



Universidad de Valladolid



**ESCUELA DE INGENIERÍAS
INDUSTRIALES**

MÁSTER EN INVESTIGACIÓN EN
INGENIERÍA DE PROCESOS Y SISTEMAS
CURSO 2011-2012

ESTUDIO DE TÉCNICAS DE
INTELIGENCIA COMPUTACIONAL
PARA LA CARACTERIZACIÓN DE
PACIENTES CON ENFERMEDAD DE
ALZHEIMER

- Trabajo de Fin de Máster -

Alumno: Francisco J. Barrientos García

Tutor: Gregorio I. Sainz Palmero



Estudio de Técnicas de Inteligencia Computacional para la Caracterización de Pacientes con Enfermedad de Alzheimer

El presente Trabajo de Fin de Máster, titulado “*Estudio de Técnicas de Inteligencia Computacional para la Caracterización de Pacientes con Enfermedad de Alzheimer*” realizado por D. Francisco J. Barrientos García, ha sido desarrollado con la colaboración del **Centro de Referencia Estatal de Atención a Personas con Enfermedad de Alzheimer y otras Demencias** (CRE-Alzheimer) del IMSERSO, que ha proporcionado las bases de datos sobre los casos y diagnósticos de la enfermedad de Alzheimer, así como el conocimiento experto utilizado para la realización del trabajo.

Resumen

La Demencia tiene un impacto creciente en la Sociedad debido al aumento de la esperanza de vida. De los distintos tipos de Demencia, la Enfermedad de Alzheimer y la Demencia Vasculosa suponen más del 90 % de los casos. Un diagnóstico temprano y acertado sobre el tipo de demencia es importante para proporcionar al paciente el tratamiento adecuado. Este trabajo se divide en dos partes. La primera profundiza en el conocimiento de las técnicas de Inteligencia Computacional y su aplicación en el campo de la demencia. Como resultado, se ha obtenido un breve estado del arte de las técnicas de IC que se están aplicando actualmente y para qué fines. La segunda parte trata con un problema real: el diagnóstico diferencial entre Demencia Vasculosa y Demencia tipo Alzheimer. Para ello, se ha realizado una selección de características y se ha confrontado con las variables usadas habitualmente por los expertos en el dominio. La conclusión es que, aunque el alcance del estudio es limitado debido al reducido número de datos disponibles, ha servido para comprobar el estado actual de la investigación en este campo y testar la utilidad de la metodología como herramienta de apoyo a la toma de decisiones en el ámbito del diagnóstico diferencial entre DV y EA.

Abstract

Dementia has a growing impact on Society due to increased life expectancy. Among different types of dementia, Alzheimer's Disease (AD) and Vascular Dementia (VD) account for over 90 % of cases. Early and accurate diagnosis is crucial to provide patient with the right treatment. This work is divided into two parts. The first explores Computational Intelligence (CI) techniques and their applicability in the field of dementia. The result obtained is a brief state of the art about IC techniques currently being applied and for what purposes. The second part deals with a real problem: differential diagnosis of vascular dementia and Alzheimer type dementia. Thus, we have made a feature selection confronting it with the variables commonly used by domain experts. Conclusion is, although the scope of the study is limited because of the size of the dataset, it has been used to check the current state of the art in this field and to test the usefulness of the methodology as decision making support tool for differential diagnosis between DV and EA.

Índice general

1. Introducción y objetivos	11
1.1. Objetivos	12
1.2. Organización de la memoria	13
I Breve estado del arte: técnicas de Inteligencia Computacional aplicadas a la Demencia	15
2. Relevancia de la Demencia en la Sociedad	17
2.1. Qué es la Demencia	17
2.2. Enfermedad de Alzheimer	19
2.2.1. Epidemiología	19
2.2.2. Factores de riesgo	21
2.2.3. Diagnóstico	22
2.3. Demencia vascular	23
3. Inteligencia Computacional	25
3.1. Sistemas de clasificación	25
3.2. Selección de Características	26
3.3. Técnicas de Inteligencia Computacional	27
3.3.1. Árboles de Decisión	27
3.3.2. Redes Neuronales	29
3.3.3. Máquinas de Vectores Soporte (SVM)	30
3.3.4. Clustering	32
3.3.5. ROCK	34
4. Breve Estado del Arte	35
4.1. Búsqueda bibliográfica	35
4.2. Análisis del estado del arte	36
II Inteligencia Computacional aplicada al diagnóstico diferencial entre Demencia Vascular y Demencia tipo Alzheimer	43
5. Caso de aplicación	45
5.1. Revisión de diagnósticos	45
5.2. Características de los casos diagnosticados	46

ÍNDICE GENERAL

6. Metodología	49
6.1. Pre-procesamiento de datos	49
6.1.1. Adquisición de datos	50
6.1.2. Inducción de particiones	50
6.2. Selección de características	50
6.2.1. Selección experta	50
6.2.2. Recuento Borda	50
6.3. Métodos empleados para la experimentación	51
6.4. Evaluación de resultados	52
7. Resultados	53
7.1. Selección de Características	53
7.1.1. Selección de variables experta	53
7.1.2. Selección de variables mediante recuento Borda	55
7.2. Inducción de particiones	56
7.3. Algoritmos de Inteligencia Computacional	57
7.3.1. Árboles de Decisión	57
7.3.2. Redes Neuronales Artificiales ANN	61
7.3.3. SVM	63
7.3.4. Clustering	64
7.4. Comparación de Resultados	70
8. Conclusiones	73
8.1. Futuras líneas de investigación	74
Bibliografía	75
III Apéndices	81
A. Tablas	83
B. Inducción de particiones	89
C. Resultados completos	93
C.1. Árboles de Decisión	93
C.1.1. J4.8	93
C.1.2. Random Forest	100
C.2. Redes Neuronales Artificiales ANN	101
C.2.1. MLP	101
C.2.2. RBF	103
C.3. SVM	107
C.4. Clustering	109
C.4.1. KMeans	109
C.4.2. EM	111
C.4.3. Hierarchical Clusterer	113
C.4.4. ROCK	115

Índice de figuras

2.1. El coste social del Alzheimer en el mundo	18
3.1. Ejemplo de sobreajuste.	28
3.2. Ejemplo del algoritmo k-means estándar.	33
4.1. Número de publicaciones y citas relacionadas con IC y demencia	37
4.2. Taxonomía de la IC aplicada a la Demencia	37
4.3. Número de referencias relacionadas con IC y demencia	41
4.4. Número de citas relacionadas con IC y demencia	42
5.1. Diagnósticos.	46
5.2. Distribución por grupos de edad.	47
5.3. Estado civil y situación familiar de los usuarios.	47
5.4. Nivel educativo.	48
5.5. Profesión.	48
6.1. Metodología.	49
7.1. Representación del recuento Borda	55
B.1. Inducción de particiones para la variable “Hachinski”.	89
B.2. Inducción de particiones para la variable “Año_Inicio”.	90
B.3. Inducción de particiones para la variable “H_Meses_Evolución”.	91
B.4. Inducción de particiones para la variable “Años_Escol”.	92

ÍNDICE DE FIGURAS

Índice de tablas

2.1. Comparación entre Demencia Vasculuar y Alzheimer	24
3.1. Comprativa entre ANN y SVM	31
7.1. Ranking de características	55
7.2. Parámetros de HFP	56
7.3. Inducción de particiones de las variables de entrada	56
7.4. Evaluación de resultados: J4.8 (Experto).	58
7.5. Matriz de confusión: J4.8 (Experto).	59
7.6. Evaluación de resultados: J4.8 (Borda).	59
7.7. Matriz de confusión: J4.8 (Borda).	60
7.8. Evaluación de resultados: Random Forest (Experto).	60
7.9. Matriz de confusión: Random Forest (Experto).	60
7.10. Evaluación de resultados: Random Forest (Borda).	61
7.11. Matriz de confusión: Random Forest (Borda).	61
7.12. Evaluación de resultados: MLP (Experto).	61
7.13. Matriz de confusión: MLP (Experto).	62
7.14. Evaluación de resultados: MLP (Borda).	62
7.15. Matriz de confusión: MLP (Borda).	62
7.16. Evaluación de resultados: RBF (Experto).	63
7.17. Matriz de confusión: RFB (Experto).	63
7.18. Evaluación de resultados: RBF (Borda).	63
7.19. Matriz de confusión: RBF (Borda).	64
7.20. Evaluación de resultados: SVM (Experto).	64
7.21. Matriz de confusión: SVM (Experto).	65
7.22. Evaluación de resultados: SVM (Borda).	65
7.23. Matriz de confusión: SVM (Borda).	66
7.24. Matriz de emparejamiento: KMeans (Experto).	66
7.25. Matriz de emparejamiento: KMeans (Borda).	66
7.26. Matriz de emparejamiento: EM (Experto).	67
7.27. Matriz de emparejamiento: EM (Borda).	68
7.28. Matriz de emparejamiento: Hierarchical Clusterer (Experto).	68
7.29. Matriz de emparejamiento: Hierarchical Clusterer (Borda).	69
7.30. Matriz de emparejamiento: ROCK (Experto).	69
7.31. Matriz de emparejamiento: ROCK (Borda).	69
7.32. Comparación de los resultados obtenidos.	71
A.1. Descripción de las variables.	83

ÍNDICE DE TABLAS

A.2. Ranking de características	86
---	----

Capítulo 1

Introducción y objetivos

El objetivo principal de este Trabajo de Fin de Máster es el estudio de las diversas técnicas de Inteligencia Computacional (IC) [54] para su aplicación en la caracterización de pacientes con enfermedad de Alzheimer. Para ello, es necesario realizar una etapa previa de recogida y reconciliación de datos de pacientes con Alzheimer. El estudio tratará de determinar qué técnica de inteligencia computacional es la más adecuada para el modelado de los pacientes, en función de las características cualitativas y cuantitativas de los datos disponibles.

La Enfermedad de Alzheimer (EA) es una enfermedad degenerativa y terminal del cerebro que provoca la pérdida debilitante de la función cognitiva y finalmente la muerte [49]. La enfermedad fue descrita por primera vez por Alois Alzheimer en 1906 y se trata de la forma más común de demencia. Demencia es un término general para la pérdida de habilidades intelectuales y sociales suficientemente graves que interfieren con las actividades diarias [4]. Debido principalmente al envejecimiento demográfico y a la incidencia de la enfermedad (en 2006 se contabilizaron más de 26 millones de enfermos en todo el mundo y se estima que este número se cuadruplicará para el año 2050), la investigación sobre la EA se ha convertido en una prioridad importante en los últimos años [11].

Existen dos formas conocidas de manifestación de la EA: la forma temprana y la avanzada. Estas dos formas se diferencian por la edad de los individuos cuando comienzan a padecer la enfermedad. En la forma avanzada, la más frecuente, los síntomas de la enfermedad aparecen en personas mayores, normalmente pasados los sesenta años y se produce un empeoramiento cognitivo gradual hasta que el paciente no puede vivir de forma independiente y tiene que apoyarse por completo en los cuidados proporcionados por los familiares y cuidadores. La forma temprana es una rara manifestación de la EA que afecta sólo al 5% de las personas que padecen la enfermedad y se desarrolla en personas con edades comprendidas entre los 30 y los 60 años [49].

Se han descrito diferentes métodos para realizar un diagnóstico y determinar el estado en el que se encuentra el paciente durante la enfermedad. Generalmente, los estadios son descritos como leves o tempranos, moderados o medios y severos o avanzados [49]. Al mismo tiempo, se han descrito diferentes escalas que se utilizan para describir la progresión de la enfermedad. La escala más utilizada es la Escala de Deterioro Global (GDS) desarrollada por el doctor Barry Reisberg en 1982 [60]. Concretamente, la escala GDS divide la progresión

de la enfermedad en siete estadios basándose en la progresión observada de los síntomas y en la progresiva incapacidad de realizar ciertas actividades por parte del paciente.

Las investigaciones que se están realizando en torno a la EA están aportando luz sobre los mecanismos que afectan a la enfermedad. Sin embargo, la causa exacta que la desencadena está todavía sin resolver [72]. De hecho, aunque se han detectado varias mutaciones genéticas en pacientes afectados, éstas no explican todos los casos y diversos investigadores piensan que una combinación de factores genéticos junto con factores ambientales incrementan el riesgo de padecer la EA [49].

Debido a la gran cantidad de incógnitas que todavía no se han resuelto, no se ha podido desarrollar un tratamiento efectivo que elimine la enfermedad. De momento, la prevención en personas sanas y la estimulación cognitiva conjuntamente con tratamientos farmacológicos en personas enfermas son los únicos medios que permiten mantener y/o retrasar el curso de la enfermedad.

1.1. Objetivos

Como se ha visto, en la actualidad no hay cura para el Alzheimer, por eso el diagnóstico precoz es de vital importancia en la demencia. En esto no se diferencia de otras muchas enfermedades. Pero en este caso es un aspecto crítico puesto que los tratamientos, tanto farmacológicos como no farmacológicos que se puedan asignar al paciente, son mucho más efectivos cuanto menor es el deterioro cognitivo, i.e. en los estadios iniciales de la enfermedad. Tan importante como el diagnóstico temprano, es que el diagnóstico sea acertado, ya que en caso contrario los tratamientos no tendrán el efecto esperado, llegando incluso a ser contraproducentes según los casos. Con la demencia, el tipo más extendido es la demencia tipo Alzheimer, pero hay otros tipos de demencias y no deben ser confundidos.

La demencia, al tratarse de una enfermedad crónica, tiene un coste socio-sanitario asociado muy alto, debido a que su tratamiento médico y asistencial puede extenderse a lo largo de varios años. A este coste económico hay que añadir el coste humano, tanto por el sufrimiento del propio paciente, como por el de sus familiares y/o cuidadores. El diagnóstico precoz y preciso, reduce esta carga asociada a la demencia, en sus dos vertientes. Por un lado, al mejorar la eficacia de los tratamientos se retrasa la evolución de la enfermedad y se reduce la dependencia del enfermo de cuidadores. Por otro lado, al mantenerse el paciente activo más tiempo, mejora su calidad de vida y la de sus familiares.

Con estas premisas, los objetivos de este trabajo tienen una doble finalidad:

- Profundizar en el conocimiento de las técnicas de Inteligencia Computacional y su aplicación en el campo de la demencia.
 - Establecer la importancia de la Demencia para la Sociedad, desde diversos puntos de vista: médico, económico, etc.
 - Realizar una breve revisión de las principales técnicas de Inteligencia Computacional.
 - Realizar un estudio del estado del arte de las aplicaciones de la IC a la demencia.

- Selección de características y su confrontación con las usadas habitualmente por los expertos para el diagnóstico diferencial entre Demencia Vasculare y Demencia tipo Alzheimer.
 - Estimar qué técnicas de IC podrían ser más interesantes, dadas las características de los datos disponibles.
 - Analizar los resultados obtenidos con las variables seleccionadas de manera automática y las de los expertos en las mismas técnicas de diagnóstico diferencial.

1.2. Organización de la memoria

En la primera parte de esta memoria se introduce la teoría sobre la que se van a sustentar los métodos y aproximaciones que se van a emplear a lo largo del trabajo. En el Capítulo 3 se tratan las técnicas de Inteligencia Computacional básicas y en el Capítulo 2 la relevancia social de la demencia. Por último, el Capítulo 4 muestra un estudio del estado del arte en cuanto a qué técnicas de IC son las más empleadas en cada campo relacionado con la demencia.

Posteriormente, en la segunda parte de este trabajo, se presenta el caso de estudio, en el Capítulo 5. Se explica el problema asociado al diagnóstico diferencial entre la demencia vascular y la demencia tipo Alzheimer y la metodología, Capítulo 6, empleada para hacer frente a este problema y cuál es la solución más adecuada, a la vista de los resultados, Capítulo 7.

En las conclusiones, Capítulo 8, se presentan los hallazgos obtenidos tras la realización de este trabajo, así como las futuras líneas de investigación y posibles mejoras que se pueden abordar si se desea continuar el estudio.

El último apartado de la memoria corresponde a los Apéndices. El Apéndice A muestra las tablas completas, que por cuestión de espacio no se han incluido en el texto de la memoria. El Apéndice B contiene los detalles del proceso de inducción de particiones de aquellas variables con un elevado número de valores. Por último, el Apéndice C detalla los resultados completos de la ejecución de los algoritmos, así como los parámetros de configuración utilizados.

Parte I

Breve estado del arte: técnicas de Inteligencia Computacional aplicadas a la Demencia

Capítulo 2

Relevancia de la Demencia en la Sociedad

En este capítulo se va a justificar la importancia de la Demencia en la Sociedad desde distintos puntos de vista y la necesidad de investigar en este campo. La Enfermedad de Alzheimer y la Demencia Vascular suponen la gran mayoría de los casos de Demencia. Dado que la carga asociada a la Demencia, tanto para el sistema sanitario como para los familiares y cuidadores, es cada vez mayor debido a la mayor esperanza de vida, un diagnóstico acertado puede tener efectos muy beneficiosos al mejorar el tratamiento de los enfermos y sus posibilidades de mantenerse activos durante más tiempo.

En el capítulo siguiente se verán las principales técnicas de Inteligencia Computacional. Posteriormente, en la segunda parte de este trabajo, se estimará cuáles podrían ser más interesantes para su aplicación a un problema real asociado a la Demencia.

2.1. Qué es la Demencia

Según el diccionario de la RAE, el término demencia deriva del latín “*dementia*” y significa: Locura, trastorno de la razón. La definición médica es un poco más precisa: Deterioro progresivo e irreversible de las facultades mentales que causa graves trastornos de conducta. La Organización Mundial de la Salud define¹ la demencia “*como un síndrome debido a una enfermedad del cerebro, generalmente de naturaleza crónica o progresiva, en la que hay déficits de múltiples funciones corticales superiores.... que repercuten en la actividad cotidiana del enfermo*”.

Entre las funciones corticales superiores que el enfermo va perdiendo figuran la memoria, el entendimiento, el juicio, el habla, el cálculo, el pensamiento, la orientación, etc. No todas se deterioran simultáneamente, sino que es un proceso continuo en el que cada vez se percibe mayor número de funciones afectadas y con progresivo mayor deterioro, siendo generalmente la memoria la primera observación de alteración que percibe el enfermo o sus parientes más próximos. Sólo se puede hablar de demencia si la alteración cognoscitiva afecta a varias

¹en la Clasificación Internacional de las Enfermedades Décima Edición (CIE-10, 1992).

Capítulo 2 : Relevancia de la Demencia en la Sociedad

funciones corticales superiores.

Un grupo de investigación del Instituto Karolinska de Suecia ha intentado calcular el coste mundial de la demencia en 2005 [5]. Este coste asciende a 315 mil millones de dólares al año, de los que 227 mil millones (72 % del total mundial) es invertido por países de renta alta y 88 mil millones (28 % del total) por países de renta baja y media. El desglose estimado de estos costes viene ilustrado en la Figura 2.1. Se puede ver que la asistencia informal (familiar) se utiliza con más frecuencia en países pobres en recursos, donde los servicios sanitarios o sociales formales disponibles son escasos. La asistencia informal asciende al 56 % de los costes en países de renta baja, al 42 % en países de renta media, y al 31 % en países de renta alta.

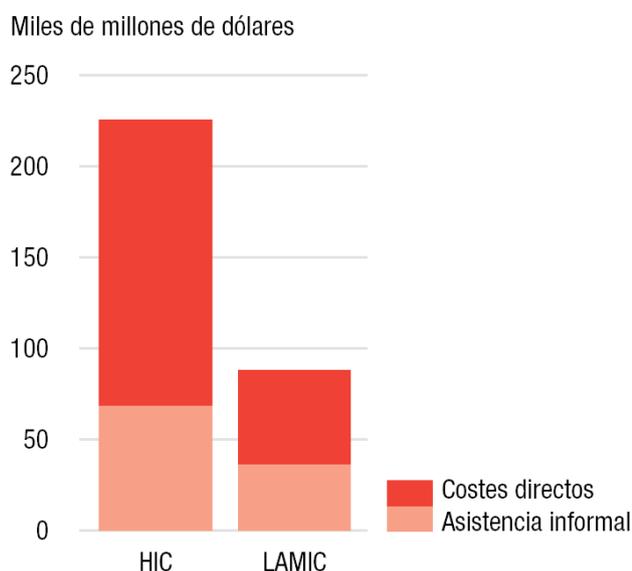


Figura 2.1: El coste social del Alzheimer y otras demencias en el mundo en 2005, atribuibles a costes directos e indirectos, en países de renta alta (HIC) y países de renta baja y media (LAMIC).

De todos los tipos de demencias, la enfermedad de Alzheimer y la demencia de tipo vascular suponen conjuntamente más del 90 % de los casos [30], aunque existen variaciones muy marcadas en distintas regiones del planeta. Por ejemplo, las demencias por enfermedades carenciales y las demencias por hidrocefalia son superiores al 30 % en muchos países africanos.

La relación entre los porcentajes de demencia Alzheimer y demencia vascular (considerada ésta como una entidad que agrupa a cualquier tipo de alteración de la circulación sanguínea a nivel cerebral: aterosclerosis, trombos, infartos, hemorragias, etc) varía muy poco entre los diferentes países desarrollados. Aproximadamente cada entidad representa un 50 % de los casos, existiendo un 10 % de patologías mixtas. Sin que se conozcan las causas hay una ligera superioridad de casos de Alzheimer en occidente, con independencia del país que se considere, frente a una mayor incidencia de casos de demencia vascular en oriente.

2.2. Enfermedad de Alzheimer

En 1906, en Alemania Alois Alzheimer describe el caso de una mujer de 51 años con deterioro cognitivo, alucinaciones auditivas, delirios, síntomas focales, incontinencia, inmovilidad y úlceras por presión.

En 1910 Krapelin llama al resultado del estudio histológico de un cerebro que presenta atrofia cerebral, aterosclerosis, pérdida neuronal y depósito amiloide fuera de la neurona y depósito de fibrillas dentro, “Enfermedad de Alzheimer”.

No fue hasta la década de los setenta cuando la enfermedad descrita por Alois Alzheimer empezó a cobrar interés. Hoy día se sabe que la degeneración neurofibrilar y las placas seniles son lesiones propias de la ancianidad y que la enfermedad de Alzheimer las comparte con otras alteraciones. Actualmente la mayor esperanza de vida o la disminución de la mortalidad han hecho que presente una alta prevalencia e incidencia.

En el 2006 se definen las demencias en diferentes escalas: DSM-IV, CIE-10, NINCDS-ADRDA, SEN, etc.

2.2.1. Epidemiología

Epidemiología descriptiva de las demencias

La epidemiología es el estudio de la distribución de la enfermedad de la población. Se puede usar para hacer estudios prospectivos (de la exposición a la enfermedad) o retrospectivos (de la enfermedad a la exposición). A su vez los estudios pueden ser:

- **Experimentales:** Se coge una muestra aleatoria ante diferentes condiciones experimentales con relación causa-efecto siendo ideal el ensayo clínico que evalúa los tratamientos nuevos. Otros serían los ensayos de campo para la prevención primaria o los ensayos comunitarios de intervención para toda la comunidad.
- **Observacionales:** Solamente pueden describir; pero no pueden investigar las causas o analizar la muestra.

Epidemiología analítica de las demencias

Se basa en el estudio de los factores de riesgo para establecer hipótesis de causa. Entre los factores de riesgo de enfermedad de Alzheimer se distinguen:

- Establecidos:
 1. Edad: La prevalencia aumenta con la edad y más después de los 60 años. También existen las formas precoces.
 2. Genética: Aumenta el riesgo 3,5 veces para familiares de primer grado. Aumenta más si en la familia hay más de un familiar afectado o si alguno tiene Síndrome de Down. Disminuye el riesgo cuanto mayor sea la edad de aparición del afectado.
 3. Mutaciones: El cromosoma 21 posee el gen que codifica la proteína precursora del amiloide, que mutado da un fragmento que se deposita entre las neuronas.

Capítulo 2 : Relevancia de la Demencia en la Sociedad

El cromosoma 14 contiene el gen que genera la proteína Presenilina. Aparece en las formas familiares más precoces que son muy graves porque son muy severas y de corta supervivencia.

El cromosoma 1 contiene el gen de la proteína Presenilina 2. Aparece en las formas familiares tardías y poco severas.

A pesar de que el estudio genético que busque estas mutaciones sea negativo, se puede padecer la enfermedad de Alzheimer por presentar poliformismos.

4. Poliformismos en el cromosoma 19: en el gen que codifica la proteína APO-E. Existen varias isoformas que se presentan en cada individuo de 2 en 2 en alelos. Se relaciona la E4 con aumento del riesgo de Enfermedad de Alzheimer y disminución de la edad de inicio. Sin embargo, la forma E2 parece ser protectora. Sujetos sin E4 pueden padecer la enfermedad y algunos sujetos con E4/E4 no desarrollan la enfermedad. Por eso no se debe hacer consejo genético en sanos familiares de los enfermos.

- Otros factores:

1. Sexo femenino: Tiene relación con su mayor supervivencia.
2. Déficit de estrógenos: Se supone que el tratamiento con estrógenos en mujeres posmenopáusicas, tiene efecto protector al actuar sobre el sistema colinérgico, sobre los factores de crecimiento nervioso en el hipocampo y al tener aquellos, efecto antioxidante. Varios estudios no han podido demostrar esa protección.
3. Traumatismos craneales: Se cree que aquellos que producen pérdida de conocimiento o son repetidos, presentan lesiones similares a la degeneración neurofibrilar del Alzheimer
4. Bajo nivel educativo: Aquellos individuos con mayor reserva funcional, siguen siendo independientes en las actividades de la vida diaria unos años más.
5. Factor de riesgo vascular: La diabetes y la hipertensión arterial.
6. Los hijos de madres con mucha edad: Mayores de 35 años, parecen tener mayor riesgo de padecer demencia.
7. Depresiones previas: Sólo en los casos en los que la depresión se deba a una causa orgánica (degenerativa).
8. Exposición al aluminio: Se ha desmentido que el aluminio que aparece depositado en las neuronas proceda del exterior.

- Posibles factores protectores del Alzheimer:

1. Tabaco: Los receptores nicotínicos cerebrales producen relajación, pero algunos estudios han visto que incrementa el riesgo de padecer demencia.
2. Uso de estrógenos.
3. Uso de antiinflamatorios no esteroideos: Parece que protegen al reducir la inflamación de las placas que producen la enfermedad.
4. Grado de reserva funcional: Los años de estudio parecen enlentecer el desarrollo de atrofia cerebral. No es concluyente.

2.2.2. Factores de riesgo

Factores genéticos

- Historia familiar: un familiar cercano (padre o hermano) que haya padecido la enfermedad de Alzheimer, constituye un factor de riesgo.
- Mutaciones causales: poseer una mutación es causa suficiente para desarrollar la enfermedad y ésta será de inicio precoz. Estas mutaciones, que sólo representan el 1% de los casos, pueden situarse en:
 - Gen de la proteína precursora del amiloide (APP).
 - Gen de la presenilina 1 (PS1).
 - Gen de la presenilina 2 (PS2).
- Genotipo APOE: La proteína APOE puede presentarse de 3 formas distintas (ApoE-e 2, ApoE-e 3, ApoE-e 4). La ApoE-e 4 constituye un factor de riesgo, que en este caso será de inicio tardío.
- Síndrome de Down: existe una relación clara entre el síndrome de Down y la enfermedad de Alzheimer. Estas personas muestran a los 40-50 años lesiones características de la enfermedad.

Factores demográficos

- Edad: es el principal factor de riesgo para la aparición de la enfermedad. A partir de los 65 años la EA aumenta exponencialmente, duplicándose cada cinco años.
- Género: las mujeres presentan una mayor incidencia.
- Etnia y nacionalidad: varía entre regiones y sociedades.
- Nutricionales: las dietas hipercalóricas y las ricas en ácidos grasos saturados, la obesidad y la ingesta excesiva de alcohol suponen un factor de riesgo.
- Medio rural: el pertenecer a familias con condiciones poco favorables, económicas, sanitarias, nutricionales, etc., puede favorecer a que la educación y maduración cerebral sea menor, limitándose por ello la reserva cerebral.
- Actitud existencial: las personas con tendencia al optimismo presentan un menor riesgo.

Antecedentes médicos

- Depresión: puede manifestarse en forma de demencia y puede ser uno de los primeros síntomas de la enfermedad.
- Menopausia, andropausia: está demostrado que el uso de terapias hormonales sustitutivas en mujeres previene ante el riesgo de padecer la enfermedad.

Capítulo 2 : Relevancia de la Demencia en la Sociedad

- Hipertensión arterial e hipercolesterolemia: ambos son factores de riesgo porque pueden producir lesiones cerebrales vasculares.
- Hiperhomocisteinemia, deficiencia en vitamina B12 y folatos.
- Toxicambientales: distintos productos tóxicos pueden contribuir a la aparición de la enfermedad, tabaco, insecticidas, fertilizantes, etc.

2.2.3. Diagnóstico

No es una enfermedad de fácil diagnóstico y en un primer momento éste se hace por eliminación, realizando múltiples pruebas.

Antecedentes

Se reúnen antecedentes detallados sobre los pacientes, incluyendo:

- Una descripción de cómo y cuándo comenzaron los síntomas.
- Una descripción del paciente y del trastorno médico, incluyendo antecedentes generales familiares.
- Una evaluación del estado emocional del paciente y de su medio ambiente.

Entorno

Recopilar información a través de los miembros de la familia o de los amigos íntimos. Las personas cercanas al paciente pueden brindar información valiosa sobre cómo ha cambiado su comportamiento y su personalidad; muchas veces, la familia y los amigos saben que algo anda mal, aún antes de que los cambios sean evidentes en las pruebas de diagnóstico.

Examen físico

Se realizan exámenes físicos y neurológicos y pruebas de laboratorio. Los exámenes de sangre y otras pruebas médicas ayudan a determinar el funcionamiento neurológico y a identificar las causas de la demencia no relacionada con el Alzheimer.

Imágenes cerebrales

Realización de una tomografía computarizada (CT) o una prueba de imágenes por resonancia magnética (MRI). Los exámenes cerebrales como éstos pueden detectar accidentes cerebrovasculares o tumores, o pueden revelar cambios en la estructura y en el funcionamiento del cerebro que indican el desarrollo de Alzheimer precoz.

