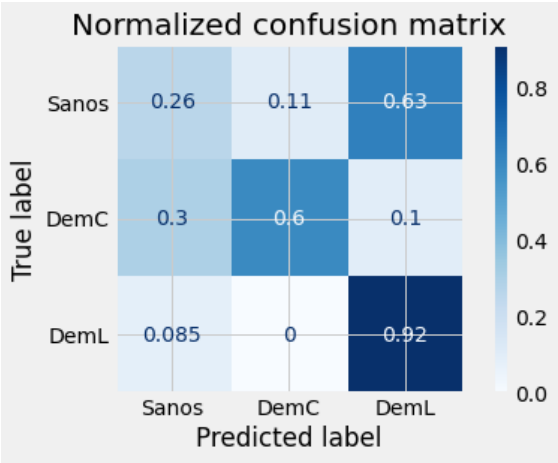


Figura 31. Matriz de confusión del árbol de decisión usando CDR para clasificar

Este modelo clasificó correctamente a 10 personas sanas, 6 personas con demencia cuestionable y a 54 personas con demencia leve.

D. Resultados del Random forest basados en el CDR

Al construir el modelo de Random forest a partir de clasificación por CDR los resultados obtenidos fueron:

a. Un nivel de accuracy del 72%
b. Un número promedio de nodos de 98
c. Un promedio de profundidad máxima de 11.
d. Posteriormente se valida el modelo calculando el AUC que arrojó un nivel de accuracy del 79%

Al calcular la matriz de confusión de este modelo, como se puede observar en la Figura 32, Se calculó que la precisión promedio del modelo es del 73%, la recuperación de precisión fue de 73% y el F1 fue de 73%.

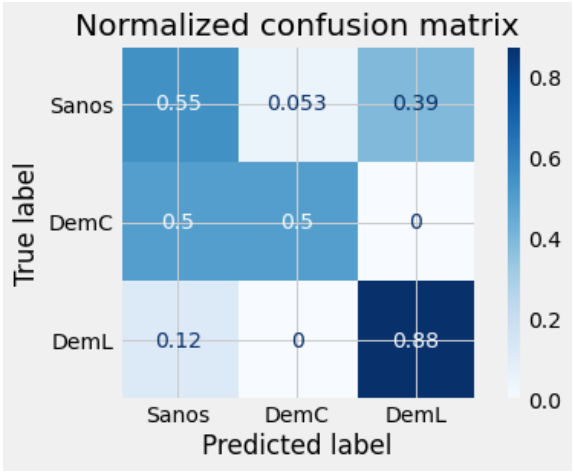


Figura 32. Matriz de confusión Random Forest usando CDR para clasificar.

Este modelo clasificó correctamente a 21 personas sanas, 6 personas con demencia cuestionable y a 5 personas con demencia leve 52.

E. Comparación de los modelos

En primer lugar se compara el rendimiento de cada modelo al clasificar a los pacientes de acuerdo al grupo al que pertenecen:

Modelo	**Árbol de Decisión**	**Random Forest**
Accuracy	0.6915887850467289	0.8317757009345794
AUC	0.7622377622377623	0.9305944055944055

Precisión	0.6374336236193713	0.7682296934633382
Recall	0.7818181818181819	0.9090909090909091
F1 score	0.7226890756302521	0.847457627118644

Tabla 2. Árbol de decisión vs. random forest usando Grupo para clasificar.

Como se puede observar en la Tabla 2, en todos los campos en los que se comparó el rendimientos entre los modelos dentro de cada campo de validación, el modelo de random forest presentó de manera constante un mejor desempeño. Es por ello que se decidió comprobar dicha información con respecto a la clasificación basada en el CDR, como se ilustra en la Tabla 3.

Modelo	Árbol de Decisión	Random Forest
Accuracy	0.6542056074766355	0.7476635514018691
AUC	0.7406542056074765	0.8107476635514019
Precisión	0.6542056074766355	0.7476635514018691
Recall	0.6542056074766355	0.7476635514018691
F1 score	0.6542056074766355	0.7476635514018691

Tabla 3. Árbol de decisión vs. random forest usando CDR para clasificar.

En la tabla 3 se puede observar nuevamente que el random forest presenta constantemente mejores resultados que el árbol de decisión a pesar de que para la clasificación con CDR los valores resultantes son menores, El rendimiento del random forest fue durante ambos casos de prueba mayor al 70% para todos los campos por lo tanto se puede confirmar la confiabilidad de las clasificaciones resultantes del random forest y es por ello que para la implementación el equipo de trabajo decidió utilizar este método para mostrarle los resultados de los pacientes y permitir un seguimiento óptimo de los mismos.