Pruebas neuropsicológicas

Las pruebas de Preguntas y Respuestas o las tareas de otro tipo que miden la memoria, las aptitudes del lenguaje, la capacidad de hacer ejercicios matemáticos y las capacidades de otro tipo relacionadas al funcionamiento cerebral, indican la clase de cambios cognitivos que ocurren en el cerebro.

Criterios

Determinados criterios se usan para determinar una “probable” enfermedad de Alzheimer. Dado que no existe ninguna prueba biológica sencilla y confiable para diagnosticar Alzheimer, el Instituto Nacional de Trastornos Neurológicos y Accidentes Cerebrovasculares y la Asociación de Alzheimer han establecido criterios para ayudar a los médicos a diagnosticar Alzheimer. Estos criterios también ayudan a los médicos a distinguir entre las formas de Alzheimer y otros tipos de demencias. La enfermedad de Alzheimer “probable” se determina cuando una persona tiene:

- Demencia confirmada mediante un examen clínico y neuropsicológico.
- Problemas en al menos dos áreas del funcionamiento mental.
- Empeoramiento progresivo de la memoria y de otra función mental.
- Ningún trastorno de la conciencia (no existen períodos en blanco).
- Comienzo de los síntomas entre los 40 y los 90 años de edad.
- Ninguna otra enfermedad que pudiera incluir demencia.
- A medida que envejecen, algunas personas desarrollan un déficit de memoria mayor de lo que se espera para su edad. Sin embargo, otros aspectos de la función cognitiva no son afectados, de manera que estas personas no satisfacen todos los criterios establecidos en el diagnóstico de Alzheimer. Por lo tanto, se dice que tienen una “deficiencia cognitiva leve” (MCI, por sus siglas en inglés *Mild Cognitive Impairment*). Cerca del 40% de estos individuos desarrollarán Alzheimer en un período de 3 años.

2.3. Demencia vascular

La demencia vascular o multi-infarto en realidad comprende todo un grupo de alteraciones cognitivas demenciales consecutivas a diversos tipos de alteraciones originadas por hemorragias, trombos, infartos, etc., en el cerebro. Se distingue de la demencia tipo Alzheimer por el modo de comienzo, las características clínicas y la evolución, como indica la Tabla 2.1. El comienzo de la enfermedad tiene lugar en la edad avanzada. El cuadro suele ser brusco, como consecuencia de un episodio isquémico aislado y el deterioro cursa con brotes pues normalmente ocurren accidentes o microaccidentes repetidos. Después de un empeoramiento, suele haber periodos de recuperación parcial, al ponerse en marcha los mecanismos de reparación cerebral. Los factores de mayor riesgo para padecer una demencia de tipo vascular son: hipertensión, diabetes, enfermedades cardiovasculares, accidentes cerebrovasculares, etc.

Dentro de este tipo de demencia podemos encontrar distintos subtipos como:

- Demencia vascular de inicio agudo. Se desarrolla de un modo rápido tras una serie de ictus debidos a trombosis vasculares, embolias o hemorragias.
- Demencia multi-infarto. Tiene un inicio más gradual, normalmente precedido de varios episodios isquémicos menores.

Capítulo 2 : Relevancia de la Demencia en la Sociedad

- Demencia vascular subcortical. Pueden existir antecedentes de hipertensión arterial y de focos de destrucción isquémica en las zonas profundas de la sustancia blanca de los hemisferios cerebrales. Por lo general, la corteza cerebral está indemne, lo cual contrasta con el cuadro clínico, que puede parecerse mucho al de una demencia de la EA.
- Demencia vascular mixta, cortical y subcortical.
- Otra demencia vascular.
- Demencia vascular sin especificación.

Tabla 2.1: Comparativa entre Demencia Vasculuar y Demencia tipo Alzheimer [4].

Tipo	Características
EA	<p>Tipo más común de demencia, entre el 60 y el 80 % de los casos.</p> <p>Dificultad para recordar nombres y eventos recientes. Apatía y depresión son también con frecuencia síntomas iniciales. Los síntomas más avanzados incluyen desorientación, confusión, cambios de caracter, dificultades para hablar, tragar, caminar.</p> <p>Nuevos criterios y guías para el diagnóstico han sido publicados en 2011. Éstos, recomiendan considerar al Alzheimer como una enfermedad que comienza antes del desarrollo de los síntomas.</p>
DV	<p>Previamente conocida como demencia post-infarto o post-apoplejía. Es menos común que el Alzheimer como causa única de demencia.</p> <p>Las alteraciones en el juicio o la incapacidad para hacer planes son más comunes como síntomas iniciales, a diferencia del Alzheimer, más relacionado con las pérdidas de memoria.</p> <p>Ocurre debido a daños en el cerebro como micro hemorragias o trombos. La localización del daño cerebral determinará como se verá afectado el individuo en su pensamiento y condición física.</p> <p>En el pasado, las evidencias de demencia vascular se usaban para descartar el diagnóstico del Alzheimer (y viceversa). esta práctica ya no se considera consistente ya que los cambios en el cerebro por ambos tipos de demencia se pueden dar simultáneamente.</p>

De lo expuesto en la Tabla 2.1, hay que destacar que en 2011 el *National Institute on Ageing* (NIA) y la *Alzheimer's Association* recomendaron nuevos criterios y guías para el diagnóstico de la Enfermedad de Alzheimer. Por tanto, todavía se está avanzando en el conocimiento de las demencias y sus características, y hay campo para la investigación en este campo.

Capítulo 3

Inteligencia Computacional

El término *Inteligencia Computacional* engloba técnicas que se basan, al menos parcialmente, en la forma de pensar del ser humano y el funcionamiento del cerebro. Con este nuevo enfoque se trata de superar las limitaciones de la computación tradicional para resolver problemas del mundo real de una manera eficiente. En primer lugar se introducirá la teoría de los sistemas de clasificación para, posteriormente mostrar las aportaciones de los sistemas inteligentes en este ámbito, en lo que se conoce como clasificadores inteligentes, que engloba técnicas tan diversas como árboles de decisión, redes neuronales, máquinas de vectores soporte, por mencionar algunas.

En el capítulo anterior se ha explicado la importancia de la Demencia para la Sociedad. En éste, se introducirán brevemente los principales algoritmos de Inteligencia Computacional. De la combinación de ambos saldrán los conceptos que se usarán en el siguiente capítulo para realizar el estudio del estado del arte.

3.1. Sistemas de clasificación

El aprendizaje automático se ocupa de la forma de hacer que el ordenador aprenda por sí solo. Generalmente, se aplica a problemas complejos, en los que no es posible acertar siempre, por lo que se debe admitir que el ordenador pueda cometer algún error en su diagnóstico, al igual que los cometen a veces los expertos humanos [67].

Si se pretende que el ordenador sea capaz de hacer determinadas predicciones, será necesario indicarle de dónde debe aprender (ejemplos), cuál es el objetivo a alcanzar (condición de parada) y qué tipo de resultados se esperan de él. En definitiva, se trata de utilizar la experiencia para realizar una serie de tareas, y para ello, esa experiencia se recoge en forma de datos que son guardados en un ordenador.

El conocimiento necesario para construir los primeros sistemas expertos se obtenía a través de la colaboración entre un experto en el dominio de la aplicación y un ingeniero del conocimiento, presentándose el problema del “cuello de botella” [19]. Este concepto ha estimulado la investigación en métodos de aprendizaje automático como medio para explicar el conocimiento [45].

Los problemas de aprendizaje automático pueden abordarse como problemas de clasificación, es decir, asignar a un objeto descrito mediante un conjunto de

características, una determinada categoría o clase. Se pueden distinguir dos tipos de clasificaciones [47]:

Supervisada Métodos que determinan, de forma automática, la clasificación que se debe asignar a un caso dado, a partir de ejemplos anteriores en los que la clase era conocida.

No Supervisada También conocida como “clustering”, enfoca la clasificación como el descubrimiento de las clases del problema. Los objetos únicamente vienen descritos por un vector de características, sin que se conozca la clase a la que pertenece cada uno de ellos.

Para que el sistema pueda tomar una decisión sobre la clasificación que va a asignar, es necesario indicarle las características que se han observado en el caso que se está estudiando actualmente, denominadas *variables predictoras*, y el sistema debe discernir entre una serie de categorías a la hora de tomar su decisión sobre la clasificación a asignar. A los distintos valores que se tienen en cuenta se les denomina *clases* o *categorías*. El conjunto de características que describe a un objeto puede ser arbitrariamente grande, por lo que en la mayoría de las ocasiones se hace necesaria una etapa previa de *Selección de Características*.

3.2. Selección de Características

Las técnicas de *Extracción y Selección de Características* pretenden reducir el tamaño del conjunto de datos, eligiendo las variables más relevantes para el caso de estudio, a partir de las características disponibles [43]. Éste es un aspecto que toma gran relevancia en multitud de áreas, especialmente en reconocimiento de patrones [42], cuando el conjunto de variables es elevado en relación al número de ejemplos disponibles.

El proceso de extracción de características, por su parte, trata de crear nuevas características basadas en las transformaciones y combinaciones de las características originales, creando con ello un nuevo espacio de representación del problema. Por lo general ambas técnicas se suelen emplear de forma conjunta.

Los objetivos generales buscados al realizar una etapa de selección de características son los siguientes [31]:

- Mejorar tanto el proceso como el modelo de un problema.
- Mejor comprensión del proceso que genera los datos.
- Mejorar la eficiencia del predictor (predictores más rápidos y efectivos en cuanto a coste de implementación y ejecución).

La realización de un proceso de selección de características permite obtener una serie de beneficios, que dependiendo del método empleado, se alcanzarán en mayor o menor medida. Estos beneficios obtenidos son los siguientes:

- Mejorar la visualización y entendimiento de la información disponible para el problema.
- Reducción de los requisitos de dimensionamiento y almacenamiento de la información.

- Reducción del tiempo de entrenamiento de los modelos, al tratar con un subconjunto más reducido de datos.
- Reducción del ruido en los datos, eliminando las características irrelevantes o redundantes.
- Aumentar el rendimiento, reduciendo la probabilidad de aparición de la maldición de la dimensionalidad (*curse of dimensionality*).

Pero también hay algunos inconvenientes. Reducir la dimensión del universo del problema puede originar clasificadores que dependan en gran medida de las características seleccionadas, por lo que una mala reducción puede provocar una pérdida en el poder discriminante del clasificador. De esto se deduce la importancia de elegir bien con qué características trabajar.

3.3. Técnicas de Inteligencia Computacional

A continuación se incluye una breve revisión de las principales técnicas de Inteligencia Computacional. Algunas de ellas serán usadas en la segunda parte de este trabajo para realizar la experimentación.

3.3.1. Árboles de Decisión

Los árboles de decisión son, quizá, uno de los paradigmas más utilizados en el mundo del aprendizaje automático, como lo demuestra el hecho de que dos de los diez algoritmos más utilizados estén basados en árboles [75]. Los factores que han influido de una manera determinante a su difusión son, entre otros, la disponibilidad de diferentes implementaciones accesibles en código abierto y algunas de sus características:

- La sencillez del modelo, ya que la idea de “árbol” es fácil de comprender.
- La explicación que aporta a la clasificación.
- La posibilidad de ser representados gráficamente.
- La rapidez a la hora de clasificar nuevos ejemplos.

J. Ross Quinlan desarrolló un método para inducir árboles de clasificación a partir de un conjunto de ejemplos, como se refleja en *C4.5: Programs for Machine Learning* [59], y al conjunto de programas que lo implementaban lo denominó **C4.5**, que era una evolución de un programa previo del mismo autor denominado **ID3** (*Iterative Dichotomizer 3*) [57]. Un árbol de decisión clasifica un ejemplo filtrándolo de manera descendente, hasta encontrar una hoja, que corresponde a la clasificación buscada.

Una cuestión importante es determinar cuál es el tamaño correcto de un árbol. En ocasiones, dependiendo del algoritmo empleado para la construcción del árbol, éste desarrolla cada rama en profundidad hasta que logra clasificar perfectamente los ejemplos de entrenamiento. Aunque parece una estrategia razonable, puede introducir complicaciones si los datos de entrenamiento están afectados por ruido, o cuando el conjunto de entrenamiento es demasiado pequeño como para considerarse un muestreo significativo del concepto objetivo.

En estos casos, pueden producirse árboles que se sobreajustan (*overfitting*) a los datos de entrenamiento. A medida que el tamaño del árbol crece, en términos del número de nodos, su precisión sobre el conjunto de entrenamiento mejora monótonicamente, pero sobre el conjunto de test primero crece y luego decae (ver Figura 3.1).

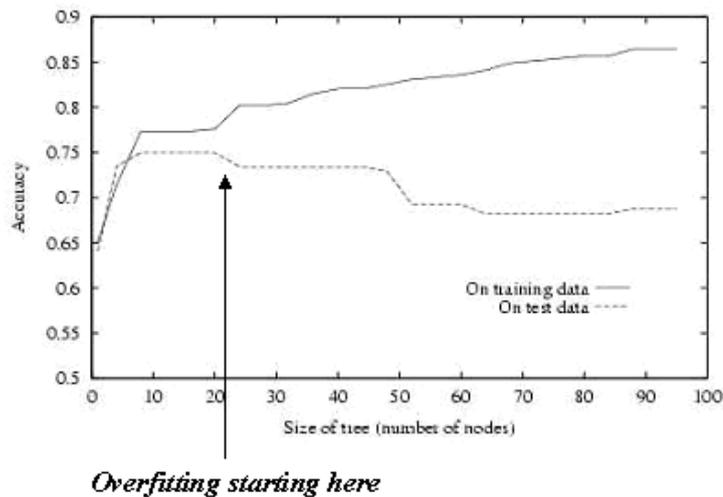


Figura 3.1: Ejemplo de sobreajuste.

Existe la posibilidad de sobreajuste, aun cuando el conjunto de entrenamiento esté libre de ruido, por ejemplo, si el conjunto de entrenamiento tiene pocos elementos. En conjuntos de entrenamiento pequeños es fácil encontrar regularidades accidentales, donde un atributo puede particionar muy bien los ejemplos dados, aunque no esté relacionado con el concepto objetivo. El fenómeno del sobreajuste puede reducir la precisión de un árbol entre un 10 y un 25 %, por lo que se han propuesto diferentes enfoques para evitarlo. Estos enfoques pueden agruparse en dos clases:

- Detener el crecimiento del árbol anticipadamente, antes de que alcance un punto donde clasifique perfectamente los ejemplos de entrenamiento.
- Dejar crecer el árbol, para después podarlo.

Aunque el primer enfoque parezca más directo, la poda posterior del árbol ha demostrado tener más éxito en la práctica. Esto se debe a la dificultad de estimar en qué momento debe detenerse el crecimiento del árbol. Un enfoque llamado *reduced-error pruning* [58] consiste en considerar cada nodo del árbol como candidato a ser podado y para ello, utiliza un conjunto de ejemplos de validación. La poda consiste en eliminar todo el subárbol que tiene como raíz el nodo en cuestión, convirtiéndolo así en una hoja, cuya clase corresponde al valor más común de los ejemplares asociados a ese nodo.

Un nodo sólo es eliminado si el árbol podado que resulta de ello, no presenta un rendimiento peor que el árbol original sobre el conjunto de validación. El efecto de esto es que los nodos que se han colocado en el árbol por coincidencias

fortuitas en los datos de entrenamiento, generalmente son eliminados debido a que las coincidencias no suelen estar presentes en el conjunto de validación.

Este método es efectivo únicamente si contamos con suficientes ejemplos, de tal forma que el conjunto de entrenamiento y el conjunto de validación sean significativos, estadísticamente hablando. De otra forma, tomar ejemplos para el conjunto de validación reduce aún más el tamaño del conjunto de entrenamiento, aumentando así la posibilidad de sobreajuste.

3.3.2. Redes Neuronales

Las Redes Neuronales son modelos de computación flexible cuyo origen se basa en estudios biológicos. Estos estudios trataban de reproducir el comportamiento del cerebro humano mediante el desarrollo de un sistema de cálculo paralelo capaz de aprender mediante la interconexión en capas de pequeñas unidades de cálculo. La descripción realizada por McCulloch-Pitts [55] de una unidad de cálculo elemental denominada neurona se considera el origen de las redes neuronales. Posteriormente, apareció la teoría de las Redes Neuronales, con desarrollos como el perceptrón de Rosenblatt [62] y el Adaline de Widrow [73]. Sin embargo, a raíz del trabajo de Minsky [46], en el que se identificó que el perceptrón no era capaz de aprender el problema XOR, este tipo de técnicas cayeron en desuso. Sin embargo, en 1986, Rumelhart [63] aplicó un concepto descubierto anteriormente por Werbos [71] en 1974, conocido como la Regla Delta, al entrenamiento de este tipo de técnicas, lo que supuso una auténtica revolución en el desarrollo de las redes neuronales.

Existe una amplia variedad de arquitecturas, funciones de activación de las neuronas y algoritmos de entrenamiento [50]. Su característica principal es su gran capacidad de aprendizaje y su consideración como aproximadores universales a funciones continuas [28]. De los múltiples modelos que existen, algunos de los más conocidos son:

MLP (Perceptrón Multicapa)

El perceptrón multicapa (MLP, del inglés *Multi-Layer Perceptron*) [63], es el modelo de red neuronal más conocido y empleado. Utilizado en problemas muy diversos, como por ejemplo, la predicción del volumen de polen en la ciudad de Granada [7] o la predicción de la demanda de electricidad [1]. Un perceptrón multicapa está constituido por un conjunto de neuronas que se comunican entre sí a través de un conjunto de conexiones ponderadas denominadas pesos [13].

La salida del sistema depende tanto de las entradas como de los pesos que se almacenan en cada conexión. Cada neurona realiza un cálculo sencillo para generar su salida. La función que se aplica se denomina función de activación, y generalmente es no lineal:

$$y_i^{[k]} = x_i^{[k+1]} = f\left(\sum_j w_{ij}^{[k]} x_j^k\right),$$

donde y_i es el valor que transmite la neurona a la capa siguiente, x_i es el valor que toma la neurona i -ésima de la capa, w_{ij} son los pesos de la neurona i de la capa k con el resto de neuronas de la capa anterior, siendo j el número de neuronas de la capa anterior, y $[k]$ identifica el número de capa k (la salida de la capa k coincide con entrada de la capa $k + 1$).

Entrenamiento: El algoritmo de aprendizaje supervisado denominado retro-propagación del error [63] es el más utilizado para ajustar redes neuronales de este tipo y se basa en la regla delta [73]. Este algoritmo puede encontrar mínimos locales, lo que producirá que la solución no sea la óptima aunque pueda ser razonablemente buena. Para evitar este tipo de problemas, se realizan modificaciones al algoritmo base, como añadir un término momento (α). Mediante la introducción del término momento se consigue, además de evitar algunos mínimos locales, acelerar el aprendizaje.

RBF (Redes Neuronales con Funciones de Base Radial)

Las Redes Neuronales con función de base radial (RBF, *Radial Basis Function*) [13] [12] son un tipo de red neuronal cuya estructura es de tres capas, y que emplea en sus neuronas un tipo de funciones específico: gaussiana o de forma acampanada. Las RBF cuentan con alimentación hacia delante y arquitectura fija de 3 capas. Utilizan combinaciones de funciones base centradas en vectores de pesos denominados centroides.

Cada neurona de la capa oculta almacena un vector de pesos denominado centroide. En estas neuronas de la capa oculta se calcula una distancia (típicamente distancia euclídea) entre el patrón de entrada y el centroide, y sobre esta distancia se aplica una función de tipo radial (normalmente una función gaussiana) de forma acampanada.

El criterio de activación de una neurona es la proximidad entre el vector de pesos y el centroide. Cuando la distancia es muy grande, la salida de la función tiende a cero y cuando la distancia es nula, se produce el máximo grado de activación. La salida de las neuronas de la capa de salida es la siguiente:

$$h_j(\vec{x}) = \Phi(\|\vec{x} - \vec{c}_j\|, \sigma),$$

donde Φ es una función de base radial, σ es el grado de influencia o “anchura” de la función, c_j es el vector centroide, y x es el patrón de entrada.

El objetivo de las funciones radiales es particionar el espacio de entrada, de manera que cada clase de patrones quede representada por su vector centroide. Visualmente, el objetivo es que la mayoría de los patrones queden recubiertos por alguna *campana* de las funciones radiales.

La salida de cada neurona de la capa de salida es calculada mediante la aplicación de una función (lineal o sigmoidea) a la suma ponderada de las salidas de la capa oculta. Esta función se denomina función de activación de la capa de salida.

3.3.3. Máquinas de Vectores Soporte (SVM)

Las máquinas de vectores soporte (o SVM, del inglés *Support Vector Machines*) son un conjunto de algoritmos de aprendizaje supervisado que, a partir de un conjunto de ejemplos de entrenamiento (muestras), puede etiquetar las clases para construir un modelo que prediga la clase de una nueva muestra. Es decir, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. De una manera más formal, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta que puede ser utilizado en problemas de

3.3 Técnicas de Inteligencia Computacional

clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta.

Según el *Top Ten Algorithms in Data Mining* [75], las máquinas de vectores soporte, incluyendo SVC (*support vector classifier*) y SVR (*support vector regressor*), se encuentran entre los métodos más robustos y precisos de todos los algoritmos de minería de datos. Las SVM se basan en la teoría del aprendizaje estadístico, requieren muy pocos ejemplos para entrenarlas y frecuentemente son insensibles al número de dimensiones.

La mayor diferencia entre las máquinas de vectores soporte y otros modelos predictivos es que las SVM no se centran en construir hipótesis que cometan pocos errores, sino que lo que pretenden es producir predicciones en las que se pueda tener mucha confianza, aún a costa de cometer ciertos errores. Tradicionalmente, la mayoría de los modelos predictivos se han centrado en lo primero, reducir al mínimo los errores cometidos por el modelo generado. Se basan en el denominado principio de Minimización del Riesgo Empírico (ERM, *Empirical Risk Minimization*). El enfoque de las máquinas de vectores soporte es diferente, no buscan reducir el riesgo empírico cometiendo pocos errores, sino que pretenden construir modelos confiables. Ese principio recibe el nombre de Minimización del Riesgo Estructural (SRM, *Structural Risk Minimization*). Las SVM buscan un modelo que estructuralmente tenga poco riesgo de cometer errores ante datos futuros.

Tabla 3.1: Comparativa entre redes neuronales artificiales y SVM [74].

ANN	SVM
Capas ocultas transforman a espacios de cualquier dimensión.	Kernels transforman a espacios de dimensión muy superior.
El espacio de búsqueda tiene múltiples mínimos locales.	El espacio de búsqueda tiene sólo un mínimo global.
El entrenamiento es costoso.	El entrenamiento es muy eficiente.
La clasificación es muy eficiente.	La clasificación es muy eficiente.
Se diseña el número de capas ocultas y nodos.	Se diseña la función kernel y el parámetro de coste C .
Muy buen funcionamiento en problemas típicos.	Muy buen funcionamiento en problemas típicos.
	Extremadamente robusto para generalización, menos necesidad de heurísticos para entrenamiento.

Uno de los principales inconvenientes de las SVM es que la fase de entrenamiento es muy costosa, debido a su complejidad computacional, lo que se traduce en que es un algoritmo inaplicable en grandes conjuntos de datos. Sin embargo, este problema se está solucionando con éxito. Una posibilidad es dividir un gran problema de optimización en una serie de problemas más pequeños, en los cuales solamente se usan un par de variables cuidadosamente seleccionadas de modo que la optimización se pueda realizar eficientemente. El proceso se repite hasta que se resuelven satisfactoriamente todos los problemas de optimización en los que se descompuso el problema original.

3.3.4. Clustering

El objetivo del clustering es la clasificación de objetos de acuerdo a similitudes entre ellos y la organización de los datos en grupos. Las técnicas de clustering se encuentran entre los métodos no supervisados, ya que no necesitan ningún tipo de información previa o identificadores de clase. El mayor potencial del clustering es la detección de algún tipo de estructura o relación oculta en los datos.

Un cluster es un grupo de objetos que son más similares entre sí que a los miembros de los otros clusters. El término “similares” debe ser entendido como similitud matemática, medida en algún sentido bien definido, por ejemplo la distancia euclídea entre dos elementos. La definición de similaridad y el método para agrupar los datos difiere según el algoritmo de clustering aplicado. Por tanto, el “mejor” algoritmo de clustering a usar dependerá de las características del conjunto de datos y el propósito de la clasificación. En ocasiones se prueba con diferentes algoritmos y se elige el que ofrezca mejores resultados.

Los algoritmos de clustering siempre intentan encontrar el mejor ajuste entre el número de clusters definidos y las formas parametrizadas de los clusters. Pero eso no significa que incluso el mejor ajuste tenga significado en absoluto, ya que o bien el número de clusters puede ser erróneo, o las formas de los clusters pueden no corresponderse con los grupos de datos.

El método denominado **k-means** (o k-medias) es un algoritmo sencillo y ampliamente utilizado. Se aplica a datos representados por puntos en un espacio vectorial de d dimensiones. Por tanto, agrupa un conjunto de vectores d -dimensionales

$$D = \{x_i | i = 1, \dots, N\},$$

donde $x_i \in \mathbb{R}^d$ denota el i -ésimo objeto o “punto de datos”. Como se ha indicado previamente, k-means es un algoritmo de clustering que particiona D en k clusters de puntos. Es decir, agrupa todos los puntos de D de manera que cada punto x_i cae en una y sólo una de las k particiones.

El valor de k es una entrada al algoritmo base. Típicamente, el valor de k está basado en criterios como conocimiento previo sobre el número de clusters presentes en D , cuántos clusters se desea obtener para la aplicación actual, o el tipo de clusters encontrados a través de la exploración/experimentación con diferentes valores de k . Cada cluster se representa por un único punto en \mathbb{R}^d . Se denomina $C = \{c_j | j = 1, \dots, k\}$ a este conjunto de representantes de los clusters, también conocidos como centroides. Como se ha mencionado antes, es necesario definir una medida de proximidad o similitud. En k-means, la medida de similitud por defecto es la distancia Euclídea. Concretamente, se trata de minimizar la siguiente función de coste no negativa:

$$Cost = \sum_{i=1}^N (\operatorname{argmin}_j \|x_i - c_j\|_2^2).$$

En otras palabras, k-means intenta minimizar el cuadrado de la distancia euclídea total entre cada punto x_i y el centroide c_j más próximo. Esta ecuación también se conoce como la función objetivo de k-means.

La investigación en el método k-means ha sido muy extensa y todavía sigue activa. Existen otras variantes, las más conocidas son k-medoid [33] y Fuzzy

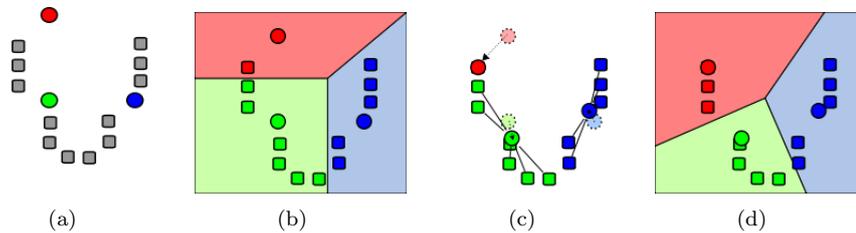


Figura 3.2: Ejemplo del algoritmo k-means estándar.

C-means [17]. La primera utiliza la mediana en lugar de la media para calcular los centroides de los clusters. De esta forma, el centroide es un punto de los pertenecientes al conjunto original, en lugar de un punto "ficticio" generado artificialmente. La segunda variante, Fuzzy C-means, se diferencia en que calcula funciones de pertenencia difusas para cada cluster, en lugar de un valor de pertenencia único y concreto.

La Figura 3.2 muestra un ejemplo sencillo del funcionamiento del algoritmo k-means estándar. En 3.2(a) se ven, resaltadas en distinto color, las k medias iniciales (en este caso, $k = 3$) escogidas aleatoriamente. Posteriormente 3.2(b), se crean los k clusters iniciales, asociando cada observación con la media más próxima. A continuación 3.2(c), el centroide de cada uno de los k clusters se convierte en la nueva media. Por último 3.2(d), se repiten los pasos 3.2(b) y 3.2(c) hasta que se alcanza la convergencia.

Otro algoritmo bastante difundido es el **Esperanza-Maximización** [15] o algoritmo **EM** (del inglés *Expectation-Maximization*) es un método iterativo para encontrar estimadores de parámetros de modelos estadísticos con la mayor probabilidad (*maximum likelihood*), donde el modelo depende de variables latentes no observables. El algoritmo EM alterna pasos de esperanza (paso E), donde se computa la esperanza de la verosimilitud mediante la inclusión de variables latentes como si fueran observables, y un paso de maximización (paso M), donde se computan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E. Los parámetros que se encuentran en el paso M se usan para comenzar el paso E siguiente, y así el proceso se repite.

En los casos en los que las ecuaciones de los modelos estadísticos no pueden ser resueltas directamente, se usa el algoritmo EM para encontrar los parámetros de máxima verosimilitud. Generalmente, estos modelos incluyen variables latentes, además de parámetros desconocidos y datos conocidos obtenidos de las observaciones. Es decir, hay valores desconocidos entre los datos o el modelo no puede ser formulado de manera más sencilla asumiendo la existencia de datos no observados adicionales.

Encontrar la solución de máxima verosimilitud requiere obtener las derivadas de las funciones de similitud con respecto a todos los valores desconocidos, i.e. los parámetros así como las variables latentes, y simultáneamente resolver las ecuaciones resultantes. En los modelos estadísticos con variables latentes, normalmente esto no es posible. En su lugar, el resultado suele ser un conjunto de ecuaciones interrelacionadas en las que la solución de los parámetros requiere los valores de las variables latentes y viceversa, pero si se sustituye un conjunto

de ecuaciones en el otro, se produce un sistema de ecuaciones sin solución. El algoritmo EM procede de forma que a partir de las observaciones encuentra la forma de resolver estos dos conjuntos de ecuaciones. Simplemente se asignan unos valores arbitrarios a uno de los dos conjuntos desconocidos, y se usan para estimar los valores del otro. A continuación se usan esos nuevos valores para encontrar una mejor estimación del primer conjunto, y así se van alternando entre ambos, hasta que los valores resultantes converjan hacia puntos fijos. El valor obtenido es un máximo de la función de similitud. En general, puede haber múltiples máximos y no hay garantía de que se encuentre el máximo global.

Para completar la lista de los métodos de clustering más populares, a continuación se explica brevemente el algoritmo de **Clustering Jerárquico** o *Hierarchical Clustering*. Este método trata de construir una jerarquía con los clusters. Las estrategias empleadas generalmente se agrupan en os clases:

- Por aglomeración: o aproximación “bottom-up”. Cada observación comienza en su propio cluster, y se van emparejando a medida que se sube en la jerarquía.
- Por división: o aproximación “top-down”. Todas las observaciones comienzan perteneciendo a un único cluster, que se va dividiendo recursivamente a medida que se desciende en la jerarquía.

Para decidir qué clusters deben ser combinados o divididos, se necesita una medida de la diferencia entre conjuntos de observaciones. En la mayoría de los métodos de clustering jerárquico, esto se logra a través del uso de una métrica adecuada (medida de la distancia entre pares de observaciones), y un criterio de enlace que especifica la diferencia de los conjuntos como una función de las distancias de las observaciones tomadas de dos en dos.

En general, las aglomeraciones y divisiones se determinan de manera egoísta y suele ser un proceso muy pesado computacionalmente, por lo que no es adecuado para conjuntos de datos de gran tamaño. Pero hay casos concretos en los que se pueden aplicar métodos optimizados. El resultado se suele presentar en forma de dendogramas.

3.3.5. ROCK

Se incluye el algoritmo ROCK [22] en una Sección a parte, ya que sus características lo hacen especialmente interesante para este trabajo. Otros algoritmos de clustering que utilizan distancias entre puntos para realizar las agrupaciones, no son apropiados cuando los atributos son binarios (verdadero/falso) o categóricos. Este es el caso de nuestro conjunto de datos, que contiene gran cantidad de variables cuyo valor es {Ausente/Presente}.

El algoritmo ROCK (*Robust Clustering Algorithm for Categorical Attributes*) utiliza un nuevo concepto de “enlace” para medir la similitud/proximidad entre dos puntos. Se basa en clustering jerárquico, pero emplea los enlaces en lugar de las distancias a la hora de combinar los clusters.

Los resultados mostrados por los autores indican que los clusters generados por ROCK son más puros, es decir, que separan mejor las instancias pertenecientes a cada clase. También se aprecian diferencias en el tamaño de los clusters, ya que los algoritmos jerárquicos tradicionales, basados en el centroide, tienden a repartir las instancias homogeneizando el tamaño de los grupos, mientras que ROCK es inmune a este efecto.

Capítulo 4

Breve Estado del Arte

El estudio del estado del arte, en cualquier disciplina, es fundamental para conocer el estado actual de la investigación en ese área. Debe recopilar y realizar un análisis crítico de los principales trabajos realizados hasta el momento para ofrecer una visión global de la investigación en el tema considerado.

Con el fin de realizar el trabajo de búsqueda bibliográfica de manera metódica, previamente debe establecerse una metodología. De esa forma, se acota el espacio de búsqueda, y se centra mejor cuales son los aspectos relevantes para esta investigación. Así se pretende evitar que la cantidad de información obtenida pueda dificultar la tarea de búsqueda y recopilación, o que ésta derive hacia aspectos de menor relevancia.

Una vez se haya completado la etapa de recogida de información, se realizará un análisis detallado, con el fin de organizarla y presentarla de una manera estructurada.

4.1. Búsqueda bibliográfica

La metodología empleada consiste en identificar las principales fuentes de información que deben ser consultadas, para no pasar por alto ninguna que pueda ser relevante. Posteriormente, se definen los términos de búsqueda que, a priori pueden proporcionar mayor cantidad de información.

Debido a la amplia implantación de Internet en el ámbito científico, la búsqueda se limitará a los recursos disponibles a través de la web. Si se encontrara alguna referencia a una fuente central para el trabajo que no estuviera accesible en la web, se procederá a buscarla por otros medios.

Para acotar el espacio de búsqueda, y dado que no se pretende hacer un estado del arte exhaustivo, se priorizan los resultados más recientes (últimos 5 años: 2007–presente), al menos en las fuentes más especializadas. De esta manera, se determinarán las líneas de investigación más actuales. En las fuentes más generalistas, se puede hacer una búsqueda sin acotación para localizar posibles referencias básicas o conocimiento previo establecido. La lista de fuentes utilizada para este estudio incluye el ISI WoK (Web of Knowledge) [61], Scopus [18], IEEE xplora [29], Google Scholar [21], Wikipedia [74] y Google [20].

Los términos de búsqueda a considerar en este trabajo deben vincular las técnicas de inteligencia computacional con la demencia, y más concretamente

con la enfermedad de Alzheimer, como por ejemplo: disease model*, alzheimer disease model*, machine learning dementia, machine learning Alzheimer database, computational intelligence dementia, computation dementia, clustering dementia, artificial neural networks dementia / ANN dementia, support vector machines dementia / SVM dementia, softcomputing dementia, fuzzy logic dementia / fuzzy systems dementia.

Es posible que, en función de los resultados obtenidos, haya que hacer algún ajuste para refinar la búsqueda, exigiendo que los términos aparezcan en determinados campos, como las palabras clave, el título, etc. En otras ocasiones habrá que ordenar los resultados obtenidos en función de la fecha de publicación, el número de citas, o la relevancia, para poder discriminar entre todos los resultados obtenidos. El criterio de “Relevancia” (*Relevance*), según la definición proporcionada por la WoK, ordena las referencias basándose en una clasificación que considera cuántos de los términos de búsqueda se encuentran en cada campo. Las referencias con mayor valoración aparecen al principio de la lista. Los campos que se consideran para esta clasificación son el título, las palabras clave (*Keywords*), el resumen y las *Keywords Plus*. Los campos “Título” y “Palabras Clave” tienen un peso ligeramente mayor que los campos “Resumen” y “Keywords Plus”.

4.2. Análisis del estado del arte

Tradicionalmente, las estrategias para generar conocimiento en medicina se han basado en observaciones realizadas en entornos clínicos o de investigación y más recientemente, el desarrollo de modelos patofisiológicos basados en biología molecular. Aunque dichas estrategias son sumamente importantes, limitan de alguna manera la tasa de generación de nuevas hipótesis. Las técnicas de aprendizaje automático y minería de datos son aproximaciones alternativas, útiles para identificar nuevas vías que perseguir, como ya es evidente en la literatura [52]. En este caso vamos a analizar qué técnicas se están empleando en la actualidad y con qué fines, para determinar si hay algún estudio previo que sea aplicable al problema objeto de este trabajo o, en caso contrario, proponer algún método que pueda proporcionar una solución válida al problema.

Es necesario incluir en este punto una nota aclaratoria. Al realizar las búsquedas entre las fuentes bibliográficas empleando términos como modelo o modelado, surgen una gran cantidad de referencias relacionadas con “modelos animales”. Estas publicaciones se refieren a experimentos realizados con animales de laboratorio, principalmente ratones, monos, perros, etc. a los que se les induce algún tipo de demencia para estudiar su evolución o para probar nuevos fármacos. Esto cae claramente fuera del ámbito de este estudio y todas esas referencias no han sido tenidas en consideración.

La Figura 4.1 muestra el número de publicaciones y citas por año en esta materia. Se puede observar que aunque el número de publicaciones tenga altibajos, hay una tendencia creciente en el número de citaciones, lo que puede suponer, según el autor, que esta línea de investigación está asentada y despierta un interés cada vez mayor entre los investigadores.

Las referencias encontradas se pueden separar en dos grandes grupos: métodos Supervisados y No Supervisados, como se explicó en la Sección 3.1. A partir de estos dos grandes grupos, se puede seguir subdividiendo y profundizando,

4.2 Análisis del estado del arte

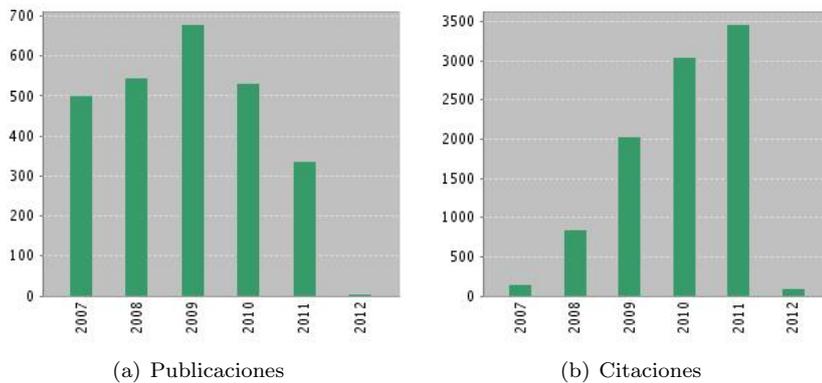


Figura 4.1: Número de publicaciones y sus correspondientes citaciones, por año, relacionadas con IC y demencia [61].

obteniendo una taxonomía como la que se propone en la Figura 4.2. En el caso de los métodos supervisados, se ha discriminado entre redes bayesianas, redes neuronales artificiales, máquinas de vectores soporte, árboles de decisión y agentes. Para los métodos no supervisados, la inmensa mayoría de las referencias están relacionadas con técnicas de clustering en alguna de sus variantes, por lo que es el único método considerado.

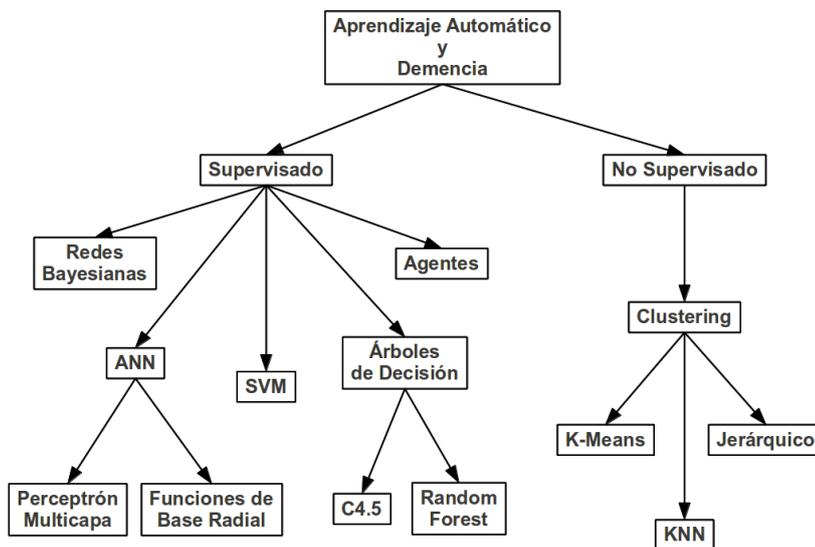


Figura 4.2: Taxonomía de la Inteligencia Computacional aplicada a la Demencia.

De la misma manera que se han agrupado las múltiples técnicas de Inteligencia Computacional en estas pocas categorías, el autor de este trabajo, ha agrupado las situaciones en las que han sido aplicadas, de acuerdo a la siguiente leyenda:

- DNP: Diagnóstico Neuropsicológico. Es un análisis funcional del cerebro

que proporciona información acerca de las áreas cerebrales alteradas, las funciones básicas afectadas, así como las repercusiones de estas alteraciones para la vida cotidiana, personal y laboral del paciente. Para la evaluación neuropsicológica se utilizan algunas pruebas especialmente diseñadas para evaluar el funcionamiento y la maduración del cerebro.

- DF: Diagnóstico Físico. Es el estudio y la práctica del conjunto de reglas y maniobras del interrogatorio clínico y exploración física de que se vale el médico ante el enfermo para valorar la integridad orgánica y funcional de los distintos órganos, aparatos y sistemas.
- DI: Diagnóstico físico mediante Imagen. Es equivalente al caso anterior, pero centrado en las imágenes médicas del cerebro. Dada la importancia de esta técnica para el diagnóstico del Alzheimer, se ha establecido como una categoría independiente.
- TTE: Tasas de Transición entre Estados. Estudia la evolución de la enfermedad y qué circunstancias intervienen para que la enfermedad evolucione y pase de un estado al siguiente.

Las Figuras 4.3 y 4.4, resumen gráficamente la información recopilada para este estudio del estado del arte. La primera Figura representa el número de publicaciones encontradas que hagan referencia a aplicaciones de la Inteligencia Computacional en el campo de la demencia. La segunda Figura muestra el número de citas recibidas por estos trabajos, y sirve a modo de indicador de la relevancia de los mismos.

El campo en el que mayor implantación han tenido las técnicas de inteligencia computacional es en el diagnóstico basado en imágenes de resonancia magnética (MRI, *Magnetic Resonance Imaging*), y entre las técnicas, la más empleada es SVM o máquinas de vectores soporte [16] [41] [70] [48]. Es especialmente interesante el estudio desarrollado por Kloepfel et al. en 2008 [36] ya que describe el éxito obtenido al emplear SVM para el diagnóstico diferencial entre sujetos con enfermedad de Alzheimer y sujetos de control, ancianos pero sanos. Además, demuestra que el método es robusto y puede ser generalizado, ya que un sistema entrenado con datos pertenecientes a pacientes de un centro, puede ser empleado para diagnosticar pacientes de otro centro distinto, lo cual tiene una gran relevancia para su futura aplicación en la práctica clínica como método de diagnóstico computerizado. Por otra parte, Vemuri et al. [68] concluye que la inclusión de información adicional, como datos demográficos o información genética, en el algoritmo de clasificación mejora ligeramente la precisión del diagnóstico.

La misma técnica, SVM, también ha sido empleada con éxito en el diagnóstico neuropsicológico [77] [51] [39] y en el estudio de tasas de transición de estados [25] [64]. En este último caso, Plant et al. [56] aplica SVM y estadísticos Bayesianos para identificar regiones del cerebro cuyo estado (visto a través de MRI) sirva para predecir el paso de MCI (*Mild Cognitive Impairment*) a demencia tipo Alzheimer. Una de las conclusiones de este estudio es que las nuevas técnicas de reconocimiento de patrones multivariable alcanzan una precisión relevante clínicamente para la predicción a priori del paso de demencia leve a Alzheimer.

Por último, dentro de las aplicaciones de SVM para el diagnóstico físico [40] [9], cabe mencionar otro trabajo de Kloepfel et al. [35] en el que compara

la precisión en el diagnóstico del Alzheimer entre varios radiólogos, con distintos grados de experiencia, y el método computerizado. Según este estudio, el método basado en SVM alcanza precisiones comparables con las de los radiólogos, pero con la ventaja de que no requiere de conocimiento experto y una vez entrenado, puede ser aplicado en diferentes centros.

Como puede apreciarse en las Figuras 4.3 y 4.4, el segundo método en cuanto a relevancia son las redes neuronales artificiales, principalmente aplicadas al diagnóstico neuropsicológico [8] [32] [38]. El trabajo más relevante es el de Pazzani et al. [53] que versa sobre la aceptación entre los expertos médicos de las reglas diagnósticas generadas mediante aprendizaje automático. Lo que el autor propone es el uso de diversas técnicas de minería de datos para descubrir patrones que puedan ser usados como apoyo en la toma de decisiones. Los expertos, en general, estaban dispuestos a emplear estas ayudas, con mayor o menor aceptación dependiendo de la técnica de minería de datos empleada para su desarrollo, y siempre que sean consistentes con el conocimiento médico existente.

El segundo grupo de investigaciones está relacionado con la aplicación de ANN al diagnóstico basado en imágenes [65] [76] [6], destacando el trabajo de Horn et al. [27]. En él, se describe la aplicación de diversas técnicas, tanto lineales como no lineales, al diagnóstico diferencial entre Alzheimer y demencia frontotemporal (FTD), a partir de imágenes SPECT procedentes de varios centros. Los resultados obtenidos por el clasificador fueron convincentes, y el método podría ayudar a los médicos en su práctica diaria, mostrándose especialmente útil cuando el diagnóstico visual es inconcluyente o cuando se trata con imágenes procedentes de otros centros diferentes al de trabajo del experto.

Este mismo estudio [27] tiene una peculiaridad, que nos sirve para enlazar con el siguiente grupo de investigaciones, basadas en clustering [69] [34] [2]. Aunque los mejores resultados se obtuvieron aplicando SVM, el autor propone como mejor solución el método basado en K-NN (*k-nearest neighbours*), que es un algoritmo de clustering. La razón para escoger éste frente al otro es que obtuvo mejor valoración al realizar la validación cruzada.

No obstante, el trabajo más relevante en el campo del diagnóstico utilizando algoritmos de clustering es el de Kovacs et al. [37]. Este trabajo defiende la necesidad de crear perfiles diagnósticos altamente personalizados, obtenidos de complementar biomarcadores asociados con procesos patogenéticos con neuroimagen y análisis genéticos. La definición de clusters de pacientes basados en los patrones de los depósitos de proteínas y la detección de proteínas modificadas, pueden tener un alto valor predictivo en los pronósticos.

En el campo de la aplicación de los árboles de decisión a la demencia, destaca el trabajo de Shankle et al. [66], que pese a no ser actual, es muy relevante por ser uno de los primeros trabajos en aplicar métodos de aprendizaje automático, no sólo a la demencia, sino en la medicina en general. Previamente, los sistemas expertos en el campo de la medicina se basaban en reglas definidas por expertos, a diferencia de los métodos de aprendizaje automático que están basados en los datos históricos almacenados. El trabajo de Shankle, mediante la aplicación de diversos algoritmos de ML, consiguió identificar un conjunto de variables, extremadamente sencillo pero preciso, para diferenciar entre enfermos de demencia leve y sujetos sanos.

Entre los trabajos más recientes, el de Maroco et al. [44] destaca por aplicar diversas técnicas de minería de datos (ANN, SVM, Árboles de Decisión y Ran-

Capítulo 4 : Breve Estado del Arte

dom Forest), para la predicción de la Demencia entre personas con incapacidad cognitiva leve (MCI). Aunque SVM obtuvo la mayor precisión, considerando otros indicadores como la especificidad y la sensibilidad, los autores eligieron random forest como la mejor solución para predecir la demencia a partir de varios test neuropsicológicos.

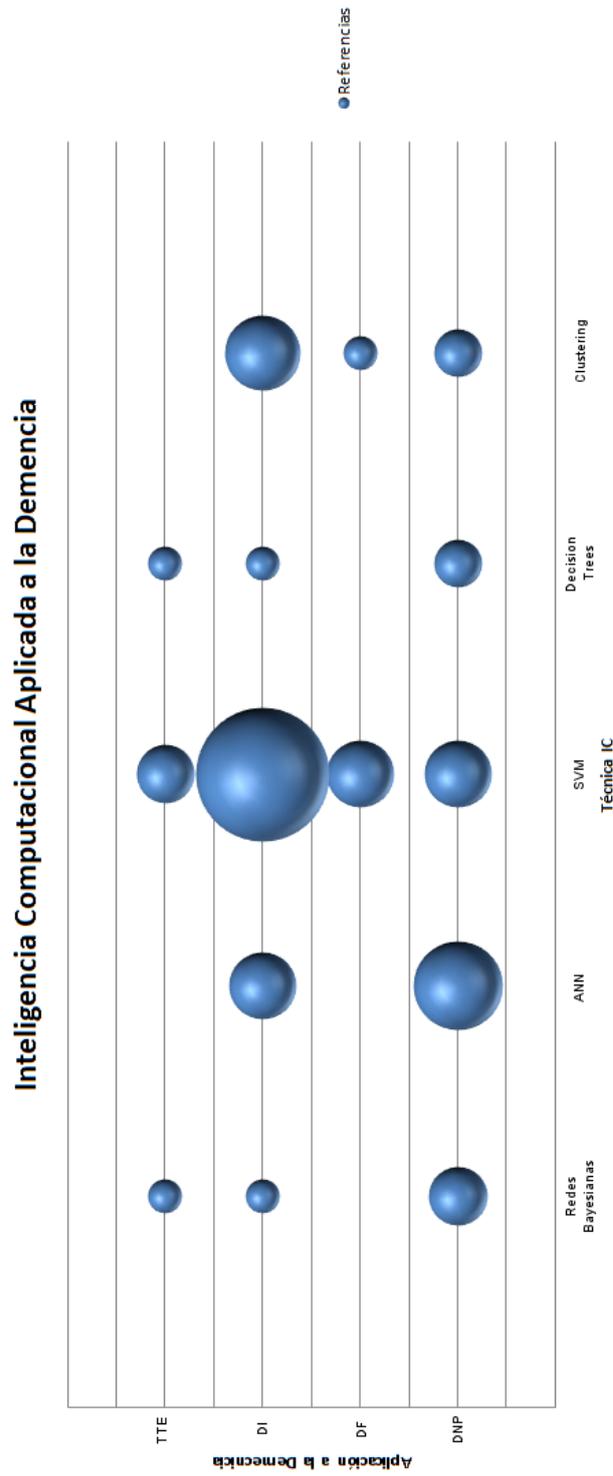


Figura 4.3: Número de referencias encontradas con relación a técnicas de Inteligencia Computacional aplicadas a la Demencia.

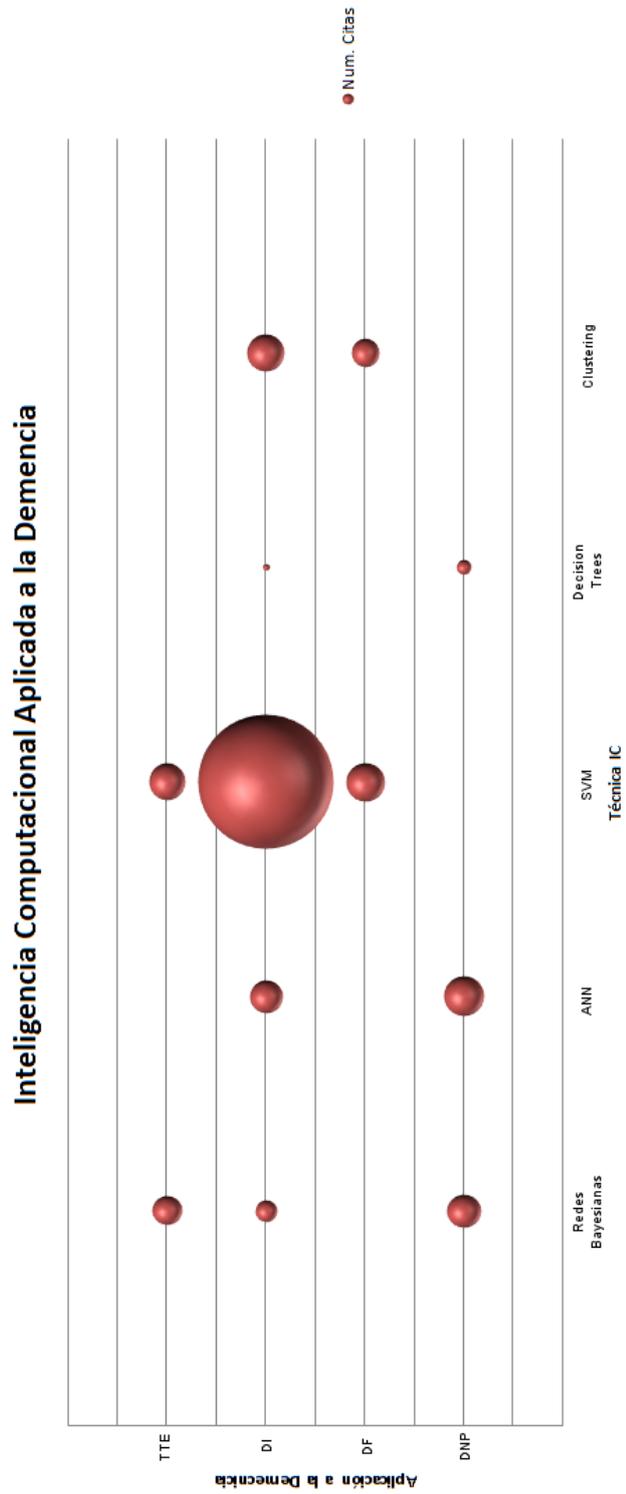


Figura 4.4: Número de citas a las referencias encontradas con relación a técnicas de Inteligencia Computacional aplicadas a la Demencia.

Parte II

Inteligencia Computacional aplicada al diagnóstico diferencial entre Demencia Vascular y Demencia tipo Alzheimer

Capítulo 5

Caso de aplicación

En este capítulo se explica el caso real relacionado con la Demencia al que se van a aplicar las técnicas estudiadas en la primera parte de este trabajo. Lo primero será justificar la necesidad de solucionar el problema planteado. Para completar el capítulo, se incluye una descripción de algunas de las variables que componen el conjunto de datos con el que se va a trabajar en los siguientes capítulos.

5.1. Revisión de diagnósticos

De lo expuesto en la Tabla 2.1 de la página 24, hay que destacar que en 2011 el *National Institute on Ageing* (NIA) y la *Alzheimer's Association* recomendaron nuevos criterios y guías para el diagnóstico de la Enfermedad de Alzheimer. Eso significa que la forma de diagnosticar las demencias todavía está en desarrollo y no hay un método establecido, contrastado y globalmente aceptado. Por eso es necesario que los terapeutas hagan un seguimiento de los pacientes y en ocasiones tengan que modificar el diagnóstico inicial, con la siguiente incidencia:

- Cerca del 30 % de los pacientes fueron re-diagnosticados
- La revisión de los diagnósticos dio lugar a más diagnósticos de los tipos:
 - Demencia vascular y mixta.
 - Enfermedad con cuerpos de Lewy

La Figura 5.1 muestra de forma gráfica estos datos. Una mejora en las tasas de acierto de los diagnósticos iniciales, podría suponer ganar tiempo para los tratamientos adecuados, y por tanto, retrasar la evolución de la enfermedad mejorando la calidad de vida del paciente y la de sus familiares. Los investigadores creen que los futuros tratamientos para frenar o detener el avance de la enfermedad y preservar las funciones cerebrales (denominados tratamientos “modificadores de la enfermedad”) serán más efectivos cuando se administren durante las fases de demencia leve (MCI) o pre-clínicas de la enfermedad [4].

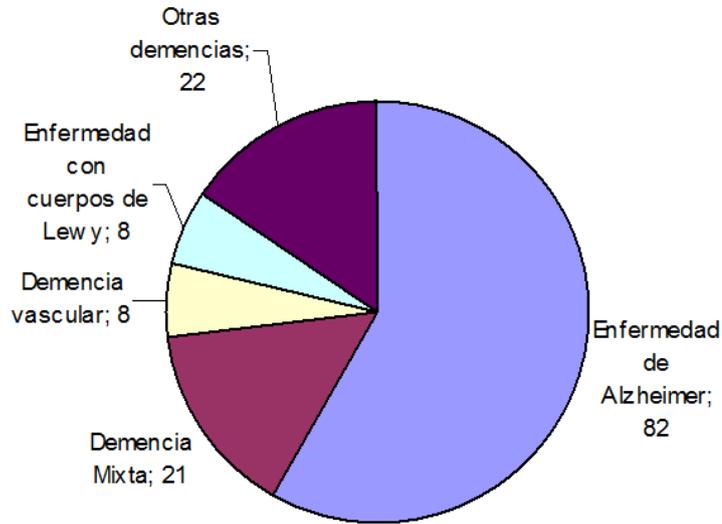


Figura 5.1: Diagnósticos.

5.2. Características de los casos diagnosticados

A continuación se muestran las características de los casos diagnosticados como demencia de diversos tipos en el CRE Alzheimer, según la información proporcionada por el propio centro. En el Apéndice A, la Tabla A.1 muestra la descripción detallada del total de las variables disponibles para los experimentos.

Datos demográficos:

- 141 usuarios hasta finales de 2010
 - 97 Mujeres (68,8 %)
 - 44 Hombres (31,2 %)
- Edad media = 79,2 años (d.t.=8,8)
 - Mujeres = 79,84 (d.t.=8,3)
 - Hombres = 77,86 (d.t.=9,7)
- Rango edades = 47-93 años
- El grupo más numeroso de usuarios se sitúa entre los 80 y 85 años de edad

La mayor proporción de mujeres que padecen Alzheimer puede deberse a su mayor esperanza de vida frente a los hombres. Varios estudios que tienen en cuenta la incidencia de la edad, no han encontrado diferencias significativas debidas al género. Por tanto, las mujeres no son más propensas a desarrollar Demencia a ninguna edad concreta [4].

La Figura 5.4 muestra la distribución del nivel educativo de los casos proporcionados por el CRE-A. Las cifras concretas son: Media de años de escolarización = 6,93 (d.t. = 4,39)

5.2 Características de los casos diagnosticados

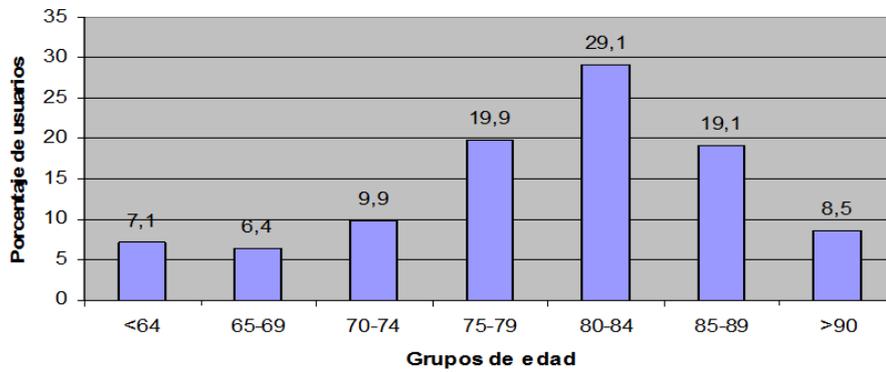


Figura 5.2: Distribución por grupos de edad.

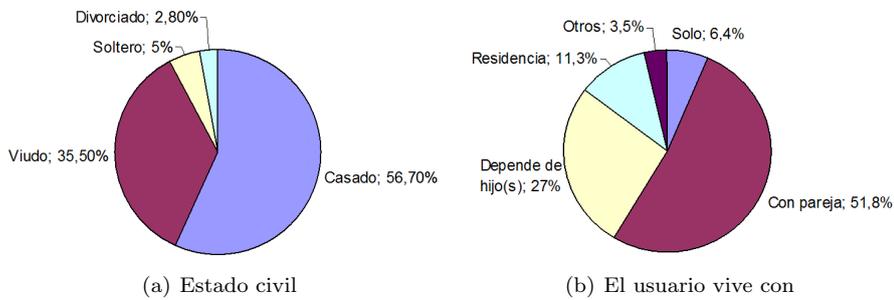


Figura 5.3: Estado civil y situación familiar de los usuarios.