H. CONCLUSIONES.

Para concluir nuestro proyecto se puede afirmar que se cumplieron todos los objetivos planteados inicialmente para el desarrollo de la plataforma.

En primer lugar se realizó la revisión sistemática de la literatura y del estado del arte en relación a herramientas de apoyo en el control de pacientes para conocer más a fondo aquello que podríamos agregar de valor en nuestra plataforma, que en este caso fue el factor de ayuda en la toma de decisiones de los doctores apoyándonos en técnicas de minería de datos. Agregado a esto realizando esta investigación pudimos determinar cuáles técnicas eran las más utilizadas en el campo médico para la clasificación de pacientes no únicamente con enfermedades relacionadas a la demencia sino a cualquier tipo de enfermedad, dentro de las cuales se encontraban los árboles de decisión pero que finalmente por cuestiones de análisis de las métricas arrojadas por el modelo terminó siendo utilizado Random Forest. Igualmente pudimos encontrar que las métricas asociadas a la matriz de confusión, curva roc y AUC son las más importantes a la hora de realizar una clasificación ya que permiten determinar el número de falsos positivos, verdaderos positivos, falsos negativos y verdaderos negativos que en este caso se ajustaban a pacientes con y sin demencia. Cabe recalcar también que dentro de la investigación para encontrar un datos que se ajustarán al proyecto, el dataset del proyecto Oasis fue el mejor encontrado de libre uso y que en la etapa de análisis y exploración de los datos arrojó los mejores resultados de porcentaje de varianza lo que aseguro una gran confiabilidad.

Teniendo en cuenta lo anterior se procedió entonces a crear la plataforma que se ajustara a las necesidades discutidas entre el grupo de trabajo y los neurólogos del hospital Universidad del Norte. Se crea entonces la plataforma web en la que se pueden tomar los datos de los pacientes que posteriormente se analizan para establecer la clasificación y los factores de riesgo de cada uno. Igualmente se crea el modulo de analitica de datos para la visualización de estos y las secciones de monitoreo y gestión de su información por parte de los doctores.

Se realizó una evaluación entre pares para determinar la calidad del desarrollo de esta plataforma. Los resultados de esta evaluación se pueden apreciar en la [Figura 32]

Figura 32. Evaluación

Nuestro modelo se compone de dos clasificaciones, una para el grupo al que pertenece el paciente y otra para el Clinical Dementia Rating (CDR) el cual es validado utilizando una validación cruzada 70-30, donde para el accuracy se arroja un resultado del 83% para el grupo y de 72% para el CDR. Pese a no ser resultados tan exactos, como expusimos anteriormente tuvimos más que nada en cuenta los resultados arrojados por el Área bajo la curva ROC (AUC) ya que esta me indica que tan probable es que un modelo clasifique bien entre una cosa u otra. Para el grupo el AUC dio un valor del 92% y para el CDR el AUC dio un valor del 81% lo cual establece una probabilidad razonable para la clasificación de los pacientes.

Finalmente dentro de los factores de riesgo asociados se tuvo en cuenta el nivel de importancia de los atributos de nuestro modelo, ya que estos establecieron las reglas de decisión de cada uno de los árboles del Random forest. donde se obtuvo lo siguiente:

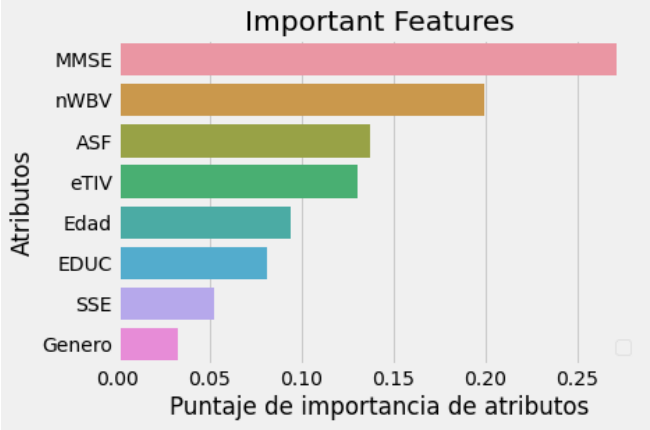


Figura 33. Importancia de los atributos

Donde se concluyo que efectivamente el factor más importante a la hora de clasificar a un paciente con demencia o no es el resultado del test que realiza, en el cual se evalúa la capacidad de memoria, la orientación espacial, la escritura y el lenguaje del individuo lo cual lógicamente tiene un alto peso a la hora de que el modelo tomará la decisión. Posterior a esto se tienen como importantes los atributos asociados a las resonancias magnéticas aplicadas a los pacientes, donde se tiene en cuenta que entre menor sea el volumen intracraneal del paciente más posibilidades tiene de padecer alguna enfermedad relacionada a la demencia. Luego por nivel de importancia se tiene la edad, el nivel de educación, el estrato socioeconómico y por último el género.