La actividad laboral más habitual entre las mujeres es ama de casa (76,3 %) mientras que entre los hombres es la de obrero especializado (43,2 %), según se muestra en la Figura 5.5.

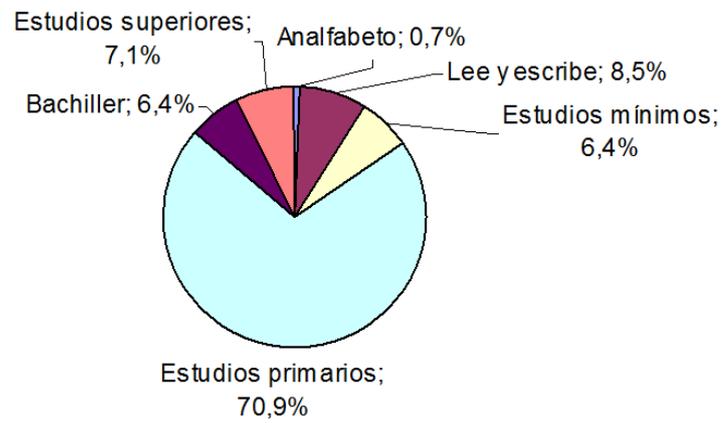


Figura 5.4: Nivel educativo.

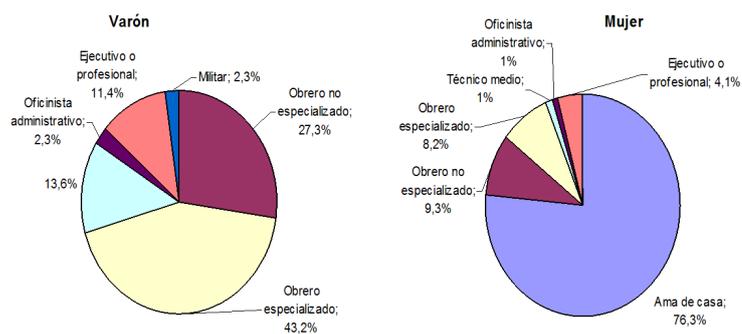


Figura 5.5: Profesión.

Capítulo 6

Metodología

En este capítulo se va a detallar la metodología empleada para la realización de los experimentos. Inicialmente, se describe el pre-procesamiento de los datos, para adecualos a las operaciones de las siguientes etapas de la metodología. A continuación se explica el método de selección de características. Posteriormente se describen las técnicas de Inteligencia Computacional que se emplearán en la experimentación y la forma de evaluar los resultados que se obtengan. En sucesivos capítulos se mostrarán los resultados concretos y las conclusiones correspondientes.

La Figura 6.1 resume, de manera gráfica, la metodología explicada. Como se puede ver, es muy sencilla y está dividida en etapas claramente diferenciadas.

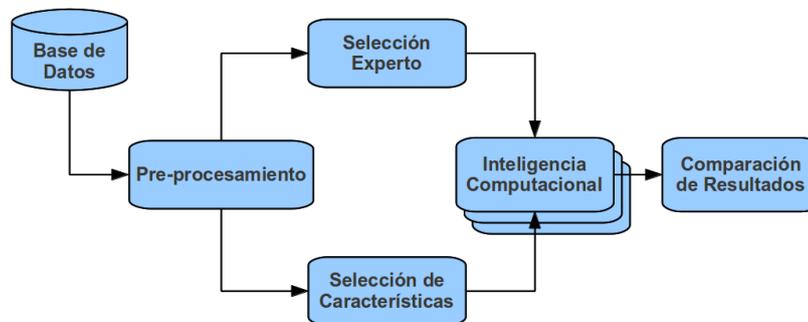


Figura 6.1: Metodología.

6.1. Pre-procesamiento de datos

Los datos disponibles para la realización de este trabajo, así como los diagnósticos, han sido suministrados por el Centro de Referencia Estatal de Atención a Personas con Enfermedad de Alzheimer y otras Demencias.

6.1.1. Adquisición de datos

El formato original en el que se encontraban los datos era el del paquete estadístico SPSS. Para la realización de los experimentos relacionados con Inteligencia Computacional, está previsto utilizar un conjunto de herramientas que incluyen Weka, Orange, RapidMiner, GUAJE o Matlab. Para evitar problemas de compatibilidad entre formatos y facilitar el posterior trabajo con los datos, lo primero es exportarlos a CSV (*comma separated values* o valores separados por comas), un formato estándar que todas estas herramientas pueden manejar sin problemas.

El conjunto de datos completo consta de 141 casos, y cada caso está formado por 121 variables. Pero no todas las variables suministradas se emplearán para la experimentación, ya que algunos datos son de carácter administrativo, para el funcionamiento del centro, pero irrelevantes para el objetivo de este trabajo. Una vez eliminadas, de forma manual, estas variables el conjunto de datos se reduce a 141 casos y 105 variables.

6.1.2. Inducción de particiones

Para la realización de esta operación, y ya que no se trata de una parte fundamental del trabajo, utilizaremos la herramienta GUAJE [3] para que calcule el número óptimo de particiones para cada variable. GUAJE genera 3 tipos diferentes de particiones difusas: *HFP* [23] [24], *Regular* (todas las particiones son iguales) y *Kmeans* [26], con un rango de etiquetas lingüísticas entre [2, 9], para cada una de ellas. Todas estas combinaciones se evalúan a través del Coeficiente de la Partición (PC) [10], la Entropía de la Partición (PE) [10] y el Índice de Chen [14]. Aquella que obtiene mejor calificación en el conjunto de los índices, es seleccionada como la mejor partición.

6.2. Selección de características

Cuando el número de características es excesivamente grande y puede comprometer la correcta ejecución de los algoritmos, ya sea por necesidades de cálculo o por restricciones de tiempo, se puede hacer una pre-selección de características. De esta forma, y sin necesidad de ser demasiado estrictos, se pueden eliminar aquellas variables menos significativas.

6.2.1. Selección experta

Para el caso del diagnóstico diferencial entre Alzheimer y Demencia Vascular, ya se dispone de una primera selección de características, suministrada por un experto en el dominio. Se compone de 26 variables, de las cuales 9 son de especial relevancia, de acuerdo con las indicaciones del experto. Esta selección se detallará en el Capítulo de Resultados (Cap. 7).

6.2.2. Recuento Borda

Tener una selección de características realizada por un experto, es algo muy valioso cuando se trata de un problema real. Por ese motivo, se va a realizar otra selección de características, mediante el método del recuento Borda. Con ello se

6.3 Métodos empleados para la experimentación

pretende confrontar el conjunto de variables seleccionado con el que suelen usar habitualmente los expertos para el diagnóstico diferencial.

El recuento Borda es un proceso electoral propuesto por el matemático francés Jean-Charles Borda en 1770. Según esta metodología, los candidatos son ordenados de acuerdo a las preferencias de cada elector. En el recuento, a cada posición de la ordenación se le atribuye una puntuación: 1 punto para el último clasificado, 2 puntos para el penúltimo, 3 para el antepenúltimo, y así sucesivamente. Este sistema tiene en consideración, no sólo la primera opción de cada elector, sino también las otras, de modo que no siempre el candidato colocado más veces en primer lugar es el vencedor.

En este caso, al no tener una votación u ordenación inicial, se emplearán métodos de scoring, como *Relief*, *Information Gain*, *Gain Ratio* y *Gini Gain*, ya que permiten ordenar y filtrar atributos por su relevancia. Posteriormente se utilizará el recuento Borda para ponderar qué variables son las más relevantes de acuerdo a las ordenaciones generadas por los métodos mencionados.

6.3. Métodos empleados para la experimentación

Para llevar a cabo la experimentación, se emplearán los métodos de Inteligencia Computacional vistos en la Sección 3.3, en concreto:

- Árboles de decisión:
 - J4.8, que es una evolución del método C4.5, previamente comentado en la Sección 3.3.
 - Random Forest, que pese a no ser considerado un método de Inteligencia Computacional, se va a incluir en la experimentación por varias razones: es un método basado en árboles de decisión, los datos disponibles son adecuados para este método y porque ha demostrado tener buenos resultados previamente en el diagnóstico de la demencia [44], como se ha explicado en el estado del arte.
- Redes neuronales artificiales:
 - Perceptrón multicapa (MLP).
 - Funciones de base radial (RBF).
- Máquinas de vectores soporte (SVM).
- Clustering:
 - KMeans.
 - Expectation Maximization.
 - Hierarchical Clusterer.
 - ROCK.

La inclusión de los algoritmos de clustering no es estrictamente necesaria, pero dado el interés en probar el funcionamiento y los resultados del método ROCK, se incluyen los demás métodos para tener más elementos de juicio.

Los parámetros de funcionamiento de cada algoritmo se ajustarán de manera individual en cada caso, como se muestra en el Apéndice C, excepto el número de clases o clusters que se establecerá en 3 (correspondientes a la Enfermedad de Alzheimer, Otras Demencias y Demencia Vascular), en aquellos casos donde se pueda especificar.

6.4. Evaluación de resultados

Tras la ejecución de los experimentos, se recopilarán determinados índices que indiquen la calidad de los resultados. Estos valores serán empleados para estimar qué técnicas podrían ser más interesantes para el diagnóstico diferencial entre el Alzheimer y la Demencia Vascular, dadas las características de los datos disponibles.

En concreto se recogerán los siguientes valores:

- CCI: *Correctly Classified Instances*, número de instancias correctamente clasificadas.
- ICI: *Incorrectly Classified Instances*, número de instancias incorrectamente clasificadas.
- MAE: *Mean Absolute Error*, error absoluto medio, que indica cómo de cerca está el pronóstico o la predicción del resultado final.
- RMSE: *Root Mean Squared Error*, error cuadrático medio, similar al anterior pero midiendo la media de los cuadrados de los errores, ya que al elevar al cuadrado se evita que los errores positivos y negativos se anulen.

Para los métodos de clustering sólo se recogerán los dos primeros índices, puesto que los algoritmos no proporcionan la información relativa a los errores.

Una forma sencilla de comprobar de un vistazo el rendimiento de un algoritmo es mediante una matriz de confusión. Cada columna de la matriz representa la clase predicha para cada instancia, mientras que cada fila representa las instancias en su clase real. El nombre de la matriz proviene del hecho de que con esta representación es muy fácil ver si el sistema está confundiendo dos clases. Cuando se trabaja con algoritmos no supervisados, en los que la clase real de las instancias puede no ser conocida, este tipo de representación se denomina *matching matrix*, que aquí he traducido como matriz de emparejamiento. Este método sí que nos permite comparar los resultados obtenidos por todos los algoritmos y servirá para apoyar los análisis que se hagan basados en CCI e ICI.

Capítulo 7

Resultados

En este capítulo se muestran los principales resultados obtenidos al aplicar la metodología detallada en el capítulo anterior. Por cuestiones de claridad, aquí sólo se muestran los datos más relevantes. En el Apéndice C se pueden consultar los resultados completos, así como los parámetros que determinan las condiciones para la experimentación. Una vez analizados los resultados, el siguiente capítulo se expondrán las conclusiones de este trabajo.

7.1. Selección de Características

Las siguientes secciones muestran los detalles de cómo se realizó la selección de características, así como el conjunto de variables resultante en cada caso.

7.1.1. Selección de variables experta

De acuerdo con los criterios del experto, de entre las variables disponibles en la base de datos (ver Tabla A.1 en la página 86), las variables que se explican en esta sección tienen mayor importancia para el diagnóstico diferencial entre Enfermedad tipo Alzheimer (EA) y la Demencia Vasculare (DV). Se muestran a continuación, agrupadas por su origen, para facilitar su comprensión, y resaltadas en color azul aquellas variables que pueden ser más relevantes.

Antecedentes personales

- Sexo: No hay muchas diferencias pero la Enfermedad de Alzheimer puede tener más prevalencia en hombres, mientras que la demencia vascular afecta más a mujeres.
- Presencia de alguno de los siguientes antecedentes personales (factores de riesgo). En principio los factores de riesgo son comunes para ambas enfermedades aunque su prevalencia pueden ser mayor en algunos casos para la DV que para la EA:
 - Hipertensión
 - Infarto Miocardio
 - Angor

Capítulo 7 : Resultados

- Cardiopatía
- **ACVA**: Si ha habido algún episodio previo de accidente cerebrovascular es indicativo de Demencia Vascular
- Diabetes
- Tabaco
- Hipercolesterolemia

Historia del deterioro

Es especialmente importante, ya que puede ser la principal diferencia entre ambos tipos de demencia.

- **Comienzo**: Agudo para la Demencia Vascular, insidioso para la Enfermedad de Alzheimer
- **Evolución**: Fluctuante para la Demencia Vascular, progresiva para la Enfermedad de Alzheimer.

Síntomas actuales

- **Síntomas Focales**: La presencia de síntomas neurológicos focales es propia de la Demencia Vascular.
- **Trastornos de la Marcha**: Los trastornos de la marcha son más habituales en la Demencia Vascular.

Otros síntomas actuales que pueden ser de mayor incidencia en la Demencia Vascular son los siguientes:

- Incontinencia urinaria o fecal
- Depresión

Exploración

- Arritmia, soplos cardíacos y pulsos abolidos: Dado que son patologías vasculares pueden aparecer en DV con mayor frecuencia que en EA.
- **Cuadro Pseudobulbar**: Puede aparecer en la DV, no así en la EA.
- Paresia facial / paraparesia: Pueden aparecer en la DV.
- Bradicinesia: También puede aparecer en la DV.
- **Marcha**: En la DV puede ser marcha apráxica, atáxica o parkinsoniana.

Escalas

- Una puntuación alta en la escala de Hachinski es indicativo de Demencia Vascular¹.

¹En nuestro caso, la base de datos tiene pocos sujetos que tengan puntuaciones en esta escala.

7.1 Selección de Características

7.1.2. Selección de variables mediante recuento Borda

Como se explicó en el Capítulo de Metodología, el conjunto de datos, una vez pre-procesado, se compone de 141 ejemplos o casos, cada uno con 109 características que lo definen (141×109). A este conjunto reducido se le han aplicado diversos métodos de scoring, como *Relief*, *Information Gain*, *Gain Ratio* y *Gini Gain*, que permiten ordenar y filtrar atributos por su relevancia. La Tabla 7.1 muestra el orden de las variables asignado por cada método de ranking y el resultado del recuento Borda de todos los criterios. La tabla completa se muestra en el Apéndice A (Tabla A.2).

Tabla 7.1: Ranking de las características, según distintos métodos de evaluación.

#	Relief	InfoGain	GainRatio	GiniGain	Recuento Borda
1	SA_Desorientacion	Año_Inicio	MAtaxica	Año_Inicio	Año_Inicio
2	SA_TIntelectuales	H_SintInicial	CPseudobulbar	H_SintInicial	H_SintInicial
3	AP_ACVA	H_Evolucion	PiramBilateral	H_Evolucion	H_Evolucion
4	Año_Inicio	Años_escol	Paraparesia	SA_Desorientacion	SA_Desorientacion
5	H_SintInicial	H_MesesEvolucion	SA_SintFocales	MMSE	SA_TIntelectuales
6	Estado_civil	H_Comienzo	ParesiaF	SA_TIntelectuales	AP_ACVA
7	H_MesesEvolucion	SA_SintFocales	SA_Hipersex	H_MesesEvolucion	MAtaxica
8	Profesion	MMSE	TembloReposo	Años_escol	H_Comienzo
9	H_Evolucion	GDS	SA_TMemoria	H_Comienzo	SA_SintFocales
10	AP_Tabaco	AP_ACVA	SA_Bulimia	SA_SintFocales	H_MesesEvolucion
11	GDS	SA_Desorientacion	SA_UrgMiccional	GDS	CPseudobulbar
12	H_Comienzo	SA_TIntelectuales	Espasticidad	Escolarizacion	Años_escol
13	SA_Agitacion	CPseudobulbar	H_SintInicial	SA_Hipersex	SA_Hipersex
14	AP_Hipertension	Escolarizacion	AP_Hepatopatía	AP_ACVA	GDS
15	Años_escol	ParesiaF	H_Evolucion	Motivo_Consulta	TembloReposo
16	SA_Hipersex	Blessed	H_Comienzo	Edad_grupos	MMSE
17	Edad_grupos	Estado_civil	AP_ACVA	TembloReposo	Estado_civil
18	Preñion	Espasticidad	PulsosAbolidos	ParesiaF	Paraparesia
19	Sexo	Edad_grupos	AF_Huntington	SA_Temblo	PiramBilateral
20	AP_Trauma	MAtaxica	SA_Desorientacion	CPseudobulbar	ParesiaF
.
.
108	AP_Otros	AP_HIV	AP_HIV	AP_HIV	AP_Otros
109	NPI_Q	AF_Down	AF_Down	AF_Down	NPI_Q

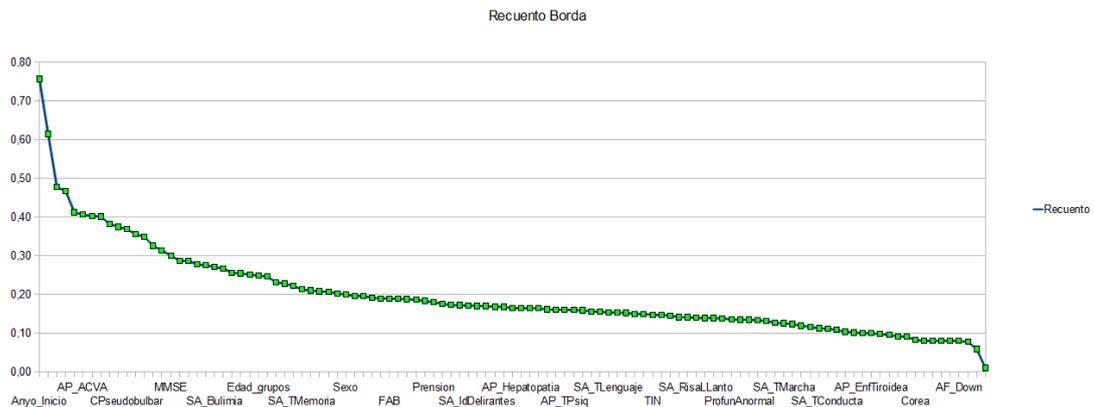


Figura 7.1: Representación del ranking de características, según el valor asignado por el recuento Borda.

Con el ranking obtenido (ver Tabla 7.1 y Tabla A.2 para la tabla completa) y de acuerdo con los valores que se muestran en la Figura 7.1, se puede apreciar que a partir de determinada variable, los valores decrecen muy lentamente y de

manera casi lineal. Es decir, el “codo” de la gráfica se encuentra aproximadamente en la variable número 22: **Esplasticidad**. Por lo que consideraremos como conjunto de variables “Borda” las 22 primeras del ranking (desde **Año_Inicio** hasta **Esplasticidad**).

7.2. Inducción de particiones

Para realizar la inducción de particiones de las variables de entrada, tal como se explicó en la Sección 6.1.2 se empleó la herramienta GUAJE. La Tabla 7.2 muestra los parámetros específicos de funcionamiento para HFP. En este caso, se dejaron los parámetros por defecto que aparecen en dicha tabla.

Tabla 7.2: Parámetros de configuración del método de inducción de particiones difusas HFP.

Parámetro	Valor
Distancia	Numérica
Distancia simplificada	false
Número de FP para distancia	-1
Número de FP para varianza	10
Criterio de fusión	distancia
Penalización de heterogeneidad	0,0
Proporción de Heterogeneidad	0,5

Esta herramienta nos proporciona tablas como las mostradas en el Apéndice B, que detallan la evaluación de todas las particiones difusas generadas para cada caso concreto. Se considera la mejor partición, aquella que minimiza la Entropía de la Partición (*PE*) [10] y maximiza el Coeficiente de la Partición (*PC*) [10] y el Índice de Chen [14]. En la tabla de resultados se muestra en color azul el mejor valor para cada índice, las casillas con fondo amarillo indican el mejor índice para cada partición y las que tienen fondo amarillo y los números en color rojo remarcan los mejores índices comparando todas las particiones.

Tabla 7.3: Resumen de los resultados obtenidos con GUAJE, para la inducción de particiones de las variables de entrada.

Variable	# Part.	Índ. Chen (máx)	Método	Cuantiles
Hachinski	3	0.80	Kmeans	(2, 6.5, 14.5)
Año_Inicio	3	0.75	Kmeans	(57, 71, 81)
H_Meses_Evolución	3	0.77	Kmeans	(44, 92, 160)
Años_Escol	3	0.79	Kmeans	(4.5, 8, 17)

La Tabla 7.3 muestra el resumen de los resultados obtenidos con GUAJE en la inducción de particiones de las variables de entrada. Esta herramienta genera automáticamente todas las combinaciones posibles entre los tres métodos (HFP [23] [24], Regular y Kmeans [26]) y las particiones posibles en el intervalo [2, 9]. Como se ha explicado anteriormente, de la comparación entre los distintos

casos y según los valores de los índices y coeficientes calculados, se obtiene el número óptimo de particiones para cada variable, que se muestra en la columna correspondiente a “# Part.” y el método empleado para generarla, columna “Método”. El valor del índice de Chen calculado se muestra en la columna “Índ. Chen”, teniendo en cuenta que los mejores valores son los más altos y el valor del índice se encuentra acotado entre $[0, 1]$.

Los datos correspondientes a la columna “Cuantiles” se han obtenido directamente de las gráficas asociadas a las particiones, como las mostradas en las figuras del Apéndice B. Todas las funciones de pertenencia obtenidas son triangulares y las particiones completas, es decir, se cubre todo el rango de valores de las características.

7.3. Algoritmos de Inteligencia Computacional

Esta Sección muestra la comparación de los resultados obtenidos por distintos métodos de clasificación, utilizando primero el conjunto de características propuestas por el experto en el área de conocimiento y después el conjunto seleccionado de manera automática. En el Apéndice C se pueden consultar las condiciones en las que se han realizado los experimentos, a través de los parámetros de configuración de los algoritmos empleados, así como los resultados completos de la ejecución de los mismos.

7.3.1. Árboles de Decisión

Utilizamos los árboles de decisión como técnica base con la que comparar el resto de resultados que se obtengan. Como se ha visto en la Sección 3.3, los árboles de decisión son un método muy difundido y ampliamente utilizado. En muchas ocasiones sirven como base para determinar cómo de bueno es otro algoritmo, puesto que si no aporta una mejora significativa para el caso concreto de estudio, no será tenido en consideración porque con toda probabilidad supondrá un mayor esfuerzo computacional que no compensará dicha mejora.

J4.8

A continuación se muestra el árbol de decisión obtenido mediante el algoritmo J4.8, con el conjunto de características seleccionadas por el experto, inducción de particiones y las variedades de demencia reducidas a tres clases. Como se puede ver, es un árbol bastante sencillo, con sólo 19 nodos, de los cuales 10 son hojas o terminales, y una profundidad de 5.

J48 pruned tree

```

-----
Hachinski=alto <= 0
|   SA_SintFocales <= 0
|   |   H_Evolucion=Fluctuante <= 0
|   |   |   H_Comienzo=Insidioso <= 0: OtrasDem (4.0/1.0)
|   |   |   H_Comienzo=Insidioso > 0
|   |   |   |   Hachinski=medio <= 0: EA (100.0/24.0)
|   |   |   |   Hachinski=medio > 0
|   |   |   |   |   AP_Hipercolest <= 0: EA (2.0)

```

Capítulo 7 : Resultados

```

| | | | | AP_Hipercolest > 0: OtrasDem (3.0)
| | H_Evolucion=Fluctuante > 0
| | | AP_Angor <= 0: OtrasDem (11.0/1.0)
| | | AP_Angor > 0: EA (3.0/1.0)
| SA_SintFocales > 0: OtrasDem (7.0)
Hachinski=alto > 0
| Bradicinesia <= 0
| | CPseudobulbar <= 0: OtrasDem (5.0)
| | CPseudobulbar > 0: DV (4.0/1.0)
| Bradicinesia > 0: DV (2.0)

```

Number of Leaves : 10

Size of the tree : 19

La Tabla 7.4 muestra resumida la información sobre la evaluación del árbol generado. Los conceptos mostrados son el número de instancias correctamente clasificadas (CCI), el número de instancias incorrectamente clasificadas (ICI), el estadístico *Kappa* que mide el acuerdo entre dos clasificaciones de N elementos en C categorías mutuamente excluyentes, el error absoluto medio (MAE) que indica cómo de cerca está el pronóstico o la predicción del resultado final, el error cuadrático medio (RMSE), y otra serie de medidas relativas del error.

Tabla 7.4: Evaluación de resultados: J4.8 (Experto).

Concepto	Valor
Correctly Classified Instances (CCI)	69.5035 %
Incorrectly Classified Instances (ICI)	30.4965 %
Kappa statistic	0.3749
Mean absolute error (MAE)	0.2584
Root mean squared error (RMSE)	0.3962
Relative absolute error	73.5652 %
Root relative squared error	94.8056 %

Una forma sencilla de comprobar de un vistazo el rendimiento de un algoritmo es mediante una matriz de confusión, como la mostrada en la Tabla 7.5. Cada columna de la matriz representa la clase predicha para cada instancia, mientras que cada fila representa las instancias en su clase real. El nombre de la matriz proviene del hecho de que con esta representación es muy fácil ver si el sistema está confundiendo dos clases. Cuando se trabaja con algoritmos no supervisados, en los que la clase real de las instancias puede no ser conocida, este tipo de representación se denomina *matching matrix*, que aquí he traducido como matriz de emparejamiento, como se puede ver en los resultados correspondientes a los algoritmos de clustering.

Si se repite el experimento, pero utilizando esta vez el conjunto de características seleccionadas mediante recuento Borda, el árbol generado mediante J4.8 es el que aparece a continuación. Es similar al anterior en cuanto a su tamaño y profundidad, pero radicalmente distinto en cuanto a los criterios de decisión en cada nodo.

J48 pruned tree

7.3 Algoritmos de Inteligencia Computacional

Tabla 7.5: Matriz de confusión: J4.8 (Experto).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	75	7	0	0
Otras Demencias	29	20	3	0
Demencia Vascular	0	4	3	0

```

SA_SintFocales <= 0
|   SA_Hipersex <= 0
|   |   H_SintInicial=Trastorno_de_Conducta <= 0
|   |   |   AP_ACVA <= 0
|   |   |   |   SA_Desorientacion <= 0: EA (87.0/16.0)
|   |   |   |   SA_Desorientacion > 0
|   |   |   |   |   SA_TIntelectuales <= 0: EA (4.0/1.0)
|   |   |   |   |   SA_TIntelectuales > 0: OtrasDem (6.0/1.0)
|   |   |   |   AP_ACVA > 0
|   |   |   |   |   MMSE=bajo <= 0
|   |   |   |   |   |   Anyos_escol=bajo <= 0: EA (4.0/2.0)
|   |   |   |   |   |   Anyos_escol=bajo > 0: OtrasDem (5.0/1.0)
|   |   |   |   |   |   MMSE=bajo > 0: EA (5.0/1.0)
|   |   |   |   |   H_SintInicial=Trastorno_de_Conducta > 0: OtrasDem (12.0/2.0)
|   |   SA_Hipersex > 0: OtrasDem (6.0)
SA_SintFocales > 0
|   Anyo_Inicio=G1 <= 0: OtrasDem (9.0)
|   Anyo_Inicio=G1 > 0: DV (3.0)

Number of Leaves : 10
Size of the tree : 19

```

La Tabla 7.6 muestra la evaluación de los resultados para este segundo caso, y se puede apreciar que el error cometido es mayor en todos sus parámetros.

Tabla 7.6: Evaluación de resultados: J4.8 (Borda).

Concepto	Valor
Correctly Classified Instances	65.2482 %
Incorrectly Classified Instances	34.7518 %
Kappa statistic	0.2964
Mean absolute error	0.2651
Root mean squared error	0.4311
Relative absolute error	75.4751 %
Root relative squared error	103.1552 %

La matriz de confusión correspondiente es la contenida en la Tabla 7.7, y se puede apreciar cómo la principal fuente del error está entre “Otras demencias” y “Alzheimer”, en ambos sentidos.

Capítulo 7 : Resultados

Tabla 7.7: Matriz de confusión: J4.8 (Borda).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	66	16	0	0
Otras Demencias	26	25	1	0
Demencia Vascular	2	4	1	0

Random Forest

Para completar los resultados, se realizan más pruebas con otros algoritmos dentro de la categoría de los árboles de decisión, aunque como en el caso de Random Forest no se pueda considerar un métodos de IC, y siempre que las características de las variables de entrada sean compatibles con el tipo de algoritmo a emplear. De esa manera, se tendrán más elementos de juicio a la hora de evaluar las soluciones. Los resultados obtenidos mediante el algoritmo **Random Forest**, con el conjunto de características seleccionadas por el experto son los mostrados en la Tabla 7.8, mientras que la matriz de confusión correspondiente es la Tabla 7.9.

Tabla 7.8: Evaluación de resultados: Random Forest (Experto).

Concepto	Valor
Correctly Classified Instances	63.8298 %
Incorrectly Classified Instances	36.1702 %
Kappa statistic	0.2631
Mean absolute error	0.2787
Root mean squared error	0.403
Relative absolute error	79.3292 %
Root relative squared error	96.4351 %

Tabla 7.9: Matriz de confusión: Random Forest (Experto).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	68	14	0	0
Otras Demencias	30	20	2	0
Demencia Vascular	1	4	2	0

La Tabla 7.10 muestra los resultados obtenidos con el algoritmo **Random Forest** y el conjunto de características seleccionadas mediante recuento Borda. La matriz de confusión correspondiente es la Tabla 7.11.

7.3 Algoritmos de Inteligencia Computacional

Tabla 7.10: Evaluación de resultados: Random Forest (Borda).

Concepto	Valor
Correctly Classified Instances	78.0142 %
Incorrectly Classified Instances	21.9858 %
Kappa statistic	0.5548
Mean absolute error	0.2297
Root mean squared error	0.3496
Relative absolute error	65.3998 %
Root relative squared error	83.6572 %

Tabla 7.11: Matriz de confusión: Random Forest (Borda).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	77	5	0	0
Otras Demencias	17	33	2	0
Demencia Vascular	0	7	0	0

7.3.2. Redes Neuronales Artificiales ANN

A continuación se recogen, de manera resumida, los resultados obtenidos mediante los algoritmos de redes neuronales artificiales. Se puede apreciar, que en general son métodos que funcionan muy bien, pero a costa de perder interpretabilidad en los resultados. También el tiempo de entrenamiento suele ser más elevado que con otros métodos, pero una vez entrenadas, funcionan de manera muy eficiente.

MLP

La Tabla 7.12 muestra los resultados obtenidos con el algoritmo *Perceptrón Multicapa* y el conjunto de características seleccionadas por el experto. La matriz de confusión correspondiente es la Tabla 7.13.

Tabla 7.12: Evaluación de resultados: MLP (Experto).

Concepto	Valor
Correctly Classified Instances	58.1560 %
Incorrectly Classified Instances	41.8440 %
Kappa statistic	0.1630
Mean absolute error	0.2841
Root mean squared error	0.4837
Relative absolute error	80.8744 %
Root relative squared error	115.7377 %

Como apunte, mencionar que este algoritmo tarda considerablemente más que el resto de algoritmos en construir el modelo y la precisión no es mejor que

Capítulo 7 : Resultados

en otros casos.

Tabla 7.13: Matriz de confusión: MLP (Experto).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	60	21	1	0
Otras Demencias	29	22	1	0
Demencia Vascular	1	6	0	0

La Tabla 7.14 muestra los resultados obtenidos con el algoritmo MLP y el conjunto de características seleccionadas mediante recuento Borda. La matriz de confusión correspondiente es la Tabla 7.15.

Tabla 7.14: Evaluación de resultados: MLP (Borda).

Concepto	Valor
Correctly Classified Instances	73.7589 %
Incorrectly Classified Instances	26.2411 %
Kappa statistic	0.4890
Mean absolute error	0.1761
Root mean squared error	0.3767
Relative absolute error	50.1298 %
Root relative squared error	90.1285 %

De acuerdo con la Tabla 7.15, la principal fuente de error es la confusión entre las clases “Enfermedad de Alzheimer” y “Otras Demencias”, pero también en sentido contrario.

Tabla 7.15: Matriz de confusión: MLP (Borda).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	67	14	1	0
Otras Demencias	16	35	1	0
Demencia Vascular	1	4	2	0

Para este otro conjunto de datos, la precisión mejora notablemente respecto al caso anterior, pero no es superior a otros algoritmos, y sigue tardando del orden de 10 veces más tiempo en generar el modelo.

RBF

La Tabla 7.16 muestra los resultados obtenidos con el algoritmo **Funciones de Base Radial** y el conjunto de características seleccionadas por el experto. La matriz de confusión correspondiente es la Tabla 7.17.

7.3 Algoritmos de Inteligencia Computacional

Tabla 7.16: Evaluación de resultados: RBF (Experto).

Concepto	Valor
Correctly Classified Instances	67.3759 %
Incorrectly Classified Instances	32.6241 %
Kappa statistic	0.3261
Mean absolute error	0.2675
Root mean squared error	0.4042
Relative absolute error	76.1481 %
Root relative squared error	96.7126 %

Tabla 7.17: Matriz de confusión: RFB (Experto).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	77	5	0	0
Otras Demencias	31	16	5	0
Demencia Vascular	0	5	2	0

De acuerdo con la Tabla 7.17, la principal fuente de error es la confusión entre las clases “Enfermedad de Alzheimer” y “Otras Demencias”.

La Tabla 7.18 muestra los resultados obtenidos con el algoritmo RBF y el conjunto de características seleccionadas mediante recuento Borda. La matriz de confusión correspondiente es la Tabla 7.19.

Tabla 7.18: Evaluación de resultados: RBF (Borda).

Concepto	Valor
Correctly Classified Instances	73.7589 %
Incorrectly Classified Instances	26.2411 %
Kappa statistic	0.4735
Mean absolute error	0.2255
Root mean squared error	0.3824
Relative absolute error	64.2011 %
Root relative squared error	91.4946 %

De acuerdo con la Tabla 7.19, la principal fuente de error es la confusión entre las clases “Enfermedad de Alzheimer” y “Otras Demencias”.

7.3.3. SVM

La Tabla 7.20 muestra los resultados obtenidos con el algoritmo Máquinas de Vectores Soporte y el conjunto de características seleccionadas por el experto. La matriz de confusión correspondiente es la Tabla 7.21.

De acuerdo con la Tabla 7.21, la principal fuente de error es la confusión entre las clases “Enfermedad de Alzheimer” y “Otras Demencias”.

Tabla 7.19: Matriz de confusión: RBF (Borda).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	77	3	2	0
Otras Demencias	25	23	4	0
Demencia Vascular	1	2	4	0

Tabla 7.20: Evaluación de resultados: SVM (Experto).

Concepto	Valor
Correctly Classified Instances	67.3759 %
Incorrectly Classified Instances	32.6241 %
Kappa statistic	0.2848
Mean absolute error	0.2175
Root mean squared error	0.4664
Relative absolute error	61.9181 %
Root relative squared error	111.5930 %
K&B Relative Info Score	4772.5666 %
K&B Information Score	0.4114 bits/instance
Class complexity order 0	1.2047 bits/instance
Class complexity scheme	350.3830 bits/instance
Complexity improvement (Sf)	-349.1782 bits/instance

La Tabla 7.22 muestra los resultados obtenidos con el algoritmo SVM y el conjunto de características seleccionadas mediante recuento Borda. La matriz de confusión correspondiente es la Tabla 7.23.

De nuevo (ver Tabla 7.23, la mayor confusión se produce entre el Alzheimer y Otras Demencias. Se puede apreciar que los resultados obtenidos con SVM son muy similares a los obtenidos con RBF. Esto probablemente es debido a que en la configuración de SVM se ha escogido un kernel tipo RBF, lo cual quiere decir que en el fondo están funcionando de la misma manera.

7.3.4. Clustering

KMeans

A continuación se muestran los resultados obtenidos con el algoritmo SimpleKMeans y el conjunto de características seleccionadas por el experto. La matriz de emparejamiento correspondiente es la Tabla 7.24. Como se mencionó antes, para los algoritmos de clustering, al tratarse de métodos no supervisados, la matriz de confusión se denomina *matching matrix*.

=== Model and evaluation on training set ===

Clustered Instances

0 56 (40%)

7.3 Algoritmos de Inteligencia Computacional

Tabla 7.21: Matriz de confusión: SVM (Experto).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	78	4	0	0
Otras Demencias	35	17	0	0
Demencia Vascular	3	4	0	0

Tabla 7.22: Evaluación de resultados: SVM (Borda).

Concepto	Valor
Correctly Classified Instances	73.0496 %
Incorrectly Classified Instances	26.9504 %
Kappa statistic	0.4188
Mean absolute error	0.1797
Root mean squared error	0.4239
Relative absolute error	51.1497 %
Root relative squared error	101.4261 %
K&B Relative Info Score	6147.5242 %
K&B Information Score	0.5300 bits/instance
Class complexity order 0	1.2047 bits/instance
Class complexity scheme	289.4468 bits/instance
Complexity improvement (Sf)	-288.2421 bits/instance

```
1      21 ( 15%)
2      64 ( 45%)
```

```
Class attribute: DxSimplificado2
```

```
Cluster 0 <-- OtrasDem
Cluster 1 <-- DV
Cluster 2 <-- EA
```

```
Incorrectly clustered instances : 70.0 49.6454 %
```

De acuerdo con la Tabla 7.24, la principal fuente de error es la confusión entre las clases “Enfermedad de Alzheimer” y “Otras Demencias”, aunque también hay un número considerable de casos de Otras Demencias que se han clasificado erróneamente como Demencia Vascular.

Resultados obtenidos por el algoritmo `SimpleKMeans`, con el conjunto de características seleccionadas mediante recuento Borda. La matriz de emparejamiento correspondiente es la Tabla 7.25.

```
=== Model and evaluation on training set ===
```

```
Clustered Instances
```

```
0      20 ( 14%)
1      53 ( 38%)
```

Capítulo 7 : Resultados

Tabla 7.23: Matriz de confusión: SVM (Borda).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	80	2	0	0
Otras Demencias	29	23	0	0
Demencia Vascular	2	5	0	0

Tabla 7.24: Matriz de emparejamiento: KMeans (Experto).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	46	34	2	0
Otras Demencias	16	21	15	0
Demencia Vascular	2	1	4	0

2 68 (48%)

Class attribute: DxSimplificado2

Cluster 0 <-- DV

Cluster 1 <-- EA

Cluster 2 <-- OtrasDem

Incorrectly clustered instances : 70.0 49.6454 %

La Tabla 7.25 muestra, como en el caso anterior, que hay bastante confusión entre todas las clases. Esto corrobora el alto número de instancias incorrectamente clasificadas obtenido.

Tabla 7.25: Matriz de emparejamiento: KMeans (Borda).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	41	39	2	0
Otras Demencias	12	26	14	0
Demencia Vascular	0	3	4	0

Expectation Maximization

A continuación se muestran los resultados obtenidos con el algoritmo Expectation Maximization y el conjunto de características seleccionadas por el experto. La matriz de emparejamiento correspondiente es la Tabla 7.26.

=== Model and evaluation on training set ===

Clustered Instances

0 37 (26%)

7.3 Algoritmos de Inteligencia Computacional

```

1      30 ( 21%)
2      74 ( 52%)

Log likelihood: -1.48487

Cluster 0 <-- OtrasDem
Cluster 1 <-- DV
Cluster 2 <-- EA

```

Incorrectly clustered instances : 64.0 45.3901 %

De acuerdo con la Tabla 7.26, la principal fuente de error es la confusión entre las clases “Enfermedad de Alzheimer” y “Otras Demencias”. También hay un gran número de Otras Demencias clasificadas como Demencia Vascular.

Tabla 7.26: Matriz de emparejamiento: EM (Experto).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	56	22	4	0
Otras Demencias	17	15	20	0
Demencia Vascular	1	0	6	0

Resultados obtenidos por el algoritmo EM, con el conjunto de características seleccionadas mediante recuento Borda. La matriz de emparejamiento correspondiente es la Tabla 7.27.

```

=== Model and evaluation on training set ===
Clustered Instances
0      78 ( 55%)
1      16 ( 11%)
2      47 ( 33%)

Log likelihood: 6.72926

Cluster 0 <-- OtrasDem
Cluster 1 <-- DV
Cluster 2 <-- EA

```

Incorrectly clustered instances : 63.0 44.6809 %

En este caso (ver Tabla 7.27), el error se ha concentrado en casos de Alzheimer clasificados como Otras Demencias.

Hierarchical Clusterer

Resultados obtenidos por el algoritmo `Hierarchical Clusterer`, con el conjunto de características seleccionadas mediante por el experto. La matriz de emparejamiento correspondiente es la Tabla 7.28.

```

=== Model and evaluation on training set ===
Clustered Instances

```

Capítulo 7 : Resultados

Tabla 7.27: Matriz de emparejamiento: EM (Borda).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	38	43	1	0
Otras Demencias	9	34	9	0
Demencia Vascular	0	1	6	0

```
0      138 ( 98%)
1       2 (  1%)
2       1 (  1%)
```

Class attribute: DxSimplificado2

```
Cluster 0 <-- EA
Cluster 1 <-- DV
Cluster 2 <-- OtrasDem
```

Incorrectly clustered instances : 56.0 39.7163 %

De acuerdo con la Tabla 7.28, la principal fuente de error es la confusión entre de casos de “Otras Demencias” incorrectamente clasificados como “Enfermedad de Alzheimer”. El error cometido es tan alto que, pese a que el resto de errores es muy bajo comparado con otros métodos, la precisión total está por debajo de estos.

Tabla 7.28: Matriz de emparejamiento: Hierarchical Clusterer (Experto).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	82	0	0	0
Otras Demencias	51	1	0	0
Demencia Vascular	5	0	2	0

Resultados obtenidos por el algoritmo `Hierarchical Clusterer`, con el conjunto de características seleccionadas mediante recuento Borda. La matriz de emparejamiento correspondiente es la Tabla 7.29.

```
=== Model and evaluation on training set ===
Clustered Instances
0      135 ( 96%)
1       5 (  4%)
2       1 (  1%)
```

Class attribute: DxSimplificado2

```
Cluster 0 <-- EA
Cluster 1 <-- OtrasDem
Cluster 2 <-- DV
```

7.3 Algoritmos de Inteligencia Computacional

Incorrectly clustered instances : 56.0 39.7163 %

La Tabla 7.29 muestra una situación muy similar al caso anterior, por lo que parece que el efecto del conjunto de variables seleccionado es mínimo para este algoritmo.

Tabla 7.29: Matriz de emparejamiento: Hierarchical Clusterer (Borda).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	82	0	0	0
Otras Demencias	50	2	0	0
Demencia Vascular	3	3	1	0

ROCK

Resultados obtenidos mediante el algoritmo ROCK, con el conjunto de características seleccionadas por el experto. La matriz de emparejamiento correspondiente es la Tabla 7.30. De acuerdo con los resultados mostrados en la Tabla 7.30, el mencionado algoritmo tiene una tasa de acierto del 61,70 %.

Tabla 7.30: Matriz de emparejamiento: ROCK (Experto).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	82	0	0	0
Otras Demencias	46	5	0	1
Demencia Vascular	3	4	0	0

Resultados obtenidos por el algoritmo ROCK, con el conjunto de características seleccionadas mediante recuento Borda. La matriz de emparejamiento correspondiente es la Tabla 7.31. De acuerdo con los resultados mostrados en la Tabla 7.31, el algoritmo tiene una tasa de acierto del 60,28 %.

Tabla 7.31: Matriz de emparejamiento: ROCK (Borda).

	Clasificado como...			
	E. Alzheimer	Otras D.	D. Vascular	<NA>
Enf. Alzheimer	81	1	0	0
Otras Demencias	41	2	2	7
Demencia Vascular	3	0	2	2

En el caso del algoritmo ROCK, las Tablas 7.30 y 7.31 muestran una situación muy silimilar a la del método Hierachical Clusterer, donde el error en general es muy bajo, pero una clase concentra un gran número de instancias mal

clasificadas, y eso rebaja mucho la precisión del método. Como se ha visto en la primera parte de esta memoria, el algoritmo ROCK está basado en métodos de clustering jerárquico, pero modificando la forma de medir las distancias entre los elementos. Para esta aplicación concreta, y a la vista de los resultados, esta forma distinta de considerar la distancia no ha tenido apenas efecto puesto que se ha obtenido una precisión muy similar a la de Hierarchical Clusterer.

7.4. Comparación de Resultados

La Tabla 7.32 compila los principales resultados obtenidos por todos los métodos probados en el marco de este trabajo. Las columnas etiquetadas como CCI e ICI se corresponden con el valor de *Correctly Classified Instances* e *Incorrectly Classified Instances*, respectivamente. El resto de abreviaturas son comúnmente utilizadas en la bibliografía y han sido empleadas a lo largo de esta memoria. Salvo la columna *Tiempo*, que indica el tiempo en segundos que ha tardado en generarse el resultado, el resto de los valores son porcentajes.

Del análisis de la Tabla 7.32 se deduce que el mejor resultado (78 % de aciertos) se ha obtenido aplicando Random Forest al conjunto de variables seleccionadas mediante recuento Borda. En el caso del peor resultado (50 % de aciertos), este se ha obtenido con KMeans independientemente del conjunto de variables utilizado.

Al revisar las tablas de confusión y las tablas de emparejamiento, se puede observar que la mayoría de los errores se concentran en los casos de Alzheimer clasificados como Otras Demencias y viceversa.

En opinión del autor, el buen resultado obtenido por Random Forest puede deberse a algún tipo de sesgo en los datos o carencia, ya que el dataset de partida no era excesivamente grande y estaba poco poblado, i.e. había bastantes campos en blanco. Generalmente, los algoritmos basados en redes neuronales (MLP y RBF, en este caso) suelen tener las precisiones más altas, lo que se cumple aquí salvo por la excepción de RF.

De la misma manera, tampoco es coherente las diferencias en la precisión al utilizar el conjunto de variables del experto o el conjunto de variables seleccionadas mediante recuento Borda. En opinión del autor, sería necesario profundizar en el estudio para ver qué factores son los que realmente están influyendo en esos resultados.

Para el caso del peor resultado, obtenido por KMeans, no hay objeción ya que al tratarse de un conjunto de datos de entrenamiento supervisado, es coherente que los algoritmos supervisados hayan obtenido en general mejores resultados que los no supervisados.

Merece especial mención el caso del algoritmo ROCK, que ha sido incluido en este trabajo de manera expresa, al tratarse de un algoritmo especialmente diseñado para trabajar con conjuntos de variables categóricas. El resultado obtenido por este método no está a la altura de lo que se esperaba al inicio del trabajo, puesto que sin ser de los peores, tampoco es superior a los otros métodos de clustering testados.

Tabla 7.32: Comparación de los resultados obtenidos.

	Variables Experto				Variables Recuento Borda					
	CCI	ICI	MAE	RMSE	Tiempo	CCI	ICI	MAE	RMSE	Tiempo
J4.8	69.50	30.50	0.26	0.40	0.02	65.25	34.75	0.27	0.43	0.10
Random Forest	63.83	36.17	0.28	0.40	0.03	78.01	21.99	0.23	0.35	0.11
Perceptrón Multicapa	58.16	41.84	0.28	0.48	1.09	73.76	26.24	0.18	0.38	2.28
Funciones de Base Radial	67.38	32.62	0.27	0.40	0.04	73.76	26.24	0.23	0.38	0.22
SVM	67.38	32.62	0.22	0.47	0.03	73.05	26.95	0.18	0.42	0.03
KMeans	50.35	49.65			0.06	50.35	49.65			0.02
Expectation Maximization	54.61	45.39				55.32	44.68			
Hierarchical Clusterer	60.28	39.72			0.91	60.28	39.72			0.71
ROCK	61.70	38.30				60.28	39.72			

Capítulo 8

Conclusiones

El objetivo de este Trabajo de Fin de Máster era doble. Por un lado realizar un estudio del estado del arte relativo a la aplicación de técnicas de Inteligencia Computacional en el campo de la demencia. Por otro, realizar una aplicación práctica de algunas de estas técnicas a un problema real, utilizando datos de casos y diagnósticos reales proporcionados por el Centro de Referencia Estatal de Atención a Personas con Enfermedad de Alzheimer y otras Demencias.

La primera parte ha demostrado que casi todas las investigaciones se centran en la aplicación de diversas técnicas, principalmente SVM, al diagnóstico de la enfermedad a través de imágenes de resonancia magnética del cerebro. No obstante, se han encontrado interesantes investigaciones que están abriendo el campo a otras técnicas y nuevas aplicaciones, como el diagnóstico neuropsicológico, menos intrusivo para el paciente y más fácil de realizar al no necesitar de complejos y costosos aparatos.

Por tanto, a la vista de estos resultados y del hecho de que los criterios y las guías para el diagnóstico no son definitivas y se revisan y actualizan cada cierto tiempo, parece que hay campo para la realización de más estudios en la línea del diagnóstico neuropsicológico, o completar algunos de los que ya están en marcha.

La segunda parte del trabajo se ha centrado en la caracterización de pacientes con enfermedad de Alzheimer. Para ello, se ha realizado una recogida y reconciliación de datos sobre casos y diagnósticos de pacientes con Alzheimer y otras demencias. Se ha conseguido determinar qué técnica de inteligencia computacional es la más apropiada para el diagnóstico diferencial entre la Demencia Vasculosa y la Demencia tipo Alzheimer, en función de las características cualitativas y cuantitativas de los datos disponibles. Las diferentes técnicas evaluadas se han comparado con el diagnóstico realizado por un experto obteniendo unos resultados, en algunos casos, muy significativos puesto que mejoraban considerablemente la evaluación experta.

De entre los algoritmos evaluados, el mejor resultado se ha obtenido mediante la aplicación de **Random Forest** al conjunto de variables seleccionadas mediante recuento Borda. En cambio, los algoritmos basados en redes neuronales artificiales han obtenidos mejores resultados en promedio, independientemente del conjunto de datos utilizado.

El peor resultado se ha obtenido aplicando **KMeans** independientemente del conjunto de datos seleccionado. En general, los algoritmos no supervisados han

obtenido peores resultados que los supervisados, pero esto puede ser debido a que el conjunto de datos era supervisado y, por tanto, es coherente que éstos hayan tenido mejores resultados.

Merece una mención especial el caso del algoritmo ROCK, que es un algoritmo de clustering. Se ha incluido en este trabajo al tratarse de un método especialmente diseñado para trabajar con variables categóricas, como lo son la mayoría en este caso. No obstante, los resultados obtenidos no han sido tan prometedores como cabía esperar, situándose a un nivel muy similar al de los otros algoritmos de clustering evaluados.

En opinión del autor de este trabajo, los resultados no son concluyentes por el limitado alcance del estudio y el reducido número de datos disponibles. No obstante, el estudio ha servido para comprobar el estado actual de la investigación en este campo y testar la metodología, con el fin de probar si es útil como herramienta de apoyo a la toma de decisiones en el ámbito del diagnóstico diferencial entre la demencia vascular y la demencia tipo Alzheimer.

8.1. Futuras líneas de investigación

Como se ha mencionado en la Sección anterior las líneas de trabajo posibles son múltiples, y por mencionar algunas, el autor propone:

- Estudiar el efecto de la combinación de ambos conjuntos de variables: las seleccionadas por el experto y las seleccionadas mediante recuento Borda, para obtener una solución intermedia que mejore la precisión del diagnóstico, reduciendo así el error.
- Aplicar otros métodos de selección de variables, e.g. OWA, y analizar si la nueva selección se comporta mejor que la realizada. En caso contrario, combinar de forma ponderada las variables propuestas por ambos métodos, en busca del conjunto óptimo de variables.
- Incluir un grupo de control en el estudio. Estaría formado por personas de edad avanzada, pero sanas. Serviría para calibrar mejor la precisión en el diagnóstico y descubrir posibles fallos en el funcionamiento de los algoritmos. También sería deseable contar con datos sobre casos y diagnósticos provenientes de otros centros y evaluados por otros expertos. De esta manera se podría determinar la capacidad de generalización del método y su posible implantación fuera del entorno de desarrollo.

Bibliografía

- [1] A. Abraham and B. Nath. A neuro-fuzzy approach for modelling electricity demand in victoria. *Applied Soft Computing*, 1(2):127–138, 2001.
- [2] Paul Aljabar, Daniel Rueckert, and William R Crum. Spectral clustering as a diagnostic tool in cross-sectional mr studies: an application to mild dementia. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 11(Pt 2):442–9, 2008. Times Cited: 0.
- [3] J.M. Alonso. Generating understandable and accurate fuzzy models in a java environment. <http://sourceforge.net/p/guajefuzzy/wiki/Home/>.
- [4] Alzheimer’s Association. 2012 Alzheimer’s Disease Facts and Figures. *Alzheimer’s & Dementia*, 8(2), 2012.
- [5] Alzheimer’s Disease International. *Informe Mundial sobre el Alzheimer - Resumen ejecutivo*. ADI, 2009.
- [6] H. Arimura, T. Magome, Y. Yamashita, and D. Yamamoto. Computer-aided diagnosis systems for brain diseases in magnetic resonance images. *Algorithms*, 2(3):925–952, 2009. Times cited: 4.
- [7] J. L. Aznarte, J. M. Benítez Sánchez, and D. Nieto Lugilde. Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. *Expert Systems with Applications*, 32(4):1218–1225, 2007.
- [8] P.G. Báez, M.A.P. Del Pino, C.F. Viadero, and C.P.S. Araujo. Artificial intelligent systems based on supervised humann for differential diagnosis of cognitive impairment: Towards a 4p-hcads. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5517 LNCS(PART 1):981–988, 2009. Times cited: 3.
- [9] Samuel R. S. Barnes, E. Mark Haacke, Muhammad Ayaz, Alexander S. Boikov, Wolff Kirsch, and Dan Kido. Semiautomated detection of cerebral microbleeds in magnetic resonance images. *MAGNETIC RESONANCE IMAGING*, 29(6):844–852, JUL 2011. Times Cited: 2.
- [10] J. C. Bezdek. *Pattern recognition with fuzzy objective functions algorithms*. Plenum Press, New York, 1981.

BIBLIOGRAFÍA

- [11] Ron Brookmeyer, Elizabeth Johnson, Kathryn Ziegler-Graham, and H. Michael Arrighi. Forecasting the global burden of alzheimer's disease. *Alzheimer's and Dementia*, 3(3):186–191, 2007. Available at: <http://works.bepress.com/rbrookmeyer/23>.
- [12] D.S. Broomhead and d. Lowe. Multivariable functional interpolation and adaptative networks. *Complex Systems*, 2:321–355, 1988.
- [13] E. Cabello Pardos. *Técnicas de reconocimiento facial mediante Redes Neuronales*. Universidad Politécnica de Madrid, 2004.
- [14] M. Y. Chen. Establishing interpretable fuzzy models from numeric data. In *Proceedings of the 4th World Congress on Intelligent Control and Automation*, pages 1857–1861, 2002.
- [15] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.
- [16] Simon Duchesne, Anna Caroli, C. Geroldi, Christian Barillot, Giovanni B. Frisoni, and D. Louis Collins. Mri-based automated computer classification of probable ad versus normal controls. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 27(4):509–520, APR 2008. Times Cited: 34.
- [17] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1974.
- [18] B.V. Elsevier. Sciverse scopus. <http://www.scopus.com/home.url>. Last visited: Dec'11.
- [19] E.A. Feigenbaum. State of the art report on machine intelligence. *Expert systems in the 1980s*, 1981.
- [20] Google. Google. <http://www.google.com>. Last visited: Dec'11.
- [21] Google. Google académico. <http://scholar.google.es/>. Last visited: Dec'11.
- [22] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Rock: a robust clustering algorithm for categorical attributes. In *Information Systems*, volume 25, pages 345–366, 2000.
- [23] S. Guillaume and B. Charnomordic. A new method for inducing a set of interpretable fuzzy partitions and fuzzy inference systems from data. In Springer-Verlag, editor, *Studies in Fuzziness and Soft Computing*, pages 148–175, Heidelberg, 2003.
- [24] S. Guillaume and B. Charnomordic. Generating an interpretable family of fuzzy partitions. *IEEE Transactions on Fuzzy Systems*, 12(3):324–335, June 2004.
- [25] S. Haller, D. Nguyen, C. Rodriguez, J. Emch, G. Gold, A. Bartsch, K.O. Lovblad, and P. Giannakopoulos. Individual prediction of cognitive decline in mild cognitive impairment using support vector machine-based analysis of diffusion tensor imaging data. *Journal of Alzheimer's Disease*, 22(1):315–327, 2010. Times cited: 10.

-
- [26] J. A. Hartigan and M. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.
- [27] Jean-Francois Horn, Marie-Odile Habert, Aurelie Kas, Zoulikha Malek, Philippe Maksud, Lucette Lacomblez, Alain Giron, and Bernard Fertil. Differential automatic diagnosis between alzheimer’s disease and fronto-temporal dementia based on perfusion spect images. *ARTIFICIAL INTELLIGENCE IN MEDICINE*, 47(2):147–158, OCT 2009. Times Cited: 8.
- [28] M. Hornik, M. Stinchcombe, and H. White. Multi-layer feed-forward networks are universal approximators. *Neural networks*, 2:359–366, 1989.
- [29] IEEE - Institute of Electrical and Electronics Engineers. Ieee xplore. <http://ieeexplore.ieee.org/Xplore/guesthome.jsp>. Last visited: Dec’11.
- [30] IMSERSO. Centro de referencia estatal de atención a personas con enfermedad de alzheimer y otras demencias. <http://www.crealzheimer.es/>. última visita: julio 2012.
- [31] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):4–37, 2000.
- [32] Sandhya Joshi, P. Deepa Shenoy, K. R. Venugopal, and L. M. Patnaik. *Evaluation of Different Stages of Dementia Employing Neuropsychological and Machine Learning Techniques*, pages 154–160. IEEE, 2009. Times Cited: 0.
- [33] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc., 1990.
- [34] Stefan Kloeppe, Ahmed Abdulkadir, Stathis Hadjidemetriou, Sabine Issleib, Lars Frings, Thao Nguyen Thanh, Irina Mader, Stefan J. Teipel, Michael Huell, and Olaf Ronneberger. A comparison of different automated methods for the detection of white matter lesions in mri data. *NEURO-IMAGE*, 57(2):416–422, JUL 15 2011. Times Cited: 1.
- [35] Stefan Kloeppe, Cynthia M. Stonnington, Josephine Barnes, Frederick Chen, Carlton Chu, Catriona D. Good, Irina Mader, L. Anne Mitchell, Ameet C. Patel, Catherine C. Roberts, Nick C. Fox, Clifford R. Jack Jr., John Ashburner, and Richard S. J. Frackowiak. Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. *BRAIN*, 131:2969–2974, NOV 2008. Times Cited: 32.
- [36] Stefan Kloeppe, Cynthia M. Stonnington, Carlton Chu, Bogdan Draganski, Rachael I. Scahill, Jonathan D. Rohrer, Nick C. Fox, Jr. Jack, Clifford R., John Ashburner, and Richard S. J. Frackowiak. Automatic classification of mr scans in alzheimers disease. *BRAIN*, 131(Part 3):681–689, MAR 2008 2008. Times Cited: 152.
- [37] Gabor G. Kovacs, Gergo Botond, and Herbert Budka. Protein coding of neurodegenerative dementias: the neuropathological basis of biomarker diagnostics. *ACTA NEUROPATHOLOGICA*, 119(4):389–408, APR 2010. Times Cited: 18.