Para definir entonces qué valor obtenido por el paciente se puede considerar de riesgo se realizó la revisión del dataset y se dividieron los datos respecto a la media.

Este proyecto permite entonces hacer un énfasis en cómo las técnicas de minería de datos pueden resultar beneficiosas en áreas como la medicina y como pueden resultar de apoyo a los doctores para hacer inferencias respecto a el estados de sus pacientes. Pese a no obtener valores de exactitud tan altos se considera que la herramienta cumple su cometido y que posteriormente mediante esta se puede llegar a generar una base de datos robusta y con mayor cantidad de datos a analizar para poder mejorar en cada una de las métricas analizadas y que así el modelo sea perfeccionado cada vez más.

I. Referencias

1. Fulton et al.(2019). Classification of Alzheimer's Disease with and without Imagery using Gradient Boosted Machines and ResNet-50. Brain Sciences, 9(9), 212. doi: 10.3390/brainsci9090212
2. Organización mundial de la salud. (2019). Demencia. Retrieved from https://www.who.int/es/news-room/fact-sheets/detail/dementia
3. Alzheimer's Association. (2019). Facts and Figures. Retrieved 11 March 2020, Retrieved 11 March 2020, from https://www.alz.org/alzheimers-dementia/facts-figures?lang=en-US
4. Alzheimer's Disease International. (2015). Dementia statistics | Alzheimer's Disease International. Retrieved 11 March 2020, from https://www.alz.co.uk/research/statistics
5. OISS. (2020). Boletín programa Iberoamericano de cooperación sobre la situación de las personas adultas mayores. Retrieved 11 March 2020, from https://oiss.org/wp-content/uploads/2020/01/Boletin-OISS-19-Alta.pdf
6. MinSalud. (2017). Boletín de Salud Mental No 3. from https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ENT/Boletin-demencia-salud-mental.pdf
7. Ertek, G., Tokdil, B., & Günayddn, b. (2018). Supplement for 'Risk Factors and Identifiers for Alzheimer's Disease: A Data Mining Analysis'. SSRN Electronic Journal. doi: 10.2139/ssrn.3151516
8. Mani, S., Dick, M., Pazzani, M., Teng, E., Kempler, D., & Taussig, I. (1999). Refinement of Neuro-psychological Tests for Dementia Screening in a Cross Cultural Population Using Machine Learning. Artificial Intelligence In Medicine, 326-335. doi: 10.1007/3-540-48720-4_35
9. Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of healthcare information management, 19(2), 65.
10. Abdullah, S., et al (2019). Assessing the predictive ability of the UPDRS for falls classification in early stage Parkinson's disease. Cornell university.
11. Alashwal, H., El Halaby, M., Crouse, J., Abdalla, A., & Moustafa, A. (2019). The Application of Unsupervised Clustering Methods to Alzheimer's Disease. Frontiers In Computational Neuroscience, 13. doi: 10.3389/fncom.2019.00031
12. Farooqui, N., & Mehra, R. (2018). Design of A Data Warehouse for Medical Information System Using Data Mining Techniques. 2018 Fifth International Conference On Parallel, Distributed And Grid Computing (PDGC). doi: 10.1109/pdgc.2018.8745864
13. Delshi Howsalya Devi, R., & Deepika, P. (2015). Performance comparison of various clustering techniques for diagnosis of breast cancer. 2015 IEEE International Conference On Computational Intelligence And Computing Research (ICCIC). doi: 10.1109/iccic.2015.7435711
14. Mahalakshmi, M., & Sundararajan, M. (2015). Tracking the student's performance in Web-based education using Scrum methodology. 2015 International Conference On Computing And Communications Technologies (ICCCT). doi: 10.1109/iccct2.2015.7292779
15. Xuesong Zhang, & Dorn, B. (2011). Agile Practices in a Small-Scale, Time-Intensive Web Development Project. 2011 Eighth International Conference On Information Technology: New Generations. doi: 10.1109/itng.2011.187
16. Bin-Hezam, R., & E., T. (2019). A Machine Learning Approach towards Detecting Dementia based on its Modifiable Risk Factors. *International Journal Of Advanced Computer Science And Applications*, *10*(8). doi: 10.14569/ijacsa.2019.0100820
17. Olegas NIAKŠU (2015). CRISP Data Mining Methodology Extension for Medical Domain. *Baltic J. Modern Computing, Vol. 3*
18. Latifah, E., Abdullah, S., & Soemartojo, S. (2018). Identifying of factor associated with parkinson's disease subtypes using random forest. Journal Of Physics: Conference Series, 1108, 012064. doi: 10.1088/1742-6596/1108/1/012064