BIBLIOGRAFÍA

- [38] Sridhar Krishnamurti, Lyle Drake, and Justin King. Neural network modeling of central auditory dysfunction in Alzheimer's disease. *NEURAL NETWORKS*, 24(6, SI):646–651, AUG 2011. Times cited: 0.
- [39] Dmitry Lagun, Cecelia Manzanares, Stuart M. Zola, Elizabeth A. Buffalo, and Eugene Agichtein. Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *JOURNAL OF NEUROSCIENCE METHODS*, 201(1):196–203, SEP 30 2011. Times Cited: 0.
- [40] C. Laske, T. Leyhe, E. Stransky, N. Hoffmann, A.J. Fallgatter, and J. Dietzsch. Identification of a blood-based biomarker panel for classification of alzheimer's disease. *International Journal of Neuropsychopharmacology*, 14(9):1147–1155, 2011. Times cited: 2.
- [41] S. Li, F. Shi, F. Pu, X. Li, T. Jiang, S. Xie, and Y. Wang. Hippocampal shape analysis of alzheimer disease based on machine learning methods. *American Journal of Neuroradiology*, 28(7):1339–1345, 2007. Times cited: 29.
- [42] H. Liu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 2005.
- [43] Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.
- [44] Joao Maroco, Dina Silva, Ana Rodrigues, Manuela Guerreiro, Isabel Santana, and Alexandre de Mendonca. Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC research notes*, 4:299, Aug 17 2011. Times Cited: 0.
- [45] D. Michie. Inductive rule generation in the context of the fifth generation. In *Proc. of the First International Workshop on Machine Learning*, pages 66–70, Monticello, Illinois, june 1983.
- [46] M. Minsky and S. Pappert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, Cambridge MA, 1969.
- [47] Sushmita Mitra and Tinku Acharya. *Data mining: multimedia, soft computing, and bioinformatics*. John Wiley, 2003.
- [48] J.H. Morra, Zhuowen Tu, L.G. Apostolova, A.E. Green, A.W. Toga, and P.M. Thompson. Comparison of adaboost and support vector machines for detecting alzheimer's disease through automated hippocampal segmentation. *Medical Imaging, IEEE Transactions on*, 29(1):30–43, jan. 2010. Times cited: 19.
- [49] U.S. Department of Health & Human Services National Institute on Ageing National Institutes of Health. Alzheimer's disease education and referral center (adear). <http://www.nia.nih.gov/Alzheimers/>.
- [50] O. Nelles. *Nonlinear System Identification*. Springer, 2001.

-
- [51] Laurence O'Dwyer, Franck Lambertson, Arun L. W. Bokde, Michael Ewers, Yetunde O. Faluyi, Colby Tanner, Bernard Mazoyer, Desmond O'Neill, Mairead Bartley, D. Ronan Collins, Tara Coughlan, David Prvulovic, and Harald Hampel. Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment. *PLOS ONE*, 7(2), FEB 23 2012. Article number: e32441 Times Cited: 0 Z9 0.
- [52] M. A. Oquendo, E. Baca-Garcia, A. Artés-Rodríguez, F. Perez-Cruz, H. C. Galfalvy, H. Blasco-Fontecilla, D. Madigan, and N. Duan. Machine learning and data mining: strategies for hypothesis generation. *Molecular Psychiatry*, JAN 10 2012. Times Cited: 0.
- [53] MJ Pazzani, S Mani, and WR Shankle. Acceptance of rules generated by machine learning among medical experts. *METHODS OF INFORMATION IN MEDICINE*, 40(5):380–385, 2001. Times cited: 35.
- [54] Witold Pedrycz. *Computational Intelligence: An Introduction*. CRC-Press, 1999.
- [55] W. Pitts and W. McCulloch. How we know universals: the perception of auditory and visual forms. *Bulletin Math. biophys*, 9:27–47, 1947.
- [56] C. Plant, S.J. Teipel, A. Oswald, C. B ohm, T. Meindl, J. Mourao-Miranda, A.W. Bokde, H. Hampel, and M. Ewers. Automated detection of brain atrophy patterns based on mri for the prediction of alzheimer's disease. *NeuroImage*, 50(1):162–174, 2010. Times cited: 20.
- [57] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, 1986.
- [58] J. Ross Quinlan. Simplifying decision trees. *Int. J. Man-Mach. Stud.*, 27(3):221–234, 1987.
- [59] J. Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
- [60] Barry Reisberg, Steven H. Ferris, Mony J. de Leon, and Thomas Crook. The global deterioration scale for assessment of primary degenerative dementia. *The American Journal of Psychiatry*, 139(9):1136–1139, Sep. 1982.
- [61] Thomson Reuters. Isi web of knowledge. <http://isiknowledge.com>. Last visited: Dec'11.
- [62] F. Rosenblatt. *Principles of Neurodynamics*. Spartan Books, 1962.
- [63] David E. Rumelhart and James L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, 1987.
- [64] G.E. Sakr, I.H. Elhajj, and H. Abou-Saad Huijjer. Support vector machines to define and detect agitation transition. *IEEE Transactions on Affective Computing*, 1(2):98–108, 2010. Times cited: 2.
- [65] A. Savio, M.T. García-Sebastián, D. Chyzyk, C. Hernández, M. Graña, A. Sistiaga, A. López de Munain, and J. Villanúa. Neurocognitive disorder detection based on feature vectors extracted from vbm analysis of structural mri. *Computers in Biology and Medicine*, 41(8):600–610, 2011. Times cited: 7.

BIBLIOGRAFÍA

- [66] WR Shankle, S Mani, MJ Pazzani, and P Smyth. Dementia screening with machine learning methods. In *INTELLIGENT DATA ANALYSIS IN MEDICINE AND PHARMACOLOGY*, volume 414 of *KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE*, pages 149–165, 1997. Times cited: 5.
- [67] Basilio Sierra-Araujo. *Aprendizaje Automático: Conceptos Básicos y Avanzados*. Pearson Prentice Hall, 2006.
- [68] Prashanthi Vemuri, Jeffrey L. Gunter, Matthew L. Senjem, Jennifer L. Whitwell, Kejal Kantarci, David S. Knopman, Bradley F. Boeve, Ronald C. Petersen, and Jr. Jack, Clifford R. Alzheimer’s disease diagnosis in individual subjects using structural mr images: Validation studies. *NEUROIMAGE*, 39(3):1186–1197, FEB 1 2008. Times Cited: 92.
- [69] Prashanthi Vemuri, Gyorgy Simon, Kejal Kantarci, Jennifer L. Whitwell, Matthew L. Senjem, Scott A. Przybelski, Jeffrey L. Gunter, Keith A. Josephs, David S. Knopman, Bradley F. Boeve, Tanis J. Ferman, Dennis W. Dickson, Joseph E. Parisi, Ronald C. Petersen, and Jr. Jack, Clifford R. Antemortem differential diagnosis of dementia pathology using structural mri: Differential-stand. *NEUROIMAGE*, 55(2):522–531, MAR 15 2011. Times Cited: 3.
- [70] Ying Wang, Yong Fan, Priyanka Bhatt, and Christos Davatzikos. High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *NeuroImage*, 50(4):1519–1535, 2010. Times cited:21.
- [71] P. Werbos. *Beyond Regression: new tools for prediction and analysis in the behavioral sciences*. Harvard University, 1974.
- [72] WHO. World health organization. <http://www.who.int/>.
- [73] B. Widrow. Generaliation and information storage in networks of adaline ‘neurons’. *Self Organizing Systems*, pages 435–461, 1959.
- [74] Wikipedia. Wikipedia. <http://www.wikipedia.org/>.
- [75] Windong Wu and Vipin Kumar. *The Top Ten Algorithms in Data Mining*. CRC Press, 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742, April 2009.
- [76] Y. Yamashita, H. Arimura, and K. Tsuchiya. Computer-aided detection of ischemic lesions related to subcortical vascular dementia on magnetic resonance images. *Academic Radiology*, 15(8):978–985, 2008. Times cited: 6.
- [77] Daoqiang Zhang and Dinggang Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer’s disease. *NEUROIMAGE*, 59(2):895–907, JAN 16 2012. Times Cited: 0.

Parte III
Apéndices

Apéndice A

Tablas

Tabla A.1: Descripción de las variables.

Nombre	Definición	Valores
N_Hist	Número de historia	Ordinal
Edad	Edad	En años
Edad.2	Edad	{1 = <60 2 = 61-70 3 = >70}
Edad.grupos	Edad por grupos	{1 = <64 2 = 65-69 3 = 70-74 4 = 75-79 5 = 80-84 6 = 85-89 7 = >90}
Sexo	Sexo	{1 = Varon 2 = Mujer}
Recurso	Recurso utilizado	{1 = Centro de día 2 = Centro de noche 3 = Residencial temporal 4 = Módulo de respiro familiar 5 = Estancias breves de familias
Fecha_Entrada	Fecha de entrada	Fecha
Fecha_Salida	Fecha de salida	Fecha
Dx	Diagnóstico inicio	{1 = Enfermedad de Alzheimer 2 = Enfermedad de Alzheimer Probable 3 = Demencia Mixta 4 = Demencia Vascular 5 = Demencia Secundaria 6 = Demencia Frontal 7 = Parkinson Demencia 8 = Deterioro Cognitivo Leve 9 = Deterioro Cognitivo Multifactorial 10 = Deterioro Cognitivo No Filiado 11 = Enfermedad con Cuerpos de Lewy 12 = Enfermedad de Huntington 13 = Enfermedad de Pick 14 = Parálisis Supranuclear Progresiva 15 = Afasia Progresiva Primaria 16 = Atrofia Cortical Posterior 17 = Demencia Degenerativa Primaria 18 = Pseudodemencia Depresiva 19 = Secuelas ACV 20 = Otros}
DxSimplificado	Diagnóstico simplificado inicio	{1 = Enfermedad de Alzheimer 2 = Demencia Mixta 3 = Demencia vascular 4 = Enfermedad con cuerpos de Lewy 5 = Otras demencias}
Dx_DxF	Dx Final = Dx Inicial	{0 = Diferente 1 = Igual}
DxFinal	Diagnóstico final	{1 = Enfermedad de Alzheimer 2 = Enfermedad de Alzheimer Probable 3 = Demencia Mixta 4 = Demencia Vascular 5 = Demencia Secundaria 6 = Demencia Frontal 7 = Parkinson Demencia 8 = Deterioro Cognitivo Leve 9 = Deterioro Cognitivo Multifactorial 10 = Deterioro Cognitivo No Filiado 11 = Enfermedad con Cuerpos de Lewy 12 = Enfermedad de Huntington 13 = Enfermedad de Pick 14 = Parálisis Supranuclear Progresiva 15 = Afasia Progresiva Primaria 16 = Atrofia Cortical Posterior}

Sigue ...

ApéndiceA : Tablas

Tabla A.1: Descripción de las variables (continuación)

Nombre	Definición	Valores
		17 = Demencia Degenerativa Primaria 18 = Pseudodemencia Depresiva 19 = Secuelas ACV 20 = Otros 21 = Demencia Fronto-Temporal 22 = EA + Demencia por meningioma 23 = Demencia Vascular Subcortical 24 = Daño Cerebral Postraumático}
DxSimplificado2	Diagnóstico simplificado final	{1 = Enfermedad de Alzheimer 2 = Demencia Mixta 3 = Demencia vascular 4 = Enfermedad con cuerpos de Lewy 5 = Otras demencias}
Motivo.Consulta	Motivo de consulta	{1 = Trastorno de memoria 2 = Trastorno de conducta 3 = Trastorno de la marcha 4 = Síntomas focales 5 = Fracaso sociolaboral 6 = Depresión 7 = Otros}
Profesion	Profesión	{1 = Ama de casa 2 = Obrero no especializado 3 = Obrero especializado 4 = Técnico medio 5 = Oficinista administrativo 6 = Ejecutivo o profesional 7 = Militar}
Especificar Ciudad Provincia Escolarizacion	Especificar profesión Ciudad de origen Provincia de origen Nivel educativo	Texto libre Texto libre Texto libre {1 = Analfabeto 2 = Lee y escribe 3 = Estudios mínimos 4 = Estudios primarios 5 = Bachiller 6 = Estudios superiores}
Años_escol Estado_civil	Años de escolarización Estado civil	Valor {1 = Soltero 2 = Casado 3 = Divorciado 4 = Viudo}
Vive_con	Vive con	{1 = Solo 2 = Con pareja 3 = Depende de hijo(s) 4 = Residencia 5 = Otros}
estancia_meses estancia_dias Activo	Estancia en el centro en meses Estancia en el centro en días Activo	Valor Valor {0 = No 1 = Si}
Completo	Completo	{0 = No 1 = Si}
Antecedentes Personales		
AP_Hipertension	Hipertensión	{0 = Ausente 1 = Presente}
AP_InfMioc	Infarto Miocardio	{0 = Ausente 1 = Presente}
AP_Angor	Angor	{0 = Ausente 1 = Presente}
AP_Cardiopatía	Cardiopatía	{0 = Ausente 1 = Presente}
AP_ACVA	ACVA	{0 = Ausente 1 = Presente}
AP_Diabetes	Diabetes	{0 = Ausente 1 = Presente}
AP_Hepatopatía	Hepatopatía	{0 = Ausente 1 = Presente}
AP_Alcohol	Alcohol	{0 = Ausente 1 = Presente}
AP_Tabaco	Tabaco	{0 = Ausente 1 = Presente}
AP_TPsiq	Trastornos Psiquiátricos	{0 = Ausente 1 = Presente}
AP_Epilepsia	Epilepsia	{0 = Ausente 1 = Presente}
AP_Down	Síndrome de Down	{0 = Ausente 1 = Presente}
AP_HIV	HIV	{0 = Ausente 1 = Presente}
AP_Trauma	Trauma Craneal	{0 = Ausente 1 = Presente}
AP_EnfTiroidea	Enfermedad Tiroidea	{0 = Ausente 1 = Presente}
AP_Hipercolest	Hipercolesterolemia	{0 = Ausente 1 = Presente}
AP_Otros	Otros	{0 = Ausente 1 = Presente}
Antecedentes Familiares		
AF_Down	Síndrome de Down	{0 = Ausente 1 = Presente}
AF_Demencia	Demencia	{0 = Ausente 1 = Presente}
AF_Parkinson	Parkinson	{0 = Ausente 1 = Presente}
AF_Huntington	Huntington	{0 = Ausente
Sigue . . .		

Tabla A.1: Descripción de las variables (continuación)

Nombre	Definición	Valores
AF_TPsiq	Trastornos Psiquiátricos	{ 1 = Presente 0 = Ausente 1 = Presente }
Historia del deterioro mental		
H_Comienzo	Forma de Comienzo	{ 1 = Agudo 2 = Subagudo 3 = Insidioso }
H_MesesEvolucion	Meses de evolución	Valor
Anys_Ev	Años de evolución	Valor
H_Evolucion	Forma de evolución	{ 1 = Progresiva 2 = Fluctuante 3 = Intermitente 4 = Estable 5 = Otra }
H_SintInicial	Síntoma Inicial	{ 1 Trastorno de Memoria 2 Trastorno de Conducta 3 Trastorno de Razonamiento 4 Trastorno Depresivo 5 Sintomatología Focal }
Síntomas Actuales		
SA_TMemoria	Trastornos de Memoria	{ 0 = Ausente 1 = Presente }
SA_Desorientacion	Desorientacion	{ 0 = Ausente 1 = Presente }
SA_TIntelectuales	Otros Trastornos Intelectuales	{ 0 = Ausente 1 = Presente }
SA_TConducta	Trastornos de Conducta	{ 0 = Ausente 1 = Presente }
SA_Depresion	Depresión	{ 0 = Ausente 1 = Presente }
SA_InhibPsic	Inhibición Psicomotora	{ 0 = Ausente 1 = Presente }
SA_IdDelirantes	Ideas Delirantes	{ 0 = Ausente 1 = Presente }
SA_Alucinaciones	Alucinaciones	{ 0 = Ausente 1 = Presente }
SA_Agitacion	Agitación	{ 0 = Ausente 1 = Presente }
SA_Confusion	Cuadros Confusionales	{ 0 = Ausente 1 = Presente }
SA_SintFocales	Síntomas Focales	{ 0 = Ausente 1 = Presente }
SA_Temblor	Temblor	{ 0 = Ausente 1 = Presente }
SA_TLenguaje	Trastornos del Lenguaje	{ 0 = Ausente 1 = Presente }
SA_TMarcha	Trastornos de la Marcha	{ 0 = Ausente 1 = Presente }
SA_IncontUrinaria	Incontinencia Urinaria	{ 0 = Ausente 1 = Presente }
SA_IncontFecal	Incontinencia Fecal	{ 0 = Ausente 1 = Presente }
SA_UrgMiccional	Urgencia Miccional	{ 0 = Ausente 1 = Presente }
SA_RisaLLanto	Risa-Llanto Espasmódico	{ 0 = Ausente 1 = Presente }
SA_Hipersex	Hipersexualidad	{ 0 = Ausente 1 = Presente }
SA_Bulimia	Bulimia	{ 0 = Ausente 1 = Presente }
SA_TSuenyo	Trastornos del Sueño	{ 0 = Ausente 1 = Presente }
SA_CEpilepticas	Crisis Epilépticas	{ 0 = Ausente 1 = Presente }
Anyo_Inicio	Año de inicio	Valor
Escala		
MEC2	MEC2	Valor
MEC	Mini-examen cognoscitivo	Valor
MMSE	Mini-Mental State Examination	Valor
MMSEesperado	MMSE Esperado	Valor
Barthel	Índice de Barthel	Valor
GDS	Global Deterioration Scale	Valor
Blessed	Blessed	Valor
Lawton	Lawton	Valor
Hachinski	Hachinski	Valor
Yesavage	Yesavage	Valor
Zarit	Zarit	Valor
CDR	CDR	Valor
FAB	FAB	Valor
Hamilton	Hamilton	Valor
NPLQ	NPLQ	Valor
TIN	TIN	Valor
Exploración General		
Arritmia	Arritmia	{ 0 = Ausente 1 = Presente }
Soplos	Soplos cardíacos	{ 0 = Ausente 1 = Presente }
Hepatomegalia	Hepatomegalia	{ 0 = Ausente 1 = Presente }
PulsosAbolidos	Pulsos abolidos	{ 0 = Ausente 1 = Presente }
Exploración Neurológica (Pares craneales)		
Papiledema	Papiledema	{ 0 = Ausente 1 = Presente }
AgVisual	P. de agudeza visual	{ 0 = Ausente

Sigue ...

Apéndice A : Tablas

Tabla A.1: Descripción de las variables (continuación)

Nombre	Definición	Valores
AgAuditiva	P. de agudeza auditiva	{0 = Ausente 1 = Presente}
DCampimet	Defectos campimétricos	{0 = Ausente 1 = Presente}
ParesiaF	Paresia Facial	{0 = Ausente 1 = Presente}
Amimia	Amimia	{0 = Ausente 1 = Presente}
CPseudobulbar	Cuadro Pseudobulbar	{0 = Ausente 1 = Presente}
Exploración Neurológica (Sistema Motor)		
RigPlastica	Rigidez Plástica	{0 = Ausente 1 = Presente}
Espasticidad	Espasticidad	{0 = Ausente 1 = Presente}
RuedaDentada	Rueda Dentada	{0 = Ausente 1 = Presente}
Oposicionismo	Oposicionismo	{0 = Ausente 1 = Presente}
PiramUnilateral	Piramidismo Unilateral	{0 = Ausente 1 = Presente}
PiramBilateral	Piramidismo Bilateral	{0 = Ausente 1 = Presente}
Paraparesia	Paraparesia	{0 = Ausente 1 = Presente}
Dismetria	Dismetría	{0 = Ausente 1 = Presente}
Exploración Neurológica (Movimientos Anormales)		
TemblorReposo	Temblor de reposo	{0 = Ausente 1 = Presente}
TemblorActitud	Temblor de actitud	{0 = Ausente 1 = Presente}
Bradicinesia	Bradicinesia	{0 = Ausente 1 = Presente}
Mioclónicas	Mioclónfas	{0 = Ausente 1 = Presente}
Corea	Corea	{0 = Ausente 1 = Presente}
Exploración Neurológica (Sensibilidad)		
SuperfAnormal	Superficie anormal	{0 = Ausente 1 = Presente}
ProfunAnormal	Profunda anormal	{0 = Ausente 1 = Presente}
Exploración Neurológica (Reflejos Línea Media)		
Glabealar	Glabealar	{0 = Ausente 1 = Presente}
Prension	Prensión	{0 = Ausente 1 = Presente}
Periorales	Periorales	{0 = Ausente 1 = Presente}
Exploración Neurológica (Marcha)		
MParetica	Parética	{0 = Ausente 1 = Presente}
MAtaxica	Atáxica	{0 = Ausente 1 = Presente}
MParkinsoniana	Parkinsoniana	{0 = Ausente 1 = Presente}
MApraxica	Apráxica	{0 = Ausente 1 = Presente}
Descripción de las variables (Fin)		

Tabla A.2: Ranking de las características, según distintos métodos de evaluación.

#	Relief	InfoGain	GainRatio	GiniGain	Recuento Borda
1	SA_Desorientacion	Año_Inicio	MAtaxica	Año_Inicio	Año_Inicio 0,76
2	SA_TIntelectuales	H_SintInicial	CPseudobulbar	H_SintInicial	H_SintInicial 0,61
3	AP_ACVA	H_Evolucion	PiramBilateral	H_Evolucion	H_Evolucion 0,48
4	Año_Inicio	Años_escol	Paraparesia	SA_Desorientacion	SA_Desorientacion 0,47
5	H_SintInicial	H_MesesEvolucion	SA_SintFocales	MMSE	SA_TIntelectuales 0,41
6	Estado_civil	H_Comienzo	ParesiaF	SA_TIntelectuales	AP_ACVA 0,41
7	H_MesesEvolucion	SA_SintFocales	SA_Hipersex	H_MesesEvolucion	MAtaxica 0,40
8	Profesion	MMSE	TemblorReposo	Años_escol	H_Comienzo 0,40
9	H_Evolucion	GDS	SA_TM Memoria	H_Comienzo	SA_SintFocales 0,38
10	AP_Tabaco	AP_ACVA	SA_Bulimia	SA_SintFocales	H_MesesEvolucion 0,37
11	GDS	SA_Desorientacion	SA_UrgMiccional	GDS	CPseudobulbar 0,37
12	H_Comienzo	SA_TIntelectuales	Espasticidad	Escolarizacion	Años_escol 0,36
13	SA_Agitacion	CPseudobulbar	H_SintInicial	SA_Hipersex	SA_Hipersex 0,35
14	AP_Hipertension	Escolarizacion	AP_Hepatopatia	AP_ACVA	GDS 0,33
15	Años_escol	ParesiaF	H_Evolucion	Motivo_Consulta	TemblorReposo 0,31
16	SA_Hipersex	Blessed	H_Comienzo	Edad_grupos	MMSE 0,30
17	Edad_grupos	Estado_civil	AP_ACVA	TemblorReposo	Estado_civil 0,29
18	Prension	Espasticidad	PulsosAbolidos	ParesiaF	Paraparesia 0,29
19	Sexo	Edad_grupos	AF_Huntington	SA_Temblor	PiramBilateral 0,28
20	AP_Trauma	MAtaxica	SA_Desorientacion	CPseudobulbar	ParesiaF 0,27
21	SA_Confusión	SA_Hipersex	SuperfAnormal	Provincia	SA_Bulimia 0,27
22	Vive_con	Hachinski	SA_Temblor	MEC	Espasticidad 0,27
23	Periorales	Vive_con	SA_TIntelectuales	SA_IdDelirantes	SA_Temblor 0,26

Segue ...

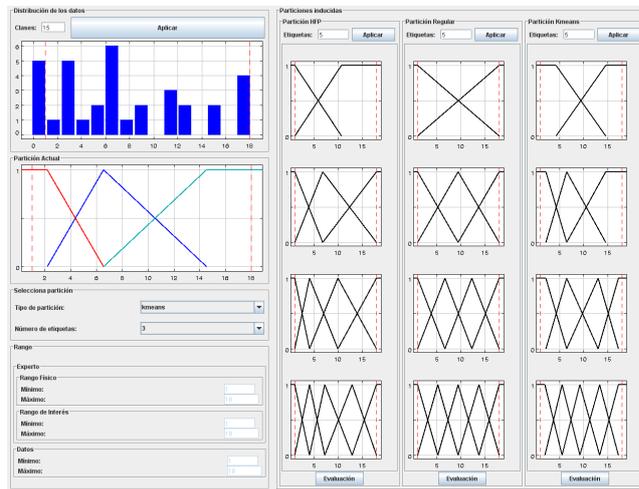
Tabla A.2: Ranking de las características, según distintos métodos de evaluación (continuación)

#	Relief	InfoGain	GainRatio	GiniGain	Recuento Borda	
24	MApraxica	Barthel	Hepatomegalia	MEC2	AP_Tabaco	0,25
25	Soplos	Motivo_Consulta	Motivo_Consulta	Hachinski	Motivo_Consulta	0,25
26	SA_TLenguaje	MEC	PiramUnilateral	FAB	Edad_grupos	0,25
27	AgAuditiva	MMSEesperado	Año_Inicio	RigPlastica	Profesion	0,25
28	TembloReposo	MEC2	SA_RisaLLanto	SA_TMemoria	Vive_con	0,23
29	SA_Temblo	PiramBilateral	ProfunAnormal	SA_Bulimia	Escolarizacion	0,23
30	SA_IncontFecal	Paraparesia	MParetica	Blessed	Provincia	0,22
31	AP_Hipercolest	Lawton	AP_Tabaco	Vive_con	SA_TMemoria	0,21
32	RuedaDentada	FAB	Dismetria	Estado_civil	SA_UrgMiccional	0,21
33	Motivo_Consulta	Provincia	Provincia	Espasticidad	RigPlastica	0,21
34	Glabelar	TembloReposo	GDS	Lawton	SA_Agitacion	0,21
35	RigPlastica	AP_Tabaco	Escolarizacion	MAtaxica	Blessed	0,20
36	AF_TPsiq	RigPlastica	Estado_civil	Profesion	Sexo	0,20
37	Provincia	SA_Temblo	Años_escol	MMSEesperado	MEC	0,20
38	SA_CEPiletpicas	Profesion	MMSE	Barthel	AP_Hipertension	0,20
39	AP_Cardiopatia	SA_TMemoria	RigPlastica	AP_TPsiq	SuperfAnormal	0,19
40	DCampimet	SA_Bulimia	H_MesesEvolucion	PiramBilateral	RuedaDentada	0,19
41	AP_Epilepsia	SA_IdDelirantes	SA_CEPiletpicas	Paraparesia	FAB	0,19
42	MParkinsoniana	MParetica	AP_Epilepsia	RuedaDentada	SA_Confusión	0,19
43	CDR	SuperfAnormal	SA_IdDelirantes	Hamilton	MEC2	0,19
44	SA_Bulimia	Hamilton	MParkinsoniana	CDR	Barthel	0,19
45	SA_SintFocales	CDR	SA_Alucinaciones	AgVisual	Hachinski	0,18
46	AgVisual	Yesavage	AgVisual	PiramUnilateral	Prension	0,18
47	MEC	MParkinsoniana	Vive_con	SA_Agitacion	AP_Trauma	0,18
48	MMSE	SA_UrgMiccional	RuedaDentada	MParetica	CDR	0,17
49	Mioclónias	SA_Alucinaciones	Sexo	AP_Tabaco	MParkinsoniana	0,17
50	Blessed	RuedaDentada	AF_TPsiq	SA_UrgMiccional	Lawton	0,17
51	Oposicionismo	Sexo	Blessed	AP_Hepatopatia	SA_IdDelirantes	0,17
52	FAB	AP_TPsiq	AP_TPsiq	SA_TMarcha	Soplos	0,17
53	TIN	PiramUnilateral	SA_Confusión	Bradicinesia	AgVisual	0,17
54	Barthel	SA_TMarcha	SA_TMarcha	Yesavage	AF_Huntington	0,17
55	Hamilton	AgVisual	Edad_grupos	SuperfAnormal	MApraxica	0,16
56	Escolarizacion	Zarit	Lawton	Amimia	AP_Hepatopatia	0,16
57	Zarit	AP_Hipertension	Hachinski	SA_Alucinaciones	Hamilton	0,16
58	Arritmia	SA_Confusión	Amimia	Zarit	PulsosAbolidos	0,16
59	MEC2	Amimia	Barthel	MParkinsoniana	SA_CEPiletpicas	0,16
60	Espasticidad	AP_Hepatopatia	CDR	TIN	AgAuditiva	0,16
61	AP_Angor	TIN	MEC2	Dismetria	AP_TPsiq	0,16
62	AP_TPsiq	Bradicinesia	MMSEesperado	SA_Depresion	PiramUnilateral	0,16
63	Amimia	SA_Agitacion	MEC	SA_TConducta	Periorales	0,16
64	Yesavage	SA_TConducta	AP_Hipertension	PulsosAbolidos	AP_Epilepsia	0,16
65	Hachinski	Prension	FAB	AF_Huntington	MParetica	0,16
66	AP_Alcohol	Dismetria	Profesion	Sexo	SA_TLenguaje	0,15
67	SA_TSueño	SA_RisaLLanto	Activo	Hepatomegalia	Zarit	0,15
68	Paraparesia	ProfunAnormal	Arritmia	Soplos	AF_TPsiq	0,15
69	Bradicinesia	Activo	Bradicinesia	SA_CEPiletpicas	Yesavage	0,15
70	Lawton	SA_Depresion	AP_Trauma	AP_Epilepsia	SA_Alucinaciones	0,15
71	SA_Alucinaciones	Hepatomegalia	SA_Agitacion	AgAuditiva	TIN	0,15
72	SA_InhibPsic	SA_CEPiletpicas	Soplos	NPLQ	Amimia	0,15
73	SA_IdDelirantes	AP_Epilepsia	AP_EnfTiroidea	SA_Confusión	SA_IncontFecal	0,15
74	PiramBilateral	PulsosAbolidos	Mioclónias	SA_RisaLLanto	Hepatomegalia	0,14
75	AF_Demencia	AF_Huntington	Hamilton	ProfunAnormal	AP_Hipercolest	0,14
76	CPseudobulbar	NPLQ	Zarit	MApraxica	SA_RisaLLanto	0,14
77	AF_Parkinson	AP_EnfTiroidea	SA_TConducta	AP_Cardiopatia	Glabelar	0,14
78	ParesiaF	Soplos	Corea	Prension	MMSEesperado	0,14
79	MAtaxica	AP_TPsiq	AP_Alcohol	AF_TPsiq	Bradicinesia	0,14
80	SuperfAnormal	Arritmia	AF_Parkinson	Arritmia	AP_Cardiopatia	0,14
81	SA_RisaLLanto	AgAuditiva	Prension	AP_Hiperten	ProfunAnormal	0,13
82	PiramUnilateral	MApraxica	AP_InfMIOC	AP_Diabetes	DCampimet	0,13
83	SA_TConducta	AP_Trauma	TIN	Activo	Arritmia	0,13
84	AP_EnfTiroidea	Glabelar	Yesavage	AP_Angor	Mioclónias	0,13
85	SA_IncontUrin	AP_Cardiopatia	TembloActitud	AP_Trauma	Dismetria	0,13
86	TembloActitud	DCampimet	DCampimet	Glabelar	SA_TMarcha	0,13
87	MParetica	DCampimet	AP_Otros	SA_TSueño	Oposicionismo	0,12
88	SA_TMarcha	AP_InfMIOC	SA_Depresion	SA_IncontFecal	AP_Angor	0,12
89	Hepatomegalia	AP_Diabetes	AgAuditiva	Mioclónias	AP_Alcohol	0,12
90	SA_Depresion	AF_Parkinson	MApraxica	TembloActitud	SA_TSueño	0,11
91	SA_UrgMiccional	SA_TSueño	AP_Cardiopatia	DCampimet	SA_TConducta	0,11
92	ProfunAnormal	Mioclónias	AP_Diabetes	AP_EnfTiroidea	SA_Depresion	0,11
93	AP_Diabetes	AP_Otros	AP_Angor	SA_TLenguaje	SA_InhibPsic	0,10
94	AF_Huntington	SA_IncontUrin	Glabelar	AP_Hipercolest	AF_Parkinson	0,10
95	Activo	AP_Angor	SA_TSueño	AF_Demencia	AF_Demencia	0,10
96	Completo	SA_TLenguaje	SA_IncontUrin	SA_InhibPsic	AP_EnfTiroidea	0,10
97	PulsosAbolidos	AF_Demencia	SA_TLenguaje	AF_Parkinson	Activo	0,10
98	Papiledema	Periorales	AF_Demencia	Periorales	TembloActitud	0,10
99	AP_Down	Corea	Oposicionismo	Corea	AP_Diabetes	0,09
100	AP_HIV	AP_Alcohol	Periorales	AP_Alcohol	SA_IncontUrin	0,09
101	AF_Down	Oposicionismo	AP_Hipercolest	AP_InfMIOC	Corea	0,08
102	Dismetria	AP_Hipercolest	SA_IncontFecal	SA_IncontUrin	Completo	0,08
103	Corea	SA_IncontFecal	SA_InhibPsic	Oposicionismo	Papiledema	0,08
104	MMSEesperado	SA_InhibPsic	NPLQ	AP_Otros	AP_Down	0,08
105	AP_InfMIOC	Completo	Completo	Completo	AP_HIV	0,08
106	SA_TMemoria	Papiledema	Papiledema	Papiledema	AF_Down	0,08
107	AP_Hepatopatia	AP_Down	AP_Down	AP_Down	AP_InfMIOC	0,08
108	AP_Otros	AP_HIV	AP_HIV	AP_HIV	AP_Otros	0,06
109	NPLQ	AF_Down	AF_Down	AF_Down	NPLQ	0,01

Ranking de las características, según distintos métodos de evaluación (Fin)

Apéndice B

Inducción de particiones



(a) Particiones inducidas para “Hachinski”.

Partición Actual			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
3	0,69102	0,45642	0,68627

Partición HFP			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,78822	0,30976	0,65762
3	0,76977	0,45902	0,71291
4	0,69320	0,45470	0,74126
5	0,69062	0,45355	0,74484

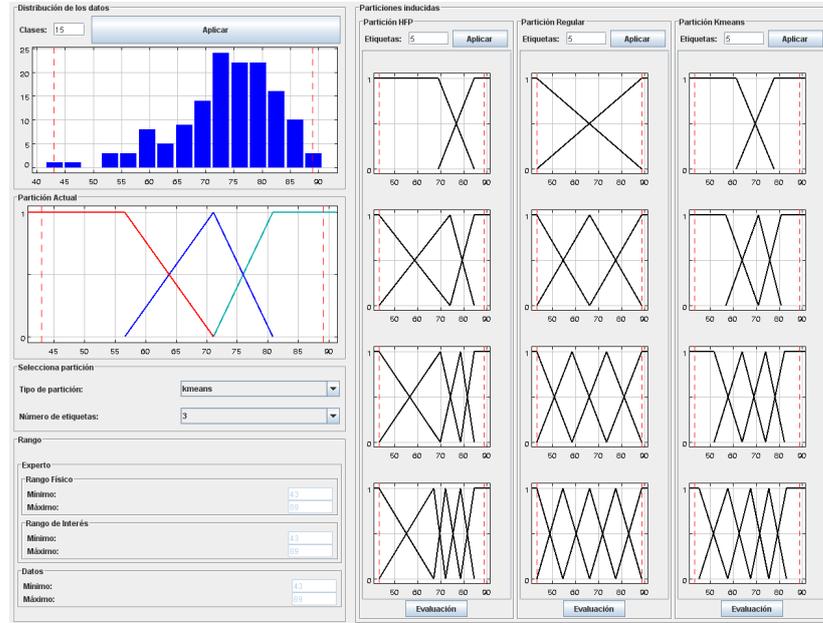
Partición Regular			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,72725	0,41405	0,60900
3	0,69102	0,45642	0,68627
4	0,72522	0,41728	0,76586
5	0,70527	0,44148	0,75630

Partición Kmeans			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,84950	0,23202	0,78919
3	0,78286	0,33217	0,80059
4	0,76995	0,35484	0,80967
5	0,76421	0,36620	0,81756

(b) Evaluación de las particiones para “Hachinski”.

Figura B.1: Inducción de particiones para la variable “Hachinski”.

ApéndiceB : Inducción de particiones



(a) Particiones inducidas para “Año.Inicio”.

Partición Actual			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
3	0,74412	0,38479	0,74816

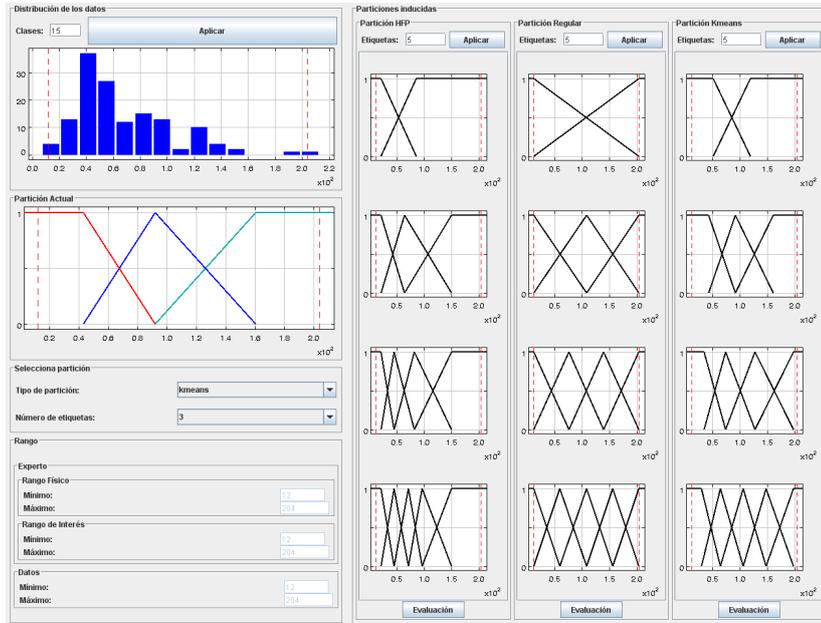
Partición HFP			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,74967	0,36819	0,61753
3	0,71509	0,42876	0,71351
4	0,69272	0,45808	0,72948
5	0,70165	0,44878	0,75630

Partición Regular			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,62854	0,54849	0,44272
3	0,62873	0,54370	0,61589
4	0,68234	0,47785	0,72079
5	0,66726	0,49805	0,72780

Partición Kmeans			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,82125	0,26884	0,73748
3	0,74412	0,38479	0,74816
4	0,72829	0,40882	0,76399
5	0,70286	0,44632	0,75675

(b) Evaluación de las particiones para “Año.Inicio”.

Figura B.2: Inducción de particiones para la variable “Año.Inicio”.



(a) Particiones inducidas para “H_Meses_Evolución”.

Partición Actual			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
3	0,63621	0,53875	0,62027

Partición HFP			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,76505	0,35024	0,64484
3	0,69878	0,45828	0,71079
4	0,69425	0,45581	0,72419
5	0,69456	0,44809	0,71411

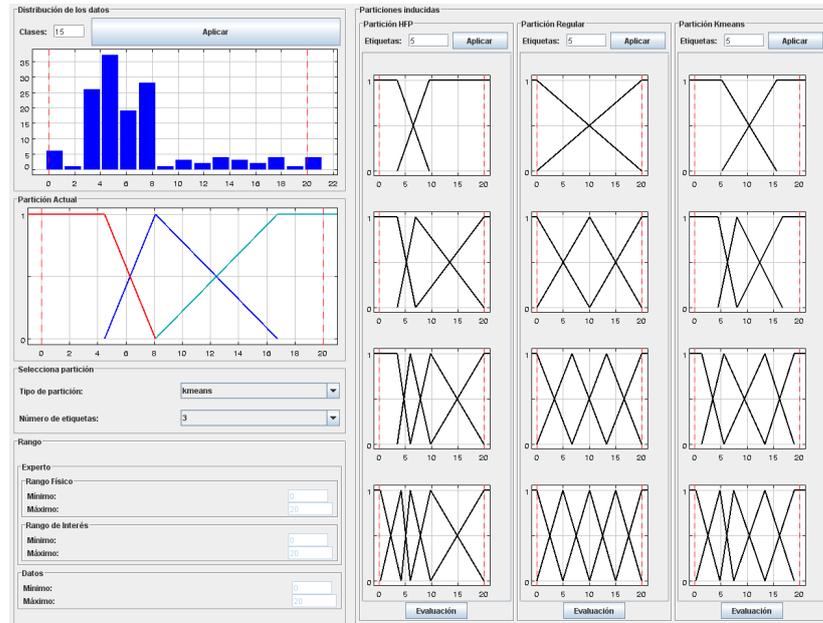
Partición Regular			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,65761	0,51402	0,49712
3	0,63621	0,53875	0,62027
4	0,65032	0,52321	0,70226
5	0,68401	0,46009	0,72061

Partición Kmeans			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,82621	0,26027	0,73542
3	0,76319	0,35720	0,76830
4	0,72783	0,41062	0,76728
5	0,70269	0,45131	0,75721

(b) Evaluación de las particiones para “H_Meses_Evolución”.

Figura B.3: Inducción de particiones para la variable “H_Meses_Evolución”.

ApéndiceB : Inducción de particiones



(a) Particiones inducidas para “Años_Escol”.

Partición Actual			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
3	0,59106	0,58447	0,53948

Partición HFP			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,72872	0,40438	0,60023
3	0,71220	0,42756	0,70523
4	0,68841	0,45224	0,71343
5	0,69688	0,44174	0,72652

Partición Regular			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,64280	0,52924	0,49007
3	0,59106	0,58447	0,53948
4	0,69585	0,46232	0,74805
5	0,74562	0,37437	0,79248

Partición Kmeans			
Etiquetas	Coefficiente de la partición (max)	Entropía de la partición (min)	Índice de Chen (max)
2	0,88396	0,18463	0,84694
3	0,78140	0,33345	0,78876
4	0,71613	0,43745	0,76920
5	0,74716	0,39718	0,80249

(b) Evaluación de las particiones para “Años_Escol”.

Figura B.4: Inducción de particiones para la variable “Años_Escol”.

Apéndice C

Resultados completos

Detalle de los resultados completos obtenidos durante la ejecución de los diferentes algoritmos empleados en este trabajo.

C.1. Árboles de Decisión

Como se ha comentado en el texto de la memoria, se comienza con los árboles de decisión para establecerlos como caso base para el resto de comparaciones.

C.1.1. J4.8

En la primera prueba se utilizó el conjunto de datos completo, puesto que los árboles por su propia naturaleza, realizan una selección de características a medida que se van formando, eligiendo primero a las variables más relevantes. Como este conjunto de datos no es adecuado para otros tipos de algoritmos, se ha ido depurando y se ha vuelto a pasar el nuevo conjunto de datos por el J4.8, para comprobar si había mucha diferencia de los resultados de unos conjuntos a otros.

Los parámetros seleccionados para la ejecución del algoritmo J4.8 fueron los siguientes:

- *C*: *confidenceFactor*. El factor de confianza se usa para la poda del árbol, de manera que cuanto menor es el valor más se poda el árbol. $C = 0,25$
- *M*: *minNumObj*. Indica el número mínimo de instancias por hoja. $M = 2$

```
=== Run information ===  
  
Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2  
Instances:   141  
Attributes:  117  
              [list of attributes omitted]  
Test mode:   10-fold cross-validation
```

ApéndiceC : Resultados completos

```

=== Classifier model (full training set) ===

J48 pruned tree
-----

TemblorReposo = Ausente
|
| Paraparesia = Ausente
| |
| | Edad_2 <= 1: Otras demencias (7.0/1.0)
| | Edad_2 > 1
| | |
| | | ParesiaF = Ausente
| | | |
| | | | SA_SintFocales = Ausente
| | | | |
| | | | | Motivo_Consulta = Trastorno de memoria
| | | | | |
| | | | | | H_Evolucion = Progresiva
| | | | | | |
| | | | | | | AP_ACVA = Ausente
| | | | | | | |
| | | | | | | | SA_Tintelectuales = Presente: Enfermedad de Alzheimer (50.0/2.0)
| | | | | | | | SA_Tintelectuales = Ausente
| | | | | | | | |
| | | | | | | | | Hachinski =
| | | | | | | | | |
| | | | | | | | | | SA_IdDelirantes = Ausente: Enfermedad de Alzheimer (21.0/3.0)
| | | | | | | | | | SA_IdDelirantes = Presente
| | | | | | | | | | |
| | | | | | | | | | | estancia_dias <= 234: Enfermedad de Alzheimer (2.0)
| | | | | | | | | | | estancia_dias > 234: Enfermedad con cuerpos de Levy (4.0/2.0)
| | | | | | | | | | |
| | | | | | | | | | | Hachinski = 17.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 15.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 13.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 1.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 6.0: Demencia Mixta (1.0)
| | | | | | | | | | | Hachinski = 12.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 3.0: Demencia Mixta (3.0)
| | | | | | | | | | | Hachinski = 18.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 9.0: Demencia Mixta (1.0)
| | | | | | | | | | | Hachinski = 5.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 4.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 2.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 11.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | Hachinski = 8.0: Enfermedad de Alzheimer (1.0)
| | | | | | | | | | |
| | | | | | | | | | | AP_ACVA = Presente
| | | | | | | | | | | |
| | | | | | | | | | | | SA_TSuenyo = Ausente: Demencia Mixta (7.0/1.0)
| | | | | | | | | | | | SA_TSuenyo = Presente: Enfermedad de Alzheimer (5.0)
| | | | | | | | | | | | H_Evolucion = Intermitente: Enfermedad de Alzheimer (2.0)
| | | | | | | | | | | | H_Evolucion = Estable: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | | H_Evolucion = Fluctuante
| | | | | | | | | | | | |
| | | | | | | | | | | | | AP_Angor = Ausente
| | | | | | | | | | | | | |
| | | | | | | | | | | | | | SA_Tintelectuales = Presente: Enfermedad con cuerpos de Levy (4.0)
| | | | | | | | | | | | | | SA_Tintelectuales = Ausente: Demencia Mixta (4.0/1.0)
| | | | | | | | | | | | | | AP_Angor = Presente: Enfermedad de Alzheimer (3.0/1.0)
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | Motivo_Consulta = Trastorno de la marcha: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | | | | | Motivo_Consulta = Trastorno de conducta
| | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | SA_Desorientacion = Presente: Enfermedad de Alzheimer (5.0/1.0)
| | | | | | | | | | | | | | | | SA_Desorientacion = Ausente: Otras demencias (3.0)
| | | | | | | | | | | | | | | | Motivo_Consulta = Sintomas focales: Enfermedad de Alzheimer (0.0)
| | | | | | | | | | | | | | | | SA_SintFocales = Presente
| | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | Recurso = Residencial temporal: Otras demencias (1.0)
| | | | | | | | | | | | | | | | | Recurso = Centro de dia: Demencia vascular (2.0)
| | | | | | | | | | | | | | | | | Recurso = Modulo de respiro familiar: Enfermedad con cuerpos de Levy (2.0/1.0)
| | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | ParesiaF = Presente: Otras demencias (5.0/1.0)
| | | | | | | | | | | | | | | | | Paraparesia = Presente: Demencia vascular (3.0)
| | | | | | | | | | | | | | | | | TemblorReposo = Presente: Otras demencias (5.0)

Number of Leaves : 36
Size of the tree : 53

```

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	85	60.2837 %
Incorrectly Classified Instances	56	39.7163 %
Kappa statistic	0.3225	
Mean absolute error	0.181	
Root mean squared error	0.3778	
Relative absolute error	73.4193 %	
Root relative squared error	108.1807 %	
Total Number of Instances	141	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.817	0.373	0.753	0.817	0.784	0.752	Enfermedad de Alzheimer
	0.5	0.067	0.579	0.5	0.537	0.646	Otras demencias
	0.25	0.053	0.222	0.25	0.235	0.57	Demencia vascular
	0.125	0.03	0.2	0.125	0.154	0.474	Enfermedad con cuerpos de Levy
	0.19	0.125	0.211	0.19	0.2	0.475	Demencia Mixta
Weighted Avg.	0.603	0.251	0.583	0.603	0.591	0.668	

C.1 Árboles de Decisión

```
=== Confusion Matrix ===

 a  b  c  d  e  <-- classified as
67  3  1  1 10 | a = Enfermedad de Alzheimer
 5 11  4  1  1 | b = Otras demencias
 4  1  2  0  1 | c = Demencia vascular
 3  1  0  1  3 | d = Enfermedad con cuerpos de Lewy
10  3  2  2  4 | e = Demencia Mixta
```

Para centrarnos en AD y DV podemos juntar el resto de casos en “Otras demencias”, a ver si así mejora un poco el error de clasificación.

Los parámetros seleccionados para la ejecución del algoritmo J4.8 fueron los siguientes:

- C : *confidenceFactor*. $C = 0,25$
- M : *minNumObj*. $M = 2$

```
=== Run information ===

Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Instances:   141
Attributes:  117
              [list of attributes omitted]
Test mode:   10-fold cross-validation
```

```
=== Classifier model (full training set) ===

J48 pruned tree
-----
MAtaxica = Ausente
|
| SA_SintFocales = Ausente
| |
| | SA_Hipersex = Ausente
| | |
| | | Edad_2 <= 1: Otras demencias (4.0)
| | | Edad_2 > 1
| | | |
| | | | TemblorReposo = Ausente
| | | | |
| | | | | H_Comienzo = Insidioso
| | | | | |
| | | | | | Hachinski =
| | | | | | |
| | | | | | | Dismetria = Ausente
| | | | | | | |
| | | | | | | | estancia_dias <= 578: Enfermedad de Alzheimer (74.0/6.0)
| | | | | | | | estancia_dias > 578: Otras demencias (6.0/1.0)
| | | | | | | | Dismetria = Presente: Otras demencias (3.0/1.0)
| | | | | | | Hachinski = 17.0: Enfermedad de Alzheimer (0.0)
| | | | | | | Hachinski = 15.0: Enfermedad de Alzheimer (0.0)
| | | | | | | Hachinski = 13.0: Demencia vascular (1.0)
| | | | | | | Hachinski = 1.0: Enfermedad de Alzheimer (5.0)
| | | | | | | Hachinski = 6.0: Otras demencias (4.0/1.0)
| | | | | | | Hachinski = 12.0: Otras demencias (2.0)
| | | | | | | Hachinski = 3.0
| | | | | | | |
| | | | | | | | AP_Hipertension = Presente: Enfermedad de Alzheimer (2.0)
| | | | | | | | AP_Hipertension = Ausente: Otras demencias (3.0)
| | | | | | | | Hachinski = 18.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | Hachinski = 9.0: Enfermedad de Alzheimer (2.0/1.0)
| | | | | | | | Hachinski = 5.0: Demencia vascular (1.0)
| | | | | | | | Hachinski = 4.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | Hachinski = 2.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | Hachinski = 11.0: Enfermedad de Alzheimer (0.0)
| | | | | | | | Hachinski = 8.0: Enfermedad de Alzheimer (0.0)
| | | | | | | H_Comienzo = Subagudo
| | | | | | | |
| | | | | | | | AP_Angor = Ausente: Otras demencias (5.0)
| | | | | | | | AP_Angor = Presente: Enfermedad de Alzheimer (3.0)
| | | | | | | | H_Comienzo = Agudo: Otras demencias (4.0)
| | | | | | | | TemblorReposo = Presente: Otras demencias (2.0)
| | | | | | | SA_Hipersex = Presente: Otras demencias (6.0)
| | | | | | SA_SintFocales = Presente
| | | | | | |
| | | | | | | Anyo_Inicio <= 66: Demencia vascular (3.0)
| | | | | | | Anyo_Inicio > 66: Otras demencias (9.0)
| MAtaxica = Presente: Demencia vascular (2.0)

Number of Leaves : 27
Size of the tree : 39
```

ApéndiceC : Resultados completos

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      91           64.539 %
Incorrectly Classified Instances    50           35.461 %
Kappa statistic                    0.2928
Mean absolute error                 0.2575
Root mean squared error            0.4539
Relative absolute error            73.3148 %
Root relative squared error        108.6027 %
Total Number of Instances          141
```

```
=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.854    0.475    0.714    0.854    0.778    0.738    Enfermedad de Alzheimer
      0.385    0.169    0.571    0.385    0.46     0.613    Otras demencias
      0.143    0.052    0.125    0.143    0.133    0.618    Demencia vascular
Weighted Avg. 0.645    0.341    0.632    0.645    0.629    0.686
```

```
=== Confusion Matrix ===

  a  b  c  <-- classified as
70 11  1 | a = Enfermedad de Alzheimer
26 20  6 | b = Otras demencias
 2  4  1 | c = Demencia vascular
```

Árbol con las características seleccionadas por el experto, pero con todas las variedades de demencia. Los parámetros seleccionados para la ejecución del algoritmo J4.8 fueron los siguientes:

- *C*: *confidenceFactor*. $C = 0,25$
- *M*: *minNumObj*. $M = 2$

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Instances: 141
Attributes: 28
           [list of attributes omitted]
Test mode:10-fold cross-validation
```

C.1 Árboles de Decisión

```

=== Classifier model (full training set) ===

J48 pruned tree
-----

CPseudobulbar = Ausente
| SA_SintFocales = Ausente
| | Hachinski =
| | | H_Comienzo = Insidioso
| | | | AP_Tabaco = Ausente: Enfermedad de Alzheimer (78.0/15.0)
| | | | AP_Tabaco = Presente
| | | | | AP_Cardiopatia = Ausente
| | | | | | N_Hist <= 71: Otras demencias (4.0)
| | | | | | N_Hist > 71: Enfermedad de Alzheimer (4.0/2.0)
| | | | | AP_Cardiopatia = Presente: Enfermedad de Alzheimer (5.0)
| | | | H_Comienzo = Subagudo
| | | | | Sexo = Mujer
| | | | | | AP_Angor = Ausente: Enfermedad con cuerpos de Lewy (3.0)
| | | | | | AP_Angor = Presente: Enfermedad de Alzheimer (2.0)
| | | | | | Sexo = Varon: Otras demencias (3.0/1.0)
| | | | | H_Comienzo = Agudo: Enfermedad con cuerpos de Lewy (2.0/1.0)
| | | | Hachinski = 17.0: Enfermedad de Alzheimer (0.0)
| | | | Hachinski = 15.0: Demencia Mixta (1.0)
| | | | Hachinski = 13.0: Enfermedad de Alzheimer (0.0)
| | | | Hachinski = 1.0: Enfermedad de Alzheimer (5.0)
| | | | Hachinski = 6.0: Demencia Mixta (5.0/1.0)
| | | | Hachinski = 12.0: Demencia Mixta (1.0)
| | | | Hachinski = 3.0
| | | | | AP_Hipertension = Presente: Enfermedad de Alzheimer (2.0)
| | | | | AP_Hipertension = Ausente: Demencia Mixta (3.0)
| | | | Hachinski = 18.0: Enfermedad de Alzheimer (0.0)
| | | | Hachinski = 9.0: Enfermedad de Alzheimer (2.0/1.0)
| | | | Hachinski = 5.0: Demencia vascular (2.0/1.0)
| | | | Hachinski = 4.0: Otras demencias (1.0)
| | | | Hachinski = 2.0: Otras demencias (1.0)
| | | | Hachinski = 11.0: Demencia Mixta (1.0)
| | | | Hachinski = 8.0: Enfermedad de Alzheimer (1.0)
| | SA_SintFocales = Presente
| | | MApraxica = Ausente: Otras demencias (8.0/2.0)
| | | MApraxica = Presente: Demencia vascular (3.0/1.0)
| CPseudobulbar = Presente: Demencia vascular (4.0/1.0)

Number of Leaves : 26
Size of the tree : 37

```

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      90          63.8298 %
Incorrectly Classified Instances    51          36.1702 %
Kappa statistic                    0.3311
Mean absolute error                 0.1852
Root mean squared error            0.3521
Relative absolute error            75.1126 %
Root relative squared error        100.8197 %
Total Number of Instances          141

```

```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.878	0.525	0.699	0.878	0.778	0.746	Enfermedad de Alzheimer
	0.409	0.059	0.563	0.409	0.474	0.662	Otras demencias
	0.25	0.023	0.4	0.25	0.308	0.703	Demencia vascular
	0.25	0.023	0.4	0.25	0.308	0.439	Enfermedad con cuerpos de Lewy
	0.238	0.058	0.417	0.238	0.303	0.533	Demencia Mixta
Weighted Avg.	0.638	0.326	0.602	0.638	0.607	0.682	

```

=== Confusion Matrix ===

 a  b  c  d  e  <-- classified as
72  2  0  2  6 | a = Enfermedad de Alzheimer
13  9  0  0  0 | b = Otras demencias
 3  3  2  0  0 | c = Demencia vascular
 4  1  0  2  1 | d = Enfermedad con cuerpos de Lewy
11  1  3  1  5 | e = Demencia Mixta

```

Resultados del árbol obtenido mediante el algoritmo J4.8, con el conjunto de características seleccionadas por el experto y las variedades de demencia reducidas a tres grupos.

ApéndiceC : Resultados completos

Los parámetros seleccionados para la ejecución del algoritmo J4.8 fueron los siguientes:

- C: *confidenceFactor*. $C = 0,25$
- M: *minNumObj*. $M = 2$

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Instances:    141
Attributes:   35
              [list of attributes omitted]
Test mode:10-fold cross-validation
```

```
=== Classifier model (full training set) ===

J48 pruned tree
-----
Hachinski=alto <= 0
| SA_SintFocales <= 0
| | H_Evolucion=Fluctuante <= 0
| | | H_Comienzo=Insidioso <= 0: OtrasDem (4.0/1.0)
| | | H_Comienzo=Insidioso > 0
| | | | Hachinski=medio <= 0: EA (100.0/24.0)
| | | | Hachinski=medio > 0
| | | | | AP_Hipercolest <= 0: EA (2.0)
| | | | | AP_Hipercolest > 0: OtrasDem (3.0)
| | | | H_Evolucion=Fluctuante > 0
| | | | | AP_Angor <= 0: OtrasDem (11.0/1.0)
| | | | | AP_Angor > 0: EA (3.0/1.0)
| | SA_SintFocales > 0: OtrasDem (7.0)
Hachinski=alto > 0
| Bradicinesia <= 0
| | CPseudobulbar <= 0: OtrasDem (5.0)
| | CPseudobulbar > 0: DV (4.0/1.0)
| | Bradicinesia > 0: DV (2.0)

Number of Leaves : 10
Size of the tree : 19
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          98           69.5035 %
Incorrectly Classified Instances        43           30.4965 %
Kappa statistic                        0.3749
Mean absolute error                    0.2584
Root mean squared error                0.3962
Relative absolute error                 73.5652 %
Root relative squared error            94.8056 %
Total Number of Instances              141
```

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.915	0.492	0.721	0.915	0.806	0.705	EA
	0.385	0.124	0.645	0.385	0.482	0.609	OtrasDem
	0.429	0.022	0.5	0.429	0.462	0.973	DV
Weighted Avg.	0.695	0.333	0.682	0.695	0.67	0.683	

```
=== Confusion Matrix ===

 a  b  c  <-- classified as
75  7  0 | a = EA
29 20  3 | b = OtrasDem
 0  4  3 | c = DV
```

C.1 Árboles de Decisión

Resultados del árbol obtenido mediante el algoritmo J4.8, con el conjunto de características seleccionadas mediante recuento Borda y las variedades de demencia reducidas a tres grupos.

Los parámetros seleccionados para la ejecución del algoritmo J4.8 fueron los siguientes:

- *C*: *confidenceFactor*. $C = 0,25$
- *M*: *minNumObj*. $M = 2$

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Instances:    141
Attributes:   54
              [list of attributes omitted]
Test mode:10-fold cross-validation
```

```
=== Classifier model (full training set) ===

J48 pruned tree
-----
SA_SintFocales <= 0
| SA_Hipersex <= 0
| | H_SintInicial=Trastorno_de_Conducta <= 0
| | | AP_ACVA <= 0
| | | | SA_Desorientacion <= 0: EA (87.0/16.0)
| | | | SA_Desorientacion > 0
| | | | | SA_TIntelectuales <= 0: EA (4.0/1.0)
| | | | | SA_TIntelectuales > 0: OtrasDem (6.0/1.0)
| | | | AP_ACVA > 0
| | | | | MMSE=bajo <= 0
| | | | | Anyos_escol=bajo <= 0: EA (4.0/2.0)
| | | | | Anyos_escol=bajo > 0: OtrasDem (5.0/1.0)
| | | | | MMSE=bajo > 0: EA (5.0/1.0)
| | | | H_SintInicial=Trastorno_de_Conducta > 0: OtrasDem (12.0/2.0)
| | | SA_Hipersex > 0: OtrasDem (6.0)
| SA_SintFocales > 0
| | Anyo_Inicio=G1 <= 0: OtrasDem (9.0)
| | Anyo_Inicio=G1 > 0: DV (3.0)

Number of Leaves : 10
Size of the tree : 19
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      92           65.2482 %
Incorrectly Classified Instances    49           34.7518 %
Kappa statistic                     0.2964
Mean absolute error                  0.2651
Root mean squared error              0.4311
Relative absolute error              75.4751 %
Root relative squared error          103.1552 %
Total Number of Instances          141
```

```
=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.805   0.475   0.702     0.805   0.75       0.704    EA
      0.481   0.225   0.556     0.481   0.515     0.642    OtrasDem
      0.143   0.007   0.5       0.143   0.222     0.57     DV
Weighted Avg.   0.652   0.359   0.638     0.652   0.637     0.674
```

```
=== Confusion Matrix ===

 a  b  c  <-- classified as
66 16  0 | a = EA
26 25  1 | b = OtrasDem
 2  4  1 | c = DV
```

ApéndiceC : Resultados completos

C.1.2. Random Forest

Variables seleccionadas por el experto. Los parámetros seleccionados para la ejecución del algoritmo **Random Forest** fueron los siguientes:

- I : *numTrees*. Número de árboles que serán generados. $I = 10$
- K : *maxDepth*. La profundidad máxima de los árboles, será ilimitada si el valor es 0. $K = 0$
- S : *seed*. Semilla para usar en la generación de números aleatorios. $S = 1$

```
=== Run information ===
Scheme:weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1
Relation: CREA_Dataset_N3_Fuzzy_Experto_Bin
Instances: 141
Attributes: 35
           [list of attributes omitted]
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 6 random features.
Out of bag error: 0.3688
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          90           63.8298 %
Incorrectly Classified Instances        51           36.1702 %
Kappa statistic                        0.2631
Mean absolute error                     0.2787
Root mean squared error                 0.403
Relative absolute error                 79.3292 %
Root relative squared error            96.4351 %
Total Number of Instances              141
```

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.829	0.525	0.687	0.829	0.751	0.752	EA
	0.385	0.202	0.526	0.385	0.444	0.652	OtrasDem
	0.286	0.015	0.5	0.286	0.364	0.735	DV
Weighted Avg.	0.638	0.381	0.618	0.638	0.619	0.714	

```
=== Confusion Matrix ===

  a  b  c  <-- classified as
68 14  0 | a = EA
30 20  2 | b = OtrasDem
 1  4  2 | c = DV
```

Variables seleccionadas mediante recuento Borda. Los parámetros seleccionados para la ejecución del algoritmo **Random Forest** fueron los siguientes:

- I : *numTrees*. Número de árboles que serán generados. $I = 10$
- K : *maxDepth*. La profundidad máxima de los árboles, será ilimitada si el valor es 0. $K = 0$
- S : *seed*. Semilla para usar en la generación de números aleatorios. $S = 1$

C.2 Redes Neuronales Artificiales ANN

```
=== Run information ===
Scheme:weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1
Relation:   CREA_Dataset_N3_Fuzzy_Borda_Bin
Instances:  141
Attributes: 54
           [list of attributes omitted]
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 10 trees, each constructed while considering 6 random features.
Out of bag error: 0.2908

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      110           78.0142 %
Incorrectly Classified Instances    31           21.9858 %
Kappa statistic                    0.5548
Mean absolute error                 0.2297
Root mean squared error             0.3496
Relative absolute error             65.3998 %
Root relative squared error        83.6572 %
Total Number of Instances          141

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.939	0.288	0.819	0.939	0.875	0.856	EA
	0.635	0.135	0.733	0.635	0.68	0.785	OtrasDem
	0	0.015	0	0	0	0.766	DV
Weighted Avg.	0.78	0.218	0.747	0.78	0.76	0.825	

```

=== Confusion Matrix ===

 a  b  c  <-- classified as
77  5  0 | a = EA
17 33  2 | b = OtrasDem
 0  7  0 | c = DV
```

C.2. Redes Neuronales Artificiales ANN

C.2.1. MLP

Los parámetros seleccionados para la ejecución del algoritmo MLP, con las variables seleccionadas por el experto, fueron los siguientes:

- L : *learningRate*. La cantidad con la que los pesos son actualizados. $L = 0,3$
- M : *momentum*. Momento aplicado a los pesos durante su actualización. $M = 0,2$
- N : *trainingTime*. En número de épocas durante las que se va a entrenar. Si el conjunto de validación no es cero, entonces se puede terminar antes el entrenamiento de la red. $N = 500$
- V : *validationSetSize*. El tamaño, en porcentaje, del conjunto de validación. El entrenamiento continuará hasta que se observe que el error en el conjunto de validación a ido siendo sucesivamente peor, o hasta que se termine el tiempo de entrenamiento. Si el valor es 0, no se usará ningún conjunto de validación y en su lugar la red será entrenada durante el número de épocas especificado. $V = 0$

ApéndiceC : Resultados completos

- *S*: *seed*. Semilla utilizada para inicializar el generador de números aleatorios. Los números aleatorios se usan para establecer los pesos iniciales de las conexiones entre los nodos, y también para mezclar los datos del conjunto de entrenamiento. $S = 0$
- *E*: *validationThreshold*. Se emplea para finalizar las pruebas de validación. El valor dictamina cuántas veces en una fila el error del conjunto de validación puede empeorar antes de dar por terminado el entrenamiento. $E = 20$
- *H*: *hiddenLayers*. Define las capas ocultas de la red neuronal. Es una lista de números positivos. 1 para cada capa oculta. La lista debe ir separada por comas. Para no tener capas ocultas, se debe poner el valor 0. Este parámetro sólo se usará si se ha seleccionado *autobuild*. Existen comodines 'a' = (atributos + clases) / 2, 'i' = atributos, 'o' = clases, 't' = atributos + clases. $H = a$

```
==== Run information ====
Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Instances: 141
Attributes: 35
           [list of attributes omitted]
Test mode:10-fold cross-validation
```

```
==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances      85          60.2837 %
Incorrectly Classified Instances    56          39.7163 %
Kappa statistic                    0.2185
K&B Relative Info Score           3537.4426 %
K&B Information Score              42.9981 bits    0.305 bits/instance
Class complexity | order 0         169.8671 bits    1.2047 bits/instance
Class complexity | scheme          330.8508 bits    2.3465 bits/instance
Complexity improvement (Sf)        -160.9836 bits   -1.1417 bits/instance
Mean absolute error                0.2752
Root mean squared error            0.4774
Relative absolute error            78.334 %
Root relative squared error        114.2426 %
Total Number of Instances         141
```

```
==== Detailed Accuracy By Class ====
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.707	0.475	0.674	0.707	0.69	0.659	Enfermedad de Alzheimer
	0.462	0.315	0.462	0.462	0.462	0.562	Otras demencias
	0.429	0	1	0.429	0.6	0.835	Demencia vascular
Weighted Avg.	0.603	0.392	0.612	0.603	0.602	0.632	

```
==== Confusion Matrix ====

 a  b  c  <-- classified as
58 24  0 | a = Enfermedad de Alzheimer
28 24  0 | b = Otras demencias
 0  4  3 | c = Demencia vascular
```

Los parámetros seleccionados para la ejecución del algoritmo MLP, con las variables seleccionadas mediante recuento Borda, fueron los siguientes:

- *L*: *learningRate*. $L = 0,3$
- *M*: *momentum*. $M = 0,2$
- *N*: *trainingTime*. $N = 500$
- *V*: *validationSetSize*. $V = 0$

C.2 Redes Neuronales Artificiales ANN

- S: *seed*. $S = 0$
- E: *validationThreshold*. $E = 20$
- H: *hiddenLayers*. $H = a$

```

=== Run information ===
Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation:   CREA_Dataset_N3_Fuzzy_Borda_Bin
Instances:  141
Attributes: 54
            [list of attributes omitted]
Test mode:10-fold cross-validation
    
```

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      104           73.7589 %
Incorrectly Classified Instances     37           26.2411 %
Kappa statistic                     0.489
Mean absolute error                  0.1761
Root mean squared error              0.3767
Relative absolute error              50.1298 %
Root relative squared error         90.1285 %
Total Number of Instances           141
    
```

```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.817	0.288	0.798	0.817	0.807	0.823	EA
	0.673	0.202	0.666	0.673	0.667	0.783	OtrasDem
	0.286	0.015	0.5	0.286	0.364	0.764	DV
Weighted Avg.	0.738	0.243	0.732	0.738	0.733	0.806	

```

=== Confusion Matrix ===

 a  b  c  <-- classified as
67 14  1 | a = EA
16 35  1 | b = OtrasDem
 1  4  2 | c = DV
    
```

C.2.2. RBF

Variables Experto

Los parámetros seleccionados para la ejecución del algoritmo RBF, con las variables seleccionadas por el experto, fueron los siguientes:

- B: *numClusters*. El número de clusters que debe generar K-Means. $B = 3$
- S: *clusteringSeed*. La semilla aleatoria que se debe pasar a K-means. $S = 1$
- R: *ridge*. Establece el valor de la cresta para la regresión lineal o logística. $R = 1,0E - 8$
- M: *maxIts*. Número máximo de iteraciones que debe realizar la regresión logística. Sólo se aplica a problemas de clasificación discreta. $M = -1$
- W: *minStdDev*. Establece la desviación estándar mínima para los clusters. $W = 0,1$

ApéndiceC : Resultados completos

```

=== Run information ===

Scheme:weka.classifiers.functions.RBFNetwork -B 3 -S 1 -R 1.0E-8 -M -1 -W 0.1
Instances: 141
Attributes: 35
          [list of attributes omitted]
Test mode:10-fold cross-validation

```

```

=== Classifier model (full training set) ===

Radial basis function network
(Logistic regression applied to K-means clusters as basis functions):

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...

```

Variable	Class	
	EA	OtrasDem
pCluster_0_0	-3681.772	-3683.3883
pCluster_0_1	357.9698	356.2842
pCluster_0_2	-3688.7872	-3690.0355
pCluster_1_0	-50.8572	-48.9445
pCluster_1_1	9732.8868	9738.4038
pCluster_1_2	18386.4202	18387.6458
pCluster_2_0	-3701.9779	-3694.8688
pCluster_2_1	-3701.3308	-3695.4228
pCluster_2_2	-3697.953	-3696.1861
Intercept	3694.0582	3694.5729

```

Odds Ratios...

```

Variable	Class	
	EA	OtrasDem
pCluster_0_0	0	0
pCluster_0_1	2.9126897794559735E155	5.398208601927918E154
pCluster_0_2	0	0
pCluster_1_0	0	0
pCluster_1_1	Infinity	Infinity
pCluster_1_2	Infinity	Infinity
pCluster_2_0	0	0
pCluster_2_1	0	0
pCluster_2_2	0	0

```

=== Stratified cross-validation ===
=== Summary ===

```

Correctly Classified Instances	95	67.3759 %
Incorrectly Classified Instances	46	32.6241 %
Kappa statistic	0.3261	
Mean absolute error	0.2675	
Root mean squared error	0.4042	
Relative absolute error	76.1481 %	
Root relative squared error	96.7126 %	
Total Number of Instances	141	

C.2 Redes Neuronales Artificiales ANN

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.939	0.525	0.713	0.939	0.811	0.722	EA
	0.308	0.112	0.615	0.308	0.41	0.605	OtrasDem
	0.286	0.037	0.286	0.286	0.286	0.655	DV
Weighted Avg.	0.674	0.349	0.656	0.674	0.637	0.676	

```
=== Confusion Matrix ===
```

a	b	c	<-- classified as
77	5	0	a = EA
31	16	5	b = OtrasDem
0	5	2	c = DV

Los parámetros seleccionados para la ejecución del algoritmo RBF, con las variables seleccionadas mediante recuento Borda, fueron los siguientes:

- *B*: *numClusters*. $B = 3$
- *S*: *clusteringSeed*. $S = 1$
- *R*: *ridge*. $R = 1,0E - 8$
- *M*: *maxIts*. $M = -1$
- *W*: *minStdDev*. $W = 0,1$

```
=== Run information ===  
Scheme:weka.classifiers.functions.RBFNetwork -B 3 -S 1 -R 1.0E-8 -M -1 -W 0.1  
Instances: 141  
Attributes: 54  
           [list of attributes omitted]  
Test mode:10-fold cross-validation
```

ApéndiceC : Resultados completos

```

=== Classifier model (full training set) ===

Radial basis function network
(Logistic regression applied to K-means clusters as basis functions):

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...

```

Variable	Class	
	EA	OtrasDem
pCluster_0_0	168.9324	128.199
pCluster_0_1	9.7651	-30.0949
pCluster_0_2	137.5895	97.5591
pCluster_1_0	-40.2935	-76.6934
pCluster_1_1	6.1951	-28.5763
pCluster_1_2	-55.9144	242.8651
pCluster_2_0	-839.1812	-840.5878
pCluster_2_1	-1219.1297	-1168.2391
pCluster_2_2	-1138.3355	-1116.9302
Intercept	388.9731	427.6967

```

Odds Ratios...

```

Variable	Class	
	EA	OtrasDem
pCluster_0_0	2.3248788767353935E73	4.74352925242103E55
pCluster_0_1	17415.2378	0
pCluster_0_2	5.679906738802212E59	2.3408164591147446E42
pCluster_1_0	0	0
pCluster_1_1	490.3178	0
pCluster_1_2	0	2.9852644042652965E105
pCluster_2_0	0	0
pCluster_2_1	0	0
pCluster_2_2	0	0

```

=== Stratified cross-validation ===
=== Summary ===

```

Correctly Classified Instances	104	73.7589 %
Incorrectly Classified Instances	37	26.2411 %
Kappa statistic	0.4735	
Mean absolute error	0.2255	
Root mean squared error	0.3824	
Relative absolute error	64.2011 %	
Root relative squared error	91.4946 %	
Total Number of Instances	141	

```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.939	0.441	0.748	0.939	0.832	0.762	EA
	0.442	0.056	0.821	0.442	0.575	0.684	OtrasDem
	0.571	0.045	0.4	0.571	0.471	0.683	DV
Weighted Avg.	0.738	0.279	0.758	0.738	0.72	0.729	

```

=== Confusion Matrix ===
      a  b  c  <-- classified as
77   3  2  |  a = EA
25  23  4  |  b = OtrasmDem
 1   2  4  |  c = DV

```

C.3. SVM

Los parámetros seleccionados para la ejecución del algoritmo SVM, con las variables seleccionadas por el experto, fueron los siguientes:

- *S*: *SVMType*. El tipo de SVM a usar: 0 = SVC (clasificación), 1 = nu-SVC (clasificación), 2 = one-class SVM (clasificación), 3 = epsilon-SVR (regresión), 4 = nu-SVR (regresión). $S = 0$
- *K*: *kernelType*. Tipo de kernel a utilizar: 0 = lineal, 1 = polinomial, 2 = función de base radial, 3 = sigmoide. $K = 2$
- *D*: *degree*. El grado del kernel. $D = 3$
- *G*: *gamma*. Gamma a utilizar, si es 0 entonces se usa $1/\text{max_index}$. $G = 0,0$
- *R*: *coef0*. Coeficiente a usar. $R = 0,0$
- *N*: *nu*. El valor de ν para nu-SVC, one-class SVM y nu-SVR. $N = 0,5$
- *M*: *cacheSize*. Tamaño de la caché en M. $BM = 40,0$
- *C*: *cost*. El parámetro de coste C de C-SVC, epsilon-SVR y nu-SVR. $C = 1,0$
- *E*: *eps*. Tolerancia del criterio de finalización. $E = 0,001$
- *P*: *loss*. Es el epsilon para la función de pérdida en epsilon-SVR. $P = 0,1$

```

=== Run information ===

Scheme:      weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1
Instances:   141
Attributes:  35
             [list of attributes omitted]
Test mode:   10-fold cross-validation

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

```

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      95          67.3759 %
Incorrectly Classified Instances    46          32.6241 %
Kappa statistic                    0.2848
K&B Relative Info Score            4772.5666 %
K&B Information Score              58.0112 bits    0.4114 bits/instance
Class complexity | order 0         169.8671 bits    1.2047 bits/instance
Class complexity | scheme          49404 bits      350.383 bits/instance
Complexity improvement (Sf)        -49234.1329 bits -349.1782 bits/instance
Mean absolute error                0.2175
Root mean squared error            0.4664
Relative absolute error            61.9181 %
Root relative squared error        111.593 %
Total Number of Instances          141

```

ApéndiceC : Resultados completos

```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.951	0.644	0.672	0.951	0.788	0.654	EA
	0.327	0.09	0.68	0.327	0.442	0.619	OtrasDem
	0	0	0	0	0	0.5	DV
Weighted Avg.	0.674	0.408	0.642	0.674	0.621	0.633	

```

=== Confusion Matrix ===

```

a	b	c	<-- classified as
78	4	0	a = EA
35	17	0	b = OtrasDem
3	4	0	c = DV

Los parámetros seleccionados para la ejecución del algoritmo SVM, con las variables seleccionadas mediante recuento Borda, fueron los siguientes:

- S: *SVMType*. $S = 0$
- K: *kernelType*. $K = 2$
- D: *degree*. $D = 3$
- G: *gamma*. $G = 0,0$
- R: *coef0*. $R = 0,0$
- N: *nu*. $N = 0,5$
- M: *cacheSize*. $M = 40,0$
- C: *cost*. $C = 1,0$
- E: *eps*. $E = 0,001$
- P: *loss*. $P = 0,1$

```

=== Run information ===

```

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1
Instances: 141
Attributes: 54
[list of attributes omitted]
Test mode: 10-fold cross-validation

```

=== Classifier model (full training set) ===

```

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

```

=== Stratified cross-validation ===
=== Summary ===

```

Correctly Classified Instances	103	73.0496 %
Incorrectly Classified Instances	38	26.9504 %
Kappa statistic	0.4188	
K&B Relative Info Score	6147.5242 %	
K&B Information Score	74.724 bits	0.53 bits/instance
Class complexity order 0	169.8671 bits	1.2047 bits/instance
Class complexity scheme	40812 bits	289.4468 bits/instance
Complexity improvement (Sf)	-40642.1329 bits	-288.2421 bits/instance
Mean absolute error	0.1797	
Root mean squared error	0.4239	
Relative absolute error	51.1497 %	
Root relative squared error	101.4261 %	
Total Number of Instances	141	

```

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.976	0.525	0.721	0.976	0.829	0.725	EA
	0.442	0.079	0.767	0.442	0.561	0.682	OtrasDem
	0	0	0	0	0	0.5	DV
Weighted Avg.	0.73	0.335	0.702	0.73	0.689	0.698	

```

=== Confusion Matrix ===
  a  b  c  <-- classified as
80  2  0 | a = EA
29 23  0 | b = OtrásDem
 2  5  0 | c = DV

```

C.4. Clustering

C.4.1. KMeans

Los parámetros seleccionados para la ejecución del algoritmo `SimpleKMeans`, con las variables seleccionadas por el experto, fueron los siguientes:

- *V*: `displayStdDevs`. Parámetro booleano para determinar si se deben mostrar, entre los resultados, las desviaciones estándar de los atributos numéricos y los contadores de los atributos nominales.
- *N*: `numClusters`. Establece el número de clusters. $N = 3$
- *A*: `distanceFunction`. Especifica la función distancia a utilizar para comparar las instancias. $A = \text{EuclideanDistance}$
- *I*: `maxIterations`. Establece el número máximo de iteraciones. $S = 500$
- *S*: `seed`. La semilla a utilizar para los números aleatorios. $S = 10$

```

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -V -N 3
      -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Instances:    141
Attributes:   35
              [list of attributes omitted]
Ignored:
              DxSimplificado2

Test mode:Classes to clusters evaluation on training data
=== Model and evaluation on training set ===

kMeans
=====

Number of iterations: 8
Within cluster sum of squared errors: 466.24032738095235
Missing values globally replaced with mean/mode

```

ApéndiceC : Resultados completos

```
=== Model and evaluation on training set ===

Clustered Instances

0      56 ( 40%)
1      21 ( 15%)
2      64 ( 45%)

Class attribute: DxSimplificado2
Classes to Clusters:

  0  1  2  <-- assigned to cluster
34  2 46 | EA
21 15 16 | OtrasDem
 1  4  2 | DV

Cluster 0 <-- OtrasDem
Cluster 1 <-- DV
Cluster 2 <-- EA

Incorrectly clustered instances : 70.0 49.6454 %
```

Los parámetros seleccionados para la ejecución del algoritmo SimpleKMeans, con las variables seleccionadas mediante recuento Borda, fueron los siguientes:

- *V*: *displayStdDevs*.
- *N*: *numClusters*. $N = 3$
- *A*: *distanceFunction*. $A = \text{EuclideanDistance}$
- *I*: *maxIterations*. $S = 500$
- *S*: *seed*. $S = 10$

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -V -N 3
      -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Instances: 141
Attributes: 54
           [list of attributes omitted]
Ignored:
           DxSimplificado2

Test mode:Classes to clusters evaluation on training data
=== Model and evaluation on training set ===

kMeans
=====

Number of iterations: 10
Within cluster sum of squared errors: 691.589900110988
Missing values globally replaced with mean/mode
```

```

=== Model and evaluation on training set ===

Clustered Instances

0      20 ( 14%)
1      53 ( 38%)
2      68 ( 48%)

Class attribute: DxSimplificado2
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  2 41 39 | EA
14 12 26 | OtrasDem
  4  0  3 | DV

Cluster 0 <-- DV
Cluster 1 <-- EA
Cluster 2 <-- OtrasDem

Incorrectly clustered instances : 70.0 49.6454 %

```

C.4.2. EM

Los parámetros seleccionados para la ejecución del algoritmo EM, con las variables seleccionadas por el experto, fueron los siguientes:

- I: *maxIterations*. Número máximo de iteraciones. $I = 100$
- N: *numClusters*. Establece el número de clusters. Con el valor -1 el número de clusters se selecciona automáticamente por validación cruzada. $N = 3$
- M: *minStdDev*. Establece la desviación estándar mínima permitida. $M = 1,0E - 6$
- S: *seed*. La semilla a utilizar para los números aleatorios. $S = 100$

```

=== Run information ===

Scheme:      weka.clusterers.EM -I 100 -N 3 -M 1.0E-6 -S 100
Instances:   141
Attributes:  35
             [list of attributes omitted]

Ignored:
             DxSimplificado2

Test mode:   Classes to clusters evaluation on training data
=== Model and evaluation on training set ===

EM
==
Number of clusters: 3

```

ApéndiceC : Resultados completos

```
Clustered Instances
0      37 ( 26%)
1      30 ( 21%)
2      74 ( 52%)

Log likelihood: -1.48487

Class attribute: DxSimplificado2
Classes to Clusters:

  0  1  2 <-- assigned to cluster
22  4 56 | EA
15 20 17 | OtrasDem
  0  6  1 | DV

Cluster 0 <-- OtrasDem
Cluster 1 <-- DV
Cluster 2 <-- EA

Incorrectly clustered instances : 64.0  45.3901 %
```

Los parámetros seleccionados para la ejecución del algoritmo EM, con las variables seleccionadas mediante recuento Borda, fueron los siguientes:

- I: *maxIterations*. $I = 100$
- N: *numClusters*. $N = 3$
- M: *minStdDev*. $M = 1,0E - 6$
- S: *seed*. $S = 100$

```
=== Run information ===

Scheme:      weka.clusterers.EM -I 100 -N 3 -M 1.0E-6 -S 100
Instances:   141
Attributes:  54
              [list of attributes omitted]
Ignored:     DxSimplificado2
Ignored:     DxSimplificado2
Test mode:   Classes to clusters evaluation on training data
=== Model and evaluation on training set ===

EM
==
Number of clusters: 3
```

```

Clustered Instances
0      78 ( 55%)
1      16 ( 11%)
2      47 ( 33%)

Log likelihood: 6.72926

Class attribute: DxSimplificado2
Classes to Clusters:

  0  1  2  <-- assigned to cluster
43  1 38 | EA
34  9  9 | OtrasDem
 1  6  0 | DV

Cluster 0 <-- OtrasDem
Cluster 1 <-- DV
Cluster 2 <-- EA

Incorrectly clustered instances : 63.0 44.6809 %

```

C.4.3. Hierarchical Clusterer

Los parámetros seleccionados para la ejecución del algoritmo `Hierarchical Clusterer`, con las variables seleccionadas por el experto, fueron los siguientes:

- *N*: `numClusters`. Número de clusters. $N = 3$
- *L*: `linkType`. Método para medir la distancia entre dos clusters. Con `SINGLE` mide la distancia mínima entre dos items cualesquiera, uno de cada cluster. $L = \text{SINGLE}$
- *P*: `printNewick`. Valor booleano que indica si el cluster debe ser impreso en formato Newick. P
- *A*: `distanceFunction`. Establece la función para calcular la distancia entre dos instancias individuales. $A = \text{EuclideanDistance}$
- *R*: `attributeIndices`. Especifica el rango de atributos sobre los que actuar. Es una lista de índices de atributos separados por comas. Se pueden utilizar comodines como `first` o `last`. $R = \text{first} - \text{last}$

```

=== Run information ===
Scheme:weka.clusterers.HierarchicalClusterer -N 3 -L SINGLE -P -D -A "weka.core.EuclideanDistance -R first-last"
Instances: 141
Attributes: 35
           [list of attributes omitted]
Ignored:
           DxSimplificado2
Test mode:Classes to clusters evaluation on training data

```

ApéndiceC : Resultados completos

```
=== Model and evaluation on training set ===

Clustered Instances
0      138 ( 98%)
1         2 (  1%)
2         1 (  1%)

Class attribute: DxSimplificado2
Classes to Clusters:

  0  1  2 <-- assigned to cluster
82  0  0 | EA
51  0  1 | OtrasDem
 5  2  0 | DV

Cluster 0 <-- EA
Cluster 1 <-- DV
Cluster 2 <-- OtrasDem

Incorrectly clustered instances : 56.0  39.7163 %
```

Los parámetros seleccionados para la ejecución del algoritmo `Hierarchical Clusterer`, con las variables seleccionadas mediante recuento Borda, fueron los siguientes:

- `N`: *numClusters*. $N = 3$
- `L`: *linkType*. $L = \text{SINGLE}$
- `P`: *printNewick*. P
- `A`: *distanceFunction*. $A = \text{EuclideanDistance}$
- `R`: *attributeIndices*. $R = \text{first-last}$

```
=== Run information ===

Scheme:weka.clusterers.HierarchicalClusterer -N 3 -L SINGLE -P -A "weka.core.EuclideanDistance" -R first-last
Instances: 141
Attributes: 54
           [list of attributes omitted]
Ignored:
           DxSimplificado2
Test mode:Classes to clusters evaluation on training data
```

```

=== Model and evaluation on training set ===

Clustered Instances
0      135 ( 96%)
1       5 (  4%)
2       1 (  1%)

Class attribute: DxSimplificado2
Classes to Clusters:

  0  1  2 <-- assigned to cluster
82  0  0 | EA
50  2  0 | OtrasDem
 3  3  1 | DV

Cluster 0 <-- EA
Cluster 1 <-- OtrasDem
Cluster 2 <-- DV

Incorrectly clustered instances : 56.0 39.7163 %

```

C.4.4. ROCK

Los parámetros seleccionados para la ejecución del algoritmo ROCK, con las variables seleccionadas por el experto, fueron los siguientes:

- *n*. Número de clusters deseados. $n = 3$
- *beta*. Umbral de distancia. Opcional porque $\beta = 1 - \theta$. $beta = 0,65$
- *theta*. Parámetro de vecindad en el rango $[0, 1)$. $theta = 0,35$
- *fun*. Función de distancia a utilizar. $fun = dist$
- *args*. Una lista de argumentos para los parámetros. $args = list(method = "binary")$

```

rockMerge: terminated with 3 clusters
data: x
beta: 0.65
theta: 0.35
fun: dist
args: list(method = "binary")
 1  2  3
133 7 1
dropping 1 clusters
computing distances ...
computing classes ...
$cl
 [1] 1  1  1  1  1  1  1  1  1  2  2  1  1  1  1  2  1  1  1  1  1
 [21] 1  1  1  1  1  1  1  1  1  1  1  1  1  2  1  1  1  1  1  1  1
 [41] 1  1  1  1  1  1  1  1  1  1  1  1  1  2  1  1  1  1  1  1  1
 [61] 1  1  1  1  2  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 [81] 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  1  1  1  1
[101] 1  1  1  1  1  1  2  1  1  <NA> 1  1  1  1  1  1  1  2  1  1  1
[121] 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
[141] 1
Levels: 1 2 <NA>

$size
 1  2 <NA>
131 9 1

```

Los parámetros seleccionados para la ejecución del algoritmo ROCK, con las variables seleccionadas mediante el recuento Borda, fueron los siguientes:

- *n*. $n = 3$

ApéndiceC : Resultados completos

- *beta. beta* = 0,5
- *theta. theta* = 0,5
- *fun. fun* = *dist*
- *args. args* = `list(method = "binary")`

```
rockMerge: terminated with 13 clusters
data: x
beta: 0.5
theta: 0.5
fun: dist
args: list(method = "binary")
  1  2  3  4  5  6  7  8  9 10 11 12 13
125 4  1  2  1  1  1  1  1  1  1  1  1
dropping 10 clusters
computing distances ...
computing classes ...
$cl
 [1] 1  1  1  1  1  1  1  1  2  2  1  1  <NA> 1  2  1  1  1  1  1  1
 [21] 1  1  1  1  1  1  1  1  1  1  1  1  1  2  1  1  1  1  1  1
 [41] 1  1  1  1  1  1  1  1  4  1  1  1  1  1  1  1  1  1  1  <NA> 1
 [61] 1  1  1  1  1  <NA> 1  <NA> 1  1  1  1  1  1  1  1  1  1  1  1
 [81] 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  <NA> <NA> 1  1
[101] 4  1  1  1  1  1  1  <NA> 1  1  <NA> 1  1  1  1  1  1  1  1  1
[121] 1  1  <NA> 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  4  1  1
[141] 1
Levels: 1 2 4 <NA>
$size
  1  2  4 <NA>
125 4  3  9
```